Title

Development of Systems Biology Strategies for Environmental Exposures in Health and Disease

Author

CHEN, YEN-WEI

Publication Date

2022

UNIVERSITY OF CALIFORNIA

Los Angeles

Development of Systems Biology Strategies for Environmental Exposures in Health and Disease

A dissertation submitted in partial satisfaction of the requirements for the degree in Doctor of Philosophy in Molecular Toxicology

by

YEN-WEI CHEN

2022

ABSTRACT OF THE DISSERTATION

Development of Systems Biology Strategies for

Environmental Exposures in Health and Disease

by

YEN-WEI CHEN

Doctor of Philosophy in Molecular Toxicology

University of California, Los Angeles, 2022

Professor Xia Yang, Co-Chair

Professor Patrick Allard, Co-Chair

Environmental exposures such as drug administration, chemical contamination, and unhealthy diets can lead to toxicity or adverse effects that impose significant health and economical burdens. Addressing the mechanisms of these effects has become a critical topic. However, the majority of the current mechanistic research efforts has focused on a narrow molecular space, for example, on changes in the transcriptome in a select tissue and species, without taking into account other types of molecular, cellular, tissue, and species information. I hypothesize that development of new tools and strategies to integrate multispecies multi-tissue multicellular information will uncover new insights on environmental exposure. Hence my research aims to develop tools and apply computational analyses to understand molecular networks of different

exposures and identify potential therapies. First, I built a tissue- and species-specific drug gene signature database for >900 drugs across tens of thousands of transcriptome datasets across human, mouse and rat models, and implemented a network-based repositioning tool to link drug signatures with different organ toxicities and disease therapeutics. This system was applied to validated for hyperlipidemia, non-alcoholic fatty liver disease and hepatoxicity. Next, I investigated multi- and trans-generational (F1 and F3) effect stemming from ethanol exposure on *C. elegans* model through single nucleus RNA-seq. I established the first comprehensive whole-organism transcriptional map of an environmental response at cell-type specific resolution. Results indicated strong alterations in metabolism, lipid transportation pathways as well as abnormal germline phenotypes among germline clusters. Finally, I investigated molecular effects of how a Western diet (high fat high sucrose) and a fructose rich diet affect metabolic regulation through single cell RNA-seq analysis of diverse cell subpopulations across different tissues (liver, adipose, hypothalamus and small intestine) in a mouse model. We identified susceptible tissues, cell types, biological pathways, and ligands mediating metabolic syndrome (*Avp*, *Apoe*, *Oxt*). These various projects involving different model systems, tissues, and cell types provided new analytical tools and revealed systems level insights on diverse types of environmental exposures.

The dissertation of YEN-WEI CHEN is approved.

Oliver Hankinson

Van Savage

Patrick Allard, Committee Co-Chair

Xia Yang, Committee Co-Chair

University of California, Los Angeles

2022

DEDICATION

This dissertation is dedicated to HSENG-CHANG, LING-LING, YEN-YU

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of Supplementary Tables

(Complete supplementary tables in separate file)

LIST OF ABBREVIATIONS

ADR	adverse drug reactions
DILI	drug-induced liver injury
CTD	Comparative Toxicogenomics Database
KEGG	Kyoto Encyclopedia of Genes and Genomes
DEG	differential expressed genes
FDR	false discovery rate
NASH	non-alcoholic steatohepatitis
NAFLD	non-alcoholic fatty liver disease
LDL	low-density lipoprotein cholesterol
TG	triglycerides
TC	total cholesterol
UC	unesterified cholesterol
GWAS	genome-wide association study
BN	Bayesian gene regulatory network
AUROC	Area under the curve of receiver operating characteristic
HMGCR	β-Hydroxy β-methylglutaryl-CoA receptor
PPAR	peroxisome proliferator-activated receptor
GPCR	G-protein coupled receptor
NMR	Nuclear Magnetic Resonance
HFHS	high fat high sucrose
FDA	Food and Drug Administration
snRNA-seq	single nuclei RNA-seq

ACKNOWLEDGEMENTS

First, I would like to express my deepest appreciation to my advisors Xia and Patrick. Co-mentorship made my PhD journey challenging but also exciting. I am extremely grateful for the guidance from both of you to push the boundaries in my research capabilities, scientific thinking, and career development. Xia, I also wish to thank you for your patience and guidance in supporting my development of analysis techniques. Patrick, I am grateful for your help in exploring different frontiers in toxicology. I am incredibly thankful for the time you both spent to ensure that I succeed, despite your busy schedules. I will forever adore the years I have spent in your labs and want you to know it that I could not have gotten to this point without you. COVID has changed lots of things, but it will never change your king mentorship and your being as great mentors.

Mom and Dad, thank you for always being supportive of me no matter what decision I make. I understand studying abroad is never an easy path. Your support is a beacon that encourages me to keep going, even during the darkest period. I love you both and could not have achieved any of this without you.

Chapter 2 is a version of **Y Chen**, Graciel D, Jessica D, Thien N, Jessica Y, Sung-min H, Peter C, Douglas A, Montgomery B, Jennifer G, Nima Z, Paul P, and Xia Y. PharmOmics: A Species- and Tissue-specific Drug Signature Database and Gene Network-based Drug Repositioning Tool. *iScience*, 2022;25(4):104052. YC was supported by UCLA Eureka fellowship and Burroughs Wellcome Fund Inter-school Training Program in Chronic Diseases. DA was supported by NIH-NCI National Cancer Institute (T32CA201160), UCLA dissertation year fellowship and UCLA Hyde fellowship. XY was funded by NIH DK104363 and DK117850. GD was supported by the National Institute of Environmental Health Sciences (T32ES015457) and

the American Diabetes Association Postdoctoral Fellowship (1-19-PDF-007-R). YC curated and analyzed data, constructed database, and designed and conducted application studies. GD, JY, and PC conducted validation experiments. JD, TXN, DH, and MB designed and implemented the PharmOmics web server. DA provided support in data curation and analysis. GA, JG, NZ, and PP assisted with data curation. YC, GD, JD, and XY wrote the manuscript. XY designed and supervised the research. All authors contributed to manuscript editing.

Chapter 3 is a version of L Truong, Y Chen, E Beltrame, R Barrere-Cain, B Panter, X Yang, P Steinberg, P Allard. Mapping of inter- and trans-generational Single-nucleus resolution mapping of the adult *C. elegans* and its application to elucidate intergenerational response to alcohol exposure. Submitted to PNAS. LT is supported by the NIH Training Grant in Genomic Analysis and Interpretation T32 HG002536; YWC is supported by the UCLA Eureka fellowship and Burroughs Wellcome Fund Inter-school Training Program in Chronic Diseases; PA is supported by NIEHS R01 ES027487, NIAAA R21 AA024889, and the Burroughs Wellcome Innovation in Regulatory Science Award. LT, RBC, and BP performed all biological experiments, YWC performed all bioinformatic analyses, XY and PA supervised all experiments and analyses. LT, RBC, YWC, XY, and PA wrote the manuscript.

Chapter 4 is a version of Y Chen, S Majid, I Ahn, G Diamante, I Cely, G Zhang, S Komzyuk, Bonnett, S Wang, D Arneson and X Yang. Multitissue Single-Cell Analysis Reveals Differential Tissue, Cellular, and Molecular Sensitivity Between Fructose and High Fat High Sucrose Diets. YWC is supported by the UCLA Eureka fellowship and Burroughs Wellcome Fund Inter-school Training Program in Chronic Diseases; GD was supported by the National Institute of Environmental Health Sciences (T32ES015457) and the American Diabetes Association Postdoctoral Fellowship (1-19-PDF-007-R); DA was supported by NIH-NCI National Cancer

Institute (T32CA201160), UCLA dissertation year fellowship and UCLA Hyde fellowship; XY is funded by NIH R01DK14363. YC conducted data analysis and visualization, SM, IA, GD, IC, GZ conducted experiment; SK, JC supported data analysis and visualization; DA supported data analysis; XY supervised all experiments and analysis. YC, IA, GZ, SK, SW and DA wrote manuscript.

Appendix consists of the following manuscript.

Appendix A is a version of J Ding, M Blencowe, T Nghiem, S Ha, Y Chen, G Li, X Yang. Mergeomics 2.0: a web server for multi-omics data integration to elucidate disease networks and predict therapeutics. Nucleic Acids Research, Volume 49, Issue W1, 2 July 2021, Pages W375–W387. X.Y. is supported by NIH R01 [NS117148, NS111378, DK117850, HL145708, HL147883, HD100298]; Montgomery Blencowe is supported by the American Heart Association Predoctoral Fellowship; Sung-min Ha is supported by the UCLA QCBio Collaboratory Postdoc Fellowship. Funding for open access charge: National Institutes of Health [DK117850, HD100298, HL145708, HL147883, NS111378, NS117148]. Y. Chen supported PharmOmics pipeline for this study.

Appendix B is a version of J Li, L Pan, W G. Pembroke, J E. Rexach, M I. Godoy, M C. Condro, A G. Alvarado, M Harteni, Y Chen, L Stiles, A Y. Chen, I B. W, X Yang, S A. Goldman, D H. Geschwind, H I. Kornblum & Y Zhang. Conservation and divergence of vulnerability and responses to stressors between human and mouse astrocytes. Nature Communications volume 12, Article number: 3958 (2021). This work is supported by the Achievement Rewards for College Scientists foundation Los Angeles Founder Chapter and the National Institute of Mental Health of the National Institutes of Health (NIH) Award T32MH073526 to M.I.G., the Dr. Sheldon and Miriam G. Adelson Medical Research Foundation to S.A.G., H.I.K., and D.H.G., the

National Institute of Neurological Disorders and Stroke of the NIH R00NS089780, R01NS109025, the National Institute of Aging of the NIH R03AG065772, National Center for Advancing Translational Science UCLA CTSI Grant UL1TR001881, UCLA Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research Innovation Award, the W.M. Keck Foundation Junior Faculty Award, the UCLA Jonsson Comprehensive Cancer Center and Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research Ablon Scholars Program, and the Friends of the Semel Institute for Neuroscience & Human Behavior Friends Scholar Award to Y.Z. J.L. and Y.Z. conceived of the project and designed the experiments. J.L. performed all experiments except those noted below. L.P. performed xenograft experiments and RNA-seq of xenografted astrocytes. M.I.G. contributed to the generation of RNA-seq libraries. M.C.C., A.G.A., and M.H. optimized xenografting conditions and assisted the xenograft experiments under the supervision of H.I.K. W.G.P., J.E.R., and D.H.G. performed WGCNA and analyzed some of the single-cell sequencing datasets. Y-W.C. and X.Y. performed mapping of xenografted RNA-seq reads to human and mouse genomes. L.S. performed Seahorse Respirometry and TMRE/MTG imaging experiments. A.Y.C. and I.B.W. procured tissue samples. S.A.G. developed the xenograft method and provided training for xenograft experiments. J.L. and Y.Z. analyzed the data and wrote the paper. All authors read the manuscript.

Appendix C is a version of E E. Noble, C A. Olson, E Davis, L Tsan, Y Chen, R Schade, C Liu, A Suarez, R B. Jones, C de La Serre, X Yang, E Y. Hsiao & S E. Kanoski. Gut microbial taxa elevated by dietary sugar disrupt memory function. Translational Psychiatry volume 11, Article number: 194 (2021). The research was supported by DK116942 and DK104897, and institutional funds to S.E.K., DK118000 and DK111158 to E.E.N., DK116558 to A.N.S., D.K. 118944 to C.M.L. C.A.O. was supported by an F31 AG064844. E.Y.H. was supported by the

ARO MURI award W911NF-17-1-0402. DK104363 to X.Y., Eureka Scholarship and BWF-CHIP Fellowship to Y.C. Y. C. conducted RNA-seq analysis and manuscript writing.

Appendix D is a version of M Blencowe; D Arneson; J Ding; Y Chen; Z Saleem; X Yang. Network modeling of single-cell omics data: challenges, opportunities, and progresses. Emerg Top Life Sci (2019) 3 (4): 379–398. D.A. is funded by the UCLA Dissertation Year Fellowship. X.Y. is funded by the National Institutes of Health Grants DK104363 and NS103088. Y.-W.C. is supported by the UCLA Hyde Fellowship. M.B., D.A., J.D., Y.-W.C., Z.S., and X.Y. drafted and edited the manuscript.

Appendix E is a version of L Shu, Q Meng, G Diamante, B Tsai, Y Chen, A Mikhail, H Luk, B Ritz, P Allard, X Yang. Prenatal Bisphenol A Exposure in Mice Induces Multitissue Multiomics Disruptions Linking to Cardiometabolic Disorders. Endocrinology, Volume 160, Issue 2, February 2019, Pages 409–429. L.S. is supported by a University of California, Los Angeles, Dissertation Year Fellowship, Eureka Scholarship, Hyde Scholarship, Burroughs Wellcome Fund Inter-School Program in Metabolic Diseases Fellowship, and the China Scholarship Council. G.D. is supported by NIEHS/National Institutes of Health Grant T32ES015457. P.A. is supported by National Institutes of Health/NIEHS Grant R01-ES02748701 and the Burroughs Wellcome Foundation. X.Y. is supported by National Institutes of Health Grant DK104363 and the Leducq Foundation. Y. C. supported analysis of RNA-seq data.

Appendix F is a version of J Camacho, L Truong, Z Kurt, Y Chen, M Morselli, G Gutierrez, M Pellegrini, X Yang and P Allard. The Memory of Environmental Chemical Exposure in *C. elegans* Is Dependent on the Jumonji Demethylases jmjd-2 and jmjd-3/utx-1. May 2018Cell Reports 23(8). P.A. is supported by NIH/NIEHS R01 ES02748701 and the Burroughs Wellcome

Foundation . J.C. received support from NIH/NIEHS T32 ES015457 Training in Molecular Toxicology, the North American Graduate Fellowship , the NSF AGEP Competitive Edge , the NSF Graduate Research Fellowship , and the Eugene-Cota Robles Fellowship . L.T. is supported by the NIH Training Grant in Genomic Analysis and Interpretation T32 HG002536 . G.G. is supported by NIH/NIEHS R25 ES02550703 . Z.K. was supported by an American Heart Association post-doctoral fellowship (17POST33670739) and the Iris Cantor-UCLA Executive Advisory Board/CTSI Pilot Award . M.M. was supported by a Dissertation Year Fellowship (University of California, Los Angeles) . X.Y. is supported by NIH/NIDDK R01 DK104363 and NIH/NINDS R21 NS103088. J.C., L.T., M.M., and G.G. performed the experiments. J.C., L.T., Z.K., Y.-W.C., M.P., X.Y., and P.A. analyzed and interpreted the results. J.C., L.T., Z.K., Y.-W.C., X.Y., and P.A. wrote the manuscript.

VITA

EDUCATION

2022	PhD Candidate, Molecular Toxicology University of California, Los Angeles
2017	Master of Science, Toxicology National Taiwan University
2015	Doctor of Pharmacy National Taiwan University

RESEARCH EXPERIENCE

2017 – Present	PhD Candidate Department of Integrative Biology and Physiology University of California, Los Angeles

Advisor: Dr. Xia Yang
Developed a PharmOmics, a species- and tissue-specific drug signature database and gene network-based drug repositioning Tool (http://mergeomics.research.idre.ucla.edu/)

- Developed a first-in-place transcriptome framework for drug prediction by establishing gene network-based drug repositioning concept and implemented web interface to allow queries for disease gene(s).
- Established pipeline to retrieve chemical signature database from public data and constructed framework in gene regulatory network-based repositioning.
- Validated gene network-based drug repositioning concept in noise robustness benchmarking, finding known drugs from hyperlipidemia, hepatitis and hyperuricemia and finding new therapeutics in non-alcoholic fatty liver disease.
- Clustered drug induced liver injury (DILI) genes based on gene regulatory network mapping to improve understanding of DILI biological mechanisms and DILI drug prediction.

Developed analysis strategies to Investigate metabolic syndrome (MetS) through multi-tissue single cell RNA-seq.

- Located key factors (biological pathways, circulating ligands, metabolites, susceptible cell types) and constructed MetS cell type response and cross-talk maps through single cell RNA-seq data analysis.

Co-advisor: Dr. Patrick Allard
Investigated trans-generational chemical exposure mechanism on *C. elegans* model through sequencing strategies

- Characterized of BPA's trans-generational chromosomal-wide epigenomic effect through analysis of histone modification Chip-seq (H3K9me3 and H3K27me3), followed by validation with transcriptome and experimental data.
- Identified germline specific ribosomal dysfunction and lipid metabolism abnormality mediating trans-generational birth dysfunction from alcohol exposure through single nucleus RNA-seq and highly sensitive pipeline.

2021 Summer graduate Internship
Amgen Translational Safety and Biological Activity department
South San Francisco, California

- Integrated human genetic, pharmacology, and various drug clinical trial knowledgebase to guide rationale selection for off-target proteins of drug safety relevance
- Established and optimized innovative pipeline using weight of evidence (expression data, species conservation, literature scores) to systematically characterize over 5,000 off-target proteins aid future hit triage and follow up
- Performed cell line selection from proteomics and transcriptomics databases to guide downstream experimentation
- Contributed to non-clinical safety strategy for new emerging modalities such as siRNA and PROTAC using systems biology approaches

2015-2017 Master student
National Taiwan University, Graduate institute of Toxicology
Taipei, Taiwan

Advisor: Dr. Hsing-Chen Tsai
Investigated mechanisms of lung cancer resistance and relapse to decitabine, an epigenetic targeting drug

- Pinpointed decitabine resistant lung cancer subtype and investigated mechanism, including hypomethylation and apoptosis response through bioinformatics analyses of DNA methylation arrays and gene expression microarray.
- Identified blood methylation biomarkers for early diagnosis of lung cancer based on blood circulating DNA
- Determined combination of DNA methylation markers for lung cancer diagnosis based on TCGA datasets and validated with clinical samples.

PEER-REVIEWED FIRST AUTHOR PUBLICATIONS

Yen-Wei Chen, Douglas Arneson, Graciel Diamente, Jennifer Garcia, Nima Zaghari, Paul Patel, Patrick Allard, and Xia Yang. PharmOmics: A Species- and Tissue-specific Drug Signature Database for Toxicity Prediction and Drug Repurposing. *iScience*, 2022;25(4):104052.

SELECTED FIRST AUTHOR MANUSCRIPTS IN Preparation (*, co-first author)

Yen-Wei Chen, Sana Majid, In Sook Ahn, Graciel Diamante, Ingrid Cely, Guanglin Zhang, Sergey Komzyuk, Jack Bonnett, Douglas Arneson, Xia Yang. Multitissue Single-Cell Analysis Reveals Differential Tissue, Cellular, and Molecular Sensitivity between Fructose and High fat high sucrose Diets. In preparation

Lisa Truong*, **Yen-Wei Chen***, Eduardo Beltrame, Rio Barrere-Cain, Blake Panter, Xia Yang, Paul Steinberg, Patrick Allard. Mapping of inter- and trans-generational Single-nucleus resolution mapping of the adult *C. elegans* and its application to elucidate intergenerational response to alcohol exposure. In preparation

Chapter 1 Introduction

Exposure to various environmental agents, such as therapeutic drugs, industrial chemicals, alcohol and diets, has been associated with different health risks. For example, drug toxicity or adverse drug reaction (ADR) sometimes leads to severe clinical manifestations such as allergies, hepatotoxicity, nephrotoxicity, and even cardiotoxicity. ADRs cause the withdrawal of numerous marketed drugs and a large proportion of failures of new drugs in clinical trials. Management of drug toxicity may cost up to 30.1 billion dollars annually in US [1]. In addition to risks related to drug treatment, diseases related to unhealthy diet (such as diet rich in fat and sugar) are also on the rise and cause significant costs to the society which is linked to $50 billion cost [2–4]. Finally, alcohol ingestion and related diseases and social burdens is also tremendous in the US with economic burden estimated up to 249 billion [5,6]. Understanding the molecular mechanisms underlying these diverse environmental exposures will offer necessary information to support risk assessment and guide the development of preventative and therapeutic strategies.

In order to conduct proper risk assessment, careful selection of species and model systems is required. For example, rat is commonly used in toxicological research as well as large toxicogenomics database such as TG-GATEs [7] and drugMatrix (https://ntp.niehs.nih.gov/DrugMatrix/index.html) due to diversified genetic backgrounds that mimic human populations [8]. Compared to rat models, the mouse is more commonly used in disease modeling [9] as well as studies examining gene by environment interactions [10] due to well-controlled genetics in inbred or recombinant strains. Furthermore, invertebrate models such as *C. elegans* also have unique advantages toxicology studies due to its tractability, short life cycle, highly evolutionary conserved genes, and easily accessible cell types including germ cells [11]. Though each of these different species and model systems carries unique advantages,

species differences in genetics, tissue functions, and metabolic and signaling pathways lead to species specific responses to environmental exposures, which complicate translational interpretation in exposure characterizations and human relevance in health and disease [12].

In addition to species related complexity, tissue and cell type specific responses also cause challenges in chemical characterization. For instance, fructose exposure acts on the brain to change food intake behaviors while inducing fat accumulation in liver in rodents [13]. Each tissue also contains numerous diverse types of cell types. For example, liver consists of two large groups of cells, hepatocytes and non-parenchymal cells. Hepatocytes also function differently according to location and nutrient gradient: hepatocytes that are located nearby the perinodal region have a higher oxygen gradient and conduct more beta-oxidation, whereas hepatocytes located in the central vein region have higher activity in lipogenesis, triglyceride synthesis and glycolysis [14]. Liver non-parenchymal cells such as Kupffer cells, sinusoidal endothelial cells and hepatocyte satellite cells function together to support hepatocytes [15]. Different cell types not only respond differently to environmental exposures but also show different crosstalk patterns with each other through ligand receptor interactions.

Last but not least, tissue/cell specific molecular alterations further increase complexity during risk assessment. Tissue specific gene expression and regulation has been documented through Genotype-Tissue Expression (GTEx) project [16] and affected mechanism characterizations in disease and exposure. For example, it has been known that mutations in *LPL*, *CETP, APOA and APOB* were found in patients with metabolic syndrome [17]. However, these mutations showed tissue specificity, with *LPL*, *CETP* highly expressed in adipose tissue while *APOA* and *APOB* are specifically expressed in liver. Furthermore, tissue specific gene expression is also

linked with adverse drug reaction [18]. Identification of tissue specific molecular mechanism is also critical in environmental exposure characterization.

The complexities from diversity of species, tissues, cell types and molecular interactions have hindered thorough investigation and understanding of exposure characterization. In recent years, systems biology concepts and methodologies have been developed to enable better characterization of disease and exposures. In particular, network biology [19] is a central concept in systems biology to provide a fundamental framework to model molecular interactions with species, tissue- and cell type information. Network modeling can facilitate better mechanism characterization [20–22] and support identification of novel therapies [23,24]. In addition to applying network concept at molecular layers such as genes, proteins [25] or metabolites [26], network concept can also be applied on cell type and tissue layer [27] with different genes, proteins and metabolites serving as cross-talking molecules between cells and tissues. Recent advances in single cell RNA sequencing (scRNA-seq) [28] [29] further enable high throughput transcriptome profiling at single cell resolution to capture complex responses to environmental exposure across different cell types, tissues and species. Despite the availability of advanced techniques and novel algorithms, applications of systems biological concepts and approaches to environmental exposure studies to derive molecular insights are still limited. Current toxicology and risk assessment tools are heavily focused on relationships between chemical structures, in vitro assay-based toxicodynamics and toxicokinetics [30–33]. Providing systems biology based insights will further support hazard identification and risk assessment in toxicology.

In order to address the aforementioned gaps, my dissertation work leverages multi-species, multi-tissue, multi-cellular systems biology concepts and methodologies to understand environmental exposures including pharmaceutical drugs, alcohol, and diets. Specifically, my

research is focused on 1) development of “PharmOmics”, a tissue- and species-specific drug gene signature database for >900 drugs across tens of thousands of transcriptome datasets and integrated with network-based repositioning tool to link drug signatures with different organ toxicities and disease therapeutics; 2) investigating multi- and trans-generational (F1 and F3) effect stemming from ethanol exposure on *C. elegans* model through single nucleus RNA-seq through establishment of the first comprehensive whole-organism transcriptional map of an environmental response at cell-type specific resolution; 3) Investigating molecular effects of how a Western diet (high fat high sucrose) and a fructose rich diet affect metabolic regulation through single cell RNA-seq analysis of diverse cell subpopulations across different tissues (liver, adipose, hypothalamus and small intestine) in a mouse model.

The methodological details of curation, algorithm implementation and performance evaluation of PharmOmics is described in **Chapter 2**. PharmOmics aims to retrieve tissue and species-specific insights of drug mechanism through integration of drug signatures and tissue specific gene regulatory network, which distinguished itself against other tools which either overlooked the network perspective or only focused on non-tissue specific protein interaction networks. I have first implemented semi-automatic pipeline which retrieved tissue- and species-specific drug gene signature database for >900 drugs across tens of thousands of transcriptome datasets across human, mouse and rat models. Followed by this, I implemented network-based drug repositioning algorithm to link drug signatures with different organ toxicities and disease therapeutics. Through in silico validation with diseases of known drug treatment (hyperlipidemia, hyperuricemia, hepatitis) and in vivo validation with diseases without drug treatment (non-alcoholic fatty liver disease), I have demonstrated that PharmOmics is able to identify disease treatment drugs and mechanistic insights. Moreover, this platform is also available from our MergeOmics webserver [34] which provides easy-to-access system for the scientific community.

In **Chapter 3**, I investigated multi- and trans-generational (F1 and F3) effects stemming from ethanol exposure on *C. elegans* model through single nucleus RNA-seq, which demonstrated the unique advantage of using *C. elegans* model to both investigating multi-generation effects efficiently as well as capturing exposure effects among cell types from whole organism level from single nucleus RNA-seq. Through implementation of tools designed for single nucleus datasets and highly sensitive differentially expressed genes detection algorithm, I established the first comprehensive whole-organism transcriptional map of an environmental response at cell-type specific resolution. Results indicated strong alterations in metabolism, lipid transportation pathways as well as abnormal germline phenotypes among germline clusters, which indicated long-lasting effect of ethanol exposure in reproductive system through non-genetic mechanisms.

In **Chapter 4**, I investigated molecular effects of how a Western diet (high fat high sucrose) and a fructose rich diet affect metabolic regulation through single cell RNA-seq analysis of diverse cell subpopulations across different tissues (liver, adipose, hypothalamus and small intestine) in a mouse model. In order to retrieve cell type interactions from different perspectives, I have conducted analysis involving traditional differential expressed genes/pathways, ligand-receptor interaction modeling and metabolite flux analysis. Through interactions via circulating ligands and metabolites, I have constructed cell type interaction mappings which helped identification of susceptible tissues, cell types, biological pathways, and ligands mediating metabolic syndrome (Avp, Apoe, Oxt). The multi-tissue systems biology investigation provided thorough comparison between two MetS inducing diets not only from cell type response itself but also from cross-talks across cell types which supported future development of biomarkers and therapeutics.

In summary, by developing a bioinformatics platform and applying systems biology to derive mechanistic insights, my studies demonstrate how systems biology strategies can be applied in different environmental exposures in order to reveal systems level insights. The platforms and molecular insights obtained can serve as foundations for future environmental exposure characterization and support future drug development as well as chemical risk assessment development.

Chapter 2 PharmOmics: A Species- and Tissue-specific Drug Signature Database and Gene Network-based Drug Repositioning Tool

Introduction

Drug development has been challenging and costly over the past decades due to the high failure rate in clinical trials [35]. Most drugs with excellent efficacy and safety profiles in preclinical studies often encounter suboptimal efficacy or safety concerns in humans [36]. This translational gap is likely attributable to our incomplete understanding of the systems level activities of drugs in individual tissues and organ systems [37] as well as the differences between humans and preclinical model systems [38].

Drug activities can be captured by gene expression patterns, commonly referred to as gene signatures. By measuring how a pharmacological agent affects the gene signature of a tissue in a particular species, we can infer the tissue-specific biological pathways involved in therapeutic processes or toxicological responses. This concept has prompted drug repositioning studies to repurpose approved drugs for new disease indications [20,23,39–42]. Similarly, gene signatures can reveal mechanisms underlying adverse drug reactions (ADRs) and be leveraged to predict ADRs as previously shown for liver and kidney toxicity [43–45].

A drug may affect different molecular processes between tissues, providing treatment effects in the desired target tissue(s) but causing toxicity or ADRs in other tissues, which can be captured in tissue-specific drug signatures. In addition, rodent models have been commonly used in toxicology and preclinical studies, yet species-specific effects of drugs have been observed [46] and may underlie the lack of efficacy or unexpected ADRs of drugs when used in humans [47]. Therefore, understanding the species-specific molecular effects of drugs is of translational importance. A detailed species- and tissue-specific drug genomic signature database will

significantly improve our understanding of the molecular networks affected by drugs at a systems level and facilitate network-based drug discovery and ADR prediction for translational medicine.

The potential of using gene signatures to facilitate target and toxicity identification has led to several major efforts in characterizing genomic signatures related to drug treatment [39,48–50]. However, none of the existing platforms offer comprehensive cross-tissue and cross-species in vivo assessments of drug activities to allow prediction of drug effects on individual tissues and translational potential based on consistencies or discrepancies between species. For instance, the Comparative Toxicogenomics Database (CTD), a literature-based resource curating chemical-to-gene/protein associations as well as chemical-to-disease and gene/protein-to-disease connections [48], lacks the cellular and tissue context of the curated interactions. More systematic, data-driven databases like CMap [39] and LINC1000 [49] focus on characterizing and cataloging the genomic footprints of more than ten thousand chemicals using in vitro cell lines (primarily cancer cell lines) to offer a global view of the molecular responses of individual cellular systems to drugs. However, in vitro cell-lines may not always capture in vivo tissue-specificity of drug activities. To move into in vivo systems, large toxicogenomics databases like TG-GATEs [7] and DrugMatrix from the National Toxicology Program of the National Institute of Environmental Health Sciences (https://ntp.niehs.nih.gov/DrugMatrix/index.html) have become available to provide unbiased transcriptome assessment for heart, muscle, liver, and kidney systems. However, information about other organ systems is limited. Efforts to analyze publicly deposited transcriptomic datasets in GEO [51]and ArrayExpress [52], which have broader tissue coverage, for individual drugs have been described [50], but systematic analyses of species- and tissue-specific effects of drugs have not been achieved.

Here, we developed a bioinformatics pipeline (**Figure 2.1**) to curate a database that contains 13,530 rat, human, and mouse transcriptomic datasets across >20 tissues covering 941 drugs. We then evaluated the tissue- and species-specificity of drug signatures as well as the performance of these signatures in gene network-based drug repositioning, toxicity prediction, and comparisons of molecular activities between tissues and species. To benchmark the performance of drug repositioning methods, we conducted in silico evaluation of the retrieval rate of known drugs for various diseases, tested method robustness using simulated disease signatures with noise, compared across existing and new methods, and conducted experimental validation of novel predictions. The drug signatures and network-based drug repositioning tool are hosted on an interactive web server, PharmOmics, to enable public access to drug signatures, integrative analyses and visualization for drug repositioning (http://mergeomics.research.idre.ucla.edu/runpharmomics.php).

Results

Construction of the PharmOmics database containing dose-, tissue- and species-stratified drug signatures

As illustrated in **Figure 2.1A**, we compiled a list of clinically relevant drugs, including 766 approved drugs from Kyoto Encyclopedia of Genes and Genomes (KEGG), the US Food and Drug Administration (FDA), European Medical Agency, and Japanese Pharmaceuticals and Medical Devices Agency, with an additional 175 chemicals from TG-GATEs [7] and DrugMatrix (https://ntp.niehs.nih.gov/DrugMatrix/index.html). The compiled drug list was queried against GEO, ArrayExpress, TG-GATEs, and DrugMatrix to identify transcriptomics datasets from human, mouse, and rat studies, which were further annotated with species, tissue, dosage, and treatment time information (**STAR Methods**). Numbers of datasets, platform information, and sample size distribution are detailed in STAR Methods. Differentially expressed genes (DEGs)

were obtained from individual datasets as “dose/time-segregated signatures” and from meta-analysis of multiple datasets for each drug or each class of drugs across treatment regimen for each tissue and each species as “meta-signatures” (**STAR Methods**). All DEGs are compiled into a drug signature database, comprised of 18,710 gene signatures. Inspection of the database indicated higher coverage for liver compared to other organs/tissues (**Figure 2.1B, 4.1C**), more rat signature sets compared to other species (**Figure 2.1C, 4.1D**), and more signatures from DrugMatrix compared to other data sources (**Figure 2.1B, 4.1D**).

Implementation of the PharmOmics web server for drug signature query and drug repositioning prediction

To allow easy data access and use of the PharmOmics database, we provide drug signature query, species and tissue comparison, drug repositioning, and drug network visualization on an open access web server Mergeomics 2.0 [34,53] (http://mergeomics.research.idre.ucla.edu; **STAR methods**). The PharmOmics web server features three main functions (**Figure 2.2A**). First, species- and tissue-stratified drug signatures and pathways for both the dose/time-segregated and meta signatures can be queried, and comparative analysis to examine similarities and differences between tissues and species for a given drug can be carried out. Second, it features a network drug repositioning tool that is based on the connectivity in a given gene network between PharmOmics drug signatures and user input genes such as a disease signature. Third, the web server offers a gene overlap-based drug repositioning tool that assesses direct overlap between drug gene signatures and user input genes. The gene overlap-based approach is similar to what has been previously implemented in other drug repositioning tools; however, the network-based repositioning approach is unique to PharmOmics. An example of network-based repositioning using a sample liver network and a sample hyperlipidemia gene set as inputs and the resulting drug predictions and network visualization of a top drug, oxymetholone, are shown

in **Figure 2.2B** and 2.**2C**. Lastly, network and gene overlap scores for hepatotoxicity and known ADR links from SIDER database are given in both network- and overlap-based analysis results to predict potential ADRs of the input signature.

Utility of PharmOmics drug signatures in retrieving known therapeutic drugs for various diseases

Drug repositioning has mainly relied on analysis of direct overlaps between drug signatures and disease genes [39,50,54]. Recently, protein-protein interaction networks have also been used for network-based drug repositioning by assessing network connectivity between disease genes and known drug targets [24]. However, it remains unclear whether tissue-specific gene regulatory networks coupled with tissue-matched drug signatures are of value for drug repositioning. To this end, we evaluated the ability of PharmOmics drug signatures to identify drugs for diseases based on network connectivity of gene signatures of diseases and drugs matched by tissue in addition to the commonly used gene overlap approach. We hypothesized that if a drug is useful for treating a disease, the drug and disease signatures likely target similar pathways and therefore would have direct gene overlaps or connect extensively in gene networks. For network-based drug repositioning, we used a network proximity measure between drug DEGs and diseases genes as previously described for protein network-based analysis [24] (**STAR methods**). Here, we used tissue-specific Bayesian gene regulatory networks (BNs) and tested the mean shortest distance between drug DEGs and disease genes. For gene overlap-based drug repositioning, we calculate the Jaccard score, gene overlap fold enrichment, and Fisher's exact test p values as measures of direct gene overlap.

The performance of PharmOmics drug repositioning was first assessed using hyperlipidemia as the test case because multiple known drugs are available as positive controls. Since

hyperlipidemia is most relevant to low density lipoprotein cholesterol (LDL) and liver tissue, we retrieved LDL causal genes and pathways in liver tissue based on genome-wide association studies (GWAS) of LDL in conjunction with genetic regulation of liver gene expression using Mergeomics (**STAR methods**) [9,22,53]. In addition to retrieving disease genes based on GWAS, a hyperlipidemia signature from CTD [48] was also used as an alternative disease signature source. For each drug with different dose and treatment durations, the signature with the highest overlap with the disease signature was used to represent the drug. Gene overlap- and network-based methods using dose/time-segregated signatures had similar overall performance as assessed by the area under receiver operating characteristics curve (~90% Area under the curve of receiver operating characteristic (AUROC); $p<0.001$) in the identification of antihyperlipidemic drugs (**Figure 2.3A, 2.3B**). The dose/time-segregated signatures performed better than the meta signatures when using network-based repositioning (**Figure 2.3C, 2.3D**). When compared to other existing drug repositioning platforms, PharmOmics was able to retrieve higher prediction rankings for the known anti-hyperlipidemia drugs (**Table 2.1**) than CMap (MergeOmics signature $p=0.0064$, CTD signature $p = 0.0056$) and L1000 (MergeOmics signature $p=0.03$, CTD signature $p < 0.001$), while showing comparable results to CREEDS (non-significant for both Mergeomics and CTD signature) based on Wilcoxon signed rank test. PharmOmics also reached better AUROC (**Figure 2.3C, 2.3D**) than CMap and L1000, as well as higher balanced accuracy, defined as (sensitivity+specificity)/2 (**Table 2.2**), than CREEDS, CMap, and L1000. These results support the capacity of PharmOmics as a complementary drug repositioning tool to existing platforms.

To provide molecular insights into the top drug predictions, we examined the disease network overlap patterns of the top drugs, lovastatin, a known anti-hyperlipidemia drug (**Figure 2.3E**), and oxymetholone, a known androgen drug with hyperlipemia ADR (FDA box warning label)

(**Figure 2.3F**). The network approach can retrieve both therapeutic drugs and drugs with ADRs because network connectivity rather than direction of change was the main consideration. Both drugs had DEGs connecting to genes related to cholesterol metabolism and peroxisome proliferator-activated receptor (PPAR) pathways in the hyperlipidemia network (**Figure 2.3E, 2.3F**). However, lovastatin DEGs had direct overlap with cholesterol biosynthesis genes such as *Hmgcr* (target of statin drugs) and *Sqle* along with more DEGs that connected to disease genes, while oxymetholone did not have *Hmgcr* and *Sqle* as DEGs and had smaller disease subnetwork overlap, suggesting key differences between the two drugs. Notably, many drug DEGs did not directly overlap with disease genes, which supports the utility of a network-based drug repositioning approach that does not require the direct retrieval of a known drug target or direct overlap of drug DEGs with disease genes.

We further evaluated the performance of PharmOmics in retrieving known drugs for several other diseases for which known therapeutic drugs are available and can serve as positive controls. Using CTD disease signatures for hepatitis, network-based repositioning obtained 79% AUROC ($p<0.001$, **Figure 2.3G**) in retrieving both steroid and non-steroid anti-inflammatory agents (prediction ranks in **Table S2.1**). We also queried type 2 diabetes signatures and found PharmOmics was able to predict PPAR gamma agonist drugs (79% AUROC, $p=0.04$, **Figure 2.3H**), but not sulfonylurea drugs which act on the pancreatic islet to enhance insulin release (prediction ranks in **Table S2.2**), due to a paucity of drug signatures in islets. Finally, we queried hyperuricemia signatures and network-based repositioning obtained 90% AUROC ($p=0.009$, **Figure 2.3I**, prediction ranks in **Table S2.3**) for detecting anti-hyperuricemia drugs. The overall performance of PharmOmics for these various diseases is better or on par with other platforms (**Figure 2.3G-I**).

We reasoned that network-based repositioning is likely more robust against missing genes in disease signatures than traditional gene overlap-based analysis. To test this, we masked part of the disease gene signatures for hyperlipidemia and hepatitis as test cases. Results showed that network-based repositioning maintained similar performance even when 50% of disease genes were masked, while gene overlap-based strategy showed decrease in performance when 20% or more genes were masked from the disease signatures (**Figure 2.3J**).

Overall, these various test cases using known therapeutic drugs as positive controls support both the utility and robustness of network-based drug repositioning for the diseases tested when drug signatures from the appropriate tissues are available.

Utility of PharmOmics to predict known and novel drugs for non-alcoholic fatty liver disease (NAFLD)

After establishing the performance of PharmOmics in drug repositioning using the case studies above where positive controls are available, we applied PharmOmics to predict potential drugs for non-alcoholic fatty liver disease (NAFLD), for which there is currently no approved drugs. Using NAFLD steatosis signatures from a published mouse study [9] and the CTD NAFLD signatures [48], we predicted PPAR alpha agonists (clofibrate, fenofibrate, bezafibrate, and gemfibrozil), HMG-CoA reductase inhibitors (lovastatin, fluvastatin, and simvastatin), a PPAR gamma agonist (rosiglitazone), and a nonsteroidal anti-inflammatory drug (aspirin) to be among the top 10% of drug candidates based on the average ranking of drugs predicted using both the mouse steatosis signature and CTD NAFLD signature (**Table S2.4**). PPAR agonists have been well supported as potential drugs for NAFLD [55–66], whereas statins showed positive association yet less literature documentation [67–70]. Aspirin was recently reported to be associated with

reducing liver fibrosis progression in a cohort association study in humans [71], but here it was predicted for liver steatosis or general NAFLD.

Next, we sought to validate the top predicted drugs in mitigating liver steatosis. We chose fluvastatin as a positive control due to its high prediction rank across different platforms (top 5% in PharmOmics, CMap, and L1000; top 20% in CREEDS; **Table S2.4**) and better efficacy compared to other statins in improving metabolic phenotypes in a methionine- and choline-deficient diet mouse model used to study non-alcoholic steatohepatitis (NASH) [69]. We also chose to test aspirin as a unique top prediction by PharmOmics (top 5%). In comparison, aspirin had much lower ranks in CREEDS (30%) and CMap (35%) and was not documented in L1000.

Fluvastatin and aspirin were tested using a mouse steatosis model induced by a high fat high sucrose (HFHS) Western diet, which has been previously used to study NAFLD ([9,72–74]. Key genes identified in this diet-induced NAFLD model [9] were known NAFLD-associated genes [75,76] and reproducible in independent human studies [77], supporting its utility as a model for this disease. Comparison between the mice in HFHS group (NAFLD) and the chow group (Control) confirmed that HFHS induced increases in hepatic triglycerides (TG), a measure of liver steatosis, without significant differences in liver weight or other lipids (**Figure 2.4**). Comparison of the fluvastatin and aspirin treated groups with the NAFLD group revealed significant treatment effects by both drugs on mitigating body weight gain (fluvastatin: $p<0.0001$, **Figure 2.5A**; aspirin: $p<0.0001$, **Figure 2.5B**), reducing adiposity (fluvastatin: $p<0.0001$, **Figure 2.5C**; aspirin: $p=0.0008$, **Figure 2.5D**), and decreasing hepatic triglycerides (TG) (fluvastatin: $p=0.0044$, **Figure 2.5E**; aspirin: $p=0.0023$, **Figure 2.5F**). Drug treatments did not significantly alter liver weight, total cholesterol (TC), and unesterified cholesterol (UC) (**Figure 2.5E**, **2.5F**; **Figure 2.6**).

We further investigated whether the effects of the drugs on NAFLD steatosis phenotypes were confounded by food and water intake. No difference was observed in food and water intake in the fluvastatin treatment group (**Figure 2.6E, 2.6F**), but in the aspirin treatment group there was a significant decrease in food intake but no difference in water intake compared to the NAFLD group (**Figure 2.6G, 2.6H**). Adjusting for food intake effects using linear regression showed that the significant effects of fluvastatin on body weight gain (p=0.0306), adiposity (p=0.0022), and hepatic TG (p=0.0190) remained significant. For aspirin, the significant effects on hepatic TG (p=0.0372) remained, but the effects on body weight gain and adiposity (p=0.0511) were no longer significant.

Our experimental validation experiments support the efficacy of both fluvastatin and aspirin in mitigating liver TG levels independent of food and water intake. Agreeing with the PharmOmics prediction ranks, the effects of fluvastatin were stronger than that of aspirin (**Figure 2.5A-F**). Moreover, visualization of the network overlaps between NAFLD signatures and drug signatures revealed more extensive disease network connections for fluvastatin than for aspirin (**Figure 2.5G, 2.5H**), and the signatures of the two drugs connected to pathways involved in NAFLD such as PPAR signaling pathways and fatty acid and steroid biosynthesis.

Utility of PharmOmics drug signatures in predicting and understanding hepatotoxicity

We further explored the potential of coupling PharmOmics drug signatures and tissue gene networks to predict liver toxicity, a major type of ADR for which both toxicity signatures and orthogonal ADR documentations from various independent databases are available for performance evaluation. We used the chemical-induced liver injury signature containing 435 genes from CTD to predict the degree of hepatotoxicity of drugs based on the overlap and liver

gene network connectivity between PharmOmics drug signatures and the CTD liver injury signature. We then used both the liver histological severity from TG-GATEs and the independent FDA drug-induced liver injury (DILI) categories (“most”, “less” – moderate/mild, and “no” DILI concern) as in silico independent validation of the predicted hepatotoxic drugs.

We found that drug ranking of hepatotoxicity from both the network-based and gene overlap-based analyses from PharmOmics increased with higher histological severity as defined by TG-GATEs (**Figure 2.7A**), supporting a positive relationship between the predicted hepatotoxicity scores and experimental hepatotoxicity measures. Next, we tested the performance of PharmOmics in predicting hepatotoxic drugs from the FDA DILI drug database. PharmOmics dose/time-segregated signatures resulted in higher performance (67% AUROC, p=0.0014) compared to the meta signatures (63% AUROC, p = 0.029) and the other platforms tested such as CREEDS, CMap, and L1000 (AUROC 50-53%; non-significant p > 0.05 for CREEDS and L1000; CMap showed significantly higher scores in drugs with lower hepatotoxicity, **Figure 2.7B-2.7C**).

Top drug predictions by PharmOmics based on the CTD hepatotoxicity signatures were wy-14643 (experimental drug with severe histological finding in TG-GATEs), dexamethasone (moderate DILI concern category in FDA and moderate histological finding in TG-GATEs), phenobarbital (moderate DILI concern), indomethacin (“most” DILI concern), and fenofibrate (moderate DILI concern). The network overlapping patterns of the top predicted drugs with the CTD liver injury genes (**Figure 2.7D**) showed that the top predicted drugs exhibited consistent targeting of the hepatotoxicity gene subnetworks.

Since CTD contains a large number (435) of curated hepatotoxicity genes, we hypothesized that this large network could be divided into subnetworks indicative of different mechanisms towards liver toxicity, which might improve toxicity prediction for drugs with different mechanisms. Therefore, we applied the Louvain clustering method to divide the liver injury network defined by the CTD hepatotoxicity genes into subnetworks and filtered out subnetworks with less than 10 genes. These subnetworks showed varying abilities in identifying drugs with DILI concerns (**Table S2.5**). The best performing hepatotoxicity subnetwork showed improved AUROC compared to the whole network (75% vs 67%; **Figure 2.7B**). Further scrutinization of the top performing subnetwork revealed that the antioxidant gene GSR, the phase 2 drug metabolizer *NAT2*, and the inflammatory response gene *IRAK1* showed the best predictability. These results suggest that the network-based toxicity prediction approach may help prioritize predictive genes, pathways, and subnetworks related to hepatotoxicity.

Utility of meta signatures to understand tissue and species specificity

To evaluate tissue and species specificity of drug signatures, we used the meta signatures, which reflect the dose/time-independent, consistent genes affected by drugs across studies in the same tissue or species. We analyzed the overlap in gene signatures for each drug across different tissues and species and visualized the results using UpSetR [78]. As shown in **Figure 2.8A**, the overlap rate in the DEGs of the same drug between tissues and organs is usually less than 5%, indicating a high variability in DEGs between tissues.

As an example, we examined atorvastatin, a HMGCR (β-Hydroxy β-methylglutaryl-CoA receptor) inhibitor, which has well understood mechanisms and has been broadly tested in different tissues under the human species label. We found that two DEGs, TSC22D3 and THBS1, involved in extracellular matrix and inflammation respectively, were shared across

tissues (**Figure 2.8B**). At the pathway level, immune related pathways were shared between blood and liver cells but not in the urogenital system (**Figure 2.8C, Table S2.6**). Unique liver pathways include steroid synthesis and drug metabolism, which is expected as the known target of statin drugs is HMGCR, the rate limiting enzyme in cholesterol biosynthesis in liver. Blood monocyte DEGs indicated changes in inflammation related pathways, while G-protein coupled receptor (GPCR) ligand binding proteins were altered in prostate cancer cells. The tissue specificity of drug meta signatures supports tissue-specific therapeutic responses and emphasizes the need for comprehensive inclusion of drug signatures from different tissue systems.

We also found evidence for high species specificity. As shown in **Figure 2.8D**, the pair-wise overlaps in DEGs between species for the same drug is generally lower than 5%. Here we chose PPAR gamma receptor agonist rosiglitazone as an example because this drug has datasets across human, rat, and mouse in PharmOmics, and its mode of action is well-studied. As shown in **Figure 2.8E** and **Figure 2.8F**, nine genes (*CPT1C, AKR1B1, VNN1, ACSM3, CD36, CPT1A, PDK4, ZNF669, ADH1C*) and several pathways (PPAR signaling and fatty acid, triacylglycerol, and ketone body metabolism) were consistently identified from liver DEGs across species (**Table S2.7**), reflecting the major species-independent pharmacological effects of rosiglitazone. Bile acid related genes were altered in rat datasets, whereas retinol metabolism and adipocytokine pathways were altered in human datasets. The species differences identified highlight the importance of investigating the physiological differences among model systems to facilitate drug design with better translational potential. Our cross-species comparative analysis also revealed that only 21% of the unique drug-tissue pairs (236 out of 1144) have data from two or more species, thus highlighting the need for systematic data generation across species to better understand between-species similarities and differences in drug actions.

Discussion and conclusion

We present PharmOmics, an open-access drug signature database along with a web interface for accessing and utilizing the signatures for various applications. PharmOmics utilizes publicly available drug-related transcriptomic datasets across multiple data repositories and provides unique tissue-, species-, and dose/time-stratified gene signatures that are more reflective of in vivo activities of drugs. We also developed a unique framework for drug repositioning based on tissue-specific gene network models. We examined the potential applications of PharmOmics for various utilities including drug repurposing, toxicity prediction, and comparison of molecular activities between tissues and species. We also carried out in silico performance assessments across drug signature databases and in vivo mouse experiments to validate select drugs from network-based predictions for liver steatosis.

Compared to the well-established CMap and LINC1000 platforms, PharmOmics focuses more on in vivo settings and likely captures physiologically relevant drug signatures to improve drug repositioning performance. Compared to a previous crowdsourcing effort which also utilizes publicly available drug datasets [50], the PharmOmics platform includes more curated databases (TG-GATEs, DrugMatrix Affymetrix, DrugMatrix Codelink datasets) and has a systematic tissue, species, and treatment regimen stratification to facilitate drug repositioning. Comparison across platforms revealed statistically significant gene signature overlaps, but the degree of overlap is low (**Figure 2.9**), supporting that these are complementary platforms. PharmOmics is also the only tool utilizing a gene network framework rather than a direct gene overlap approach. We believe that the increased coverage of in vivo datasets, consideration of tissue-, species-, and dose specificity, and the use of a network approach all contribute to the improved performance

of PharmOmics. However, in cases where tissues, networks, and doses are not available in PharmOmics, existing platforms have advantages.

The use of tissue annotation with BRENDA Tissue Ontology helps normalize organ labels and improves comparability of datasets. The tissue- and species-specific analyses implemented in PharmOmics allows for comprehensive molecular insight into the actions of drug molecules in individual tissues and species. Our results support that different species have unique drug responses in addition to shared features. Therefore, drug responses obtained in animal models require caution when translating to humans. This notion agrees with the long-observed high failure rate of drug development that has primarily relied on preclinical animal models and argues for greater consideration and understanding of inter-species differences in drug actions.

In addition to tissue and species stratification, we also provide detailed dose/time-segregated drug signatures, which can help better understand the dose- and time-dependent effects of drugs through gene signature and pathway comparisons offered through our web server. By contrast, the meta-analysis signatures capture the consistent genes and pathways across treatment regimens, which likely represent core, dose/time-independent mechanisms, and can help address the sample size issue of individual datasets since most drug treatment datasets carried out to date are of small sample size. Repositioning with meta signatures also significantly shortens the computing time in network-based repositioning applications. For instance, computation using 1251 human meta signatures can be completed in 40 minutes, whereas using ~14,000 dose/time-segregated signatures can take 4 hours.

Previous drug repositioning studies support the utility of a protein network-based approach for drug repositioning. Here we show that combining the drug transcriptomic signatures in

PharmOmics with tissue-specific gene regulatory networks and gene signatures of diseases can predict potential therapeutic avenues and tissue toxicity. Compared to other platforms, the use of tissue- and species-specific drug signatures along with network biology is a unique feature of PharmOmics, which enables drug prioritization based on network proximity rather than direct gene overlaps. We demonstrate in various applications that network-based analysis had more robust performance to that of gene overlap-based analysis. Moreover, network-based repositioning offers molecular and mechanistic insights into the therapeutic or toxic effects of drugs. For instance, different NAFLD network overlapping patterns were observed between fluvastatin and aspirin which reflect different drug mechanisms for the same disease phenotype that can be explored further.

In conclusion, we have established a new drug signature database, PharmOmics, across different dosage, species and tissues, which captures the systems level in vivo activities of drug molecules. In addition, we demonstrate the possible means to integrate these signatures with network biology to address drug repositioning needs for disease treatment and to predict and characterize toxicity. Finally, our study tested the concept of tissue-matched drug repositioning and supports consideration of the tissue context of disease and drugs in the improvement of drug repositioning performance, and repositioning efforts will be further expanded when more tissue specific disease and drug signatures are available. PharmOmics has the potential to complement other available drug signature databases to accelerate drug development and toxicology research. It should be noted that we aim to position PharmOmics as a data-driven tool for hypothesis generation. Integration with known drug characteristics to select drug candidates and design follow up experiments are still essential.

Limitations of study

There are several limitations in this study. First, our computational pipeline may not be able to identify all drug datasets from GEO and ArrayExpress database and currently does not accommodate RNA sequencing datasets (~10% of retrieved drug datasets). Variations in annotations of drug names, sample size, definition of treatment vs control groups, and tissue/cell line labeling across datasets make it challenging to design a fully automated pipeline to curate drug datasets. Another issue is that deposited RNA sequencing datasets are in non-standardized formats, with some as raw counts and others as normalized counts such as FPKM and RPKM, making a streamlined and standardized analysis of these datasets difficult. We are currently processing RNA sequencing datasets and will add these to PharmOmics in the future. It is therefore crucial for public data repositories to offer clear definitions and instructions for metadata generation in order to standardize terms and data processing procedures across datasets to facilitate future data reuse. Secondly, the coverage of tissue, species, and treatment regimens across drugs is unbalanced, preventing a thorough comparison across tissues, species, dosages, and treatment windows. We will continue to update our PharmOmics database periodically to include more datasets as they become available to increase the coverage of datasets and drug signatures. Thirdly, the sample sizes for drug treatment studies tend to be small (majority with n=3/group or less). This is an intrinsic limitation of existing drug studies and is a common challenge to existing drug databases including TG-GATEs, DrugMatrix, CMap, L1000, and CREEDS. This fact highlights the need for systematic efforts to increase sample sizes in drug genomic studies. To mitigate the sample size concern and reduce the reliance on individual studies, we implemented a meta-analysis strategy to aggregate drug signatures across studies to derive meta signatures. However, this strategy removes dosage- and time-dependent effects. We offer both options in our database to mitigate sample size concerns through meta-analysis while retaining dose and time regimen information through the

dose/time-segregated analysis. Fourth, our network-based applications are currently limited in the coverage of high-quality tissue specific regulatory networks and computing power. We will continue to expand and improve the tissue networks and computing environment in our web server. Lastly, systematic validation efforts are needed to substantiate the value of drug repositioning tools like PharmOmics. Thus far, we utilized both in silico performance assessments and in vivo experiments to validate our predictions in limited settings. As with the other existing platforms such as CMap and L1000, future application studies and community-based validation efforts are necessary to further assess the value of PharmOmics.

Methods

Curation of tissue- and species-specific drug transcriptomic datasets

A total of 941 drugs, including 766 FDA approved drugs from KEGG, FDA, European Medical Agency, and Japanese Pharmaceuticals and Medical Devices Agency, and 175 chemicals from TG-GATEs and DrugMatrix were queried against GEO, ArrayExpress, TG-GATEs, and DrugMatrix to identify datasets. Duplicated datasets between data repositories were removed. We developed a semi-automated pipeline combining automated search with manual checking to identify relevant datasets for drug treatment. The automated process first extracts datasets containing drug generic names or abbreviations and then inspects the potential datasets for availability of both drug treatment and control labels in the constituent samples. Labels identified by the automated process were also manually checked to validate the labels and remove potential false detections. Only datasets with n>=3/group in both drug treatment and control groups were included in our downstream analyses. Although a larger sample size is desired, the majority (77.7%) of drug transcriptome datasets for the dose/time segregated signature database have n=3/group, 21.9% datasets have n=2/group, and <1% datasets have n>3/group (**Figure 2.10A**). It should be noted that this sample size is used in all major drug/chemical

signature databases, including CMap, L1000, TG-GATEs and DrugMatrix, in order to cover different chemicals and time and dose regimens. GEO/ArrayExpress datasets showed larger sample size variation compared to dedicated toxicogenomics databases (**Figure 2.10B**). Currently our gene signatures were obtained from microarray datasets since RNA-seq datasets were not standardized in the GEO/ArrayExpress platform and different normalization methods will require a different downstream processing pipeline. The 1460 microarray datasets for 342 drugs from GEO/ArrayExpress were from Affymetrix (55%), Illumina (25%) and Agilent (20%) platforms; the 5370 DrugMatrix datasets for 655 drugs and chemicals contained Affymetrix and Codelink microarrays; the 6700 datasets for 169 drugs and chemicals from TG-GATEs mainly used Affymetrix microarrays. Affymetrix and Illumina microarrays provided similar transcriptome coverages while Codelink platform is an older design which only covered around 6000 genes. Agilent microarrays are two-color compared to the other three platforms which used single-color arrays.

Obtaining drug treatment signatures stratified by species and tissues

Species and tissue labels were retrieved based on the metadata of each dataset. Tissue names were standardized based on the BRENDA Tissue Ontology [79]. We implemented a search procedure to climb the ontology tree structure to determine the suitable tissue annotations. This was done by first building a priority list of widely used tissues/organs in toxicological research using the BRENDA Tissue Ontology tree system. Tissue/organ priority order was set to "kidney", "liver", "pancreas", "breast", "ovary", "adipose tissue", "cardiovascular system", "nervous system", "respiratory system", "urogenital system", "immune system", "hematopoietic system", "skeletal system", "integument" (endothelial and skin tissue), "connective tissue", "muscular system", "gland", "gastrointestinal system", and "viscus" (other non-classified tissue). Tissue terms relevant to each of these tissues or organs were curated from the ontology tree

into a tissue/organ ontology table. Next, we looked up terms from our tissue/organ ontology table in the Cell/Organ/Tissue column of the metadata in each transcriptomic dataset. If a tissue/organ term was not found, we searched the title and summary columns of the metadata as well to retrieve additional information. When the search returned multiple tissue terms (for example, breast cancer cell line may be categorized as both epithelial and breast organ), we used the term with the highest priority as described above. We prioritized the tissue terms based on the relevance to toxicology to make tissue assignments unique for each dataset to reduce ambiguity. Manual checking was conducted to confirm the tissue annotation for each dataset.

For each gene expression dataset from GEO and ArrayExpress, normalized data were retrieved, and quantile distribution of the values was assessed. When a dataset was not normally distributed, log2-transformation using GEO2R [51] was applied. For gene expression datasets from Codelink microarrays (DrugMatrix), quantile normalization was conducted. For Affymetrix microarrays (DrugMatrix and TG-GATEs), GCRMA [80]normalization was conducted. To identify differentially expressed genes (DEGs) representing drug signatures, two different strategies were used. First, the widely used DEG analysis method LIMMA [81] was applied to obtain dose and time segregated signatures under false discovery rate (FDR) < 0.05. To overcome the low sample size issue and obtain “consensus” drug signatures for a drug/chemical, LIMMA was also applied to datasets where multiple doses and treatment durations were tested, and treatment effects were derived by combining dose/time experiments for the same drug/chemical in each study. Second, we leveraged different studies for the same drugs or chemicals in the same tissue and species to derive meta-analysis signatures. To address heterogeneity in study design, platforms, sample size, and normalization methods across different data repositories, we applied the characteristic direction method from the

GeoDE package to derive consistent DEGs for each drug across different data sources. GeoDE was designed to accommodate heterogenous datasets that have differing parameters and outputs between treatment and control groups. It uses a “characteristic direction” measure to identify biologically relevant genes and pathways. The normalized characteristic directions for all genes were then transformed into a non-parametric rank representation. Subsequently, gene ranks of a particular drug from the same tissue/organ system and the same organism were aggregated across datasets using the Robust Rank Aggregation method [82], a statistically controlled process to identify drug DEGs within each tissue for each species. Robust Rank Aggregation provides a non-parametric meta-analysis across different ranked lists to obtain commonly shared genes across datasets, which avoids statistical issues associated with heterogeneous datasets. It computes a null distribution based on randomized gene ranks and then compares the null distribution with the empirical gene ranks to obtain a p-value for each gene. The robust rank aggregation process was done for the upregulated and downregulated genes separately to obtain DEGs for both directions under Bonferroni-adjusted p-value < 0.01, a cutoff implemented in the Robust Rank Aggregation algorithm. To obtain species-level signatures for each drug, we further aggregated DEGs across different organs tested for each drug within each species.

Pathway analysis of individual drug signatures was conducted using Enrichr [83] by intersecting each signature with pathways or gene sets from KEGG [84] and gene ontology biological process (GOBP) terms [85]. Gene signatures were defined as FDR < 0.05 for dose/time segregated signatures and Bonferroni-adjusted p-value < 0.01 for meta-analysis signatures. In addition, pathway enrichment analysis based on network topology analysis [86] was conducted using Bioconductor package ROntoTools [87]. Pathways at FDR < 0.05 were considered significant in both methods.

We curated 14,366 drug signatures segregated by treatment dosage and duration, tissue, and species, covering 719 drugs and chemicals, among which 554 are FDA approved. In addition, our meta signatures is a consensus of 4,344 signatures covering 551 drugs across treatment regimens. In total, the entire database is based on 13,530 rat, human, and mouse transcriptomic datasets across >20 tissue or organ systems across 941 drugs and chemicals from GEO, ArrayExpress, DrugMatrix, and TG-GATEs to derive drug signatures. The toxicogenomics databases TG-GATEs and DrugMatrix mainly contain liver and kidney datasets from rats, while public data repositories GEO and ArrayExpress contain datasets with broader tissue and species coverage (**Figure 2.1B**). Overall, the rat datasets are mainly from liver and kidney whereas human and mouse datasets also contained signatures from other tissues and organs such as breast and the nervous system (**Figure 2.1C**). There is also a species bias between the data repositories; GEO covered more mouse and human datasets, TG-GATEs mainly has human and rat datasets, and DrugMatrix curated more rat datasets (**Figure 2.1D**).

Curation of gene networks

We used tissue-specific networks, for example Bayesian gene regulatory networks (BNs) of mouse liver constructed using a previously established method [88,89] based on transcriptomic and genetic data from different mouse liver transcriptomic datasets [21,90–93]. For each data set, 1,000 BNs with different random seeds were reconstructed using Monte Carlo Markov Chain simulation and the model with the best fit for each network was determined. In the resulting set of 1,000 networks, edges appearing in over 30% of the networks were included in a consensus network. This practice has been found to produce experimentally supported regulatory relations between genes [88,89]. The union of nodes and edges from BNs of multiple mouse or human studies were used as tissue-specific networks.

Curation of drug signatures from CMap, LINC1000, and CREEDS for comparison with PharmOmics

To compare PharmOmics with other established drug signature platforms for drug repurposing, we downloaded signatures from L1000FWD [94] (http://amp.pharm.mssm.edu/l1000fwd/download_page) which were well annotated for matched drug signature overlapping comparison. For CREEDS [50] (http://amp.pharm.mssm.edu/CREEDS/) repositioning, the web-based Enrichr [83] tool was used to query disease signatures to their DrugMatrix library, and outputs based on "combined score" implemented by Enrichr were used. Finally, CMap repositioning test were completed through query from the website directly (https://clue.io/) and rank based CMap scoring was used. For CMap and L1000 results which are based on in vitro cell lines, results from all cell lines were summarized to represent common usage of in vitro studies. For CREEDS results where in vivo studies were available, only the corresponding tissues were included for comparability with PharmOmics. We compared PharmOmics with the CREEDS, CMap and L1000 at the regimens that showed the best performance in drug repurposing analysis in each platform.

Curation of disease gene signatures for drug repositioning

To test the potential of PharmOmics drug signatures for drug repositioning, we curated disease gene signatures for hyperlipidemia and NAFLD. Hyperlipidemia was chosen as a test disease because numerous positive control drugs are available to assess the performance of PharmOmics in retrieving the known drugs compared to other existing drug repositioning tools. NAFLD was chosen as another test case since no effective drugs are currently available for this condition and our predictions may help guide future drug development.

The hyperlipidemia signatures were derived from two resources: i) genes and pathways identified by the Mergeomics pipeline [53] based on low-density lipoprotein cholesterol (LDL) genome-wide association study (GWAS) summary statistics data [95], and ii) genes based on mechanistic and therapeutic evidence collected by the Comparative Toxicogenomics Database (CTD) [48] under Mesh ID D006949. These two different resources represent disease gene signatures derived from either GWAS inference or a literature-based system. NAFLD gene signatures were retrieved from i) studies of NAFLD mouse model [9] from a large systems genetics cohort comprised of hundreds of mice from ~100 genetically diverse strains, and ii) CTD gene signature under Mesh ID D065626.

As additional test cases, we also retrieved gene signatures for chemical induced liver injury under CTD Mesh ID D056486, for hepatitis under CTD Mesh ID D006527, for hyperuricemia under CTD Mesh ID D033461 and for type 2 diabetes under CTD Mesh ID D003924.

Measurement of similarity between signatures of drugs, ADRs and diseases

We used two different methods to determine similarities between two signatures (e.g., a drug signature vs. a disease or ADR signature, or a drug signature vs. signature of another drug). The first method is based on signature overlaps and uses a signed Jaccard score based on upregulated genes from the first signature set (a1), upregulated genes from the second signature set (b1), downregulated genes from the first signature set (a2) and downregulated genes from the second signature set (b2). The Jaccard score was defined in the following formula:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$$signed\ Jaccard\ score = J(a1,b1) + J(a2,b2) - J(a1,b2) - J(a2,b1)$$

If there is no direction from the disease signature (a), Jaccard score was determined based on a simple overlap between disease signature (a) and drug signatures (b) without considering direction (b1, b2).

The second method to determine similarities between two signatures is based on a distance measure derived from the mean of shortest path lengths between network key drivers of a drug gene signature (A) and a disease signature (B) in a given Bayesian gene regulatory network (BN) based on a key driver analysis (see below for details). This distance measure is adapted from a previous study using protein interaction networks [24].

$$\text{distance}(\mathrm{B}, \mathrm{A}) = \frac{1}{\|\mathrm{A}\|} \sum_{\mathrm{a} \in \mathrm{A}} \min \mathrm{b} \in \mathrm{B}\ \text{distance}(\mathrm{b}, \mathrm{a})$$

To reduce variations in result, only signatures with more than 10 genes were included in analysis. To obtain a null distribution for shortest path lengths, we permuted genes with the same degree as the drug/disease/ADR genes in each network 1,000 times and calculated a z-score based on the mean and standard error of the null distribution.

Comparison of gene signatures from different species used gene symbol conversion based on ortholog information from HGNC consortium [96]. ROCR package [97] were used to assessed the performance of the gene overlap based or network based methods in drug repositioning or ADR prediction.

Comparison of PharmOmics with existing drug signature platforms

To assess the degree of agreement in drug signatures between the PharmOmics database and existing platforms, we compared PharmOmics with the CREEDS [50] and L1000FWD [54]

databases, for which drug signatures are accessible. As shown in **Figure 2.9**, both the PharmOmics dose/time-segregated signatures and the meta signatures showed better concordance with the two existing platforms than the agreement between CREEDS and L1000FWD, as reflected by higher overlap fold enrichment score and lower statistical p values. These platforms have differences in the datasets and analytical strategies and therefore are complementary. Due to the lack of full access to CMap signatures, we were not able to systematically compare PharmOmics against CMap.

PharmOmics web server implementation

To allow easy data access and use of PharmOmics, we have created a freely accessible web tool deployed on the same Apache server used to host Mergeomics [53], a computational pipeline for integrative analysis of multi-omics datasets to derive disease-associated pathways, networks, and network regulators (http://mergeomics.research.idre.ucla.edu).

The PharmOmics web server features three functions (**Figure 2.2A**). First, it allows queries for species- and tissue-stratified drug signatures and pathways for both the dose/time-segregated and meta signatures. Details of statistical methods (e.g., LIMMA vs characteristic direction), signature type (dose/time-segregated vs meta), and datasets used are annotated. The drug query also includes a function for DEG and pathway signature comparisons between user-selected species and tissues which can be visualized and downloaded. Second, it features a network drug repositioning tool that is based on the connectivity of drug signatures in PharmOmics to user input genes such as a disease signature. This tool requires a list of genes and a gene network that can be chosen from our preloaded gene regulatory networks if relevant or a custom upload (see Applications below for details in implementation). In order to keep reasonable computation time and memory requirement of network repositioning on dose/time

segregated signatures, we implemented on the web server the option to run repositioning with a maximum of 500 genes for each drug signature, which were defined by their FDR value regardless of directionality. In the output, Z-score and p-value results of network repositioning are displayed and available for download. In addition, we list the overlapping genes between drug signatures in the given network and the input genes, the drug genes with direct connections to input genes through one-edge extension, and input genes with one-edge connections to drug genes in the downloadable results file. The output page also provides network visualization which details the genes affected by a drug and their overlap with and direct connections to user input genes using Cytoscape.js. The network nodes and edges files are also available for download and can be used on Cytoscape Desktop. An example of the web interface of the input submission form and results display of the network repositioning tool using a sample liver network and a sample hyperlipidemia gene set is shown in **Figure 2.2B** and **2.2C**. Lastly, the web server offers a gene overlap-based drug repositioning tool that assesses direct overlap between drug gene signatures and user input genes. Gene overlap-based drug repositioning requires a single list of genes or separate lists of upregulated and downregulated genes and outputs the Jaccard score, odds ratio, Fisher's exact test p-value, within-species rank, and gene overlaps for drugs showing matching genes with the input genes. This gene overlap-based approach is similar to what was implemented in other drug repositioning tools, but the network-based repositioning approach is unique to PharmOmics.

Experimental methods for NAFLD drug validation

Eight week old mice underwent dietary treatment with fluvastatin and aspirin purchased from Cayman Chemicals (Ann Arbor, MI). The target intake concentrations of fluvastatin and aspirin were 15mg/kg and 80 mg/kg, respectively, which were chosen based on doses used in previous studies that did not show toxicity [69,98]. These experimental diets were then administered for 10

weeks. The average fluvastatin intake was 14.98 mg/kg/day, and the average aspirin intake was 79.67 mg/kg/day. During drug treatment, metabolic phenotypes such as body weight, body fat and lean mass composition were monitored weekly. Fat and lean mass were measured with Nuclear Magnetic Resonance (NMR) Bruker minispec series mq10 machine (Bruker BioSpin, Freemont, CA). At the end of treatment, mice were sacrificed after a 4 hour fasting period and livers from all groups were weighed, flash frozen, and stored at −80 °C until lipid analysis. For metabolic phenotypes measured at multiple time points (body weight gain and adiposity), differences between groups were analyzed using a 2-way ANOVA followed by Sidak's multiple comparisons test.

Hepatic lipid quantification

Hepatic lipids were extracted using the Folch method [99]. Briefly, frozen liver tissues were homogenized in methanol, and then chloroform was added to each sample to obtain a 2:1 mixture of chloroform and methanol. Samples were then incubated overnight at 4C. Following incubation samples were filtered and magnesium chloride was added to the filtrate and centrifuged. The resulting aqueous phase and soluble proteins were aspirated, and the remaining organic phase was evaporated using nitrogen gas. The dried lipids were dissolved in a Triton X-100 solution. The samples were stored in −80°C prior to analysis. The lipid extracts were analyzed by the UCLA GTM Mouse Transfer Core for triglyceride (TG), total cholesterol (TC), unesterified cholesterol (UC), and phospholipids (PL) levels by colorimetric assay [100,101]. Depending on data normality, the groups were analyzed using either a two-sided t-test or Mann-Whitney test.

Quantification and statistical analysis

Data representation, dispersion and precision measures can be viewed in the figure legends. For in vivo experimental data comparisons using two-way ANOVA (with Sidak post-hoc analysis), t-test and Mann-Whitney test, Prism v8 was used for analysis. Significance level $p < 0.05$ is noted using an asterisk *. For repositioning score two group comparison was performed by Wilcoxon signed rank test in R 4.0.2. Significance levels $p < 0.05$, $p < 0.01$ and $p < 0.001$ are noted using asterisks *, **, and ***, respectively. Multiple group comparison was performed by Kruskal-Wallis test followed by post-hoc pairwise Wilcoxon signed rank test in R 4.0.2. Significance levels $p < 0.05$, $p < 0.01$ and $p < 0.001$ are noted using asterisks *, **, and ***, respectively). Figures were generated by Prism v8 for in vivo experimental data, R default plot for ROC curves and histograms, and R ggplot2 [102] for boxplots. Sample sizes can be viewed in the figure legends. The statistics used in the bioinformatics analysis was described in the individual method sections above.

Data and code availability

- All data, including indexed dataset catalog, pre-computed drug signatures and pre-computed pathway enrichments for individual drugs are deposited to and accessible through the PharmOmics web server (http://mergeomics.research.idre.ucla.edu/runpharmomics.php). We also implemented functions for same-tissue between-species comparison and same-species between-tissue comparison and comparison result download. In addition, network-based drug repositioning analysis and gene overlap-based drug repositioning analysis using all drug signatures are available at http://mergeomics.research.idre.ucla.edu/runpharmomics.php.

- Code for PharmOmics repositioning is available at https://github.com/XiaYangLabOrg/pharmomics.

- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Experimental Model and Subject Details-animals

Since drug repositioning was done using steatosis gene signatures, we validated the predicted drugs using a diet-induced steatosis mouse model which has been previously [9,72–74] used to study NAFLD. Briefly, seven-week old C57BL/6J male mice were purchased from the Jackson Laboratory (Bar Harbor, ME). Mice were maintained on a 12-hour light/dark cycle environment at UCLA and were given ad libitum access to food and water. After a one week acclimation period mice were randomly assigned to four experimental groups (n=7-9/group) on different diets/treatments: regular chow diet (Control) (Lab Rodent Diet 5053, St. Louis, MO), high fat high sucrose (HFHS) diet (Research Diets-D12266B, New Brunswick, NJ) to induce hepatic steatosis, a key NAFLD phenotype, HFHS diet with Fluvastatin treatment (NAFLD + Flu), and HFHS diet with aspirin treatment (NAFLD + Asp). All animal experiments were done under the protocol approved by the UCLA institutional animal care and use committee (IACUC).

Tables

Table 2.1. Prediction percentile of FDA approved antihyperlipidemic drug based on hyperlipidemia signatures from MergeOmics (MO) pipeline and CTD database across different platforms tested.

Platform	**PharmOmics dose/time seg network**		**PharmOmics dose/time seg Jaccard**		**PharmOmics meta**		**CREEDS**		**CMap**		**L1000**	
Disease gene signature	MO	CTD	MO	CTD	MO	CTD	MO	CTD	MO	CTD	MO	CTD
Atorvastatin	0.951	0.794	0.981	0.957	0.498	0.316	0.989	0.82	0.913	0.164	0.962	0.668
Bezafibrate	0.856	0.995	0.901	0.982	0.981	0.932	0.571	0.95	0.332	0.561	0.394	0.755
Cerivastatin	0.989	0.848	0.995	0.962	0.798	0.719	0.986	0.836	0.879	0.516	0.967	0.761
Clofibrate	0.965	0.97	0.802	0.927	0.951	0.992	0.737	0.986	0.196	0.291	0.31	0.615
Clofibric acid	0.93	0.58	0.949	0.892	NA	NA	NA	NA	NA	NA	NA	NA
Fenofibrate	0.984	0.986	0.908	0.883	0.954	0.954	0.797	0.943	0.121	0.108	NA	NA
Fluvastatin	1	0.997	1.000	0.924	0.97	0.985	1	0.815	0.905	0.963	0.958	0.514
Gemfibrozil	0.992	0.962	0.984	0.873	0.787	0.844	0.9	0.712	0.677	0.612	0.363	0.591
Lovastatin	0.995	0.984	0.986	0.986	0.905	0.43	0.993	0.632	0.972	0.084	0.992	0.979
Nafenopin	0.726	0.943	0.472	0.864	NA	NA	0.431	0.712	NA	NA	NA	NA
Niacin	0.192	0.873	0.821	0.309	0.137	0.711	0.719	0.343	0.671	0.171	0.107	0.307
Pravastatin	0.894	0.339	0.911	0.862	NA	NA	0.979	0.854	0.829	0.669	0.592	0.717
Simvastatin	0.949	0.935	0.856	0.992	0.916	0.909	0.996	0.9	0.972	0.951	0.987	0.843
Ciprofibrate	NA	NA	NA	NA	NA	NA	NA	NA	0.685	0.998	0.292	0.272
Ezetimibe	NA	NA	NA	NA	NA	NA	NA	NA	0.905	0.982	0.657	0.269
Probucol	NA	NA	NA	NA	NA	NA	NA	NA	0.552	0.115	0.018	0.529
Rosuvastatin	NA	NA	NA	NA	NA	NA	NA	NA	0.913	0.056	0.905	0.464
Median	**0.951**	**0.943**	**0.911**	**0.924**	**0.911**	**0.876**	**0.94**	**0.828**	**0.829**	**0.516**	**0.624**	**0.603**
Mean	**0.879**	**0.862**	**0.890**	**0.878**	**0.79**	**0.779**	**0.841**	**0.792**	**0.701**	**0.483**	**0.607**	**0.592**
Total number of drugs	**369**	**369**	**369**	**369**	**263**	**263**	**281**	**281**	**934**	**934**	**867**	**867**

Table 2.2. Comparison of drug repositioning performance between PharmOmics and other existing platforms for hyperlipidemia.

Drug pool for each database was limited to FDA approved drugs to match the drug selection criteria in PharmOmics to make results comparable. Significance were defined at the recommended cutoffs for each platform: z-score < -2.33 in PharmOmics, overlap BH adjusted p < 0.05 in CREEDS, L1000 and PharmOmics dose/time segregated Jaccard, and connection score > 95 or < -95 in CMap query system. For CMap and L1000, drug signatures from all cell lines (CMap_all or L1000_all) or from the hyperlipidemia relevant liver cell line HepG2(CMap_HEPG2 or L1000_HEPG2) were used.

Drug signature platform	**Total FDA listed drugs**	**Significant drugs (% total)**	**Known hyperlipidemia drugs**	**Significant hyperlipidemia drugs (% known drugs)**	**Balanced accuracy** (sensitivity+ specificity/2)
PharmOmics meta_liver	263	33 (12.5%)	10	6 (60%)	74.7%
PharmOmics dose/time segregated_liver - network	369	29 (7.9%)	13	9 (69.2%)	81.8%
PharmOmics dose/time segregated_liver - Jaccard	369	171 (46.3%)	13	12 (92.3%)	73.8%
CMap	934	264 (28.6%)	15	8 (53.3%)	62.7%
CMap_HEPG2	667	135 (20.3%)	13	1 (7.7%)	43.6%
L1000	867	428 (49.3%)	14	8 (57.1%)	54.1%
L1000_HEPG2	153	37 (24.2%)	5	0 (0%)	37.5%
CREEDS_liver	281	257 (91.4%)	12	12 (100%)	54.5%

Table S2.1. Prediction percentile of steroid and non-steroid anti-inflammatory drugs based on hepatitis signatures from CTD database across different platforms tested

Table S2.2. Prediction percentile of FDA approved anti-diabetic drug based on type2 diabetes signatures from CTD database across different platforms tested

Table S2.3. Prediction percentile of FDA approved gout treatment drug based on hyperuricemic signatures from CTD database across different platforms tested.

Table S2.4. Network repositioning result for non-alcoholic fatty liver disease based on genetic pathways obtained from studies of female and male mice.

Table S2.5. Submodule repositioning result based on signatures from CTD chemical induced liver injury

Table S2.6. Cross-tissue comparison of Atorvastatin Pathways.

Table S2.7. Cross-species comparison of Rosuvastatin Pathways.

Figures

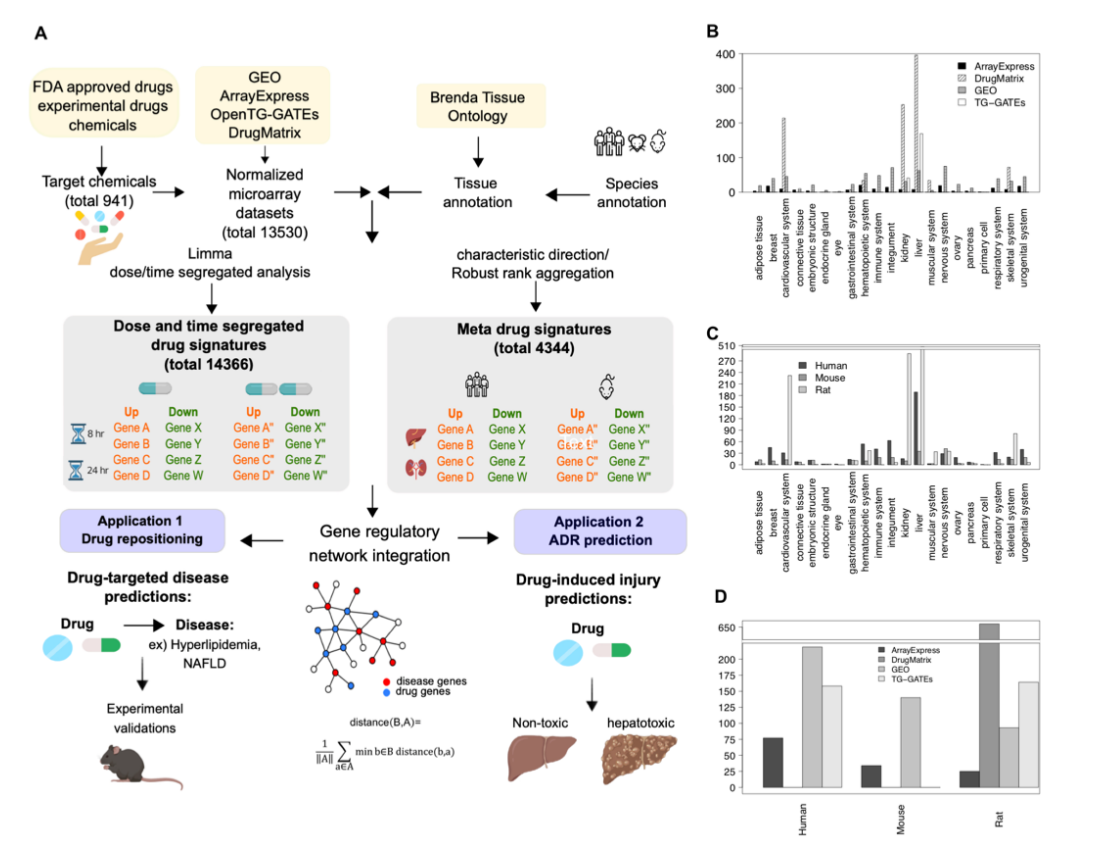
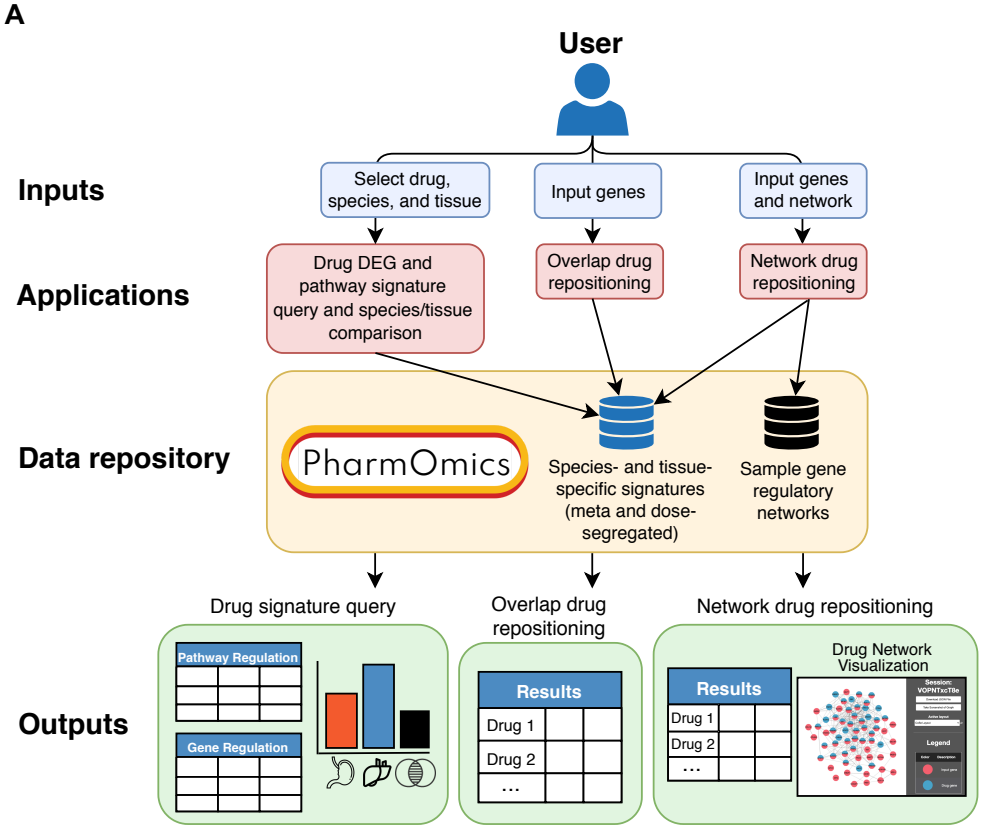


Figure 2.1 PharmOmics data processing pipeline and database summary.

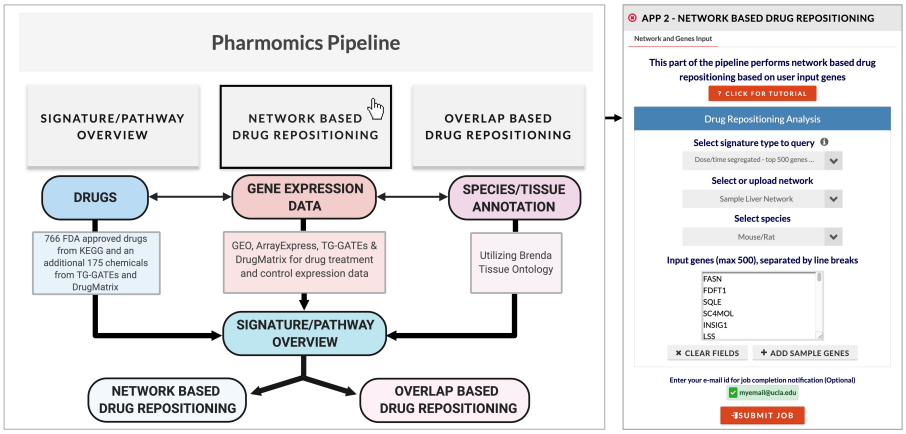
(A) FDA approved drugs were searched against GEO, ArrayExpress, TG-GATEs, and DrugMatrix data repositories. Additional experimental drugs and chemicals from TG-GATEs and DrugMatrix were also included. Datasets were first annotated with tissue and species information, followed by retrieval of dose/time-segregated using LIMMA [81]or meta-analysis drug signatures using GeoDE [94] and Robust Rank Aggregation [82]. These signatures were used to conduct drug repositioning analysis and hepatotoxicity prediction based on either direct gene overlaps or a gene network-based approach. (B) Summary of available datasets based on data

sources and tissues. Y-axis indicates unique dataset counts, and X-axis indicates tissue and data resources. (C) Summary of available datasets based on tissues and species. Y-axis indicates unique dataset counts, and X-axis indicates tissue and species. (D) Summary of available datasets based on data sources and species. Y-axis indicates unique dataset counts, and X-axis indicates data resources and species.

A


User
Inputs
Select drug, species, and tissue
Input genes
Input genes and network
Applications
Drug DEG and pathway signature query and species/tissue comparison
Overlap drug repositioning
Network drug repositioning
Data repository
PharmOmics
Species- and tissue-specific signatures (meta and dose-segregated)
Sample gene regulatory networks
Drug signature query
Overlap drug repositioning
Network drug repositioning
Outputs
Pathway Regulation
Gene Regulation
Results
Drug 1
Drug 2
...
Results
Drug 1
Drug 2
...
Drug Network Visualization

B


Pharmomics Pipeline
SIGNATURE/PATHWAY OVERVIEW
NETWORK BASED DRUG REPOSITIONING
OVERLAP BASED DRUG REPOSITIONING
DRUGS
GENE EXPRESSION DATA
SPECIES/TISSUE ANNOTATION
766 FDA approved drugs from KEGG and an additional 175 chemicals from TG-GATEs and DrugMatrix
GEO, ArrayExpress, TG-GATEs & DrugMatrix for drug treatment and control expression data
Utilizing Brenda Tissue Ontology
SIGNATURE/PATHWAY OVERVIEW
NETWORK BASED DRUG REPOSITIONING
OVERLAP BASED DRUG REPOSITIONING
APP 2 - NETWORK BASED DRUG REPOSITIONING
Network and Genes Input
This part of the pipeline performs network based drug repositioning based on user input genes
CLICK FOR TUTORIAL
Drug Repositioning Analysis
Select signature type to query
Select or upload network
Sample Liver Network
Select species
Mouse/Rat
Input genes (max 500), separated by line breaks
FASN
FDFT1
SQLE
SC4MOL
INSIG1
LSS
CLEAR FIELDS
ADD SAMPLE GENES
Enter your e-mail id for job completion notification (Optional)
myemail@ucla.edu
SUBMIT JOB

C

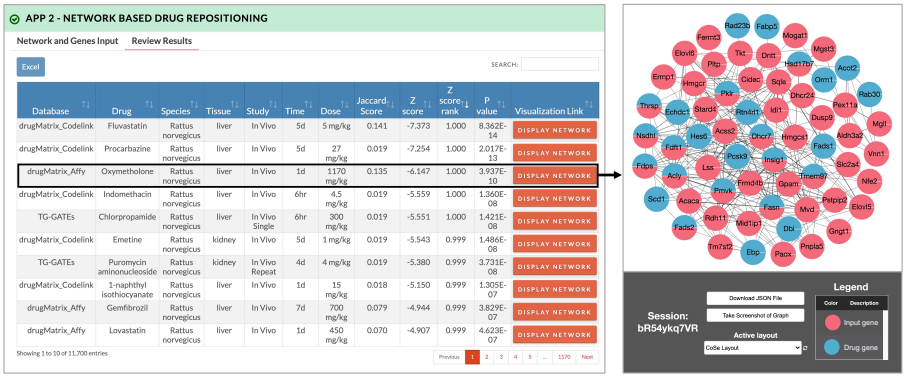

APP 2 - NETWORK BASED DRUG REPOSITIONING
Network and Genes Input
Review Results
Excel
SEARCH:
Database
Drug
Species
Tissue
Study
Time
Dose
Jaccard Score
Z score
Z score rank
P value
Visualization Link
drugMatrix_Codelink Fluvastatin Rattus norvegicus liver In Vivo 5d 5 mg/kg 0.141 -7.373 1.000 8.362E-14 DISPLAY NETWORK
drugMatrix_Codelink Procarbazine Rattus norvegicus liver In Vivo 5d 27 mg/kg 0.019 -7.254 1.000 2.017E-13 DISPLAY NETWORK
drugMatrix_Affy Oxymetholone Rattus norvegicus liver In Vivo 1d 1170 mg/kg 0.135 -6.147 1.000 3.937E-10 DISPLAY NETWORK
drugMatrix_Codelink Indomethacin Rattus norvegicus liver In Vivo 6hr 4.5 mg/kg 0.019 -5.559 1.000 1.360E-08 DISPLAY NETWORK
TG-GATEs Chlorpropamide Rattus norvegicus liver In Vivo Single 6hr 300 mg/kg 0.019 -5.551 1.000 1.421E-08 DISPLAY NETWORK
drugMatrix_Codelink Emetine Rattus norvegicus kidney In Vivo 5d 1 mg/kg 0.019 -5.543 0.999 1.486E-08 DISPLAY NETWORK
TG-GATEs Puromycin aminonucleoside Rattus norvegicus kidney In Vivo Repeat 4d 4 mg/kg 0.019 -5.380 0.999 3.731E-08 DISPLAY NETWORK
drugMatrix_Codelink 1-naphthyl isothiocyanate Rattus norvegicus liver In Vivo 1d 15 mg/kg 0.018 -5.150 0.999 1.305E-07 DISPLAY NETWORK
drugMatrix_Affy Gemfibrozil Rattus norvegicus liver In Vivo 7d 700 mg/kg 0.079 -4.944 0.999 3.829E-07 DISPLAY NETWORK
drugMatrix_Affy Lovastatin Rattus norvegicus liver In Vivo 1d 450 mg/kg 0.070 -4.907 0.999 4.623E-07 DISPLAY NETWORK
Showing 1 to 10 of 11,700 entries
Previous
Next
Session: bR54ykq7VR
Download JSON File
Take Screenshot of Graph
Active layout
CoSe Layout
Legend
Color
Description
Input gene
Drug gene

Figure 2.2 PharmOmics web server implementation.

(A) PharmOmics web server outline. The web server hosts drug signature and pathway queries, between-tissue and -species drug signature comparisons, and network-based and gene overlap-based drug repositioning. Users can query, download, and perform drug repositioning using all species- and tissue-specific meta and dose/time-segregated signatures. Interactive results tables and network visualizations are displayed on the website and available for download. (B) User interface of network drug repositioning web tool using sample hyperlipidemia gene set and sample mouse Bayesian gene regulatory network. Inputs to network drug repositioning includes i) signature type to query (meta-analyzed, dose/time-segregated with top 500 genes per signature, or dose/time-segregated with all genes), ii) network (custom upload or select a sample network), iii) species (relating to the species of the network being used), and iv) genes. In the example case we choose dose/time-segregated signatures using top 500 genes, a sample liver network, mouse/rat species, and the sample hyperlipidemia gene set (loaded from “Add sample genes”). If human gene symbols are provided with the “Mouse/Rat” species selection, the genes will be converted to mouse/rat symbols. (C) After the job is complete, the results file is displayed on the website and available for download. Subnetworks of top ranked drugs can be visualized using the “Display Network” button which will load an interactive display of the subnetwork topology for a select drug. For example, the oxymetholone drug signature in rat liver is a top hit, and the drug network is shown on the right. Additional data in the downloadable results file include the genes that are both a drug gene and an input gene in the network, drug genes that are directly connected (first neighbor) to input genes, and input genes directly connected to drug genes.

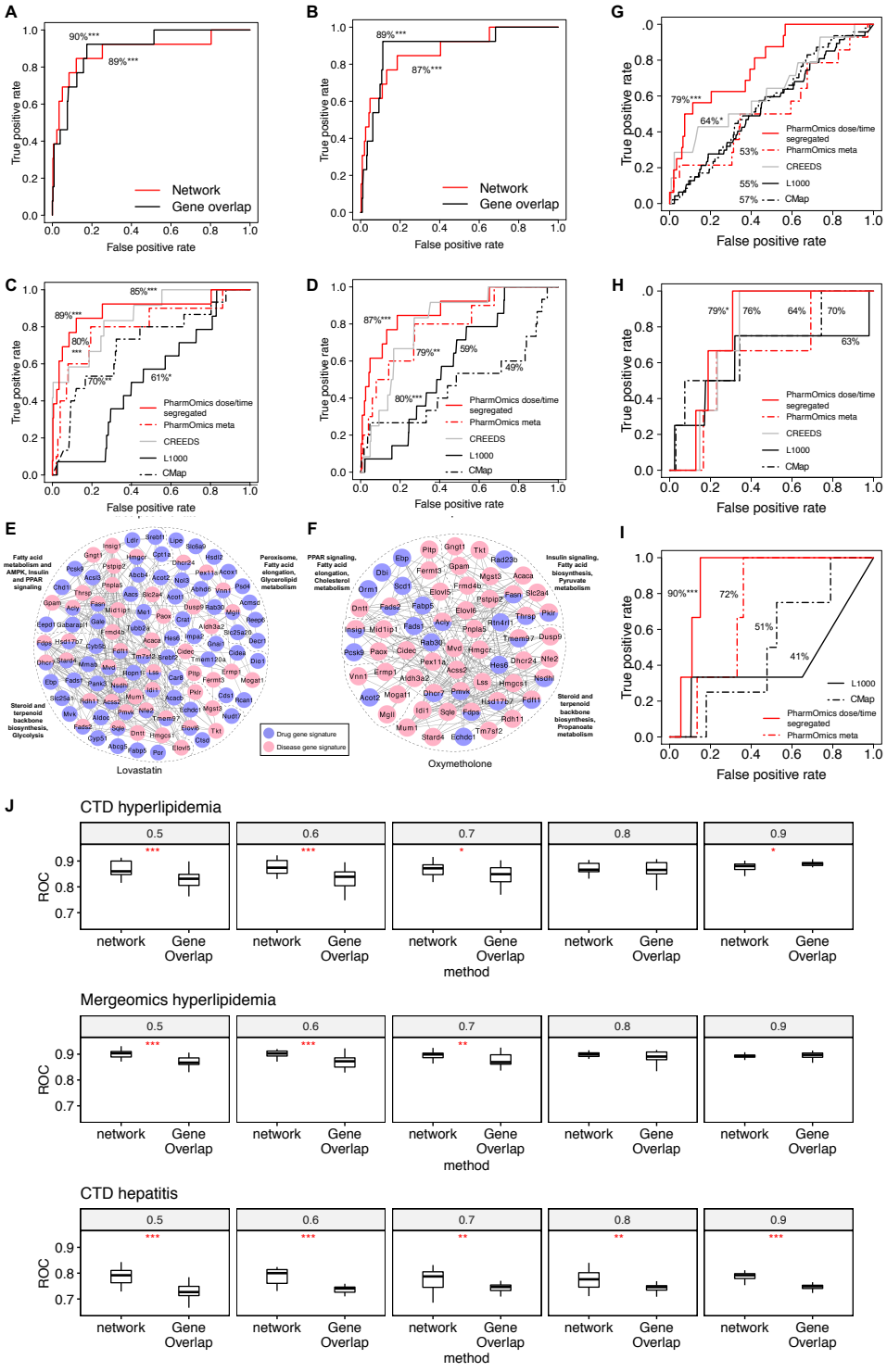


Figure 2.3 Drug repositioning using PharmOmics for diseases with known therapeutics.

(A and B) Area under the curve of receiver operating characteristics (AUROC) plots for network-based repositioning and gene overlap-based repositioning in identifying anti-hyperlipidemia drugs (total n = 369, target n=13) using (A) Mergeomics hyperlipidemia signature or (B) CTD hyperlipidemia signature. (C and D) Comparison of drug repositioning performance between PharmOmics network-based approach with CREEDS (total n = 281, target n=12), using the "combined score" generated by the enrichment analysis tool implemented in Enrichr), L1000 (total n = 867, target n=14), and CMap query system (total n = 934, target n=15) using (C) Mergeomics hyperlipidemia signature and (D) CTD hyperlipidemia signature to identify anti-hyperlipidemic drugs. For drugs with multiple datasets with different doses and treatment times, only the best performing signature was used. (E and F) Drug-hyperlipidemia subnetwork based on Mergeomics hyperlipidemia signature (red) and drug signature (blue) showing first neighbor (direct) connections using (E) lovastatin and (F) oxymetholone signatures. Direct overlapping genes between disease and drug signatures are network nodes colored with both red and blue. (G - I) Comparison of drug repositioning performance between PharmOmics network-based approach with L1000, CREEDS and CMap query system using CTD signatures for hepatitis (G), type 2 diabetes (H), and hyperuricemia (I) to identify steroid and non-steroidal anti-inflammatory drugs (n=16 in PharmOmics, n=14 in CREEDS, n=47 in CMap, n=47 in L1000) (G), PPAR gamma agonists (n=11 in PharmOmics, n=9 in CREEDS, n=13 in CMap, n=13 in L1000) (H), and anti-hyperuricemia drugs (n=3 in PharmOmics, n=4 in CMap, n=3 in L1000) (I), respectively. Note that in (I), CREEDS was not included due to lack of anti-hypouricemic drugs. (J) Boxplot showing AUROC performance with different proportion of original disease signatures used after masking disease genes. For each proportion, 20 random sampling of original disease signature was conducted to obtain AUROC in identifying disease related drugs. Wilcoxon signed rank test was used to calculate significance in difference between gene overlap-based

AUROC and network-based AUROC. *, **, *** indicates $p < 0.05$, $p < 0.01$ and $p < 0.001$ repectively from Wilcoxon signed rank test. See also Table S2.1, S2.2, S2.3

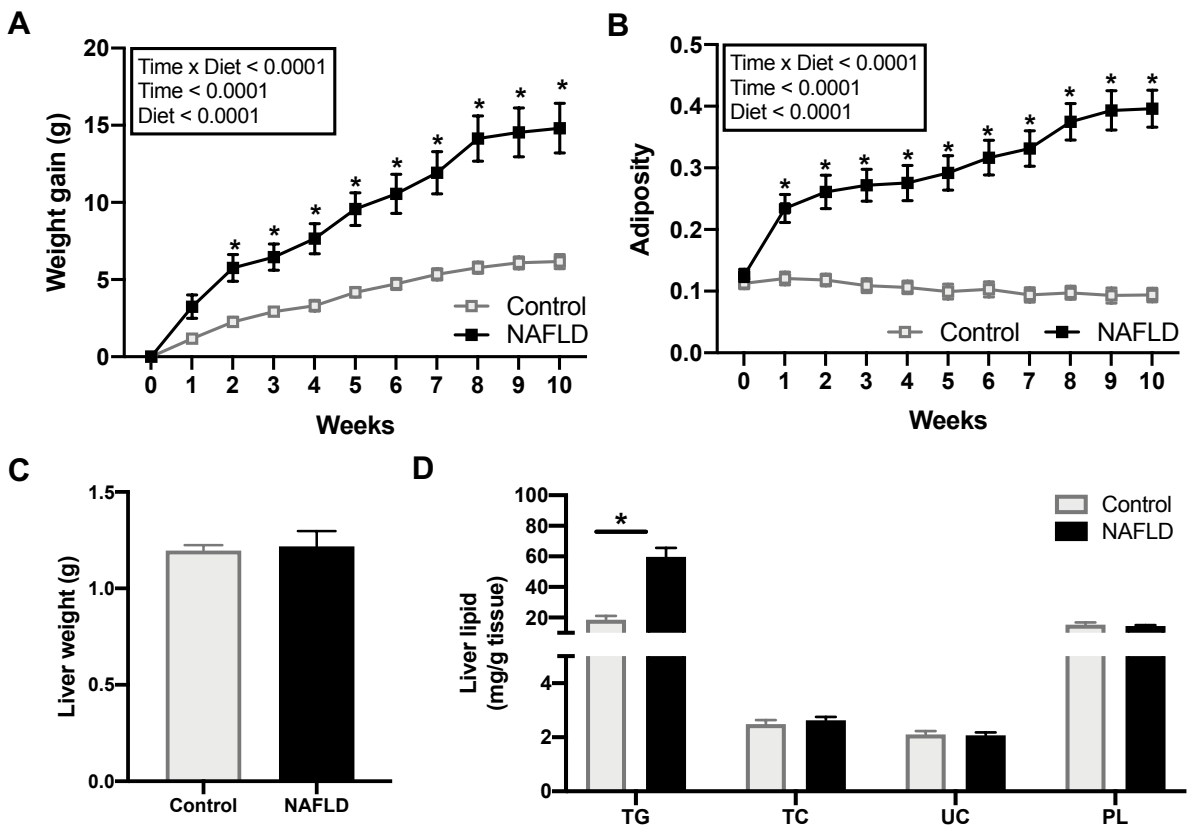


Figure 2.4. Effects of high fat high sucrose diet on body composition and liver lipids in C57BL/6J mice.

(A) Time course of body weight gain of mice on control and high fat high sucrose (HFHS) diet for 10 weeks. (B) Time course of adiposity of mice on control and HFHS diet for 10 weeks. (A and B) Data are represented as mean +/- SEM and was analyzed by two-way ANOVA followed by Sidak post-hoc analysis to examine treatment effects at individual time points. (C) Bar plot of liver weight in mice on a control and HFHS diet. (D) Bar plot of hepatic lipid levels in mice on a control and HFHS diet (D). Triglyceride (TG), Total Cholesterol (TC), Unesterified Cholesterol (UC), Phospholipid (PL). (C and D) Data are represented as mean +/- SEM and was analyzed using either the two-sided t-test or Mann-Whitney test. P value <0.05 was considered significant and is denoted by an asterisk (*). Sample size n = 8-9/group. Control diet (Control); HFHS diet (NAFLD).

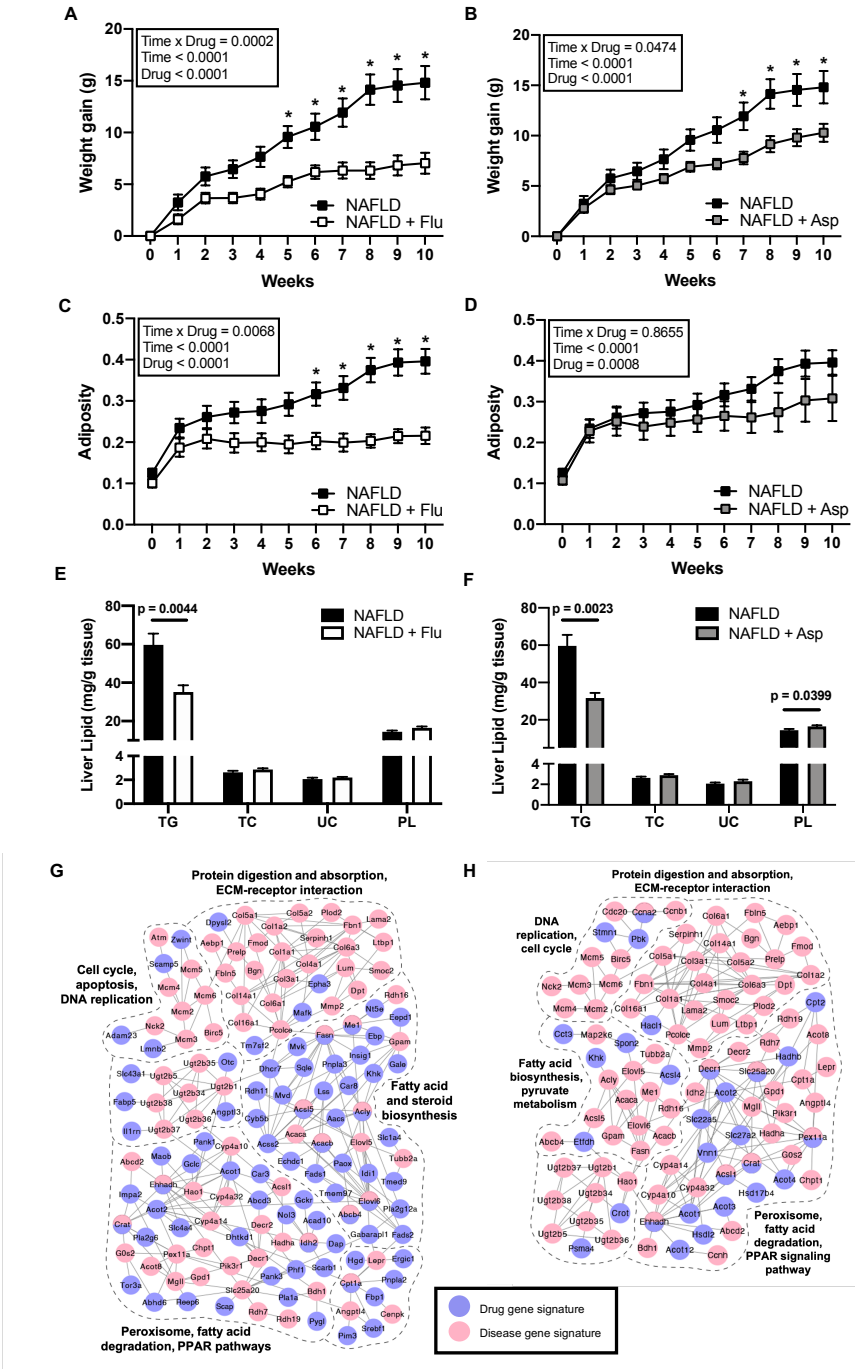


Figure 2.5 *In vivo* validation of top predicted drugs fluvastatin and aspirin on preventing NAFLD phenotypes in a diet-induced NAFLD mouse model.

Mouse groups include C57BL/6J mice fed a high fat high sucrose (HFHS) diet to induced liver steatosis (NAFLD), HFHS with fluvastatin (NAFLD + Flu), and HFHS with aspirin (NAFLD +

Asp). (A and B) Time course of body weight gain in NAFLD mice treated with fluvastatin (A) or aspirin (B) over 10 weeks. (C and D) Time course of fat mass and muscle mass ratio (adiposity) in mice treated with fluvastatin (C) or aspirin (D) over 10 weeks. (A-D) Data are represented as mean +/- SEM and were analyzed by two-way ANOVA followed by Sidak post-hoc analysis to examine treatment effects at individual time points. P value < 0.05 was considered significant and is denoted by an asterisk (*). (E and F) Quantification of lipids in the liver of mice on fluvastatin (E) or aspirin (F) treatment for 10 weeks. Triglyceride (TG), Total Cholesterol (TC), Unesterified Cholesterol (UC), Phospholipid (PL). Data are represented as mean +/- SEM and were analyzed using two-sided t-test. P value < 0.05 was considered significant and is denoted by an asterisk (*). Sample size n = 7-8/group. (G) Gene network view of fluvastatin gene signatures overlapping with NAFLD disease signatures. (H) Gene network view of aspirin gene signatures overlapping with NAFLD disease signatures. See also Figure 2.4, 4.6 and Table S2.4

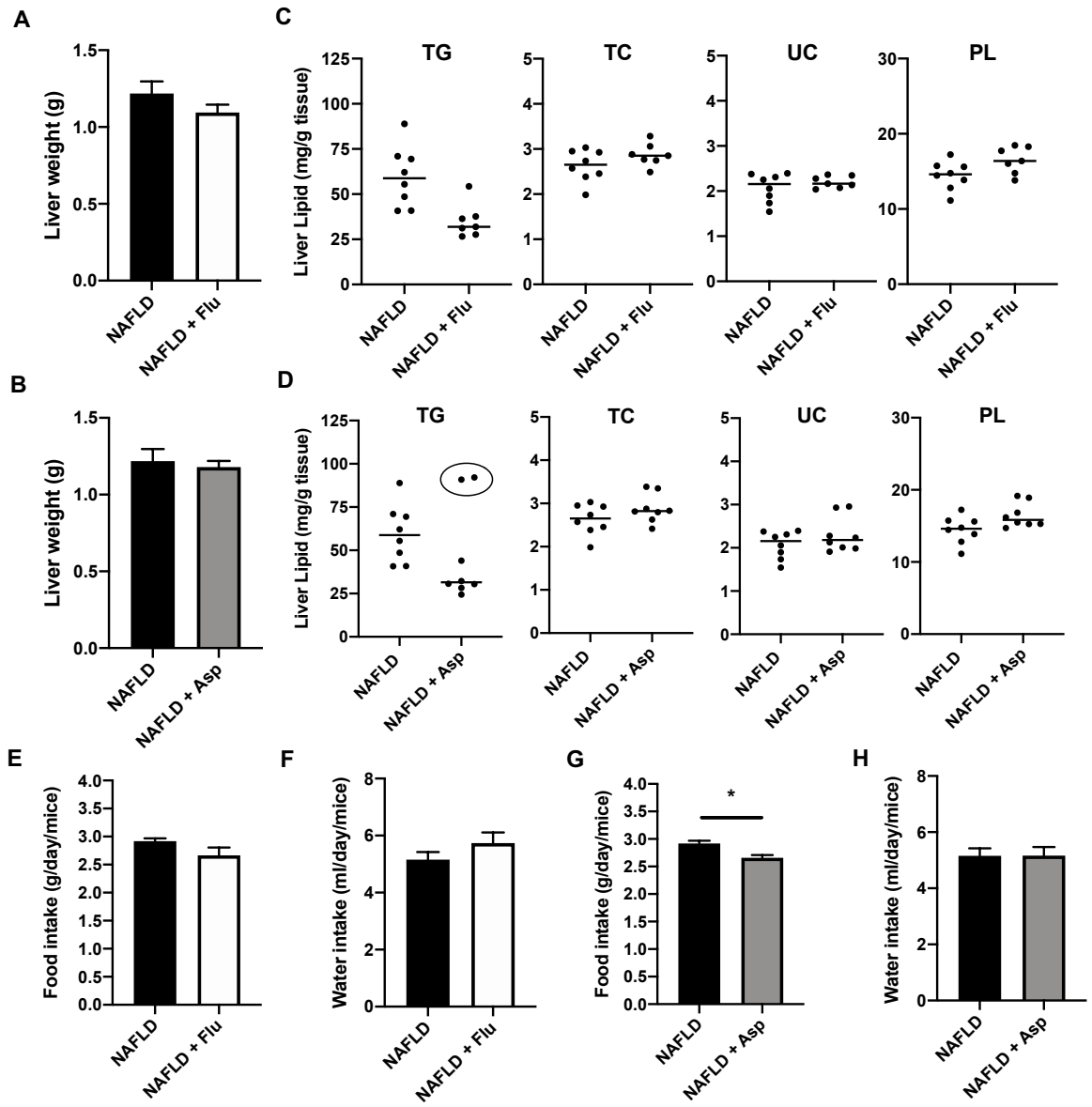


Figure 2.6 Liver weight, food intake, and water intake in C57BL/6J mice on a HFHS diet with or without fluvastatin or aspirin.

(A and B) Bar plot of liver tissue weight in mice on a HFHS diet with or without fluvastatin (A) or aspirin (B). (C and D) Dotplots of lipid levels in the liver of mice on fluvastatin (C) or aspirin (D) treatment for 10 weeks. Triglyceride (TG), Total Cholesterol (TC), Unesterified Cholesterol (UC), Phospholipid (PL). Circle indicates the identified outliers using the ROUT method on Graphpad (Prism v8), which were removed from the subsequent analysis. (E and F) Bar plot of food (E)

and water (F) intake in mice on a HFHS diet with or without fluvastatin. (G and H) Bar plot of food (G) and water (H) intake in mice on a HFHS diet with or without aspirin. (A, B and E-H) Data are represented as mean +/- SEM. All data was analyzed using a two-sided t-test. P value <0.05 was considered significant and is denoted by an asterisk (*). Sample size n = 7-8/group. HFHS group (NAFLD); HFHS with Fluvastatin (NAFLD + Flu); HFHS with Aspirin (NAFLD + Asp).

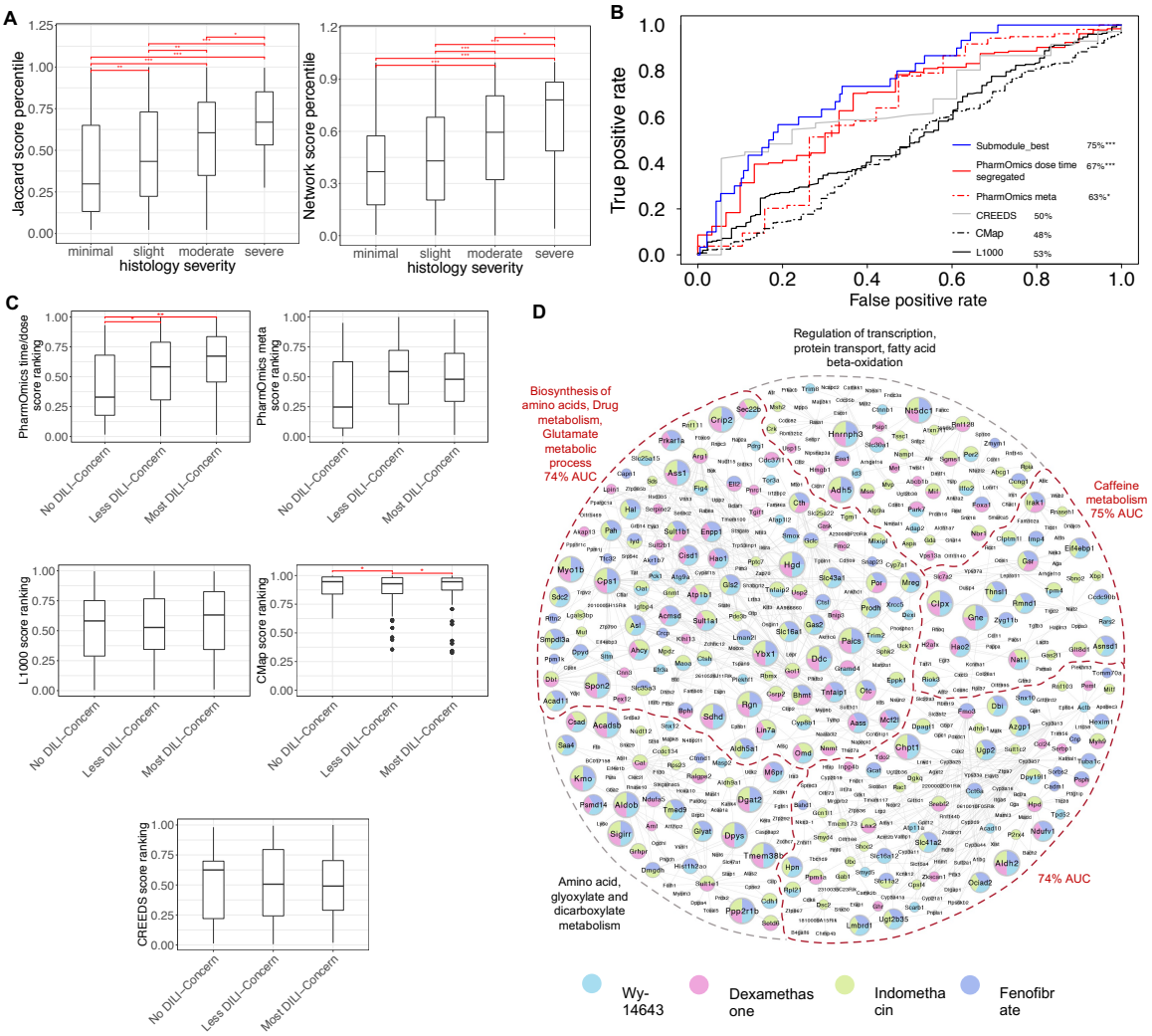


Figure 2.7 Utility of PharmOmics drug signatures in hepatotoxicity prediction.

The analysis was based on matching between PharmOmics drug signatures and hepatotoxicity signatures of drug induced liver injury (DILI) curated from comparative toxicogenomics database (CTD). (A) Boxplots of gene overlap-based hepatotoxicity ranking (left) and network-based hepatoxicity ranking (right) by PharmOmics, across four categories of liver injury histological severity defined by the independent TG-GATEs database (x-axis) (all doses included, n=205 in "minimal" category, n=221 in "slight" category, n=147 in "moderate" category, n=37 in "severe"

category). (B) ROC curves comparing PharmOmics with other tools in predicting hepatotoxic drugs from the FDA DILI drug database. For PharmOmics, three sets of tests were performed, where dose/time-segregated drug signatures, meta signatures, or a hepatotoxicity subnetwork was used. Significance were calculated by comparing “no DILI-concern” category (n= 30 in PharmOmics dose/time segregated signatures, n=19 in PharmOmics meta, signatures, n=94 in CMap, n=88 in L1000, n=18 in CREEDS) vs “less DILI-concern” plus “most DILI-concern” categories (n= 185 in PharmOmics dose/time segregated signatures, n=156 in PharmOmics meta signatures, n=276 in CMap, n=251 in L1000, n=142 in CREEDS). (C) Hepatotoxicity signature matching scores from various drug repositioning platforms across three different DILI drug categories. For drugs with multiple dose and time points, only the best score was used. PharmOmics scores are derived from network-based matching; CMap scores were derived from the CMap online query plarform; L1000 scores are from Jaccard scores from the L1000 plarform; CREEDS scores are from the combined scores derived from enrichr platform. Boxplots show interquartile range (IQR) and median values (line inside the box). IQR was defined as between 25th (Q1) and 75th (Q3) percentile. The upper and lower bars indicate the points within Q3 + 1.5*IQR and Q1 – 1.5*IQR, respectively. (D) Liver hepatotoxicity network based on CTD hepatotoxicity genes and its overlap with drug signatures of 4 of the top 5 predicted drugs by PharmOmics which had >50 drug signature genes. Phenobarbital was among the top 5 drugs but was not included in the figure due to its small DEG size. Colors of the network nodes denote the different drugs targeting the genes. The top 3 predictive subnetworks are labeled in red. Kruskal-Wallis test followed by post-hoc pairwise Wilcoxon signed rank test was used for statistics in A and C and Wilcoxon signed rank test was used to calculate significance for B. *, **, *** indicates $p < 0.05$, $p < 0.01$ and $p < 0.001$, respectively. See also Table S2.5

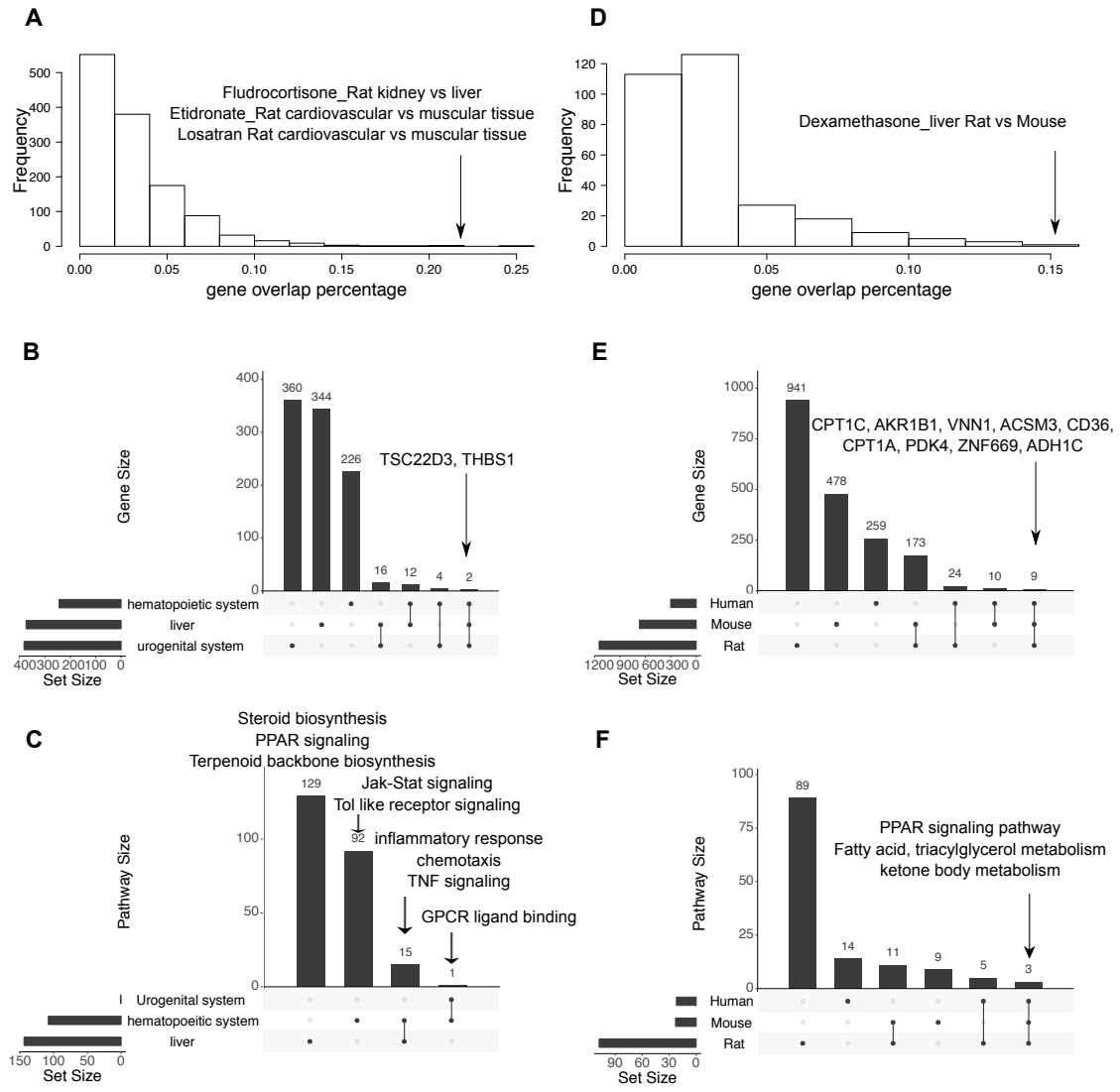


Figure 2.8 Cross-tissue and cross-species comparisons of drug signatures in PharmOmics.

(A) Distribution of drug signature overlap percentages between tissue pairs in matching species from PharmOmics meta database. Arrow points to the pairs of tissues for drugs with high overlap in gene signatures. (B) Upset plot of cross-tissue comparison for atorvastatin signatures genes. Y-axis indicates number of genes. (C) Upset plot of cross-tissue comparison for pathways enriched in atorvastatin signatures. Y-axis indicates number of pathways. (D) Distribution of drug signature overlap percentages between pairs of species for matching

tissues from PharmOmics meta signature database. Arrow points to the species pair with high gene signature overlap for a matching drug. (E) Upset plot of cross-species comparison for rosiglitazone liver gene signatures. (F) Upset plot of cross-species comparison for pathways enriched in rosiglitazone liver signatures. Pairs of tissues with shared drug signature genes or pathways are connected with black vertical lines in the bottom portion of the Upset plots.

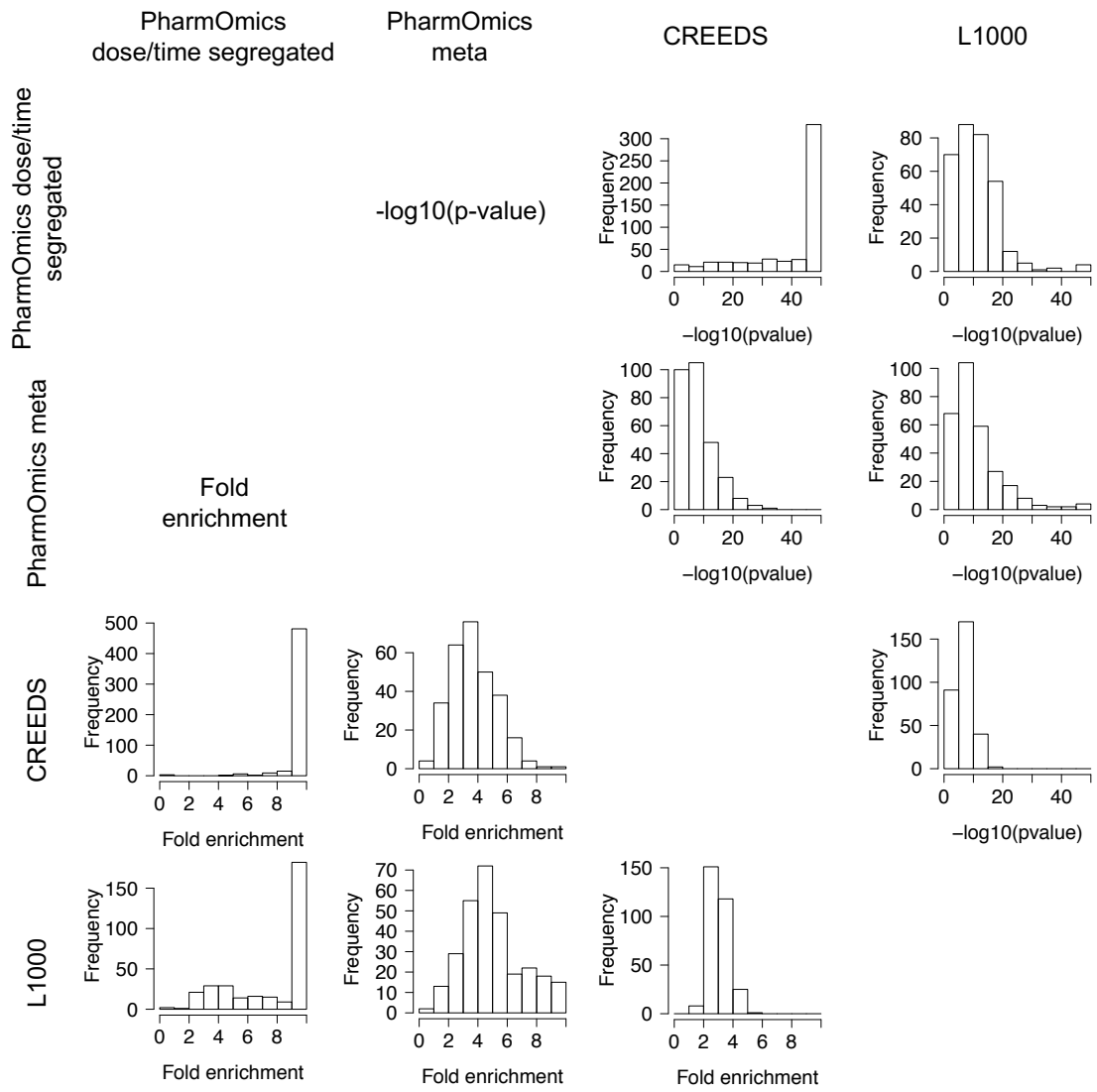


Figure 2.9 Comparison of drug signatures between PharmOmics and existing drug signature databases CREEDS and L1000.

For each drug in PharOmics database shared among other databases, only the overlap scores for the best matched signatures between two databases are used. Lower left triangular matrix represents the histogram of the gene overlap fold enrichment scores calculated using hypergeometric test, and the upper right triangular matrix represents the histogram of the -log10(pvalue) from Fisher's exact test.

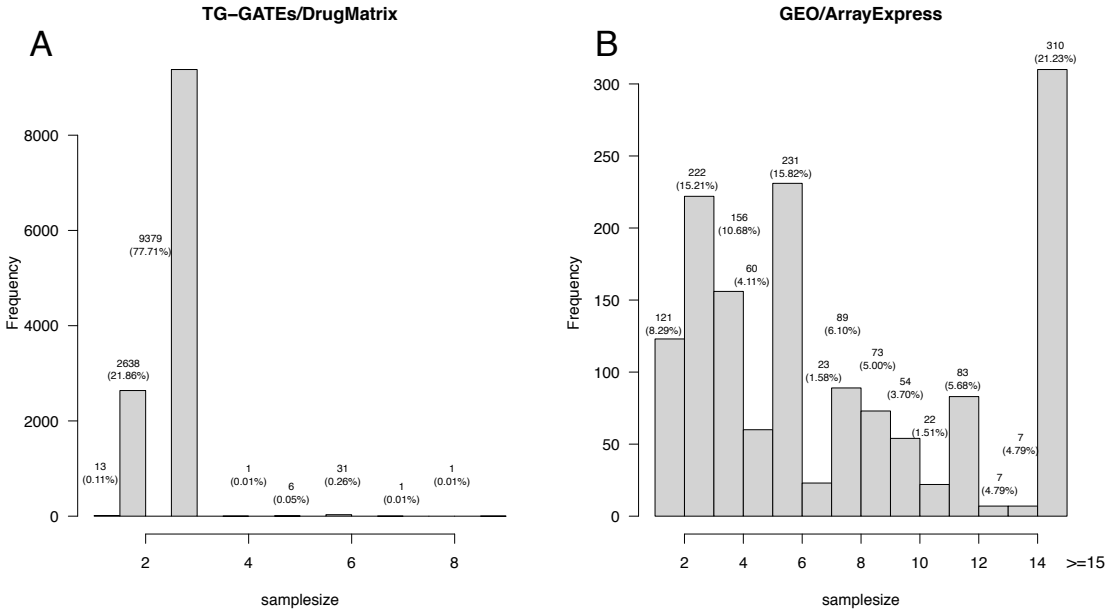


Figure 2.10 Histogram of sample size distribution among different PharmOmics signature databases.

(A) Sample size distribution of datasets from TG-GATEs and DrugMatrix. (B) sample size distribution of datasets from GEO/Arrayexpress database. Datasets with sample size <3/group were excluded from downstream analysis.

Chapter 3 Single-nucleus resolution of the adult *C. elegans* and its application to elucidate intergenerational response to alcohol exposure.

Introduction

In mammals, *in utero* exposure to alcohol is associated with Fetal Alcohol Syndrome Disorders (FASD) which hosts an array of well-characterized morphological, neurological, and reproductive deficits [103]. FASD is also known for its long lasting effect which passes through multiple generations. FASD caused structural and functional anomalies involving the reproductive system, the central nervous system, craniofacial morphogenesis, the heart, kidney, liver, and gastrointestinal system [103,104]. However, although *in utero* alcohol exposure clearly impacts the function of multiple organ systems, a comprehensive understanding of all organs, tissues, and cell types that are the most affected by alcohol remains lacking

In addition to impacting the health across multiple generations, mounting evidence in various model systems, such as mice, rats, *Drosophila*, and *C. elegans*, indicate that at least some exposure-related adverse reproductive and neurobehavioral features also extend beyond the F1 generation and are detectable in F3 progeny. For instance, a rat model of late gestational ethanol exposure demonstrated that not only F1, but also F2 and F3 individuals show an average increase in ethanol intake by 50%[105]. Moreover, preconception exposure is sufficient to cause increased alcohol intake in the offspring, together with signs of spatial learning and memory deficits [106]. Notably, the impact of *in utero* ethanol exposure on alcohol and substance use across several generations is also observed in the broader context of several established multi- and transgenerational models which showed impacts in various cognitive, behavioral, or physical endpoints[107,108].

C. elegans is a simplified but highly advantageous model for studying the multi-generation effects of alcohol exposure and is the most used invertebrate species for modeling FASD [109]. Direct exposure to ethanol causes a variety of dose- and duration-dependent outcomes similar to those elicited in mammals, such as growth and fertility impairments, neuro-depressive effects, increased alcohol preference, disinhibition, and withdrawal. These effects all involve similar cellular and neurological pathways in *C. elegans* and mammals[109]. *C. elegans* is also highly conserved in both ethanol's metabolism [110] and reproductive system, with two gonads opening into a common uterus where embryos initiate their development, provides suitable modeling for *in utero* exposure to alcohol.

Recently, the combination of single-cell RNA sequencing (scRNA-seq) technologies and the tractability of the model organism *C. elegans*, with its well-established differentiation lineages and stages has enabled the layering of transcriptional data with developmental events at both embryonic and larval (L2) stages [111,112]. This has led to the identification of gene expression changes that track the development of 502 preterminal and terminal cell types in embryos [112] and the characterization of 27 distinct cell types at various stages[111]. Furthermore, we and others have shown that *C. elegans* is also a powerful model for the study of multi- and trans-generational responses to environmental stimuli [11,113–116]. However, single cell transcriptomic approaches have yet to be applied to the characterization of environmental exposures, including alcohol, at the whole organism level and across generations.

Here, we used a single nucleus RNA sequencing (snRNA-seq) approach to maximize the isolation of diverse cell types that were previously too sensitive to be captured by traditional single cell RNA sequencing, including neuronal cell types with long processes as well as

syncytial organs such as the germline, followed by RNA-seq. We applied this approach to examine the transcriptional impact of parental (P0) exposure to physiologically low doses of ethanol that do not cause overt tissue toxicity on the F1 offspring (inter-generational exposure) as well as on the F3 generation (trans-generational exposure). Mechanical extraction and isolation of adult *C. elegans* nuclei followed by snRNA-seq identified a large number of distinct cell types that resolved into both known and novel functional categories. We also demonstrate that this powerful method can provide robust insights into the effect of intergenerational ethanol exposure at the tissue-, cell -specific resolution and identify the cells, and molecular pathways that are most impacted by such an exposure.

Material and methods

Culture conditions and strains

The sperm-defective strain *JK560 fog-1(q253)* which prevents self-fertilization was used for sequencing experiments. Worms were cultured on standard nematode growth medium (NGM) plates streaked with single colony OP50 *E. coli* and maintained at 20°C. The generation of worms to be collected for single-nucleus analysis was moved to 25°C at the L1 stage and grown at 25°C until collection. Worms were synchronized using a 10µm nylon mesh filter (NY1102500 EDM Millipore and SX00025000 EDM Millipore) which only allows L1 stage worms to pass through. The L1 worms were kept for 62hrs at 25°C and then washed with five rounds of M9 buffer and centrifuged at 1,300g for 1 minute to form a pellet. After washing, worms were spun in a rotator with 1mL of M9 for 30 minutes to remove OP50 from the worms' gut. The worms were then allowed to settle by gravity for 5 minutes and the final compact worm pellet volume was adjusted to 30µL.

C. elegans ethanol exposure

For ethanol exposure, a population of gravid adult worms and embryos was bleached. The embryos were then plated on standard OP50 seeded NGM plates and allowed to grow to the L4 stage (approximately 50hrs post bleaching). *C. elegans* were exposed for 48 hours in a liquid culture containing M9 buffer solution, standard OP50 bacteria (10mg/mL), and ethanol at a final concentration of 0.05% and 0.5% in 15mL conical tubes. Water is used for the control exposure. Following the either ethanol or water exposure, the progeny of the exposed P0 generation were obtained using gravid adult bleaching. The synchronized F1 and F3 population was then plated on standard OP50-seeded NGM plates and grown for 16hrs at 20°C at which time all F1 and F3 were at the L1 stage, and plates were then transferred to 25°C for 48hrs into adulthood. Finally, we also collected unexposed P0 group to support cluster identification.

Single-nucleus dissociation

All single-nucleus dissociation steps were done at 4°C. A compact 30µL pellet of adult JK560 *C. elegans* (approximately 4,000 worms) was transferred to a prechilled Dounce homogenizer (Z378623-1EA Sigma) and homogenized for 10 strokes with 400µL of ice cold FA lysis buffer (50mM HEPES/NaOH pH 7.5, 1mM EDTA, 0.1% Triton X-100, 150mM NaCl, protease inhibitor 0.5X (Roche 11697498001), RNase inhibitor 0.2U/µL (Thermo Fisher 10777019), and RNase free water). Worms were homogenized for 10 strokes in a 1.5mL Wheaton Dounce homogenizer and for an additional 20 strokes (350µL FA buffer) with an Eppendorf Dounce homogenizer (06-434 Fisher) in a corkscrew fashion. Between each set of 10 homogenization strokes, debris was pelleted at 100g for 1 minute and supernatant containing nuclei was collected and pooled in a fresh 1.5mL low retention microcentrifuge tube.

After homogenization, the pooled supernatant containing nuclei was centrifuged at 100g for 1 minute to pellet remaining debris. The top 900µL of supernatant with nuclei was transferred to a

fresh 1.5mL low retention microcentrifuge tube and washed twice with filtered PBS BSA-1% ((AM9624 Thermo Fisher) 0.22µm pressure filter (Thermo Scientific 03-377-26, Fisher SLGP033RS)). Nuclei were pelleted at 500g for 4 minutes. After the final wash, nuclei were resuspended in 750-850µL PBS-BSA 1% and then filtered using a 40µm Flowmi tip filter (BAH136800040-50EA Sigma Aldrich).

Nuclei integrity was verified by staining single-nuclei isolations with DAPI and observing the nuclei under a fluorescent microscope. The nuclei did not have a frayed appearance and were compact, indicating that they were intact. Nuclei extractions were performed at 4°C in a timely fashion to prevent cellular transcription during the dissociation process. On average, a total of 1,200 nuclei was obtained per batch of 4,000 worms. Flow cytometry was used to ensure optimal nuclei concentration (700-1200 events/µl).

Library Preparation and sequencing

Library preparation was performed by UCLA Technology Center for Genomics & Bioinformatics. Nuclei were isolated into single droplets and barcoded using the 10X Chromium Next GEM single cell 3' reagent kit. Followed by sequencing with 50bp long paired end reads with the NovaSeq 6000.

Single-nuclei dataset preprocessing

snRNA-seq reads were demultiplexed and aligned to the ENSEMBL ce10 *C. elegans* transcriptome to generate gene expression matrices using CellRanger (10x Genomics). The reference transcriptome was converted to accommodate pre-mRNA alignment by replacing “transcript” to “exon” in the annotation GTF file. Followed by CellRanger preprocessing, data were cleaned of ambient RNA by SoupX [117] and doublets by Diem [118]. Low-quality droplets

were filtered out according to the following criteria: 1) gene number less than 300 or more than 8000, 2) unique molecular identifier (UMI, indicates unique number of gene reads before PCR amplification process) count less than 500 or more than 40000, 3) mitochondrial RNA percentage > 15% per cell, and 4) ribosomal RNA >20% per cell.

Identification of cell clusters

R Seurat 3.1.5 [119] package was used for normalization, cell type identification, marker identification and batch effect correction for snRNA-seq data for all 31 sample groups. snRNA-seq data was log-normalized. The top 2000 variable genes were selected as representative features, then gene expression was corrected with UMI counts, mitochondrial gene percentage and ribosomal RNA percentage in preparation for clustering analysis. Canonical correlation analysis (CCA) was applied across different batches and treatment conditions to mitigate batch effects in cluster identification. Cell clusters were identified using the Louvain algorithm [120]. We included all treatment groups for unsupervised clustering since increased cell numbers increase the power to identifying smaller cell types [121]. Cluster specific gene markers were detected by Wilcoxon Rank Sum test [122] for cell type identification. To reduce biases from ethanol treatment in identifying cluster identifies, only cells from water and unexposed group were included unless cells from these two groups consist of less than 20% of the during cluster identification. Furthermore, for each cluster, the gene had to be expressed in at least 25% of the cells of the given cluster and there had to be at least a 0.25 log fold change in gene expression compared to other cells. Cell cluster identity was determined based on the overlap between highly expressed genes in each cluster with known cell type marker genes obtained from literature, Nematode Expression Pattern Database (NEXTDB) [123] as well as tissue enrichment analysis from wormbase [124] (**Table 3.1**). Log-normalized expression levels in t-SNE (t-distributed stochastic neighbor embedding) plot projections were used to visualize cell clusters in two

dimensions and dot heatmaps were used to visualize marker expression across different cell types.

Differential gene expression and pathway analyses

The Monocle [125] pipeline was used in order to identify DEGs across different cell types, generations and dose levels. Four different monocle models were created to assess DEGs in F1_0.05 (F1 of 0.05% ethanol exposure group), F1_0.5 (F1 of 0.5% ethanol exposure group), F3_0.05 (F3 ofr 0.05% ethanol exposure group) and F3_0.5 (F3 of 0.5% ethanol exposure group) condition. For each condition (generation and dose level), only cell types with more than 10 cells in each group were included. For genes expressed in more than 20% of cells for each cell type, a negative binomial model was fitted based on raw counts to normalize data, followed by fitting a generalized linear model to retrieve the dietary exposure effect with batch effects corrected as follows:

$$Gene\ expression = b1*batch+b2*ethanol+b3*gene\ count+b4*UMI\ count$$

The batch term was only included for conditions F1_0.05 and F3_0.05 where two batches of water and 0.05% ethanol samples were involved, for the F1_0.5 and F3_0.5 conditions this term was not used since only water and 0.5% ethanol samples of same batch were considered. The b2 coefficient was used to estimate dietary exposure effects. Statistical p-value was obtained using a likelihood ratio test against the null model where the dietary exposure term was not included. Significant DEGs were defined as genes with Benjamini & Hochberg corrected FDR < 0.05.

The DEGs were then subject to pathway annotation analysis. Only cell types with no fewer than

20 DEGs were included in this analysis. Gene ontology analysis was conducted using the clusterprofiler package [126] with the *C. elegans* gene ontology biological pathway (GOBP), molecular function (GOMF) database [85] and wormbase phenotype database [127]. Enrichment P values were corrected using the Benjamini–Hochberg method and a significance threshold of FDR < 0.05 was utilized. We also filtered significantly enriched pathways with less than 2 overlapped genes. For significantly enriched pathways, fold changes were calculated by averaging the fold changes of the pathway genes between treatment and control nuclei. For wormbase phenotypes, we also retrieved annotations for each phenotype by querying EBI OLS (ontology lookup service) API. Annotations from first level (nematode phenotype, physiology phenotype and anatomical phenotype) were not utilized since these terms were too general to provide any meaningful interpretation. We then selected the top 20 most common annotations and compared their proportion in original database with our enrichment results.

Euclidean distance-based measurement of cell type sensitivity

We used Euclidean distance to identify cell types that are sensitive to ethanol exposure. We only included cell types with at least 10 cells per treatment (ethanol and control) group per batch. For each gene, the expression distance between nuclei for water and ethanol treatment groups were squared and summed, then the square root was taken. In order to avoid potential biases caused by genes that are either highly expressed or not expressed, expression values were normalized to z-scores and only the top 1,000 expressed genes were used. To account for variability in expression characteristics for each cell type, treatment labels were permutated 1000 times to calculate the null distribution for each individual cell type. P values were calculated between the observed Euclidean distance and the null distribution for each cell type and adjusted with the Benjamini & Hochberg method.

To visualize the differences between water and ethanol treated nuclei for individual cell types, the fold change (FC) in the Euclidean distance of the ethanol treatment group compared with the water treatment group for each cell type was normalized by dividing the empirical Euclidean distance by the median Euclidean distance of the null distribution per cell type. The log10(FC) vs. -log10(adjusted p value) of each cell type was then plotted to visualize and rank the vulnerable cell types in ethanol treatment. For the 0.05% ethanol treatment group where two batches were involved, FDR and log10(FC) will be averaged.

Statistical Analysis

Unless otherwise mentioned, statistical analysis was conducted by R/3.5.1 (R Core Team).

Results

Single-nucleus preparation and snRNA-seq

Intact nuclei were isolated from adult *fog-1*(q253) *C. elegans* raised at the restrictive temperature of 25°C [128]. Since the focus of our study was to characterize adult tissue response to ethanol, we used a sperm-defective strain to prevent self-fertilization and the resulting crowding of our snRNA-seq data with embryonic cell types (see material and methods section). Briefly, worms were synchronized using 10µm filters and allowed to grow to day one of adulthood before mechanical nuclear extraction (**Figure 3.1A**). Single nucleus RNA-seq library preparation was performed using the 10X Genomics Chromium system followed by 50 PE sequencing using the Illumina Novaseq 6000 platform. In total, we generated transcriptomic data for 81,267 nuclei, each with more than 500 transcripts derived from 31 groups collected in 5 distinct batches. On average, 2,181 unique molecular identifiers (UMIs) and 992 genes were detected per nucleus with high sequencing depth (90.3% average sequencing depth).

snRNA-seq reads were demultiplexed and aligned to the ENSEMBL ce10 *C. elegans* transcriptome to generate gene expression matrices using CellRanger (10x Genomics) . To prevent the inclusion of empty droplets and to correct for ambient RNA contamination, we also applied Diem [118] and SoupX [117], respectively. Diem identifies empty droplets through modeling semi-supervised expectation maximization and outperforms other methods in snRNA-seq. Since Diem only filtered empty droplets and didn't correct expression levels for remaining droplets, we combined Diem with SoupX, which models snRNA-seq contamination levels and corrects expression levels for the remaining droplets. Using these stringent parameters, we obtained transcriptomic data from 46904 droplets representing a median of 2577 UMIs and 1266 genes (**Figure 3.1B**). A total of 31 discrete clusters were identified following batch/group effect correction using canonical correlation analysis (CCA) in Seurat v3 followed by the application of the Louvain clustering algorithm [120,129] (**Figure 3.1C**). Log-normalized expression levels in t-SNE (t-distributed stochastic neighbor embedding) plot projections were used to visualize cell clusters in two dimensions and dot heatmaps were used to visualize cell type specific marker expression across different cell types.

To facilitate unbiased cell type annotation, top differentially expressed markers (FDR<0.05) in each cluster was used to identify potential cell types. Cluster identity was identified based on matching top markers with Nematode Expression Pattern Database (NEXTDB) [123], literature annotation and tissue enrichment analysis from wormbase [124]. Highly confident clusters based on all evidences were shown in **Table 3.1**.

SnRNA-seq revealed organism-wide impact of inter-generational exposure to ethanol

We first applied snRNA-seq to identify the transcriptional outcome of a 48-hour (L4 to the end of day 1 of adulthood) parental exposure to two concentrations of ethanol (0.05% and 0.5%) or

water control on the F1 adult progeny. These doses were chosen to capture both low levels of ethanol easily reached in human populations and in pregnant women in particular [130]. We first compared cell-type proportions in the F1 following parental ethanol exposure and observed that broadly similar cell-type distributions were observed across all treatment conditions (**Figure 3.2A**). However, we observed a significant number of Differentially Expressed Genes (DEGs) (FDR<0.05) between ethanol treatment and water. Across all F1 clusters from the 0.05% ethanol exposure condition, we identified a total of 1,223 DEGs, including 583 consistently upregulated DEGs, 520 consistently downregulated DEGs, and 120 DEGs that were differentially up or down regulated in cluster specific ways (i.e. upregulated in some clusters but downregulated in other clusters). Surprisingly, compared to the 0.05% ethanol treatment, exposure to the higher 0.5% ethanol concentration resulted in fewer DEGs identified in the F1s: a total of 948 DEGs, including 430 uniformly upregulated DEGs, 407 uniformly downregulated DEGs, and 111 DEGs that were either up- or down-regulated in a cluster-specific fashion. A detailed Venn diagram shows that 35 DEGs were shared across all conditions (**Figure 3.2B**).

Pathway analysis of the union of all cell type specific DEGs revealed the enrichment of some Gene Ontology (GO) functional categories that align with alcohol metabolism such as carboxylic acid metabolic process driven by the presence in our DEG list of several aldehyde dehydrogenases (**Table S3.1**), which catalyze the final step of ethanol metabolism from acetaldehyde into acetate. Other GO categories that are shared across all exposure conditions include: small molecule biosynthetic process, which is largely enriched in fatty acid metabolism-related genes, and defense response. In addition, embryo development ending in birth or the egg hatching pathway was shared across F105 (F1 of 0.5% ethanol treatment group), F305 (F3 of 0.5% ethanol treatment group), F1005 (F1 of 0.05% ethanol treatment group) (**Table S3.2**).

These results indicated our low dose ethanol exposure treatment, though not causing changes strong enough on body proportion, still induced overall ethanol related responses in worm body.

SnRNA-seq revealed cell type specific impact of inter-generational exposure to ethanol

We also conducted cluster-specific DEG analysis to investigate cell type specific effects on F1. Cluster-specific analyses did not reveal significant changes in cell type proportions at any treatment dose nor generation (**Figure 3.2A**). However, cluster-resolved DEG analysis indicated clearly distinct transcriptional responses to parental ethanol exposure between cell types. While some genes were consistently up-regulated (*atp-6*, *nduo-6*) or down-regulated (*vit-5*) across all clusters between ethanol and water treatment, most DEGs showed cell type-specificity, highlighted by the low overlap of the top DEGs per cluster (**Figure 3.3**). To rank order the F1 clusters by sensitivity to ethanol exposure, we employed a Euclidean distance analysis [131,132], which estimates the degree of global transcriptomic shifts between exposure and control groups . Several clusters (1, 15, 29) with a strong germline identity showed the most significant transcriptomic shifts at the F1 generation under the 0.5% ethanol exposure condition (**Figure 3.4A**). Cluster 1 shows a gene signature suggestive of mid-pachytene; cluster 15 of late-pachytene where pachytene indicates third stage of the prophase of meiosis during germline development; cluster 29 of oocyte. Other cluster categories that appeared most affected included clusters related to muscle function such as cluster 2 (musculature) and cluster 17 (striated muscle cells). The degree of the transcriptomic shift was much less pronounced following 0.05% ethanol exposure compared to 0.5% ethanol, suggestive of a dose-dependent transcriptomic response across cell types.

We hypothesized that while most DEGs are cell type-specific, genes implicated in ethanol metabolism may show a more uniform response across clusters. Thus, we investigated the

expression of genes involved in ethanol metabolism, including 3 distinct alcohol dehydrogenases (*sodh-1*, H24K24.3, ZK829.7) and 10 aldehyde dehydrogenases (*alh-3*, *-4*, *-7* through *-13*), whose expression was detectable in our datasets (**Figure 3.5**). Contrary to our expectations, of the 13 genes examined, only 5 showed significant changes in expression (FDR < 0.05) and did so in a cluster-dependent fashion. For example, *sodh-1* was upregulated in cluster 13 and cluster 18 under the 0.05% exposure condition but was downregulated in cluster 2 and cluster 27 at 0.5%. Notably, the cell types showing the highest increase in ethanol metabolism genes were not the cell types that were the least sensitive to ethanol and vice versa, suggesting that upregulation of ethanol metabolism does not necessarily protect a tissue from the inter-generational impact of exposure.

We then inspected the top DEGs across cell types and conditions in order to identify the genes most significantly affected by ethanol (**Figure 3.4B**). Results included changes in ribosomal related genes (*rrn-3.1*, *rpl-10*, *rrn-2.1*, *epl-41.2*), cytochrome related genes (ctc-3, *ctb-1*, *ctb-2*) and vitellogenin related genes (*vit-2*, *vit-3*, *vit-1*, *vit-6*); all were commonly altered across different cell types. In hypodermis cells (cluster 6), 0.05% ethanol lead to downregulation of ribosomal, cytochrome and v vitellogenin genes. In mid pachytene (cluster 1), 0.5% ethanol resulted in downregulation of ribosomal and vitellogenin related genes and upregulation of cytochrome related genes; in early pachytene (cluster 12) 0.5% ethanol resulted in downregulation of ribosomal related genes and in late pachytene (cluster 15) it upregulated cytochrome related genes. Top DEG analysis showed 0.5% ethanol treatment affected germline cytochrome, ribosome and vitellogenin functions in F1 offspring.

Next, we interrogated the F1 data to identify whether the clusters with significant alterations (FDR<0.05) in our Euclidean analysis shared differentially enriched pathways (**Figure 3.4C** and

3.4D). GO analysis for molecular function and biological pathway revealed that ribosomal function and lipid metabolism under molecular pathways category and aging and reproduction under biological pathways category were significantly enriched across several clusters (**Figure 3.4C, Table S3.3 for 0.05% group and S3.4 for 0.5% group**). We also examined which DEG were responsible for the significantly enriched pathways related to reproductive function and lipid transportation, hence we investigated DEG overlapping patterns in pathway "embryo development ending in birth or egg hatching" that showed specificity in oocyte cluster (**Figure 3.6A**) and "lipid transporter activity" (**Figure 3.6B**) which was highly shared pathways across different germline clusters. Results indicated that 0.5% alcohol affected several germline genes such as *grd-5*, *mex-1* and *ani-1* in the oocyte cluster related to reproductive system pathway (**Figure 3.6A**) and vitellogenin related genes were responsible for lipid transporter activity (**Figure 3.6B**). Vitellogenin is also noted for functional relationship with post-embryonic development regulation as well as mediating intergenerational effects [133]. Hence, we think pathway enrichment analysis indicated low dose ethanol exposure affected several reproductive functions.

Since the 2 most sensitive clusters based our Euclidean Distance analysis displayed germline identity (**Figure 3.4A**), we examined whether reproduction-related phenotypes were significantly over-represented in our dataset. We utilized wormbase which documented different phenotypes and genes associated with different phenotypes. Through comparison of cell type specific DEGs with Wormbase phenotype database, we have shown several of top shared phenotypes across sensitive cell types were related to germline functions (**Figure 3.4E, Table S3.5** showed phenotype enrichment result for F1 0.05% group and Table S3.6 showed enrichment result for F1 0.5% group). To determine that germline functions enrichment is not due to higher proportion of germline related annotations in the Wormbase phenotype database, we further compared the

proportion of phenotype categories from our enrichment results with the Wormbase phenotypes. Our dataset had a significantly higher proportion of phenotypes from reproductive system development and cell development category among both treatment groups (**Figure 3.4F**). A closer look at overlapping genes related to “oocyte number decreased” (**Figure 3.6C**) and “early larval arrest” (**Figure 3.6D**) indicated ribosomal genes (*rps* and *rpl* family) as main overlapping genes which corroborated with top DEGs shared across different clusters under 0.5% ethanol treatment.

Characterization of the transgenerational impact of ethanol at single nucleus resolution

To capture the transgenerational effect of a P0 exposure to ethanol, we performed snRNA-seq on the F3 generation (**Figure 3.7**). Across all clusters of the F3s stemming from a P0 0.05% ethanol exposure, a total of 798 unique DEGs satisfied an FDR < 0.05: 366 were constantly upregulated, 369 were constantly downregulated, and 63 DEGs were differentially up or down regulated in cluster specific ways (i.e. upregulated in some clusters but downregulated in other clusters). For 0.5% ethanol, a total of 918 unique DEGs were identified at an FDR < 0.05, 402 were upregulated, 422 were downregulated, and 94 DEGs were differentially up or down regulated in cluster specific ways.

Cluster specific analysis also revealed cell type specific responses to ethanol in the F3 generation. Sensitivity analysis by Euclidean distance indicated more clusters had significant Euclidean distance alteration on 0.5% ethanol exposure compared to 0.05% (**Figure 3.7A**). In addition, sensitivity analysis also revealed cell types that are sensitive and insensitive to ethanol treatment, with mid-pachytene germline cluster (cluster 1) being the most sensitive cell type.

A closer look at top shared DEGs (**Figure 3.7B**) among cell types with significantly altered sensitivity revealed genes from the ATP pathway (*atp-6*, *ctc-1*, *ctc-2*, *ctc-3*) and collagen related genes. We also observed top DEGs were generally downregulated for both concentrations of ethanol exposure in the F3 generation, compared to upregulating effects in F1 (**Figure 3.4B**) Pathway analysis using GO biological pathway and molecular function enrichment analysis (**Figure 3.7C** and **3.7D, Table S3.7** for significant pathway enrichment in F3 under 0.05% **and S3.8** for significant pathway enrichment in F3 under 0.5%) indicated significant (FDR < 0.05) alteration of lipid transporter activity in mid pachytene (cluster 1), upregulation of muscle related pathways in body wall musculature (cluster 2) and early pachytene cells (cluster 12). Examination of overlapping DEGs in pathways of interest indicated vitellogenin related genes (*vit-1 to vit-6*) as major DEGs related to “lipid transporter activity” (**Figure 3.8A**) which showed long lasting ethanol exposure effects under 0.5% involving germline and muscle cell types. We also examined another commonly shared F3 pathway “structural constituent of cuticle” among different cell types (**Figure 3.8B**) and showed collagen gene alteration as major factor related to this pathway.

Finally, we also repeated enrichment analysis with wormbase phenotype database through documented genes associated with different phenotypes (**Figure 3.7E**). Result indicated showed strong alterations in different phenotypes involving narrowed rachis structure, pachytene region organization alteration and organism morphology variation in 0.5% ethanol treated F3, although they were not as consistent as 0.5% ethanol treated F1 (**Figure 3.4E**). We have identified “pachytene region organization variant downregulated” in cluster 1 which is also found in F1, indicating potentially lasting ethanol exposure effect in F3. We also identified body wall muscle morphology variant and “paralyzed” phenotypes altered in cluster 2 (body wall muscle), which might indicate a potential muscle effect of 0.5% ethanol exposure. A closer look

of overlapping genes related to enriched phenotype “germ cell compartment morphology variant” in mid pachytene (cluster 1) showed ribosomal genes (*rps* and *rpl* family) were main overlapping genes (**Figure 3.8C**), which is also found in overlapping genes related to “oocyte number decreased” under F1 (**Figure 3.6C**). We also found that the phenotype “paralyzed” (**Figure 3.8D**) which was altered in germline related clusters (1, 12) was related to actin (*act-1*), Tropomyosin (*lev-11*) and *unc* gene family. A survey of phenotype annotation category still indicated phenotypes with germline related functional annotation were altered compared to overall database proportion (**Figure 3.7F**). Through these DEG, pathway and phenotype analyses, I have shown that ethanol treatment can induce strong genetic alterations for up to 3 generations and involve ribosomal function and lipid transportation function.

Discussion

We have developed a single-nucleus RNA-seq approach in the adult *C. elegans* nematode that identifies a large number of known cell types while providing in-depth transcriptional information about them, and can be applied to achieve a nuanced understanding of physiological responses to environmental cues.

Our methods generate robust numbers of genes per nucleus, even when compared to mammalian studies [134]. Additionally, single-cell and single-neuron sequencing are very well correlated and single-nucleus sequencing has the advantage of removing confounding transcripts from the mitochondrial genome [111].

Our approach has several limitations. By working in a *fog-1* mutant background, we were not able to identify sperm cells, a necessary trade-off to avoid the production of embryos. It is possible that the absence of *fog-1* alters the transcriptional landscape of the germline. However,

fog-1(q253) was chosen specifically because of the normal morphology and staging of the hermaphrodite germline in *fog-1* mutants [128] which we validated by DAPI staining and cluster analysis.

Nonetheless, we were successful in *in silico* validating some findings of our ethanol studies through documented gene-phenotype relationships, such as the “oocyte number reduction” phenotype in F1 and “germ cell compartment variant” phenotype in F3, showing that ethanol affected ribosomal and reproductive phenotypes across generations. These results demonstrate that a low dose of ethanol at the parental generation has a significant impact on the offspring’s oocyte function and embryonic viability. Direct ethanol exposure has been known to cause aneuploidy in mammals [135–137], however the underlying molecular mechanisms as well as cell type specificity underlying the effect of transgenerational exposure has not been fully investigated. Here, we reveal the effects of low-dose ethanol on the reproductive system across multiple generations i) organism-wide response involving alcohol metabolism regulation through union of all DEGs ii) alteration of global gene expression pattern through Euclidean distance based sensitivity analysis iii) alteration of several molecular pathways crucial for proper germline development, including mitochondrial function [138,139], ribosomal functions [140,141] and lipid transportation[142,143] iv) alteration of genes connected with germline dysfunction phenotypes. Together, snRNA-seq of the adult *C. elegans* represents a powerful method for the comprehensive identification of cell types in the nematode and for probing the multigenerational transcriptional impact of physiological and environmental changes.

Tables

Table 3.1. Putative cell type annotation based on literature, NEXTDB and tissue enrichment analysis (TEA). Germline related clusters are bold.

cluster number	putative identify	markers	Source	TEA
1	**mid-pachytene**	*rad-50/M116.5/ppw-2/msh-6*	NEXTDB	germ line
2	body wall musculature	*myo-3/tnt-2/unc-27*	NEXTDB	striated muscle
3	spermatheca	*snf-9/ssp-37/ule-3*	NEXTDB	spermatheca
4	uterine epithelial cells	*pes-23/ifa-1/C35B1.4*	NEXTDB	hermaphrodite
6	hypodermis	*sqt-3*	NEXTDB	epithelial system
12	**early pachytene**	*syp-3/Gld-1/Glp-1/*	NEXTDB	Psub1 Embryonic founder cell
13	pharyngeal gland cells	*phat-3/Y8A9A.2/phat-8*	NEXTDB	pharynx
15	**late- pachytene**	*Y9D1A.1/clp-3/daf-2*	NEXTDB	germ line
17	striated muscle cells	*pde-4/dyb-1*	NEXTDB	striated muscle
18	pharyngeal muscle cells	*pqn-31/tnc-2/pqn-94*	NEXTDB	pharynx
19	proximal gonadal sheath	*skpo-1/tbh-1*	NEXTDB	somatic gonad
20	amphid and phasmid sheath cells	*F53F4.13/T02B11.3*	Bacaj et al[144].	amphid sheath cell
23	**mitotic zone germline**	*wdr-5.1/alg-5/F11E6.7*	NEXTDB	AB Embryonic founder cell
25	distal gonadal sheath	*skpo-1/tbh-1/lim-7*	NEXTDB, Killian et al[145].	somatic gonad
28	coelomycete cell	*cup-4*	Cao et al[111].	coelomic system
29	**oocytes**	*Y9D1A.1/clp-3/emb-5*	NEXTDB	Abprapap embryonic cell

Table S3.1. Differentially enriched pathways after ethanol treatment based on union of all cell type specific DEGs across different conditions

Table S3.2. Shared differentially enriched pathways after ethanol treatment based on union of all cell type specific DEGs across different conditions

Table S3.3. All significantly enriched GOBP and GOMF pathways in F1 after 0.05% ethanol treatment

Table S3.4. All significantly enriched GOBP and GOMF pathways in F1 after 0.5% ethanol treatment

Table S3.5. All significantly enriched wormbase phenotypes in F1 after 0.05% ethanol treatment

Table S3.6. All significantly enriched wormbase phenotypes in F1 after 0.5% ethanol treatment

Table S3.7. All significantly enriched GOBP and GOMF pathways in F3 after 0.05% ethanol treatment

Table S3.8. All significantly enriched GOBP and GOMF pathways in F3 after 0.5% ethanol treatment

Table S3.9. All significantly enriched wormbase phenotypes in F3 after 0.05% ethanol treatment

Table S3.10. All significantly enriched wormbase phenotypes in F3 after 0.5% ethanol treatment

Figures

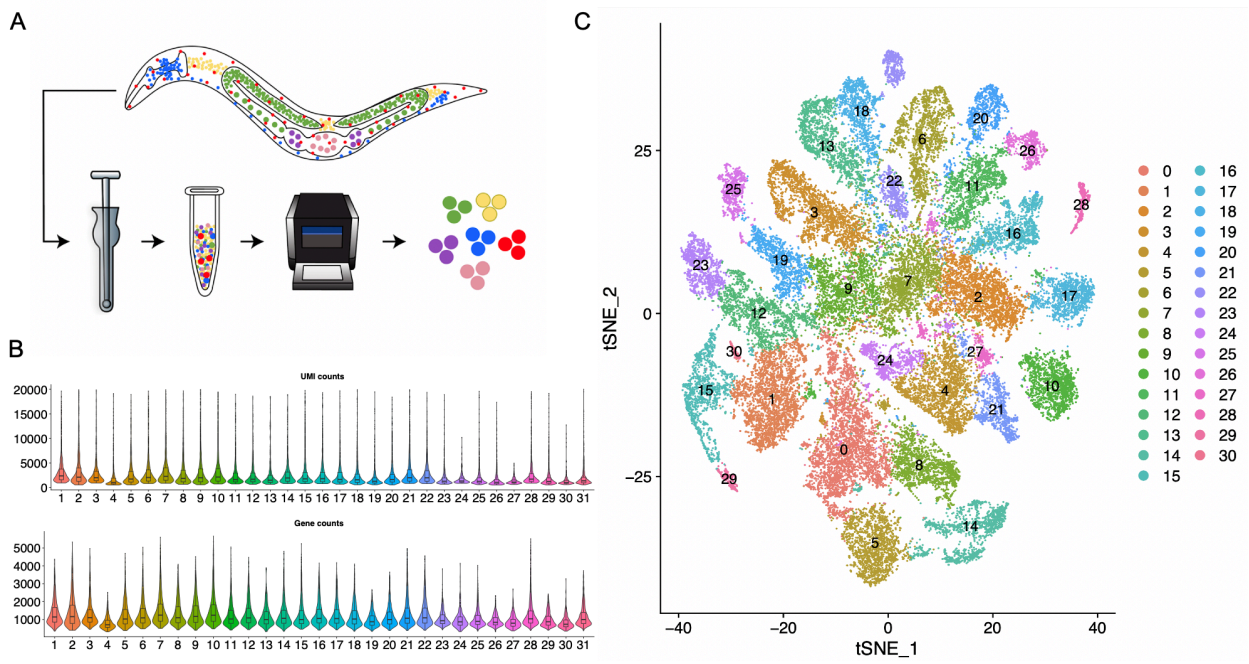


Figure 3.1. snRNA-seq identifies distinct cell and functional categories in the *C. elegans* adult hermaphrodite.

A. Experimental flow for single-nucleus isolation and snRNA-seq. Worms were homogenized in Eppendorf, followed by nuclei extraction and sequencing. B. Gene counts across 9 single-nuclei studies. Samples 1-2 : group 1, unexposed; samples 3-10: group 2, water F1 (samples 7,8), water F3 (samples 9,10), 0.05% ethanol F1 (samples 3,4), 0.05% ethanol F3 (samples 5,6); samples 11-13: group3, non-treated; samples 14-22: group 4, water F3 (samples 20-22), 0.05% ethanol F3 (samples 14-16), 0.5% ethanol F3 (samples 17-19); samples 23-31: group 5, water F1 (samples 29-31), 0.05% ethanol F1 (samples 23-25), 0.5% ethanol F3 (samples 26-28). Groups indicating samples collected from different batches. C. t-distributed stochastic neighbor embedding (t-SNE) plot of cells from all the samples with clustering labels identified by combination of unsupervised Louvain clustering.

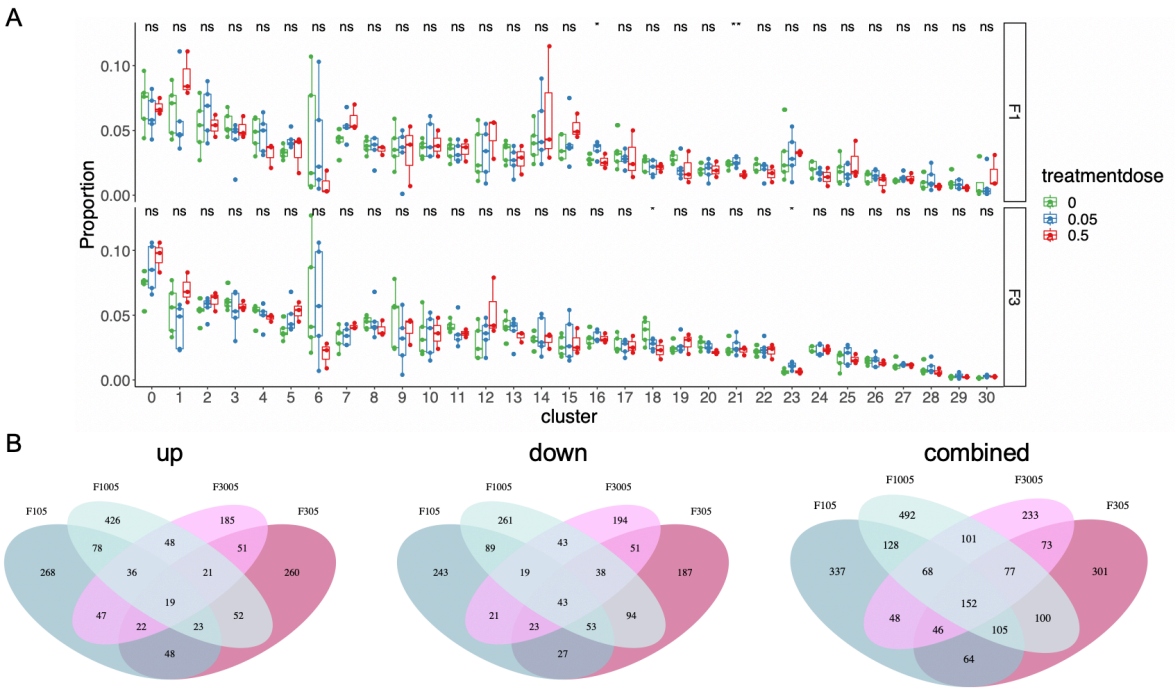


Figure 3.2. Organism-wide multi- and trans- generational low-dose effect of ethanol on *C. elegans*.

A. Proportion distribution plot of all F1 and F3 samples colored by different treatment dose, each dot represents one sample. X-axis indicates cluster number assigned by Louvain clustering and Y-axis indicates cells of that cluster divided by all cells from that specific sample. All conditions were non-significant based on the post-hoc Tukey statistic. B. Venn diagram based on the union of DEGs across all the cell types (A) Upregulated DEGs only (B) Downregulated genes only (C) all DEGs. Key: F105 (0.5% ethanol F1 generation), F1005 (0.05% ethanol F1), F305 (0.5% ethanol F3), F3005 (0.05% ethanol F3).

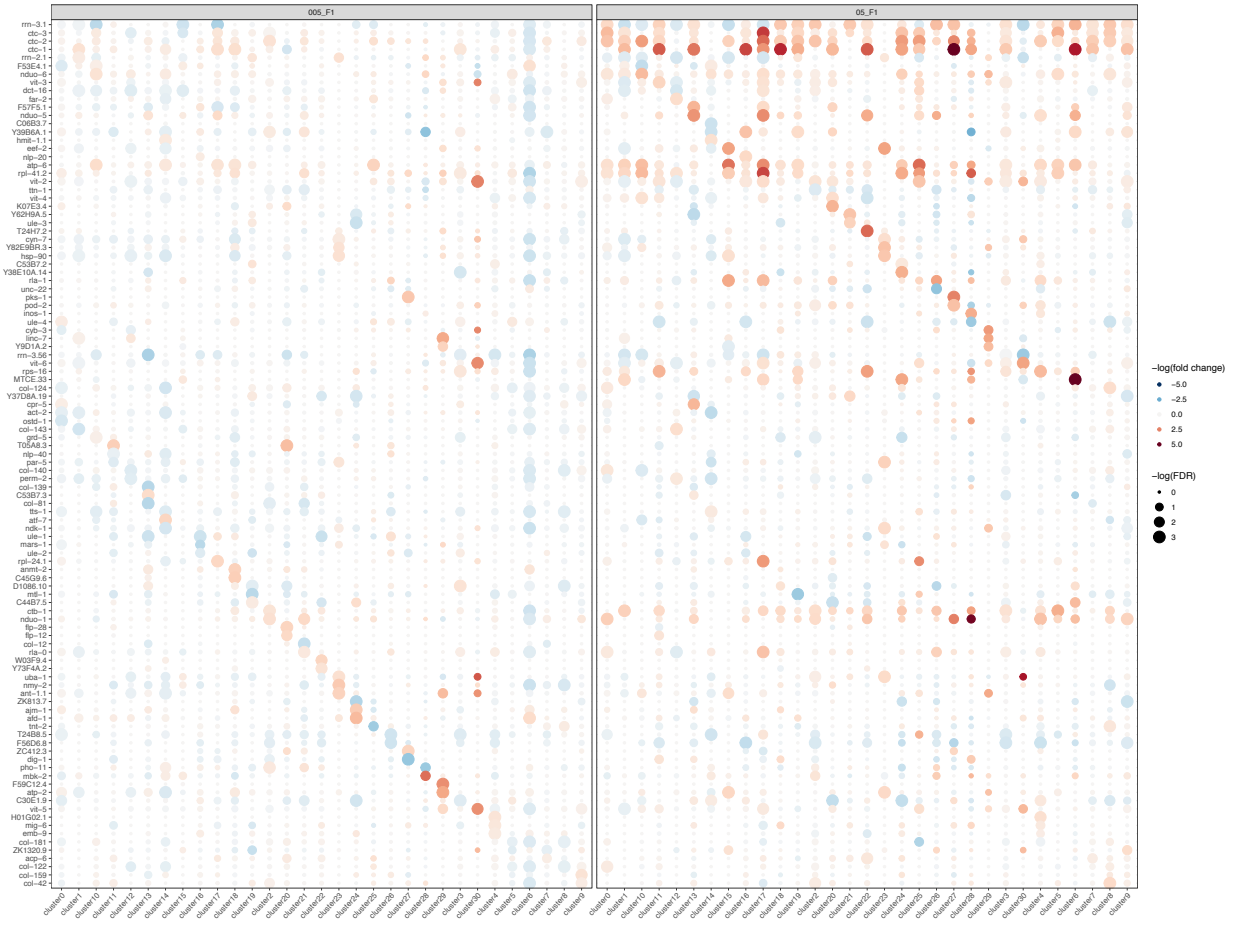


Figure 3.3. Dot heatmap of top 3 differential expressed genes across clusters in F1.

X-axis indicates different cell types and Y-axis indicates top5 differentially expressed genes after ethanol treatment across clusters ranked by monocle based FDR. The plot is divided by doses (0.05% on the left and 0.5% on the right). The size of the dot is correlated to –log(FDR) of differential expression p-value and the color is representing direction and scale of fold change (upregulation is shown in red and downregulation is shown in blue).

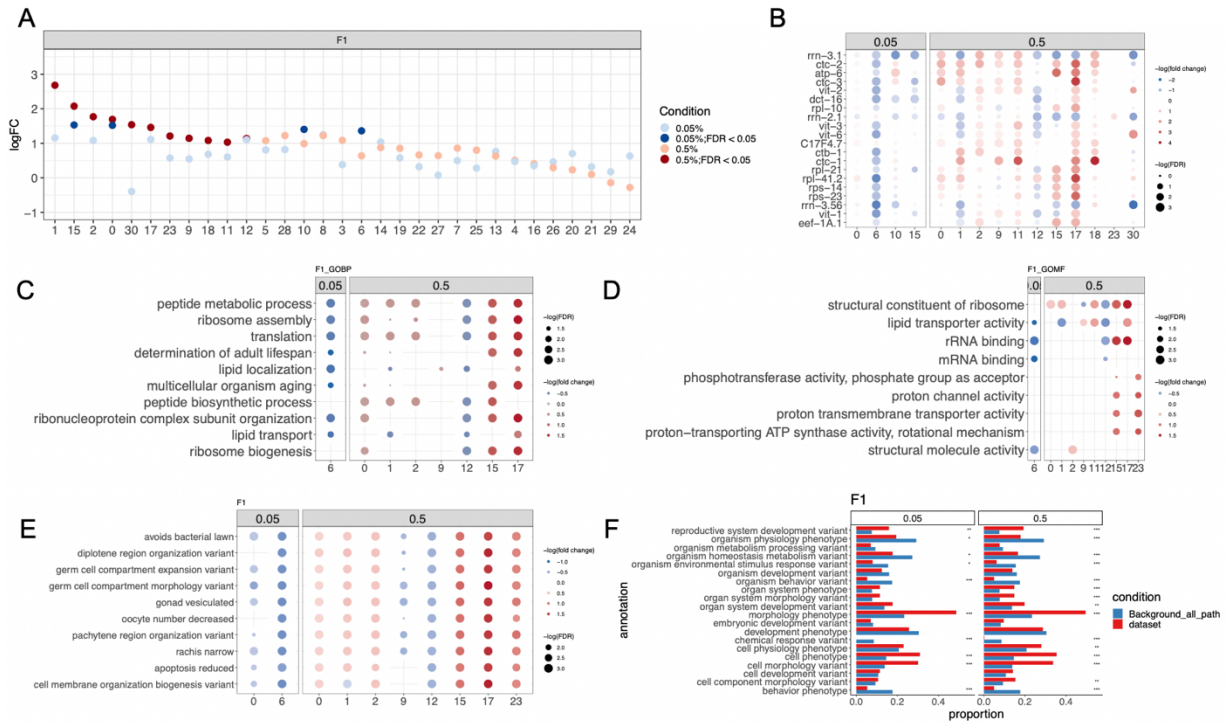


Figure 3.4. Analysis of ethanol exposure effects on first generation (F1).

A. Euclidean distance sensitivity analysis of all the cell clusters. X-axis indicated cluster number and y-axis indicates log fold change compared to Euclidean distance obtained by permuting treatment labels 1000 times. Significance was accessed based on comparing Euclidean distance against 1000 random permutated labels. B. Dot heatmap of top shared DEGs across cell types with significantly altered Euclidean distance metric. Dot size corresponded to -log(FDR) and dot color corresponded to -log(fold change) retrieved by differential gene analysis, only significant DEGs were plotted. C. Dot heatmap of top shared gene ontology biological pathway (GOBP) pathways across cell types with significantly altered Euclidean distance metric. D. Dot heatmap of top shared gene ontology molecular function (GOMF) pathways across cell types with significantly altered Euclidean distance metric. Dot size corresponded to -log(FDR) obtained from enrichment analysis and dot color corresponded -log(median fold change) of overlapping genes in each pathway. E. Dot heatmap of top shared wormbase phenotype across cell types with significantly altered Euclidean distance metric. Dot size corresponded to -

log(FDR) obtained from enrichment analysis and dot color corresponded -log(median fold change) of overlapping genes in each pathway. F. Bar plot showing the proportion of top wormbase phenotype annotations from all enriched pathways (“dataset”) and wormbase phenotype database (“Background_all_path”). For each wormbase phenotype from the original database we retrieved the corresponding wormbase phenotype annotations by querying EBI OLS API, followed by selecting the top 20 shared phenotypes. Annotations from the first level (nematode phenotype, physiology phenotype and anatomical phenotype) were too general to form meaningful interpretations. Proportions were calculated based on the proportion of annotations among all enriched pathways (“dataset”) and the wormbase phenotype database (“Background_all_path”). Fisher’s exact test was used to compare proportions between the two conditions in each annotation category.

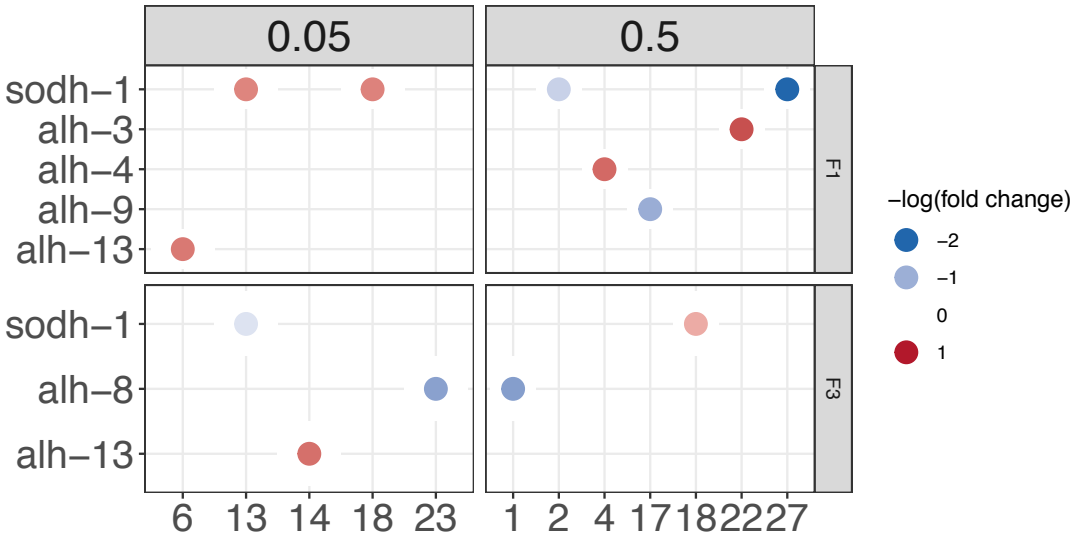


Figure 3.5 Dot heatmap of ethanol metabolism related genes across different clusters.

Different panels indicated dose and generation. Only DEGs (ie FDR < 5%) were plotted as dot.

Color indicated –log10(fold change).

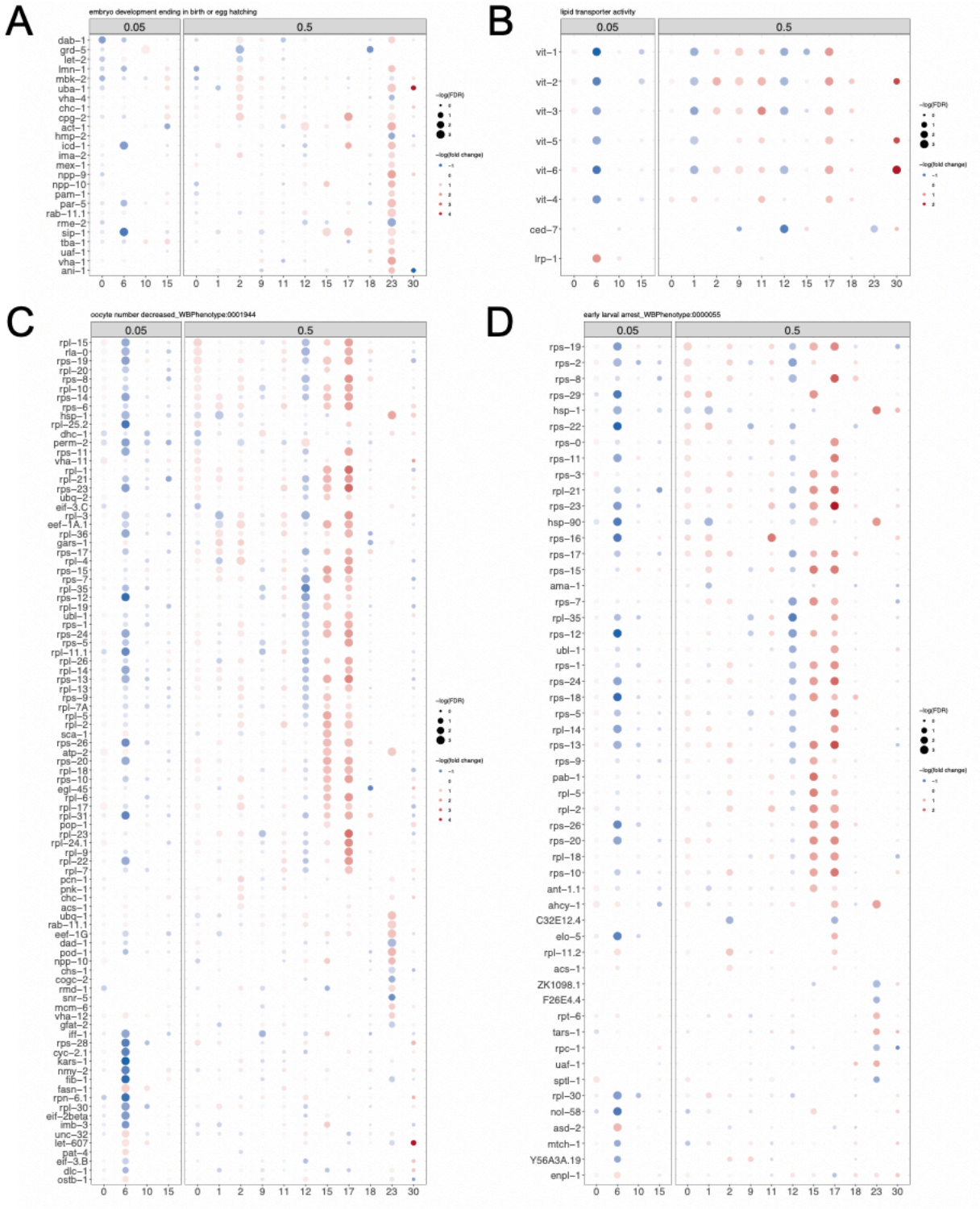


Figure 3.6. Dot heatmap of overlapping F1 DEGs under selected significantly enriched pathways and phenotypes across different clusters

A. DEGs overlapped with pathway “embryo development ending in birth or egg hatching” B. DEGs overlapped with pathway “lipid transporter activity” C. DEGs overlapped with phenotype “oocyte number decreased” D. DEGs overlapped with phenotype “early larval arrest”. DEGs (ie FDR < 5%) were plotted as dot and Color indicated –log10(fold change).

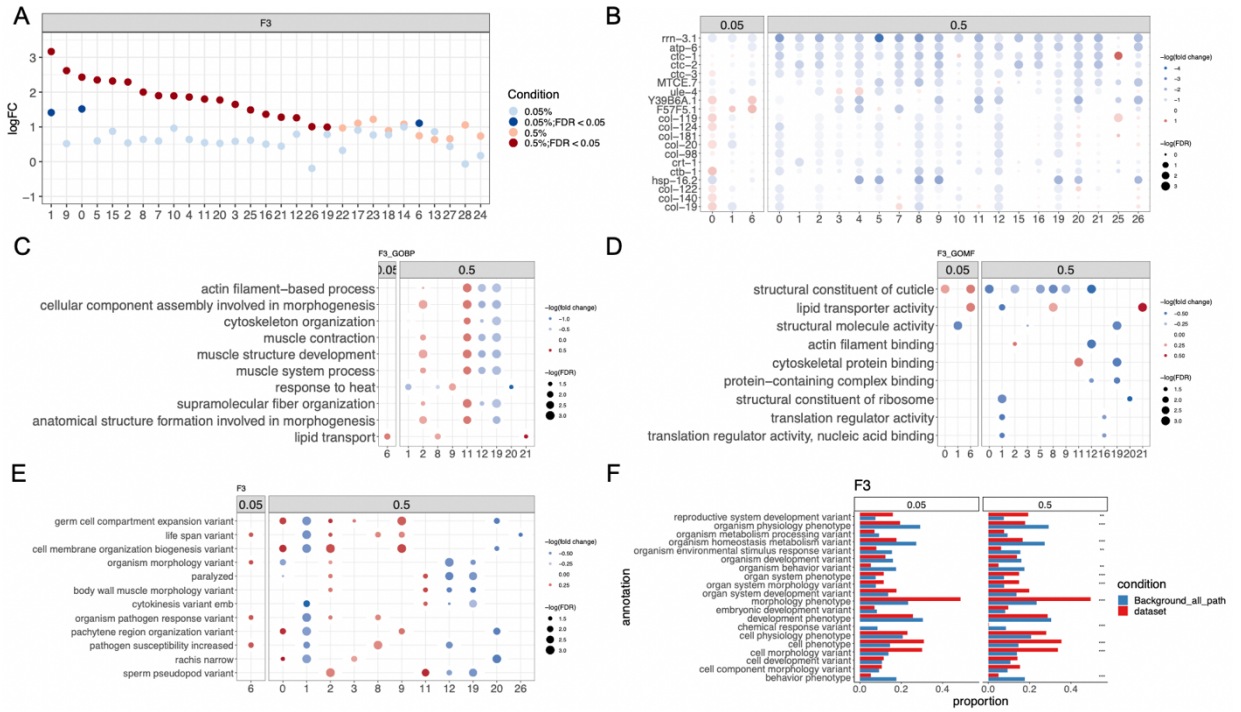


Figure 3.7. Analysis of ethanol exposure effects on the third generation (F3).

A. Euclidean distance sensitivity analysis of all the cell clusters. X-axis indicated cluster number and y-axis indicated log fold change compared to Euclidean distance obtained by permuting treatment labels. Significance was accessed based on comparing Euclidean distance against 1000 random permutated labels. B. Dot heatmap of top shared DEGs across cell types with significantly altered Euclidean distance metric. Dot size corresponded to -log(FDR) and dot color corresponded to -log(fold change) retrieved by differential gene analysis, only significant DEGs were plotted. C. Dot heatmap of top shared gene ontology biological pathway (GOBP) pathways across cell types with significantly altered Euclidean distance metric. D. Dot heatmap of top shared gene ontology molecular function (GOMF) pathways across cell types with significantly altered Euclidean distance metric. Dot size corresponded to -log(FDR) obtained from enrichment analysis and dot color corresponded -log(median fold change) of overlapping genes in each pathway. E. Dot heatmap of top shared wormbase phenotype across cell types with significantly altered Euclidean distance metric. Dot size corresponded to -log(FDR)

obtained from enrichment analysis and dot color corresponded -log(median fold change) of overlapping genes in each pathway. F. Bar plot showing the proportion of top wormbase phenotype annotations from all enriched pathways (“dataset”) and wormbase phenotype database (“Background_all_path”). For each wormbase phenotype from original database we retrieved corresponding wormbase phenotype annotations by querying EBI OLS API, followed by selecting top 20 shared phenotypes. Annotations from first level (nematode phenotype, physiology phenotype and anatomical phenotype) since these terms were too general to make interpretations. Proportion were calculated based on proportion of annotations among all enriched pathways (“dataset”) and wormbase phenotype database (“Background_all_path”). Fisher exact test was used to compare proportions between two conditions in each annotation category.

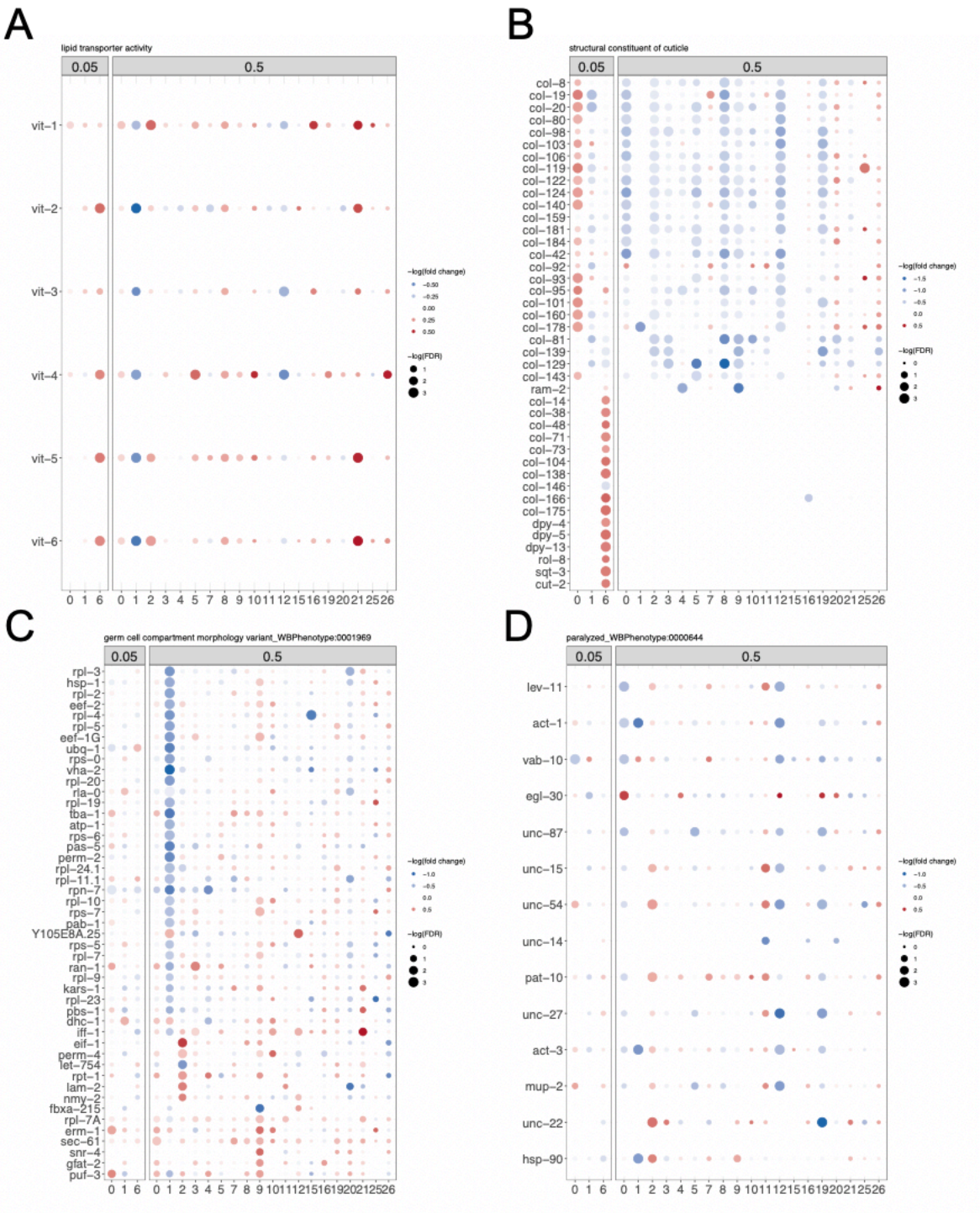


Figure 3.8 Dot heatmap of F3 DEGs overlapped in selected significantly enriched pathways and phenotypes across different clusters, faceted by dose

A. DEGs overlapped with pathway “lipid transporter activity” B. DEGs overlapped with pathway

“structural constituent of cuticle” C. DEGs overlapped with phenotype “germ cell compartment morphology variant” D. DEGs overlapped with phenotype “paralyzed”. DEGs (ie FDR < 5%) were plotted as dot and Color indicated –log10(fold change).

Chapter 4 Multi-tissue Single-Cell Analysis Reveals Differential Tissue, Cellular, and Molecular Sensitivity Between Fructose and High Fat High Sucrose Diets

Introduction

Metabolic syndrome (MetS) is a common complex disorder comprising diverse symptoms such as hyperlipidemia, obesity, insulin resistance, and hypertension. Previous research has highlighted the importance of genetics and lifestyle in the development of MetS[146–148]. In particular, modern diets rich in high fat, sucrose, and/or fructose content, are crucial risk factors for MetS. The rodent high fat high sucrose diet (HFHS) resembles the classic Western diet[149], whereas the high fructose diet has only become more heavily studied in the context of MetS over the past decade in both epidemiological studies[150,151] and molecular mechanistic investigations[152–154] due to its presence in corn syrup, commonly used addictive in soda and snacks. The two MetS risk diets possess different metabolic characteristics[10]. The metabolism of HFHS involves the breakdown of sucrose into glucose (processed by hexokinase) and fructose (metabolized by ketohexokinase (KHK)) as well as fat metabolism (lipolysis and fatty acid oxidation) and HFHS diet primarily promotes adiposity and insulin resistance. In contrast, the fructose diet mainly requires intestinal uptake of fructose by GLUT5[155] and clearance via KHK in both the small intestine and liver[156,157]. Fructose diet was mainly reported to cause dyslipidemia and hyperglycemia[10]. Additionally, mouse strains with different genetic background respond differentially to the two diets[10,149]. For example, C57BL/6J (B6) mice are highly sensitive to HFHS-induced weight and adiposity gain but are relatively resistant to fructose-induced weight and adiposity gain. DBA/2J (DBA) mice, on the other hand, show the reverse trend. These differences support that MetS induced by different risk diets are likely mediated by different mechanisms and represent different MetS subtypes.

Numerous tissues are involved in the pathophysiology of MetS, including the hypothalamus (central control of energy homeostasis and metabolism)[158], liver (lipid and glucose metabolism)[159], adipose tissue (energy storage, immune and endocrine functions)[160], and the small intestine (nutrient absorption and gut microbial interactions)[159,161]. In line with this, the contribution of nutrition source on MetS is expected to involve multi-tissue multi-cellular mechanisms, which is supported by studies for both HFHS[153,162,163] and fructose[13,164,165] diet studies. To date, studies on the role of HFHS and fructose in MetS were conducted mainly in bulk tissues, which mask cell-type specific responses of each of the heterogenous tissues. A holistic understanding of MetS pathogenesis at cell type resolution has been limited by the paucity of systemic investigation across different tissues and cell types.

In the present study, we employed single-cell RNA-seq (scRNA-seq) to characterize cell-type specific transcriptomic changes across different tissues relevant to metabolic regulation, including the hypothalamus, liver, adipose tissue, and small intestine. By simultaneously investigating all cell types across multiple tissues in two diet-induced MetS models, our study offers a better understanding of cell-type specific responses found in both HFHS- and fructose-induced MetS and reveals potential mechanistic differences conferred by the different MetS diets. Our results pinpoint sensitive tissues, cell types, cell-type specific pathways, and cell-cell communications networks across tissues altered in MetS that are either uniquely responsive to fructose and HFHS, or shared between diets. Our multi-tissue single cell data and cross-tissue/cell type gene networks for two common risk diets for MetS serve as a rich resource for the metabolism community.

Methods

Diet-induced MetS models

As shown in the overall study design in **Figure 4.1A**, eight-week-old B6 male mice were treated with chow diet with water (control group, n=6), chow diet with water containing 15% fructose mimicking the prevalent uptake route of fructose through liquid drinks (Fructose group, n=6), and a high fat high sucrose diet containing 16.8% kcal protein, 51.4% kcal carbohydrate, and 31.8% kcal fat with water mimicking the Western diet (Research Diets- D12266B, NJ, USA; HFHS group, n=6) for 11 weeks. The diets were changed once a week and water or fructose solutions were changed twice a week. Food and liquid were accessed *ad libitum* and the intake amount for each was measured. After 11 weeks, mice were sacrificed, and tissues including hypothalamus, gonadal adipose tissue, whole liver, and the small intestine (including duodenum, jejunum, and ileum) were collected for scRNAseq using Drop-seq[121].

Phenotypic characterization of MetS

Body weight was measured weekly using a scale and body mass composition (fat mass, lean mass) was determined using a Bruker Minispec Mq10 NMR analyzer (Bruker BioSpin, Fremont, CA). To measure glucose tolerance, animals were fasted overnight by transferring mice to clean cages before intraperitoneal glucose tolerance test (IPGTT) at the 4th and 10th week of HFHS or fructose treatment. For IPGTT, 20% glucose was injected intraperitoneally into each mouse at 2g glucose/kg body weight. Blood glucose levels from tail vein were measured at 0, 15, 30, 90, and 120 min after glucose injection using an AlphaTrak portable blood glucose meter (Abbott Laboratories, North Chicago, IL, USA). Area under the curve (AUC) across 0-120min was calculated to measure the degree of glucose intolerance. For lipid profiling, blood samples from fasted mice were collected through retroorbital bleeding at 11 weeks before sacrifice. Plasma total cholesterol (TC), high-density lipoprotein (HDL), triglycerides (TG), glucose, and insulin were measured by enzymatic colorimetric assays at the UCLA GTM Mouse Transfer Core as previously described[166]. Low-density lipoprotein (LDL) was calculated as LDL = TC -

HDL - (TG/5). Differences in the phenotypes between treatment groups were assessed using One-way ANOVA, except for body weight, relative fat mass and relative lean mass across various timepoints where two-way ANOVA was used.

scRNA-seq, data preprocessing, and quality control

Tissues were collected, cells were lysed, suspended and barcoded using optimized protocols for hypothalamus[121], gonadal adipose stromal vascular fraction (SVF)[167], small intestine[168], liver non-parenchymal cells (NPCs)[169] and hepatocytes (http://mouselivercells.com/procedure.html) and version 3.1 of the online Drop-seq protocol (http://mccarrolllab.com/download/905/). For each tissue, n=3-5 independent biological replicates were processed from different mice in each diet group. As tissue dissociation and the Drop-seq procedure were time-consuming, only n=2/group was processed each day for each tissue. Sequencing libraries were prepared according to the Drop-seq protocol and sequenced using Illumina HiSeq4000 with 50bp paired-end setting at 20-50k reads/cell.

The fastq files of the Drop-seq sequencing data were processed to digital expression gene matrices using the Drop-seq pipeline (https://github.com/Hoohm/dropSeqPipe) and DropEST[170]. Fastq files were converted to BAM format and cell and molecular barcodes were tagged. Reads corresponding to low quality barcodes were removed. Next, any occurrence of the Illumina adapter sequence or polyA tails found in the reads were trimmed. These cleaned reads were converted back to fastq format to be aligned to the mouse reference genome mm10 using STAR-2.5.0c[171]. Reads which overlapped with exons and introns were tagged using a mm10 RefFlat annotation file with the tool DropEST[170]. The transcript counts of each cell were normalized by the total number of Unique Molecular Identifiers (UMIs) for that cell. These values were then multiplied by 10,000 and log transformed. Single cells were identified from using the

following thresholds: gene counts from 200 to 3000, mitochondrial transcripts less than 10%, and ribosomal transcripts less than 20%. Samples that did not meet the quality control cutoffs were removed from downstream analysis, yielding quality data for hypothalamus (n=2 control, 2 fructose, 2 HFHS samples, single batch), gonadal adipose SVF (n=2 control, 2 fructose, 2 HFHS samples, single batch), small intestine (n=2 control, 2 fructose, 2 HFHS samples, single batch), liver NPCs (n=3 control, 5 fructose, 4 HFHS samples, two batches)[169] and hepatocytes (n=2 control, 4 fructose, 4 HFHS samples, two batches).

scRNA-seq data analysis

R Seurat 4.0.2[172] package was used for gene expression normalization and differential expression analysis. Briefly, scRNA-seq data were first log-normalized. Canonical correlation analysis (CCA) was applied across different samples to mitigate technical variation in cell cluster identification. Cell clusters were identified based on shared nearest neighbor (SNN) modularity optimization[120] and visualized using t-SNE, t-SNE was used to better visualize subcluster patterns (hepatocytes, neurons, enterocytes, adipose progenitor cells). Wilcoxon signed rank test based on the Seurat Findmarker function was used to identify highly expressed marker genes for each cluster. Cell cluster identity was determined based on the overlap between highly expressed marker genes in each cluster with known cell type marker genes obtained from the literature for liver, adipose SVF, small intestine and hypothalamus (full marker list and references in **Table S4.1)**.

Euclidean distance-based measurement of cell type sensitivity

Euclidean distances between gene expression profile groupings were used to identify cell types that were sensitive to each diet[131]. Only cell types with > 10 cells in each of the MetS diet group and control group were included. Euclidean distance was calculated by taking the sum of the

square of gene expression differences between groups of cells across genes, followed by taking the square root. In order to avoid potential bias caused by genes that are either highly expressed or non-expressed, expression values were normalized to z-scores and only the top 1,000 expressed genes were used. To account for variabilities in expression characteristics for each cell type, background Euclidean distance was calculated based on permutated treatment labels. P-values were calculated by comparing the Euclidean distance based on real data and that from 5000 permutations, followed by adjustment for multiple testing with the Benjamini-Hochberg[173] method across all cell clusters analyzed.

Differential gene expression analysis

To identify differentially expressed genes (DEGs) induced by fructose or HFHS diet, two different methods were used. For adipose, small intestine and hypothalamus, where only a single batch of data was involved, a pipeline involving the Wilcoxon rank sum test was used to obtain conservative DEGs that were consistent across comparisons involving individual samples. Briefly, for cell types where a gene is expressed in more than 10% of cells in both groups involved in a comparison, cells from each tissue sample in a treatment group (i.e., fructose or HFHS) were compared against cells from all samples in the control diet group or vice versa. For instance, to obtain DEGs affected by HFHS in small intestine scRNAseq data, cells from each control sample were compared against cells from all cells in the HFHS samples, and cells from each HFHS sample were compared against cells from all cells in the control group. Within each comparison, the Wilcoxon Rank Sum test was used in the Seurat Findmarkers function. Subsequently, the Metap[174] package in R was applied to meta-analyze the adjusted p-values from the Findmarkers output across all comparisons for each treatment and cell type. Significant DEGs were defined as genes with the same direction of change in 75% of all comparison pairs and meta-analysis adjusted p-value < 0.05. This pipeline is more conservative

than the traditional approach which pools cells from all samples in each treatment group, since we only select DEGs that show consistent changes across comparisons for each diet treatment.

For liver samples where two different batches were involved, the monocle[125] pipeline was used to correct for batch effect. For genes expressed in more than 25% of cells in each cell type, a negative binomial model was first used to normalize count data, followed by fitting a generalized linear model (GLM) to retrieve dietary exposure effect with batch effects corrected as follows:

$$Gene\ expression\ =\ b1 * batch + b2 * dietary\ exposures + b3 * UMI\ counts + b4 * Gene\ Counts$$

The b2 coefficient obtained was used to estimate dietary exposure effects. Statistical p-value was obtained using a likelihood ratio test against the null model where the dietary exposure term was not included:

$$Gene\ expression\ =\ b1 * batch + b3 * UMI\ counts + b4 * Gene\ Counts$$

Significant DEGs were defined as genes with Benjamini-Hochberg corrected false discovery rate (FDR) < 0.05. Cell types with more than 20 DEGs were subject to pathway enrichment analysis using the enrichr[83] package and the GO[85] and KEGG[84] pathway databases. Significant pathways were defined as pathways with Benjamini-Hochberg FDR < 0.05 and more than 3 overlapping genes between DEGs and pathways.

Enrichment of cell type specific DEGs for disease-associated SNPs from human genome wide association studies (GWAS)

To understand the relationship between cell type specific DEGs in our mouse models and different human diseases, the Marker Set Enrichment Analysis (MSEA) procedure in the Mergeomics pipeline was used[53,175]. The full summary statistics for human GWAS studies for various metabolic traits or diseases was curated and used in the analysis (**Table S4.2**). Only SNPs within 50kb of the human orthologs of the mouse DEGs were used. SNPs were further trimmed based on linkage disequilibrium (LD), by keeping only one eSNP for each LD block (defined as LD r^2> 0.7) using the Marker Dependency Filtering (MDF) function in Mergeomics. 1000Genome phase 1 data was used to obtain LD block data through the PLINK2 tool[176]. The LD pruned SNPs were mapped to each DEG set and the disease GWAS association P-values of the corresponding SNPs were extracted from each disease GWAS summary statistics. A modified chi-squared statistic was used for the enrichment analysis by comparing GWAS p-values of the SNP set mapped to a given DEG set against SNP sets that were mapped from randomly generated gene sets across a range of quantile-based GWAS p-value cutoffs using MSEA. The definition of the modified chi-square statistic used in MSEA is:

$\chi = \sum_{i=1}^{n} \frac{O_i - E_i}{\sqrt{E_i} + \kappa}$, where *O* and *E* are the numbers of the observed and estimated positive findings above a GWAS p-value cutoff defined by the i-th quantile, respectively; n is the number of quantile points (10 points were identified ranging from the top 50% to top 99.9% of signals based on the GWAS P-value rankings), and κ = 1 was a stability parameter that diminishes artefacts for small SNP sets with low expected counts. For the MSEA procedure, an FDR<5% cut-off was used, which was estimated by the Benjamini-Hochberg method[173] to identify significantly enriched DEG sets for each disease GWAS.

Ligand-receptor interaction network analysis

In order to characterize potential long range interactions mediated by secreted ligand-receptor pairs between the cell types of different metabolic tissues, we implemented Nichenet[177], which curates ligand-receptor interactions from various publicly available resources. The Nichenet ligand-receptor model was integrated with cell type DEGs identified for each diet to find potentially activated ligand-receptor pairs. Differentially expressed ligands identified by DEG analysis from source cells and differentially expressed receptors identified by Nichenet in target cells were used to assess ligand-receptor interaction. Secreted ligands were selected based on the UniProt database[178] and human protein atlas[179] to exclusively focus on ligands that fit the long range interaction model. Ligand-receptor interaction networks between cell types were visualized using cytoscape[180].

Analysis of metabolic intake and reaction flux through single cell flux estimation analysis (scFEA)

In order to infer flux of metabolic intermediates in fructose, glucose and fatty acid metabolism, we applied scFEA[26] which combines metabolic network analysis with deep learning methods to quantify metabolic pathway activities based on scRNAseq data. scFEA contains curated metabolic modules, however, the fructose metabolic pathway was not covered. We manually added the fructose metabolism module based on the KEGG database by adding fructose intake (*Slc2a5* and *Slc2a2*[181]), fructose to fructose 1-phosphate (F1P) (*Khk*), F1P to Glyceraldehyde + Glycerone-phosphate (synonym of dihydroxyacetone phosphate, DHAP) (*Aldoa*, *Aldoc*, *Aldob*, *Aldoart1*, *Aldoart2*), Glyceraldehyde to glyceraldehyde 3-phosphate (G3P) (*Tkfc*) and Glycerone-phosphate to G3P metabolic steps (*Tpi1*). Log-normalized gene expression counts from scRNAseq were used as input. Since the flux estimate of this method is a relative measurement within each tissue, we used [median(treatment) – median(control))/median(control)] to represent median fold change in the treatment group

compared to the control group. We focused on the nutrients (glucose, fatty acid and fructose) and metabolic steps (glycolysis, fatty acid metabolism and fructolysis) that are most relevant to HFHS and fructose diets.

Data availability

Raw fastq and raw gene expression count matrix were deposited to GEO while Interactive dataset visualization for each tissue is available from single cell portal (https://singlecell.broadinstitute.org/single_cell) under SCP1403 (small intestine), SCP1404 (liver NPC and hepatocytes), SCP1405 (adipose SVF), SCP1406 (hypothalamus) and SCP1407 (hypothalamus neurons subset).

Results

Fructose and HFHS diets induced different aspects of MetS

15% fructose solution or a HFHS diet for 11 weeks induced a variety of MetS-related phenotypes in B6 mice (**Figure 4.1**). HFHS diet caused a significant increase in body weight from 6 to 11 weeks of treatment, and significant fat mass gain started at half a week until 11 weeks, without significant changes in lean mass (**Figure 4.1B-D**). Glucose tolerance was impaired at week 4 and week 11 of HFHS treatment. In addition, plasma triglycerides (TG), plasma total cholesterol (TC), high-density lipoprotein (HDL), low-density lipoprotein (LDL), un-esterified cholesterol (UC) and insulin levels were significantly elevated by HFHS (**Figure 4.1E-G**). Conversely, 15% fructose in drinking water did not cause significant changes in body weight, plasma TG or glucose tolerance, but significantly raised plasma levels of TC, HDL, and UC (**Figure 4.1E-G**). Notably, mice provided with either the HFHS diet or fructose exhibited decreased food intake compared to controls; however, the total caloric intake of mice consuming HFHS or fructose was maintained at a similar level as the control group (**Figure**

4.1H). Our results confirmed previous reports of hyperlipidemia, hyperglycemia and obesity induced by the HFHS diet[10], and hyperlipidemia in the absence of body weight or body composition changes in B6 mice upon fructose consumption[182,183]. Therefore, the two diets altered different aspects of MetS pathologies in B6 mice, which may represent different subtypes of diet-induced MetS.

scRNA-seq identified expected cell types across metabolic tissues

In order to understand action mechanisms of MetS inducing diets in a cell-type specific manner, we dissected multiple metabolic tissues at the end of the 11-week dietary treatments from the same mice that had undergone phenotypic characterization. scRNA-seq data was obtained for small intestine, adipose stromal vascular fraction (SVF) , liver NPCs, liver hepatocytes, and hypothalamus for mice fed chow (control), HFHS, and fructose (n=2-5/group after quality control, **Figure 4.2A-D**). We identified distinct cell clusters based on tSNE plots for the small intestine (**Figure 4.2E**), adipose SVF (**Figure 4.2F**), liver (**Figure 4.2G**), and hypothalamus (**Figure 4.2H**). Cell clusters were then annotated with cell types based on the expression patterns of known cell type markers for small intestine (**Figure 4.3A-B**), adipose SVF (**Figure 4.3C-D**), liver (**Figure 4.3E-F**), and hypothalamus (**Figure 4.3G-H**). In addition to the major cell types known to constitute each of the tissues, our scRNAseq data also revealed hepatocyte subtypes based on zonation patterns, subtypes of adipocyte progenitor cells (APCs), and neuronal subtypes within the hypothalamus (**Figure 4.3I-J**) based on the expression patterns of known subtype markers. Therefore, our scRNA-seq data retrieved the expected cell types and subtypes for the metabolic tissues examined.

Fructose and HFHS diets induced different cell type specific response based on cell proportion and global transcriptomic changes

We first investigated whether fructose and HFHS treatments altered cell type proportions (general overview in **Figure 4.4A-D** for small intestine, adipose SVF, liver and hypothalamus, respectively). We found that fructose treatment increased the proportion of T cells in the small intestine (15.0% in fructose samples compared to 6.7% in controls and 2.4% in HFHS samples). By contrast, HFHS treatment increased macrophage populations in adipose SVF (For M1 macrophage 2% in control, 0.5% in fructose and 7.6% in HFHS. For M2 macrophage 4.6% in control, 0.7% in fructose and 11.7% in HFHS)

We further hypothesized that cell types sensitive to fructose or HFHS diet will undergo more robust transcriptomic changes. To measure transcriptome level changes between treatment and control cells Euclidean distance was used to quantify the scale of transcriptional differences between treatment and control groups for each cell type. As shown in **Figure 4.4E**, several hypothalamic cell types, including glutamatergic neurons, tanycytes and Myelinating oligodendrocytes showed high sensitivity to fructose; liver dendritic cells and adipose SVF macrophages had higher sensitivity to HFHS; small intestine proximal enterocytes, hepatocytes, adipose SVF APCs, and hypothalamic GABAergic neurons were sensitive to both diets.

Overall, the various analyses above support differential cellular sensitivity to different MetS risk diets despite some similarities.

Identification of DEGs altered by MetS risk diets in individual cell types

To understand the molecular changes in individual cell types induced by each diet, we identified DEGs using meta analysis of p-values from treatment-control pairs for each tissue (small intestine, hypothalamus, and adipose SVF) and a GLM based model for liver to account for batch effects (**Table 4.1**; **Table S4.3**).

Comparing all significant DEGs between diets for cell types with more than 10 DEGs, (**Figure 4.5**), we found significant overlaps in DEGs between diets for the majority of cell types, especially small intestine proximal enterocytes, liver pericentral hepatocytes, liver periportal hepatocytes and hypothalamic astrocytes compared to other cell types. Compared to this, macrophages and APC cell types in SVF showed higher numbers of HFHS specific DEGs while neurons in hypothalamus showed higher numbers of fructose specific DEGs, while these cell types still share significant proportions of overlapped DEGs. DEG overlapping analysis revealed differential cell type responses to diets in the perspective of DEGs.

We further examined top DEGs for each cell type to understand the most significant molecular changes involved in MetS induced by HFHS and fructose (top 3 DEGs in **Table 4.1** and top DEG heatmaps in **Figure 4.6**). For example, in the small intestine, *Apoa1*, an apolipoprotein related to cholesterol flux, and *mt-Rnr2* (encoding humanin, a mitochondrial derived peptide related to metabolic control[184]) were the most significant DEGs in proximal enterocytes and goblet cells, and were upregulated in response to both fructose and HFHS treatments (**Figure 4.6A)**. In adipose SVF, *mt-Rnr2* (humanin) and *Malat1* (a long non-coding RNA related to glucose metabolism[185]), were the top genes in both APCs and macrophages, where fructose upregulated and HFHS diet downregulated both genes (**Figure 4.6B)**. In liver hepatocytes, *Car3*, a gene encoding carbonic anhydrase III and known as a nutritionally regulated biomarker[186], was the top upregulated DEG after fructose treatment; *Cyp3a11,* a cytochrome enzyme related to drug metabolism[178]*,* was the top downregulated DEG after HFHS diet. Both diets upregulated *Fabp1*, a gene encoding fatty acid binding protein related to cholesterol uptake in hepatocytes[178] and *Mup20*, a gene encoding male pheromone which was also found to be affected by liver injury[187] (**Figure 4.6C)**. In liver NPCs, *Malat1* was the top DEG

downregulated by both fructose and HFHS diets. In the hypothalamus, *Malat1* was upregulated and *Nnat* (a gene encoding the receptor for the tridecapeptide neurotensin related to metabolic regulation[188]), was the top DEG downregulated by fructose in most hypothalamic cell types (astrocytes, myelinating oligodendrocytes, and all neuron clusters). *Ube3a*, a ubiquitin-protein ligase, was found to be the top upregulated DEG in all neuronal clusters after fructose treatment. *Copg2*, a gene related to intracellular protein transport[178] was the top upregulated DEG in both glutamatergic and GABAergic neurons by both fructose and HFHS diets (**Figure 4.6D)**. Notably, *mt-Rnr2* and *Malat1* were consistently top DEGs across many cell types for both diets. The importance of many of these top DEGs in metabolism in diverse tissues and cell types support the broad metabolic regulatory effects of both MetS risk diets.

Cell type specific DEGs are associated with human diseases

To understand the disease relevance of the cell-type specific DEG sets affected by fructose and HFHS, we integrated the DEGs with human disease GWAS data for various metabolic traits and disease, which provide association between the DEGs affected by MetS risk diets and human cardiometabolic diseases as GWAS implicates potential disease causal genes (**Table S4.2**).

The cell type specific DEGs altered by fructose (**Figure 4.4F)** and HFHS (**Figure 4.4G**) showed significant enrichment statistics to human cardiometabolic diseases or traits. For example, both fructose- and HFHS-induced DEGs from hepatocytes, adipose SVF APCs and small intestine proximal enterocytes were enriched for GWAS signals for TG, LDL, and TC. Both fructose- and HFHS diet-induced DEGs in hypothalamic neurons were enriched for GWAS associations with anorexia and BMI-related traits, supporting previous studies[189,190] that documented that both diets affected pathways regulating food intake and energy balance. Finally, we also found different cell type specificity between diets for disease association. Notably, HFHS-induced

DEGs from liver NPCs (Kupffer cells, endothelial cells, dendritic cells) were enriched for GWAS associations with HbA1c and the diabetic disposition index. Additional HFHS DEGs from immune cell types (adipose SVF macrophages, liver NKT cells, liver B cells, and small intestine T cells) were also enriched for MetS associated genetic signals which were not found for DEGs affected by fructose diet.

Pathway analysis of DEGs revealed tissue and cell type specific biological processes affected by MetS risk diets

We performed pathway analysis on the sets of DEGs to retrieve over-represented pathways and biological processes from individual cell types in response to the two MetS-inducing diets (top select pathways in **Table 4.1;** pathway enrichment statistics in **Table S4.4**).

First, we generated a Venn Diagram to obtain an overview of the pathways enriched in both HFHS and Fructose diet (**Figure 4.7, Table S4.5**). We found that glycolysis/gluconeogenesis was enriched in small intestine proximal enterocytes specifically in response to fructose, while HFHS treatment lead to enrichment in gastric acid secretion; fat digestion and absorption pathways were shared by both diets. In liver periportal hepatocytes, the citrate cycle (TCA cycle) was specific to the fructose diet, fatty acid alpha-oxidation was specific to HFHS diet and metabolic pathways were shared by both diets. We have found some cell types showed large pathway numbers shared by both diets, such as small intestine proximal enterocytes sharing fat digestion and absorption, liver periportal hepatocytes sharing metabolic pathways and pericentral hepatocytes sharing PPAR signaling pathway. We also identified cell types with larger numbers of pathways enriched in HFHS diet, such as small intestine goblet (e.g. renin-angiotensin system) and T cells (e.g. carbohydrate digestion and absorption) and all SVF APC cell types (e.g ECM-receptor interaction), as well as cell types with larger number of pathways

enriched in fructose diet, such as hypothalamus glutamatergic neurons (e.g. long-term potentiation).

We also inspected pathways related to tissue specific functions. In the liver, translation, metabolic, and inflammatory pathways were altered by both MetS diets but the cellular specificity and direction of change in the pathways were not always consistent between diets (**Figure 4.8A**). For instance, fructose-treated mice exhibited a downregulation of translation pathways in four major liver cell types (hepatocytes, Kupffer cells, endothelial cells, and dendritic cells), whereas HFHS downregulated translation pathways in hepatocytes and Kupffer cells while upregulating these pathways in endothelial cells. Fructose treatment resulted in downregulation of several metabolic pathways such as triglyceride homeostasis (detailed in **Figure 4.8B**) and gluconeogenesis was observed in hepatocytes; in contrast, the HFHS diet upregulated metabolic pathways in hepatocytes, including triglyceride homeostasis, carbon metabolism, and PPAR signaling pathways. Inflammatory pathways (e.g., Antigen processing and presentation) were also downregulated in hepatocytes by fructose and HFHS diets while upregulated in periportal hepatocytes under HFHS treatment. Finally, antigen processing and apoptosis pathways in liver Kupffer cells were upregulated in response to HFHS diet but were downregulated in response to fructose diet.

In previous studies, bulk profiling of adipose SVF cells from mice fed a HFHS diet revealed alterations in extracellular matrix (ECM), inflammation and apoptosis[22,191,192]. We have now resolved these findings at the cell-type level (**Figure 4.8C)**. ECM related pathways were upregulated in most of the cell types. In addition, “Negative regulation of apoptotic process” were upregulated in APC subtypes and downregulated in macrophage subtypes. Finally, the changes in inflammatory pathways showed similar directionality in APCs and macrophages.

TNF signaling pathway was downregulated while lysosome and inflammatory response pathways were upregulated. Compared to HFHS, the fructose diet induced fewer alterations in adipose SVF. In APC subtypes, ribosomal pathways were upregulated and the “Negative regulation of apoptotic process” process was downregulated by fructose. Further examination of DEGs involved in the negative regulation of apoptotic pathways indicated that fructose and HFHS diet induced different DEGs within this pathway (**Figure 4.8D)**. Of note, HFHS diet elicited a larger number of DEGs, higher sensitivity and larger enriched pathway number compared to fructose diet, which indicated diet-specific responses in adipose SVF cell populations.

Among the major small intestine cell types, proximal enterocytes and goblet cells displayed a strong and upregulating responses to both fructose and HFHS diets in nutrient absorption and lipid transport process, including carbohydrate digestion and absorption, fat digestion and absorption, cholesterol absorption, and chylomicron assembly (**Figure 4.8E)**. Proximal enterocytes also showed upregulation of the renin-angiotensin pathway and downregulation of oxytocin pathways under both fructose and HFHS treatment. In addition, HFHS diet treatment uniquely upregulated chylomicron assembly, fat digestion and carbon metabolism in T cells.

Finally, hypothalamic neurons showed distinct pathway alteration patterns between diets. While pathways related to GABAergic synapse, glutamatergic synapse function and oxytocin signaling were shared by both diets with different directionality; downregulated in fructose treatment while upregulated in HFHS treatment, as exemplified by genes from the oxytocin signaling pathway (**Figure 4.8F-G**). Fructose diet induced additional pathways in oxidative phosphorylation and Vasopressin−regulated water reabsorption (**Figure 4.8F**). In addition, genes from pathways shared by two diets showed different directionality. Fructose treatment downregulated these

pathways while HFHS treatment upregulated them, as exemplified by genes from the oxytocin signaling pathway (**Figure 4.8G**). Both risk diets downregulated mitochondrial electron transport, suggesting downregulation of ATP synthesis.

In summary, our pathway analysis revealed key similarities and differences in the molecular processes perturbed by fructose and HFHS diets across tissues and cell types. This was demonstrated by inflammatory and ECM related pathways in APC cell types in adipose SVF affect by HFHS. In addition, MetS diets showed different directionality in hepatocyte metabolic functions and neuron synapse pathways though they are shared in both MetS diets. Through single cell analysis, we were able to investigate how molecular processes in different tissues was perturbed by MetS diets at cell type resolution, which was missing in bulk tissue RNA analysis.

MetS risk diet induced cell type specific alterations in nutrient uptake and metabolic flow

We hypothesize that the genes and pathways altered in individual tissues and cell types reflect nutrient and metabolic flux changes under HFHS and fructose diets. We applied single cell Flux Estimation Analysis (scFEA) to infer the intake and metabolic activities in major cell types of different tissues. We used our scRNAseq data and the curated metabolic gene module information (**Methods**) as inputs to scFEA, which used machine learning models to estimate metabolite intake and flux from genes in metabolic gene modules. Flux outputs were further compared across conditions within cell type to estimate changes induced by MetS diets.Results indicated strong cell type specific alterations in nutrient uptake and metabolism. For example, small intestine proximal and distal enterocytes showed a robust increase in fatty acid intake which was estimate by *Slc27a1-6* and fructose intake which was estimate by *Slc2a5* under both diets (**Figure 4.9A**). These cell types also showed increased fructose metabolism as indicated

by increased fructose to Fructose-1-Phosphate (F1P) flux estimated by *Khk* under HFHS treatment and increased F1P to Glyceraldehyde and Glycerone−phosphate flow which was estimated by *Aldoa, Aldoc, Aldob, Aldoart1, Aldoart2* under both fructose and HFHS treatments. In comparison, there was little alteration related to glucose metabolism and intake, which is in line with the role of enterocytes in changing activities of fatty acid and fructose absorption and metabolism[156]. Compared to enterocytes, few changes were found in small intestine macrophages, indicating a limited role for inflammatory cells in absorbing and metabolizing nutrients in the small intestine.

In liver (**Figure 4.9B**), both periportal and pericentral hepatocytes showed increased intake of fatty acids under both fructose and HFHS diet which was estimated by *Slc27a1-6*, and increased pyruvate to acetyl-CoA flow which was estimated by *Dlat, Dld, Pdha1, Pdha2, Pdhb* under both dietary treatments. The conversion of acetyl-CoA to fatty acid, which was estimated by *Fasn, Acaca, Acsl1*, etc. was inferred to be increased in periportal hepatocytes under fructose treatment, whereas HFHS increased this conversion in both periportal and pericentral hepatocytes. These results suggest both diets lead to increased energy storage. We also did not see any changes in fructose intake in hepatocytes and only a small increase in fructose metabolic flux under HFHS treatment. Instead, we observed the increased glucose intake in both diets, which suggests fructose conversion to glucose in the small intestine[156]. When fructose intake does not exceed the metabolic capacity of the small intestine, hepatocytes played a smaller role[156,193]. Compared to the alterations in hepatocytes, Kupffer cells showed minor changes in the metabolism of nutrients, which indicated minor roles of inflammatory cells in liver nutrient metabolism.

In adipose SVF (**Figure 4.9C**), HFHS diet increased fatty acid intake in both APCs and macrophage cell types, along with increases in fatty acid and acetyl-CoA flux activities in APCs. Compared to HFHS, the fructose diet increased the activity of the first step of glycolysis (Glucose to G6P) in APCs but not the other steps, which suggests alterations in the glycolytic flux and a potential build-up of G6P. We also noted there was little alteration in both fructose intake and metabolic flux in SVF cells, suggesting a minor role of adipose SVF in fructose metabolism.

In the hypothalamus (**Figure 4.9D**), the fructose diet increased GABA intake in astrocytes while decreasing fatty acid intake in both GABAergic and glutamatergic neurons. Compared to this, the HFHS diet increased glucose intake in both neuron clusters and decreased fatty acid intake in glutamatergic neurons. While fatty acid intake was decreased, the metabolic flux showed limited alteration, which suggests fatty acids were not actively utilized for energy related roles in neuronal cells[194]. Both fructose and HFHS diets also decreased the rate of two steps in the glycolysis pathway (3-phosphoglycerate to pyruvate and pyruvate to Acetyl−CoA) in neurons, which could result in potential glucose and intermediate G3P build up in neurons. Finally the lower availability of direct energy substrate, acetyl-CoA, may limit energy production in TCA cycle of neurons, which may result in the downregulation of the mitochondrial electron transport pathway in neurons as shown in **Figure 4.8**.

We further conducted metabolic pathway enrichment analysis in order to compare with scFEA results (**Figure 4.9E**). Results indicated that both diets increased in fat digestion and absorption pathways in small intestine enterocytes. Under both diets, liver hepatocytes decreased gluconeogenesis and increased fatty acid metabolism pathway. However, fructose treatment increased glycogen synthesis pathways and bile acid synthesis/secretion, while HFHS

increased fatty acid biosynthetic process. SVF APC cells decreased differentiation to fat cell under both diets. Pathway analysis indicated enterocytes as a major fatty acid absorption site while hepatocytes conducting glucose and fatty acid metabolic activities, which corroborated with the intake and flux activities we have observed from **Figure 4.9A-B**. It is reported that fatty acid transporter protein (encoded by *Slc27a*) transports bile acids and long-chain FA[195]. The upregulation of bile acid biosynthesis and secretion in pericentral hepatocytes under fructose treatment (**Figure 4.9E**) suggests the increased fatty acid intake (**Figure 4.1A**) under fructose treatment is partly mediated with bile acids. Furthermore, pathway analysis also identified gluconeogenesis and glycogen synthesis activities which were not well captured from flux analysis.

The scRNA-seq metabolic flux analysis, combined with pathway enrichment analysis revealed specific cell types with potential metabolic alterations which may be direct effectors of the diets. Both fructose and HFHS diets indicated fructose metabolic alterations in enterocytes, fatty acid metabolic alterations in hepatocytes and glycolysis alterations in neurons. HFHS induced additional fatty acid intake in APCs. Through the uptake and metabolism of nutrients in the diets, these cells can trigger changes in the pathways of other cell types nearby or remote metabolic tissues.

Network analysis inferred key ligand-receptor interactions between the hypothalamus and peripheral metabolic tissues that are affected by fructose and HFHS diets

Tissue crosstalk plays a major role in the pathophysiology of MetS, a systemic disease[196,197]. This inter-tissue communication is often mediated by circulating ligands, such as leptin and ghrelin, hormones that can be secreted from one tissue and bind to receptors in another tissue[198]. Therefore, identifying novel circulating factors mediating tissue crosstalk will enable us

to better understand the molecular mechanisms underlying MetS. We hypothesized that MetS-inducing diets would lead to changes in ligand secretion and circulation, resulting in responses in downstream tissues. To investigate potential ligand-receptor pairs involved in the systemic regulation of dietary response in MetS, we focused on long range interactions between cell types across different tissues using Nichenet[177] (**Figure 4.10**). In particular, we used DEGs from each tissue and cell type to identify ligand-receptor pairs that were affected by each diet to elucidate how secreted ligands from a source tissue/cell type affects receptor-mediated functions in other tissues and cell types. It is important to note that all interactions discussed below are inferred by the scRNAseq data and Nichenet.

In small intestine, proximal enterocytes and goblet cells had the largest number of DEGs that encode ligands secreted from the small intestine for both fructose (5 differentially expressed ligands out of 7 total detected ligands; **Figure 4.10A**) and HFHS diets (6 DEGs out of 8 total detected ligands; **Figure 4.10B**). The ligands *Apoa1* were the top ligand interacting with APC cells, macrophages, astrocytes, enterocytes and liver NPC cells (Kupffer, sinusoidal endothelial cells and B cells) after HFHS treatment. *Saa1* was the top ligand interacting with APC cells, hepatocytes, oligodendrocytes and enterocytes after fructose treatment. Of note, *Saa1* was a unique ligand interacting with hepatocytes in the fructose treatment condition. In addition, T cell ligand secretion (*Ccl5*, *Ccl25*, *Apoa1*, *Apob*) was only altered by HFHS diet, which indicated stronger inflammatory response after HFHS treatment in SI.

In adipose SVF (**Figure 4.10C-D**), HFHS treatment affected more DEGs encoding secreted ligands across major cell types than fructose treatment (19 vs 4 ligands). Notably, one of the ligands is encoded by *Apoe,* the cholesterol carrier ligand secreted from mesothelial, endothelial, macrophage subtypes (M1, M2), APC subtypes (Hsd11b1, Pi16), and T cells. We

also identified DEGs encoding inflammation related ligands (*C3*, *Tnf*, *Il1b*, *Il1rn* and *Igf1*), which are secreted from endothelial, a macrophage subtype (M2), and APC subtypes(Hsd11b1, Agt, Pi16) and interacted with not only cell types from SVF but also various cell types from liver, small intestine and hypothalamus. These results indicated that the HFHS diet likely induces a widespread inflammatory response that originates from adipose SVF cells and propagates to other metabolic tissues.

In liver, both fructose (14 altered ligands; **Figure 4.10E**) and HFHS (18 altered ligands; **Figure 4.10F**) significantly affected DEGs encoding ligands. The main ligands related to cholesterol transport (*Apoa1*, *Apoe*, *Apob*) as well as inflammatory function (*C3*, *Igf1* and *Trf*) from periportal hepatocytes, pericentral hepatocytes and Kupffer cells were altered by fructose and HFHS diets. For metabolic ligands altered by fructose diet, *Apoe* and *Apob* were mostly interacting with receptors in multiple small intestine and liver cell types while *Apoa1* was interacting with receptors in adipose SVF, hypothalamus and liver cell types. For HFHS diet, *Apob* from periportal hepatocytes was interacting with small intestine goblet and proximal enterocytes and liver cell types; *Apoe* was interacting with small intestine, adipose SVF and liver pericentral hepatocytes; *Apoa1* was interacting with different types of cells from all tissues. For inflammatory ligands such as *Il1b, C3, Igf1* and *Trf*, the fructose diet altered receptor activities for these ligands mostly in the small intestine and liver, while HFHS diet affected more receptor activities in adipose SVF. Finally, *Fga,* a gene encoding a fibrinogen subunit, was secreted from pericentral and periportal hepatocytes and targeted adipose APCs and hypothalamic GABAergic and glutamatergic neurons in response to the fructose diet.

In the hypothalamus, fructose (**Figure 4.10G**) resulted in more ligand alterations compared to HFHS (7 vs 4 ligands; **Figure 4.10H**). Among these, known neuropeptides, such as *Oxt*, *Avp*

and *Apoe* exhibited altered expression in hypothalamic neurons and supporting cell types and interacted with periportal and pericentral hepatocytes, proximal enterocytes and goblet cell types. *Gal*, a less studied ligand encoding Galanin and related to metabolic function[199], was secreted from GABAergic neurons and interacted with receptors in hepatocytes and adipose APC cells. We found that more cell types responded to the 4 ligands altered by HFHS than responded to the 7 ligands altered by fructose. This is likely due to more extensive alterations in the corresponding receptors in target peripheral tissues (liver, small intestine and adipose SVF) under HFHS treatment.

Several ligands identified in the analysis are known MetS-related ligands. For example, *Avp* (vasopressin) [200,201] was altered by both fructose and HFHS in hypothalamic neurons. Apolipoproteins (*Apoa1*, *Apob* and *Apoe*) was affected in proximal enterocytes, GABAergic and glutamatergic neurons, pericentral and periportal hepatocytes by both fructose and HFHS diet. Inflammatory ligands (*C3*, *Igf1*, *Tnf, Il1b, Il1rn, etc.*) were affected in adipose macrophages and APCs by HFHS treatment and corroborates the inflammatory response induced by HFHS diet[153,202,203]. Importantly, novel ligands such as *Gal* and *Fga*, which were not known to play a role in MetS, were also identified.

When comparing the ligands identified in our study with a recently published serum proteome study for cardiometabolic disorders[204], we found that the ligands identified by our study had both lower p-values in association to T2D (p= 0.017 Wilcoxon signed-rank test) and a higher proportion of significant ligands with T2D association p-value < 0.05 compared to the whole serum proteome results (23.8% vs 9.9%, p = 0.007 Fisher's exact test). This supports the potential clinical relevance of our scRNAseq-based results. Finally, our analysis also identified novel interacting cell types mediated by these ligands. For example, hypothalamic *Avp,* which

was previously reported to interact with the liver[205], was also shown to interact with distal enterocytes, goblet cells and T cells in the small intestine in our analysis, which warrants future investigation.

Discussion

Both fructose and HFHS diets have long been associated with MetS[13,146,162,206], and bulk tissue studies of these diets have revealed the involvement of multiple tissues and diverse molecular pathways [156,207,208] in MetS pathophysiology induced by these risk diets. As metabolic tissues are highly heterogeneous in cell type composition, single cell studies are necessary to elucidate the cellular landscape of MetS diets. Recent single cell studies focused on the effect of high fat diet on adipose tissue[209,210]. However, there is limited knowledge of the cellular landscape resulting from the differential effects between fructose and high fat diets across metabolic tissues. Our study fills this gap by applying single cell analysis across multiple tissues and dietary treatments.

Our study confirmed previous findings that the two MetS risk diets induced distinct metabolic phenotypic responses[10,211]. HFHS increased weight and fat mass, elevated plasma lipids and insulin levels, and impaired glucose tolerance in B6 mice, all of which was not observed except high plasma lipids in fructose-fed B6 mice. These results are consistent with previous studies which show that HFHS diet consumption is accompanied by higher body weight and more severe hyperglycemia and hyperinsulinemia than diets with only one component, either fat or sucrose[212–214]. Both diets affected lipid traits. The phenotypic similarities and differences suggest both shared and distinct mechanisms of the two MetS risk diets. Our single cell analysis supports this hypothesis and shows that diverse cell types such as liver hepatocytes, endothelial cells, macrophages, hypothalamic neurons, intestine enterocytes, and adipose

APCs across different tissues exhibit differential responses to the two MetS diets (**Table S4.1**). For instance, adipose SVF APCs and macrophages were found to respond more dramatically to HFHS diet, which agrees with the phenotypic changes for HFHS. The hypothalamic GABAergic and glutamatergic neurons were more responsive to fructose, and the fructose-responsive DEGs were previously linked to anorexia and BMI regulation in human GWAS, which may agree the known role of hypothalamus in regulating food intake and energy balance[189,190].

Further investigation of DEGs and pathway analysis revealed common and unique alterations induced by the two MetS diets, from mt-Rnr2 gene involving metabolic function to pathways in fatty acid absorption and glycolysis to genes. Among the shared DEGs between diets, *mt-Rnr2* (humanin) was a top DEG in small intestine proximal enterocytes, goblet cells, adipose SVF APC cells and macrophages, while *Malat1* was a top DEG in periportal and pericentral hepatocytes, various adipose SVF APC subtypes and macrophages subtypes. *Mt-Rnr2* is involved in the regulation of energy expenditure[184]. *Malat1* has been associated with atherosclerosis[215] and inflammation in endothelial cells[216] as well as hepatic steatosis[217]. Despite the presence of these shared DEGs, numerous DEGs were specific to each diet. Our pathway analysis also uncovered how functional pathways were differentially altered between cell types and treatments. For instance, metabolic pathways including PPAR signaling, fatty acid metabolism and cholesterol homeostasis were downregulated by fructose but upregulated by HFHS in hepatocytes; inflammatory pathways were induced in adipose APCs and macrophages under HFHS treatment; neuropeptide pathways were affected in hypothalamic neurons by both fructose and HFHS treatment, however a larger number of DEGs and pathways were found for fructose diet. These findings highlight the heterogeneity of cellular responses in each tissue to different diets and the importance of using single cell technologies to decipher such heterogeneity to reveal precise cellular and gene targets of individual risk diets.

Our multitissue single cell studies also offer the opportunity to explore cross-tissue cell-cell communication. We conducted ligand-receptor interaction network analysis in order to identify secreted ligands which show differential expression in a given cell type and their potential downstream responding cell types with differential expression of the corresponding receptors. The network analysis recapitulated genes encoding known metabolic regulators, such as *Avp* (vasopressin) [205,218], *Apoe* (apolipoprotein E)[219] and *C3* (complement factor 3)[220]. Some of the ligands (e.g., *Apoe, C3*, *Il1rn*) identified in our analysis were also found in a recently human serum proteome study[204] for playing a role in clinical outcomes of metabolic disorders. We also identified novel ligands which were less investigated, including *Fga*[221,222] and *Gal*[199,223]. *Fga* encodes for the fibrinogen alpha chain which acts as the alpha component of fibrinogen and is known for its wound repair function. The association between of fibrinogen and MetS was investigated in previous studies[222,224], but the functional role of *Fga* in MetS has not been explored. *Gal* encodes a neuropeptide which is related to cognitive function and endocrine regulation[225]. While *Gal* was previously shown to be a potential MetS biomarker[226], its physiological function in metabolism is not well understood.

In addition to known and novel ligands, we also identified the potential ligand sources and interacting cell types. For example, vasopressin and oxytocin were secreted mainly from the hypothalamus neurons (GABAergic and Glutamatergic) and supporting cells (Astrocytes, oligodendrocytes precursor) and targeted cell types in liver and small intestine. *Apoe* was secreted from liver, hypothalamus and adipose SVF cell types and targeted cell types in all tissues. *Fga* was secreted from liver cell types and targeted adipose APCs and hypothalamic neuronal cell types in response to the fructose diet. *Gal* was secreted from the hypothalamus GABAergic neuron and targeted cell types in the liver and adipose SVF under HFHS diet. The

cell-type level fine map of both the known and novel ligands and the potential cell-cell interactions across tissues revealed by our analysis warrants further experimental testing to investigate their causal and functional relationships with MetS.

To explore how the two diets may trigger differential cellular and molecular sensitivity in the four metabolic tissues, we further used our scRNA-seq data to carry out metabolic activity analysis to infer how different MetS diets can alter metabolic flow in specific cell types which may in turn affect molecular pathways in various cell types across tissues. This analysis showed that small intestine enterocytes not only played a role in fatty acid and fructose uptake but also metabolized fructose, in contradiction to the commonly held belief that the liver is the key tissue that metabolizes fructose. Our finding is support by a recent publication[156] based on isotope tracing which indicated that fructose can be fully metabolized in the small intestine if the fructose quantity was within small intestine's metabolic capability. In addition, both diets increased fatty acid uptake as well as fatty acid metabolism by the liver, with stronger effects observed for HFHS diet, which may explain the robust increase in triglycerides levels as a result of HFHS treatment. Metabolic activity analysis also showed increased fatty acid uptake in both adipose APCs and macrophages under HFHS treatment, which we hypothesize is linked with strong alterations in inflammatory, ECM and PPAR signaling pathways identified in these cell types. In the hypothalamus, fatty acid uptake was decreased in neuronal cells by both diets. It is documented that fatty acids are used as signaling molecules that regulates energy balance in hypothalamus[194,227] that could potentially contribute to altered energy intake behavior. Through this analysis, we also showed metabolic activity is highly cell type specific, where supporting cell types such as small intestine macrophages, liver Kupffer cells and adipose SVF endothelial cells showed very limited alterations in metabolic activity compared to intestinal enterocytes and hepatocytes, as expected. Furthermore, we found limited alterations in fructose metabolic

activity metabolic activity in cell types that do not compromise the small intestine, which indicated that under our experimental design, fructose from both HFHS and fructose diets were more likely metabolized in the small intestine and conferred their effects through increasing fatty acid levels and related metabolic activities[228].

Integrating our findings from the phenotypic characterization, cell type specific DEGs and pathways, cross-cell-type interaction, and the nutrient uptake and metabolic flow analysis across cell types from different tissues, we propose the following mechanistic models under either fructose or HFHS diet. Under fructose diet (**Figure 4.11**) fructose induced enhanced fructose and fatty acid absorption activity in small intestine enterocytes. The increased fatty acid and fructose level traversed across different organs, causing alterations of hepatocyte PPAR signaling and triglyceride metabolism, as altering adipose APC inflammatory pathway and ECM pathways, and inducing anorexia phenotypes and reducing glycolysis flux in hypothalamic neurons. These fructose-induced alterations might be mediated by key promoted ligand secretion (such as *Apoe*, *Fga* and *C3* from hepatocytes and *Avp* from neurons) and potentially further enhanced metabolic dysfunctions, especially high plasma cholesterol. In HFHS diet (**Figure 4.11**) also induced induced enhanced fructose and fatty acid absorption activity in small intestine enterocytes. Circulating glucose and fatty acid further caused PPAR, carbon metabolism and fatty acid metabolic upregulations in hepatocytes. Furthermore, excessive fatty acid concentrations can also induce robust inflammatory response and ECM dysregulation in APC cells and M2 macrophage. Neurons in hypothalamus is also showing alterations in food intake GWAS correlations and synapse function. Ligands (*Avp* and *Gal* from neurons, *Il1m* from APC cells, *C3* and *Apoe* from APC and hepatocytes) were secreted during the process while potentially further mediating metabolic dysfunction including obesity, glucose tolerance impairment, and dyslipidemia.

Our study has several limitations. To better capture a variety of tissues and treatments, a larger sample size per group than the present design is preferred. However, single cell studies using similar sample size have been shown to capture critical disease mechanisms[132,229–231]. Secondly, our network analysis focused on ligand receptor interaction network, which may miss important interactions compared to other purely data-driven algorithms[232]. Finally, our DEG and pathway analyses as well as network analysis revealed numerous hypotheses that require experimental validation.

The combination of all our analyses enabled us to build a model of the B6 response to two MetS risk diets in the cell type level across hypothalamus, liver, adipose, and small intestine in response to two MetS risk diets. While HFHS induced robust changes in adipose APCs and macrophages, fructose had more profound effects on hypothalamic cell populations, particularly neurons. We also found that *Malat1* and *mt-Rnr2* were frequently altered across multiple cell types by both diets, supporting these as highly responsive gene markers of diet-induced MetS. In addition, we integrated DEGs with the ligand receptor database to identify several known and novel circulating regulators mediating cell-cell interactions in MetS, including *Avp*, *Apoe*, *Oxy, Fga* and *Gal* as well as their source and target cell types. We also integrated cell type specific DEGs with both biological pathways and human GWAS to understand the potential relations between MetS diets and human metabolic diseases. The analysis revealed the specific cell types related to MetS, such as hepatocytes, which exhibited altered lipid metabolism. This is likely related to plasma TG and HDL levels which are affected by both diets. Feeding HFHS also further increased plasma LDL levels. Finally, through metabolic flux estimation, we were able to identify cell types that are likely to be directly affected by the nutrients in the different diets and further trigger other molecular pathways in these and additional cell types. For instance, small

intestine proximal enterocytes were shown to be particularly important in metabolizing fructose and changing metabolic pathways when fed either HFHS or fructose diet; under HFHS diet, adipose SVF APCs absorbed fatty acids and increased inflammatory response; liver hepatocytes uptook and metabolized circulating fatty acids and changed fatty acid and carbohydrate metabolic functions upon both HFHS or fructose diet; hypothalamus neurons decreased fatty acid intake and glycolysis activities, accompanied by altered neuropeptide and energy balance related functions under both HFHS or fructose diet. This study supports further development of treatments targeting sensitive cell types and key regulators in MetS subtypes induced by different dietary risks.

Conclusion

Through application of scRNA-seq in small intestine, liver, adipose SVF and hypothalamus with two MetS related diets, our study enabled the elucidation of the roles of individual cell types in MetS by pinpointing the genes and pathways that are altered in these cell types and revealing the inter-tissue cell-cell crosstalk network. The use of scRNA-seq enabled detailed comparison between fructose and HFHS which induced different metabolic dysfunctions and cell type specific effects. Our datasets and findings could serve as a rich resource to expediate future nutrigenomic and mechanistic studies of MetS.

Tables

Table 4.1. Summary of differential expressed genes and pathways in selected cell types.

tissue name	*Cell types*	*treatment*	*DEG#*	*Overlap DEG#/ Pathway#*	*top selected DEGs*	*selected top pathways*
Small intestine	Proximal enterocytes	Fructose	670	417/75	*Apoa1 Actb Usp4*	Fat digestion and absorption, Oxytocin signaling pathway, Renin-angiotensin system
		HFHS	627		*Fgfr3 Sepp1 Slc5a1*	Fat digestion and absorption, PPAR signaling pathway, Carbohydrate digestion and absorption
Adipose	APC_Pi16	Fructose	150	74/17	*mt-Rnr2 Malat1 Gm26809*	Extracellular matrix disassembly, I-kappaB kinase/NF-kappaB signaling, Negative regulation of G2/M transition of mitotic cell cycle
		HFHS	338		*Malat1 Fabp4 Lyz2*	ECM-receptor interaction, TNF signaling pathway, PPAR signaling pathway
	APC_Hsd11b1	Fructose	180	82/17	*mt-Rnr2 Malat1 Cxcl1*	Negative regulation of apoptotic process, I-kappaB kinase/NF-kappaB signaling, extracellular matrix disassembly
		HFHS	424		*Malat1 Lyz2 Cd74*	ECM-receptor interaction, PI3K-Akt signaling pathway, TNF signaling pathway
Liver	Periportal hepatocytes	Fructose	272	189/68	*Car3 Mup17 Mup9*	PPAR signaling pathway, Very-low-density lipoprotein particle assembly, Glycolysis/Gluconeogenesis
		HFHS	409		*Cyp3a11 Fabp1 Mup20*	PPAR signaling pathway, Cholesterol homeostasis, Neutrophil degranulation
	Sinusoidal endothelial cells	Fructose	47	30/11	*Abi1 Gm26924 Malat1*	Ribosome, Oxytocin signaling pathway
		HFHS	135		*Fos Hbb-bs Iigp1*	Type I interferon signaling pathway, Negative regulation of apoptotic process, Vascular endothelial growth factor receptor signaling pathway
Hypothalamus	Glutamatergic neurons	Fructose	627	93/18	*mt-Rnr2 Fgf14 Malat1*	Oxidative phosphorylation, Glutamatergic synapse, Long-term potentiation
		HFHS	170		*Copg2 Lrba Dlg2*	Glutamatergic synapse, Oxytocin signaling pathway, Calcium signaling pathway
	GABAergic neurons	Fructose	713	167/55	*Malat1 Oxt Avp*	GABAergic synapse, Long-term potentiation, Thyroid hormone synthesis
		HFHS	307		*Cntnap2 Tbc1d9 Ube3a*	GABAergic synapse, Long-term potentiation, Oxytocin signaling pathway
	Astrocyte	Fructose	77	36/1	*Malat1 mt-Rnr2 Nrxn1*	Regulation of axon extension, positive regulation of potassium ion transmembrane transporter activity, cholesterol catabolic process
		HFHS	56		*Trpm3 Nrxn1 Gpc5*	Retrograde endocannabinoid signaling, Serotonergic synapse, Ion transmembrane transport

Table S4.1. Markers used for annotation across different cell types

Table S4.2. GWAS studies used for MergeOmics analysis on cell type specific DEGs

Table S4.3. Differential expressed gene statistics across all tissues and dietary treatments

Table S4.4. Differentially enriched pathways based on cell type specific DEGs across different tissues and dietary treatments

Table S4.5. Unique and shared cell type specific pathways in each cell type across different dietary treatments

Figures

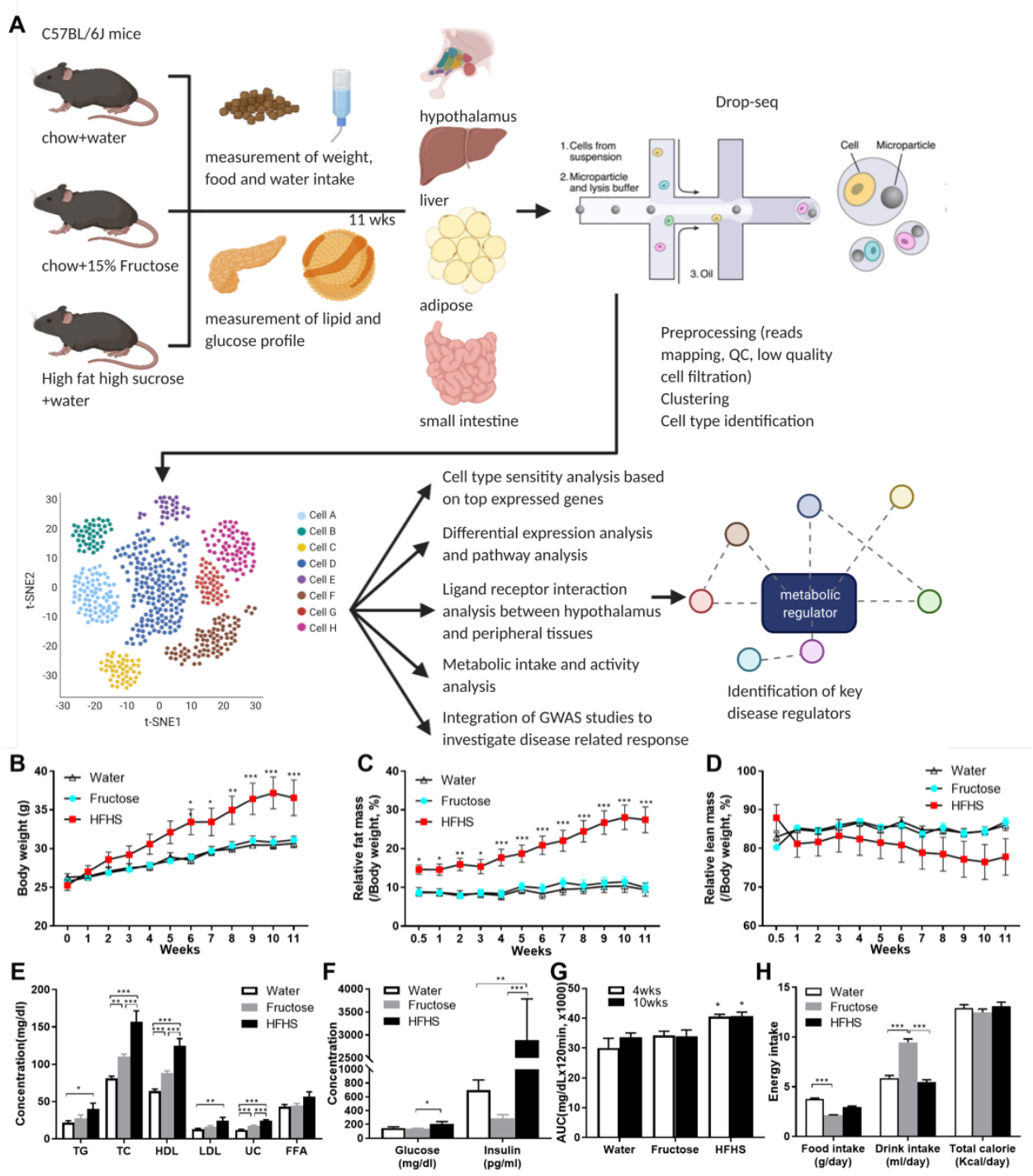


Figure 4.1. Study design and phenotypic analysis of mice.

(A) Overview of study. Male C57BL/6J mice were treated with chow+water (Control), chow+15%

fructose water (Fruc) or high fat high sucrose + water (HFHS). After 11 weeks of dietary treatment, hypothalamus, liver, gonadal adipose stromal vascular fraction and small intestine were collected and sequenced by Drop-seq. Data were processed and cell types were identified. This was followed by cell type sensitivity analysis, differential gene and pathway analysis, GWAS integration analysis, ligand-receptor network analysis, and metabolic flux (intake and activity) analysis. (B-G) Phenotypic analysis of mice treated with various diets. Cumulative change in body weight (B), relative fat mass (C), and relative lean mass (D) in mice fed normal Chow diet with water,15% fructose, or HFHS over 11 weeks. * denotes $P < 0.05$, ** denotes $P < 0.01$, and *** denotes $P < 0.001$ by Two-way ANOVA, followed by Sidak post-hoc analysis. Fasting plasma lipids (E), glucose and insulin (F) levels in response to fructose or HFHS consumption. Glucose tolerance determined using IPGTT was conducted at 4 and 10 weeks shown as area under the curve (AUC) (G). (H) Calori intake of mice fed 3 different diets. Data are expressed as means $\pm$ SEMs. * denotes $P < 0.05$, ** denotes $P < 0.01$, and *** denotes $P < 0.001$ by two-sided Student's *t*-test. Sample n=6/group.

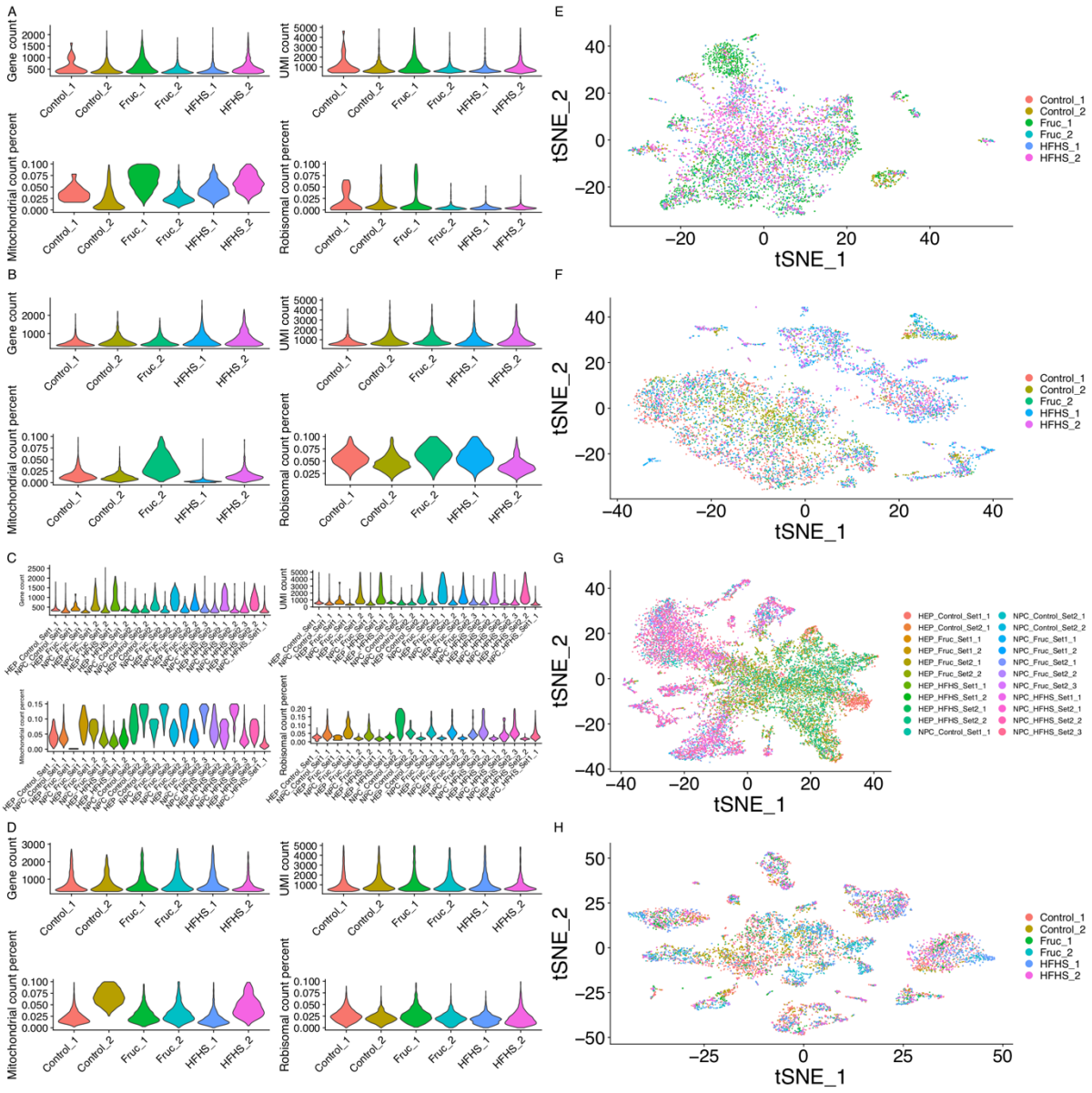


Figure 4.2. QC metrics and sample batch effect corrections visualization.

(A-D) QC metrics of filtered datasets, showing gene counts (top left), UMI counts (top right), mitochondrial gene percentage (bottom left) and ribosomal gene percentage (bottom right) in each sample used in analysis for (A) Small intestine (B) SVF (C) Liver (D) Hypothalamus. (E-H) tSNE plot of all cells, colored by samples after batch effect correction for (E) Small intestine (F) SVF (G) Liver (H) Hypothalamus

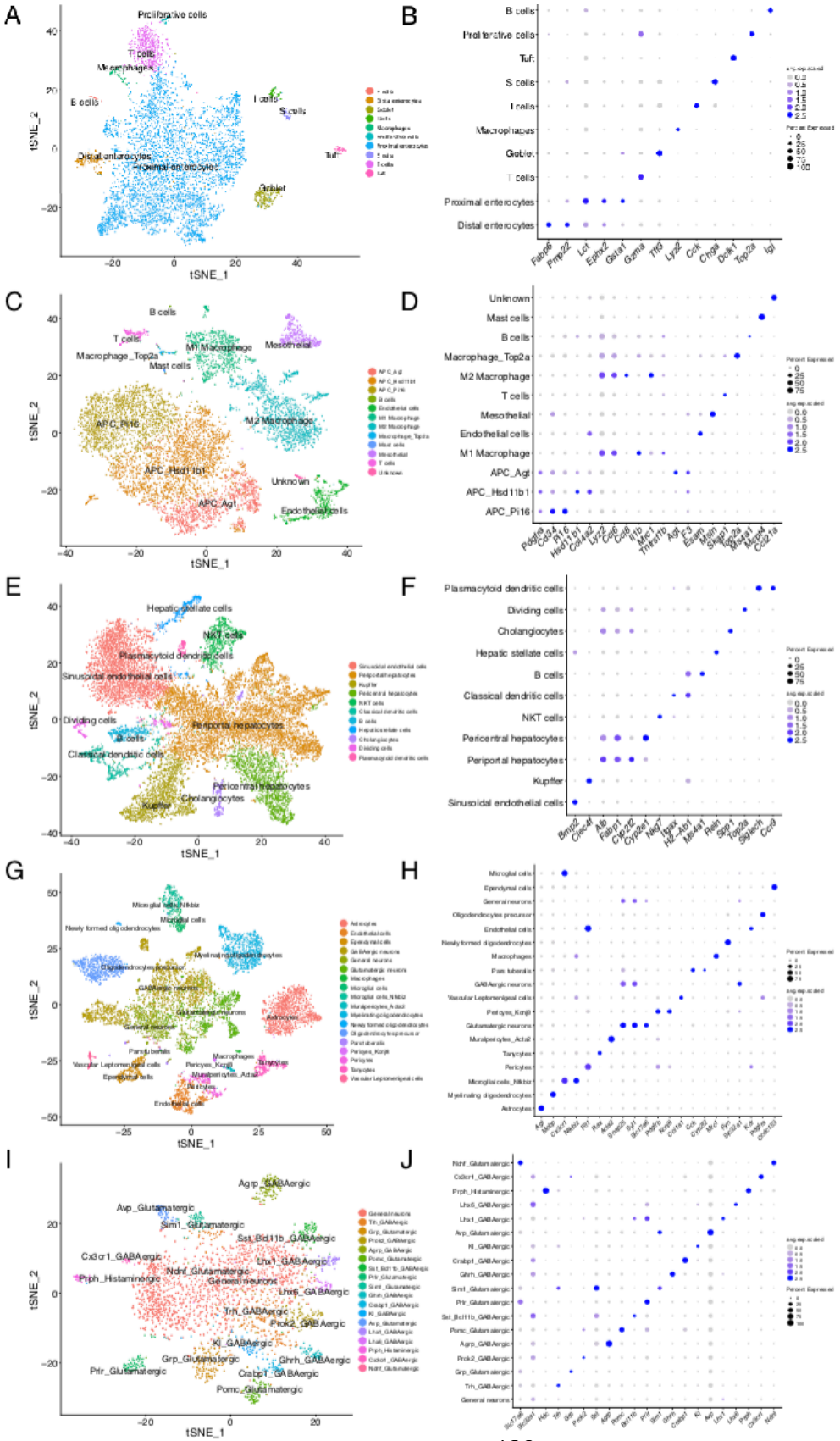

A
Proliferative cells
T cells
Macrophages
B cells
T cells
S cells
Tuf
Distal enterocytes
Proximal enterocytes
Goblet
tSNE_1
tSNE_2
B
B cells
Proliferative cells
Tuf
S cells
T cells
Macrophages
Goblet
T cells
Proximal enterocytes
Distal enterocytes
Fabp6
Pmp22
Lct
Epha2
Gata1
Gzma
Tff3
Lyz2
Cck
Chga
Dclk1
Top2a
Igj
C
B cells
T cells
M1 Macrophage
Mesothelial
Macrophage_Top2a
Mast cells
APC_Pi16
M2 Macrophage
APC_Hsd11b1
Unknown
APC_Agt
Endothelial cells
tSNE_1
tSNE_2
D
Unknown
Mast cells
B cells
Macrophage_Top2a
M2 Macrophage
T cells
Mesothelial
Endothelial cells
M1 Macrophage
APC_Agt
APC_Hsd11b1
APC_Pi16
Pdgfra
Cd34
Pi16
Hsd11b1
Col4a2
Lyz2
Cd6
Ccl6
Il1b
Mrc1
Tnfrsf1b
Agt
F3
Esam
Msln
Sfrp1
Top2a
Mki67
Mcpt4
Ccl21a
E
Hepatic stellate cells
NKT cells
Plasmacytoid dendritic cells
Sinusoidal endothelial cells
Dividing cells
Periportal hepatocytes
B cells
Classical dendritic cells
Pericentral hepatocytes
Cholangiocytes
Kupffer
tSNE_1
tSNE_2
F
Plasmacytoid dendritic cells
Dividing cells
Cholangiocytes
Hepatic stellate cells
B cells
Classical dendritic cells
NKT cells
Pericentral hepatocytes
Periportal hepatocytes
Kupffer
Sinusoidal endothelial cells
Bmp2
Clec4f
Alb
Fabp1
Cxcl12
Cyp2e1
Ntn1
Hpx
H2-Ab1
Mt4a1
Retn
Spp1
Top2a
Selectn
Ccl3
G
Microglial cells_Nfkbiz
Microglial cells
Newly formed oligodendrocytes
Myelinating oligodendrocytes
Oligodendrocytes precursor
GABAergic neurons
General neurons
Glutamatergic neurons
Astrocytes
Pars tuberalis
Macrophages
Vascular Leptomeningeal cells
Pericytes_Kcnj8
Mural/pericytes_Actа2
Ependymal cells
Tanycytes
Endothelial cells
tSNE_1
tSNE_2
H
Microglial cells
Ependymal cells
General neurons
Oligodendrocytes precursor
Endothelial cells
Newly formed oligodendrocytes
Macrophages
Pars tuberalis
GABAergic neurons
Vascular Leptomeningeal cells
Pericytes_Kcnj8
Glutamatergic neurons
Mural/pericytes_Actа2
Tanycytes
Pericytes
Microglial cells_Nfkbiz
Myelinating oligodendrocytes
Astrocytes
I
Agrp_GABAergic
Avp_Glutamatergic
Sim1_Glutamatergic
Sst_Bcl11b_GABAergic
Cxcl3cr1_GABAergic
Lhx1_GABAergic
Ndnf_Glutamatergic
Prph_Histaminergic
General neurons
Lhx6_GABAergic
Trh_GABAergic
Prok2_GABAergic
Kl_GABAergic
Grp_Glutamatergic
Ghrh_GABAergic
Ptrr_Glutamatergic
Crabp1_GABAergic
Pomc_Glutamatergic
tSNE_1
tSNE_2
J
Ndnf_Glutamatergic
Cxcl3cr1_GABAergic
Prph_Histaminergic
Lhx6_GABAergic
Lhx1_GABAergic
Avp_Glutamatergic
Kl_GABAergic
Crabp1_GABAergic
Ghrh_GABAergic
Sim1_Glutamatergic
Ptrr_Glutamatergic
Sst_Bcl11b_GABAergic
Pomc_Glutamatergic
Agrp_GABAergic
Prok2_GABAergic
Grp_Glutamatergic
Trh_GABAergic
General neurons

Figure 4.3. Identification of cell types in scRNAseq datasets from small intestine, adipose SVF, liver, and hypothalamus

(A, C, E, G, I) tSNE with cell type annotation (B, D, F, H, J) marker heat map. (A, B) Small intestine. (C, D) adipose SVF. (E, F) Liver. (G, H) Hypothalamus. (I, J) Hypothalamus neuronal subtypes. In marker dot heatmaps B, D, F, H, J, dot size indicates % of cells in each cluster with detectable marker expression (plotted as Z-score transformed value) and dot color corresponds average expression of genes in expressed cells. Sample size n=2-5/tissue.

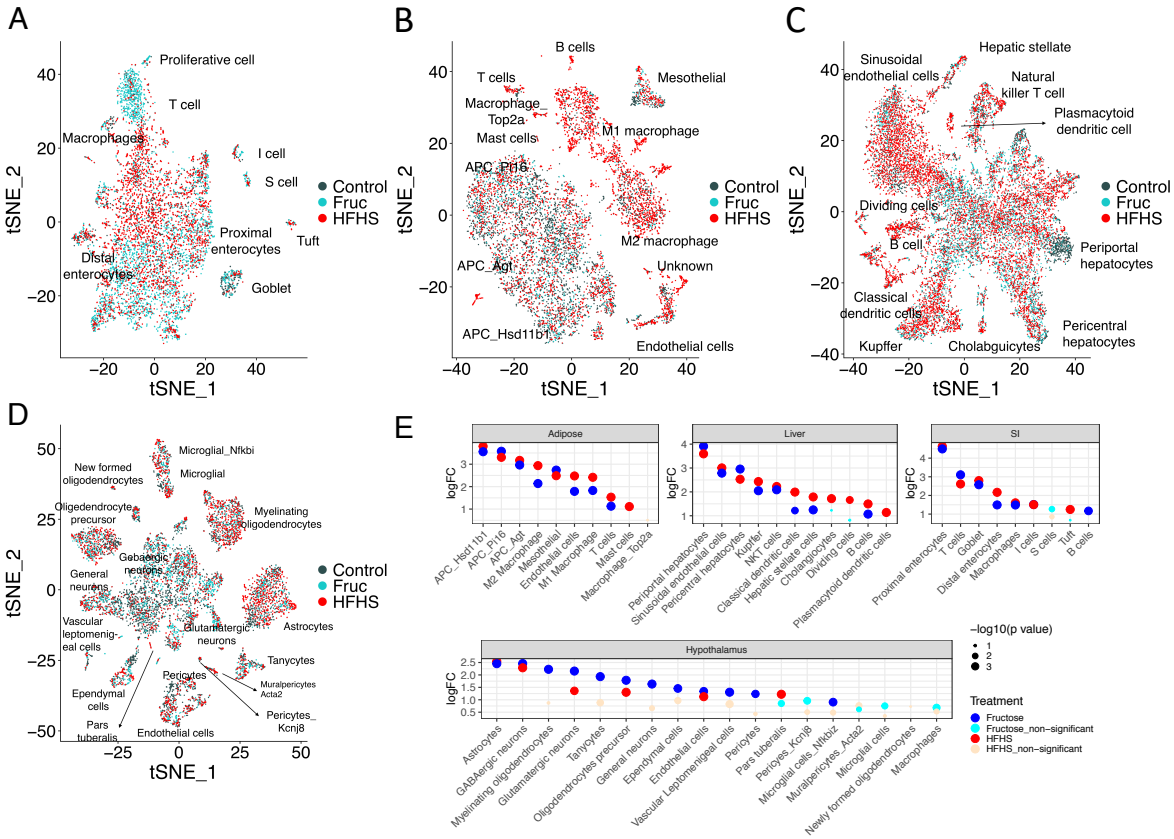

A
Proliferative cell
T cell
Macrophages
I cell
S cell
Proximal enterocytes
Tuft
Distal enterocytes
Goblet
tSNE_1
tSNE_2
Control
Fruc
HFHS
B
B cells
T cells
Mesothelial
Macrophage_Top2a
M1 macrophage
Mast cells
APC_Pi16
M2 macrophage
APC_Agt
Unknown
APC_Hsd11b1
Endothelial cells
C
Sinusoidal endothelial cells
Hepatic stellate
Natural killer T cell
Plasmacytoid dendritic cell
Dividing cells
B cell
Periportal hepatocytes
Classical dendritic cells
Pericentral hepatocytes
Kupffer
Cholabguicytes
D
Microglial_Nfkbi
New formed oligodendrocytes
Microglial
Oligodendrocyte precursor
Myelinating oligodendrocytes
Gabaergic neurons
General neurons
Vascular leptomenig-eal cells
Glutamatergic neurons
Astrocytes
Tanycytes
Pericytes
Muralpericytes Acta2
Ependymal cells
Pericytes_ Kcnj8
Pars tuberalis
Endothelial cells
E
Adipose
Liver
SI
Hypothalamus
logFC
−log10(p value)
Treatment
Fructose
Fructose_non−significant
HFHS
HFHS_non−significant
F
Fructose
Color key
−log(FDR)
SVF_APC_Hsd11b1
SVF_APC_Pi16
SVF_APC_Agt
SI_Proximal enterocytes
Liver_Periportal hepatocytes
Liver_Pericentral hepatocytes
SI_T cells
SI_Goblet
Hyp_General neurons
Hyp_Myelinating oligodendrocytes
Hyp_Ependymal cells
Liver_Sinusoidal endothelial cells
Hyp_Astrocytes
Hyp_Tanycytes
Liver_NKT cell
Liver_Classical dendritic cells
Liver_Cholangiocytes
Hyp_Endothelial cells
Hyp_GABAergic neurons
Hyp_Glutamatergic neurons
Hyp_Oligodendrocytes precursor
Liver_Kupffer
IIBDGC_IBD
IIBDGC_UC
GLGC_TG
GLGC_HDL
GLGC_TC
GLGC_LDL
CONVERGE_MDD
GCAN_AnorexiaNervosa
GIANT_BMI
EGG_BirthWeight
GIANT_WHRadjBMI
GIANT_HIPadjBMI
IIBDGC_CD
IGAP_AlzheimerDisease
PGC_Schizophrenia
SSGAC_Neuroticism
SSGAC_DepressiveSymptoms
MAGIC_FastingInsulinMain
DIAGRAM_T2D
MAGIC_HbA1C
GIANT_WCadjBMI
MAGIC_Incr30
EGG_ChildhoodBMI
G
HFHS
SVF_Endothelial cells
Liver_Kupffer
SI_Proximal enterocytes
SVF_M2 Macrophage
SVF_Mesothelial
SI_Distal enterocytes
Liver_B cells
SVF_APC_Pi16
SVF_APC_Hsd11b1
SVF_APC_Agt
SVF_M1 Macrophage
Liver_Periportal hepatocytes
Liver_Pericentral hepatocytes
SI_Goblet
SI_T cells
Liver_Dividing cells
Liver_NKT cell
Liver_Classical dendritic cells
Liver_Sinusoidal endothelial cells
Liver_Hepatic stellate cells
Hyp_Myelinating oligodendrocytes
Hyp_Astrocytes
Hyp_General neurons
Hyp_Glutamatergic neurons
Hyp_Oligodendrocytes precursor
Hyp_GABAergic neurons
IIBDGC_IBD
IIBDGC_UC
GLGC_TG
GLGC_TC
GLGC_LDL
MAGIC_Incr30
MAGIC_FastingInsulinMain
GIANT_WCadjBMI
DIAGRAM_T2D
GIANT_BMI
CONVERGE_MDD
GCAN_AnorexiaNervosa
SSGAC_DepressiveSymptoms
SSGAC_Neuroticism
PGC_Schizophrenia
IGAP_AlzheimerDisease
EGG_ChildhoodBMI
MAGIC_HbA1C
MAGIC_DI
EGG_BirthWeight
IIBDGC_CD
GLGC_HDL
GIANT_WHRadjBMI
GIANT_HIPadjBMI

Figure 4.4. HFHS and fructose diets induced transcriptomic alternations with differential tissue and cell type specificity

(A, B, C, D) tSNE figure with cell type annotation colored by different dietary treatments. (A) Small intestine, (B) SVF, (C) Liver and (D) Hypothalamus. (E) Euclidean distance representation of differential cell type sensitivity at transcriptome level after MetS diet treatment. The fold change (FC) of the Euclidean distance of either Fructose or HFHS treatment group compared with control group in each cell type was calculated by dividing the empirical Euclidean distance of that cell type by the median Euclidean distance of the permutation-based null distribution of that same cell type. The null distribution is also used to calculate p-values by comparing with empirical Euclidean distance. P-values were adjusted for each cell type by using the Benjamini & Hochberg method[173]. (F, G) Heatmap of GWAS enrichment results from Mergeomics which integrated full summary statistics of human GWAS of various diseases/traits with (F) fructose DEGs and (G) HFHS DEGs. Each heatmap tile corresponds to -log(FDR) from enrichment analysis between cell type DEGs and GWAS disease/trait. Only significantly enriched results (FDR < 5%) were colored. GWAS, genome-wide association study.

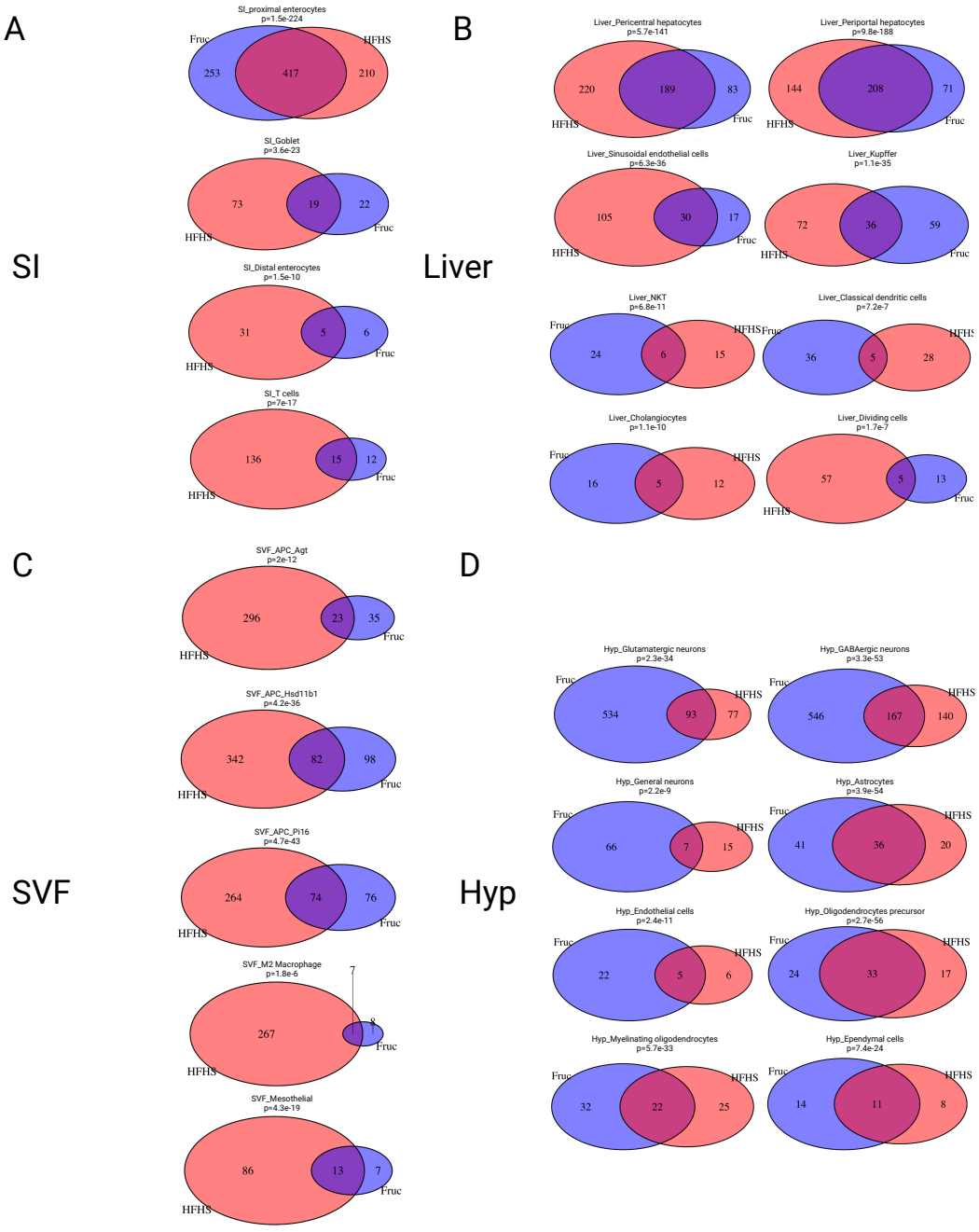
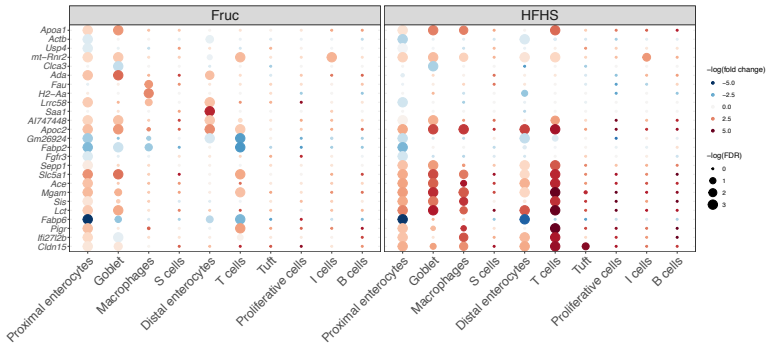
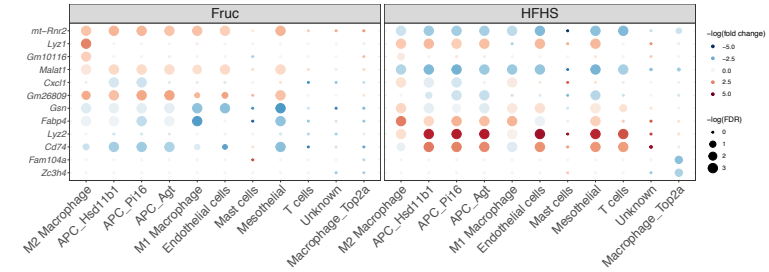
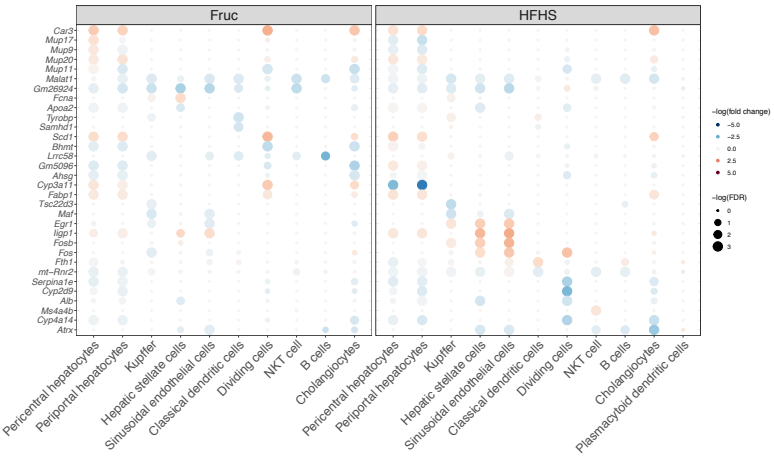


Figure 4.5. Venn diagram of DEGs shared in HFHS and fructose diets

(A) Cell types in small intestine (B) cell types in liver (C) cell types in SVF (D) cell types in hypothalamus. Only cell types with more than 10 DEGs in both dietary treatments were included, p-value was calculated based on fisher exact test.


Fruc
HFHS
Apoa1
Actb
Usp4
mt−Rnr2
Ctca3
Ada
Fau
H2−Aa
Lrrc58
Saa1
AI747448
Apoc2
Gm26924
Fabp2
Fgf13
Sept1
Slc5a1
Ace
Mgam
Sis
Lct
Fabp6
Pigr
Ifi27l2b
Cldn15
Proximal enterocytes
Goblet
Macrophages
S cells
Distal enterocytes
T cells
Tuft
Proliferative cells
I cells
B cells
−log(fold change)
−5.0
−2.5
0.0
2.5
5.0
−log(FDR)
0
1
2
3


Fruc
HFHS
mt−Rnr2
Lyz1
Gm10116
Malat1
Cxcl1
Gm26809
Gsn
Fabp4
Lyz2
Cd74
Fam104a
Zc3h4
M2 Macrophage
APC_Hsd11b1
APC_Pi16
APC_Agt
M1 Macrophage
Endothelial cells
Mast cells
Mesothelial
T cells
Unknown
Macrophage_Top2a
−log(fold change)
−5.0
−2.5
0.0
2.5
5.0
−log(FDR)
0
1
2
3


Fruc
HFHS
Car3
Mup17
Mup9
Mup20
Mup11
Malat1
Gm26924
Fcna
Apoa2
Tyrobp
Samhd1
Scd1
Bhmt
Lrrc58
Gm5096
Ahsg
Cyp3a11
Fabp1
Tsc22d3
Maf
Egr1
Igp1
Fosb
Fos
Fth1
mt−Rnr2
Serpina1e
Cyp2d9
Alb
Ms4a4b
Cyp4a14
Atrx
Pericentral hepatocytes
Periportal hepatocytes
Kupffer
Hepatic stellate cells
Sinusoidal endothelial cells
Classical dendritic cells
Dividing cells
NKT cell
B cells
Cholangiocytes
Plasmacytoid dendritic cells
−log(fold change)
−5.0
−2.5
0.0
2.5
5.0
−log(FDR)
0
1
2
3

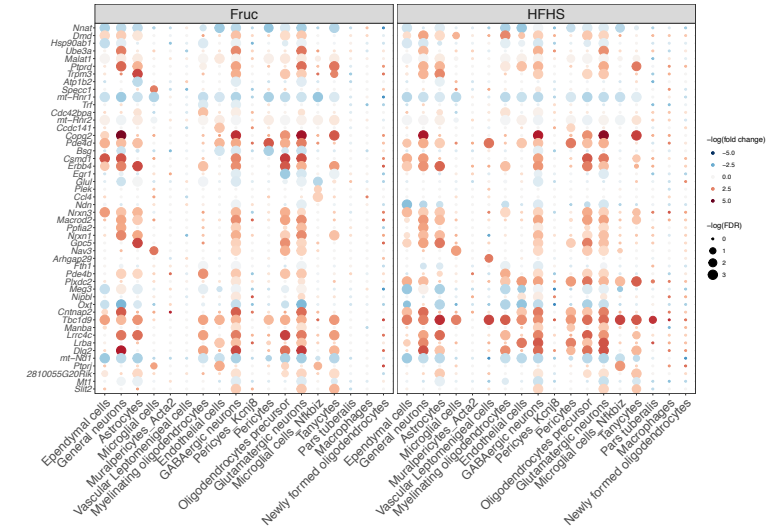

Fruc
HFHS
Nnat
Dmd
Hspa9ab1
Ube3a
Malat1
Ptprd
Trpm3
Atp1b2
Specc1
mt−Rnr1
Trf
Cdc42bpa
mt−Rnr2
Ccdc141
Copg2
Pde4dip
Bsg
Camk1
Erbb4
Egr1
Gfap
Plek
Col4
Ndn
Nrxn3
Macrod2
Pvrl3
Nrxn1
Gpc5
Nav3
Arhgap29
Fth1
Pde4b
Plxdc2
Meg3
Nipbl
Dst
Cntnap2
Tbc1d9
Manba
Lrrc4c
Lrba
Dlg2
mt−Nd1
Ptprj
2810055G20Rik
Mt1
Slit2
Ependymal cells
General neurons
Astrocytes
Microglial cells
Murabericytes_Acta2
Vascular Leptomeningeal cells
Myelinating oligodendrocytes
Endothelial cells
GABAergic neurons
Pericytes_Kcnj8
Pericytes
Oligodendrocytes precursor
Glutamatergic neurons
Microglial cells_Nfkbiz
Tanycytes
Para tuberalis
Newly formed oligodendrocytes
Macrophages
Murabericytes_Actg2
−log(fold change)
−5.0
−2.5
0.0
2.5
5.0
−log(FDR)
0
1
2
3

Figure 4.6. Dot heatmap of top expressed genes across different cell types.

Top 3 differentially expressed genes were plotted for (A) Small intestine. (B) SVF (C) Liver (D) Hypothalamus. Dot color corresponds to -log(fold change) and dot size corresponds to -log(FDR) estimated by Seurat Wilcoxon method (Small intestine, SVF and hypothalamus) or monocle negative binomial GLM modeling (liver).

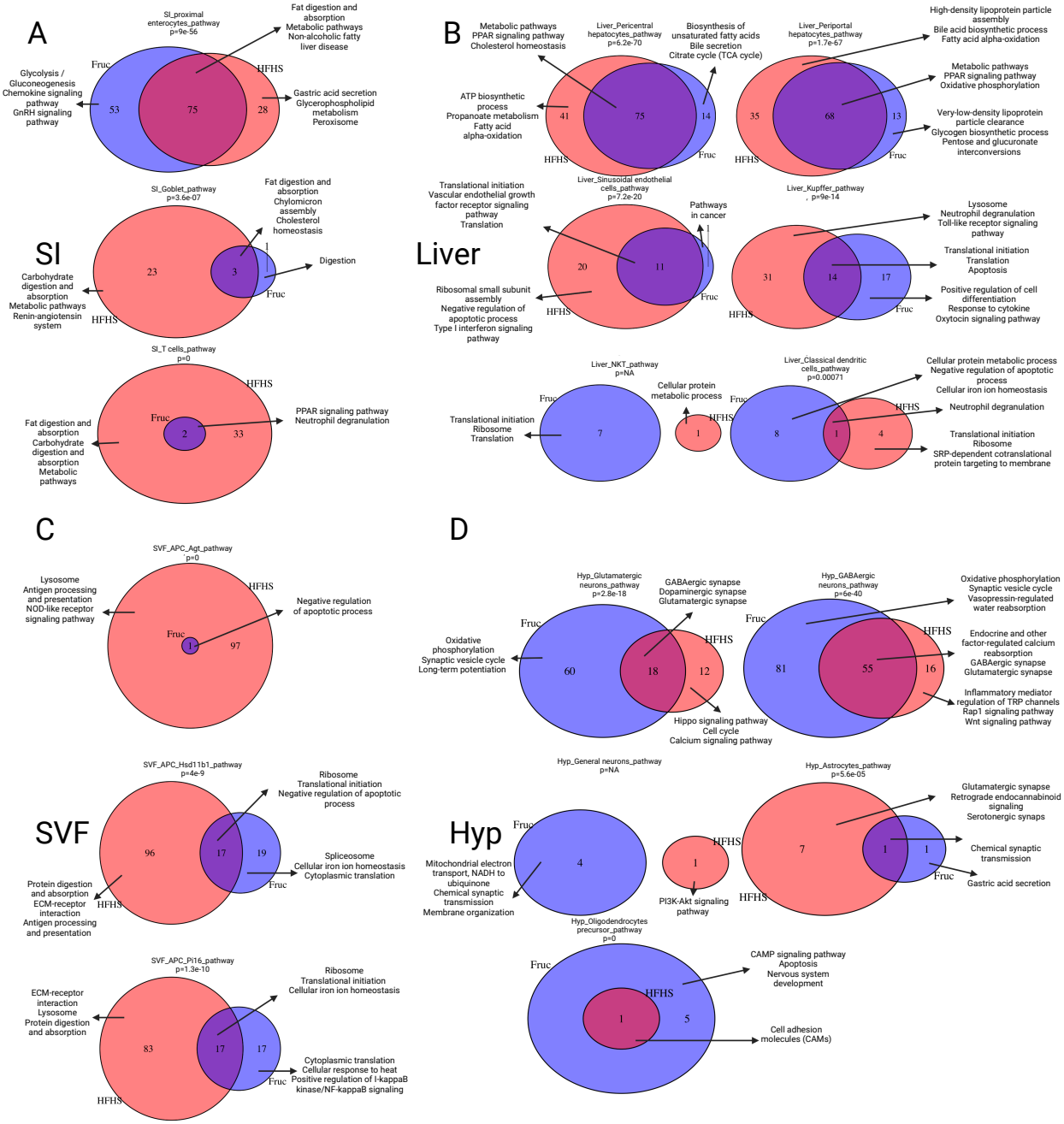


Figure 4.7. Venn diagram of enriched pathways shared in HFHS and fructose diets

(A) Cell types in small intestine (B) cell types in liver (C) cell types in SVF (D) cell types in hypothalamus. Selected representative pathways were shown in text box. Only cell types with at least 1 enriched pathway in both dietary treatments were included, p-value was calculated based on fisher exact test.

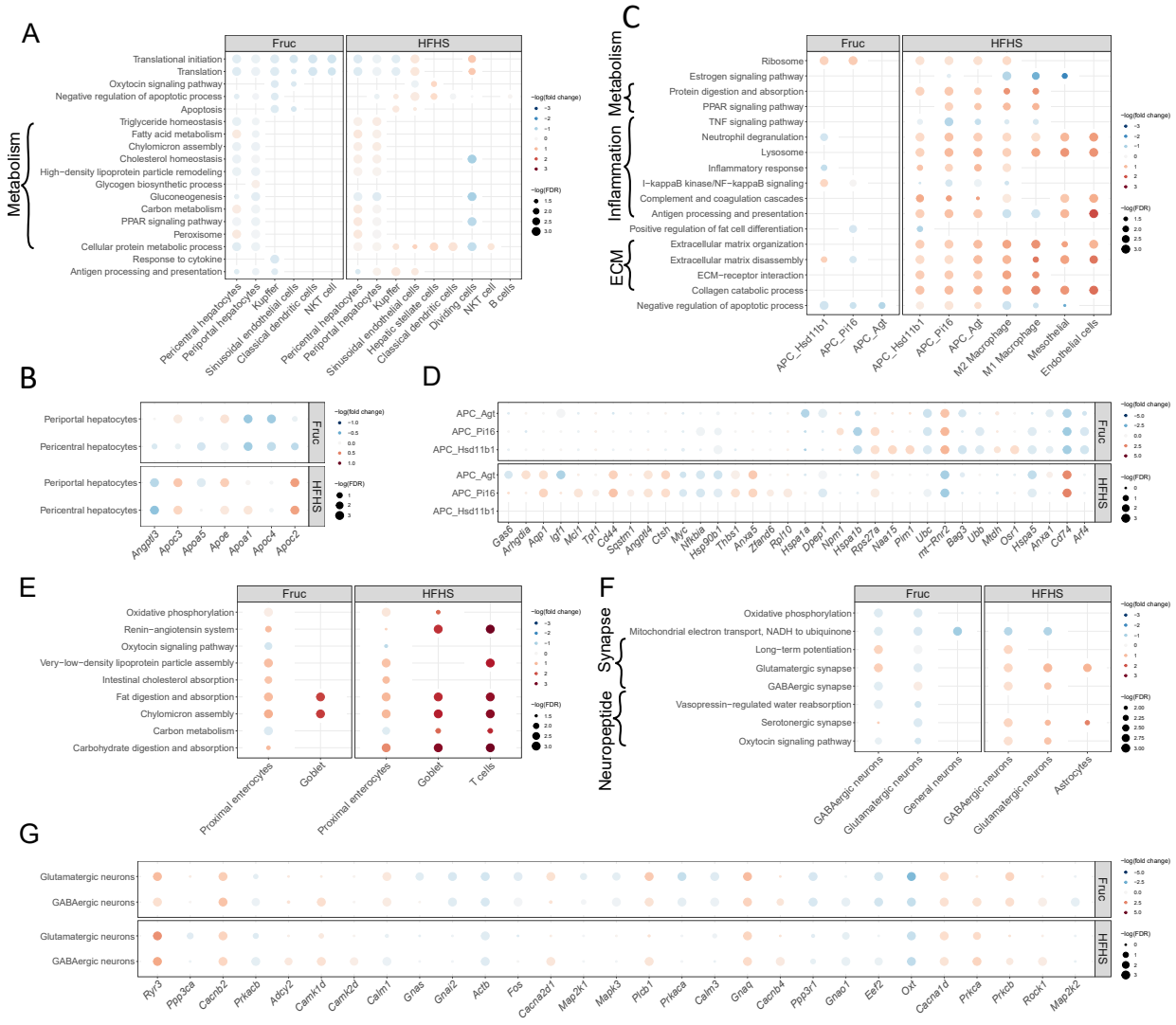


Figure 4.8. Top pathways and genes affected by fructose and HFHS diets in individual cell types.

(A, C, E, F) Pathway dot heatmap of top selected differentially enriched pathways (**B, D, G**) Gene dot heat map showing all overlapped differentially expressed genes from selected pathway. (A) liver (B) pathway “Triglyceride homeostasis” pathway in liver hepatocytes (C) SVF (D) pathway "negative regulation of apoptotic process" in SVF APCs (E) small intestine (F) hypothalamus (G) pathway "Oxytocin signaling pathway" in hypothalamus neurons. For pathway dot heatmaps in A, C, E and F, Dot color corresponds to -log(median fold change of

pathway overlapped genes) and dot size corresponds to -log(FDR) estimated by enrichr pathway database. For dot heatmaps in B, D and G, dot color corresponds to -log(fold change) and dot size corresponds to -log(FDR) estimated by Seurat Wilcoxon method (Small intestine, SVF and hypothalamus) or monocle negative binomial GLM modeling (liver).

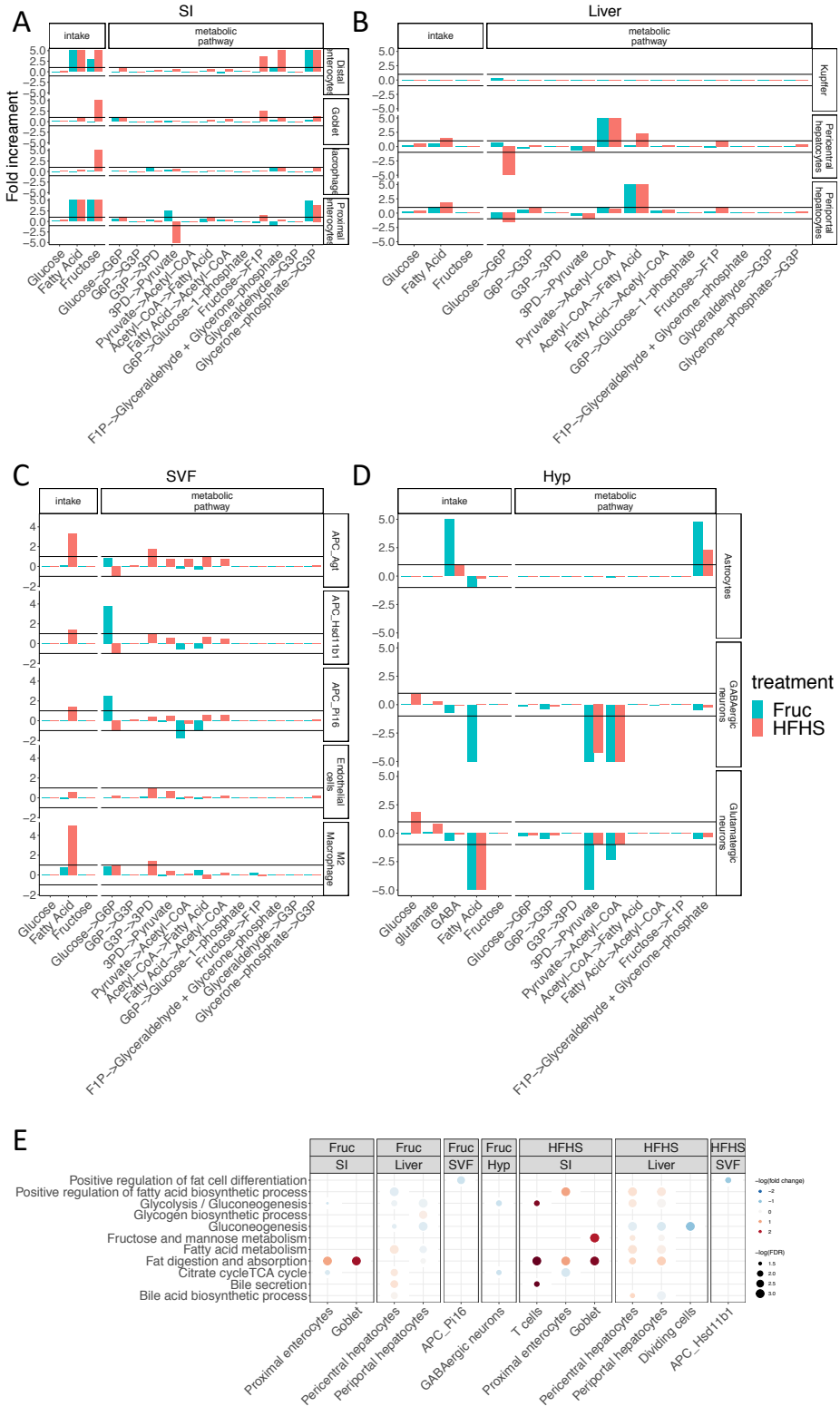

A
SI
intake
metabolic pathway
Fold increament
Distal enterocytes
Goblet
Macrophage
Proximal enterocytes
Glucose
Fatty Acid
Fructose
Glucose-->G6P
G6P-->G3P
G3P-->3PD
3PD-->Pyruvate
Pyruvate-->Acetyl-CoA
Acetyl-CoA-->Fatty Acid
Fatty Acid-->Acetyl-CoA
Acetyl-CoA-->1-phosphate
Fructose-->F1P
Glycerone-phosphate
G6P-->Glucose-1-phosphate
F1P-->Glyceraldehyde + Glycerone-phosphate
Glyceraldehyde-->G3P
Glycerone-phosphate-->G3P
B
Liver
Kupffer
Pericentral hepatocytes
Periportal hepatocytes
C
SVF
APC_Agt
APC_Hsd11b1
APC_Pi16
Endothelial cells
M2 Macrophage
D
Hyp
Astrocytes
GABAergic neurons
Glutamatergic neurons
glutamate
GABA
treatment
Fruc
HFHS
E
Fruc
SI
Liver
SVF
Hyp
HFHS
Positive regulation of fat cell differentiation
Positive regulation of fatty acid biosynthetic process
Glycolysis / Gluconeogenesis
Glycogen biosynthetic process
Gluconeogenesis
Fructose and mannose metabolism
Fatty acid metabolism
Fat digestion and absorption
Citrate cycleTCA cycle
Bile secretion
Bile acid biosynthetic process
Proximal enterocytes
Goblet
Pericentral hepatocytes
Periportal hepatocytes
APC_Pi16
GABAergic neurons
T cells
Dividing cells
APC_Hsd11b1
-log(fold change)
-log(FDR)

Figure 4.9. Metabolic flux analysis inferred cell type specific and dietary specific flux alterations. (A-D) Metabolic flux analysis of (A) Small intestine, (B) Liver, (C) SVF and (D) Hypothalamus. X-axis indicates intake flux estimate of different metabolites as well as metabolic flux estimate for selected glycolysis, fatty acid metabolism and fructose metabolism pathways. Y-axis indicates median fold increment (median(treatment) – median(control))/median(control). Sub-panel labels indicate different metabolic flux (x-axis) or different cell types (y-axis). G6P, Glucose-6-Phosphate; G3P, Glyceraldehyde 3-phosphate; 3PD, 3-Phosphoglycerate; F1P, Fructose-1-Phosphate. (E) Pathway dot heatmap of top selected differentially enriched pathways related to metabolic functions. Dot color corresponds to -log(median fold change of pathway overlapped genes) and dot size corresponds to -log(FDR) estimated by enrichr pathway database.

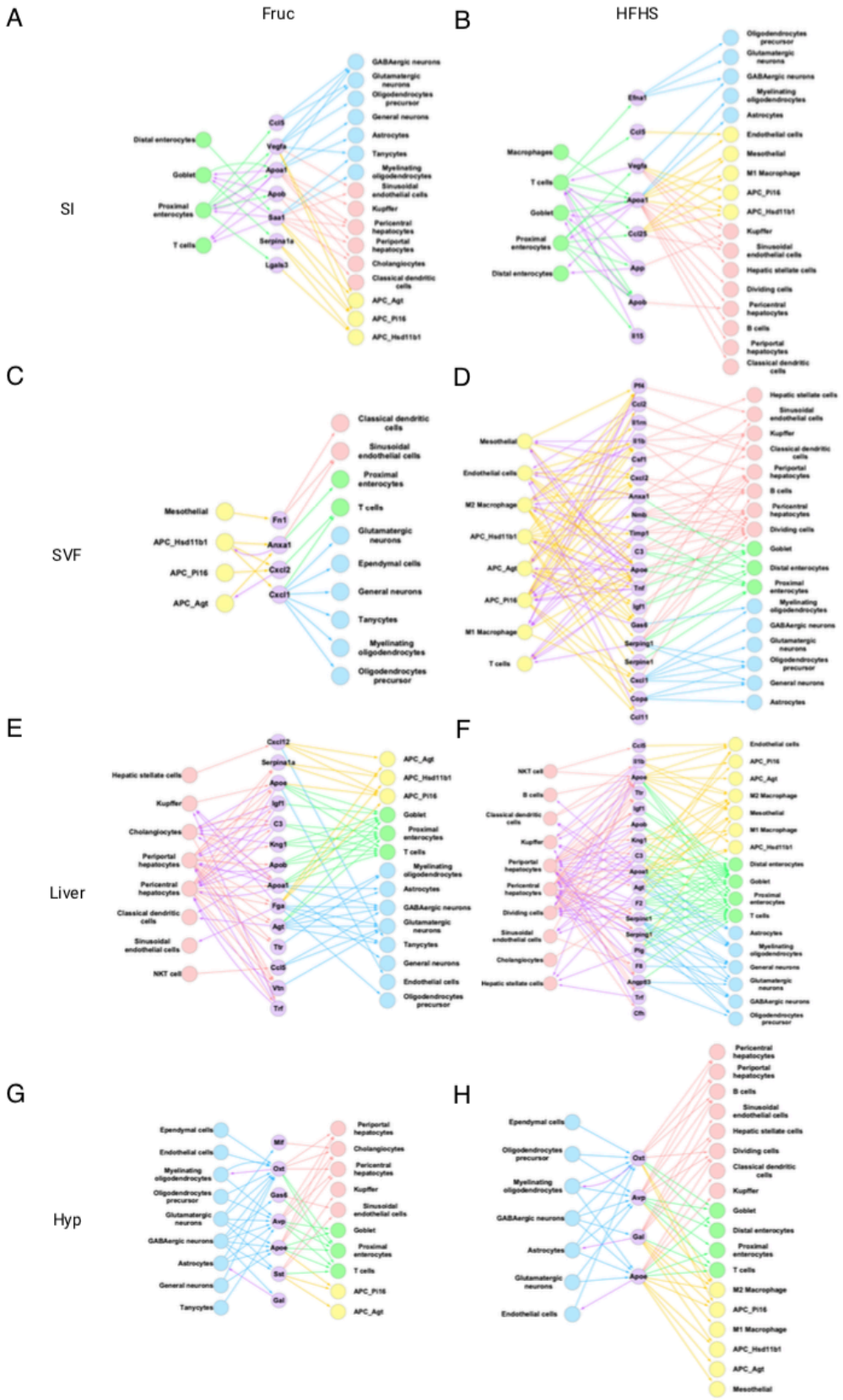

A
Fruc
SI
Distal enterocytes
Goblet
Proximal enterocytes
T cells
Ccl5
Vegfa
Apoa1
Apob
Saa1
Serpina1a
Lgals3
GABAergic neurons
Glutamatergic neurons
Oligodendrocytes precursor
General neurons
Astrocytes
Tanycytes
Myelinating oligodendrocytes
Sinusoidal endothelial cells
Kupffer
Pericentral hepatocytes
Periportal hepatocytes
Cholangiocytes
Classical dendritic cells
APC_Agt
APC_Pi16
APC_Hsd11b1
B
HFHS
Macrophages
T cells
Goblet
Proximal enterocytes
Distal enterocytes
Ehd1
Ccl5
Vegfa
Apoa1
Ccl25
App
Apob
Il15
Oligodendrocytes precursor
Glutamatergic neurons
GABAergic neurons
Myelinating oligodendrocytes
Astrocytes
Endothelial cells
Mesothelial
M1 Macrophage
APC_Pi16
APC_Hsd11b1
Kupffer
Sinusoidal endothelial cells
Hepatic stellate cells
Dividing cells
Pericentral hepatocytes
B cells
Periportal hepatocytes
Classical dendritic cells
C
SVF
Mesothelial
APC_Hsd11b1
APC_Pi16
APC_Agt
Fn1
Anxa1
Cxcl2
Cxcl1
Classical dendritic cells
Sinusoidal endothelial cells
Proximal enterocytes
T cells
Glutamatergic neurons
Ependymal cells
General neurons
Tanycytes
Myelinating oligodendrocytes
Oligodendrocytes precursor
D
Mesothelial
Endothelial cells
M2 Macrophage
APC_Hsd11b1
APC_Agt
APC_Pi16
M1 Macrophage
T cells
Pf4
Ccl2
S100a
S100b
Calr
Cxcl2
Anxa1
Nmb
Timp1
C3
Apoe
Tnf
Igf1
Gas6
Serping1
Serpine1
Cxcl1
Cops8
Ccl11
Hepatic stellate cells
Sinusoidal endothelial cells
Kupffer
Classical dendritic cells
Periportal hepatocytes
B cells
Pericentral hepatocytes
Dividing cells
Goblet
Distal enterocytes
Proximal enterocytes
Myelinating oligodendrocytes
GABAergic neurons
Glutamatergic neurons
Oligodendrocytes precursor
General neurons
Astrocytes
E
Liver
Hepatic stellate cells
Kupffer
Cholangiocytes
Periportal hepatocytes
Pericentral hepatocytes
Classical dendritic cells
Sinusoidal endothelial cells
NKT cell
Cxcl12
Serpina1a
Apoe
Igf1
C3
Kng1
Apob
Apoa1
Fga
Agt
Ttr
Ccl5
Vtn
Tnf
APC_Agt
APC_Hsd11b1
APC_Pi16
Goblet
Proximal enterocytes
T cells
Myelinating oligodendrocytes
Astrocytes
GABAergic neurons
Glutamatergic neurons
Tanycytes
General neurons
Endothelial cells
Oligodendrocytes precursor
F
NKT cell
B cells
Classical dendritic cells
Kupffer
Periportal hepatocytes
Pericentral hepatocytes
Dividing cells
Sinusoidal endothelial cells
Cholangiocytes
Hepatic stellate cells
Ccl5
S100a
Apoe
Ttr
Igf1
Apob
Kng1
C3
Apoa1
Agt
F2
Serpine1
Serping1
Plg
F9
Angptl3
Tnf
Cfh
Endothelial cells
APC_Pi16
APC_Agt
M2 Macrophage
Mesothelial
M1 Macrophage
APC_Hsd11b1
Distal enterocytes
Goblet
Proximal enterocytes
T cells
Astrocytes
Myelinating oligodendrocytes
General neurons
Glutamatergic neurons
GABAergic neurons
Oligodendrocytes precursor
G
Hyp
Ependymal cells
Endothelial cells
Myelinating oligodendrocytes
Oligodendrocytes precursor
Glutamatergic neurons
GABAergic neurons
Astrocytes
General neurons
Tanycytes
Mif
Ost
Gas6
Avp
Apoe
Sst
Gal
Periportal hepatocytes
Cholangiocytes
Pericentral hepatocytes
Kupffer
Sinusoidal endothelial cells
Goblet
Proximal enterocytes
T cells
APC_Pi16
APC_Agt
H
Ependymal cells
Oligodendrocytes precursor
Myelinating oligodendrocytes
GABAergic neurons
Astrocytes
Glutamatergic neurons
Endothelial cells
Ost
Avp
Gal
Apoe
Pericentral hepatocytes
Periportal hepatocytes
B cells
Sinusoidal endothelial cells
Hepatic stellate cells
Dividing cells
Classical dendritic cells
Kupffer
Goblet
Distal enterocytes
Proximal enterocytes
T cells
M2 Macrophage
APC_Pi16
M1 Macrophage
APC_Hsd11b1
APC_Agt
Mesothelial

Figure 4.10. Long range ligand-receptor analysis between small intestine, SVF, liver and hypothalamus.

Ligand receptor interaction network with ligands exhibiting altered expression from small intestine cell types to other cell types after (A) fructose and (B) HFHS treatment. Ligand receptor interaction network with ligands exhibiting altered expression from adipose SVF cell types to other cell types after (C) fructose and (D) HFHS treatment. Ligand receptor interaction network with ligands exhibiting altered expression from liver cell types to other cell types after (E) fructose and (F) HFHS treatment. Ligand receptor interaction network with ligands exhibiting altered expression from hypothalamus cell types to other cell types after (G) fructose and (H) HFHS treatment. Nodes represent ligands (purple color, middle layer) which were DEGs in source cell types (left layer) or target cell types (right layer) with corresponding receptors as DEGs. Cell types of different tissues are labeled with different colors: green - small intestine, blue - hypothalamus, pink – liver, yellow – adipose SVF.

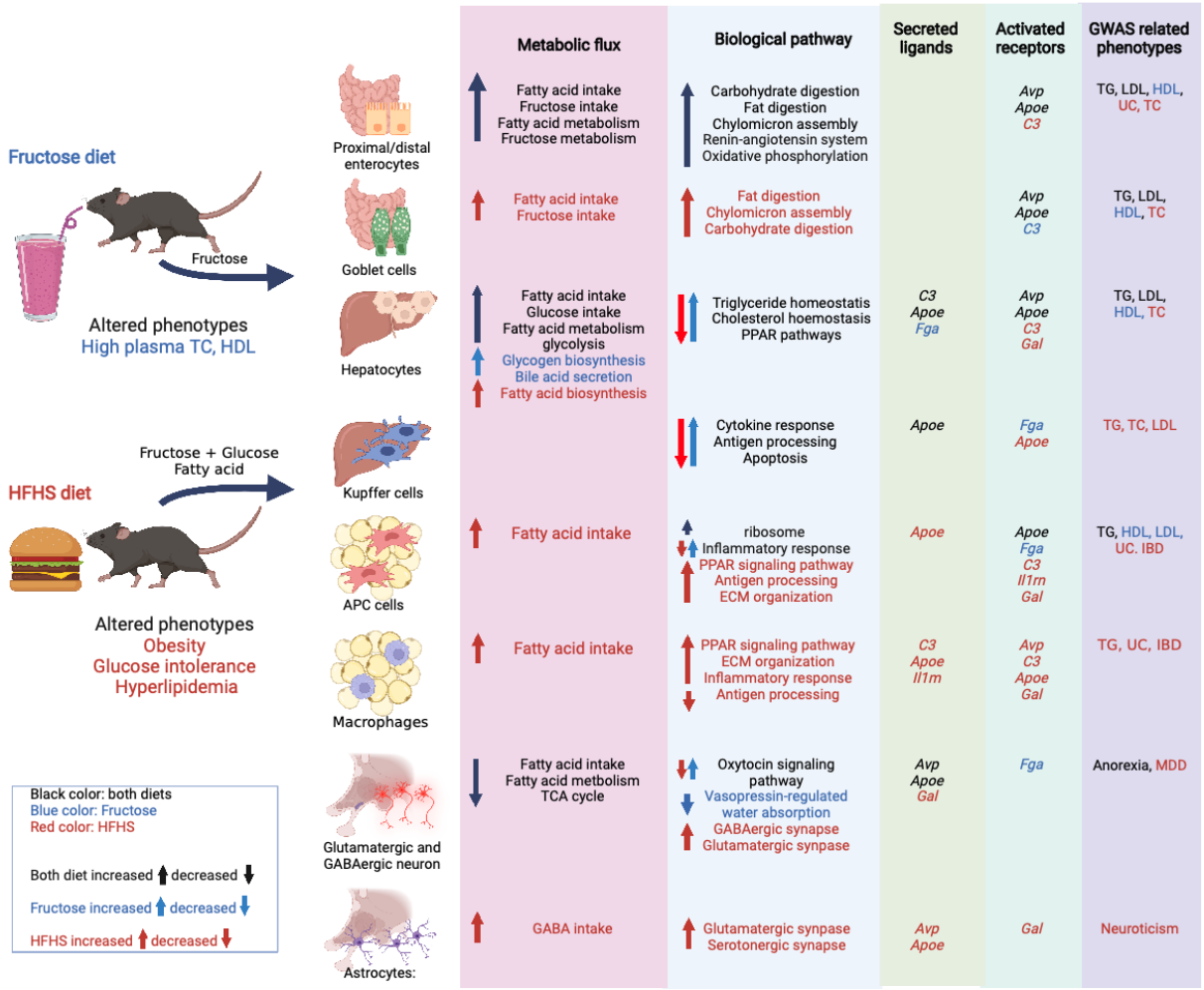


Figure 4.11. Integrated summary of all the analysis.

Summary of phenotype, metabolic flux, biological pathway, secreted ligands, activated receptors and GWAS related phenotypes changes among critical cell types induced by Fructose (blue) and HFHS (red) diets. Directionality of metabolic flux and biological pathway changes were illustrated with arrow.

Chapter 5. Conclusion and future direction

Implementation of systems biology strategies supports the characterization of different exposure effects

Environmental exposures impose significant health burden in human populations, yet the molecular mechanisms are difficult to dissect due to the complex effects on biological systems encompassing molecules, cells, tissues, and species. To meet the challenges, systems biology concepts and approaches have emerged to delineate the complex interaction networks across molecular layers [233,234], cell types [27], and tissues or organ systems [197]. However, the application of systems biology in exposure characterization has been limited. In this dissertation, I applied systems biology concepts with three different projects covering different environmental exposure schemes in daily life, including drug mechanism, transgenerational alcohol exposure, and harmful dietary exposure. Through these projects, I developed a novel gene network-based computational tool to predict drug-disease or drug-toxicity relationships and characterized molecular cross-talks across cell types and tissues under different exposure scenarios. These projects helped derive systems level understanding of genes and pathways affected by pharmaceutical drug entities, alcohol exposure, and dietary exposures in different tissues and species.

PharmOmics as a novel systems biology tool for drug characterization, drug repositioning, and toxicity prediction

In **Chapter 2**, I established a new drug signature database, PharmOmics, for ~1000 pharmaceutical drugs across different drug dosage, tissues, and species. I also demonstrated means to integrate these signatures with network biology to address drug repositioning needs for disease treatment and to predict and characterize toxicity. Finally, this study examined and

validated the concept of matching tissue specificity during drug repositioning, which improved repositioning performance as well as indicating tissue specific mechanism insights. I have established PharmOmics as a potential complementary drug signature database to accelerate drug development and toxicology research.

The PharmOmics system had tremendous potential to be further expanded to improve the breadth and depth of the tool. On breadth side, drug signatures can be expanded with multiomics datasets such as RNA-seq, epigenomics, proteomics, metabolomics, and microbiome datasets from public data repositories or with literature-based signature from comparative toxicogenomics database (CTD). Our initial analysis showed discrepancy between signatures extracted from CTD and transcriptomics studies, thereby supporting the need to incorporate different types of molecular signatures into the PharmOmics framework. In addition to expanding omics layers and types of molecular signatures, tissue specific networks need to be expanded to support broader application of drug repositioning. On depth side, network based repositioning can further be extended to consider drug combinations [235], which should support identifications of drug combinations in disease treatment and toxicities related to drug interactions. Therefore, PharmOmics provided a highly expandable framework which can be actively updated to incorporate multiomics signatures as well as advanced network based analytical algorithms.

Broadening applications of *C. elegans* tool for multi- and trans- generational risk assessment

In **Chapter 3**, the Allard lab developed a single nuclei RNA-sequencing technique fir *C. elegans* in order to capture trans-generational ethanol exposure across different cell types. Through combination of sensitive pipelines, I have shown cell type specific effect of low dose ethanol

exposure across generations. Furthermore, I have identified potential pathways involving mitochondria, ribosome and lipid metabolism dysfunction in germline clusters, which can be further correlated with wormbase related reproductive disorder phenotypes. This project established *C. elegans* as a model for transgenerational toxicity research and, when combined with cell type specific information, it becomes a system that can help pin down key cell types and mechanisms that may be involved in the mammalian response.

Through the application of single nuclei RNA-seq and bioinformatics tools, I was able to provide evidences indicating single nucleus *C. elegans* as a trans-generational risk assessment model. Further analysis involving gene regulatory network methods as well as involving more datasets should be able substantiate this concept. For example, constructing gene regulatory network based on single nuclei datasets can identify potential key regulators which can improve the characterization of the mechanism. Gene regulatory networks can either be using correlation modules such as WGCNA [236] or GENIE3 [237] which do not require prior knowledge. In addition to constructing gene regulatory network from data, adding 1-2 additional dosage points and conducting dose response analysis per cell types would also be helpful to locate pathways mediating low dose ethanol exposure. This strategy was applied previous in Per- and poly-fluoroalkyl substances (PFAS) characterization with in vitro model and successfully obtained mechanistic insights across different PFAS[238]. Finally, our study single nucleus pipeline showed the potential of using C. elegans as a model for trans-generational risk assessment, which can be further expanded to different environmental chemicals such as bisphenol-A and PFAS. Through comparison with different published resources, it will enable me to further understand advantages and disadvantages of this system compared to other in vitro and in vivo models.

Identification of novel metabolic syndrome mechanisms, biomarkers and treatments based on multi-tissue single cell RNA-seq

In **Chapter 4**, single cell RNA-seq was applied across the small intestine, liver, adipose SVF and hypothalamus of two groups of mice that were fed two different MetS related diets. To understand the effects of these diets, I performed multiple complementary analyses. Using the scRNA-seq data, I not only identified differentially expressed genes and pathways and thereby established the roles of individual cell-type specific responses, but also utilized ligand-receptor and metabolic flux analysis to establish an intercellular crosstalk network . I have shown that the use of scRNA-seq data enabled the comprehensive comparison of the effects of fructose and HFHS diets, which induced different metabolic dysfunction and cell-type specific effects. This dataset and findings serve as a rich resource to accelerate future biomarker and mechanistic studies of MetS.

I established a cell-type interaction map for different dietary treatments in order to characterize their relationships with metabolic syndrome. This study revealed cell-type specific responses as well as ligand interactions which can serve as biomarkers and lead to the development of novel treatments. This study can further be expanded through inference of gene regulatory networks through GENIE3 or SCENIC [240]. By integrating gene regulatory network modules associated with identified ligand-receptor interactions, I am able to further identify cell type specific responses and investigate differences between fructose and HFHS diets. Finally, through combination of cell type specific DEGs and cell type specific gene regulatory network through PharmOmics framework, I can further identify treatments targeting different cell types. Our preliminary results showed that even with tissue specific network, we can still find metabolic drugs targeting hepatocytes and anti-inflammatory drugs targeting liver Kupffer cells. This result

can further be extended through construction of cell type specific regulatory networks. Through the identification of cell-type specific drug treatment, I can practice the concept of precision medicine at the cell-type specific level, which has the potential to improve prediction performance compared to bulk level drug prediction.

Systems biology as a future main direction for environmental exposure research

In this dissertation, I have shown the advantages of applying a systems biology strategy to three different exposure scenarios. Through the incorporation of biological networks, I was able to improve prediction of drug-disease relationships, infer molecular mechanisms throughgene regulatory network models, and construct cell-type cross talk models to elucidate inter-tissue communication. I have shown the benefit of applying of novel cutting age high-throughput genomic technologies, such as high throughput sequencing, single cell sequencing, network biology, to the fields of modern toxicology and pharmacology. Through the careful selection of experimental models, high throughput methods, and customized analysis pipelines, I unveiled molecular and cellular mechanisms that were previously unexplored, supported the development of drugs, and characterized the response to toxicants. Given the exploding numbers of chemicals and drugs need to be screened nowadays for their vast mechanisms and toxicities, a more efficient pipeline based on high throughput techniques should be considered. I have demonstrated three scenarios where systems biology strategies performed well compared to traditional methods.

However, systems biology and high throughput techniques are not without drawbacks. They might identify too many potential targets which are hard to validate all of them, even with the conservative tools I have already applied in the analysis. Hence in my future direction, I have proposed different directions to further establish the accuracy and robustness of systems

biology in environmental exposure, including expanding PharmOmics gene signature and algorithm, application of single nuclei C. elegans model with more dose and chemicals, as well as integrating cell type specific response with biological regulatory networks to find cell type specific treatments. Finally, close collaboration with experimental biologist is also critical for developing and validating these concepts. These future directions are potential strategies which can further reduce the bias and improve the applicability of using systems biology methods as risk assessment tools.

Appendix

Appendix A

Published online 28 May 2021 *Nucleic Acids Research, 2021, Vol. 49, Web Server issue* ***W375–W387***
https://doi.org/10.1093/nar/gkab405

Mergeomics 2.0: a web server for multi-omics data integration to elucidate disease networks and predict therapeutics

Jessica Ding [1,2,†], Montgomery Blencowe [1,2,†], Thien Nghiem[1,†], Sung-min Ha[1], Yen-Wei Chen[1,3], Gaoyan Li[1,2] and Xia Yang [1,2,3,4,5,*]

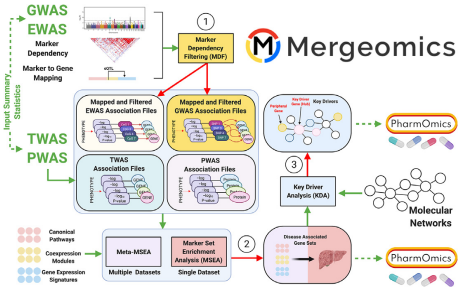
[1]Department of Integrative Biology and Physiology, University of California, Los Angeles, 610 Charles E. Young Drive East, Los Angeles, CA 90095, USA, [2]Interdepartmental Program of Molecular, Cellular and Integrative Physiology, University of California, Los Angeles, 610 Charles E. Young Drive East, Los Angeles, CA 90095, USA, [3]Interdepartmental Program of Molecular Toxicology, University of California, Los Angeles, 610 Charles E. Young Drive East, Los Angeles, CA 90095, USA, [4]Interdepartmental Program of Bioinformatics, University of California, Los Angeles, 610 Charles E. Young Drive East, Los Angeles, CA 90095, USA and [5]Institute for Quantitative and Computational Biosciences, University of California, Los Angeles, 610 Charles E. Young Drive East, Los Angeles, CA 90095, USA

Received February 28, 2021; Revised April 28, 2021; Editorial Decision April 30, 2021; Accepted May 02, 2021

ABSTRACT

The Mergeomics web server is a flexible online tool for multi-omics data integration to derive biological pathways, networks, and key drivers important to disease pathogenesis and is based on the open source Mergeomics R package. The web server takes summary statistics of multi-omics disease association studies (GWAS, EWAS, TWAS, PWAS, etc.) as input and features four functions: Marker Dependency Filtering (MDF) to correct for known dependency between omics markers, Marker Set Enrichment Analysis (MSEA) to detect disease relevant biological processes, Meta-MSEA to examine the consistency of biological processes informed by various omics datasets, and Key Driver Analysis (KDA) to identify essential regulators of disease-associated pathways and networks. The web server has been extensively updated and streamlined in version 2.0 including an overhauled user interface, improved tutorials and results interpretation for each analytical step, inclusion of numerous disease GWAS, functional genomics datasets, and molecular networks to allow for comprehensive omics integrations, increased functionality to decrease user workload, and increased flexibility to cater to user-specific needs. Finally, we have incorporated our newly developed drug repositioning pipeline PharmOmics for prediction of potential drugs targeting disease processes that were identified by Mergeomics. Mergeomics is freely accessible at http://mergeomics.research.idre.ucla.edu and does not require login.

GRAPHICAL ABSTRACT



INTRODUCTION

The advent of omics technologies has made significant strides in unveiling various disease-associated genetic and epigenetic variants, genes, proteins and metabolites. The ever-growing source of multi-omics datasets available including genomics, epigenomics, transcriptomics, proteomics and metabolomics now presents a new challenge of integrating these different data types for more meaningful and holistic interpretation of complex diseases. To conduct a comprehensive investigation of disease pathogenesis, we

*To whom correspondence should be addressed. Tel: +1 310 206 1812; Fax: +1 310 206 9184; Email: xyang123@ucla.edu
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

must consider multiple omics layers that contribute to biological complexity (1). The computational pipeline Mergeomics was developed to meet the need for multi-omics integration and functional interpretation to obtain mechanistic understanding. Mergeomics provides flexibility to incorporate the full spectrum of summary statistics (not just top hits) of individual layers of omics or multi-omics data simultaneously along with diverse functional genomics data across data types, studies and species. As such, genome-wide association studies (GWAS) as well as epigenome- (EWAS), transcriptome- (TWAS), proteome- (PWAS) and metabolome-wide association studies (MWAS) can all be accommodated.

The development of our Mergeomics tool follows the philosophy of utilizing a systems biology approach to unravel the complex interactions across molecular domains as well as cell types, tissues and organ systems that occur in disease. In particular, we are guided by the omnigenic disease model (2), which states that a large proportion of the genome likely contributes to disease pathogenesis through molecular interactions both within and between tissues. Utilizing this data-driven analysis considering the interactions among different omics layers and tissue contexts will uncover global maps to identify critical targets in disease pathogenesis, which can be followed by experimental approaches to investigate the detailed events that occur through the predicted molecules or pathways.

With the abundance of omics data available, it is unsurprising that various tools or methods have been developed to better integrate and interpret these datasets (3–5). These tools can be broadly categorized into two application categories: multi-omics biomarker predictions of diseases or subtypes (i.e. uncovering correlative or predictive but not necessarily disease-causing features) or mechanistic understanding of disease pathogenesis (i.e. regulators, molecular interactions and processes involved in disease development). Mergeomics focuses on mechanistic modeling but not predictive modeling. In terms of approaches, fusion (such as PFA (6), SNF (7), PSDF (8)), Bayesian (e.g. iCluster (9), PSDF (8), BCC (10)), correlation, multivariate (e.g. MFA (11), IntegrOmics (12), MixOmics (13)), pathway and network methods (PARADIGM (14), SNF (7), iOmicsPASS (15), MiBiOmics (16), Lemon-Tree (17), PaintOmics (18), NetICS (19), Metascape (20)) have been implemented (3–5). Mergeomics falls within the network method category that mainly focuses on understanding disease pathogenesis through uncovering multiple molecular targets within biological processes important to disease. The benefit of a network approach over other integrative options is its ability to provide biological interpretability, which is reliant not on the identification of latent structures through mathematical deconvolution but on the utilization of prior information based on molecular interactions, which can help provide clear targetable options (e.g. genes) in disease. Compared to other tools, Mergeomics not only accommodates diverse data types (GWAS, EWAS, TWAS, PWAS, MWAS) from different sources, studies, or species for a given disease, but also considers relationships between omics layers through functional genomics such as expression quantitative trait loci (eQTLs), molecular pathways, and tissue-specific gene regulatory networks to derive disease networks and predict therapeutics. Mergeomics also uses full summary statistics, not raw data or lists of top associations, as input, thereby reducing the need for raw data processing and harmonization and for pre-determining a specific cutoff to call for significant markers. Mergeomics has the ability to conduct pathway analysis and model gene regulatory networks, protein-protein interaction networks, and transcription factor networks in order to predict and visualize network regulators of disease. These unique features help maximize the utility of existing datasets and overcome limitations of other tools which utilize a narrower range of multi-omics data sources, do not provide mechanistic interpretations, or require programming skills with no intuitive web server for ease of use.

Since the release of the open source Mergeomics R package (https://bioconductor.org/packages/release/bioc/html/Mergeomics.html) (21) and web server in 2016 (22), this tool has been used to model a diverse set of diseases including cardiometabolic disorders such as non-alcoholic fatty liver disease (23), cardiovascular disease (24–26) and type 2 diabetes (27), autoimmunity including psoriasis (28) and rheumatoid arthritis (29), alcohol dependence (30), brain injury (31), Sjogren's syndrome (32) and environmental contributions to disease (33–35). Importantly, multiple validations of molecular predictions from Mergeomics with *in silico*, *in vitro* and *in vivo* studies highlight the validity and causal nature of the disease network predictions (23,27–28,31,35–40). Due to increasing demand for multi-omics integration and interpretation from scientists with different areas of expertise, we have implemented major revisions and improvements on the Mergeomics web server. Specifically, we have redesigned the user interface, simplified workflows, offered detailed tutorials and case studies, and provided more datasets and network models for utilization. The Mergeomics 2.0 web server offers the scientific community much-improved accessibility to our pipeline, caters to each user's specific goals in multi-omics studies, and addresses a broad range of biological questions, particularly emphasizing a mechanistic understanding of disease pathogenesis and prediction of potential therapeutics based on mechanistic understanding.

OVERVIEW AND UPDATES ON THE CORE FUNCTIONS OF MERGEOMICS

Overview of core functions

Mergeomics 2.0 features four core functions as previously implemented in version 1.0 with an addition of a new function. First, we provide a preprocessing tool, Marker Dependency Filtering (MDF) to remove omics marker redundancies such as linkage disequilibrium (LD) between single nucleotide polymorphisms (SNPs). Second, Marker Set Enrichment Analysis (MSEA) is used to identify omics-informed disease processes through the integrations of omics markers such as SNPs with functional genomics, canonical pathways, or co-expression networks. Third, Meta-MSEA runs MSEA on multiple datasets and conducts pathway/network level meta-analysis to retrieve consistent disease processes informed across datasets. Fourth, Key Driver Analysis (KDA) pinpoints network regulators

of disease processes based on the topology of biological networks. In Mergeomics 2.0, we added a new functional module called PharmOmics, which takes as input multi-omics-informed disease pathways or networks from Mergeomics to match with drug signatures to predict potential therapeutic drugs.

Introduction of PharmOmics into Mergeomics 2.0

We have recently developed a novel species- and tissue-specific network-based drug repositioning tool, PharmOmics, which is based on *in vivo* molecular studies of drugs (41). PharmOmics is a complementary drug repositioning tool to other existing tools, such as CMap (42) and LINCS L1000 (43), which are mostly based on *in vitro* cell line data. We provide two drug repositioning methods: network-based drug repositioning and gene overlap-based drug repositioning. Network-based drug repositioning ranks drugs based on the degree of connectivity of genes influenced by drug treatments to disease gene signatures in a given gene network model (44). Gene overlap-based drug repositioning is based on the degree of direct overlap between drug genes and disease genes. Users can directly input their disease pathway results from MSEA (genes from disease pathways are used as input) or KDA (genes from the disease network or significant key drivers (KDs) are used as input). For both MSEA and KDA, specific gene sets can be input into drug repositioning for a more refined analysis. As PharmOmics is based on gene expression studies, inputs are limited to genes or proteins. Users can also input their genes of interest into PharmOmics for drug repositioning analysis without running any other functions in Mergeomics.

Flexible workflows using the core functions

Each of the main functions of Mergeomics described above can be utilized as a standalone analysis tool or can be combined into a multi-step workflow with several different cases as portrayed in Figure 1. There are four cases or starting points that a user has the option to select. In case one, the user has one GWAS dataset and is prompted first to run MDF where they provide their association dataset, mapping data (e.g. SNP to gene), and marker dependency data (LD in the case of GWAS) to retrieve corrected SNP associations and mapping files. The MDF step is optional if the user does not wish to correct for LD, although we highly recommend this correction to avoid statistical artefacts due to LD. These results along with a gene set are fed into MSEA to uncover disease-associated pathways, which can be further analyzed in KDA to identify key regulators or PharmOmics for drug repositioning. In case two, the user has EWAS, TWAS, PWAS or MWAS data, and they are led to MSEA, where MDF and marker mapping are optional. As in the GWAS path, results from MSEA can be carried to KDA or PharmOmics. In case three, the user has multiple omics datasets and utilizes Meta-MSEA, which will run MSEA on each dataset and then conduct a meta-analysis across datasets to retrieve consistent biological processes, which can be input into PharmOmics or KDA. Finally, in case four, the user has a gene set and network of interest and can directly run KDA, which will provide KD genes and a subnetwork visualization of the top KDs, and the KDs or subnetwork can be input into PharmOmics to predict drugs.

Update on Marker Dependency Filtering (MDF)

MDF prepares input files for MSEA by correcting for dependency between omics markers and is an optional function. This preprocessing step is most commonly used for GWAS data to correct for LD between SNPs and filter out redundant SNPs, which is critical for removing redundant association signals that can result in statistical and biological artefacts in downstream analysis. Another purpose of MDF is to link the SNPs to potential downstream genes based on functional evidence, such as tissue-specific eQTLs. Correcting for dependency between other omics markers is currently seldom used. However, this feature can be utilized to correct for dependency between other types of markers (methylation sites, transcripts, etc.), if desired. MDF uses as input an association file which details markers (e.g. SNPs) and their disease association strengths (e.g. $-\log_{10}$ *P*-values or effect size, note that *P*-values are prohibited as MDF ranks larger values as stronger association strength, which is opposite of *P*-values), a mapping file used for marker to gene mapping (e.g. SNPs are mapped to genes to be enriched for gene sets), and a marker dependency file indicating the dependency between markers (e.g. LD between SNPs, to remove redundant markers) (Figure 2). The resulting corrected association and mapping files are then used as input to MSEA. MDF also allows for the selection of a top percentage of markers (50% or 25% recommended) to be considered in the analysis which reduces noise from low signal markers.

Updates to MDF include an increased number of marker to gene mapping options such as the addition of all available tissue-specific Genotype-Tissue Expression project (GTEx) (45) cis-eQTLs and splicing QTLs (cis-sQTLs) (Table 1), the ability to combine up to five mapping options, and the inclusion of LD files for all 26 populations from 1000 Genomes (1000G) (46) and methylation disequilibrium from EWAS software 2.0 (47). For analysis starting from GWAS data, MDF is a default preprocessing step, but we have included the option to skip MDF. For analytical paths starting from other omics data, users have the option to add MDF if needed.

Update on Marker Set Enrichment Analysis (MSEA)

In MSEA, full summary statistics of omics markers such as SNPs from GWAS, epigenetic sites from EWAS, genes from TWAS, proteins from PWAS, or metabolites from MWAS and their disease association values are taken as input and are integrated with functional genomics, canonical pathways, or co-expression networks to retrieve disease-associated pathways and networks. MSEA calculates and summarizes enrichment of disease/trait omics markers in sets of functionally related genes, such as canonical pathways and co-expression networks, across a range of statistical cutoffs in the full summary statistics using a chi-square-like statistic and then uses permutation to determine statistical *P*-values for the enrichment. We emphasize the importance to provide the association strength of the given

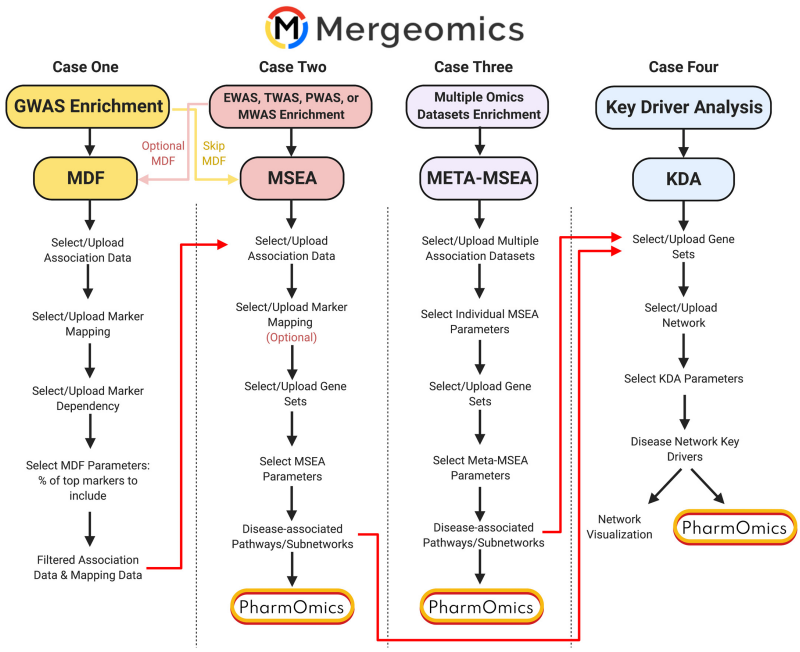


Figure 1. Workflow of Mergeomics. We provide four options on the web server to tailor to the user's data type. Case One: Individual GWAS analysis. For GWAS datasets we advise utilizing the MDF function; however, we also provide the ability to skip MDF and directly run MSEA and follow the workflow to PharmOmics or KDA. Case Two: Individual EWAS, TWAS, PWAS or MWAS analysis. In this case, we directly start at MSEA without MDF; however, we also provide the ability to utilize the MDF function if needed. From here the user can feed the MSEA results into PharmOmics or KDA. Case Three: Multi-omics analysis. If the user has multiple omics of the same type (e.g. two GWAS) or different types (e.g. TWAS and EWAS), they can utilize the Meta-MSEA function to derive disease-associated pathways and can input their results into PharmOmics or KDA. Case Four: A gene list(s) to run KDA. The user in this case can upload their gene sets of interest and upload or select a network to derive KD genes and visualize top KD subnetworks. The disease subnetwork or significant KDs can be fed into PharmOmics for drug repositioning.

marker wherein a larger number reflects greater association such as $-\log_{10}$ *P*-values or effect size to avoid incorrect downstream analysis and interpretability.

MSEA is able to analyze diverse data types, and each has different considerations of inputs which was partly described in the above MDF section (Figure 2). The output from MSEA can be interpreted as omics-informed disease pathways or networks. If GWAS is used, MSEA results can imply causal disease processes since GWAS carries causal inference. For other omics data, the MSEA results can only be interpreted as disease-associated processes but may or may not be causal. Considering GWAS along with other omics data, in our opinion, is a useful way to identify causal genes and processes. We also advise the user to take care in their interpretation of the names or annotations of pathways deemed to be significant (FDR < 0.05) as some can be misleading. Attention to the genes enriched in a given pathway derived from the input dataset should be checked in the gene details output file to confirm whether the pathway name is indeed appropriate as the genes may be more suitable or representative of another biological process. A user can conclude the analysis with results from MSEA or use the MSEA results as input to KDA with a user-defined statistical cutoff to identify network KDs of the disease processes based on molecular network topology.

In Mergeomics 2.0, we added the ability to use disease-associated gene sets derived from MSEA as input to PharmOmics for drug repositioning analysis, selecting either specific gene sets or by false discovery rate (FDR) or *P*-value threshold, to pinpoint drugs whose gene signatures align with those of the disease-associated gene sets identified by MSEA.

Update on Meta-MSEA

Meta-MSEA allows for integration of multiple datasets of the same omics type (e.g. two or more GWAS datasets) or multiple omics types (e.g. GWAS, EWAS, TWAS) and runs MSEA for each omics dataset followed by a meta-analysis. This integration reveals consistencies and differences in biological perturbation across different omics types or different studies of the same omics type.

In Mergeomics 2.0, we improved the guidance of running Meta-MSEA in regard to the differences in preprocessing of the different types of omics data. In addition, we have increased the flexibility of this analysis to allow for spe-

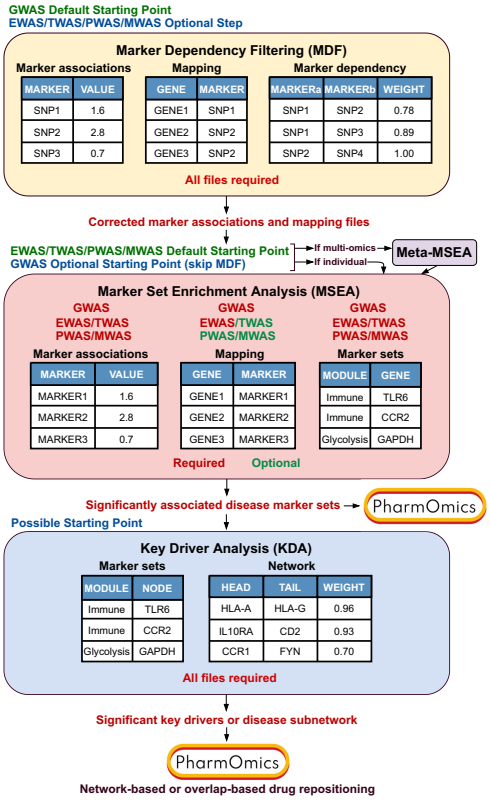


Figure 2. Mergeomics pipeline inputs. MDF is the default starting point for GWAS analysis and is an optional step for EWAS/TWAS/PWAS/MWAS. MDF requires marker-disease associations, a marker-gene mapping file, and a marker dependency file. Users with GWAS data can also skip MDF and run MSEA directly. MDF produces corrected marker-disease associations and marker-gene mapping files containing independent markers that are used for MSEA. For MSEA, required files for all datasets are the marker-disease associations and marker sets (pathway/modules). The marker to gene mapping file is required for GWAS and EWAS and optional for MWAS, TWAS and PWAS. Disease-associated marker sets from MSEA can be fed into KDA, which requires gene sets and a network. KDA can also be a starting point of analysis. Disease-associated gene sets from MSEA or KDs and disease subnetwork from KDA can be fed into PharmOmics for drug repositioning.

cific inputs and parameters for each association data. After each individual omics dataset is added, the user will be able to review which datasets have been successfully uploaded and their individual MSEA parameters with the option to add additional datasets or delete certain datasets, providing an easy way to track all the different inputs. As in the results generated from the individual MSEA, significantly associated gene sets from Meta-MSEA can be used as input to KDA or PharmOmics drug repositioning. We have also implemented user-defined individual MSEA FDR cutoffs to KDA in that the disease-associated pathways must pass all individual MSEA FDR cutoffs as well as the meta-FDR to be used in KDA, allowing the user to focus on the most consistent and robust disease processes across different datasets. In addition, we now provide heterogeneity statistics from Cochran's Q test to indicate the variability between datasets.

Update on Key Driver Analysis (KDA)

KDA identifies essential regulators of disease-associated pathways and networks, which are then visualized in the web browser using Cytoscape.js (Figure 3). KDA results can also be downloaded as network files ready to be used on Cytoscape Desktop for further customization of the network visualization. A Chi-square-like statistic, $\chi = \frac{O-E}{\sqrt{E-\kappa}}$, is used to identify genes (KDs) that are connected to a significantly larger number of disease-associated genes than what is expected by random chance. O and E represent the observed and expected numbers of disease-associated genes in a hub subnetwork, and E is estimated by $\frac{N_k N_p}{N}$ where N_p is the disease gene set size, N_k is the hub degree, and N is the full network order. KDs represent prioritized disease regulatory genes based on network topology. In numerous recent applications of Mergeomics, top KDs have been shown to be causal for diseases based on experimental evidence (23,27,36), thereby supporting their importance. KDA can be utilized as a follow up analysis to MSEA or Meta-MSEA, and it can also be used as an independent analysis using a gene list of interest and a given network as inputs. For instance, the user can upload a list of curated disease genes and choose or upload a relevant network to run KDA to identify how the disease genes interact in the network and whether there are key hub nodes in the network that regulate the disease genes.

In Mergeomics 2.0, we added the ability to visualize input gene overlap with a given network, if any, in the case that no KDs were found. The user can therefore be better informed on the reason for the lack of KD hits based on the distribution and connectivity of the input genes in the network. If few input genes are in the network or the input genes are widely dispersed in the network, KDs may not be identified. We have additionally increased the number of sample tissue-specific networks (Table 1). As we have done similarly with MSEA and Meta-MSEA, disease subnetworks or significant KDs from KDA can be used directly for PharmOmics drug repositioning, and users can further customize which processes in the subnetwork are used in drug repositioning for a more focused analysis.

DATA AND SAMPLE INPUT UPDATES

We have significantly augmented the amount of Mergeomics-ready sample files with commonly used datasets and will continue to actively update sample files to enrich data resources on a monthly basis.

In Mergeomics 2.0, we include over 20 GWAS datasets from a broader range of diseases from metabolic syndrome

Table 1. Sample resources on Mergeomics web server. Complete list in Supplementary Tables S1–S4

General data category	Data type	Specifics	Citation
Association data	GWAS	Alzheimer's disease	(71)
		Attention deficit hyperactivity disorder	(72)
		Alcohol dependence	(73)
		Body mass index	(74)
		Breast cancer	(75)
		Coronary artery disease	(76)
		Fasting glucose	(77)
		Heart failure	(78)
		High density lipoproteins (HDL)	(79)
		Low density lipoproteins (LDL)	(79)
		Major depressive disorder	(80)
		Parental lifespan	(81)
		Parkinson's disease	(82)
		Psoriasis	(83)
		Severe illness in COVID-19	(84)
		Schizophrenia	(85)
		Stroke	(86)
		Systemic lupus erythematosus	(87)
		Type 2 diabetes	(88)
		Total cholesterol	(79)
		Triglycerides	(79)
	EWAS	Birth weight	(89)
		Maternal anxiety	(90)
		Social communication	(91)
		Psoriasis	(62,63)
Marker mapping	Chromosomal distance	10kb, 20kb, 50kb	(46)
	Regulome	RegulomeDB (ENCODE)	(92)
	eQTL	49 tissue types	(45)
	sQTL	49 tissue types	(45)
Marker dependency	Linkage disequilibrium	26 populations at $r^2 > 0.5$ and >0.7	(46)
	Methylation disequilibrium	$r^2 > 0.5$	(47)
Marker sets	Canonical (knowledge based)	KEGG	(50)
		Reactome	(51)
		BioCarta	(52)
		MSigDB	(49)
		GO	(53)
		BioPlanet	(55)
		WikiPathways	(54)
	Data-driven (co-expression)	24 tissue specific modules (WGCNA/MEGENA)	(45,56–57)
Networks	Gene regulatory human and mouse composite (Bayesian)	Adipose, blood, brain, kidney, liver, muscle	(58,93–98)
	Gene regulatory (GIANT)	Adipose, blood, brain, kidney, liver, muscle	(61)
	Protein-protein interaction	STRING	(59)
	Transcription factor-target (FANTOM5)	Adipose, blood, brain, kidney, liver, muscle	(60)

to psychiatric disorders (Table 1; detailed data sources and links in Supplementary Table S1). For omics dependency filtering options, we have added the full array of LD data from 26 human populations studied in 1000G (46) with LD above 0.5 and 0.7 for SNP filtering to remove redundant SNPs in high LD and have also provided an example methylation disequilibrium data file for correction of EWAS data. For SNP to gene mapping options, we have added all tissue-specific cis-eQTL and cis-sQTL mapping files from the GTEx version 8 (q-value < 0.05) (45), which inform on the SNPs associated with gene expression level changes (eQTL) or differential splicing (sQTL). In addition, we offer ENCODE regulatory gene mapping (48) and various chromosomal location-based mapping options (Table 1; Supplementary Table S2). Moreover, we have increased the number of curated pathways from version 1 to include all gene sets from Molecular Signatures Database (MSigDB) (49) such as KEGG (50), Reactome (51), Biocarta (52) canonical pathways, chemical and genetic perturbation, microRNA and transcription factor targets, and cell type marker signatures, Gene Ontology (53), Wikipathways (54) and Bioplanet (55), among others (Table 1; Supplementary Table S3). To complement knowledge-based pathways, we include our data-driven tissue-specific co-expression network modules utilizing GTEx transcriptome datasets and co-expression network construction tools MEGENA (56) and WGCNA (57) (Table 1; details of data sources, methods, and parameters used to construct networks in Supplementary Table S3). Finally, we have constructed tissue-specific Bayesian gene regulatory networks (58) and include them as sample networks on the web server. We also provide human protein-protein interaction networks (59), transcription factor networks (60) and GIANT networks (61) (Table 1; Supplementary Table S4). Sample files are available to download from our sample resources page (http://mergeomics.research.idre.ucla.edu/samplefiles.php), and further clarification on correct formatting of input data is detailed on the web server and in Figure 2.

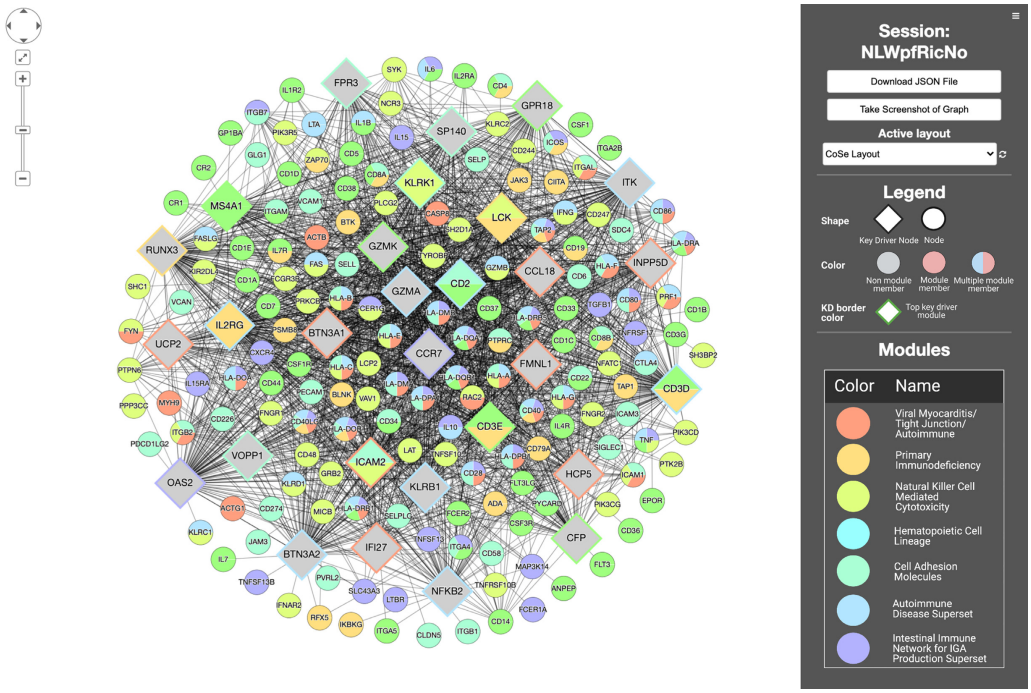


Figure 3. Top KDs network visualization. Screenshot of the in-browser interactive network visualization (using Cytoscape.js) directed from the KDA results page. The colors of the nodes represent member genes of a disease-associated pathway. The diamond shaped nodes represent KD genes, where the border color represents the top pathway that is regulated by the KD. If a node has multiple colors, it is part of two or more disease-associated pathways, and if a node is grey, it does not belong to the disease pathways (non-member genes) but is present in the input network.

GENERAL UPDATES

We have completely redesigned the user interface for a much more intuitive guidance of the use of the pipeline for different omics data types. To start the pipeline, users are presented with four workflow options in regard to their data: (i) GWAS, (ii) EWAS, TWAS, PWAS or MWAS, (iii) multiple of the same or different types of omics data and (iv) a gene set list (user can run KDA or PharmOmics). The separation of GWAS from other omics datasets is for the additional need to correct for LD and link SNPs to candidate genes through MDF, which is not required or is optional for other omics datasets. For EWAS, a marker to gene mapping file is required if the user uploads epigenetic markers such as CpG probes. For MWAS, a metabolite to gene mapping file is optional but not required if the user uses metabolite sets as the marker sets to be tested. Marker mapping is not needed for TWAS and PWAS as the markers (genes and proteins) match the gene sets. This workflow design clearly delineates what is needed for each specific data type, which is more intuitive for the user. We have also improved the fluidity and presentation of the pipeline workflow as each collapsible step appears below the previous in a vertical format so that the user can revisit input files, parameters, and results of previous steps in the pipeline and choose to rerun a step at any point in the pipeline. A workflow map with navigation links is also generated on the left sidebar to help visualize the steps taken and downstream paths.

We have improved the system that allows users to return to their session where results of analyses can be revisited or continued onto the next step using a unique tracking ID number that is valid for up to 48 hours after the start of their session. The user can also choose to have their results emailed upon completion of the analysis, which is not mandatory but is recommended because the tracking ID allows the user to reload their session and retrieve completed jobs in case a crash occurs. Because later steps of the pipeline, KDA and PharmOmics, can be run independently, downloadable result files from MSEA and KDA can be uploaded directly to the desired next step in the analysis (e.g. MSEA to KDA/PharmOmics or KDA to PharmOmics).

In addition, we have improved case-specific responsiveness of the web server to better inform the user such as error-checking of user uploaded files to ensure the file is formatted correctly and providing feedback on user results such as whether the results are substantial enough to be used in the next step of the analysis. Across all applications of Merge-

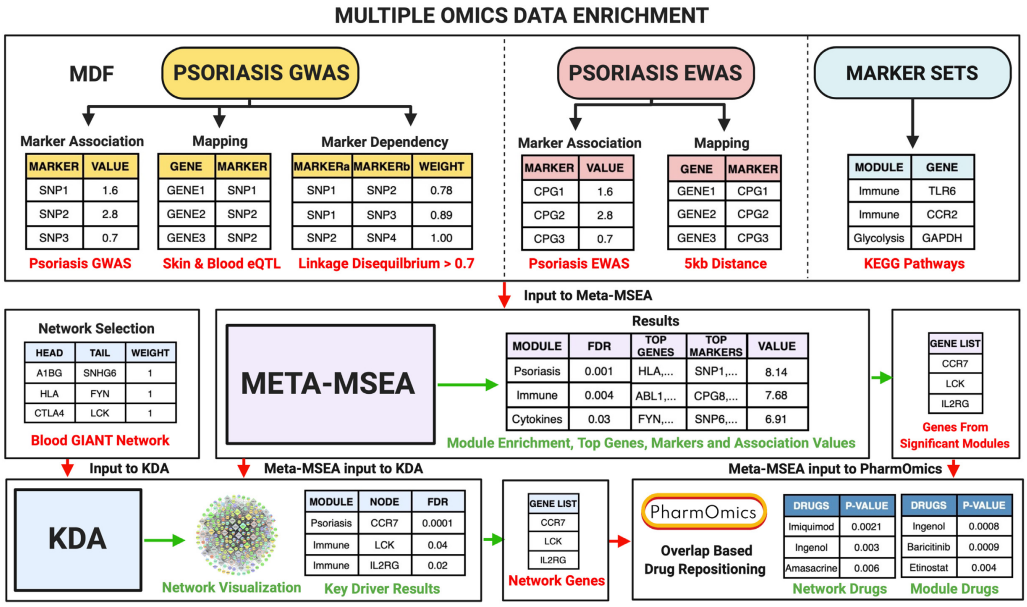


Figure 4. Meta-MSEA use case study overview. To showcase the function and output of the web server, we utilized multiple human psoriasis GWAS and EWAS data and ran the multiple omics data workflow (Case 3 in Figure 1, Meta-MSEA). Firstly, we uploaded the psoriasis GWAS data, mapped the SNPs to genes using a combined skin and blood eQTL file, and filtered for LD > 0.7 to remove redundant SNPs in LD. Next, we uploaded our psoriasis EWAS association datasets and mapped the CpG sites to genes based on a 5 kb distance. Finally, we uploaded KEGG pathways with a psoriasis control set. Pathway enrichment results are produced, and each pathway's top genes, markers, and corresponding association values are displayed. Psoriasis-associated pathways are used as input into KDA as well as PharmOmics drug repositioning (using genes from significant pathways/modules). In the KDA, along with the Meta-MSEA input, we chose the blood GIANT network option and ran the KDA providing KD results and visualization (Figure 3) and additionally utilized the network genes as an input into PharmOmics. Finally, two sets of drug repositioning results were produced using gene overlap-based drug repositioning in PharmOmics: one based on the genes of significant pathways from the Meta-MSEA results and the other based on the KDA subnetwork genes.

omics 2.0 we have provided an improved review of analysis inputs and parameters and new interactive tables with pagination, sorting, and search features (Figure 5). We also implemented real-time runtime analysis output and progress updates, and this job log including any errors that occurred is available for download at the conclusion of the analysis. Finally, we have improved multi-device usage including on tablets and phones such that it can be appropriately viewed on different screen sizes. We further improved the tutorial to explain input file preparation, parameter setting, and the underlying methods of each computational function and provide video tutorials to demonstrate the different pipeline options.

USE CASE: IDENTIFYING PATHOGENIC PATHWAYS AND NETWORKS FOR PSORIASIS BASED ON MULTI-OMICS DATA

The use case described here utilizes publicly available GWAS and EWAS data to perform Meta-MSEA and subsequently KDA to find pathogenic pathways and regulators of psoriasis (Figure 4). All data used in this example are provided as sample data on the web server which can be downloaded (http://mergeomics.research.idre.ucla.edu/samplefiles.php). GWAS of psoriasis was obtained from dbGAP database (www.ncbi.nlm.nih.gov/gap) with accession phs000019.v1.p1, and two EWAS of psoriasis were obtained from GEO (GSE31835 and GSE63315) (62,63). For preprocessing of the GWAS data, we use the top 50% of SNPs ranked by $-\log_{10}$ *P*-value and correct for LD between SNPs using MDF with the psoriasis GWAS summary statistics as the marker associations, combined skin and blood eQTLs as the SNP to gene mapping, and the 1000G CEU LD structure containing SNPs with $r^2 > 0.7$ as the marker dependency file. For the EWAS data, CpG sites are mapped to adjacent genes within 5 kb. Next, we chose canonical pathways from the KEGG database and a positive control gene set from the NHGRI-EBI GWAS catalog (64) for psoriasis as the pathways or marker sets to be examined. We ran Meta-MSEA across the GWAS and two EWAS datasets. At the conclusion of Meta-MSEA, a set of results files and a summary table display are generated on the webpage detailing the pathways ranked by meta *P*-value and their top markers and corresponding mapped genes (Figure 5A; Supplementary Table S5).

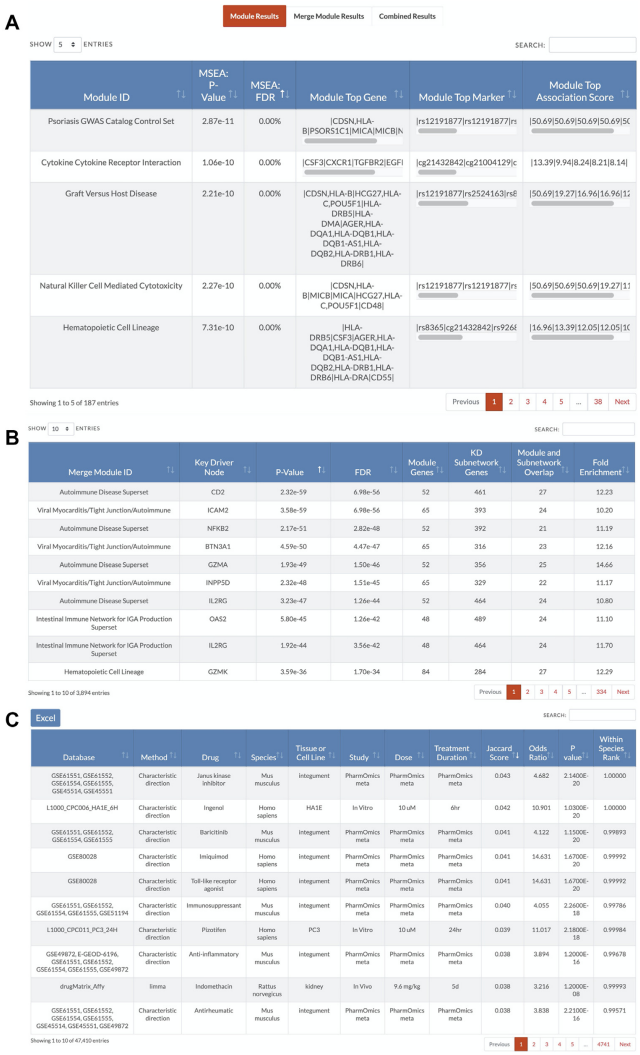


Figure 5. Output files from Meta-MSEA, KDA, and PharmOmics based on the case study of psoriasis outlined in Figure 4. Tables are interactive with pagination, search, and sort functions. Result files are downloadable from links on the webpage above the output tables (not shown). (**A**) Example Meta-MSEA output from the psoriasis use case. The table shown details the significance of association of each pathway/module and the top markers and corresponding association strengths that contributed to the module association. There are two additional tables which can be displayed by clicking on the tabs to the right of 'Module Results' at the top. The second table shows the significance and details of merged modules after merging redundant pathways (termed 'Supersets'), and these non-overlapping gene sets are used as input to KDA. The third table shows the individual significance values for each omics dataset included in this Meta-MSEA of one GWAS and two EWAS of psoriasis. (**B**) Example KDA output from the psoriasis use case. The table shown records the significance of KDs, the pathways/modules that they regulate based on network topology, and details of the local subnetwork such as the number of KD subnetwork genes and number of pathway/module gene overlap with the KD subnetwork. Merged pathways/modules are represented by the term 'Superset', which means they are comprised of multiple redundant (significant gene overlap) pathways. (**C**) Example PharmOmics drug repositioning output using a gene overlap-based analysis between disease pathways and drug signatures. Gene overlap-based drug repositioning queries all tissue- and species-specific meta-analyzed and dose/time segregated gene signatures of drugs in our PharmOmics database as well as all L1000 drug signatures. The table shown gives the dataset source of the drug signature, the method of differential gene expression analysis, details of the drug study including species, tissue or cell line, whether the study was done *in vitro* or *in vivo*, the dose and time regimen, the Jaccard score, and statistical significance of the gene overlap between the input psoriasis related genes from Meta-MSEA and the drug signatures.

As shown in Figure 5A, 'Cytokine cytokine receptor interaction', 'Graft versus host disease' and 'Natural killer cell mediated cytotoxicity' were three of the top pathways identified among others. Following Meta-MSEA, KDA was run with default parameters using non-redundant supersets (pathways that were merged due to significant overlap in gene members) significantly associated with psoriasis from Meta-MSEA and a blood GIANT Bayesian gene regulatory network (61) (chosen due to the relevance of the immune system to psoriasis) to identify KDs of the disease related gene sets. At the conclusion of the KDA, a table is produced on the webpage listing the KDs and significance of enrichment of psoriasis-associated gene sets in their network neighborhood (Figure 5B; Supplementary Table S6). For example, *ICAM2* is identified as the KD for the viral myocarditis/tight junction/autoimmune pathway, and *CD2* is identified as a KD for the Autoimmune Disease Superset. By default, the top five KDs and their local subnetworks from each gene set is included in the interactive subnetwork visualization in the browser (Figure 3).

With addition of the PharmOmics pipeline to the Mergeomics web server, we ran two drug repositioning analyses: one directly from the MSEA results and the other considering the whole subnetwork derived from the KDA (Figure 5C; Supplementary Table S7). In this case study, we do not consider gene expression direction changes (upregulation or downregulation) in psoriasis and therefore will simply be utilizing genes involved in disease without considering if they are protective or pathogenic; thus, our predicted drug list will contain drugs that can induce as well as drugs that can potentially treat psoriasis. In addition, PharmOmics interrogates all drug signatures regardless of the tissue or species, and the user can choose to focus on the relevant drug studies for their given dataset. For example, we mainly focused on drugs that were studied in integument tissue, due to its relevance to psoriasis. In the top 10 repositioned drugs derived from psoriasis associated gene sets from Meta-MSEA, we find 8/10 to have prior association with a role in psoriasis pathogenesis (Imiquimod (65)) or treatment including broad options suggesting classes of drugs such as anti-inflammatory, immunosuppressant, JAK inhibitors, and anti-rheumatic drugs and more specific options such as Baricitinib (66), Ingenol (67), and Etinostat (68) (Figure 5C; Supplementary Table S7). Similarly, using the psoriasis subnetwork from the KDA highlights Imiquimod and Ingenol within the top 10 drugs, and the remainder of the results are broad categories such as JAK inhibitors, anti-inflammatory drugs, and anti-rheumatic drugs (Supplementary Table S8), each of which are actively being investigated in the treatment of psoriasis (69,70). The predicted drugs can form new hypotheses for experimental testing.

FUTURE DIRECTIONS

The web server will continue to actively incorporate the most up-to-date public resources including multi-omics association data, functional genomics data such as eQTLs or protein QTLs (pQTLs), knowledge-based pathways, gene co-expression networks, and gene regulatory networks on a monthly basis. We will also include single cell networks when available to understand the gene regulatory connections within a given cell type or between cell types rather than across a whole tissue, which will offer higher resolution molecular mechanisms of disease pathogenesis. Cell type level association data derived from single cell omics studies can be used in the current platform. We will also continue incorporating additional analytical functions into the web server such as different forms of meta-analysis that can be conducted within the Meta-MSEA tool as well as adding new features to better accommodate analysis of data types that are currently not considered or well tested, such as gut microbiome and spatial transcriptomics data.

CONCLUSION

Thanks to advancements in technologies, the number of multi-omics data (GWAS, EWAS, TWAS, PWAS and MWAS) increases exponentially. The systems biology approach to interrogate multi-tissue multi-omics data has become a promising method to understand biology in a data-driven way and sheds light on the hidden mechanisms. However, the computational knowledge and skills required to perform such integrative analysis are often considered as a hurdle to many biologists. Therefore, the Mergeomics web server was developed to lower this barrier to enable fellow researchers to dive into multi-omics systems biology. The current update, Mergeomics 2.0, is a versatile web-based tool that provides multi-omics data integration using a pathway- and network-based approach. The improvements we made support a wide range of pre-calculated networks and data for all steps of the pipeline to fulfill a variety of needs and research purposes. In addition, the new user interface presents a more intuitive and flexible environment that greatly improves its ease of use. In addition to a detailed tutorial, each step of the pipeline contains embedded guidance to facilitate the user experience. We believe that the Mergeomics 2.0 and systematics approach applied here will accelerate our understanding of complex diseases and guide therapeutics.

DATA AVAILABILITY

Sample resources are available on our sample resources page on the Mergeomics web server (http://mergeomics.research.idre.ucla.edu/samplefiles.php), and the R package for Mergeomics can be found on (https://bioconductor.org/packages/release/bioc/html/Mergeomics.html).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Dr Yuqi Zhao for guiding the use case example. We thank Yanning Zuo, Russell Littman, Neil Hsu, Jenny Cheng, I-Hsin Tseng, Caden McQuillen, Yutian Zhao, Zara Saleem, Hengjian Li and Carissa Zhu for their suggestions and testing of the web server.

FUNDING

X.Y. is supported by NIH R01 [NS117148, NS111378, DK117850, HL145708, HL147883, HD100298]; Montgomery Blencowe is supported by the American Heart Association Predoctoral Fellowship; Sung-min Ha is supported by the UCLA QCBio Collaboratory Postdoc Fellowship. Funding for open access charge: National Institutes of Health [DK117850, HD100298, HL145708, HL147883, NS111378, NS117148].
Conflict of interest statement. None declared.

REFERENCES

1. Yang,X. (2020) Multitissue multiomics systems biology to dissect complex diseases. *Trends Mol. Med.*, **26**, 718–728.
2. Boyle,E.A., Li,Y.I. and Pritchard,J.K. (2017) An expanded view of complex traits: from polygenic to omnigenic. *Cell*, **169**, 1177–1186.
3. Subramanian,I., Verma,S., Kumar,S., Jere,A. and Anamika,K. (2020) Multi-omics data integration, interpretation, and its application. *Bioinform Biol. Insights*, **14**, 1177932219899051.
4. Graw,S., Chappell,K., Washam,C.L., Gies,A., Bird,J., Robeson,M.S. 2nd and Byrum,S.D. (2020) Multi-omics data integration considerations and study design for biological systems and disease. *Mol. Omics*, **17**, 170–185.
5. Huang,S., Chaudhary,K. and Garmire,L.X. (2017) More is better: recent progress in multi-omics data integration methods. Front Genet., **8**, 84.
6. Shi,Q., Zhang,C., Peng,M., Yu,X., Zeng,T., Liu,J. and Chen,L. (2017) Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data. *Bioinformatics*, **33**, 2706–2714.
7. Wang,B., Mezlini,A.M., Demir,F., Fiume,M., Tu,Z., Brudno,M., Haibe-Kains,B. and Goldenberg,A. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.
8. Yuan,Y., Savage,R.S. and Markowetz,F. (2011) Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput. Biol.*, **7**, e1002227.
9. Shen,R., Mo,Q., Schultz,N., Seshan,V.E., Olshen,A.B., Huse,J., Ladanyi,M. and Sander,C. (2012) Integrative subtype discovery in glioblastoma using iCluster. *PLoS One*, **7**, e35236.
10. Lock,E.F. and Dunson,D.B. (2013) Bayesian consensus clustering. *Bioinformatics*, **29**, 2610–2616.
11. de Tayrac,M., Le,S., Aubry,M., Mosser,J. and Husson,F. (2009) Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: multiple factor analysis approach. *BMC Genomics*, **10**, 32.
12. Le Cao,K.A., Gonzalez,I. and Dejean,S. (2009) integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics*, **25**, 2855–2856.
13. Rohart,F., Gautier,B., Singh,A. and Le Cao,K.A. (2017) mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.*, **13**, e1005752.
14. Vaske,C.J., Benz,S.C., Sanborn,J.Z., Earl,D., Szeto,C., Zhu,J., Haussler,D. and Stuart,J.M. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**, i237–i245.
15. Koh,H.W.L., Fermin,D., Vogel,C., Choi,K.P., Ewing,R.M. and Choi,H. (2019) iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. *NPJ Syst Biol Appl*, **5**, 22.
16. Zoppi,J., Guillaume,J.F., Neunlist,M. and Chaffron,S. (2021) MiBiOmics: an interactive web application for multi-omics data exploration and integration. *BMC Bioinformatics*, **22**, 6.
17. Bonnet,E., Calzone,L. and Michoel,T. (2015) Integrative multi-omics module network inference with Lemon-Tree. *PLoS Comput. Biol.*, **11**, e1003983.
18. Hernandez-de-Diego,R., Tarazona,S., Martinez-Mira,C., Balzano-Nogueira,L., Furio-Tari,P., Pappas,G.J. Jr and Conesa,A. (2018) PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Res.*, **46**, W503–W509.
19. Dimitrakopoulos,C., Hindupur,S.K., Hafliger,L., Behr,J., Montazeri,H., Hall,M.N. and Beerenwinkel,N. (2018) Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics*, **34**, 2441–2448.
20. Zhou,Y., Zhou,B., Pache,L., Chang,M., Khodabakhshi,A.H., Tanaseichuk,O., Benner,C. and Chanda,S.K. (2019) Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.*, **10**, 1523.
21. Shu,L., Zhao,Y., Kurt,Z., Byars,S.G., Tukiainen,T., Kettunen,J., Orozco,L.D., Pellegrini,M., Lusis,A.J., Ripatti,S. *et al.* (2016) Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems. *BMC Genomics*, **17**, 874.
22. Arneson,D., Bhattacharya,A., Shu,L., Makinen,V.P. and Yang,X. (2016) Mergeomics: a web server for identifying pathological pathways, networks, and key regulators via multidimensional data integration. *BMC Genomics*, **17**, 722.
23. Chella Krishnan,K., Kurt,Z., Barrere-Cain,R., Sabir,S., Das,A., Floyd,R., Vergnes,L., Zhao,Y., Che,N., Charugundla,S. *et al.* (2018) Integration of multi-omics data from mouse diversity panel highlights mitochondrial dysfunction in non-alcoholic fatty liver disease. *Cell Syst.*, **6**, 103–115.
24. Chen,L., Yao,Y., Jin,C., Wu,S., Liu,Q., Li,J., Ma,Y., Xu,Y. and Zhong,Y. (2019) Integrative genomic analysis identified common regulatory networks underlying the correlation between coronary artery disease and plasma lipid levels. *BMC Cardiovasc. Disord.*, **19**, 310.
25. Hartman,R.J.G., Owsiany,K., Ma,L., Koplev,S., Hao,K., Slenders,L., Civelek,M., Mokry,M., Kovacic,J.C., Pasterkamp,G. *et al.* (2021) Sex-stratified gene regulatory networks reveal female key driver genes of atherosclerosis involved in smooth muscle cell phenotype switching. *Circulation*, **143**, 713–726.
26. Liu,Y., Lu,P., Wang,Y., Morrow,B.E., Zhou,B. and Zheng,D. (2019) Spatiotemporal gene coexpression and regulation in mouse cardiomyocytes of early cardiac morphogenesis. *J. Am. Heart Assoc.*, **8**, e012941.
27. Shu,L., Chan,K.H.K., Zhang,G., Huan,T., Kurt,Z., Zhao,Y., Codoni,V., Tregouet,D.A., Cardiogenics,C., Yang,J. *et al.* (2017) Shared genetic regulatory networks for cardiovascular disease and type 2 diabetes in multiple populations of diverse ethnicities in the United States. *PLoS Genet.*, **13**, e1007040.
28. Zhao,Y., Jhamb,D., Shu,L., Arneson,D., Rajpal,D.K. and Yang,X. (2019) Multi-omics integration reveals molecular networks and regulators of psoriasis. *BMC Syst. Biol.*, **13**, 8.
29. Jung,S.M., Park,K.S. and Kim,K.J. (2020) Deep phenotyping of synovial molecular signatures by integrative systems analysis in rheumatoid arthritis. *Rheumatology (Oxford)*, doi:10.1093/rheumatology/keaa751.
30. Drake,J., McMichael,G.O., Vornholt,E.S., Cresswell,K., Williamson,V., Chatzinakos,C., Mamdani,M., Hariharan,S., Kendler,K.S., Kalsi,G. *et al.* (2020) Assessing the role of long noncoding rna in nucleus accumbens in subjects with alcohol dependence. *Alcohol. Clin. Exp. Res.*, **44**, 2468–2480.
31. Meng,Q., Zhuang,Y., Ying,Z., Agrawal,R., Yang,X. and Gomez-Pinilla,F. (2017) Traumatic brain injury induces genome-wide transcriptomic, methylomic, and network perturbations in brain and blood predicting neurological disorders. *EBioMedicine*, **16**, 184–194.
32. Min,H.K., Moon,S.J., Park,K.S. and Kim,K.J. (2019) Integrated systems analysis of salivary gland transcriptomics reveals key molecular networks in Sjogren's syndrome. *Arthritis Res. Ther.*, **21**, 294.
33. Diamante,G., Cely,I., Zamora,Z., Ding,J., Blencowe,M., Lang,J., Bline,A., Singh,M., Lusis,A.J. and Yang,X. (2021) Systems toxicogenomics of prenatal low-dose BPA exposure on liver metabolic pathways, gut microbiota, and metabolic health in mice. *Environ. Int.*, **146**, 106260.
34. Zhang,G., Byun,H.R., Ying,Z., Blencowe,M., Zhao,Y., Hong,J., Shu,L., Chella Krishnan,K., Gomez-Pinilla,F. and Yang,X. (2020) Differential metabolic and multi-tissue transcriptomic responses to fructose consumption among genetically diverse mice. *Biochim. Biophys. Acta Mol. Basis Dis.*, **1866**, 165569.
35. Shu,L., Meng,Q., Diamante,G., Tsai,B., Chen,Y.W., Mikhail,A., Luk,H., Ritz,B., Allard,P. and Yang,X. (2019) Prenatal bisphenol a

variants and shared genetic basis in two large cohorts. *Nat. Commun.*, **11**, 4423.
76. Nikpay,M., Goel,A., Won,H.-H., Hall,L.M., Willenborg,C., Kanoni,S., Saleheen,D., Kyriakou,T., Nelson,C.P., Hopewell,J.C. *et al.* (2015) A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, **47**, 1121–1130.
77. Manning,A.K., Hivert,M.F., Scott,R.A., Grimsby,J.L., Bouatia-Naji,N., Chen,H., Rybin,D., Liu,C.T., Bielak,L.F., Prokopenko,I. *et al.* (2012) A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.*, **44**, 659–669.
78. Shah,S., Henry,A., Roselli,C., Lin,H., Sveinbjörnsson,G., Fatemifar,G., Hedman,Å.K., Wilk,J.B., Morley,M.P., Chaffin,M.D. *et al.* (2020) Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nat. Commun.*, **11**, 163.
79. Willer,C.J., Schmidt,E.M., Sengupta,S., Peloso,G.M., Gustafsson,S., Kanoni,S., Ganna,A., Chen,J., Buchkovich,M.L., Mora,S. *et al.* (2013) Discovery and refinement of loci associated with lipid levels. *Nat. Genet.*, **45**, 1274–1283.
80. Coleman,J.R.I., Peyrot,W.J., Purves,K.L., Davis,K.A.S., Rayner,C., Choi,S.W., Hübel,C., Gaspar,H.A., Kan,C., Van der Auwera,S. *et al.* (2020) Genome-wide gene-environment analyses of major depressive disorder and reported lifetime traumatic experiences in UK Biobank. *Mol. Psychiatry*, **25**, 1430–1446.
81. Timmers,P.R., Mounier,N., Lall,K., Fischer,K., Ning,Z., Feng,X., Bretherick,A.D., Clark,D.W. and eQTLGen ConsortiumeQTLGen Consortium and Shen,X. *et al.* (2019) Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *Elife*, **8**, e39856.
82. Blauwendraat,C., Heilbron,K., Vallerga,C.L., Bandres-Ciga,S., von Coelln,R., Pihlstrom,L., Simon-Sanchez,J., Schulte,C., Sharma,M., Krohn,L. *et al.* (2019) Parkinson's disease age at onset genome-wide association study: Defining heritability, genetic loci, and alpha-synuclein mechanisms. *Mov. Disord.*, **34**, 866–875.
83. Nair,R.P., Duffin,K.C., Helms,C., Ding,J., Stuart,P.E., Goldgar,D., Gudjonsson,J.E., Li,Y., Tejasvi,T., Feng,B.J. *et al.* (2009) Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nat. Genet.*, **41**, 199–204.
84. Pairo-Castineira,E., Clohisey,S., Klaric,L., Bretherick,A.D., Rawlik,K., Pasko,D., Walker,S., Parkinson,N., Fourman,M.H., Russell,C.D. *et al.* (2020) Genetic mechanisms of critical illness in COVID-19. *Nature*, **591**, 92–98.
85. Schizophrenia Working Group of the Psychiatric Genomics, C. (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421–427.
86. Hahn,J., Fu,Y.P., Brown,M.R., Bis,J.C., de Vries,P.S., Feitosa,M.F., Yanek,L.R., Weiss,S., Giulianini,F., Smith,A.V. *et al.* (2020) Genetic loci associated with prevalent and incident myocardial infarction and coronary heart disease in the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium. *PLoS One*, **15**, e0230035.
87. Wang,Y.F., Zhang,Y., Lin,Z., Zhang,H., Wang,T.Y., Cao,Y., Morris,D.L., Sheng,Y., Yin,X., Zhong,S.L. *et al.* (2021) Identification of 38 novel loci for systemic lupus erythematosus and genetic heterogeneity between ancestral groups. *Nat. Commun.*, **12**, 772.
88. Fuchsberger,C., Flannick,J., Teslovich,T.M., Mahajan,A., Agarwala,V., Gaulton,K.J., Ma,C., Fontanillas,P., Moutsianas,L., McCarthy,D.J. *et al.* (2016) The genetic architecture of type 2 diabetes. *Nature*, **536**, 41–47.
89. Kupers,L.K., Monnereau,C., Sharp,G.C., Yousefi,P., Salas,L.A., Ghantous,A., Page,C.M., Reese,S.E., Wilcox,A.J., Czamara,D. *et al.* (2019) Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight. *Nat. Commun.*, **10**, 1893.
90. Sammallahti,S., Cortes Hidalgo,A.P., Tuominen,S., Malmberg,A., Mulder,R.H., Brunst,K.J., Alemany,S., McBride,N.S., Yousefi,P., Heiss,J.A. *et al.* (2021) Maternal anxiety during pregnancy and newborn epigenome-wide DNA methylation. *Mol. Psychiatry*, doi:10.1038/s41380-020-00976-0.
91. Rijlaarsdam,J., Cecil,C.A.M., Relton,C.L. and Barker,E.D. (2021) Epigenetic profiling of social communication trajectories and co-occurring mental health problems: a prospective, methylome-wide association study. *Dev. Psychopathol.*, doi:10.1017/S0954579420001662.
92. Boyle,A.P., Hong,E.L., Hariharan,M., Cheng,Y., Schaub,M.A., Kasowski,M., Karczewski,K.J., Park,J., Hitz,B.C., Weng,S. *et al.* (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.*, **22**, 1790–1797.
93. Emilsson,V., Thorleifsson,G., Zhang,B., Leonardson,A.S., Zink,F., Zhu,J., Carlson,S., Helgason,A., Walters,G.B., Gunnarsdottir,S. *et al.* (2008) Genetics of gene expression and its effect on disease. *Nature*, **452**, 423–428.
94. Derry,J.M., Zhong,H., Molony,C., MacNeil,D., Guhathakurta,D., Zhang,B., Mudgett,J., Small,K., El Fertak,L., Guimond,A. *et al.* (2010) Identification of genes and networks driving cardiovascular and metabolic phenotypes in a mouse F2 intercross. *PLoS One*, **5**, e14319.
95. Wang,S.S., Schadt,E.E., Wang,H., Wang,X., Ingram-Drake,L., Shi,W., Drake,T.A. and Lusis,A.J. (2007) Identification of pathways for atherosclerosis in mice: integration of quantitative trait locus analysis and global gene expression data. *Circulation Res.*, **101**, e11–e30.
96. Yang,X., Schadt,E.E., Wang,S., Wang,H., Arnold,A.P., Ingram-Drake,L., Drake,T.A. and Lusis,A.J. (2006) Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res.*, **16**, 995–1004.
97. Schadt,E.E., Molony,C., Chudin,E., Hao,K., Yang,X., Lum,P.Y., Kasarskis,A., Zhang,B., Wang,S., Suver,C. *et al.* (2008) Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.*, **6**, e107.
98. Tu,Z., Keller,M.P., Zhang,C., Rabaglia,M.E., Greenawalt,D.M., Yang,X., Wang,I.M., Dai,H., Bruss,M.D., Lum,P.Y. *et al.* (2012) Integrative analysis of a cross-loci regulation network identifies App as a gene regulating insulin secretion from pancreatic islets. *PLoS Genet.*, **8**, e1003107.

exposure in mice induces multitissue multiomics disruptions linking to cardiometabolic disorders. *Endocrinology*, **160**, 409–429.
36. Blencowe,M., Ahn,I.S., Saleem,Z., Luk,H., Cely,I., Makinen,V.P., Zhao,Y. and Yang,X. (2021) Gene networks and pathways for plasma lipid traits via multitissue multiomics systems analysis. *J. Lipid Res.*, **62**, 100019.
37. Zhao,Y., Blencowe,M., Shi,X., Shu,L., Levian,C., Ahn,I.S., Kim,S.K., Huan,T., Levy,D. and Yang,X. (2019) Integrative genomics analysis unravels tissue-specific pathways, networks, and key regulators of blood pressure regulation. *Front Cardiovasc Med*, **6**, 21.
38. Hui,S.T., Kurt,Z., Tuominen,I., Norheim,F., Richard,C.D., Pan,C., Dirks,D.L., Magyar,C.E., French,S.W., Chella Krishnan,K. *et al.* (2018) The genetic architecture of diet-induced hepatic fibrosis in mice. *Hepatology*, **68**, 2182–2196.
39. Meng,Q., Ying,Z., Noble,E., Zhao,Y., Agrawal,R., Mikhail,A., Zhuang,Y., Tyagi,E., Zhang,Q., Lee,J.H. *et al.* (2016) Systems nutrigenomics reveals brain gene networks linking metabolic and brain disorders. *EBioMedicine*, **7**, 157–166.
40. Makinen,V.P., Civelek,M., Meng,Q., Zhang,B., Zhu,J., Levian,C., Huan,T., Segre,A.V., Ghosh,S., Vivar,J. *et al.* (2014) Integrative genomics reveals novel molecular pathways and gene networks for coronary artery disease. *PLos Genet.*, **10**, e1004502.
41. Chen,Y.-W., Diamante,G., Ding,J., Nghiem,T.X., Yang,J., Ha,S.-m., Cohn,P., Arneson,D., Blencowe,M., Garcia,J. *et al.* (2021) PharmOmics: a species- and tissue-specific drug signature database and online tool for drug repurposing. bioRxiv doi: https://doi.org/10.1101/837773, 30 March 2021, preprint: not peer reviewed.
42. Lamb,J. (2007) The Connectivity Map: a new tool for biomedical research. *Nat. Rev. Cancer*, **7**, 54–60.
43. Subramanian,A., Narayan,R., Corsello,S.M., Peck,D.D., Natoli,T.E., Lu,X., Gould,J., Davis,J.F., Tubelli,A.A., Asiedu,J.K. *et al.* (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, **171**, 1437–1452.
44. Cheng,F., Desai,R.J., Handy,D.E., Wang,R., Schneeweiss,S., Barabasi,A.L. and Loscalzo,J. (2018) Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat. Commun.*, **9**, 2691.
45. Consortium,G.T. (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, **369**, 1318–1330.
46. Auton,A., Brooks,L.D., Durbin,R.M., Garrison,E.P., Kang,H.M., Korbel,J.O., Marchini,J.L., McCarthy,S., McVean,G.A. and Abecasis,G.R. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
47. Xu,J., Zhao,L., Liu,D., Hu,S., Song,X., Li,J., Lv,H., Duan,L., Zhang,M., Jiang,Q. *et al.* (2018) EWAS: epigenome-wide association study software 2.0. *Bioinformatics*, **34**, 2657–2658.
48. Consortium,E.P. (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*, **306**, 636–640.
49. Liberzon,A., Birger,C., Thorvaldsdottir,H., Ghandi,M., Mesirov,J.P. and Tamayo,P. (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.*, **1**, 417–425.
50. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
51. Fabregat,A., Jupe,S., Matthews,L., Sidiropoulos,K., Gillespie,M., Garapati,P., Haw,R., Jassal,B., Korninger,F., May,B. *et al.* (2018) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.
52. Nishimura,D. (2001) BioCarta. *Biotech. Softw. Internet Rep.*, **2**, 117–120.
53. The Gene Ontology,C. (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
54. Slenter,D.N., Kutmon,M., Hanspers,K., Riutta,A., Windsor,J., Nunes,N., Melius,J., Cirillo,E., Coort,S.L., Digles,D. *et al.* (2018) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.*, **46**, D661–D667.
55. Huang,R., Grishagin,I., Wang,Y., Zhao,T., Greene,J., Obenauer,J.C., Ngan,D., Nguyen,D.T., Guha,R., Jadhav,A. *et al.* (2019) The NCATS BioPlanet - an integrated platform for exploring the universe of cellular signaling pathways for toxicology, systems biology, and chemical genomics. Front. Pharmacol., **10**, 445.
56. Song,W.M. and Zhang,B. (2015) Multiscale embedded gene co-expression network analysis. *PLoS Comput. Biol.*, **11**, e1004574.
57. Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.
58. Zhu,J., Lum,P.Y., Lamb,J., GuhaThakurta,D., Edwards,S.W., Thieringer,R., Berger,J.P., Wu,M.S., Thompson,J., Sachs,A.B. *et al.* (2004) An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet. Genome Res.*, **105**, 363–374.
59. Szklarczyk,D., Gable,A.L., Nastou,K.C., Lyon,D., Kirsch,R., Pyysalo,S., Doncheva,N.T., Legeay,M., Fang,T., Bork,P. *et al.* (2021) The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, **49**, D605–D612.
60. Marbach,D., Lamparter,D., Quon,G., Kellis,M., Kutalik,Z. and Bergmann,S. (2016) Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods*, **13**, 366–370.
61. Greene,C.S., Krishnan,A., Wong,A.K., Ricciotti,E., Zelaya,R.A., Himmelstein,D.S., Zhang,R., Hartmann,B.M., Zaslavsky,E., Sealfon,S.C. *et al.* (2015) Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.*, **47**, 569–576.
62. Roberson,E.D., Liu,Y., Ryan,C., Joyce,C.E., Duan,S., Cao,L., Martin,A., Liao,W., Menter,A. and Bowcock,A.M. (2012) A subset of methylated CpG sites differentiate psoriatic from normal skin. *J. Invest. Dermatol.*, **132**, 583–592.
63. Gu,X., Nylander,E., Coates,P.J., Fahraeus,R. and Nylander,K. (2015) Correlation between reversal of DNA methylation and clinical symptoms in psoriatic epidermis following narrow-band UVB phototherapy. *J. Invest. Dermatol.*, **135**, 2077–2083.
64. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
65. van der Fits,L., Mourits,S., Voerman,J.S., Kant,M., Boon,L., Laman,J.D., Cornelissen,F., Mus,A.M., Florencia,E., Prens,E.P. *et al.* (2009) Imiquimod-induced psoriasis-like skin inflammation in mice is mediated via the IL-23/IL-17 axis. *J. Immunol.*, **182**, 5836–5845.
66. Menter,A., Disch,D., Clemens,J., Janes,J., Papp,K. and Macias,W. (2014) *J. Am. Acad. Dermatol.*, **70**, AB162.
67. Martin,G., Strober,B.E., Leonardi,C.L., Gelfand,J.M., Blauvelt,A., Kavanaugh,A., Stein Gold,L., Berman,B., Rosen,T. and Stockfleth,E. (2016) Updates on psoriasis and cutaneous oncology: Proceedings from the 2016 MauiDerm meeting based on presentations by. *J. Clin. Aesthet. Dermatol.*, **9**, S5–S29.
68. McLaughlin,F. and La Thangue,N.B. (2004) Histone deacetylase inhibitors in psoriasis therapy. *Curr. Drug Targets Inflamm. Allergy*, **3**, 213–219.
69. Kwatra,S.G., Dabade,T.S., Gustafson,C.J. and Feldman,S.R. (2012) JAK inhibitors in psoriasis: a promising new treatment modality. *J. Drugs Dermatol.*, **11**, 913–918.
70. Rendon,A. and Schäkel,K. (2019) Psoriasis pathogenesis and treatment. *Int. J. Mol. Sci.*, **20**, 1475.
71. Marioni,R.E., Harris,S.E., Zhang,Q., McRae,A.F., Hagenaars,S.P., Hill,W.D., Davies,G., Ritchie,C.W., Gale,C.R., Starr,J.M. *et al.* (2018) GWAS on family history of Alzheimer's disease. *Translational Psychiatry*, **8**, 99.
72. Middeldorp,C.M., Hammerschlag,A.R., Ouwens,K.G., Groen-Blokhuis,M.M., Pourcain,B.S., Greven,C.U., Pappa,I., Tiesler,C.M.T., Ang,W., Nolte,I.M. *et al.* (2016) A genome-wide association meta-analysis of attention-deficit/hyperactivity disorder symptoms in population-based pediatric cohorts. *J. Am. Acad. Child Adolesc. Psychiatry*, **55**, 896–905.
73. Olfson,E. and Bierut,L.J. (2012) Convergence of genome-wide association and candidate gene studies for alcoholism. *Alcohol Clin. Exp. Res.*, **36**, 2086–2094.
74. Locke,A.E., Kahali,B., Berndt,S.I., Justice,A.E., Pers,T.H., Day,F.R., Powell,C., Vedantam,S., Buchkovich,M.L., Yang,J. *et al.* (2015) Genetic studies of body mass index yield new insights for obesity biology. *Nature*, **518**, 197–206.
75. Rashkin,S.R., Graff,R.E., Kachuri,L., Thai,K.K., Alexeeff,S.E., Blatchins,M.A., Cavazos,T.B., Corley,D.A., Emami,N.C., Hoffman,J.D. *et al.* (2020) Pan-cancer study detects genetic risk

nature COMMUNICATIONS

Appendix B

ARTICLE

https://doi.org/10.1038/s41467-021-24232-3 OPEN

Conservation and divergence of vulnerability and responses to stressors between human and mouse astrocytes

Jiwen Li[1], Lin Pan[1], William G. Pembroke[2], Jessica E. Rexach[2], Marlesa I. Godoy[1], Michael C. Condro[1], Alvaro G. Alvarado[1], Mineli Harteni[1], Yen-Wei Chen[3], Linsey Stiles[4], Angela Y. Chen[5], Ina B. Wanner[1,6], Xia Yang[3,7,8,9], Steven A. Goldman[10,11], Daniel H. Geschwind[1,2,12], Harley I. Kornblum[1,6,9,13] & Ye Zhang[1,6,8,9,14]✉

Astrocytes play important roles in neurological disorders such as stroke, injury, and neurodegeneration. Most knowledge on astrocyte biology is based on studies of mouse models and the similarities and differences between human and mouse astrocytes are insufficiently characterized, presenting a barrier in translational research. Based on analyses of acutely purified astrocytes, serum-free cultures of primary astrocytes, and xenografted chimeric mice, we find extensive conservation in astrocytic gene expression between human and mouse samples. However, the genes involved in defense response and metabolism show species-specific differences. Human astrocytes exhibit greater susceptibility to oxidative stress than mouse astrocytes, due to differences in mitochondrial physiology and detoxification pathways. In addition, we find that mouse but not human astrocytes activate a molecular program for neural repair under hypoxia, whereas human but not mouse astrocytes activate the antigen presentation pathway under inflammatory conditions. Here, we show species-dependent properties of astrocytes, which can be informative for improving translation from mouse models to humans.

[1] Department of Psychiatry and Biobehavioral Sciences, Semel Institute for Neuroscience and Human Behavior, David Geffen School of Medicine at the University of California, Los Angeles, CA, USA. [2] Department of Neurology, David Geffen School of Medicine at the University of California, Los Angeles, CA, USA. [3] Department of Integrative Biology and Physiology, University of California, Los Angeles, CA, USA. [4] Department of Endocrinology, David Geffen School of Medicine at the University of California, Los Angeles, CA, USA. [5] Department of Obstetrics and Gynecology, University of California, Los Angeles, CA, USA. [6] Intellectual and Developmental Disabilities Research Center at UCLA, Los Angeles, CA, USA. [7] Institute for Quantitative and Computational Biosciences at UCLA, Los Angeles, CA, USA. [8] Brain Research Institute at UCLA, Los Angeles, CA, USA. [9] Molecular Biology Institute at UCLA, Los Angeles, CA, USA. [10] Center for Translational Neuromedicine and Department of Neurology, University of Rochester Medical Center, Rochester, NY, USA. [11] Center for Translational Neuromedicine, University of Copenhagen Faculty of Health and Medical Sciences, Copenhagen, Denmark. [12] Department of Human Genetics, David Geffen School of Medicine at the University of California, Los Angeles, CA, USA. [13] Department of Pediatrics, David Geffen School of Medicine at the University of California, Los Angeles, CA, USA. [14] Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at UCLA, Los Angeles, CA, USA. ✉email: yezhang@ucla.edu

Mice are one of the most widely used experimental animals in biomedical research due to the ease with which they can be genetically manipulated and the many paradigms that translate well between species. However, humans and mice differ greatly in body size, life span, ecological niche, behavior, and pathogenic challenges. Many mouse models of neurodegenerative disorders exhibit milder neuron degeneration phenotypes compared with human patients[1–4]. Mouse models of ischemic stroke can often achieve full functional recovery[5], whereas human patients frequently have irreversible functional deficits. These limitations represent a key barrier in translational research, as over 90% of neurological drug candidates with promising animal data failing in human clinical trials[6]. Therefore, a full understanding of the cellular and molecular differences between the human and mouse brain is urgently needed.

Astrocytes are critical for many aspects of development and function in the central nervous system (CNS)[7–23]. Most of our knowledge on the biology of astrocytes is based on studies using mouse astrocytes. Aside from human astrocytes being larger and morphologically more complex than mouse astrocytes[24,25], little is known about the similarities and differences between human and mouse astrocytes, particularly their responses to disease-relevant perturbations. Because of this knowledge gap, it is challenging to harness the knowledge gained from mouse astrocytes to elucidate the biology of human astrocytes and their roles in neurological disorders.

In this study, we systematically examined human astrocytes under three conditions: acutely purified, cultured without serum, and xenografted into mouse brains. We found extensive conservation between human and mouse astrocyte transcriptomes, but also identified important differences between mouse and human astrocytes that were maintained across all three conditions. We identified striking differences in the cell survival, mitochondrial physiology, and molecular responses of human and mouse astrocytes under oxidative stress, hypoxia, inflammatory cytokine treatment, and simulated viral infections. These findings reveal important mechanistic differences between human and mouse astrocytes and provide insight into how mouse models of neurodegeneration and stroke can be improved to achieve better translation to humans.

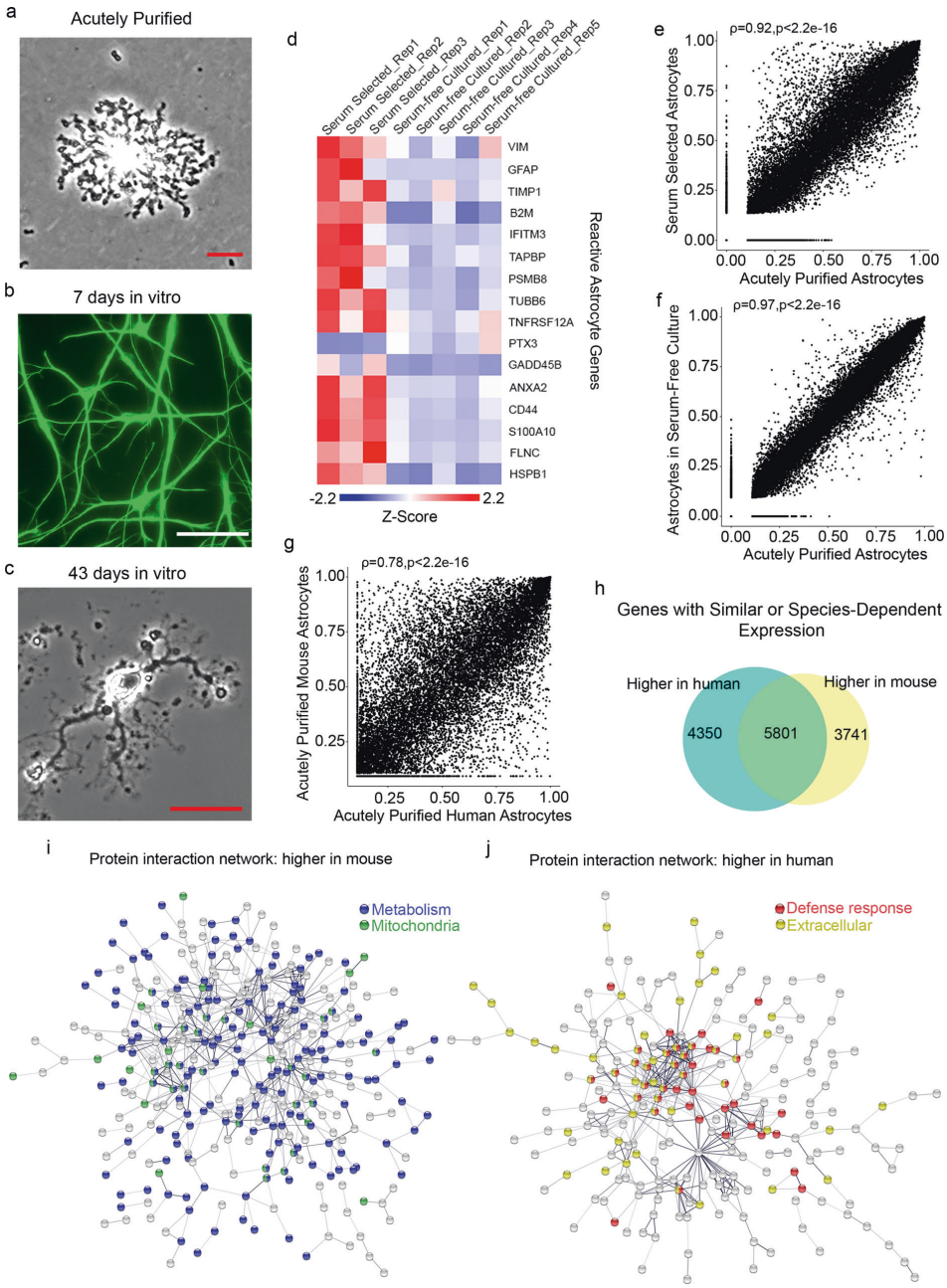
Results

Immunopanned astrocytes exhibit resting transcriptome profiles. We recently developed an immunopanning method for the acute purification of human astrocytes and a serum-free chemically defined medium that keeps human astrocytes healthy for at least six weeks in vitro (Fig. 1a–c)[26]. Here, we tested whether immunopanned human astrocytes resemble resting or reactive astrocytes by RNA sequencing (RNA-seq). We assessed the expression of genes previously found to be induced by stroke and inflammation in mouse astrocytes[27]. The expression of these genes was significantly lower in cultured immunopanned astrocytes than in serum-selected astrocytes (average fold change = 0.18; false discovery rate (FDR) = 0.032; Fig. 1d). In addition, we identified reactive astrocyte genes induced in inflammatory conditions in humans (see below) and again found lower expression of these genes in cultured immunopanned astrocytes than in serum-selected astrocytes (Supplementary Fig. 1). Both immunopanned and serum-selected human astrocyte cultures exhibited high expression of astrocyte-specific genes and low or undetectable levels of genes specific to neurons, microglia, oligodendrocyte precursor cells, or endothelial cells (Supplementary Fig. 2).

To examine the extent to which immunopanned human astrocytes could model in vivo astrocytes, we performed immunopanning purification of human astrocytes and harvested RNA (1) immediately after purification to capture the in vivo gene signature (referred to as acutely purified thereafter) and (2) after 4–6 days of culturing in our serum-free chemically defined medium (referred to as serum-free cultured thereafter). We then performed RNA-seq and compared the transcriptomes of acutely purified and cultured human astrocytes. We found that the gene expression from the serum-free astrocyte cultures more closely resembled acutely purified astrocytes than astrocytes obtained using the traditional serum-selected method (Spearman's correlation = 0.97 vs. 0.92; these correlation coefficients are significantly different; $p < 0.0001$; Fig. 1e, f, Supplementary Fig. 3 and Supplementary Data 1 and 2). We performed principal component analysis (PCA) and found that acutely purified astrocytes and serum-free cultures of astrocytes are more similar to each other than to serum-selected astrocytes (Supplementary Fig. 4). Overall, immunopanned human astrocytes recapitulate the expression of the majority of genes expressed by astrocytes in vivo and therefore represent a useful platform for studying human astrocyte biology.

Species-dependent astrocytic gene expression. We compared the acutely purified human astrocytes described above with corresponding mouse transcriptome data that we previously collected[26,28]. The overall gene expression profiles showed conservation between human and mouse astrocytes (Spearman's correlation $\rho = 0.78$; Fig. 1g). However, thousands of genes exhibited significant differences in expression between species (8091 genes, FDR < 0.05; genes with percentile rankings in the top two-thirds of both species were included; Fig. 1h; Supplementary Fig. 5). To pinpoint the genes and pathways that differed between human and mouse astrocytes, we analyzed protein-interaction networks and gene ontology (GO) terms. We found that the genes expressed at higher levels by mouse compared to human astrocytes were enriched in multiple GO terms associated with metabolism (Fig. 1i and Supplementary Data 3). In contrast, genes expressed at higher levels by human compared to mouse astrocytes were enriched in a single GO term, defense response (Fig. 1j and Supplementary Data 3). We analyzed the subcellular localization of proteins encoded by genes differentially expressed between human and mouse astrocytes. Interestingly, mouse astrocytes showed higher expression of genes associated with the compartment mitochondria, whereas human astrocytes showed higher expression of genes assigned to extracellular space (Fig. 1i, j), including secreted cytokines. The top hub genes with the most protein-protein interactions with other genes in the network include *IL6*, a cytokine involved in inflammation, and *TLR4*, a Toll-like family receptor that mediates responses to bacterial lipopolysaccharide, in the genes with higher expression in humans, and *Ndufa7* and *Ndufb7*, mitochondrial respiratory chain components, in the genes with higher expression in mouse (Supplementary Data 4). To assess whether gene expression differences between human and mouse astrocytes can be found in other independent datasets, we analyzed single-cell and single-nucleus RNA-seq datasets from human and mouse brains[29]. We found that the human-mouse expression differences determined from our immunopanned astrocyte bulk RNA-seq data correlate with the human-mouse expression differences derived from single-cell RNA-seq data ($r^2 = 0.35$; Supplementary Fig. 6), confirming that our approach of using acutely purified astrocytes from human and mouse can recapitulate the species differences in astrocytes observed in vivo. The human-mouse gene expression differences we identified are consistent across 15 mouse strains (Supplementary Fig. 7).

The human-specific gene signature is intrinsically programmed. The higher expression of defense response genes by human astrocytes could be a result of either intrinsic properties or



differences in external factors (e.g., neuronal or glial cell types, systemic or environmental variations) between human and mouse samples. To assess differences between mouse and human astrocytes when exposed to equivalent external environments, we transplanted human astrocytes into mouse brains and compared them with the neighboring host mouse astrocytes[30–32]. We purified primary human fetal astrocytes, and then injected them into the brains of neonatal mice (Fig. 2a). We aged the xenografted chimera mice for about 8 months, and then confirmed widespread distribution of human astrocytes in the host mouse brains (Fig. 2b–e). We then purified all astrocytes (human and mouse) from the chimeric mice by immunopanning and

Fig. 1 Comparison of astrocyte transcriptomes in vivo and in vitro and between human and mouse. a An astrocyte bound to an anti-HepaCAM antibody-coated petri dish during immunopanning purification. RNA was extracted immediately after the cells stuck to the dish. These samples are referred to as acutely purified. Scale bar: 10 μm. **b** Astrocytes in serum-free culture stained with anti-GFAP antibodies. Scale bar: 50 μm. **c** A bright-field image of an astrocyte in serum-free culture. Scale bar: 20 μm. **d** Expression of reactive astrocyte marker genes in serum-selected and serum-free cultures of human astrocytes. Z-score is calculated as (RPKM—average RPKM across all samples)/standard deviation. Genes with FDR < 0.1 between serum-selected and serum-free cultures and RPKM > 1.5 are shown. **e**, **f** Scatter plots and Spearman's correlation coefficients (ρ) of gene expression between cultured and acutely purified human astrocytes using the serum-selected culture method and our serum-free culture method. For each condition, gene expression across 3–5 patient samples was averaged. Only protein-coding genes were included. Two-tailed t-test. $p < 2.2\times10^{-16}$. **g** Scatter plot and Spearman's correlation of gene expression between acutely purified human and mouse astrocytes. Two-tailed t-test. $p < 2.2\times10^{-16}$. Human brain tissue was derived from donors of different ages (8–63 years). Mouse brain tissue was derived from postnatal and adult mice. Additional information is provided in Supplementary Data 8. **h** Number of genes with similar or species-dependent expression. Genes with percentile rankings in the top two-thirds were included in this analysis to eliminate those not expressed or expressed at very low levels. Percentile rankings of the expression of each gene were compared across human and mouse astrocyte samples and differences were tested by Welch's T-test followed by post-hoc multiple comparison adjustment using the Benjamini and Hochberg FDR method[124–126]. FDR < 0.05. **i**, **j** Protein interaction networks of genes expressed at higher levels by mouse astrocytes than human astrocytes (**i**) and at higher levels by human astrocytes than mouse astrocytes (**j**) (percentile ranking difference > 0.4). FDR < 0.05. Blue: genes associated with the GO term metabolism. Green: genes associated with the cellular component mitochondria. Red: genes associated with the GO term defense response. Yellow: genes associated with the cellular component extracellular.

performed RNA-seq. We exploited DNA sequence differences between human and mouse genes in order to separate sequencing reads of human vs. mouse origin at the mapping step. This approach allowed us to obtain the transcriptome profile of human astrocytes grafted in a host mouse brain (Supplementary Data 5).

To test whether the human-specific astrocyte gene signature is intrinsically programmed or induced by other cell types in the human brain environment, we compared gene expression differences between human and mouse astrocytes using both the acutely purified and the xenograft/host dataset. We reasoned that human-mouse astrocyte differences would be attenuated in the xenograft model if the astrocyte differences were driven by human-specific environmental factors. We calculated gene expression differences between human and mouse astrocytes based on the acutely purified dataset (hmDiff_acute) and the chimera dataset (hmDiff_chimera) (Fig. 2f; Supplementary Fig. 8). The heatmaps (Fig. 2f; Supplementary Fig. 8) show differentially expressed genes based on these two datasets. If the species-specific gene expression patterns were determined by the host environment, then differentially expressed genes across species based on the acutely purified dataset would not be differentially expressed in the xenografted dataset (i.e., the xenografted column of the heatmap would appear white). Instead, we observed similar patterns of species-specific gene expression across the acutely purified and xenografted datasets. We found a positive correlation between hmDiff_acute and hmDiff_chimera (Pearson's correlation = 0.60 for all genes; Pearson's correlation = 0.85 for genes with percentile differences > 0.4; Fig. 2f and Supplementary Fig. 8). We also calculated gene expression correlations between transplanted human astrocytes and acutely purified human/mouse astrocytes and found that transplanted human astrocytes resemble human astrocytes more than mouse astrocytes (i.e., the correlation coefficient of transplanted human vs. acutely purified human was significantly higher than that of transplanted human vs. acutely purified mouse; $p < 0.0001$; Supplementary Fig. 9a, b). These analyses suggest that the human-specific astrocyte gene signature is largely intrinsically programmed, with only minor environmental contributions by neurons and other cell types in the human brain.

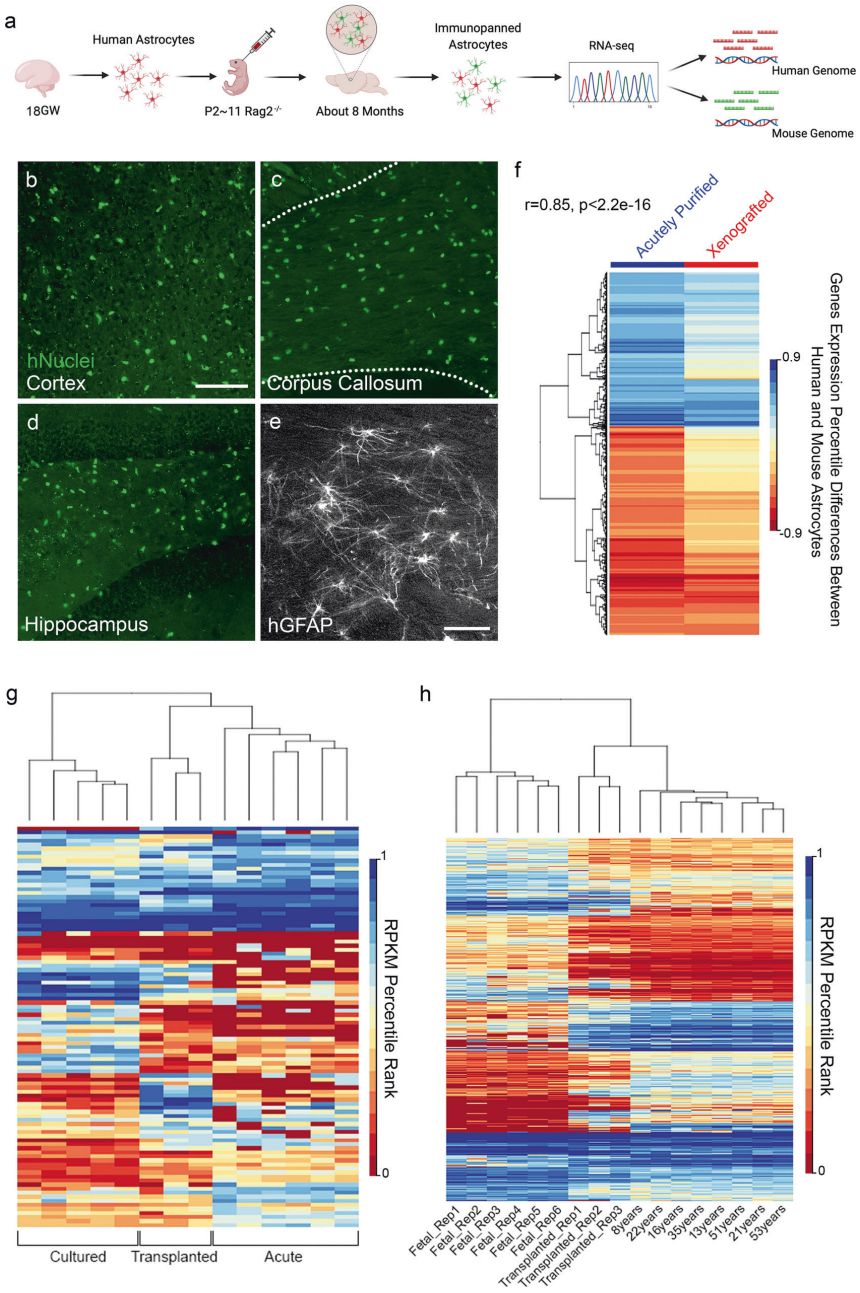
One challenge that has limited human astrocyte research is the difficulty in obtaining mature cells for experimental manipulations, due to (1) the limited availability of fresh healthy adult human brain tissue and (2) the restriction that stem cell-derived human astrocytes mostly resemble developing stages[33–35]. Here, we found that certain genes that are expressed in acutely purified astrocytes in vivo, but lost in culture, are regained in xenografted astrocytes (Fig. 2g). In fact, the xenografted human astrocytes were able to reach mature stages that are difficult to access in in vitro models (Fig. 2h; Supplementary Fig. 9c–e), allowing us to observe the persistence of mouse and human transcriptomic differences across a broad developmental range. Therefore, in addition to identifying consistent species differences in astrocyte transcriptomic profiles across acutely purified, cultured, and xenografted conditions, we established the xenograft model as a much-needed platform for studying mature human astrocytes in vivo.

To assess how the introduction of human astrocytes may change host mouse astrocytes, we compared the transcriptome of host mouse astrocytes (this study) with naïve mouse astrocytes from a similar age[26] and found differentially expressed genes (Supplementary Data 6 and 7).

Species-dependent susceptibility to oxidative stress. To examine responses of human and mouse astrocytes to environmental perturbations, we treated human and mouse astrocytes with several disease-relevant stimuli—including oxidative stress, hypoxia, simulated viral infection, and an inflammatory cytokine—and evaluated the responses.

Oxidative stress is produced by reactive oxygen species (ROS) such as peroxides, superoxide, hydroxyl radical, singlet oxygen, and alpha-oxygen. ROS are byproducts of normal metabolism in most cell types in the body. Importantly, during pathogen invasion, tissue damage, and inflammation, immune cells such as neutrophils and macrophages produce high levels of ROS that help fend off pathogen infections but may also damage healthy cells in infected tissue. In the brain, oxidative stress is a key pathological process underlying neurodegenerative disorders (such as Alzheimer's disease, Parkinson's disease, Huntington's disease, and amyotrophic lateral sclerosis), stroke, and traumatic injury.

To examine responses of human and mouse astrocytes to oxidative stress, we purified human and mouse astrocytes from developmentally equivalent stages [gestational week 17–20 for human brains and postnatal day 1–3 (P1–3) for mouse brains; see Methods for details on developmental stage matching]. Astrocytes have been shown to be regionally heterogeneous[36]. Therefore, whenever possible, we matched anatomical locations in human and mouse brains. We used whole cerebral cortex for astrocyte purification for all mouse samples and a subset of human samples with a clearly identifiable cerebral cortex. In cases where identification of cerebral cortex was difficult due to tissue



fragmentation, we selected, to the best of our ability, fragments most likely to be cortex (large thin sheets).

We performed immunopanning purification to obtain human and mouse astrocytes, plated them at similar densities, and cultured them using identical growth media. To examine responses of human and mouse astrocytes to oxidative stress, we treated cells cultured 3 days in vitro (div) with 100 μM H_2O_2. We then examined cell survival by staining with the live-cell dye calcein-AM and the dead-cell dye ethidium homodimer, 18 h after treatment onset (Fig. 3a). We found that human astrocytes were much more susceptible than mouse astrocytes to oxidative stress (survival rates: 0.29 ± 0.01 for human astrocytes and

Fig. 2 The human-specific astrocyte gene signature is intrinsically programmed. a Experimental design. Gestational week 18 primary human astrocytes were purified and injected into the brains of neonatal immunodeficient Rag2-knockout mice. After about 8 months, we purified astrocytes from xenografted mouse brains by immunopanning. The astrocytes from both human grafts and mouse hosts were sequenced together and reads were mapped to human and mouse genomes, respectively. GW, gestational week. P, postnatal day. **b**–**d** Xenografted human cells in host mouse brains stained with an antibody against human nuclei (green). Scale bar: 100 μm. Dashed lines delineate the corpus callosum. **e** Xenografted human astrocytes in host mouse brains stained with an anti-GFAP antibody that only reacts with human GFAP but not mouse GFAP. Scale bar: 50 μm. **f** Species differences in gene expression (shown as percentile ranking in human minus percentile ranking in mouse) in xenografted and acutely purified astrocytes highly correlate. Genes with percentile rankings > 0.33, species differences with FDR < 0.05, and species differences in percentile ranking > 0.4 are shown. r, Pearson's correlation coefficient. **g** Non-supervised hierarchical clustering of gene expression in serum-free cultures, acutely purified astrocytes, and transplanted human astrocytes. Genes with significant differences between cultured and acutely purified astrocytes (FDR < 0.05, fold change > 4, average RPKM > 1) are shown. **h** Non-supervised hierarchical clustering of gene expression of acutely purified astrocytes from patients of different ages and transplanted human astrocytes. Genes with significant differences between age groups (fetal, child, adult; FDR < 0.05, fold change > 2, maximum RPKM > 1) are shown.

0.54 ± 0.01 for mouse astrocytes; Fig. 3b–d; $p < 0.001$; data represent average ± SEM unless otherwise noted). Since plating density, H_2O_2 concentration, and treatment duration may affect cell survival, we tested different combinations of these conditions and again found that human astrocytes were much more susceptible than mouse astrocytes to oxidative stress (Supplementary Fig. 10).

Species differences in mitochondrial respiration. Mitochondria are both sources and targets of ROS: the mitochondrial respiration chain produces ROS, and ROS (either endogenous or exogenous) can damage mitochondrial function. Furthermore, mitochondria play important roles in cell death. Therefore, to identify the cellular mechanisms underlying the striking difference in the susceptibilities of human and mouse astrocytes to oxidative stress, we examined mitochondrial metabolism in these cells. We purified human and mouse astrocytes, cultured them under the same conditions and assessed mitochondrial metabolism.

Although human astrocytes are larger than mouse astrocytes in vivo[25] and in vitro[26], the basal respiration rate per mouse astrocyte (oxygen consumption rate, OCR) is almost twice as high compared with human astrocytes (mouse: 1.41±0.11 pmol/min per 1000 cells; human: 0.71±0.09 pmol/min per 1000 cells; $p < 0.05$; Fig. 4a). Furthermore, respiration for ATP production is also higher in mouse than human astrocytes (mouse: 1.02±0.06 pmol/min per 1000 cells; human 0.58±0.08 pmol/min per 1000 cells; $p < 0.05$. Fig. 4b). Human and mouse astrocytes were plated at similar densities and cultured with identical media for all metabolic experiments (Supplementary Fig. 11).

These differences in the mitochondrial respiration rates in human and mouse astrocytes raised the question of whether energy substrates are utilized differently in these cells. Glucose is the predominant energy substrate in healthy brains. Through glycolysis, glucose becomes pyruvate, leading to two alternative metabolic pathways (Fig. 4j): (1) Pyruvate can be converted to acetyl-CoA, enter the tricarboxylic acid cycle, and eventually be converted to substrates of oxidative phosphorylation and produce ATP. This process occurs intracellularly within astrocytes. (2) Pyruvate can be converted to lactate and exported to the extracellular space. Neurons can take in lactate and use it as an energy substrate, although the astrocyte-neuron-lactate-shuttle hypothesis remains controversial. To examine the usage of glucose by the two alternative pathways in human and mouse astrocytes, we used Seahorse Respirometry's pH-sensitive electrodes to measure the extracellular acidification rate (ECAR), an approximate measure of lactate production and the glycolysis rate, and compared the ECAR to OCR, an approximate measure of the oxidative phosphorylation rate. We found that the OCR/ECAR ratio was higher in mouse astrocytes than in human astrocytes (Fig. 4c). Thus, mouse astrocytes may utilize a larger proportion of glucose for oxidative phosphorylation, which provides energy for astrocytes themselves, whereas human astrocytes utilize a larger proportion of glucose for lactate production, which may serve as an energy substrate for neurons.

Having identified metabolic differences between human and mouse astrocytes in unperturbed conditions, we next examined changes in mitochondrial metabolism and physiology under oxidative stress in human and mouse astrocytes. We treated the astrocytes with 100 μM H_2O_2 and measured OCR 1, 3, and 5 h after treatment onset. While mouse astrocytes exhibited a small increase, human astrocytes exhibited a substantial reduction of OCR under oxidative stress (Fig. 4f, g; mouse: 0 hr 1.41 ± 0.11, 5 hr 1.66 ± 0.12, $p < 0.05$; human: 0 hr 0.71 ± 0.09, 3 hr 0.53 ± 0.08, $p < 0.05$). Similarly, we found that ATP-linked OCR is stable in mouse astrocytes but reduced in human astrocytes under oxidative stress (Fig. 4h, i). Therefore, mitochondria from mouse astrocytes are highly resilient to oxidative damage; these organelles may work harder as an adaptive response to oxidative damage and, as a result, produce more ATP for cellular protective pathways (see section on the detoxification pathway below). In contrast, mitochondria from human astrocytes are quickly damaged and cannot keep up with the cellular energy demand when exposed to oxidative stress.

To further examine the physiological status of mitochondria under oxidative stress, we performed fluorescence imaging using tetramethylrhodamine ethyl ester (TMRE), a dye sensitive to the membrane potential across the mitochondrial inner membrane[37]. We found that the mitochondrial membrane potential remained largely stable in mouse astrocytes but depolarized quickly in human astrocytes (Fig. 4d, e; mouse 0.81 ± 0.02 at 1 hr, 0.74 ± 0.03 at 3 hr, not significant; human 0.77 ± 0.04 at 1 hr, 0.48 ± 0.03 at 3 hr, $p < 0.05$). Therefore, mitochondria in human astrocytes are more susceptible to oxidative damage than those in mouse astrocytes.

Species-dependent expression of detoxification pathway genes. The species differences in oxidative stress susceptibility may be because mouse astrocytes have evolved adaptive mechanisms, such as more efficient detoxification pathways, under high ROS conditions that render protection against oxidative stress. To test this hypothesis, we examined the function of the peroxisome, an organelle involved in ROS detoxification[38]. We blocked mitochondrial oxidation with Antimycin A, which binds and inactivates Complex III[39], to focus on non-mitochondrial oxygen consumption, which has a large contribution from peroxisomal oxidation[40]. We found that the non-mitochondrial oxygen consumption rate was higher in mouse astrocytes than in human astrocytes (Fig. 5c), consistent with the possibility that peroxisome oxidation operates faster in mouse astrocytes than human astrocytes. To evaluate molecular differences in ROS detoxification pathways, we purified human

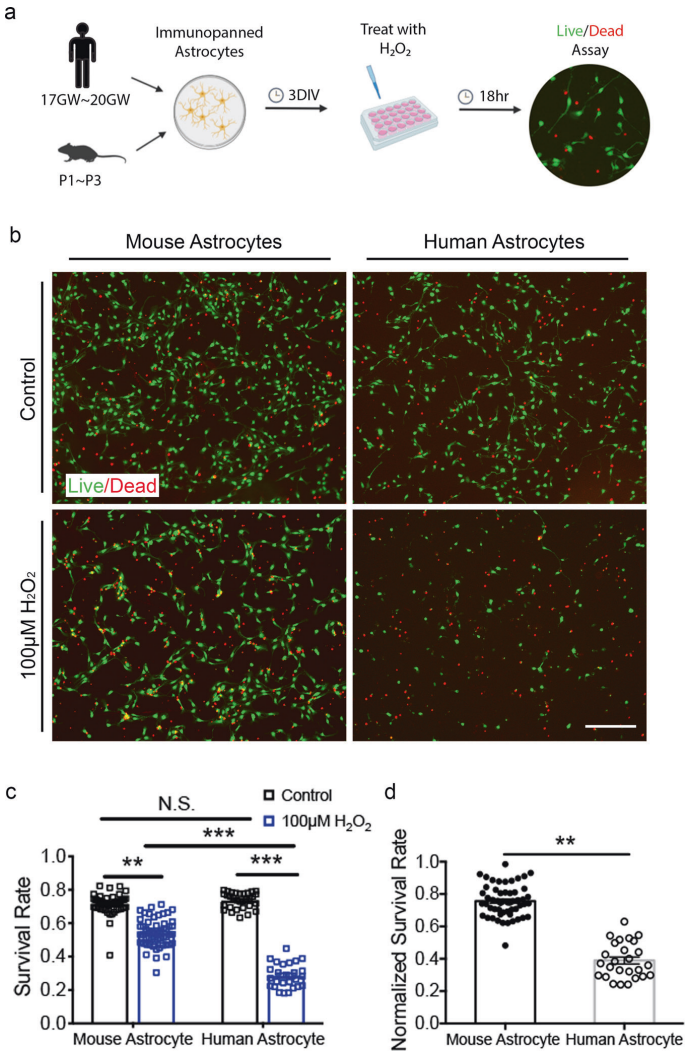
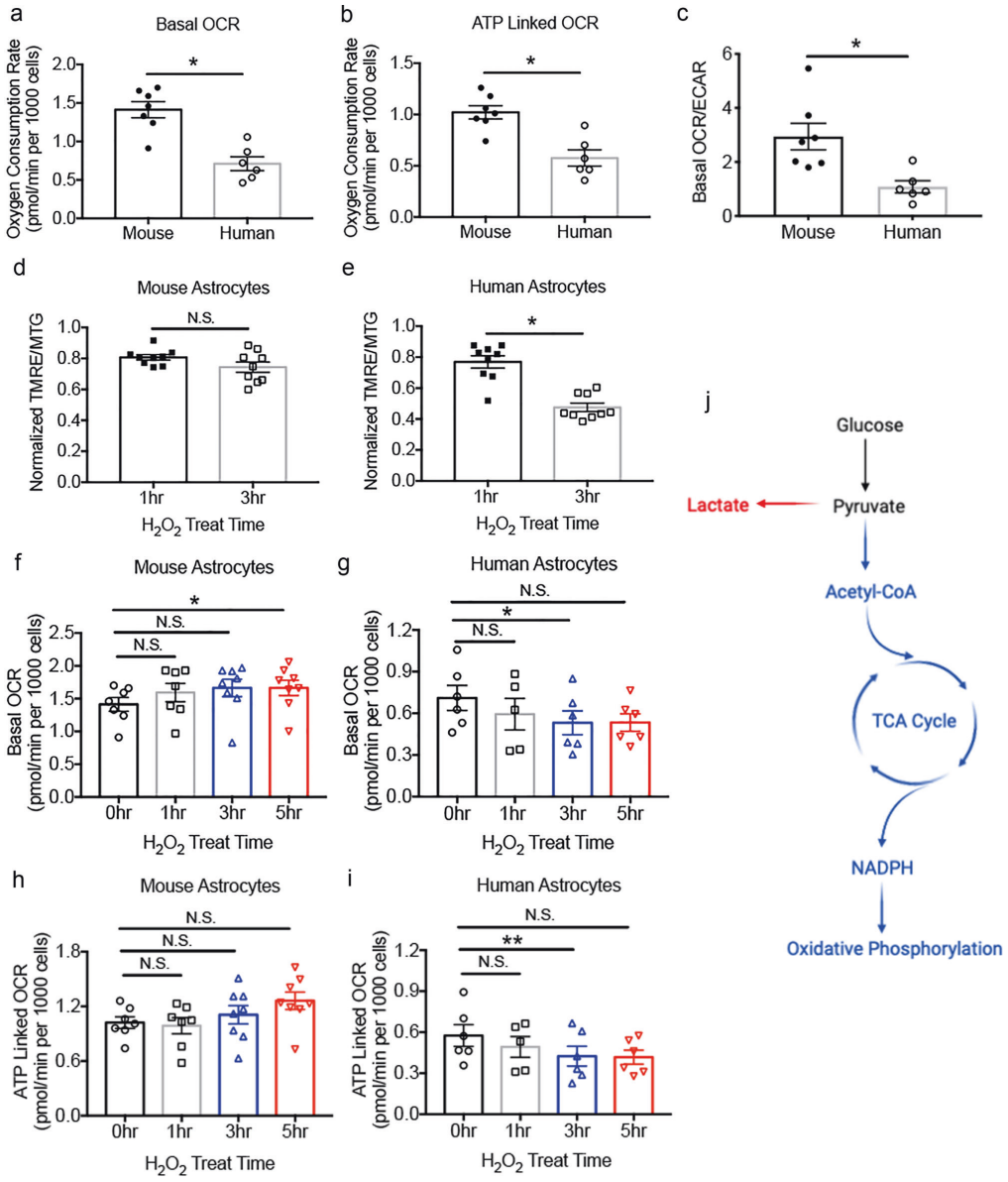


Fig. 3 Human astrocytes are more susceptible to oxidative stress than mouse astrocytes. a Experimental design. Hr, hour. **b** Human and mouse astrocytes treated with H_2O_2 or medium control stained with the live cell dye calcein-AM (green) and the dead cell dye ethidium homodimer (red). Scale bar: 200 μm. **c** Survival rate. Mouse astrocytes: $N = 50$ images from 12 cultures treated with medium control and 54 images from 12 cultures treated with H_2O_2 generated from 6 litters of mice. Human astrocytes: $N = 31$ images from 7 cultures treated with medium control and 26 images from 6 cultures treated with H_2O_2 generated from 3 patients. Data are presented as mean ± SEM in all figures, unless otherwise indicated. Mouse astrocytes: control vs. H_2O_2, $p = 0.0039$. H_2O_2-treated mouse astrocytes vs. H_2O_2-treated human astrocytes, $p = 0.0002$. Human astrocytes: control vs. H_2O_2, $p < 0.0001$. $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ in all figures. Two-way analysis of variance (ANOVA) with Tukey's test for multiple comparisons. N.S., not significant. The *p*-values were calculated using average results from each litter of mice and each patient as independent observations, unless otherwise indicated. **d** Survival rate of astrocytes treated with H_2O_2 normalized to the survival rate of medium control-treated cells. Replicate numbers N are defined in (**c**). $p = 0.0096$. Two-tailed unpaired Welch's *t*-test.

and mouse astrocytes by immunopanning[26,28,41] and performed RNA-seq immediately after purification and after 5–6 days of culture in serum-free conditions (this study). We found that the gene encoding a major peroxisomal ROS detoxification enzyme, catalase[42], is expressed at 3–6-fold higher levels by mouse astrocytes than by human astrocytes [reads per kilobase per million mapped reads (RPKM): acutely purified, 2.04±0.24 for human astrocytes and 5.67±0.38 for mouse astrocytes; in vitro, 1.14±0.10 for human astrocytes and 7.60±0.71 for mouse astrocytes; Fig. 5a]. An additional molecular pathway, the pentose phosphate pathway, produces NADPH, which neutralizes ROS[43]. The rate-limiting step of the



pentose phosphate pathway is catalyzed by glucose-6-phosphate dehydrogenase (G6PD), an important antioxidant enzyme[44]. Using RNA-seq, we found that *G6PD* gene expression is 2–10-fold higher in mouse astrocytes than in human astrocytes (RPKM: acutely purified, 0.17±0.05 for human astrocytes and 2.14±0.48 for mouse astrocytes; in vitro, 2.05 ±0.32 for human astrocytes and 4.39±0.32 for mouse astrocytes; Fig. 5b). The expression levels of *CAT* and *G6PD* were consistently higher in 15 strains of mice compared to humans (Supplementary Fig. 12). We also explored other major detoxification pathways and found generally comparable expression by human and mouse astrocytes. Taken together, higher amounts of catalase and G6PD may protect mouse astrocytes from oxidative stress (Fig. 5d, e).

As we performed all our in vitro functional experiments using developing astrocytes, we next obtained RNA-seq data from adult human and mouse astrocytes (Supplementary Data 8). Notably, the species differences persisted throughout development and adulthood (Fig. 5a, b). Oxidative stress is a core pathological process in a range of neurological conditions, including neurodegenerative disorders

Fig. 4 Mitochondrial metabolism differences between human and mouse astrocytes. a Basal oxygen consumption rate (OCR) of mouse and human astrocytes. Each data point (circle or square) represents one well of astrocyte culture prepared from one human patient or one litter of 8–10 mice throughout this figure. Mouse astrocytes: N = 7 cultures generated from 4 litters of mice in **a**–**c**. Human astrocytes: $N = 6$ wells of cultured cells generated from 3 patients in **a**–**c**. $p = 0.0365$. Two-tailed unpaired Welch's t-test. The *p*-values were calculated using average results from each litter of mice and each patient as independent observations, unless otherwise indicated. **b** OCR linked to ATP production in the presence of oligomycin. $p = 0.0497$. Two-tailed unpaired Welch's t-test. **c** The ratio of OCR to extracellular acidification rate (ECAR). $p = 0.0168$. Two-tailed unpaired Welch's t-test. **d**, **e** Tetramethylrhodamine, ethyl ester (TMRE) fluorescence (reporting mitochondrial membrane potential) normalized by MitoTracker Green (MTG, a general mitochondrial dye) fluorescence. Data represent H_2O_2-treated conditions normalized to medium control-treated conditions. $N = 9$ wells of cultured cells generated from 3 litters of mice and 4 patients. Mouse astrocytes: 3 hr vs. 1 hr, $p = 0.3959$. Human astrocytes: 3 hr vs. 1 hr, $p = 0.0373$. Two-tailed unpaired Welch's t-test. N.S., not significant. **f**, **g** Basal OCR of astrocytes treated with 100 μM H_2O_2. 0 hr and 1 hr mouse astrocytes: $N = 7$ wells of cultured cells generated from 4 litters of mice. 3 hr and 5 hr mouse astrocytes: $N = 8$ wells of cultured cells generated from 4 litters of mice. 0 hr, 3 hr, and 5 hr human astrocytes: $N = 6$ wells of cultured cells generated from 3 patients. 1 hr human astrocytes: $N = 5$ wells of cultured cells generated from 3 patients. Mouse astrocytes: 0 hr vs. 1 hr, $p = 0.1883$; 0 hr vs. 3 hr, $p = 0.4246$; 0 hr vs. 5 hr, $p = 0.0285$. Human astrocytes: 0 hr vs. 1 hr, $p = 0.0954$; 0 hr vs. 3 hr, $p = 0.0235$; 0 hr vs. 5 hr, $p = 0.1837$. One-way repeated measure ANOVA with Dunnett's multiple comparison test. **h**, **i** ATP-linked OCR of astrocytes treated with 100 μM H_2O_2. The replicate numbers are defined in (**f**–**g**). Mouse astrocytes: 0 hr vs. 1 hr, $p = 0.8885$; 0 hr vs. 3 hr, $p = 0.7597$; 0 hr vs. 5 hr, $p = 0.2198$. Human astrocytes: 0 hr vs. 1 hr, $p = 0.1639$; 0 hr vs. 3 hr, $p = 0.0016$; 0 hr vs. 5 hr, $p = 0.2583$. One-way repeated measure ANOVA with Dunnett's multiple comparison test. **j** Diagram of glucose metabolism. TCA, tricarboxylic acid. NADPH, nicotinamide adenine dinucleotide phosphate.

such as Alzheimer's disease, Parkinson's disease, Huntington's disease, and amyotrophic lateral sclerosis. Mouse models of neurodegenerative disorders often have milder phenotypes compared to human patients[1–4]. Our findings suggest that differences in astrocytic responses to oxidative stress may contribute to the increased resiliency of mouse models of neurodegeneration compared to human patients (see Discussion).

To determine whether the lower expression of *catalase* and *G6pd* in humans is an astrocyte-specific attribute or a more general difference across species, we analyzed a single-cell RNA-seq dataset of human and mouse brains[45] to examine the expression of these genes in all major cell types of the brain. We found lower expression of catalase and G6pd in humans than in mice in glutamatergic neurons, GABAergic neurons, oligodendrocyte precursor cells, and oligodendrocytes (Supplementary Fig. 12). Therefore, lower expression of catalase and G6pd is generally observed across multiple cell types in the human brain compared to the mouse brain.

Hypoxia induces a molecular program for neural growth in mice. Adult mouse models of ischemic stroke often achieve full functional recovery[5], whereas adult human stroke patients usually have irreversible functional deficits. Dozens of neural protective drug candidates that improved recovery in mouse models of stroke have failed to show benefits in human clinical trials[46]. Hypoxia is a key physical change in ischemic stroke. Responses of mouse astrocytes to hypoxia have been closely examined previously but responses of human astrocytes are largely unknown.

We exposed human and mouse astrocytes to hypoxia (Fig. 6a) and found that human and mouse astrocytes exhibited similarly high levels of cell survival and had normal healthy morphology under hypoxic and control conditions. We then performed RNA-seq of all treated and control human and mouse astrocytes. To assess transcriptional responses, we used a combination of differential expression and weighted gene co-expression network analysis (WGCNA) (Methods; Supplementary Fig. 13 and Supplementary Data 9).

When we examined the extent to which hypoxia-induced genes are shared between human and mouse astrocytes, we found that 3.4% of (11 out of 322) genes downregulated in human astrocytes were also downregulated in mouse astrocytes and 5.3% of (7 out of 132) genes upregulated in human astrocytes were also upregulated in mouse astrocytes, demonstrating partial conservation of hypoxic responses between human and mouse (Fig. 6d, e; upregulated overlap: 13.8-fold higher than expected by chance; $p = 3.65e{-}07$; downregulated overlap: 6.0-fold higher than expected by chance; $p = 7.50e{-}07$; FDR < 0.05; fold change > 1.5; average RPKM > 1; overlap: genes meeting all three criteria in both species; see also Supplementary Fig. 14). GO terms and KEGG pathway analyses revealed that genes upregulated in both human and mouse astrocytes were enriched in the GO term hypoxia response and the hypoxia inducible factor 1 (HIF1) pathway (Fig. 6f and Supplementary Data 2). Interestingly, astrocytes from both species upregulated genes involved in glycolysis and positive regulation of mitochondrial autophagy. Glycolysis provides an alternative pathway to generate energy without oxygen and autophagy of idling mitochondria may conserve resources within cells.

Despite the partial conservation of hypoxic responses between human and mouse, hypoxia induced stronger molecular changes in mouse astrocytes relative to human astrocytes in terms of the number of differentially expressed genes (454 in mouse vs. 52 in human; fold change > 1.5, FDR < 0.05, average RPKM > 1; Fig. 6b–e) and effect size (Supplementary Fig. 15). Genes upregulated by hypoxia in mouse, but not human, astrocytes were enriched in GO terms such as nervous system development, neurogenesis, neuron differentiation, and axon guidance (Fig. 6g and Supplementary Data 2) and included genes encoding the growth factor Ndnf, morphogen Bmp4, axon guidance molecule Epha5, and cell adhesion molecule Cadm3 (Fig. 6h–o and Supplementary Data 10, 11). Furthermore, we identified a module (grey60) upregulated by hypoxia in mouse astrocytes but not in human astrocytes (Supplementary Data 9 and Supplementary Fig. 13). This module is involved in development and cell adhesion, corroborating our finding that hypoxia induces a molecular program that aids neural repair specifically in mouse astrocytes. These differences may contribute to the differences in functional recovery and responses to drug candidates between human patients and mouse models of stroke.

Among the genes downregulated by hypoxia, we found that genes associated with the GO terms amino acid transmembrane transport and cellular response to nutrient levels were enriched in both human and mouse astrocytes. In contrast, the GO terms L-glutamate transmembrane transport and circadian rhythm were only enriched in downregulated genes in human astrocytes, and the terms cell cycle and electron transport chain were only enriched in downregulated genes in mouse astrocytes (Supplementary Data 2 and Supplementary Fig. 16).

Inflammatory signals induce antigen presentation pathways in humans. Many viruses, such as human immunodeficiency virus,

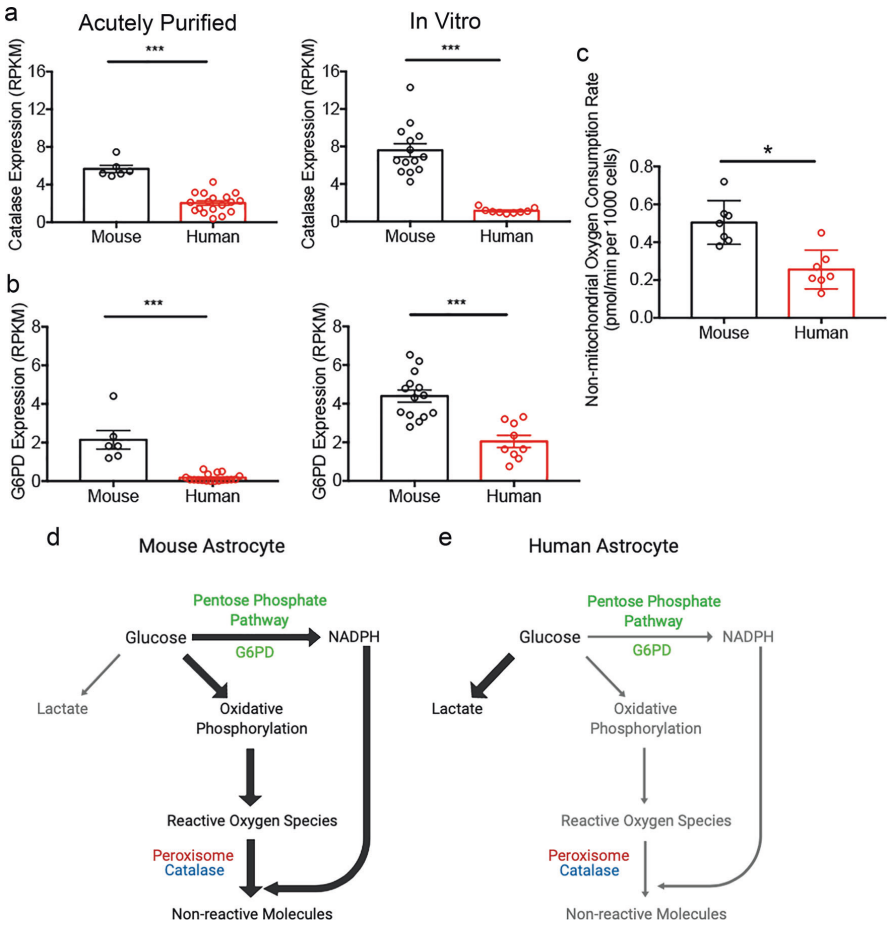


Fig. 5 Detoxification pathway differences between human and mouse astrocytes. a, **b** Expression of ROS detoxification pathway genes catalase (**a**) and glucose-6-phosphate dehydrogenase (G6pd in human/G6pdx in mouse) (**b**) by acutely purified astrocytes and serum-free cultures of astrocytes determined by RNA-seq. $N = 6$ litters of mice and 18 human patients for acutely purified samples. $N = 14$ litters of mice and 9 human patients in vitro. Acutely purified: Catalase, $p < 0.0001$; G6PD, $p < 0.0001$, two-tailed Mann-Whitney test. Serum-free cultures: Catalase, $p < 0.0001$; G6PD, $p < 0.0001$. Two-tailed unpaired Welch's t-test, unless otherwise indicated. Samples include children and adult patients as well as developing and adult mice. The ages of the patients and mice are listed in Supplementary Data 8. **c** Non-mitochondrial OCR measured in the presence of antimycin-A. $N = 7$ wells of cultured cells from each species generated from 4 litters of mice and 3 patients. $p = 0.0126$. Two-tailed unpaired Welch's t-test. The *p*-values were calculated using average results from each litter of mice and each patient as independent observations. **d**, **e** Model of glucose metabolism and detoxification pathways in human and mouse astrocytes. The widths of the arrows represent the rate of the metabolic processes. Mouse astrocytes have higher rates of oxidative phosphorylation, which presumably produce more ROS than human astrocytes. The higher abundance of detoxification pathway genes and the higher peroxisomal activity in mouse compared to human may protect the cells against oxidative damage.

new world alpha viruses, and some flaviviruses (e.g., Zika virus), are capable of infecting CNS cells, inducing neuroinflammatory responses, and causing acute or long-lasting neurological deficits[47,48]. Astrocytes, along with microglia, are CNS-resident cells that modulate neuroinflammation. However, responses of human astrocytes to viral infections and their consequences to CNS homeostasis and function are poorly understood. In addition to viral infections, neuroinflammation is a core pathological component of a range of neurological conditions such as traumatic injury, stroke, neurodegeneration, and aging. tumor necrosis factor alpha (TNFα) is a major pro-inflammatory cytokine involved in neuroinflammation that induces reactivity of mouse astrocytes. Although researchers have long assumed that TNFα induces similar changes in human astrocytes, no study has compared the effect of this key pro-inflammatory cytokine on human and mouse astrocytes.

We exposed human and mouse astrocytes to the viral mimetic double-stranded RNA, poly I:C, or TNFα (Fig. 7a). Both human and mouse astrocytes exhibited similarly high levels of cell survival and had normal healthy morphology under treatment and control conditions. We then performed RNA-seq of all treated and control human and mouse astrocytes. In contrast to

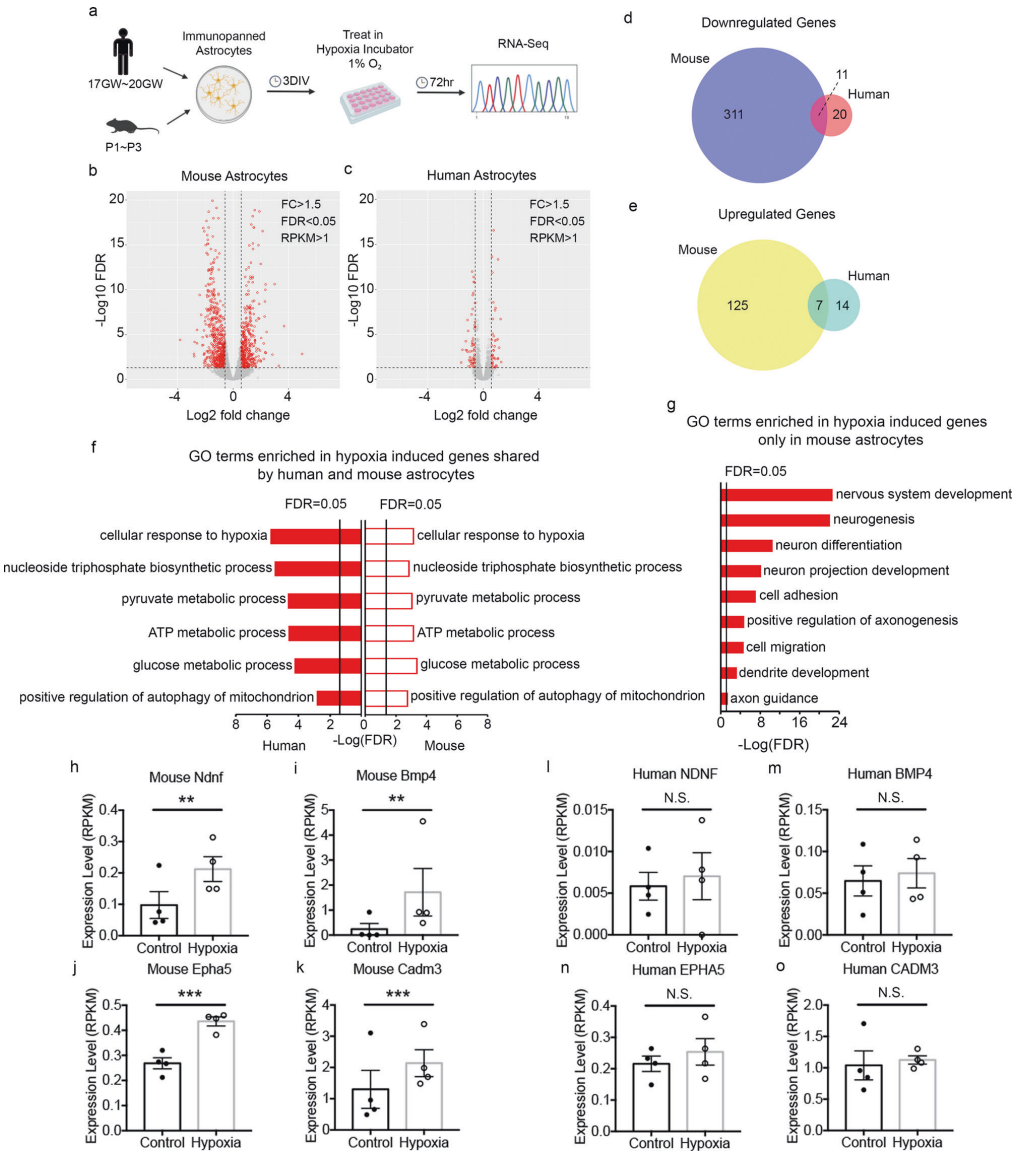


Fig. 6 Molecular responses of human and mouse astrocytes to hypoxia. a Experimental design. **b**, **c** Volcano plots of genes that significantly differ between hypoxia and control conditions. Each red dot represents a significantly different gene. FDR, false discovery rate. FC, fold change. RPKM, Reads Per Kilobase of transcript, per Million mapped reads. **d**, **e** The number of significantly up or downregulated genes in hypoxia-treated human and mouse astrocytes. Genes with FDR < 0.05, fold change > 1.5, and average RPKM of control or treated groups > 1 are shown. Genes in the overlapping regions are those meeting all three criteria in both species. **f** Top shared gene ontology (GO) terms enriched in hypoxia-induced genes in human and mouse astrocytes ranked by FDR. **g** Development-associated GO terms enriched only in hypoxia-induced genes in mouse but not human astrocytes. **h**–**o** Expression of genes associated with the GO term nervous system development in control and hypoxia-treated human and mouse astrocytes. $N = 4$ litters of mice and 4 human patients. Mouse: Ndnf, $p = 0.002$; Bmp4, $p = 0.0015$; Epha5, $p = 0.0003$; Cadm3, $p = 0.0001$. Multiple comparison-adjusted p values were calculated by the DESeq2 package. N.S., not significant.

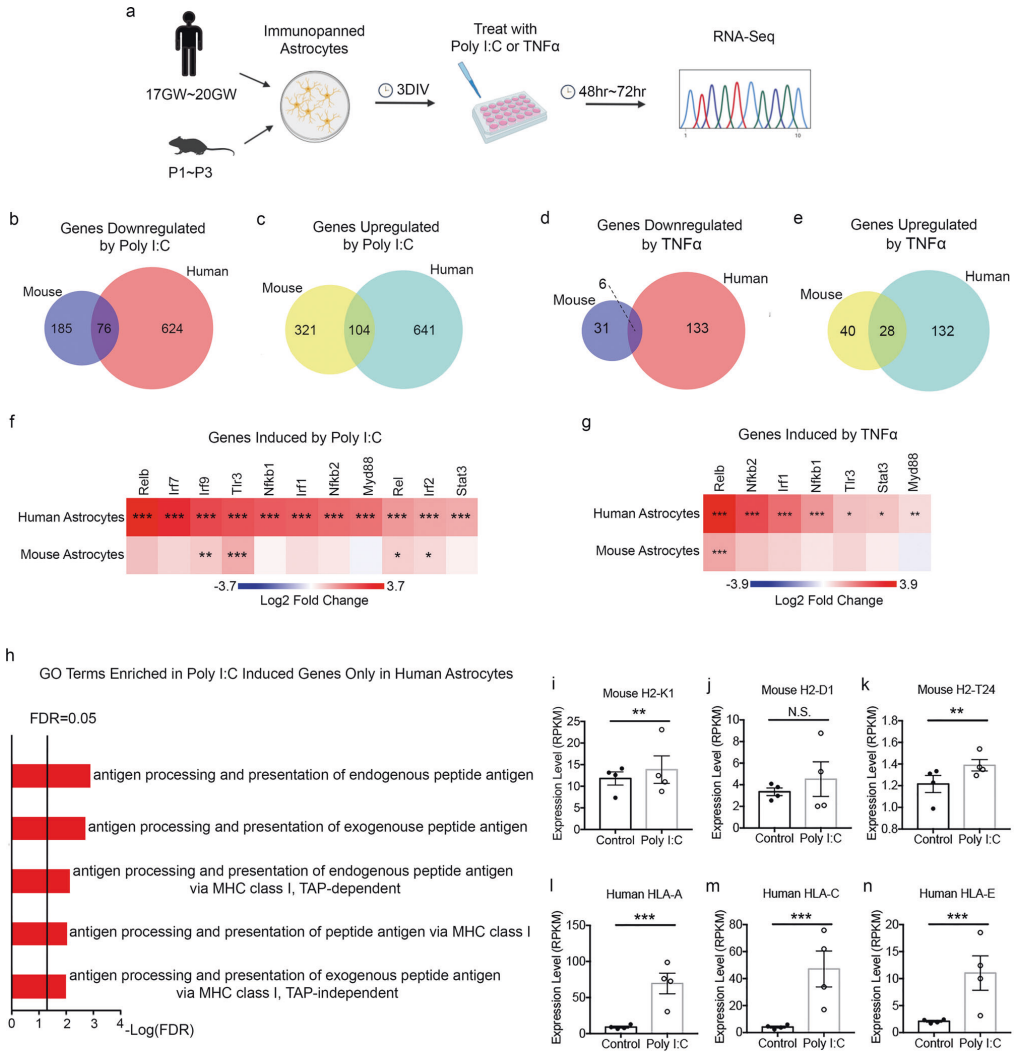


Fig. 7 Molecular responses of human and mouse astrocytes to poly I:C and TNFα. a Experimental design. **b**–**e** The number of significantly up or downregulated genes in poly I:C- and TNFα-treated human and mouse astrocytes. Genes with FDR < 0.05, fold change > 1.5, and average RPKM of control or treated groups > 1 are shown. **f**, **g** Fold changes of *Tlr3, NFκB*, and interferon response pathway genes in poly I:C- and TNFα-treated human and mouse astrocytes. Asterisks represent significance determined by DESeq2. **h** Selected antigen presentation-related GO terms enriched in poly I:C-induced genes only in human astrocytes. **i**–**n** Expression of the top 3 highest-expressing MHC Class I antigen presentation genes in poly I:C-treated and control human and mouse astrocytes. $N = 4$ litters of mice and 4 human patients. Mouse: H2-K1, $p = 0.0051$; H2-T24, $p = 0.0019$. Human: HLA-A, $p < 0.0001$; HLA-C, $p < 0.0001$; HLA-E, $p < 0.0001$. Multiple comparison-adjusted p values were calculated by the DESeq2 package. N.S., not significant.

hypoxia, we found that both poly I:C and TNFα induced stronger molecular responses in human astrocytes relative to mouse astrocytes in terms of the number of differentially expressed genes (Fig. 7b-e and Supplementary Fig. 17a, b, d, e) and effect size (Supplementary Fig. 17c, f). We next examined the extent to which the poly I:C- and TNFα-induced gene changes were shared between human and mouse astrocytes. We found that a significant proportion (10.9%; 76 out of 700) of the down-regulated genes in human astrocytes were also downregulated in mouse astrocytes and a significant proportion (14.0%; 104 out of 745) of upregulated genes in human astrocytes were also upregulated in mouse astrocytes under poly I:C treatment (Fig. 7b, c; upregulated: 1.7-fold higher than expected by chance; $p = 2.79e{-}07$; downregulated: 2.2-fold; $p = 4.89e{-}12$). A significant proportion (4.3%; 6 out of 139) of genes downregulated by TNFα in human astrocytes were also downregulated in mouse astrocytes and a significant proportion (17.5%; 28 out of 160) genes upregulated by TNFα in human astrocytes were also upregulated in mouse astrocytes,

reflecting partial conservation between human and mouse (Fig. 7d, e; upregulated: 13.1-fold; $p = 2.53e{-}25$; downregulated: 5.1-fold; $p = 9.85e{-}04$). We also observed a significant correlation of fold change in mouse vs. human (poly I:C, 0.236. TNFα 0.192; Supplementary Fig. 14). GO term and KEGG pathway enrichment analysis of the conserved genes (Supplementary Data 2) revealed enrichment for genes involved in responses to cytokines and other organisms.

Furthermore, we examined genes induced by poly I:C or TNFα only in human astrocytes. These genes showed enriched GO terms associated with antigen processing and presentation of peptide antigen via major histocompatibility complex (MHC) class I (Fig. 7h and Supplementary Fig. 17h). The expression levels of the three highest expressing MHC Class I genes in human astrocytes (*HLA-A, HLA-C*, and *HLA-E*) and mouse astrocytes (*H2-K1, H2-D1*, and *H2-T24*) are shown in Fig. 7i–n, Supplementary Fig. 17h–m, and Supplementary Fig. 18. These genes showed modest or no increase in poly I:C- or TNFα-treated mouse astrocytes but a consistent and robust increase in human astrocytes. Genes encoding additional MHC Class I-interacting antigen processing and presenting proteins, such as *Tap1, Tap2*, and *ICAM1*, showed similarly robust increases in human astrocytes but no change in mouse astrocytes treated with poly I:C (Supplementary Data 10, 11 and Supplementary Figs. 19 and 20). Interestingly, human induced pluripotent stem cell-derived astrocytes also increased the expression of MHC Class I genes upon TNFα treatment[49], demonstrating the value of stem cell-derived astrocytes in studying species-specific features of human astrocytes.

Among the genes downregulated by poly I:C treatment, the GO term virion assembly was enriched in both human and mouse, potentially revealing a conserved defensive response to viral infections. In addition, poly I:C treatment induced downregulation of genes associated with cell cycle and CNS development only in human astrocytes and genes associated with response to hydrogen peroxide only in mouse astrocytes. No GO term was enriched in genes downregulated by TNFα-treatment in either human or mouse, likely because of the small number of genes downregulated by TNFα in mouse astrocytes. Similar to poly I:C treatment, TNFα induced downregulation of genes associated with cell cycle and CNS development only in human astrocytes. Additionally, TNFα induced downregulation of genes associated with cell communication and glycerolipid metabolism only in mouse astrocytes (Supplementary Data 2 and Supplementary Fig. 16).

To identify coregulated gene networks changing under poly I:C or TNFα treatment, we performed WGCNA and identified a module (black) upregulated in human, but not mouse, astrocytes under both treatment conditions (Supplementary Data 9 and Supplementary Fig. 13). This module is involved in inflammatory responses to double-stranded RNA. The network analyses corroborated the finding that, at the specific dosage of poly I:C or TNFα we used, human astrocytes showed stronger inflammatory responses compared to mouse astrocytes.

Signaling pathways downstream of poly I:C treatment have been well characterized in multiple cell types of the immune system. Poly I:C binds the Toll-like receptor 3 (Tlr3), a pattern-recognition receptor located in endosomes that recognizes the danger signal. TLR3 signals through an adapter protein, Myd88, which activates the nuclear factor kappa B (NFκB) signaling pathway (e.g., Rel, Relb, Nfkb1, Nfkb2). NFκB activation and nuclear translocation, in turn, activates the interferon signaling pathway (interferon responsive genes include *Irf1, Irf2, Ifr7*, and *Irf9*). The NFκB signaling pathway cross-talks with Stat3, the phosphorylation of which is involved in astrocyte reactivity. We found that all of the above-mentioned molecules are strongly upregulated after poly I:C treatment in human astrocytes but showed modest or no upregulation in mouse astrocytes (Fig. 7f and Supplementary Fig. 21a). We found a similar pattern of stronger activation of these genes in TNFα-treated human astrocytes compared to mouse astrocytes (Fig. 7g and Supplementary Fig. 21b).

To assess whether cell death affected our transcriptome analyses, we examined cell survival in acutely purified and cultured human and mouse astrocytes. By trypan blue staining of dead cells, we found close to 100% survival of acutely purified astrocytes from both humans and mice. Astrocyte cultures always have a small proportion of dead cells, but we did not observe differences in cell survival between human and mouse astrocytes (Supplementary Fig. 22). To determine whether cell death may have affected our RNA-seq analyses, we examined the expression of cell death-associated genes[50] in our RNA-seq data. We found low or no expression of these genes in all conditions tested (acutely purified, serum-selected culture, serum-free culture, xenograft, host, hypoxia-, poly I:C-, TNFα-treated, and untreated control astrocytes from both human and mouse; Supplementary Data 12). None of the cell death-associated genes were differentially expressed in any treatments we performed. Therefore, cell death is unlikely to compromise RNA-seq analyses under the conditions we tested.

We next examined whether astrocytes treated with various challenges secrete signals that differentially affect neuronal attributes. We treated cultured human and mouse astrocytes with hypoxia and TNFα, collected astrocyte conditioned medium (ACM), and applied the ACM to mouse cortical neurons. We did not observe differences in neuronal survival, process outgrowth, or NFκB activation between any groups of ACM-treated neurons. We further assessed whether neurons exhibited molecular changes in response to hypoxia or TNFα-treated human and mouse ACM by performing RNA-seq of the treated neurons. We found that neurons treated with hypoxia-mouse ACM showed downregulation of non-coding RNAs such as *Rn7sk* and *Gm24187*. Neurons treated with hypoxia-human ACM showed downregulation of non-coding RNAs such as *Rn7sk, Bc1*, and *Gm24187* (Supplementary Data 13). No protein-coding genes exhibited significant gene expression differences between ACM treatment groups. TNFα-treated human and mouse ACM did not induce significant gene expression changes in neurons. In these experiments, we did not test contact-dependent astrocyte-neuron interactions, which may be interesting to investigate in future studies.

Poly I:C and TNFα induce common transcriptional responses. We next investigated whether different types of perturbations induce a shared core astrocyte reactivity program vs. distinct programs specific to each perturbation. We found that, in both species, very few genes were induced by all three stimuli (i.e., hypoxia, poly I:C, and TNFα; Supplementary Fig. 23). Poly I:C and TNFα induced many common gene changes, but these genes differed greatly from the hypoxia-induced genes, which was corroborated by WGCNA results (Supplementary Fig. 13).

Comparison of treatment-induced changes with neurological diseases. To compare hypoxia-, poly I:C-, and TNFα-induced changes of cultured human and mouse astrocytes with neurological disease-associated changes in human patients and mouse models in vivo, we analyzed single-cell RNA-seq datasets of Alzheimer's disease[51,52], multiple sclerosis[53–55], and healthy control patients and bulk RNA-seq data of glioblastoma-associated astrocytes[56] (see Methods for details). Interestingly, we found that poly I:C- and TNFα-treated human astrocytes exhibit shared gene expression changes with astrocytes from both

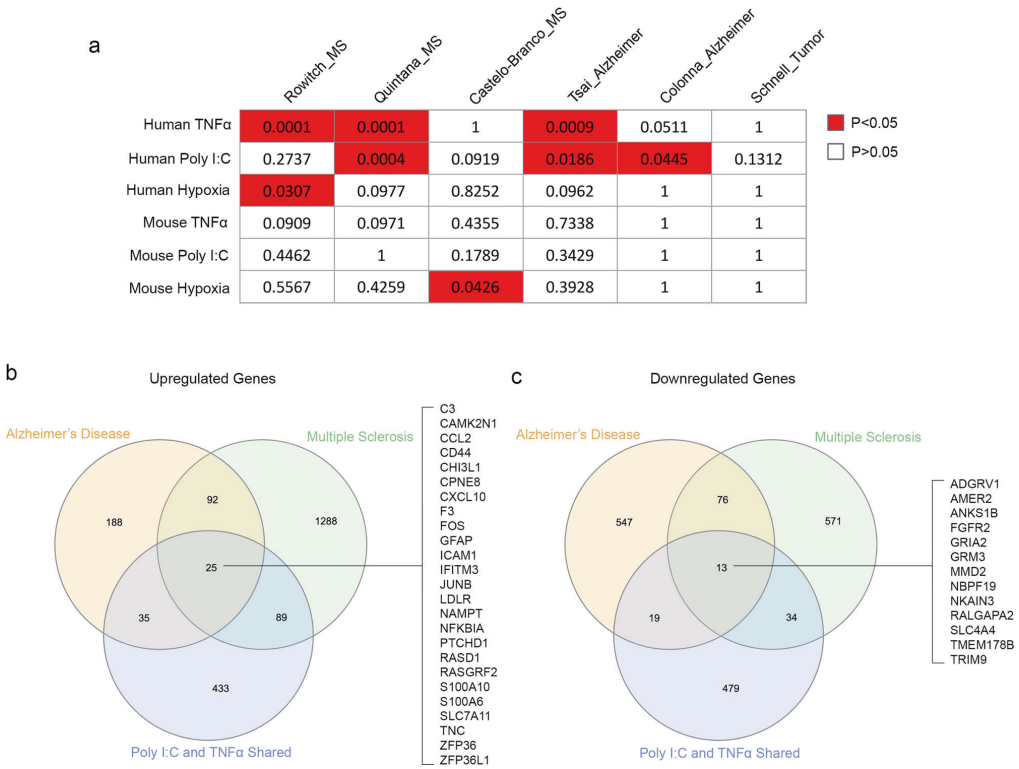


Fig. 8 Core inflammatory astrocyte genes shared in Alzheimer's disease, multiple sclerosis, and poly I:C and TNFα treatments in humans. We analyzed single-cell RNA-seq data from Alzheimer's disease patients, multiple sclerosis patients, and healthy controls. We compared the differentially regulated genes in astrocytes in these diseases with the hypoxia-, poly I:C-, and TNFα-induced genes we identified in cultures of human and mouse astrocytes. **a** Significant concordant gene expression changes are present in disease conditions in vivo and in astrocyte treatments in vitro. To test whether treatment A and disease B exhibited concordant gene expression changes, we counted the number of genes in the following four categories: (1) upregulated in treatment A and upregulated in disease B; (2) upregulated in treatment A and downregulated in disease B; (3) downregulated in treatment A and downregulated in disease B; and (4) downregulated in treatment A and upregulated in disease B. We used the number of genes in each of the four categories in a contingency table and used two-sided Fisher's exact test to detect significant concordance. MS, multiple sclerosis. **b**, **c** We found 38 core inflammatory astrocyte genes shared by Alzheimer's disease, multiple sclerosis, and poly I:C and TNFα treatment in humans, 25 of which are upregulated (**b**) and 13 are downregulated (**c**) in diseases and treatments.

Alzheimer's disease and multiple sclerosis patients (Fig. 8). We found that 38 genes showed consistent changes in astrocytes in Alzheimer's disease, multiple sclerosis, poly I:C treatment, and TNFα treatment (25 genes were upregulated and 13 genes were downregulated by diseases and treatments; Fig. 8b, c). The upregulated genes include those involved in interferon (*IFITM3*), NFκB (*NFKBIA*), and cytokine (*CCL2, CXCL10*) signaling, immediately early genes (*FOS, JUNB*), and calcium signaling modulators (*S100A6, S100A10*). The downregulated genes include one encoding a protein that interacts with amyloid-β precursor protein (*ANKS1B*) and two encoding glutamate receptors (*GRIA2, GRM3*), suggesting changes in astrocytic responses to synaptic glutamate release in multiple neurological disorders. Because both poly I:C and TNFα can induce inflammatory changes, and there are inflammatory changes in Alzheimer's disease and multiple sclerosis, these 38 genes are likely a core group of signature inflammatory astrocyte genes in humans. Using these genes as markers may facilitate the identification of inflammatory-reactive astrocytes in multiple diseases in the future. In contrast to poly I:C- and TNFα-treated human astrocytes, we did not detect any significant correlation between gene expression changes of poly I:C- or TNFα-treated mouse astrocytes and Alzheimer's disease or multiple sclerosis patients, highlighting species-dependent gene signatures in astrocyte reactivity (Fig. 8a).

We detected a weak correlation of gene expression changes in hypoxia-treated human astrocytes with a small subset of disease datasets. Therefore, hypoxia-induced changes are likely distinct from changes of astrocytes in Alzheimer's disease or multiple sclerosis. We did not observe any correlation of glioblastoma-associated astrocyte gene expression changes with any treatment-induced changes in human or mouse astrocytes.

We next compared our identified treatment-induced gene expression changes with gene expression changes in the astrocytes of two mouse models of neurological disorders, bacterial endotoxin lipopolysaccharide-induced inflammation and ischemic stroke[27], also referred to as A1 and A2 astrocytes in the literature[57,58]. We did not observe any significant correlation

between any of the treatments and A1 or A2-specific gene expression changes (Supplementary Fig. 24).

Astrocyte heterogeneity and hypoxia-, poly I:C-, and TNFα-induced changes. To examine astrocyte heterogeneity, we assessed NFκB activation at the single-cell level in poly I:C-treated human and mouse astrocytes. We performed immunostaining with an antibody against the NFκB component p65, which is localized at the nuclei when NFκB signaling is activated[59]. Upon poly I:C treatment, a larger subpopulation of human astrocytes compared to mouse astrocytes exhibited NFκB activation (Supplementary Fig. 25; mouse: 0.9±0.4% in control, 2.9±0.5% in poly I:C-treated; human: 1.0±0.2% in control, 13.9±1.7% in poly I:C-treated; mouse poly I:C-treated vs. human poly I:C-treated: $p < 0.001$), revealing species-dependent properties of astrocyte subpopulation dynamics.

To assess whether hypoxia, poly I:C, or TNFα may induce changes of previously characterized astrocyte subpopulations, we compared our treatment-induced gene expression changes with astrocyte subpopulation markers from single-cell RNA-seq studies[51,54,60,61]. We found significant concordance between poly I:C- and TNFα-treated human astrocytes with astrocyte cluster 3 reported by Tsai and colleagues[51] and anti-correlated gene expression between poly I:C- and TNFα-treated human astrocytes with astrocyte cluster 1 reported by Schwartz and colleagues[61] (Supplementary Fig. 26). Tsai et al. astrocyte cluster 3 expresses well-known reactive astrocyte markers, such as *glial fibrillary acidic protein (GFAP), IFITM3*, and *CD44*. These observations suggest that poly I:C and TNFα treatment increase the subpopulation of astrocytes with reactive characteristics, which could result from dynamic gene expression changes and/or selective proliferation/depletion of subpopulations. In contrast to poly I:C and TNFα treatment, we did not observe concordant gene expression changes between hypoxia treatment and any reported astrocyte subpopulations. We did not detect concordant gene expression between any of our treatments with any astrocyte subpopulations reported by Regev and colleagues[60].

Discussion

In this study, we evaluated the conservation and divergence of astrocytic responses to disease-relevant perturbations between human and mouse. We used methods for isolating, culturing, and stimulating resting/homeostatic astrocytes from developmentally matched human and mouse astrocytes and applied equivalent, controlled experimental paradigms for direct comparison. We identified several important differences between both resting and reactive human and mouse astrocytes: (1) The rates of mitochondrial resting state respiration differed between mouse and human astrocytes. (2) Human astrocytes were more susceptible to oxidative stress than mouse astrocytes, potentially contributing to the observed differences in neurodegeneration between mouse models and human patients. (3) Hypoxia induced a pro-growth molecular program in mouse but not human astrocytes, potentially underlying the greater functional recovery that occurs in mouse models of ischemic stroke compared to human stroke patients. (4) Poly I:C and TNFα induced antigen-presenting genes in human but not mouse astrocytes.

Utilizing knowledge on the conservation and divergence of human and mouse astrocytes for translational research

Identifying conserved and divergent cellular processes. We found extensive conservation in gene expression levels between human and mouse astrocytes in some cellular processes and divergence in others. For example, genes with similar expression levels in human and mouse astrocytes include those involved in mRNA metabolic processes, intracellular transport, and glial cell differentiation, whereas mitochondrial metabolism and cytokine signaling genes are divergent across species. Therefore, findings on mRNA metabolic processes, intracellular transport, and glial cell differentiation using mouse models may be readily translatable to humans. By contrast, more caution must be taken before extrapolating mitochondrial and cytokine findings from mouse models to human patients.

"Humanizing" mouse models of diseases. Mouse models of neurodegeneration often have less severe defects compared to human patients[1–4]. Oxidative stress is a critical pathological process in neurodegeneration. Our finding of greater resilience of mouse astrocytes to oxidative stress compared to human astrocytes suggests that reducing detoxification activities in mouse models of neurodegeneration (for example, using Catalase heterozygous or knockdown mice) may improve the resemblance of these models to human patients.

Improving neural repair in humans by investigating repair mechanisms in mice. Mouse models of ischemic stroke typically exhibit spontaneous functional recovery[5,62], whereas human patients often have limited functional recovery and permanent disabilities. We showed that hypoxia induced HIF1 pathway activation, increases in glycolysis, and stimulation of autophagy in both species. However, hypoxia induced a pro-growth molecular program specifically in mouse astrocytes; human astrocytes were able to sense an oxygen shortage and make adaptive changes but stopped short of activating the pro-growth program. Investigating how signal transduction occurs in mouse astrocytes that links hypoxia to neuronal growth genes may lead to therapies that activate the pro-neuronal growth program in human astrocytes.

Species differences in energy metabolism in astrocytes and other cell types. A few genes are associated with the expansion of the cerebral cortex in human evolution[63–68]; one such gene encodes a protein targeted to mitochondria, implicating metabolic changes in human brain evolution[69]. It is unclear whether the species differences in metabolism that we identified are specific to astrocytes, but we found consistent species-dependent expression of genes associated with reactive oxygen species detoxification in multiple cell types in the brain. Other studies have reported transcriptome and developmental differences between human and mouse brains[29,70–72]. However, very few studies have compared the respiration rates of human and mouse cells. One study that compared the metabolism of human and mouse muscle cells reported mixed results[73]. At the organism level, our results are consistent with the observation that smaller mammals typically have higher metabolic rates per unit body weight than larger mammals[74].

Mitochondrial and energy metabolism changes are important in the pathogenesis of many neurological disorders. For example, many genes associated with Parkinson's disease risk are involved in mitochondrial function[75], a large set of genes involved in metabolism are induced after traumatic brain injury[76], and impairment of glycolysis-derived metabolites in astrocytes contributes to cognitive deficits in Alzheimer's disease[77]. Previous studies have not directly compared the mitochondrial function and energy metabolism between human and mouse for any cell type of the central nervous system, to the best of our knowledge. Our discovery of mitochondrial and energy metabolism differences between human and mouse cells should be taken into consideration in translational research.

Potential species-specific interactions between astrocytes and neurons. Astrocytes are important for the development and function of neurons. Previous studies have shown that transplantation of human astrocytes into mouse brains affects neuronal function and learning and memory[30]. In this study, we compared the impact of secreted signals from human and mouse astrocytes on neurons and did not observe significant species differences. It is likely that contact-dependent interactions between astrocytes and neurons exhibit species-specific attributes. It is also possible that the concentration of astrocyte-conditioned media we used is not high enough to induce detectable species-dependent effects.

Potential limitations of the study. All in vitro experiments in this study were performed using developing human and mouse astrocytes. Therefore, we do not recommend extrapolation of our conclusions to adult and/or aging contexts without further investigation. Although it is important to directly compare adult human and mouse astrocytes, it is challenging to obtain large numbers of fresh healthy brain tissue donations from adults to characterize astrocytic responses to disease-relevant stimuli with sufficient statistical power. Nevertheless, we performed RNA-seq of astrocytes purified from healthy brain tissue donated from adults (Supplementary Data 8)[26]. Analysis of adult astrocyte RNA-seq data showed human-mouse divergent pathways consistent with our in vitro findings, suggesting that the potential species differences are similar between developing and adult stages (Fig. 5a, b).

Methods

Lead contact and materials availability. Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Ye Zhang (yezhang@ucla.edu). This study did not generate new unique reagents.

Experimental animals. All animal experimental procedures were approved by the Chancellor's Animal Research Committee at the University of California, Los Angeles (UCLA) and were conducted in compliance with national and state laws and policies. The research protocol for the transplantation of human cells into host mouse brains to create chimera model was approved by the Chancellor's Animal Research Committee at UCLA and were conducted in compliance with national and state laws and policies. We used C57BL6 mice group-housed in standard cages (2–3 per cage). Rooms were maintained on a 12-h light/dark cycle at 20–26 °C, 30–70% humidity. Euthanasia and preparation of primary cultures of astrocytes were performed during the light cycle. For each astrocyte culture batch, 8–10 mixed-sex pups at P1–3 from 1 to 2 L were combined.

Human tissue samples. Fetal human brain tissue without identifiable personal information was obtained following elective pregnancy termination with exemption determination from the UCLA Office of the Human Research Protection Program. All donors have provided informed consent. The consent forms indicate that the donation is voluntary, refusal to donate tissue will not affect the donor's medical care or their relationship with their physicians, the donated material will be used for purposes of education, research, or for the advancement of medical science, and that there will be no payment to the donor. The next of keen were not involved in providing informed consent. Samples from patients with genetic disorders such as Down's syndrome were excluded from the study when known. Gestational week 17–20 brain tissue was immersed in 4 °C Dulbecco's phosphate buffered saline (PBS, Gibco, 14040182) and transferred to the lab for tissue dissociation. In cases with largely intact brain tissue, we used whole cerebral cortex for astrocyte purification. In cases with fragmented tissue, we used fragments most likely to be cerebral cortex (typically large thin sheets). We included both female and male brain tissue. Sample sizes are noted in figure legends for each experiment. Results on astrocytes from children and adults was obtained by analyzing a previously published dataset[26].

Primary cell culture. Primary astrocyte cultures from humans and mice were generated by immunopanning and were maintained in a humidified 37°C incubator with 10% CO_2 (see method details below). Cells from both females and males were used.

Immunopanning purification of astrocytes. To examine responses of human and mouse astrocytes to stressors, we purified human and mouse astrocytes from developmentally equivalent stages, to the best of our knowledge. In humans, astrocytogenesis starts during the second trimester and continues through the third trimester[78–80]. Human astrocytes reach maturity roughly around one year of age as determined by gene expression[70–72], although their physiological and functional maturation timeline is unclear. In mice, astrocytogenesis starts at the perinatal period (embryonic day 17.5 [E17.5] of a 19-day gestation) and peaks between P0 and P14[81]. Mouse astrocytes reach maturity roughly around one month of age as determined by morphology and gene expression[82]. A single-cell RNA-seq study of human and mouse brains found that molecular features of gestational week 16–20 human brains are similar to P0-P5 mouse brains[83]. Therefore, we purified astrocytes from gestational week 17–20 human brains and P1–3 mouse brains. Within those age ranges, we did not observe age-dependent differences in any of the assays we tested.

We started astrocyte purification experiments using human and mouse brain tissue within similar postmortem intervals. For human samples, we received tissue within 30 min to 1 h postmortem. We then performed a very simple dissection procedure that takes <3 min. For mouse samples, we combined one to two litters of 8–10 mice to get enough cells for each experiment. We combined both male and female mice for all experiments, although sex may not have been equally represented in each litter of mice. Cerebral cortex dissection from all the mice typically took ~45 min before we started the astrocyte purification experiments. We purified human and mouse astrocytes according to a previously published immunopanning protocol[26,28,41]. Briefly, we coated three 150 mm-diameter petri dishes first with species-specific secondary antibodies and then with an antibody against CD45 (BD550539, both human and mouse, 10 µl antibody in 12 ml buffer per panning plate), a hybridoma supernatant against the O4 antigen (mouse, 4 ml hybridoma supernatant in 8 ml buffer per panning plate) or an antibody against CD90 (BD550402, human, 20 µl antibody in 12 ml buffer per panning plate), and an antibody against HepaCAM (R&D Systems, MAB4108, 10 µl antibody in 12 ml buffer per panning plate), respectively. We dissected cerebral cortices from human and mouse in PBS and removed meninges. We then dissociated the tissue with 6 units/ml papain at 34.5°C for 45 min. We mechanically triturated the tissue with 5 ml serological pipets in the presence of a trypsin inhibitor solution. We then depleted microglia/macrophages, oligodendrocyte precursor cells, and neurons from the single-cell suspension by incubating the suspension sequentially on the CD45, O4 (for mouse), or CD90 (for human) antibody-coated petri dishes. We incubated the single-cell suspension on the HepaCAM antibody-coated petri dish. After washing away nonadherent cells with PBS, we lifted astrocytes bound to the HepaCAM antibody-coated petri dish using trypsin and plated them on poly-D-lysine-coated plastic coverslips in a serum-free medium containing Dulbecco's modified Eagle's medium (DMEM) (Life Technologies, 11960069), Neurobasal (Life Technologies, 21103049), sodium pyruvate (Life Technologies 11360070), glutamine (Life Technologies, 25030081), N-acetyl cysteine (Sigma, A8199), and heparin-binding EGF-like growth factor (Sigma, E4643). For most of the H_2O_2, TNFα, hypoxia, and poly I:C treatment experiments, with exceptions detailed below, astrocytes were plated on 24-well culture plates at 75–100k per well. Human and mouse astrocyte cultures had similar final densities for every type of experiment. For high-density cultures for H_2O_2 treatment, 30k astrocytes were plated in a 50 µl droplet in the middle of pre-dried poly-D-lysine-coated plastic coverslips on 24-well plates. After allowing the cultures to settle for 20 min at 37°C, additional media were added. For Seahorse Respiration Assays, astrocytes were plated at 100–250k/well in Agilent Seahorse 96-well cell culture microplates (cat#101085-004). For TMRE/MTG imaging, astrocytes were plated at 25–50k/well on dark-walled flat-bottom 96-well assay plates (Corning, cat#3603). For poly I:C treatment, astrocytes were plated directly on poly-D-lysine-coated 24-well culture plates (Fisher, cat#08-772-1) without coverslips because poly I:C addition often causes cell to float away from the coverslips. To purify xenografted human astrocytes and host mouse astrocytes from adult host mouse brains, we dissociated whole brains using 20 units/ml papain, depleted microglia/macrophages, oligodendrocytes, and oligodendrocyte precursor cells with anti-CD45 antibody-, GalC hybridoma supernatant-, and O4 hybridoma supernatant-coated plates, respectively. Three consecutive plates with the same antibody were used for depletion of each cell type. We then collected astrocytes with anti-HepaCAM antibody-coated plates. The general procedures we used for the purification of human and mouse astrocytes are based on a previously developed method for purifying rat astrocytes[41], although we used different versions of antibodies for the isolation of cells from different species.

Serum-selection purification of astrocytes. Human brain tissue was dissociated into single-cell suspensions as described above and plated on poly-D-lysine-coated 25 cm^2 culture flasks (VWR, cat#10861-672) in DMEM (Gibco, cat#11960044) with 10% fetal bovine serum (Gibco, cat#16140071) and 2 mM glutamine. After 4-6 days, we vigorously shook off the cells in the top layer (neurons and other glia) and left the astrocytes on the bottom layer. We then harvested astrocytes for RNA-seq.

RNA-seq. We purified total RNA using the miRNeasy Mini kit (Qiagen, cat#217004) and analyzed RNA concentration and integrity with TapeStation (Agilent) and Qubit. All samples showed RNA integrity numbers higher than 8.4.

We then generated cDNA using the Nugen Ovation V2 kit (Nugen), fragmented the cDNA using a Covaris sonicator, and generated sequencing libraries using the Next Ultra RNA Library Prep kit (New England Biolabs) with 9–10 cycles of PCR amplification. We sequenced the libraries with Illumina HiSeq 4,000 and NovaSeq sequencers and obtained 16.3 ± 5.7 million (mean ± standard deviation) 50 bp and 100 bp single-endreads per sample.

RNA-seq data analysis. We mapped sequencing reads to human genome hg38 and mouse genome MM10 using the STAR package and HTSEQ to obtain raw counts. We then used the EdgeR-Limma-Voom packages in R to obtain RPKM values. We calculated differential gene expression with the DESeq2 package. Statistical significance of the overlap between two groups of genes was determined using http://nemates.org/MA/progs/overlap_stats.cgi. Significance of the difference between two correlation coefficients was calculated using http://vassarstats.net/rdiff.html?.

Comparison of transcriptomes of acutely purified astrocytes. We mapped RNA-seq data from our previously obtained acutely purified human and mouse astrocyte datasets[26,28] as described above. The ages of the samples is described in Supplementary Data 8. We calculated percentile rankings of RPKM values of each gene in each human and mouse astrocyte sample. We excluded genes with maximal percentile rankings across all samples < 0.33, as these genes are not expressed or very lowly expressed in all samples. We then performed Welch's T-test between human and mouse samples and multiple-comparison post-hoc adjustment using the FDR method. Genes with FDR values < 0.05 and human-mouse percentile ranking differences > 0.4 were used for GO and cellular component analyses using string-db.org. Test gene lists were compared to background gene lists including all genes expressed at RPKM > 0.05 in astrocytes.

Comparison of human data to transcriptome data of 14 mouse strains. To test whether the human-mouse astrocytic gene expression differences are specific to the C57/BL6 strain we used, we compared our data to an RNA-seq study of mouse hippocampus from multiple strains (data are available from 15 strains)[84]. Notably, the Neuner study[84] and our study differ in technical details. Therefore, to avoid the impact of technical batch effects in the comparison of our human data to mouse data from 15 different strains, we took advantage of the fact that both studies performed RNA-seq of the C57/BL6 mouse strain. We divided the expression of each gene from our human samples by the average expression in our C57/BL6 mouse samples to obtain normalized expression of each gene from each sample. Similarly, we divided the expression of each gene in each of the 14 strains (other than C57/BL6) from the Neuner study by the average expression in the C57/BL6 strain determined by the Neuner study to obtain normalized expression of each gene from each sample in the Neuner study. We then compared normalized expression in our study to normalized expression in the Neuner study. We avoided direct comparison of expression levels (e.g., RPKM/ fragments per kilobase per million mapped reads (FPKM)/transcripts per million (TPM)) across studies because it would be complicated by technical batch effects.

Comparison of single-cell RNA-seq data of human and mouse astrocytes. To validate our observed human mouse astrocyte gene expression differences, we utilized single-cell expression data derived from human and mouse cortex[29]. For human and mouse, respectively, we utilized the available trimmed-mean and median expression TPM values for all genes in each of their identified cell-type clusters. To calculate the human-mouse expression fold-change difference, we calculated the mean expression of astrocyte clusters and compared the mean expression between species. This fold-change species difference was compared to the fold-change species difference calculated using acutely purified astrocytes from both human and mouse.

WGCNA. Expression values from human and mouse were merged into a single expression matrix using only one-to-one human-mouse orthologues. Genes were retained if they had > 20% non-zero values and were subsequently log2 (+0.001) transformed. We combined all conditions in the analyses. We removed expression variation unrelated to the effect of treatment using the linear regression model "expr ~ (1 | replicate)". This maintained differences within each replicate pair, capturing the effect of a treatment, but regressed out differences between replicate pairs such as basal species differences or technical differences such as sequencing batch. Network analysis was performed through WGCNA using biweight mid-correlation (bicor) to reduce sensitivity to outliers. A soft threshold power of 18 was chosen to achieve scale-free topology ($r^2 > 0.8$). The topological overlap matrix was hierarchically clustered and modules were defined using a minimum module size of 50 and deepSplit cut of 2. Module-trait correlations were used to assess whether a module was significantly associated with a particular treatment in a particular species.

H_2O_2 treatment. We treated human and mouse astrocytes cultured in 24-well plates with 100–500 μM H_2O_2 (Sigma, cat#95321-100 ML) and performed the cell survival assay, Seahorse respiration assay, and mitochondrial membrane potential assay described below.

Cell survival assay. We incubated human and mouse astrocytes with the live cell dye calcein-AM and the dead cell dye ethidium homodimer using the LIVE/DEAD™ Viability Kit (Invitrogen, cat#L3224) for 10 min at room temperature protected from light and imaged the cells with an Evos FL Auto 2 inverted fluorescence microscope (Invitrogen) with a 10× lens.

Seahorse respiration assay. We cultured human and mouse astrocytes with the media detailed above and changed it to Seahorse assay medium with 10 mM glucose, 2 mM glutamine, 1 mM pyruvate, and 5 mM HEPES on the day of the Seahorse respiration assay. We used an Agilent Seahorse XFe96 Analyzer to measure oxygen concentration and extracellular pH changes. We first measured basal oxygen consumption rates in unperturbed conditions. We then added oligomycin to inhibit ATP-synthase (mitochondrial complex IV). The differences between the basal and oligomycin conditions reflect the amount of oxygen consumption used for ATP production. We next added carbonyl cyanide-4 (trifluoromethoxy) phenylhydrazone (FCCP), an uncoupling agent that collapses the proton gradient and disrupts the mitochondrial membrane potential. As a result, electron flow through the electron transport chain is uninhibited, and oxygen consumption by complex IV reaches the maximum amount. Lastly, we added antimycin A to block complex III and shut down mitochondrial respiration. In the presence of antimycin A, the measured respiration rate represents non-mitochondrial respiration, with major contributions from peroxisomes. We took measurements every 5 min for 3–4 data points per condition. We sequentially added 2 μM oligomycin, 0.5 μM and 0.9 μM FCCP, and 2 μM antimycin A. After taking measurements, we stained cells with DAPI and counted the number of cells in each sample. Results were then normalized by cell number. We used the Agilent Seahorse Wave software to analyze Seahorse assay data.

Mitochondrial membrane potential assay. We loaded cultured human and mouse astrocytes with 14 nM TMRE, 200 nM MTG, and 1 μg/ml Hoechst for 45 min, treated the cells with 100 μM H_2O_2, and then measured fluorescence at 1 and 3 h after H_2O_2 treatment. After staining, the cells were washed three times with culture medium containing 14 nM TMRE to remove extra MTG and Hoechst dyes. TMRE and MTG fluorescence were imaged with an Operetta High-Content Imaging System (PerkinElmer). Fluorescence intensity after H_2O_2 treatment was normalized to untreated control.

Hypoxia treatment. We first cultured immunopanned human and mouse astrocytes at atmospheric oxygen concentrations for three days. We then cultured them at 1% oxygen for three days. Control cells were cultured at atmospheric oxygen concentrations for 6 days. We then harvested RNA for RNA-seq.

Poly I:C treatment. We cultured immunopanned human and mouse astrocytes for three days. We then added 200 μg/ml poly I:C (Sigma, cat#P1530-25MG) to the culture medium and cultured the cells for an additional three days. We then harvested RNA for RNA-seq.

TNFα treatment. We treated human and mouse astrocytes cultured for 3 days with 30 ng/ml TNFα for 48 h and harvested the cells for RNA-seq. We treated mouse astrocytes with TNFα from human (Cell Signaling Technology, 8902SF) and mouse (Cell Signaling Technology, 5178SF) sources and sequenced them in separate experiments. A similar number of genes were induced in mouse astrocytes by TNFα from human and mouse sources. We used cells treated with human TNFα for subsequent analyses.

Transplantation of human astrocytes into host mouse brains. We transplanted human astrocytes into host mouse brains according to published protocols[30–32]. Briefly, we purified human astrocytes as described above under the serum-selection purification of astrocytes section. We then injected 100,000 cells per μl, 1 μl per injection, and 4 injections per mouse at age P2–11. We used Rag2 immunodeficient mice to avoid graft rejection. The mice were maintained in autoclaved cages with autoclaved food and water in a pathogen-free facility.

Mapping xenograft reads to combined human-mouse reference genome. RNA-seq data were assessed for quality parameters using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and then trimmed with Trim_galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). The RNA-seq reads were then mapped to an in silico combined human-mouse reference genome. Briefly, reference genome and gene annotation files of human (hg38) and mouse (mm10) were downloaded from GENCODE. Human chromosomes were tagged as "chr" and mouse chromosomes were renamed as "m.chr". The two fasta files for human and mouse were then concatenated and indexed using STAR aligner, allowing only one top-scored locus to be mapped if multiple mappings occur. Benchmarking results showed low false-alignment rates in both pure human

(0.74%) and pure mouse RNA-seq (2.74%). After alignment, bam files were separated based on the "chr" (human) and "m.chr" (mouse) labels, followed by read counting using Rsubread to obtain the corresponding count matrix.

Non-supervised hierarchical clustering. We performed clustering in R using the hclust() function.

Comparison of host vs. naïve mouse astrocytes. We compared the transcriptomes of host mouse astrocytes (average age 8 months) to naïve mouse astrocytes of a similar age (7 or 9 months) from our earlier study[26]. The percentile ranking of each gene was calculated based on RPKM and differences in percentile were tested by the t-test with post-hoc Bonferroni correction for multiple comparisons.

Comparison with human patients and mouse models of disease. To analyze Alzheimer's disease- and multiple sclerosis-associated changes, we used single-cell RNA-seq datasets[51,52,54,55]. For glioblastoma-associated changes, we used a bulk RNA-seq dataset from purified astrocytes[56]. For single-cell RNA-seq datasets, we only used differentially expressed genes in the astrocyte clusters. The comparison of our treatment-induced signature from bulk RNA-seq data and published disease signature from single-cell RNA-seq data must be conducted carefully to avoid technical bias. There are major differences in the sample sources, sample collection methods, and sequencing parameters between datasets. Notably, single-cell and bulk RNA-seq data differ substantially in dynamic range. Therefore, direct comparison of counts or RPKM/FPKM/TPM between single-cell and bulk RNA-seq datasets may be problematic. To perform comparisons with minimal complications from technical variants, we compared the overlap of differentially expressed gene lists from our treatment study and published disease studies. If treatment A and disease B induce similar gene signature changes, we expect to find significantly more genes changed in concordant directions in A and B compared to genes changed in opposite directions in A and B. If a treatment and a disease do not induce similar gene signature changes, we expect to find similar numbers of genes changing in concordant vs. opposite directions in these two conditions as predicted by chance. To test whether treatment A and disease B exhibited concordant gene expression changes, we counted the number of genes in the following four categories: (1) upregulated in treatment A and upregulated in disease B; (2) upregulated in treatment A and downregulated in disease B; (3) downregulated in treatment A and downregulated in disease B; and (4) downregulated in treatment A and upregulated in disease B. We used the number of genes in each of the four categories in a contingency table and used Fisher's exact test to detect significant concordance. We used the lists of genes differentially expressed by astrocyte clusters between disease and control patients from the published disease studies, which used the statistical tests and parameters detailed in these publications[51,52,54,55]. Specifically, the following gene lists were used for this analysis: Mathys et al.[51], Supplementary Data 2, astrocyte cluster, no pathology vs. pathology; Zhou et al.[52], Supplementary Data 4, DEG tab, astrocyte cluster, Alzheimer's disease vs. control; Schirmer et al.[53], Supplementary Data 6, astrocyte cluster, multiple sclerosis vs. control; Wheeler et al.[54], Supplementary Data 10 and 12, astrocyte cluster, multiple sclerosis vs. control; and Heiland et al.[56], Fig. 1b, tumor vs. control.

Comparison with astrocyte subpopulation markers. We analyzed four previously published single-cell RNA-seq datasets from humans with subclusters of astrocytes[51], using a similar methodology as described above for comparison with disease datasets. When the number of genes was <1000, we used Fisher's exact test; when the number of genes was equal to or more than 1000, we used Chi-square test. We next performed Bonferroni correction for multiple comparisons to identify concordant gene expression between each treatment and each astrocyte subcluster. The following astrocyte subcluster differentially expressed gene lists were used in this analysis: Mathys et al.[51], Supplementary Data 6; Wheeler et al.[54], Supplementary Data 8; Habib et al.[60], Supplementary Data 8; and Habib et al.[61], Supplementary Data 2.

ACM treatment of neurons. We plated primary human and mouse astrocytes purified by immunopanning as described above in high-density cultures on 6-well plates. To obtain TNFα-treated ACM, we added 30 ng/ml TNFα to the astrocytes at 3 div and harvested ACM at 6 div. To obtain hypoxia-treated ACM, we cultured the astrocytes in atmospheric (21%) oxygen for 3 days and moved the cultures to an incubator with 1% oxygen for 3 days and collected ACM at the end of the treatment. We also collected untreated control ACM. We concentrated ACM with centrifuge tubes with 3 kilodalton filters (Thermo Scientific, 88525) and spun them at 6000–8000 g for 3–4 h at 4 °C and stored the ACM in single-use aliquots at −80° C. We generated primary cortical neuron cultures from E17 mice, added ACM (150 µg total protein/ml) at 0 div and then added AraC (5 µM) at 1 div to eliminate contaminating astrocytes. We harvested the neurons at 6 div, collected RNA, generated sequencing libraries, and performed sequencing and data analyses as described above.

Principal component analysis. The R package ggplot2 was used for principal component analysis with logRPKM as the input using all default settings.

Immunohistochemistry and immunocytochemistry. Mice were anesthetized with isoflurane and transcardially perfused with PBS followed by 4% paraformaldehyde (PFA). Brains were removed and further fixed in 4% PFA at 4°C overnight. The brains were washed with PBS and cryoprotected in 30% sucrose at 4°C for two days before being immersed in optimal cutting temperature compound (Fisher, cat#23-730-571) and stored at −80°C. Brains were sectioned on a cryostat (Leica) and 30 µm floating sections were blocked and permeablized in 10% donkey serum with 0.2% Triton X-100 in PBS and then stained with primary antibodies against human nucleus protein (Chemicon, cat#MAB1281, dilution 1:500) and human GFAP (Sternberger, cat#SMI21, dilution 1:500) at 4°C overnight. Sections were washed three times with PBS and incubated for 2 h at room temperature with secondary antibodies followed by three additional PBS washes. The sections were then mounted on Superfrost Plus microscope slides (Fisher, cat#12-550-15) and covered with mounting medium (Fisher, cat#H1400NB) and glass coverslips.

For immunocytochemistry of cultured cells, we fixed and permeablized astrocytes with 4% PFA and 0.2% Triton-X100 in PBS. After blocking in 10% donkey serum, we stained astrocytes with primary antibody against NFκB p65 (1:200; Cell Signaling Technology, Cat#8242) and fluorescent secondary antibodies (Invitrogen). After three washes in PBS, we stained the cells with DAPI (Thermo Scientific, Cat#62248) and imaged them using an Evos FL Auto 2 inverted fluorescence microscope (Invitrogen) with 10x and 20x lenses. We used Photoshop CS5 and FIJI to process images.

Statistical analysis and reproducibility. The numbers of patients, animals, and replicates are described in figures and figure legends. Experiments shown in the figures were repeated independently for the times listed below with similar results: Fig. 1a, 60 times. 1b, 10 times. 1c, twice. 2b-e, 12 times. 3b, 6 times with mouse samples and 3 times with human samples. Supplementary Fig. 10b, 7 times with mouse samples and 4 times with human samples. 25a, twice with mouse samples and 3 times with human samples. RNA-seq data were analyzed as described in the RNA-seq section above. For all non-RNA-seq data and RNA-seq data comparisons between species, analyses were conducted using RStudio (Version 1.3.1093) and Prism 8 software (Graphpad). Normality of data was tested by the Shapiro-Wilke test. For data with a normal distribution, Welch's two-sided t test was used for two-group comparisons and a one-way ANOVA was used for multi-group comparisons. For data that deviate from the normal distribution, the Mann-Whitney test was used. Data from technical replicates from the same patient or the same litter of mice were averaged and used as a single biological replicate in statistical analyses. An estimate of the variation in each group is indicated by the standard error of the mean (SEM) or standard deviation (SD). $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data supporting the findings of this study are provided within the paper and its supplementary information. A source data file is provided with this paper. We deposited all RNA-seq data to the Gene Expression Omnibus repository under accession number GSE147870. All additional information will be made available upon reasonable request to the authors. Source data are provided with this paper.

Code availability

All codes used in this study have been previously published[85,86] and are available (STAR, HTseq, DESeq2, Limma, EdgeR). No custum code was generated in this study.

Received: 22 June 2020; Accepted: 27 May 2021;
Published online: 25 June 2021

References

1. Sasaguri, H. et al. APP mouse models for Alzheimer's disease preclinical studies. *EMBO J.* **36**, 2473–2487 (2017).
2. Bezard, E., Yue, Z., Kirik, D. & Spillantini, M. G. Animal models of Parkinson's disease: Limits and relevance to neuroprotection studies. *Mov. Disord.* **28**, 61–70 (2013).
3. Masliah, E. et al. Dopaminergic loss and inclusion body formation in alpha-synuclein mice: implications for neurodegenerative disorders. *Science* **287**, 1265–1269 (2000).
4. Arnold, E. S. et al. ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43. *Proc. Natl Acad. Sci. U. S. A.* **110**, E736–E745 (2013).

5. Manwani, B. et al. Functional recovery in aging mice after experimental stroke. *Brain. Behav. Immun.* **25**, 1689–1700 (2011).
6. Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. & Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnol.* **32**, 40–51 (2014).
7. Pfrieger, F. W. & Barres, B. A. Synaptic efficacy enhanced by glial cells in vitro. *Science* **277**, 1684–1687 (1997).
8. Ullian, E. M., Sapperstein, S. K., Christopherson, K. S. & Barres, B. A. Control of synapse number by Glia. *Sci. (80-)* **291**, 657–661 (2001).
9. Blanco-Suarez, E., Liu, T.-F., Kopelevich, A. & Allen, N. J. Astrocyte-secreted chordin-like 1 drives synapse maturation and limits plasticity by increasing synaptic GluA2 AMPA receptors. *Neuron* **100**, 1116–1132.e13 (2018).
10. Ma, Z., Stork, T., Bergles, D. E. & Freeman, M. R. Neuromodulators signal through astrocytes to alter neural circuit activity and behaviour. *Nature* **539**, 428–432 (2016).
11. Huang, Y. H., Sinha, S. R., Tanaka, K., Rothstein, J. D. & Bergles, D. E. Astrocyte glutamate transporters regulate metabotropic glutamate receptor-mediated excitation of hippocampal interneurons. *J. Neurosci.* **24**, 4551–4559 (2004).
12. Parpura, V. et al. Glutamate-mediated astrocyte–neuron signalling. *Nature* **369**, 744–747 (1994).
13. Pascual, O. et al. Astrocytic purinergic signaling coordinates synaptic networks. *Science* **310**, 113–116 (2005).
14. Nedergaard, M. Direct signaling from astrocytes to neurons in cultures of mammalian brain cells. *Science* **263**, 1768–1771 (1994).
15. Kelley, K. W. et al. Kir4.1-dependent astrocyte-fast motor neuron interactions are required for peak strength. *Neuron* **98**, 306–319.e7 (2018).
16. Chung, W.-S. et al. Astrocytes mediate synapse elimination through MEGF10 and MERTK pathways. *Nature* **504**, 394–400 (2013).
17. Stogsdill, J. A. et al. Astrocytic neuroligins control astrocyte morphogenesis and synaptogenesis. *Nature* **551**, 192–197 (2017).
18. Anderson, M. A. et al. Astrocyte scar formation aids central nervous system axon regeneration. *Nature* **532**, 195–200 (2016).
19. Yu, X. et al. Reducing astrocyte calcium signaling in vivo alters striatal microcircuits and causes repetitive behavior. *Neuron* **99**, 1170–1187.e9 (2018).
20. Molofsky, A. V. et al. Astrocyte-encoded positional cues maintain sensorimotor circuit integrity. *Nature* **509**, 189–194 (2014).
21. Eroglu, Ç. et al. Gabapentin receptor α2δ-1 is a neuronal thrombospondin receptor responsible for excitatory CNS synaptogenesis. *Cell* **139**, 380–392 (2009).
22. Allen, N. J. et al. Astrocyte glypicans 4 and 6 promote formation of excitatory synapses via GluA1 AMPA receptors. *Nature* **486**, 410–414 (2012).
23. Farhy-Tselnicker, I. et al. Astrocyte-secreted glypican 4 regulates release of neuronal pentraxin 1 from axons to induce functional synapse formation. *Neuron* **96**, 428–445.e13 (2017).
24. Oberheim, N. A. et al. Uniquely hominid features of adult human astrocytes. *J. Neurosci.* **29**, 3276–3287 (2009).
25. Oberheim, N. A., Wang, X., Goldman, S. & Nedergaard, M. Astrocytic complexity distinguishes the human brain. *Trends Neurosci.* **29**, 547–553 (2006).
26. Zhang, Y. et al. Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. *Neuron* **89**, 37–53 (2016).
27. Zamanian, J. L. et al. Genomic analysis of reactive astrogliosis. *J. Neurosci.* **32**, 6391–6410 (2012).
28. Zhang, Y. et al. An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J. Neurosci.* **34**, 11929–11947 (2014).
29. Hodge, R. D. et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
30. Han, X. et al. Forebrain engraftment by human glial progenitor cells enhances synaptic plasticity and learning in adult mice. *Cell Stem Cell* **12**, 342–353 (2013).
31. Windrem, M. S. et al. A competitive advantage by neonatally engrafted human glial progenitors yields mice whose brains are chimeric for human glia. *J. Neurosci.* **34**, 16153–16161 (2014).
32. Windrem, M. S. et al. Neonatal chimerization with human glial progenitor cells can both remyelinate and rescue the otherwise lethally hypomyelinated shiverer mouse. *Cell Stem Cell* **2**, 553–565 (2008).
33. Krencik, R., Weick, J. P., Liu, Y., Zhang, Z.-J. & Zhang, S.-C. Specification of transplantable astroglial subtypes from human pluripotent stem cells. *Nat. Biotechnol.* **29**, 528–534 (2011).
34. Tchieu, J. et al. NFIA is a gliogenic switch enabling rapid derivation of functional human astrocytes from pluripotent stem cells. *Nat. Biotechnol.* **37**, 267–275 (2019).
35. Sloan, S. A. et al. Human astrocyte maturation captured in 3D cerebral cortical spheroids derived from pluripotent stem cells. *Neuron* **95**, 779–790.e6 (2017).
36. Chai, H. et al. Neural circuit-specialized astrocytes: transcriptomic, proteomic, morphological, and functional evidence. *Neuron* **95**, 531–549 (2017).
37. Crowley, L. C., Christensen, M. E. & Waterhouse, N. J. Measuring mitochondrial transmembrane potential by TMRE staining. *Cold Spring Harb. Protoc.* https://doi.org/10.1101/pdb.prot087361 (2016).
38. Dringen, R., Pawlowski, P. G. & Hirrlinger, J. Peroxide detoxification by brain cells. *J. Neurosci. Res.* **79**, 157–165 (2005).
39. Ma, X. et al. Mitochondrial electron transport chain complex III is required for antimycin A to inhibit autophagy. *Chem. Biol.* **18**, 1474–1481 (2011).
40. Nordgren, M. & Fransen, M. Peroxisomal metabolism and oxidative stress. *Biochimie* **98**, 56–62 (2014).
41. Foo, L. C. et al. Development of a method for the purification and culture of rodent astrocytes. *Neuron* **71**, 799–811 (2011).
42. Petriv, O. I. & Rachubinski, R. A. Lack of peroxisomal catalase causes a progeric phenotype in caenorhabditis elegans. *J. Biol. Chem.* **279**, P19996–20001 (2004).
43. Xu, Y. et al. Glucose-6-phosphate dehydrogenase-deficient mice have increased renal oxidative stress and increased albuminuria. *FASEB J.* **24**, 609–616 (2010).
44. Ho, H. Y., Cheng, M. L. & Chiu, D. T. Y. Glucose-6-phosphate dehydrogenase - From oxidative stress to cellular functions and degenerative diseases. *Redox Report* **12**, 109–118 (2007).
45. Bakken, T. E. et al. Single-cell RNA-seq uncovers shared and distinct axes of variation in dorsal LGN neurons in mice, non-human primates and humans. bioRxiv 2020.11.05.367482 (2020). https://doi.org/10.1101/2020.11.05.367482
46. Minnerup, J., Sutherland, B. A., Buchan, A. M. & Kleinschnitz, C. Neuroprotection for stroke: current status and future perspectives. *Int. J. Mol. Sci.* **13**, 11753–11772 (2012).
47. Michalicová, A., Bhide, K., Bhide, M. & Kováč, A. How viruses infiltrate the central nervous system *Acta Virol.* **61**, 393–400 (2017).
48. Pellegrini, L. et al. SARS-CoV-2 infects the brain choroid plexus and disrupts the blood-CSF barrier in human brain organoids. *Cell Stem Cell* https://doi.org/10.1016/j.stem.2020.10.001 (2020).
49. Perriot, S. et al. Human induced pluripotent stem cell-derived astrocytes are differentially activated by multiple sclerosis-associated cytokines. *Stem Cell Reports.* **11**, 1199–1210 (2018).
50. Sharma, D., Kim, M. S. & D'Mello, S. R. Transcriptome profiling of expression changes during neuronal death by RNA-Seq. *Exp. Biol. Med.* **240**, 242–251 (2015).
51. Mathys, H. et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **571**, 332–337 (2019).
52. Zhou, Y. et al. Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer's disease. *Nat. Med.* https://doi.org/10.1038/s41591-019-0695-9 (2020).
53. Schirmer, L. et al. Neuronal vulnerability and multilineage diversity in multiple sclerosis. *Nature* **573**, 75–82 (2019).
54. Wheeler, M. A. et al. MAFG-driven astrocytes promote CNS inflammation. *Nature* **578**, 593–599 (2020).
55. Jäkel, S. et al. Altered human oligodendrocyte heterogeneity in multiple sclerosis. *Nature* **566**, 543–547 (2019).
56. Henrik Heiland, D. et al. Tumor-associated reactive astrocytes aid the evolution of immunosuppressive environment in glioblastoma. *Nat. Commun.* **10**, 2541 (2019).
57. Liddelow, S. A. et al. Neurotoxic reactive astrocytes are induced by activated microglia. *Nature* **541**, 481–487 (2017).
58. Guttenplan, K. A. et al. Neurotoxic reactive astrocytes drive neuronal death after retinal injury. *Cell Rep.* **31**, 107776 (2020).
59. Chang, C. C., Zhang, J., Lombardi, L., Neri, A. & Dalla-Favera, R. Mechanism of expression and role in transcriptional control of the proto-oncogene NFKB-2/LYT-10. *Oncogene* **9**, 923–933 (1994).
60. Habib, N. et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).
61. Habib, N. et al. Disease-associated astrocytes in Alzheimer's disease and aging. *Nat. Neurosci.* **23**, 701–706 (2020).
62. Ito, M. et al. RNA-sequencing analysis revealed a distinct motor cortex transcriptome in spontaneously recovered mice after stroke. *Stroke* **49**, 2191–2199 (2018).
63. Mekel-Bobrov, N. et al. Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens. *Science* **309**, 1720–1722 (2005).
64. Florio, M. et al. Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* **347**, 1465–1470 (2015).
65. Long, K. R. et al. Extracellular matrix components HAPLN1, lumican, and collagen I cause hyaluronic acid-dependent folding of the developing human neocortex. *Neuron* **99**, 702–719.e6 (2018).
66. Kalebic, N. et al. Neocortical expansion due to increased proliferation of basal progenitors is linked to changes in their morphology. *Cell Stem Cell* **24**, 535–550.e9 (2019).

67. Wang, X., Tsai, J.-W., LaMonica, B. & Kriegstein, A. R. A new subtype of progenitor cell in the mouse embryonic neocortex. *Nat. Neurosci.* **14**, 555–561 (2011).
68. Hansen, D. V., Lui, J. H., Parker, P. R. L. & Kriegstein, A. R. Neurogenic radial glia in the outer subventricular zone of human neocortex. *Nature* **464**, 554–561 (2010).
69. Namba, T. et al. Human-specific ARHGAP11B acts in mitochondria to expand neocortical progenitors by glutaminolysis. *Neuron* **105**, 867–881.e9 (2020).
70. Johnson, M. B. et al. Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* **62**, 494–509 (2009).
71. Kang, H. J. et al. Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–489 (2011).
72. Miller, J. A. et al. Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199–206 (2014).
73. Jacobs, R. A., Díaz, V., Meinild, A., Gassmann, M. & Lundby, C. The C57Bl/6 mouse serves as a suitable model of human skeletal muscle mitochondrial function. *Exp. Physiol.* **98**, 908–921 (2013).
74. Perlman, R. L. Mouse models of human disease: an evolutionary perspective. *Evol. Med. public Heal.* **2016**, 170–176 (2016).
75. Billingsley, K. J. et al. Mitochondria function associated genes contribute to Parkinson's Disease risk and later age at onset. npj Park. *Dis* **5**, 8 (2019).
76. Arneson, D. et al. Single cell molecular alterations reveal target cells and pathways of concussive brain injury. *Nat. Commun.* **9**, 3894 (2018).
77. Le Douce, J. et al. Impairment of glycolysis-derived l-serine production in astrocytes contributes to cognitive deficits in Alzheimer's disease. *Cell Metab.* **31**, 503–517.e8 (2020).
78. Choi, B. H. & Lapham, L. W. Radial glia in the human fetal cerebrum: A combined golgi, immunofluorescent and electron microscopic study. *Brain Res* **148**, 295–311 (1978).
79. Roessmann, U. & Gambetti, P. Astrocytes in the developing human brain. *Acta Neuropathol.* **70**, 308–313 (1986).
80. Elder, G. A. & Major, E. O. Early appearance of type II astrocytes in developing human fetal brain. *Dev. Brain Res.* **42**, 146–150 (1988).
81. Molofsky, A. V. & Deneen, B. Astrocyte development: a guide for the perplexed. *Glia* **63**, 1320–1329 (2015).
82. Bushong, E. A., Martone, M. E. & Ellisman, M. H. Maturation of astrocyte morphology and the establishment of astrocyte domains during postnatal hippocampal development. *Int. J. Dev. Neurosci.* **22**, 73–86 (2004).
83. Zhong, S. et al. Decoding the development of the human hippocampus. *Nature* **577**, 531–536 (2020).
84. Neuner, S. M., Heuer, S. E., Huentelman, M. J., O'Connell, K. M. S. & Kaczorowski, C. C. Harnessing genetic complexity to enhance translatability of Alzheimer's disease mouse models: a path toward precision medicine. *Neuron* **101**, 399–411 (2019).
85. Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res* **47**, e47 (2019).
86. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

Acknowledgements

We thank Baljit Khakh, Mark Sharpley, Michael Sofroniew, Jill Haney, Ajit Divakaruni for advice. We thank the UCLA Mitochondria Core, the Center for AIDS Research Core, the Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, UCLA BioSequencing Core Facility for their services, Kory Hamane, Mahnaz Akhavan and Suhua Feng for their technical support. This work is supported by the Achievement Rewards for College Scientists foundation Los Angeles Founder Chapter and the National Institute of Mental Health of the National Institutes of Health (NIH) Award T32MH073526 to M.I.G., the Dr. Sheldon and Miriam G. Adelson Medical Research Foundation to S.A.G., H.I.K., and D.H.G., the National Institute of Neurological Disorders and Stroke of the NIH R00NS089780, R01NS109025, the National Institute of Aging of the NIH R03AG065772, National Center for Advancing Translational Science UCLA CTSI Grant UL1TR001881, UCLA Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research Innovation Award, the W.M. Keck Foundation Junior Faculty Award, the UCLA Jonsson Comprehensive Cancer Center and Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research Ablon Scholars Program, and the Friends of the Semel Institute for Neuroscience & Human Behavior Friends Scholar Award to Y.Z.

Author contributions

J.L. and Y.Z. conceived of the project and designed the experiments. J.L. performed all experiments except those noted below. L.P. performed xenograft experiments and RNA-seq of xenografted astrocytes. M.I.G. contributed to the generation of RNA-seq libraries. M.C.C., A.G.A., and M.H. optimized xenografting conditions and assisted the xenograft experiments under the supervision of H.I.K. W.G.P., J.E.R., and D.H.G. performed WGCNA and analyzed some of the single-cell sequencing datasets. Y-W.C. and X.Y. performed mapping of xenografted RNA-seq reads to human and mouse genomes. L.S. performed Seahorse Respirometry and TMRE/MTG imaging experiments. A.Y.C. and I. B.W. procured tissue samples. S.A.G. developed the xenograft method and provided training for xenograft experiments. J.L. and Y.Z. analyzed the data and wrote the paper. All authors read the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-24232-3.

Correspondence and requests for materials should be addressed to Y.Z.

Peer review information *Nature Communications* thanks Andras Lakatos and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix C

Noble et al. *Translational Psychiatry* (2021)11:194
https://doi.org/10.1038/s41398-021-01309-7

Translational Psychiatry

ARTICLE Open Access

Gut microbial taxa elevated by dietary sugar disrupt memory function

Emily E. Noble[1], Christine A. Olson[2], Elizabeth Davis[3], Linda Tsan[3], Yen-Wei Chen[2], Ruth Schade[1], Clarissa Liu[3], Andrea Suarez[3], Roshonda B. Jones[3], Claire de La Serre[1], Xia Yang[2], Elaine Y. Hsiao[2] and Scott E. Kanoski[3]

Abstract

Emerging evidence highlights a critical relationship between gut microbiota and neurocognitive development. Excessive consumption of sugar and other unhealthy dietary factors during early life developmental periods yields changes in the gut microbiome as well as neurocognitive impairments. However, it is unclear whether these two outcomes are functionally connected. Here we explore whether excessive early life consumption of added sugars negatively impacts memory function via the gut microbiome. Rats were given free access to a sugar-sweetened beverage (SSB) during the adolescent stage of development. Memory function and anxiety-like behavior were assessed during adulthood and gut bacterial and brain transcriptome analyses were conducted. Taxa-specific microbial enrichment experiments examined the functional relationship between sugar-induced microbiome changes and neurocognitive and brain transcriptome outcomes. Chronic early life sugar consumption impaired adult hippocampal-dependent memory function without affecting body weight or anxiety-like behavior. Adolescent SSB consumption during adolescence also altered the gut microbiome, including elevated abundance of two species in the genus *Parabacteroides* (*P. distasonis* and *P. johnsonii*) that were negatively correlated with hippocampal function. Transferred enrichment of these specific bacterial taxa in adolescent rats impaired hippocampal-dependent memory during adulthood. Hippocampus transcriptome analyses revealed that early life sugar consumption altered gene expression in intracellular kinase and synaptic neurotransmitter signaling pathways, whereas *Parabacteroides* microbial enrichment altered gene expression in pathways associated with metabolic function, neurodegenerative disease, and dopaminergic signaling. Collectively these results identify a role for microbiota "dysbiosis" in mediating the detrimental effects of early life unhealthy dietary factors on hippocampal-dependent memory function.

Introduction

The gut microbiome has recently been implicated in modulating neurocognitive development and consequent functioning[1–4]. Early life developmental periods represent critical windows for the impact of indigenous gut microbes on the brain, as evidenced by the reversal of behavioral and neurochemical abnormalities in germ free rodents when inoculated with conventional microbiota during early life, but not during adulthood[5–7]. Dietary factors are a critical determinant of gut microbiota diversity and can alter gut bacterial communities, as evident from the microbial plasticity observed in response to pre- and probiotic treatment, as well as the "dysbiosis" resulting from consuming unhealthy, yet palatable foods that are associated with obesity and metabolic disorders (e.g., Western diet; foods high in saturated fatty acids and added sugar)[8]. In addition to altering the gut microbiota, consumption of Western dietary factors yields long-lasting memory impairments, and these effects are more pronounced when consumed during early life developmental periods vs. during

Correspondence: Elaine Y. Hsiao (ehsiao@ucla.edu) or
Scott E. Kanoski (kanoski@usc.edu)
[1]University of Georgia, Athens, GA, USA
[2]University of California, Los Angeles, CA, USA
Full list of author information is available at the end of the article

adulthood[9–11]. Whether diet-induced changes in specific bacterial populations are functionally related to altered early life neurocognitive outcomes, however, is poorly understood.

The hippocampus, which is well known for its role in spatial and episodic memory and more recently for regulating learned and social aspects of food intake control[12–17], is particularly vulnerable to the deleterious effects of Western dietary factors[9,18,19]. During the juvenile and adolescent stages of development, a time when the brain is rapidly developing, consumption of diets high in saturated fat and sugar[20–22] or sugar alone[23–26] impairs hippocampal function while in some cases preserving memory processes that do not rely on the hippocampus. While several putative underlying mechanisms have been investigated, the precise biological pathways linking dietary factors to neurocognitive dysfunction remain largely undetermined[11]. Here we aimed to determine whether sugar-induced alterations in gut microbiota during early life are causally related to hippocampal-dependent memory impairments observed during adulthood.

Methods and materials

Experimental subjects

Juvenile male Sprague Dawley rats (Envigo; arrival postnatal day (PN) 26–28; 50–70 g) were housed individually in standard conditions with a 12:12 light/dark cycle. All rats had ad libitum access to water and Lab Diet 5001 (PMI Nutrition International, Brentwood, MO; 29.8 % kcal from protein, 13.4% kcal from fat, 56.7% kcal from carbohydrate), with modifications where noted. Treatment group sizes for Aim 1 experiments are derived from power analyses conducted in Statistica Software (V7) based on our published data, pilot data, and relevant publications in the literature. All experiments were performed in accordance with the approval of the Animal Care and Use Committee at the University of Southern California.

Experiment 1

Twenty-one juvenile male rats (PN 26–28) were divided into two groups with equal bodyweight and given ad libitum access to (1) 11% weight-by-volume (w/v) solution containing monosaccharide ratio of 65% fructose and 35% glucose in reverse osmosis-filtered water (SUG; $n = 11$) or 2) or an extra bottle of reverse osmosis-filtered water (CTL; $n = 10$). This solution was chosen to model commonly consumed sugar-sweetened beverages (SSBs) in humans in terms of both caloric content and monosaccharide ratio[27]. In addition, all rats were given ad libitum access to water and standard rat chow. Food intake, solution intake, and body weights were monitored thrice-weekly except were prohibited due to behavioral testing. At PN 60, rats underwent Novel Object in Context (NOIC) testing, to measure hippocampal-dependent episodic contextual memory. At PN 67 rats underwent anxiety-like behavior testing in the Zero Maze, followed by body composition testing at PN 70 and an intraperitoneal glucose tolerance test (IP GTT) at PN 84. All behavioral procedures were run at the same time each day (4–6 h into the light cycle). Investigators were blind to animal groups when scoring the behavioral tasks such that the scorers did not know which animal was in which group. Fecal and cecal samples were collected prior to sacrifice at PN 104.

In a separate cohort of juvenile male rats ($n = 6$/group) animals were treated as above, but on PN day 60 rats were tested in the Novel Object Recognition (NOR) and Open Field (OF) tasks, with two days in between tasks. Animals were sacrificed and tissue punches were collected from the dorsal hippocampus on PN day 65. Tissue punches were flash-frozen in a beaker filled with isopentane and surrounded dry ice and then stored at −80 °C until further analyses.

Experiment 2

Twenty-three juvenile male rats (PN 26–28) were divided into two groups of equal bodyweight and received a gavage twice daily (12 h apart) for 7 days (only one treatment was given on day 7) of either (1) saline (SAL; $n = 8$), or (2) a cocktail of antibiotics consisting of Vancomycin (50 mg/kg), Neomycin (100 mg/kg), and Metronidazole (100 mg/kg) along with supplementation with 1 mg/mL of ampicillin in their drinking water (ABX; $n = 15$), which is a protocol modified from[28]. Animals were housed in fresh, sterile cages on Day 3 of the antibiotic or saline treatment, and again switched to fresh sterile cages on Day 7 after the final gavage. All animals were maintained on sterile, autoclaved water and chow for the remainder of the experiment. Rats in the ABX group were given water instead of ampicillin solution on Day 7. Animals in the ABX group were further subdivided to receive either gavage of a 1:1 ratio of *Parabacteroides distasonis* and *Parabacteroides johnsonii* (PARA; $n = 8$) or saline (SAL; $n = 7$) thirty-six hours after the last ABX treatment. To minimize potential contamination, rats were handled minimally for 14 days. Cage changes occurred once weekly at which time animals and food were weighed. Experimenters wore fresh, sterile PPE, and weigh boxes were cleaned with sterilizing solution in between each cage change. On PN 50 rats were tested in NOIC, on PN 60 rats were tested in NOR, on PN 62 rats were tested in the Zero Maze, followed by OF on PN 64. Investigators were blind to animal groups when scoring the behavioral tasks such that the scorers did not know which animal was in which group when timing the behavior (NOIC, NOR, Zero Maze, OF). On PN 73 rats were given an IP GTT, and on PN 76 body composition was tested. Rats were sacrificed at PN 83 and dorsal

hippocampus tissue punches and cecal samples were collected. Tissue punches were flash-frozen in a beaker filled with isopentane and surrounded by dry ice and cecal samples were placed in microcentrifuge tubes embedded in dry ice. Samples were subsequently stored at −80 °C until further analyses.

IP glucose tolerance test (IP GTT)

Animals were food-restricted 24 h prior to IP GTT. Immediately prior to the test, baseline blood glucose readings were obtained from the tail tip and recorded by a blood glucose meter (One-touch Ultra2, LifeScan Inc., Milpitas, CA). Each animal was then intraperitoneally (IP) injected with dextrose solution (0.923 g/ml by body weight) and tail tip blood glucose readings were obtained at 30, 60, 90, and 120 min after IP injections, as previously described[23].

Zero Maze

The Zero Maze is an elevated circular track (63.5 cm fall height, 116.8 cm outside diameter), divided into four equal-length sections. Two sections were open with 3 cm high curbs, whereas the 2 other closed sections contained 17.5 cm high walls. Animals are placed in the maze facing the open section of the track in a room with ambient lighting for 5 min while the experimenter watches the animal from a monitor outside of the room. The experimenter records the total time spent in the open sections (defined as the head and front two paws in open arms), and the number of crosses into the open sections from the closed sections.

The novel object in context task

NOIC measures episodic contextual memory based on the capacity for an animal to identify which of two familiar objects it has never seen before in a specific context. Procedures were adapted from prior reports[29,30]. Briefly, rats are habituated to two distinct contexts on subsequent days (with the habituation order counterbalanced by the group) for 5-min sessions: Context 1 is a semi-transparent box (15 in. W × 24 in. L × 12 in. H) with orange stripes and Context 2 is a grey opaque box (17 in. W × 17 in. L × 16 in. H) (Context identify assignments counterbalanced by the group), each context is in a separate dimly lit room, which is obtained using two desk lamps pointed toward the floor. Day 1 of NOIC begins with each animal being placed in Context 1 containing two distinct similarly sized objects placed in opposite corners: a 500 ml jar filled with blue water (Object A) and a square glass container (Object B) (Object assignments and placement counterbalanced by the group). On day 2 of NOIC, animals are placed in Context 2 with duplicates of one of the objects. On NOIC day 3, rats are placed in Context 2 with Objects A and Object B. One of these objects is not novel to the rat, but its placement in Context 2 is novel. All sessions are 5 min long and are video recorded. Each time the rat is placed in one of the contexts, it is placed with its head facing away from both objects. The time spent investigating each object is recorded from the video recordings by an experimenter who is blinded to the treatment groups. Exploration is defined as sniffing or touching the object with the nose or forepaws. The task is scored by calculating the time spent exploring the Novel Object to the context divided by the time spent exploring both Objects A and B combined, which is the novelty or "discrimination index". Rats with an intact hippocampus will preferentially investigate the object that is novel to Context 2, given that this object is a familiar object yet is now presented in a novel context, whereas hippocampal inactivation impairs the preferential investigation of the object novel to Context 2[29].

Novel object recognition

The apparatus used for NOR is a grey opaque box (17 in. W × 17 in. L × 16 in. H) placed in a dimly lit room, which is obtained using two desk lamps pointed toward the floor. Procedures are adapted from ref. [31]. Rats are habituated to the empty arena and conditions for 10 min on the day prior to testing. The novel object and the side on which the novel object is placed are counterbalanced by the group. The test begins with a 5-min familiarization phase, where rats are placed in the center of the arena, facing away from the objects, with two identical copies of the same object to explore. The objects were either two identical cans or two identical bottles, counterbalanced by the treatment group. The objects were chosen based on preliminary studies which determined that they are equally preferred by Sprague Dawley rats. Animals are then removed from the arena and placed in the home cage for 5 min. The arena and objects are cleaned with 10% ethanol solution, and one of the objects in the arena is replaced with a different one (either the can or bottle, whichever the animal has not previously seen, i.e., the "novel object"). Animals are again placed in the center of the arena and allowed to explore for 3 min. Time spent exploring the objects is recorded via video recording and analyzed using Any-maze activity tracking software (Stoelting Co., Wood Dale, IL).

Open Field

OF measures general activity level and also anxiety-like behavior in the rat. A large gray bin, 60 cm (L) × 56 cm (W) is placed under diffuse even lighting (30 lux). A center zone is identified and marked in the bin (19 cm L × 17.5 cm W). A video camera is placed directly overhead and animals are tracked using AnyMaze Software (Stoelting Co., Wood Dale, IL). Animals are placed in the center of the box facing the back wall and allowed to

explore the arena for 10 min while the experimenter watches from a monitor in an adjacent room. The apparatus is cleaned with 10% ethanol after each rat is tested.

Body composition

Body composition (body fat, lean mass) was measured using LF90 time-domain nuclear magnetic resonance (Bruker NMR minispec LF 90II, Bruker Daltonics, Inc.).

Bacterial transfer

P. distasonis (ATCC 8503) was cultured under anaerobic conditions at 37 °C in Reinforced Clostridial Medium (RCM, BD Biosciences). *P. johnsonii* (DSM 18315) was grown in anaerobic conditions in PYG medium (modified, DSM medium 104). Cultures were authenticated by full-length 16S rRNA gene sequencing. For bacterial enrichment, 10^9 colony-forming units of both *P. distasonis* and *P. johnsonii* were suspended in 500 µL pre-reduced PBS and orally gavaged into antibiotic-treated rats. When co-administered, a ratio of 1:1 was used for *P. distasonis* and *P. johnsonii*.

Gut microbiota DNA extraction and 16s rRNA gene sequencing in sugar-fed and control rats

All samples were extracted and sequenced according to the guidelines and procedures established by the Earth Microbiome Project[32]. DNA was extracted from fecal and cecal samples using the MO BIO PowerSoil DNA extraction kit. Polymerase chain reaction (PCR) targeting the V4 region of the 16S rRNA bacterial gene was performed with the 515F/806R primers, utilizing the protocol described in Caporaso et al.[33]. Amplicons were barcoded and pooled in equal concentrations for sequencing. The amplicon pool was purified with the MO BIO UltraClean PCR Clean-up kit and sequenced by the 2 × 150 bp MiSeq platform at the Institute for Genomic Medicine at UCSD. All sequences were deposited in Qiita Study 11255 as raw FASTQ files. Sequences were demultiplexed using Qiime-1 based "split libraries" with the forward reads only dropping. Demultiplexed sequences were then trimmed evenly to 100 bp and 150 bp to enable comparison to other studies for meta-analyses. Trimmed sequences were matched to known OTUs at 97% identity.

Gut microbiota DNA extraction and 16S rRNA gene sequencing for *Parabacteroides*-enriched and control rats

Total bacterial genomic DNA was extracted from rat fecal samples (0.25 g) using the Qiagen DNeasy PowerSoil Kit. The library was prepared following methods from (Caporaso et al.[33]). The V4 region (515F–806R) of the 16S rDNA gene was PCR amplified using individually barcoded universal primers and 30 ng of the extracted genomic DNA. The conditions for PCR were as follows: 94 °C for 3 min to denature the DNA, with 35 cycles at 94 °C for 45 s, 50 °C for 60 s, and 72 °C for 90 s, with a final extension of 10 min at 72 °C. The PCR reaction was set up in triplicate, and the PCR products were purified using the Qiaquick PCR purification kit (QIAGEN). The purified PCR product was pooled in equal molar concentrations quantified by nanodrop and sequenced by Laragen, Inc. using the Illumina MiSeq platform and 2 × 250 bp reagent kit for paired-end sequencing. Amplicon sequence variants (ASVs) were chosen after denoising with the Deblur pipeline. Taxonomy assignment and rarefaction were performed using QIIME2-2019.10.

Hippocampal RNA extraction and sequencing

Hippocampi from rats treated with or without sugar or *Parabacteroides* were subject to RNA-seq analysis. Total RNA was extracted according to the manufacturer's instructions using RNeasy Lipid Tissue Mini Kit (Qiagen, Hilden, Germany). Total RNA was checked for degradation in a Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA). Quality was very high for all samples, and libraries were prepared from 1 µg of total RNA using a NuGen Universal Plus mRNA-seq Library Prep Kit (Tecan Genomics Inc., Redwood City, CA). Final library products were quantified using the Qubit 2.0 Fluorometer (Thermo Fisher Scientific Inc., Waltham, MA, USA), and the fragment size distribution was determined with the Bioanalyzer 2100. The libraries were then pooled equimolarly, and the final pool was quantified via qPCR using the Kapa Biosystems Library Quantification Kit, according to the manufacturer's instructions. The pool was sequenced in an Illumina NextSeq 550 platform (Illumina, San Diego, CA, USA), in Single-Read 75 cycles format, obtaining about 25 million reads per sample. The preparation of the libraries and the sequencing were performed at the USC Genome Core (http://uscgenomecore.usc.edu/).

RNA-seq quality control

Data quality checks were performed using the FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and low-quality reads were trimmed with Trim_Galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). RNA-seq reads passing quality control were mapped to *Rattus novegicus* transcriptome (Rnor6) and quantified with Salmon[34]. Salmon directly mapped RNA-seq reads to Rat transcriptome and quantified transcript counts. Txiimport[35] was used to convert transcript counts into gene counts. Potential sample outliers were detected by principal component analysis (PCA) and one control and one treatment sample from the *Parabacteroides* experiment were deemed outliers (Fig. S1) and removed.

Identification of differentially expressed genes (DEGs)

DESeq2[36] were used to conduct differential gene expression analysis between sugar treatment and the corresponding controls or between *Parabacteroides* treatment and the corresponding controls. Low-abundance genes were filtered out and only those having a mean raw count > 1 in more than 50% of the samples were included. Differentially expressed genes were detected by DESeq2 with default settings. Significant DEGs were defined as Benjamini–Hochberg (BH) adjusted false-discovery rate (FDR) < 0.05. For heatmap visualization, genes were normalized with variance stabilization transformation implemented in DESeq2, followed by calculating a *z*-score for each gene.

Pathway analyses of DEGs

For the pathway analyses, DEGs at an unadjusted *P* value < 0.01 were used. Pathway enrichment analyses were conducted using enrichr[37] by intersecting each signature with pathways or gene sets from KEGG[38], gene ontology biological pathways, cellular component, molecular function[39], and Wikipathways[40]. Pathways at FDR < 0.05 were considered significant. Unless otherwise specified, R 3.5.2 was used for the analysis mentioned in the RNA sequencing section.

Additional statistical methods

Data are presented as means ± SEM. For analytic comparisons of body weight, total food intake, and chow intake, groups were compared using repeated-measures ANOVA in Prism software (GraphPad Inc., version 8.0). Taxonomic comparisons from 16S rRNA sequencing analysis were analyzed by analysis of the composition of microbiomes (ANCOM). When significant differences were detected, Sidak post-hoc test for multiple comparisons was used. The area under the curve for the IP GTT testing was also calculated using Prism. All other statistical analyses were performed using Student's two-tailed unpaired *t* tests in excel software (Microsoft Inc., version 15.26). Normality was confirmed prior to the utilization of parametric testing. For all analyses, statistical significance was set at $P < 0.05$. A predetermined criterion for exclusion was utilized and was based on the Grubbs Outlier Test (Prism, Graphpad Inc.) using alpha = 0.05.

Results

Early life sugar consumption impairs hippocampal-dependent memory function

Results from the NOIC task, which measures hippocampal-dependent episodic contextual memory function[30], reveal that while there were no differences in total exploration time of the combined objects on days 1 or 3 of the task (Fig. 1A, B), animals fed sugar solutions in early life beginning at PN 28 had a reduced capacity to discriminate an object that was novel to a specific context when animals were tested during adulthood (PN 60), indicating impaired hippocampal function (Fig. 1C). Conversely, animals fed sugar solutions in early life performed similarly to those in the control group when tested in the novel object recognition task (NOR) (Fig. 1D), which tests object recognition memory independent of context. Notably, when performed using the current methods with a short duration between the familiarization phase and the test phase, NOR not hippocampal-dependent but instead is primarily dependent on the perirhinal cortex[30,41–43]. These data suggest that early life dietary sugar consumption impairs performance in hippocampal-dependent contextual-based recognition memory without affecting performance in perirhinal cortex-dependent recognition memory independent of context[23].

Elevated anxiety-like behavior and altered general activity levels may influence novelty exploration independent of memory effects and may therefore confound the interpretation of behavioral results. Thus, we next tested whether early life sugar consumption affects anxiety-like behavior using two different tasks designed to measure anxiety-like behavior in the rat: the elevated zero mazes and the OF task, the latter of which also assesses levels of general activity[44]. Early life sugar consumption had no effect on time spent in the open area or in the number of open area entries in the zero maze (Fig. 1E, F). Similarly, early life sugar had no effect on distance traveled or time spent in the center zone in the OF task (Fig. 1G, H). Together these data suggest that habitual early life sugar consumption did not increase anxiety-like behavior or general activity levels in the rats.

Early life sugar consumption impairs glucose tolerance without affecting total caloric intake, body weight, or adiposity

Given that excessive sugar consumption is associated with weight gain and metabolic deficits[45], we tested whether access to a sugar solution during the adolescent phase of development would affect food intake, body weight gain, adiposity, and glucose tolerance in the rat. Early life sugar consumption had no effect on body composition during adulthood (Fig. 1I, Fig. S2A, B). Early life sugar consumption also had no effect on body weight or total kcal intake (Fig. 1J, K), which is in agreement with the previous findings[23,26,46]. Animals steadily increased their intake of the 11% sugar solution throughout the study (Fig. 1L) but compensated for the calories consumed in the sugar solutions by reducing their intake of dietary chow (Fig. S2C). However, animals that were fed sugar solutions during adolescence showed impaired peripheral glucose metabolism in an IP GTT (Fig. S2D).

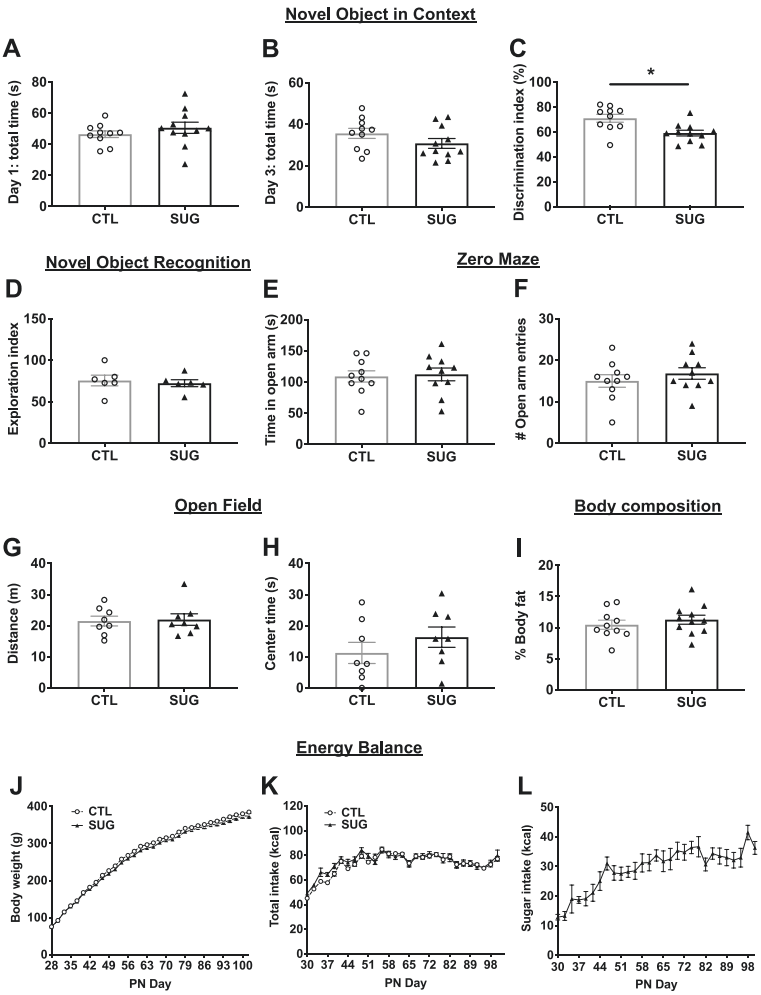


Fig. 1 Early life sugar consumption negatively impacts hippocampal-dependent memory function. A, **B** Early life sugar consumption had no effect on total exploration time on days 1 (familiarization) or day 3 (test day) of the Novel Object in Context (NOIC) task. **C** The discrimination index was significantly reduced by early life sugar consumption, indicating impaired hippocampal function ($P < 0.05$, $n = 10,11$; two-tailed, type 2 Student's *t* test). **D** There were no differences in exploration index in the Novel Object Recognition (NOR task) ($n = 6$; two-tailed, type 2 Student's *t* test). **E**, **F** There were no differences in time spent in the open arm or the number of entries into the open arm in the Zero Maze task for anxiety-like behavior ($n = 10$; two-tailed, type 2 Student's *t* test). **G**, **H** There were no differences in distance traveled or time spent in the center arena in the Open Field task ($n = 8$; two-tailed, type 2 Student's *t* test). **I** There was no differences in body fat % during adulthood between rats fed early life sugar and controls ($n = 10,11$; two-tailed, type 2 Student's *t* test). **J**, **K** Body weights and total energy intake did not differ between the groups ($n = 10,11$; two-way repeated-measures ANOVA), despite (**L**) increased kcal consumption from sugar-sweetened beverages in the sugar group. CTL = control, SUG = sugar, PN = post-natal day; data shown as mean ± SEM.

Gut microbiota is impacted by early life sugar consumption

Principal component analyses of 16 s rRNA gene sequencing data of fecal samples revealed a separation between the fecal microbiota of rats fed early life sugar and controls (Fig. 2A). Results from LEfSe analysis identified differentially abundant bacterial taxa in fecal samples that were elevated by sugar consumption. These include the family *Clostridiaceae* and the genus *02d06*

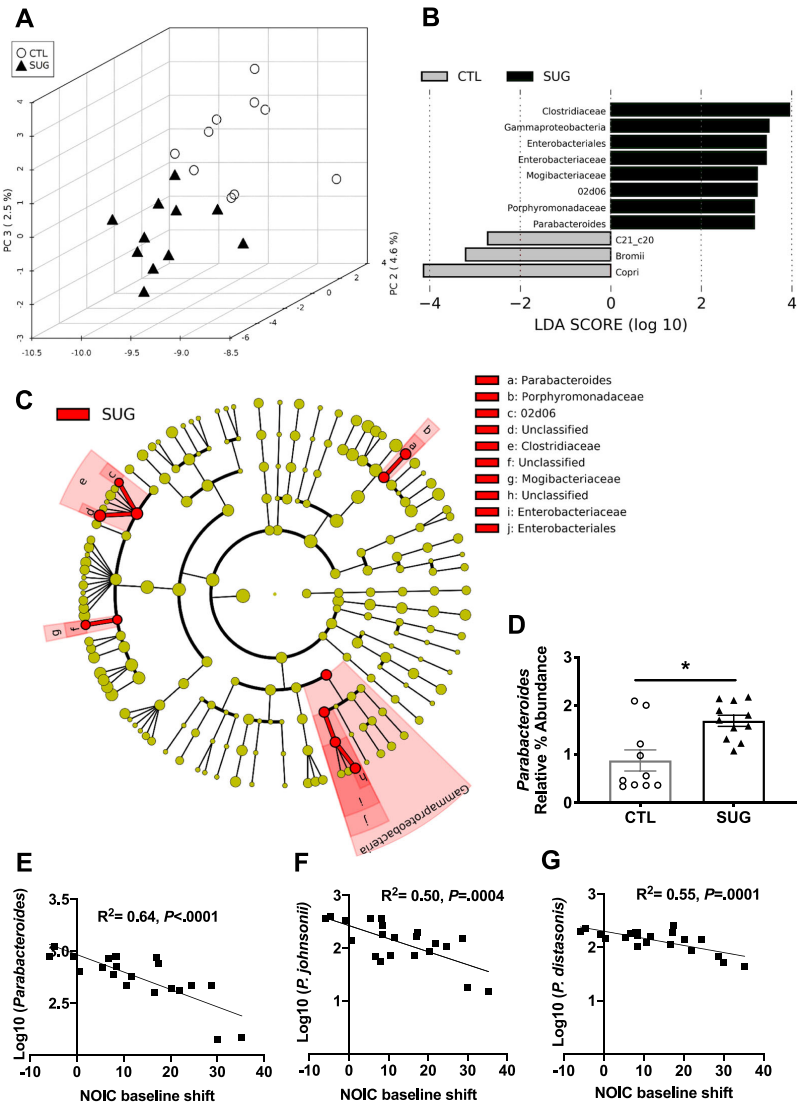


Fig. 2 Effect of adolescent sugar consumption on the gut microbiome in rats. A Principal component analysis showing separation between fecal microbiota of rats fed early life sugar or controls ($n = 11, 10$; dark triangles = sugar, open circles = control). **B** Results from LEfSe analysis showing Linear Discriminate Analysis (LDA) scores for microbiome analysis of fecal samples of rats fed early life sugar or controls. **C** A cladogram representing the results from the LEfSe analysis with the class as the outermost taxonomic level and species at the innermost level. Taxa in red are elevated in the sugar group. **D** Relative % abundance of fecal *Parabacteroides* are significantly elevated in rats fed early life sugar ($P < 0.05$; $n = 11, 10$, two-tailed, type 2 Student's t test). **E** Linear regression of log normalized fecal *Parabacteroides* counts against the shift from baseline performance scores in the novel object in context task (NOIC) across all groups tested ($n = 21$). **E**–**G** Linear regression of the most abundant fecal *Parabacteroides* species against shift from baseline performance scores in NOIC across all groups tested ($n = 21$). *$P < 0.05$; data are shown as mean ± SEM.

within *Clostridiaceae*, the family *Mogibacteriaceae*, the family *Enterobacteriaceae*, the order *Enterobacteriales*, the class of *Gammaproteobacteria*, and the genus *Parabacteroides* within the family *Porphyromonadaceae* (Fig. 2B, C). In addition to an elevated % relative abundance of the genus *Parabacteroides* in animals fed early life sugar (Fig. 2D), log-transformed counts of the *Parabacteroides* negatively correlated with performance scores in the NOIC memory task (Fig. 2E). Of the additional bacterial populations significantly affected by sugar treatment, regression analyses did not identify any other genera as being significantly correlated to NOIC memory performance. Within *Parabacteroides*, levels of two operational taxonomic units (OTUs) that were elevated by sugar negatively correlated with performance in the NOIC task, identified as taxonomically related to *P. johnsonii* and *P. distasonis* (Fig. 2F, G). The significant negative correlation between NOIC performance and *Parabacteroides* was also present within each of the diet groups alone, but when separated out by diet group only *P. distasonis* showed a significant negative correlation for each diet group ($P < 0.05$), whereas *P. johnsonii* showed a nonsignificant trend in both the control and sugar groups ($P = 0.06$, and $P = 0.08$, respectively; Fig. S3A–C). The abundance of other bacterial populations that were affected by sugar consumption was not significantly related to memory task performance.

There was a similar separation between groups in bacteria analyzed from cecal samples (Fig. S4A). LEfSe results from cecal samples show elevated *Bacilli*, *Actinobacteria*, *Erysipelotrichia*, and *Gammaproteobacteria* in rats fed early life sugar, and elevated *Clostridia* in the controls (Fig. S4B, C). Abundances at the different taxonomic levels in fecal and cecal samples are shown in (Figs. S5 and S6). Regression analyses did not identify these altered cecal bacterial populations as being significantly correlated to NOIC memory performance.

Early life Parabacteroides enrichment impairs memory function

To determine whether neurocognitive outcomes due to early life sugar consumption could be attributable to elevated levels of *Parabacteroides* in the gut, we experimentally enriched the gut microbiota of naïve juvenile rats with two *Parabacteroides* species that exhibited high 16S rRNA sequencing alignment with OTUs that were increased by sugar consumption and were negatively correlated with behavioral outcomes in rats fed early life sugar. *P. johnsonii* and *P. distasoni* species were cultured individually under anaerobic conditions and transferred to a group of antibiotic-treated young rats in a 1:1 ratio via oral gavage using the experimental design described in Methods and outlined in Fig. 3A, and from ref. [28]. To confirm *Parabacteroides* enrichment, 16SrRNA sequencing was performed on rat fecal samples for SAL–SAL, ABX-SAL, and ABX-PARA groups. Alpha diversity was analyzed using observed OTUs (Fig. 3B), where both ABX-SAL and ABX-PARA fecal samples have significantly reduced alpha diversity when compared with SAL–SAL fecal samples, suggesting that antibiotic treatment reduces microbiome alpha diversity. Further, either treatment with antibiotics alone or antibiotics followed by *Parabacteroides* significantly alters microbiota composition relative to the SAL–SAL group (Fig. 3C). Taxonomic comparisons from 16S rRNA sequencing analysis were analyzed by analysis of the composition of microbiomes (ANCOM). Differential abundance on relative abundance at the species level (Fig. 3D) was tested across samples hypothesis-free. Significant taxa at the species level were corrected for using FDR-corrected P values to calculate W in ANCOM. Comparing all groups resulted in the highest W value of 144 for the *Parabacteroides* genus, which was enriched in ABX-PARA fecal samples after bacterial gavage with an average relative abundance of 55.65% (Fig. 3E). This confirms successful *Parabacteroides* enrichment for ABX-PARA rats post-gavage when compared to either ABX-SAL (average relative abundance of 5.47%) or ABX-SAL rats (average relative abundance of 0.26%).

All rats treated with antibiotics showed a reduction in food intake and body weight during the initial stages of antibiotic treatment, however, there were no differences in body weight between the two groups of antibiotic-treated animals by PN50, at the time of behavioral testing (Fig. S7A–C). Similar to a recent report[47], Parabacteroides enrichment in the present study impacted body weight at later time points. Animals who received *P. johnsonii* and *P. distasonis* treatment showed reduced body weight 40 days after the transfer, with significantly lower lean mass (Fig. S7D–F). There were no differences in percent body fat between groups, nor were there significant group differences in glucose metabolism in the IPGTT (Fig. S7G). Importantly, the body weights in the ABX-PARA group did not significantly differ from the ABX-SAL control group at the time of behavioral testing.

Results from the hippocampal-dependent NOIC memory task showed that while there were no differences in total exploration time of the combined objects on days 1 or 3 of the task, indicating similar exploratory behavior, animals enriched with *Parabacteroides* showed a significantly reduced discrimination index in the NOIC task compared with either control group (Fig. 4A–C), indicating impaired performance in hippocampal-dependent memory function. When tested in the perirhinal cortex-dependent NOR task[30], animals enriched with *Parabacteroides* showed impaired object recognition memory compared with the antibiotic-treated control group as

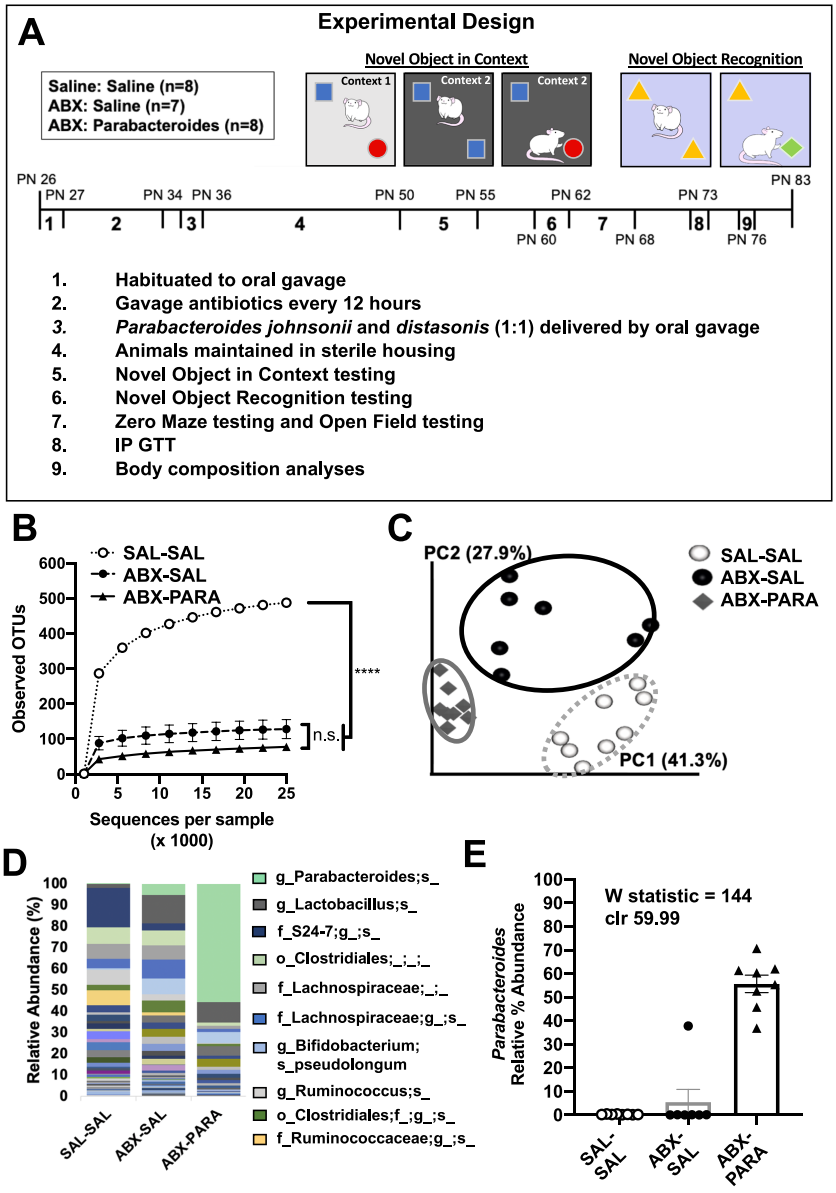


Fig. 3 (See legend on next page.)

(see figure on previous page)

Fig. 3 Intestinal *Parabacteroides* is enriched by antibiotic treatment and oral gavage of *P. distasonis* and *P. johnsonii*. **A** Schematic showing the timeline for the experimental design of the *Parabacteroides* transfer experiment. **B** Alpha diversity based on 16S rRNA gene profiling of fecal matter ($n = 7–8$) represented by observed operational taxonomic units (OTUs) for a given number of sample sequences. **C** Principal coordinates analysis of weighted UniFrac distance based on 16S rRNA gene profiling of feces for SAL–SAL, ABX-SAL, and ABX-PARA enriched rats ($n = 7–8$). **D** Average taxonomic distributions of bacteria from 16S rRNA gene sequencing data of feces for SAL–SAL, ABX-SAL, and ABX-PARA enriched animals ($n = 7–8$). **E** Relative abundances of *Parabacteroides* in fecal microbiota for SAL–SAL, ABX-SAL, and ABX-PARA enriched animals ($n = 7–8$) (ANCOM). PN post-natal day, IP GTT intraperitoneal glucose tolerance test. Data are presented as mean ± S.E.M. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. n.s. not statistically significant, SAL–SAL rats treated with saline, ABX-SAL rats treated with antibiotics followed by sterile saline gavage, ABX-PARA rats treated with antibiotics followed by a 1:1 gavage of *Parabacteroides distasonis* and *Parabacteroides johnsonii*.

indicated by a reduced novel object exploration index (Fig. 4D). These findings show that, unlike sugar-fed animals, *Parabacteroides* enrichment impaired perirhinal cortex-dependent memory processes in addition to hippocampal-dependent memory.

Results from the zero maze showed no differences in time spent in the open arms nor in the number of open arm entries for the *Parabacteroides*-enriched rats relative to controls (Fig. 4E, F), indicating that the enrichment did not affect anxiety-like behavior. Similarly, there were no differences in distance traveled or time spent in the center arena in the OF test, which is a measure of both anxiety-like behavior and general activity in rodents (Fig. 4G, H). Together these data suggest that *Parabacteroides* treatment negatively impacted both hippocampal-dependent perirhinal cortex-dependent memory function without significantly affecting general activity or anxiety-like behavior.

Early life sugar consumption and Parabacteroides enrichment alter hippocampal gene expression profiles

To further investigate how sugar and *Parabacteroides* affect cognitive behaviors, we conducted transcriptome analysis of the hippocampus samples. Figure S1A, C shows the results of principal component analysis revealing moderate separation based on RNA sequencing data from the dorsal hippocampus of rats fed sugar in early life compared with controls. Gene pathway enrichment analyses from RNA sequencing data revealed multiple pathways significantly affected by early life sugar consumption, including four pathways involved in neurotransmitter synaptic signaling: dopaminergic, glutamatergic, cholinergic, and serotonergic signaling pathways. In addition, several gene pathways that also varied by sugar were those involved in kinase-mediated intracellular signaling: cGMP-PKG, RAS, cAMP, and MAPK signaling pathways (Fig. 5A, Table S1).

Analyses of individual genes across the entire transcriptome using a stringent FDR criterion further identified 21 genes that were differentially expressed in rats fed early life sugar compared with controls, with 11 genes elevated and 10 genes decreased in rats fed sugar compared to controls (Fig. 5B). Among the genes impacted, several genes that regulate cell survival, migration, differentiation, and DNA repair were elevated by early life sugar access, including *Faap100*, which encodes an FA core complex member of the DNA damage response pathway[48], and *Eepd1*, which transcribes an endonuclease involved in repairing stalled DNA replication forks, stressed from DNA damage[49]. Other genes associated with endoplasmic reticulum stress and synaptogenesis were also significantly increased by sugar consumption, including *Klf9*, *Dgkh*, *Neurod2*, *Ppl*, and *Kirrel1*[50–53].

Several genes were reduced by dietary sugar, including *Tns2*, which encodes tensin 2, important for cell migration[54], *RelA*, which encodes an NF/kB complex protein that regulates the activity-dependent neuronal function and synaptic plasticity[55], and *Grm8*, the gene for the metabotropic glutamate receptor 8 (mGluR8). Notably, reduced expression of the mGluR8 receptor may contribute to the impaired neurocognitive functioning in animals fed sugar, as mGluR8 knockout mice show impaired hippocampal-dependent learning and memory[56].

Figure S1A, B, D shows the results of the principal component analysis of dorsal hippocampus RNA sequencing data indicating a moderate separation between rats enriched with *Parabacteroides* and controls. Gene pathway analyses revealed that early life *Parabacteroides* treatment, similar to effects associated with sugar consumption, significantly altered the genetic signature of dopaminergic synaptic signaling pathways, though differentially expressed genes were commonly affected in opposite directions between the two experimental conditions (Fig. S8). *Parabacteroides* treatment also impacted gene pathways associated with metabolic signaling. Specifically, pathways regulating fatty acid oxidation, rRNA metabolic processes, mitochondrial inner membrane, and valine, leucine, and isoleucine degradation were significantly affected by *Parabacteroides* enrichment. Other pathways that were influenced were those involved in neurodegenerative disorders, including Alzheimer's disease and Parkinson's disease, though most of the genes affected in these pathways were mitochondrial genes (Fig. 5D, Table S2).

At the level of individual genes, dorsal hippocampal RNA sequencing data revealed that 15 genes were

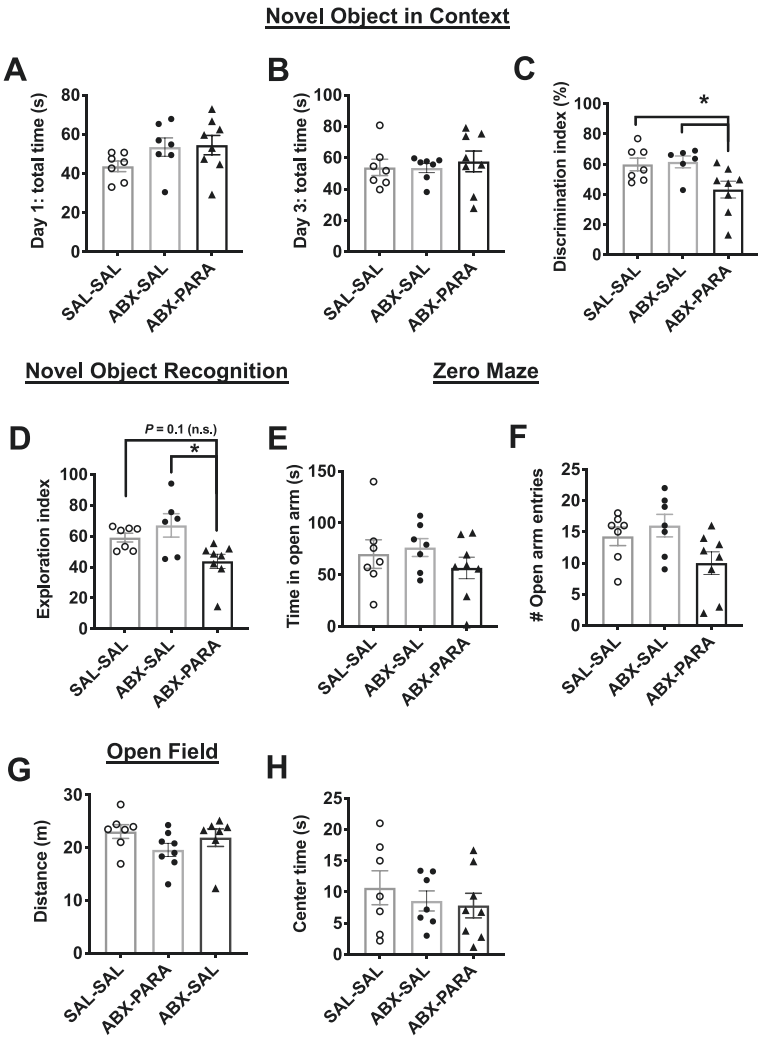


Fig. 4 Early life enrichment with *Parabacteroides* negatively impacts neurocognitive function. A, **B** Early life enrichment with a 1:1 ratio of *P. johnsonii* and *P. distasonis* had no effect on total exploration time in the Novel Object in Context (NOIC) task. **C** Discrimination index was significantly reduced by enrichment with *P. johnsonii* and *P. distasonis*, indicating impaired hippocampal function ($n = 7,8$; $F_{(2, 19)} = 4.92$; $P < 0.05$, one-way ANOVA with Tukey's multiple comparison test). **D** There was a significant reduction in the exploration index in the Novel Object Recognition (NOR task), indicating impaired perirhinal cortex function ($n = 7,8$; $F_{(2, 19)} = 3.61$; $P < 0.05$, one-way ANOVA with Tukey's multiple comparison test). **E**, **F** There were no differences in time spent or a number of entries into the open arm by animals with *P. johnsonii* and *P. distasonis* enrichment in the Zero Maze task for anxiety-like behavior ($n = 7,8$; one-way ANOVA). **G**, **H** There were no differences in distance traveled or time spent in the center arena in the Open Field task ($n = 7,8$; one-way ANOVA). SAL–SAL saline–saline control, ABX-SAL antibiotics-saline control, ABX-PARA antibiotics-*P. johnsonii* and *P. distasonis* enriched, PN post-natal day; data shown as mean ± SEM; $^*P < 0.05$.

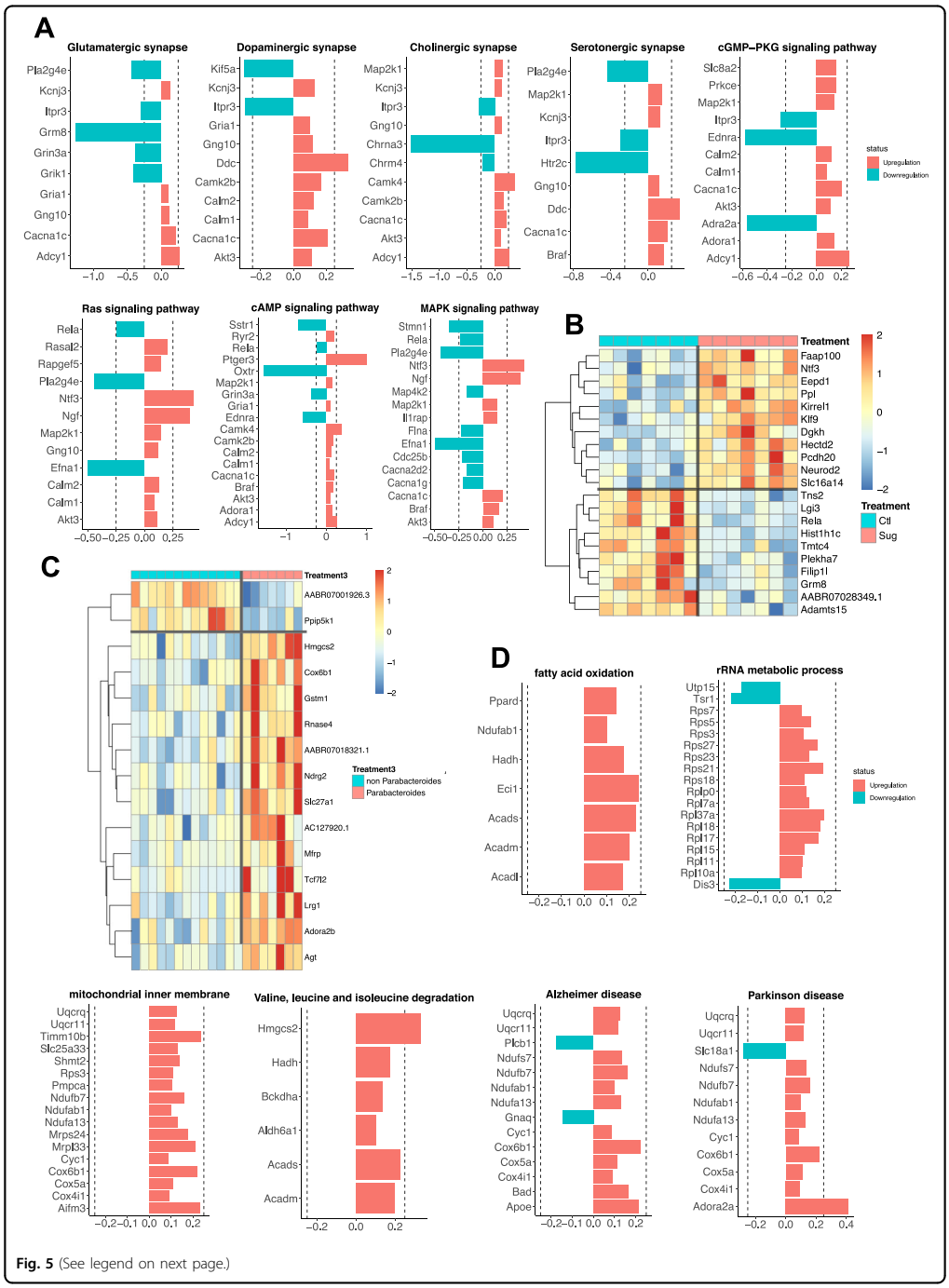


Fig. 5 (See legend on next page.)

(see figure on previous page)
Fig. 5 Effect of early life sugar or targeted *Parabacteroides* enrichment on hippocampal gene expression. A Pathway analyses for differentially expressed genes (DEGs) at a P value < 0.01 in hippocampal tissue punches from rats fed early life sugar compared with controls. Upregulation by sugar is shown in red and downregulation by sugar in blue. **B** A heatmap depicting DEGs that survived the Benjamini–Hochberg corrected FDR of P < 0.05 in rats fed early life sugar compared with controls. Warmer colors (red) signify an increase in gene expression and cool colors (blue) a reduction in gene expression by treatment (CTL control, SUG early life sugar; $n = 7$/group). **C** A heatmap depicting DEGs that survived the Benjamini–Hochberg corrected FDR of $P < 0.05$ in rats with early life *Parabacteroides* enrichment compared with combined control groups. Warmer colors (red) signify an increase in gene expression and cool colors (blue) a reduction in gene expression by treatment ($n = 7, 14$). **D** Pathway analyses for differentially expressed genes (DEGs) at a P value < 0.01 in rats enriched with *Parabacteroides* compared with combined controls. Upregulation by *Parabacteroides* transfer is shown in red and downregulation in blue. The dotted line indicates ±0.25 log2 fold change.

differentially expressed in rats enriched with *Parabacteroides* compared with controls, with 13 genes elevated and two genes decreased in the *Parabacteroides* group compared with controls (Fig. 5C). Consistent with results from gene pathway analyses, several individual genes involved in metabolic processes were elevated by *Parabacteroides* enrichment, such as *Hmgcs2*, which is a mitochondrial regulator of ketogenesis and provides energy to the brain under metabolically taxing conditions or when glucose availability is low[57], and *Cox6b1*, a mitochondrial regulator of energy metabolism that improves hippocampal cellular viability following ischemia/reperfusion injury[58]. *Parabacteroides* enrichment was also associated with incased expression of *Slc27A1* and *Mfrp*, which are each critical for the transport of fatty acids into the brain across capillary endothelial cells[59,60].

Discussion

Dietary factors are a key source of gut microbiome diversity[28,46,61–63] and emerging evidence indicates that diet-induced alterations in the gut microbiota may be linked with altered neurocognitive development[28,63–65]. Our results identify species within the genus *Parabacteroides* that are elevated by habitual early life consumption of dietary sugar and are negatively associated with hippocampal-dependent memory performance. Further, targeted microbiota enrichment of *Parabacteroides* perturbed both hippocampal- and perirhinal cortex-dependent memory performance. These findings are consistent with previous literature in showing that early life consumption of Western dietary factors impairs neurocognitive outcomes[10,11], and further suggest that altered gut bacteria due to excessive early life sugar consumption may functionally link dietary patterns with cognitive impairment.

Our previous data show that rats are not susceptible to habitual sugar consumption-induced learning and memory impairments when 11% sugar solutions are consumed ad libitum during adulthood, in contrast to effects observed in the present and previous study in which the sugar is consumed during early life development[23]. It is possible that habitual sugar consumption differentially affects the gut microbiome when consumed during adolescence vs. adulthood. However, a recent report showed that adult consumption of a high fructose diet (35% kcal from fructose) promotes gut microbial "dysbiosis" and neuroinflammation and cell death in the hippocampus, yet without impacting cognitive function[66], suggesting that perhaps neurocognitive function is more susceptible to gut microbiota influences during early life than during adulthood. Indeed, several reports have identified early life critical periods for microbiota influences on behavioral and neurochemical endpoints in germ-free mice[5,75]. However, the age-specific profile of sugar-associated microbiome dysbiosis and neurocognitive impairments remains to be determined.

Given that the adolescent rats consuming SSBs compensated for these calories by consuming less chow, it is possible that reduced nutrient (e.g., dietary protein) consumption may have contributed to the deficits in hippocampal function. However, we think this is unlikely, as adolescent SSB access did not produce any substantial nutrient deficiency that would restrict growth, as evidenced by the similarities in body weight between the experimental and control group. Furthermore, prior studies that directly examined the effects of adolescent caloric (and thereby nutrient) restriction on learning and memory in rats found that there were no differences in hippocampal-dependent memory function when rats were restricted by ~40% from PN 25 to PN 67[67], Importantly, the parameters in this study closely match those in the present study, as our adolescent SSB access was given over a similar developmental period prior to behavioral testing, and produced a ~40% reduction in total chow kcal consumption. Thus, it is likely that excessive sugar consumption and not nutrient deficiency led to memory deficits, although future work is needed to more carefully examine these variables independently.

While our study reveals a strong negative correlation between levels of fecal *Parabacteroides* and performance in the hippocampal-dependent contextual episodic memory NOIC task, as well as impaired NOIC performance in rats given access to a sugar solution during adolescence, sugar intake did not produce impairments in

the perirhinal cortex-dependent NOR memory task. This is consistent with our previous report in which rats given access to an 11% sugar solution during adolescence were impaired in hippocampal-dependent spatial memory (Barne's maze procedure), yet were not impaired in a nonspatial task of comparable difficulty that was not hippocampal-dependent[23]. Present results revealing that early life sugar consumption negatively impacts hippocampal-dependent contextual-based object recognition memory (NOIC) without influencing NOR memory performance is also consistent with previous reports using a cafeteria diet high in both fat content and sugar[68,69]. On the other hand, enrichment of *P. johnsonii* and *P. distasonis* in the present study impaired memory performance in both tasks, suggesting a broader impact on neurocognitive functioning with this targeted bacterial enrichment approach.

Gene pathway analyses from dorsal hippocampus RNA sequencing identified multiple neurobiological pathways that may functionally connect gut dysbiosis with memory impairment. Early life sugar consumption was associated with alterations in several neurotransmitter synaptic signaling pathways (e.g., glutamatergic and cholinergic) and intracellular signaling targets (e.g., cAMP and MAPK). A different profile was observed in *Parabacteroides*-enriched animals, where gene pathways involved with metabolic function (e.g., fatty acid oxidation and branched-chain amino acid degradation) and neurodegenerative disease (e.g., Alzheimer's disease) were altered relative to controls. Given that sugar has effects on bacterial populations in addition to *Parabacteroides*, and that sugar consumption and *Parabacteroides* treatment differentially influenced peripheral glucose metabolism and body weight, these transcriptome differences in the hippocampus are not surprising. However, gene clusters involved with dopaminergic synaptic signaling were significantly influenced by both early life sugar consumption and *Parabacteroides* treatment, thus identifying a common pathway through which both diet-induced and gut bacterial infusion-based elevations in *Parabacteroides* may influence neurocognitive development. Though differentially expressed genes were commonly affected in opposite directions in *Parabacteroides* enriched animals compared with early life sugar treated animals, it is possible that perturbations to the dopamine system play a role in the observed cognitive dysfunction. For example, while dopamine signaling in the hippocampus has not traditionally been investigated for mediating memory processes, several recent reports have identified a role for dopamine inputs from the locus coeruleus in regulating hippocampal-dependent memory and neuronal activity[70,71]. Interestingly, endogenous dopamine signaling in the hippocampus has recently been linked with regulating food intake and food-associated contextual learning[72], suggesting that dietary effects on gut microbiota may also impact feeding behavior and energy balance-relevant cognitive processes.

It is important to note that comparisons between the gene expressional analyses in the *Parabacteroides* enrichment and sugar consumption experiments should be made cautiously given that there were slight differences in the timing of the hippocampus tissue harvest between the two experiments (PN 65 for sugar consumption vs. PN 83 for the *Parabacteroides* enrichment). Further, future work is needed to determine whether differences in gene expression observed in each experiment translates to differential expression at the protein level. It is also worth emphasizing that the levels of *Parabacteroides* conferred by our enrichment study were substantially higher than in the dietary sugar study, and thus it is not surprising that *Parabacteroides* enrichment would confer a different impact on host physiology, hippocampal gene expression, and neurocognition compared to *Parabacteroides* elevations associated with SSB consumption. Regardless of these caveats in comparing the two models, our data extend the field by highlighting a specific bacterial population that (1) is capable of negatively impacting neurocognitive development when experimentally enriched, and (2) is elevated by early life consumption of dietary sugar with levels correlating negatively with hippocampal-dependent memory performance.

Many of the genes that were differentially upregulated in the hippocampus by *Parabacteroides* enrichment were involved in fat metabolism and transport. Thus, it is possible that *Parabacteroides* conferred an adaptation in the brain, shifting fuel preference away from carbohydrate toward lipid-derived ketones. Consistent with this framework, *Parabacteroides* were previously shown to be upregulated by a ketogenic diet in which carbohydrate consumption is drastically depleted and fat is used as a primary fuel source due. Furthermore, enrichment of *Parabacteroides merdae* together with *Akkermansia muciniphila* was protective against seizures in mice[28]. It is possible that *P. distasonis* reduces glucose uptake from the gut, enhances glucose clearing from the blood, and/or alters nutrient utilization in general, an idea further supported by the recent finding that *P. distasonis* is associated with reduced diet- and genetic-induced obesity and hyperglycemia in mice[47].

The present findings produce several opportunities for further mechanistic investigation. For example, how do diet-induced alterations in gut bacteria impact the brain? Several possible mechanisms have been investigated and proposed, such as impaired gut barrier function and endotoxemia[63,73], perhaps related to altered short-chain fatty acid production[66,74]. Moreover, it is well-known that the liver is negatively impacted by excessive fructose consumption[75], and emerging evidence highlights a gut microbiome–liver axis with crosstalk via bile acids and

cytokines[76]. It is possible that dietary sugar-induced microbiota changes alter the hepatic–gut axis, thus contributing to altered cognitive function. Indeed, an altered bile acid profile due to gut microbiota-produced bile acid secondary metabolites is associated with cognitive dysfunction in Alzheimer's Disease in humans[77].

Taken together, our collective results provide insight into the neurobiological mechanisms that link early life unhealthy dietary patterns with altered gut microbiota changes and neurocognitive impairments. Currently, probiotics, live microorganisms intended to confer health benefits, are not regulated with the same rigor as pharmaceuticals but instead are sold as dietary supplements. Our findings suggest that gut enrichment with certain species of *Parabacteroides* is potentially harmful to neurocognitive mnemonic development. These results highlight the importance of conducting rigorous basic science analyses on the relationship between diet, microorganisms, brain, and behavior prior to widespread recommendations of bacterial microbiome interventions for humans.

Acknowledgements
We thank Alyssa Cortella for contributing the rodent artwork. We thank Caroline Szjewski, Lekha Chirala, Vaibhav Konanur, Sarah Terrill, and Ted Hsu for their critical contributions to the research. The research was supported by DK116942 and DK104897, and institutional funds to S.E.K., DK118000 and DK111158 to E.E.N., DK116558 to A.N.S., D.K. 118944 to C.M.L. C.A.O. was supported by an F31 AG064844. E.Y.H. was supported by the ARO MURI award W911NF-17-1-0402. DK104363 to X.Y., Eureka Scholarship and BWF-CHIP Fellowship to Y.C.

Author details
[1]University of Georgia, Athens, GA, USA. [2]University of California, Los Angeles, CA, USA. [3]University of Southern California, Los Angeles, CA, USA

Data availability
All data are available upon request. The 16S rRNI microbiome sequencing data are available through Qiita (ID 13651 and 11255) and the RNA sequencing data are available through the NCBI Gene Expression Omnibus, GSE150091.

Conflict of interest
The authors declare no competing interests.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41398-021-01309-7.

Received: 26 January 2021 Revised: 18 February 2021 Accepted: 2 March 2021
Published online: 31 March 2021

References

1. Vuong, H. E., Yano, J. M., Fung, T. C. & Hsiao, E. Y. The microbiome and host behavior. *Annu. Rev. Neurosci.* **40**, 21–49 (2017).
2. Noble, E. E., Hsu, T. M. & Kanoski, S. E. Gut to brain dysbiosis: mechanisms linking western diet consumption, the microbiome, and cognitive impairment. *Front. Behav. Neurosci.* **11**, 9 (2017).
3. Lach, G. et al. Enduring neurobehavioral effects induced by microbiota depletion during the adolescent period. *Transl. Psychiatry* **10**, 382 (2020).
4. Morais, L. H. et al. Enduring behavioral effects induced by birth by caesarean section in the mouse. *Curr. Biol.* **30**, 3761–3774 e3766 (2020).
5. Neufeld, K. A., Kang, N., Bienenstock, J. & Foster, J. A. Effects of intestinal microbiota on anxiety-like behavior. *Commun. Integr. Biol.* **4**, 492–494 (2011).
6. Sudo, N. et al. Postnatal microbial colonization programs the hypothalamic-pituitary-adrenal system for stress response in mice. *J. Physiol.* **558**, 263–275 (2004).
7. Diaz Heijtz, R. et al. Normal gut microbiota modulates brain development and behavior. *Proc. Natl Acad. Sci. USA* **108**, 3047–3052 (2011).
8. Cryan, J. F. et al. The microbiota-gut-brain axis. *Physiol. Rev.* **99**, 1877–2013, https://doi.org/10.1152/physrev.00018.2018 (2019).
9. Kanoski, S. E. & Davidson, T. L. Western diet consumption and cognitive impairment: links to hippocampal dysfunction and obesity. *Physiol. Behav.* **103**, 59–68 (2011).
10. Noble, E. E., Hsu, T. M., Liang, J. & Kanoski, S. E. Early-life sugar consumption has long-term negative effects on memory function in male rats. *Nutr. Neurosci.* https://doi.org/10.1080/1028415X.2017.1378851 (2019).
11. Noble, E. E. & Kanoski, S. E. Early life exposure to obesogenic diets and learning and memory dysfunction. *Curr. Opin. Behav. Sci.* **9**, 7–14 (2016).
12. Hsu, T. M. et al. Hippocampus ghrelin receptor signaling promotes socially-mediated learned food preference. *Neuropharmacology* **131**, 487–496 (2018).
13. Hsu, T. M. et al. A hippocampus to prefrontal cortex neural pathway inhibits food motivation through glucagon-like peptide-1 signaling. *Mol. Psychiatry* **23**, 1555–1565 (2018).
14. Hsu, T. M. et al. Hippocampus ghrelin signaling mediates appetite through lateral hypothalamic orexin pathways. *Elife* https://doi.org/10.7554/eLife.11190 (2015).
15. Kanoski, S. E., Fortin, S. M., Ricks, K. M. & Grill, H. J. Ghrelin signaling in the ventral hippocampus stimulates learned and motivational aspects of feeding via PI3K-Akt signaling. *Biol. Psychiatry* **73**, 915–923 (2013).
16. Davidson, T. L. et al. Contributions of the hippocampus and medial prefrontal cortex to energy and body weight regulation. *Hippocampus* **19**, 235–252 (2009).
17. Kanoski, S. E. & Grill, H. J. Hippocampus contributions to food intake control: mnemonic, neuroanatomical, and endocrine mechanisms. *Biol. Psychiatry* **81**, 748–756 (2017).
18. Davidson, T. L., Sample, C. H. & Swithers, S. E. An application of Pavlovian principles to the problems of obesity and cognitive decline. *Neurobiol. Learn. Mem.* **108**, 172–184 (2014).
19. Baym, C. L. et al. Dietary lipids are differentially associated with hippocampal-dependent relational memory in prepubescent children. *Am. J. Clin. Nutr.* **99**, 1026–1032 (2014).
20. Valladolid-Acebes, I. et al. Spatial memory impairment and changes in hippocampal morphology are triggered by high-fat diets in adolescent mice. Is there a role of leptin? *Neurobiol. Learn Mem.* **106**, 18–25 (2013).
21. Boitard, C. et al. Impairment of hippocampal-dependent memory induced by juvenile high-fat diet intake is associated with enhanced hippocampal inflammation in rats. *Brain Behav. Immun.* **40**, 9–17 (2014).
22. Boitard, C. et al. Juvenile, but not adult exposure to high-fat diet impairs relational memory and hippocampal neurogenesis in mice. *Hippocampus* **22**, 2095–2100 (2012).
23. Hsu, T. M. et al. Effects of sucrose and high fructose corn syrup consumption on spatial memory function and hippocampal neuroinflammation in adolescent rats. *Hippocampus* **25**, 227–239 (2015).
24. Kendig, M. D., Boakes, R. A., Rooney, K. B. & Corbit, L. H. Chronic restricted access to 10% sucrose solution in adolescent and young adult rats impairs spatial memory and alters sensitivity to outcome devaluation. *Physiol. Behav.* **120**, 164–172 (2013).
25. Reichelt, A. C., Killcross, S., Hambly, L. D., Morris, M. J. & Westbrook, R. F. Impact of adolescent sucrose access on cognitive control, recognition memory, and parvalbumin immunoreactivity. *Learn Mem.* **22**, 215–224 (2015).
26. Noble, E. E., Hsu, T. M., Liang, J. & Kanoski, S. E. Early-life sugar consumption has long-term negative effects on memory function in male rats. *Nutr. Neurosci.* **22**, 273–283 (2019).
27. Walker, R. W., Dumke, K. A. & Goran, M. I. Fructose content in popular beverages made with and without high-fructose corn syrup. *Nutrition* **30**, 928–935 (2014).
28. Olson, C. A. et al. The gut microbiota mediates the anti-seizure effects of the ketogenic diet. *Cell* **173**, 1728–1741 e1713 (2018).

29. Martinez, M. C., Villar, M. E., Ballarini, F. & Viola, H. Retroactive interference of object-in-context long-term memory: role of dorsal hippocampus and medial prefrontal cortex. *Hippocampus* **24**, 1482–1492 (2014).
30. Balderas, I. et al. The consolidation of object and context recognition memory involve different regions of the temporal lobe. *Learn Mem.* **15**, 618–624 (2008).
31. Beilharz, J. E., Maniam, J. & Morris, M. J. Short exposure to a diet rich in both fat and sugar or sugar alone impairs place, but not object recognition memory in rats. *Brain Behav. Immun.* **37**, 134–141 (2014).
32. Thompson, L. R. et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
33. Caporaso, J. G. et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).
34. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419, https://doi.org/10.1038/nmeth.4197 (2017).
35. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521 (2015).
36. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
37. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–97 (2016).
38. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
39. The Gene Ontology, C. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45**, D331–D338 (2017).
40. Slenter, D. N. et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* **46**, D661–D667 (2018).
41. Aggleton, J. P. & Brown, M. W. Contrasting hippocampal and perirhinal cortex function using immediate early gene imaging. *Q. J. Exp. Psychol. B* **58**, 218–233 (2005).
42. Albasser, M. M., Davies, M., Futter, J. E. & Aggleton, J. P. Magnitude of the object recognition deficit associated with perirhinal cortex damage in rats: Effects of varying the lesion extent and the duration of the sample period. *Behav. Neurosci.* **123**, 115–124 (2009).
43. Cohen, S. J. & Stackman, R. W. Jr. Assessing rodent hippocampal involvement in the novel object recognition task. A review. *Behav. Brain Res.* **285**, 105–117 (2015).
44. Sestakova, N., Puzserova, A., Kluknavsky, M. & Bernatova, I. Determination of motor activity and anxiety-related behaviour in rodents: methodological aspects and role of nitric oxide. *Interdiscip. Toxicol.* **6**, 126–135 (2013).
45. Goran, M. I. et al. The obesogenic effect of high fructose exposure during early development. *Nat. Rev. Endocrinol.* **9**, 494–500 (2013).
46. Noble, E. E. et al. Early-life sugar consumption affects the rat microbiome independently of obesity. *J. Nutr.* **147**, 20–28 (2017).
47. Wang, K. et al. *Parabacteroides distasonis* alleviates obesity and metabolic dysfunctions via production of succinate and secondary bile acids. *Cell Rep.* **26**, 222–235 e225 (2019).
48. Ling, C. et al. FAAP100 is essential for activation of the Fanconi anemia-associated DNA damage response pathway. *EMBO J.* **26**, 2104–2114 (2007).
49. Kim, H. S. et al. Endonuclease EEPD1 is a gatekeeper for repair of stressed replication forks. *J. Biol. Chem.* **292**, 2795–2804 (2017).
50. Zucker, S. N. et al. Nrf2 amplifies oxidative stress via induction of Klf9. *Mol. Cell* **53**, 916–928 (2014).
51. Yasuda, S. et al. Diacylglycerol kinase eta augments C-Raf activity and B-Raf/C-Raf heterodimerization. *J. Biol. Chem.* **284**, 29559–29570 (2009).
52. Murdoch, H. et al. Periplakin interferes with G protein activation by the melanin-concentrating hormone receptor-1 by binding to the proximal segment of the receptor C-terminal tail. *J. Biol. Chem.* **280**, 8208–8220 (2005).
53. Gerke, P. et al. Neuronal expression and interaction with the synaptic protein CASK suggest a role for Neph1 and Neph2 in synaptogenesis. *J. Comp. Neurol.* **498**, 466–475 (2006).
54. Chen, H., Duncan, I. C., Bozorgchami, H. & Lo, S. H. Tensin1 and a previously undocumented family member, tensin2, positively regulate cell migration. *Proc. Natl Acad. Sci. USA* **99**, 733–738 (2002).
55. O'Mahony, A. et al. NF-kappaB/Rel regulates inhibitory and excitatory neuronal function and synaptic plasticity. *Mol. Cell Biol.* **26**, 7283–7298 (2006).
56. Gerlai, R., Adams, B., Fitch, T., Chaney, S. & Baez, M. Performance deficits of mGluR8 knockout mice in learning tasks: the effects of null mutation and the background genotype. *Neuropharmacology* **43**, 235–249 (2002).
57. Shao, X. et al. HMG-CoA synthase 2 drives brain metabolic reprogramming in cocaine exposure. *Neuropharmacology* **148**, 377–393 (2019).
58. Yang, S., Wu, P., Xiao, J. & Jiang, L. Overexpression of COX6B1 protects against I/Rinduced neuronal injury in rat hippocampal neurons. *Mol. Med Rep.* **19**, 4852–4862 (2019).
59. Ochiai, Y. et al. The blood-brain barrier fatty acid transport protein 1 (FATP1/SLC27A1) supplies docosahexaenoic acid to the brain, and insulin facilitates transport. *J. Neurochem.* **141**, 400–412 (2017).
60. Kautzmann, M. I. et al. Membrane-type frizzled-related protein regulates lipidome and transcription for photoreceptor function. *FASEB J.* **34**, 912–929 (2020).
61. David, L. A. et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).
62. de La Serre, C. B. et al. Propensity to high-fat diet-induced obesity in rats is associated with changes in the gut microbiota and gut inflammation. *Am. J. Physiol. Gastrointest. Liver Physiol.* **299**, G440–448 (2010).
63. Bruce-Keller, A. J. et al. Obese-type gut microbiota induce neurobehavioral changes in the absence of obesity. *Biol. Psychiatry* **77**, 607–615 (2015).
64. Leigh, S. J., Kaakoush, N. O., Westbrook, R. F. & Morris, M. J. Minocycline-induced microbiome alterations predict cafeteria diet-induced spatial recognition memory impairments in rats. *Transl. Psychiatry* **10**, 92 (2020).
65. Leigh, S. J., Kaakoush, N. O., Bertoldo, M. J., Westbrook, R. F. & Morris, M. J. Intermittent cafeteria diet identifies fecal microbiome changes as a predictor of spatial recognition memory impairment in female rats. *Transl. Psychiatry* **10**, 36 (2020).
66. Li, J. M. et al. Dietary fructose-induced gut dysbiosis promotes mouse hippocampal neuroinflammation: a benefit of short-chain fatty acids. *Microbiome* **7**, 98 (2019).
67. Alamy, M., Errami, M., Taghzouti, K., Saddiki-Traki, F. & Bengelloun, W. A. Effects of postweaning undernutrition on exploratory behavior, memory and sensory reactivity in rats: implication of the dopaminergic system. *Physiol. Behav.* **86**, 195–202 (2005).
68. Kendig, M. D., Westbrook, R. F. & Morris, M. J. Pattern of access to cafeteria-style diet determines fat mass and degree of spatial memory impairments in rats. *Sci. Rep.* **9**, 13516 (2019).
69. Yang, Y. et al. Early-life high-fat diet-induced obesity programs hippocampal development and cognitive functions via regulation of gut commensal Akkermansia muciniphila. *Neuropsychopharmacology* **44**, 2054–2064 (2019).
70. Takeuchi, T. et al. Locus coeruleus and dopaminergic consolidation of everyday memory. *Nature* **537**, 357–362 (2016).
71. Kempadoo, K. A., Mosharov, E. V., Choi, S. J., Sulzer, D. & Kandel, E. R. Dopamine release from the locus coeruleus to the dorsal hippocampus promotes spatial learning and memory. *Proc. Natl Acad. Sci. USA* **113**, 14835–14840 (2016).
72. Azevedo, E. P. et al. A role of Drd2 hippocampal neurons in context-dependent food intake. *Neuron* **102**, 873–886 e875 (2019).
73. Ou, Z. et al. Protective effects of Akkermansia muciniphila on cognitive deficits and amyloid pathology in a mouse model of Alzheimer's disease. *Nutr. Diabetes* **10**, 12 (2020).
74. Hu, L. et al. High salt elicits brain inflammation and cognitive dysfunction, accompanied by alternations in the gut microbiota and decreased SCFA production. *J. Alzheimers Dis.* **77**, 629–640 (2020).
75. Stanhope, K. L. Sugar consumption, metabolic disease and obesity: the state of the controversy. *Crit. Rev. Clin. Lab. Sci.* **53**, 52–67 (2016).
76. Cerdo, T., Dieguez, E. & Campoy, C. Early nutrition and gut microbiome: interrelationship between bacterial metabolism, immune system, brain structure, and neurodevelopment. *Am. J. Physiol. Endocrinol. Metab.* **317**, E617–E630 (2019).
77. MahmoudianDehkordi, S. et al. Altered bile acid profile associates with cognitive impairment in Alzheimer's disease—an emerging role for gut microbiome. *Alzheimers Dement.* **15**, 76–92 (2019).

Emerging Topics in Life Sciences (2019) **3** 379–398
https://doi.org/10.1042/ETLS20180176

Appendix D

Review Article

Network modeling of single-cell omics data: challenges, opportunities, and progresses

Montgomery Blencowe[1], Douglas Arneson[1,2], Jessica Ding[1], Yen-Wei Chen[1,3], Zara Saleem[1] and Xia Yang[1,2,3,4]

[1]Department of Integrative Biology and Physiology, University of California, Los Angeles, 610 Charles E. Young Drive East, Los Angeles, CA 90095, U.S.A.; [2]Bioinformatics Interdepartmental Program, University of California, Los Angeles, 610 Charles E. Young Drive East, Los Angeles, CA 90095, U.S.A.; [3]Molecular Toxicology Program, University of California, Los Angeles, 610 Charles E. Young Drive East, Los Angeles, CA 90095, U.S.A.; [4]Institute for Quantitative and Computational Biosciences, University of California, Los Angeles, 610 Charles E. Young Drive East, Los Angeles, CA 90095, U.S.A.

Correspondence: Xia Yang (xyang123@ucla.edu)

Single-cell multi-omics technologies are rapidly evolving, prompting both methodological advances and biological discoveries at an unprecedented speed. Gene regulatory network modeling has been used as a powerful approach to elucidate the complex molecular interactions underlying biological processes and systems, yet its application in single-cell omics data modeling has been met with unique challenges and opportunities. In this review, we discuss these challenges and opportunities, and offer an overview of the recent development of network modeling approaches designed to capture dynamic networks, within-cell networks, and cell–cell interaction or communication networks. Finally, we outline the remaining gaps in single-cell gene network modeling and the outlooks of the field moving forward.

Introduction

Network modeling has long been employed as a powerful tool to understand and interpret complex biological systems, with networks themselves serving both as a computational framework and a major data type. Networks depict biological systems as nodes and edges, where nodes represent biological entities such as genes, proteins, metabolites, phenotypic traits, cells, environmental exposures, or even gut bacteria, and edges represent the relationships between nodes such as regulator–effector connections, statistical correlations, physical binding, and enzymatic or metabolic reactions (Figure 1A). As the amount and types of biological data continue to grow at an exponential rate, so too do the number and types of biological networks including protein–protein interaction networks [1], metabolic networks [2], genetic interaction networks [3], gene/transcriptional regulatory networks (GRNs) [4], and cell signaling networks [5]. While different network models possess inherent strengths and limitations depending on their underlying assumptions, they share the common feature of being graphical models which describe information flow in biological systems to help understand and interpret the underlying biological processes.

Network modeling has seen extensive applications over the past decades to help understand key biological processes and regulators of health and disease. In particular, the enormous complexity of human physiology and pathophysiology demands a systems level understanding of how biological molecules interact within individual cells and tissues and between cells and tissues to maintain homeostasis, and how perturbations of these interactions lead to diseases. The omnigenic disease model, which states that all genes interacting in networks can contribute to complex diseases, is increasingly recognized and accepted [6]. These conceptual frameworks match perfectly with the capacity of network biology, and, therefore, it is not surprising to witness the increasing use of network modeling approaches in essentially all fields of biology. For example, many genetic variants can influence disease, each through very small effects which make biological interpretation difficult.

Received: 16 April 2019
Revised: 7 June 2019
Accepted: 24 June 2019

Version of Record published:
8 July 2019

PORTLAND PRESS

Emerging Topics in Life Sciences (2019) **3** 379–398
https://doi.org/10.1042/ETLS20180176

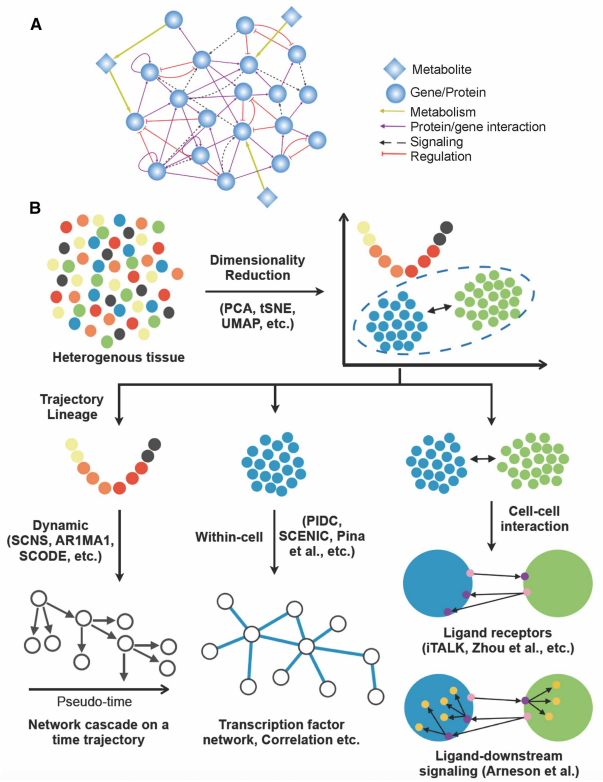


Figure 1. Overview of molecular networks and single-cell network modeling approaches.

(**A**) Various types of network nodes and edges (connections) in a molecular network. (**B**) Concepts and example methods of current single-cell network modeling approaches. Transcriptomes of single cells from tissue samples are first processed and clustered using dimension reduction techniques, followed by cell identity determination using known cell markers. Cell populations in various dynamic states can be ordered using pseudo-time and trajectory analyses, and the pseudo-time information can be used in dynamic network modeling. Gene networks within a given cell population and cell–cell communication networks can also be reconstructed based on various assumptions and algorithms.

These convoluted genetic effects can be better understood through their relationships in transcriptional and signaling networks and biological pathways [6,7]. Our group and others have leveraged network models for the interpretation of genetic causes of complex diseases [8–20]. Similarly, networks can be used to understand the molecular cascades involved in various environmental causes of diseases [21–23]. For example, by integrating genetic association with tissue-specific GRNs, Chella Krishnan et al. [20] found that numerous genetic variants associated with nonalcoholic fatty liver disease affect diverse biological pathways ranging from lipid metabolism, immune system, cell cycle, transcriptional regulation to insulin signaling, Notch signaling, and oxidative phosphorylation that interact in GRNs in the liver and adipose tissues. Based on network topology, they identified key regulators involved in mitochondrial function at the center of disease pathways and subnetworks. In another study, network modeling of genetic risks of cardiovascular disease and type 2 diabetes using tissue-specific GRNs revealed shared and disease-specific networks and regulators [10]. In a systematic effort,

Emerging Topics in Life Sciences (2019) **3** 379–398
https://doi.org/10.1042/ETLS20180176

Greene et al. [16] constructed 144 tissue/cell-specific networks and used these networks to predict and understand lineage-specific responses to IL1B stimulation, tissue-specific activities and functions of LEF1, connection of Parkinson's disease with other diseases, and the tissue context and genes involved in hypertension.

While network-based approaches have furthered our understanding of complex diseases, it is important to note that the majority of the network methodologies and applications have primarily relied upon omics data derived from bulk tissues. At the bulk tissue level, many methodologies and algorithms have been developed for network modeling, primarily focusing on predicting GRNs within [24–28] and between tissues [29–31] with reasonable accuracy. However, bulk tissues like non-parenchymal cells in the liver are comprised of heterogeneous cell populations including Kupffer cells, sinusoidal endothelial cells, and hepatocyte satellite cells, all with distinct functions associated with unique gene regulatory profiles [32]. Given the heterogeneous nature of tissues, bulk tissue networks primarily represent the average activities of all cell populations which can be dominated by the most abundant cell types. Therefore, tissue networks cannot capture the unique behaviors of individual cell populations and how cells interact to perform higher-level tissue functions.

The recent explosion of high-throughput single-cell omics technologies brings exciting possibilities to model dynamic, within-, and between-cell gene networks to elucidate the processes underlying cell development, functional state, and cell–cell communications that are missing from bulk tissue networks. These single-cell omics technologies give us the unprecedented ability to examine the transcriptional, protein, and epigenomic profiles at single-cell resolution, which are necessary to tease apart the regulatory and functional relationships of biomolecules within individual cells or cell types, and between cell populations. In theory, similar framework and methodologies that have been used for bulk tissue network modeling could be extended to single-cell data to uncover the regulatory mechanisms governing functions within and between cells. However, as suggested by Chen and Mar in their recent study [33], the models for bulk tissue may not be well suited to overcome the unique challenges introduced by single-cell data.

Here, we discuss the existing network modeling approaches developed for bulk tissue omics data, the unique challenges imposed by single-cell omics data for use in network modeling, the recent development of approaches for network models which make use of single-cell data, and their key underlying algorithms, advantages, and disadvantages. Lastly, we discuss the remaining gaps to be overcome and where we see the field headed to achieve a more efficient and accurate modeling of gene regulatory networks based on single-cell omics data.

Commonly used GRN modeling approaches for bulk tissue data

Common GRN methods that have been developed and optimized for bulk tissue data are generally based on correlation, regression, ordinary differential equations (ODEs), mutual information, Gaussian graphical models, and Bayesian approaches [24–28]. For example, a correlation-based method named weighted gene coexpression network analysis (WGCNA) is the most commonly used methodology [34,35]. WGCNA is used to find clusters (or modules) of genes which are highly correlated and usually represent tightly regulated genes involved in similar biological pathways or functions. Although coexpression-based methods are computationally efficient and less dependent on assumptions, these methods mainly group genes involved in similar functions or under similar regulation, but cannot infer directionality or direct regulatory relations and require integration with other information to facilitate interpretability [28]. Regression-based methods such as GENIE3 resolve networks by determining the most predictive subset of genes for each network gene based on regression models [36]. These methods perform well for linear cascades but not for feed-forward loops. For methods based on mutual information, such as ARACNE [37] and CLR [38], network structure is determined by the degree of dependencies between pairs of genes. These mutual information-based network methods can infer directionality and potential causality, and can predict feed-forward loops more accurately but have limited performance with linear cascades. Bayesian network (BN) modeling approaches offer flexible frameworks to incorporate and integrate multi-omics data as prior information to infer causal and directional gene–gene interactions [39,40]. A BN encodes conditional dependencies between genes, where each gene is determined by the values of its parent nodes. To improve accuracy, BNs search through the multivariate space of possible graphs which comes at the expense of higher computation cost [25,41] and the lack of guarantee that optimal topology can be detected. The different commonly used GRN inference algorithms come with different pros and cons, and the integration of multiple methods can compensate for the disadvantages inherent to each method and provide a

Emerging Topics in Life Sciences (2019) **3** 379–398
https://doi.org/10.1042/ETLS20180176

better interpretation of the data [27]. It is important to note that these methods were optimized for bulk tissue-level data that generally conform to standard data distributions and have few missing values.

Single-cell technologies and data structures

Leveraging the recently developed single-cell technologies, we are now able to examine the transcriptional (DropSeq [42], inDrop [43], 10× Genomics, SmartSeq v4, Mars-Seq [44], Seq-Well [45], SPLiT-seq [46], sci-RNA-seq [47]), protein (CITE-seq [48]), and epigenomic profiles such as open chromatin (scATAC-seq [49]) and methylation landscapes [50]. These single-cell technologies bring about exciting possibilities to explore biology at an unprecedented resolution and scale. The most popular and widely available technologies for interrogating single cells in a high-throughput manner are single-cell RNA sequencing (scRNAseq). Typically, these high-throughput single-cell transcriptome technologies are based on counting transcript fragments from the 3′-end which are then aligned to reference genomes. The resulting data structure which aggregates gene counts for each single cell is called a digital gene expression (DGE) matrix. For other data types, similar cell by marker (e.g. protein, chromatin location, and methylation sites) matrices form the main data structures. Although the projection of single-cell epigenome onto single-cell transcriptome [51] has been performed, the integration of multi-omics data for GRN modeling has not been attempted to our knowledge and is a future direction for methodological development. Multi-omics data can be incorporated in many ways, including constructing a single network with edge confidence extrapolated across omics layers and multiple networks from individual omics layers with interactions between the layers drawn from correlative relations or known functional relevance. For example, open chromatin located in the promoter or enhancer regions of particular genes would allow a directed edge to be drawn between the scATAC-seq and scRNAseq layers; proteome data may help infer interactions between proteins and provide information on regulatory proteins such as transcription factors (TFs) and epigenomic regulators that regulate the transcriptome and epigenome. In this review, we will focus on scRNAseq data as these are the most abundant single-cell data type investigated for GRN modeling.

Performance of existing GRN approaches designed for bulk tissue data in single-cell network modeling

Recently, Chen and Mar [33] evaluated the ability of five generalized network reconstruction methods that were commonly used for bulk tissue data in network construction using both empirical and simulated single-cell data. The methods used in their analyses included partial correlation, BN, GENIE3, ARACNE, and CLR [36–38,52]. Using precision-recall and receiver operating characteristic curves to evaluate whether each method can accurately recapitulate a reference network, it was found that all methods failed to perform significantly better than random generation in both simulated and experimental single-cell datasets. Additionally, there was limited overlap in network predictions between methods. These findings suggest a lack of generalizability and applicability of the existing methods for network construction based on single-cell data. However, caution is needed to interpret such comparison results as the validity of the gold standard reference network used and the quality evaluation metric can significantly influence the comparison results.

Unique challenges, opportunities, and considerations in network modeling of scRNAseq data

The potential lack of performance of existing methods can result from the unique challenges related to data sparsity, distribution, and increased data dimension and capacity [53–55]. First, for scRNAseq using recent high-throughput platforms, due to the miniscule amount of mRNA present in a single-cell and current technological limitations, most entries in DGE matrices are zeros which result in a very sparse matrix, making the direct extension of methods designed for bulk tissue data difficult. Importantly, although these zeros can be a result of stochastic gene expression in individual cells (biological zeros), they do not necessarily mean the absence of mRNA molecules but rather the result of low technical sensitivity for moderate to lowly expressed genes, termed dropouts. Of note, read count-based scRNAseq is zero-inflated, whereas scRNAseq incorporating unique molecular identifier (UMI) counts has been found to possess 'non-zero-inflated' features resulting in different distributions compared with the read count-based technologies [56]. The discrepancies in the underlying data distributions in read count and UMI-based scRNAseq demand the implementation of novel methods that consider these different technologies in the future [57,58].

In an attempt to assign values to dropouts, many single-cell imputation methods such as MAGIC [59], scImpute [60], DrImpute [61], SAVER [62], BISCUIT [63], ScUnif [64], PBLR [65], and deepImpute [66] have been developed in recent years; however, the performance of these methods varies significantly [67]. In benchmarking, scImpute and DrImpute succeeded in simulated data but failed when faced with non-collinear empirical data, while both SAVER and BISCUIT could only consistently impute dropouts with near-zero values [67]. Additionally, the primary metrics used to measure performance (e.g. rand index or mutual information) benchmark the ability of these methods to define cell clusters; it is unclear how these imputed values may affect network structures. Without a consensus and experimental validation of the results from these imputation methods, caution is needed when using imputed data for network construction. A straightforward and intuitive approach taken by Han et al. [68] used subsets of cells from the same cell type and averaged the non-zero values of each gene across cells from each subset to derive a 'supercell' gene expression matrix, which is less inflated with zero values and may prove more biologically relevant. It is important to note that this practice will reduce the cell numbers and sacrifice statistical power.

A second challenge, related to the dropout issue in single-cell data, is the nonstandard data distribution patterns. The high number of dropout values significantly skews the data distribution from unimodal such as a Gaussian distribution to multimodal distributions, which violate the statistical assumptions underlying most of the classic GRN modeling approaches. Careful assessment of the data distribution patterns and tailoring of appropriate statistical methods are needed for single-cell network construction. Several statistical methods such as zero-inflated factor analysis (ZIFA) [69] and ZINB-WaVE (Zero-Inflated Negative Binomial-based Wanted Variation Extraction) [70] have been developed to specifically model the zero-inflated single-cell data distribution. ZIFA is a dimensionality reduction method based on the assumption that lowly expressed genes are more likely to result in dropouts than those that are highly expressed. ZIFA extends factor analysis by incorporating a model of the dropout rate as an exponential decay based on the mean non-zero expression. However, there are limitations to ZIFA in that it models strictly zero measurements and cannot account for near-zero values. Additionally, ZIFA has an underlying linear transformation framework; however, nonlinear dimensionality reduction techniques like t-SNE and UMAP have been demonstrated to be useful in the interpretation of single-cell data, so the extension of zero-inflation modeling to these nonlinear approaches could be useful. ZINB-WaVE is another dimensionality reduction technique which uniquely models the count nature of scRNAseq data and offers normalization using a sample-level intercept and flexible gene-level and sample-level covariate incorporation to address batch effects and sequence composition effects (e.g. gene length or GC content). To address the zero inflation and overdispersion of the data, ZINB-WaVE modifies a standard negative binomial distribution which does not fit the data well, with a term that gives the probability of observing a 0 instead of an actual count. Although ZINB-WaVE is primarily demonstrated as a dimensionality reduction technique for single-cell data, the authors suggest that the low-dimensional representation can be used in downstream analyses like clustering or pseudo-time. Recently, Townes et al. [58] found that multinomial methods outperform other current practices in feature selection and dimension reduction. Consideration of these alternative statistical methods in GRN inference may prove useful. It should be noted that these statistical methods have been developed for read count data and may not be suitable for UMI-based single-cell datasets as they have different underlying data distributions which are not zero-inflated.

Thirdly, it is essential that the field masters the ability to correct for confounding factors and extrapolate data acquired from multiple experiments into one common atlas. Challenges arise as the composition of cellular data remains variable across batches and studies, and even when batches contain the same cell types, the cell number and transcriptional states of individual cell types can vary significantly due to procedural noise (tissue dissociation, sorting, and reagent batches), scRNAseq platforms (e.g. 10× Genomics vs. Dropseq), and chemistry versions (version 2 vs. 3 of 10× Genomics). Much like using batch correction within the bulk tissue setting to adjust confounding factors, the integration of data sets produced by different experiments or even labs are invaluable, as it bolsters the statistical strength and reproducibility. Methods originally developed for bulk tissue batch correction such as limma [71] and ComBat [72] have been applied in the batch correction of single-cell data [73–76]; however, there have been studies which demonstrate the limitations of applying these methods developed for bulk data to both simulated and real single-cell data [77]. Recently, significant progress has been made in this area, yielding methods developed specifically for single-cell batch correction such as canonical correlation analysis (CCA) [78] and mnnCorrect [77] and methods for cell-type identification based on a labeled reference dataset such as scmap [79] and singleR [80]. However, it is important to proceed to downstream analyses like GRN construction with caution after applying a batch correction method to single-

Emerging Topics in Life Sciences (2019) **3** 379–398
https://doi.org/10.1042/ETLS20180176

cell data, and it is necessary to understand the underlying algorithms and assumptions. Methods like CCA and mnnCorrect only leverage the highly variable genes which are shared across datasets for integration and return a corrected gene expression matrix which only contains the variable genes used for integration. These genes primarily define cell-type-specific markers, and the process of CCA inherently introduces dependencies between genes and violates the assumptions of statistical tests used for downstream analysis like differential expression, so the authors of CCA caution against using CCA for more than conserved cell-type identification across datasets. Broadly speaking, batch correction methods developed for bulk data perform more poorly in batch correction, whereas methods developed for single-cell data are more accurate at clustering cell types from different batches but may not be extended to downstream analyses. Therefore, there is a need for the development of methods which can do both.

Lastly, compared with bulk tissue data which usually consists of experimental group IDs, sample IDs, and feature measurements, single-cell data also present increased dimensionality and data volume by adding tens of cell types and thousands of cells from each sample. Such increases in dimensionality and data volume not only make network modeling more complex and computationally expensive, but also bring new possibilities from the biological perspective that are beyond the capacity of existing methods. In addition to the typical question of how genes are organized and interacting in a network, one can address many new provocative questions. For example, what defines a cell type? How are genes organized in each cell type? How different are the network architectures between cell types? What are the relationships between cells — do they come from the same or different lineages and how do the lineages evolve? Are there different states of the same cell type? What gene regulatory circuitry determines a cell state? How do cells transition from one state to another? Which cells communicate with one another to determine higher-level functions, and through which genes and pathways do they communicate? Many of these new questions are not considered or readily addressable by the existing methods designed for bulk tissues. In addition to offering the opportunities to answer these important questions, the cell–cell variability or heterogeneity in the large numbers of cells measured in each sample also provides sufficient information to construct within-sample or profile-specific networks [55]. Such networks describe the GRN of a single biological sample, which is not possible for bulk tissue profiling data. In other words, the ability to exploit the large cell number dimension allows GRNs to be constructed for each sample based on its constituent cell profiles, which can be used to derive consensus networks across samples to enhance accuracy.

Recent methods for scRNAseq GRN modeling

Recognizing the need for new GRN modeling methods for single-cell data, many approaches have been recently developed primarily based on scRNAseq data. We categorize these methods based on fundamental biological questions (dynamic modeling, within-cell networks, and cell–cell interaction networks; Figure 1B), followed by the specific biological assumptions (e.g. TF to target interactions, ligand–receptor interactions) and algorithms (e.g. coexpression, regression, ODEs, Bayesian, and Boolean), as summarized in Table 1.

The most straightforward algorithm is coexpression in which the likelihood of a gene interacting with another depends on the strength of their pairwise correlation coefficients. Though computationally tractable, most of these methods do not provide directionality and likely infer functional relatedness rather than direct regulation. More complex methods include ODEs, Boolean networks, and BNs, each with advantages and limitations, as discussed earlier. Boolean networks require discretization of gene expression values and apply Boolean functions to describe regulatory interactions, which likely result in oversimplification. ODE-based methods involve using linear, nonlinear, or piecewise differential equations to model the dynamic nature of mRNA content in a continuous, rather than discrete, manner. A BN is a directed acyclic graph (DAG) that integrates prior information to guide its gene–gene interaction predictions and is probabilistic in nature. Lastly, information theory measures describe statistical dependencies between biological entities and include entropy, the notion that information is quantified based on the uncertainty of a random variable, and mutual information, in which the observation of one random variable can inform on or reduce the uncertainty of another random variable. This measure produces more general correlations that allow capturing of nonlinear dependencies and is commonly employed in network inference.

Of note, with new methodologies being developed at rapid speed, it is not possible to exhaustively document all available methodologies. Here, we highlight the broad categories for single-cell GRN modeling and discuss example methods to illustrate the concepts and make note of their advantages and potential limitations. We

Emerging Topics in Life Sciences (2019) **3** 379–398
https://doi.org/10.1042/ETLS20180176

Table 1 Summary of single-cell network modeling approaches

Part 1 of 3

Category	Example methods	Underlying biological assumption	Algorithmic basis	Advantages	Limitations
Dynamic network **(extensively reviewed in refs [53–55])**	SCNS [81]	Single-gene changes between cell transition states can inform on gene regulatory relations	Boolean	Does not rely on prior knowledge. Has a web UI. Resulting models are executable and can be used to make predictions	Need data discretization; limit to small numbers of genes; regulatory relations need to follow Boolean rules
	SCODE [82]	TF expression dynamics (pseudo-time) and TF regulatory relations (GENIE3)	ODE; Bayesian model selection	Estimate relational expression efficiently using linear regression; reduction of time complexity; fast algorithm	Need dimension reduction first for computing speed and memory feasibility; assumes that all cells are on the same trajectory; optimization is computationally intractable
	GRISLI [83]	Variability in scRNAseq data caused by cell cycle, states, etc. allows the inference of pseudo-time associated with each individual cell	ODE	Makes no restrictive assumption on the gene network structure; can consider multiple trajectories; fast algorithm	Has to estimate the velocity of each individual cell using information from neighbors
	SINCERITIES [84]	Changes in the expression of a TF will alter the expression of target genes	Ridge regression and partial correlation analysis	Low computational complexity and able to handle large-scale data	Requires scRNAseq data at multiple time points. Restricted to TFs and their targets to infer edges
	Scribe [85]	Cell ordering can be improved with time-series or cell velocity estimations	RDI	Outperforms other pseudo-time methods given time-series data. Can be applied to any data type if the data structure is appropriate	Requires time-ordered gene expression profiles or velocity estimation from introns and exons
	AR1MA1-VBEM [40]	The cell differentiation process or response to external stimulus reveals the hierarchical structure of the transcriptome	First-order autoregressive moving-average and variational Bayesian expectation-maximization	Weighted interactions between genes along psuedotime. Model used accounts for noisy data	Data are expressed as fold changes between timepoints/ conditions or scaled by housekeeping genes
	SCINGE [86]	Learned target regulator genes can be used to assign each cell to their progress along a trajectory	Granger causality	Smooths irregular pseudo-times and missing expression values	Near random performance for predicting targets of individual regulators
	SopitSC [87]	Similarities between whole transcriptomes of single cells can be used to order them	Cells ordered by minimum paths on weighted cluster-to-cluster graph derived from cell similarity matrix	Includes comprehensive single-cell workflow; leverages information from other parts of the workflow to improve performance	Cannot be run with other tools, have run the full workflow to get pseudo-time inference
Within-cell or cell population network	SCENIC [88]	TF target-based regulation	Combining TF regulatory relations (GENIE3) with TF-binding motif analysis	Robust against dropouts, get a TF score for individual cells (no averaging of cells).	Limited to TF-based relations

Continued

PORTLAND PRESS

Emerging Topics in Life Sciences (2019) **3** 379–398
https://doi.org/10.1042/ETLS20180176

Table 1 Summary of single-cell network modeling approaches

Part 2 of 3

Category	Example methods	Underlying biological assumption	Algorithmic basis	Advantages	Limitations
	Pina et al. [89]	TFs drive lineage commitment	Odds ratio for on/off gene associations and spearmen correlation for expression levels associations	Robust to dropouts	Based on single-cell multiplex qRT-PCR, may be difficult to extend the method to sparse single-cell data (selected 44 genes to test)
	Iacono et al. [90]	Coexpression is regulated by TFs, cofactors, and signaling molecules which can be captured with gene–gene correlations	Pearson correlation using *z*-score-transformed counts	Can compute correlations at the single-cell level and it is robust to dropouts and noise inherent to single-cell data	Networks are very dense (some have millions of significant edges)
	PIDC [39,91]	Gene regulatory information reflected in dependencies in the expression patterns of genes	Partial information decomposition using gene trios	Compared with correlation, captures more complicated gene dependencies	Networks are influenced by data discretization, choice of mutual information estimator, method developed for sc-qPCR data, may not be extendable to higher throughput and sparser scRNAseq data
	Jackson et al. [92]	Deletion of TFs combined with experimental conditions allows for the inference of gene relationships	MTL to leverage cross-dataset commonalities and incorporate prior knowledge	Does not require sophisticated normalization of single-cell data or imputation. Able to combine multiple conditions/datasets for more accurate inference. TF deletions give strong causal link to affected genes	Requires single-cell data with TF deletions and/or environmental perturbations
	Wang et al. [93]	Gene perturbations allow for inference of causal relationships	Scoring of conditional independence test to identify optimal DAG	Gives causal relationships between genes	Requires interventional data. No loops allowed in DAG
	ACTION [94]	Functional identity of cells is determined by a weak, but specifically expressed set of genes which are mediated by TFs	Kernel-based cell similarity and geometric approach to identify primary functions	Robust to dropout and does not require averaging. Identifies functions unique to cell types	Requires TFs and their targets. Only provides TF-driven networks
	SINCERA [95]	TF target-based regulation	First-order conditional dependence on gene expression to construct a DAG	Key TFs identified using multiple importance metrics	Only considers TFs and their targets. Requires genes/TFs to be DEGs or expressed in >80% of cells

Continued

Emerging Topics in Life Sciences (2019) **3** 379–398
https://doi.org/10.1042/ETLS20180176

Table 1 Summary of single-cell network modeling approaches

Part 3 of 3

Category	Example methods	Underlying biological assumption	Algorithmic basis	Advantages	Limitations
Cell–cell communication network	iTALK [96]	Ligand–receptor interactions	Threshold ranked list of genes from two cell types for ligand–receptor pairs	Allows for the inference of directionality of interaction	Requires curation of ligand–receptor interactions (not all interactions are known). Average expression at the cell-type level (no longer single cell). Cannot reveal novel interactions beyond known ligand–receptor knowledge
	Zhou et al. [97]	Ligand–receptor interactions	Expression of ligand and corresponding receptor more than three standard deviations greater than the mean	Allows for the inference of directionality of interaction	Requires curation of ligand–receptor interactions (not all interactions are known). Average expression at the cell-type level (no longer single cell)
	Kumar et al. [98]	Ligand–receptor interactions	Product of the average expression of ligand and corresponding receptor	Allows for the inference of directionality of interaction. Interaction score gives the strength of interaction (rather than just significance)	Requires curation of ligand–receptor interactions (not all interactions are known). Average expression at the cell-type level (no longer single cell)
	Arneson et al. [99]	Ligand to downstream signaling	Coexpression of ligand genes in source cells with other genes in target cells	Use secreted ligands as a guidance for directional inference between cell populations	Gene expression is summarized to the cell population level and coexpression is at the sample level, requiring large sample sizes
	SoptSC [87]	Ligand–receptor interactions	Likelihood estimate of the interaction between two cells based on expression of the ligand, receptor, and downstream pathway target genes (including expression direction). Consensus signaling network derived from all cells in each cluster	Incorporates target genes of pathways and their directionality. Computes interaction likelihood at the single-cell level and summarizes across all cells in the cluster for higher confidence	Requires curation of ligand–receptor interactions and their downstream pathways
	scTensor [100]	Ligand–receptor interactions	Tensor decomposition with cell–cell interactions as hypergraphs	Allows L–R pairs to function across multiple cell-type pairs (not restricted to a single-cell-type pair), which is more reflective of underlying biology	Requires curation of ligand–receptor interactions. Averages single cells to the cell-type level

Emerging Topics in Life Sciences (2019) **3** 379–398
https://doi.org/10.1042/ETLS20180176

also exclude methods that were developed based on data from older low-throughput single-cell platforms such as single-cell qPCR, which do not share the same challenges as sparse high-throughput scRNAseq.

Dynamic networks

To date, the majority of the scRNAseq-based GRN modeling approaches were designed to address dynamic cell-state transition (Figure 1B), as scRNAseq data contain information from asynchronous cell populations which show temporal dynamics, allowing for the mapping of cellular transitions on a pseudo-time scale [101,102]. Common models for expression dynamics or pseudo-time estimates assume that cellular changes (i.e. development, activation, and deactivation) progress along a continuous curve or an idealized tree and that each intermediate stage is short and captured through the sequencing of large numbers of cells. Under these assumptions, computational modeling can infer the trajectories of cellular dynamics, which can be derived based on known regulatory relations such as TF target information, similarities in gene expression, and RNA velocity represented by immature and mature mRNA content [60]. However, it is important to note that the simultaneous presence of various cellular states at a given snapshot does not represent real time course for the inference of sequential or lineage information. Therefore, incorporating pseudo-time may not necessarily improve GRN construction.

To date, more than 50 methods have been developed for trajectory inference to derive pseudo-time information, and these have been reviewed and compared previously [101,102]. Pseudo-time ordering lends directionality and interaction-type information for dynamic GRN modeling [40,82,83,85,86,103,104]. Such pseudo-time information is integrated with commonly used network construction algorithms outlined above such as correlation [88,94,95], ODE [82,83,103,104], Boolean [105,106], BN [39,40], information theory [91], and other methods [107].

Many of the dynamic GRN methods have been extensively reviewed by others [53–55], and we only discuss a few examples of the different categories here. A Boolean network method, SCNS, is based on single-gene changes between ordered cells where cells have been discretized into an on/off state [81]. Another method SCODE uses a linear ODE, a pseudo-time estimation that assumes all cells are on the same trajectory, and a TF-based framework to model TF dynamics that captures regulatory relationship between genes [82]. Building on this, GRISLI was recently developed, which uses a similar approach to SCODE, but considers multiple cell trajectories, does not assume a network structure, and has faster computation times [83]. GRISLI first estimates the velocity of each cell, followed by solving a sparse regression problem to relate the gene expression of a cell to its velocity profile to estimate the GRN. An information theory-based method, SINCERITIES, utilizes Granger Causality for directionality information and quantifies temporal changes in the expression of each gene between two subsequent (pseudo)timepoints [82]. Changes in TF expression are used to predict changes in corresponding genes in the next time window using ridge regression, with edge direction and sign inferred using partial correlation analysis on the expression of every gene pair. SCINGE also uses kernel-based Granger Causality regression on ordered single-cell data to predict regulator-target gene interactions and then ranks the predicted interactions by aggregating the regression results [86]. An additional method is PIPER [107], which uses local Poisson graphical modeling to more effectively capture network changes during cellular differentiation and highlight the key TFs that drive these changes. A BN inference approach, AR1MA1-VBEM (Variational Bayesian Expectation-Maximization), applies a first-order autoregressive moving-average (AR1MA1) model to fit the time-series with a linear model that represents observations as combinations of the data at the previous timepoint and a noise term, and uses a VBEM framework that utilizes variational calculus to optimize the marginal likelihood and posterior distributions of network models [40]. Scribe [85] is another recently developed method, which uses restricted directed information (RDI) [108] to infer causal GRNs by borrowing linked time-series data or inferred cell velocity from intronic (indicative of immature RNA) and exonic reads. The authors demonstrate that Scribe outperforms other pseudo-time methods when true time-series data are available; however, the performance of all methods suffers dramatically when temporal information for the measurements is lost. Interestingly, Deshpande et al. [86] recently compared various methods and found that incorporating pseudo-time does not always lead to better performance but can hurt network reconstruction in certain cases. As discussed earlier, this is likely due to issues in the assumptions of pseudo-time methods.

Within-cell population networks

The second category of methods focuses on modeling GRNs within-cell populations without considering cell trajectories or dynamics. These methods include coexpression and TF-based [88,94,95], coexpression and

TF-independent [89,90,109], and information theory [91] (Table 1 and Figure 1B). This is in line with the basic concepts underlying GRN modeling of gene–gene interactions for a tissue, except here single-cell data are modeled for specific cell populations.

Similar to dynamic network modeling, the simplest approach for modeling GRNs within-cell populations is based on coexpression. Here, the coexpression methods are divided into two groups: those which utilize prior information in the form of TFs and those that do not. For the methods which are TF-independent, the likelihood of a gene interacting with another depends on the strength of their pairwise correlation coefficients and all possible gene pairs are considered. In the TF-based methods, genes are grouped into modules based on those with the strongest pairwise correlation coefficients with the different TFs or are segregated into potential interactions based on prior literature or motif evidence. A more sophisticated approach for defining GRNs within-cell populations, which can capture nonlinear gene dependencies, is partial information decomposition which is derived from information theory. Here, the information provided by pairs of genes is used to quantify unique, shared, and synergistic information about a third gene across all sets of three genes to infer a network structure.

Several correlation-based methods have been developed that compare the gene expression patterns between known or predicted TFs and target genes, or between all genes. For instance, SCENIC couples gene coexpression with TF-binding motif analyses of modules of coexpressed genes to identify GRN modules, predict TF regulators, and identify single-cell level activity of putative TF targets (called regulons) [88]. The activity of the regulons can be used to cluster cell types, compare network conservation, and identify important cell states and GRNs involved in disease. Another method SINCERA is a full analytical pipeline for processing scRNAseq data. It first identifies candidate TFs and their targets for each cell type [95]. The interactions between two TFs or a TF and a target gene are then determined using first-order conditional dependence on gene expression [110], and the key TFs in each GRN are identified by integrating six different node importance metrics. An additional coexpression-based GRN method, ACTION, uses a novel archetype orthogonalization approach to construct cell-type-specific GRNs based on the key assumption that the functional identity of a cell is determined by a set of weak, but specifically expressed genes which are mediated by a set of TFs [94]. ACTION describes each cell as a set of 'cellular functions' in high dimensional space and the number of these functions is determined using a non-parametric approach. The genes which are unique to each cellular function are determined using orthogonalization and the role of TFs in controlling the genes in these cellular functions is assessed. The TF and associated target genes within a cellular function serve to constitute the network.

Pina et al. [89] and, more recently, Iacono et al. [90] also utilize coexpression but build global GRNs that are not limited to TF target relations [89,90]. The former calculates pairwise Spearman rank correlations between all sets of genes across cells within a cell type to infer cell-type GRNs in hematopoiesis, and significant pairwise associations were identified using the odds ratio of linearly transformed expression data. Iacono et al. [109] used a Pearson correlation-based method which first transforms the expression values using bigSCale to derive a z-score for each gene using a probabilistic model to account for noise and variability inherent to single-cell data. Pairwise correlations of z-scores are used to construct GRNs. The use of z-scores boosts the number of significant gene-to-gene correlations.

To reveal complex gene dependencies not afforded by simpler correlation strategies, GRN inference methods have employed techniques from information theory. Specifically, PIDC uses partial information decomposition to find the unique information provided by any pair of two genes across all other possible genes [80]. The confidence of an edge between two genes is the sum of the scores of those genes across all other genes in the set. This multivariate information approach makes use of the large sample size present in single-cell analyses to identify nonlinear dependencies between pairs of genes by leveraging a third gene.

Cell–cell communication networks

The basic functions of a given heterogeneous tissue are determined not only by the activities of individual cell types within the tissue, but also by the intimate communications and coordination among cell populations. For instance, neurons and astrocytes interact to ensure essential brain functions [111], and immune cells interact with adipocytes in the adipose tissue to regulate energy metabolism and thermogenesis [112]. As such, cell–cell communication is a critical biological question yet has not been comprehensively addressed due to the previous lack of high-throughput, high-resolution singe-cell data. The unique ability of single-cell approaches to simultaneously capture numerous cells of diverse cell types makes it feasible to model cell–cell communication networks (Table 1 and Figure 1B). The underlying assumption for modeling such networks is that

Emerging Topics in Life Sciences (2019) **3** 379–398
https://doi.org/10.1042/ETLS20180176

communications between cells can be captured by molecular patterns measured in individual cell populations. For example, a pair of communicating cells may express genes and proteins involved in a particular function (e.g. one expressing a ligand and another expressing the corresponding receptor to trigger signaling pathways) in a coordinated fashion.

Early attempts to model cell–cell communication networks have been primarily based on the concept of gene coexpression with or without the consideration of ligand–receptor interaction information. The underlying assumption is that gene correlation patterns between cells reflect true biological interactions. The validity of this assumption has been supported by evidence at the level of tissue–tissue interactions. For instance, gene coexpression between brain regions can recapitulate the functionally derived interactions of the mouse brain connectome [113], and gene coexpression between five different mouse tissues revealed novel endocrine factors mediating the communications which are subsequently validated with experiments [30].

Coexpression methods were quickly adapted to single-cell data when Han et al. built cell–cell connections based on the similarities in gene expression profiles across cell types [68]. However, such networks more likely reflect the similarities between cell types rather than interactions or communications. To modify the classical coexpression framework, ligand–receptor-based methods have been proposed which rely on the assumption that a significant portion of cell–cell communication occurs via the release of chemical molecules from one cell that bind to receptors of another cell. Utilizing this assumption allows ligand–receptor-based methods to construct reliable biology-based directed networks. However, it comes at the expense of heavily limiting the set of potential genes in an inherently sparse data modality. It is important to note that coexpression-based analyses typically utilize Pearson's correlation coefficient, which may not be suitable for read-based single-cell datasets due to the zero-inflated nature and unique distribution patterns. When using coexpression-based analysis on single-cell data, it is important that data transformation and appropriate statistics are taken into consideration.

There are several methods illustrating cell–cell communication via ligand–receptor interactions. Zhou et al. [97] curated a list of >25000 known ligand–receptor pairs to examine their changes in the transcriptomes of ∼4000 melanoma cells. To determine if a pair of cells were communicating, the ligand and corresponding receptor had to be expressed above a certain tunable threshold in the two cell types. Similarly, Kumar et al. [98] focused on ∼1800 literature-based ligand–receptor pairs, but implemented a different scoring scheme that considers the product of average receptor expression and average ligand expression in the respective cell types under examination. Ported as an R package with a data visualization tool, iTALK is another new ligand–receptor interaction-based network construction method [96]. iTALK identifies ligand–receptor pairs (from a database of >2600 pairs) between two cell types by interrogating the list of ranked genes derived from average expression (single timepoint/condition) or differentially expressed genes (multiple timepoints/conditions) for each cell type and the list of ligand–receptor pairs in the iTALK database. Additionally, iTALK is able to use metadata (e.g. timepoints, groups, and cohorts) to find cell–cell interaction changes by identifying differentially expressed ligand–receptor pairs. Similarly, Smillie et al. [114] have used thousands of literature-supported receptor–ligand interactions from the FANTOM5 database to identify cell–cell interactions by requiring that genes are cell marker genes or differentially expressed genes to call significant interactions between cells. In most ligand–receptor approaches, ligand–receptor pairs are restricted to communicating cell types; however, in scTensor, Tsuyuzaki et al. [100] took a more flexible approach where no such restrictions are made. In scTensor, cell–cell interactions are represented as hypergraphs which describe directed edges of ligand–receptor pairs determined using tensor decomposition. A recent method proposed by Vento-Tormo et al. [115] also considers secreted molecules as well as cell-surface molecules and uses a permutation-based approach to find enriched ligand–receptor pairs between cell types. To achieve this, the authors developed CellPhoneDB, a public repository of ligand–receptor interactions curated from public resources of protein–protein interactions, which includes the subunit composition of ligands and receptors to fully represent their interactions. For proteins which are comprised of multiple subunits, expression of all subunits is required to infer accurate interactions.

The above methods all focus exclusively on ligand–receptor pairs which heavily limit the putative genes to sets of literature-curated gene pairs which can inform on cell–cell communication. Previously, a less restrictive modeling approach that dissects tissue–tissue communication networks [30] based on the coexpression of genes encoding secreted peptides from a source tissue and all genes in a target tissue has been developed. Arneson et al. [99] adopted this concept to build cell–cell communication network maps in the hippocampus of sham mice versus mice with traumatic brain injury, revealing extensive rewiring of networks in brain injury. This method infers connections between cells based on the assumption that one cell communicates with

another cell by secreting signaling molecules that bind to their receptors on the target cell to trigger downstream molecular events in the target cell. As such, it is likely that coexpression exists between the genes that encode secreted signaling molecules (i.e. the ligands) in the source cell type and the receptors as well as the downstream pathway genes in the target cell type. Additional methods can broaden the scope of cell–cell interactions beyond ligand–receptor-based relationships by considering the patterns among all expressed genes between cell types, although the biological interpretation of this approach is less straightforward.

Hybrid methods

Although most GRN methods tackle either dynamic or within-cell or between-cell networks, Wang et al. [87] have proposed SoptSC, a unifying framework to conduct single-cell analysis starting from gene expression matrices to basic analytical workflows (e.g. normalization, clustering, dimensionality reduction, and identifying cell marker genes) and subsequently to infer both cell–cell communication networks and pseudotemporal ordering/lineage. The key premise underlying SoptSC is that the structured cell-to-cell similarity matrix can help improve the network inference steps. The similarity matrix is also used for pseudotemporal ordering by finding the shortest path between cells on a weighted cluster-to-cluster graph. To infer cell–cell signaling networks, the likelihood estimate of the interaction between two cells is calculated based on the expression of ligand–receptor pairs and the directionality of downstream pathway target genes. A consensus network between clusters/cell types is generated by summarizing the probability of signaling between all cells of any two cell types.

Gene perturbation networks

All of the above methods utilize assumptions regarding information flow such as TF cascades and ligand–receptor relationships, without direct causal information. Single-cell data containing gene perturbation information are extremely useful for providing causal information for GRN construction, as targeted perturbation of a gene is the source or trigger of downstream responses of other genes. A method proposed by Jackson et al. [92] leverages gene deletion mutants. Specifically, they pooled 72 different yeast strains across 12 different genotypes (TF deletions) and 11 different conditions to generate scRNAseq data for 38 000 cells. In addition to the expression data, this method uses prior information from TF targets and biophysical parameters like TF activity and mRNA decay rates to construct a GRN using a multitask learning (MTL) framework [116]. This allows for the integration of information across different conditions and experiments that explains the relationships between the TF perturbations and the observed gene expression changes. By directly deleting TFs, the authors have created a valuable dataset which could serve as a useful benchmark for other single-cell network inference methods. Leveraging single-cell data from Perturb-seq [117], which combines CRISPR/Cas9-mediated gene perturbation with single-cell sequencing to generate high-throughput interventional gene expression data, Wang et al. [93] proposed an algorithm for inferring causal DAGs. The algorithm is based on Greedy SP which restricts the permutation-based DAG search space, and potential network scores are evaluated using the Greedy Interventional Equivalence Search [118]. To further extend this research on causal network inference, Wang et al. [119] introduced a method which could identify differences between DAGs inferred from different datasets. The same group has also demonstrated that soft interventions used in Perturb-seq, such as those that cause partial disruption of gene dependencies (e.g. RNAi or CRISPR-mediated gene activation), provide the same amount of causal information as hard interventions (e.g. CRISPR/Cas9-mediated gene deletions), which result in complete disruptions, despite being less invasive [120].

Performance assessment of single-cell GRN modeling methods

Chen and Mar recently applied a few single-cell network modeling methods including SCENIC [88], SCODE [82], and PIDC [39,91] to both simulated and empirical single-cell datasets to assess their ability to capture known network interactions. They found that there was low agreement between methods. However, as each method has unique assumptions and may not be designed to capture similar interactions, agreement between methods is not necessarily appropriate to assess performance. Another comparison study that examined the performance of multiple network inference methods that incorporate pseudo-time information, such as SCINGE, SCODE, and SINCERITIES, also indicates that many regulator-target predictions can be near random for each of the methods tested [86].

Emerging Topics in Life Sciences (2019) **3** 379–398
https://doi.org/10.1042/ETLS20180176

These findings call for the refinement of single-cell network modeling approaches and comprehensive revaluation of the performance of existing single-cell GRN methods. On the other hand, the ligand–receptor framework that is driven by a biological assumption along with data-driven gene coexpression appears to be promising for cell–cell communication network modeling. For example, modeling with this approach to scRNAseq data recapitulated known cell–cell interactions within the hippocampus [99].

Remaining gaps and future directions

Single-cell multi-omics profiling technologies are rapidly evolving, bringing revolutionary forces to improve our understanding of the basic unit of life, the cell, as well as the cross-talks between cells in physiological and pathological conditions. Major progresses have been made to more accurately classify cell types, correct for confounding factors, and delineate cell lineages and cell-state transitions. However, these advances are not sufficient to bring a complete understanding of the regulatory machinery underlying the functions of individual cell populations and cell–cell interactions that determine higher-level tissue functions. Existing methodologies to model gene networks that were optimized for bulk tissue data either perform poorly for single-cell data or cannot accommodate the new biological questions brought about by single-cell data, and methods that efficiently and accurately model the outpour of single-cell data into comprehensive GRN maps are still in infancy. In particular, novel network methods that are designed to address the unique challenges of single-cell data, such as data sparsity, multimodal distribution, and higher dimensionality, are still in great need. The data sparsity issue can be addressed through the improvement of single-cell technologies to enhance signal capture, or by more accurate imputation methods that are supported by strong experimental validation data. These efforts will help mitigate the issues associated with nonstandard data distribution that limits the utilization of existing network methodologies. Alternatively, methods built on more appropriate statistics and algorithms that can better accommodate dropout values and the unique data distribution are warranted.

Another critical and less highlighted gap in network modeling of single-cell data is the missing spatial information to restrain the modeling space. Many of the current high-throughput single-cell sequencing methods lack the ability to maintain the spatial identity of individual cells, which reduces one's ability to resolve cell networks accurately, particularly during development where development layers are in close proximity. Various high-throughput fluorescence in situ hybridization (FISH) methods have been developed as tools to resolve spatial information [121–128]. The spatial distances between pairs of single cells can be used as a prior to construct more sophisticated and accurate network models under the assumption that cells which are located closer together are more likely to communicate. This assumption is supported by the recent discovery of localization of ligand-producing cells directly adjacent to target cells expressing the corresponding receptor [129]. Another key advantage of single-molecule FISH-based methods is that they are extremely quantitative and do not suffer from dropouts which plague high-throughput single-cell sequencing-based methods. The absence of dropouts allows for accurate single-cell level interrogation of network predictions. With the spatial single-cell methods, it is also possible to combine phenotypes (i.e. behavior) with cellular activation (i.e. cFos) to integrate into the model under the assumption that cells which are active during a particular phenotype or stimulus are more likely to be communicating. This approach has been previously used by Moffitt et al. [130] to identify sets of neurons activated during parenting. Therefore, coupling single-cell sequencing approaches with high-throughput single-molecule imaging has enormous potential to improve network modeling at single-cell resolution. Despite the potential, there are limitations and complications involved in using spatial data to construct GRNs. First, cell segmentation of single-molecule FISH-based methods is non-trivial and GRN construction is impossible without it. Additionally, a single image carries limited representation of the dynamic cellular landscape. In fact, many of these technologies can only achieve the imaging depth of a single cell, so it is essentially a two-dimensional snapshot at a given time which may not capture cellular dynamics outside of the imaged plane and time frame.

At present, the majority of the methods are designed for scRNAseq, and methods incorporating other single-cell omics scales (genetic, epigenetic, and protein) are needed [55]. This faces the same challenge that has been encountered by bulk tissue GRN inference, and recent progresses in multi-omics integration and modeling may offer guidance for single-cell multi-omics modeling [131–134].

Lastly, the accuracy of predicted networks from empirical data is difficult to assess, as high-throughput validation through perturbing predicted regulators in single cells *in vivo* is more challenging than that for whole-body knockout or knockdown. On a positive note, new high-throughput gene perturbation technologies such as Perturb-seq in combination with scRNAseq have the potential to generate insight into true casual

relationships between genes and cells. Data from such platforms can serve as more appropriate benchmarking datasets to assess the predictions of existing network methods by testing how well each method can retrieve the true regulatory or interactive relations known from the perturbation-response experiments. Along the same line, use of known, experimentally validated gene–gene, cell–cell circuits from the literature can be used to benchmark the methods. Even in the absence of validated network connections, a community-based approach can be employed to improve the network performance *in silico* by combining multiple inferred networks from various methods to obtain consensus networks. This approach has been shown to be invaluable for improving the quality of the predicted networks [27,91,135,136].

In summary, we are entering a golden era where biological discoveries can be made at an unprecedented resolution and throughput. Network modeling of single-cell multi-omics data represents one of the key tools to unlock the convoluted molecular mechanisms underlying pathophysiology and guide precision medicine. Despite numerous challenges, the field is rapidly evolving and ample opportunities for methodological innovations await to more accurately depict the molecular maps of cells in health and disease.

Summary

- Single-cell omics data offer unique challenges and opportunities for molecular network modeling.
- Significant progress has been made to dissect the dynamic, within-cell, and between-cell gene regulatory networks.
- Performance of current methods await further evaluation.
- Significant gaps remain in the development of network modeling approaches that can accommodate unique statistical challenges, diverse omics domains, and spatial information.

Abbreviations

AR1MA1, autoregressive moving-average; BN, Bayesian network; CCA, canonical correlation analysis; DAG, directed acyclic graph; DGE, digital gene expression; FISH, fluorescence in situ hybridization; GIES, Greedy Interventional Equivalence Search; GRNs, gene/transcriptional regulatory networks; MTL, multitask learning; ODE, ordinary differential equation; RDI, restricted directed information; TFs, transcription factors; UMI, unique molecular identifier; VBEM, Variational Bayesian Expectation-Maximization; WGCNA, weighted gene coexpression network analysis; ZIFA, zero-inflated factor analysis.

Author Contribution

M.B., D.A., J.D., Y.-W.C., Z.S., and X.Y. drafted and edited the manuscript.

Funding

D.A. is funded by the UCLA Dissertation Year Fellowship. X.Y. is funded by the National Institutes of Health Grants DK104363 and NS103088. Y.-W.C. is supported by the UCLA Hyde Fellowship.

Competing Interests

The Authors declare that there are no competing interests associated with the manuscript.

References

1 Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J. et al. (2014) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 https://doi.org/10.1093/nar/gku1003

2 Terzer, M., Maynard, N.D., Covert, M.W. and Stelling, J. (2009) Genome-scale metabolic networks. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **1**, 285–297 https://doi.org/10.1002/wsbm.37

3 Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G. et al. (2016) A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, aaf1420 https://doi.org/10.1126/science.aaf1420

Emerging Topics in Life Sciences (2019) **3** 379–398
https://doi.org/10.1042/ETLS20180176

4 Vlaic, S., Conrad, T., Tokarski-Schnelle, C., Gustafsson, M., Dahmen, U., Guthke, R. et al. (2018) Modulediscoverer: identification of regulatory modules in protein–protein interaction networks. *Sci. Rep.* **8**, 433 https://doi.org/10.1038/s41598-017-18370-2
5 Thurley, K., Wu, L.F. and Altschuler, S.J. (2018) Modeling cell-to-cell communication networks using response-time distributions. *Cell Syst.* **6**, 355–367.e5 https://doi.org/10.1016/j.cels.2018.01.016
6 Boyle, E.A., Li, Y.I. and Pritchard, J.K. (2017) An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 https://doi.org/10.1016/j.cell.2017.05.038
7 Costanzo, M., Kuzmin, E., van Leeuwen, J., Mair, B., Moffat, J., Boone, C. et al. (2019) Global genetic networks and the genotype-to-phenotype relationship. *Cell* **177**, 85–100 https://doi.org/10.1016/j.cell.2019.01.033
8 Mäkinen, V.-P., Civelek, M., Meng, Q., Zhang, B., Zhu, J., Levian, C. et al. (2014) Integrative genomics reveals novel molecular pathways and gene networks for coronary artery disease. *PLoS Genet.* **10**, e1004502 https://doi.org/10.1371/journal.pgen.1004502
9 Shu, L., Zhao, Y., Kurt, Z., Byars, S.G., Tukiainen, T., Kettunen, J. et al. (2016) Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems. *BMC Genomics* **17**, 874 https://doi.org/10.1186/s12864-016-3198-9
10 Shu, L., Chan, K.H.K., Zhang, G., Huan, T., Kurt, Z., Zhao, Y. et al. (2017) Shared genetic regulatory networks for cardiovascular disease and type 2 diabetes in multiple populations of diverse ethnicities in the United States. *PLoS Genet.* **13**, e1007040 https://doi.org/10.1371/journal.pgen.1007040
11 Kurt, Z., Barrere-Cain, R., LaGuardia, J., Mehrabian, M., Pan, C., Hui, S.T. et al. (2018) Tissue-specific pathways and networks underlying sexual dimorphism in non-alcoholic fatty liver disease. *Biol. Sex Differ.* **9**, 46 https://doi.org/10.1186/s13293-018-0205-7
12 Zhao, Y., Blencowe, M., Shi, X., Shu, L., Levian, C., Ahn, I.S. et al. (2019) Integrative genomics analysis unravels tissue-specific pathways, networks, and key regulators of blood pressure regulation. *Front. Cardiovasc. Med.* **6**, 21 https://doi.org/10.3389/fcvm.2019.00021
13 Zhao, Y., Jhamb, D., Shu, L., Arneson, D., Rajpal, D.K. and Yang, X. (2019) Multi-omics integration reveals molecular networks and regulators of psoriasis. *BMC Syst. Biol.* **13**, 8 https://doi.org/10.1186/s12918-018-0671-x
14 Calabrese, G.M., Mesner, L.D., Stains, J.P., Tommasini, S.M., Horowitz, M.C., Rosen, C.J. et al. (2017) Integrating GWAS and co-expression network data identifies bone mineral density genes SPTBN1 and MARK3 and an osteoblast functional module. *Cell Syst.* **4**, 46–59.e4 https://doi.org/10.1016/j.cels.2016.10.014
15 Carlin, D.E., Demchak, B., Pratt, D., Sage, E. and Ideker, T. (2017) Network propagation in the cytoscape cyberinfrastructure. *PLoS Comput. Biol.* **13**, e1005598 https://doi.org/10.1371/journal.pcbi.1005598
16 Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S. et al. (2015) Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47**, 569–576 https://doi.org/10.1038/ng.3259
17 Krishnan, A., Zhang, R., Yao, V., Theesfeld, C.L., Wong, A.K., Tadych, A. et al. (2016) Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat. Neurosci.* **19**, 1454–1462 https://doi.org/10.1038/nn.4353
18 Yoon, S., Nguyen, H.C.T., Yoo, Y.J., Kim, J., Baik, B., Kim, S. et al. (2018) Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2. *Nucleic Acids Res.* **46**, e60 https://doi.org/10.1093/nar/gky175
19 Hui, S.T., Kurt, Z., Tuominen, I., Norheim, F., C Davies, R., Pan, C. et al. (2018) The genetic architecture of diet-induced hepatic fibrosis in mice. *Hepatology* **68**, 2182–2196 https://doi.org/10.1002/hep.30113
20 Chella Krishnan, K., Kurt, Z., Barrere-Cain, R., Sabir, S., Das, A., Floyd, R. et al. (2018) Integration of multi-omics data from mouse diversity panel highlights mitochondrial dysfunction in non-alcoholic fatty liver disease. *Cell Syst.* **6**, 103–115.e7 https://doi.org/10.1016/j.cels.2017.12.006
21 Meng, Q., Ying, Z., Noble, E., Zhao, Y., Agrawal, R., Mikhail, A. et al. (2016) Systems nutrigenomics reveals brain gene networks linking metabolic and brain disorders. *EBioMedicine* **7**, 157–166 https://doi.org/10.1016/j.ebiom.2016.04.008
22 Meng, Q., Zhuang, Y., Ying, Z., Agrawal, R., Yang, X. and Gomez-Pinilla, F. (2017) Traumatic brain injury induces genome-wide transcriptomic, methylomic, and network perturbations in brain and blood predicting neurological disorders. *EBioMedicine* **16**, 184–194 https://doi.org/10.1016/j.ebiom.2017.01.046
23 Shu, L., Meng, Q., Diamante, G., Tsai, B., Chen, Y.-W., Mikhail, A. et al. (2018) Prenatal bisphenol A exposure in mice induces multitissue multiomics disruptions linking to cardiometabolic disorders. *Endocrinology* **160**, 409–429 https://doi.org/10.1210/en.2018-00817
24 Bansal, M., Belcastro, V., Ambesi-Impiombato, A. and di Bernardo, D. (2007) How to infer gene networks from expression profiles. *Mol. Syst. Biol.* **3**, 78 https://doi.org/10.1038/msb4100158
25 Chai, L.E., Loh, S.K., Low, S.T., Mohamad, M.S., Deris, S. and Zakaria, Z. (2014) A review on the computational approaches for gene regulatory network construction. *Comput. Biol. Med.* **48**, 55–65 https://doi.org/10.1016/j.compbiomed.2014.02.011
26 Karlebach, G. and Shamir, R. (2008) Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.* **9**, 770–780 https://doi.org/10.1038/nrm2503
27 Marbach, D., Costello, J.C., Kuffner, R., Vega, N.M., Prill, R.J., Camacho, D.M. et al. (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804 https://doi.org/10.1038/nmeth.2016
28 De Smet, R. and Marchal, K. (2010) Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.* **8**, 717 https://doi.org/10.1038/nrmicro2419
29 Dobrin, R., Zhu, J., Molony, C., Argman, C., Parrish, M.L., Carlson, S. et al. (2009) Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome Biol.* **10**, R55 https://doi.org/10.1186/gb-2009-10-5-r55
30 Seldin, M.M., Koplev, S., Rajbhandari, P., Vergnes, L., Rosenberg, G.M., Meng, Y. et al. (2018) A strategy for discovery of endocrine interactions with application to whole-body metabolism. *Cell Metab.* **27**, 1138–1155.e6 https://doi.org/10.1016/j.cmet.2018.03.015
31 Talukdar, H.A., Foroughi Asl, H., Jain, R.K., Ermel, R., Ruusalepp, A., Franzen, O. et al. (2016) Cross-tissue regulatory gene networks in coronary artery disease. *Cell Syst.* **2**, 196–208 https://doi.org/10.1016/j.cels.2016.02.002
32 Malik, R., Selden, C. and Hodgson, H. (2002) The role of non-parenchymal cells in liver growth. *Semin. Cell Dev. Biol.* **13**, 425–431 https://doi.org/10.1016/S1084952102001301
33 Chen, S. and Mar, J.C. (2018) Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics* **19**, 232 https://doi.org/10.1186/s12859-018-2217-z
34 Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 https://doi.org/10.1186/1471-2105-9-559

35 Zhang, B. and Horvath, S. (2005) A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 https://doi.org/10.2202/1544-6115.1128
36 Huynh-Thu, V.A., Irrthum, A., Wehenkel, L. and Geurts, P. (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* **5**, e12776 https://doi.org/10.1371/journal.pone.0012776
37 Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R.D. et al. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7**, S7 https://doi.org/10.1186/1471-2105-7-S1-S7
38 Greenfield, A., Madar, A., Ostrer, H. and Bonneau, R. (2010) DREAM4: combining genetic and dynamic information to identify biological networks and dynamical models. *PLoS ONE* **5**, e13397 https://doi.org/10.1371/journal.pone.0013397
39 Filippi, S. and Holmes, C.C. (2017) A Bayesian nonparametric approach to testing for dependence between random variables. *Bayesian Anal.* **12**, 919–938 https://doi.org/10.1214/16-BA1027
40 Sanchez-Castillo, M., Blanco, D., Tienda-Luna, I.M., Carrion, M. and Huang, Y. (2017) A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics* **34**, 964–970 https://doi.org/10.1093/bioinformatics/btx605
41 Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G. et al. (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* **5**, e8 https://doi.org/10.1371/journal.pbio.0050008
42 Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M. et al. (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 https://doi.org/10.1016/j.cell.2015.05.002
43 Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V. et al. (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 https://doi.org/10.1016/j.cell.2015.04.044
44 Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I. et al. (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 https://doi.org/10.1126/science.1247651
45 Gierahn, T.M., Wadsworth, II, M.H., Hughes, T.K., Bryson, B.D., Butler, A., Satija, R. et al. (2017) Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 https://doi.org/10.1038/nmeth.4179
46 Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z. et al. (2018) Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 https://doi.org/10.1126/science.aam8999
47 Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R. et al. (2017) Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 https://doi.org/10.1126/science.aam8940
48 Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H. et al. (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 https://doi.org/10.1038/nmeth.4380
49 Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P. et al. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 https://doi.org/10.1038/nature14590
50 Karemaker, I.D. and Vermeulen, M. (2018) Single-cell DNA methylation profiling: technologies and biological applications. *Trends Biotechnol.* **36**, 952–965 https://doi.org/10.1016/j.tibtech.2018.04.002
51 Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M. et al. (2018) Comprehensive integration of single cell data. *bioRxiv* https://doi.org/10.1016/j.cell.2019.05.031
52 Scutari, M. (2010) Learning Bayesian networks with the bnlearn R package. *J. Stat. Softw.* **35**, 22 https://doi.org/10.18637/jss.v035.i03
53 Babtie, A.C. and Stumpf, M.P.H. (2017) How to deal with parameters for whole-cell modelling. *J. R. Soc. Interface* **14**, 20170237 https://doi.org/10.1098/rsif.2017.0237
54 Fiers, M., Minnoye, L., Aibar, S., Bravo Gonzalez-Blas, C., Kalender Atak, Z. and Aerts, S. (2018) Mapping gene regulatory networks from single-cell omics data. *Brief. Funct. Genomics* **17**, 246–254 https://doi.org/10.1093/bfgp/elx046
55 Todorov, H., Cannoodt, R., Saelens, W. and Saeys, Y. (2019) Network inference from single-cell transcriptomic data. *Methods Mol. Biol.* **1883**, 235–249 https://doi.org/10.1007/978-1-4939-8882-2_10
56 Svensson, V. (2019) Droplet scRNA-seq is not zero-inflated. *bioRxiv*, 582064 https://doi.org/10.1101/582064
57 Hafemeister, C. and Satija, R. (2019) Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *bioRxiv*, 576827 https://doi.org/10.1101/576827
58 Townes, F.W., Hicks, S.C., Aryee, M.J. and Irizarry, R.A. (2019) Feature selection and dimension reduction for single cell RNA-seq based on a multinomial model. *bioRxiv*, 574574 https://doi.org/10.1101/574574
59 van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J. et al. (2018) Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729.e27 https://doi.org/10.1016/j.cell.2018.05.061
60 Li, W.V. and Li, J.J. (2018) An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* **9**, 997 https://doi.org/10.1038/s41467-018-03405-7
61 Gong, W., Kwak, I.-Y., Pota, P., Koyano-Nakagawa, N. and Garry, D.J. (2018) Drimpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics* **19**, 220 https://doi.org/10.1186/s12859-018-2226-y
62 Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R. et al. (2018) SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* **15**, 539–542 https://doi.org/10.1038/s41592-018-0033-z
63 Prabhakaran, S., Azizi, E., Carr, A. and Pe'er, D. (2016) Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. *JMLR Workshop Conf. Proc.* **48**, 1070–1079 PMID:29928470
64 Zhu, L., Lei, J., Devlin, B. and Roeder, K. (2017) A unified statistical framework for single cell and bulk sequencing data. *bioRxiv*, 206532 https://doi.org/10.1101/206532
65 Zhang, L. and Zhang, S. (2018) PBLR: an accurate single cell RNA-seq data imputation tool considering cell heterogeneity and prior expression level of dropouts. *bioRxiv*, 379883 https://doi.org/10.1101/379883
66 Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X. and Garmire, L.X. (2018) Deepimpute: an accurate, fast and scalable deep neural network method to impute single-cell RNA-seq data. *bioRxiv*, 353607 https://doi.org/10.1101/353607
67 Zhang, L. and Zhang, S. (2018) Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 1 https://doi.org/10.1109/TCBB.2018.2848633

68 Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S. et al. (2018) Mapping the mouse cell atlas by Microwell-Seq. *Cell* **172**, 1091–1107.e17 https://doi.org/10.1016/j.cell.2018.02.001
69 Pierson, E. and Yau, C. (2015) ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16**, 241 https://doi.org/10.1186/s13059-015-0805-z
70 Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. and Vert, J.-P. (2018) A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9**, 284 https://doi.org/10.1038/s41467-017-02554-5
71 Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. et al. (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 https://doi.org/10.1093/nar/gkv007
72 Johnson, W.E., Li, C. and Rabinovic, A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 https://doi.org/10.1093/biostatistics/kxj037
73 Messmer, T., von Meyenn, F., Savino, A., Santos, F., Mohammed, H., Lun, A.T.L. et al. (2019) Transcriptional heterogeneity in naive and primed human pluripotent stem cells at single-cell resolution. *Cell Rep.* **26**, 815–824.e4 https://doi.org/10.1016/j.celrep.2018.12.099
74 Grive, K.J., Hu, Y., Shu, E., Grimson, A., Elemento, O., Grenier, J.K. et al. (2019) Dynamic transcriptome profiles within spermatogonial and spermatocyte populations during postnatal testis maturation revealed by single-cell sequencing. *PLoS Genet.* **15**, e1007810 https://doi.org/10.1371/journal.pgen.1007810
75 Loo, L., Simon, J.M., Xing, L., McCoy, E.S., Niehaus, J.K., Guo, J. et al. (2019) Single-cell transcriptomic analysis of mouse neocortical development. *Nat. Commun.* **10**, 134 https://doi.org/10.1038/s41467-018-08079-9
76 Guo, M., Du, Y., Gokey, J.J., Ray, S., Bell, S.M., Adam, M. et al. (2019) Single cell RNA analysis identifies cellular heterogeneity and adaptive responses of the lung at birth. *Nat. Commun.* **10**, 37 https://doi.org/10.1038/s41467-018-07770-1
77 Haghverdi, L., Lun, A.T., Morgan, M.D. and Marioni, J.C. (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421 https://doi.org/10.1038/nbt.4091
78 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. and Satija, R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 https://doi.org/10.1038/nbt.4096
79 Kiselev, V.Y., Yiu, A. and Hemberg, M. (2018) Scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 https://doi.org/10.1038/nmeth.4644
80 Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A. et al. (2019) Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163 https://doi.org/10.1038/s41590-018-0276-y
81 Woodhouse, S., Piterman, N., Wintersteiger, C.M., Göttgens, B. and Fisher, J. (2018) SCNS: a graphical tool for reconstructing executable regulatory networks from single-cell genomic data. *BMC Syst. Biol.* **12**, 59 https://doi.org/10.1186/s12918-018-0581-y
82 Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M.S., Ko, S.B., Gouda, N. et al. (2017) SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* **33**, 2314–2321 https://doi.org/10.1093/bioinformatics/btx194
83 Aubin-Frankowski, P.-C. and Vert, J.-P. (2018) Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference. *bioRxiv*, 464479 https://doi.org/10.1101/464479
84 Papili Gao, N., Ud-Dean, S.M., Gandrillon, O. and Gunawan, R. (2017) SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* **34**, 258–266 https://doi.org/10.1093/bioinformatics/btx575
85 Qiu, X., Rahimzamani, A., Wang, L., Mao, Q., Durham, T., McFaline-Figueroa, J.L. et al. (2018) Towards inferring causal gene regulatory networks from single cell expression measurements. *bioRxiv*, 426981 https://doi.org/10.1101/426981
86 Deshpande, A., Chu, L.-F., Stewart, R. and Gitter, A. (2019) Network inference with granger causality ensembles on single-cell transcriptomic data. *bioRxiv*, 534834 https://doi.org/10.1101/534834
87 Wang, S., MacLean, A.L. and Nie, Q. (2018) SoptSC: similarity matrix optimization for clustering, lineage, and signaling inference. *bioRxiv*, 168922
88 Aibar, S., González-Blas, C.B., Moerman, T., Imrichova, H., Hulselmans, G., Rambow, F. et al. (2017) SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083 https://doi.org/10.1038/nmeth.4463
89 Pina, C., Teles, J., Fugazza, C., May, G., Wang, D., Guo, Y. et al. (2015) Single-cell network analysis identifies DDIT3 as a nodal lineage regulator in hematopoiesis. *Cell Rep.* **11**, 1503–1510 https://doi.org/10.1016/j.celrep.2015.05.016
90 Iacono, G., Massoni-Badosa, R. and Heyn, H. (2019) Single-cell transcriptomics unveils gene regulatory network plasticity. *Genome Biol.* **20**, 110 https://doi.org/10.1186/s13059-019-1713-4
91 Chan, T.E., Stumpf, M.P.H. and Babtie, A.C. (2017) Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst.* **5**, 251–267.e3 https://doi.org/10.1016/j.cels.2017.08.014
92 Jackson, C.A., Castro, D.M., Saldi, G.-A., Bonneau, R. and Gresham, D. (2019) Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. *bioRxiv*, 581678 https://doi.org/10.1101/581678
93 Wang, Y., Solus, L., Yang, K. and Uhler, C. (2017) Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems*, pp. 5822–5831
94 Mohammadi, S., Ravindra, V., Gleich, D.F. and Grama, A. (2018) A geometric approach to characterize the functional identity of single cells. *Nat. Commun.* **9**, 1516 https://doi.org/10.1038/s41467-018-03933-2
95 Guo, M., Wang, H., Potter, S.S., Whitsett, J.A. and Xu, Y. (2015) SINCERA: a pipeline for single-cell RNA-seq profiling analysis. *PLoS Comput. Biol.* **11**, e1004575 https://doi.org/10.1371/journal.pcbi.1004575
96 Wang, Y., Wang, R., Zhang, S., Song, S., Jiang, C., Han, G. et al. (2019) iTALK: an R Package to characterize and illustrate intercellular communication. *bioRxiv*, 507871 https://doi.org/10.1101/507871
97 Zhou, J.X., Taramelli, R., Pedrini, E., Knijnenburg, T. and Huang, S. (2017) Extracting intercellular signaling network of cancer tissues using ligand-receptor expression patterns from whole-tumor and single-cell transcriptomes. *Sci. Rep.* **7**, 8815 https://doi.org/10.1038/s41598-017-09307-w
98 Kumar, M.P., Du, J., Lagoudas, G., Jiao, Y., Sawyer, A., Drummond, D.C. et al. (2018) Analysis of single-cell RNA-seq identifies cell-cell communication associated with tumor characteristics. *Cell Rep.* **25**, 1458–1468.e4 https://doi.org/10.1016/j.celrep.2018.10.047
99 Arneson, D., Zhang, G., Ying, Z., Zhuang, Y., Byun, H.R., Ahn, I.S. et al. (2018) Single cell molecular alterations reveal target cells and pathways of concussive brain injury. *Nat. Commun.* **9**, 3894 https://doi.org/10.1038/s41467-018-06222-0

100 Tsuyuzaki, K., Ishii, M. and Nikaido, I. (2019) Uncovering hypergraphs of cell-cell interaction from single cell RNA-sequencing data. *bioRxiv*, 566182 https://doi.org/10.1101/566182
101 Saelens, W., Cannoodt, R., Todorov, H. and Saeys, Y. (2019) A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 https://doi.org/10.1038/s41587-019-0071-9
102 Babtie, A.C., Chan, T.E. and Stumpf, M.P. (2017) Learning regulatory models for cell development from single cell transcriptomic data. *Curr. Opin. Syst. Biol.* **5**, 72–81 https://doi.org/10.1016/j.coisb.2017.07.013
103 Ocone, A., Haghverdi, L., Mueller, N.S. and Theis, F.J. (2015) Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics* **31**, i89–i96 https://doi.org/10.1093/bioinformatics/btv257
104 Wei, J., Hu, X., Zou, X. and Tian, T. (2016) *Inference of genetic regulatory network for stem cell using single cells expression data. 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE
105 Lim, C.Y., Wang, H., Woodhouse, S., Piterman, N., Wernisch, L., Fisher, J. et al. (2016) BTR: training asynchronous Boolean models using single-cell expression data. *BMC Bioinformatics* **17**, 355 https://doi.org/10.1186/s12859-016-1235-y
106 Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A.J., Tanaka, Y., Wilkinson, A.C. et al. (2015) Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.* **33**, 269–276 https://doi.org/10.1038/nbt.3154
107 Mukherjee, S., Carignano, A., Seelig, G. and Lee, S.I. (2018) Identifying progressive gene network perturbation from single-cell RNA-seq data. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2018**, 5034–5040
108 Rahimzamani, A. and Kannan, S. (2016) *Network inference using directed information: the deterministic limit. 2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE
109 Iacono, G., Mereu, E., Guillaumet-Adkins, A., Corominas, R., Cusco, I., Rodriguez-Esteban, G. et al. (2018) bigSCale: an analytical framework for big-scale single-cell data. *Genome Res.* **28**, 878–890 https://doi.org/10.1101/gr.230771.117
110 Lèbre, S. (2009) Inferring dynamic genetic networks with low order independencies. *Stat. Appl. Genet. Mol. Biol.* **8**, 1–38 https://doi.org/10.2202/1544-6115.1294
111 Bélanger, M., Allaman, I. and Magistretti, P.J. (2011) Brain energy metabolism: focus on astrocyte-neuron metabolic cooperation. *Cell Metab.* **14**, 724–738 https://doi.org/10.1016/j.cmet.2011.08.016
112 Rajbhandari, P., Thomas, B.J., Feng, A.-C., Hong, C., Wang, J., Vergnes, L. et al. (2018) IL-10 signaling remodels adipose chromatin architecture to limit thermogenesis and energy expenditure. *Cell* **172**, 218–233.e17 https://doi.org/10.1016/j.cell.2017.11.019
113 Fulcher, B.D. and Fornito, A. (2016) A transcriptional signature of hub connectivity in the mouse connectome. *Proc. Natl Acad. Sci. U.S.A.* **113**, 1435–1440 https://doi.org/10.1073/pnas.1513302113
114 Smillie, C.S., Biton, M., Ordovas-Montanes, J., Sullivan, K.M., Burgin, G., Graham, D.B. et al. (2018) Rewiring of the cellular and inter-cellular landscape of the human colon during ulcerative colitis. *bioRxiv*, 455451 https://doi.org/10.1101/455451
115 Vento-Tormo, R., Efremova, M., Botting, R.A., Turco, M.Y., Vento-Tormo, M., Meyer, K.B. et al. (2018) Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* **563**, 347–353 https://doi.org/10.1038/s41586-018-0698-6
116 Castro, D.M., De Veaux, N.R., Miraldi, E.R. and Bonneau, R. (2019) Multi-study inference of regulatory networks for more accurate models of gene regulation. *PLoS Comput. Biol.* **15**, e1006591 https://doi.org/10.1371/journal.pcbi.1006591
117 Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L. et al. (2016) Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866.e17 https://doi.org/10.1016/j.cell.2016.11.038
118 Hauser, A. and Bühlmann, P. (2012) Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.* **13**, 2409–2464
119 Wang, Y., Squires, C., Belyaeva, A. and Uhler, C. (2018) Direct estimation of differences in causal graphs. In *Advances in Neural Information Processing Systems*, pp. 3770–3781
120 Yang, K.D., Katcoff, A. and Uhler, C. (2018) Characterizing and learning equivalence classes of causal dags under interventions. *preprint arXiv*, 180206310
121 Femino, A.M., Fogarty, K., Lifshitz, L.M., Carrington, W. and Singer, R.H. (2003) Visualization of single molecules of mRNA in situ. *Methods Enzymol.* **361**, 245–304 https://doi.org/10.1016/S0076-6879(03)61015-3
122 Shah, S., Lubeck, E., Zhou, W. and Cai, L. (2017) seqFISH accurately detects transcripts in single cells and reveals robust spatial organization in the hippocampus. *Neuron* **94**, 752–758.e1 https://doi.org/10.1016/j.neuron.2017.05.008
123 Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S. and Zhuang, X. (2015) RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 https://doi.org/10.1126/science.aaa6090
124 Lovatt, D., Ruble, B.K., Lee, J., Dueck, H., Kim, T.K., Fisher, S. et al. (2014) Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. *Nat. Methods* **11**, 190–196 https://doi.org/10.1038/nmeth.2804
125 Simone, N.L., Bonner, R.F., Gillespie, J.W., Emmert-Buck, M.R. and Liotta, L.A. (1998) Laser-capture microdissection: opening the microscopic frontier to molecular analysis. *Trends Genet.* **14**, 272–276 https://doi.org/10.1016/S0168-9525(98)01489-9
126 Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wahlby, C. et al. (2013) In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857–860 https://doi.org/10.1038/nmeth.2563
127 Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Ferrante, T.C., Terry, R. et al. (2015) Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.* **10**, 442–458 https://doi.org/10.1038/nprot.2014.191
128 Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R. et al. (2019) Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 https://doi.org/10.1126/science.aaw1219
129 Eng, C.-H.L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y. et al. (2019) Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235–239 https://doi.org/10.1038/s41586-019-1049-y
130 Moffitt, J.R., Bambah-Mukku, D., Eichhorn, S.W., Vaughn, E., Shekhar, K., Perez, J.D. et al. (2018) Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 https://doi.org/10.1126/science.aau5324
131 Kim, M. and Tagkopoulos, I. (2018) Data integration and predictive modeling methods for multi-omics datasets. *Mol. Omics* **14**, 8–25 https://doi.org/10.1039/C7MO00051K

Emerging Topics in Life Sciences (2019) **3** 379–398
https://doi.org/10.1042/ETLS20180176

132 Zarayeneh, N., Ko, E., Oh, J.H., Suh, S., Liu, C., Gao, J. et al. (2017) Integration of multi-omics data for integrative gene regulatory network inference. *Int. J. Data Min. Bioinform.* **18**, 223–239 https://doi.org/10.1504/IJDMB.2017.087178
133 Huang, S., Chaudhary, K. and Garmire, L.X. (2017) More is better: recent progress in multi-omics data integration methods. *Front. Genet.* **8**, 84 https://doi.org/10.3389/fgene.2017.00084
134 Arneson, D., Shu, L., Tsai, B., Barrere-Cain, R., Sun, C. and Yang, X. (2017) Multidimensional integrative genomics approaches to dissecting cardiovascular disease. *Front. Cardiovasc. Med.* **4**, 8 https://doi.org/10.3389/fcvm.2017.00008
135 Marbach, D., Prill, R.J., Schaffter, T., Mattiussi, C., Floreano, D. and Stolovitzky, G. (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl Acad. Sci. U.S.A.* **107**, 6286–6291 https://doi.org/10.1073/pnas.0913357107
136 Hill, S.M., Heiser, L.M., Cokelaer, T., Unger, M., Nesser, N.K., Carlin, D.E. et al. (2016) Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat. Methods* **13**, 310 https://doi.org/10.1038/nmeth.3773

Appendix E

RESEARCH ARTICLE

Prenatal Bisphenol A Exposure in Mice Induces Multitissue Multiomics Disruptions Linking to Cardiometabolic Disorders

Le Shu,[1,2]* Qingying Meng,[1]* Graciel Diamante,[1] Brandon Tsai,[1] Yen-Wei Chen,[3] Andrew Mikhail,[1] Helen Luk,[1] Beate Ritz,[4,5] Patrick Allard,[3,5] and Xia Yang[1,2,3,6]

[1]Department of Integrative Biology and Physiology, University of California, Los Angeles, Los Angeles, California 90095; [2]Molecular, Cellular, and Integrative Physiology Interdepartmental Program, University of California, Los Angeles, Los Angeles, California 90095; [3]Molecular Toxicology Interdepartmental Program, University of California, Los Angeles, Los Angeles, California 90095; [4]Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, California 90095; [5]Institute for Society and Genetics, University of California, Los Angeles, Los Angeles, California 90095; and [6]Institute for Quantitative and Computational Biosciences, University of California, Los Angeles, Los Angeles, California 90095

ORCiD numbers: 0000-0002-3971-038X (X. Yang).

The health impacts of endocrine-disrupting chemicals (EDCs) remain debated, and their tissue and molecular targets are poorly understood. In this study, we leveraged systems biology approaches to assess the target tissues, molecular pathways, and gene regulatory networks associated with prenatal exposure to the model EDC bisphenol A (BPA). Prenatal BPA exposure at 5 mg/kg/d, a dose below most reported no-observed-adverse-effect levels, led to tens to thousands of transcriptomic and methylomic alterations in the adipose, hypothalamus, and liver tissues in male offspring in mice, with cross-tissue perturbations in lipid metabolism as well as tissue-specific alterations in histone subunits, glucose metabolism, and extracellular matrix. Network modeling prioritized main molecular targets of BPA, including *Pparg*, *Hnf4a*, *Esr1*, *Srebf1*, and *Fasn* as well as numerous less studied targets such as *Cyp51* and long noncoding RNAs across tissues, *Fa2h* in hypothalamus, and *Nfya* in adipose tissue. Lastly, integrative analyses identified the association of BPA molecular signatures with cardiometabolic phenotypes in mouse and human. Our multitissue, multiomics investigation provides strong evidence that BPA perturbs diverse molecular networks in central and peripheral tissues and offers insights into the molecular targets that link BPA to human cardiometabolic disorders. **(*Endocrinology* 160: 409–429, 2019)**

A central concept in the Developmental Origins of Health and Disease states that adverse environmental exposure during early developmental stages is an important determinant for later-onset adverse health outcomes, even in the absence of continuous exposure in adulthood (1–3). Bisphenol A (BPA) is one of the most prevalent environmental metabolic disruptors identified to date, with widespread exposure in human populations, and likely plays a role in Developmental Origins of Health and Disease (3–5). BPA is used in the production of synthetic polymers, including epoxy resins and polycarbonates and, with its advantageous mechanical properties, is ubiquitously found in everyday goods such as plastic bottles and inner coating of canned foods (6, 7).

ISSN Online 1945-7170

Received 16 September 2018. Accepted 13 December 2018.
First Published Online 18 December 2018

*L.S. and Q.M. contributed equally to this study.

Abbreviations: BMI, body mass index; BN, Bayesian network; BPA, bisphenol A; DEG, differentially expressed gene; DMC, differentially methylated CpG; EDC, endocrine-disrupting chemical; FDR, false discovery rate; FFA, free fatty acid; GEO, Gene Expression Omnibus; GWAS, genome-wide association studies; KD, key driver; lncRNA, long noncoding RNA; MetS, metabolic syndrome; MSEA, Marker Set Enrichment Analysis; NIEHS, National Institute for Environmental Health Sciences; PPAR, peroxisome proliferator–activated receptor; qPCR, quantitative RT-PCR; RNA-seq, RNA-sequencing; RRBS, reduced representation bisulfite sequencing; SNP, single nucleotide polymorphism; TF, transcription factor; TG, triglyceride; wKDA, Weighted Key Driver Analysis.

The ability of BPA to leach from these everyday products is a primary route of human exposure (8). BPA has been shown to have the ability to disrupt endocrine signaling important for many biological functions and has been linked to body weight, obesity, insulin resistance, diabetes, metabolic syndrome (MetS), and cardiovascular diseases in both human epidemiologic and animal studies (9–17). Importantly, it has been suggested that the developing fetus is particularly vulnerable to BPA exposure (9, 18). Intrauterine growth retardation has been consistently observed after developmental BPA exposure at intake doses below the suggested human safety level and has been associated with low birth weight, elevated adult fat weight, and altered glucose homeostasis (9, 19–22). As a precaution, BPA has been banned from baby products in Europe, Canada, and the United States. However, BPA is still in use in nonbaby products, renewing concerns about the continuous exposure of populations in addition to the description of its ability to influence health outcomes, including obesity and MetS, over several generations (23–26). Together, these lines of evidence support an intriguing hypothesis that BPA may have been a contributing factor to the rise of MetS and cardiometabolic diseases worldwide in the past decades (27–29).

Despite numerous studies connecting BPA with adverse health outcomes, there remain ample conflicting findings, as summarized by the European Food Safety Agency (30), the BPA Joint Emerging Science Working Group of the US Food and Drug Administration (31), and the recent National Toxicology Program report, CLARITY-BPA, in which functions of multiple organs were examined (32). Although inconsistencies across studies might be attributable to nonmonotonic dose response, exposure window difference, and varying susceptibility among testing models (14, 33), there are also several additional layers of complexity and challenges hindering the full dissection of the biological effects of BPA. First, previous studies examining BPA in various cell types and tissues suggest a broad impact on biological systems (25, 34–36). Second, BPA has been found to modulate multidimensional molecular events, such as gene expression and epigenetic changes, which are functionally important for processes such as metabolism and immune response (37–42). However, due to most studies being designed to focus on one factor at a time as well as noncomparable study designs, it is difficult to directly compare effects across tissues or types of molecular data to derive the molecular rules of sensitivity to BPA exposures. These research gaps in our understanding of the pleiotropy of endocrine-disrupting chemicals and toxicant biological actions necessitated the establishment of the National Institute for Environmental Health Sciences (NIEHS) TaRGET consortium (43) and a more recent call for the research community to systemically interrogate multiple -omics in multiple tissues to accelerate the discovery of key biological fingerprints of environmental exposure (44).

Here, we address some of the aforementioned limitations of past studies by using a highly integrative approach. We conducted a multitissue, multiomics systems biology study to examine the systems-level influence of prenatal BPA exposure using modern integrative genomics and network modeling approaches in a mouse model (Fig. 1A). We first used next-generation sequencing technologies to characterize perturbations in both the transcriptome and the epigenome across three

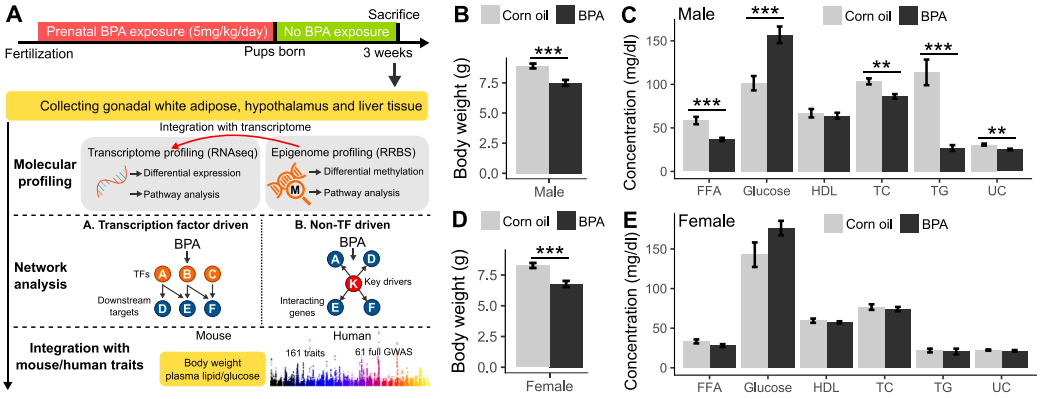


Figure 1. Overall study design and the measurements of metabolic traits in male and female offspring. (A) Framework of multiomics approaches to investigate the impact of prenatal BPA exposure. (B and C) Comparison of body weight, serum lipids, and glucose level in male mice between BPA and control groups at weaning age. (D and E) Comparison of body weight, serum lipids, and glucose level in female mice between BPA and control groups at weaning age. n = 9 to 13 mice (3 to 4 litters from different dams) per group. $^{**}P < 0.01$; $^{***}P < 0.001$ by two-sided Student *t* test. FFA, free fatty acid; HDL, high-density lipoprotein; TC, total cholesterol; TG, triglyceride; UC, unesterified cholesterol.

tissues (white adipose tissue, hypothalamus, and liver) in mouse offspring that had experienced *in utero* exposure to BPA. These tissues were chosen due to their important role in energy and metabolic homeostasis. The hypothalamus is the central regulator of endocrine and metabolic systems and plays a critical role in the regulation of nutrient and energy sensing, feeding, and energy expenditure (45); liver is critical for glucose and cholesterol homeostasis (46); and white adipose tissue is essential for energy and lipid storage, serves as a endocrine organ secreting numerous hormones related to metabolic regulation, and contributes to inflammatory processes by releasing various cytokines (47, 48). These tissues interact with one another to coordinately regulate metabolism and energy balance. Based on mounting evidence that genes operate in highly complex tissue-specific regulatory networks, we hypothesized that prenatal BPA exposure induces genomic and epigenomic reprogramming in the offspring by affecting the organization and functions of tissue-specific gene networks (49–52). Using both transcription factor (TF) and Bayesian networks (BNs), we modeled the dynamics of transcriptomic and epigenomic signatures and predicted potential regulators that govern the actions of BPA. Furthermore, the transcriptome, epigenome, and network information were layered upon metabolic phenotypes such as body weight, adiposity, circulating lipids, and glucose levels in the mouse offspring to evaluate disease association. Lastly, to assess the relevance of the BPA molecular targets identified in our mouse model for human diseases, we applied integrative genomics to bridge the mouse molecular signatures and genetic disease association data from human studies. Our study represents a comprehensive data-driven, systems-level investigation of the molecular and health impact of BPA.

Materials and Methods

Ethics statement

All animal experiments were performed in accordance with the Institutional Animal Care and Use Committee guidelines. Animal studies and procedures were approved by the Chancellor's Animal Research Committee of the University of California, Los Angeles.

Mouse model of prenatal BPA exposure

Inbred C57BL/6J mice were maintained on a special diet (5V01; LabDiet, St. Louis, MO), certified to contain <150 ppm estrogenic isoflavones, and housed under standard housing conditions (room temperature 22°C to 24°C) with a 12:12-hour light/dark cycle before mating at 8 to 10 weeks of age. Upon mating, female mice were randomly assigned to either the BPA treatment group or the control group. From 1-day postconception to 20 days postconception, BPA (Sigma-Aldrich, St. Louis, MO) dissolved in corn oil was administered to pregnant female mice via oral gavage (mimicking the common exposure route in humans) at 5 mg/kg/d on a daily basis. The dosage is situated below most reported no-observed-adverse-effect levels according to toxicity testing (https://comptox.epa.gov/dashboard/dsstoxdb/results?search=Bisphenol+A) and was typically used in previous studies (25, 53–55). We chose this dosage as a proof-of-concept for our systems biology study design and to facilitate comparison with previous studies. Control mice were fed the same amount of corn oil as vehicle. We chose corn oil over other solvents for BPA because BPA is water insoluble, and corn oil was found to be the least toxic compared with other commonly used solvents (56), but cannot exclude potential confounding from the combined effects of corn oil and BPA. Polycarbonate-free water bottles and cages were used to minimize any unintended exposure to BPA. Both parents and offspring from each treatment were maintained on a low-phytoestrogen special diet (5V01; LabDiet). Offspring in the vehicle- and BPA-treated groups were derived from three and four litters by different dams, respectively, to help assess and adjust for litter effects (57).

Characterization of cardiometabolic phenotypes and tissue collection

Male and female offspring were examined for a spectrum of metabolic phenotypes. In the male set, the control group had n = 9, and the BPA group had n = 11. For females, the control group had n = 9, and the BPA group had n = 13. There were two to three mice from each of the three to four different litters for each treatment group (57). We chose the weaning age to investigate early molecular and phenotypic changes in the offspring, which may predispose the offspring to late-onset diseases. Body weight of offspring was measured daily from postnatal day 5 up to the weaning age of 3 weeks. Mice were fasted overnight before being euthanized, and plasma samples were collected through retro-orbital bleeding. Serum lipid and glucose traits, including total cholesterol, high-density lipoprotein (HDL) cholesterol, unesterified cholesterol, triglycerides (TGs), free fatty acids (FFAs), and glucose, were measured by enzymatic colorimetric assays at the University of California, Los Angeles GTM Mouse Transfer Core as previously described (50). The liver, hypothalamus, and gonadal white adipose tissues were collected from each animal. The whole hypothalamus was collected by first carefully dissecting out the brain and placing it onto an ice-cold dissection board with the ventral side up. Using curved forceps, the tissue was held down, and the whole hypothalamus was gently pinched and spooned out. For the liver dissection, all liver lobes were first dissected out, then a portion of the distal part of the right lobe was collected for molecular profiling, and the rest of the liver was stored separately. For white adipose tissue, we chose the gonadal depot mainly due to its similarity to abdominal fat, established relevance to cardiometabolic risks, tissue abundance, and the fact that it is the most well-studied adipose tissue in mouse models. The gonadal fat depot was carefully dissected around the gonads, avoiding contamination of the gonads. After each tissue collection, the samples were flash frozen in liquid nitrogen and stored at −80°C. All mouse experiments were conducted in accordance with and approved by the Institutional Animal Care and Use Committee at University of California, Los Angeles.

Paired-end RNA-sequencing and data analysis

A total of 18 RNA samples were isolated from gonadal adipose, hypothalamus, and liver tissues (n = 3 per group per tissue; for each group, mice were randomly selected from litters

of different dams in independent cages) from male offspring using the AllPrep DNA/RNA Mini Kit (Qiagen GmbH, Hilden, Germany). The sample size was chosen based on previous RNA-sequencing (RNA-seq) studies that demonstrate sufficient reproducibility (50, 58–61). We focused on profiling male tissues because of stronger phenotypes observed in males (Fig. 1). Samples were processed for library preparation using the TruSeq RNA Library Preparation Kit (Illumina, San Diego, CA) for poly-A selection, fragmentation, and reverse transcription using random hexamer primers to generate first-strand cDNA. Second-strand cDNA was generated using RNase H and DNA polymerases, and sequencing adapters were ligated using the Illumina Paired-End sample prep kit. Library products of 250-bp to 400-bp fragments were isolated, amplified, and sequenced. Paired-end read sequencing was performed on an Illumina Hiseq2500 System. After quality control using FastQC (62), the HISAT-StringTie pipeline (63) was used for sequence alignment and transcript assembly. Identification of differentially expressed genes (DEGs) was conducted using DEseq2 (64). Sequenced reads were trimmed for adaptor sequence, masked for low-complexity or low-quality sequence, and then mapped to GRCm38/mm10 whole genome using HISAT v0.1.6. Default paired-end reads alignment parameters for HISAT were used with option -p 8. To account for multiple testing, we used the q-value method (65). After excluding genes with extremely low expression levels (fragments per kilobase of transcript per million mapped reads <1), only DEGs demonstrating differential expression comparing the BPA and control groups per tissue at a false discovery rate (FDR) $<5\%$ were used for biological pathway analysis, network analysis, and phenotypic data integration, as described later. The RNAseq quality matrix showing the number of sequencing reads and mapping rate for each sample is provided in an online repository (57). The number of reads aligned with genome for each sample varies between 30 million and 65 million, which satisfies the recommended number of reads needed for differential expression profiling (61).

Quantitative RT-PCR analysis

RNA from male and female liver, hypothalamus, and adipose tissue samples (n = 3 per tissue per treatment of males and n = 5 per tissue per treatment of females) were extracted using the MiRNeasy Mini Kit purchased from Qiagen following the manufacturer's instruction. Concentration and quality of the RNA were measured using the Thermo Fisher Scientific NanoDrop instrument (Thermo Fisher Scientific, Waltham, MA). cDNA synthesis was performed using the High-Capacity cDNA Reverse Transcription Kit from Applied Biosystems (Waltham, MA) following the manufacturer's protocol with minor modification by adding RNaseOUT Recombinant Ribonuclease Inhibitor (20 U/μL) from Invitrogen (Carlsbad, CA). Following the addition of reverse-transcription components, samples were incubated at the following thermocycler conditions: 10 minutes at 25°C, 120 minutes at 37°C, and then 5 minutes at 85°C. Upon completion, the cDNA was stored at −20°C until quantitative RT-PCR (qPCR) was performed. qPCR was done using the PowerUP SYBR Green Master Mix from Applied Biosystems. For both male and female liver and hypothalamus tissues, 20 ng of cDNA was used for the reaction. For the male and female adipose samples, 10 ng of cDNA was used due to low concentration. For all primers, a final concentration of 0.5 μM was used in the reaction. The qPCR reaction was run using the manufacturer's instructions followed by a melt-curve analysis. All primer sets displayed a single peak demonstrating specificity. PCR products were also run on a 1.5% agarose gel to validate appropriate amplicon size. Primer sequences are listed in an online repository (57). *Gapdh* was used as a housekeeping gene to quantify relative expression levels using the $\Delta\Delta$ threshold cycle analysis. Statistics was performed on the Δ threshold cycle using a t test.

Reduced representation bisulfite sequencing and data analysis

We constructed reduced representation bisulfite sequencing (RRBS) libraries for 18 DNA samples from adipose, hypothalamus, and liver tissues from male offspring (n = 3 per group per tissue from the same set of tissues chosen for transcriptome analysis described previously). The DNA samples were quantified using the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific), and 100 ng of DNA was used for library preparation. After digestion of the DNA with the *Msp*I enzyme, samples underwent an end-repair and adenylation process, followed by adapter ligation using the TruSeq barcode adapter (Illumina), size selection using AMPure Beads (Beckman Coulter, Brea, CA), and bisulfite treatment using the Epitect Kit (Qiagen). Bisulfite-treated DNA was then amplified using the TruSeq Library Prep Kit (Illumina) and sequenced with the Illumina Hiseq2500 System. Bisulfite-converted reads were processed and aligned to the reference mouse genome (GRCm38/mm10 build) using the bisulfite aligner BSMAP (66). We then used MOABS (67) for methylation ratio calling and identification of differentially methylated CpGs (DMCs). FDR was estimated using the q-value approach. Loci with methylation level changes of $>5\%$ between BPA and control groups and FDR $<5\%$ for each tissue were considered statistically significant DMCs. To annotate the locations of the identified DMCs in relation to gene regions and repetitive DNA elements accessed from the UCSC Genome Browser, we used the Bioconductor package "annotatr" (68). Specifically, gene regions were categorized into: (i) 1 to 5 kb upstream of the transcription start site; (ii) promoter (<1 kb upstream of the transcription start site); (iii) 5′ untranslated region; (iv) exons; (v) introns; and (vi) 3′ untranslated region. The "annotatr" package was also used to annotate DMCs for known long noncoding RNAs (lncRNAs) based on GENCODE Release M16. Overrepresentation of DMCs within each category was calculated using a one-sided Fisher exact test. We further evaluated the link between DEGs and their local DMCs (DMCs annotated as any of the six above-mentioned gene regions) by correlating the methylation ratio of DMCs with the expression level of DEGs.

Pathway, network, and disease association analyses of DEGs and DMCs using the Mergeomics R package

To investigate the functional connections among the BPA-associated DEGs or DMCs (collectively referred to as molecular signatures of BPAs) and to assess the potential association of BPA-affected genes with diseases in human populations, we used the Mergeomics package (69), an open-source bioconductor R package (https://bioconductor.org/packages/devel/bioc/html/Mergeomics.html) designed to perform various integrative analyses in multiomics studies. Mergeomics consists of two main libraries, Marker Set Enrichment Analysis (MSEA) and Weighted Key Driver Analysis (wKDA). In the current study, we used MSEA to assess: (i) whether known biological processes, pathways, or TF targets were enriched for BPA

molecular signatures as a means to annotate the potential functions or regulators of the molecular signatures; and (ii) whether the BPA signatures demonstrate enrichment for disease associations identified in human genome-wide association studies (GWAS) of various complex diseases (57). wKDA leverages gene network topology (interactions or regulatory relations among genes) and edge weight (strength or reliability of interactions and regulatory relations) information of graphical gene networks to predict potential key regulators of a given group of genes—in this case, the BPA-associated DEGs (57). Both MSEA and wKDA were built around χ^2-like statistics (57) that yield robust findings that have been experimentally validated (51, 52, 69). Details of each usage of the Mergeomics package are discussed later.

Functional annotation of DEGs and DMCs

To infer the functions of the DEGs and DMCs affected by BPA, we used MSEA to annotate the DEGs or local genes adjacent to the DMCs with known biological pathways curated from the Kyoto Encyclopedia of Genes and Genomes (70) and Reactome (71). In brief, we extracted the differential expression *P* values of genes in each pathway from the differential expression or methylation analyses and compared these *P* values against the null distribution of *P* values from random gene sets with matching gene numbers. If genes in a given pathway collectively show more significant differential expression or differential methylation *P* values compared with random genes based on a χ^2-like statistic, we annotate the DEGs or DMCs using that pathway (57). DEGs and DMCs can have multiple overrepresented pathways.

Identification of TF hot spots perturbed by BPA

To dissect the regulatory cascades of BPA, we first assessed whether BPA-associated DEGs were downstream targets of specific TFs. The hypothesis behind this analysis is that BPA first affects TFs, which in turn regulate the expression of downstream genes. We used TF regulatory networks for adipose, brain, and liver tissue retrieved from the FANTOM5 database (72). Note that only a whole brain (instead of hypothalamus) TF network was available, which may only partially represent hypothalamic gene regulation. Each TF network was processed to keep the edges with high confidence (57). To identify TFs for which targets were perturbed by BPA, the downstream nodes of each TF in the network were pooled as the target genes for that TF. We then assessed the enrichment for BPA exposure–related DEGs among the target genes of each TF using MSEA. TFs with FDR <5% were considered statistically significant. Cytoscape software was used for TF network visualization (73).

BN and wKDA to identify potential non-TF regulators

To further identify non-TF regulators that sense BPA and then perturb downstream genes, we used BNs of adipose, hypothalamus, and liver tissues constructed from genetic and transcriptomic data from several large-scale mouse and human studies (57). wKDA was used to identify network key drivers (KDs), which are defined as network nodes for which neighboring subnetworks are significantly enriched for BPA-associated DEGs. Briefly, wKDA takes gene set G (*i.e.*, BPA DEGs) and directional gene network N (*i.e.*, BNs) as inputs. For every gene K in network N, neighboring genes within 1-edge distance were tested for enrichment of genes in G using χ^2-like statistics followed by FDR assessment by permutation (57). Network genes that reached FDR <5% were reported as potential KDs.

Association of BPA DEGs and DMCs with mouse phenotypes and human diseases/traits

To assess whether the BPA molecular signatures were related to phenotypes examined in the mouse offspring, we calculated the Pearson correlation coefficient among expression level of DEGs, methylation ratio of DMCs, and the measurement of metabolic traits. For human diseases or traits, we accessed the GWAS catalog database (74) and collected the lists of candidate genes reported to be associated with 161 human traits/diseases ($P < 1e\text{-}5$). These genes were tested for enrichment of the BPA DEGs and DMCs in our mouse study using MSEA. We further curated all publicly available full summary statistics for 61 human traits/diseases from various public repositories (57). This allowed us to apply MSEA to comprehensively assess the enrichment for human disease association among BPA transcriptomic signatures using the full spectrum of large-scale human GWAS. For each tissue-specific gene signature, we used the single nucleotide polymorphisms (SNPs) within a 50-kb chromosomal distance as the representing SNPs for that gene. The trait/disease association *P* values of the SNPs were then extracted from each GWAS and compared with the *P* values of SNPs of random sets of genes to assess whether the BPA signatures were more likely to show stronger disease association in human GWAS (57). This strategy has been successfully used in our previous animal model studies to assess the connection of genes affected by environmental perturbations such as diets and trauma to various human diseases (50, 59).

Data availability

Supplemental methods, figures, and tables are available at Figshare (doi.org/10.6084/m9.figshare.7451069.v2) (57). RNA-seq and RRBS data have been submitted to the Gene Expression Omnibus (GEO) under accession numbers GSE121603 (for RNA-seq) and GSE121604 (for RRBS).

Results

Prenatal BPA exposure induces lower body weight and alterations in cardiometabolic phenotypes

We exposed pregnant C57BL/6J mice to BPA during gestation via oral gavage at the dosage of 5 mg/kg/d and examined the male and female offspring for a spectrum of metabolic phenotypes at weaning age. Compared with the control group, both male and female offspring from the BPA group showed significantly lower body weight (Fig. 1B and 1D). There were also considerable decreases in serum lipid parameters and an increase in serum glucose level in males (Fig. 1C), but not in females (Fig. 1E). The phenotypic differences between BPA and control groups are not the results of litter effect, as offspring from different dams in each group showed similar patterns (57).

Prenatal BPA exposure induces tissue-specific transcriptomic alterations in male weaning offspring

To explore the molecular basis underlying the potential health impact of prenatal BPA exposure, we collected three key metabolic tissues including white adipose tissue, hypothalamus, and liver from male offspring (due to the stronger observed phenotypes) at 3 weeks. We used RNA-seq to profile the transcriptome and identified 86, 93, and 855 DEGs in the adipose tissue, hypothalamus, and liver tissue, respectively, at FDR $<5\%$ (57). This supports the ability of prenatal BPA exposure to induce large-scale transcriptomic disruptions in offspring, with the impact appearing to be more prominent in liver. The DEGs show distinct expression patterns between the control and BPA groups, and samples within each group generally agree with one another on the upregulation or downregulation (Fig. 2A). The DEGs were highly tissue specific, with only 12 out of the 86 adipose DEGs and 16 out of the 93 hypothalamus DEGs being found in liver. Interestingly, the hypothalamic DEGs are predominantly upregulated in the BPA group, whereas the other two tissues did not show such direction bias (57). Only one gene, *Cyp51* (sterol 14-α demethylase), was shared across all three tissues but with different directional changes (upregulated in hypothalamus and liver and downregulated in adipose) (Fig. 2B).

Replication of the DEG signatures using both qPCR and independent studies

To validate the identified DEGs in the RNA-seq analysis of the male samples, we selected 22 genes (14 from the liver, 4 from the hypothalamus, and 4 from the adipose tissue) for qPCR analysis. We found that the majority (19 out of 22; 86%) of the genes tested in the male samples were significantly altered in the BPA samples in our qPCR data (Fig. 3A, 3C, and 3E). All 22 genes tested via qPCR showed consistent directions in expression changes as observed in our RNA-seq analysis (Fig. 3), supporting the accuracy of our RNA-seq data.

Next, to evaluate the effect of BPA in different sexes, we also analyzed the expression of the same 22 select genes in the female cohort. We found that only one gene, *Lpl*, in the liver was significantly affected in the female offspring exposed to BPA (Fig. 3B). The liver expression of *Rgs16*, *Msm01*, *Pparg*, and *Mup3* in the exposed females also showed very similar trends to the BPA male group (Fig. 3B). These subtle changes in expression levels are consistent with the weaker phenotypic data observed in females (Fig. 1D and 1E).

To further assess the reproducibility of the differential expression signatures identified in our study, we examined our DEG signatures using independent expression profiling data deposited on the GEO (57). We identified three GEO datasets related to BPA exposure in mice: two from GSE26728 (75) and one from GSE43977 (76) (Fig. 4A). These publicly available liver transcriptome datasets were derived from studies of BPA exposure during adulthood, as we were not able to identify other publicly available datasets with the same *in utero* exposure condition tested in our exposure paradigm, making a direct replication difficult. However, we reasoned that if core mechanisms exist for BPA regardless of experimental conditions, consistent signals should be derived. We compared the differential expression signatures from the three existing liver studies against ours and found limited consistency in BPA signatures across datasets, even for the two datasets that were originated from the same study (GSE26728) (Fig. 4C). These results support that BPA has condition-specific activities. Nevertheless, 10% of our DEGs were replicated in the other GEO datasets ($P < 1e\text{-}4$ compared with random expectation via a permutation analysis) (Fig. 4C). *Srebf1*, encoding a key TF in lipid metabolism, was consistent across all four datasets, along with numerous additional genes consistent in two or more studies (Fig. 4B).

Due to the major difference in the exposure window between our study and the previously noted publicly available GEO datasets, we expanded our search by collecting the BPA signatures that were reported in published mouse studies on developmental BPA exposure. We collected 24 unique DEGs from 7 mouse studies (22, 77–82), which evaluated BPA effects on liver during development (57) and found that 6 (*Cyp17a1*, *Fasn*, *Fdps Gstt3*, *Pparg*, and *Scd1*) of the 24 DEGs were also significantly affected in our study. Similarly, we compared our hypothalamic DEGs with those derived from three mouse and rat studies with similar BPA exposure window and found three overlapping genes with our study (*Akt2*, *Hip1r*, and *Ndufb7*) (83–85). Recent adipose tissue gene expression studies of developmental BPA exposure revealed very few DEGs, and none overlapped with those in our study (22, 37, 40, 86). These comparisons support certain consistencies in liver and hypothalamus DEGs between studies but limited overlaps for adipose DEGs.

Functional annotation of DEGs in adipose, hypothalamus, and liver tissues

To better understand the biological implications of the BPA exposure–related DEGs in individual tissues, we evaluated the enrichment of DEGs for known biological pathways and functional categories using the Mergeomics package (69) (Fig. 2C–2E; full results in an online repository) (57). We observed strong enrichment for pathways related to lipid metabolism (lipid transport,

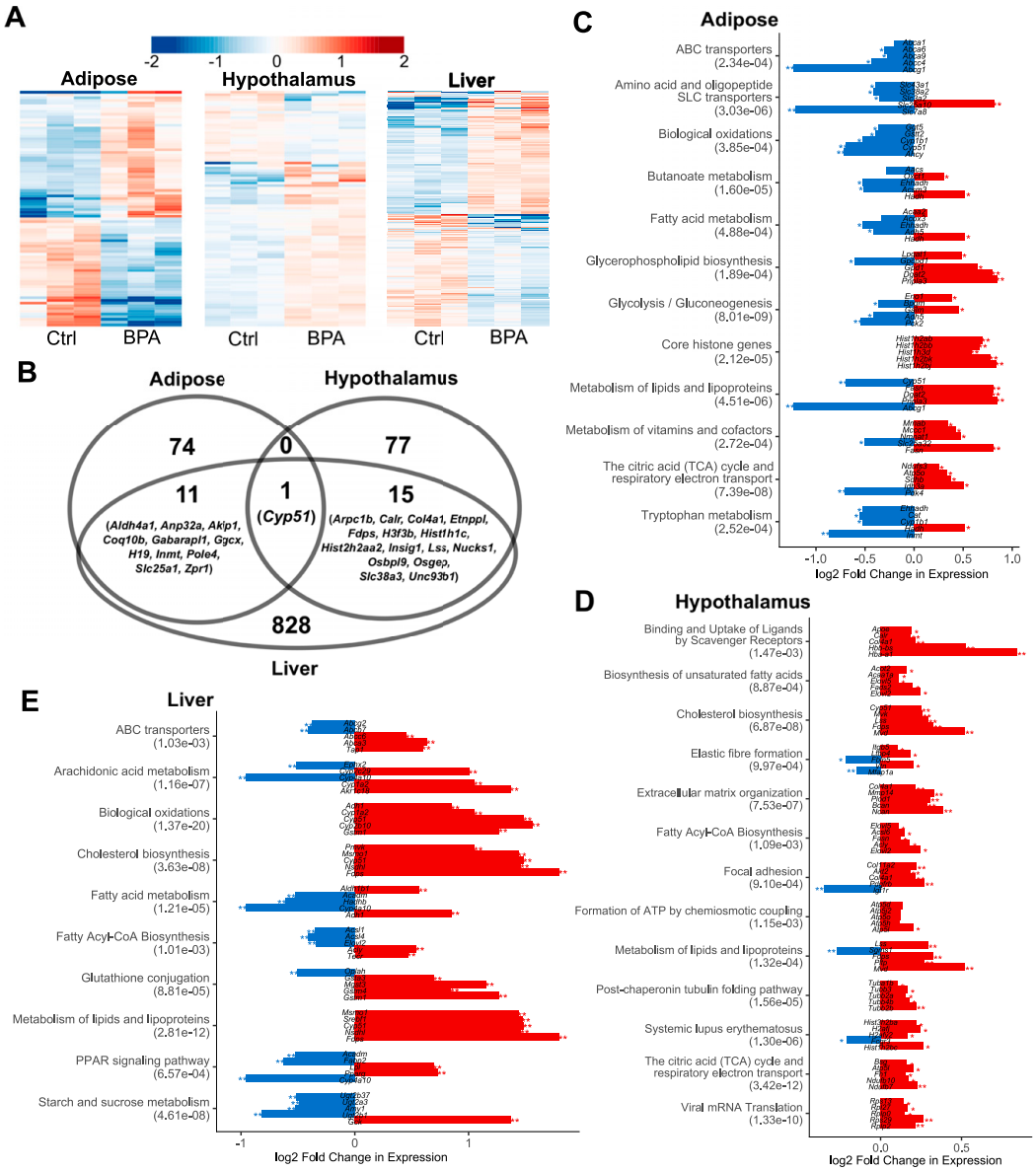


Figure 2. Prenatal BPA exposure induced transcriptomic alterations in adipose, hypothalamus, and liver. (A) Heat map of expression changes in adipose, hypothalamus, and liver for the DEGs affected by BPA. Color indicates fold change of expression, with red and blue indicating upregulation and downregulation by BPA, respectively. (B) Venn diagram demonstrating tissue-specific and shared DEGs between tissues. (C–E) Significantly enriched pathways (FDR <5%) among DEGs from each tissue. Enrichment *P* value (shown in parentheses following the name of functional annotation) is determined by MSEA. The fold change and statistical significance for the top five DEGs in each pathway are shown. $*P < 0.05$; **FDR <5% in differential expression analysis using DEseq2.

fatty acid metabolism, and cholesterol biosynthesis) and energy metabolism (biological oxidation and tricarboxylic acid cycle) across all three tissues. Most of these pathways appeared to be upregulated in all three tissues, except that genes involved in biological oxidation in adipose tissue were downregulated (Fig. 2C–2E). Individual tissues also showed perturbations of unique pathways: peroxisome proliferator–activated receptor

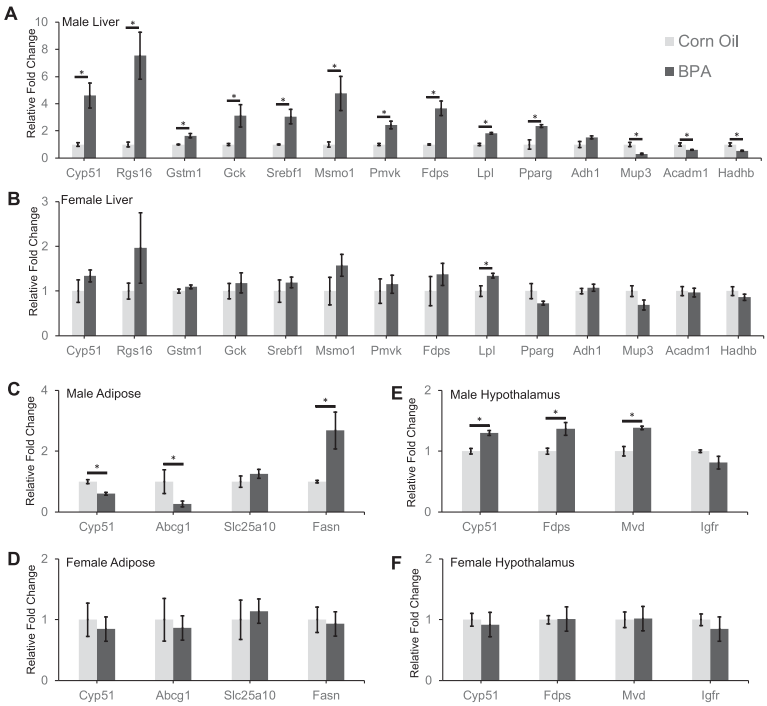


Figure 3. qPCR analysis on identified DEGs in male and female offspring. Relative fold change in expression levels of 14 genes in liver samples of (A) male and (B) female offspring. Relative fold change in expression of four genes in gonadal adipose samples of (C) male and (D) female offspring. Relative fold change in expression of four genes in hypothalamus samples of (E) male and (F) female offspring. Data presented as mean ± SE of independent replicates. n = 3 per tissue per group for males and n = 5 per tissue per group for females. *$P < 0.05$ by two-sided Student t test.

(PPAR) signaling and arachidonic acid pathways were altered in liver; extracellular matrix–related processes were enriched among hypothalamic DEGs; and core histone genes were upregulated in adipose DEGs (Fig. 2C–2E). In addition, TG biosynthesis and glucose metabolism pathways were also moderately enriched among adipose DEGs, whereas few changes were seen for genes involved in adipocyte differentiation (57).

Next, we compared the enriched pathways from our DEGs with those identified from independent GEO datasets as described above to evaluate whether distinct study-specific signatures could converge onto similar biological processes. The replicated pathways across studies include steroid hormone biosynthesis, retinol metabolism, and fatty acid metabolism, suggesting that these processes were consistently influenced by BPA under varying exposure windows and dosages (Fig. 4D). At FDR <5%, 56.1% of the significant pathways in our study were replicated in one or more independent studies ($P < 1e-4$ compared with random expectation via a permutation test; Fig. 4E). Pairwise comparison revealed relatively higher overlap ratios between our study and individual independent studies than between the previous studies, despite the greater similarity in the study design among the previous studies (Fig. 4E).

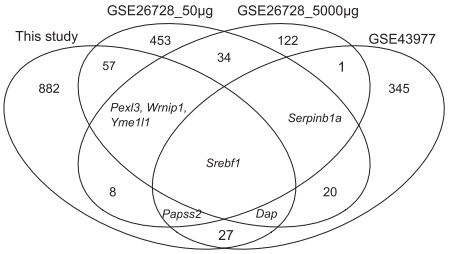
Prenatal BPA exposure induces tissue-specific epigenetic alterations in male weaning offspring

Consistent with the observed gene expression disruptions at the transcriptomic level, we observed numerous methylomic alterations using RRBS, which characterizes DNA methylation states of millions of potential epigenetic sites at single-base resolution. At FDRs <5%, 5136, 104, and 476 DMCs were found in adipose, hypothalamus, and liver tissues, respectively (57). The DMCs show distinct expression patterns between the control and BPA groups, and samples within each group generally agree with one another on the upregulation or downregulation (Fig. 5A). When comparing our adipose methylation signatures with a previous study (40), we were able to replicate five out of seven peak hypomethylated genes and six out of nine peak hypermethylated genes. Interestingly, BPA induced local methylation changes in *Gm26917* and *Yam1*, two lncRNAs with no previously known link to BPA, consistently across three tissues (Fig. 5B).

A

Study	Tissue	Organism	Exposure window	Exposure route	Study type	Dosage
Our study	Adipose/Hypothalamus/Liver	Mouse	Prenatal	Feed	*In vivo*	5mg/kg/day
GSE26728_50µg	Liver	Mouse	Adult	Food	*In vivo*	50µg/kg/day
GSE26728_5000µg						5000µg/kg/day
GSE43977	Liver	Mouse	Adult	Feed	*In vivo*	5000 ppm

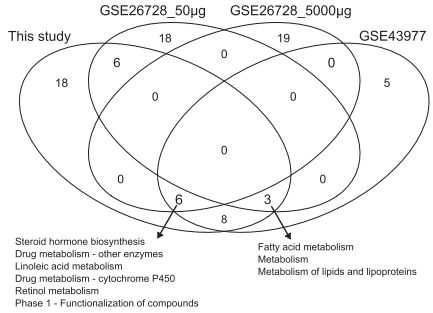
B



C

	This study	GSE26728_50µg	GSE26728_5000µg	GSE43977
This study (10.0%, p < 1e-4)	-	6.3%	1.3%	3.1%
GSE26728_50µg (14.6%, p < 1e-4)	10.9%	-	6.8%	4.0%
GSE26728_5000µg (8.8%, p= 0.38)	7.6%	22.8%	-	2.3%
GSE43977 (13.1%, p < 1e-4)	7.6%	5.8%	1.0%	-

D



E

	This study	GSE26728_50µg	GSE26728_5000µg	GSE43977
This study (56.1%, p < 1e-4)	-	22.0%	14.6%	41.5%
GSE26728_50µg (33.3%, p < 1e-4)	33.3%	-	0.0%	11.1%
GSE26728_5000µg (24.0%, p = 7e-4)	24.0%	0.0%	-	24.0%
GSE43977(77.3%, p < 1e -4)	77.3%	13.6%	27.3%	-

Figure 4. Comparison of the liver DEGs and their functional annotations against publicly available datasets relevant to BPA exposure in GEO. (A) Descriptions of the study design of different datasets. (B) Venn diagram of the DEGs identified in different datasets. DEGs were determined by Limma at $P < 0.01$. (C) The percentage of DEGs from the datasets in each row header that are replicated by the datasets in each column header. Numbers in parentheses indicate the percentage of DEGs that are replicated by at least one independent study and the significance of the replication percentage determined by permutation test. (D) Venn diagram of the functional annotations for the DEGs identified in different datasets. Functional annotations were determined by MSEA at FDR <5%. (E) The percentage of functional annotations from the datasets in each row header that are replicated by the datasets in each column header. Numbers in parentheses indicate the percentage of annotations that are replicated by at least one independent study and the significance of the replication percentage determined by permutation test.

The majority of the DMCs are located in intergenic regions (32% to 38%), followed by introns (31% to 37%) and exons (13% to 15%), but there is a paucity of DMCs in the promoter region (3% to 5%) (57). Contrary to predictions that promoter regions may be more prone to epigenetic changes, we found that within-gene and intergenic methylation alterations in DNA methylation are more prevalent, a pattern consistently observed in previous epigenomic studies (50, 87). In addition, 5.0%, 8.6%, and 8.1% DMCs overlap with repetitive DNA elements in adipose, hypothalamus, and liver, respectively, recapitulating a previous report of the interaction between BPA and repetitive DNA (88).

For DMCs that are located within or adjacent to genes, we further tested whether the local genes adjacent to those DMCs show enrichment for known functional categories. Unlike DEGs, top processes enriched for DMCs concentrated on intracellular and extracellular communication and signaling-related pathways such as axon guidance, extracellular matrix organization, and nerve growth factor signaling (Fig. 5C; full results in an online repository) (57). The affected genes in these processes are related to cellular structure, cell adhesion, and cell migration, indicating that these functions may be particularly vulnerable to BPA-induced epigenetic modulation.

Potential regulatory role of DMCs in transcriptional regulation of BPA-induced DEGs

To explore the role of DMCs in regulating DEGs, we evaluated the connection between transcriptome and methylome by correlating the expression level of DEGs with the methylation ratio of their local DMCs. For the DEGs in adipose, hypothalamus, and liver tissue, we identified 42, 36, and 278 local DMCs for which methylation ratios were significantly correlated with gene expression. At a global level, compared with non-DEGs, DEGs are more likely to contain local correlated DMCs (57). A closer look into the expression-methylation correlation by different chromosomal regions further revealed a context-dependent correlation pattern (Fig. 5D). In adipose and liver, the 3% to 5% of DMCs in promoter regions tend to show significant enrichment for negative correlation with DEGs, whereas gene body methylations for DEGs are more likely to show signficant enrichment for positive correlation with gene expression. In hypothalamus, however, positive correlations between DEGs and DMCs are more prevalent across different gene regions. In addition, liver DMCs within lncRNAs were uniquely enriched for negative correlation with lncRNA expression, although the lack of a reliable mouse lncRNA target database prevented us from further investigating whether downstream targets of the lncRNAs were enriched in the DEGs. Specific examples of DEGs showing significant correlation with local DMCs include adipose DEG *Slc25a1* (solute carrier family 25 member 1; involved in TG biosynthesis), hypothalamic DEG *Mvk* (mevalonate kinase; involved in cholesterol biosynthesis), and liver DEG *Gm20319* (an lncRNA with unknown function) (57). These results support a role of BPA-induced

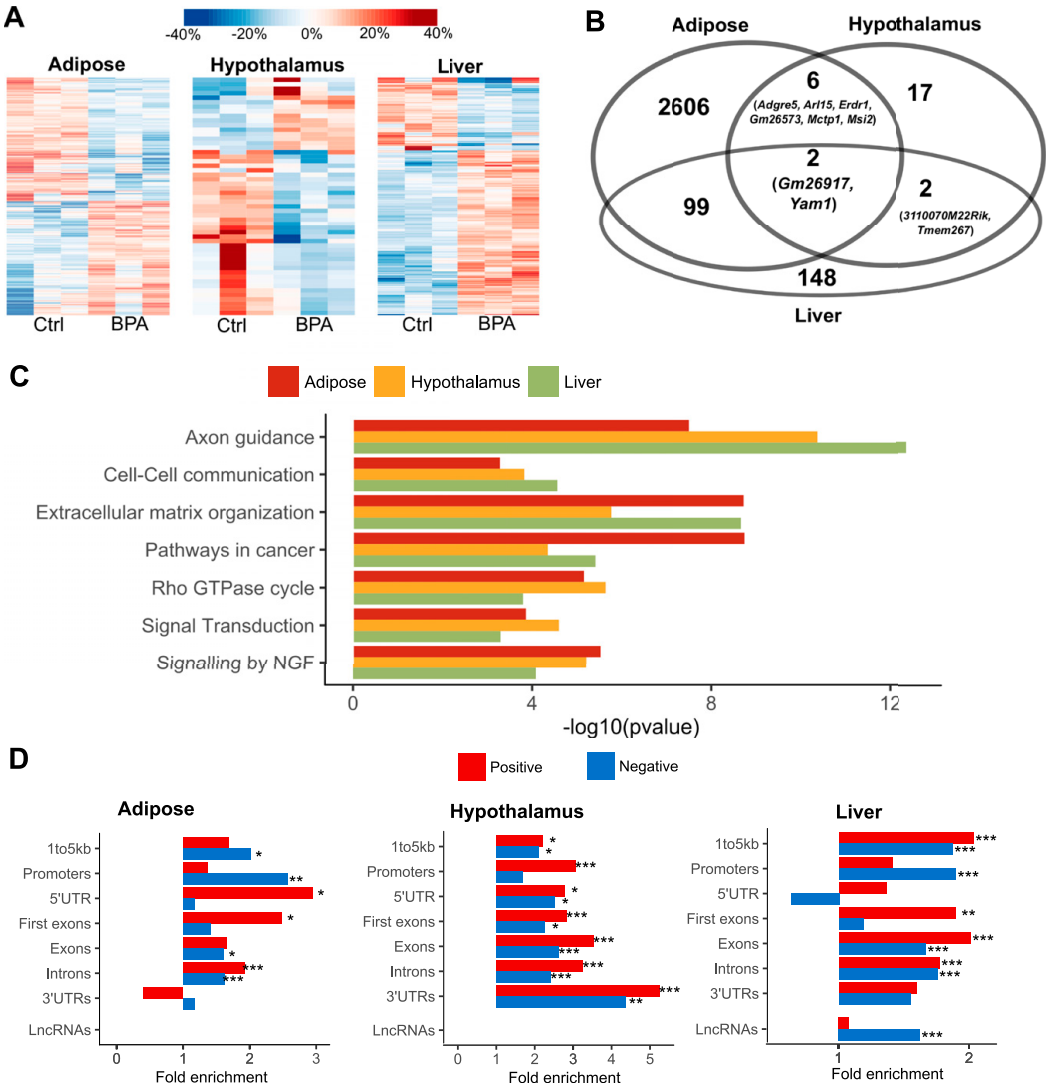


Figure 5. Prenatal BPA exposure induced methylomic level alteration in adipose, hypothalamus, and liver. (A) Heat map of methylation level changes for the DMCs. Color indicates change in methylation ratio, with red and blue indicating upregulation and downregulation by BPA, respectively. (B) Venn diagram of genes with local DMCs between tissues shows tissue-specific and shared genes mapped to DMCs. (C) Significantly enriched pathways that satisfied FDR <1% across DMCs from adipose, hypothalamus, and liver tissues. Enrichment *P* value is determined by MSEA. (D) Fold enrichment for positive correlations (red bars) or negative correlations (blue bars) between DMCs and local DEGs, assessed by different gene regions. $*P < 0.05$; $**P < 0.01$; $***P < 0.0001$; enrichment *P* values were determined using Fisher exact test. Ctrl, control; NGF, nerve growth factor; UTR, untranslated region.

differential methylation in altering the expression levels of adjacent genes.

Pervasive influence of prenatal BPA exposure on the liver TF network

BPA is known to bind to diverse types of nuclear receptors such as estrogen receptors and PPARs that function as TFs, thus influencing the action of downstream genes (89, 90). PPARg in particular has been shown to be a target of BPA in mouse and human and mechanistically linking BPA exposure with its associated effect on weight gain and increased adipogenesis (91–93). To explore the TF regulatory landscape underlying BPA exposure based on our genome-wide data, we leveraged

tissue-specific TF regulatory networks from the FANTOM5 project (72) and integrated them with our BPA transcriptome profiling data. No TF was found to be differentially expressed in adipose tissue, whereas 1 TF (Pou3f1) and 14 TFs (such as Esrra, Hnf1a, Pparg, Tcf21, and Srebf1) were found to be differentially expressed in hypothalamus and liver, respectively. Due to the temporal nature of TF action, changes in TF levels may precede the downstream target genes and not be reflected in the transcriptomic profiles measured at the time of euthanization. Therefore, we further curated the target genes of TFs from FANTOM5 networks and tested the enrichment for the target genes of each TF among our tissue-specific DEGs (57). This analysis confirmed that BPA perturbs the activity of the downstream targets for estrogen receptors Esrrg (P = 1.4e-3; FDR 1.9%) and Esrra (P = 0.03; FDR 13%) in liver, as well as Esr1 in both adipose (P = 7.2e-3; FDR 10.6%) and liver (P = 7.2e-3; FDR 4.7%). Targets of Pparg were also perturbed in liver (P = 4.1e-3; FDR 3.8%). Therefore, we demonstrated that our data-driven network modeling is able to not only recapitulate results from previous *in vitro* and *in vivo* studies showing that BPA influences estrogen signaling and PPAR signaling (90), but also uniquely point to the tissue specificity of these BPA target TFs.

In addition to these expected TFs, we identified 14 adipose TFs and 61 liver TFs for which target genes were significantly enriched for BPA DEGs at FDR <5%. Many of these TFs showed much stronger enrichment for BPA DEGs among their downstream targets than the estrogen receptors (57). The adipose TFs include nuclear TF Y subunit α (Nfya) and fatty acid synthase (Fasn), both implicated in adipocyte energy metabolism (94). The liver TFs include multiple genes from the hepatocyte nuclear factors family and the CCAAT-enhancer-binding proteins family, which are critical for liver development and function, suggesting a pervasive influence of BPA on liver TF regulation.

We further extracted the subnetwork containing 89 unique downstream targets of the significant liver TFs that are also liver DEGs. This subnetwork showed significant enrichment for genes involved in metabolic pathways such as steroid hormone biosynthesis and fatty acid metabolism. The regulatory subnetwork for the top liver TFs (FDR <5%) revealed a highly interconnected TF subnetwork that potentially senses BPA exposure and in turn governs the expression levels of their targets (Fig. 6A), with Pparg and Hnf4 among the core TFs. Some of the TFs in this network, including Esr1, Esrrg, Foxp1, and Tcf7l1, also had local DMCs identified in our study, indicating that BPA may perturb this liver TF subnetwork via local modification of DNA methylation of key TFs.

Identification of potential non-TF regulators governing BPA induced molecular perturbations

To further identify regulatory genes that mediate the action of BPA on downstream targets through non-TF mechanisms, we leveraged data-driven tissue-specific BNs generated from multiple independent human and mouse studies (57). These data-driven networks are complementary to the TF networks used previously and have proven valuable for accurately predicting gene-gene regulatory relationships and novel KDs that were experimentally validated (49–52, 95). KDs were defined as network nodes for which surrounding subnetworks are significantly enriched for BPA exposure–related DEGs. At FDR <1%, we identified 21, 1, and 100 KDs in adipose, hypothalamus, and liver, respectively (57). The top KDs in adipose (top 5 KDs *Acss2*, *Pc*, *Agpat2*, *Slc25a1*, and *Acly*), hypothalamus (*Fa2h*), and liver (top 5 KDs *Dhcr7*, *Aldh3a2*, *Fdft1*, *Mtmr11*, and *Hmgcr*) were involved in cholesterol, fatty acid, and glucose metabolism processes. In addition, three KDs—*Acss2* (acetyl-coenzyme A synthetase 2), *Acat2* (acetyl-coenzyme A acetyltransferase 2), and *Fasn* (fatty acid synthase)—were involved in the upregulation of DEGs in both adipose and liver, despite the fact that few DEG signatures overlap across tissues (Fig. 6B). These KDs are consistent with the observed increased expression of several genes implicated in lipogenesis, including *Fasn*, and help explain the liver accumulation of TGs when mice are exposed to BPA (75). Together, these results indicate that BPA may engage certain common regulators that have tissue-specific targets. The distinct upregulatory pattern within the subnetworks of individual KDs supports the potential functional importance of KDs in orchestrating the action of downstream genes. These KDs, along with the TFs from the previous analysis, may represent regulatory targets that transmit the *in vivo* biological effects of BPA.

BPA transcriptomic and methylomic signatures are related to metabolic traits in mice

To assess the relationship between the BPA molecular signatures and metabolic traits in the mouse model, the DEGs and DMCs from individual tissues were tested for correlation with the measured metabolic traits: body weight, FFAs, total cholesterol, HDL cholesterol, TGs, and blood glucose. At $P < 0.05$, over two-thirds of tissue-specific DEGs and >60% of DMCs were identified to be correlated with at least one metabolic trait (Fig. 7A and 7B). Notably, liver DEGs exhibited stronger correlation with FFAs and TGs, whereas adipose DEGs were uniquely associated with glucose level, which is consistent with the pathway annotation results for these tissues. In contrast, liver DMCs showed stronger correlations with metabolic traits than those from adipose and hypothalamus tissues.

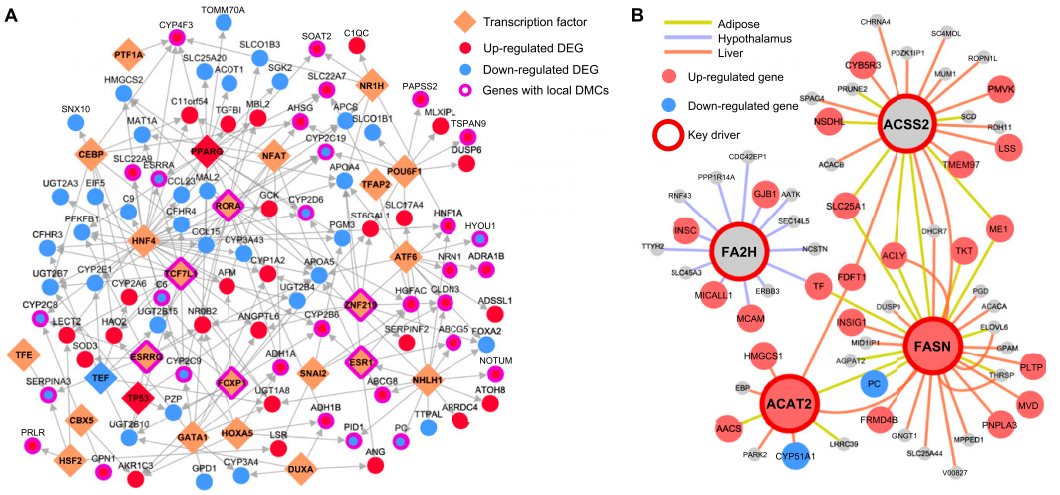


Figure 6. TFs and KDs orchestrate BPA-induced gene expression level changes. (A) Liver TF regulatory networks for the top ranked TFs (FDR <5%) based on enrichment of liver DEGs among TF downstream targets. Network topology was based on FANTOM5. For TFs with >20% overlapping downstream targets, only the TF with the lowest FDR is shown. (B) Gene-gene regulatory subnetworks (BNs) for cross-tissue KDs. Network topology was based on BN modeling of each tissue using genetic and transcriptome datasets from mouse and human populations (57). For each tissue, if two or more datasets were available for a given tissue, a network for each dataset was constructed, and a consensus network was derived by keeping only the high-confidence network edges between genes (edges appearing in two or more studies).

Cross-examination of correlation across gene expression, DNA methylation, and metabolic traits revealed 35 consistent DEG-DMC-trait associations (3 in adipose, 4 in hypothalamus, and 28 in liver) (57). For example, in adipose tissue, *Fasn* (also a perturbed TF hot spot in adipose and a shared KD in adipose and liver) was correlated with its exonic DMC at chr11:120816457, and both were correlated with TG level; in hypothalamus, *Igf1r* (insulin-like growth factor 1 receptor) was correlated with its intronic DMC at chr7:68072768, and both were correlated with blood glucose level; in liver, *Adh1* (alcohol dehydrogenase 1A) was correlated with its intronic DMC at chr3:138287690, and both were correlated with body weight (Fig. 7C). These results suggest that BPA alters local DMCs of certain genes to regulate gene expression, which may in turn regulate distinct metabolic traits.

Relevance of BPA signature to human complex traits/diseases

Human observational studies have associated developmental BPA exposure with a wide variety of human diseases ranging from cardiometabolic diseases to neuropsychiatric disorders (15, 16, 96). Large-scale human GWAS offer an unbiased view of the genetic architecture for various human traits/diseases, and intersections of the molecular footprints of BPA in our mouse study with human disease risk genes can help infer the potential disease-causing properties of BPA in humans. From the GWAS Catalog (74), we collected associated genes for 161 human traits/diseases (traits with <50 associated genes were excluded) and evaluated the enrichment for the trait-associated genes among DEG and DMC signatures. At FDR <5%, no trait was found to be significantly enriched for BPA DEGs. Surprisingly, despite the differences among tissue-specific DMCs (Fig. 5B), 19 out of the 161 traits showed consistently strong enrichment for DMCs across all 3 tissues at FDR <1%. The top traits include body mass index (BMI) and type 2 diabetes (Table 1). As DNA methylation status is known to determine long-term gene expression pattern instead of immediate dynamic gene regulation, the BMI and diabetes-associated genes may be under long-term programming by BPA-induced differential methylation, thereby affecting later disease risks.

The previous analysis involving the GWAS catalog focused only on small sets of the top candidate genes for various diseases and may have limited statistical power. To improve the statistical power, we curated the full summary statistics from 61 human GWAS that are publicly available (covering millions of SNP-trait associations in each GWAS), which enabled us to extend the assessment of disease association by considering additional human disease genes with moderate to low effect sizes (see "Materials and Methods"). This analysis showed that DEGs from all three tissues exhibited consistent enrichment for genes associated with lipid traits such as TGs, low-density lipoprotein cholesterol, and

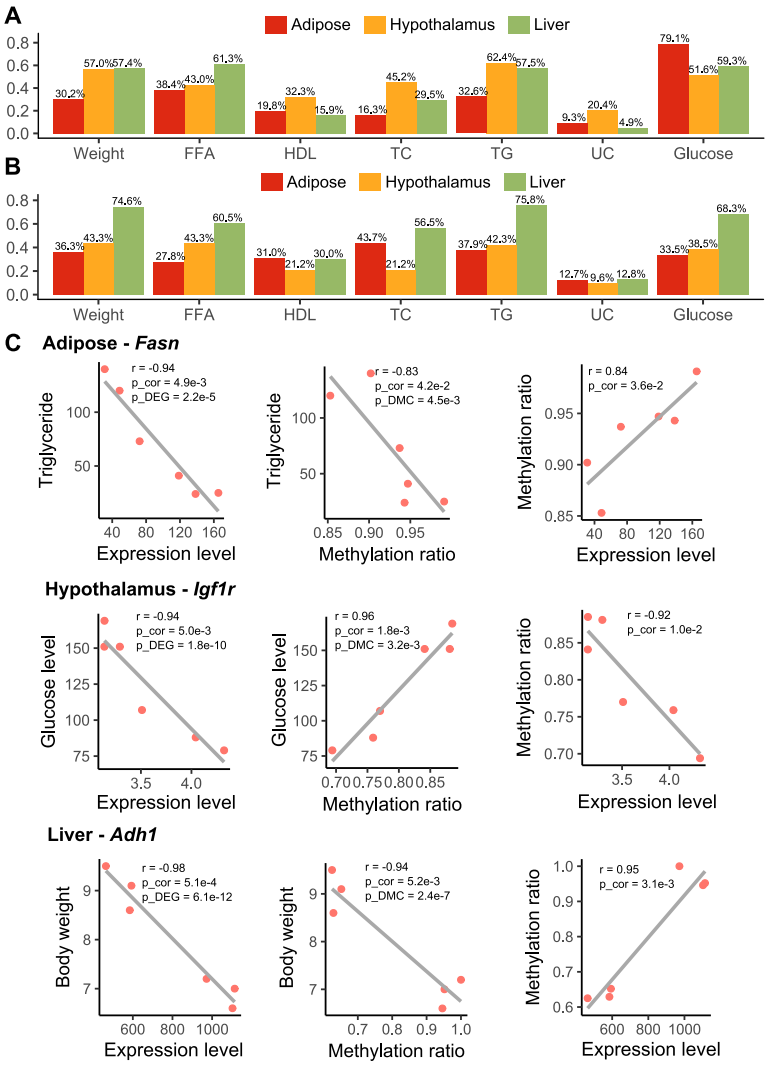


Figure 7. Correlation among gene expression, DNA methylation, and metabolic traits. (A) Percentage of tissue-specific DEGs that are correlated with metabolic traits ($P < 0.05$). (B) Percentage of tissue-specific DMCs that are correlated with metabolic traits ($P < 0.05$). (A and B) *P* values were determined using Pearson correlation test. (C) Pairwise correlation among expression level, methylation ratio, and metabolic profiles (TGs, glucose level, and body weight) for *Fasn*, *Igf1r*, and *Adh1*. p_cor, *P* value was determined using Pearson correction test; p_DEG was determined using differential expression test; p_DMC was determined using differential methylation test. Each dot represents a mouse. TC, total cholesterol; UC, unesterified cholesterol.

HDL cholesterol (Fig. 8A–8C). Interestingly, enrichment for birth weight and birth length was also observed for hypothalamus and liver signatures, respectively. Liver DEGs were also significantly associated with coronary artery disease, inflammatory bowel disease, Alzheimer's disease, and schizophrenia. Top DEGs driving the inflammatory bowel disease association involve immune and inflammatory response genes (*PSMB9*, *TAP1*, and *TNF*), whereas association with Alzheimer's disease and schizophrenia involves genes related to cholesterol homeostasis (*APOA4*, *ABCG8*, and *SOAT2*) and mitochondrial function (*GCDH*, *PDPR*, and *SHMT2*), respectively. These results suggest that tissue-specific targets of BPA are connected to diverse human complex diseases through both the central nervous system and peripheral tissues.

Table 1. Top 5 Human Traits for Which Associated Genes in GWAS Are Enriched for DMCs Across Adipose, Hypothalamus, and Liver at FDR <1% in MSEA

	Adipose		Hypothalamus		Liver	
Human Trait	***P***	**FDR, %**	***P***	**FDR, %**	***P***	**FDR, %**
Obesity-related traits	1.28E-16	0.00	3.03E-15	0.00	2.71E-19	0.00
BMI	1.30E-13	0.00	3.74E-07	0.00	9.66E-12	0.00
Postbronchodilator FEV1/FVC ratio	8.17E-09	0.00	1.45E-08	0.00	3.67E-07	0.00
Type 2 diabetes	1.21E-05	0.03	8.97E-09	0.00	0.001243	0.92
Platelet distribution width	8.16E-08	0.00	7.62E-05	0.16	5.20E-05	0.12

Abbreviations: FEV1, forced expiratory volume in 1 s; FVC, forced vital capacity.

Discussion

This multitissue, multiomics integrative study represents a systems biology investigation of prenatal BPA exposure. By integrating systematic profiling of the transcriptome and methylome of multiple metabolic tissues with phenotypic trait measurements, large-scale human association datasets, and network analysis, we uncovered insights into the molecular regulatory mechanisms underlying the health effects of prenatal BPA exposure. Specifically, we identified tens to thousands of tissue-specific DEGs and DMCs involved in diverse biological functions such as metabolic pathways (oxidative phosphorylation/tricarboxylic acid cycle, fatty acid, cholesterol, glucose metabolism, and PPAR signaling), extracellular matrix, focal adhesion, and inflammation (arachidonic acid) with DMCs partially explaining the regulation of DEGs. Network analysis helped reveal potential regulatory circuits post-BPA exposure and pinpointed both tissue-specific and cross-tissue regulators of BPA activities, including TFs such as estrogen receptors, Pparg, Srebf1, and Hnf1a, and non-TF KDs such as *Fasn*. We also identified previously under-studied targets such as *Cyp51* and lncRNAs across tissues, *Fa2h* in hypothalamus, and *Nfya* in adipose tissue. Furthermore, the BPA gene signatures and the predicted regulators were found to be linked to a wide spectrum of disease-related traits in both mouse and human.

Although our multitissue, multiomics design limits the number of biological replicates we could have for each group, the large-scale disruption we observed in the transcriptome and methylome was consistent with previous reports (37, 40, 78, 97), with a number of differential genes and methylation signals replicating previous findings. Our qPCR results further strengthen the validity of our data. Additionally, we focused our analyses on evaluating the aggregated behavior of BPA signatures using both pathway analysis and network modeling to reduce the potential noise and false positives at an individual gene level, because the random chance to have multiple genes in the same pathway to be false positives is much lower. Indeed, we found a generally higher replication rate of the biological pathways between our study and the other studies than replication between the previous studies. Moreover, our unique study design of examining multiomics in multiple tissues in parallel yields higher comparability when integrating the results between data types and across tissues, as they were from the same set of animals and were profiled in the same conditions. For instance, the much larger numbers of DEGs and more coherent network perturbations revealed in the liver tissue compared with the other two tissues derived from the same set of animals suggest that liver might be a more sensitive target tissue for BPA than the other tissues, although adipose and hypothalamus also appear to be important targets.

Across all three tissues at the transcriptome level, we found that lipid metabolism– and energy homeostasis–related processes were consistently perturbed, with the scale of perturbation being strongest in liver. This aligns well with the signficant changes in the plasma lipid profiles we observed in the offspring, the reported perturbation of lipid metabolism in fetal murine liver (78), and the reported susceptibility for nonalcoholic fatty liver diseases following BPA exposure (79, 98, 99). The only shared gene across tissues, *Cyp51*, encodes a protein that catalyzes metabolic reactions including cholesterol and steroid biosynthesis and biological oxidation (100) and is a critical regulator for testicular spermatogenesis (101). The consistent alteration of *Cyp51* across tissues suggests that this gene is a general target of BPA, with the potential to alter functions related to cholesterol, hormone, and energy metabolism. The liver signature replicated across our and previous studies (75, 76) and a top ranked TF regulator in our TF analysis, *Srebf1*, is a main regulator of lipid homeostasis, again supporting that metabolism is a central target of BPA (75, 76). We also revealed an intriguing link between BPA and lncRNAs across tissues, for which functional importance in developmental processes, disease progression, and response to BPA exposure was increasingly recognized yet underexplored (102). Our molecular data provide intriguing lncRNA

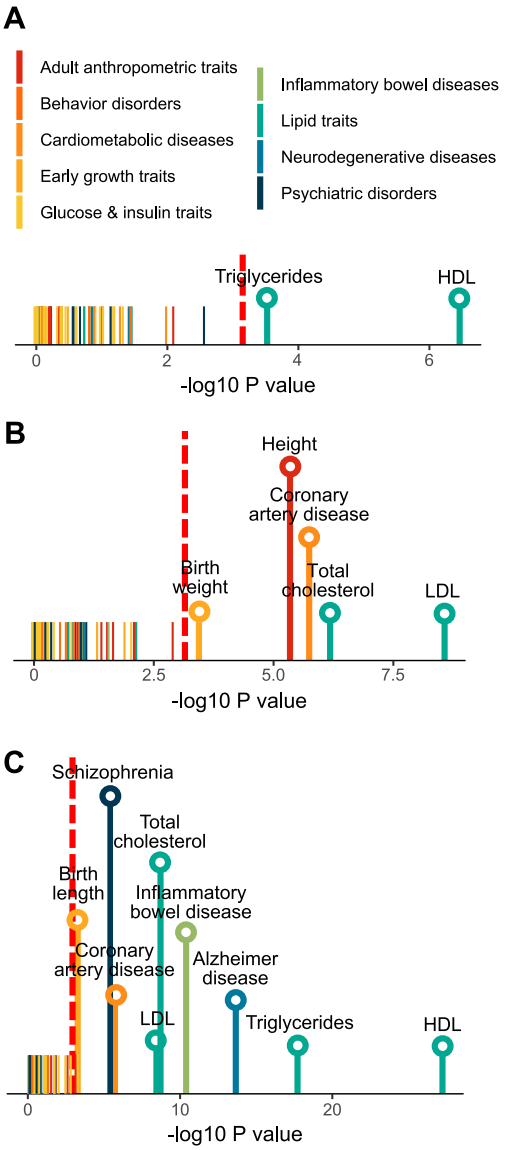


Figure 8. Association of differential expression signatures from (A) adipose, (B) hypothalamus, and (C) liver with 61 human traits/diseases, color-coded into nine primary disease categories. *P* values are determined using MSEA. Red dashed line indicates the cutoff for Bonferroni-corrected $P = 0.05$. Names of traits/diseases for which *P* values did not pass Bonferroni-corrected cutoff are not shown. LDL, low-density lipoprotein.

candidates such as *Gm20319*, *Gm26917*, and *Yam1* for future in-depth functional analyses.

For adipose tissue, clusters of genes responsible for core histones were found to be uniquely altered. Along with the strong adipose-specific differential methylation status, our results revealed gonadal adipose tissues as an especially vulnerable site for BPA-induced epigenetic reprogramming. Besides, developmental BPA exposure has been previously suggested to influence white adipocyte differentiation (86, 103, 104). However, the adipocyte differentiation pathway was not significantly enriched in our study. This is consistent with the report by Angle *et al.* (104), in which increased adipocyte number is only found in mouse offspring with prenatal BPA exposure at 5 μg/kg/d and 500 μg/kg/d, but not 5 mg/kg/d. Additionally, we found significant enrichment for TG biosynthesis and glucose metabolism genes at the differential methylation sites, suggesting that prenatal BPA exposure may affect fat storage and glucose homeostasis in the adipose tissue. Although in this study we mainly investigate gonadal adipose tissue as a surrogate for abdominal fat in the context of metabolic disorders, the information may be useful for exploring the relationship between this fat depot and the gonad.

Concerning the hypothalamus, our study uses next-generation sequencing technology to simultaneously investigate the effect of BPA on the transcriptome and DNA methylome. Hypothalamus is an essential brain region that regulates the endocrine system, peripheral metabolism, and numerous brain functions. We identified BPA-induced DEGs and DMCs that were enriched for extracellular matrix–related processes such as axon guidance, focal adhesion, and various metabolic processes. These hypothalamic pathways have been previously associated with metabolic (50, 51) and neurodegenerative diseases (50, 105), and they could underlie the reported disruption of hypothalamic melanocortin circuitry after BPA exposure (106). Our study highlights the hypothalamus as a critical target for BPA. However, our study used the whole hypothalamus containing heterogeneous nuclei, and future studies to examine individual hypothalamic nuclei such as the arcuate nucleus in the mediobasal hypothalamus will offer better resolution of the specific nuclei and cell types that may be targets of BPA.

By interrogating both the transcriptome and DNA methylome in matching tissues, we were able to directly assess both global and specific correlative relationships between DEGs and DMCs. Specifically, we found that DEGs are more likely to have correlated DMCs in the matching tissue, a trend that persists in nonpromoter regions. Our results corroborate previous findings regarding the importance of gene body methylation in disease etiology (107, 108). Given that >90% of DMCs were found in nonpromoter regions, closer investigation of the regulatory circuits involving these regions may unveil new insights into BPA response (87). Due to the severe multiple testing penalty that limits the statistical

power to assess all pairwise correlations, our analysis was restricted to analyzing local relationships between the two molecular scales. For DMCs with no marked correlation to local gene expression, the underlying reasons could be: (i) the long-range gene regulation by DMCs through three-dimensional organization of nucleus, and (ii) the long-term impact on expression changes by DMCs, which is likely missed in our analysis in which DNA methylation and gene expression are measured at the same time point.

Known as an endocrine-disrupting chemical, BPA has been speculated to exert its primary biological action by modifying the activity of hormone receptors, including estrogen receptors, PPARg, and glucocorticoid receptors (90). Indeed, the activity for the downstream targets of Pparg and three estrogen and estrogen-related receptors were found to be disrupted in the liver by prenatal BPA exposure. More importantly, our unbiased data-driven analysis revealed many additional TFs and non-TF regulatory genes that also likely mediate BPA effects. In fact, many of the identified TF targets of BPA, such as Fasn, Srebf1, and several hepatic nuclear factors, showed much higher ranking in our regulator prediction analyses. In liver, a tightly interconnected TF subnetwork was highly concentrated with BPA-affected genes involved in metabolic processes such as cytochrome P450 system (*Cyp3a25*, *Cyp2a12*, and *Cyp1a2*), lipid (*Apoa4*, *Abcg5*, and *Soat2*), and glucose (*Hnf1a*, *Adra1b*, and *Gck*) regulation, with extensive footprints of altered methylation status in the TFs and other subnetwork genes. Therefore, our results support a widespread impact of BPA on liver transcriptional regulation, and the convergence of differential methylation and gene expression in this TF subnetwork implies that BPA perturbs this subnetwork via epigenetic regulation of the TFs, which in turn trigger transcriptomic alterations in downstream genes. In hypothalamus, we identified *Fa2h* as the strongest KD. This enzyme is highly expressed in the brain and is important for the production of sphingolipids containing 2-hyroxylated fatty acids, the most abundant lipid components of the myelin sheath. Mice lacking *Fa2h* have impaired myelin maintenance (109), and mutations in human *FA2H* have been associated with neurodegeneration (110), hereditary spastic paraplegia (111), and autism (112). In adipose, we discovered a regulatory axis governed by *Nfya* and *Fasn* that are known regulators of fatty acid metabolism and adipogenesis. NF-YA is a histone-fold domain protein that binds to the inverted CCAAT element in the *Fasn* promoter (94, 113), and both *Nfya* and *Fasn* were found to be significantly perturbed by BPA in our study. Moreover, *Fasn* also serves as a cross-tissue KD, governing distinct groups of upregulated lipid metabolism genes in adipose and liver post-BPA exposure, supporting its role in mediating the BPA-induced lipid dysregulation at the systemic level. A previous study has also shown BPA-induced effects on *Fasn* methylation after perinatal exposure (97). The significant correlation of gene expression and methylation for *Fasn* with TG level furthers implicates its role as a network-level regulator and biomarker for BPA-induced lipid dysregulation. Our observation of *Fasn* is consistent with evidence suggesting its susceptibility to methylation perturbation under obesogenic feeding (114) and its causal functional importance for fatty liver diseases (52, 115). The causal regulatory role of these genes in BPA activities warrants future testing via genetic manipulation studies, such as knocking down or overexpressing the predicted regulators to examine their ability to modulate BPA activities.

One unique aspect of this study is the linking of the molecular landscape of prenatal BPA exposure to traits/diseases in both mouse and human. In our mouse study, the observed changes in body weight, lipid profiles, and glucose level are highly concordant with the functions of the molecular targets. For instance, prenatal BPA exposure perturbs both the expression levels and local DNA methylation status of *Fasn*, *Igf1r*, and *Adh1*. These DEGs and their local DMCs also significantly correlate with phenotypic outcomes, thus serving as examples of how DNA methylation and gene regulation bridge the gap between BPA exposure and phenotypic manifestation. To further enhance the translatability of our findings from mouse to human, we searched for human diseases linked to the BPA-affected genes. An intriguing discovery is the prominent overrepresentation of differential methylation signals in adipose, hypothalamus, and liver within known genes related to obesity and type 2 diabetes, supporting that BPA may affect obesity and diabetes risk through systemic reprogramming of DNA methylation. More sophisticated analyses incorporating the BPA differential gene expression and the full statistics of human GWAS corroborated the observed connection between prenatal BPA exposure and lipid homeostasis (116), birth weight (117), and coronary artery disease (15) reported in observational studies. Moreover, our findings suggest the involvement of prenatal BPA exposure in the development of inflammatory bowel syndrome, schizophrenia, and Alzheimer's disease. These associations warrant future investigations.

Designed as the discovery phase of a comprehensive investigation of *in vivo* BPA activities, our study opens numerous future lines of investigation. First, our current molecular studies focused on male tissues because of the stronger phenotypes observed in males. Our phenotypic examination and qPCR experiments on females support much subtler changes in females, and future studies will

require larger sample sizes to uncover female-specific biology. Secondly, our study design does not address whether the observed BPA genomic effects are direct or indirect, and radiotracing or substrate-binding experiments are needed to elucidate this question. Thirdly, the causal link between the genomic effects observed and the phenotypes that result from BPA exposure is not established, and genetic perturbation experiments are required to test the causal roles of the predicted regulators of BPA actions. Gene annotation accuracy may also affect the results and interpretation. Lastly, we tested *in utero* BPA exposure at one dose via oral gavage, which can cause prenatal stress and confound the results, and examined phenotypes and molecular profiles only at weaning age as a proof-of-concept for our systems biology framework. Considering that the effects of early-life exposure to BPA are highly variable and dependent on factors such as the dose, window, route (*e.g.*, using food as an alternative), and frequency of exposure as well as genetic background, age, and sex (14), future studies testing these additional variables using large sample sizes are necessary to generate a comprehensive understandings of BPA risks under various exposure conditions.

In summary, our study represents a multitissue, multiomics integrative investigation of prenatal BPA exposure. The systems biology framework we applied revealed how BPA triggers cascades of regulatory circuits involving numerous TFs and non-TF regulators that coordinate diverse molecular processes within and across core metabolic tissues, thereby highlighting that BPA exerts its biological functions via much more diverse targets than previously thought. As such, our findings offer a comprehensive systems-level understanding of tissue sensitivity and molecular perturbations elicited by prenatal BPA exposure and offer promising candidates for targeted mechanistic investigation as well as much-needed network-level biomarkers of prior BPA exposure. The strong influence of BPA on metabolic pathways and cardiometabolic phenotypes merits its characterization as a general metabolic disruptor posing systemic health risks.

Acknowledgments

We thank Zhe Ying for assistance in collecting mice hypothalamus tissue and Dr. Guanglin Zhang for assistance in the RRBS experiments.

Financial Support: L.S. is supported by a University of California, Los Angeles, Dissertation Year Fellowship, Eureka Scholarship, Hyde Scholarship, Burroughs Wellcome Fund Inter-School Program in Metabolic Diseases Fellowship, and the China Scholarship Council. G.D. is supported by NIEHS/National Institutes of Health Grant T32ES015457. P.A. is supported by National Institutes of Health/NIEHS Grant R01-ES02748701 and the Burroughs Wellcome Foundation. X.Y. is supported by National Institutes of Health Grant DK104363 and the Leducq Foundation.

Correspondence: Xia Yang, PhD, Department of Integrative Biology and Physiology, University of California, Los Angeles, Terasaki Life Sciences Building 2000D, Los Angeles, California 90095. E-mail: xyang123@ucla.edu.

Disclosure Summary: The authors have nothing to disclose.

References

1. Barouki R, Gluckman PD, Grandjean P, Hanson M, Heindel JJ. Developmental origins of non-communicable disease: implications for research and public health. *Environ Health*. 2012;**11**:42.
2. Boekelheide K, Blumberg B, Chapin RE, Cote I, Graziano JH, Janesick A, Lane R, Lillycrop K, Myatt L, States JC, Thayer KA, Waalkes MP, Rogers JM. Predicting later-life outcomes of early-life exposures. *Environ Health Perspect*. 2012;**120**(10):1353–1361.
3. Heindel JJ, Vandenberg LN. Developmental origins of health and disease: a paradigm for understanding disease cause and prevention. *Curr Opin Pediatr*. 2015;**27**(2):248–253.
4. Calafat AM, Ye X, Wong LY, Reidy JA, Needham LL. Exposure of the U.S. population to bisphenol A and 4-tertiary-octylphenol: 2003-2004. *Environ Health Perspect*. 2008;**116**(1):39–44.
5. Haugen AC, Schug TT, Collman G, Heindel JJ. Evolution of DOHaD: the impact of environmental health sciences. *J Dev Orig Health Dis*. 2015;**6**(2):55–64.
6. Tsai WT. Human health risk on environmental exposure to bisphenol-A: a review. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev*. 2006;**24**(2):225–255.
7. Sun C, Leong LP, Barlow PJ, Chan SH, Bloodworth BC. Single laboratory validation of a method for the determination of Bisphenol A, Bisphenol A diglycidyl ether and its derivatives in canned foods by reversed-phase liquid chromatography. *J Chromatogr A*. 2006;**1129**(1):145–148.
8. Vandenberg LN, Hauser R, Marcus M, Olea N, Welshons WV. Human exposure to bisphenol A (BPA). *Reprod Toxicol*. 2007;**24**(2):139–177.
9. Rubin BS, Paranjpe M, DaFonte T, Schaeberle C, Soto AM, Obin M, Greenberg AS. Perinatal BPA exposure alters body weight and composition in a dose specific and sex specific manner: The addition of peripubertal exposure exacerbates adverse effects in female mice. *Reprod Toxicol*. 2017;**68**:130–144.
10. Hao M, Ding L, Xuan L, Wang T, Li M, Zhao Z, Lu J, Xu Y, Chen Y, Wang W, Bi Y, Xu M, Ning G. Urinary bisphenol A concentration and the risk of central obesity in Chinese adults: A prospective study. *J Diabetes*. 2018;**10**(6):442–448.
11. Beydoun HA, Khanal S, Zonderman AB, Beydoun MA. Sex differences in the association of urinary bisphenol-A concentration with selected indices of glucose homeostasis among U.S. adults. *Ann Epidemiol*. 2014;**24**(2):90–97.
12. Teppala S, Madhavan S, Shankar A. Bisphenol A and metabolic syndrome: Results from NHANES. *Int J Endocrinol*. 2012;**2012**:598180.
13. Mouneimne Y, Nasrallah M, Khoueiry-Zgheib N, Nasreddine L, Nakhoul N, Ismail H, Abiad M, Koleilat L, Tamim H. Bisphenol A urinary level, its correlates, and association with cardiometabolic risks in Lebanese urban adults. *Environ Monit Assess*. 2017;**189**(10):517.
14. Wassenaar PNH, Trasande L, Legler J. Systematic review and meta-analysis of early-life exposure to bisphenol A and obesity-related outcomes in rodents. *Environ Health Perspect*. 2017;**125**(10):106001.

15. Han C, Hong YC. Bisphenol A, hypertension, and cardiovascular diseases: epidemiological, laboratory, and clinical trial evidence. *Curr Hypertens Rep*. 2016;**18**(2):11.
16. Ranciere F, Lyons JG, Loh VH, Botton J, Galloway T, Wang T, Shaw JE, Magliano DJ. Bisphenol A and the risk of cardiometabolic disorders: a systematic review with meta-analysis of the epidemiological evidence. *Environ Health*. 2015;**14**:46.
17. Stahlhut RW, Myers JP, Taylor JA, Nadal A, Dyer JA, vom Saal FS. Experimental BPA exposure and glucose-stimulated insulin response in adult men and women. *J Endocrine Soc*. 2018;**2**(10):1173–1187.
18. Liu J, Yu P, Qian W, Li Y, Zhao J, Huan F, Wang J, Xiao H. Perinatal bisphenol A exposure and adult glucose homeostasis: identifying critical windows of exposure. *PLoS One*. 2013;**8**(5):e64143.
19. Ryan KK, Haller AM, Sorrell JE, Woods SC, Jandacek RJ, Seeley RJ. Perinatal exposure to bisphenol-a and the development of metabolic syndrome in CD-1 mice. *Endocrinology*. 2010;**151**(6):2603–2612.
20. Miyawaki J, Sakayama K, Kato H, Yamamoto H, Masuno H. Perinatal and postnatal exposure to bisphenol A increases adipose tissue mass and serum cholesterol level in mice. *J Atheroscler Thromb*. 2007;**14**(5):245–252.
21. Rubin BS, Soto AM. Bisphenol A: Perinatal exposure and body weight. *Mol Cell Endocrinol*. 2009;**304**(1-2):55–62.
22. García-Arevalo M, Alonso-Magdalena P, Rebelo Dos Santos J, Quesada I, Carneiro EM, Nadal A. Exposure to bisphenol-A during pregnancy partially mimics the effects of a high-fat diet altering glucose homeostasis and gene expression in adult male mice. *PLoS One*. 2014;**9**(6):e100214.
23. Manikkam M, Tracey R, Guerrero-Bosagna C, Skinner MK. Plastics derived endocrine disruptors (BPA, DEHP and DBP) induce epigenetic transgenerational inheritance of obesity, reproductive disease and sperm epimutations. *PLoS One*. 2013;**8**(1):e55387.
24. Susiarjo M, Xin F, Bansal A, Stefaniak M, Li C, Simmons RA, Bartolomei MS. Bisphenol A exposure disrupts metabolic health across multiple generations in the mouse. *Endocrinology*. 2015;**156**(6):2049–2058.
25. Bansal A, Rashid C, Xin F, Li C, Polyak E, Duemler A, van der Meer T, Stefaniak M, Wajid S, Doliba N, Bartolomei MS, Simmons RA. Sex- and dose-specific effects of maternal bisphenol A exposure on pancreatic islets of first- and second-generation adult mice offspring. *Environ Health Perspect*. 2017;**125**(9):097022.
26. Camacho J, Truong L, Kurt Z, Chen YW, Morselli M, Gutierrez G, Pellegrini M, Yang X, Allard P. The memory of environmental chemical exposure in C. elegans is dependent on the Jumonji demethylases jmjd-2 and jmjd-3/utx-1. *Cell Reports*. 2018;**23**(8):2392–2404.
27. Baillie-Hamilton PF. Chemical toxins: a hypothesis to explain the global obesity epidemic. *J Altern Complement Med*. 2002;**8**(2):185–192.
28. Heindel JJ. Endocrine disruptors and the obesity epidemic. *Toxicol Sci*. 2003;**76**(2):247–249.
29. Newbold RR, Padilla-Banks E, Jefferson WN, Heindel JJ. Effects of endocrine disruptors on obesity. *Int J Androl*. 2008;**31**(2):201–208.
30. European Food Safety Authority. Scientific opinion on the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs. *EFSA J*. 2015;**13**(1):3978.
31. Bisphenol A (BPA) Joint Emerging Science Working Group; Department of Health and Human Services. 2014 updated review of literature and data on bisphenol A (CAS RN 80-05-7). https://www.fda.gov/downloads/food/ingredientspackaginglabeling/foodadditivesingredients/ucm424071.pdf. Published 6 June 2014. Accessed 8 October 2017.
32. National Toxicology Program. NTP research report on the CLARITY-BPA core study: a perinatal and chronic extended-dose-range study of bisphenol A in rats. NTP RR 9. https://doi.org/10.22427/NTP-RR-9. Published September 2018. Accessed 16 November 2018.
33. Beronius A, Johansson N, Rudén C, Hanberg A. The influence of study design and sex-differences on results from developmental neurotoxicity studies of bisphenol A: implications for toxicity testing. *Toxicology*. 2013;**311**(1-2):13–26.
34. Ariemma F, D'Esposito V, Liguoro D, Oriente F, Cabaro S, Liotti A, Cimmino I, Longo M, Beguinot F, Formisano P, Valentino R. Low-dose bisphenol-A impairs adipogenesis and generates dysfunctional 3T3-L1 adipocytes. *PLoS One*. 2016;**11**(3):e0150762.
35. Ben-Jonathan N, Hugo ER, Brandebourg TD. Effects of bisphenol A on adipokine release from human adipose tissue: Implications for the metabolic syndrome. *Mol Cell Endocrinol*. 2009;**304**(1-2):49–54.
36. Olsvik PA, Skjærven KH, Søfteland L. Metabolic signatures of bisphenol A and genistein in Atlantic salmon liver cells. *Chemosphere*. 2017;**189**:730–743.
37. Lejonklou MH, Dunder L, Bladin E, Pettersson V, Rönn M, Lind L, Waldén TB, Lind PM. Effects of low-dose developmental bisphenol A exposure on metabolic parameters and gene expression in male and female Fischer 344 rat offspring. *Environ Health Perspect*. 2017;**125**(6):067018.
38. Anderson OS, Kim JH, Peterson KE, Sanchez BN, Sant KE, Sartor MA, Weinhouse C, Dolinoy DC. Novel epigenetic biomarkers mediating bisphenol A exposure and metabolic phenotypes in female mice. *Endocrinology*. 2017;**158**(1):31–40.
39. Ma Y, Xia W, Wang DQ, Wan YJ, Xu B, Chen X, Li YY, Xu SQ. Hepatic DNA methylation modifications in early development of rats resulting from perinatal BPA exposure contribute to insulin resistance in adulthood. *Diabetologia*. 2013;**56**(9):2059–2067.
40. Taylor JA, Shioda K, Mitsunaga S, Yawata S, Angle BM, Nagel SC, Vom Saal FS, Shioda T. Prenatal exposure to bisphenol A disrupts naturally occurring bimodal DNA methylation at proximal promoter of fggy, an obesity-relevant gene encoding a carbohydrate kinase, in gonadal white adipose tissues of CD-1 mice. *Endocrinology*. 2018;**159**(2):779–794.
41. Faulk C, Kim JH, Jones TR, McEachin RC, Nahar MS, Dolinoy DC, Sartor MA. Bisphenol A-associated alterations in genome-wide DNA methylation and gene expression patterns reveal sequence-dependent and non-monotonic effects in human fetal liver. *Environ Epigenet*. 2015;**1**(1):dvv006.
42. Nahar MS, Kim JH, Sartor MA, Dolinoy DC. Bisphenol A-associated alterations in the expression and epigenetic regulation of genes encoding xenobiotic metabolizing enzymes in human fetal liver. *Environ Mol Mutagen*. 2014;**55**(3):184–195.
43. Wang T, Pehrsson EC, Purushotham D, Li D, Zhuo X, Zhang B, Lawson HA, Province MA, Krapp C, Lan Y, Coarfa C, Katz TA, Tang WY, Wang Z, Biswal S, Rajagopalan S, Colacino JA, Tsai ZT, Sartor MA, Neier K, Dolinoy DC, Pinto J, Hamanaka RB, Mutlu GM, Patisaul HB, Aylor DL, Crawford GE, Wiltshire T, Chadwick LH, Duncan CG, Garton AE, McAllister KA, Bartolomei MS, Walker CL, Tyson FL; TaRGET II Consortium. The NIEHS TaRGET II Consortium and environmental epigenomics. *Nat Biotechnol*. 2018;**36**(3):225–227.
44. Messerlian C, Martinez RM, Hauser R, Baccarelli AA. 'Omics' and endocrine-disrupting chemicals - new paths forward. *Nat Rev Endocrinol*. 2017;**13**(12):740–748.
45. López M, Nogueiras R, Tena-Sempere M, Diéguez C. Hypothalamic AMPK: a canonical regulator of whole-body energy balance. *Nat Rev Endocrinol*. 2016;**12**(7):421–432.
46. Rui L. Energy metabolism in the liver. *Compr Physiol*. 2014;**4**(1):177–197.
47. Choe SS, Huh JY, Hwang IJ, Kim JI, Kim JB. Adipose tissue remodeling: its role in energy metabolism and metabolic disorders. *Front Endocrinol (Lausanne)*. 2016;**7**:30.
48. Coelho M, Oliveira T, Fernandes R. Biochemistry of adipose tissue: an endocrine organ. *Arch Med Sci*. 2013;**9**(2):191–200.

49. Mäkinen VP, Civelek M, Meng Q, Zhang B, Zhu J, Levian C, Huan T, Segrè AV, Ghosh S, Vivar J, Nikpay M, Stewart AF, Nelson CP, Willenborg C, Erdmann J, Blakenberg S, O'Donnell CJ, März W, Laaksonen R, Epstein SE, Kathiresan S, Shah SH, Hazen SL, Reilly MP, Lusis AJ, Samani NJ, Schunkert H, Quertermous T, McPherson R, Yang X, Assimes TL; Coronary ARtery DIsease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) Consortium. Integrative genomics reveals novel molecular pathways and gene networks for coronary artery disease. *PLoS Genet*. 2014;**10**(7):e1004502.
50. Meng Q, Ying Z, Noble E, Zhao Y, Agrawal R, Mikhail A, Zhuang Y, Tyagi E, Zhang Q, Lee JH, Morselli M, Orozco L, Guo W, Kilts TM, Zhu J, Zhang B, Pellegrini M, Xiao X, Young MF, Gomez-Pinilla F, Yang X. Systems nutrigenomics reveals brain gene networks linking metabolic and brain disorders. *EBioMedicine*. 2016;**7**:157–166.
51. Shu L, Chan KHK, Zhang G, Huan T, Kurt Z, Zhao Y, Codoni V, Trégouët DA, Yang J, Wilson JG, Luo X, Levy D, Lusis AJ, Liu S, Yang X; Cardiogenics Consortium. Shared genetic regulatory networks for cardiovascular disease and type 2 diabetes in multiple populations of diverse ethnicities in the United States. *PLoS Genet*. 2017;**13**(9):e1007040.
52. Chella Krishnan K, Kurt Z, Barrere-Cain R, Sabir S, Das A, Floyd R, Vergnes L, Zhao Y, Che N, Charugundla S, Qi H, Zhou Z, Meng Y, Pan C, Seldin MM, Norheim F, Hui S, Reue K, Lusis AJ, Yang X. Integration of multi-omics data from mouse diversity panel highlights mitochondrial dysfunction in non-alcoholic fatty liver disease. *Cell Syst*. 2018;**6**(1):103–115.e7.
53. Dolinoy DC, Huang D, Jirtle RL. Maternal nutrient supplementation counteracts bisphenol A-induced DNA hypomethylation in early development. *Proc Natl Acad Sci USA*. 2007;**104**(32):13056–13061.
54. Susiarjo M, Sasson I, Mesaros C, Bartolomei MS. Bisphenol A exposure disrupts genomic imprinting in the mouse. *PLoS Genet*. 2013;**9**(4):e1003401.
55. Bromer JG, Zhou Y, Taylor MB, Doherty L, Taylor HS. Bisphenol-A exposure in utero leads to epigenetic alterations in the developmental programming of uterine estrogen response. *FASEB J*. 2010;**24**(7):2273–2280.
56. Kitchin KT, Ebron MT. Further development of rodent whole embryo culture: solvent toxicity and water insoluble compound delivery system. *Toxicology*. 1984;**30**(1):45–57.
57. Shu L, Meng Q, Diamante G, Tsai B, Chen Y-W, Mikhail A, Luk H, Ritz B, Allard P, Yang X. Data from: Prenatal bisphenol A exposure in mice induces multitissue multiomics disruptions linking to cardiometabolic disorders. figshare 2018. Deposited 12 November 2018. http://doi.org/10.6084/m9.figshare.7451069.v2.
58. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol*. 2014;**32**(9):903–914.
59. Meng Q, Zhuang Y, Ying Z, Agrawal R, Yang X, Gomez-Pinilla F. Traumatic brain injury induces genome-wide transcriptomic, methylomic, and network Perturbations in brain and blood predicting neurological disorders. *EBioMedicine*. 2017;**16**:184–194.
60. Chen Y, Shu L, Qiu Z, Lee DY, Settle SJ, Que Hee S, Telesca D, Yang X, Allard P. Exposure to the BPA-substitute Bisphenol S causes unique alterations of germline function. *PLoS Genet*. 2016;**12**(7):e1006223.
61. Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*. 2014;**30**(3):301–304.
62. Andrews S. Fast QC: a quality control tool for high throughput sequence data. Available at: www.bioinformatics.babraham.ac.uk/projects. Accessed 10 December 2015.
63. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc*. 2016;**11**(9):1650–1667.
64. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;**15**(12):550.
65. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*. 2003;**100**(16):9440–9445.
66. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*. 2009;**10**(1):232.
67. Sun D, Xi Y, Rodriguez B, Park HJ, Tong P, Meong M, Goodell MA, Li W. MOABS: model based analysis of bisulfite sequencing data. *Genome Biol*. 2014;**15**(2):R38.
68. Cavalcante RG, Sartor MA. annotatr: genomic regions in context. *Bioinformatics*. 2017;**33**(15):2381–2383.
69. Shu L, Zhao Y, Kurt Z, Byars SG, Tukiainen T, Kettunen J, Orozco LD, Pellegrini M, Lusis AJ, Ripatti S, Zhang B, Inouye M, Mäkinen VP, Yang X. Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems. *BMC Genomics*. 2016;**17**(1):874.
70. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000;**28**(1):27–30.
71. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky V, Song H, Williams M, Birney E, Hermjakob H, Stein L, D'Eustachio P. The Reactome pathway knowledgebase. *Nucleic Acids Res*. 2014;**42**(Database issue):D472–D477.
72. Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat Methods*. 2016;**13**(4):366–370.
73. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;**13**(11):2498–2504.
74. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*. 2014;**42**(Database issue):D1001–D1006.
75. Marmugi A, Ducheix S, Lasserre F, Polizzi A, Paris A, Priymenko N, Bertrand-Michel J, Pineau T, Guillou H, Martin PG, Mselli-Lakhal L. Low doses of bisphenol A induce gene expression related to lipid synthesis and trigger triglyceride accumulation in adult mouse liver. *Hepatology*. 2012;**55**(2):395–407.
76. Melis JP, Derks KW, Pronk TE, Wackers P, Schaap MM, Zwart E, van Ijcken WF, Jonker MJ, Breit TM, Pothof J, van Steeg H, Luijten M. In vivo murine hepatic microRNA and mRNA expression signatures predicting the (non-)genotoxic carcinogenic potential of chemicals. *Arch Toxicol*. 2014;**88**(4):1023–1034.
77. Meng Z, Wang D, Yan S, Li R, Yan J, Teng M, Zhou Z, Zhu W. Effects of perinatal exposure to BPA and its alternatives (BPS, BPF and BPAF) on hepatic lipid and glucose homeostasis in female mice adolescent offspring. *Chemosphere*. 2018;**212**:297–306.
78. Ilagan Y, Mamillapalli R, Goetz LG, Kayani J, Taylor HS. Bisphenol-A exposure in utero programs a sexually dimorphic estrogenic state of hepatic metabolic gene expression. *Reprod Toxicol*. 2017;**71**:84–94.
79. Shimpi PC, More VR, Paranjpe M, Donepudi AC, Goodrich JM, Dolinoy DC, Rubin B, Slitt AL. Hepatic lipid accumulation and Nrf2 expression following perinatal and peripubertal exposure to bisphenol A in a mouse model of nonalcoholic liver disease. *Environ Health Perspect*. 2017;**125**(8):087005.
80. Susiarjo M, Xin F, Stefaniak M, Mesaros C, Simmons RA, Bartolomei MS. Bile acids and tryptophan metabolism are novel pathways involved in metabolic abnormalities in BPA-exposed pregnant mice and male offspring. *Endocrinology*. 2017;**158**(8):2533–2542.
81. Wang D, Zhu W, Yan S, Meng Z, Yan J, Teng M, Jia M, Li R, Zhou Z. Impaired lipid and glucose homeostasis in male mice

offspring after combined exposure to low-dose bisphenol A and arsenic during the second half of gestation. *Chemosphere*. 2018; **210**:998–1005.
82. Nishizawa H, Imanishi S, Manabe N. Effects of exposure in utero to bisphenol A on the expression of aryl hydrocarbon receptor, related factors, and xenobiotic metabolizing enzymes in murine embryos. *J Reprod Dev*. 2005;**51**(5):593–605.
83. Johnson SA, Spollen WG, Manshack LK, Bivens NJ, Givan SA, Rosenfeld CS. Hypothalamic transcriptomic alterations in male and female California mice (*Peromyscus californicus*) developmentally exposed to bisphenol A or ethinyl estradiol. *Physiol Rep*. 2017;**5**(3):e13133.
84. Arambula SE, Belcher SM, Planchart A, Turner SD, Patisaul HB. Impact of low dose oral exposure to bisphenol A (BPA) on the neonatal rat hypothalamic and hippocampal transcriptome: a CLARITY-BPA Consortium study. *Endocrinology*. 2016;**157**(10): 3856–3872.
85. Cheong A, Johnson SA, Howald EC, Ellersieck MR, Camacho L, Lewis SM, Vanlandingham MM, Ying J, Ho SM, Rosenfeld CS. Gene expression and DNA methylation changes in the hypothalamus and hippocampus of adult rats developmentally exposed to bisphenol A or ethinyl estradiol: a CLARITY-BPA consortium study. *Epigenetics*. 2018;**13**(7):704–720.
86. Somm E, Schwitzgebel VM, Toulotte A, Cederroth CR, Combescure C, Nef S, Aubert ML, Hüppi PS. Perinatal exposure to bisphenol A alters early adipogenesis in the rat. *Environ Health Perspect*. 2009;**117**(10):1549–1555.
87. Lou S, Lee HM, Qin H, Li JW, Gao Z, Liu X, Chan LL, Kl Lam V, So WY, Wang Y, Lok S, Wang J, Ma RC, Tsui SK, Chan JC, Chan TF, Yip KY. Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. *Genome Biol*. 2014; **15**(7):408.
88. Faulk C, Kim JH, Anderson OS, Nahar MS, Jones TR, Sartor MA, Dolinoy DC. Detection of differential DNA methylation in repetitive DNA of mice and humans perinatally exposed to bisphenol A. *Epigenetics*. 2016;**11**(7):489–500.
89. Acconcia F, Pallottini V, Marino M. Molecular mechanisms of action of BPA. *Dose Response*. 2015;**13**(4):1559325815610582.
90. MacKay H, Abizaid A. A plurality of molecular targets: The receptor ecosystem for bisphenol-A (BPA). *Horm Behav*. 2018; **101**:59–67.
91. Wang J, Sun B, Hou M, Pan X, Li X. The environmental obesogen bisphenol A promotes adipogenesis by increasing the amount of 11β-hydroxysteroid dehydrogenase type 1 in the adipose tissue of children. *Int J Obes*. 2013;**37**(7):999–1005.
92. Ahmed S, Atlas E. Bisphenol S- and bisphenol A-induced adipogenesis of murine preadipocytes occurs through direct peroxisome proliferator-activated receptor gamma activation. *Int J Obes*. 2016;**40**(10):1566–1573.
93. Vafeiadi M, Roumeliotaki T, Myridakis A, Chalkiadaki G, Fthenou E, Dermitzaki E, Karachaliou M, Sarri K, Vassilaki M, Stephanou EG, Kogevinas M, Chatzi L. Association of early life exposure to bisphenol A with obesity and cardiometabolic traits in childhood. *Environ Res*. 2016;**146**:379–387.
94. Nishi-Tatsumi M, Yahagi N, Takeuchi Y, Toya N, Takarada A, Murayama Y, Aita Y, Sawada Y, Piao X, Oya Y, Shikama A, Masuda Y, Kubota M, Izumida Y, Matsuzaka T, Nakagawa Y, Sekiya M, Iizuka Y, Kawakami Y, Kadowaki T, Yamada N, Shimano H. A key role of nuclear factor Y in the refeeding response of fatty acid synthase in adipocytes. *FEBS Lett*. 2017; **591**(7):965–978.
95. Zhang B, Gaiteri C, Bodea LG, Wang Z, McElwee J, Podtelezhnikov AA, Zhang C, Xie T, Tran L, Dobrin R, Fluder E, Clurman B, Melquist S, Narayanan M, Suver C, Shah H, Mahajan M, Gillis T, Mysore J, MacDonald ME, Lamb JR, Bennett DA, Molony C, Stone DJ, Gudnason V, Myers AJ, Schadt EE, Neumann H, Zhu J, Emilsson V. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell*. 2013;**153**(3):707–720.
96. Inadera H. Neurological effects of bisphenol A and its analogues. *Int J Med Sci*. 2015;**12**(12):926–936.
97. Kim JH, Sartor MA, Rozek LS, Faulk C, Anderson OS, Jones TR, Nahar MS, Dolinoy DC. Perinatal bisphenol A exposure promotes dose-dependent alterations of the mouse methylome. *BMC Genomics*. 2014;**15**(1):30.
98. Ke ZH, Pan JX, Jin LY, Xu HY, Yu TT, Ullah K, Rahman TU, Ren J, Cheng Y, Dong XY, Sheng JZ, Huang HF. Bisphenol A exposure may induce hepatic lipid accumulation via reprogramming the DNA methylation patterns of genes involved in lipid metabolism. *Sci Rep*. 2016;**6**(1):31331.
99. Yang S, Zhang A, Li T, Gao R, Peng C, Liu L, Cheng Q, Mei M, Song Y, Xiang X, Wu C, Xiao X, Li Q. Dysregulated autophagy in hepatocytes promotes bisphenol A-induced hepatic lipid accumulation in male mice. *Endocrinology*. 2017;**158**(9):2799–2812.
100. Lewinska M, Juvan P, Perse M, Jeruc J, Kos S, Lorbek G, Urlep Z, Keber R, Horvat S, Rozman D. Hidden disease susceptibility and sexual dimorphism in the heterozygous knockout of Cyp51 from cholesterol synthesis. *PLoS One*. 2014;**9**(11):e112787.
101. Keber R, Rozman D, Horvat S. Sterols in spermatogenesis and sperm maturation. *J Lipid Res*. 2013;**54**(1):20–33.
102. Karlsson O, Baccarelli AA. Environmental health and long non-coding RNAs. *Curr Environ Health Rep*. 2016;**3**(3): 178–187.
103. Vom Saal FS, Nagel SC, Coe BL, Angle BM, Taylor JA. The estrogenic endocrine disrupting chemical bisphenol A (BPA) and obesity. *Mol Cell Endocrinol*. 2012;**354**(1-2):74–84.
104. Angle BM, Do RP, Ponzi D, Stahlhut RW, Drury BE, Nagel SC, Welshons WV, Besch-Williford CL, Palanza P, Parmigiani S, vom Saal FS, Taylor JA. Metabolic disruption in male mice due to fetal exposure to low but not high doses of bisphenol A (BPA): evidence for effects on body weight, food intake, adipocytes, leptin, adiponectin, insulin and glucose regulation. *Reprod Toxicol*. 2013; **42**:256–268.
105. Vercruysse P, Vieau D, Blum D, Petersén Å, Dupuis L. Hypothalamic alterations in neurodegenerative diseases and their relation to abnormal energy metabolism. *Front Mol Neurosci*. 2018;**11**:2.
106. MacKay H, Patterson ZR, Abizaid A. Perinatal exposure to low-dose bisphenol-A disrupts the structural and functional development of the hypothalamic feeding circuitry. *Endocrinology*. 2017;**158**(4):768–777.
107. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*. 2012;**13**(7):484–492.
108. Patil V, Ward RL, Hesson LB. The evidence for functional non-CpG methylation in mammalian cells. *Epigenetics*. 2014;**9**(6): 823–828.
109. Zöller I, Meixner M, Hartmann D, Büssow H, Meyer R, Gieselmann V, Eckhardt M. Absence of 2-hydroxylated sphingolipids is compatible with normal neural development but causes late-onset axon and myelin sheath degeneration. *J Neurosci*. 2008;**28**(39):9741–9754.
110. Zaki MS, Selim L, Mansour L, Mahmoud IG, Fenstermaker AG, Gabriel SB, Gleeson JG. Mutations in FA2H in three Arab families with a clinical spectrum of neurodegeneration and hereditary spastic paraparesis. *Clin Genet*. 2015;**88**(1):95–97.
111. Liao X, Luo Y, Zhan Z, Du J, Hu Z, Wang J, Guo J, Hu Z, Yan X, Pan Q, Xia K, Tang B, Shen L. SPG35 contributes to the second common subtype of AR-HSP in China: frequency analysis and functional characterization of FA2H gene mutations. *Clin Genet*. 2015;**87**(1):85–89.
112. Scheid I, Maruani A, Huguet G, Leblond CS, Nygren G, Anckarsäter H, Beggiato A, Rastam M, Amsellem F, Gillberg IC, Elmaleh M, Leboyer M, Gillberg C, Betancur C, Coleman M, Hama H, Cook EH, Bourgeron T, Delorme R. Heterozygous FA2H mutations in autism spectrum disorders. *BMC Med Genet*. 2013;**14**(1):124.

113. Oldfield AJ, Yang P, Conway AE, Cinghu S, Freudenberg JM, Yellaboina S, Jothi R. Histone-fold domain protein NF-Y promotes chromatin accessibility for cell type-specific master transcription factors. *Mol Cell*. 2014;**55**(5):708–722.
114. Gracia A, Elcoroaristizabal X, Fernández-Quintela A, Miranda J, Bediaga NG, M de Pancorbo M, Rimando AM, Portillo MP. Fatty acid synthase methylation levels in adipose tissue: effects of an obesogenic diet and phenol compounds. *Genes Nutr*. 2014;**9**(4):411.
115. Hui ST, Parks BW, Org E, Norheim F, Che N, Pan C, Castellani LW, Charugundla S, Dirks DL, Psychogios N, Neuhaus I, Gerszten RE, Kirchgessner T, Gargalovic PS, Lusis AJ. The genetic architecture of NAFLD among inbred strains of mice. *eLife*. 2015;**4**:e05607.
116. Dallio M, Masarone M, Errico S, Gravina AG, Nicolucci C, Di Sarno R, Gionti L, Tuccillo C, Persico M, Stiuso P, Diano N, Loguercio C, Federico A. Role of bisphenol A as environmental factor in the promotion of non-alcoholic fatty liver disease: in vitro and clinical study. *Aliment Pharmacol Ther*. 2018;**47**(6):826–837.
117. Veiga-Lopez A, Kannan K, Liao C, Ye W, Domino SE, Padmanabhan V. Gender-specific effects on gestational length and birth weight by early pregnancy BPA exposure. *J Clin Endocrinol Metab*. 2015;**100**(11):E1394–E1403.

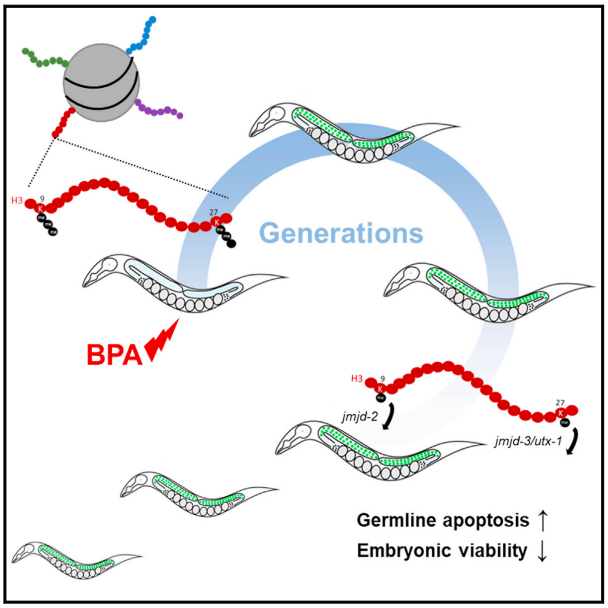
Appendix F

Article

Cell Reports

The Memory of Environmental Chemical Exposure in *C. elegans* Is Dependent on the Jumonji Demethylases *jmjd-2* and *jmjd-3/utx-1*

Graphical Abstract



Authors

Jessica Camacho, Lisa Truong, Zeyneb Kurt, ..., Matteo Pellegrini, Xia Yang, Patrick Allard

Correspondence

pallard@ucla.edu

In Brief

Little is known about the mechanisms of inheritance of artificial environmental exposures. Camacho et al. describe the transgenerational reproductive dysfunctions caused by ancestral exposure to the model environmental compound Bisphenol A, and they provide a role for the regulation of repressive histone marks by histone demethylases in this process.

Highlights

- Bisphenol A elicits a 5-generation germline array desilencing effect in *C. elegans*
- The desilencing response tracks with germline apoptosis and embryonic lethality
- Ancestrally exposed F3 germlines show a dramatic reduction in H3K9me3 and H3K27me3
- JMJD-2 and JMD-3/UTX-1 demethylases are required for BPA's transgenerational effects

Data and Software Availability

GSE113187
GSE113266

Camacho et al., 2018, Cell Reports *23*, 2392–2404
May 22, 2018
https://doi.org/10.1016/j.celrep.2018.04.078

CellPress

OPEN ACCESS
CellPress

Cell Reports
Article

The Memory of Environmental Chemical Exposure in *C. elegans* Is Dependent on the Jumonji Demethylases *jmjd-2* and *jmjd-3/utx-1*

Jessica Camacho,[1] Lisa Truong,[2] Zeyneb Kurt,[3] Yen-Wei Chen,[1] Marco Morselli,[4] Gerardo Gutierrez,[1,5] Matteo Pellegrini,[4] Xia Yang,[1,3,6,7,8] and Patrick Allard[1,8,9,10,*]

[1]Molecular Toxicology Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA 90095, USA
[2]Human Genetics and Genomic Analysis Training Program, University of California, Los Angeles, Los Angeles, CA 90095, USA
[3]Department of Integrative Biology and Physiology, University of California, Los Angeles, Los Angeles, CA 90095, USA
[4]Molecular, Cell and Developmental Biology Department, University of California, Los Angeles, Los Angeles, CA 90095, USA
[5]Department of Environmental and Occupational Health, California State University, Northridge, CA 91330, USA
[6]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA 90095, USA
[7]Institute for Quantitative and Computational Biosciences, University of California, Los Angeles, Los Angeles, CA 90095, USA
[8]Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095, USA
[9]Institute for Society and Genetics, University of California, Los Angeles, Los Angeles, CA 90095, USA
[10]Lead Contact
*Correspondence: pallard@ucla.edu
https://doi.org/10.1016/j.celrep.2018.04.078

SUMMARY

How artificial environmental cues are biologically integrated and transgenerationally inherited is still poorly understood. Here, we investigate the mechanisms of inheritance of reproductive outcomes elicited by the model environmental chemical Bisphenol A in *C. elegans*. We show that Bisphenol A (BPA) exposure causes the derepression of an epigenomically silenced transgene in the germline for 5 generations, regardless of ancestral response. Chromatin immunoprecipitation sequencing (ChIP-seq), histone modification quantitation, and immunofluorescence assays revealed that this effect is associated with a reduction of the repressive marks H3K9me3 and H3K27me3 in whole worms and in germline nuclei in the F3, as well as with reproductive dysfunctions, including germline apoptosis and embryonic lethality. Furthermore, targeting of the Jumonji demethylases JMJD-2 and JMJD-3/UTX-1 restores H3K9me3 and H3K27me3 levels, respectively, and it fully alleviates the BPA-induced transgenerational effects. Together, our results demonstrate the central role of repressive histone modifications in the inheritance of reproductive defects elicited by a common environmental chemical exposure.

INTRODUCTION

The elicitation and inheritance of phenotypes from environmental cues have been the subject of intense research and debate. Best understood is the transfer of biological information triggered by natural exposures, such as temperature, hyperosmotic stress, diet, or starvation, thanks to research advances in a variety of model systems from plants to rodents (reviewed in Heard and Martienssen, 2014). Recent reports have shown that the heritability of effects elicited by such natural cues across generations is conditioned by changes in the epigenome, or the molecular tags that alter gene expression and that are mitotically and/or meiotically heritable but do not entail a change in DNA sequence (Wu and Morris, 2001). These mechanisms include small RNA-based pathways (Gapp et al., 2014; Rechavi et al., 2014; Zhong et al., 2013) as well as through the regulation of the complex collection of covalent modifications of histone proteins (Gaydos et al., 2014; Greer et al., 2014; Kishimoto et al., 2017; Klosin et al., 2017; Siklenka et al., 2015). By contrast, the transgenerational inheritance of man-made environmental chemicals has remained controversial, particularly in mammalian settings. Several rodent studies have indicated that a one-generation parental (P)0 exposure to compounds, such as the fungicide Vinclozolin (Anway et al., 2005), or to mixtures of plastic compounds, such as Bisphenol A (BPA) and phthalates (Manikkam et al., 2013), is sufficient to cause a transgenerational decrease in the number and quality of germ cells in F3 and F4 adults, and it correlates with an alteration of DNA methylation patterns (Anway et al., 2005, 2006). However, some of these studies have been challenged (Heard and Martienssen, 2014; Hughes, 2014), have not provided a clear mechanism of inheritance, and have not explored the involvement of other epigenetic marks besides DNA methylation, such as histone modifications.

The nematode *Caenorhabditis elegans* has proven to be a valuable model system to study the effects of environmental exposures on the epigenome due to its ability to respond to a variety of stressors (Kishimoto et al., 2017; Klosin et al., 2017; Rechavi et al., 2014; Rudgalvyte et al., 2017). Here, we exploited the tractability of *C. elegans* to study the transgenerational impact of chemical exposure on reproductive function and dissect its underlying mechanisms of inheritance. These experiments were

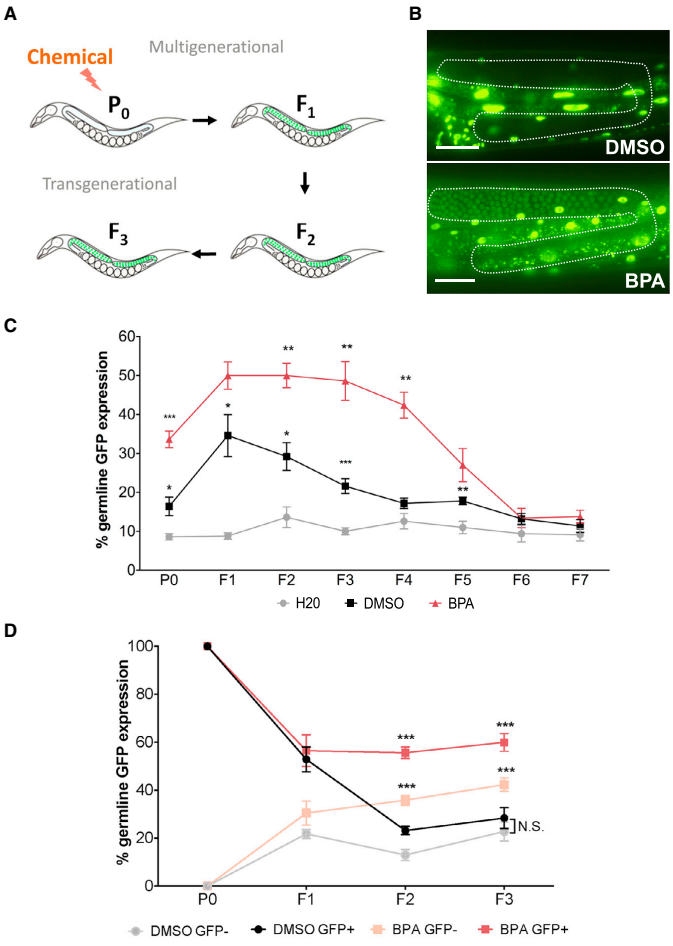


Figure 1. BPA Exposure Elicits a Transgenerational Desilencing of a Repetitive Array

(A) Exposure scheme. Nematodes are exposed to the chemicals of interest for 48 hr at the parental (P0) generation. Worms carrying the integrated array *pkIs1582* [*let-858*::GFP; *rol-6*(*su1006*)] express GFP in all somatic nuclei but silence the array in the germline. This strain is used to monitor the array desilencing over multiple generations.

(B) Representative example of silenced (top) and desilenced (bottom) *pkIs1582* array expression in F3 germlines (dashed lines). Scale bar, 50 μm.

(C) Percentage of worms displaying germline desilencing (y axis) at each generation (x axis). n = 5–10, 30 worms each; *p ≤ 0.05, **p ≤ 0.01, and ***p ≤ 0.001. Significance is indicated for BPA versus DMSO above the BPA line and DMSO versus water above the DMSO line.

(D) Lineage analysis of the germline desilencing response. Worms were sorted following exposure at the P0 generation based on their germline GFP expression. Their progeny was then followed and examined for 3 additional generations. n = 5–10, 30 worms each; ***p ≤ 0.001. BPA is compared to DMSO within each GFP status category (e.g., BPA/GFP+ versus DMSO/GFP+). All data are represented as mean ± SEM.

greatly facilitated by the nematode's short generation time, approximately 4 days at 20°C; its well-characterized distribution and regulation of chromatin marks (Bessler et al., 2010; Ho et al., 2014; Liu et al., 2011); and its ability to silence repetitive transgenes in the germline via repressive histone modifications in a fashion similar to the silencing of repetitive elements in mammalian germ cells (Kelly and Fire, 1998; Liu et al., 2014). Using these features, we investigated the mechanism of transgenerational inheritance following exposure to the model environmental chemical BPA. BPA is a widely used, high-production volume plastic manufacturing chemical highly prevalent in human samples (Vandenberg et al., 2010). We show that ancestral BPA exposure causes a histone 3, lysine 9 (H3K9) and a histone 3, lysine 27 (H3K27) trimethylation-dependent transgenerational chromatin-desilencing response in the germline that spans five generations and is associated with germline dysfunction and elevated progeny lethality.

RESULTS

Germline Transgene Desilencing following Chemical Exposure

To capture single, multi-, and transgenerational environmental effects stemming from chemical exposure, we used a germline desilencing reporter (Kelly et al., 1997). The assay that we developed (Figure 1A) is based on the strain NL2507 carrying an integrated low-complexity, highly repetitive array composed of a transgene coding for a fusion product between nuclear-localized LET-858 and GFP (*pkIs1582[let-858::GFP; rol-6(su1006)]*). This transgene is expressed in somatic cells, but it is transcriptionally silenced in the germline (Figure 1B) via accumulation of the repressive marks H3K9me3 and H3K27me3 (Kelly and Fire, 1998; Schaner and Kelly, 2006).

We first tested the reporter NL2507 strain in a chemical assay by using a variety of well-characterized inhibitors of chromatin-modifying enzymes (Figure S1). All drug exposures were performed at the P0 generation for 48 hr, encompassing the window of L4 stage to day 1 of adulthood. Drug responses were compared to the vehicle DMSO in the context of which a low rate of desilencing is observed (14.3% ± 1.6%). Following treatment with all tested inhibitors of H3K9 or H3K27 demethylases,

of non-selective methyltransferases or demethylases, as well as of histone acetyltransferases, the transgene expression remained silenced at levels comparable to the DMSO control. Conversely, HDAC inhibitors or methyltransferase inhibitors against either H3K9 or H3K27 all led to an increase in *pkIs1582* germline expression, with exposure to the class I HDAC inhibitor sodium butyrate and the SAM and EZH2 inhibitor 3-Deazaneplanocin A (DZnep) showing the highest levels of desilencing at P0, 32.5% ± 3.1% and 38.2% ± 1.9%, respectively ($p \leq 0.0001$ for both). Together, these results indicate that the desilencing of the *pkIs1582* array may serve as a sensitive and relevant indicator of chromatin mark-regulated transcriptional modulation.

BPA Exposure Causes a Heritable, Transgenerational Chromosomal Array-Desilencing Response

BPA was chosen as a test compound in the array-desilencing assay based on several lines of evidence that include changes in H3K27 histone methyltransferase Enhancer of Zeste homolog 2 (EZH2) expression (Bhan et al., 2014) and decreases in H3K9me3 levels in post-natal mouse oocytes (Trapphoff et al., 2013) and in H3K9 and H3K27 methylation levels in a variety of somatic cell types (Doherty et al., 2010; Singh and Li, 2012; Yeo et al., 2013).

First, we tested a range of BPA concentrations (10, 50, 100, and 500 μM), chosen based on previous dose-response analyses (Chen et al., 2016), to identify the lowest dose that led to a maximal desilencing effect. We initially performed the exposures at a single generation (P0) at L4 stage for 48 hr. We observed a dose-response relationship of the germline array desilencing across generations, reaching saturation at 100 μM (45.0% ± 3.3% desilencing at the F3, $p \leq 0.001$) (Figure S2A). We also tested additional 48-hr exposure windows, including from L1 to L4 (Figure S2B) and from day 0 of adulthood (24 hr post-L4) to day 2 (Figure S2C). In all cases, we observed a significant desilencing of the germline array in the F3, although the generational kinetics varied between exposure windows and none reached the maximum F3 desilencing levels achieved by the L4-to-day 1 exposure window (Figure S2A). Thus, for all subsequent experiments, we exposed the worms to a single 100-μM BPA dose from L4 to day 1. This external dose is below previously characterized *C. elegans* doses measured by gas chromatography-mass spectrometry (GC-MS) to lead to an internal BPA concentration within human physiological range (Chen et al., 2016).

We then examined the rate of array desilencing over six generations following the single P0 generation BPA exposure at 100 μM (Figure 1C). The solvent control DMSO led to a pronounced elevation in desilencing in F1 animals (34.6% ± 5.4% of worms display GFP expression in their germline) compared to water alone (8.6% ± 0.8%). However, GFP levels in the DMSO group sharply declined after the F1 generation and were statistically indistinguishable from the water control at the F4 generation. This effect of DMSO is likely due to its described positive activity in DNA relaxation, transcription enhancement, and promotion of an active chromatin state (Iwatani et al., 2006; Juang and Liu, 1987; Kim and Dean, 2004). By contrast, BPA exposure led to a dramatic increase in desilencing in the F1 generation (50.0% ± 3.5%). This BPA-induced desilencing rate was consistently higher than DMSO's and remained that way until the F5 generation. These results therefore indicate a potent transgenerational desilencing response stemming from BPA exposure and spanning 5 generations (P0–F4).

To determine whether most of the desilencing effect observed in the first transgenerational (F3) generation is primarily caused by descendants of strong P0 responders, we performed a series of lineage studies where individual P0 worms were segregated based on their germline GFP expression following BPA or DMSO exposure. Worms that showed germline desilencing at P0 following BPA exposure gave rise to F1, F2, and F3 progenies with a high rate of desilencing, nearing 60% (Figure 1D). By contrast, DMSO-exposed animals, whether silenced or desilenced at P0, showed a reduced rate of desilencing in the F2 and F3 generations, nearing 20%. Surprisingly, BPA-treated but GFP-negative P0 worms gave rise to progeny showing a higher rate of desilencing at each subsequent generation, such that there was a statistically significant difference when compared to DMSO in the F2 and F3 generations. In the latter, the proportion of descendants of BPA-exposed but GFP-negative P0s showing germline desilencing reached 42.3% ± 2.8% ($p \leq 0.01$ versus DMSO/GFP−). Interestingly, the mating of ancestrally exposed F1 hermaphrodites with unexposed males did not rescue the germline desilencing response, indicating that the primary mode of inheritance of BPA's effect is through the female germline (Figure S2D).

Collectively, these findings identify a matrilineal transgenerational inheritance of a repetitive array-desilencing response that is only partially conditioned by the ancestral (P0) response to BPA exposure.

BPA Exposure Causes a Transgenerational Alteration of the Germline Transcriptome

To investigate the impact of ancestral BPA exposure on the germline and distinguish it from that of DMSO, which also led to a mild transgenerational germline desilencing in the F3 compared to water, we performed RNA sequencing (RNA-seq) analysis on isolated F3 germlines. We identified a total of 264 transcripts that were differentially up- or downregulated at $p \leq 0.05$ in F3 germlines ancestrally exposed to BPA compared to DMSO, with 152 transcripts having a fold induction ≤ 0.5 or ≥ 1.5 (Table S1; Figure S3A). There was little overlap between the transcripts that were differentially expressed in all 3 groups, BPA versus DMSO, BPA versus water, and DMSO versus water (Figure S3B), suggesting that DMSO's transgenerational impact on the germline transcriptome is mostly distinct from that of BPA. A gene ontology analysis of the functional categories represented by the differentially expressed transcripts also highlighted the lack of overlap between the different treatment group comparisons. Interestingly, however, the second most represented functional category in the BPA versus DMSO group was reproduction, which was not represented in the DMSO versus water group (Figure S3C). This category includes 61 genes, many of them normally expressed in the germline tissue and essential for germline function (Table S2). These results therefore suggest that ancestral BPA exposure may deregulate reproductive processes by altering the germline transcriptome.

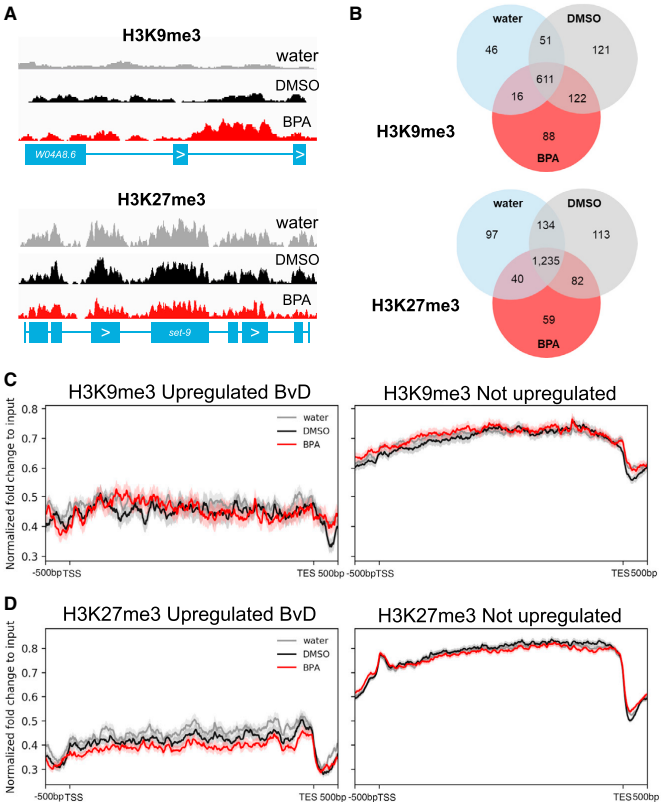


Figure 2. BPA-Induced Transgenerational Reduction in H3K9me3 and H3K27me3 Identified by ChIP-Seq

(A) Examples of ChIP-seq gene plots for H3K9me3 and H3K27me3 from F3 nematodes.

(B) Venn diagram from genes with peak calling in each of the treatment groups.

(C) Average H3K9me3 histone modification fold enrichment signals from gene bodies of either silenced upregulated genes (left panel) or silenced non-upregulated genes (right panel) after BPA treatment. Lightly shaded regions indicate the SE.

(D) Average H3K27me3 histone modification fold enrichment signals from gene bodies of either silenced upregulated genes (left panel) or silenced non-upregulated genes (right panel) after BPA treatment. Lightly shaded regions indicate the SE.

Ancestral BPA Exposure Leads to a Deregulation of Repressive Histone Marks in F3 Nematodes

Several recent reports in *C. elegans* have implicated various histone modifications as important mediators of a variety of environmental effects across generations (Kishimoto et al., 2017; Klosin et al., 2017). We therefore assessed whether BPA exposure in P0 worms could lead to observable changes in the chromatin of F3 worms. To this aim, we performed chromatin immunoprecipitation sequencing (ChIP-seq) in whole adult worms at the F3 generation ancestrally exposed to BPA, DMSO, and water. Just as for the RNA-seq analysis, these experiments were performed on a large population of worms that were not selected based on their GFP expression. We focused our analysis on two repressive marks, H3K9me3 and H3K27me3, which have both been previously implicated in chromatin silencing in the germline of a wide range of species as well as in the repression of low-complexity transgenes in the *C. elegans* germline (Bessler et al., 2010; Greer et al., 2014; Leung et al., 2014; Liu et al., 2014; Schaner and Kelly, 2006; Towbin et al., 2012).

We first mined the ChIP-seq data to identify genes with significantly altered H3K9me3 and H3K27me3 levels (see the Experimental Procedures; Figures 2A and 2B). Among the three conditions, water, DMSO, and BPA, we identified between 3,740 and 4,951 broad peaks for H3K9me3 and between 19,019 and 21,741 for H3K27me3 (Table S3). A total of 1,055 and 1,780 genes were associated with broad peak calls, i.e., showed enrichment in their gene bodies, for H3K9me3 and H3K27me3, respectively. The majority of these peak calls were shared among all three treatment groups, although the BPA treatment group generated 88 and 59 unique peaks for H3K9me3 and H3K27me3, respectively (Figure 2B). The gene ontology (GO) analysis of biological processes at false discovery rate (FDR) < 0.05 and $p < 0.001$ for the genes associated with a loss of H3K27me3 broad peaks in BPA samples compared to DMSO confirmed the relevance of the epigenomic effect detected, as the second most prominent GO category was related to the response to steroid hormone stimulus, in line with BPA's well-described estrogenic activity (Table S4).

Next we compared the ChIP-seq and RNA-seq datasets by examining the levels of H3K9me3 and H3K27me3 under all 3 treatment conditions in genes that either had a low expression level in DMSO (first quartile, i.e., silenced genes) and were not upregulated or were upregulated >2-fold based on the RNA-seq data. As expected, we found that upregulated genes had on average 40%–50% lower H3K9me3 and H3K27me3 compared to their not-upregulated counterparts (Figures 2C and 2D). The levels and distributions of the marks were consistent with their described patterns in the *C. elegans* larval chromatin, where both H3K9me3 and H3K27me3 predominantly occupy the gene body of silenced genes (Ho et al., 2014). Comparing the three treatment groups, we did not observe a difference in H3K9me3 based on expression levels, perhaps due to

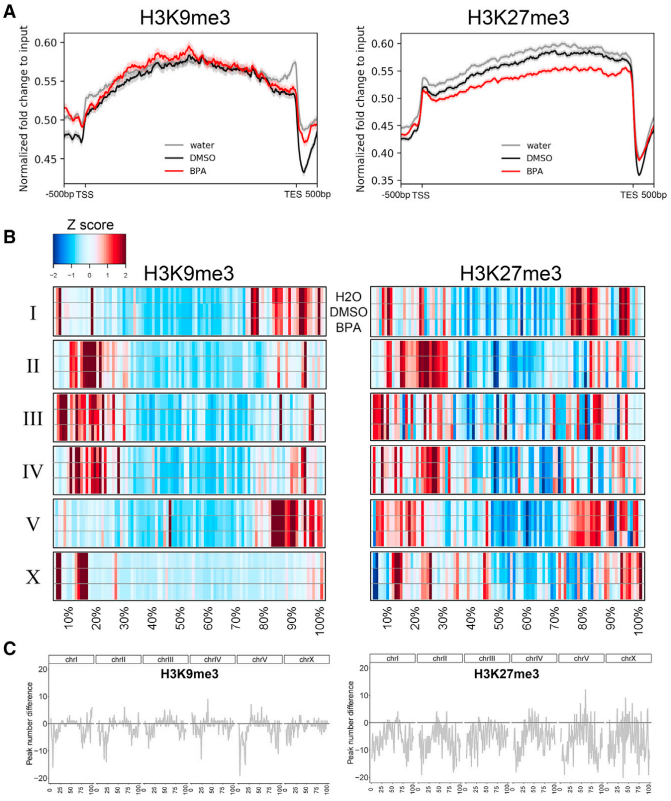


Figure 3. BPA Treatment Causes Transgenerational Intra-chromosomal Redistribution of Histone Modifications

(A) Average H3K9me3 (left) and H3K27me3 (right) histone modification fold enrichment signals from gene bodies of all genes. Shaded regions indicate SE.

(B) Heatmap of averaged H3K9me3 (left) and H3K27me3 (right) histone modification fold enrichment signals in 100 sub-regions across all chromosomes. *Z* scores were calculated on averaged values in each chromosome and sample.

(C) Difference in unique peak-calling numbers between BPA and DMSO from H3K9me3 (left) and H3K27me3 (right) along all chromosome sub-regions. The y axis indicates unique peak numbers calculated by BPA minus DMSO by region.

the tissue sources used for the two datasets (whole worms for ChIP-seq and isolated germlines for RNA-seq). However, we observed a decrease in H3K27me3 in the BPA treatment group compared to DMSO and water for genes that were upregulated (Figure 2D, lightly shaded area indicates SE). These results were similar for all genes, irrespective of expression level, where H3K27me3 was significantly reduced in the gene body compared to DMSO and water groups (Figure 3A).

Finally, we asked whether ancestral BPA exposure might not only affect H3K9me3 and H3K27me3 gene body levels but also their distribution along the chromosome axes. To this aim, we calculated the average fold enrichment of each mark over input by 1% increments along all 6 chromosomes. The data were normalized using a *Z* score for each individual chromosome and treatment group to allow the visualization of the marks' redistribution (Figure 3B). For each 1% increment, we also identified the number of peaks that were present in BPA but absent in DMSO (Figure 3C). These two complementary chromosome-wide analyses revealed a reduction of both marks from the distal chromosomal regions, largely heterochromatic (Garrigues et al., 2015), and a slight enrichment in the chromosome centers when comparing BPA to DMSO (Figures 3B and 3C). It also suggested a decrease of the marks' levels on the X chromosome. We validated the decrease in the levels of the marks by performing a multiplex histone post-translation modification (PTM) quantitation assay on pooled F3 whole-worm extracts (Table S5). The assay revealed a 25%–33% decrease in H3K9 mono-, di-, and trimethylation and a more pronounced 29%–56% decrease in H3K27 di- and trimethylation at the F3 generation in BPA-exposed P0 nematodes compared to DMSO. Conversely, another histone modification, H3K36me3, remained largely unchanged. Together, these results indicate a potent transgenerational impact of BPA on the chromatin, altering both the levels of the two repressive marks H3K9me3 and H3K27me3 as well as their distribution along chromosomal axes.

Ancestral BPA Exposure Leads to a Deregulation of Repressive Histone in the Germline

A transgenerational effect implies that the epigenomic alterations described above must also occur in the germline in order to be inherited. We therefore performed immunofluorescence against H3K9me3 and H3K27me3 in dissected germlines of the NL2507 strain containing the integrated *pkIs1582* transgene at the F3, when desilencing is pronounced, and at the F7, when germline desilencing has returned to control levels. At the pachytene stage of the F3 germline, we observed significant 26% and 24% reductions in global H3K9me3 and H3K27me3 levels, respectively, between BPA and DMSO (Figures 4A and 4B). By contrast, no significant differences were observed between water and DMSO. A similar decrease of total nuclear levels of these marks was seen in the strain PD7271, where the transgene is episomally maintained (*ccEx7271*): 23.3% and 34.6% reductions for H3K9me3 and H3K27me3, respectively (Figure S4). At the F7 generation, the germline levels of H3K9me3 and

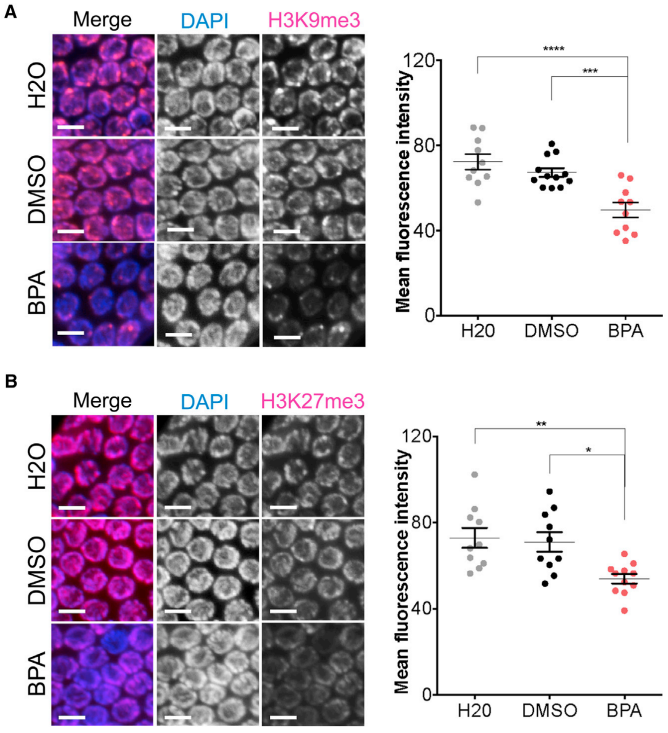


Figure 4. Ancestral BPA Exposure Decreases H3K9me3 and H3K27me3 Levels in F3 Germlines

(A and B) Immunofluorescence images of mid-to-late pachytene germline nuclei from F3 worms ancestrally exposed to DMSO or BPA and stained for H3K9me3 (A) or H3K27me3 (B). DAPI is represented in blue and the histone mark of interest in magenta in the merge. All images shown were selected representative images of the mean values obtained after quantification of all germline nuclei from that exposure group. The corresponding fluorescence intensity quantification is shown on the right panels. n = 11–12 worms, 10 nuclei per worm; $^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$, and $^{****}p \leq 0.0001$, one-way ANOVA with Sidak correction. Scale bar, 5 μm. All data are represented as mean ± SEM.

H3K27me3 in the BPA group were statistically indistinguishable from DMSO controls (Figure S5).

The use of the PD7271 *ccEx7271* array-bearing strain also allowed us to separately examine the levels of repressive modifications on the autosomes; the X chromosomes, which tend to lay apart from the rest of the chromosomes during the pachytene stage in hermaphrodites (Schaner and Kelly, 2006); and the extrachromosomal array (Figures 5A and 5B). We observed marked decreases in both H3K9me3 and H3K27me3 on autosomes (24.8% and 34.3%, respectively), X chromosomes (25.3% and 41.5%), and the extrachromosomal array (39.6% and 51.3%). We examined whether the trend toward a larger decrease of these marks on the X chromosomes compared to autosomes was significant by measuring the X:A ratio for each germline nucleus (Figure 5C). F3 germline nuclei showed a significant X:A ratio decrease in H3K27me3 levels when ancestrally exposed to BPA compared to DMSO (0.98 versus 1.09, respectively, a 10% decrease; p = 0.03), while H3K9me3 showed a trend toward a decreased X:A ratio between DMSO and BPA. Consistent with these results and with the described role of H3K27me3 in X silencing in the germline (Bender et al., 2006; Gaydos et al., 2012), we observed a modest but significant (p = 0.01) 2.36% increase in overall X-related genes with fragments per kilobase of transcript per million (FPKM) > 1 in our F3 germline RNA-seq data (Figure 5D).

Taken together, these experiments indicate a broad transgenerational impact on the germline chromatin of F3 nematodes not only confined to the repetitive arrays but also affecting the autosomes and the X chromosomes.

BPA Exposure Elicits a Transgenerational Increase in Embryonic Lethality and Germline Dysfunction

Next, we examined whether the transgenerational alteration of the germline chromatin was associated with reproductive defects. For these and all subsequent experiments, we chose to only compare BPA to DMSO, as BPA is dissolved in DMSO and the RNA-seq and ChIP-seq data indicated chromatin and expression BPA signatures distinct from those of DMSO. While the number of embryos produced was not dependent on ancestral exposure (Figure 6A), we observed a significant 85% (D = 3.83 and B = 7.07) increase in embryonic lethality in F3 worms ancestrally exposed to BPA when compared to DMSO (Figure 6B). We also examined the rate of embryonic lethality at the F7, a generation at which desilencing is not observed. Surprisingly, a trend between DMSO and BPA was still apparent even if it did not reach significance (86%, D = 3.58 and B = 6.67) (Figure 6B). The F3 embryonic lethality defect was not caused by the spurious expression of the *pkIs1582* transgene in the germline, as it was also observed in wild-type (N2) worms (Figure S6). Additionally, we assessed whether the increased embryonic lethality correlated with the transgene desilencing by separately assessing the embryonic survival of GFP-negative and GFP-positive F3 worms' progeny (Figure 6C). We observed a significantly higher level of embryonic lethality in the offspring of GFP-positive F3 worms ancestrally exposed to BPA when compared to both GFP-negative/BPA F3 offspring and GFP-positive/DMSO F3 offspring.

Finally, we monitored germline health by measuring the induction of germline apoptosis using acridine orange staining

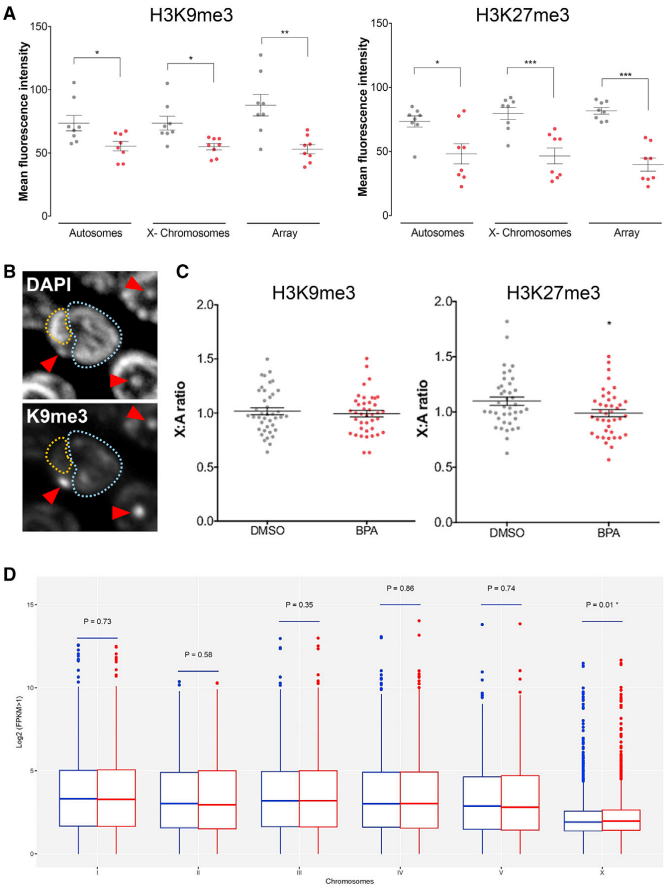


Figure 5. Ancestral BPA Exposure Leads to a Sharp Decrease in H3K9me3 and H3K27me3 on Autosomes, X Chromosomes, and an Extrachromosomal Array and an Up-regulation of X-Linked Genes

(A) Quantification of H3K9me3 and H3K27me3 levels on autosomes, X chromosomes, and an extrachromosomal array in the F3 generation following P0 exposure to either DMSO or BPA. Gray, DMSO; red, BPA. n = 8 worms, 5 nuclei per worm; $^{*}p \leq 0.05$, $^{**}p \leq 0.01$, and $^{***}p \leq 0.001$.

(B) DAPI- (top) and H3K9me3- (bottom) stained nuclei. The colored dashed lines identify the autosomes (blue) and the X chromosomes (orange). The red arrowheads identify the extrachromosomal array that is enriched in H3K9me3.

(C) Fluorescence intensity quantification of H3K9me3 and H3K27me3 levels is shown on the right. Gray is the X:A ratio for DMSO and red for BPA. n = 8 worms, 5 nuclei per worm; $^{*}p \leq 0.05$.

(D) Gene expression data from dissected F3 germlines showing all transcripts with FKPM > 1 following ancestral DMSO (blue) or BPA (red) exposure. X-linked genes show a modest but significant overall 2.36% increase in expression ($p = 0.01$). All data are represented as mean ± SEM.

(Gartner et al., 2008) at the late prophase stage, when synapsis and recombination-dependent checkpoint activation results in programmed germline nuclear culling (Bhalla and Dernburg, 2005; Gartner et al., 2008). We observed a significant increase in germline apoptosis in F3 worms ancestrally exposed to BPA when compared to DMSO (Figures 6D and 6E), which was lost at the F7. Thus, together, these results show that ancestral BPA exposure elicits a clear transgenerational reproductive dysfunction effect. They also indicate that BPA-induced transgenerational effects mostly resolve by the F7.

Jumonji Histone Demethylase Activity Is Required for the Inheritance of BPA-Induced Transgenerational Effects

Since BPA exposure at the P0 generation was correlated with a decrease in repressive histone modifications in the germline of the F3 worms, we hypothesized that BPA's effects may be dependent on levels of these marks and on the activity of the enzymes that regulate them. This hypothesis was partially supported by the RNA-seq data from which 7 differentially expressed chromatin factors were identified: *sir-2.4*, ZK1127.3, *sop-2*, TO7E3.3, *met-2*, *jmjd-1.2*, and *set-26* (Table S1). MET-2, a SET domain histone H3 lysine 9 histone methyltransferase (HMTase) (Bessler et al., 2010), was significantly downregulated, while *set-26*, another H3K9 methyltransferase (Greer et al., 2014), was represented by two functionally equivalent transcript isoforms, one upregulated and one downregulated. Therefore, to functionally implicate the dysregulation of H3K9me3 and H3K27me3 in BPA's transgenerational outcomes, we attempted to rescue its effects by genetically or chemically modulating several histone demethylases after the initial P0 exposure but prior to the F3 (Figures 7A and S8A).

We first assessed whether the deregulation of repressive H3-lysine methylation marks by BPA is required for the transgenerational inheritance of BPA-induced effects. To this end, we used a feeding RNAi strategy to downregulate the expression of *jmjd-2* (H3K9me3/H3K36me3 histone lysine demethylase [KDM]) (Greer et al., 2014; Whetstine et al., 2006) or *jmjd-3/utx-1* (H3K27me3 KDM) (Agger et al., 2007), and we monitored two hallmarks of BPA's transgenerational effects, namely, the germline array desilencing as well as the increase in embryonic lethality. When compared to control RNAi, the downregulation of *jmjd-2* or *jmjd-3/utx-1* at the F1-to-F2 transition was sufficient to increase the levels of H3K9me3 and H3K27me3, respectively,

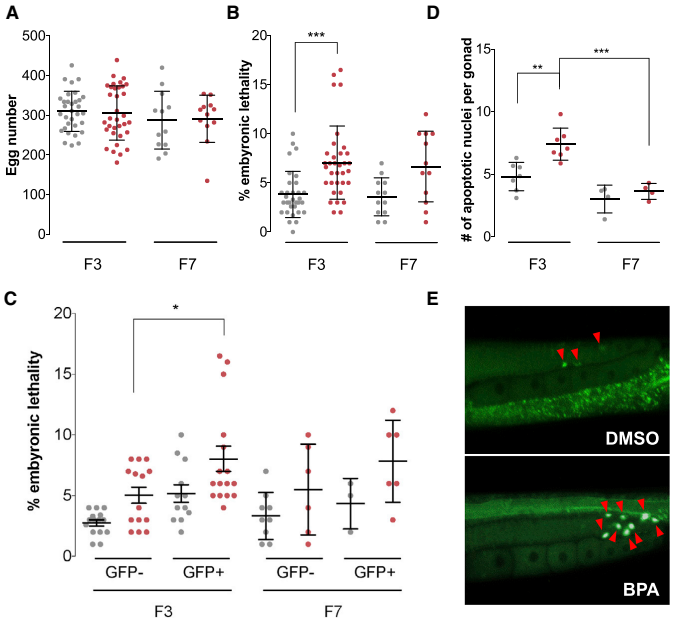


Figure 6. Transgenerational Impact of BPA on Fertility

(A) Number of eggs produced by F3 or F7 worms following P0 exposure to DMSO control (gray) or BPA (red).

(B) Percentage of lethality of embryos generated by F3 or F7 worms ancestrally exposed to either DMSO control or BPA. n = 23–33; ***p ≤ 0.001, two-way ANOVA.

(C) Embryonic lethality of F3 or F7 worms' progeny based on the GFP expression in the germline of F3 or F7 worms. n = 10; *p ≤ 0.05, two-way ANOVA.

(D) Number of apoptotic nuclei per gonadal arms of F3 or F7 worms. n = 7 repeats, 20 worms each; **p ≤ 0.01 and ***p ≤ 0.001, two-way ANOVA.

(E) Representative examples of acridine orange-stained F3 nematodes following P0 DMSO or BPA exposure. All data are represented as mean ± SEM.

in the F3 germlines (Figure 7B; quantification shown in Figure S7A). Also, while the control RNAi conditions slightly elevated the rates of desilencing and embryonic lethality compared to no-RNAi conditions, the downregulation of either *jmjd-2* or *jmjd-3/utx-1* led to a complete rescue of BPA-induced responses in the F3, except for the embryonic lethality effect under *jmjd-2* RNAi conditions, which was strongly reduced but did not reach significance (Figure 7C). Interestingly, single RNAi against *jmjd-3* or *utx-1* dramatically increased the proportion of desilenced germlines under both ancestral DMSO and BPA exposures, suggesting a partial compensation between *jmjd-3* and *utx-1* in the *C. elegans* germline (Figure S7B). This increase is similar to that of RNAi against the H3K27 HMT Polycomb Group complex member *mes-6* or against the SET domain H3K36 HMT *mes-4*, which functions to limit H3K27me3 spreading away from silenced chromatin (Figure S7B) (Gaydos et al., 2012).

We further implicated the deregulation of H3K9me3 and H3K27me3 as central to BPA's transgenerational effects by performing drug rescue experiments using the KDM4/JMJD-2 inhibitor IOX-1 (King et al., 2010), which has been shown to elevate H3K9me3 levels *in vitro* and in cell culture settings, (Hu et al., 2016; King et al., 2010; Schiller et al., 2014), and the potent selective Jumonji JMJD-3/UTX-1 H3K27 demethylase inhibitor GSK-J4 (Kruidenier et al., 2012). We first examined whether a combination of the two histone demethylase inhibitors would be sufficient to decrease the germline array desilencing and embryonic lethality effects. The co-treatment of the F1 generation with 100 μM IOX-1 and 100 μM GSK-J4 led to a significant reduction in BPA-induced array desilencing and embryonic lethality by 15.8% and 27.0%, respectively (Figure S8B). Finally, we tested the effect of the two inhibitors independently. Remarkably, F1 exposure to either IOX-1 or GSK-J4 was sufficient to suppress the elevation in array desilencing and embryonic lethality in P0 BPA-exposed worms compared to DMSO (Figure S8C). Thus, two distinct means of rescuing BPA's transgenerational effects, by RNAi or chemical inhibitors, indicate that the activity of either JMJD-2 or JMD-3/UTX-1 is required for the inheritance of BPA-induced reproductive effects.

DISCUSSION

In the present study, we aimed to characterize the molecular mechanisms of memory of environmental exposures using BPA as a model chemical. We showed that ancestral BPA exposure leads to a transgenerational decrease in the germline levels of H3K9me3 and H3K27me3 dependent on the activity of the JMJD-2 and JMJD-3/UTX-1 demethylases. Interestingly, our results indicate that, while the overt germline desilencing effect lasts only up to 5 generations, some modest impacts on reproduction extend at least until the F7 generation. These results therefore suggest that the transgenerational impact of BPA may differ depending on the type of genetic loci examined, with repetitive loci, such as the transgene, being less affected than other loci controlling *C. elegans* reproductive function.

We found that modulation of either JMJD-2 or JMJD-3/UTX-1 activity, chemically or genetically, is sufficient to dramatically reduce the inheritance of transgenerational effects. While JMJD-2 acts as both an H3K9me3 and H3K36me3 demethylase, the ability of *jmjd-2* RNAi to rescue desilencing's effects is likely caused by its action on H3K9me3, as H3K36me3 is considered an active mark in the *C. elegans* germline (Gaydos

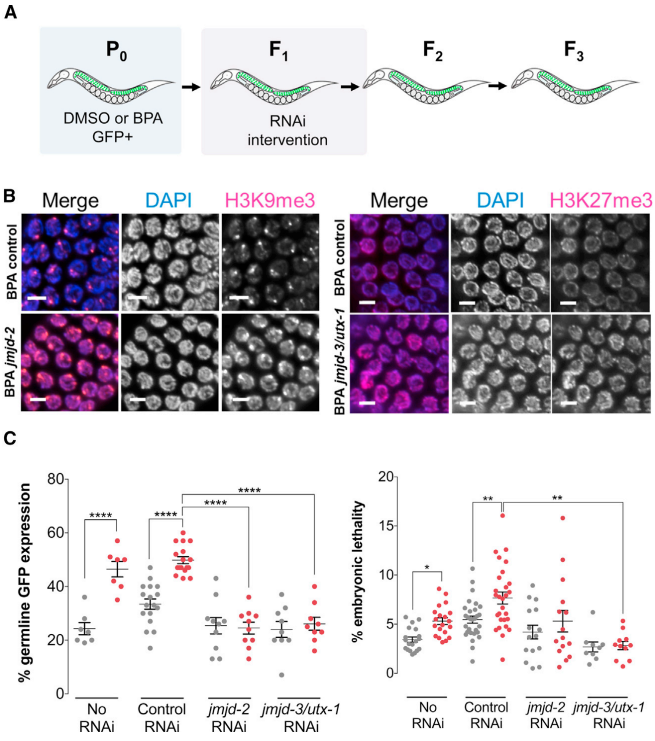


Figure 7. *jmjd-2* and *jmjd-3/utx-1* Demethylases Are Required for BPA-Induced Transgenerational Response

(A) Exposure and rescue experimental scheme. Following exposure to DMSO or BPA at the P0 generation, the progeny of GFP-positive P0 worms was collected and subjected to feeding RNAi until the F2. F3 worms were then collected and analyzed.

(B) Immunofluorescence images of mid-to-late pachytene germline nuclei from F3 worms ancestrally exposed to BPA and GFP-positive at the P0, stained for H3K9me3 or H3K27me3. DAPI is represented in blue and the histone mark of interest in magenta in the merge. All images shown were selected representative images of the mean values obtained after quantification of all germline nuclei from that exposure group (Figure S7A). Scale bar, 5 μm.

(C) RNAi rescue of ancestral DMSO- (gray) or BPA- (red) induced effects following either no F1 treatment, empty vector control, *jmjd-2*, or *jmjd-3/utx-1* feeding RNAi. n = 7–17 repeats, 30 worms each for desilencing assay and n = 4–8 repeats, 3–4 worms each for the embryonic lethality assay; *p ≤ 0.05, **p ≤ 0.01, and ****p ≤ 0.0001, two-way ANOVA. All data are represented as mean ± SEM.

et al., 2012) and RNAi against *jmjd-2* increases its levels (Whetstine et al., 2006), which is inconsistent with the observed decrease in BPA-induced desilencing in *jmjd-2* RNAi F3 animals. Our results thus suggest a cooperation between H3K9me3 and H3K27me3 for proper chromatin silencing in the *C. elegans* germline. Such cooperation is understood in mammalian embryonic stem cells (ESCs) to emerge from the interaction between Jarid2/Jumonji and Polycomb Repressive Complex 2 (PRC2) (Pasini et al., 2010; Peng et al., 2009) and to be important for heterochromatin formation and/or maintenance through PRC2's effect on increasing the binding efficiency of HP1 to H3K9me3 (Boros et al., 2014). In *C. elegans*' embryonic or larval chromatin, there is a strong overlap between H3K27me3 and H3K9me3 at genome-wide levels (Garrigues et al., 2015; Ho et al., 2014). This overlap is particularly significant at chromosomal arms of heterochromatic nature as well as lamina-associated domains (Ho et al., 2014), something also observed in our data (Figure 3B). In the *C. elegans* meiotic germline, the overlap between H3K27me3 and H3K9me3 chromosomal distribution is likely to be high, as H3K27me3 distribution is greater than that of H3K9me3 (Bender et al., 2004; Bessler et al., 2010; Schaner and Kelly, 2006).

Our results are consistent with previous observations in mouse germ cells, where exposure of growing oocytes to low BPA concentrations decreased H3K9me3 levels (Trapphoff et al., 2013). However, the effect of BPA may also be context dependent, as an increase in EZH2 expression and, consequently, an elevation of H3K27me3 was detected in mammary tissues following BPA exposure (Doherty et al., 2010). Our work suggests that, at least in *C. elegans*, the tight regulation of H3K9 and H3K27 methylation is central to the epigenetic memory of ancestral exposures. It will be crucial to examine how histone-based epimutations may be inherited across generations in mammalian models, since the mammalian epigenome undergoes two distinct waves of reprogramming, once in the primordial germ cells (PGCs) and a second time after fertilization in the pre-implantation embryo (reviewed in Tang et al., 2016). During the first reprogramming in PGCs, there is a wide fluctuation in H3K9me2 level, which becomes depleted (Seki et al., 2005), and in H3K27me3 level, which is gradually enriched globally (Hajkova et al., 2008). However, H3K9me3 is maintained in a dotted pattern in the pericentric heterochromatic regions as well as on endogenous retroviruses (Liu et al., 2014; Seki et al., 2005). Thus, H3K9me3 could serve in mammals as a molecular mediator of exposure memory in the germline.

The centrality of H3K9me3 in the inheritance of natural environmental effects has recently been further highlighted in *C. elegans*, where temperature-mediated alteration of transgene expression was detected for up to 14 generations (Klosin et al., 2017). However, other environmental cues, such as starvation or hyperosmosis, have been shown, depending on the studies,

to require small RNA-based mechanisms and/or H3K4 trimethylase activity (Kishimoto et al., 2017; Rechavi et al., 2014). While these pathways may be mechanistically related, it will be necessary to examine whether a unifying mechanism of environmental inheritance can be identified, especially as we also identified a requirement for the regulation of H3K27 methylation for the transgenerational inheritance of BPA's exposure. Finally, our findings on the transgenerational memory of exposure to the model toxicant BPA and its impact on the germline's epigenome and reproduction also raise important questions for human risk from exposure, as our work identified transgenerational reproductive effects even in the absence of such a response in the earlier generations and at BPA concentrations lower than those previously characterized and that yielded internal concentrations close to those found in human reproductive tissues (Chen et al., 2016; Schönfelder et al., 2002; Vandenberg et al., 2010).

In conclusion, we have uncovered a transgenerational effect on reproduction stemming from exposure to the environmental chemical BPA and mediated in part by a deregulation of repressive histone modifications. These findings, therefore, highlight the need to comprehensively examine the effect of our chemical environment on the unique context of the germline epigenome, and they also offer interventional means to prevent the transmission of such effects across generations.

EXPERIMENTAL PROCEDURES

Culture Conditions and Strains

Standard methods of culturing and handling of *C. elegans* were followed (Stiernagle, 2006). Worms were maintained on nematode growth medium (NGM) plates streaked with OP50 *E. coli*, and all experiments were performed at 20°C (at 25°C, a pronounced desilencing of pkIs1582 is observed in the germline). Strains used in this study were obtained from the *C. elegans* Genetics Center (CGC) and include the following: NL2507 (*pkIs1582*[*let-858*::GFP; *rol-6*(*su1006*)]), PD7271 (*pha-1(e2123)* III*; ccEx7271*), and N2 (wild-type).

Chemical Exposure and GFP Scoring

The exposure and GFP germline desilencing assessments were performed as previously described (Lundby et al., 2016). Briefly, all chemicals tested were obtained from Sigma-Aldrich and were dissolved in DMSO to a stock concentration of 100 mM. Worms were synchronized by bleaching an adult population of the strain of interest, plating the eggs, and allowing the synchronized population to reach L4 larval stage (approximately 50 hr). These were then collected and incubated for 48 hr in 50 μL OP50 bacteria, 500 μL M9, and 0.5 μL of the chemical of interest for a final chemical concentration of 100 μM. After 48 hr, the worms were collected and allowed to recover on NGM plates for 1–2 hr (mixed population) or immediately plated as individual worms to separately labeled 35-mm seeded NGM plates (GFP+/− population sorting) and recovered there. Worms were scored for germline GFP expression using a Nikon H600L microscope at 40× magnification.

Apoptosis Assay and Embryonic Lethality Assessment

Apoptosis assay was performed by acridine orange staining on synchronized adult hermaphrodites collected at 20–24 hr post-L4, as previously described (Allard and Colaiácovo, 2011; Chen et al., 2016). Embryonic lethality was performed by monitoring the numbers of embryos produced by each worm of each day of its reproductive life and subsequent larvae hatched from these embryos. The ratio of the latter measure by the former and multiplied by 100 generates the rate of embryonic lethality.

Chemical Rescue

F1 L4 larvae were obtained from DMSO- or BPA-exposed GFP-positive P0 worm populations, and they were exposed for 48 hr to the chemical rescue drugs IOX-1 and GSK-J4 dissolved in DMSO to a stock concentration of 100 mM. In combination treatments, one drug was prepared at a higher concentration so that the final DMSO concentration never exceeded 0.11%. The exposed F1 adult worms were then allowed to recover on NGM plates, and their offspring were followed until the F3 generation for GFP scoring and embryonic lethality assessment.

RNAi Experiments

Worms were exposed to RNAi by feeding (Kamath and Ahringer, 2003) with *E. coli* strains containing either an empty control vector (L4440) or expressing double-stranded RNA. RNAi constructs against *jmjd-2*, *jmjd-3*, *utx-1*, *mes-4*, and *mes-6* were obtained from the Ahringer RNAi library and sequence verified. P0 worms were exposed to BPA or DMSO for 48 hr following the procedure described above. For *jmjd-2* and *jmjd-3*/*utx-1* RNAi, F1 adult worms from GFP-positive P0 worms were placed on plates of *E. coli* containing an empty control vector (L4440) or expressing double-stranded RNA to lay overnight. F2 worms were grown on RNAi bacteria from hatching until the first day of adulthood, at which point they were transferred to non-RNAi OP50 plates. The subsequent generation (F3) was collected at adulthood (24 hr post-L4) for further analysis. For *mes-4* and *mes-6* RNAi, the same procedure was followed but from the F2 to F3 generation to circumvent their associated maternal sterility phenotype.

Immunofluorescence

Immunofluorescence images were collected at 0.5-μm z intervals with an Eclipse Ni-E microscope (Nikon) and a cooled charge-coupled device (CCD) camera (model CoolSNAP HQ, Photometrics) controlled by the NIS Elements AR system (Nikon). The images presented and quantified are projections approximately halfway through 3D data stacks of *C. elegans* gonads, which encompass entire nuclei. Images were subjected to 3D landweber deconvolution analysis (5 iterations) with the NIS Elements AR analysis program (Nikon). H3K27me3 and H3K9me3 quantification in mid-late pachytene germ cell nuclei was performed with the ImageJ software. F3 worms were staged at L4, and gonad dissection and immunofluorescence were performed 20–24 hr post-L4, as previously described (Chen et al., 2016). Primary antibodies were used at the following dilutions: rabbit α-H3K9me3, 1:500 (Abcam); and mouse α-H3K27me2me3, 1:200 (Active Motif). Secondary antibodies were used at the following dilutions: Cy3 α-rabbit, 1:700; and TxRed α-mouse, 1:200, (Jackson ImmunoResearch).

Germline RNA Amplification and RNA-Seq Analysis

Total RNA was extracted from needle-dissected gonads of F3 adult worms obtained from a mixed population of H_2O-, DMSO-, and BPA-exposed P0 nematodes. The experiments were performed on 4 biological replicates of 30 gonads each that were processed through the NucleoSpin RNA XS, Macherey Nagel kit. cDNA was synthesized using the SMART-Seq v4 Ultra Low Input RNA Kit for sequencing, amplified 10×, and purified using agentcourt AMPure beads.

Nextera XT Library Prep Kit was used to prepare the sequencing libraries from 1 ng cDNA. Single-end sequencing at 50-bp length was performed on an Illumina Hiseq 4000 system (Illumina, CA, USA), and a total of ~350 million reads was obtained for 12 samples (3 treatment groups × 4 replicates/group). Data quality checks were performed using the FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). RNA-seq reads passing quality control (QC) were analyzed using a pipeline comprised of HISAT (Kim et al., 2015), StringTie (Pertea et al., 2015), and Ballgown (Frazee et al., 2015) tools. HISAT was used to align reads against the *C. elegans* genome to discover the locations from which the reads originated and to determine the transcript splice sites. Then, StringTie was used to assemble the RNA-seq alignments into potential transcripts. Ballgown was used to identify the transcripts and genes that were differentially expressed between the BPA and DMSO groups, between the BPA and control (water) groups, and between the DMSO and control groups. FPKMs for each transcript were obtained by Ballgown and used as the expression measure. We filtered out the low-abundance transcripts and kept those having a mean FPKM > 1 across all samples. To test the transcriptional impact of BPA on individual chromosomes, we applied a Student's t test to determine whether the differences in the mean

log2(FPKM + 1) values between the BPA and DMSO groups were significant for all transcripts with FPKM > 1 on each chromosome. $p \leq 0.05$ was considered significant.

ChIP-Seq and Multiplex PTM Assay

Histone modification H3K9me3 and H3K27me3 ChIP-seq data were generated as a service by Active Motif using their in-house antibodies from 3 biological repeats of frozen F3 nematode populations, with 200 μL worms per sample repeat. The sequencing data were obtained through Illumina Nextseq and mapped to ce10 genome by Burrows-Wheeler Aligner (BWA) algorithm (Li and Durbin, 2009). Following pooling of the sequencing data per exposure category (Yang et al., 2014), the data were normalized to input and million reads to produce a signal track file by MACS2 (Zhang et al., 2008). For chromosome-wide mark distribution analysis, each chromosome was divided into 100 sub-regions and average fold enrichment score per base in sub-regions. We normalized signals with *Z* score for each chromosome and each sample.

For gene body histone modification analysis, deepTools (Ramírez et al., 2014) was utilized to obtain aggregated signal from −500 bp of the upstream transcription start site (TSS) to +500 bp of the downstream transcription end site (TES). We first summarized genes with multiple transcripts into a single gene by the one with the most significant difference from BPA and DMSO from RNA-seq results. Silenced genes were defined as genes expressed in the lowest 25% (Q1, 1,801 genes) of all genes in the DMSO group, and upregulated genes were defined as silenced genes upregulated more than 2-fold after BPA treatment (244 genes) based on RNA-seq results. We called peaks by MACS2 broad peak function with q value = 0.1 (cutoff). Broad peak is used as a peak-calling category when analyzing data for protein-DNA association with broader DNA coverage, such as for H3K9me3 and H3K27me3. It joins nearby narrower peak calling into one broader peak. To compare differential peak, unique peak method was used to compare BPA and DMSO samples (Steinhauser et al., 2016). Non-overlapping broad peaks called by MACS2 were defined as unique peaks. Unique peaks from BPA and DMSO in 100 sub-regions along each chromosome were compared. We further define peaked genes as genes with any peak calling in gene body region. Unless specified, analyses were conducted by R 3.4.0 (R Core Team, 2017) and Bioconductor (Huber et al., 2015).

The multiplex PTM quantitation assay was also generated as service by Active Motif on a Luminex platform, and it was performed on pooled samples (totaling 100 μL) generated from 3–4 individual repeats per exposure condition.

Statistical Analyses

Unless indicated otherwise, an unpaired t test assuming unequal variance with Welch's correction was applied. For multi-group comparisons, a one-way ANOVA with Sidak correction or two-way ANOVA was used.

DATA AND SOFTWARE AVAILABILITY

The accession numbers for the ChIP-seq and RNA-seq data reported in this paper are GEO: GSE113187 and GSE113266.

SUPPLEMENTAL INFORMATION

Supplemental Information includes eight figures and five tables and can be found with this article online at https://doi.org/10.1016/j.celrep.2018.04.078.

ACKNOWLEDGMENTS

P.A. is supported by NIH/NIEHS R01 ES02748701 and the Burroughs Wellcome Foundation. J.C. received support from NIH/NIEHS T32 ES015457 Training in Molecular Toxicology, the North American Graduate Fellowship, the NSF AGEP Competitive Edge, the NSF Graduate Research Fellowship, and the Eugene-Cota Robles Fellowship. L.T. is supported by the NIH Training Grant in Genomic Analysis and Interpretation T32 HG002536. G.G. is supported by NIH/NIEHS R25 ES02550703. Z.K. was supported by an American Heart Association post-doctoral fellowship (17POST33670739) and the Iris Cantor-UCLA Executive Advisory Board/CTSI Pilot Award. M.M. was supported by a Dissertation Year Fellowship (University of California, Los Angeles). X.Y. is supported by NIH/NIDDK R01 DK104363 and NIH/NINDS R21 NS103088.

AUTHOR CONTRIBUTIONS

J.C., L.T., M.M., and G.G. performed the experiments. J.C., L.T., Z.K., Y.-W.C., M.P., X.Y., and P.A. analyzed and interpreted the results. J.C., L.T., Z.K., Y.-W.C., X.Y., and P.A. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 7, 2017
Revised: March 15, 2018
Accepted: April 17, 2018
Published: May 22, 2018

REFERENCES

Agger, K., Cloos, P.A., Christensen, J., Pasini, D., Rose, S., Rappsilber, J., Issaeva, I., Canaani, E., Salcini, A.E., and Helin, K. (2007). UTX and JMJD3 are histone H3K27 demethylases involved in HOX gene regulation and development. Nature *449*, 731–734.

Allard, P., and Colaiácovo, M.P. (2011). Mechanistic insights into the action of Bisphenol A on the germline using C. elegans. Cell Cycle *10*, 183–184.

Anway, M.D., Cupp, A.S., Uzumcu, M., and Skinner, M.K. (2005). Epigenetic transgenerational actions of endocrine disruptors and male fertility. Science *308*, 1466–1469.

Anway, M.D., Leathers, C., and Skinner, M.K. (2006). Endocrine disruptor vinclozolin induced epigenetic transgenerational adult-onset disease. Endocrinology *147*, 5515–5523.

Bender, L.B., Cao, R., Zhang, Y., and Strome, S. (2004). The MES-2/MES-3/MES-6 complex and regulation of histone H3 methylation in C. elegans. Curr. Biol. *14*, 1639–1643.

Bender, L.B., Suh, J., Carroll, C.R., Fong, Y., Fingerman, I.M., Briggs, S.D., Cao, R., Zhang, Y., Reinke, V., and Strome, S. (2006). MES-4: an autosome-associated histone methyltransferase that participates in silencing the X chromosomes in the C. elegans germ line. Development *133*, 3907–3917.

Bessler, J.B., Andersen, E.C., and Villeneuve, A.M. (2010). Differential localization and independent acquisition of the H3K9me2 and H3K9me3 chromatin modifications in the Caenorhabditis elegans adult germ line. PLoS Genet. *6*, e1000830.

Bhalla, N., and Dernburg, A.F. (2005). A conserved checkpoint monitors meiotic chromosome synapsis in Caenorhabditis elegans. Science *310*, 1683–1686.

Bhan, A., Hussain, I., Ansari, K.I., Bobzean, S.A., Perrotti, L.I., and Mandal, S.S. (2014). Histone methyltransferase EZH2 is transcriptionally induced by estradiol as well as estrogenic endocrine disruptors bisphenol-A and diethylstilbestrol. J. Mol. Biol. *426*, 3426–3441.

Boros, J., Arnoult, N., Stroobant, V., Collet, J.F., and Decottignies, A. (2014). Polycomb repressive complex 2 and H3K27me3 cooperate with H3K9 methylation to maintain heterochromatin protein 1α at chromatin. Mol. Cell. Biol. *34*, 3662–3674.

Chen, Y., Shu, L., Qiu, Z., Lee, D.Y., Settle, S.J., Que Hee, S., Telesca, D., Yang, X., and Allard, P. (2016). Exposure to the BPA-Substitute Bisphenol S Causes Unique Alterations of Germline Function. PLoS Genet. *12*, e1006223.

Doherty, L.F., Bromer, J.G., Zhou, Y., Aldad, T.S., and Taylor, H.S. (2010). In utero exposure to diethylstilbestrol (DES) or bisphenol-A (BPA) increases EZH2 expression in the mammary gland: an epigenetic mechanism linking endocrine disruptors to breast cancer. Horm. Cancer *1*, 146–155.

Frazee, A.C., Pertea, G., Jaffe, A.E., Langmead, B., Salzberg, S.L., and Leek, J.T. (2015). Ballgown bridges the gap between transcriptome assembly and expression analysis. Nat. Biotechnol. *33*, 243–246.

Gapp, K., Jawaid, A., Sarkies, P., Bohacek, J., Pelczar, P., Prados, J., Farinelli, L., Miska, E., and Mansuy, I.M. (2014). Implication of sperm RNAs in transgenerational inheritance of the effects of early trauma in mice. Nat. Neurosci. *17*, 667–669.

Garrigues, J.M., Sidoli, S., Garcia, B.A., and Strome, S. (2015). Defining heterochromatin in C. elegans through genome-wide analysis of the heterochromatin protein 1 homolog HPL-2. Genome Res. *25*, 76–88.

Gartner, A., Boag, P.R., and Blackwell, T.K. (2008). Germline survival and apoptosis. WormBook1-20.

Gaydos, L.J., Rechtsteiner, A., Egelhofer, T.A., Carroll, C.R., and Strome, S. (2012). Antagonism between MES-4 and Polycomb repressive complex 2 promotes appropriate gene expression in C. elegans germ cells. Cell Rep. *2*, 1169–1177.

Gaydos, L.J., Wang, W., and Strome, S. (2014). Gene repression. H3K27me and PRC2 transmit a memory of repression across generations and during development. Science *345*, 1515–1518.

Greer, E.L., Beese-Sims, S.E., Brookes, E., Spadafora, R., Zhu, Y., Rothbart, S.B., Aristizábal-Corrales, D., Chen, S., Badeaux, A.I., Jin, Q., et al. (2014). A histone methylation network regulates transgenerational epigenetic memory in C. elegans. Cell Rep. *7*, 113–126.

Hajkova, P., Ancelin, K., Waldmann, T., Lacoste, N., Lange, U.C., Cesari, F., Lee, C., Almouzni, G., Schneider, R., and Surani, M.A. (2008). Chromatin dynamics during epigenetic reprogramming in the mouse germ line. Nature *452*, 877–881.

Heard, E., and Martienssen, R.A. (2014). Transgenerational epigenetic inheritance: myths and mechanisms. Cell *157*, 95–109.

Ho, J.W., Jung, Y.L., Liu, T., Alver, B.H., Lee, S., Ikegami, K., Sohn, K.A., Minoda, A., Tolstorukov, M.Y., Appert, A., et al. (2014). Comparative analysis of metazoan chromatin organization. Nature *512*, 449–452.

Hu, Q., Chen, J., Zhang, J., Xu, C., Yang, S., and Jiang, H. (2016). IOX1, a JMJD2A inhibitor, suppresses the proliferation and migration of vascular smooth muscle cells induced by angiotensin II by regulating the expression of cell cycle-related proteins. Int. J. Mol. Med. *37*, 189–196.

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods *12*, 115–121.

Hughes, V. (2014). Epigenetics: The sins of the father. Nature *507*, 22–24.

Iwatani, M., Ikegami, K., Kremenska, Y., Hattori, N., Tanaka, S., Yagi, S., and Shiota, K. (2006). Dimethyl sulfoxide has an impact on epigenetic profile in mouse embryoid body. Stem Cells *24*, 2549–2556.

Juang, J.K., and Liu, H.J. (1987). The effect of DMSO on natural DNA conformation in enhancing transcription. Biochem. Biophys. Res. Commun. *146*, 1458–1464.

Kamath, R.S., and Ahringer, J. (2003). Genome-wide RNAi screening in Caenorhabditis elegans. Methods *30*, 313–321.

Kelly, W.G., and Fire, A. (1998). Chromatin silencing and the maintenance of a functional germline in Caenorhabditis elegans. Development *125*, 2451–2456.

Kelly, W.G., Xu, S., Montgomery, M.K., and Fire, A. (1997). Distinct requirements for somatic and germline expression of a generally expressed Caenorhabditis elegans gene. Genetics *146*, 227–238.

Kim, A., and Dean, A. (2004). Developmental stage differences in chromatin subdomains of the beta-globin locus. Proc. Natl. Acad. Sci. USA *101*, 7028–7033.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360.

King, O.N., Li, X.S., Sakurai, M., Kawamura, A., Rose, N.R., Ng, S.S., Quinn, A.M., Rai, G., Mott, B.T., Beswick, P., et al. (2010). Quantitative high-throughput screening identifies 8-hydroxyquinolines as cell-active histone demethylase inhibitors. PLoS ONE *5*, e15535.

Kishimoto, S., Uno, M., Okabe, E., Nono, M., and Nishida, E. (2017). Environmental stresses induce transgenerationally inheritable survival advantages via germline-to-soma communication in Caenorhabditis elegans. Nat. Commun. *8*, 14031.

Klosin, A., Casas, E., Hidalgo-Carcedo, C., Vavouri, T., and Lehner, B. (2017). Transgenerational transmission of environmental information in *C. elegans*. Science *356*, 320–323.

Kruidenier, L., Chung, C.W., Cheng, Z., Liddle, J., Che, K., Joberty, G., Bantscheff, M., Bountra, C., Bridges, A., Diallo, H., et al. (2012). A selective jumonji H3K27 demethylase inhibitor modulates the proinflammatory macrophage response. Nature *488*, 404–408.

Leung, D., Du, T., Wagner, U., Xie, W., Lee, A.Y., Goyal, P., Li, Y., Szulwach, K.E., Jin, P., Lorincz, M.C., and Ren, B. (2014). Regulation of DNA methylation turnover at LTR retrotransposons and imprinted loci by the histone methyltransferase Setdb1. Proc. Natl. Acad. Sci. USA *111*, 6690–6695.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Liu, T., Rechtsteiner, A., Egelhofer, T.A., Vielle, A., Latorre, I., Cheung, M.S., Ercan, S., Ikegami, K., Jensen, M., Kolasinska-Zwierz, P., et al. (2011). Broad chromosomal domains of histone modification patterns in C. elegans. Genome Res. *21*, 227–236.

Liu, S., Brind'Amour, J., Karimi, M.M., Shirane, K., Bogutz, A., Lefebvre, L., Sasaki, H., Shinkai, Y., and Lorincz, M.C. (2014). Setdb1 is required for germline development and silencing of H3K9me3-marked endogenous retroviruses in primordial germ cells. Genes Dev. *28*, 2041–2055.

Lundby, Z., Camacho, J., and Allard, P. (2016). Fast Functional Germline and Epigenetic Assays in the Nematode Caenorhabditis elegans. Methods Mol. Biol. *1473*, 99–107.

Manikkam, M., Tracey, R., Guerrero-Bosagna, C., and Skinner, M.K. (2013). Plastics derived endocrine disruptors (BPA, DEHP and DBP) induce epigenetic transgenerational inheritance of obesity, reproductive disease and sperm epimutations. PLoS ONE *8*, e55387.

Pasini, D., Cloos, P.A., Walfridsson, J., Olsson, L., Bukowski, J.P., Johansen, J.V., Bak, M., Tommerup, N., Rappsilber, J., and Helin, K. (2010). JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. Nature *464*, 306–310.

Peng, J.C., Valouev, A., Swigut, T., Zhang, J., Zhao, Y., Sidow, A., and Wysocka, J. (2009). Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. Cell *139*, 1290–1302.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295.

R Core Team (2017). R: A language and environment for statistical computing (R Foundation for Statistical Computing).

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, W187–W191.

Rechavi, O., Houri-Ze'evi, L., Anava, S., Goh, W.S.S., Kerk, S.Y., Hannon, G.J., and Hobert, O. (2014). Starvation-induced transgenerational inheritance of small RNAs in C. elegans. Cell *158*, 277–287.

Rudgalvyte, M., Peltonen, J., Lakso, M., and Wong, G. (2017). Chronic MeHg exposure modifies the histone H3K4me3 epigenetic landscape in Caenorhabditis elegans. Comp. Biochem. Physiol. C Toxicol. Pharmacol. *191*, 109–116.

Schaner, C.E., and Kelly, W.G. (2006). Germline chromatin. WormBook, 1–14.

Schiller, R., Scozzafava, G., Tumber, A., Wickens, J.R., Bush, J.T., Rai, G., Lejeune, C., Choi, H., Yeh, T.L., Chan, M.C., et al. (2014). A cell-permeable ester derivative of the JmjC histone demethylase inhibitor IOX1. ChemMedChem *9*, 566–571.

Schönfelder, G., Wittfoht, W., Hopp, H., Talsness, C.E., Paul, M., and Chahoud, I. (2002). Parent bisphenol A accumulation in the human maternal-fetal-placental unit. Environ. Health Perspect. *110*, A703–A707.

Seki, Y., Hayashi, K., Itoh, K., Mizugaki, M., Saitou, M., and Matsui, Y. (2005). Extensive and orderly reprogramming of genome-wide chromatin

modifications associated with specification and early development of germ cells in mice. Dev. Biol. *278*, 440–458.

Siklenka, K., Erkek, S., Godmann, M., Lambrot, R., McGraw, S., Lafleur, C., Cohen, T., Xia, J., Suderman, M., Hallett, M., et al. (2015). Disruption of histone methylation in developing sperm impairs offspring health transgenerationally. Science *350*, aab2006.

Singh, S., and Li, S.S. (2012). Epigenetic effects of environmental chemicals bisphenol A and phthalates. Int. J. Mol. Sci. *13*, 10143–10153.

Steinhauser, S., Kurzawa, N., Eils, R., and Herrmann, C. (2016). A comprehensive comparison of tools for differential ChIP-seq analysis. Brief. Bioinform. *17*, 953–966.

Stiernagle, T. (2006). Maintenance of C. elegans. WormBook, 1–11.

Tang, W.W., Kobayashi, T., Irie, N., Dietmann, S., and Surani, M.A. (2016). Specification and epigenetic programming of the human germ line. Nat. Rev. Genet. *17*, 585–600.

Towbin, B.D., González-Aguilera, C., Sack, R., Gaidatzis, D., Kalck, V., Meister, P., Askjaer, P., and Gasser, S.M. (2012). Step-wise methylation of histone H3K9 positions heterochromatin at the nuclear periphery. Cell *150*, 934–947.

Trapphoff, T., Heiligentag, M., El Hajj, N., Haaf, T., and Eichenlaub-Ritter, U. (2013). Chronic exposure to a low concentration of bisphenol A during follicle culture affects the epigenetic status of germinal vesicles and metaphase II oocytes. Fertil. Steril. *100*, 1758–1767.e1.

Vandenberg, L.N., Chahoud, I., Heindel, J.J., Padmanabhan, V., Paumgartten, F.J.R., and Schoenfelder, G. (2010). Urinary, circulating, and tissue biomonitoring studies indicate widespread exposure to bisphenol A. Environ. Health Perspect. *118*, 1055–1070.

Whetstine, J.R., Nottke, A., Lan, F., Huarte, M., Smolikov, S., Chen, Z., Spooner, E., Li, E., Zhang, G., Colaiacovo, M., and Shi, Y. (2006). Reversal of histone lysine trimethylation by the JMJD2 family of histone demethylases. Cell *125*, 467–481.

Wu, C.T., and Morris, J.R. (2001). Genes, genetics, and epigenetics: a correspondence. Science *293*, 1103–1105.

Yang, Y., Fear, J., Hu, J., Haecker, I., Zhou, L., Renne, R., Bloom, D., and McIntyre, L.M. (2014). Leveraging biological replicates to improve analysis in ChIP-seq experiments. Comput. Struct. Biotechnol. J. *9*, e201401002.

Yeo, M., Berglund, K., Hanna, M., Guo, J.U., Kittur, J., Torres, M.D., Abramowitz, J., Busciglio, J., Gao, Y., Birnbaumer, L., and Liedtke, W.B. (2013). Bisphenol A delays the perinatal chloride shift in cortical neurons by epigenetic effects on the Kcc2 promoter. Proc. Natl. Acad. Sci. USA *110*, 4315–4320.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhong, S.H., Liu, J.Z., Jin, H., Lin, L., Li, Q., Chen, Y., Yuan, Y.X., Wang, Z.Y., Huang, H., Qi, Y.J., et al. (2013). Warm temperatures induce transgenerational epigenetic release of RNA silencing by inhibiting siRNA biogenesis in Arabidopsis. Proc. Natl. Acad. Sci. USA *110*, 9171–9176.

References

1 Sultana J, Cutroneo P, Trifirò G. Clinical and economic burden of adverse drug reactions. *J Pharmacol Pharmacother* 2013;**4**:S73–7. https://doi.org/10.4103/0976-500X.120957.

2 Afshin A, Sur PJ, Fay KA, Cornaby L, Ferrara G, Salama JS, *et al.* Health effects of dietary risks in 195 countries, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* 2019;**393**:1958–72. https://doi.org/10.1016/S0140-6736(19)30041-8.

3 Jardim TV, Mozaffarian D, Abrahams-Gessel S, Sy S, Lee Y, Liu J, *et al.* Cardiometabolic disease costs associated with suboptimal diet in the United States: A cost analysis based on a microsimulation model. *PLOS Med* 2019;**16**:e1002981.

4 Scarborough P, Bhatnagar P, Wickramasinghe KK, Allender S, Foster C, Rayner M. The economic burden of ill health due to diet, physical inactivity, smoking, alcohol and obesity in the UK: an update to 2006–07 NHS costs. *J Public Health (Bangkok)* 2011;**33**:527–35. https://doi.org/10.1093/pubmed/fdr033.

5 Sacks JJ, Gonzales KR, Bouchery EE, Tomedi LE, Brewer RD. 2010 National and State Costs of Excessive Alcohol Consumption. *Am J Prev Med* 2015;**49**:e73–9. https://doi.org/10.1016/j.amepre.2015.05.031.

6 Griswold MG, Fullman N, Hawley C, Arian N, Zimsen SRM, Tymeson HD, *et al.* Alcohol use and burden for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* 2018;**392**:1015–35. https://doi.org/10.1016/S0140-6736(18)31310-2.

7 Igarashi Y, Nakatsu N, Yamashita T, Ono A, Ohno Y, Urushidani T, *et al.* Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res* 2015;**43**:D921-7. https://doi.org/10.1093/nar/gku955.

8 *Microbial Status and Genetic Evaluation of Mice and Rats: Proceedings of the 1999*

US/Japan Conference. Washington (DC); 2000.

9 Chella Krishnan K, Kurt Z, Barrere-Cain R, Sabir S, Das A, Floyd R, *et al.* Integration of Multi-omics Data from Mouse Diversity Panel Highlights Mitochondrial Dysfunction in Non-alcoholic Fatty Liver Disease. *Cell Syst* 2018;**6**:103-115.e7. https://doi.org/10.1016/j.cels.2017.12.006.

10 Wong SK, Chin K-Y, Suhaimi FH, Fairus A, Ima-Nirwana S. Animal models of metabolic syndrome: a review. *Nutr Metab (Lond)* 2016;**13**:65. https://doi.org/10.1186/s12986-016-0123-9.

11 Weinhouse C, Truong L, Meyer JN, Allard P. Caenorhabditis elegans as an emerging model system in environmental epigenetics. *Environ Mol Mutagen* 2018;**59**:560–75. https://doi.org/10.1002/em.22203.

12 Prior H, Haworth R, Labram B, Roberts R, Wolfreys A, Sewell F. Justification for species selection for pharmaceutical toxicity studies. *Toxicol Res (Camb)* 2020;**9**:758–70. https://doi.org/10.1093/toxres/tfaa081.

13 Johnson RJ, Perez-Pozo SE, Sautin YY, Manitius J, Sanchez-Lozada LG, Feig DI, *et al.* Hypothesis: could excessive fructose intake and uric acid cause type 2 diabetes? *Endocr Rev* 2009;**30**:96–116. https://doi.org/10.1210/er.2008-0033.

14 Halpern KB, Shenhav R, Matcovitch-Natan O, Toth B, Lemze D, Golan M, *et al.* Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* 2017;**542**:352–6. https://doi.org/10.1038/nature21065.

15 Malik R, Selden C, Hodgson H. The role of non-parenchymal cells in liver growth. *Semin Cell Dev Biol* 2002;**13**:425–31.

16 Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, *et al.* Genetic effects on gene expression across human tissues. *Nature* 2017;**550**:204–13. https://doi.org/10.1038/nature24277.

17 Taylor JY, Kraja AT, de Las Fuentes L, Stanfill AG, Clark A, Cashion A. An overview of the genomics of metabolic syndrome. *J Nurs Scholarsh an Off Publ Sigma Theta Tau Int Honor Soc Nurs* 2013;**45**:52–9. https://doi.org/10.1111/j.1547-5069.2012.01484.x.

18 Áine D, Marie V, Amanda D, Hong-Hee W, L. RJ, S. FI, *et al.* Tissue-specific genetic features inform prediction of drug side effects in clinical trials. *Sci Adv* 2022;**6**:eabb6242. https://doi.org/10.1126/sciadv.abb6242.

19 Chuang H-Y, Hofree M, Ideker T. A decade of systems biology. *Annu Rev Cell Dev Biol* 2010;**26**:721–44. https://doi.org/10.1146/annurev-cellbio-100109-104122.

20 Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011;**12**:56–68. https://doi.org/10.1038/nrg2918.

21 Tu Z, Keller MP, Zhang C, Rabaglia ME, Greenawalt DM, Yang X, *et al.* Integrative analysis of a cross-loci regulation network identifies App as a gene regulating insulin secretion from pancreatic islets. *PLoS Genet* 2012;**8**:e1003107. https://doi.org/10.1371/journal.pgen.1003107.

22 Shu L, Chan KHK, Zhang G, Huan T, Kurt Z, Zhao Y, *et al.* Shared genetic regulatory networks for cardiovascular disease and type 2 diabetes in multiple populations of diverse ethnicities in the United States. *PLoS Genet* 2017;**13**:e1007040. https://doi.org/10.1371/journal.pgen.1007040.

23 Guney E, Menche J, Vidal M, Barábasi A-L. Network-based in silico drug efficacy screening. *Nat Commun* 2016;**7**:10331. https://doi.org/10.1038/ncomms10331.

24 Cheng F, Desai RJ, Handy DE, Wang R, Schneeweiss S, Barabási A-L, *et al.* Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat Commun* 2018;**9**:2691. https://doi.org/10.1038/s41467-018-05116-5.

25 Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, *et al.* The STRING database in 2017: quality-controlled protein–protein association networks, made broadly

accessible. *Nucleic Acids Res* 2017;**45**:D362–8. https://doi.org/10.1093/nar/gkw937.

26 ALGHAMDI N, Chang W, Dang P, Lu X, Wan C, Gampala S, *et al.* A graph neural network model to estimate cell-wise metabolic flux using single-cell RNA-seq data. *Genome Res* 2021. https://doi.org/10.1101/gr.271205.120.

27 Blencowe M, Arneson D, Ding J, Chen Y-W, Saleem Z, Yang X. Network modeling of single-cell omics data: challenges, opportunities, and progresses. *Emerg Top Life Sci* 2019;**3**:379–98. https://doi.org/10.1042/ETLS20180176.

28 Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* 2018;**50**:96. https://doi.org/10.1038/s12276-018-0071-8.

29 Schwartzman O, Tanay A. Single-cell epigenomics: techniques and emerging applications. *Nat Rev Genet* 2015;**16**:716.

30 Bell S, Abedini J, Ceger P, Chang X, Cook B, Karmaus AL, *et al.* An integrated chemical environment with tools for chemical safety testing. *Toxicol Vitr* 2020;**67**:104916. https://doi.org/https://doi.org/10.1016/j.tiv.2020.104916.

31 Kavlock R, Chandler K, Houck K, Hunter S, Judson R, Kleinstreuer N, *et al.* Update on EPA's ToxCast Program: Providing High Throughput Decision Support Tools for Chemical Risk Management. *Chem Res Toxicol* 2012;**25**:1287–302. https://doi.org/10.1021/tx3000939.

32 Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res* 2021;**49**:D1388–95. https://doi.org/10.1093/nar/gkaa971.

33 Dimitrov SD, Diderich R, Sobanski T, Pavlov TS, Chankov G V, Chapkanov AS, *et al.* QSAR Toolbox - workflow and major functionalities. *SAR QSAR Environ Res* 2016;**27**:203–19. https://doi.org/10.1080/1062936X.2015.1136680.

34 Ding J, Blencowe M, Nghiem T, Ha S, Chen Y-W, Li G, *et al.* Mergeomics 2.0: a web

server for multi-omics data integration to elucidate disease networks and predict therapeutics. *Nucleic Acids Res* 2021. https://doi.org/10.1093/nar/gkab405.

35 Morgan S, Grootendorst P, Lexchin J, Cunningham C, Greyson D. The cost of drug development: a systematic review. *Health Policy* 2011;**100**:4–17. https://doi.org/10.1016/j.healthpol.2010.12.002.

36 Yu HWH. Bridging the translational gap: collaborative drug development and dispelling the stigma of commercialization. *Drug Discov Today* 2016;**21**:299–305. https://doi.org/https://doi.org/10.1016/j.drudis.2015.10.013.

37 Lin Z, Will Y. Evaluation of Drugs With Specific Organ Toxicities in Organ-Specific Cell Lines. *Toxicol Sci* 2011;**126**:114–27. https://doi.org/10.1093/toxsci/kfr339.

38 Denayer T, Stöhr T, Van Roy M. Animal models in translational medicine: Validation and prediction. *New Horizons Transl Med* 2014;**2**:5–11. https://doi.org/https://doi.org/10.1016/j.nhtm.2014.08.001.

39 Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* 2017;**171**:1437-1452.e17. https://doi.org/10.1016/j.cell.2017.10.049.

40 Cheng F, Desai RJ, Handy DE, Wang R, Schneeweiss S, Barabási A-L, *et al.* Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat Commun* 2018;**9**:2691. https://doi.org/10.1038/s41467-018-05116-5.

41 Hall CJ, Wicker SM, Chien A-T, Tromp A, Lawrence LM, Sun X, *et al.* Repositioning drugs for inflammatory disease – fishing for new anti-inflammatory agents. *Dis Model Mech* 2014;**7**:1069.

42 Corbett A, Pickett J, Burns A, Corcoran J, Dunnett SB, Edison P, *et al.* Drug repositioning for Alzheimer' s disease. *Nat Rev Drug Discov* 2012;**11**:833. https://doi.org/10.1038/nrd3869.

43 Godoy P, Bolt HM. Toxicogenomic-based approaches predicting liver toxicity in vitro. *Arch Toxicol* 2012;**86**:1163–4. https://doi.org/10.1007/s00204-012-0892-5.

44 Low Y, Uehara T, Minowa Y, Yamada H, Ohno Y, Urushidani T, *et al.* Predicting Drug-Induced Hepatotoxicity Using QSAR and Toxicogenomics Approaches. *Chem Res Toxicol* 2011;**24**:1251–62. https://doi.org/10.1021/tx200148a.

45 AbdulHameed MD, Ippolito DL, Stallings JD, Wallqvist A. Mining kidney toxicogenomic data by using gene co-expression modules. *BMC Genomics* 2016;**17**:790. https://doi.org/10.1186/s12864-016-3143-y.

46 Toutain P-L, Ferran A, Bousquet-Melou A. Species differences in pharmacokinetics and pharmacodynamics. *Handb Exp Pharmacol* 2010:19–48. https://doi.org/10.1007/978-3-642-10324-7_2.

47 Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters 2018. https://doi.org/10.1093/biostatistics/kxx069.

48 Davis AP, Grondin CJ, Johnson RJ, Sciaky D, King BL, McMorran R, *et al.* The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Res* 2017;**45**:D972–8. https://doi.org/10.1093/nar/gkw838.

49 Wang Z, Clark NR, Ma'ayan A. Drug-induced adverse events prediction with the LINCS L1000 data. *Bioinformatics* 2016;**32**:2338–45. https://doi.org/10.1093/bioinformatics/btw168.

50 Wang Z, Monteiro CD, Jagodnik KM, Fernandez NF, Gundersen GW, Rouillard AD, *et al.* Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nat Commun* 2016;**7**:12846.

51 Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 2013;**41**:D991–5. https://doi.org/10.1093/nar/gks1193.

52 Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, *et al.* ArrayExpress update—simplifying data submissions. *Nucleic Acids Res* 2015;**43**:D1113–6. https://doi.org/10.1093/nar/gku1057.

53 Shu L, Zhao Y, Kurt Z, Byars SG, Tukiainen T, Kettunen J, *et al.* Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems. *BMC Genomics* 2016;**17**:874. https://doi.org/10.1186/s12864-016-3198-9.

54 Wang Z, Lachmann A, Keenan AB, Ma'ayan A. L1000FWD: fireworks visualization of drug-induced transcriptomic signatures. *Bioinformatics* 2018;**34**:2150–2.

55 Liss KHH, Finck BN. PPARs and nonalcoholic fatty liver disease. *Biochimie* 2017;**136**:65–74. https://doi.org/10.1016/j.biochi.2016.11.009.

56 Abd El-Haleim EA, Bahgat AK, Saleh S. Resveratrol and fenofibrate ameliorate fructose-induced nonalcoholic steatohepatitis by modulation of genes expression. *World J Gastroenterol* 2016;**22**:2931–48. https://doi.org/10.3748/wjg.v22.i10.2931.

57 Ratziu V, Charlotte F, Bernhardt C, Giral P, Halbron M, Lenaour G, *et al.* Long-term efficacy of rosiglitazone in nonalcoholic steatohepatitis: results of the fatty liver improvement by rosiglitazone therapy (FLIRT 2) extension trial. *Hepatology* 2010;**51**:445–53. https://doi.org/10.1002/hep.23270.

58 Ratziu V, Giral P, Jacqueminet S, Charlotte F, Hartemann-Heurtier A, Serfaty L, *et al.* Rosiglitazone for nonalcoholic steatohepatitis: one-year results of the randomized placebo-controlled Fatty Liver Improvement with Rosiglitazone Therapy (FLIRT) Trial. *Gastroenterology* 2008;**135**:100–10. https://doi.org/10.1053/j.gastro.2008.03.078.

59 Zhang N, Lu Y, Shen X, Bao Y, Cheng J, Chen L, *et al.* Fenofibrate treatment attenuated chronic endoplasmic reticulum stress in the liver of nonalcoholic fatty liver disease mice. *Pharmacology* 2015;**95**:173–80. https://doi.org/10.1159/000380952.

60 Laurin J, Lindor KD, Crippin JS, Gossard A, Gores GJ, Ludwig J, *et al.* Ursodeoxycholic

acid or clofibrate in the treatment of non-alcohol-induced steatohepatitis: a pilot study. *Hepatology* 1996;**23**:1464–7. https://doi.org/10.1002/hep.510230624.

61 Fernandez-Miranda C, Perez-Carreras M, Colina F, Lopez-Alonso G, Vargas C, Solis-Herruzo JA. A pilot trial of fenofibrate for the treatment of non-alcoholic fatty liver disease. *Dig Liver Dis* 2008;**40**:200–5. https://doi.org/10.1016/j.dld.2007.10.002.

62 Hertz R, Bar-Tana J. Peroxisome proliferator-activated receptor (PPAR) alpha activation and its consequences in humans. *Toxicol Lett* 1998;**102**–**103**:85–90.

63 Kawaguchi K, Sakaida I, Tsuchiya M, Omori K, Takami T, Okita K. Pioglitazone prevents hepatic steatosis, fibrosis, and enzyme-altered lesions in rat liver cirrhosis induced by a choline-deficient L-amino acid-defined diet. *Biochem Biophys Res Commun* 2004;**315**:187–95. https://doi.org/10.1016/j.bbrc.2004.01.038.

64 Nan Y-M, Fu N, Wu W-J, Liang B-L, Wang R-Q, Zhao S-X, *et al.* Rosiglitazone prevents nutritional fibrosis and steatohepatitis in mice. *Scand J Gastroenterol* 2009;**44**:358–65. https://doi.org/10.1080/00365520802530861.

65 Neuschwander-Tetri BA, Brunt EM, Wehmeier KR, Oliver D, Bacon BR. Improved nonalcoholic steatohepatitis after 48 weeks of treatment with the PPAR-gamma ligand rosiglitazone. *Hepatology* 2003;**38**:1008–17. https://doi.org/10.1053/jhep.2003.50420.

66 Torres DM, Jones FJ, Shaw JC, Williams CD, Ward JA, Harrison SA. Rosiglitazone versus rosiglitazone and metformin versus rosiglitazone and losartan in the treatment of nonalcoholic steatohepatitis in humans: a 12-month randomized, prospective, open- label trial. *Hepatology* 2011;**54**:1631–9. https://doi.org/10.1002/hep.24558.

67 Pastori D, Polimeni L, Baratta F, Pani A, Del Ben M, Angelico F. The efficacy and safety of statins for the treatment of non-alcoholic fatty liver disease. *Dig Liver Dis* 2015;**47**:4–11. https://doi.org/https://doi.org/10.1016/j.dld.2014.07.170.

68 Sigler MA, Congdon L, Edwards KL. An Evidence-Based Review of Statin Use in Patients

With Nonalcoholic Fatty Liver Disease. *Clin Med Insights Gastroenterol* 2018. https://doi.org/10.1177/1179552218787502.

69 Park HS, Jang JE, Ko MS, Woo SH, Kim BJ, Kim HS, *et al.* Statins Increase Mitochondrial and Peroxisomal Fatty Acid Oxidation in the Liver and Prevent Non-Alcoholic Steatohepatitis in Mice. *Diabetes Metab J* 2016;**40**:376–85. https://doi.org/10.4093/dmj.2016.40.5.376.

70 Bravo M, Raurell I, Hide D, Fernández-Iglesias A, Gil M, Barberá A, *et al.* Restoration of liver sinusoidal cell phenotypes by statins improves portal hypertension and histology in rats with NASH. *Sci Rep* 2019;**9**:20183. https://doi.org/10.1038/s41598-019-56366-2.

71 Simon TG, Henson J, Osganian S, Masia R, Chan AT, Chung RT, *et al.* Daily Aspirin Use Associated With Reduced Risk For Fibrosis Progression In Patients With Nonalcoholic Fatty Liver Disease. *Clin Gastroenterol Hepatol* 2019;**17**:2776-2784.e4. https://doi.org/10.1016/j.cgh.2019.04.061.

72 Chella Krishnan K, Floyd RR, Sabir S, Jayasekera DW, Leon-Mimila P V, Jones AE, *et al.* Liver Pyruvate Kinase Promotes NAFLD/NASH in Both Mice and Humans in a Sex-Specific Manner. *Cell Mol Gastroenterol Hepatol* 2021;**11**:389–406. https://doi.org/10.1016/j.jcmgh.2020.09.004.

73 Hui ST, Parks BW, Org E, Norheim F, Che N, Pan C, *et al.* The genetic architecture of NAFLD among inbred strains of mice. *Elife* 2015;**4**:e05607. https://doi.org/10.7554/eLife.05607.

74 Norheim F, Chella Krishnan K, Bjellaas T, Vergnes L, Pan C, Parks BW, *et al.* Genetic regulation of liver lipids in a mouse model of insulin resistance and hepatic steatosis. *Mol Syst Biol* 2021;**17**:e9684. https://doi.org/https://doi.org/10.15252/msb.20209684.

75 Chakravarthy M V, Lodhi IJ, Yin L, Malapaka RR V, Xu HE, Turk J, *et al.* Identification of a physiologically relevant endogenous ligand for PPARalpha in liver. *Cell* 2009;**138**:476–

88. https://doi.org/10.1016/j.cell.2009.05.036.

76 Wu J, Wang C, Li S, Li S, Wang W, Li J, *et al.* Thyroid hormone-responsive SPOT 14 homolog promotes hepatic lipogenesis, and its expression is regulated by Liver X receptor α through a sterol regulatory element-binding protein 1c–dependent mechanism in mice. *Hepatology* 2013;**58**:617–28. https://doi.org/https://doi.org/10.1002/hep.26272.

77 Lee S, Zhang C, Liu Z, Klevstig M, Mukhopadhyay B, Bergentall M, *et al.* Network analyses identify liver-specific targets for treating liver diseases. *Mol Syst Biol* 2017;**13**:938. https://doi.org/10.15252/msb.20177703.

78 Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 2017;**33**:2938–40. https://doi.org/10.1093/bioinformatics/btx364.

79 Gremse M, Chang A, Schomburg I, Grote A, Scheer M, Ebeling C, *et al.* The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res* 2011;**39**:D507–13. https://doi.org/10.1093/nar/gkq968.

80 Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *J Am Stat Assoc* 2004;**99**:909–17. https://doi.org/10.1198/016214504000000683.

81 Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47–e47. https://doi.org/10.1093/nar/gkv007.

82 Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* 2012;**28**:573–80. https://doi.org/10.1093/bioinformatics/btr709.

83 Kuleshov M V, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids*

Res 2016;**44**:W90–7. https://doi.org/10.1093/nar/gkw377.

84 Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**:D353–61. https://doi.org/10.1093/nar/gkw1092.

85 Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res* 2017;**45**:D331-d338. https://doi.org/10.1093/nar/gkw1108.

86 Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, *et al.* A systems biology approach for pathway level analysis. *Genome Res* 2007;**17**:1537–45. https://doi.org/10.1101/gr.6202607.

87 Voichita C, Ansari S, Draghici S. ROntoTools: R Onto-Tools suite 2020.

88 Zhu J, Wiener MC, Zhang C, Fridman A, Minch E, Lum PY, *et al.* Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput Biol* 2007;**3**:e69. https://doi.org/10.1371/journal.pcbi.0030069.

89 Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, *et al.* Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet* 2008;**40**:854–61. https://doi.org/10.1038/ng.167.

90 Derry JM, Zhong H, Molony C, MacNeil D, Guhathakurta D, Zhang B, *et al.* Identification of genes and networks driving cardiovascular and metabolic phenotypes in a mouse F2 intercross. *PLoS One* 2010;**5**:e14319. https://doi.org/10.1371/journal.pone.0014319.

91 Wang SS, Schadt EE, Wang H, Wang X, Ingram-Drake L, Shi W, *et al.* Identification of pathways for atherosclerosis in mice: integration of quantitative trait locus analysis and global gene expression data. *Circ Res* 2007;**101**:e11-30. https://doi.org/10.1161/CIRCRESAHA.107.152975.

92 Yang X, Schadt EE, Wang S, Wang H, Arnold AP, Ingram-Drake L, *et al.* Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res*

2006;**16**:995–1004. https://doi.org/10.1101/gr.5217506.

93 Zhong H, Beaulaurier J, Lum PY, Molony C, Yang X, MacNeil DJ, *et al.* Liver and Adipose Expression Associated SNPs Are Enriched for Association to Type 2 Diabetes. *PLOS Genet* 2010;**6**:e1000932.

94 Clark NR, Hu KS, Feldmann AS, Kou Y, Chen EY, Duan Q, *et al.* The characteristic direction: a geometrical approach to identify differentially expressed genes. *BMC Bioinformatics* 2014;**15**:79. https://doi.org/10.1186/1471-2105-15-79.

95 Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* 2013;**45**:1274–83. https://doi.org/10.1038/ng.2797.

96 Tweedie S, Braschi B, Gray K, Jones TEM, Seal RL, Yates B, *et al.* Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res* 2021;**49**:D939–46. https://doi.org/10.1093/nar/gkaa980.

97 Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics* 2005;**21**:3940–1. https://doi.org/10.1093/bioinformatics/bti623.

98 Zhu K, Hu M, Yuan B, Liu J-X, Liu Y. Aspirin attenuates spontaneous recurrent seizures in the chronically epileptic mice. *Neurol Res* 2017;**39**:744–57. https://doi.org/10.1080/01616412.2017.1326657.

99 FOLCH J, LEES M, SLOANE STANLEY GH. A simple method for the isolation and purification of total lipides from animal tissues. *J Biol Chem* 1957;**226**:497–509.

100 Warnick GR. Enzymatic methods for quantification of lipoprotein lipids. *Methods Enzymol* 1986;**129**:101–23. https://doi.org/10.1016/0076-6879(86)29064-3.

101 Hedrick CC, Castellani LW, Warden CH, Puppione DL, Lusis AJ. Influence of mouse apolipoprotein A-II on plasma lipoproteins in transgenic mice. *J Biol Chem*

1993;**268**:20676–82.

102 Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, *et al.* ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics 2020.

103 Caputo C, Wood E, Jabbour L. Impact of fetal alcohol exposure on body systems: A systematic review. *Birth Defects Res Part C Embryo Today Rev* 2016;**108**:174–80. https://doi.org/https://doi.org/10.1002/bdrc.21129.

104 La Vignera S, Condorelli RA, Balercia G, Vicari E, Calogero AE. Does alcohol have any effect on male reproductive function? A review of literature. *Asian J Androl* 2013;**15**:221–5. https://doi.org/10.1038/aja.2012.118.

105 Nizhnikov ME, Popoola DO, Cameron NM. Transgenerational Transmission of the Effect of Gestational Ethanol Exposure on Ethanol Use-Related Behavior. *Alcohol Clin Exp Res* 2016;**40**:497–506. https://doi.org/10.1111/acer.12978.

106 Hollander J, McNivens M, Pautassi RM, Nizhnikov ME. Offspring of male rats exposed to binge alcohol exhibit heightened ethanol intake at infancy and alterations in T-maze performance. *Alcohol* 2019;**76**:65–71. https://doi.org/10.1016/j.alcohol.2018.07.013.

107 Lam MK-P, Homewood J, Taylor AJ, Mazurski EJ. Second generation effects of maternal alcohol consumption during pregnancy in rats. *Prog Neuro-Psychopharmacology Biol Psychiatry* 2000;**24**:619–31. https://doi.org/https://doi.org/10.1016/S0278-5846(00)00097-X.

108 Yohn NL, Bartolomei MS, Blendy JA. Multigenerational and transgenerational inheritance of drug exposure: The effects of alcohol, opiates, cocaine, marijuana, and nicotine. *Prog Biophys Mol Biol* 2015;**118**:21–33. https://doi.org/10.1016/j.pbiomolbio.2015.03.002.

109 Davis JR, Li Y, Rankin CH. Effects of Developmental Exposure to Ethanol on Caenorhabditis elegans. *Alcohol Clin Exp Res* 2008;**32**:853–67. https://doi.org/https://doi.org/10.1111/j.1530-0277.2008.00639.x.

110 Alaimo JT, Davis SJ, Song SS, Burnette CR, Grotewiel M, Shelton KL, *et al.* Ethanol metabolism and osmolarity modify behavioral responses to ethanol in C. elegans. *Alcohol Clin Exp Res* 2012;**36**:1840–50. https://doi.org/10.1111/j.1530-0277.2012.01799.x.

111 Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science (80-)* 2017. https://doi.org/10.1126/science.aam8940.

112 Packer JS, Zhu Q, Huynh C, Sivaramakrishnan P, Preston E, Dueck H, *et al.* A lineage-resolved molecular atlas of C. Elegans embryogenesis at single-cell resolution. *Science (80-)* 2019. https://doi.org/10.1126/science.aax1971.

113 Camacho J, Truong L, Kurt Z, Chen Y-W, Morselli M, Gutierrez G, *et al.* The Memory of Environmental Chemical Exposure in C. elegans Is Dependent on the Jumonji Demethylases jmjd-2 and jmjd-3/utx-1. *Cell Rep* 2018;**23**:2392–404. https://doi.org/10.1016/j.celrep.2018.04.078.

114 Camacho J, Allard P. Histone Modifications: Epigenetic Mediators of Environmental Exposure Memory. *Epigenetics Insights* 2018;**11**:2516865718803641–2516865718803641. https://doi.org/10.1177/2516865718803641.

115 Kelly WG. Transgenerational epigenetics in the germline cycle of Caenorhabditis elegans. *Epigenetics Chromatin* 2014;**7**:6. https://doi.org/10.1186/1756-8935-7-6.

116 Kishimoto S, Uno M, Okabe E, Nono M, Nishida E. Environmental stresses induce transgenerationally inheritable survival advantages via germline-to-soma communication in Caenorhabditis elegans. *Nat Commun* 2017;**8**:14031. https://doi.org/10.1038/ncomms14031.

117 Young MD, Behjati S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* 2020;**9**:giaa151. https://doi.org/10.1093/gigascience/giaa151.

118 Alvarez M, Rahmani E, Jew B, Garske KM, Miao Z, Benhammou JN, *et al.* Enhancing droplet-based single-nucleus RNA-seq resolution using the semi-supervised machine learning classifier DIEM. *Sci Rep* 2020;**10**:11019. https://doi.org/10.1038/s41598-020-67513-5.

119 Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, *et al.* Comprehensive Integration of Single-Cell Data. *Cell* 2019. https://doi.org/10.1016/j.cell.2019.05.031.

120 Waltman L, Van Eck NJ. A smart local moving algorithm for large-scale modularity-based community detection. *Eur Phys J B* 2013. https://doi.org/10.1140/epjb/e2013-40829-0.

121 Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 2015;**161**:1202–14. https://doi.org/10.1016/j.cell.2015.05.002.

122 Wilcoxon F. Individual Comparisons by Ranking Methods BT - Breakthroughs in Statistics: Methodology and Distribution. In: Kotz S, Johnson NL, editors. New York, NY: Springer New York; 1992. p. 196–202.

123 Maeda I, Kohara Y, Yamamoto M, Sugimoto A. Large-scale analysis of gene function in Caenorhabditis elegans by high-throughput RNAi. *Curr Biol* 2001;**11**:171–6. https://doi.org/10.1016/S0960-9822(01)00052-5.

124 Angeles-Albores D, N. Lee RY, Chan J, Sternberg PW. Tissue enrichment analysis for C. elegans genomics. *BMC Bioinformatics* 2016;**17**:366. https://doi.org/10.1186/s12859-016-1229-9.

125 Qiu X, Hill A, Packer J, Lin D, Ma Y-A, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. *Nat Methods* 2017;**14**:309–15. https://doi.org/10.1038/nmeth.4150.

126 Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological

themes among gene clusters. *OMICS* 2012;**16**:284–7. https://doi.org/10.1089/omi.2011.0118.

127 Davis P, Zarowiecki M, Arnaboldi V, Becerra A, Cain S, Chan J, *et al.* WormBase in 2022—data, processes, and tools for analyzing Caenorhabditis elegans. *Genetics* 2022;**220**:iyac003. https://doi.org/10.1093/genetics/iyac003.

128 Barton MK, Kimble J. fog-1, a regulatory gene required for specification of spermatogenesis in the germ line of Caenorhabditis elegans. *Genetics* 1990.

129 Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573-3587.e29. https://doi.org/https://doi.org/10.1016/j.cell.2021.04.048.

130 Burd L, Blair J, Dropps K. Prenatal alcohol exposure, blood alcohol concentrations and alcohol elimination rates for the mother, fetus and newborn. *J Perinatol* 2012;**32**:652–9. https://doi.org/10.1038/jp.2012.57.

131 Arneson D, Zhang G, Ying Z, Zhuang Y, Byun HR, Ahn IS, *et al.* Single cell molecular alterations reveal target cells and pathways of concussive brain injury. *Nat Commun* 2018. https://doi.org/10.1038/s41467-018-06222-0.

132 Liu W, Venugopal S, Majid S, Ahn IS, Diamante G, Hong J, *et al.* Single-cell RNA-seq analysis of the brainstem of mutant SOD1 mice reveals perturbed cell types and pathways of amyotrophic lateral sclerosis. *Neurobiol Dis* 2020. https://doi.org/10.1016/j.nbd.2020.104877.

133 Perez MF, Lehner B. Vitellogenins - Yolk Gene Function and Regulation in Caenorhabditis elegans. *Front Physiol* 2019;**10**:1067. https://doi.org/10.3389/fphys.2019.01067.

134 Selewa A, Dohn R, Eckart H, Lozano S, Xie B, Gauchat E, *et al.* Systematic Comparison of High-throughput Single-Cell and Single-Nucleus Transcriptomes during Cardiomyocyte

Differentiation. *Sci Rep* 2020. https://doi.org/10.1038/s41598-020-58327-6.

135 Kaufman MH, Bain IM. Influence of ethanol on chromosome segregation during the first and second meiotic divisions in the mouse egg. *J Exp Zool* 1984. https://doi.org/10.1002/jez.1402300217.

136 Hunt PA. Ethanol-induced aneuploidy in male germ cells of the mouse. *Cytogenet Genome Res* 1987. https://doi.org/10.1159/000132333.

137 Kaufman MH. Ethanol-induced chromosomal abnormalities at conception. *Nature* 1983. https://doi.org/10.1038/302258a0.

138 Oh KH, Sheoran S, Richmond JE, Kim H. Alcohol induces mitochondrial fragmentation and stress responses to maintain normal muscle function in Caenorhabditis elegans. *FASEB J* 2020;**34**:8204–16. https://doi.org/10.1096/fj.201903166R.

139 Charmpilas N, Tavernarakis N. Mitochondrial maturation drives germline stem cell differentiation in Caenorhabditis elegans. *Cell Death Differ* 2020;**27**:601–17. https://doi.org/10.1038/s41418-019-0375-9.

140 Voutev R, Killian DJ, Hyungsoo Ahn J, Hubbard EJA. Alterations in ribosome biogenesis cause specific defects in C. elegans hermaphrodite gonadogenesis. *Dev Biol* 2006;**298**:45–58. https://doi.org/https://doi.org/10.1016/j.ydbio.2006.06.011.

141 Mercer M, Jang S, Ni C, Buszczak M. The Dynamic Regulation of mRNA Translation and Ribosome Biogenesis During Germ Cell Development and Reproductive Aging . *Front Cell Dev Biol* 2021.

142 Branicky R, Desjardins D, Liu J-L, Hekimi S. Lipid transport and signaling in Caenorhabditis elegans. *Dev Dyn* 2010;**239**:1365–77. https://doi.org/https://doi.org/10.1002/dvdy.22234.

143 Watts JL, Browse J. Dietary manipulation implicates lipid signaling in the regulation of germ cell maintenance in C. elegans. *Dev Biol* 2006;**292**:381–92.

https://doi.org/10.1016/j.ydbio.2006.01.013.

144 Bacaj T, Tevlin M, Lu Y, Shaham S. Glia are essential for sensory organ function in C. elegans. *Science (80-)* 2008. https://doi.org/10.1126/science.1163074.

145 Killian DJ, Hubbard EJA. Caenorhabditis elegans germline patterning requires coordinated development of the somatic gonadal sheath and the germ line. *Dev Biol* 2005;**279**:322–35. https://doi.org/https://doi.org/10.1016/j.ydbio.2004.12.021.

146 Mendrick DL, Diehl AM, Topor LS, Dietert RR, Will Y, La Merrill MA, *et al.* Metabolic Syndrome and Associated Diseases: From the Bench to the Clinic. *Toxicol Sci* 2017;**162**:36–42. https://doi.org/10.1093/toxsci/kfx233.

147 Rochlani Y, Pothineni NV, Kovelamudi S, Mehta JL. Metabolic syndrome: pathophysiology, management, and modulation by natural compounds. *Ther Adv Cardiovasc Dis* 2017;**11**:215–25. https://doi.org/10.1177/1753944717711379.

148 Lyssiotis CA, Cantley LC. F stands for fructose and fat. *Nature* 2013;**502**:181.

149 Rodríguez-Correa E, González-Pérez I, Clavel-Pérez PI, Contreras-Vargas Y, Carvajal K. Biochemical and nutritional overview of diet-induced metabolic syndrome models in rats: what is the best choice? *Nutr Diabetes* 2020;**10**:24. https://doi.org/10.1038/s41387-020-0127-4.

150 Johnson RK, Lichtenstein AH, Anderson CAM, Carson JA, Després J-P, Hu FB, *et al.* Low-Calorie Sweetened Beverages and Cardiometabolic Health: A Science Advisory From the American Heart Association. *Circulation* 2018;**138**:e126–40. https://doi.org/10.1161/CIR.0000000000000569.

151 Malik VS, Popkin BM, Bray GA, Després J-P, Hu FB. Sugar-sweetened beverages, obesity, type 2 diabetes mellitus, and cardiovascular disease risk. *Circulation* 2010;**121**:1356–64. https://doi.org/10.1161/CIRCULATIONAHA.109.876185.

152 Nakagawa T, Hu H, Zharikov S, Tuttle KR, Short RA, Glushakova O, *et al.* A causal role

for uric acid in fructose-induced metabolic syndrome. *Am J Physiol Renal Physiol* 2006;**290**:F625-31. https://doi.org/10.1152/ajprenal.00140.2005.

153 Yang Z-H, Miyahara H, Takeo J, Katayama M. Diet high in fat and sucrose induces rapid onset of obesity-related metabolic syndrome partly through rapid response of genes involved in lipogenesis, insulin signalling and inflammation in mice. *Diabetol Metab Syndr* 2012;**4**:32. https://doi.org/10.1186/1758-5996-4-32.

154 Burchfield JG, Kebede MA, Meoli CC, Stöckli J, Whitworth PT, Wright AL, *et al.* High dietary fat and sucrose results in an extensive and time-dependent deterioration in health of multiple physiological systems in mice. *J Biol Chem* 2018;**293**:5731–45. https://doi.org/10.1074/jbc.RA117.000808.

155 Patel C, Douard V, Yu S, Gao N, Ferraris RP. Transport, metabolism, and endosomal trafficking-dependent regulation of intestinal fructose absorption. *FASEB J Off Publ Fed Am Soc Exp Biol* 2015;**29**:4046–58. https://doi.org/10.1096/fj.15-272195.

156 Jang C, Hui S, Lu W, Cowan AJ, Morscher RJ, Lee G, *et al.* The Small Intestine Converts Dietary Fructose into Glucose and Organic Acids. *Cell Metab* 2018;**27**:351-361.e3. https://doi.org/10.1016/j.cmet.2017.12.016.

157 Kim M-S, Krawczyk SA, Doridot L, Fowler AJ, Wang JX, Trauger SA, *et al.* ChREBP regulates fructose-induced glucose production independently of insulin signaling. *J Clin Invest* 2016;**126**:4372–86. https://doi.org/10.1172/JCI81993.

158 Jais A, Brüning JC. Hypothalamic inflammation in obesity and metabolic disease. *J Clin Invest* 2017;**127**:24–32. https://doi.org/10.1172/JCI88878.

159 Traber MG, Buettner GR, Bruno RS. The relationship between vitamin C status, the gut-liver axis, and metabolic syndrome. *Redox Biol* 2019;**21**:101091. https://doi.org/10.1016/j.redox.2018.101091.

160 Kahn CR, Wang G, Lee KY. Altered adipose tissue and adipocyte function in the

pathogenesis of metabolic syndrome. *J Clin Invest* 2019;**129**:3990–4000. https://doi.org/10.1172/JCI129187.

161 Dabke K, Hendrick G, Devkota S. The gut microbiome and metabolic syndrome. *J Clin Invest* 2019;**129**:4050–7. https://doi.org/10.1172/JCI129194.

162 Samuel VT, Shulman GI. Mechanisms for insulin resistance: common threads and missing links. *Cell* 2012;**148**:852–71. https://doi.org/10.1016/j.cell.2012.02.017.

163 Wakil SJ, Abu-Elheiga LA. Fatty acid metabolism: target for metabolic syndrome. *J Lipid Res* 2009;**50**:S138–43. https://doi.org/https://doi.org/10.1194/jlr.R800079-JLR200.

164 Taskinen M-R, Packard CJ, Borén J. Dietary Fructose and the Metabolic Syndrome. *Nutrients* 2019;**11**:1987. https://doi.org/10.3390/nu11091987.

165 Hannou SA, Haslam DE, McKeown NM, Herman MA. Fructose metabolism and metabolic disease. *J Clin Invest* 2018;**128**:545–55. https://doi.org/10.1172/JCI96702.

166 Meng Q, Ying Z, Noble E, Zhao Y, Agrawal R, Mikhail A, *et al.* Systems Nutrigenomics Reveals Brain Gene Networks Linking Metabolic and Brain Disorders. *EBioMedicine* 2016;**7**:157–66. https://doi.org/10.1016/j.ebiom.2016.04.008.

167 Majka SM, Miller HL, Helm KM, Acosta AS, Childs CR, Kong R, *et al.* Chapter Fifteen - Analysis and Isolation of Adipocytes by Flow Cytometry. In: Macdougald OABT-M in E, editor. *Methods Adipose Tissue Biol. Part A*, vol. 537. Academic Press; 2014. p. 281–96.

168 Kremski VC, Varani L, DeSaive C, Miller P, Nicolini C. Crypt cell isolation in the small intestine of the mouse. *J Histochem Cytochem* 1977;**25**:554–9. https://doi.org/10.1177/25.7.894003.

169 Mederacke I, Dapito DH, Affò S, Uchinami H, Schwabe RF. High-yield and high-purity isolation of hepatic stellate cells from normal and fibrotic mouse livers. *Nat Protoc* 2015;**10**:305.

170 Petukhov V, Guo J, Baryawno N, Severe N, Scadden DT, Samsonova MG, *et al.* dropEst:

pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. *Genome Biol* 2018;**19**:78. https://doi.org/10.1186/s13059-018-1449-6.

171 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21. https://doi.org/10.1093/bioinformatics/bts635.

172 Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;**36**:411.

173 Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B* 1995. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

174 Michael Dewey. metap: meta-analysis of significance values 2020.

175 Arneson D, Bhattacharya A, Shu L, Mäkinen V-P, Yang X. Mergeomics: a web server for identifying pathological pathways, networks, and key regulators via multidimensional data integration. *BMC Genomics* 2016;**17**:722. https://doi.org/10.1186/s12864-016-3057-8.

176 Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;**4**:7. https://doi.org/10.1186/s13742-015-0047-8.

177 Browaeys R, Saelens W, Saeys Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat Methods* 2020;**17**:159–62. https://doi.org/10.1038/s41592-019-0667-5.

178 UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**:D480–9. https://doi.org/10.1093/nar/gkaa1100.

179 Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, *et al.* A pathology atlas of the human cancer transcriptome. *Science (80-)* 2017;**357**:eaan2507.

https://doi.org/10.1126/science.aan2507.

180 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504. https://doi.org/10.1101/gr.1239303.

181 Helsley RN, Moreau F, Gupta MK, Radulescu A, DeBosch B, Softic S. Tissue-Specific Fructose Metabolism in Obesity and Diabetes. *Curr Diab Rep* 2020;**20**:64. https://doi.org/10.1007/s11892-020-01342-8.

182 Glendinning JI, Breinager L, Kyrillou E, Lacuna K, Rocha R, Sclafani A. Differential effects of sucrose and fructose on dietary obesity in four mouse strains. *Physiol Behav* 2010;**101**:331–43. https://doi.org/10.1016/j.physbeh.2010.06.003.

183 Montgomery MK, Fiveash CE, Braude JP, Osborne B, Brown SHJ, Mitchell TW, *et al.* Disparate metabolic response to fructose feeding between different mouse strains. *Sci Rep* 2015;**5**:18474. https://doi.org/10.1038/srep18474.

184 Kim S-J, Xiao J, Wan J, Cohen P, Yen K. Mitochondrially derived peptides as novel regulators of metabolism. *J Physiol* 2017;**595**:6613–21. https://doi.org/https://doi.org/10.1113/JP274472.

185 Malakar P, Stein I, Saragovi A, Winkler R, Stern-Ginossar N, Berger M, *et al.* Long Noncoding RNA MALAT1 Regulates Cancer Glucose Metabolism by Enhancing mTOR-Mediated Translation of TCF7L2. *Cancer Res* 2019;**79**:2480 LP – 2493. https://doi.org/10.1158/0008-5472.CAN-18-1432.

186 Renner SW, Walker LM, Forsberg LJ, Sexton JZ, Brenman JE. Carbonic anhydrase III (Car3) is not required for fatty acid synthesis and does not protect against high-fat diet induced obesity in mice. *PLoS One* 2017;**12**:e0176502–e0176502. https://doi.org/10.1371/journal.pone.0176502.

187 Lu K, Liu G, Yang L, Liu F, Gao L, Shi J, *et al.* Sustainable inflammation transforms

hepatic cells by causing oxidative stress injury and potential epithelial-mesenchymal transition. *Int J Oncol* 2016;**49**:971–80. https://doi.org/10.3892/ijo.2016.3580.

188 Brown J, Sagante A, Mayer T, Wright A, Bugescu R, Fuller PM, *et al.* Lateral Hypothalamic Area Neurotensin Neurons Are Required for Control of Orexin Neurons and Energy Balance. *Endocrinology* 2018;**159**:3158–76. https://doi.org/10.1210/en.2018-00311.

189 Timper K, Brüning JC. Hypothalamic circuits regulating appetite and energy homeostasis: pathways to obesity. *Dis Model Mech* 2017;**10**:679–89. https://doi.org/10.1242/dmm.026609.

190 Coll AP, Farooqi IS, O'Rahilly S. The hormonal control of food intake. *Cell* 2007;**129**:251–62. https://doi.org/10.1016/j.cell.2007.04.001.

191 Feng D, Tang Y, Kwon H, Zong H, Hawkins M, Kitsis RN, *et al.* High-fat diet-induced adipocyte cell death occurs through a cyclophilin D intrinsic signaling pathway independent of adipose tissue inflammation. *Diabetes* 2011;**60**:2134–43. https://doi.org/10.2337/db10-1411.

192 Ruiz-Ojeda FJ, Méndez-Gutiérrez A, Aguilera CM, Plaza-Díaz J. Extracellular Matrix Remodeling of Adipose Tissue in Obesity and Metabolic Diseases. *Int J Mol Sci* 2019;**20**:4888. https://doi.org/10.3390/ijms20194888.

193 Jang C, Wada S, Yang S, Gosis B, Zeng X, Zhang Z, *et al.* The small intestine shields the liver from fructose-induced steatosis. *Nat Metab* 2020;**2**:586–93. https://doi.org/10.1038/s42255-020-0222-9.

194 Tse EK, Salehi A, Clemenzi MN, Belsham DD. Role of the saturated fatty acid palmitate in the interconnected hypothalamic control of energy homeostasis and biological rhythms. *Am J Physiol Metab* 2018;**315**:E133–40. https://doi.org/10.1152/ajpendo.00433.2017.

195 Mihalik SJ, Steinberg SJ, Pei Z, Park J, Kim DG, Heinzer AK, *et al.* Participation of two

members of the very long-chain acyl-CoA synthetase family in bile acid synthesis and recycling. *J Biol Chem* 2002;**277**:24771–9. https://doi.org/10.1074/jbc.M203295200.

196 Priest C, Tontonoz P. Inter-organ cross-talk in metabolic syndrome. *Nat Metab* 2019;**1**:1177–88. https://doi.org/10.1038/s42255-019-0145-5.

197 Oishi Y, Manabe I. Organ System Crosstalk in Cardiometabolic Disease in the Age of Multimorbidity . *Front Cardiovasc Med* 2020:64.

198 Srikanthan K, Feyh A, Visweshwar H, Shapiro JI, Sodhi K. Systematic Review of Metabolic Syndrome Biomarkers: A Panel for Early Detection, Management, and Risk Stratification in the West Virginian Population. *Int J Med Sci* 2016;**13**:25–38. https://doi.org/10.7150/ijms.13800.

199 Poritsanos NJ, Mizuno TM, Lautatzis M-E, Vrontakis M. Chronic increase of circulating galanin levels induces obesity and marked alterations in lipid metabolism similar to metabolic syndrome. *Int J Obes* 2009;**33**:1381–9. https://doi.org/10.1038/ijo.2009.187.

200 Nakamura K, Velho G, Bouby N. Vasopressin and metabolic disorders: translation from experimental models to clinical use. *J Intern Med* 2017;**282**:298–309. https://doi.org/https://doi.org/10.1111/joim.12649.

201 Melander O. Vasopressin, from Regulator to Disease Predictor for Diabetes and Cardiometabolic Risk. *Ann Nutr Metab* 2016;**68(suppl 2**:24–8. https://doi.org/10.1159/000446201.

202 den Hartigh LJ, Wang S, Goodspeed L, Ding Y, Averill M, Subramanian S, *et al.* Deletion of serum amyloid A3 improves high fat high sucrose diet-induced adipose tissue inflammation and hyperlipidemia in female mice. *PLoS One* 2014;**9**:e108564–e108564. https://doi.org/10.1371/journal.pone.0108564.

203 Collins KH, Paul HA, Hart DA, Reimer RA, Smith IC, Rios JL, *et al.* A High-Fat High-Sucrose Diet Rapidly Alters Muscle Integrity, Inflammation and Gut Microbiota in Male

Rats. *Sci Rep* 2016;**6**:37278. https://doi.org/10.1038/srep37278.

204 Ritchie SC, Lambert SA, Arnold M, Teo SM, Lim S, Scepanovic P, *et al.* Integrative analysis of the plasma proteome and polygenic risk of cardiometabolic diseases. *Nat Metab* 2021. https://doi.org/10.1038/s42255-021-00478-5.

205 Andres-Hernando A, Jensen TJ, Kuwabara M, Orlicky DJ, Cicerchi C, Li N, *et al.* Vasopressin mediates fructose-induced metabolic syndrome by activating the V1b receptor. *JCI Insight* 2021;**6**:. https://doi.org/10.1172/jci.insight.140848.

206 Pan Y, Kong L-D. High fructose diet-induced metabolic syndrome: Pathophysiological mechanism and treatment by traditional Chinese medicine. *Pharmacol Res* 2018;**130**:438–50. https://doi.org/https://doi.org/10.1016/j.phrs.2018.02.020.

207 Bursać BN, Vasiljević AD, Nestorović NM, Veličković NA, Vojnović Milutinović DD, Matić GM, *et al.* High-fructose diet leads to visceral adiposity and hypothalamic leptin resistance in male rats--do glucocorticoids play a role? *J Nutr Biochem* 2014;**25**:446–55. https://doi.org/10.1016/j.jnutbio.2013.12.005.

208 Rask-Madsen C, Kahn CR. Tissue-specific insulin signaling, metabolic syndrome, and cardiovascular disease. *Arterioscler Thromb Vasc Biol* 2012;**32**:2052–9. https://doi.org/10.1161/ATVBAHA.111.241919.

209 Sárvári AK, Van Hauwaert EL, Markussen LK, Gammelmark E, Marcher A-B, Ebbesen MF, *et al.* Plasticity of Epididymal Adipose Tissue in Response to Diet-Induced Obesity at Single-Nucleus Resolution. *Cell Metab* 2021;**33**:437-453.e5. https://doi.org/https://doi.org/10.1016/j.cmet.2020.12.004.

210 Rajbhandari P, Arneson D, Hart SK, Ahn IS, Diamante G, Santos LC, *et al.* Single cell analysis reveals immune cell–adipocyte crosstalk regulating the transcription of thermogenic adipocytes. *Elife* 2019;**8**:e49501. https://doi.org/10.7554/eLife.49501.

211 Zhang G, Byun HR, Ying Z, Blencowe M, Zhao Y, Hong J, *et al.* Differential metabolic

and multi-tissue transcriptomic responses to fructose consumption among genetically diverse mice. *Biochim Biophys Acta Mol Basis Dis* 2020;**1866**:165569. https://doi.org/10.1016/j.bbadis.2019.165569.

212 Surwit RS, Kuhn CM, Cochrane C, McCubbin JA, Feinglos MN. Diet-induced type II diabetes in C57BL/6J mice. *Diabetes* 1988;**37**:1163–7. https://doi.org/10.2337/diab.37.9.1163.

213 Surwit RS, Feinglos MN, Rodin J, Sutherland A, Petro AE, Opara EC, *et al.* Differential effects of fat and sucrose on the development of obesity and diabetes in C57BL/6J and A/J mice. *Metabolism* 1995;**44**:645–51. https://doi.org/10.1016/0026-0495(95)90123-x.

214 Black BL, Croom J, Eisen EJ, Petro AE, Edwards CL, Surwit RS. Differential effects of fat and sucrose on body composition in A/J and C57BL/6 mice. *Metabolism* 1998;**47**:1354–9. https://doi.org/10.1016/s0026-0495(98)90304-3.

215 Li H, Zhao Q, Chang L, Wei C, Bei H, Yin Y, *et al.* LncRNA MALAT1 modulates ox-LDL induced EndMT through the Wnt/β-catenin signaling pathway. *Lipids Health Dis* 2019;**18**:62. https://doi.org/10.1186/s12944-019-1006-7.

216 Puthanveetil P, Chen S, Feng B, Gautam A, Chakrabarti S. Long non-coding RNA MALAT1 regulates hyperglycaemia induced inflammatory process in the endothelial cells. *J Cell Mol Med* 2015;**19**:1418–25. https://doi.org/https://doi.org/10.1111/jcmm.12576.

217 Yan C, Chen J, Chen N. Long noncoding RNA MALAT1 promotes hepatic steatosis and insulin resistance by increasing nuclear SREBP-1c protein stability. *Sci Rep* 2016;**6**:22640. https://doi.org/10.1038/srep22640.

218 Taveau C, Chollet C, Waeckel L, Desposito D, Bichet DG, Arthus M-F, *et al.* Vasopressin and hydration play a major role in the development of glucose intolerance and hepatic steatosis in obese rats. *Diabetologia* 2015;**58**:1081–90. https://doi.org/10.1007/s00125-015-3496-9.

219 Shen L, Tso P, Woods SC, Clegg DJ, Barber KL, Carey K, *et al.* Brain Apolipoprotein E: an Important Regulator of Food Intake in Rats. *Diabetes* 2008;**57**:2092 LP – 2098. https://doi.org/10.2337/db08-0291.

220 Liu Z, Tang Q, Wen J, Tang Y, Huang D, Huang Y, *et al.* Elevated serum complement factors 3 and 4 are strong inflammatory markers of the metabolic syndrome development: a longitudinal cohort study. *Sci Rep* 2016;**6**:18713. https://doi.org/10.1038/srep18713.

221 Marfà S, Jimenez W. Fibrinogen α-Chain as a Serum Marker of Liver Disease BT - Biomarkers in Liver Disease. In: Preedy VR, editor. Dordrecht: Springer Netherlands; 2016. p. 1–20.

222 Imperatore G, Riccardi G, Iovine C, Rivellese AA, Vaccaro O. Plasma Fibrinogen: A New Factor of the Metabolic Syndrome: A population-based study. *Diabetes Care* 1998;**21**:649 LP – 654. https://doi.org/10.2337/diacare.21.4.649.

223 Leibowitz SF, Akabayashi A, Wang J. Obesity on a high-fat diet: role of hypothalamic galanin in neurons of the anterior paraventricular nucleus projecting to the median eminence. *J Neurosci* 1998;**18**:2709–19. https://doi.org/10.1523/JNEUROSCI.18-07-02709.1998.

224 Onat A, Özhan H, Erbilen E, Albayrak S, Küçükdurmaz Z, Can G, *et al.* Independent prediction of metabolic syndrome by plasma fibrinogen in men, and predictors of elevated levels. *Int J Cardiol* 2009;**135**:211–7. https://doi.org/10.1016/j.ijcard.2008.03.054.

225 Lang R, Gundlach AL, Holmes FE, Hobson SA, Wynick D, Hökfelt T, *et al.* Physiology, Signaling, and Pharmacology of Galanin Peptides and Receptors: Three Decades of Emerging Diversity. *Pharmacol Rev* 2015;**67**:118 LP – 175. https://doi.org/10.1124/pr.112.006536.

226 Alotibi MN, Alnoury AM, Alhozali AM. Serum nesfatin-1 and galanin concentrations in the adult with metabolic syndrome. Relationships to insulin resistance and obesity. *Saudi*

Med J 2019;**40**:19–25. https://doi.org/10.15537/smj.2019.1.22825.

227 Le Foll C. Hypothalamic Fatty Acids and Ketone Bodies Sensing and Role of FAT/CD36 in the Regulation of Food Intake . *Front Physiol* 2019:1036.

228 Herman MA, Samuel VT. The Sweet Path to Metabolic Demise: Fructose and Lipid Synthesis. *Trends Endocrinol Metab* 2016;**27**:719–30. https://doi.org/10.1016/j.tem.2016.06.005.

229 Jiang L, Chen T, Sun S, Wang R, Deng J, Lyu L, *et al.* Nonbone Marrow CD34+ Cells Are Crucial for Endothelial Repair of Injured Artery. *Circ Res* 2021;**129**:e146–65. https://doi.org/10.1161/CIRCRESAHA.121.319494.

230 Scalzo RL, Foright RM, Hull SE, Knaub LA, Johnson-Murguia S, Kinanee F, *et al.* Breast Cancer Endocrine Therapy Promotes Weight Gain With Distinct Adipose Tissue Effects in Lean and Obese Female Mice. *Endocrinology* 2021;**162**:. https://doi.org/10.1210/endocr/bqab174.

231 Stancill JS, Kasmani MY, Khatun A, Cui W, Corbett JA. Single-cell RNA sequencing of mouse islets exposed to proinflammatory cytokines. *Life Sci Alliance* 2021;**4**:e202000949. https://doi.org/10.26508/lsa.202000949.

232 Armingol E, Officer A, Harismendy O, Lewis NE. Deciphering cell–cell interactions and communication from gene expression. *Nat Rev Genet* 2021;**22**:71–88. https://doi.org/10.1038/s41576-020-00292-x.

233 Gosak M, Markovič R, Dolenšek J, Slak Rupnik M, Marhl M, Stožer A, *et al.* Network science of biological systems at different scales: A review. *Phys Life Rev* 2018;**24**:118–35. https://doi.org/https://doi.org/10.1016/j.plrev.2017.11.003.

234 Yu D, Kim M, Xiao G, Hwang TH. Review of biological network data and its applications. *Genomics Inform* 2013;**11**:200–10. https://doi.org/10.5808/GI.2013.11.4.200.

235 Cheng F, Kovács IA, Barabási A-L. Network-based prediction of drug combinations. *Nat*

Commun 2019;**10**:1197. https://doi.org/10.1038/s41467-019-09186-x.

236 Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**:559. https://doi.org/10.1186/1471-2105-9-559.

237 Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS One* 2010;**5**:e12776.

238 Rowan-Carroll A, Reardon A, Leingartner K, Gagné R, Williams A, Meier MJ, *et al.* High-Throughput Transcriptomic Analysis of Human Primary Hepatocyte Spheroids Exposed to Per- and Polyfluoroalkyl Substances as a Platform for Relative Potency Characterization. *Toxicol Sci* 2021;**181**:199–214. https://doi.org/10.1093/toxsci/kfab039.

239 Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* 2019;**37**:773–82. https://doi.org/10.1038/s41587-019-0114-2.

240 Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 2017;**14**:1083–6. https://doi.org/10.1038/nmeth.4463.