UNIVERSITY OF CALIFORNIA SAN DIEGO

Exploitation of Metadata in Molecular Genomics Studies

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

in

Bioinformatics and Systems Biology

by

Jamison M McCorrison

Committee in charge:

Professor Nicholas J Schork, Chair
Professor Vikas Bansal, Co-Chair
Professor Hannah Carter
Professor Sirivash Mirarab
Professor Gene Yeo

2020

The Dissertation of Jamison M McCorrison is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California San Diego

2020

DEDICATION

To my mother and father, who have never stopped believing in what I can accomplish, and who would never hesitate to make my world a better place. I can never thank you enough.

To the one and only Dr. Katie Guffey, my life partner and the source of so much joy in my life. You are the light that made this possible.

EPIGRAPH

“One of the symptoms of an approaching nervous breakdown
is the belief that one's work is terribly important.”
-Bertrand Russell

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my Advisor (Prof. Nicholas Schork) for his years of mentorship and for taking the role as co-chair of my committee. It has been an honor to work on such a great variety of projects with such incredible minds. The way you have affected research in those you inspire, and those you have taught, will change the world forever.

I would like to acknowledge my co-chair Prof. Vikas Bansal for his support as the co-chair of my committee. My many thanks also go out to the other members of committee, and to the many members of the Bioinformatics and Systems Biology department at the University of California whose instruction and friendship has been key to advancing my research.

I would also like to acknowledge the diverse co-hort that made up each focus of my study, providing me with incredible access to large and complex data sets of tremendous interest. This thesis was made possible through the interest of open-minded researchers willing to try experimental methods crossing over from other types of studies. This work is done in honor of their academic spirit and the hope that others will follow this path of exploring experimental data to its deepest extent.

Chapter 1.2, in full, is a reprint of the material as it appears in JMIR Ment Health. 2019. Association Between Improvement in Baseline Mood and Long-Term Use of a Mindfulness and Meditation App: Observational Study. Athanas AJ, McCorrison JM, Smalley S, Price J, Grady J, Campistron J, Schork NJ. 2019. The dissertation author was a primary investigator and co-lead author of this paper.

Chapter 2.2, in full, is a reprint of the material as it appears in Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. 2016. Krishnaswami SR, Grindberg RV, Novotny M, Venepally P, Lacar B, Bhutani K, Linker SB, Pham S, Erwin JA, Miller JA, Hodge R, McCarthy JK, Kelder M, McCorrison J, Aevermann BD, Fuertes FD, Scheuermann RH, Lee J, Lein ES, Schork N, McConnell MJ, Gage FH, Lasken RS. 2016. The dissertation author was a primary investigator and author of this paper.

Chapter 2.4, in part, is currently being prepared for submission for publication of the material. Multifactorial Quality Control Analysis for Single Cell Transcriptomic Profiling. 2020. McCorrison J, Rangan A, Schork NJ. The dissertation author was the primary investigator and lead author of this paper.

Chapter 3.2, in part, is currently being prepared for submission for publication of the material. Multi-reference Genome-wide RNA-sequence Analysis of 49 Bird Species identifies Transcripts Associated with Avian Longevity. 2020. McCorrison J, Chan AP, Choi Y, Ding K, Pickering A, Pawlikowska L,Norden-Krichmar T, Evans D, Schork NJ, Miller RA. The dissertation author was the primary investigator and lead author of this paper.

Chapter 4.2, in full, is a reprint of the material as it appears in Genetic Support for Longevity-Enhancing Drug Targets: Issues, Preliminary Data, and Future Directions. 2019. McCorrison J, Girke T, Goetz LH, Miller R, Schork NJ. The dissertation author was a primary investigator and lead author of this paper.

VITA

B.Sc. 2008	Bioinformatics & Molecular Biology, Rensselaer Polytechnic Institute Minor: Electronic Arts
Ph.D. 2020	Bioinformatics & Systems Biology, University of California San Diego Dissertation Title: Exploitation of Metadata in Molecular Genomic Studies Chair: Nicholas Schork

PUBLICATIONS

Hodge RD, Miller JA, Novotny M, Kalmbach BE, Ting JT, Bakken TE, Aevermann BD, Barkan ER, Berkowitz-Cerasano ML, Cobbs C, Diez-Fuertes F, Ding SL, McCorrison J, Schork NJ, Shehata SI, Smith KA, Sunkin SM, Tran DN , Venepally P, Yanny AM, Steemers FJ, Phillips JW, Bernard A, Koch C, Lasken RS, Scheuermann RH, Lein ES. **Transcriptomic evidence that von Economo neurons are regionally specialized extratelencephalic-projecting excitatory neurons.** *Nature Communications* 11, 1172 (2020). https://doi.org/10.1038/s41467-020-14952-3

McCorrison J, Girke T, Goetz LH, Miller R, Schork NJ. **Genetic Support for Longevity-Enhancing Drug Targets: Issues, Preliminary Data, and Future Directions.** *The Journals of Gerontology*: Series A, Volume 74, Issue Supplement 1, December 2019, Pages S61–S71, doi: 10.1093/gerona/glz206.

Athanas AJ, McCorrison JM, Smalley S, Price J, Grady J, Campistron J, Schork NJ. **Association Between Improvement in Baseline Mood and Long-Term Use of a Mindfulness and Meditation App: Observational Study.** *JMIR Ment Health*. 2019 May 8;6(5):e12617. doi: 10.2196/12617.

Boldog E, Bakken TE, Hodge RD, Novotny M, Aevermann BD, Baka J, Bordé S, Close JL, Diez-Fuertes F, Ding SL, Faragó N, Kocsis AK, Kovács B, Maltzer Z, McCorrison JM, Miller JA, Molnár G, Oláh G, Ozsvár A, Rózsa M, Shehata SI, Smith K, Sunkin SM, Tran DN, Venepally P, Wall A, Puskás LG, Barzó P, Steemers FJ, Schork NJ, Scheuermann RH, Lasken RS, Lein ES & Tamás G**. Transcriptomic and morphophysiological evidence for a specialized human cortical GABAergic cell type.** Aug 27, 2018. *Nature Neuroscience*. 21, pages1185–1195 (2018). doi: 10.1038/s41593-018-0205-2.

Bakken T, Cowell L, Aevermann BD, Novotny M, Hodge R, Miller JA, Lee A, Chang I, McCorrison J, Pulendran B, Qian Y, Schork NJ, Lasken TS, Lein ES, and Scheuermann RH. **Cell type discovery and representation in the era of high-content single cell phenotyping.** Dec 21, 2017. *BMC Bioinformatics*. 2017; 18(Suppl 17): 559.s. doi: 10.1186/s12859-017-1977-1.

Wright MS, McCorrison J, Gomez AM, Beck E, Harkins D, Shankar J, Mounaud S, Segubre-Mercado E, Mojica AMR, Bacay B, Nzenze SA, Kimaro SZM, Adrian P, Klugman KP, Lucero MG, Nelson KE, Madhi S, Sutton GG, Nierman WC, Losada L. **Strain Level Streptococcus Colonization Patterns during the First Year of Life.** *Front Microbiol*. 2017 Sep 6;8:1661. doi: 10.3389/fmicb.2017.01661.

Aevermann B, McCorrison J, Venepally P, Hodge R, Bakken T, Miller J, Novotny M, Tran DN, Diez-Fuertes F, Christiansen L, Zhang F, Steemers F, Lasken RS, Lein ED, Schork N, Scheuermann RH. **Production of a preliminary qualiy control pipeline for single nuclei RNA-Seq and its application in the analysis of cell type diversity post-mortem human brain neocortex.** *Pac Symp Biocomput*. 2017;22:564-575. doi: 10.1142/9789813207813_0052.

Danaher RJ, Fouts DE, Chan AP, Choi Y, DePew J, McCorrison JM, Nelson KE, Wang C, Miller CS. **HSV-1 clinical isolates with unique in vivo and in vitro phenotypes and insight into genomic differences.** *J Neurovirol*. 2017 Apr;23(2):171-185. doi: 10.1007/s13365-016-0485-9.

Krishnaswami SR, Grindberg RV, Novotny M, Venepally P, Lacar B, Bhutani K, Linker SB, Pham S, Erwin JA, Miller JA, Hodge R, McCarthy JK, Kelder M, McCorrison J, Aevermann BD, Fuertes FD, Scheuermann RH, Lee J, Lein ES, Schork N, McConnell MJ, Gage FH, Lasken RS. **Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons.** *Nature Protocols.* 2016 Mar;11(3):499-524. doi: 10.1038/nprot.2016.015.

Sizova MV1, Chilaka A, Earl AM, Doerfert SN, Muller PA, Torralba M, McCorrison JM, Durkin AS, Nelson KE, Epstein SS. **High-quality draft genome sequences of five anaerobic oral bacteria and description of Peptoanaerobacter stomatis gen. nov., sp. nov., a new member of the family Peptostreptococcaceae.** *Stand Genomic Sci.* 2015 Jul 18;10:37. doi: 10.1186/s40793-015-0027-8.

McCorrison JM, Venepally P, Singh I, Fouts DE, Lasken RS, and Methé BA. **NeatFreq: reference-free data reduction and coverage normalization for De Novo sequence assembly.** *BMC Bioinformatics* November 2014, 15:357. doi:10.1186/s12859-014-0357-3

Ó Cuív P, Klaassens ES, Durkin AS, Harkins DM, Foster L, McCorrison J, Torralba M, Nelson KE, and Morrison M**. Draft Genome Sequence of Enterococcus faecium PC4.1, a Clade B Strain Isolated from Human Feces.** *Genome Announc* February 2014 2:1 e00022-1. doi: 10.1128/genomeA.00022-14.

Vipond J, Kane J, Hatch G, McCorrison J, Nierman W, and Losada L**. Sequence determination of Burkholderia pseudomallei NCTC 13392 colony morphology variants.** *Genome Announc.* November/December 2013 vol. 1 no. 6 e00925-13. doi: 10.1128/genomeA.00925-13.

Varga J, Losada L, Zelazny A, Kim M, McCorrison J, Brinkac L, Sampaio E, Greenberg D, Singh I, Heiner C, Ashby M, Nierman W, Holland S and Goldberg J. Draft genome sequence determination of Burkholderia ceocepacia ET12 lineage strains K56-2 and BC7.

Genome Announc. September/October 2013 vol. 1 no. 5 e00841-13. doi: 10.1128/genomeA.00841-13.

DePew J, **Zhou B**, **McCorrison JM**, **Wentworth DE**, **Purushe J**, **Koroleva G** and **Fouts DE**. **Sequencing viral genomes from a single isolated plaque.** *Virology Journal*, June 2013, 10:181 doi:10.1186/1743-422X-10-181. doi: 10.1186/1743-422X-10-181.

Shkoporov AN, Efimov BA, Khokhlova EV, Chaplin AV, Kafarskaya LI, Durkin AS, McCorrison J, Torralba M, Gillis M, Sutton G, Weibel DB, Nelson KE, Smeianov VV. **Draft Genome Sequences of Two Pairs of Human Intestinal Bifidobacterium longum subsp. longum Strains, 44B and 1-6B and 35B and 2-2B, Consecutively Isolated from Two Children after a 5-Year Time Period.** *Genome Announc.* May/June 2013 vol. 1 no. 3 e00234-13. doi: 10.1128/genomeA.00234-13.

Páraic Ó Cuív, Eline S. Klaassens, A. Scott Durkin, Derek M. Harkins, Les Foster, Jamison McCorrison, Manolito Torralba, Karen E. Nelson, and Mark Morrison. **Draft Genome Sequence of Enterococcus faecalis PC1.1, a Candidate Probiotic Strain Isolated from Human Feces.** *Genome Announc.* January/February 2013 vol. 1 no. 1 e00160-12

Huttenhower C et al. (Human Microbiome Project Consortium) **Structure, function and diversity of the healthy human microbiome**. *Nature*, June 2012; 486 (7402): 207. doi: 10.1038/nature11234.

Methé BA et al. (Human Microbiome Project Consortium) **A framework for human microbiome research**. *Nature*, June 2012; 486 (7402): 215. doi: 10.1038/nature11209.

Young ND, Debellé F, Oldroyd GE, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KF, Gouzy J, Schoof H, Van de Peer Y, Proost S, Cook DR, Meyers BC, Spannagl M, Cheung F, De Mita S, Krishnakumar V, Gundlach H, Zhou S, Mudge J, Bharti AK, Murray JD, Naoumkina MA, Rosen B, Silverstein KA, Tang H, Rombauts S, Zhao PX, Zhou P, Barbe V, Bardou P, Bechner M, Bellec A, Berger A, Bergès H, Bidwell S, Bisseling T, Choisne N, Couloux A, Denny R, Deshpande S, Dai X, Doyle JJ, Dudez AM, Farmer AD, Fouteau S, Franken C, Gibelin C, Gish J, Goldstein S, González AJ, Green PJ, Hallab A, Hartog M, Hua A, Humphray SJ, Jeong DH, Jing Y, Jöcker A, Kenton SM, Kim DJ, Klee K, Lai H, Lang C, Lin S, Macmil SL, Magdelenat G, Matthews L, McCorrison J, Monaghan EL, Mun JH, Najar FZ, Nicholson C, Noirot C, O'Bleness M, Paule CR, Poulain J, Prion F, Qin B, Qu C, Retzel EF, Riddle C, Sallet E, Samain S, Samson N, Sanders I, Saurat O, Scarpelli C, Schiex T, Segurens B, Severin AJ, Sherrier DJ, Shi R, Sims S, Singer SR, Sinharoy S, Sterck L, Viollet A, Wang BB, Wang K, Wang M, Wang X, Warfsmann J, Weissenbach J, White DD, White JD, Wiley GB, Wincker P, Xing Y, Yang L, Yao Z, Ying F, Zhai J, Zhou L, Zuber A, Dénarié J, Dixon RA, May GD, Schwartz DC, Rogers J, Quétier F, Town CD, Roe BA.. **The Medicago Genome Provides Insight into the Evolution of Rhizobial Symbioses.** *Nature*, November 2011; doi:10.1038/nature10625.

Cuív PÓ, Klaassens ES, Durkin AS, Harkins DM, Foster L, McCorrison J, Torralba M, Nelson KE, Morrison M. **Draft Genome Sequence of Bacteroides vulgatus PC510, a Strain Isolated**

from Human Feces. *Genome Announcements.* doi: 10.1128/JB.05256-11.

Cuív PÓ, Klaassens ES, Durkin AS, Harkins DM, Foster L, McCorrison J, Torralba M, Nelson KE, Morrison M. **Draft Genome Sequence of Turicibacter sanguinis PC909, Isolated from Human Feces.** *Journal of Bacteriology.* March 2011; p. 1288-1289, Vol. 193, No. 5. doi: 10.1128/JB.01328-10.

Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, Rusch DB, Mitreva M, Sodergren E, Chinwalla AT, Feldgarden M, Gevers D, Haas BJ, Madupu R, Ward DV, Birren BW, Gibbs RA, Methe B, Petrosino JF, Strausberg RL, Sutton GG, White OR, Wilson RK, Durkin S, Giglio MG, Gujja S, Howarth C, Kodira CD, Kyrpides N, Mehta T, Muzny DM, Pearson M, Pepin K, Pati A, Qin X, Yandava C, Zeng Q, Zhang L, Berlin AM, Chen L, Hepburn TA, Johnson J, McCorrison J, Miller J, Minx P, Nusbaum C, Russ C, Sykes SM, Tomlinson CM, Young S, Warren WC, Badger J, Crabtree J, Markowitz VM, Orvis J, Cree A, Ferriera S, Fulton LL, Fulton RS, Gillis M, Hemphill LD, Joshi V, Kovar C, Torralba M, Wetterstrand KA, Abouellleil A, Wollam AM, Buhay CJ, Ding Y, Dugan S, FitzGerald MG, Holder M, Hostetler J, Clifton SW, Allen-Vercoe E, Earl AM, Farmer CN, Liolios K, Surette MG, Xu Q, Pohl C, Wilczek-Boney K, Zhu D. **A Catalog of Reference Genomes from the Human Microbiome.** *Science*, March 2010; *doi:* 10.1126/science.1183605.

MANUSCRIPTS SUBMITTED FOR PUBLICATION OR IN PREPARATION

McCorrison J, Chan AP, Choi Y, Ding K, Pickering A, Pawlikowska L, Norden-Krichmar T, Evans D, Schork NJ, Miller RA. **Multi-reference Genome-wide RNA-sequence Analysis of 49 Bird Species identifies Transcripts Associated with Avian Longevity.** (Under submission to Nature Methods, 4/29/2020.)

McCorrison J, Rangan A, Schork NJ. **Multifactorial Quality Control Analysis for Single Cell Transcriptomic Profiling**. **(**Presented in draft.)

Chan AP, Choi Y McCorrison J, Schork NJ. **De novo transcriptome assembly, annotation, and comparison of 52 avian species.** (Current title, analysis not shown.)

Athanas, AJ, McCorrison J, Campistron J, Bender N, Price J, Smalley S, Schork NJ. **Driving Factors in Emotional State Transitions with Use of Mindfulness and Mediation App: Observational Study.** (Current title, analysis not shown.)

Kleinstein SE, McCorrison J, Ahmed A, Van Dyke TE, Freire M. **Transcriptomics of Diabetic and Healthy Human Neutrophils.** medRxiv 19011353; doi: https://doi.org/10.1101/19011353

FIELDS OF STUDY

Major Field: Bioinformatics and Systems Biology

Studies in Molecular Genomics, Bioinformatics, Clinical Informatics, and Laboratory Methods Development

Professors Nicholas J Schork and Vikas Bansal

ABSTRACT OF THE DISSERTATION

Exploitation of Metadata for Molecular Genomics Studies

by

Jamison M McCorrison

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2020

Professor Nicholas J Schork, Chair
Professor Vikas Bansal, Co-Chair

There is a great deal of interest in analyzing very large data sets in the biomedical sciences. This is due to the availability of high-throughput assays, such as DNA sequencing technologies and high-resolution imaging devices, advances in data storage and high-performance computing, and analytic techniques rooted in artificial intelligence and machine learning. However, many modern data sets are constructed from individual component data sets which create issues for data harmonization and scientific integration. ‘Metadata,’ i.e., data about the data within component data sets, can be used to facilitate integration and drawing inferences from the combined data sets, but requires care and is sensitive to how those data can be used. Metadata also arises in many situations in which the combination of data sets has more subtle and nuanced aspects to it, such as in analyzing species differences in evolutionary studies, where the species data are often collected independently with different techniques, making it important to know what specific protocols and techniques were used in order to organize and enable relevant comparisons and avoid batch effects, false positives, and other phenomena associated with heterogeneous data sets. I describe the application of statistical methods in four different contexts in which metadata are available. First, I describe an analysis involving the classification of emotions recorded as part of a digital therapeutic implemented in smart phone app designed to reduce stress. Meta data arise when considering the sources and settings of individual data collections. Second, I consider an analysis relating fibroblast transcriptomes to longevity across 49 avian species, where each species has a unique genome, but only a subset of species actually have available reference genomes. Third, I describe studies exploring variation in single cell gene expression patterns from studies of the human brain using expression profiles generated with different protocols and which have different quality control profiles. Fourth, I consider the analysis of genetically-mediated drug targets for longevity in which information from different sources is used to make more compelling and comprehensive statements of the candidacy of any one gene for drug development. I also consider general themes about the use of metadata in contemporary biomedical sciences and discuss areas for future research.

INTRODUCTION

Any attempt to reliably and systematically characterize and interpret signals from large biomedical data sets critically depends on an ability to control and accommodate different sources of variation that could potentially impede the detection of those signals. Identifying the sources of variation themselves may be non-trivial, since they may be associated with the technical aspects of the collection of the data, and not necessarily of focus in the more downstream analyses of those data. For example, the various steps in a protocol for preparing samples for a gene expression assay which, if done incorrectly, or in different ways across a broader set of studies, could lead to erroneous gene expression values or simply create noise that could mask any signals in the data.

Since the collection of large data sets is often done in batches, typically involves any number of data collection devices or groups of individuals, and possibly pursued with different quality control standards, there is often information about the data collection process itself, or some other aspects of the data collection process and units of observation other than the observations themselves, that could impact the interpretation of relevant data analyses. Thus, information about the data that are collected, i.e., ‘data about the data’, or ‘metadata’, should be considered in relevant analyses to ensure valid and appropriate interpretation of the results. This is particularly relevant for studies making use of high-throughput, data intensive assays like DNA or RNA sequencing, where the information about, e.g., the flow cells used in the sequencing reactions, the compartments in the flow cell that contain relevant sequencing reads for each unit of observation, the technician performing the DNA preparation, etc. may all impact the reliability of the sequencing runs. Issues arising from problems associated with phenomena like this may be compounded further if they apply to multiple units of observation collected at different times. For example, if many sequencing runs are pursued and, e.g., multiplexing is used (i.e., sequence data for a unit of observation is generated using different flow cells simultaneously or in series), then the 'demultiplexing' task of recovering all the relevant sequence data from those flow cells can exploit batch-specific meta-data that, if ignored, could lead to unreliable results for all the relevant units of observation. In this thesis I will describe a multi-focus review of sample-specific metadata, and the means by which disparate samples may be compared through multivariate statistical methods.

The focus of my doctoral thesis project is to showcase specific analytical methods and general approaches for exploiting metadata in studies involving large biomedical data sets. I consider four different settings that involve meta-data in different contexts. My analyses suggest that the use of meta-data can accommodate shortcomings in the experimental design of a study, problems caused by inconsistent biological sampling, or data harmonization issues that arise in the analysis of aggregated data sets. The first analysis setting I consider involves data collected on individuals' moods through a digital app designed to reduce stress by offering a choice of meditations based on user emotional selection. Mood data is collected both prior to and after the individual user pursues the meditation. I use analytical methods in this project to reduce the total number of moods an individual could record, which are highly variable and numerous, and match them to metadata about moods which were developed based on predefined, expert-opinion groupings of moods. These methods have their roots in microbiome studies, assays of bacterial species apparent within a targeted sample, in which the presence or lack of presence of a species identified in microbiome studies may be evaluated as a reduced variable based on previous characterizations of microbial species' genes and the different phyla and clades they are associated with. I use the mood classifications from this analysis to explore how the meditations offered by the app impact or anticipate changes in individual user moods after repeated use of the app.

The second analysis setting I consider involves testing the relationship between individual transcripts measured in 49 bird species using an RNA-sequencing protocol and the lifespan of those bird species. There are a number of thorny issues in such an analysis that involve metadata. For example, the lifespans of the bird species come from different sources and are based on different analyses, some involving different numbers of individual birds samples to represent a species to determine them. In addition, not all bird species studied have reference genomes, making the choice of which references to use for read mapping and transcript abundance calculations complicated. We must rely on phylogenetic information about the relationships of each query species to each bird species with a reference genome. Finally, in classifying the transcripts into orthologous groups for direct comparisons of transcript abundances across the species, specifications involving how to identify orthologous groups of transcripts need to be determined.

The third analysis setting I consider involves the identification of cell types within human neurons using a unique single cell RNA-sequencing protocol. The single cell data were generated over a period of time in different experiments that all exhibited different quality control profiles. Very detailed information about how each experiment was conducted, as well as metrics capturing the quality of the data generated, were recorded and used to identify and control for potential batch effects when the data were aggregated and analyzed collectively. As a byproduct of my analyses, a general methodology for reducing the likelihood that single cell RNA sequencing experiments will suffer from artifacts was developed.

The fourth and last analysis setting I consider involves evaluating the evidence that genetic variants associated with human longevity are good drug targets as well as evidence that current drugs hypothesized to impact longevity have genetic support – that is, that the gene targets of those drugs harbor variations that are associated with longevity. The different data sources associated with the information I used to address these questions are all 'metadata'-based' since they merely reflect the results of different studies (e.g., genetic association studies, pharmacologic studies exploring drug targets, etc.). I find that most of the variants associated with longevity are not necessarily good drug targets given a lack of consensus on the 'druggability' in the pharmacology community and that most drugs hypothesized to influence longevity – or shown to influence longevity in a non-human species – are not supported by genetic information.

The individual manuscripts resulting from my research illustrate that the incorporation of metadata into large-scale biomedical studies exposes underlying complexities in the interpretation of the data analyses introduced by experimental protocols, ancillary data of relevance to the primary data, and varying quality of the data that is aggregated for an analysis. However, the use of appropriate analytical methods can overcome these complexities to a high degree and help improves the interpretation of the results of the studies. The methods I developed and applied are broad, and at least some aspect of them may be applied to a wide variety of focused experimental and meta-analysis settings. As more and more data are generated by different laboratories, under different experimental conditions, or in different yet complementary contexts, the need to be sensitive to how metadata can be exploited will become commonplace.

CHAPTER 1: DIGITAL THERAPEUTIC REDUCES BASELINE DEPRESSION IN TEST POPULATION

1.1: INTRODUCTION

Low cost clinical intervention is now becoming fiscally viable through the use of digital therapeutics, particularly as the shifting tides of FDA approval mean that these tools can be considered as funded alternatives to research with traditional medications. [1] In this analyses, we leverage a smart phone application, 'Stop.Breath.Think', and the collection of user self-assessed qualitative status before and after a digital intervention. In this case, the app itself introduces complexities in our collection of qualitative status because our comparable terms, words describing human emotions, were selected via a set of guided methods (e.g. first selecting an emoji showing 5 general emotions, then providing a non-randomized list of emotion terms within the 5 general emotion classes). In addition, the application captures emotions in only 1-5 of 115 emotions before and after selection.

Prior emotional classification techniques have been shown previously. The Yale Mood Meter provides an example of a 2-dimensional self-evaluation, representing correlated emotions in a matrix format with correlated terms presented in an adjacent fashion. [2] I note prior 2-dimensional composition rendering using the Theyer 2-d emotion model, which leverages human response to musical input to gauge interaction with clinical emotional terms. [3] Recent advancements in emotional classification include a 3-dimensional representation of the emotional space that can be captured in terms of human neurochemistry. [4] Seeking to gather a credible understanding of our user's emotional classification before, and after, a digital therapeutic interaction, while only comparing very small numbers of shared terms is a complicating factor. Simple Euclidean distance would fail to accurately render the effects of rare emotion selection as representing particular states. Some of these rare selections could be due to a term being nondescriptive, or the word simply being at the end of a list.

To compare rarely occurring selections, and predict whether a user, or group of users was canonically reacting to our digital interactions in significantly correlated ways, I looked to common methods in microbiome studies. My concurrent work in this field included the assay of bacteria present, or not present, within the human oral- and nasalpharyngal microbiome during the initial 24 months of life, focusing on the canonical response of infants to the Streptococcus vaccines. Clinical insights from measurements of swings in bacterial abundance in the infant microbiome over the first 12 months of the oralpharyngal microbiome assay were discussed in my co-authored manuscript:

Wright MS, McCorrison J, Gomez AM, Beck E, Harkins D, Shankar J, Mounaud S, Segubre-Mercado E, Mojica AMR, Bacay B, Nzenze SA, Kimaro SZM, Adrian P, Klugman KP, Lucero MG, Nelson KE, Madhi S, Sutton GG, Nierman WC, Losada L. **Strain Level Streptococcus Colonization Patterns during the First Year of Life.** *Front Microbiol*. 2017 Sep 6;8:1661. doi: 10.3389/fmicb.2017.01661.

Comparison of microbial abundances leverages sparse abundance matrices by nature of the strongly differing compositions of bacteria identified in the healthy human gut. [5] In fact, a common observation within populations of gut bacteria from various world-wide locations have identified massive swings in the composition of species apparent at the genus level which can predict human diet, geography, and health. [6,7,8] These populations have even been found to form “enterotypes”, or groupings of canonical sets of bacterial flora which define some set of stable co-occurring bacterial abundance ‘types’ in major populations. [9]

My experimental longitudinal nasalpharyngal analyses of infants in 2 distinct geographical locations (Phillippines and South Africa) during the first 2 years of life (analysis not shown) automated common methods for gut microbiome processing utilizing Bray-Curtis distance for sample-sample distance comparison. [10] This is a method looking only at the distance of co-occurrence, or within the small co-selections between terms within the large sparse graph. After a series of normalization stages, I found that this method was directly applicable to co-occurrence of emotional selections from the digital therapeutic. We further validated with nonsupervised clustering with Principal Coordinates Analysis (PCoA) and Permutation around Medoids (PAM) in the Ape gut microbiome used to commonly identify these types of naturally co-occurring emotional selections. [11, PAM] In much the way that metadata (e.g. time since pneumococcal vaccination (PCV)) can be applied to predict the disappearance of bacteria (e.g. lack of Streptococcus pneumoniae), we hypothesized that we could apply our metadata (e.g. all users; emotional state co-selection) to determine how well we could predict a different clinical outcome (e.g. progression away from a baseline anxious or depressed state.)

1.2: ASSOCIATION BETWEEN IMPROVEMENT IN BASELINE MOOD AND LONG-TERM USE OF A MINDFULNESS AND MEDITATION APP: OBSERVATIONAL STUDY

See published work, co-lead-authored* with Argus Athanas, Ph.D.c., reproduced in this chapter:

Athanas AJ*, ***McCorrison JM*,*** Smalley S, Price J, Grady J, Campistron J, Schork NJ. **Association Between Improvement in Baseline Mood and Long-Term Use of a Mindfulness and Meditation App: Observational Study.** *JMIR Ment Health*. 2019 May 8;6(5):e12617. doi: 10.2196/12617.

Original Paper

Association Between Improvement in Baseline Mood and Long-Term Use of a Mindfulness and Meditation App: Observational Study

Argus J Athanas[1*], BSc; Jamison M McCorrison[2,3*], BSc; Susan Smalley[4], PhD; Jamie Price[5], JD; Jim Grady[5], BA; Julie Campistron[5], MBA; Nicholas J Schork[1,3,6,7], PhD

[1]Department of Biomedial Informatics, University of California San Diego, San Diego, CA, United States

[2]Department of Bioinformatics and Systems Biology, University of California San Diego, San Diego, CA, United States

[3]J Craig Venter Institute, San Diego, CA, United States

[4]Department of Psychiatry, University of California Los Angeles, Los Angeles, CA, United States

[5]Stop, Breathe & Think, Los Angeles, CA, United States

[6]The Translational Genomics Research Institute (TGen), Department of Quantitative Medicine, Phoenix, AZ, United States

[7]The City of Hope/Translational Genomics Research Institute IMPACT Center, Duarte, CA, United States

[*]these authors contributed equally

Corresponding Author:
Nicholas J Schork, PhD
The Translational Genomics Research Institute (TGen)
Department of Quantitative Medicine
445 N 5th St
Phoenix, AZ, 85004
United States
Phone: 1 (602) 343 8400
Email: nschork@tgen.org

Abstract

Background: The use of smartphone apps to monitor and deliver health care guidance and interventions has received considerable attention recently, particularly with regard to behavioral disorders, stress relief, negative emotional state, and poor mood in general. Unfortunately, there is little research investigating the long-term and repeated effects of apps meant to impact mood and emotional state.

Objective: We aimed to investigate the effects of both immediate point-of-intervention and long-term use (ie, at least 10 engagements) of a guided meditation and mindfulness smartphone app on users' emotional states. Data were collected from users of a mobile phone app developed by the company Stop, Breathe & Think (SBT) for achieving emotional wellness. To explore the long-term effects, we assessed changes in the users' basal emotional state before they completed an activity (eg, a guided meditation). We also assessed the immediate effects of the app on users' emotional states from preactivity to postactivity.

Methods: The SBT app collects information on the emotional state of the user before and after engagement in one or several mediation and mindfulness activities. These activities are recommended and provided by the app based on user input. We considered data on over 120,000 users of the app who collectively engaged in over 5.5 million sessions with the app during an approximate 2-year period. We focused our analysis on users who had at least 10 engagements with the app over an average of 6 months. We explored the changes in the emotional well-being of individuals with different emotional states at the time of their initial engagement with the app using mixed-effects models. In the process, we compared 2 different methods of classifying emotional states: (1) an expert-defined a priori mood classification and (2) an empirically driven cluster-based classification.

Results: We found that among long-term users of the app, there was an association between the length of use and a positive change in basal emotional state (4% positive mood increase on a 2-point scale every 10 sessions). We also found that individuals who were anxious or depressed tended to have a favorable long-term emotional transition (eg, from a sad emotional state to a happier emotional state) after using the app for an extended period (the odds ratio for achieving a positive emotional state was 3.2 and 6.2 for anxious and depressed individuals, respectively, compared with users with fewer sessions).

Conclusions: Our analyses provide evidence for an association between both immediate and long-term use of an app providing guided meditations and improvements in the emotional state.

(JMIR Ment Health 2019;6(5):e12617) doi:10.2196/12617

KEYWORDS

mental health; smartphone; emotional well-being; mindfulness

Introduction

Background

Behavioral conditions, neuropsychiatric diseases, and poor general mental health are seen as major contributors to morbidity, mortality, and lost productivity on a global scale. However, these factors are often overlooked in discussions about the current state of health care, which tend to focus on physical well-being [1]. Many studies suggest that mental health can play a large role in physical health, recovery from disease, and ultimately productivity and, therefore, should receive greater attention [2-4]. Unfortunately, there are serious questions about how mental health can be promoted and, in instances when it is called for, how relevant interventions can be prescribed and deployed efficiently in a cost-effective manner [5-7]. This is especially true given the number of people who may actually benefit from such interventions [8]. In light of this, there is enthusiasm for the development of smartphone apps that can not only monitor an individual's health—both physical and mental—but also deliver content designed to help coach them through difficult times or provide a needed intervention. In fact, many smartphone apps have been developed, or are under development, to aid in health care via, for example, image-based diagnostics, glucose monitoring for diabetes, and physical fitness promotion [9,10]. For mental health management and intervention, there is growing enthusiasm for the development of smartphone platforms that provide guidance on mindfulness and meditation as a way of relieving stress and promoting mental health and well-being. Many of the resulting platforms have been or are undergoing testing in clinical studies [11-15].

The use of mobile phone apps in combating or mediating behavioral conditions, stress, negative emotional states, and elevating mood is also consistent with directions that public health and regulatory officials are considering. In fact, evidence is mounting from clinical trials showing that smartphone apps can be effective in a variety of settings. Agencies such as the US Food and Drug Administration (FDA) have created, and in instances passed, legislation allowing the filing and approval of mobile health apps as approved health technologies on the same level as in vitro diagnostics and drugs. Pear Therapeutics was one of the first companies to have a smartphone app for addiction approved for use by the FDA in 2016 [16]. Many other commercial and academic groups are developing smartphone apps for a wide variety of conditions that go beyond the simple direct-to-consumer market by seeking regulatory approval for their use in clinical contexts [17-19]. Unfortunately, not enough time has elapsed since the introduction of smartphone-based intervention apps to provide insight into their long-term repeated effects as well as their effects in real-world settings (ie, outside of clinical trials) [20-22].

Objectives

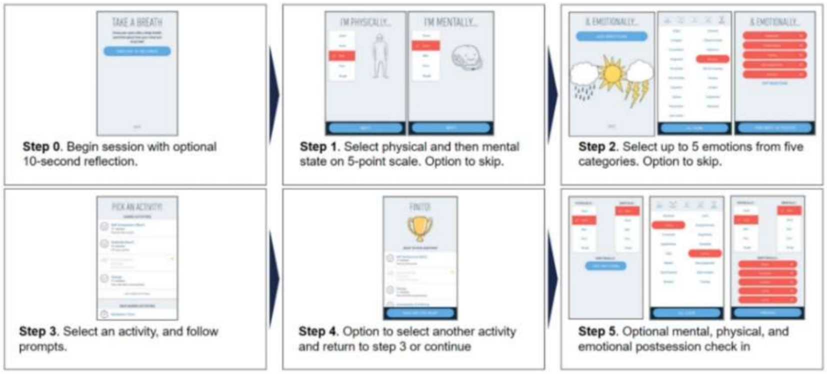
Stop, Breathe & Think (SBT) has developed a smartphone app that provides guided meditations and mindfulness activities to promote self-awareness coaching to interested users. As noted, mindfulness and meditation have been shown to improve affect and mood and promote healthy thought patterns [23,24]. The SBT app prompts users before and after they are guided through meditation and mindfulness activities to provide an emotional, mental, and physical *check-in*, thereby allowing an assessment of an individual user's emotional state and mood pre- versus postactivity in real time. As repeated uses of the app by SBT users are archived, longitudinal information on its users with regard to their long-term engagement with the app is retained. This allows further analysis of the influence of repeated engagements with the app on an individual user's basal mood over time in real-world settings. We pursued such an analysis using data from SBT users who had at least 10 engagements with the app. The SBT app allows users to choose from more than 100 unique emotions to reflect their emotional state at the time they use the app. These emotions cover a range of human emotions including anger, remorse, anxiety, calmness, and enthusiasm. Users are guided through meditations that they can choose from based on an algorithm developed by SBT. We focused our analyses on the *baseline* (or *basal*) emotional state of a user, before he or she engaged in a guided meditation or mindfulness activity and were primarily interested in the long-term and repeated use effects of the SBT app on this baseline emotional state. Essentially, we wanted to ask the question if the continued use of the app lifted the spirits of the user over time. We were particularly interested in users who tended to pick emotions associated with depression and anxiety when engaging with the app before meditating.

Methods

The Basic Stop, Breathe & Think App

The SBT app is a multiplatform (ie, iOS, Android, and Alexa) app designed to guide users through meditations and mindfulness activities to alleviate stress, anxiety, and depression and improve the sense of well-being. Upon opening the app, a user can participate in an optional 10-second reflection period. After this optional reflection period, users describe their current mood, emotional state, and physical health by choosing from a number of emotions; the SBT app then provides suggestions for specific meditation and mindfulness activities. The user can choose from among the suggested activities after being asked to endorse up to 5 different characterizations of their mood and emotional state. A user can choose not to provide any input regarding their mood, emotional state, and physical health and simply engage in an activity.

Figure 1. Stop, Breathe & Think user interface and stages of interaction with the app. Users are provided several ways in which they can record their current emotional state both pre- and postactivity. These emotional check-ins are optional, but the intuitive and simple selection process makes it easy for most users to enter at least some emotional status information.



Step 0. Begin session with optional 10-second reflection.

Step 1. Select physical and then mental state on 5-point scale. Option to skip.

Step 2. Select up to 5 emotions from five categories. Option to skip.

Step 3. Select an activity, and follow prompts.

Step 4. Option to select another activity and return to step 3 or continue

Step 5. Optional mental, physical, and emotional postsession check in

Figure 1 provides a schematic of an individual session and the corresponding points where user information is collected.

It should be understood that all information collected with the SBT app is volunteered by users as stated and defined in the SBT user licensing agreement and privacy policy. In addition, for purposes of our data analyses, all the data we obtained from SBT were anonymized and put into a Health Insurance Portability and Accountability Act (HIPAA)-compliant format such that users could not be reidentified. Functionality and delivery of the SBT app and service varies from device and platform implementation (eg, Alexa, Android, and Web browser). Therefore, to avoid batch effects, we focused on users who were exclusively on an iOS platform and started using the app after SBT provided its last major version of the app (05/01/2016). Users had to have completed at least 10 sessions or engagements with the app, with a minimum of 6 of those sessions including pre- and postactivity emotion selections. The SBT app content is in English and to avoid translation errors and alternative interpretations of the language used in the SBT app, we restricted our analyses to individuals from native English-speaking countries: the United States, United Kingdom, Canada, and Australia. An additional filter was used, restricting users' ages to between 12 and 100 years.

Emotional Check-ins Pre- and Postactivity Score

The SBT app allows the user to endorse between 1 and 5 emotional states out of a possible 115, before and after engagement in a guided meditation or mindfulness activity (or series of activities if they choose to engage in more than 1 activity during a session). This emotional *check-in* involves selecting an initial emoticon and then choosing from a list of emotions within subgroups of terms that closest characterize the user's current emotional state. These 115 emotions were chosen for the app based on internal SBT research and user requests. All emotions were classified as positive, neutral, or negative and given corresponding scores of 1, 0, and −1, respectively. All emotions and their corresponding scores are provided in Multimedia Appendix 1. As users can select up to 5 emotions, an average emotional score was calculated for both pre- and postactivity and standardized to a range from −1 (all negative emotions) to 1 (all positive emotions). Our analysis explored (1) trends in the preactivity emotional score over repeated uses of the app while accounting for the covariates as well as serial correlation between sessions and (2) trends in changes of the emotional scores before and after an activity over repeated uses of the app.

Clustering of Emotions

In addition to treating the preactivity emotion scores and changes in emotion scores pre- and postactivity as dependent variables and time, sex, and age covariates as independent variables, we also explored the patterns among the emotion endorsements to see if there was evidence for obvious clusters of emotions that could reflect the same general emotional state. We leveraged principal coordinates analysis (PCoA) and the nonsupervised clustering technique, *Partitioning Around Medoids* (PAM), for these analyses [25]. We pursued these analyses as it is arguable that some users may see a subset of the emotions as synonymous and hence only choose one among many possible choices to describe their emotional state at the time to avoid redundancy, whereas other users might see those same subsets of emotions as complementary and reflecting different aspects of their mood. In addition, other users may preferentially select emotions based on their location in the selection list or choose a set of rare emotions that are infrequently selected by other users to differentiate their emotions.

The distance between the emotions was calculated using the Bray-Curtis distance measure [26]. To determine the optimal number of nonsupervised emotion clusters in 2-dimensional PCoA component space, we selected the number of clusters with the largest silhouette score. Once we identified the optimal number of clusters, emotions were then assigned to one of the identified clusters.

An individual's emotional status was also summarized in terms of the relative *distances* (using the Euclidean distance measure) between pre- and postactivity states. The distances between an individual's emotional status and the medoid of the closest associated emotion cluster were calculated as well. Emotions were labeled with clinical categories, associating each of them with either anxiety, depression, anger, or happiness (Multimedia Appendix 1). Ultimately, using distances between emotional states and emotional clusters allowed us to build models relating the number of times users engaged with the app to gross changes in emotional states defined by the emotion clusters.

Statistical Analyses to Identify Long-Term Changes in Emotional State

To assess the effects of the continued use of the app on the preactivity emotional state, we used Linear Mixed-Effects (LME) models and Generalized Linear Models (GLMs) as implemented in the lme4 package in R [27]. These analysis techniques can accommodate serial correlations among emotions over time and also account for both fixed (eg, sex) and random effects (eg, variation in preactivity emotional state or the degree to which use of the app changed the preactivity emotional state over time). We pursued different analyses to evaluate changes in the preactivity emotional state over time, including a model that considered the effect of the emotional states possessed by individuals at their first engagement with the app. These analyses considered both the emotion scores as the dependent variables as well as the use of the emotions as defined by the cluster analysis clinical labels as dependent variables. We also tested the effect of repeated uses of the app on the change in the emotional state pre- to postactivity by treating the ratio of pre- to postemotion score as a dependent variable.

We included several covariates in our analyses and tested them for their effects on the emotional state: session index (ie, 1 as the first use and 2 as the second use—which captures the repeated use of the app), gender, age, country of origin, subscription status, and whether the user remained anonymous (ie, did not fill out information in his or her account—which may indicate a fake or disengaged user). As there is large variability in the number of completed sessions and the distribution of the number of uses of the app per individual has an extreme right skew, we applied a $\log_{10}$ transformation to the session index variable. This transformation markedly improved the normality of the session index as a variable (data not shown). LME models were fit, and the features associated with the preactivity emotional state as the dependent variable were selected using a forward stepwise selection procedure based on the Akaike Information Criteria. Similar models were fit with the pre- to postactivity emotional state ratio as the dependent variable. GLMs were fit to the data when changes in emotion categories (ie, based on clinical or cluster analysis labels) were taken as the dependent variable.

Results

Defining the Dataset

After all the duration, quality, platform, and country filters were applied, 13,393 users remained (10,082 females, 2187 males, and 1124 undeclared sex). The average age of the users was 32.3 (SD 13.5) years, with 31.7 (SD 13.3) years for females, 34.6 (SD 13.4) years for males, and 33.3 (SD 15.0) years for undeclared participants. Collectively, the users completed 569,961 sessions with the app, with 302,514 of these sessions having emotional check-in data, with an average of 42.6 sessions and 22.6 emotional check-ins per user. Multimedia Appendix 2 provides a histogram depicting the distribution of the length of time users engaged with the app. Multimedia Appendix 3 shows average period between app uses given the total length of engagement for users.

Cluster Analysis of the Emotions

The use of the silhouette scores based on the PCoA and PAM analyses suggested that there were likely 8 clusters of emotions [28]. As noted, the relative distances between pre- and postactivity emotional states and the distances between each user's emotional state and the closest associated emotion cluster were calculated. In addition, each of the 115 emotions that could be endorsed was assigned to one of the emotion clusters (see Multimedia Appendix 1). Using these cluster labels, we calculated the mean orientation of each cluster and the relative distance of each individual's emotional scores both pre- and postactivity from these means. These distances were compared with the other emotion scores we calculated and were highly correlated with them (Figure 2). Figure 3 provides a graphical depiction of the results of the clustering using the first 2 principal coordinates obtained from our analyses.

Figure 2. Average emotional score versus cluster centroid distances correlation matrix represented as a heat map. As an example for interpreting the numbers in the matrix, a −0.90 correlation between the preactivity emotion score (x-axis Average Pre Emo Score label) and positivity cluster (y-axis Dist positivity label) shows that users who score higher on the preactivity emotional score had a shorter distance of their selected emotions to the centroid of the positive emotion cluster. Note that labels with Dist reflect distance measures derived from the cluster analyses (eg, Dist Anxiety reflects the distance of a user's emotional score from the anxiety cluster mean) and Emo reflects a specified emotional cluster.

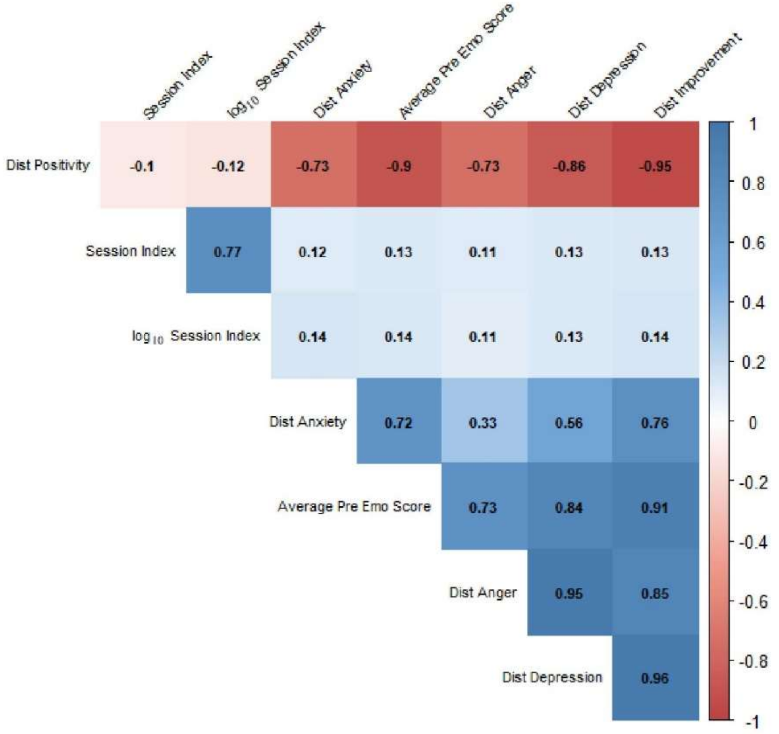
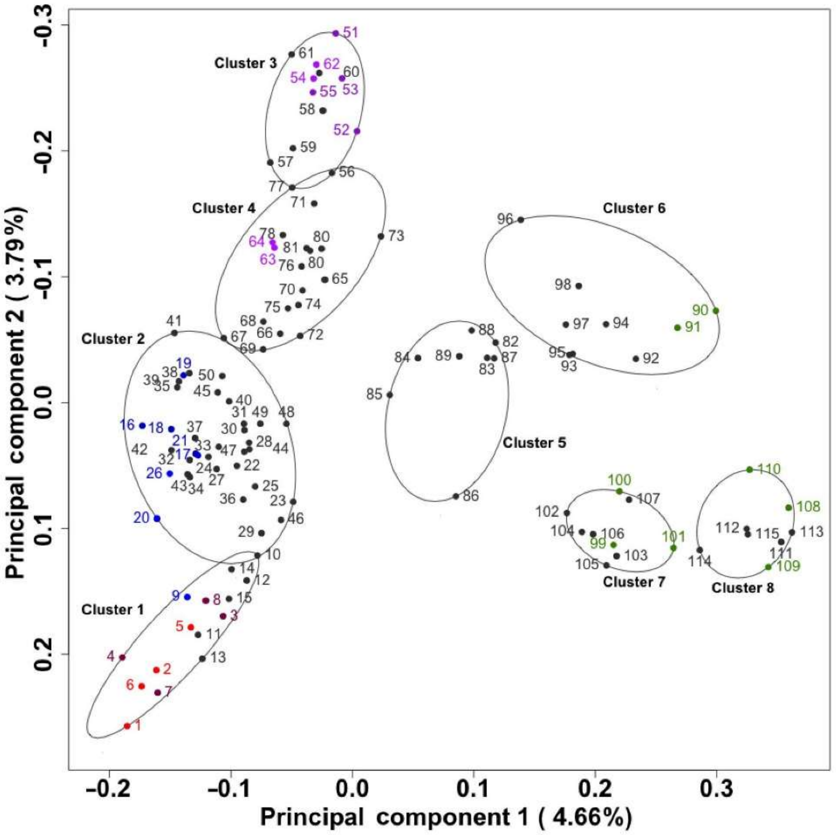


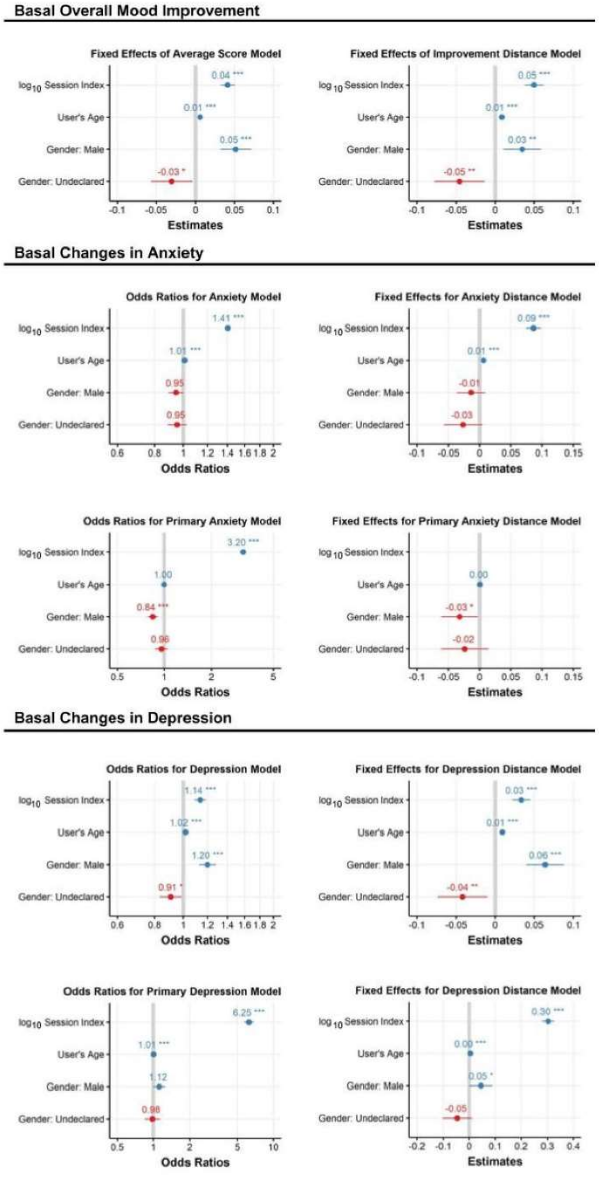
Figure 3. Emotion clustering using both pre- and postactivity emotion endorsements. The points in the plot reflect positions in the first 2 principal components defined by the Bray-Curtis distance between each pre- and postactivity emotional selection. The 8 circular clusters encompassing the emotions were defined by a permutation around medoids analysis technique, in which 8 clusters maximized the average cluster silhouette scores. Cluster boundaries are drawn on the smallest region including all underlying emotions. Emotions are labeled by clinical association such that terms clinically associated with anger are in red and pink, depression in blue, anxiety in purple, and happiness in green.



Mixed-Effects Modeling: Long-Term Use Effect on Preactivity Mood and Emotional State

Using the average preactivity emotional scores, as well as the cluster-based distance measures, as dependent variables, we fit linear-mixed models with session, as well as the important covariates, as independent variables, while accommodating serial correlation emotions. The results using the average preactivity emotional state scores suggest that a statistically significant relationship exists between the number of uses of the app (ie, session index) and the preactivity emotional state, with an elevation in mood (ie, increase in positive emotions) occurring with repeated use of the app. Adjusting for scale, users experience a 2% improvement in mood after their first session, a 4% increase after their 10th session, and a 6% increase after their 100th session. The clinical relevance of this improvement in mood needs to be investigated further. We found that males have an average 2.5% higher (improved) preactivity mood than females and that older users have a more positive mood than younger users. Additional analyses suggested that repeated use of the app resulted in specific improvements in levels of anxiety and depression. After the first 10 sessions with the app—which on average corresponded to a 63.4-day period—users were 82% more likely to report no anxious emotions and 28% more likely to report no depressive emotions. This effect was even more pronounced when we only examined users whose first emotion endorsement reflected anxiety (440%) or depression (1050%). Figure 4 depicts the effect size and statistical significance of the estimated regression coefficients for the analysis models with the average emotional score in the left panels and cluster-based emotion similarity scores in the right panels. The statistical significance (ie, *P* values) were calculated using a Wald-Z statistic approximation. Models fit using a subset of users who reported anxious or depressed emotions in their first session with the app are labeled as *primary* models. The session index is consistently associated with improvements in mood, suggesting, again, that repeated use of the app positively impacts mood.

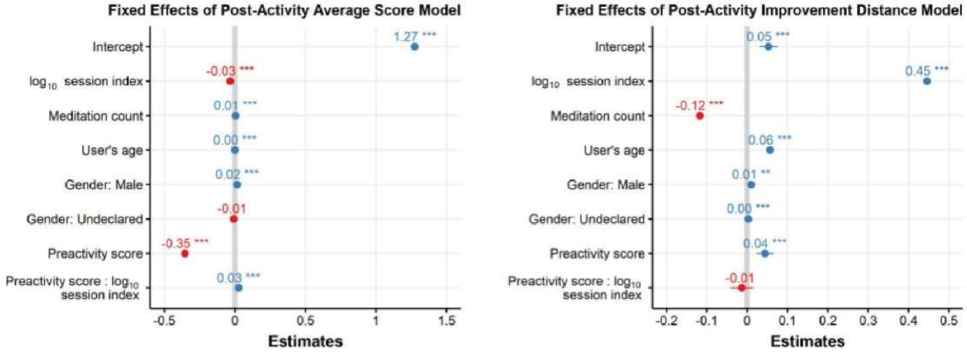
Figure 4. Linear mixed-effects regression coefficient estimates, their SEs, and P values (<.001***, <.01**, and <.05*) for models with the preactivity emotional state as the dependent variable. Analyses with the emotion scoring method as the dependent variable are on the left panels and analyses using distances from clustering as the dependent variable are on the right panels. Generalized Linear Model logit regression models were used with a binary dependent variable indicating if the emotion terms endorsed at a session reflected anxiety (middle panels) or reflected depression (bottom panels).



Mixed-Effects Modeling: Pre- Versus Postactivity Mood or Emotional State

We also fit models that considered the ratio of preactivity to postactivity emotional scores as the dependent variable. Figure 5 plots the regression coefficients resulting from the fits of these models with the ratio of average emotional score pre- to postactivity as the dependent variable (top panel) and the ratio of the distances between the emotions based on the clustering (bottom panel). The results suggest that repeated use of the app leads to increases in improvement of the mood/emotional state achieved through a meditation or mindfulness activity—or rather that the activities seem to lead to larger improvements in mood as the user has more engagements with the app.

Figure 5. Linear mixed-effects regression coefficient estimates, their SEs and P values (<.001***, <.01**, and <.05*) for models with pre- to postactivity change in the emotional state as the dependent variable. An analysis with the standardized change in emotion score pre- to postactivity as the dependent variable is reflected in the top panel, and proximity to the positive emotional clusters as the dependent variable is reflected in the bottom panel.



Discussion

Principal Findings

Our analyses show that repeated engagements with the SBT app are associated with an improvement in users' emotional states over time. In the absence of a randomized control trial, it is difficult to say with certainty that there is a direct causal relationship between the use of the SBT app and emotional state; however, given the large diverse sample size, we believe that the impact of unmeasured covariates on our results (such as external events in the users' lives) is likely to be small, although potential biases in the users of the app may exist. The effect we observed is more pronounced for users who often endorse anxiety or depression when capturing their emotional state at their initial uses. We also found that age and sex covariates are associated with the basal mood or emotional state. Ultimately, our analyses suggest the possibility that guided meditations and mindfulness activities have the potential to be effective ways of reducing anxiety, depression, and stress and ultimately elevating mood, although the ultimate clinical significance of the improvements in the emotional state that we observed needs to be explored. Our analyses did reveal other interesting phenomena. For example, although a minority in our study, males tended to have higher baseline emotional scores and responded better to the SBT app than females. The age of a user was also found to be a significant correlate of the basal emotional state, with older users generally endorsing more positive emotions.

Limitations of the Study

Our analyses are not without limitations, the first and foremost being that there is no control group and comparator app. This makes it difficult to definitively state that guided meditation and mindfulness activities are causally related or responsible for the increase in baseline mood or emotional state over time. However, given the sample size and magnitude of the effect, the significant change in emotional state after immediate and prolonged use of the app suggests that it has potential as an intervention. Another limitation is that all the information we analyzed was self-reported without any oversight by a third party. There could be users who did not follow instructions and entered erroneous emotions to expedite engagement with the meditations. Many of the individuals we did include in our analyses did not record emotions for each and every one of their sessions, resulting in many incomplete observations. Finally, a potential limitation with our analyses is that there could have been a heavy selection bias among the individuals using the app in the sense that they were motivated enough to download it and use it. Thus, this may be an indication that they could be predisposed to responding positively to the app.

Broad Emotional State Transitions

Our use of the emotion clusters and similarity scoring of emotions based on our cluster analyses of those emotions allowed us to explore how often individual users transitioned from one broad set of analogous and almost synonymous emotions to another. On the basis of these analyses, we found evidence that, in general, individual users' emotional states move from negative to positive over repeated uses of the app. We find that anxiety-prone and more depressed individuals benefit from the app more than others. These findings, as with the analyses, need to be verified in more controlled settings, such as randomized control trials, but again suggest that there is promise for the app and related apps in clinical and public health settings.

Future Directions

There are a number of questions that deserve attention beyond those that we addressed with our data. For example, the number of uses of the app may not reflect the total length of time the app was used (eg, a user could engage with the app intensely over a short period of time or stretch their use out over a longer period of time). Assessing the impact of the number of uses versus length of time on outcomes could provide a more detailed insight into the benefits of the app. In addition, it would be good to see if a companion study designed especially for adolescent populations also has a positive effect on their emotions [29]. In addition, special clinical populations may benefit from the app (eg, clinically depressed individuals and individuals with addictions). It would be of value to explore analyses that focus on the impact of large-scale social stressors (eg, school shootings, national election results, and natural disasters) on the use of the app as well as its effects on mood in the wake of stress-inducing events. Geolocation data on users could better define such exposures to social stressors should they be location specific (eg, a natural disaster in a particular state). Finally, as emphasized, it would be ideal to test the utility of the app in bona fide clinical trials to determine which aspects of the app are causally related to improvements in mood and emotional state as well as identifying subgroups of individuals that appear to respond best to particular activities.

As more and more attention is given to the delivery of health care and health maintenance strategies through devices such as smartphones, robots, and telemedicine communications, greater sensitivity to the nuanced effects of these devices should motivate studies of them that are pursued in a comprehensive manner. Such sensitivity and more elaborate studies could also lead to more efficient and sophisticated deployment of these devices and help combat the need for expensive and logistically challenging visits to health care providers.

Conflicts of Interest

Within Stop Breathe & Think, SS and NJS are advisory consultants, JP and JC are cofounders, and JG is an employee. SS, NJS, JP, JC, and JG all hold equity in Stop Breathe & Think.

Multimedia Appendix 1

Assignment and scores for Stop, Breathe & Think selectable emotions.

[XLSX File (Microsoft Excel File), 11KB - mental_v6i5e12617_app1.xlsx]

Multimedia Appendix 2

Histogram of time from first to last recorded session for users with at least ten sessions and six emotional check-ins. On average users participated in sessions with the app over a period of 180 days, with a median use of 119 days, and maximum of 702 days.

[PNG File, 14KB - mental_v6i5e12617_app2.png]

Multimedia Appendix 3

Average user period between sessions. On average a user will interact with the app at least once every 6.34 days, and the majority of users complete at least two sessions per month.

[PNG File, 13KB - mental_v6i5e12617_app3.png]

References

1. GBD 2016 DALYsHALE Collaborators. Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 2017 Sep 16;390(10100):1260-1344 [FREE Full text] [doi: 10.1016/S0140-6736(17)32130-X] [Medline: 28919118]
2. Asherson P, Kosmas C, Patel C, Doll H, Joseph A. Health-related quality of life and work productivity of adults with ADHD: a UK web-based survey. Eur Psychiatry 2017 Apr 14;41:S352. [doi: 10.1016/j.eurpsy.2017.02.332]
3. Aitken LM, Burmeister E, McKinley S, Alison J, King M, Leslie G, et al. Physical recovery in intensive care unit survivors: a cohort analysis. Am J Crit Care 2014 Dec 31;24(1):33-40. [doi: 10.4037/ajcc2015870]
4. Goetzel RZ, Hawkins K, Ozminkowski RJ, Wang S. The health and productivity cost burden of the "top 10" physical and mental health conditions affecting six large U.S. employers in 1999. J Occup Environ Med 2003 Jan;45(1):5-14. [doi: 10.1097/00043764-200301000-00007] [Medline: 12553174]
5. Musiat P, Tarrier N. Collateral outcomes in e-mental health: a systematic review of the evidence for added benefits of computerized cognitive behavior therapy interventions for mental health. Psychol Med 2014 Nov;44(15):3137-3150. [doi: 10.1017/S0033291714000245] [Medline: 25065947]
6. Insel TR. Assessing the economic costs of serious mental illness. Am J Psychiatry 2008 Jun;165(6):663-665. [doi: 10.1176/appi.ajp.2008.08030366] [Medline: 18519528]
7. Smit F, Cuijpers P, Oostenbrink J, Batelaan N, de Graaf R, Beekman A. Costs of nine common mental disorders: implications for curative and preventive psychiatry. J Ment Health Policy Econ 2006 Dec;9(4):193-200. [Medline: 17200596]
8. Remes O, Brayne C, van der Linde R, Lafortune L. A systematic review of reviews on the prevalence of anxiety disorders in adult populations. Brain Behav 2016 Dec;6(7):e00497 [FREE Full text] [doi: 10.1002/brb3.497] [Medline: 27458547]

9. Iacoviello BM, Steinerman JR, Klein DB, Silver TL, Berger AG, Luo SX, et al. Clickotine, a personalized smartphone app for smoking cessation: initial evaluation. JMIR Mhealth Uhealth 2017 Apr 25;5(4):e56 [FREE Full text] [doi: 10.2196/mhealth.7226] [Medline: 28442453]
10. Iacoviello BM, Murrough JW, Hoch MM, Huryk KM, Collins KA, Cutter GR, et al. A randomized, controlled pilot trial of the Emotional Faces Memory Task: a digital therapeutic for depression. NPJ Digit Med 2018;1 [FREE Full text] [doi: 10.1038/s41746-018-0025-5] [Medline: 30854473]
11. Champion L, Economides M, Chandler C. The efficacy of a brief app-based mindfulness intervention on psychosocial outcomes in healthy adults: a pilot randomised controlled trial. PLoS One 2018;13(12):e0209482 [FREE Full text] [doi: 10.1371/journal.pone.0209482] [Medline: 30596696]
12. Lindsay EK, Chin B, Greco CM, Young S, Brown KW, Wright AG, et al. How mindfulness training promotes positive emotions: dismantling acceptance skills training in two randomized controlled trials. J Pers Soc Psychol 2018 Dec;115(6):944-973. [doi: 10.1037/pspa0000134] [Medline: 30550321]
13. Economides M, Martman J, Bell MJ, Sanderson B. Improvements in stress, affect, and irritability following brief use of a mindfulness-based smartphone app: a randomized controlled trial. Mindfulness (N Y) 2018;9(5):1584-1593 [FREE Full text] [doi: 10.1007/s12671-018-0905-4] [Medline: 30294390]
14. Bostock S, Crosswell AD, Prather AA, Steptoe A. Mindfulness on-the-go: effects of a mindfulness meditation app on work stress and well-being. J Occup Health Psychol 2019 Feb;24(1):127-138. [doi: 10.1037/ocp0000118] [Medline: 29723001]
15. Coulon SM, Monroe CM, West DS. A systematic, multi-domain review of mobile smartphone apps for evidence-based stress management. Am J Prev Med 2016 Jul;51(1):95-105. [doi: 10.1016/j.amepre.2016.01.026] [Medline: 26993534]
16. Waltz E. Pear approval signals FDA readiness for digital treatments. Nat Biotechnol 2018 Dec 6;36(6):481-482. [doi: 10.1038/nbt0618-481] [Medline: 29874220]
17. Kvedar JC, Fogel AL, Elenko E, Zohar D. Digital medicine's march on chronic disease. Nat Biotechnol 2016 Mar 10;34(3):239-246. [doi: 10.1038/nbt.3495] [Medline: 26963544]
18. Harrison V, Proudfoot J, Wee PP, Parker G, Pavlovic DH, Manicavasagar V. Mobile mental health: review of the emerging field and proof of concept study. J Ment Health 2011 Dec;20(6):509-524. [doi: 10.3109/09638237.2011.608746] [Medline: 21988230]
19. Olff M. Mobile mental health: a challenging research agenda. Eur J Psychotraumatol 2015;6:27882 [FREE Full text] [doi: 10.3402/ejpt.v6.27882] [Medline: 25994025]
20. Blonde L, Khunti K, Harris SB, Meizinger C, Skolnik NS. Interpretation and impact of real-world clinical data for the practicing clinician. Adv Ther 2018 Nov;35(11):1763-1774 [FREE Full text] [doi: 10.1007/s12325-018-0805-y] [Medline: 30357570]
21. Berger ML, Sox H, Willke RJ, Brixner DL, Eichler H, Goettsch W, et al. Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. Value Health 2017 Dec;20(8):1003-1008 [FREE Full text] [doi: 10.1016/j.jval.2017.08.3019] [Medline: 28964430]
22. Monti S, Grosso V, Todoerti M, Caporali R. Randomized controlled trials and real-world data: differences and similarities to untangle literature data. Rheumatology (Oxford) 2018 Dec 1;57(57 Suppl 7):vii54-vii58. [doi: 10.1093/rheumatology/key109] [Medline: 30289534]
23. Baer R. Mindfulness training as a clinical intervention: a conceptual and empirical review. Clin Psychol 2003;10(2):125-143. [doi: 10.1093/clipsy.bpg015]
24. Zoogman S, Goldberg S, Hoyt W, Miller L. Mindful Well-Being. 2015. Mindfulness interventions with youth: A meta-analysis URL: http://www.mindful-well-being.com/wp-content/uploads/2014/07/Zoogman-et-al-2014-meta-anlysis.pdf [accessed 2019-04-11] [WebCite Cache ID 77YDPwxOO]
25. Zoogman S, Goldberg SB, Hoyt WT, Miller L. Institut de recherche pour le développement. 2014 Jan 15. Mindfulness Interventions with Youth: A Meta-Analysis URL: http://hal.ird.fr/ird-01887318/document [accessed 2019-04-11] [WebCite Cache ID 77YE7TtR6]
26. Dixon, Philip. VEGAN, a package of R functions for community ecology. J Veg Sci 2003;14(6):927-930. [doi: 10.1111/j.1654-1103.2003.tb02228.x]
27. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Soft 2015;67(1). [doi: 10.18637/jss.v067.i01]
28. Reynolds AP, Richards G, de la Iglesia B, Rayward-Smith VJ. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. J Math Model Algor 2006 Mar 3;5(4):475-504. [doi: 10.1007/s10852-005-9022-1]
29. Monto M, McRee N, Deryck F. Nonsuicidal self-injury among a representative sample of US adolescents, 2015. Am J Public Health 2018 Aug;108(8):1042-1048. [doi: 10.2105/AJPH.2018.304470] [Medline: 29927642]

Abbreviations

FDA: Food and Drug Administration
GLM: Generalized Linear Model

LME: Linear Mixed-Effects
PAM: Partitioning Around Medoids
PCoA: Principal Coordinates Analysis
SBT: Stop, Breathe & Think

Edited by J Torous; submitted 26.10.18; peer-reviewed by R Vanderkruik, U Subnis; comments to author 11.12.18; revised version received 02.02.19; accepted 24.03.19; published 02.05.19

Please cite as:
Athanas AJ, McCorrison JM, Smalley S, Price J, Grady J, Campistron J, Schork NJ
Association Between Improvement in Baseline Mood and Long-Term Use of a Mindfulness and Meditation App: Observational Study
JMIR Ment Health 2019;6(5):e12617
URL: https://mental.jmir.org/2019/5/e12617/
doi:10.2196/12617
PMID:

Chapter 1.2, in full, is a reprint of the material as it appears in JMIR Ment Health. 2019. Association Between Improvement in Baseline Mood and Long-Term Use of a Mindfulness and Meditation App: Observational Study. Athanas AJ, McCorrison JM, Smalley S, Price J, Grady J, Campistron J, Schork NJ. 2019. The dissertation author was a primary investigator and co-lead author of this paper.

1.3: FUTURE WORK

Interpreting our data using the Bray-Curtis distance matrix and using our principal components to represent variance in our analysis space is complicated. Understanding users in the context of the recurrent pre- and post-intervention ‘location’ in the ‘emotional vacuum’ defined by previous observations, we next seek to predict user outcome trajectories given their pre-mediation selection. Correlating emotions with other clinical outcomes or longitudinal events is the subject of my co-authored paper (analysis not shown):

Athanas, AJ, McCorrison J, Campistron J, Bender N, Price J, Smalley S, Schork NJ. **Driving Factors in Emotional State Transitions with Use of Mindfulness and Mediation App: Observational Study.** (Current title, analysis not shown.)

This work is applicable to many fields and the methods discussed can take on many alternate delivery mechanisms and applications. Some social media applications are commonly utilized to track these types of co-selection terms, and associated metadata with interventions (e.g. ads) to produce an intended outcome (e.g. a user clicking on those ads). The same tools are widely applicable as larger data sets become available in microbiome assays, studies of emotional classification, studies of digital therapeutic response, and other analyses requiring comparison of many samples with a sparse, but widely co-occurring, sample co-selection distance matrix.

CHAPTER 2: ISOLATING AND CHARACTERIZING NOVEL NEURONAL CELL TYPES IN THE HUMAN FRONTAL CORTEX

2.1. INTRODUCTION

The ability to sequence individual cells is contributing to a revolution in the understanding of bacterial cell types that cannot be cultured and therefore previously could not be amplified to sufficient protein abundances for sequencing. [1] Prior advances in single cell research have primarily focused on the advancement of the understanding of the human microbiome by, for example, allowing individual isolation of non-culture-amplifying species composing greater than 75% of the gut composition. [2] Because single cell amplification is commonly used when low biomass environmental samples are collected, exponential variation in coverage is inherent to single cell amplification protocols. [3] The severity and contribution of this bias to resulting informatics analysis is understood but normalization methodologies have been limited to the context of reference-free bacterial assembly. [4,5,6]

Previous research that worked from the roadmap of classically studied single cell bacterial models was often slow paced and time-consuming because of its dependence on low-throughput lab methodologies and multiple rounds of validation with existing bacterial models. In recent years, reductions of MDA reaction volume improved the specificity of template amplification and reduced bias [7,8]. More recent methods have been refined for high throughput handling of individual cells including “sensitive, highly-multiplexed single cell RNA-seq” with SmartSeq2 [9]. The resulting landscape allows for the rapid amplification and sequencing of an incredible number of poorly studied bacteria as well as individual cells isolated from eukaryotes. However, there is not currently a well-documented understanding of the bias contributions to transcriptomic analysis in either space. This project approaches the ability to use single cell amplification methods to both detect expression of transcripts within tissue types in the human brain and quantify the bias contributions inherent to individual sample preparations, lab methodologies, and informatics protocols.

I collaborated with a team of researchers to collect individual cells across specific geographic regions of the human brain. We used these cells to better understand of variation across known cell types, predict or define new cell types, and define their canonical expression patterns. This work is a collaborative effort with the Allen Institute of Brain Science (AIBS) and is an extension of their work transcriptomic profiling of cell specificity within pools of whole cells isolated from each roughly-defined geographic region of the human brain (Figure 8D). [10] The AIBS group had already placed a large effort in understanding diversity between tissue types in mice as part of the Mouse Genome Atlas which was leveraged to validate candidate cell types using pre-defined mouse neural marker genes. [11]

Neurons are highly interconnected, and considerable damage must be done to their extensions to separate them by physical means such as laser-capture micro dissection. Likewise, dispersion of cells by proteolytic degradation of surface proteins places the cells under stress and substantially alters gene expression. The isolation of small quantities of RNA from within the nucleus frequently results in low yield or biases within amplification and interpretation of downstream sequencing. It has previously been shown in mice that single cell RNA-seq (scRNA-seq) analysis is a successful tool for elucidation of sub-types within cells in the mouse cortex [12,13] despite these complexities.

From two post-mortem human donors, collaborators have collected samples from different layers of the cerebral cortex. One such layer-specific extraction was performed on tissue related to development in dementia in apes and humans [14] and another within the temporal cortex, associated with visual and audible comprehension (Tamas, Univ. Szegad, Hungary). [15,16] This latter layer was selected in the hopes of collecting rare GABA-ergic neurons present in the tissue. Analysis was separated into 1) three batches of validation studies, totaling 720 samples, during which lab protocols were refined for publication and 2) the preparation of approximately 2,000 single nuclei single neuron samples using the idealized production pipeline. Preliminary quality control evaluation during the "validation studies" revealed a lack of correlation between standard RNA quality metrics for whole cell projects. The RNA Identity Number (RIN), an evaluation of the ratio of 28s and 18s peaks showed an unexpected lack of correlation to the successful extraction of non-fragmented (full length) cDNA after amplification. Likewise, correlations between quantitative PCR (qPCR) for the expression of common housekeeping genes (ActB, GAPDH) and Picogreen values (denoting successful generation of dsDNA) appeared to be batch-specific.

The use of both wet lab and dry lab metrics for the production of a QC classification model using random forest machine learning appears to be an effective objective strategy for the quality control of low input, highly-amplified samples, providing further insights into the data features that are most useful for identifying quality outliers.

Aevermann B, McCorrison J, Venepally P, Hodge R, Bakken T, Miller J, Novotny M, Tran DN, Diezfuertes F, Christiansen L, Zhang F, Steemers F, Lasken RS, Lein ED, Schork N, Scheuermann RH. **Production of a preliminary quality control pipeline for single nuclei RNA-Seq and its application in the analysis of cell type diversity in the post-mortem human brain neocortex.** *Pac Symp Biocomput*. 2017;22:564-575. doi: 10.1142/9789813207813_0052.

We found that there appear to be at least two classes of failed samples, and that the metrics useful in identifying each are different. Failed samples with a second peak in the percentage of GC content plot apparently due to reads derived from the ERCC spike-in control are identified by metrics like the percentage of exact duplicates and percentage of unique reads, presumably due to the fact that a relatively small number of transcripts derived from the ERCC control are responsible for a significant proportion of the total reads obtained from those samples.

The successful amplification of unsheared ssRNA, as represented by bioAnalyzer traces, was best represented by the 3' bias within alignments to all highly detected transcripts (Figure 6). Partially degraded RNA (from freezing, RNAse degradation, etc.) resulted in deeper sequencing coverage for the 3' end of transcripts only when degraded products contain the polyA tail required for amplification. (Figure 6) 3' bias has now been adopted as a simple pass/fail metric as an easy informatics control to quantify lab-based amplification failures. Validation of the protocol and initial summaries of visual interpretations of the results were published in Nature. [17,18] Our standard laboratory workflow for single nuclei RNA-seq, constructed in tandem with a laboratory team to present optimized performance in our quality analyses, is summarized in the following chapter.

2.2. . USING SINGLE NUCLEI FOR RNA-SEQ TO CAPTURE THE TRANSCRIPTOME OF POSTMORTEM NEURONS

See published work, reproduced in this chapter:

Krishnaswami SR, Grindberg RV, Novotny M, Venepally P, Lacar B, Bhutani K, Linker SB, Pham S, Erwin JA, Miller JA, Hodge R, McCarthy JK, Kelder M, ***McCorrison J***, Aevermann BD, Fuertes FD, Scheuermann RH, Lee J, Lein ES, Schork N, McConnell MJ, Gage FH, Lasken RS. **Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons.** *Nature Protocols.* 2016 Mar;11(3):499-524. doi: 10.1038/nprot.2016.015.

Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons

Suguna Rani Krishnaswami[1,9], Rashel V Grindberg[2,9], Mark Novotny[1], Pratap Venepally[3], Benjamin Lacar[4], Kunal Bhutani[1], Sara B Linker[4], Son Pham[4], Jennifer A Erwin[4], Jeremy A Miller[5], Rebecca Hodge[5], James K McCarthy[1], Martin Kelder[4], Jamison McCorrison[1], Brian D Aevermann[1], Francisco Diez Fuertes[1,6], Richard H Scheuermann[1], Jun Lee[7], Ed S Lein[5], Nicholas Schork[1], Michael J McConnell[8], Fred H Gage[4] & Roger S Lasken[1]

[1]J. Craig Venter Institute, La Jolla, California, USA. [2]Institute of Microbiology, ETH Zurich, Zurich, Switzerland. [3]J. Craig Venter Institute, Rockville, Maryland, USA. [4]Salk Institute for Biological Studies, La Jolla, California, USA. [5]Allen Institute for Brain Science, Seattle, Washington, USA. [6]Centro Nacional de Microbiología, Instituto de Salud Carlos III, Madrid, Spain. [7]LeGene Biosciences, San Diego, California, USA. [8]Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, Virginia, USA. [9]These authors contributed equally to this work. Correspondence should be addressed to R.S.L. (rlasken@jcvi.org) or R.V.G. (grindbergr@ethz.ch).

Published online 18 February 2016; doi:10.1038/nprot.2016.015

A protocol is described for sequencing the transcriptome of a cell nucleus. Nuclei are isolated from specimens and sorted by FACS, cDNA libraries are constructed and RNA-seq is performed, followed by data analysis. Some steps follow published methods (Smart-seq2 for cDNA synthesis and Nextera XT barcoded library preparation) and are not described in detail here. Previous single-cell approaches for RNA-seq from tissues include cell dissociation using protease treatment at 30 °C, which is known to alter the transcriptome. We isolate nuclei at 4 °C from tissue homogenates, which cause minimal damage. Nuclear transcriptomes can be obtained from postmortem human brain tissue stored at −80 °C, making brain archives accessible for RNA-seq from individual neurons. The method also allows investigation of biological features unique to nuclei, such as enrichment of certain transcripts and precursors of some noncoding RNAs. By following this procedure, it takes about 4 d to construct cDNA libraries that are ready for sequencing.

INTRODUCTION

Methods for carrying out RNA-seq from single cells[1–5] are dramatically affecting many research fields, including the study of cellular development, the identification of cell types and states, the exploration of human disease and the development of stem cell technologies. The gene expression repertoires of individual cell types are revealed as opposed to the averaging of all transcriptomes obtained from bulk tissue. However, cells of the central nervous system (CNS) have been under-studied, partly because of the difficulty of isolating intact whole cells. Neurons are highly interconnected, and considerable damage must be done to their extensions to separate them by physical means such as laser-capture microdissection. An intracellular tagging method called TIVA uses RNA extracted from single cells, but it is limited to small numbers of cells[6]. Extraction of cytoplasmic content by a glass microcapillary[7,8] or by laser-capture microdissection[9] is of low throughput. An alternative, high-throughput approach is to disperse the cells and to isolate them by FACS. This approach has been recently reported for neurons isolated from brain tissue[10,11]. However, dispersion of cells by proteolytic degradation of surface proteins places the cells under stress, which substantially alters gene expression[12].

We have developed an alternative approach that takes advantage of the low levels of mRNA contained in the nucleus of the cell[13], and it avoids harsh treatment that would perturb gene expression. Through extensive comparisons of nuclear and cellular transcriptomes, we demonstrated that nuclei can substitute for whole cells in most RNA-seq applications[13]. For the majority of genes, nuclei yielded expression signatures that were very similar to those obtained from whole-cell controls. Furthermore, some transcripts that are known to be enriched in the nucleus on the basis of earlier bulk RNA studies[14–17] were also confirmed to be enriched in single nuclei, adding confidence to the accuracy of data. Here we provide a detailed protocol based on our previously published method[13] for RNA-seq using nuclei from brain tissue or cells, which can be used to obtain global transcriptomes from neurons, glia and other cell types. Although it is described here for brain tissue, it should also be applicable to any tissue type in which dissociation of whole cells would require harsh treatments and the consequent alteration of the transcriptome.

Development of the protocol

Many methods are available for the isolation of nuclei; however, the literature spans decades, and it typically lacks detailed information on the quality of RNA obtained, focusing instead on accessing intact DNA for chromatin preparation or for assaying the nuclear protein content[18]. We therefore developed an approach to meet the need for isolating individual nuclei for use in RNA analysis, which we have successfully applied to cultured neuroprogenitor cells and fresh mouse brain tissue[13]. The protocol detailed here includes two main modifications to the published method. First, we now consider cleanup by sucrose-iodixanol gradient centrifugation[18] to be necessary only if cell debris is likely to interfere with immunostaining; it is therefore included in the PROCEDURE as an optional step with the default approach to subject the filtered crude homogenate directly to FACS[19]. Second, we now use Smart-seq2 for cDNA synthesis[3] (instead of the method by Tang *et al.*[5]), which is reported to improve synthesis of full-length cDNA via a template switching mechanism for synthesis of the second-strand cDNA[4].

Overview of the procedure

Our experimental workflow (**Fig. 1**) begins with tissue homogenization in the presence of a detergent to lyse the cell membrane, determination of the number of nuclei obtained

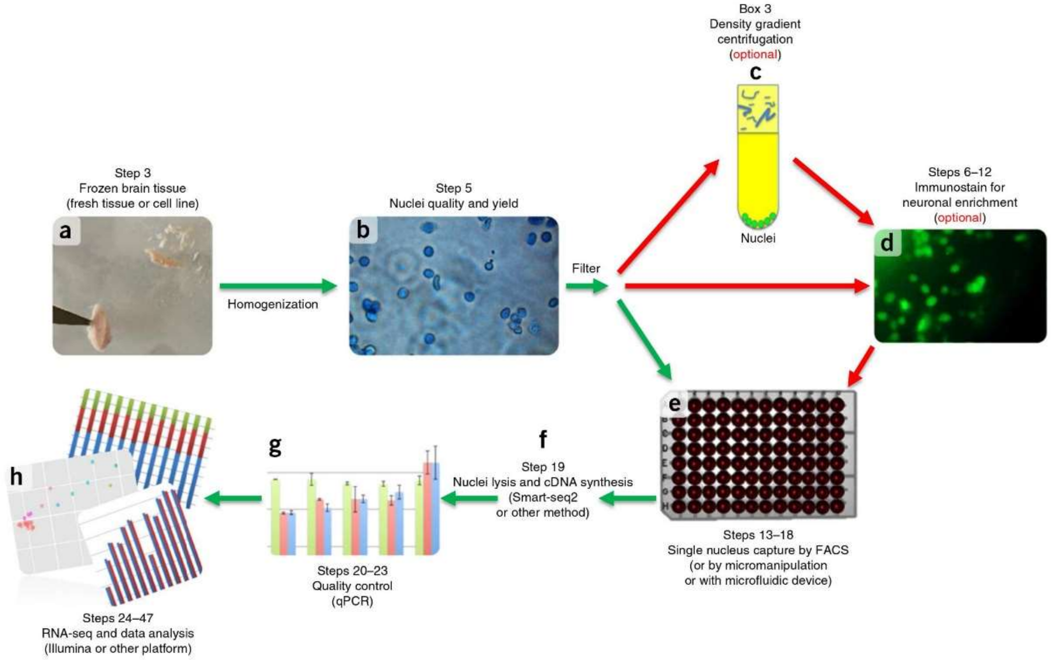


Figure 1 | Single nuclei isolation experimental workflow. Dounce homogenization in lysis buffer is used to disrupt cellular membranes for fresh or frozen tissue (**a**). Nuclei quality and yield is determined by hemocytometer count (**b**). (**c–e**) Nuclei and cellular debris are filtered for optional purification and immunostaining steps (density gradient centrifugation (**c**) or staining for neuronal enrichment (**d**)), or for FACS sorting (**e**). (**f**,**g**) Subsequently, lysis of the nuclei and cDNA synthesis is carried out using either published methods[3] or commercial kits (SMARTer, Clontech) (**f**), and it is quality-controlled for size distribution using a Bioanalyzer (Agilent) and the presence of several transcripts by qPCR (**g**). (**h**) Sequencing and data analysis confirm single nucleus transcriptome capture. Step numbers indicate the corresponding step numbers in the PROCEDURE section. Graphs in **g** and **h** are for illustrative purposes only.

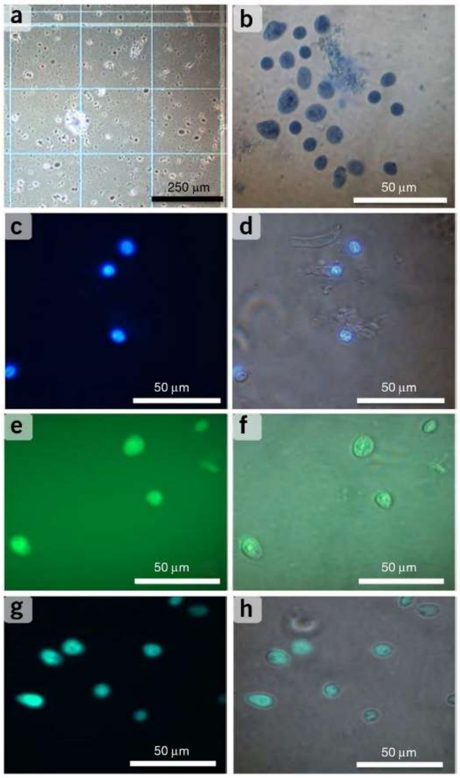
with a hemocytometer (Steps 1–5) and FACS (Steps 13–18). The nuclei are lysed and cDNA is synthesized, amplified (Step 19) and tested in quantitative PCR (qPCR) quality control assays to indicate successful capture of the transcriptome by assaying several housekeeping and tissue-specific genes (Steps 20–23). Samples that pass quality control assays are used in downstream sequencing library preparation (Step 24) and RNA-seq (Step 25). A series of bioinformatic analyses then follow to assess sequence quality (Steps 26–28), mapping and expression (Steps 29–34), variation (Steps 35–38), gene coverage (Steps 39–41), intron and exon coverage (Steps 42–46), and the classification of cell types (Step 47). The main stages of the protocol are discussed in more detail below.

Tissue handling and homogenization to release nuclei. In general, the initial quality and methods used to handle postmortem brain specimens will affect the quality of the RNA-seq data. RIN scores (RNA integrity number[20] ranging from 1 to 10) for specimens are often provided by the brain banks; however, we also determined RIN scores in our laboratory and sometimes found differences, possibly because the specimens had been stored for long periods of time and then taken through a thawing step in our laboratory. The RIN scores that we determined were used to evaluate the starting quality of the frozen specimens. We selected specimens with a RIN value of ≥7.

We chose Dounce homogenization to handle the very small tissue dissections often required to investigate various brain regions. Dounce homogenization[21] with a nonionic surfactant, Triton X-100, is used to lyse the cell membrane and release nuclei. The detergent can also permeabilize and lyse the nuclear envelope, but only under harsh conditions for an extended period of time[22]. Sufficient Triton X-100 is included in the homogenization step to facilitate the release of nuclei, allowing them to remain intact, and to permit optimal antibody[18] staining and isolation by FACS without forming aggregates. Hoechst stain is added to the homogenization lysis buffer to identify nuclei during FACS.

Before proceeding with FACS, the overall quality of the nuclei and number obtained should be determined using fluorescence photomicrography (after Dounce homogenization and again after the sucrose-iodixanol gradient centrifugation if that optional step is performed). High-resolution electron microscopy has been used for assessing the integrity of the nuclei and purity of the preparation, but it will be impractical for most laboratories[19]. Light microscopy can be used to assess whether the outer cell membranes are lysed, and whether the suspension contains encumbering amounts of non-nuclear material. A phase-contrast light microscope should be used at each stage of the nuclear isolation procedure to evaluate the yield, purity and integrity of nuclei, which can be visualized and scored with a hemocytometer (**Fig. 2a**). Nuclei will stain with trypan blue, and the nucleolus

Figure 2 | Quality control of nuclei isolation. (**a**,**b**) Nuclei were obtained from the human prefrontal cortex and extracted via Dounce homogenization; they were stained with 0.2% (vol/vol) trypan blue, counted on a hemocytometer (**a**), placed on a slide and microscopically examined for morphological quality and yield (**b**). (**c**,**d**) By using epifluorescence microscopy, nuclei were stained with DNA intercalating dye Hoechst 33342 (10 ng μl^{-1}) (**c**), with blue fluorescent nuclei images overlaid with the bright-field image to identify intact nuclei (**d**). (**e**) After cell strainer filtration, nuclei were stained with NeuN-Alexa Fluor 488–conjugated antibody (0.01 mg ml^{-1}) to identify intact neuronal nuclei. (**f**) The fluorescent image was overlaid with the bright-field image to further distinguish nuclei derived from neuronal versus non-neuronal cells. (**g**,**h**) By using FACS, cells were sorted onto a microscope slide and imaged for NeuN fluorescence (**g**) and overlaid in bright field (**h**) to confirm FACS sorting conditions.



can often be identified (**Fig. 2b**). Fluorescent labels, Hoechst for DNA and a neuronal nuclei marker, NeuN, can be used together for facile detection of nuclei derived from neurons (**Fig. 2c–h**).

Staining and FACS. To enable sorting of nuclei derived from neurons, nuclei can be immunostained with an antibody specific to NeuN, a nuclear membrane protein (**Supplementary Fig. 1**), before filtering the homogenate to remove large aggregated debris and subjecting it to FACS. Software gating on the FACS (**Fig. 3**) uses a series of doublet discrimination gates (**Fig. 3a–c**) to isolate single nuclei from any remaining aggregated nuclei, followed by a nuclear staining gate using Hoescht and NeuN labeling to isolate single neuronal nuclei (**Fig. 3d**,**e**). Alternatively, nuclei from all cell types can be sorted by using nuclear staining with either Hoechst or propidium iodide (PI) (**Fig. 3d**,**f**). Single nuclei are sorted into lysis buffer containing ERCC (External RNA Consortium Control) spike-in RNA standards (Ambion), which allow the sensitivity of transcript detection to be determined. Following FACS, single nuclei can be verified to be free of the debris particles and aggregated nuclei by microscopic observation (**Figs. 2g**,**h** and **3h**,**i**).

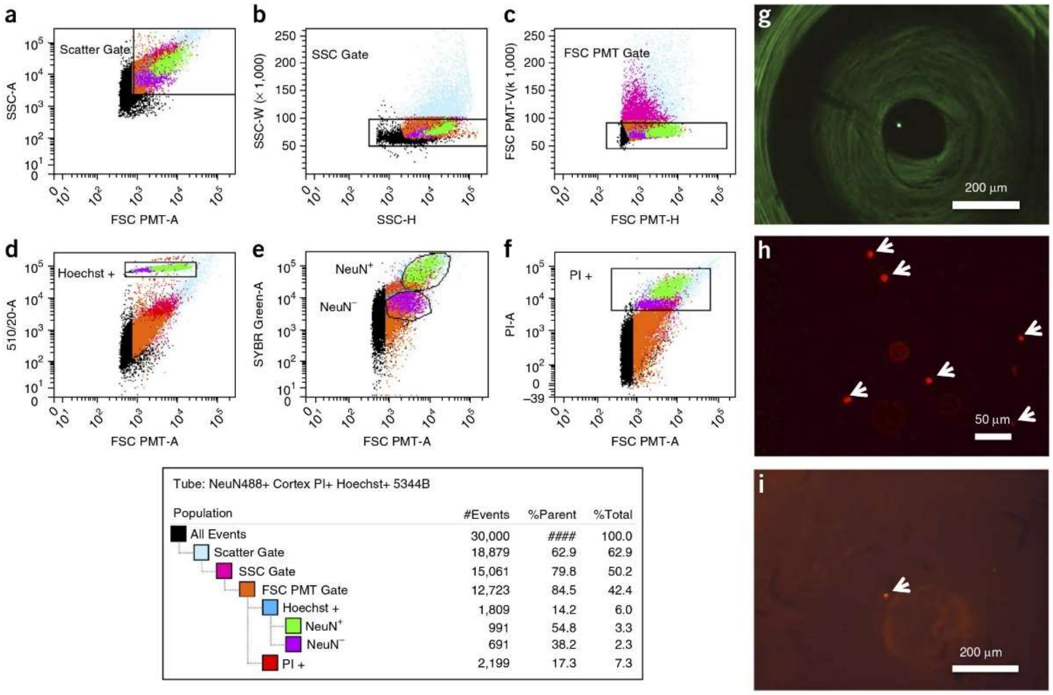
Lysis of nuclei, cDNA preparation and quality control. We do not provide detailed procedural information for nuclear lysis and cDNA preparation. Instead, we refer users to the Smart-seq2 protocol[3], which we now use because it generates a higher percentage of full-length cDNAs[4]. We follow the protocol exactly for lysis of the nuclei, but we have made two modifications for cDNA preparation: first, the cDNA is amplified by PCR for 21 cycles instead of 18 to compensate for the lower amount of RNA in a nucleus compared with a whole cell; second, the template-switching oligonucleotide (TSO) primer described in Picelli *et al.*[3] is modified by 5′ biotinylation[11]. We have recently confirmed (M.N. and R.S.L., unpublished data) observations by others that this modification reduces non-specific amplification caused by synthesis of TSO concatemers.

Before investing time and funds in RNA-seq, we carry out quality control assays by qPCR for targeted gene products. We use reporter housekeeping genes (*ACTB* and *GAPDH*), as well as high-, medium- and low-copy ERCC spike-in control qPCR assays (Thermo Fisher). In addition, assays targeting genes specific for neuronal nuclei of interest are recommended.

Preparation of sequencing library and sequencing. For procedural details for preparing sequencing libraries, we refer users to the Fluidigm C1 manual (C1 System for mRNA-Seq, part no. 100-7168 available at https://www.fluidigm.com/documents; select 'C1 System for mRNA Seq' to download the PDF automatically). We use the Illumina Nextera XT library preparation kit and perform multiplexed paired-end sequencing of barcoded libraries using an Illumina MiSeq system. **Figure 4** shows an example of the quality of the cDNA and sequencing library. **Supplementary Table 1** shows a summary of a typical sequencing experiment. The cDNA insert size of the sequencing library is 250–500 bp, and the read-length of paired-end sequences is 150 bases. A read-depth of 1.5–2.0 × 10^6 has been previously shown to be adequate for the detection of saturating levels of RNA expression in single cells[23].

Data analysis. The sequence reads are analyzed for quality and pre-processed to remove artifacts that fail to map to the genome (**Box 1**). A substantial number of reads contain Smart-seq2 primer and adapter sequences and their concatemers. In addition, deep sequencing yields many duplicate sequences of abundant transcripts that will reduce the ability to detect low-copy transcripts. Duplicate sequences cannot be removed, as removal would preclude accurate quantification of RNA expression. However, it is imperative that the levels of sequence duplication across samples are evaluated to examine its potential impact on the detection of low-copy transcripts.

After quality assessment and trimming, we perform analysis of RNA expression using the RSEM package[24], as described in **Box 2** and Steps 29–31. The trimmed sequencing reads are mapped to the human and ERCC spike-in transcript reference sequences.



Tube: NeuN488+ Cortex PI+ Hoechst+ 5344B

Population	#Events	%Parent	%Total
■ All Events	30,000	####	100.0
□ Scatter Gate	18,879	62.9	62.9
■ SSC Gate	15,061	79.8	50.2
■ FSC PMT Gate	12,723	84.5	42.4
■ Hoechst +	1,809	14.2	6.0
🟩 NeuN+	991	54.8	3.3
■ NeuN−	691	38.2	2.3
🟥 PI +	2,199	17.3	7.3

Figure 3 | FACS of single nuclei. Nuclei triple-stained with NeuN-Alexa Fluor 488–conjugated antibody (0.01 mg ml^{-1}; EMD Millipore), Hoechst 33342 (10 ng ml^{-1}) and PI (1 μM) were filtered through a 35-μm cell strainer and loaded onto a custom FACS ARIA II flow sorter (Becton Dickinson) equipped with a forward scatter photomultiplier tube. (**a–d**) Doublet discrimination gating was used to isolate single nuclei (**a–c**) and intact nuclei determined by subgating on Hoechst 33342 (**d**). (**a**) Particles smaller than nuclei (black dots) are eliminated with an area plot of forward scatter (FSC-PMT-A) versus side scatter (SSC-A), with gating for nuclei-sized particles inside the gate (box). (**b**,**c**) Plots of height versus width in the side scatter and forward scatter channels, respectively, are used for doublet discrimination with gating to exclude aggregates of two or more nuclei. (**e**,**f**) Subsequent plots and gating discern NeuN-Alexa Fluor488–conjugated antibody (**e**) and PI-stained nuclei (**f**). The resultant hierarchical color key ensures that only single nuclei that are positive or negative for staining with the NeuN antibody (NeuN+ and NeuN−) are passed through each gating condition. (**g**) Yellow fluorescent 10- to 14-μm polystyrene microspheres (Spherotech) were used to determine the accuracy and precision of microplate targeting, and they were confirmed by microscopic imaging of single spheres in a 384-well microplate. (**h**,**i**) Subsequent FACS gating of labeled nuclei (arrows) was confirmed via imaging on a microscope slide (**h**), as well as within individual wells of a 384-well microplate (**i**).

The sequencing depth observed for a given transcript quantitatively reflects the number of mRNA template molecules obtained from the lysed nucleus. The total number of genes detected for each nucleus and the percentage of reads mapped to the genome and ERCC spike-in controls is determined (**Fig. 5**). The sensitivity of the detection of RNA expression across different samples is analyzed by evaluating the expression of both ERCC spike-in control transcripts (**Fig. 6**) and the human mRNA at different levels of abundance (**Fig. 7**). All sequencing data—even from high-quality RNA—will show some level of 3′ bias in the coverage, because the reverse transcriptase (RT) will fail to produce full-length cDNA for some proportion of the transcripts, resulting in little or no coverage for the 5′ end of these RNAs. Even though the Smart-seq2 method disfavors incomplete cDNA strand synthesis, some cDNA that is only partially extended is still generated. In addition, 3′ bias will be indicative of mRNA damage due to RNase degradation, shearing or hydrolysis, which might occur during tissue handling, storage or processing of the nuclei. Partially degraded RNA will result in deeper sequence coverage for the 3′ end of transcripts, as only those degradation products that contain the 3′ polyA tail will be converted to cDNA (**Fig. 8a**). To confirm that any 3′ bias observed in cDNA from nuclei is not due to RNA degradation, we compare the sequence coverage with that of a high-quality control RNA from the same tissue (**Fig. 8b**). Sequence coverage of introns and exons is used to ensure that the sequences are derived from mRNA rather than from genomic DNA (**Fig. 9**), which is not removed from the nuclear extracts.

The primary goal of many single-cell or single-nuclei sequencing pipelines is the classification and characterization of known and potentially novel cell types, and several strategies have been presented for such analyses of hundreds to many thousands of cells[11,25,26]. For the small number of nuclei analyzed here, we developed an approach based on a straightforward application of dimensionality reduction (principal coordinate analysis), *k*-means clustering and manual inspection of canonical cell type marker genes (**Fig. 10**), which can be reproduced using the code provided as **Supplementary Methods**.

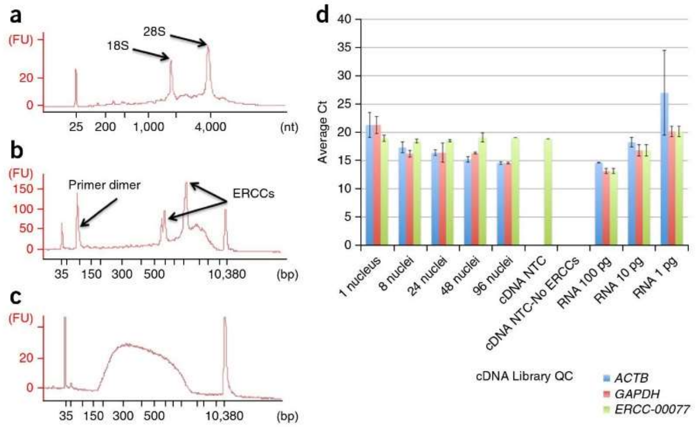


Figure 4 | qPCR and Bioanalyzer quality control analysis of total mRNA, single-nucleus cDNA synthesis and a single-nucleus NexteraXT RNA-seq library. Total RNA from ~2–3 mm^3 section of total human prefrontal cortex tissue was purified using a Qiagen RNeasy mini kit, quantified by Nanodrop spectrophotometry and diluted to 5 ng μl^{-1}. (**a**) The mRNA quality was determined using RIN values by loading 1 μl onto an Agilent RNA pico chip and run on the Agilent Bioanalyzer. (**b**,**c**) Representative example using a single nucleus for Smart-seq2 cDNA synthesis followed by PCR amplification (**b**; 1 μl) and a Nextera XT sequencing library (**c**; 1 μl) were also analyzed. (**b**) After AMPure bead purification of the cDNA, a size range of ~150 bp to 7 kbp is expected, with the majority of fragments in the 1–3 kb range. After AMPure bead purification of each Nextera XT library, a size range of ~200 bp to 1 kbp is expected. The hash marks on the *x* axis are 35, 50, 100, 150, 200, 300, 400, 500, 600, 700, 1,000, 2,000, 3,000, 7,000 and 10,000, with lane marker peaks seen at 35 and 10,380 bp. Separately, Smart-seq2 synthesis of cDNA and PCR was performed on single nuclei (*n* = 24), and on pools of 8 nuclei (*n* = 4), 24 nuclei (*n* = 4), 48 nuclei (*n* = 2), 96 nuclei (*n* = 2) and duplicates of 100 pg, 10 pg and 1 pg total RNA from the prefrontal cortex, to serve as technical replicates to reveal artifactual noise level due to technical causes such as variation in pipetting and temperature differences between PCR block wells. NTCs are used to detect nonspecific cDNA amplification derived from contaminants in the reaction components or introduced during handling. (**d**) Quality control qPCR of cDNA was performed in 10-μl reactions using ABI TaqMan gene expression assays for *GAPDH, ACTB* and ERCC-00077. qPCR cycle threshold (Ct) values were plotted for comparison with single nuclei Cts, typically ranging between 15 and 25. Note that Cts increase by about 3 cycles per tenfold increase in input RNA template, as expected from the doubling rate of DNA in PCR.

Advantages and limitations

The key strengths of our protocol are as follows:

- The use of nuclei for RNA-seq avoids the difficulties involved in obtaining undamaged whole neurons.
- Alteration of the transcriptome by treatment with proteases is avoided. The clinical samples and isolated nuclei are maintained at 4 °C until they are ready for use in cDNA synthesis.
- We have demonstrated RNA-seq from nuclei isolated by micromanipulation[13] and FACS.
- The technical and biological variation is similar for whole cells and nuclei[13]. For most transcripts, the nuclear and whole-cell expression profiles were similar, and therefore nuclei can generally be substituted for whole cells to define cell lineage, state or type populations, for example, by principal component analysis.
- Nuclear transcriptomes will provide insights into how they differ from cytoplasmic transcriptomes such as enrichment of certain transcripts in nuclei[13] and regulatory processes controlling the rate of transcription[27].

Box 1 | Sequence analysis: evaluation of sequence quality and preprocessing

(A) Assessment of sequence quality

Illumina sequences obtained from each sample (nucleus) are analyzed by fastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to evaluate sequence yield, base quality, GC profile, *k*-mer distribution and primer contamination. A computer grid environment or a multiprocessor (CPU) Unix workstation is required for processing large numbers of samples simultaneously.

(B) Evaluation of sequence duplication

To assess the extent of unique transcript representation and any skewed PCR bias in the fragments represented in cDNA libraries, the degree of read duplication is analyzed. However, the duplicated RNA-seq reads are not removed, as it will preclude the accurate estimation of transcript abundance (expression). The fastx_collapser tool (http://hannonlab.cshl.edu/fastx_toolkit/commandline.html) is used with Phred 33 base quality score offset to calculate the absolute number of identical reads (duplicates) in the input sample .fastq sequences. The program accepts only one sequence file as input. Multiple sequence files require iterative processing by a shell script.

(C) Trimming of adapters, primers and low-quality bases

The Trimmomatic tool (http://www.usadellab.org/cms/?page=trimmomatic) is used to trim the adapter and/or primer sequences present in adapters_primers.txt (**Supplementary Note**) from the ends of PE input.sample.fastq sequences to facilitate their successful mapping to the reference transcriptome. The program, executed using eight threads per job, performs the following: trims the end bases below a Phred quality score of 3 or any bases in a 4-base-wide sliding window when the average quality per base drops below 15; clips adapters/primers from the sequences by allowing two seed mismatches, requiring a minimum of 30 matches in palindromic mode and a minimum of ten matches in nonpalindromic (simple) mode between the read sequence and the adapters/primers. Any sequences trimmed from the original length of 150 bases to shorter than 60 bases are removed from the output.

Box 2 | Sequence mapping and RNA expression analysis

(A) Preparation of the reference genome

The trimmed sequencing reads are mapped to the transcripts derived from the human reference genome (GRCh37). The reference .fasta is prepared by the concatenation of GRCh37 human genome .fasta, the ERCC RNA spike-in .fasta and .fasta files for other marker (GFP) genes (RSEM_GRCh37_ERCC_GFP_RNASpikes.fa). The reference index files required by Bowtie2 mapping program and the transcript-specific reference sequences are generated from the GRCh37_ERCC_GFP_RNASpikes.fa and the corresponding annotation (GRCh37_ERCC_GFP_RNASpikes.gtf) files by the 'rsem-prepare-reference' command available in RSEM expression analysis software (http://deweylab.biostat.wisc.edu/rsem/).

(B) Mapping and the calculation of expression values

The 'rsem-calculate-expression' command from the RSEM expression analysis software is used to map paired-end reads to the reference transcripts (RSEM_GRCh37_ERCC_GFP_RNASpikes.transcripts.fa). The RNA expression values at gene and isoform levels are calculated using the expectation-maximization (EM) algorithm as implemented by the RSEM program. Multiple threads of eight or more are used to generate alignments mapped to genomic coordinates (sample_name.genome.bam), while tagging reads with nonunique alignments (--tag), calculating 95% credibility intervals (--calc-ci) and posterior mean estimates (--calc-pme), allowing insertions in the range of 1–500 bases (--fragment-length-min/max) and estimating the read start position distribution (--estimate-rspd). The text entries shown in parentheses in the preceding lines indicate the command's options. The output files are prefixed with sample_name.

(C) Determination of the sensitivity of expression analysis

The ERCC spike-in transcripts available from Life Technologies (https://www.lifetechnologies.com/order/catalog/product/4456740) are added to the reverse transcriptase mix along with sample RNA before the cDNA amplification. The individual ERCC spike-in mRNAs (http://tools.lifetechnologies.com/content/sfs/manuals/cms_095046.txt), which are present at a wide range of low to high molar concentrations in the reaction mixture, facilitate the determination of the lower threshold of detection sensitivity of transcript expression in terms of copy numbers.

The main limitations are as follows:

- Cytoplasmic mRNA concentrations are directly rate limiting for protein synthesis, and thus whole cells may possibly give a more direct indication of downstream biological functions dependent on the proteome. Use of nuclei might result in loss of some information contained in cytoplasmic mRNA; however, for frozen brain tissue, whole cells have tended to generate poor-quality cDNA, and they may not be an option.
- Nuclei are generally fragile compared with whole cells, and some loss can be expected at each stage of an isolation procedure[18].
- The small amounts of mRNA present in nuclei may necessitate optimization of the number of PCR cycles required to obtain sufficient cDNA for use in sequencing depending on the experimental needs. We amplified the nuclear cDNA with 21 cycles because of low amounts of RNA in the nucleus, compared with 18 cycles for whole cells[3]. However, some low-copy transcripts may still be more difficult to detect in nuclei. Furthermore, increasing the cycle number could introduce some amplification bias in the library by compressing expression values for high-copy transcripts.
- Cytoplasmic transcripts are not detectable, nor are small noncoding RNAs (ncRNAs) and other short sequence mRNAs lacking polyA tails. The low amounts of RNA contained in a nucleus may also prevent the detection of some ncRNAs.

Applications

Nuclear and cytoplasmic transcriptomes are likely to differ in many ways, and a more comprehensive analysis is needed to determine the advantages and limitations of using nuclei for transcriptomic studies. Some studies of specific nuclear functions

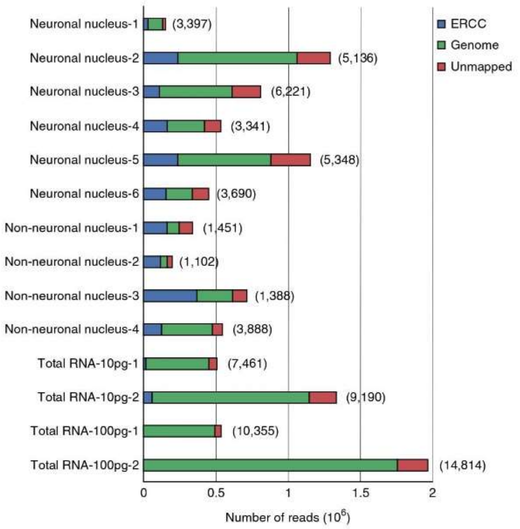


Figure 5 | Overall characteristics of mapping and expression. The sequencing reads for ten individual nuclei were split into three groups: 'ERCC', 'Genome' and 'Unmapped' on the basis of their mapping using the RSEM software. On average, 417,964, 183,278 and 941,644 reads were mapped to the genome for each neuronal nucleus, non-neuronal nucleus and total RNA sample, respectively. The numbers in the parenthesis indicate the number of genes with a TPM value >0 for the sample. It is clear that our sequencing did not reach saturation for some samples, as there is a high correlation between the number of reads mapped to the genome and the number of genes expressed. The high number of genes detected for Total RNA also reflects the pooling of RNA from multiple cells, which captures all genes expressed in the population.

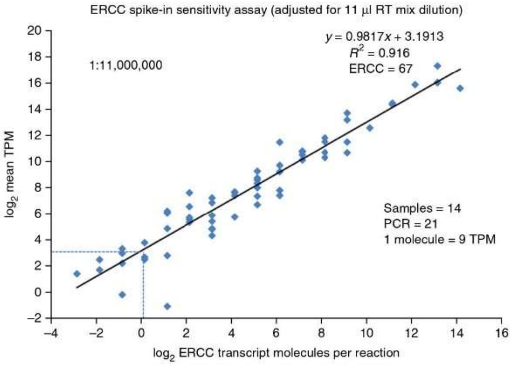


Figure 6 | Behavior of ERCC spike-in controls, sensitivity and detection limit estimation. The number of ERCC spike-in transcript molecules, diluted 1.1 × 10^7 fold from the original stock in the final RT-mix, are plotted against the average TPM expression values across all 14 samples using $\log_2$ scale for both axes. The 1.1 × 10^7-fold dilution (PROCEDURE Step 13 and INTRODUCTION) is greater than that recommended by the ERCC spike-in manufacturer, who had optimized it for use with nanogram quantities of RNA in microarray studies. The low levels of RNA in a single nucleus necessitate the greater dilution in order to avoid high percentages of sequencing reads devoted to ERCC spike-ins. However, some of the lower-copy transcript species present in the ERCC spike-in stock are consequently diluted to <1 copy per Smart-seq2 reaction tube. ERCC spike-in transcripts with expression in at least one of the 14 nuclei were considered (ERCC n = 67 of 92) with regression equation $y = 0.9817x + 3.1913$ and $R^2 = 0.916$. The RNA released from the lysed nuclei plus the added ERCC spike-in controls were amplified to 21 PCR cycles. The detection threshold for a single ERCC spike-in transcript molecule is shown to be approximately equivalent to 9 TPM RNA expression units (1 molecule = 9 TPM, as indicated by the intersection of the dashed lines).

may be enhanced by directly accessing nuclei—for example, studies of the regulation of transcriptional activation mediated by transcription factors, promoters, enhancers, epigenetic modifications and other mechanisms that control synthesis of mRNA. Critical control of cellular development and function occur at this level of regulation. Some processing of ncRNAs may also require analysis via nuclei such as initial rates of primary miRNA synthesis. The polyA tail of this ncRNA species allowed measurement of cDNAs produced by polyT priming[13], whereas the polyA tail is removed before transport of this RNA to the cytoplasm. In general, we anticipate that the nuclear transcriptome will have some advantages for investigating the regulatory processes controlling transcription rates. In contrast, the concentration of cytoplasmic mRNA reflects transport from the nucleus and various rates of mRNA processing and degradation. The cytoplasmic mRNAs serve as the template for ribosomes and the formation of the proteome, and thus they may have advantages in some studies.

RNA-seq analysis of human neurons is particularly challenging. For acute surgically derived tissues, the isolation of intact living neurons has been proven to be difficult, although a recent report demonstrated feasibility[10]. Similarly, technical challenges including cell isolation, RNA quality and glial transcript contamination have hindered progress in profiling single neurons from frozen postmortem tissues (R.H. and E.S.L., unpublished data). The use of nuclei avoids these obstacles. Furthermore, protease treatment to disperse whole cells, as done in recent studies of single neurons[10,11], is known to profoundly alter gene expression[12]. We have recently observed additional examples in which protease treatment altered gene expression. Unexpected *Fos* activation was found in almost all of the cells dissociated by protease from a mouse brain region that is reported to have low *Fos* expression and which lacked Fos protein based on antibody staining before protease treatment. No such activation was observed using the nuclei isolation protocol, which is performed at 4 °C and without the use of proteases. Importantly, we are able to detect *Fos* activation using the nuclei isolation protocol when mice have been exposed to environmental stimuli, which are known to induce *Fos*[28]. These observations suggest that caution is needed in interpreting transcriptomes

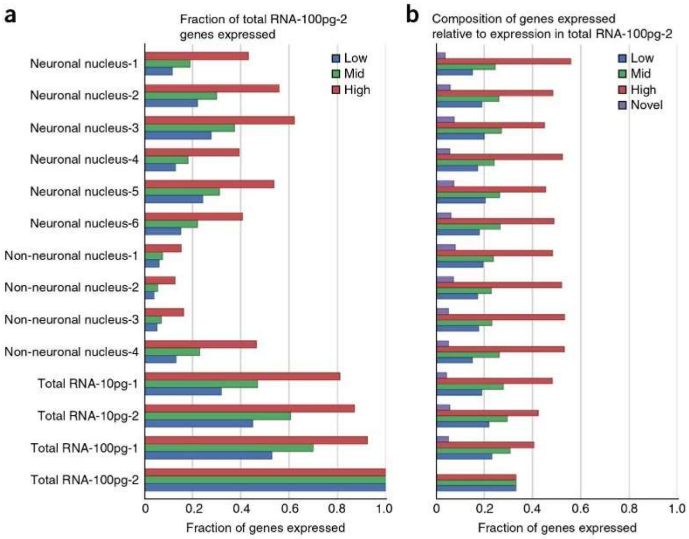


Figure 7 | Biological variation and technical noise stratified by relative expression of genes. The genes that are expressed in bulk Total RNA-100pg-2 (see **Supplementary Table 1**) were stratified equally into low, mid and high expressers based on their TPM values (4,292 genes per category). Low genes had TPM values between 0.01 and 7.44, mid genes had TPM values between 7.45 and 25.97 and high genes had TPM values >25.98 (**a**). For the 4,292 genes of each category, the graph shows the fraction found in each sample. By definition, Total RNA-100pg-2 has 100%, 100% and 100% representation for low, mid and high (**b**). Each gene that is expressed in the sample is labeled by its expression in the bulk RNA sample. The fraction of low-, mid- and high-expressed genes, as well as novel genes that were not found in the bulk control, was quantified.

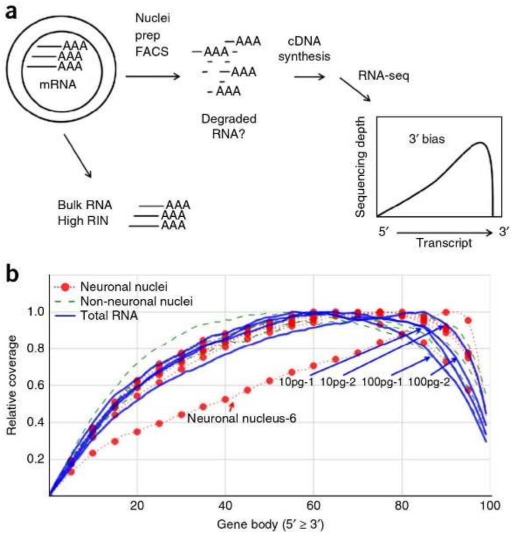


Figure 8 | The use of 3′ bias as a quality control assay for cDNA. (**a**) Total (bulk) RNA derived from tissue is confirmed to have a high RIN score before isolation of nuclei. Partial degradation of the RNA might occur during the preparation of nuclei by Dounce homogenization (nuclei prep) or FACS of the individual nuclei. If the mRNA is degraded by hydrolysis, shearing or RNases, truncated mRNA species could be created, and those containing the polyA sequence at the 3′ end of the transcripts might produce cDNA. This would generate greater RNA-seq coverage of the 3′ end of transcripts (3′-bias) compared with the high-quality bulk RNA. Gene body coverage across 4,292 highly expressed genes was calculated by RseqC. The relative coverage is defined as coverage at a base / maximum coverage across the gene. (**b**) The total RNA samples are indicated (two replicates of 10 pg and 100 pg RNA each; **Supplementary Table 1**). As these total RNA controls are all from a single RNA purification from bulk tissue, they would have identical coverage profiles in the ideal case. The minor differences indicate the level of technical variation accumulated from all of the reaction steps. The single nuclei have very similar 3′ bias to the total RNA controls, demonstrating that little damage was done to the RNA during the processing of nuclei. Neuronal nucleus 6 (**Supplementary Table 1**) is indicated, and it diverges from normal behavior. It may be an example of partially degraded mRNA being obtained from the nucleus and the resulting truncated cDNA; however, we believe that it is actually attributable to its low number of reads mapping to the genome, which must be taken into consideration for this analysis. We have recently confirmed that partially degraded total mRNA, which is formed experimentally by heating in the presence of sodium acetate, results in a commensurate increase in 3′ bias, demonstrating that this analysis can quantitatively detect RNA damage (M.N. and R.S.L., unpublished data).

from protease-treated cells. As the majority of accessible human brain specimens are obtained from frozen archives and collections, the use of nuclei may provide the best option that is currently available for RNA-seq from neurons.

The number of different cell types in the brain remains poorly understood. Cell 'type' implies stable characteristics, such as the synthesis of a particular neurotransmitter, and these cells have generally arisen by differentiation through developmental pathways, although the steps in these processes and their reversibility are not completely understood. It will be important to identify the abundance and functions of all the cell types in the brain. It also remains unclear how to define cell 'states,' which may simply reflect a range of intermediate functional activities rather than being discrete cell types. RNA-seq from individual brain cells will be crucial in resolving these questions. Moreover, RNA-seq will be a powerful new method for investigating the genomics and biochemistry of individual brain cells in a way that is not possible with bulk RNA. New computational methods are rapidly being introduced that will enable discovery of metabolic and regulatory pathways and investigation of brain function at the most basic levels of cell and systems biology. The use of nuclei to obtain transcriptomes from large numbers of cells has the potential to be a powerful new tool in neuroscience to investigate both normal and disease processes.

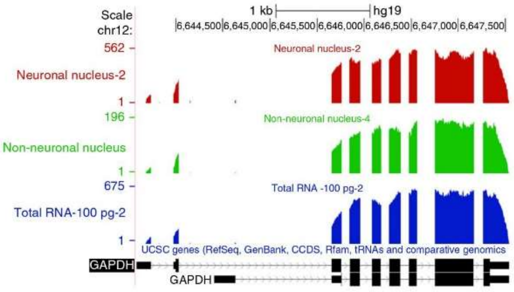


Figure 9 | Read depth across the *GAPDH* gene. University of California at Santa Cruz (UCSC) genome browser snapshot of custom bedGraph tracks detailing the coverage across the *GAPDH* gene for neuronal nucleus 2, non-neuronal nucleus 4 and total RNA 100pg-2 samples (**Supplementary Table 1**). The lack of coverage across introns indicates that most of the *GAPDH* transcripts sequenced were spliced transcripts for all three types of sample types. The position of exons is indicated by the black rectangles in the genomic map at the bottom.

Experimental design

Starting material. We have used cultured neuroprogenitor cells and fresh mouse brain tissue[13], and we include an example using frozen human brain (ANTICIPATED RESULTS), as the source for nuclei. When brain tissue can be used fresh without freezing (as for laboratory animals or when fresh human biopsies are available), we have elected to cool the sample to 4 °C and to use it for isolation of single nuclei as soon as possible. However, frozen brain tissue performed well, by producing full-length cDNAs and informative transcriptomes (ANTICIPATED RESULTS). Methods that cross-link mRNA, such as paraformaldehyde fixation of tissues, will severely limit the ability to produce full-length cDNA. When a sufficient quantity of tissue specimen is available for extraction of bulk RNA, we suggest determining the RNA quality before proceeding with single-nuclei isolation (**Box 3**). We selected tissues with RIN values ≥7, as these can be obtained from many brain archives. We have not carefully evaluated RNA of poorer quality. However, if RNA with a RIN score of <7 is all that is available, it should be tested and it may still yield valuable data. In general, we selected samples with the highest RIN available.

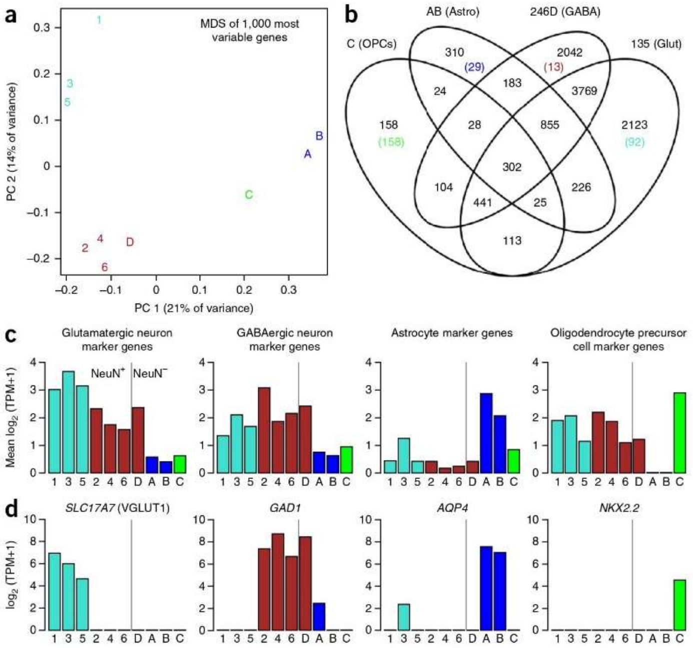


Figure 10 | Nuclei captured from several neuronal and glial cell types. Nuclei cluster into four discrete groups. (**a**) Multidimensional scaling (MDS) plot of 10 nuclei (**Supplementary Table 1**) based on the first two principal coordinates (PC, *x* and *y* axes). Labels 1–6 are the NeuN+ cells 1–6, and A–D correspond to NeuN– cells 1–4, and they are color-coded based on *k*-means clustering with *n* = 4. (**b**) Venn diagram showing the number of genes expressed in at least one cell in each group. The number of cells expressed in all cells of one cluster and no cells in any other cluster are shown in parentheses, and they are color-coded as in **a**. Cell clusters correspond to discrete cell types based on known marker genes. (**c**) Average expression of marker genes for glutamatergic neurons, GABAergic neurons[36], astrocytes and oligodendrocyte precursor cells[37] is shown for each cell, and it is color-coded as in **a**. Cells to the right and left of the vertical bar are the NeuN+ and NeuN– cells collected by FACS, respectively. (**d**) Canonical marker genes for glutamatergic neurons (*SLC17A7*), GABAergic neurons (*GAD1*), astrocytes (*AQP4*) and oligodendrocyte precursor cells (*NKX2.2*) are expressed as expected, based on cell type. Axes and colors for **d** are the same as those in **c**.

Homogenization. Nuclei were obtained by Dounce homogenization of ~2–3 mm^3 of human brain tissue for use in FACS sorting. In general, the Dounce step does not give quantitative recovery of nuclei because they are fragile and easily damaged. Some large pieces of tissue remained after this step; however, additional Dounce strokes appeared to destroy free nuclei even as more were released from the tissue. About 60,000 intact nuclei were obtained based on a hemocytometer count. If smaller amounts of tissue must be used, micromanipulation can be considered as a means to isolate a small number of nuclei[13].

The Dounce homogenization of tissues should be optimized for each specimen. Samples containing a mixture of cell types or samples from connective tissues and intracellular fibrous material may require more strokes. However, note that although more thorough homogenization (by increasing the number of strokes) will release more nuclei, it will also increase the number of damaged nuclei. The Triton X-100 used in this protocol is compatible with RNA-seq methods that use specific cell type enrichment via surface protein labeling. When immunostaining is not required, substitution of NP-40 for Triton X-100 has been suggested as a means to reduce loss of nuclei[21], although we have not verified this for use in single-nuclei RNA-seq.

Immunostaining. We immunostain nuclei with an antibody specific to NeuN, a nuclear membrane protein that is specific for neurons. In combination with Hoescht stain, which stains all nuclei, this allowed separation of nuclei by FACS into neuronal and non-neuronal populations. We have not found suitable alternative neuronal markers for FACS; staining for proteins within the nucleus would require permeabilization and fixation steps, which is incompatible with RNA-seq.

Isolating individual nuclei. We isolated individual nuclei by FACS; however, other methods can be used. We have also demonstrated the use of micromanipulation to isolate individual nuclei for use in RNA-seq[13]. Micromanipulation has the advantage of allowing inspection of nuclear morphology and fluorescent labeling with a microscope, and of providing confirmation that a single nucleus was added to the reaction well for cDNA synthesis. Micromanipulation may be an advantage for confirming the identity of nuclei from rare cell types that are not easily enriched by FACS. Another option is a microfluidic approach such as the C1 Single-Cell Autoprep System (Fluidigm), which can be used to isolate single nuclei from bulk preparations of adult human neurons (M.N., R.S.L. and M. Ray (of Fluidigm), unpublished observations). Similar to intact cells, some optimization of the nuclei loading conditions, including varying concentration, for each tissue type may be needed to maximize the nuclei captured per run. This instrument generally requires that at least 2,000 cells or nuclei be loaded onto the integrated fluidic circuit for optimal performance.

RNA-seq cDNA synthesis and sequencing platform. Smart-seq2[3] was used here to synthesize double-stranded cDNA; however, other methods can be used[1,4,5]. Previously[13], we successfully used the method by Tang *et al.*[5]. Any sequencing method is acceptable if it is well suited for the short cDNA library inserts. We have tested SOLiD sequencing (Life Technologies)[13] and Illumina sequencing (ANTICIPATED RESULTS) with comparable results.

Sample controls. It is important to include no-template controls (NTCs) in each experiment. Very low amounts of contaminating DNA or RNA, which are present in the Smart-seq2 reagents,

Box 3 | Sample quality assessment of tissue and cultured cells. ● TIMING 1 h

▲ CRITICAL Sample processing procedures vary widely depending on the sample type, and they can affect the quality of the RNA that can be obtained. For human postmortem brain, fresh mouse brain or cultured cells, we recommend determining the RNA quality by assessing the integrity of the bulk sample before proceeding with single nuclei isolation. If sufficient sample is not available, the tissues can be used directly for nuclei isolation.

1. For tissue samples, place a sterile Petri dish and scalpel on dry ice to chill. Transfer the brain sample to the Petri dish using sterile and RNase-free forceps. Remove a section of ~2–3 mm^3 using the scalpel. For cultured cells, collect them by trypsinization and centrifugation. Remove the supernatant and resuspend the cells in 1× cold PBS. Pellet the cells with centrifugation at 2,000*g* for 15 min. Repeat resuspension and centrifugation two more times. The pelleted cells can be kept at –80 °C for up to 3 months or they can be processed immediately.
2. Follow the Qiagen RNeasy mini kit's recommended protocol to isolate total RNA from either tissue or pelleted cells.
3. Assess the integrity of the total RNA on an Agilent Bioanalyzer (or similar device) using an RNA 6,000 pico chip as per the manufacturer's recommendation.

▲ CRITICAL STEP Where possible, it is recommended to proceed with single nuclei isolation using samples that have a RIN value of ≥7.

for example, can be sufficient to compete with the small amount of targeted material from a single cell or nucleus. NTCs, which receive water instead of the sorted nucleus, should not support cDNA synthesis. If some bacterial reads are obtained, they are possibly derived from contaminants in the reagents. If human sequence is obtained from the NTCs, contamination introduced in the laboratory is likely. We also use an aliquot of the FACS effluent (lacking a nucleus) as a negative control[13] to demonstrate that the sort buffer cannot support cDNA synthesis owing to free RNA or DNA released from the homogenized tissue. Any robust cell line easily maintained in the laboratory can be used as a positive control to demonstrate typical performance and to detect a loss of efficiency due to poor reagents, for example. The use of the same positive control in all experiments is helpful, as typical RNA content and the number of genes expressed may differ among cell types. Also consider using cell lines that express specific marker genes of interest as positive controls for comparison with the brain tissues. It is helpful to include technical replicates in experiments in which purified RNA is used as the template. Technical sources of variation include degree of success in synthesizing cDNA and constructing Nextera libraries. The technical variation contributes noise that interferes with the desired detection of biological differences.

Spike-in controls. An extrinsically added spike-in RNA is used as a positive control for the reverse transcriptase (RT) reaction. We used the ERCC spike-ins[29], a set of 96 different microbial mRNAs. These are present in a range of concentrations, allowing the determination of the sensitivity and range for the detection of transcripts. The concentration of ERCC spike-ins added is adjusted for various applications so that they will contribute a smaller percentage of the reads compared with the experimental specimen. We have adjusted the dilution of the ERCC spike-in stock commensurate with the small amount of RNA in a human cell nucleus. The dilution can be adjusted if it is found that too many or too few reads are obtained. Changing the dilution will alter which of the 96 mRNA species represent <10 molecules, the lower limit for detection. The dilution of 1:1.1 × 10^7 in the RT reaction results in ERCC-00077 being present at 2.2 copies per reaction. Approximately 50 of the 92 species are detected by sequencing at this dilution, and the remaining 42 are not detected, as they are added at <1 copy per reaction. Failure to detect ERCC spike-in controls in RNA-seq indicates a failed Smart-seq2 reaction. Detection of ERCC spike-ins but failure to detect cellular transcripts indicates failed recovery of RNA from the nuclei. An unexpectedly high proportion of ERCC spike-in sequencing reads relative to cellular transcripts also indicates poor recovery of RNA from the nucleus or that the cell was relatively quiescent and had low RNA content.

qPCR controls for cDNA quality. The quality of the gene expression information in the cDNA libraries can be assessed before investing time and expense in DNA sequencing by using TaqMan qPCR for a limited number of transcripts. We have found that cycle thresholds for housekeeping genes typically range between 15 and 30 cycles, depending on the amount of available mRNA in the nucleus and the original sample RIN. Control samples with 8, 24, 48 and 96 pooled nuclei should have correspondingly lower cycle thresholds. Total RNA controls from the same tissue sample ranging from 1 to 100 pg should also have progressively lower cycle thresholds. We have generally discarded cDNAs that lack all of the housekeeping genes tested for by qPCR. However, the pass/fail criteria are not easily defined, and they must be developed for each specific study. Caution should be exercised in discarding samples simply because certain transcripts are not detected, as transcription rates are highly variable through time, even for constitutive genes. qPCR for transcripts that are diagnostic for a cell type and other specialized characteristics can also be very useful in prescreening before investing in RNA-seq. However, where an unbiased sampling of a cell population is desired, it is important to weigh the benefits of selecting for specific transcripts against the risk of systematically biasing the selection.

In addition to sorting single nuclei, pools of 8, 24, 48 and 96 nuclei, for example, can serve as positive controls for cDNA synthesis. The pools also reveal the full range of transcripts in a cell population (the pan-transcriptome), and they can serve to validate detection of differentially expressed transcripts in the individual nuclei. The sequencing depth for a given transcript from a single nucleus can be compared with the sequencing depth from a pool of nuclei. For example, a transcript found at high copy number, but only in a small percentage of nuclei, should be commensurately low in the pools.

MATERIALS

REAGENTS

- Tissue sample. We have successfully used cultured neuroprogenitor cells and fresh mouse brain tissue[13] and frozen human prefrontal cortex brain obtained from the US National Institutes of Health (NIH) NeuroBioBank located at the University of Maryland as an example here (ANTICIPATED RESULTS). The quality of the initial sample can be checked before isolating nuclei, as described in **Box 3**. **! CAUTION** An Institutional Review Board approval may be required to obtain, process and place samples on a flow-sorting instrument. Precautions to protect the user include standard personal protective equipment, but potentially also a protective laminar flow hood for the flow cytometer if biohazardous sample material is to be used.
- RNaseZap RNase decontamination solution (Ambion, cat. no. AM9780)
- Nuclease-free water (Ambion, cat. no. AM9932)
- β-Mercaptoethanol, 14.3 M (Sigma, cat. no. M6250-100 ml) **! CAUTION** This is a combustible liquid. It is toxic if swallowed or if inhaled. It is very hazardous in case of skin contact (permeator) and ingestion. Severe overexposure can result in death. It causes skin irritation, and it may cause an allergic skin reaction. It also causes serious eye damage. Avoid contact with skin and eyes. Avoid inhalation of vapor or mist, and handle it while you are wearing appropriate personal protective equipment.
- Complete, EDTA-free (Roche, cat. no. 11873580001)
- Sucrose (Sigma, cat. no. S0389-500G)
- Potassium chloride, 2 M (Ambion buffer kit, cat. no. 9010)
- Tris buffer, pH 8.0, 1 M (Ambion buffer kit, cat. no. 9010)
- Magnesium chloride, 1 M (Ambion buffer kit, cat. no. 9010)
- EDTA, 0.5 M (Ambion buffer kit, cat. no. 9010)
- RNase inhibitor, cloned (40 U μl^{-1}; Ambion, cat. no. AM2682)
- Hoechst 33342, trihydrochloride, trihydrate (10 mg ml^{-1}; Molecular Probes, cat. no. H3570) **! CAUTION** This compound is harmful if swallowed. It causes skin irritation, and it may cause respiratory irritation. It is suspected of causing genetic defects; handle it while you are wearing appropriate personal protective equipment.
- Propidium iodide (PI; 1.0 mg ml^{-1}; (Molecular Probes, cat. no. P3566) **! CAUTION** This compound is harmful if swallowed. It causes skin irritation, and it may cause respiratory irritation. It is suspected of causing genetic defects; handle it while you are wearing appropriate personal protective equipment.
- DAPI (1.0mg ml^{-1}; Molecular Probes, cat. no. 62248) **! CAUTION** DAPI is harmful if swallowed. It causes skin irritation, and it may cause respiratory irritation. It is suspected of causing genetic defects; handle it while you are wearing appropriate personal protective equipment.
- Triton X-100 (Sigma-Aldrich, cat. no. T8787-100ML) **! CAUTION** Triton X-100 is harmful if swallowed, and it causes serious eye damage; handle it while you are wearing appropriate personal protective equipment.
- dNTP mix (10 mM each; Thermo Fisher, cat. no. 18427-088)
- Superscript II reverse transcriptase (Thermo Fisher, cat. no. 18064-014)
- Betaine (BioUltra ≥99.0%; Sigma-Aldrich, cat. no. 61962)
- KAPA HiFi HotStart ReadyMix (2×; KAPA Biosciences, cat. no. KK26010)
- Ethanol, molecular biology grade (Sigma-Aldrich, cat. no. E7023-500 ml)
- Agencourt Ampure XP beads (Beckman Coulter, cat. no. A63881)
- Adapter oligos (See Synthesis of cDNA, Step 19). All oligos except the LNA-modified TSO were ordered from IDT (https://www.idtdna.com), and they were HPLC-purified. LNA-modified TSO was ordered from Exiqon (http://www.exiqon.com/), and it was HPLC-purified. TSO (5′-biotin-AAGCAGTGGTATCAACGCAGAGTACATrGrG+G-3′); oligo-dT30VN (5′-biotin–AAGCAGTGGTATCAACGCAGAGTACT30VN-3′); ISPCR oligo (5′-biotin-AAGCAGTGGTATCAACGCAGAGT-3′)
- UltraPure BSA (50 mg ml^{-1}; Ambion, cat. no. AM2616)
- Trypan blue (0.4%; Sigma-Aldrich, cat. no. T8154)
- ERCC spike-in mix 1 (Ambion, cat. no. 4456740)
- RNase-free PBS, pH 7.4 (Ambion, cat. no. AM9625)
- 0.5% RNase-free BSA (Ambion, cat. no. AM2616)
- RNasin Plus RNase inhibitor (Promega, cat. no. N2615)
- Mouse IgG1k (BD Pharmingen, cat. no. 554121)
- Mouse monoclonal anti-NeuN antibody (Millipore, cat. no. MAB377)
- Goat anti-mouse Alexa Fluor 594–conjugated secondary antibody (Life Technologies, cat. no. A11005)
- DAPI (Life Technologies, cat. no. D1306)
- Yellow fluorescent polystyrene microspheres, 10 μm (Spherotech, cat. no. FP-10052-2)
- Perfecta ROX FastMix (Quanta Bioscience, cat. no. 95077-05K)
- TaqMan gene expression real-time PCR assay (Thermo Fisher)
- RNeasy Mini Kit (50) (Qiagen, cat. no. 74104)
- Quant-iT PicoGreen dsDNA assay kit (Molecular Probes, cat. no. P11496)
- Agilent RNA 6000 pico kit (Agilent Technologies, cat. no. 5067-1513)
- Agilent high-sensitivity DNA kit (Agilent Technologies, cat. no. 5067-4626)
- Nextera XT DNA library preparation kit, 96 samples (Illumina, cat. no. FC-131-1096)
- Nextera XT 96-index kit (Illumina, cat. no. FC-131-1002)
- MiSeq reagent kit v2, 300-cycles PE (Illumina, cat. no. MS-102-2002)

Software for sequence quality assessment

- FASTX (http://hannonlab.cshl.edu/fastx_toolkit/download.html)
- fastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).
- RSeQC[30,31] (http://rseqc.sourceforge.net/) can be used as an alternative to FASTX and fastQC

Software for sequence trimming

- Trimmomatic (http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.33.zip)
- Alternatively, Cutadapt[32] (https://cutadapt.readthedocs.org/en/stable/) can be used

Software for sequence alignment

- Bowtie2 (http://sourceforge.net/projects/bowtie-bio/files/bowtie2/)
- SAM tools (http://sourceforge.net/projects/samtools/files/samtools/)

Software for RNA expression analysis

- RSEM (http://deweylab.biostat.wisc.edu/rsem/). Alternatives to RSEM include Tophat2 (ref. 33) (https://ccb.jhu.edu/software/tophat/index.shtml), Cufflinks[33] (http://cole-trapnell-lab.github.io/cufflinks/) and Star[34] (https://code.google.com/p/rna-star/)

Software for data analysis

- R (https://cran.r-project.org/)
- Python and related packages (https://www.python.org/)
- IPython (http://ipython.org/)
- Pandas (http://pandas.pydata.org/)
- Matplotlib (http://matplotlib.org/)
- Seaborn (http://stanford.edu/~mwaskom/software/seaborn/)
- Bedtools (http://bedtools.readthedocs.org/en/latest/)
- IGV (http://www.broadinstitute.org/igv/)

EQUIPMENT

- Dounce homogenizer, 1 ml (Wheaton, cat. no. 357538)
- Sterile forceps (VWR, cat. no. 89259-946)
- Sterile Petri dish (VWR, cat. no. 25384-070)
- Sterile scalpel (Miltex, cat. no. 4-410)
- BD FACS-ARIA II Flow sorter with an automated cell deposit unit
- BD Falcon tube with a cell strainer cap (Becton Dickinson, cat. no. 352235)
- Falcon polystyrene conical tube (50 ml, BD Biosciences, cat. no. 352095)
- Inverted fluorescence microscope Olympus IX70
- Hemocytometer (Hausser Scientific, cat. no. 1483)
- Teflon-coated multi-well glass slides (Electron Microscopy Sciences, cat. no. 63430-04)
- 96-well black Fluortrac micro plate (VWR, cat. no. 82050728)
- 384-well plates (Phenix Research Products, cat. no. MPC-384HS4NH-C)
- 96-well plates (Eppendorf, twin.tec PCR plate 96, skirted, colorless, cat. no. D156224K)
- 8-strip, nuclease-free, 0.2 ml, thin-walled PCR tubes with caps (Eppendorf, cat. no. 951010022)
- Microcentrifuge Safe-Lock tubes (Eppendorf, cat no. 022363344)
- Multichannel pipettes and filter tips (Rainin LTS pipette set, 1–10 μl; 2–20 μl; 20-200 μl)
- DynaMag-96 side skirted magnetic rack (Thermo Fisher, cat. no. 12027)
- MicroAmp clear adhesive film (Applied Biosystems, cat. no. 4306311)
- MicroAmp optical adhesive film (Applied Biosystems, cat. no. 4311971)
- Thermal cycler (Applied Biosystems 9700)
- Fluorometer (Molecular Dynamics Flexstation 3)
- Spectrophotometer (Thermo Fisher, Model: NanoDrop ND-1000)
- Agilent 2100 Bioanalyzer (Agilent Technologies)
- Refrigerated centrifuge (Eppendorf, Model: Centrifuge 5804 R)
- C1 system for RNA-seq manual: 'Using C1 to Generate Single-Cell cDNA Libraries for mRNA Sequencing Protocol' (Fluidigm Part No. 100-7168, https://www.fluidigm.com/documents)

- DNA sequencing instrument ▲ CRITICAL A compatible Illumina DNA sequencing instrument (MiSeq, NextGen 500, HiSeq 2000, HiSeq 2500) is necessary to complete sequencing of the Nextera XT libraries, as the barcodes and sequencing adapters are designed for the Illumina sequencing platform.
- 64-bit computer running Linux with 4 GB of RAM (16 GB preferred)

REAGENT SETUP

Nuclei isolation medium #1 (NIM1) Combine the following components. ▲ CRITICAL This buffer should be made in advance, and it can be stored in a 50-ml conical tube at 4 °C for up to 6 months.

Component	Volume (µl)	Final concentration (mM)
1.5 M sucrose	2,500	250
1 M KCl	375	25
1 M MgCl$_2$	75	5
1 M Tris buffer, pH 8.0	150	10
Nuclease-free water	11,900	—
Total volume	15,000	—

Nuclei isolation medium #2 (NIM2) The following reagents should be combined in a 15-ml conical tube and placed at 4 °C or on ice for immediate use and then discarded.

Component	Volume (µl)	Final concentration
NIM1	4,895	
1 mM DTT	5	1 µM
50× protease inhibitor	100	1×
Total volume	5,000	

Homogenization buffer Combine the following reagents. ▲ CRITICAL This buffer should be made in a 5-ml conical tube, protected from light, and it should be placed at 4 °C or on ice for immediate use and then discarded.

Component	Volume (µl)	Final concentration
NIM2	1,452/1,453.5 (w/woPI)	1×
RNaseIn 40 U µl^{-1}	15	0.4 U µl^{-1}
Superasin 20 U µl^{-1}	15	0.2 U µl^{-1}
Triton X-100 10% (v/v)	15	0.1% (v/v)
PI (optional for FACS)	1.5/0 (w/wo PI)	1 µM
DAPI (optional for FACS)	1.5/0 (w/wo PI)	1 µM
Hoechst 33342	1.5/0 (w/wo PI)	10 ng ml^{-1}
Total volume	1,500	

Iodixanol medium (IDM) The following reagents should be combined in a 50-ml conical tube, and the medium can be stored at 4 °C for up to 6 months.

Component	1× volume (µl)	Final concentration (mM)
1.5 M sucrose	2,500	250
1 M KCl	2,250	150
1 M MgCl$_2$	450	30
1 M Tris buffer, pH 8.0	900	60
Nuclease-free water	8,900	—
Total volume	15,000	—

Iodixanol dilutions The following reagents should be combined, according to final concentration, in 50-ml conical tubes, and they can be stored at 4 °C for up to 6 months.

Component	1× volume (µl)	Final concentration
Iodixanol 60% (vol/vol)	12,500	50% vol/vol
IDM	2,500	—
Total volume	15,000	—
Component	**1× volume (µl)**	**Final concentration**
Iodixanol 60%	7,250	29% vol/vol
IDM	7,750	—
Total volume	15,000	—

Nuclei storage buffer (NSB) The following reagents should be combined in a 50-ml conical tube, and the buffer can be stored at 4 °C for up to 6 months.

Component	1× volume (µl)	Final concentration (mM)
Sucrose	0.855 g	166.5
1 M MgCl$_2$	50	5
1 M Tris buffer, pH 8.0	500	10
Nuclease-free water	14,450	—
Total volume	15,000	—

EQUIPMENT SETUP

FACS For high-throughput single-nuclei isolation by flow cytometry, the operator should be familiar with standard doublet discrimination gating and instrument settings for sorting single nuclei events. In preparation for sorting single nuclei into 384-well microplates for cDNA synthesis, accuracy and precision of sorting single events in a plate can be confirmed by targeting the bottom of each microplate well with 10-µm yellow fluorescent polystyrene microspheres and by inverting the plate for direct imaging on an inverted fluorescence microscope. Typically, 16 wells on both ends of the plate are targeted for spatial precision and >95% accuracy for a single bead. For nuclei sorting, staining in 1 µM DAPI, Hoechst 33342 or PI is suitable. The choice of stain depends on the number and type of antibody fluorophores used for the detection of the cell type of interest. Targeting and confirming sorted nuclei on a microscope slide and in microplate wells is also recommended. **Figure 3** shows the FACS gating strategy.

Computational requirements The protocol requires experience in running commands in UNIX (LINUX) shell environment. Experience with running Python and Perl language scripts is also required. C++, Perl, Python, Java and R programs are required to be installed. Prerequisite software is listed in the Reagents section. Users who do not have programming experience can use Galaxy analysis portal (https://usegalaxy.org/), which is an open, web-based platform, to execute most of the programs and commands described in this protocol, including those mentioned under alternate analysis packages. It allows the user to specify parameters and to run tools and workflows almost exactly as described under the PROCEDURE section of this protocol or modify some of the steps in the analysis in accordance with their preference. For more specific details on how to use this software, the user can access the site https://wiki.galaxyproject.org/. Data: requirements vary according to experimental goals .Sequence type: Illumina or other sequencing platforms that generate short reads (50–250 bases). Sequence format: .fastq or .fasta. Reference genome: .fasta, index and .gtf or .gff files.

Directory structure Choose or create a directory in which analysis is performed (RUNDIR). Save sequence files and reference .fasta, index and annotation (.gtf or .gff) files to SEQDIR and REFDIR, respectively. Trimmed reads are also copied to SEQDIR (these can be symlinks to files located elsewhere). The programs and individual commands described under the PROCEDURE section below are assumed to be available in the RUNDIR either as symlinks to the executables or copies of the installed binary files and scripts.

PROCEDURE

Nuclei isolation ● TIMING 1–2 h

▲ **CRITICAL** Keep the workstation and tools free of RNases by thoroughly cleaning with RNaseZap solution before the experiment.

1| Prepare nuclei isolation media 1 and 2 (NIM1 and NIM2) and homogenization buffer, and place them on ice.
▲ **CRITICAL STEP** NIM1 can be prepared and stored at 4 °C for up to 6 months. NIM2 and homogenization buffer should be freshly prepared.

2| Precool the Dounce homogenizer and pestles on ice. Once it is cooled, fill the homogenizer with 1.0 ml of cold homogenization buffer and keep it on ice.

3| If you are using tissue, transfer the sample to a Petri dish (on ice) and cut out a (2–3 mm^3) section using a chilled scalpel. Immediately transfer the tissue section into the precooled Dounce homogenizer. If you are using cultured cells, place 250 μl of cells (collected and resuspended in 1 × 10^6 cells per ml of 1× cold PBS) into the Dounce homogenizer.

4| Homogenize the tissue or cells with five strokes of the loose pestle, followed by 10–15 strokes of the tight pestle.
▲ **CRITICAL STEP** To reduce heat caused by friction, the Dounce homogenization should be performed on ice with gentle strokes, and care should be taken to avoid foaming. The mortar should be immersed in ice. The precooled homogenization buffer is an important aid in heat reduction during homogenization.

5| Filter the homogenate through a BD Falcon tube with a cell strainer cap; this filters out debris larger than 35 μm. Estimate the number of intact nuclei by staining a 10-μl aliquot of the filtered homogenate with trypan blue (10 μl), by loading it onto a hemocytometer and viewing it under a light microscope. At this point, nuclei can either be immunostained for neuronal markers (Optional Steps 6–12) to enrich for neuronal nuclei during FACS or they can be subjected directly to FACS (Steps 13–18) based on double discrimination only.
▲ **CRITICAL STEP** We obtained ~6 × 10^4 nuclei per milliliter from 2–3 mm^3 of frozen normal human cortical brain tissue. **Figure 2** shows a typical amount of debris present and varying sizes (7–10 μm) of nuclei from prefrontal cortical tissue.
▲ **CRITICAL STEP** For frozen human brain tissues, we recommend proceeding directly to FACS (Step 13), after filtering the homogenate, without further purification, as the nuclei have been subjected to freezing, and additional purification steps may cause further RNA damage. For fresh brain tissues, an additional iodixanol centrifugation-based purification may be helpful depending on the experiment. In general, each purification step results in lower yields of nuclei, and adjusting the starting material is desirable according to the downstream application.
? TROUBLESHOOTING

(Optional) Neuronal nuclei immunostaining ● TIMING 1–1.5 h

▲ **CRITICAL** The anti-NeuN antibody can be used to enrich for nuclei originating from neurons. We chose a dual-antibody staining strategy that first tags the nuclei with an unconjugated mouse anti-NeuN antibody, followed by a goat anti-mouse Alexa Fluor 594–conjugated secondary antibody. Mouse IgG1k detected by goat anti-mouse Alexa Fluor 594 serves as an isotype control for FACS to ensure specificity of the NeuN antibody (see **Supplementary Fig. 1** for the expected level of staining).

6| After homogenization and filtering (Step 5), concentrate the nuclei by centrifugation (1,000*g* for 8 min at 4 °C), and remove the supernatant. Resuspend in 500–1,000 μl of staining buffer (RNase-free PBS, pH 7.4, with 0.5% (wt/vol) RNase-free BSA and 0.2 U μl^{-1} of RNasin Plus RNase inhibitor).

7| Incubate the sample for 15 min on ice to allow for blocking of nonspecific binding with 0.5% (wt/vol) BSA. Remove 100 μl of the sample to a new tube for isotype control staining, and keep the remainder of the sample for staining with mouse anti-NeuN antibody.

8| For the isotype control sample, add purified mouse IgG1k to the tube at a final dilution of 1:5,000. For NeuN staining, add mouse monoclonal anti-NeuN antibody to the tube at a final dilution of 1:5,000. Incubate the samples on a tube rotator for 30 min at 4 °C.

9| Wash the samples by adding 500 μl of staining buffer to each tube and inverting the tubes several times. Spin the samples for 5 min at 400*g* in a refrigerated (4 °C) centrifuge to pellet nuclei.

10| Resuspend the pelleted nuclei in 500–1,000 μl of staining buffer, and add goat anti-mouse Alexa Fluor 594–conjugated secondary antibody to each tube at a final dilution of 1:5,000. Incubate the samples for 30 min on a tube rotator at 4 °C.

11| Wash the samples by adding 500 μl of staining buffer to each tube and by inverting the tubes several times. Spin the samples for 5 min at 400*g* in a refrigerated (4 °C) centrifuge to pellet nuclei.

12| Resuspend nuclei in 500–1,000 μl of staining buffer, and add DAPI at a final concentration of 1 μg μl^{-1} to each tube. Proceed directly to FACS (Steps 13–18).

Nuclei FACS sorting ● TIMING 2–3 h

13| Prepare lysis buffer by adding the following reagents to a 1.5-ml Eppendorf tube, and then place it on ice.

Component	1× volume (μl)	Final concentration
10% (vol/vol) Triton X-100	20	0.2% (vol/vol)
RNase inhibitor 40 U μl^{-1}	50	2 U μl^{-1}
ERCC spike-in mix 1, 1:2,000	1	1:2 × 10^6
Nuclease-free water	929	—
Total volume	1,000	

▲ CRITICAL STEP The lysis buffer should be freshly made for each experiment.

14| Prepare 96- or 384-well thin-walled PCR plates by adding 2 μl of lysis buffer to each well.

15| Prepare the FACS instrument for daily FACS setup, testing and droplet delay optimization.
▲ CRITICAL STEP We recommend adhering to the FACS manufacturer's instructions that the droplet stream be optimized for timing delay, with any satellite droplets merged by the fifth drop after the droplet breakoff. Failure to optimize the droplet breakoff may result in a charge placed on the satellite droplet instead of the droplet of interest.

16| Prepare FACS plots for doublet discrimination gating according to the manufacturer's recommendation to prevent sorting of doublets, triplets and further groupings of attached nuclei. Adjust the instrument software parameters to enable single-cell stringency. Load a small amount of sample into the instrument to confirm gating, and arrange gates on the FACS plots as needed. For samples that have been immunostained, sort populations for both $NeuN^+$ and $NeuN^-$ with the $NeuN^+$ population clearly distinguished with Alexa Fluor 488 fluorescence. If an unbiased nuclei population is desired, sorting may be completed using the $DAPI^+$ population.

17| Confirm FACS parameter settings for single nuclei sorting before sorting the actual samples. Confirmation can be achieved by targeting of the plate using 10-μm yellow fluorescent polystyrene microspheres or similar (Equipment Setup).
▲ CRITICAL STEP We recommend that even experienced FACS users complete a series of practice sorts (with single-cell sort instrument parameter settings) of microspheres before the actual sample sorting in order to confirm that the sorting is accurately timed and that the plate is properly targeted. Day-to-day variability in both of these parameters necessitates these precautionary steps to ensure efficient and accurate single nuclei sorting. Accuracy of microsphere sorting is determined by direct imaging of the microspheres at the bottom of the inverted microplate well (**Fig. 3g**). An accuracy of no less than 95% single microsphere sorting is recommended. For 384-well microplate sorting, the microscope objective often does not possess the dynamic focal range required to image the bottom of the well. A simple loosening of the objective for a few turns will bring the bottom of the well and the microsphere into focus. For 96-well plates, a custom objective with a long working distance for focal range may be required.

18| Proceed to FACS of sample nuclei. We recommend keeping the overall event rate for particles to 200–2,000 events per second on the FACS instrument to prevent swamping of the detectors that may result in a poor sorting accuracy. Depending on the concentration of nuclei, dilution of the sample may be required.
▲ CRITICAL STEP Before microplate sorting, a final confirmation of single nuclei sorting onto a slide for direct imaging of sorted single nuclei is recommended. Sorting into ~1 μl of NSB on a microscope slide can be sufficient to locate, count and image single nuclei. If the nuclei or a subpopulation of the nuclei are found to be difficult to distinguish from other particles, consider performing iodixanol density gradient centrifugation (**Box 4**) before proceeding with sorting of the rest of the sample.
? TROUBLESHOOTING
■ PAUSE POINT Plates with FACS-sorted nuclei can be sealed with a MicroAmp Thermo-Seal lid, frozen on dry ice and stored at −80 °C. Otherwise, proceed with lysis and reverse transcription immediately (Step 19).

Sequencing library preparation ● TIMING 2 h

24| Use cDNA preparations (from Step 19) that pass quality control (Step 23) to prepare a sequencing library; we use the Illumina Nextera XT library prep kit and follow the instructions in the Fluidigm C1 manual (see INTRODUCTION and MATERIALS). We start at page 35 of the manual with dilution of the cDNA and proceed through tagmentation, PCR amplification and AMPure XP bead cleanup, with the modifications for nuclei indicated in the table below. Determine the quality of the final pooled Nextera XT libraries, for example, by using the high-sensitivity DNA kit for Agilent Bioanalyzer according to the manufacturer's recommendations.

Modification no.	Page and Step in Fluidigm C1 manual	Modification	Reason
1	Page 41–43, Pool and Cleanup	Purify each of the Nextera XT reactions individually (not as a single pool) and Elute each individual reaction in 17 µl of Low TE (10:0.1) and quantify each with PicoGreen	Individual purification, elution and quantification of Nextera XT libraries allows for the exclusion of failed sequencing library preps in the final RNA-seq pool
2	Page 43, Repeat Cleanup Step	Pool the samples; note the starting volume of the pool. Perform cleanup using AMPure XP beads and elute with the same volume as used when pooled	A pool is generated from 3 ng from each individual library. The library should not include libraries that failed amplification

cDNA Sequencing: sequence type and yield ● TIMING 24 h

25| Subject the libraries to paired-end (preferable) or single-end sequencing on a suitable Illumina NGS platform (MiSeq, HiSeq and NextSeq); aim to generate 2–5 million reads per sample with a read length of 100–150 bases. Data are generated in .fastq format. Example sequencing statistics are provided in **Supplementary Table 1**.

RNA-seq analysis: sequence quality assessment and preprocessing ● TIMING variable

26| *Sequence quality assessment*. Evaluate sequence files from each nucleus (sample) from Step 25 using the fastQC tool for sequence yield, base quality, GC profile, *k*-mer distribution, contamination and so on. A computer grid environment should be used for processing a large number of samples simultaneously. The prototype command used is shown below. Note that the fastqc version available to the user can differ from the one shown here.

```
$ java -Xmx1500m -cp RUNDIR/fastqc_v0.10.1/FastQC/sam-
1.32.jar:fastqc_v0.10.1/FastQC/jbzip2-0.9.jar:fastqc_v0.10.1/FastQC/
-Dfastqc.nogroup=true uk.ac.babraham.FastQC.FastQCApplication
SEQDIR/input.sample.fastq.gz
```

27| *Sequence duplication*. Determine the degree of sequence duplication in the input data. Use the fastx_collapser tool to calculate the absolute number of identical reads (duplicates) in the input sample fastq sequences (from Step 25). Use correct base quality score offset (-Q). Process multiple sequence files iteratively (the program accepts only one sequence file as input).

```
$ RUNDIR/fastx_collapser -Q 33 -v -i SEQDIR/input.sample.fastq
1>/dev/null 2>input.sample.fastq.duplicate_summary.txt
```

28| *Sequence trimming*. Use the trimmomatic program to perform trimming of input paired-end or single-end .fastq reads (from Step 25) to remove adapter/primer sequences and low-quality end bases. The adapters and primers used in the commands below are shown in the **Supplementary Note**.

If sequences are paired-end only:

```
$ java -jar Trimmomatic-0.32/trimmomatic-0.32.jar PE -threads 8 –
phred33 -trimlog input.sample.fastq.trim.log input.sample.R1.fastq.gz
input.sample.R2.fastq.gz input.sample_trimmed.R1.fastq
input.sample_trimmed.S1.fastq input.sample_trimmed.R2.fastq
input.sample_trimmed.S2.fastq
ILLUMINACLIP:adapters.primers.txt:2:30:10 LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:60
```

If sequences are single-end only:

```
$ java -jar Trimmomatic-0.32/trimmomatic-0.32.jar SE -threads 8
-phred33 -trimlog input.sample.fastq.trim.log
input.sample.single.fastq.gz input.sample_trimmed.single.fastq
ILLUMINACLIP:adapters.primers.txt:2:30:10 LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:60
```

RNA-seq analysis: sequence mapping and expression analysis by RSEM ● TIMING variable

29| *Preparation of the reference genome*. Index the reference genome and transcript fasta files for mapping the trimmed reads to the reference genome using bowtie2 program. Use reference genome annotation file (GTF) for the generation of indexes for individual transcripts. Choose a prefix for naming the index files used in the mapping.

```
$ RUNDIR/rsem-prepare-reference --gtf
REFDIR/GRCh37_ERCC_GFP_RNASpikes.gtf --bowtie2
REFDIR/GRCh37_ERCC_GFP_RNASpikes.fa RSEM_GRCh37_ERCC_GFP_RNASpikes
```

30| *Calculating expression values*. Map paired-end reads that survive trimming (Step 28) to the reference transcripts, and calculate gene- and isoform-level expression values using expectation-maximization algorithm, as implemented by the RSEM program.

```
$ RUNDIR/rsem-calculate-expression --bowtie2 -p 8 --tag MA:i:2
-fragment-length-min 1 --fragment-length-max 500 --output-genome-bam
--calc-pme --calc-ci --estimate-rspd --time --paired-end
SEQDIR/input.sample.R1.fastq SEQDIR/input.sample.R2.fastq
RSEM_GRCh37_ERCC_GFP_RNASpikes sample_name
```

31| *Sensitivity assay of transcript expression*. To determine the lower threshold and the dynamic range of detection sensitivity across high to low copy numbers of RNA expression using ERCC spike-in transcripts, first convert the ERCC RNA spike-in molar concentrations (http://tools.lifetechnologies.com/content/sfs/manuals/cms_095046.txt) to number of molecules after adjusting for 1:1.1 × 10^7 dilutions used in preparing the final reaction mixture. Then, calculate mean transcripts per million (TPM) values from nuclei (samples) for each of the 92 ERCC spike-in transcripts expressed at >0 TPM in at least one sample. Finally, generate a regression plot after transforming the number of ERCC spike-in molecules (x axis) and the mean TPM values (y axis) on log2 scale (**Fig. 6** and **Supplementary Table 2**).

32| *Extract supplementary methods and load IPython Notebook*. Download the SupplementaryMethods.zip file (**Supplementary Methods**) and extract its content. It contains files for Steps 32–46 and Step 47 in the folders 'steps 32–46' and 'step47', respectively. For ease of use, Steps 32–46 are present in the accompanying IPython notebook (data_analysis.ipynb). The notebook also makes calls to the supplementary file (helpers.py) to parse and process the data generated. In what follows, all directions for the notebook appear as IN>. Note that it is not necessary unless directed to change the commands in the notebook; one may execute a code block by pressing control+enter. The commands are duplicated here for completeness and for alternate workflows. This pipeline is also available online at https://github.com/Schork-Lab/np_single_nucleus_rnaseq/

```
Download and move to directory with the SupplementaryMethods.zip
$ unzip SupplementaryMethods.zip
$ cd steps32-46
$ ipython notebook
```

```
In the browser window that opens, click on data_analysis.ipynb
```

33| *Load libraries and change paths*. The script begins by loading the necessary libraries. If the libraries cannot be loaded, please use the Python Package Index to download them, and restart the IPython notebook. Before beginning the analysis, it is necessary to set several paths that follow from the Directory Structure. These include paths to the .bam files (bam_dir) and RSEM-generated genes.results file (rsem_dir). These paths follow from the analysis until Step 30. In addition, if tools samtools, bedtools and geneBody_coverage.py are not in the system path, please include full paths to them.

```
IN> #Python libraries
import os

# Python packages
import pandas as pd
import seaborn as sns

# User modules
import helpers

# Figure styles
sns.set_context('notebook')
sns.set_style("white")

IN> data_path = "/home/kunal/tscc_projects/lasken/data/"
bam_dir = data_path
rsem_dir = data_path
out_dir = os.path.join(data_path, "out")
if not os.path.exists(out_dir): os.mkdir(out_dir)
path_to_samtools = 'samtools'
path_to_genebody_coverage = 'geneBody_coverage.py'
path_to_bedtools = 'bedtools'
```

34| *Calculate and plot overall mapping statistics*. Calculate the number of reads mapped to the genome, mapped to the ERCC spike-ins or that remain unmapped using the samtools idxstats tool. Python is used to generate the necessary Unix commands, and they are executed within the IPython environment. Load the resulting files into Python and generate a stacked barplot.

```
IN> for fn in os.listdir(bam_dir):
if fn.endswith('.genome.sorted.bam'):
      out_file = os.path.join(out_dir,
                          1.    fn.replace('.bam','.idxstats'))
      in_file = os.path.join(bam_dir, fn)
      samtools_cmd = "%s idxstats %s > %s" % \
      (path_to_samtools, in_file, out_file)
      print "Running samtools idxstats for file: %s" % fn
!$samtools_cmd

mapped_df = helpers.load_mapped_data(out_dir).sort()
ax = mapped_df.plot(kind= 'barh', stacked=True)
ax.set_xlabel('Number of Reads')
```

? TROUBLESHOOTING

RNA-seq analysis: biological and technical variation ● TIMING variable

35| *Load and parse TPM values generated by RSEM*. RSEM generates a genes.results file with several quantitative measures of a gene's expression. For all samples, load and parse these files to extract only the TPM column, and then merge all the files into a single matrix. For this matrix, the rows are gene ids, the columns are sample ids and the value of each cell is the TPM value for that gene in that sample. Filter out all genes that are only expressed in one sample or zero samples. Also, filter out ERCC spike-in contigs' expression from the TPM matrix.

```
IN> tpm_df = helpers.filter_df(helpers.load_tpms(rsem_dir),
                              genes_only=True,
                              expressed_in_multiple=False)
```

36| *Calculate and plot counts of genes expressed in single nuclei relative to bulk RNA*. Divide a chosen control sample's set of expressed genes into 'low', 'mid' and 'high' designations on the basis of their quantiles of expression. The default values used for the low- expressed genes are those that are in the quantile up to 0.33, the values for mid-expressed genes are: 0.33 to 0.67, and the values for high-expressed genes are: 0.68 to 1. For each sample, count how many genes are designated as low, mid, high, or novel through set intersections.

```
IN> control = 'Total RNA-100pg-2' # Set control sample name
low, mid, high = helpers.get_low_mid_high_genes(tpm_df[control])
expressed_df = helpers.calculate_relative_expression(tpm_df, low, mid, high)
```

37| *Plot relative expression in single nuclei compared with bulk RNA*. Create two plots: one plot details which fraction of the control sample's genes is expressed in each sample (**Fig. 7a**). The other plot details the relative composition of the genes expressed in each sample to the control sample (**Fig. 7b**).

```
IN> cols = ['Low', 'Mid', 'High']
control_values = expressed_df[cols].ix[control]
fraction_df = expressed_df[cols].astype(float)/control_values
ax = fraction_df.plot(kind= 'barh')
ax.set_title('Fraction of %s Genes Expressed' % control)
ax.set_xlabel('Fraction of Genes Expressed')
```

```
IN> composition_df = expressed_df.apply(lambda x:
x.astype(float)/x.sum(),
                                 1.    axis=1)
cols = ['Low', 'Mid', ′High', 'Novel']
ax = composition_df[cols].plot(kind= 'barh')
ax.set_title('Composition of Genes Expressed \nRelative to Expression
in %s′ % control,loc= 'left')
ax.set_xlabel('Fraction of Genes Expressed')
```

38| *Plot pairwise correlation of expression across all samples*. Calculate pairwise Spearman's correlation for all samples. Stratify the correlation matrices by low, mid and high genes based on their expression in the previously defined control sample. Plot the resulting matrices as heat maps.

```
IN> for genes, gene_type in zip([low, mid, high], ['Low', 'Mid′, 'High']):
            fig = plt.figure(figsize=(8,8))
            ax = fig.add_subplot(111)
ax = sns.corrplot(tpm_df.ix[genes.index].sort(axis=1),
method= 'spearman', ax=ax, diag_names=False,
cmap_range=(0, 1), cbar=True)
ax.set_title('Correlation Stratified by %s Expression in
%s' % (gene_type, control))
sns.despine()
```

RNA-seq analysis: quality based on coverage across the gene body ● TIMING variable

39| *Create bed file of highly expressed genes*. To gain a better idea of the quality of the transcripts being sequenced, focus on transcripts that are highly expressed. Create a .bed file to be used in other tools that only has the highly expressed transcripts based on the control sample.

```
IN> gtf_file = os.path.join(data_path,
                     'reference',
                     'GRCh37_ERCC_GFP_RNASpikes.gtf')
high_gtf = gtf_file.replace('.gtf', '.high_expressed.gtf')
print "Subsetting %s to only highly expressed genes as %s" %
(gtf_file, high_gtf)

helpers.subset_gtf_by_genes(high_gtf, gtf_file, list(high.index))
high_bed = high_gtf.replace('.gtf', '.bed')
print "Converting %s to %s" % (high_gtf, high_bed)
!perl gtf2bed.pl $high_gtf > $high_bed
```

40| *Calculate coverage across the gene body*. For the highly expressed transcripts, calculate their coverage across the length of the gene body using RseqC's geneBodyCoverage.py tool. Use Python to generate the command that includes all the sample .bam files, as well as the highly expressed genes .bed file. Run this command through the IPython shell.

```
IN> rnaseqc_prefix = os.path.join(out_dir,
"rnaseqc_high_coverage_control")
sample_files = [os.path.join(bam_dir, fn) for fn in
os.listdir(bam_dir)
              if fn.endswith('.genome.sorted.bam')]
in_files = ", ".join(sample_files)
rnaseq_c_cmd = " %s -i %s --refgene %s --out-prefix %s" % \
              (path_to_genebody_coverage, in_files, high_bed,
               rnaseqc_prefix)
fns = ", ".join([os.path.basename(fn) for fn in sample_files])
print "Running gene body coverage for sample files: %s" % fns
!rnaseq_c_cmd
```

41| *Plot relative coverage across the gene body*. Load in the previously generated geneBodyCoverage files from Step 40 using a helper function. The helper function defines the normalized coverage as the (coverage − minimum coverage)/(maximum coverage − minimum coverage). Plot the data and set appropriate labels.

```
IN> rnaseqc_file = rnaseqc_prefix+ '.geneBodyCoverage.txt'
normalized_df, coverage_df =
helpers.load_gene_body_coverage(rnaseqc_file)
```

```
ax = normalized_df.plot()
ax.set_xlabel("Gene Body (5' -> 3')")
ax.set_ylabel("Relative Coverage")
```

RNA-seq analysis: quality based on intron and exon coverage ● TIMING variable

42| *Align reads using TopHat2*. As the RSEM program maps sequences to only exons in the annotated reference transcripts, use TopHat2 program for mapping reads to both exons and introns in the reference genomic sequence. Generate the appropriate index files needed for bowtie2 mapper, which is executed by TopHat2. Run the following commands in sequence to generate a .bam alignment file with sequences mapped to exons and introns.

```
$ RUNDIR/bowtie2-build REFDIR/GRCh37_ERCC_GFP_RNASpikes.fa
GRCh37_ERCC_GFP_RNASpikes

$ RUNDIR/samtools faidx REFDIR/GRCh37_ERCC_GFP_RNASpikes.fa

$ RUNDIR/tophat2 -p 8 --library-type fr-unstranded -G
REFDIR/GRCh37_ERCC_GFP_RNASpikes.gtf GRCh37_ERCC_GFP_RNASpikes
SEQDIR/input.sample.R1.fastq.gz SEQDIR/input.sample.R2.fastq.gz
```

43| *Inspect in IGV*. Open IGV Viewer, and load in the .bam file. Manually zoom in and out of large housekeeping genes such as GAPDH to inspect whether only spliced transcripts are being sequenced.

44| *Create intron and exon .bed files*. Set paths for the names and locations of the intron and exon .bed files. Use the accompanying create_intron_exon_beds.sh to create intron and exon .bed files based on the provided GTF file.

```
IN> intronic_bed = os.path.join(data_path,
                'reference',
                'GRCh37_ERCC_GFP_RNASpikes.gtf.introns.bed')
exonic_bed = os.path.join(data_path,
                'reference',
                'GRCh37_ERCC_GFP_RNASpikes.gtf.exons.bed')
!sh create_intron_exon_beds.sh $gtf_file $exonic_bed $intronic_bed
```

45| *Calculate coverage overlaps with exons and introns*. Set the path to the TopHat2-generated .bam file (from Step 42). Use bedtools command bamtobed in conjunction with the bedtool coverage command to look at the coverage across introns and exons of the sample .bam file.

```
sample_tophat_bam = '/path/to/tophat.aligned.bam'
intronic_out = sample_tophat_bam.replace('.bam', '.intronic_coverage')
intronic_cmd = "%s bamtobed -splitD -i %s | awk
\'BEGIN{OFS=\"\\t\"}$1=\"chr\"$1\' | %s coverage -a - -b %s > %s" %
(path_to_bedtools, sample_tophat_bam, path_to_bedtools, intronic_bed,
intronic_out)
print "Creating intronic coverage file"
!$intronic_cmd
print

exonic_out = sample_tophat_bam.replace('.bam', '.exonic_coverage')
exonic_cmd = "%s bamtobed -splitD -i %s | awk
```

```
\'BEGIN{OFS=\"\\t\"}$1=\"chr\"$1\' | %s coverage -a - -b %s > %s" %
(path_to_bedtools, sample_tophat_bam, path_to_bedtools, exonic_bed,
exonic_out)
print "Creating exonic coverage file"
!$exonic_cmd
```

46| *Load and plot exonic versus intronic coverage*. Load the generated bedtools coverage files from Step 45 for the introns and exons. Select regions that have at least 1 read mapping to them and that are at least 1 kb long. Plot the differences between the intronic and exonic regions, as shown in **Supplementary Figure 2**.

```
IN> intronic_df = helpers.load_bedtools_coverage(intronic_out,
                                                 min_reads=1,
                                                 min_length=100)
exonic_df = helpers.load_bedtools_coverage(exonic_out, 1, 100)
fig = helpers.plot_bedtools_coverage(intronic_df, exonic_df)
```

Sample classification ● TIMING variable

47| Cell type classification for assessing how well nuclear and brain cell RNA matches: R code and additional files required to reproduce this step are provided in the folder 'step47' (**Supplementary Methods**). Convert Ensembl Gene identifiers into current gene symbols using BioMart (http://www.ensembl.org/biomart/martview; downloaded 1/26/15 (ref. 26)), and exclude all transcripts without a current gene symbol. Convert TPM values to log scale (offsetting by 1). Cluster cells by identifying the 1,000 genes with the highest variability, finding the Pearson's correlation distance, performing multidimensional scaling to identify the first four principal coordinates and running *k*-means clustering with $K = 4$ on these principal coordinates. Calculate the number of genes expressed in each cluster for comparison. Determine the cell type of each cluster by collecting lists of marker genes for known brain cell types[24,35], by determining the expression levels of these sets of genes in each nuclei, assigning cell type based on high expression of markers and confirming cell type classification based on nearly exclusive enrichment of individual canonical marker genes.

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 1**.

TABLE 1 | Troubleshooting table.

Step	Problem	Possible reason	Solution
5	Low nuclei yield	Poor-quality tissue	Obtain intact tissue with a low number of freeze-thaw cycles
		Lack of nuclei in tissue	Microscopically assess the density of nuclei in the tissue
		Inadequate cell lysis to release the nuclei	Use appropriate concentration of detergents, salt and sucrose in the cell lysis buffer
			Optimize the number of Dounce strokes
			Use a homogenizer with appropriate clearance level to release the nuclei
			Use chilled buffers and homogenizers, and execute the entire procedure at 4 °C
		Improper centrifugation may cause the cellular debris to sediment, which may alter the yield of pure nuclei	Optimize the density gradient for nuclei isolation and the speed of centrifugation to your tissue type
18	Poor recovery of single nuclei from FACS	Targeting of FACS for single nuclei isolation is compromised	Optimize FACS conditions and determination of sorting gates
			Sort nuclei onto a glass slide and visualize them under the microscope

(continued)

TABLE 1 | Troubleshooting table (continued).

Step	Problem	Possible reason	Solution
23	Failure of qPCR assays	No nuclei in the wells (if using FACS)	Optimize single-nucleus targeting into wells of the microtiter plate prior to FACS
		Low-quality RNA obtained from the lysed nuclei	Use a sample with a high RIN value
		mRNA degradation	Keep the workstation and tools free of RNases by thoroughly cleaning with RNaseZap. Do this daily or before each experiment
		Inefficient cDNA synthesis	Use fresh dNTPs
			Keep all reagents on ice and minimize the freeze-thaw cycles of sensitive items
		Reverse transcription failure	Check all cDNA synthesis steps using ERCC spike-in as a positive control
34	Excessive DNA sequencing reads failing to map to the reference genome	Concatemer formation from the TSO primer of the Smart-seq2 method	Be certain to use the 5′ biotin–modified TSO primer[11] (as done in step 19 and discussed in the INTRODUCTION) rather than the unmodified version used in Picelli *et al.*[3]

● **TIMING**

Steps 1–5, nuclei isolation: 1–2 h
Steps 6–12, (optional) neuronal nuclei immunostaining: 1–1.5 h
Steps 13–18, nuclei FACS sorting: 2–3 h
Step 19, cDNA synthesis by Smart-seq2: 1 d
Steps 20–23, qPCR and TaqMan analysis: 3 h
Step 24, sequencing library preparation: 2 h
Step 25, cDNA sequencing: sequence type and yield: 24 h
Steps 26–46, RNA-seq analysis: sequence quality assessment and preprocessing: variable
Step 47, sample classification: variable
Box 3, sample quality assessment of tissue and cultured cells: 1 h
Box 4, density gradient centrifugation: 1 h (optional)

ANTICIPATED RESULTS

This protocol enables the FACS-based isolation of single nuclei suitable for RNA sequencing. The use of a neuron-specific antibody for staining allows comparison of transcriptomes from neurons and other cell types. The RNA-seq data can be used to determine cell types based on the profiles of the genes expressed.

Figures 3–10 are generated from an RNA-seq experiment on single nuclei isolated from frozen normal human cortical brain samples obtained from the NIH NeuroBioBank located at the University of Maryland, where they were stored at −80 °C. The brain specimens had been collected and deposited at NeuroBioBank up to several hours after death. The nuclei were stained with NeuN-Alexa Fluor 488–conjugated antibody and sorted using FACS gating parameters designed to distinguish neurons and non-neurons (**Fig. 3a**–**f**). The sorting accuracy and precision for single nuclei was verified by sorting beads into 384-well plates and viewing them under the microscope (**Fig. 3g**). **Figure 3h**,**i** shows PI-stained nuclei.

Bioanalyzer analysis of the quality of cDNA library synthesis and amplification by Smart-seq2 gave typical results (**Fig. 4**). After AMPure bead purification of the cDNA library, a size range of ~150 bp to 7 kbp is expected, with the majority of fragments in the 1- to 3-kb range (**Fig. 4b**). Primer dimers in the size range of ~100 bp make up a small minority of the total cDNA, and therefore no further purification after Ampure bead cleanup is necessary before library prep. The primer dimers are further reduced in the purification of the library prep (**Fig. 4c**) and in a second purification of the pooled library for Illumina sequencing. After Nextera XT purification, the typical size range of the library is 200–1,000 bp (**Fig. 4c**). If necessary, libraries may be pooled and further purified before sequencing to get the optimal library insert size for maximum read depth, but with the expectation of some loss of material.

Detection of gene expression (Steps 29 and 30)

Of the ten nuclei sequenced in this example, six were identified as neuronal on the basis of FACS for the NeuN protein and four were non-neuronal (**Fig. 7**). Note that the percentage of reads mapping to ERCC spike-in controls, the genome and unmapped reads can vary widely depending on the starting amount of mRNA derived from the nucleus and the amount of artifactual PCR products such as primer dimers that are created. The number of genes expressed also varies widely among single nuclei, most likely owing to variation in the mRNA content of phenotypically different cell types, as well as technical sources of variation caused by insufficient lysis of the nuclei and suboptimal cDNA synthesis (see 'Sample controls' in the INTRODUCTION for comments on use of technical replicates to evaluate experimental noise). The number of genes detected ranged from 1,102 to 6,221 (**Fig. 7**). The range was higher for total RNA as expected, as a population of different cell types is represented by this RNA template. Failure to detect many genes expressed from a single nucleus may indicate poor yields of cDNA and lack of sensitivity for low-copy transcripts. However, caution must be used in this conclusion, as the cells may simply have been relatively quiescent. More genes will be expressed in pools of multiple nuclei reflecting the full range of genes expressed in the cell population. This can also serve as an important validation for genes detected in single nuclei. In general, the genes expressed in the pools should represent the sum of all genes detected in the individual nuclei. The level of expression should also agree between pools and individual nuclei. For example, a gene that is expressed at a high level but in only a small percentage of nuclei should appear at a commensurately low level in the pools. Expression signatures are nearly identical between nuclei and whole cells over a wide range of RPKM values; however, a subset of transcripts known to be enriched in nuclei, on the basis of bulk-RNA extractions, was confirmed as enriched in the individual nuclei[13].

Sensitivity of detection (Step 31)

The detection sensitivity of the RNA expression analysis is determined by adding ERCC spike-in control transcripts of various concentrations to the lysis buffer (Step 13) used to release RNA from the nuclei. The ERCC spike-ins are processed along with sample RNA through the RT reaction and subsequent cDNA amplification and sequencing (**Box 2c**). The limit of detection for ERCC spike-in transcripts should be <10 copies, as observed in the example provided here (**Fig. 6** and **Supplementary Table 2**). Expression of a single-copy ERCC spike-in transcript can be detected at an approximate threshold value of nine TPM (intersection with the *y* axis, **Fig. 6**). A failure to generate cDNA for the ERCC spike-ins would indicate failure of the Smart-seq2 reaction, for example, because of inactive RT. If the ERCC spike-ins generate the expected amount of cDNA but cellular transcripts are not detected, then the transcripts were lost at some stage of the process, probably because of degradation of RNA in the cell resulting from improper handling or storage, failure to successfully sort the nuclei into the wells or failure to completely lyse the nucleus.

When compared with the transcripts expressed at high and medium levels, those with a low level of expression show a greater degree of variation relative to the pattern of expression seen in the control total RNA samples (**Fig. 7b** and **Supplementary Fig. 3**). It is possible that the lack of expression for low-copy transcripts reflects real biological phenomena. For example, low-copy transcripts may be more likely to be variably expressed if they tend to be involved in regulatory or other nonconstitutive functions. However, we suspect that at least some of the effect results from variable sensitivity below 10 transcript copies.

Assessing 3′ bias (Steps 40 and 41)

3′ bias can be indicative of damaged RNA, as well as poor activity from the RT, and it is a source of noise in RNA-seq experiments (**Fig. 8a**). The graph output details the relative coverage across the gene body from the 5′ end to the 3′ end for the highly expressed genes in the example given (**Fig. 8b**). An almost square wave should be observed, showing uniform relative coverage across the gene body with drop-offs near the 5′ and 3′ end because of end effects in the Nextera tagmentation reaction for library construction. If the plot is highly skewed to the 3′ end relative to the plot for control RNA of high RIN value, it is indicative of poor RNA quality. In this example, a nearly identical 3′ bias was found in cDNA from nuclei and the control-purified RNA (**Fig. 8b**), confirming that the cDNAs from single nuclei were predominantly full length. Recently, we have confirmed that damaged mRNA controls, generated by heating in the presence of sodium acetate, quantitatively generate 3′ bias in the sequence coverage (M.N. and R.S.L., unpublished data).

Analysis of exon and intron coverage (Steps 42–46)

Most or all of the detected transcripts will be fully spliced with relatively uniform coverage across exon/exon junctions but not intron/exon junctions[13]. In the example shown here, exons are fully covered by at least one read, whereas only a few intronic regions have their entire length covered fully by reads (**Fig. 9** and **Supplementary Fig. 2**). Although many reads map to intronic regions[13], the source of these reads is not clear. The length of an exon does not seem to show a correlation with the extent to which the exon is covered. Only small introns (<10 kb) show full coverage across their entire length. The absence of intronic reads and the relatively even sequence coverage across exon/exon junctions confirms an earlier finding[13] that most or all of the transcripts obtained from the nuclear lysates have been spliced. Intronic reads were detected, but these were not present evenly across exon/intron junctions (**Fig. 9** and **Supplementary Fig. 2**), as should be observed if unspliced transcripts were detected.

Cell type classification (Step 47)

The RNA-seq data can be used to verify that specific cell types have been enriched by FACS. In the example shown here, the presence of the NeuN protein (**Fig. 5**, nuclei labeled neuronal), a neuron-specific nuclear marker, based on antibody labeling during FACS of nuclei, was consistent with the RNA-seq detection of NeuN transcript in half of the nuclei labeled with anti-NeuN antibody and none of the NeuN-negative nuclei (**Supplementary Table 1**). In cases in which the cell is positive for a protein marker based on FACS, but the transcript is not detected, it is possible that the protein is longer lived than the transcript. Transcription tends to occur in bursts, and it does not exactly reflect protein concentrations. Alternatively, some nuclei may be spuriously identified as positive during FACS.

Gene expression values from nuclei can be used to identify cell types[13]. In the example given here, gene expression was analyzed from the ten postmortem human nuclei to evaluate the identities and characteristics of the cells. To do so in an unbiased manner, we first identified the 1,000 annotated genes with the highest variability across the 10 nuclei, and we then clustered the nuclei into four groups using *k*-means clustering (**Fig. 10a**). All of the NeuN$^+$ nuclei (labeled 1–6 in **Fig. 10**) and one of the NeuN$^-$ (D) nuclei were found in two clusters that contained a large number of overlapping genes (**Fig. 10b** and **Supplementary Table 3**), whereas the remaining three NeuN$^-$ nuclei (A–C) clustered separately, suggesting that our FACS strategy is highly accurate, but not perfect, at separating nuclei from different cell types. To further characterize these nuclei, we measured the average expression levels of known marker genes for different brain cell types, on the basis of two studies that transcriptionally profiled pure cell populations in mouse (**Fig. 10c** and **Supplementary Table 4**). The remaining two clusters of predominantly NeuN$^+$ nuclei both showed high expression for neuronal markers, but they showed different levels of inhibitory and excitatory marker genes[36]. The remaining two clusters of NeuN$^-$ nuclei showed lower expression of neuronal markers but high expression of markers for specific glial populations[37]: astrocytes and oligodendrocyte precursor cells (**Fig. 10c** and **Supplementary Table 4**). Expression patterns of specific marker genes for these cell types confirm these cell type classifications (**Fig. 10d**). Overall, we found that the ten nuclei profiled by RNA-seq came from four distinct brain cell types.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS R.S.L. was funded by NIH-1RC1 HG005471, NIH-2R01 HG003647; R.S.L. and F.H.G. by Transformative R01MH095741; F.H.G. by the Jeffry M. and Barbara Picower Foundation, the Mathers Foundation, the McDonnell Foundation and the National Institute of Mental Health (NIMH); and M.J.M. by a Crick-Jacobs Junior Fellowship. K.B. is supported by the National Science Foundation Graduate Research Fellowship under grant no. DGE-1144086. The authors thank the Allen Institute for Brain Science founders, Paul G. Allen and Jody Allen, for their vision, encouragement and support. Research was supported in part by the Allen Institute for Brain Science.

AUTHOR CONTRIBUTIONS R.S.L., R.V.G., M.J.M., S.R.K., F.H.G., E.S.L. and J.L. contributed to design of the project. S.R.K., M.N., B.J.L., J.A.E., R.H. and J.K.M. carried out experiments. P.V., K.B., S.B.L., S.P., M.K., J.M., J.K.M., N.S., B.D.A., F.D.-F. and R.H.S. contributed to sequencing data analysis. R.S.L., R.V.G., S.R.K. and M.N. wrote the manuscript with contributions from all the authors.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
2. Kurimoto, K., Yabuta, Y., Ohinata, Y. & Saitou, M. Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis. *Nat. Protoc.* **2**, 739–752 (2007).
3. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
4. Ramskold, D. *et al.* Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
5. Tang, F. *et al.* RNA-seq analysis to capture the transcriptome landscape of a single cell. *Nat. Protoc.* **5**, 516–535 (2010).
6. Lovatt, D. *et al.* Transcriptome *in vivo* analysis (TIVA) of spatially defined single cells in live tissue. *Nat. Methods* **11**, 190–196 (2014).
7. Citri, A., Pang, Z.P., Sudhof, T.C., Wernig, M. & Malenka, R.C. Comprehensive qPCR profiling of gene expression in single neuronal cells. *Nat. Protoc.* **7**, 118–127 (2012).
8. Qiu, S. *et al.* Single-neuron RNA-seq: technical feasibility and reproducibility. *Front. Genet.* **3**, 124 (2012).
9. Lovatt, D., Bell, T. & Eberwine, J. Single-neuron isolation for RNA analysis using pipette capture and laser capture microdissection. *Cold Spring Harb. Protoc.* doi:10.1101/pdb.prot072439 (2015).
10. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. USA* **112**, 7285–7290 (2015).
11. Zeisel, A. *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
12. Huang, H.L. *et al.* Trypsin-induced proteome alteration during cell subculture in mammalian cells. *J. Biomed. Sci.* **17**, 36 (2010).
13. Grindberg, R.V. *et al.* RNA-sequencing from single nuclei. *Proc. Natl. Acad. Sci. USA* **110**, 19802–19807 (2013).
14. Barthelson, R.A., Lambert, G.M., Vanier, C., Lynch, R.M. & Galbraith, D.W. Comparison of the contributions of the nuclear and cytoplasmic compartments to global gene expression in human cells. *BMC Genomics* **8**, 340 (2007).
15. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149–1154 (2005).
16. Schwanekamp, J.A. *et al.* Genome-wide analyses show that nuclear and cytoplasmic RNA levels are differentially affected by dioxin. *Biochim. Biophys. Acta* **1759**, 388–402 (2006).
17. Trask, H.W. *et al.* Microarray analysis of cytoplasmic versus whole cell RNA reveals a considerable number of missed and false positive mRNAs. *RNA* **15**, 1917–1928 (2009).
18. Jiang, Y., Matevossian, A., Huang, H.S., Straubhaar, J. & Akbarian, S. Isolation of neuronal chromatin from brain tissue. *BMC Neurosci.* **9**, 42 (2008).
19. Birnie, G.D. Isolation of nuclei from animal cells in culture. *Methods Cell Biol.* **17**, 13–26 (1978).
20. Schroeder, A. *et al.* The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* **7**, 3 (2006).
21. Dounce, A.L., Witter, R.F., Monty, K.J., Pate, S. & Cottone, M.A. A method for isolating intact mitochondria and nuclei from the same homogenate, and the influence of mitochondrial destruction on the properties of cell nuclei. *J. Biophys. Biochem. Cytol.* **1**, 139–153 (1955).
22. Hymer, W.C. & Kuff, E.L. Isolation of nuclei from mammalian tissues through the use of Triton X-100. *J. Histochem. Cytochem.* **12**, 359–363 (1964).
23. Wu, A.R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* **11**, 41–46 (2014).

24. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
25. Macosko, E.Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
26. Usoskin, D. *et al.* Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* **18**, 145–153 (2015).
27. Rabani, M. *et al.* High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell* **159**, 1698–1710 (2014).
28. Lacar, B. *et al.* Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nat. Commun.* (in the press).
29. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
30. DeLuca, D.S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).
31. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
32. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *Bioinformatics Action* **17**, 2 (2013).
33. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
34. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
35. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
36. Sugino, K. *et al.* Molecular taxonomy of major neuronal classes in the adult mouse forebrain. *Nat. Neurosci.* **9**, 99–107 (2006).
37. 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).

Chapter 2.2, in full, is a reprint of the material as it appears in Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. 2016. Krishnaswami SR, Grindberg RV, Novotny M, Venepally P, Lacar B, Bhutani K, Linker SB, Pham S, Erwin JA, Miller JA, Hodge R, McCarthy JK, Kelder M, McCorrison J, Aevermann BD, Fuertes FD, Scheuermann RH, Lee J, Lein ES, Schork N, McConnell MJ, Gage FH, Lasken RS. 2016. The dissertation author was a primary investigator and author of this paper.

2.3. WHAT IS A NEURAL CELL TYPE?

The methods described in the previous chapter have yielded a vast array of biological insight, providing methods to target samples in many tissues. [1] The development of methods to account for metadata and remove bias-associated with noise has provided a novel tool to examine variability between individual neurons, the first requirement to interpreting inter-neuron signaling and cell lineage determination. By focusing on the frontal cortex of the human brain, we focus on tissue with known geographic association with memory formation and other traits that are fundamentally characteristic of humans as opposed to rodents and other small host species. These methods characterize single cell samples through transcriptomic assay (e.g. as selected by marker genes separating clustered transcriptomic abundances from > 1700 individually sequenced single nuclei from individual human neurons) and find groups of cells with unique marker protein expression that may be identified within the specific tissues they were isolated in, and may also unique features in their morphology (shape.) The methods by which high throughput sequencing and high throughput, high-content cytometry may be leveraged to rapidly identify cell types is defined as a statistical and ontological strategy in the following manuscript (analysis not shown).

Bakken T, Cowell L, Aevermann BD, Novotny M, Hodge R, Miller JA, Lee A, Chang I, McCorrison J, Pulendran B, Qian Y, Schork NJ, Lasken TS, Lein ES, and Scheuermann RH. **Cell type discovery and representation in the era of high-content single cell phenotyping.** Dec 21, 2017. *BMC Bioinformatics*. 2017; 18(Suppl 17): 559.s. doi: 10.1186/s12859-017-1977-1.

Perhaps the most exciting novel finding of this study is the identification of a rare cell types within tissue that had previously been poorly characterized in a whole cell context. Within one of these regions we targeted a cell type of unique morphology which appeared to match with a strongly correlated transcriptomic cluster in our analysis set. The resulting biomarkers matched cell staining assays in the tissue of interest and the resulting cells were further isolated for synaptic potentiation, or the evaluation of their correlated expression of current across axonic bounds. The resulting analysis describes this novel cell type, the 'rosehip' cell named for its rosehip-like axonal boutons, noting its unique characteristics, its potential role in complex functions in the brain associated with social and memory function in the *Homo sapiens*, and the way these methods can be applied for other cell typing assays subject to high scrutiny.

Boldog E, Bakken TE, Hodge RD, Novotny M, Aevermann BD, Baka J, Bordé S, Close JL, Diez-Fuertes F, Ding SL, Faragó N, Kocsis AK, Kovács B, Maltzer Z, McCorrison JM, Miller JA, Molnár G, Oláh G, Ozsvár A, Rózsa M, Shehata SI, Smith K, Sunkin SM, Tran DN, Venepally P, Wall A, Puskás LG, Barzó P, Steemers FJ, Schork NJ, Scheuermann RH, Lasken RS, Lein ES & Tamás G**. Transcriptomic and morphophysiological evidence for a specialized human cortical GABAergic cell type.** Aug 27, 2018. *Nature Neuroscience*. 21, pages1185–1195 (2018). doi: 10.1038/s41593-018-0205-2.

We have used the same methods, though with less canonical evidence, to identify other neural 'cell types', or hypothetical cell type targets of interest. Another notable example has been published, detailing the identification of subcerebreal excitatory neurons from our human samples which are shared in mice, the 'von Economo' neurons found in layer 5 of the fronto-insular cortex.

Hodge RD, Miller JA, Novotny M, Kalmbach BE, Ting JT, Bakken TE, Aevermann BD, Barkan ER, Berkowitz-Cerasano ML, Cobbs C, Diez-Fuertes F, Ding SL, McCorrison J, Schork NJ, Shehata SI, Smith KA, Sunkin SM, Tran DN , Venepally P, Yanny AM, Steemers FJ, Phillips JW, Bernard A, Koch C, Lasken RS, Scheuermann RH, Lein ES. **Transcriptomic evidence that von Economo neurons are regionally specialized extratelencephalic-projecting excitatory neurons.** *Nature Communications* 11, 1172 (2020). https://doi.org/10.1038/s41467-020-14952-3

While these analyses have become influential due to the advancements they have made possible in the immediate evaluation of previously-hidden signal, there is much room left to improve upon the high throughput nature of isolating and sequencing samples, interpreting their inter-sample distance and co-clustering of those samples, and making the best use of metadata captured throughout the data collection process.

There are two ways to approach recursive clustering once we are able to identify particular patterns of transcripts that are highly correlated with quality profiles. One is to normalize based on their separation in quality-profile-specific matrix component space (via UMAP/tSNE/PCA on their quality matrices.) The second is to normalize based on a quality profile label applied to their supervised or unsupervised cluster ID. In this study we describe 'bias modes', groups of quality control-predictive metadata components for each sample, which describing correlated trends in those metadta components' prediction of variance in sample transcript abundance. The goal of this study can be redefined as the development of a systematic approach for cataloging protocol-specific biases. In Chapter 3.2 I identified the most critical portions of an experimental protocol where the collection of unbiased data can be improved, much like when critical evolutionary breakpoints for lifespan are identified in a phylogenetic tree in the avian study (Chapter 2.2). Here I provide a recipe which can be used on any single cell data set with sample specific metadata to make informed decisions about the level of their unsupervised or supervised confidence of sample clustering based on sample transcriptomic variance.

2.4. Leveraging Biclustering Elucidates Variation in Sample Clustering Descriptive of Laboratory, Sequencing, and Informatics Biases

See un-published work, a draft currently being prepared for submission, reproduced in this chapter:

McCorrison J, Rangan A, Schork NJ. **Multifactorial Quality Control Analysis for Single Cell Transcriptomic Profiling.**

MULTIFACTORIAL QUALITY CONTROL ANALYSIS FOR SINGLE CELL TRANSCRIPTOMIC PROFILING

Jamison McCorrison B.S.[1], Aaditya Rangan, Ph.D.[2], Nicholas J. Schork, Ph.D.[3]

The University of California San Diego, La Jolla, CA (JM); Courant Institute of Mathematical Sciences, New York University; Flatiron Institute Center for Computational Mathematics, Department of Quantitative Medicine and Systems Biology, The Translational Genomics Research Institute (TGen; LHG, NJS), City of Hope/TGen IMPACT Center (NJS), and Departments of Psychiatry and Family Medicine and Public Health, University of California San Diego (NJS)

Address correspondence to:
Nicholas J. Schork, Ph.D.
J. Craig Venter Institute
La Jolla, CA 92075
nschork@jcvi.org

ABSTRACT

Single cell analyses are beginning to reveal how individual cell types contribute to the state of a tissue, any functional consequences it may exhibit, and transcription networks shared between host species. Complications in this type of single cell RNA-sequencing (scRNA-seq) assay can be complex since they are designed to interrogate the transcriptomes of thousands of individual cells at once, and can be overly sensitive to a wide range of laboratory conditions and settings in which the assays themselves are performed. We describe a comprehensive approach to imputing, normalizing, and comparing the clustering of single cell transcriptomic sequencing populations using biclustering, leveraging metadata from the quality control (QC) assessments of the laboratory, sequencing, and informatics processing. We also describe four methods for the cunsupervised or supervised evaluation of QC terms significantly correlated with sample- or cluster-specific transcriptomic variance. We illustrate the utility of sample-specific, transcript-level abundance normalization by optimizing our data formatting to replicate expectations of variation across the data set.. We use these methods to identify potential marker genes during cell typing and to delineate rare expression patterns at the exonic level in a set of human frontal cortex samples.

INTRODUCTION

As a result of isolating individual cells, gene expression profiles of individual cell types are revealed as opposed to the averaging of all transcriptomes obtained from localized bulk tissue. However, the identification of novel transcript abundances within individual human nuclei is complicated by the difficulty of localized isolation of targets. Identifying rare targets in tissue often breaks the cell cytoplasm during physical isolation individual cells with many intertangled axonic structures, the very physical components that allow for cell signaling between cells of various types. Error is also introduced from the measurement limitations when using the small quantities of proteins within just a single isolated cell.

Recent advances in automated high throughput handling of individual cells has included “sensitive, highly-multiplexed single cell RNA-seq” with SmartSeq2 [1]. By leveraging the isolation of the individual nuclei of human neuron samples, single nuclei single cell RNA-seq (sc-sn-RNA-seq) allows researchers to investigate the transcriptomic abundances observed specifically from individual cells isolated in eukaryotes. However, coverage biases and unexpected predictive correlations to non-exonic coverage events have been observed. For example, partially degraded RNA (e.g. from freezing, RNAse degradation) results in deeper sequencing coverage for the 3’ end of transcripts only when degraded products contain the polyA tail required for amplification.

Through the use of a random forest classifier trained on a human-driven qualitative assessment, the evaluation of a large collection of human neurons and their associated laboratory covariates was able to provide a predictive assessment of the binary pass/fail status of individual sample-preparation quality [2]. These methods have allowed for the elucidation of high-quality cell type classifications derived through transcriptomics and through downstream laboratory analysis, including the semi-supervised clustering methods described by Bakken et. al, which is the subject of the supervised analysis example in this manuscript [3].

The evaluation of a single cell preparation bias by leveraging quality control metadata and abundance values using silhouette conditions across TSNE-clustered nearest neighbors has been approached previously [4]. Visual interpretation of clusters of single cells grouped by transcriptomic abundance similarity is commonly performed, allowing analysts to pursue variations in biological data, and sample data, of further interest for research or omission. [5] Methods for rapid visualization of these silhouette scores in pre-clustered TSNE results have been previously published. [6] Statistical tools exist for the comparison of samples for preprocessing and gene selections, as well as unsupervised clustering and integrated analysis amongst resulting subsets of samples. [7] Newer tools are already available which leverage covariate-derived linear modeling to evaluate continuous or discrete trends between population-wide sample variation to enhance the predictability of single cell RNA seq analyses. [8]. Previous literature using this dataset described unexpected non-coding sequence coverage events which appeared within both intragenic and intronic regions of our single nuclei samples. [9] Under further investigation, these predictive contiguous non-random coverage events appeared to represent either: 1) true novel expression of a new transcript or transcript isoform, 2) non-coding sequences which will not survive replication but which remains predictive of cell type, 3) simple overlapping annotation error, 4) transcripts dropped or differing from the gene model, 5) simple mis-mapping and multi-mapping error, or 6) as tested below, problematic RNA capture resulting from an aberration in or consequence of an experimental lab and informatics protocol.

We test the correlation between variation in our quality control metrics in both exonic and intronic coverage events in the context of the GRch38 Human Reference Sequence model by RefSeq, providing a methodology by which large scale investigation of observed abundance and metadata in tandem may reveal systematic biases in the preparation of cell lines and the laboratory configurations of the preceding sample preparation pipelines themselves. Whereas many existing strategies are 'fully global' (ignoring cluster-information), they do not pinpoint variance between neighboring pairs of supervised or unsupervised clusters (e.g. potential cell types of known type, or undefined cell types of interest). [10] The degree to which the variation between the samples in any one cluster, and their transcript abundance vectors, is predicted by the transcript abundance or by covariate terms, or “bias modes” (correlated groups of covariate terms), is also not known.

We recommend a set of reproducible strategies which show the effects of biclustering transcript and quality metadata in a way that is not ‘fully’ local, since our determination of QC-clusters themselves involves transcript-data and QC-data, but which can shed light on certain local structure, without being lead astray by more ‘global’ population-wide trends.. (Figure 1) While such a method is underpowered when cluster-sizes are small, we show that leveraging localized information to identify covariates that are systematic drivers of cluster separation provides insight into the identification of novel clusters and potential marker genes. We suggest the latter as an ideal output for this type of analysis in the single cell context where laboratory methods for cell type validation (e.g. immunostaining) may help identify rare types in experimental target tissues. We focus on significant transcriptomic correlations lost in the noisy signal we observe in our experimental data and, conversely, the identification of false positives (e.g. to prioritize targets during costly laboratory tissue and layer specific targeting and confirmation trials.)

Our recipe stands upon 3 foundations which we show are critical to the evaluation of any single cell transcriptomic cell typing assay. We start with sample preparation; the imputation, normalization, and comparison of clusters must be performed in a way which attempts to maximize the recovery of lost signal while minimizing any imposed structure on our abundance matrices. Next, we identify statistically significant bias modes using both unsupervised techniques (e.g. only the sample transcript metadata and quality control metadata) and supervised techniques when possible (e.g. with base knowledge from prior study, cell staining, marker gene expression, etc.). Last, we leverage covariate-corrected unsupervised clustering of snRNA-seq data (single nucleus RNA-sequencing data) by using the covariates (metadata) to ‘correct’ the genetic data (transcript abundances) using linear modeling techniques and groups of correlated quality control metrics (‘bias modes’) predicting bias-associated variance (e.g. measured by co-clustering after correction.) We apply the proposed method to a large scRNA-seq (single cell RNA-sequencing) study of nuclear RNA harvested from neurons in different layers of the human brain and investigate the novel inferences that can be made by analyzing the expression inside and outside of the annotated coding regions of the human reference genome.

RESULTS

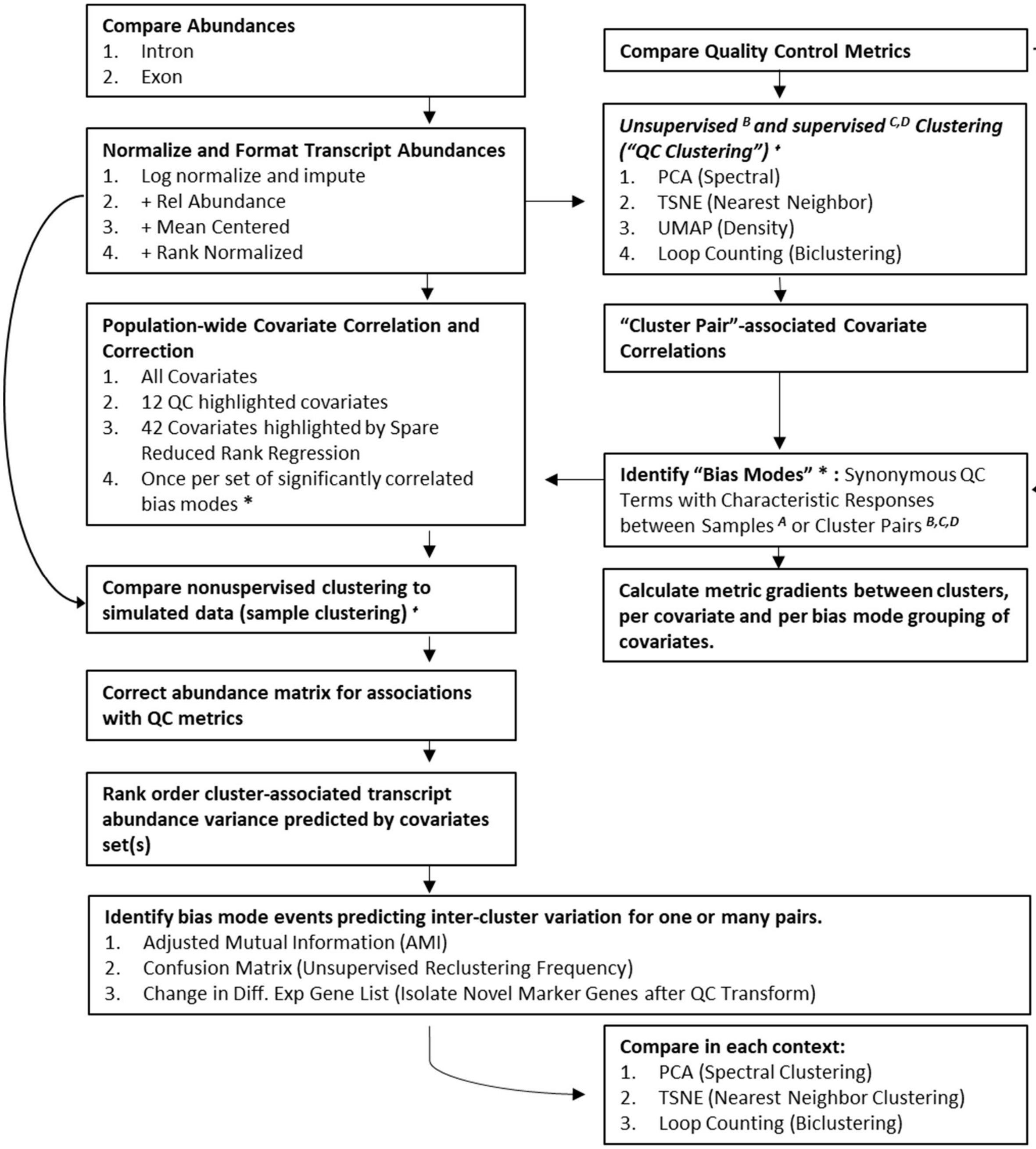
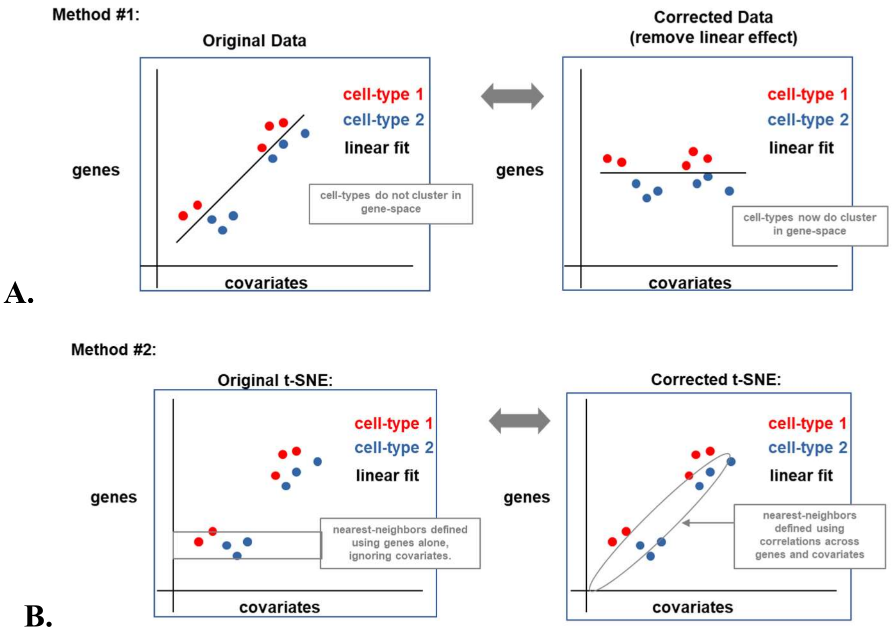
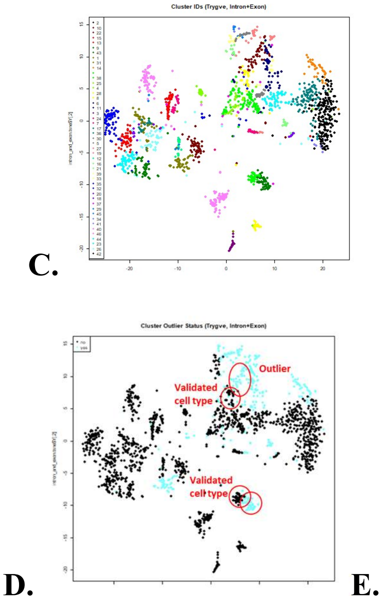

Compare Abundances
1. Intron
2. Exon
Normalize and Format Transcript Abundances
1. Log normalize and impute
2. + Rel Abundance
3. + Mean Centered
4. + Rank Normalized
Population-wide Covariate Correlation and Correction
1. All Covariates
2. 12 QC highlighted covariates
3. 42 Covariates highlighted by Spare Reduced Rank Regression
4. Once per set of significantly correlated bias modes *
Compare nonusupervised clustering to simulated data (sample clustering) *
Correct abundance matrix for associations with QC metrics
Rank order cluster-associated transcript abundance variance predicted by covariates set(s)
Compare Quality Control Metrics
Unsupervised B and supervised C,D Clustering ("QC Clustering") *
1. PCA (Spectral)
2. TSNE (Nearest Neighbor)
3. UMAP (Density)
4. Loop Counting (Biclustering)
"Cluster Pair"-associated Covariate Correlations
Identify "Bias Modes" * : Synonymous QC Terms with Characteristic Responses between Samples A or Cluster Pairs B,C,D
Calculate metric gradients between clusters, per covariate and per bias mode grouping of covariates.
Identify bias mode events predicting inter-cluster variation for one or many pairs.
1. Adjusted Mutual Information (AMI)
2. Confusion Matrix (Unsupervised Reclustering Frequency)
3. Change in Diff. Exp Gene List (Isolate Novel Marker Genes after QC Transform)
Compare in each context:
1. PCA (Spectral Clustering)
2. TSNE (Nearest Neighbor Clustering)
3. Loop Counting (Biclustering)

Figure 2.4.1. Analysis Flowgram (Unsupervised Clustering). Sub-experimental arms: A) Unsupervised (Sample x QC), e.g. you have clusters and you want to identify population-wide QC terms of relevance (within a relative monoculture.) **B)** Unsupervised ([Sample x QC] + [Sample x Transcript]), e.g. people need a recipe for unsupervised clustering with transcript-level interpretation of laboratory and sequencing bias. **C)** Supervised ([Sample x QC] + [Sample x Transcript] + Sample Cluster Assignments), e.g. you want to target hard-to find drug targets. **D)** Supervised (Sample x QC+ Sample Cluster Assignments): e.g. you have clusters and you want to identify QC terms of relevance to sampling/clustering quality to improve future studies.


Method #1:
Original Data
Corrected Data
(remove linear effect)
cell-type 1
cell-type 2
linear fit
genes
covariates
cell-types do not cluster in gene-space
cell-types now do cluster in gene-space
A.
Method #2:
Original t-SNE:
Corrected t-SNE:
nearest-neighbors defined using genes alone, ignoring covariates.
nearest-neighbors defined using correlations across genes and covariates
B.



E.

Quality profile similarity	Expression profile similarity	Hypothesis
↑ 2H	↑	Collapse clusters, or identify small expression outlier.
↑ 2L	↑	Highly correlated bias cases.
↑ 2H	↓	Supported cluster separation. * *Novel exp. pattern at high quality suggest possible novel cell type, transitory type, or uncaptured bias.*
↑ L	↓	Cluster separation explained by metrics (1) lab (2) seq (3) ifx (4) assigned type (5) intron/intergenic-specific coverage event
↓ L	↑	High/low quality cluster pair identified.
↓	↓	Model outlier

Figure 2.4.2. Spectral ordination and clustering of samples in transcriptomic space. A-B) A cartoon showing interpretations of transcriptomic and quality driven interactions. **C)** TSNE Clustering of Cell Abundances (Colors Shown, Top, Intron+Exon Abundance) and our Biological Insight on Outliers varying from Cell Typing Expectations. **D)** The denoted 'outlier' clusters highlighted due to unexpected transcriptomic marker performance vs. laboratory validation. **E)** A decision tree highlighting theoretical outcomes associated with biclustering correlation events of varied use for downstream biological inference. 'Quality profile similarity', 2H = both clusters are 'high quality', with small amount of variation predicted by quality terms associated with low quality outcomes. 2L = both clusters are 'low quality'. L = one cluster is 'low' quality. H = one cluster is 'low' quality'.

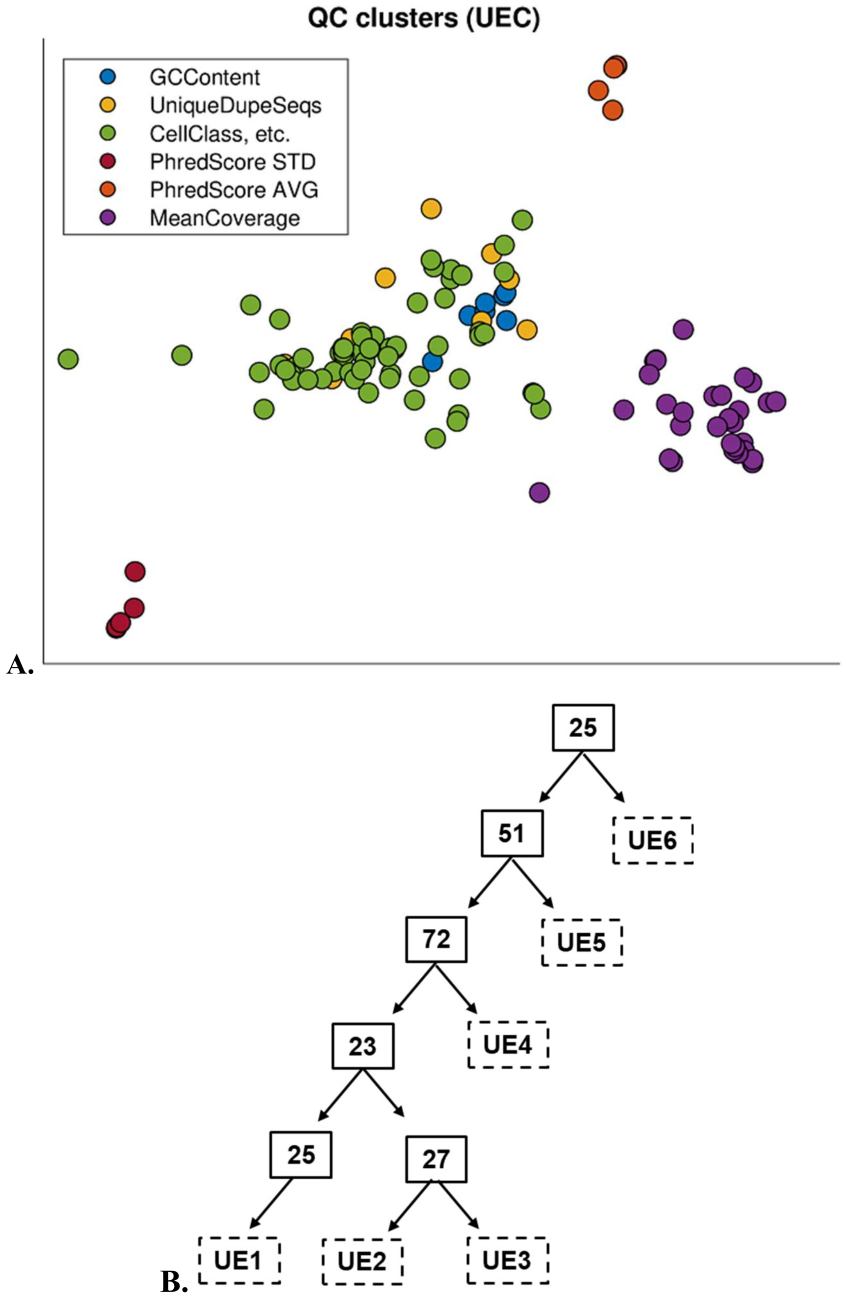


Figure 2.4.3. A) The 6 QC-clusters from Table 1 found after accounting for the exon-data. To start with, we project each of the 127 QCs (considered as a 1781-dimensional vector) onto the dominant 2 left-principal-components of the 1781-by-127 matrix [S^(-1)*U*C], as described in the text above. We then color and label the clusters using some of the terms that commonly appear amongst their QC-labels (note, however, that the clusters themselves were determined using only the QC-values along with the transcript-data). Note that, while these clusters do not necessarily look distinct when projected onto these 2 principal components, they exhibit distinct correlations (with a statistical significance < 0.0015) in the full space. **B.)** The pvalue tree for the QC-clusters found after accounting for the exon transcript data. The integer values shown are -log(p-value) for each split. The null-hypothesis involves 'right-spinning' the QC-matrix (as described in the text).

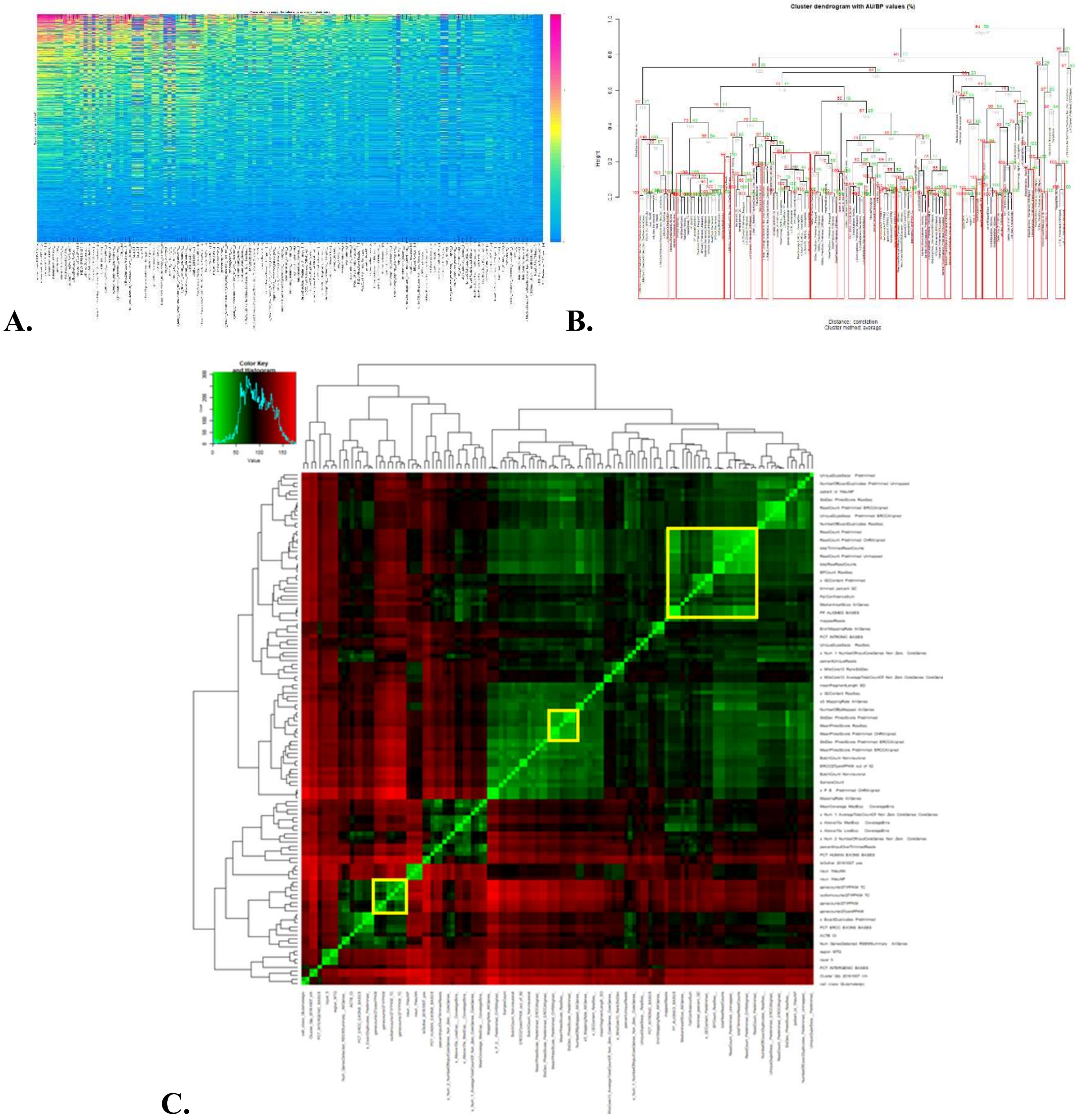


Figure 2.4.4. Linear modelling for investigative analysis of specific terms. A) Cluster Pair x Covariate Z-score matrix. **B)** Pvclust hierarchical clustering with selection of significant covariates subgroups ("bias modes") for which cluster pairs exhibit significantly correlated (au p-value > 0.99) response to those terms en masse. **C)** The euclidean distance matrix representing distance between covariates' cluster-pair Z-score matrix terms, illustrating covariates with correlated response predicting sub-classification of population-wide cluster pair relationships with 3 significant bias modes highlighted in yellow. **Top:** Cluster 6 with bootstrap AU p-value = 1 (< 10e-10) and Std Err = 0 representing "the overall sequencing abundance of comparable human gene sequence." **Middle:** PVClust Cluster 23 with bootstrap AU p-value = 0.998 and Std Err = 0.005 representing "Sequencer error (PHRED Score Variability)." **Bottom:** PVClust Cluster 10 with bootstrap AU p-value = 1 (< 10e-10) and Std Err = 0 representing "the number of comparable genes observed (e.g. the ability to detect less-expressed genes?)"

Table 2.4.1. Significantly Correlated "Bias Mode" Events. P(E)=Exon-derived p-value for the Experimental Arm. P(I)=Intron-derive p-value for the Experimental Arm. As in Figure 1: **A)** Unsupervised (Sample x QC), e.g. you have clusters and you want to identify population-wide QC terms of relevance (within a relative monoculture.) **B)** Unsupervised ([Sample x QC] + [Sample x Transcript]), e.g. people need a recipe for unsupervised clustering with transcript-level interpretation of laboratory and sequencing bias. **C)** Supervised ([Sample x QC] + [Sample x Transcript] + Sample Cluster Assignments), e.g. you want to target hard-to-find drug targets. **D)** Supervised (Sample x QC+ Sample Cluster Assignments): e.g. you have clusters and you want to identify QC terms of relevance to sampling/clustering quality to improve future studies. A full list of QC terms and their assignments is provided in Supplemental Table 1.

Cluster #1: "Cluster GC Percentage" p=10e-25
• GC Percentage: All sequences. • GC Percentage: Trimmed sequence (Only reads aligned to ERCCs, Only reads aligned to human reference, and Unmapped reads.)
Cluster #2: "Batch, Depth of Sample Capture" p=10e-27
• Batch • Iterative sample ID • Percent of reads trimmed • Mean fragment length (Value, Std. Deviation in Value) • Percent of all reads mapped to Human Reference • % of Unique Duplicate Sequences: All reads • % of Unique Duplicate Sequences: Trimmed reads (Only reads aligned to human reference, and unmapped reads.) • % of Mitochondrial Core Genes at greater than zero expression abundance. • Median Insert Size of Paired End sequences • Unique
Cluster #3: "Population-wide Predictors of Sample Quality" p=10e-27
• Cell class, based on marker genes (Excitatory, inhibitory, glia) • Outlier status, based on laboratory and transcriptomic validation [3] • Cell type, based on marker genes (GABAergic, Glutamergic, Non-neuronal) • Neun-positive sorting percentage (likelihood of gathering only nucleic content) • Brain region (Frontal insular cortex, Middle temporal gyrus) • Brain layer (1,5) • RandomForest Pass/Fail Confidence Score • cDNA PicoGreen Concentration (Quantity of double stranded DNA during protocol assay) [1] • Marker gene abundance (ACTB, Custom Set 1 [3], Custom set 2 [3], 13 neural mitochondrial marker genes) • ERCC Count of Ladder Sequence • Percentage of non-duplicate input reads (All Reads, Trimmed Reads, Trimmed Paired Ends) • Percentage of reads maintaining paired end relationships after trimming. • Number of, and percentage of, genes present (Greater than {0,1} FPKM) • Pecentage of isoforms present (Greater than {0,1} FPKM) • Pecentage of ERCC barcodes present (Greater than {0,1} FPKM) • Percentage of trimmed reads mapped to human reference • Percentage of trimmed reads mapped to human reference in each region type (Exons, Introns, Intergenic) • Percentage of reads in coverage bins (High, Medium) • Mapping rate (All genes, End 1, End 2, 3' end, 5' end)

Table 2.4.1 Significantly Correlated “Bias Mode” Events (Continued.)

Cluster 4: “Sequencing Quality (Across the length of the read)” p=10e-72
• Standard Deviation in Phred score (All reads, reads aligned to ERCC sequence, reads aligned to human reference, unmapped reads)
Cluster 5: “Sequencing Quality (Across the full span of the read)” p=10e-51
• Mean Phred score (All reads, reads aligned to ERCC sequence, reads aligned to human reference, unmapped reads)
Cluster 6: “Depth of Sequencing (including ERCC ladder sequence)” p=10e-25
• Total input reads (All, trimmed) • Number reads mapped to human reference • Total input bases (All, trimmed) • Number of duplicate reads before trimming • Mean coverage of expression bin (high, medium, low) • % of Samples above 15x in Low coverage bins • Count of non-zero abundance mitochondrial core genes • Number of reads and basepairs mapped to group (genes, ERCC ladder sequence, human reference sequence) • % of unmapped exact duplicate sequences.

Table 2.4.2. Cluster Confidence Matrix. Change in Signif. Diff. Exp. Clusters before and after correction using Z rank 1. The probability of observing our sample-cluster overlaps vs. Ho of baseline assumption of these overlaps in a random shuffling of sample labels (e.g. the 'repackaged probability' of seeing collective overlaps with AIBS clusters that are as rare or rarer than our observations. Compare to Suppolemental Table 3 usiing Z rank 2. Low = Poorly conserved clustering (e.g., the cluster fell apart.) High = Clustering converged toward centroid.

QC	None	All	SRRR	RF	UE1	UE2	UE3	UE4	UE5	UE6
# of Clusters	23	62	60	7	1	37	58	34	62	35
-logP: Population	616	447	1614	330	0	443	1473	947	1153	606
-logP: Per-cluster										
1	41	63	333	188	0	116	332	157	252	75
2	14	25	34	0	0	0	24	13	0	28
3	62	0	16	1	0	1	45	37	22	4
4	14	35	52	6	0	0	21	17	3	1
5	0	25	71	5	0	2	48	4	14	18
6	1	11	49	6	0	2	20	4	6	10
7	57	3	6	1	0	1	55	42	4	1
8	14	0	42	5	0	0	39	22	0	4
9	0	22	40	1	0	1	5	4	2	4
10	55	0	30	1	0	1	36	17	4	0
11	0	95	52	5	0	33	63	59	43	57
12	4	0	64	5	0	67	80	42	103	58
13	58	59	67	15	0	3	64	37	63	26
14	0	1	6	1	0	12	33	32	52	20
15	0	0	3	1	0	12	31	11	37	36
16	11	57	62	7	0	15	62	45	48	35
17	10	0	12	0	0	0	6	12	2	0
18	43	0	2	0	0	0	17	17	3	2
19	45	38	47	52	0	2	41	44	56	5
20	0	0	18	1	0	1	1	0	0	1
21	0	0	0	0	0	27	13	2	28	15
22	0	1	16	0	0	4	1	0	1	12
23	6	1	137	7	0	60	153	56	104	28
24	0	0	4	1	0	6	13	2	19	5
25	26	0	1	0	0	1	6	14	2	0
26	30	0	12	0	0	0	15	14	2	0
27	14	0	3	0	0	1	5	11	2	2
28	0	14	24	0	0	3	1	1	2	4
29	0	0	1	0	0	4	3	3	14	3
30	17	0	4	0	0	0	3	10	1	0
31	0	0	0	0	0	5	1	1	12	2
32	10	0	1	0	0	1	3	3	1	0
33	5	0	0	1	0	3	1	2	3	3

Table 2.4.2. Cluster Confidence Matrix (Continued.)

QC	None	All	SRRR	RF	UE1	UE2	UE3	UE4	UE5	UE6
-logP: Per-cluster										
34	4	80	134	12	0	3	51	30	31	27
35	0	2	3	0	0	1	0	3	0	0
36	0	0	0	0	0	4	2	0	9	1
37	11	1	16	10	0	0	4	13	0	1
38	2	0	0	0	0	0	3	2	0	0
39	0	0	3	0	0	4	2	0	1	5
40	0	2	2	9	0	0	3	3	3	0
41	1	0	0	0	0	0	2	0	0	0
42	117	79	170	14	0	61	188	139	203	67
43	5	1	90	9	0	6	24	6	6	8
44	1	0	48	1	0	25	46	19	70	30
45	12	52	101	2	0	0	23	24	6	9
46	3	20	53	0	0	1	12	12	14	10
47	0	0	0	0	0	1	0	0	0	0

Table 2.4.3. Variation in Cluster-specific Markers.

(Not included in this draft.)

DISCUSSION

Correlation amongst Representations of Potential Transcriptomic Cell Types. The ability to sequence individual cells is contributing to a revolution in the understanding of bacterial cell types that cannot be cultured and therefore previously could not be amplified to sufficient protein abundances for sequencing. [11] Single cell amplification is commonly used when low biomass environmental samples are collected, exponential variation in coverage is inherent to single cell amplification protocols. [12] The severity and contribution of this bias to resulting informatics analysis is partially understood but normalization methodologies have been limited to the context of reference-free bacterial assembly. [13,14,15]

Human frontal cortex sampling and cell type diversity. For this study, cells of the central nervous system were gathered specifically because they were under-studied, partly because of the difficulty of isolating intact whole cells and successfully interpreting the minute signals within. This defines some the expectations for variation within our cohort. Excitatory cells are known to be higher in overall transcription, and thus their transcriptomic reads will be more easily sampled. [3] This increases the likelihood of capturing lesser expressed abundance trends associated with rare and difficult to study cell types, compared to Inhibitory and Glial cell types. Using the normalized abundances for the intron- and exon-specific abundances for each sample versus the human reference, we were able to recluster the original AIBS clusters in 'nearest neighbor' space using TSNE, approximating the clusters identified in that study using the intron and exon abundances together (Figure 2c, Table 2, Supplemental Figure 1). By exploring intronic coverage significance in cell typing before and after quality control correction, we admit that there are flaws in the upstream laboratory processing pipeline which can cause false positives or false negatives during cell typing, and attempt to make sense of novel (e.g. previously undefined in RefSeq) human expression events in the rare isolated single cells of the human brain.

Sample Prep: Imputation. When considering data from differing cell types, we must consider that we inherently lose signal from lesser expressing cell types (e.g. inhibitory and glial cell types) in comparison to their higher expressing, more-easily studied counterparts. To identify the rare variations representing potential novel signal patterns in the noise, we use imputation to account for lost signal from drop out due to biological or laboratory signature loss. Since we are preforming a subsequent analysis which is influenced by correlations between imputed terms, we also ensure that our imputation does not introduce significant spurious structure into the dataset. We approach this task by noting that the principal-components of the original data (i.e., without imputation) capture a certain amount of variance; following imputation, the variance captured by the same number of principal-components should be similar.

Sample Prep: Normalization and comparison. We use mean-centering after log-normalization when searching for clusters which are strongly related to the euclidean distance because that process hinges on correlations. However, when later measuring the amount of variance predicted (Z-scores) for each transcript across each AIBS-cluster-pair (in the context of determining candidate marker-genes for each AIBS-cluster-pair) we use the rank-normalized transcript data (since we calculate our Z-scores based on comparing the AUC for each transcript across each AIBS-cluster-pair with the distribution of analogous AUCs obtained under a permutation-test.

Population-wide quality control metric correlation. For each quality metric, the direction of association with transcript abundance was provided based on insights from the preliminary studies, and the feedback of the team that developed the single cell protocol. 12 of the 127 quality metric used in the study were highlighted by a human-trained Random Forest implementation predicting of pass/fail status [2]. Sparse reduced rank regression highlighted a different subset of 12 quality metrics as being significantly predictive. (Supplemental Table 1) Previous observations noted 3’ bias as an easily measured indicator of biases caused by a failure to capture high quantities of full length RNA during sequencing, making an apparent fail case identifier in prior RandomForest modelling [2,3]. Biases associated with the length of mRNA can be captured as a function of other quality terms (e.g. dealing with BP counts aligned, 3’ bias scores, and, most clearly, bioAnalyzer trace results). The length-associated observations were compared to expectations in random re-sampling and no strong correlation was identified between the lengths of our genes and the covariates at the population level. (Supplemental Figure 2)

As a comparison to common population-wide methods by Finak et. al, we also conducted a test of bimodality across samples using MAST, as described, and noting only a minority of genes display a strong bimodality across samples. [8] The vast majority of transcripts displayed either a unimodal or monotonic distribution across samples. Only in rare cases (less that 5%) does the distribution look like a gaussian plus a spike, or like a sum of 2 gaussians.

We hypothesize that bimodal relationships to quality control metrics, or groups of quality control metrics, exist not only across the population, for metrics which define variation that effects all samples (e.g. sequencing quality score), but also among subsets of the graph defined by cluster-cluster interactions. We decided to take our novel approach focusing both on an unsupervised interpretation of the data, and a comparison to an analysis using pre-defined or unsupervised sample clustering. We hypothesize that we should expect more than bimodality in the data clustered by transcriptomic abundance clustering, or through biological validation, because we evaluate interactions between many highly differing (e.g. performing different functions across tissues and layer depths of sampling), and highly similar (e.,g, performing similar functions within tissues and layers, or across them) cell types. We believe that studying the degree of cluster separation defined by clear relationships to our covariates can provide cost effective methods for detecting rare signals in lowly expressed genes and avoid false positives resulting in mischaracterization during transcriptomic cell typing. (Supplemental Figure 3)

Accounting for cell type variability in a meta-cell typing transcriptomic assay. As described above, high and low expression, and drop out of any expressed signal for a gene, are expected to exist not only as a result of cell class (e.g. exhibitory versus inhibitory) but also due to implicit biases of sampling (e.g. ability to capture unsheared RNA from a frozen cell). Here we describe application of our methods to a single nuclei single neuron data set utilizing the previous AIBS clustering methods, 47 ground-truth sample-clusters, and prior knowledge from the study describing those clusters. (Figure 2c) This prior knowledge included specific clusters of interest, labelled as 'outliers', due to their proximity to well defined clusters with known cell markers in TSNE nearest neighbor space, while lacking the same (or sufficient) characteristic markers suitable for staining or other forms of cell type validation. (Figure 2d) The goal of neural studies attempts to describe inter-tissue signaling and variation, and quiescence. Hand-curated evaluation of some clusters of interest revealed relationships that separated outlier clusters with their most significant metadata correlations matching either RandomForest-defined pass/fail terms, or, by contrast, tissue-specific markers (Supplemental Figure 7a-b). We experimented with several different unsupervised clustering algorithms, including schemes based on simple spectral clustering, t-sne and umap, as well as a scheme based on 'loop-counting' [16] which is similar to message-passing [17], spectral clustering [18] and the 'large-average-submatrix' method of [19]. We compared the performance of these methods on the 'planted bicluster problem', and found that some methods are more sensitive than others. (Supplemental Figure 5)

Subsets of Metadata Terms with Correlated Cluster-Pair-specific or Sample-Pair specific Response. Within this paradigm we employ two 'unsupervised' methodologies to look for QC-clusters. First (a) we directly look for QC-clusters within the sample-by-QC-matrix. Second (b) we project and rescale the sample-by-QC-matrix onto the left-principal components of the sample-by-transcript matrix before searching for clusters. Note that method-(a) attempts to find clusters of redundant QCs that are ‘universal’, in the sense that they would apply to any and all analyses (regardless of the transcripts). By contrast, method-(b) attempts to find clusters of QCs that are redundant within the context of the transcripts – these QC-clusters will apply only when considering that particular set of transcripts (e.g., Exons), and would not generalize to other sets of transcripts (e.g., Introns) (Supplemental Figure 9). To summarize our analysis, we found no statistically significant ‘universal’ QC-clusters using method-(a), but we did find 6 strongly significant QC-clusters with respect to the Exon-transcripts using method-(b). We noted that the 1000 sampled transcripts are strongly correlated with one another across the samples, while the QCs are not so strongly clustered in the sample-space.

Unsupervised Bias Modes. We made use of 2 unsupervised methodologies, discussed in the methods as methods A-D. Our first unsupervised method, described in the methods as “A”, is completely unsupervised: A QC-cluster found in this context would represent a subset of QCs that are correlated across all (or a significant subset) of samples. We don't find any statically significant QC-clusters in this context. In our second unsupervised method, described in the methods as “B”, a QC-cluster represents a subset of QCs that are correlated in a subset of samples, relative to the distribution of transcripts across those same samples. We do find statistically significant QC-clusters in this context (when we use the Exon transcript data). (Figure 4a-b) The significant QC clusters found in this case, clusters where the p-value was lower than 0.0015, are described in Table 1, where items in parentheses refer to multiple filters on the same QC term from various stages of the pre-processing stages. The orientation of these qc terms within the projection of each of the 127 QCs (considered as a 1781-dimensional vector) onto the dominant 2 left-principal-components of the 1781-by-127 matrix [S^(-1)*U’*C] is provided in Figure 3.

Supervised Bias Modes. We also considered two 'supervised' methodologies for clustering the QCs which make use of the ground-truth sample-clusters mentioned above. In methodology-(c): we generalize our unsupervised methodology-(b) as follows: Once again we project the QC-data onto the principal components of the transcript-matrix, except this time instead of using the standard principal-components of the transcript-matrix, we use 'sample-cluster-supervised' principal-components instead. To do this, we determine 'supervised' principal components of the transcript-data by finding directions in sample-space which optimize a cost-function which rewards (i) high inter-cluster distances and (ii) low intra-cluster distances. As this cost-function requires a single parameter defining the ratio between terms (i) and (ii), we scan over this ratio, searching for the 'best' supervised principal-components (i.e., those which produce the most statistically significant QC-clusters). Despite this exhaustive search, we were unable to locate any statistically significant QC-clusters using this method (for the Exon data), though relationships were found using simple linear relationships (Figure 5d-h). In methodology-(d), we search for QC-clusters using only the ground-truth AIBS sample-clusters, without considering the transcript-data. This amounts to searching for QC-clusters within the Z-score matrix which records the level of differential expression of each QC-term across each sample-cluster-pair. (Supplemental Figure 6) We were also unable to find statistically significant clusters using this method.

Practical investigation using linear modelling in subsets of cluster-pair interactions. Results from the linear model comparison of unsupervised terms identified comparable bias mode clusters, but the analyses were not significant. (Figure 3b-c) The results before and after correction were compared to present the effects of bias terms in providing false correlation across samples. (Supplemental Figure 12)

Correcting abundance vectors derived from all QC terms, and subsets of QC terms. Given the imputed, clustered abundance matrix, and bias modes from supervised and unsupervised contexts, we proceeded with the hypothesis that "correcting" our sample-specific abundance matrix for transcript-level correlations with quality metrics would identify alternative composition of our samples in some component space, and unsupervised clusters derived from their adjacency. Such a method differs from population wide searches for bimodal separations in factored or continuous metadata variables by leveraging the expected transcriptomic correlations due to the meta-analysis of many diverse cell types. [3,8] We show that when we correct for covariate-derived noise, we may observe an uneven distribution of effects across cluster-pair relationships, providing insight into which correlated transcriptomic abundance events have a higher probability of being a true signal (potential novel type) as opposed to a marker of a cell type whose representation has been bifurcated (or separated moreso) by, e.g. batch or preparation biases (Figure 2a-b, Supplemental Figure 3). While we utilized a dynamic selection of columns to incorporate from our Ʒ cluster-pair-covariate rank matrix, all observed analyses reduced to a single column component. We believe this satisfies our requirement that our normalization corrects our abundance matrix, given the influence of our covariates, while maintaining the variance described by samples' cluster centroids and their outcomes. In other worse, we ensure that inter-cluster variance is transcriptomic noise is minimized so metadata terms with shared effects on clustered samples should be a non-true signal captured by 1 rank.

Variation in the experimental cohort before and after correction exceeds statistical random sampling. Summaries of clustering confusion matrix results using linear models provided insights into variability related to individual QC terms, and set of QC terms which were not necessarily clustered together in a significant way, suggesting that such methods could be used to identify correlations with any group of metadata – even if the results will be handled noting their lesser significance. The resulting cluster terms widely matched expectations from our biclustering unsupervised methods, but with greater granularity in the separation of our terms. (Tables 2-3). Our experimental methods for linear model-based analysis of QC term subsets revealed wide differences when evaluating an experimental 'bias mode' list revealed vast difference in the ability to cluster samples before and after accounting for variance specifically associated with those terms, revealing greater granularity in separation across the component space defined in nearest neighbor space. (Figure 5d-h)

Conclusion. In summary, we present a novel recipe for the biclustering of single cell transcriptomic data profiling a variety of cell types, and exhibit the function and utility for future research on an example data set sampled from the individual nuclei isolated from individual neurons in the human frontal cortex. Our recipe stands upon 3 foundations critical to unsupervised biclustering for a single cell transcriptomic cell typing assay with informed insight on laboratory and sequencing bias: abundance table imputation, unsupervised clustering, and covariate-correction. We further summarize the effects of data normalization on variation in the matrix. We believe that future work related to biclustering in this field holds great potential as the ability to sequence individual cells is rapidly contributing to clinical advancements and our understanding of neural cell types at an individual level. This research may help accelerate the understanding of neural development, network regeneration, and memory formation within the human brain while also providing insight into the laboratory methods themselves. It is our hope that, as shown in prior study, critical evaluation of metadata from precursor stages of a protocol will be leveraged more in the future, providing insight into canonical biases. Furthermore, through critical evaluation of these biases, and the specific genes targeted, it is our hope that methods will be refined to help identify even more difficult to target individual cell types.

METHODS

Sample Information. All samples were prepared following the protocol described by Lasken et. al in Nature Methods. [3] Samples, captured from 2 post-mortem human frontal cortexes following cardiac arrest, were cut into slabs for study and stored frozen for approximately 2 years. In short, samples were selected based on the layer and region of the brain, near regions highlighted for containing unusually shaped cell types. After thawing samples, cells were selected based on RIN count and subjected to FACS sorting for nucleus isolation and a SmartSeq2-based amplification protocol to generate a Nextera library for sequencing.

Quantification and Normalization. Illumina 2500 paired end sequencing reads were trimmed to remove low quality sequence and primer contamination using Trimmomatic [20]. The resulting reads were sent to FastQC for read based quality control metric calculation. [21] Reads were aligned to 3 versions of the GrCH38 reference: the version derived from the genes' exon coordinates, another from the intronic coordinates, and a third using the entire gene bounds. Resulting counts were normalized using RSEM to Transcripts per Million (TPM). The resulting alignments were submitted to RNAseqQC and resulting metrics stored for downstream analysis. [22] Additional metrics were curated based on laboratory and biological insights. We utilized prior knowledge about our samples in the form of multiple insights. Cluster associations and outlier associations were derived from in depth recursive cell typing and associated marker gene identification and [3] Random Forest confidence of pass fail was incorporated from the analysis described using the same data set. [2]

Sample Prep: Imputation and normalization. Abundance tables were curated in intron and exon specific contexts. Genes with > 90% missing data are removed. Non-zero reads were log-normalized, producing a data-matrix ‘D’ with missing values. To fill in the missing values of D we apply an svd-based imputation-scheme. For the purposes of notation, let the index-set “S” correspond to the missing entries of D. Thus, D(S) is currently undefined.

The first step is to choose a dimension ‘d’ which refers to the number of principal components we will use to impose structure on the missing data in D. By comparing the spectrum of D with that of a random-matrix (i.e., the marchenko-pastur distribution) we see that the top 16 principal components of D are large. Thus, we set d=16 for the following process.

The first phase of our algorithm is to use the first d-principal-components of D to recover the structure of D(S). We begin by creating a (temporary) matrix E by first copying D, and then filling in each missing entry with randomly drawn values from the same column (i.e., each entry of E(S) is filled in by using from the non-missing samples associated with the same transcript). We calculate the dominant d principal-components of E, and then use these principal-components to construct a d-rank approximation to E, denoted by F. We then look at the entries of F which correspond to the missing entries of D (i.e., the entries of F(S)) and update E by replacing E(S) with F(S). We then return to the calculation of the dominant d principal-components of E, iterating until E converges.

Once E has converged, we have finished the first phase of our imputation. The missing entries in E(S) have been filled in a manner consistent with the dominant d-dimensional structure of E itself. However, at this stage the entries in E(S) are usually ‘too correlated’. That is, they exhibit an artificially high level of correlation which is not exhibited by the non-missing entries of E.

To correct for this artificially high correlation, we first calculate the singular-value-decomposition U*S*V' of E. The matrix V will be N-by-M (where N is the number of transcripts, and M is the number of samples). We then 'spin' the principal vectors of V corresponding to dimensions d+1, d+2, etc. This is done simply by replacing the final (M-d) columns of V with a random orthonormal set of vectors drawn from the same span (i.e., ensuring that they are perpendicular to the first d columns of V). We then construct the d-rank 'leading' approximation to E by using the first d-principal components of E. We'll denote this leading approximation by F1. We also construct the (M-d)-rank 'trailing' approximation to the residual of F1 by using the final (M-d) principal components of E, replacing the usual matrix V with the randomly oriented 'spun' version of V produced above. We'll denote this trailing approximation by F2. We then produce a surrogate matrix G by adding together alpha*F1 + beta*F2/p, where p is the square-root of the fraction of D that is filled – i.e,. p=sqrt(M*N-|S|). We then randomly permute the rows of G (corresponding to shuffling the samples) producing a matrix H. We then pretend as though H(S) is missing, and impute the values of H(S) in the same way that we first imputed the values of E(S). Finally, we measure the principal-values of H, and compare them to the principal-components of E that we observed during our first pass. We optimize alpha and beta so that the l2-norm of the difference in principal-values observed in the principle-values of H and E is as small as possible. Once we have found the optimal alpha and beta, we define J = alpha*F1 + beta*F2/p to be our imputed version of D.

We have designed this algorithm so that it functions well when presented with a large random matrix with a single 'spike' [23] that has been 'perforated' at random. In this case the optimal alpha and beta are both equal to 1. The more strongly the data deviates from the spike-model, the farther away from 1 we expect alpha and beta to be. For both our Exon and Intron data-sets the optimal alpha and beta were quite close to 1 (i.e. between 0.9 and 1.1). In addition, the principal-components of our surrogate H are very close to those of E. Together, these metrics indicate that our imputation algorithm has successfully captured the d-dimensional structure of the missing data without introducing spurious correlations.

The resulting imputed abundance matrix was converted to relative abundances (by dividing by the total), and using the centered log normalized. We considered log-centering after normalization, to ensure that each column has 0-mean, but we decided this this was not necessary since we automatically mean-center when clustering, and mean-centering doesn't affect the results when we rank-normalize for downstream marker-gene analysis.

Sample Prep: Comparison. We compared the AIBS clusters to the unsupervised clustering of the Exon data (both pre- and post-covariate-correction) by evaluating their linear residual and gathered the negative of the log of the p-value, rounded to the nearest integer, in Table 2. Roughly speaking, anything above 500 or so is very good, and anything above 1000 or so means that the AIBS cluster was mostly recapitulated by the unsupervised clustering. (Table 2, Supplemental Figure 11) After applying our unsupervised approach, we observed sample-clusters that – overall – coincided rather strongly with those of the AIBS-sample-clusters. This consistency reinforces the validity of both the ground-truth labels, as well as our unsupervised methodology. In terms of unsupervised clustering algorithms, we have included results comparing six methodologies.

First, we evaluate the 'half-loop' method described in our prior research by Rangan et al. [16]. This is an iterative method similar to message-passing [17], spectral clustering [18], and the 'large-average-submatrix' method of [19]. While this method allows for several internal approximations, such as binarization, to 'cut corners' and speed up the computation, we ran this algorithm in its ‘exact’ mode, with no approximations used. Consequently, this method has no free parameters.

Second, we leverage principal-component projection followed by ‘isosplit5’. [24] This method involves first projecting the samples onto the first ‘n_rank’ left-principal-components (note that ‘n_rank’ is a parameter we must specify), and then applying isosplit5 () with the default parameters (i.e., ‘K_init=200’ and ‘isocut_threshold=1.0’). We applied this method for n_rank ranging from 1 to 6.

Third, ‘exact’ t-sne, followed by isosplit5: This method involves first using the ‘exact’ mode of ‘fast_tsne’ [25] with either ‘n_rank=1’ or ‘n_rank=2’, and then using isosplit5 (as in #2 above) to cluster the resulting arrangement of points. We applied this method for n_rank=1,2. Fourth, we use ‘fast’ t-sne followed by isosplit5: This is equivalent to method three with the exception that we use the option ‘theta=0.5’ in fast-tsne, corresponding to the default ‘fast’ approximation. Fifth, we leverage umap, followed by isosplit5: This method involves first using umap (with default parameters), followed by isosplit5 (as in #2 above). [26] Last we leverage umap, followed by hdbscan: This method involves first using umap (with default parameters), followed by hdbscan (with ‘minpts=10’, and the remaining parameters set to default values). [27]

Subsets of Metadata Terms with Correlated Cluster-Pair-specific or Sample-Pair specific Response. For the purpose of this manuscript we define the term “bias mode” to mean any grouping of metadata (qc terms) for which there is a significantly correlated response across a subset of the population, as represented by sample-sample interactions (completely supervised [what does this mean?]). The QC’s themselves are quite correlated – for example most of the variance of the 1781-by-127 sample-by-QC matrix (which doesn’t include transcripts) is captured by the first 14-15 principal-components (specifically, 62% of the variance is captured by the first 15 components). By this measurement, in comparison to a random matrix, the QC-matrix is astoundingly correlated. To identify our 'bias modes', or QC-clusters, we are specifically looking for subsets of the 127 QCs that are more correlated than one would expect within this 14-15 dimensional representation of variation across the population. In other words, we are looking for clusters of 'redundant' QCs that are more correlated than 'chance', given the observed correlations across all the QCs. The null hypothesis we use is modeled by drawing a random set of 127 vectors with the same principal-components as the original QC-matrix. We can easily draw a trial from this null-hypothesis by 'right spinning' the sample-by-QC-matrix: i.e., by right-multiplying the 1781-by-127 sample-by-QC-matrix by a random 127-by-127 orthonormal matrix. Then, we check to see how strongly clustered those 127 random vectors are. What we are specifically looking for are QC-clusters within the original sample-by-QC-matrix that are more strongly correlated than the typical QC-clusters found within the randomly spun data.

We decorrelated the transcripts by first calculate the singular-value-decomposition U*S*V' of the 2-by-1000 transcript-matrix. Then we left-multiply both the 2-by-1000 transcript-matrix as well as the 2-by-127 QC-matrix by S^(-1)*U'. This is equivalent to the commonly used 'mahalanobis' rescaling. [28] With this rescaling the transcript-data is uncorrelated (bottom left subplot), while the QC-data is now strongly clustered (bottom-right subplot). Note that the QC-clusters are more evident *after* rescaling the sample-space to 'correct' for the correlations across the transcripts. Our methodology-(a) corresponds to trying to find QC-clusters in the top-right subplot of Supplemental Figure 9, whereas methodology-(b) corresponds to clustering the bottom-right subplot.

Population-wide quality control metric correlation. To assess the influence of various covariates on a specified set of sample-clusters, we calculate the Z-score matrix ‘Z’, which is of size [number of covariates] -by- [number of sample-cluster-pairs]. Given a particular covariate ‘j’ and a particular sample-cluster-pair ‘(k1,k2)’, the value of Z(j,k1,k2) is obtained as follows. First we measure the AUC associated with covariate j between sample-cluster-pairs k1 and k2.Then we assess the statistical significance of this AUC by calculating the AUC for a large number of label-permuted trials, and then estimating the Z-score of the original AUC from step one with respect to the mean and standard-deviation of the distribution of AUCs under the null-hypothesis in the AUC label-permuted trials. Note that, when z>0, the one-sided p-value associated with the z-score from step can be simply calculated as: log(p) = log(0.5) + erfcln(z/sqrt(2)).

Unsupervised Bias Modes. Our evaluation of clusters of quality terms with correlated effects across subsets of our abundance matrix took on two unsupervised methods. First, “A”, the completely unsupervised look at the QC-matrix alone, which is of size #-samples by #-QCs. We try and cluster this matrix which does *not* consider either the transcript data or any sample-clustering. A cluster found in this context would represent a subset of QCs that are correlated across all (or a significant subset) of samples. Second, “B”, the unsupervised use of transcript-data, looking at the QC-matrix which is of size #-samples by #-QCs as well as the transcript matrix, which is of size #-samples by #-Genes/Reads. We then project the QCs onto the principal components defined by the transcripts, and then try and cluster the resulting QC-projections. Due to rank disparity, a naive linear model was not ideal for this implementation since we fit the QCs perfectly, and found no residual.

Supervised Bias Modes. Our evaluation of clusters of quality terms with correlated effects across subsets of our abundance matrix also considered two supervised methods C: semi-supervised: use transcript-data as well as ground-truth sample-clusters (i.e., the AIBS-clusters). First define (supervised) principal components of the transcripts which best separate the ground-truth sample-clusters. And then project the QCs onto those principal components, and then (finally) try and cluster the resulting QC-projections. Supervised Bias Modes. D: fully-supervised: use only the ground-truth sample-clusters (but not the transcript-data) to define a sample-cluster-by-QC matrix. Then search for QC-clusters in that matrix.

Practical investigation using linear modelling in subsets of cluster-pair interactions. Visual interpretation of quality metric associations and cluster pair interactions is provided as a means of understanding simple linear relationships identified between cluster pairs and quality-associated bias metrics. The matrix of Z values for each cluster pair's relationships with each quality metric (e.g. piece of metadata) was alternatively evaluated by looking at the Euclidean distance measured across all terms. The resulting metric-metric distance values, defining the level of correlation in linear prediction of variance across all cluster pairs, were clustered hierarchically with pvclust to identify the most significant groupings of variables associated with these linear interactions. [29] (Figure 3b-c) The recursive bootstrapping analyses uses in this simple linear method are not memory efficient, and becomes computationally intractable when evaluating non-supervised 'effect groups' of correlated cluster pairs with similar response to all, or some, of our metadata terms. (Supplemental Figure 12a-b) These 'effect groups' were also calculated, again using pvclust but over a subset of example metadata components, to illustrate how to identify clusters with the greatest level of their inter-cluster relationships defined by a particular 'bias mode'. Clusters within 'effect groups' had counts added to their 'effect histogram' and the total scores were attributed to individual samples, and their clusters, to note the clusters which canonically perform in a similar fashion with regards to these metadata terms. (Supplemental Figure 12c-d).

Correcting abundance vectors derived from all QC terms, and subsets of QC terms. For clarity, we'll use our Exon data as an example to illustrate our method for correcting for covariates. In this case the data-set involves M=1781 samples, N=15137 exonic transcripts, and L=127 covariates. The data-matrices involved are the M-by-N (imputed) transcript matrix 'A', as well as the M-by-L rank-normalized covariate matrix 'C'. For notational purposes, we will also use '1' to refer to the constant M-by-1 vector of all ones.

We first use a version of linear regression to solve for the (1+M)-by-N coefficient matrix z_J, such that A is approximately equal to [1 C]*z_J. The regression we use is referred to as 'reduced rank regression' and produces a sequence of (1+M)-by-1 vectors 'u_j' and N-by-1 vectors 'v_j' for j=1,2,…,(1+L). [30] At each step J in this sequence, the vectors u_j and v_j are chosen to minimize the frobenius-norm of the difference (A – [1 C]($S_{j=1:J}$ u_j *v'_j)). Given the sequence of vectors from step #1a, the full z matrix is formed by summing over j: $z_J = S_{j=1:J}$ u_j * v'_j. Note that if [1+C] is full rank (i.e., rank 1+L) then the vectors u_j will be chosen such that [1 C]*($S_{j=1:J}$ u_j) is equal to U(:,1:J)*S(1:J,1:J), where U*S*V' is a singular-value-decomposition of A. Furthermore, as long as [1 C] is full rank then the vectors v_j will coincide with the columns of V. If [1 C] is rank-deficient then these equalities will not hold. Now, for any rank J, we can calculate the M-by-N residual matrix R_J = A – [1 C]*z_J. We use the residual R_J as a covariate-corrected version of the transcript-matrix A. By varying J, we can increase the level to which we correct for the covariates (note that R_0 = A). For our summary table we choose the rank J to be 'full' – i.e., determined by the rank of the covariate-matrix [1 C]. For the example given here (i.e., where C consists of all L=127 covariates) we choose J=12. For different populations of covariates we vary J accordingly. For example, for the 12 ‘RRR’ covariates we use J=12. In the case of the QC-clusters described above, we use J=1 (as the covariates within each of these QC-clusters are presumed to act in a similar way across the sample-population).

Now, for each pair of pre-defined or unsupervised clusters (‘cluster pairs’), we can measure the differential-expression of any of these 'covariate-corrected' genes from any of these data-arrays (e.g., the differential-expression of any particular column of raw, imputed, normalized, and/or covariate-corrected data). We provide corrected data that is actually the residual of the fit of our covariates onto the abundance set (i.e., the residual between the original data and the model).] (C_corrected = A_original – C_model). Moreover, we can apply clustering techniques (e.g., t-SNE, Biclustering) to the rows of these 'covariate-corrected' data-arrays (e.g., use t-SNE to cluster the rows of the abundance matrix.)

Comparing the results of covariate correction. In order to search for marker-genes associated with a specified set of sample-clusters $\{S_k\}_{k=1..K}$, we first calculate the Z-score matrix ‘Z’, which is of size [number of transcripts] -by- [number of sample-cluster-pairs]. Given a particular transcript ‘j’ and a particular sample-cluster-pair ‘(k1,k2)’, the value of Z(j,k1,k2) is obtained as follows. First, we measure the AUC associated with transcript j between sample-cluster-pairs k1 and k2. Then we assess the statistical significance of this AUC by calculating the AUC for a large number of label-permuted trials, and then estimating the Z-score of the original AUC from step #1a with respect to the mean and standard-deviation of the distribution of AUCs under the null-hypothesis in step #1b1. Note that, when z>0, the one-sided p-value associated with the z-score from #1b can be simply calculated as: log(p) = log(0.5) + erfcln(z/sqrt(2)).

Note that this Z-score matrix will depend on the transcripts used. If we use the original transcripts 'A', we will calculate Z(A). On the other hand, if we use the covariate-corrected transcripts R_J, we will calculate $Z(R_J)$.Once we have calculated both Z(A) and $Z(R_J)$, we can simply calculate the correlation between the two. Moreover, we can step through each of the sample-clusters S_k. For each sample-cluster k we can calculate the correlation 'c_k' between submatrices of Z(A) and $Z(R_J)$ corresponding to those sample-cluster-pairs that include sample-cluster k. (Table 2) Note that this latter process produces a single number (i.e., the correlation c_k) for each sample-cluster k. The sample-clusters for which c_k is low can be considered as strongly affected by the covariate-correction associated with R_J. Conversely, those sample-clusters for which c_k is high can be thought of as 'robust' with respect to the covariates associated with R_J.

As noted in the variables above (Figure 2), we compared the performance across all combinations of 1) intron- or exon-abundances, 2) 1 of our 4 data preparation methods, 3) using all (127), or some 12 RF-derived [2], 12 SRRR-derived [30], and MIN to MAX bias mode-specific) covariates.

We also conducted a preliminary analysis using 1 through R zeta components in the correction of the abundance matrix. An example comparison of the results correcting using different counts of terms from Z is provided in Table 2 and Supplemental Table 3. The results were compared in multiple ways, and the results compared to identify the most informative methodology for comparison versus comparison in simulated data. As in previous examples comparing normalized and imputed data, we compare our example results in terms of both adjusted mutual information (AMI), calculated by evaluating the 'repackaged probability' of seeing collective overlaps with the pre-defined AIBS clusters that are as rare or rarer than our observations. In other words, these are the probabilities of observing our sample-cluster overlaps vs. our baseline assumption of these overlaps occurring in a random shuffling of sample labels. We also calculate the differentially expressed genes, genes surpassing log(FoldChange) > *(variable undefined in draft)* and p-value > *(variable undefined in draft)* using *(variable undefined in draft)*, across cluster pair associations before and after to delineate potential marker genes of interest. (Table 3). We provide an example of this implementation on the human frontal cortex samples, highlighting variation in cluster pairs which vary from defined cell types in correlated ways, but which the cause of their independent clustering is not understood.

Chapter 2.4, in part, is currently being prepared for submission for publication of the material. Multifactorial Quality Control Analysis for Single Cell Transcriptomic Profiling. 2020. McCorrison J, Rangan A, Schork NJ. The dissertation author was the primary investigator and lead author of this paper.

2.5: FUTURE WORK

The improvement of laboratory methods in single cell sequencing is rapid, and novel methods are naturally taken as a way of getting around biological barriers associated with targeting different cell types, in different tissues. As shown above, quality error may be introduced not only through sequencing, but simply through sample storage (e.g. by freezing and unfreezing.) Advanced cell typing assays inherently target the most difficult to study cells, the ones which will be the most interesting to understand due to their unknown function and possible correlated effects in signaling with better-understood neuronal cell types.

To begin to studying more complicated trends like time series neural development and memory formation within hosts, and comparing hosts, we must begin by getting a strong functional foundation of the variety apparent in these regions of the brain, and to what degree our analysis of lesser expressing cell types (e.g. inhibitory cells) is subject to data loss.

As we once saw in the metagenomic context, single cell sequencing is moving cell typing research from the population wide evaluation of rare targets in unstudied tissues to the isolation of the rarest signals. This is performed through extreme preparation methods that stretch the limits of the ability to capture RNA from the host cell and to amplify it and interpret the degree to which any signal was confounded. Leveraging longitudinal assays with this type of information will be absolutely essential when cell type delineation in organs. Furthermore, as we approach the development of organ-like tissues which could be used for high throughput assay and clinical interventions, organoids, huge opportunities remain to leverage metadata-aware biclustering to improve the cost and throughput of experimental protocols. These methods have further applications in the development of synthetic materials, biomasses for fuel/food, and other industrial applications.

CHAPTER 3: ELUCIDATING LONGEVITY-ASSOCIATED OUTCOMES FROM FIBROBLAST-DERIVED TRANSCRIPTOMIC SEQUENCING IN A REFERENCE FREE CONTEXT

3.1. INTRODUCTION

As we once saw in the metagenomic context, where individually bacteria were slowly isolated and sequenced one by one as high throughput sequencing improved, single cell sequencing is already being used for the population wide evaluation of rare targets in unstudied tissues. The goal of many research studies is the isolation of the rarest signals, those previously unstudied due to their difficulty of isolation, via extreme preparation methods that stretch the limits of the ability to capture signal from small quantities of RNA. Standard methods of transcriptomics analysis have been complicated by the lack of available reference information on novel cell types. Further difficulties are imposed when interpreting true signal from false noise during expression events which are not captured by the defined reference coding region sequence. Complications arise in the following chapter when comparing abundances because of these reference accuracies but also because of, in some cases, a complete lack of defined references for query species. To compare abundances amongst species without defined references, we must create interpretation terms (e.g. defined groups of proteins with similar function) translating our query-specific transcript-specific abundance to a representation that can be compared across all of our query species. We linked our query-specific transcripts together using a mapping to the best defined references for each query species in literature (e.g. using ortholog groups associated with reference-specific genes, assigned to each transcript). While we admit that leveraging orthologs as indicators of gene- or isoform-specific expression can lead to spurious abundances, we found in preliminary study that such a method is far more sensitive than the comparable use of a single best-defined reference (e.g. the domesticated chicken.)

In the following chapter, we compare the performance of our abundance when we use alignment to our best defined references, and also via de novo recapitulation of the sequences and re-alignment to determine abundances. Leveraging our metadata against the abundances is subject to a series of additional complications which we address at length. Most critically, we compare common modelling methods to evaluate trends in transcriptomic abundance which account for expected variance in transcriptomic expression associated with evolutionary divergence events (e.g. nodes of the phylogenetic tree) and phylogenetic relationships between species (e.g. branch lengths of the tree) defined from literature differently. We also discuss our metadata itself, critically evaluating its influence on our results when gathered from various sources, or normalized for transformation in our models which account for phylogeny contrast in different ways.

This study follows analysis by collaborators from the Longevity Consortium evaluating the relationships between longevity events, events dictating extreme high or low deviations from our expectation of lifespan given mass, using linear modelling versus metadata to predict outcomes in rodents and in canines. Additional studies provide comparable targets in additional host species, and which account for longevity using different measurement types (e.g. variants) and host species. We show that the complications of eukaryotic, reference-free analysis in a diverse cohort can be accounted for, with caveats requiring individual investigation of highlighted results associated with a given phenotype, and identify potential longevity associated targets shared across host species.

3.2. MULTI-REFERENCE GENOME-WIDE RNA-SEQUEQNE ANALYSIS OF 49 BIRD SPECIES IDENTIFIES TRANSCRIPTS ASSOCIATED WITH AVIAN LONGEVITY

See un-published work, a draft currently being prepared for submission, reproduced in this chapter:

McCorrison J, Chan AP, Choi Y, Ding K, Pickering A, Pawlikowska L,Norden-Krichmar T, Evans D, Schork NJ, Miller RA. **Multi-reference Genome-wide RNA-sequence Analysis of 49 Bird Species identifies Transcripts Associated with Avian Longevity.**

MULTI-REFERENCE GENOME-WIDE RNA-SEQUENCE ANALYSIS OF 49 BIRD SPECIES IDENTIFIES TRANSCRIPTS ASSOCIATED WITH AVIAN LONGEVITY

Jamison McCorrison, Agnes Chan, Yongwook Choi, Andrew Pickering, Ludmilla Pawlikowska, Trina Norden-Krichmar, Kuan-Fu Ding, Daniel Evans, Richard A. Miller, Nicholas J. Schork

University of California, San Diego, La Jolla, CA (JM, KFD, NJS); University of Michigan, Ann Arbor, MI (AP, RAM); University of California, San Francisco, San Francisco, CA (LP); University of California, Irvine, Irvine, CA (TNK); California Pacific Medical Center, San Francisco, CA (DE); The Translational Genomics Research Institute, Phoenix, AZ (NJS)

Address correspondence to:

Nicholas J. Schork, Ph.D.
Quantitative Medicine
The Translational Genomics Research Institute (TGen)
445 North Fifth Street
Phoenix, AZ 85004
nschork@tgen.org

and

Richard A. Miller, M.D., Ph.D.
Department of Pathology
University of Michigan
Ann Arbor, MI 48103
millerr@umich.edu

ABSTRACT

Bird species exhibit great variation in lifespan, raising the question as to whether or not this variation can be attributed to inherent DNA sequence and/or gene expression differences among them. In order to identify genes whose expression levels correlate with lifespan across bird species, we characterized the transcript abundance profiles of fibroblasts obtained from 49 species exhibiting great variation in their maximum lifespans using RNA-sequencing protocols. Due to the fact that reference genomes were available for only 20 of the species, we used reference genomes for each species from the 20 species with references based on their phylogenetic or transcriptomic distances. We also contrasted reference-guided transcript abundance calculations with abundances determined from de novo assembly of each species transcriptome. We correlated transcript abundance levels with the maximum lifespans of the 49 bird species, controlling for both phylogenetic relationships as well as differences in body size. We also identified the human orthologs of the most strongly associated transcripts and ultimately found evidence for 63 human gene equivalents whose abundance levels correlated with bird lifespan at FDR-adjust p-value < 0.05. These associated transcripts are known to mediate important biological processes, including organ morphology relating to intestinal and gonad development, and carcinogen markers in the stomach, liver, and intestine.

KEYWORDS

Longevity Transcriptomics Avian Bird Orthology Aging Lifespan Ontology RNAseq

BACKGROUND

The identification of genes contributing to lifespan has received a great deal of recent attention, in part because of the belief that their identification could lead to insights into, e.g., nutritional or pharmacological intervention targets for enhancing longevity, possibly by slowing the aging process and age-related disease onset.[1,2,3,4,5] Unfortunately, the identification of genes that influence lifespan in a way that could lead to longevity-enhancing interventions has been elusive, due, most likely, to the number and complexity of such genes. One strategy for potentially overcoming this complexity involves exploring within-species variation in lifespan and the genes and genetic variations that might contribute to it using well-controlled study designs. For example, large-scale genome-wide association studies (GWAS) have been pursued in humans that have common ancestral origins to identify genetic variants associated with human lifespan, as have studies of specific strains of mice that exhibit intra (and inter) strain variation in lifespan.[6] In addition, highly controlled gene manipulation studies, such as those involving knock-out and transgene protocols, have been pursued in studies involving yeast, worms, and flies, to determine if the manipulation of specific genes affects the longevity of those species.[7,8,9]

Complementary and more recent approaches to the identification of genes contributing to lifespan involve exploring genetic similarities and differences across multiple species that exhibit variation in lifespan – the intuition being that if, for example, a gene's increased expression level is necessary or sufficient for extending lifespan in one species, then that gene and its orthologs in other species should be expressed at relatively high levels in long-lived, as opposed to short-lived, species. In this context, a recent study by Ma et al.[10] considered characterizing the transcriptomes of 16 mammalian species exhibiting a wide range in lifespans using RNA sequencing (‘RNA-seq’) technologies and protocols. By exploiting special bioinformatics and statistical methods to normalize the expression levels across the species, as well as control for the phylogenetic relationships between the species, the authors identified genes whose expression levels were associated with lifespan across the 16 species. Many of these genes were known to be involved processes relevant to aging, such as DNA repair and metabolism. In addition, by focusing on multiple species, and by characterizing the gene orthologs across those species, and the methods for identifying those gene orthologs, the authors provide a useful resource for researchers investigating genes in other species to enable comparison of results with theirs. [11]

We conducted a study exploring the relationship between lifespan and transcript abundance levels quantified from RNA-seq protocols using fibroblasts obtained from 49 different bird species known to exhibit wide variation in their lifespans. Unfortunately, unlike other species, there are not, to date, universally accepted avian reference genomes for the majority of the species we considered in our analyses. We therefore adopted two strategies to leverage as many reliable and available avian reference genomes as we could. First, we used the reference genome from the species closest in the phylogenetic distance to each of the 49 species we studied to assign and quantify transcripts. Second, we compared the performance of the analysis using reference-guided transcript abundances to transcript abundances obtained from the *de novo* reconstruction of transcripts for each species coupled with the identification of orthologous gene groups derived from these *de novo* transcripts. [Chan, Choi, McCorrison, Pickering, Pawlikowska, Norden-Krichmar, Ding, Evans, Miller and Schork; (manuscript in preparation)]

We tested the association of each set of orthologous gene (i.e., transcript) group to maximum life span (MLS) and body size-corrected MLS while correcting for the phylogenetic relationships of the species using the method of phylogenetic contrasts [12] and multivariate distance matrix regression (MDMR; [13, 14, 15]. Note that for these analyses we considered maximum lifespan and body size data from different sources [16.] We did this since there is not consensus on the MLS and body size information for all species w studied. We also identified the human orthologs of the associated genes using OrthoDB [17]. Finally, we conducted pathway enrichment analysis to identify common processes and genetic networks that are influenced by the associated genes. Ultimately, our strategies and workflows for identifying transcripts whose abundances are associated with MLS in birds provides a recipe for conducting similar analyses across additional sets of diverse species. In addition, our list of associated genes and pathways can be compared with the results of different types studies seeking to identify genes associated with MLS. We believe that our study is largest to date to explore the avian transcriptome and should motivate additional studies investigating evolutionarily conserved processes and pathways contributing to lifespan.

RESULTS

A graphical representation of the phylogenetic relationships between the 49 species we studied, as well as their MLS and body size values are depicted in Figure 1. Color coding of the species indicates which reference genome was used for each of these 49 species to assess transcript abundances from the RNA-sequencing reads. Note that some of the 49 species were assigned use of the same reference genome whereas other reference genomes were used with only one or two species. Note also that the long-lived species, even corrected for body size, occur in different sub branches of the phylogeny.

The relationship between MLS and body size across the 49 species is provided in Figure 2. Note that some of the species appear to be more outlying relative to others (e.g., Turkey, Ruffed Grouse and Rock Dove) given what is otherwise a fairly linear relationship between log body mass and MLS. The residuals from the regression of log body mass on MLS were considered in our association analyses with transcripts.

After transcripts abundances were quantified by either mapping reads to chosen reference genomes or counting them from the *de novo* transcript assembly for each species, we determined orthologous groups of transcripts using OrthoDB [17]. We then tested the association of each orthologous group (OG) to MLS and MLS-corrected for body size using a simple linear model, the phylogenetic contrast method of Felsenstein, as implemented in the R module CAPER [12,18], and MDMR [13,14,15]. Note that the MLS and body size values were obtained from different sources so we considered these different values in our analyses. Table 1 provides the results of these association analyses and suggests that the transcript quantification methods, the different analytical methods, and the choice of MLS/body size values can make a difference in the number of associated OGs.

In order to reduce the number of associations, we considered only those OGs obtained with the reference genome alignments that had known human orthologs based on the OrthoDB [17]. Table 2 provides the results for the most significantly associated OGs exhibiting either a positive or negative abundance association with MLS. It can be seen from Table 2 that there is consistency in the associations of these OGs with the OG abundances derived from the *de novo* assemblies. Figure 3 provides volcano plots summarizing all the associations involving the reference-guided and *de novo* assembled OGs and again suggests reasonable agreement between the two strategies. These analysis results give us confidence that our results were robust to the nuances surrounding transcript assembly or OG determination. Figures 4-6 provide examples of individual OGs and their associations with MLS using the different analytical methods. These figures also provide information about the human gene orthologs of the associated OGs as well as information about the pathway involvement of the human orthologous genes.

We next took the most significantly associated OGs from the reference-guided and *de novo* assembled approaches and performed pathway enrichment analyses on each set of associated OGs using the Ingenuity Software Suite [19]. Table 3 summarizes the results. Tables 4a-4c break down the results of these pathway enrichment analyses and provide information on the most significantly enriched processes, cellular components, molecular functions, respectively. Table 5 lists the most significant diseases and/or functional associations of the enriched pathways. Finally, we identified the 63 OGs that were most significantly associated with the residual of natural log of mass on MLS (FDR-adjusted p-value cutoff 0.05) and subjected them to pathway analyses. Figures 7 and 8 depict the results. Many of these OGs had identifiable human orthologs and 10 of them were also found to be associated with the residual of natural log of mass on lifespan across mammalian species, including EVI5L, MRPL37, PWWP2A, THOC5, and WHAMM, based on the findings by Ma et al [10].

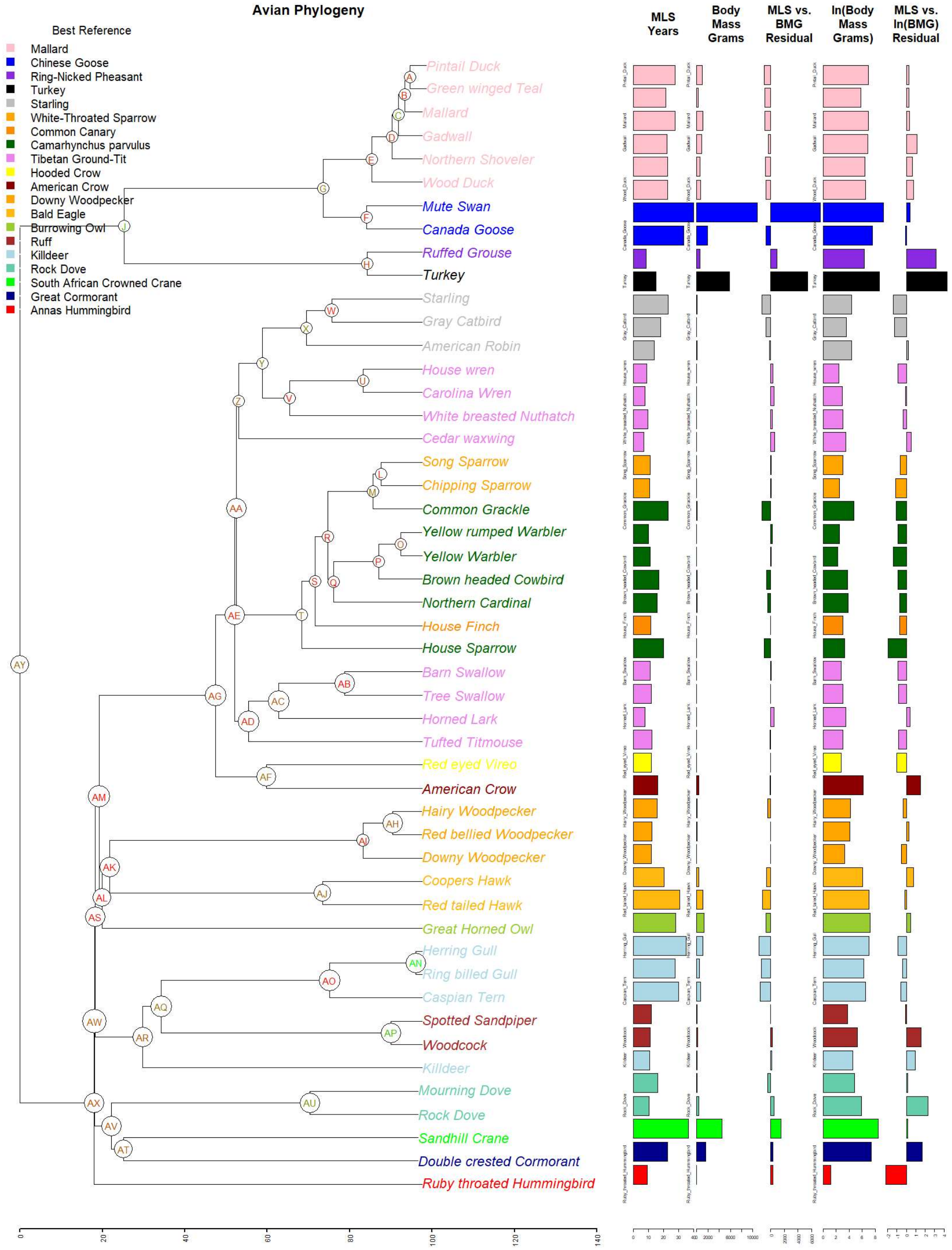


Figure 3.2.1. Phylogenetic relationships among the 49 species. Leaves = query species, colored by best reference assigned. Nodes = evolutionary breakpoints, colored by caper-derived phylogenetic contrast. Bar plots = metadata for each row, as defined in Supplemental Table 1, using metadata context "BMG (A)".

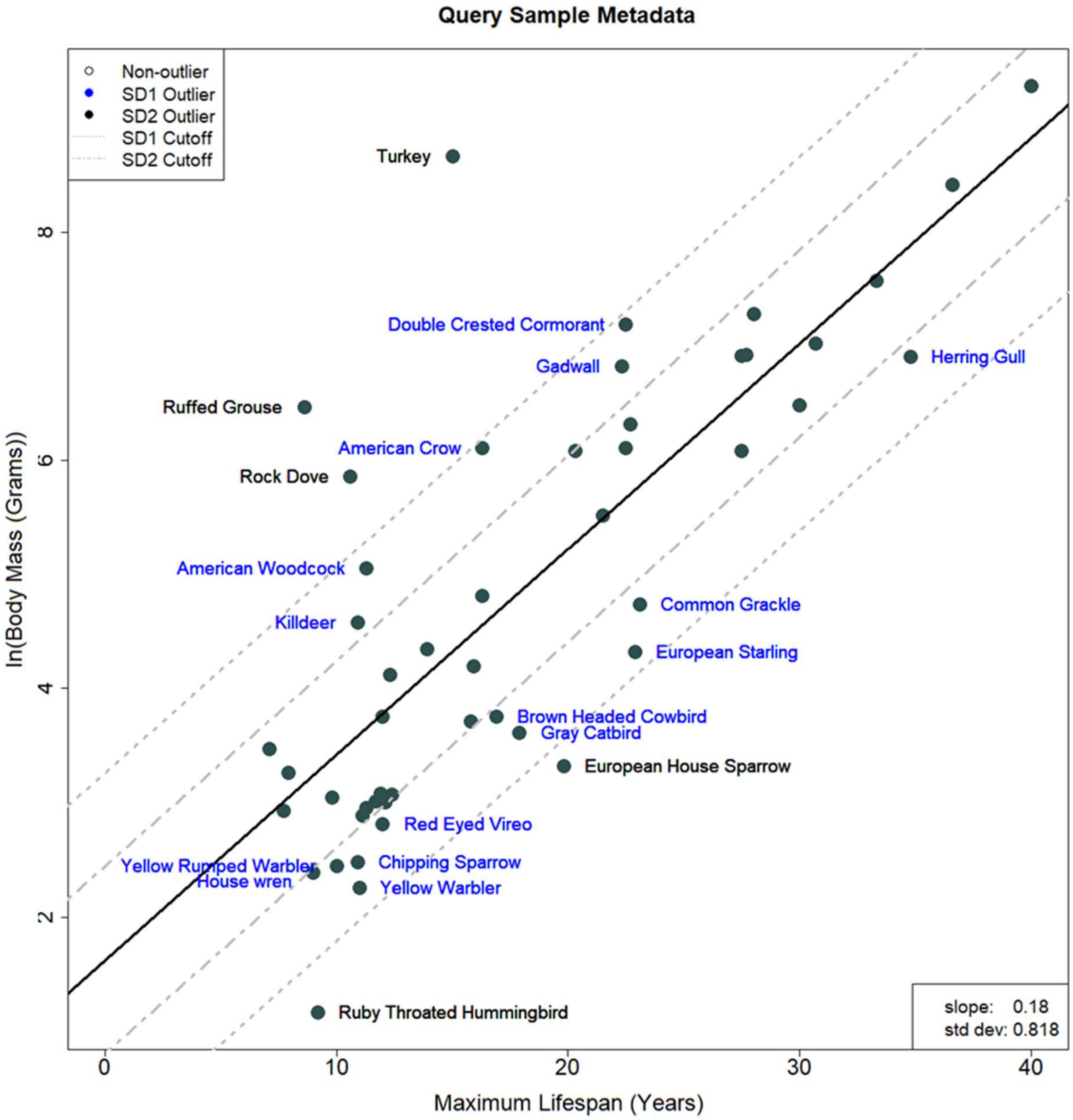


Figure 3.2.2: Relationship of MLS to log (base e)-normalized body mass in grams across all 49 species. Dotted and dashed lines indicate 1 and 2 standard deviation distances from the regression line based on residual values from the regression of MLS on log body mass. Metadata provided in Supplemental Table 5a (Columns = "Dependent Variable: BMG (A)", "Independent Variable: MLS (A)"). This plot is recreated in alternative metadata contexts in Supplemental Figure 2.

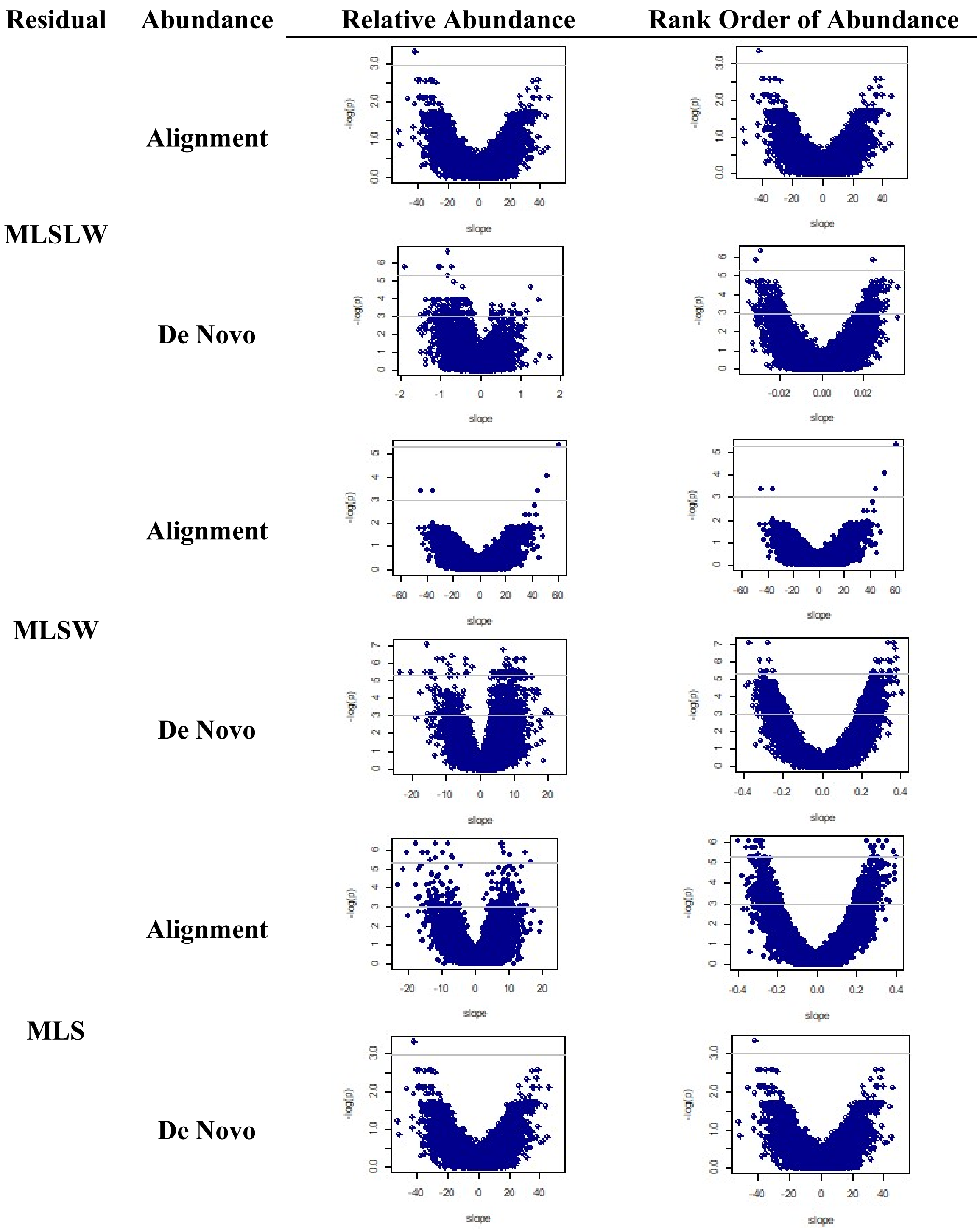
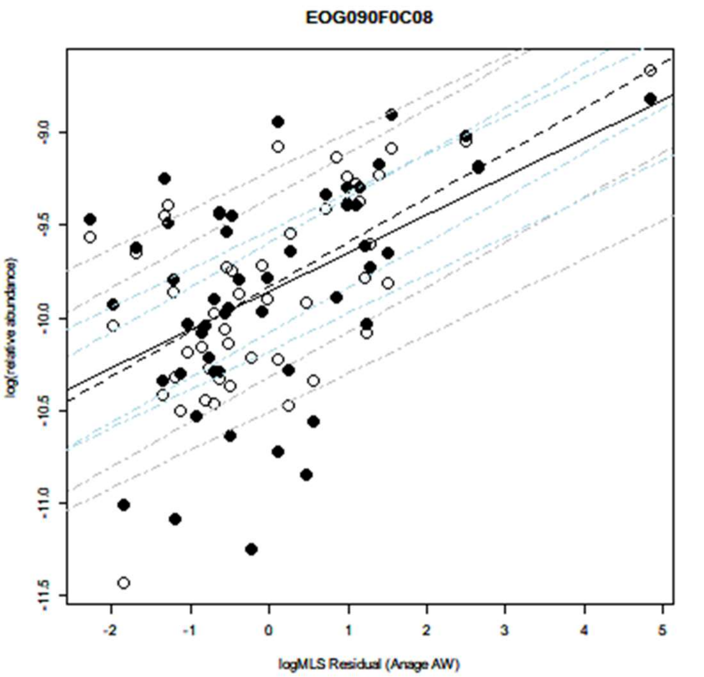


Figure 3.2.3. Volcano plots constrasting MLS associations with the reference-guided and *de novo* assembled OGs. Number of unique ortholog groups (Y, negative natural log of raw p-value) and their associated phylogenetically contrast-adjusted slope (X, calculated using the caper model package), A) CAPER, MLSLW, log(Rel Abs), (BMG (A)), B) CAPER, MLSW log(Rel Abs) , (BMG (A)), C) CAPER, MLS log(Rel Abs) , (BMG (A)), D) CAPER, MLSLW, Rank(Abs), (BMG (A)) , E) CAPER, MLSW Rank (Abs) , (BMG (A)) , F) CAPER, MLS Rank (Abs) , (BMG (A). Compare to metadata context (AW (A) in Supplementary Figure.

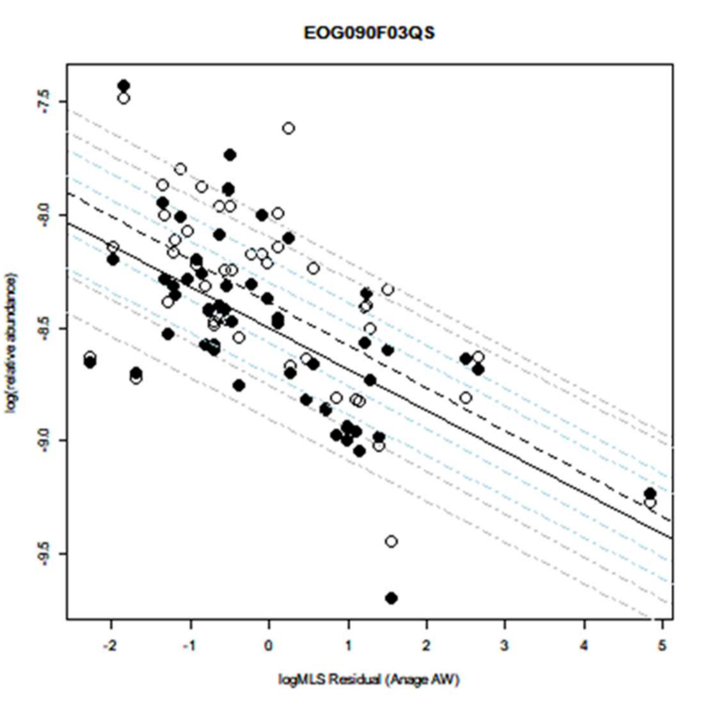


Query-specific Symbols	PAIP2 (Q5ZJS6) Polyadenylate-binding protein-interacting protein 2	
HuRef Equivalents	PAIP2 (Immune function, restricting viral synthesis and replication)	
Biological Processes	Many	GO:0007283,GO:0007613,GO:0045947,GO:1900271
Molecular Functions	Many	GO:0003729,GO:0030371
Cellular Components	GO:0005737	cytoplasm
Interpro Domains	IPR009818	Ataxin-2, C-terminal
Unclassified GO Terms	GO:0005515,GO:0006417	

	BMG (O)	BMG (A)	AW (A)
LM Residuals (CCC rho.q)	0.910189	0.9254402	0.9262045
Phylogeny Contrasts (CCC rho.q)	0.5510748	0.8322029	0.8513352

	LM		CAPER		MDMR (Mash)		MDMR (Phylo)	
BMG (A)	**Align**	**De novo**	**Align**	**De novo**	**Align**	**De novo**	**Align**	**De novo**
Slope	1.362371	1.239269	1.362371	1.239269	NA	NA	NA	NA
P	0.000521	2.49E-06	1.83E-06	1.54E-05	0.012	0.14	0.014	0.198
FDR(P)	0.009794	0.000532	0.005292	0.009905	0.018536	0.146522	0.024231	0.204554
AW (A)	**Align**	**De novo**	**Align**	**De novo**	**Align**	**De novo**	**Align**	**De novo**
Slope	1.362371	1.239269	1.362371	1.239269	NA	NA	NA	NA
P	0.000547	2.76E-06	1.27E-06	2.04E-05	0.008	0.152	0.016	0.196
FDR(P)	0.009514	0.000492	0.003886	0.015741	0.015906	0.159081	0.026981	0.203471

Figure 3.2.4. Example positively associated ortholog group with defined human gene symbol PAIP2.

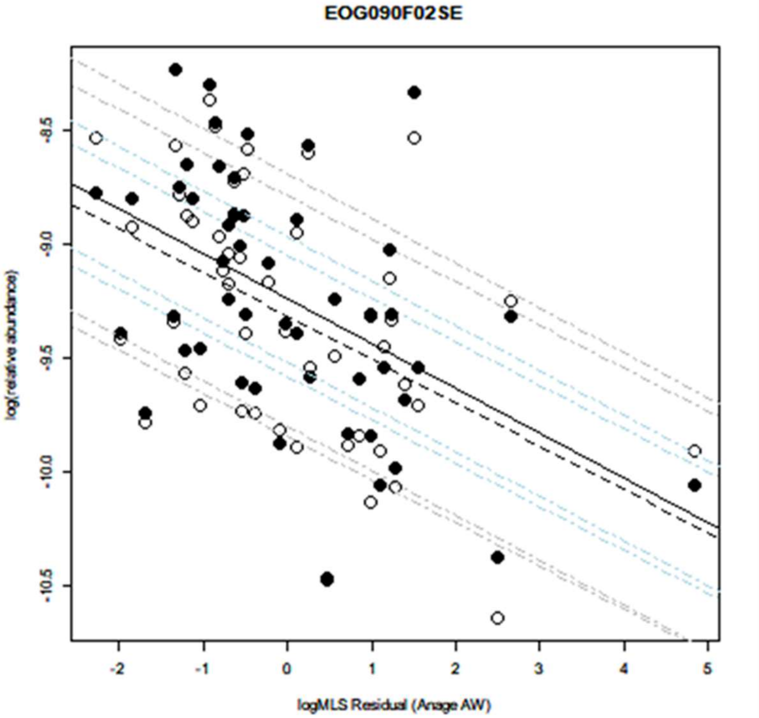


Query-specific Symbols	TMEM87A (transmembrane protein 87A)	
HuRef Equivalents	TMEM87A (Cellular membrane transporter identified in golgi apparatus, localized in liver tissue)	
Biological Processes	NA	None
Molecular Functions	Many	GO:0003729,GO:0030371
Cellular Components	GO:0016021	Integral Component of membrane
Interpro Domains	IPR009637	Lung seven transmembrane receptor-like
Unclassified GO Terms	GO:0005575	

	BMG (O)	BMG (A)	AW (A)
LM Residuals (CCC rho.q)	0.9726184	0.947102	0.9506649
Phylogeny Contrasts (CCC rho.q)	**0.5864688**	**0.8197971**	**0.8211864**

	LM		CAPER		MDMR (Mash)		MDMR (Phylo)	
BMG (A)	Align	De novo	Align	De novo	Align	De novo	Align	De novo
Slope	-1.980223	-1.902287	-1.980223	-1.902287	NA	NA	NA	NA
P	1.54E-06	1.12E-06	5.3388E-07	1.3369E-06	0.668	0.336	0.576	0.246
FDR(P)	4.93E-04	4.04E-04	0.00308261	0.00323283	0.6682315	0.3395876	0.5777009	0.2518
AW (A)	Align	De novo	Align	De novo	Align	De novo	Align	De novo
Slope	-1.980223	-1.902287	-1.980223	-1.902287	NA	NA	NA	NA
P	3.14E-06	2.05E-06	2.0192E-06	6.8307E-06	0.636	0.328	0.546	0.264
FDR(P)	6.07E-04	4.78E-04	0.00388635	0.00986008	0.6365512	0.3322582	0.5488517	0.2703203

Figure 3.2.5. Example negatively associated ortholog group with defined human gene symbol TMEM87A.



Query-specific Symbols	GCLC (F1NQZ9) Glutamate-cysteine ligase catalytic subunit	
HuRef Equivalents	GCLC (Increased levels related to the development of hepatocellular carcinoma (liver cirrhosis))	
Biological Processes	Many	GO:0006534,GO:0006536,GO:0006749,GO:0006750, GO:0006790,GO:0006979,GO:0008637,GO:0009058, GO:0009725,GO:0019852,GO:0031397,GO:0032436, GO:0034641,GO:0046685,GO:0050880,GO:0051900, GO:2001237
Molecular Functions	Many	GO:0000287,GO:0004357,GO:0016595,GO:0016874, GO:0043531,GO:0046982,GO:0050662
Cellular Components	Many	GO:0005829,GO:0017109
Interpro Domains	IPR004308	Glutamate-cysteine ligase catalytic subunit
Unclassified GO Terms	GO:0003674,GO:0008150,GO:0009408,GO:0009410,GO:0043066, GO:0045454,GO:0045892	

	BMG (O)	BMG (A)	AW (A)
LM Residuals (CCC rho.q)	0.9541201	0.951162	0.9509247
Phylogeny Contrasts (CCC rho.q)	0.4988913	0.6059601	0.6630369

	LM		CAPER		MDMR (Mash)		MDMR (Phylo)	
BMG (A)	**Align**	**De novo**	**Align**	**De novo**	**Align**	**De novo**	**Align**	**De novo**
Slope	-0.80999	-0.83848	-0.80999	-0.83848	NA	NA	NA	NA
P	0.012611	9.83E-05	1.78E-05	2.33E-07	0.008	0.02	0.016	0.04
FDR(P)	0.06711	0.004268	0.010292	0.001345	0.016722	0.026614	0.025684	0.048562
AW (A)	**Align**	**De novo**	**Align**	**De novo**	**Align**	**De novo**	**Align**	**De novo**
Slope	-0.80999	-0.83848	-0.80999	-0.83848	NA	NA	NA	NA
P	0.010376	8.51E-05	1.75E-05	7.37E-07	0.008	0.024	0.018	0.044

Figure 3.2.6. Example positively associated ortholog group with defined human gene symbol GCLC.

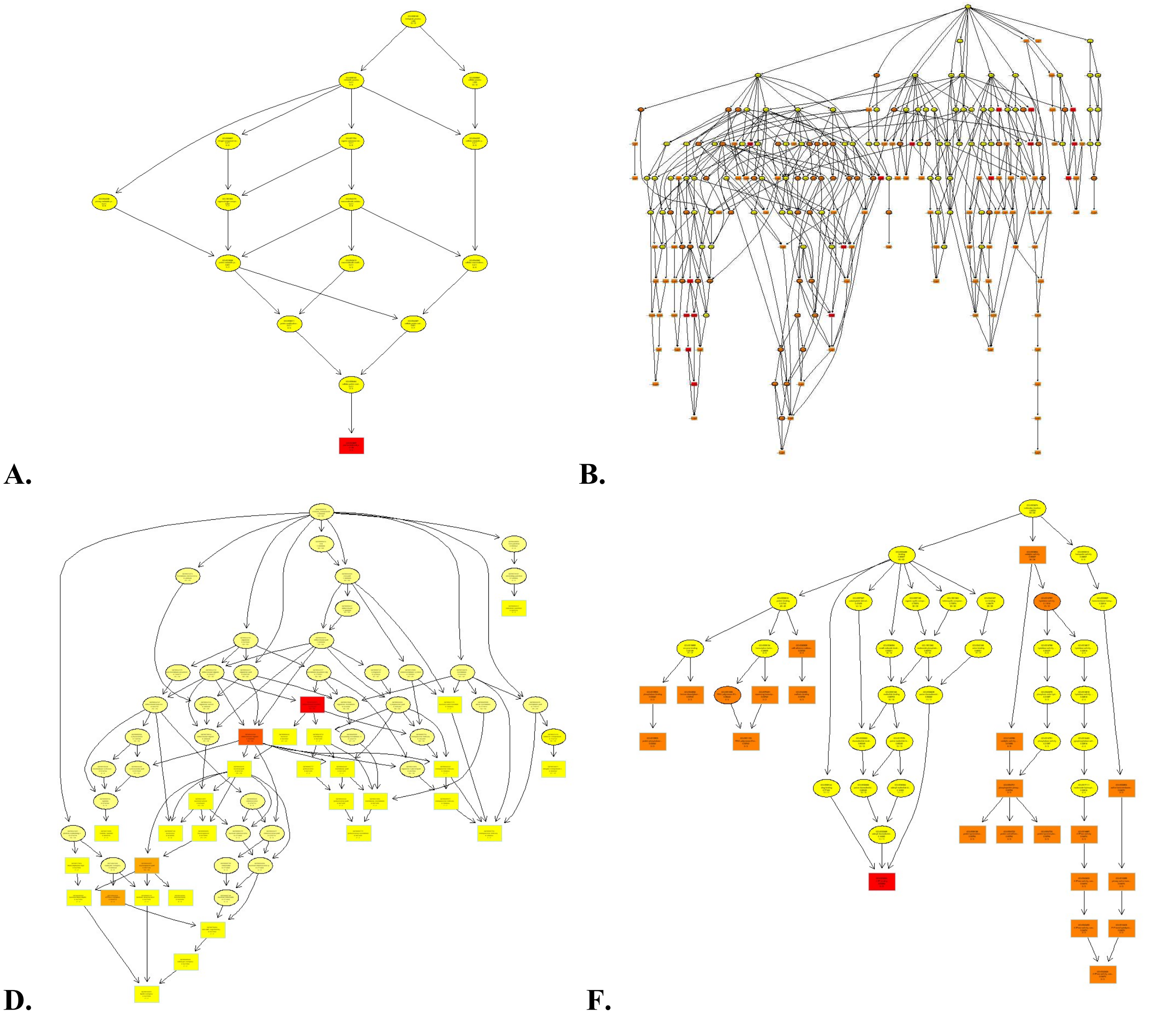


Figure 3.2.7. TopGO Gene Ontology Results. Metadata = BMG (A). GO terms significantly associated with significant OGs from both Alignment and De Novo abundance contexts. (A,C,E) Positive association with residual on lifespan. (B) Negative association with residual on lifespan. **A-B)** Biological processes, cutoff = KS 0.05. **D.)** Cellular components, cutoff = KS 0.05, *negative-only*. **E-F.)** Molecular functions, cutoff = KS 0.05, *negative-only*.

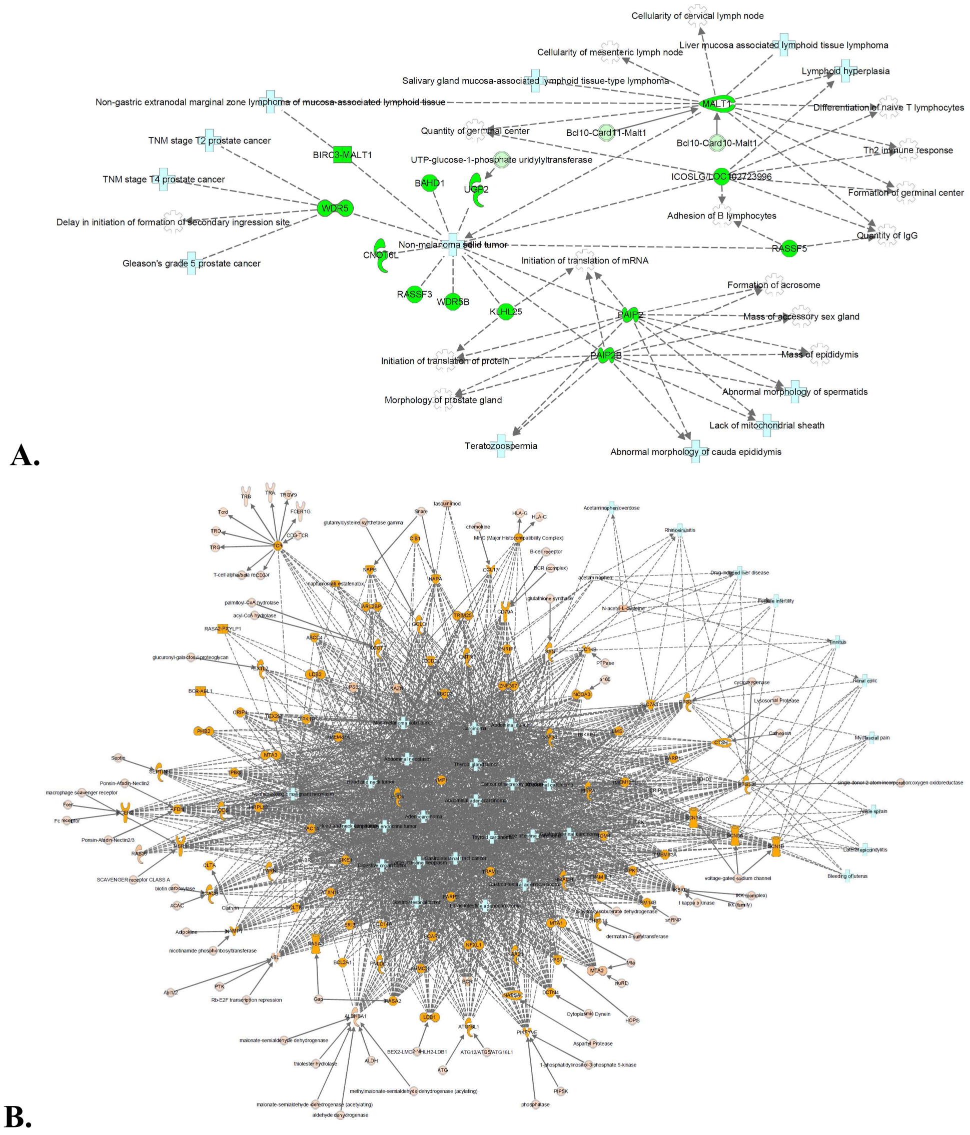

Cellularity of cervical lymph node
Liver mucosa associated lymphoid tissue lymphoma
Cellularity of mesenteric lymph node
Lymphoid hyperplasia
Salivary gland mucosa-associated lymphoid tissue-type lymphoma
Non-gastric extranodal marginal zone lymphoma of mucosa-associated lymphoid tissue
MALT1
Differentiation of naive T lymphocytes
Quantity of germinal center
Bcl10-Card11-Malt1
Bcl10-Card10-Malt1
TNM stage T2 prostate cancer
BIRC3-MALT1
Th2 immune response
UTP-glucose-1-phosphate uridylyltransferase
ICOSLG/LOC102723996
TNM stage T4 prostate cancer
BAHD1
UGP2
Formation of germinal center
WDR5
Adhesion of B lymphocytes
Delay in initiation of formation of secondary ingression site
Quantity of IgG
Non-melanoma solid tumor
RASSF5
CNOT6L
Initiation of translation of mRNA
Gleason's grade 5 prostate cancer
Formation of acrosome
RASSF3
WDR5B
KLHL25
PAIP2
Mass of accessory sex gland
Initiation of translation of protein
PAIP2B
Mass of epididymis
Abnormal morphology of spermatids
Morphology of prostate gland
Lack of mitochondrial sheath
Teratozoospermia
Abnormal morphology of cauda epididymis
A.
B.

Figure 3.2.8. Ingenuity Pathway Figure (hand-curated). Metadata = BMG (A). Significant OGs shared in Alignment and De Novo abundance contexts with Ingenuity Pathway equivalents (Human, Mouse, Rat)**.** Data used for Figure examples: Model = CAPER, Dependent = MLSLW, Independent = LG, Metadata = BMG (A), FDR p-value cutoff = 0.5. IP pathway selection cutoff = 5e-8. Direction of association = **A)** Positive (Dark Green=Positive Association with Residual on Lifespan, Light Green = IP Linked Associated Target). **B)** Negative (Dark Orange=Positive Association with Residual on Lifespan, Light Orange = IP Linked Associated Target). Some of the most significant Disease and Function terms, those with greater than 100 connections to genes were omitted for visualization (see Supplemental Table). Associations between Human reference genes co-defined for significant OGs (dark green, dark orange) and diseases (light blue) and functions (white) are indicated with a dashed line. Connections between these genes and their linked targets from the IP database are provided with a solid line (Table 1, Table 2) Compare to Table 5.

Table 3.2.1. Number significant OG for each model at Cutoff = FDR-adjusted P-value = 0.005. The number of OGs significant at threshold for each model (e.g. “CAPER”). Abundance type : DN = De Novo, Al = Align. Directionality of association: Slope of association after phylogeny correction is positive (+) or negative (-). Me = Metadata: BMG (A) = Body Mass (Grams), updated with AnAge, A = Adult Weight (Grams), source from AnAge. D = Dependent variable: A = MLS = Maximum lifespan, B = MLSW = residual of BMG (Body mass in grams) on MLS, C = MLSLW = residual of the natural log of the BMG on MLS.. In = Independent variable: G = relative gene abundance, RG = rank-ordered abundance, LG = natural log of abundance. (Table on next page.)

Table 3.2.1. Number significant OG for each model at Cutoff = FDR-adjusted P-value = 0.005. Legend on previous page

Me	D	In	LM						CAPER						MDMR MASH						MDMR PHYLO					
			DN		Al		Sh		DN		Al		Sh		DN		Al		Sh		DN		Al		Sh	
			-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+
BMG (A)	A	G	5	31	6	30	4	29	4	22	4	22	4	22	16	27	17	26	12	22	0	1	0	1	0	1
		RG	0	0	0	0	0	0	15	9	14	10	10	5	33	35	27	41	14	22	1	1	1	1	0	0
		LG	18	23	26	15	17	14	8	11	12	7	8	7	15	31	17	29	14	28	0	2	1	1	0	1
	B	G	0	14	0	14	0	14	0	2	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0
		RG	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
		LG	0	0	0	0	0	0	2	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
	C	G	30	41	24	47	22	39	0	0	0	0	0	0	59	46	57	48	51	40	10	12	6	16	5	11
		RG	0	0	0	0	0	0	0	2	0	2	0	2	40	46	40	46	20	26	8	7	8	7	4	3
		LG	137	40	128	49	122	34	1	0	1	0	1	0	76	35	70	41	62	27	19	3	17	5	16	2
AW (A)	A	G	7	34	2	39	2	34	1	13	1	13	0	12	16	50	17	49	12	45	0	1	0	1	0	1
		RG	0	0	0	0	0	0	0	0	0	0	0	0	51	56	53	54	27	30	0	0	0	0	0	0
		LG	41	31	54	18	39	16	21	11	26	6	20	5	28	63	36	55	23	50	0	1	0	1	0	1
	B	G	1	37	1	37	0	36	3	16	3	16	3	16	0	0	0	0	0	0	0	0	0	0	0	0
		RG	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
		LG	2	5	3	4	1	3	14	13	14	13	7	6	1	0	1	0	1	0	0	0	0	0	0	0
	C	G	37	44	31	50	29	42	0	0	0	0	0	0	64	67	63	68	52	56	10	17	8	19	6	15
		RG	0	0	0	0	0	0	1	0	0	1	0	0	43	44	46	41	25	23	3	7	6	4	2	3
		LG	145	48	139	54	130	39	2	1	2	1	2	1	87	35	79	43	69	25	22	5	20	7	19	4
Shared	A	G	0	8	0	8	0	8	0	1	0	1	0	1	3	10	4	9	2	8	0	0	0	0	0	0
		RG	0	0	0	0	0	0	0	0	0	0	0	0	6	7	8	5	4	3	0	0	0	0	0	0
		LG	4	2	6	0	4	0	2	2	2	2	2	2	3	8	4	7	3	7	0	0	0	0	0	0
	B	G	0	14	0	14	0	14	0	2	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0
		RG	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
		LG	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	C	G	29	39	24	44	22	37	0	0	0	0	0	0	47	41	45	43	40	36	7	11	4	14	3	10
		RG	0	0	0	0	0	0	0	0	0	0	0	0	30	32	28	34	15	19	2	6	4	4	1	3
		LG	133	39	124	48	119	34	1	0	1	0	1	0	65	24	59	30	53	18	15	2	13	4	12	1

Table 3.2.2. A table of the most significant OGs for the Alignment-derived abundance models, reduced to only those with defined human reference equivalents. Top OGs for which all 49 query species exhibited non-zero expression and which are the most significant in predictive of maximum lifespan based on the MLSLW phylogeny-adjusted model.

Top OGs with Shared Huref Common Symbol				CCC rho.q		LM		C		MDMR (Mash)		MDMR (Phylo)	LM		CAPER		MDMR (Mash)		MDMR (Phylo)	
Sym	OG	S	P	LM	Ph	A	DN	DN	A	DN	A	DN	A	DN	A	DN	A	DN	A	DN
Positive Association with Longevity																				
PAIP2	EOG090F0C08	1.36	0.005	0.925	0.832	0.010	0.001	0.010	0.019	0.147	0.024	0.205	0.010	0.000	0.004	0.016	0.016	0.159	0.027	0.203
RASSF3	EOG090F099I	0.92	0.011	0.943	0.376	0.002	0.023	0.349	0.538	0.162	0.431	0.149	0.002	0.026	0.026	0.426	0.544	0.159	0.441	0.158
KLHL25	EOG090F03CE	0.66	0.012	0.988	0.260	0.096	0.132	0.250	0.017	0.017	0.027	0.024	0.121	0.154	0.017	0.277	0.017	0.016	0.032	0.026
MALT1	EOG090F02FF	0.88	0.023	0.945	0.550	0.023	0.011	0.056	0.052	0.018	0.061	0.031	0.026	0.010	0.031	0.053	0.052	0.019	0.071	0.034
WDR5	EOG090F075P	0.92	0.023	0.969	0.153	0.170	0.308	0.566	0.020	0.020	0.035	0.031	0.181	0.278	0.030	0.541	0.023	0.025	0.041	0.034
BAHD1	EOG090F022R	1.08	0.025	0.967	0.156	0.001	0.006	0.126	0.141	0.022	0.094	0.027	0.001	0.006	0.029	0.167	0.145	0.023	0.107	0.028
CNT6L	EOG090F03WS	1.02	0.025	0.928	0.804	0.017	0.012	0.447	0.023	0.043	0.033	0.056	0.023	0.015	0.041	0.509	0.027	0.044	0.032	0.063
UGP2	EOG090F04H0	0.60	0.026	0.998	0.368	0.297	0.206	0.037	0.020	0.029	0.031	0.041	0.282	0.198	0.043	0.054	0.023	0.033	0.039	0.047
ICOSLG, LICOS	EOG090F07AA	0.39	0.030	0.999	0.694	0.657	0.769	0.026	0.017	0.017	0.024	0.024	0.690	0.816	0.018	0.019	0.016	0.016	0.026	0.026
ATOX1	EOG090F0CIK	0.57	0.036	0.932	0.700	0.069	0.967	0.296	0.025	0.017	0.035	0.024	0.084	0.959	0.052	0.405	0.025	0.016	0.039	0.026
None defined.	EOG090F09MR	0.55	0.039	0.987	0.526	0.913	0.540	0.067	0.017	0.020	0.024	0.027	0.920	0.529	0.036	0.063	0.016	0.023	0.026	0.027
MRPL44	EOG090F083E	1.08	0.049	0.927	0.870	0.773	0.656	0.121	0.017	0.017	0.024	0.024	0.808	0.702	0.073	0.211	0.016	0.016	0.026	0.026
RPL21	EOG090F0B88	1.19	0.049	0.999	0.503	0.050	0.004	0.019	0.019	0.093	0.024	0.049	0.061	0.004	0.078	0.019	0.019	0.098	0.026	0.047

Table 3.2.2. Most significant OGs for the Alignment-derived abundance models (Continued, 2 of 3.)

Top OGs with Shared Huref Common Symbol				CCC rho.q		LM		C	MDMR (Mash)		MDMR (Phylo)		LM		CAPER		MDMR (Mash)		MDMR (Phylo)	
Sym	OG	S	P	LM	Ph	A	DN	DN	A	DN	A	DN	A	DN	A	DN	A	DN	A	DN
Negative Association with Longevity																				
TMEM87A	EOG090F03QS	-1.98	0.003	0.947	0.820	0.000	0.000	0.003	0.668	0.340	0.578	0.252	0.001	0.000	0.004	0.010	0.637	0.332	0.549	0.270
CHST14	EOG090F07EP	-0.81	0.007	0.987	0.878	0.641	0.467	0.176	0.017	0.017	0.024	0.024	0.579	0.451	0.010	0.214	0.016	0.016	0.026	0.026
AFDN, MLLT4	EOG090F00EM	-1.38	0.007	0.903	0.699	0.001	0.001	0.019	0.017	0.017	0.024	0.024	0.001	0.001	0.010	0.019	0.016	0.016	0.026	0.026
KAZN	EOG090F07E1	-0.68	0.007	0.995	0.619	0.001	0.038	0.132	0.066	0.031	0.057	0.033	0.000	0.037	0.011	0.179	0.074	0.035	0.069	0.036
MTA1, MTA3	EOG090F02S8	-0.89	0.008	0.858	0.658	0.001	0.435	0.176	0.241	0.027	0.418	0.045	0.001	0.451	0.010	0.190	0.269	0.031	0.407	0.043
GCLC	EOG090F02SE	-0.94	0.008	0.990	0.953	0.009	0.015	0.005	0.025	0.018	0.024	0.024	0.007	0.011	0.004	0.004	0.023	0.017	0.026	0.026
MARCH1, MARCH8	EOG090F07QP	-0.97	0.010	0.972	0.773	0.069	0.073	0.007	0.017	0.029	0.024	0.024	0.073	0.088	0.017	0.017	0.016	0.029	0.026	0.030
ATXN1L	EOG090F03JS	-1.21	0.010	0.951	0.606	0.374	0.690	0.339	0.017	0.017	0.024	0.024	0.392	0.641	0.017	0.250	0.016	0.016	0.026	0.026
ARMC10	EOG090F0AQK	-0.81	0.010	0.988	-0.101	0.067	0.004	0.001	0.017	0.027	0.026	0.049	0.058	0.004	0.011	0.004	0.016	0.031	0.028	0.053
EHD3, EHD4	EOG090F0503	-0.44	0.011	0.785	0.688	0.000	0.000	0.090	0.033	0.027	0.053	0.045	0.000	0.000	0.032	0.187	0.040	0.035	0.057	0.049
IKBKE, TBK1	EOG090F0359	-0.57	0.013	0.995	0.861	0.407	0.402	0.019	0.017	0.017	0.024	0.024	0.404	0.417	0.017	0.030	0.016	0.016	0.026	0.026
ALDH6A1	EOG090F041N	-0.69	0.017	0.999	0.993	0.246	0.342	0.021	0.017	0.017	0.024	0.024	0.265	0.364	0.034	0.040	0.016	0.016	0.026	0.026
ERI3	EOG090F089Q	-1.09	0.021	0.962	0.756	0.053	0.419	0.127	0.017	0.017	0.026	0.024	0.046	0.367	0.017	0.116	0.016	0.016	0.027	0.026
MRPL13	EOG090F0A3O	-0.86	0.021	0.939	0.656	0.043	0.169	0.275	0.056	0.061	0.059	0.076	0.031	0.142	0.014	0.168	0.066	0.074	0.059	0.079

Table 3.2.2. Most significant OGs for the Alignment-derived abundance models (Continued, 3 of 3.)

Top OGs with Shared Huref Common Symbol				CCC rho.q		LM		C	MDMR (Mash)		MDMR (Phylo)		LM		CAPER		MDMR (Mash)		MDMR (Phylo)	
Sym	OG	S	P	LM	Ph	A	DN	DN	A	DN	A	DN	A	DN	A	DN	A	DN	A	DN
Negative Association with Longevity																				
MSRA	EOG090F0CAU	-0.50	0.021	0.959	0.205	0.008	0.064	0.193	0.087	0.029	0.110	0.033	0.006	0.070	0.013	0.203	0.086	0.033	0.109	0.038
GMDS	EOG090F06BO	-0.57	0.021	0.987	0.664	0.514	0.942	0.905	0.017	0.017	0.024	0.024	0.414	0.974	0.011	0.752	0.016	0.016	0.026	0.026
TMEM63A	EOG090F022V	-0.96	0.021	0.929	0.635	0.000	0.010	0.113	0.000	0.000	0.000	0.000	0.000	0.008	0.017	0.101	0.000	0.000	0.000	0.000
C2CD2L	EOG090F04AY	-0.75	0.021	0.959	0.314	0.018	0.063	0.019	0.058	0.041	0.065	0.033	0.023	0.079	0.017	0.033	0.056	0.037	0.059	0.032
ABR, BCR, RAB36	EOG090F00T9	-1.32	0.021	0.835	0.783	0.000	0.000	0.019	0.056	0.119	0.055	0.106	0.000	0.000	0.032	0.038	0.060	0.129	0.055	0.121
BCAR3, TRNAR-UCU	EOG090F01TE	-0.69	0.021	0.996	0.966	0.037	0.020	0.019	0.048	0.059	0.033	0.037	0.039	0.022	0.026	0.019	0.058	0.066	0.032	0.034
None defined.	EOG090F07XQ	-0.50	0.021	0.938	0.492	0.017	0.304	0.109	0.093	0.022	0.057	0.031	0.017	0.330	0.030	0.189	0.104	0.027	0.071	0.032
ACCS	EOG090F040J	-0.71	0.022	0.990	0.857	0.140	0.339	0.146	0.000	0.017	0.000	0.024	0.137	0.379	0.014	0.108	0.000	0.016	0.000	0.026
PIKFYVE	EOG090F007L	-1.20	0.022	0.993	0.542	0.196	0.374	0.019	0.000	0.000	0.000	0.024	0.203	0.381	0.017	0.016	0.000	0.016	0.000	0.026
ABCD2	EOG090F02LH	-0.86	0.023	0.976	0.462	0.057	0.115	0.026	0.029	0.020	0.039	0.024	0.063	0.112	0.017	0.029	0.029	0.025	0.034	0.026
NFXL1	EOG090F01W5	-0.88	0.023	0.963	0.718	0.011	0.069	0.021	0.486	0.075	0.447	0.072	0.009	0.058	0.017	0.019	0.520	0.086	0.458	0.085
TRIM25	EOG090F039M	-0.39	0.023	0.985	0.645	0.174	0.181	0.022	0.023	0.020	0.043	0.043	0.147	0.136	0.016	0.016	0.025	0.023	0.045	0.043
TOM1	EOG090F052R	-0.75	0.023	0.956	0.450	0.006	0.011	0.095	0.056	0.069	0.073	0.106	0.006	0.007	0.017	0.045	0.056	0.084	0.077	0.111

Table 3.2.3. BMG (A): Number significant GO terms for each CAPER model at each cutoff. Independent variable = LG, natural log of relative gene abundance. MLS = Maximum lifespan, MLSW = residual of BMG (Body mass in grams) on MLS, MLSLW = residual of the natural log of the BMG on MLS.

Tool: Database	Independent variable	De Novo			Alignment			Shared		
		Pos	Neg	All	Pos	Neg	All	Pos	Neg	All
Ingenuity Pathways: Diseases and Functions	MLSLW	6	200	206	6	200	206	6	200	206
TopGO: Biological Processes	MLS	22	32	42	21	46	42	21	32	40
	MLSW	0	95	108	0	95	108	0	95	108
	MLSLW	5	1	8	5	1	8	5	1	8
TopGO: Cellular Components	MLS	10	4	12	2	8	12	2	4	4
	MLSW	0	5	5	0	5	5	0	5	5
	MLSLW	0	11	11	0	11	11	0	11	11
TopGO: Molecular Processes	MLS	5	1	8	5	1	8	5	1	8
	MLSW	0	11	11	0	11	11	0	11	11
	MLSLW	0	21	30	0	25	30	0	21	25

Table 3.2.4a. Significant Biological Processes. CAPER, Dependent = MLSLW, Independent = LG, Metadata = BMG (A). Significant order shown only for significant OG set (p-value cutoff shown).

Shared Order	Direction of Association	Align-Only Significant	De Novo-Only Significant	GO.ID	Term	classic KS	elimKS
1	All	1	1	GO:0031331	positive regulation of cellular cataboli...	0.0055	0.0055
2		3	3	GO:0071310	cellular response to organic substance	0.0114	0.0114
3		5	5	GO:0001818	negative regulation of cytokine producti...	0.012	0.012
4		6	6	GO:0006511	ubiquitin-dependent protein catabolic pr...	0.012	0.012
5		NA	NA	GO:0019941	modification-dependent protein catabolic...	0.012	0.012
6		NA	NA	GO:0043632	modification-dependent macromolecule cat...	0.012	0.012
7		NA	NA	GO:0051603	proteolysis involved in cellular protein...	0.012	0.012
8		4	4	GO:0009605	response to external stimulus	0.0124	0.0124
9		7	7	GO:0007050	cell cycle arrest	0.0144	0.0144
10		8	8	GO:0045786	negative regulation of cell cycle	0.0144	0.0144
11		9	9	GO:0045930	negative regulation of mitotic cell cycl...	0.0144	0.0144
12		NA	NA	GO:0044281	small molecule metabolic process	0.0145	0.0145
13		10	10	GO:0007346	regulation of mitotic cell cycle	0.0155	0.0155
14		11	11	GO:0010564	regulation of cell cycle process	0.0155	0.0155
15		12	12	GO:0022402	cell cycle process	0.0155	0.0155
16		13	13	GO:0044770	cell cycle phase transition	0.0155	0.0155
17		14	14	GO:0044772	mitotic cell cycle phase transition	0.0155	0.0155
18		15	15	GO:0045787	positive regulation of cell cycle	0.0155	0.0155
19		2	2	GO:0055086	nucleobase-containing small molecule met...	0.0155	0.0155
20		16	16	GO:0090068	positive regulation of cell cycle proces...	0.0155	0.0155
21		17	17	GO:1901987	regulation of cell cycle phase transitio...	0.0155	0.0155
22		18	18	GO:1901990	regulation of mitotic cell cycle phase t...	0.0155	0.0155
23		19	19	GO:1903047	mitotic cell cycle process	0.0155	0.0155
24		20	20	GO:0016567	protein ubiquitination	0.0175	0.0175
25		21	21	GO:0032446	protein modification by small protein co...	0.0175	0.0175
26		22	22	GO:0034097	response to cytokine	0.0175	0.0175
27		23	23	GO:0070647	protein modification by small protein co...	0.0175	0.0175
NA		24	NA	GO:1901135	carbohydrate derivative metabolic proces...		
NA		25	NA	GO:1901137	carbohydrate derivative biosynthetic pro...		

Table 3.2.4a. Significant Biological Processes. (Continued.)

Shared Order	Direction of Association	Align-Only Significant	De Novo-Only Significant	GO.ID	Term	classicKS	elimKS
NA	**Pos**			None			
1	**Neg**	3	3	GO:0048522	positive regulation of cellular process	0.00449	0.0045
2		2	2	GO:0009605	response to external stimulus	0.00488	0.0049
3		1	1	GO:0071310	cellular response to organic substance	0.00499	0.005
4		5	5	GO:0009893	positive regulation of metabolic process	0.00659	0.0066
5		4	4	GO:0002684	positive regulation of immune system pro...	0.00754	0.0075
6		9	9	GO:0006950	response to stress	0.00869	0.0087
7		6	6	GO:0009890	negative regulation of biosynthetic proc...	0.00939	0.0094
8		7	7	GO:0031396	regulation of protein ubiquitination	0.01039	0.0104
9		NA	34	GO:1903320	regulation of protein modification by sm...	0.01039	0.0104
10		NA	NA	GO:0001816	cytokine production	0.01211	0.0121
11		NA	NA	GO:0001817	regulation of cytokine production	0.01211	0.0121
12		8	8	GO:0001818	negative regulation of cytokine producti...	0.01211	0.0121
13		NA	NA	GO:0016567	protein ubiquitination	0.01211	0.0121
14		NA	NA	GO:0032446	protein modification by small protein co...	0.01211	0.0121
15		NA	NA	GO:0070647	protein modification by small protein co...	0.01211	0.0121
16		10	10	GO:0045934	negative regulation of nucleobase-contai...	0.0126	0.0126
17		11	11	GO:0051253	negative regulation of RNA metabolic pro...	0.0126	0.0126
18		12	12	GO:0034097	response to cytokine	0.01763	0.0176
19		14	14	GO:0045892	negative regulation of transcription, DN...	0.01893	0.0189
20		15	15	GO:1902679	negative regulation of RNA biosynthetic ...	0.01893	0.0189
21		16	16	GO:1903507	negative regulation of nucleic acid-temp...	0.01893	0.0189
22		NA	NA	GO:0031325	positive regulation of cellular metaboli...	0.01981	0.0198
23		NA	NA	GO:0010629	negative regulation of gene expression	0.01986	0.0199
NA		13	NA	GO:0055086	nucleobase-containing small molecule met...		
NA		17	NA	GO:0009896	positive regulation of catabolic process		
NA		18	NA	GO:0031331	positive regulation of cellular cataboli...		

Table 3.2.4b. Significant Cellular Components. CAPER, Dependent = MLSLW, Independent = LG, Metadata = BMG (A)

Shared Order	Direction of Association	Align-Only Significant	De Novo-Only Significant	GO.ID	Term	classicKS	elimKS
1	**All**	1	1	GO:0043231	intracellular membrane-bounded organelle	0.00014	0.0011
2		4	4	GO:0005654	nucleoplasm	0.00256	0.0026
3		2	2	GO:1904949	ATPase complex	0.009	0.009
4		5	5	GO:0044446	intracellular organelle part	0.00243	0.0182
5		11	11	GO:0016021	integral component of membrane	0.02249	0.0225
6		7	7	GO:0000118	histone deacetylase complex	0.03384	0.0338
7		8	8	GO:0016581	NuRD complex	0.03384	0.0338
8		9	9	GO:0070603	SWI/SNF superfamily-type complex	0.03384	0.0338
9		10	10	GO:0090545	CHD-type complex	0.03384	0.0338
10		NA	NA	GO:0044451	nucleoplasm part	0.03875	0.0388
11		6	6	GO:0016604	nuclear body	0.04199	0.042
12		13	13	GO:0005777	peroxisome	0.04513	0.0451
13		14	14	GO:0005778	peroxisomal membrane	0.04513	0.0451
14		15	15	GO:0031903	microbody membrane	0.04513	0.0451
15		16	16	GO:0042579	microbody	0.04513	0.0451
16		17	17	GO:0044438	microbody part	0.04513	0.0451
17		18	18	GO:0044439	peroxisomal part	0.04513	0.0451
NA		3	NA	GO:0005634	nucleus		
NA		12	NA	GO:0031981	nuclear lumen		
NA		19	NA	GO:0005912	adherens junction		
NA		20	NA	GO:0044428	nuclear part		
NA	**Pos**	NA	NA	None			
1	**Neg**	1	1	GO:0043231	intracellular membrane-bounded organelle	0.000012	0.000039
2		19	19	GO:0044446	intracellular organelle part	0.00013	0.00057
3		4	4	GO:0044451	nucleoplasm part	0.0072	0.0072
4		3	3	GO:1904949	ATPase complex	0.00957	0.00957
5		2	2	GO:0005654	nucleoplasm	0.00068	0.02109
6		NA	NA	GO:0044428	nuclear part	0.0018	0.02379
7		NA	NA	GO:0031981	nuclear lumen	0.00099	0.02699
8		NA	22	GO:0005730	nucleolus	0.02959	0.02959

Table 3.2.4b. Significant Cellular Components (Continued.)

Shared Order	Direction of Association	Align-Only Significant	De Novo-Only Significant	GO.ID	Term	classicKS	elimKS
9	**Neg (ctd.)**	6	6	GO:0000118	histone deacetylase complex	0.03781	0.03781
10		7	7	GO:0016581	NuRD complex	0.03781	0.03781
11		8	8	GO:0017053	transcriptional repressor complex	0.03781	0.03781
12		9	9	GO:0070603	SWI/SNF superfamily-type complex	0.03781	0.03781
13		10	10	GO:0090545	CHD-type complex	0.03781	0.03781
14		11	11	GO:0090568	nuclear transcriptional repressor comple...	0.03781	0.03781
15		NA	NA	GO:0005634	nucleus	0.00457	0.0426
16		NA	21	GO:0016021	integral component of membrane	0.04659	0.04659
17		12	12	GO:0005777	peroxisome	0.04739	0.04739
18		13	13	GO:0005778	peroxisomal membrane	0.04739	0.04739
19		14	14	GO:0031903	microbody membrane	0.04739	0.04739
20		15	15	GO:0042579	microbody	0.04739	0.04739
21		16	16	GO:0044438	microbody part	0.04739	0.04739
22		17	17	GO:0044439	peroxisomal part	0.04739	0.04739
23		5	5	GO:0016604	nuclear body	0.04891	0.04891
24		18	NA	GO:0005912	adherens junction		

Table 3.2.4c. Significant Molecular Functions. CAPER, Dependent = MLSLW, Independent = LG, Metadata = BMG (A)

Shared Order	Direction of Association	Align-Only Significant	De Novo-Only Significant	GO.ID	Term	classicKS	elimKS
1	**All**	1	1	GO:0005524	ATP binding	0.0038	0.0038
2		2	2	GO:0140096	catalytic activity, acting on a protein	0.0169	0.0169
3		3	3	GO:0019902	phosphatase binding	0.0234	0.0234
4		4	4	GO:0019903	protein phosphatase binding	0.0234	0.0234
5		5	5	GO:0045296	cadherin binding	0.0373	0.0373
6		6	6	GO:0050839	cell adhesion molecule binding	0.0373	0.0373
7		7	7	GO:0001103	RNA polymerase II repressing transcripti...	0.0374	0.0374
8		8	8	GO:0042826	histone deacetylase binding	0.0374	0.0374
9		9	9	GO:0070491	repressing transcription factor binding	0.0374	0.0374
10		10	10	GO:0015399	primary active transmembrane transporter...	0.0467	0.0467
11		11	11	GO:0015405	P-P-bond-hydrolysis-driven transmembrane...	0.0467	0.0467
12		12	12	GO:0016887	ATPase activity	0.0467	0.0467
13		13	13	GO:0022804	active transmembrane transporter activit...	0.0467	0.0467
14		14	14	GO:0042623	ATPase activity, coupled	0.0467	0.0467
15		15	15	GO:0042626	ATPase activity, coupled to transmembran...	0.0467	0.0467
16		16	16	GO:0043492	ATPase activity, coupled to movement of ...	0.0467	0.0467
NA	**Pos**	NA	NA	None			
1-16	**Neg**	1-16	1-16	See 'All'			

Table 3.2.5. Significant Ingenuity Pathway Diseases or Functions. CAPER, Dependent = MLSLW, Independent = LG, Metadata = BMG (A)

Order Sh	Dir	P C u	RO A	RO DN	Categories	Diseases or Functions	P Sh	P A	P DN	# Mol Sh	# Mol AL	# Mol DN
1	**Pos**	**5.00E-09**	1	3	Gene Expression, Protein Synthesis	Initiation of translation of mRNA	0.00000113	1.13 E-06	5.19 E-06	3	3	3
2			2	1	Cell Morphology, Endocrine System Disorders, Organ Morphology, Organismal Injury and Abnormalities, Reproductive System Development and Function, Reproductive System Disease	Lack of mitochondrial sheath	0.00000125	1.25 E-06	3.36 E-06	2	2	2
3			3	4	Endocrine System Disorders, Organ Morphology, Organismal Injury and Abnormalities, Reproductive System Development and Function, Reproductive System Disease	Abnormal morphology of cauda epididymis	0.0000035	3.50 E-06	9.41 E-06	2	2	2
4			4	11	Protein Synthesis	Initiation of translation of protein	0.0000212	2.12 E-05	9.63 E-05	3	3	3
5			5	8	Cancer, Developmental Disorder, Hematological Disease, Immunological Disease, Organismal Injury and Abnormalities	Non-gastric extranodal marginal zone lymphoma of mucosa-associated lymphoid tissue	0.0000288	2.88 E-05	7.73 E-05	2	2	2
6			6	12	Organ Morphology, Organismal Development, Reproductive System Development and Function	Mass of epididymis	0.0000437	4.37 E-05	1.17 E-04	2	2	2
25			27	2	Cancer, Organismal Injury and Abnormalities	Non-melanoma solid tumor	4.50E-04	4.50 E-04	3.9E -06	13	13	21

Table 3.2.5. Significant Ingenuity Pathway Diseases or Functions. (Continued, 2 of 3)

Order Sh	Dir	P Cu	RO A	RO DN	Categories	Diseases or Functions	P Sh	P A	P DN	# Mol Sh	# Mol AL	# Mol DN
39	**Pos**	**5.00E-09**	41	5	Cancer, Organismal Injury and Abnormalities	Malignant solid tumor	9.09 E-04	9.09 E-04	1.2E -05	13	13	21
NA			NA	6	Carbohydrate Metabolism, Nucleic Acid Metabolism, Small Molecule Biochemistry	Metabolism of nucleoside diphosphate sugar	NA	NA	3.1E -05	NA	NA	2
1	**Neg**	**5.00E-09**	1	1	Cancer, Organismal Injury and Abnormalities	Tumorigenesis of tissue	8.25 E-13	8.25 E-13	8.3E -13	104	104	104
2			2	2	Cancer, Organismal Injury and Abnormalities	Cancer	6.15 E-12	8.25 E-13	8.3E -13	111	111	111
3			3	3	Auditory Disease	Tinnitus	2.44 E-11	8.25 E-13	8.3E -13	6	6	6
4			4	4	Cancer, Organismal Injury and Abnormalities	Cancer of secretory structure	1.14 E-10	8.25 E-13	8.3E -13	86	86	86
5			5	5	Cancer, Organismal Injury and Abnormalities	Head and neck tumor	1.25 E-10	8.25 E-13	8.3E -13	84	84	84
6			6	6	Cancer, Organismal Injury and Abnormalities	Head and neck carcinoma	7.58 E-10	8.25 E-13	8.3E -13	81	81	81
7			7	7	Cancer, Organismal Injury and Abnormalities	Adenocarcinoma	1.08 E-09	8.25 E-13	8.3E -13	88	88	88
8			8	8	Cancer, Organismal Injury and Abnormalities	Nonhematologic malignant neoplasm	1.13 E-09	8.25 E-13	8.3E -13	99	99	99
9			9	9	Organismal Injury and Abnormalities, Renal and Urological Disease	Renal colic	1.69 E-09	8.25 E-13	8.3E -13	6	6	6
10			10	10	Cancer, Organismal Injury and Abnormalities	Carcinoma	1.99 E-09	8.25 E-13	8.3E -13	97	97	97
11			11	11	Cancer, Organismal Injury and Abnormalities	Extracranial solid tumor	2.23 E-09	8.25 E-13	8.3E -13	102	102	102
12			12	12	Cancer, Endocrine System Disorders, Organismal Injury and Abnormalities	Nonpituitary endocrine tumor	2.7E -09	8.25 E-13	8.3E -13	78	78	78
13			13	13	Cancer, Organismal Injury and Abnormalities	Abdominal neoplasm	3.19 E-09	8.25 E-13	8.3E -13	94	94	94

Table 3.2.5. Significant Ingenuity Pathway Diseases or Functions. (Continued, 3 of 3)

Or der Sh	Dir	P Cu	RO A	RO DN	Categories	Diseases or Functions	P Sh	P A	P DN	# Mol Sh	# Mol AL	# Mol DN
14	**Neg**	**5.00E-09**	14	14	Gastrointestinal Disease, Hepatic System Disease, Organismal Injury and Abnormalities	Drug-induced liver disease	3.43 E-09	8.25 E-13	8.3E -13	6	6	6
15			15	15	Cancer, Endocrine System Disorders, Organismal Injury and Abnormalities	Thyroid gland tumor	3.72 E-09	8.25 E-13	8.3E -13	77	77	77
16			16	16	Connective Tissue Disorders, Organismal Injury and Abnormalities, Skeletal and Muscular Disorders	Ankle sprain	4.66 E-09	8.25 E-13	8.3E -13	5	5	5

DISCUSSION

We have pursued what we believe is the largest study to date investigating the relationship between gene expression levels (via transcript abundances obtained from RNA-seq assays) and lifespan in different bird species. We focused on fibroblast gene expression patterns in 49 bird species that are known to exhibit a wide range in lifespans and developed a workflow and analysis strategy processing and analyzing the RNA-seq data using reference genomes chosen for each species based on their proximity to a species with a well-established reference genome. We also developed strategies for determining orthologous sets of genes across the bird species as well as humans. To ensure that our findings are robust given that not all species we studied had a reference genome, we also assessed transcript abundances using a *de novo* assembly workflow. Similar strategies have been used in other studies of new or poorly characterized species. [10] We also used multiple association analysis strategies as well as different sources for MLS and body size information for each species, noting that our study differs from previous analyses by leveraging the residual of species longevity on body mass in grams (e.g. where 'negative' associations with the residual be interpreted by their inverse correlation to extended lifespan, so 'positive' with regards to the relationship of their abundance, and its effects in transmission networks resulting in extreme longevity.) (Supplemental Figure 9)

Because of the branch lengths in the phylogenetic relationships among the species we studied (Figure 1), it was possible that some genes have been lost or have undergone substantial changes (e.g., at DNA sequence level or via duplications resulting in paralogies and multiple orthologies) as the species evolved. We therefore needed to be sensitive about making claims about orthologous transcripts across the 49 species, and account for expected transcriptomic variation using insights into the evolutionary divergence amongst them. In addition, because of uncertainties in the actual MLS and average body size of each species, the use of multiple sources for them made sense. Finally, since the exact phylogenetic relationships between the species is uncertain, we choose to assess the association of each OG to MLS using different methodologies.

In terms of assessing orthology among transcripts identified from the reference-guided and *de novo* assembly analyses, we tried to leverage different levels of stringency for making claims about orthologous proteins amongst using the OrthoMCL tool [12] Relaxing criteria for stringency with OrthoMCL, or even the number of species considered in the identified of common orthologous transcripts, led to larger sets of transcripts we could associate with longevity at the cost of sensitivity towards rare transcripts and isoform-specific analyses unique within queries and/or their best references. Even the "strictest" clustering rate resulted in a lower correlation across our OGs throughout the data set, when analyzed by linear concordance and CCC rho. [13] For this reason, we decided to use the predefined GigaDB resources as the most accurate for curating comparable summed abundance values. [20] This consistency was also apparent in downstream analyses where our GO ontology searches based on significant terms, highlighting targeted networks related to the source tissue and with consistent terms amongst abundance derivation methods.

Preliminary observations of concordance between the OrthoDB clusters exhibited generally strong correlations across the natural log normalized relative abundance vectors, particularly in deeply covered transcripts. Examples of discordance, measured using linear model residual values and CCC rho q-value, provided a degree of confidence in OG specificity performance and the ability to recapitulate expected linear relationships between OG abundance and the residual of body mass on lifespan across the diverse avian phylogeny (Supplemental Figure 1a-b,e-f). These trends in specificity were more dramatic when representing abundance in rank ordered format (Supplemental Figure 1b, 1d).

Although we tabulated transcript counts across different subsets of the 49 species, we emphasize that we put our focus on the transcripts that we were confident were present across all species. A '0.0 abundance' assignment for a transcript for a particular species could have resulted for one of four reasons. First, the transcript may not be expressed in fibroblasts in that species, though it could have been observed as non-zero expression level in another tissue. This would amount to the transcript having a true 0.0 abundance level assignment for that species. Second, the gene could have been lost due to evolutionary changes between one species and another, creating genomic separation between that species and other species. This would also result in the transcript having a true 0.0 abundance level assignment for that species. Third, the gene could be lost due to a lack of similarity between the observed expressed transcripts and the best available reference genome. Fourth, the RNA-seq assay could have been problematic (e.g., read count from the sequencing was too low to interrogate the transcript). These last two 0.0 abundance assignments would not, necessarily, be biologically relevant. It is unknown which of these four reasons could explain that missing transcript abundance, especially without additional information (e.g., DNA sequencing and genome assembly results for the species). With the hypothesis that some drop-out of abundance signal may be expected due to limitations in the sample preparation pipeline, we limited our sample analysis to those OGs expressed across the cohort, and by pursuing concurrent analysis using the relative, absolute, and rank-ordered representation of expression.

To account for expected transcriptomic variation within sub-clades of our novel avian cohort, we first validated expectations from prior study of a linear correlation between our species' body mass and longevity, accounting for variations in our metadata from various sources. (Figure 2, Supplemental Figure 2). We compared the performance of a simple linear model to the phylogenetic contrast method described by Felsenstein et. al in 1985 and implemented as 'Caper' [12,18]. We also considered a multivariate linear regression model leveraging principal components. We further explored the use of MDMR with a representation of our sample-sample distance matrix using the apparent transcriptome dissimilarity using the MASH tool and via the apparent phylogenetic distance in our source tree. Comparable results for top OGs in all other models are summarized in Table 2 and Supplemental Table 10 as well as in Figure 4-6 and Supplemental Figure 2. We see variable performance amongst the clade in response to each correction method. Results for all models are provided in Supplemental Data Zip.

Many OGs were found to be significant at each of our cutoff tests. The transcripts that exhibited the strongest association with either MLS, MLSW or MLSLW were identified as both associated with positive and negative prediction of longevity vs. the expectations set by species mass, with more negatively predictive OGs identified at extreme significance across all models (Figure 3.) Many of the genes identified by the abundance-derived tests overlapped with significant OGs from our other models, and many of those had defined reference specific gene identifiers with human orthologs. A majority of the same OGs were identified as significant in the "BMG (A)" and "AW (A)" metadata contexts when utilizing the log of body mass, while the variations between metadata inputs dramatically altered the significant gene count shared in the other contexts, with and without using rank-ordered abundance. (Table 1) Several novel significant markers were identified in only one of the two metadata context, including KLHL25 and CHST14. (Table 2) Both metadata contexts performed similarly in the de novo abundance context, though several novel significant markers were identified in the latter context, including UGP2 and ARMC10. (Supplemental Figure 6a)

The significant defined genes from the MLSLW abundance-derived models were involved in a number of TopGO Biological Processes, including “GO:0002684: positive regulation of immune system process” (Table 4a, Supplemental Table 12a). A number of immune regulatory genes were highlighted in our analysis including PAIP2 (Figure 4), the most significant positively correlated gene in our study, and a gene known to regulate viral synthesis and replication in humans. [21]. Significantly associated GO Cellular Components terms such as “GO:0043231: intracellular membrane-bounded organelle” and “GO:0060271 cilium assembly” were highlighted due to genes like the significantly negatively correlated TMEM87A (Figure 5), a Golgi-resident membrane protein which has been highlighted for its associated with intracellular signalling via retrograde transport [22]. The most significant GO terms also include a number of carcinogenic response terms, including “GO:004852 positive regulation of cellular process”, which includes a strongly negative correlated OG defined with symbol ARMC10 (Supplemental Figure 6a), known to decrease activity of the p53 tumor suppressor [23]. Generally, we see wide number of immune response and cell signaling terms which significantly associated with our differentially expressed OG terms.

Significantly associated Cellular Components included a wide array of general fibroblast-specific transcription network activity, but specifically highlighted several complexes expected to vary with longevity association. (Table 4b, Supplemental Table 12b) For example the NuRD complex, a chromatin signaling pathway for cancer regulation in fibroblast cells, and the

SWI/SNF complex, another a chromatin remodeling complex marker controlling development in fibroblasts were both significantly associated with deviations from expectations of longevity [24,25]. The CHD chromatin remodeling regulation network was also significantly associated, which is consistent with significantly associated Molecular Functions like ATPase activity and other markers of transcription regulation. [26] (Table 4c, Supplemental Table 12c)

Ingenuity Pathways' Diseases and Functions highlighted a number of terms relating to tumor development and translation regulation. Significant terms in the "BMG (A)" set were more general, denoting primarily terms associated with tumors in soft tissue and also reproductive morphology (Table 5) while significant terms in the "AW (A)" metadata set included more tissue-specific terms relating to fibroblast cell morphology, especially in the positively associated group (Supplemental Table 13). The most significant grouping of terms seen, both in terms of p-value and their network connectivity, are terms negatively associated with stomach, renal, and lower intestinal cancers including "Cancer, Gastrointestinal Disease, Organismal Injury and Abnormalities" disease terms "Digestive organ" and "Digestive system", and "Organismal Injury and Abnormalities, Renal and Urological Disease" disease term "Renal colic" (Figure 8b, Supplemental Figure 8b). This relationship is shown in example OG with human reference equivalent GCLC, a known liver cirrhosis marker negatively associated with longevity expectation in our study (Figure 6) [cite].

Additional associated OGs can be found for the MLSLW studies in Table 5 and for all models and contexts in the Supplemental Data. This includes a number of other head and neck tumor networks, reproductive terms, and thyroid markers. Some of the most significantly associated OGs were not highlighted as part of a network, but highlight interesting targets like

KLHDC1, whose mutation has been known to lead to progressive blindness in humans. [27] (Supplemental Figure 6b)

Our analyses are a first step in the incorporation of studies of avian longevity into broader studies of genes that contribute to longevity in general. Our studies could benefit from more complete reference genome and transcriptome assemblies and annotations, and from leveraging more complex methods of identifying orthologous relationships to remove false positives caused by a lack of specificity in calculating characterizing transcript abundances. However, we believe our study of birds and the genes that influence their lifespans is important for a number of reasons. First, they have a very long and complicated evolution, creating a great deal of diversity at the phenotype and genetic levels. Second, they are, for at least many bird species, accessible and amenable to study. Third, their lifespans are not incredibly long (like mice, worms and flies), making it possible to study them longitudinally in relatively short periods of time. Fourth, given their divergence from mammalian and other species, their study could provide novel insights into highly conserved mechanisms of aging. However, there are some other disadvantages to our study of birds as well. As noted, birds have not been well studied historically, especially at the genomic level, creating a need for more sophisticated cross-species orthology assessments both within the avian clade and between birds and other species. In addition, we focused on fibroblasts, which may not be the best tissue to study for aging. Despite this, given the very extensive evolutionary histories of bird species, the fact that we identified transcripts strongly associated with lifespan across 49 diverse bird species suggests the existence of shared or conserved factors contributing to avian lifespan. In addition, given that some of these transcripts were associated with genes found to be associated with longevity in mammalian species, and that many of them are known to contribute to immune processes that have been implicated in other species [28,29], our results suggest the existence of elements of conserved processes contributing to longevity across many organisms in nature.

METHODS

Metadata validation and sample exclusion. 3 query species from the sequenced cohort were omitted due to metadata error and inconsistency. First, this included the ring-necked pheasant, for which the best data in the wild are short-term. Second, the Yellow-throated warbler was omitted due to the small number of source material, from only 4 band recoveries and not four known lifespans. Last, the Ostrich was omitted due to the absence of either wild or captive ostrich populations.

RNA-Sequencing (RNA-seq). We pursued whole transcriptome sequencing (WTS) via RNA-seq using standard Illumina HiSeq 2000 sequencing technology and protocols on RNA harvested from fibroblasts obtained on 49 bird species with maximal lifespans ranging from 7.1 to 40 years, as listed in Supplemental Table 1 (8 to 70 years in alternate metadata source, Supplemental Table 5a). Collaborators at X caught species of interest in the wild and fibroblasts isolated from each was expanded in cultures for 2-3 generations so that a 3 or 4th passage cell line could be used, wherein mutations are unlikely to occur. Fibroblast cell lines were derived from sun-protected skin samples of adult birds, as described in earlier papers, and cryopreserved at passages *(details under review for current draft).* [30] Samples were stored in low oxygen and thawed aliquots were expanded to produce approximately *(details under review for current draft)* x 10^6 cells, which were lysed and sent as frozen pellets to the University of California, San

Francisco UCSF for sequencing. RNA was extracted using Ribo-Zero and ScriptSeqv2 library prep followed by placement on 1 lane for each sample on the HiSeq 2000. [31]

Species-Specific Best Comparable Reference Selection. A poor choice of reference genomes, or a lack of accepted and validated de novo transcript assembly and quantification strategies, can adversely affect RNA-seq studies seeking to compare transcript abundances across previously under studied species. [30,32] 42 of the 49 species we studied did not have available individual reference genomes that could be used for transcript assignment and abundance calculations for the RNA-seq analyses using alignment-derived abundance (Supplemental Table 1). When queries did not have finished reference genomes available, we assigned "best defined references", and sample-sample distance, in two ways. The first method leveraged the pairwise distance analysis resulting in the phylogenetic tree discussed above (2012) to identify 10 high quality reference genomes (Supplemental Table 2a, Supplemental Table 2b, analysis not shown). A second method leveraged MASH, to calculate query samples' kmer distribution similarity amongst the transcriptomes of the 49 query species and the 74 most recent references available in GigaDB public resources at the time of publication (Supplemental Figure 3a). [33, 20] A 49x49 sample-sample Euclidean distance matrix based on this 49x74 MASH calculation is provided in Supplemental Table 9b (Supplemental Figure 3b,). Principal component analysis of each sample-sample distance matrix for the 49 query species was conducted to revealed which query species' variance was most highly correlated with each component and the amount of variance associated with each ordinal component. (Supplemental Figure 4a, Supplemental Figure 4b, Supplemental Table 7)

Phylogenetic variance estimation. The phylogenetic distances between our 49 available avian query species (Figure 1, Supplemental Table 9a) was derived from contrasts derived from the multiple sequence alignment analysis described by Jetz et. al. and made available in birdtree [34]. As described above, this source was also used to define the best matching reference for each of the species (Supplemental Table 1). We leveraged the cophenetic.phylo function in the ape package in R to compute the pairwise distances and render using (Figure 1). [35] A sample-sample distance matrix equivalent to the phylogenetic tree derived from this analysis was generated using R package 'adephylo' (Supplemental Table 9a, Supplemental Figure 3c). [36] Principal component analysis of the sample-sample phylogenetic distance matrix for the 49 query was conducted for comparison to the MASH equivalent **(**Figure 5a, Figure 5b**).** Multivariate linear model analysis with ANOVA was used to derive the percentage of population variance described by each principle component, and the association of different subsets of the avian query clade with these components (Supplemental Table 8, Supplemental Table 7, Supplemental Figure 4b-c).

Transcript Alignment to References Genomes and Computing RPKM Transcript Abundance Counts. The paired end sequences derived from RNA-seq runs (forward-reverse, 100 base pairs (bp)) from each of the 50 species were submitted to quality trimming to remove chaff that could result in misalignment using AdapterRemoval and Trimmomatic for read preparation [37,38]. Alignment of the sequencing reads from each of the 50 species to the best matched reference was conducted using HISAT2.[39] Stringtie was used to summarize and transform the counts into abundances and their RPKM values. [40] The resulting output was summarized into a tabular format for comparison across the 49 species obtained with the best-matched reference for each.

De Novo Query Species Contigs and Computing Contig-derived Abundance Counts. A companion paper (Chan et al., in preparation) describes the results of a de novo assembly of transcripts for the 50 species considered here. In brief, the transcriptomic sequencing reads from each of the 50 query avian species were also used to build scaffolded contigs independent of their best reference following the same pipeline of AdapterRemoval and Trimmomatic for read preparation followed by Trinity for assembly [41]. Following assembly, the sequencing reads were aligned to the transcriptomic scaffolds using bowtie2 and the resulting counts were normalized with RSEM [42,43].

Omission of samples with human contaminant identified. Chan et. al. observed that American tree sparrow showed a different species distribution from other birds (i.e. much higher number of human protein matches using the SwissProt tool across the transcriptomic set, with ~150% of the median human matches and ~50% of the median of chicken matches, the closest expected reference in this test. [44] This species was omitted, resulting in the 49 query species documented in the final study. (Supplemental Table 3)

Identifying Orthologous Transcripts Across the 49 Species. In the absence of a single comparable reference which could be used to accurately summarize abundances of all 49 query species, we summed the transcript abundances assigned to orthologous protein groups amongst the references which have been made publicly available, as defined in the GigaDB avian clade. When utilizing our alignment-based approach, each query-specific best-reference-specific transcript abundance was assigned that references' transcript identifier as well as its associated reference-specific protein identifier. These protein identifiers, representing all of the 20 best references derived from the MASH-based transcriptomic kmer similarity analysis, were comparable via their assignment to the defiend OrthoDB clusters in GigaDB. In the de novo-derived abundance context, scaffolded contig-specific abundances were attributed to OrthoDB clusters by running a protein BLAST against the OrthODB. OrthoDB version ID conversions, and the association with associated human reference-specific common symbols was conducted with OrthoDB resources *(version omitted in current draft)*. The count of query species expressing each OG in our data set was calculated and only those OGs with greater than zero abundance for every one of our 49 species was retained for further study. **(**Supplemental Table 4**)** Alternative ortholog clustering (not shown) was pursued using the 2012 birdtree reference definitions, reference-to-reference protein BLAST to identify sequence similarity [45], and strictly clustered ortholog groups defined using cd-hit-est and OrthoMCL. [46,12]

Correlation analysis. We tested the correlation between our alignment- and de novo-derived abundance terms and species mass residual on longevity to evaluate the reproducibility of our abundances, their orthologous clustering, and relationships between metadata contexts. The linear association between the summed transcript abundances associated with each ortholog group was conducted using linear fit and additionally via Lin's correlation coefficient (CCC) rho.q value, revealing disparity between high and low correlation ortholog groups. (Supplemental Figure 1, Table 2, Supplemental Table 10, Figure 4-6, Supplemental Figure 6) The associated correlation term has been provided alongside each OG to each table. Linear fit p-values and R2 terms were also calculated for each OG and are provided in the full tabular format attached in the Supplemental Data.

Testing the Association Between Maximum Lifespan and Weight. It is well known that there is a strong association between maximum lifespan (MLS) and body size across species. To verify this and to determine the degree to which we would have to control for body size in assessing correlations between transcript abundances and MLS, we explored the relationship between weight and MLS using linear regression. Linear residual values retained from these regressions (MLSW) were retained and used as dependent variables in exploring the association between transcript abundance and MLS. (Supplemental Table 5, Supplemental Figure 2a, Supplemental Figure 2c) We also calculated the residuals between MLS and the logarithm (base e) of mass (MLSLW) across the 49 species while correcting for the expected variance in expression defined by the phylogenetic relationships in our literature-derived tree. (Figure 2, Supplemental Figure 2b, nodes) [12,18] Species of exceptional interest or exhibiting genetic longevity traits were highlighted for comparative analysis to the rest of the population based on their orientation outside of 1 or 2 standard deviations from a linear regression line. (Table 1, Figure 2)

Metadata selection and imputation. The linear and phylogeny contrast model residuals of our query-specific avian mass and lifespan were compared in the context of metadata sourced in two ways. First, “O”: novel metadata collected by authors responsible for sample capture, designating wild longevity metadata and traits with sources cited in Supplemental Table 1. Second, “A”: metadata from the AnAge provided for 74 total avian species. [16] AnAge (“A”) provided 2 contexts for weight that could be tested. The first, “Body mass in grams” (BMG) matched the curated set described as “O” above but was populated for only 22 of our 49 query species. The second, “Adult weight in grams” (AW) was populated for all 49 query species in AnAge “A”. AnAge also had maximum lifespan data populated for those 47 species. We tested the use of metadata for each query species calculated the resulting residual of mass on lifespan in each of the regression contexts discussed above (MLS, MLSW, MLSLW) as shown in Supplemental Table 5a, iterating over 1) O: All curated metadata, 2) O+A: All curated metadata, updated with AnAge when populated, and 3) All AnAge metadata. The corresponding caper-derived phylogeny contrast residuals were calculated in Supplemental Table 5b.

Metadata validation via alternative phylogeny adjustment comparison. Given the wide swing in residual performance in both model types, we attempted to validate our selection for publication review via validation of expectations from our known sample dissimilarity. We leveraged multi-dimensional multivariate regression (MDMR) to identify how well the residuals in our comparisons of body mass and lifespan, representative of the deviation of each query from expectations of longevity, were predicted by the phylogenetic tree-derived distance matrix and by the MASH-derived kmer dissimilarity matrix versus the GigaDB avian reference set. **(**Supplemental Table 6**)** Based on these results, we favored the use of the "BMG (A)" and "AW (A)" metadata contexts for in depth review, with all results provided in the Supplemental Data.

Testing Associations Between Common Transcripts and MLS. For each metadata and abundance context, we calculated the linear association of the relative abundance of each transcript present in the 49 species to MLS, weight-adjusted MLS (MLSW, the residual of a linear regression of query bird mass on lifespan) and log weight-adjusted MLS (MLSLW, the residual of a linear regression of the natural log of the query species mass on lifespan). Abundance vectors were normalized to relative abundance within each sample using the R library 'vegan'. [47] Alternative abundance vectors tested the use of rank-ordered abundance, or relative abundance without log normalization (Supplemental Data). To control for phylogenetic relationships between the species, we favored the phylogenetic contrast methods developed by Felsenstein et al., which is commonly used in cross-species analyses as well intra-species comparisons involving subclades or species strains, and implemented as R library 'caper' [12,18]. We compared the phylogenetic contrast modelling approach to the alternate baseline expectations of variance defined by the birdtree-sourced phylogenetic distance matrix and the MASH-derived transcriptome kmer similarity distance matrix for each ortholog group. Given the large number of transcripts tested for association with MLS, MLSW and MLSLW, we also used false discovery rate (FDR) techniques to minimize likely false positive results. [48] P-values for all models and OGs are provided in the Supplemental Zip File along with their slope of association, R2 values, and OG metadata. Significant OGs of interest were selected based on their significance of association, the correlation between the alignment and de novo-derived abundance vectors, and a confirmation in the significance of performance across the contrast and multidimensional multivariate regression modelling techniques. Example cases for visualization were selected based on their prediction of variance in the data set and plotted with trendlines representing each model in Figures 4-6 and Supplemental Figure 6 along with summaries of their abundance correlation terms (in alignment vs. de novo contexts) and the performance in our alternative models accounting for population wide variation based on phylogenetic or transcriptomic sample dissimilarity.

Gene Ontology analyses. P-values and directionality from each ortholog group was assigned to each of that OG's clustered reference-specific proteins using definitions from the reference genomes, and significant OG lists were selected based on a scale of FDR-adjusted cutoffs (Table 2, Supplemental Tables 10, Figure 3, Supplemental Figure 5). GO ontology terms associate with each significant OG were captured from the OrthoDB resources and provided in the Supplemental Data. We also tested for the enrichment of biological process (BP), molecular function (MF), and cellular component (CC) terms among the transcripts exhibiting associations with each model using the topGO package in R. [49] Transcripts were mapped to ENTREZ gene IDs using biomaRt. [50] Network interpretation was conducted across significant transcripts highlighted by cutoff at an FDR-adjusted p-value < 0.05 with ordered preference for genes with greater absolute influence (positive or negative) on lifespan. Fisher's exact tests and the

Kolmogorov-Smirnov test were implemented using R package 'ALL' [51] The results were compared amongst the metadata contexts, the abundance formatting, and the model used **(**Tables 4a-c, Supplemental Tables 12a-c**).** Significant pathway terms below ClassiKS cutoff 0.05, or the most significant term if not exist below the threshold, and their differentially expressed gene associations, were summarized as examples in Figure 7 using R library 'Rgraphviz' **(**Figure 5a-b, Supplemental Figures 7a-d**).** [52] The significant term lists for each GO enrichment type are provided in the Supplemental Zip File.

Ingenuity Pathway analyses. Significant lists of gene symbols below FDR-adjusted cutoff 0.05, for each model, were also submitted for additional pathway analysis using the Ingenuity Pathway Analysis (IPA) services' Functional Analysis to identify significantly correlated Diseases and Functions. [19] Specifically, each list of significant OG terms' defined common symbols were searched against IPA Genes and Targets database, and all matching terms were used as the query list for functional search. The count of all terms surpassing p-value thresholds was calculated for Table 3 and Supplemental Table 11, as for the GO terms described in the previous section. The most significant Ingenuity Pathway terms were summarized at various thresholds to denote their most significant terms (Table 5, Supplemental Table 13). Those significant terms with less than 100 network connections were highlighted in Figure 8 and Supplemental Figure 8.

Chapter 3.2, in part, is currently being prepared for submission for publication of the material. Multi-reference Genome-wide RNA-sequence Analysis of 49 Bird Species identifies Transcripts Associated with Avian Longevity. McCorrison J, Chan AP, Choi Y, Ding K, Pickering A, Pawlikowska L,Norden-Krichmar T, Evans D, Schork NJ, Miller RA. The dissertation author was the primary investigator and lead author of this paper.

3.3: FUTURE WORK: USING THE DE NOVO RECONSTRUCTION OF QUERY REFERENCE GENOMES AS REFERENCE

There are many insights to be gained from this unique avian fibroblast sampling of 52 total bird species, where 3 species were omitted from the previous study due to dubious metadata but remained applicable to de novo analyses. Additional insights into conserved genes apparent in separate sub-clades of our query species populations, variation in genes between clades, and other indicators of genetic diversity are discussed in our drafted manuscript.

Chan AP, Choi Y McCorrison J, Schork NJ. **De novo transcriptome assembly, annotation, and comparison of 52 avian species.** (Current title, analysis not shown.)

I believe this research will only be improved upon over time as more references become available, or all species are able to be compared using isoform-specific abundances and comparison in terms of their known orthologous terms, rather than using their nearest phylogeny- or transcriptomics-derived adjacent neighbors. The ability to compare correlated genes, and networks of genes, expressed with correlation to phylogenetic outcomes of interest is of great value for the identification of potential drug targets and to the understanding of network utility across diverse evolutionary bounds. By studying these divergent co-evolution events, or conserved events, in a de novo context, we may begin to remove some of the biases we encounter in our measurement assay. Until that time, I believe that leveraging the use of multiple techniques, aligning reads to a best reference and summing over comparable orthologous terms versus a de novo technique and orthologous terms, and placing confidence upon the most consistent results, is the best method to pursue the evaluation of shared networks of interest across our diverse host species.

CHAPTER 4: LONGEVITY-ENHANCING DRUG TARGETS

4.1. INTRODUCTION

In the previous chapters, I identified trends in genes which were predictive of phenotypic outcomes and correlated them with other networks of genes known to perform functions in tandem (e.g. transcription networks.) This work has highlighted the ability to find targets in experimental species which can be correlated to host species of clinical interests, particularly, humans. However, additional insights are available if these types of significant lists of terms are compared to public databases for context. In the following chapter, we pursue an investigation of hypothetical gene targets, their overlapping functional variants, and the search for longevity-enhancing drugs that may be correlated with these points of clinical intervention.

The next chapter focuses on genetic variants associated with human longevity from many studies, with focus on longevity-associated phenotypic outcomes and sourced from various tissues and analysis types. We evaluate variants, genes, and variants which are located in those genes, which are considered high quality drug targets. Conversely, we evaluate current drugs hypothesized to impact longevity, and the genes associated with those drugs. The different data sources associated with the information I used to address these questions are all 'metadata'-based' since they merely reflect the results of different studies (e.g., genetic association studies, pharmacologic studies exploring drug targets, etc.). I find that most of the variants associated with longevity are not necessarily good drug targets, given a lack of consensus on the 'druggability' in the pharmacology community. Conversely, most drugs hypothesized to influence longevity – or shown to influence longevity in a non-human species – are not supported by genetic information. However, I believe this work lays a strong foundation for methods to expand the automated detection of potential drug targets in common studies, including those leveraging advanced insight using the exploitation of metadata.

4.2. GENETIC SUPPORT FOR LONGEVITY-ENHANCING DRUG TARGETS: ISSUES, PRELIMINARY DATA, AND FUTURE DIRECTIONS

See published work, reproduced in this chapter:

McCorrison J, Girke T, Goetz LH, Miller R, Schork NJ. **Genetic Support for Longevity-Enhancing Drug Targets: Issues, Preliminary Data, and Future Directions.** *The Journals of Gerontology*: Series A, Volume 74, Issue Supplement 1, December 2019, Pages S61–S71, doi: 10.1093/gerona/glz206.

THE GERONTOLOGICAL SOCIETY OF AMERICA®

Journals of Gerontology: Biological Sciences
cite as: *J Gerontol A Biol Sci Med Sci,* 2019, Vol. 74, No. S1, S61–S71
doi:10.1093/gerona/glz206

OXFORD

Healthy Longevity 2019: Supplement Article

Genetic Support for Longevity-Enhancing Drug Targets: Issues, Preliminary Data, and Future Directions

Jamison McCorrison, BS,[1] Thomas Girke, PhD,[2] Laura H. Goetz, MD,[3,4] Richard A. Miller, PhD,[5,6] and Nicholas J. Schork, PhD[3,7,8,9,*]

[1]Graduate Program in Bioinformatics and Systems Biology, University of California–San Diego. [2]Institute for Integrative Genome Biology, University of California, Riverside. [3]Department of Quantitative Medicine and Systems Biology, The Translational Genomics Research Institute (TGen), Phoenix, Arizona. [4]Department of Medical Oncology, City of Hope National Medical Center, Duarte, California. [5]Department of Pathology and [6]Glenn Center for the Biology of Aging, University of Michigan, Ann Arbor. [7]Department of Population Sciences, City of Hope National Medical Center, Duarte, California. [8]Department of Psychiatry and [9]Department of Family Medicine and Public Health, University of California–San Diego.

*Address correspondence to: Nicholas J. Schork, PhD, Quantitative Medicine and Systems Biology, The Translational Genomics Research Institute (TGen), An Affiliate of the City of Hope National Medical Center, 445 North Fifth Street, Phoenix, AZ 85004. E-mail: nschork@tgen.org

Received: July 22, 2019; Editorial Decision Date: August 5, 2019

Decision Editor: David Le Couteur, MBBS, FRACP, PhD

Abstract

Interventions meant to promote longevity and healthy aging have often been designed or observed to modulate very specific gene or protein targets. If there are naturally occurring genetic variants in such a target that affect longevity as well as the molecular function of that target (eg, the variants influence the expression of the target, acting as "expression quantitative trait loci" or "eQTLs"), this could support a causal relationship between the pharmacologic modulation of the target and longevity and thereby validate the target at some level. We considered the gene targets of many pharmacologic interventions hypothesized to enhance human longevity and explored how many variants there are in those targets that affect gene function (eg, as expression quantitative trait loci). We also determined whether variants in genes associated with longevity-related phenotypes affect gene function or are in linkage disequilibrium with variants that do, and whether pharmacologic studies point to compounds exhibiting activity against those genes. Our results are somewhat ambiguous, suggesting that integrating genetic association study results with functional genomic and pharmacologic studies is necessary to shed light on genetically mediated targets for longevity-enhancing drugs. Such integration will require more sophisticated data sets, phenotypic definitions, and bioinformatics approaches to be useful.

Keywords: Longevity, Mortality, Human Aging

The identification of interventions, such as nutritional supplements, specific diets, and drugs that can reduce age-related disease risk and enhance longevity, is receiving a great deal of attention. The reasons for this are not just rooted in an age-old fascination with mortality, but also the belief that it might be possible to slow the aging process, simultaneously reducing the risk of many age-related diseases and morbidities, maintaining health, and ultimately increasing longevity (1–5). However, the identification of relevant targets for the development of longevity-enhancing drugs, such as specific genes or proteins, is complicated by the fact that human longevity and the aging process are complex and ultimately influenced by a number of genetic and nongenetic factors (3,6–9). This makes it difficult to identify compelling longevity-enhancing drug targets because the effects of any one potential gene or protein target could be obscured by the effects of others.

There are strategies to identify longevity-enhancing drug targets that overcome this complexity, and many have been used with some level of success. For example, many researchers have studied longevity in nonhuman species since relevant experiments can be performed in ways that are not feasible in humans. These studies range from the comparison of, for example, genes and their expression levels across species exhibiting variation in life span (10–13), exploring natural variation in life span among individual animals within a species (14–16), using contrived gene manipulation techniques (such as knocking out a gene or controlling its expression using various genetic engineering strategies) and assessing their effects on life span (17,18), or simply screening drugs against individual animals to see which have a positive effect (19,20). The biological insights into the aging processes from these studies have been varied, with many offering important observations on evolutionarily conserved processes involved in aging. However, given the fundamental differences between humans and other species at the molecular and physiologic levels, it is still an open question as to whether these insights can be readily translated into findings that can form the basis for human longevity-enhancing interventions (21).

An alternative to identifying drug targets involving nonhuman species is to use human genetic studies, in particular genome-wide association studies (GWAS), which seek to identify naturally occurring DNA sequence variants that are associated with, for example, longevity, healthspan, and susceptibility/resistance to diseases. A number of GWAS have been pursued to date that have focused on human longevity, healthspan, and protection from disease (3,14,22–26). Unfortunately, the results of many of these studies have not been replicated, in part due to the multifactorial nature of human longevity, but also due to difficulties in assembling relevant cohorts necessary for such studies (eg, large numbers of very long-lived individuals; large cohorts with longitudinal data reflecting health trajectories over time, etc.) (27).

Despite complications with many GWAS initiatives, it has been shown that a drug designed to modulate or affect a gene or protein target, which harbors variants associated with the specific disease that the drug was designed to treat, actually yields better outcomes during the drug development process than a drug that targets a gene that does not harbor such variants (28–31). This is plausible because naturally occurring genetic variations that have an impact on a phenotype of relevance must work through some mechanism that could, in theory, be modulated pharmacologically (32). Many success stories exist in which drugs have been developed that target or modulate genes harboring variants associated with a specific disease (see, eg, research on the development of Ivacaftor for cystic fibrosis (33) and *PCSK9* inhibitors for treating hypercholesterolemia (34)). In fact, these successes are consistent with, and have motivated, the growing interest in tailoring medicine to individuals' genetic (and other) profiles via precision medicine initiatives (35,36).

Identifying drug targets based on genetic association study results is not trivial, however, because the mere association of a genetic variant with a phenotype, especially one as complex as longevity, is insufficient. One must identify the actual molecular mechanism or process through which the variant alters the phenotype before a genetic association can reveal a viable drug target. Unfortunately, many variants found to be associated with longevity-related phenotypes via GWAS—as well as most other phenotypes—do not actually reside in genes or their more obvious surrounding regulatory elements. As an example, these variants may reside in intergenic regions or regulatory elements, which are not well characterized, making their immediate functional effects hard to discern (37). In addition, the genes that harbor variants are not always found to be amenable to pharmacologic modulation (ie, they might not be "druggable") (38). Finally, many variants associated with a disease or trait, even those in genes that are thought to be druggable, do not implicate or suggest specific or obvious mechanisms for pharmacologic modulation. For example, it might not be obvious whether or not a variant causes overexpression of a gene in a specific tissue of relevance whose pharmacologic inhibition would lead to consistent and favorable phenotypic outcomes (39). As a result, the use of genetic information to identify or prioritize drug targets is likely to require integrated approaches which draw on the insights from a number of disciplines beyond genetics, including molecular, systems and evolutionary biology, genomics, pharmacology, and chemoinformatics (see the Discussion section for more detail) (2,3,27,40).

One approach to determining whether a variant found to be associated with, for example, longevity is likely to reveal a viable drug target is to determine whether that variant is also known to influence, or correlate with, a molecular phenotype that could be amenable to pharmacologic modulation in a tissue of relevance. For example, if the associated variant is known to affect the expression level of a gene (ie, is an "expression quantitative trait loci" or "eQTL") or the abundance of a particular protein (ie, is a "pQTL") in muscle, cardiac, or brain cells, there is a possibility that the variant influences this molecular phenotype in a causal pathway leading from the variant to the longevity phenotype. Evidence for a causation would make that molecular phenotype a logical longevity-enhancing drug target (41). In fact, databases are available that catalog variants that have been found to be eQTLs, pQTLs, or other molecular phenotypes, as in the GTex database (42). In addition, statistical strategies have been developed to test the hypothesis that, for example, an eQTL, or other molecular (or "intermediate") phenotype, is in a causal pathway leading from the relevant variant to an overt, clinically meaningful, phenotype like longevity (43).

We surveyed the available literature and interrogated a number of resources focused on associations with genetic variants, and their known effects, to determine whether the research community might be able to exploit information about genetic associations involving longevity and longevity-related phenotypes to identify possible longevity-enhancing drug targets. We pursued this in two ways. We first identified a number of drugs thought to be candidates for enhancing longevity in humans based on their effects on longevity in nonhuman species, their mechanisms of action, and/or their impacts on age-related diseases. We then determined whether there was evidence that those drugs modulate or target genes harboring variants associated with human longevity-associated phenotypes and/or a potential mechanism amenable to pharmacologic modulation (eg, if the variants are known eQTLs or pQTLs.)

We also identified individual variants found to be associated with human longevity, healthspan, and disease protection based on GWAS (3,22–25,44). We then determined whether these longevity phenotype-associated variants, or more precisely the alternate alleles at the locus harboring the variants, are associated with age-related disease phenotypes. We also determined whether these variants reside in druggable genes, are known QTLs (eQTLs, pQTLs, etc.), or are in linkage disequilibrium (LD > 0.8) with variants that are QTLs and/or associated with other phenotypes. We cataloged eQTLs in LD with the sourced eQTLs, as well as variants associated with longevity phenotypes because they give an indication of how complex a regulatory setting a target gene may be operating within. For example, if perturbations within a gene induced by naturally occurring sequence variants have ripple effects involving a number of other genes, then modulating that gene pharmacologically could affect multiple pathways or molecular networks, for better or worse. For genes harboring associated variants, we also determined whether there were pharmacology studies published in the chemoinformatics literature suggesting that a drug or compound exhibited activity against genes whose DNA sequences were homologous to that gene (45,46). Activity against a homologous gene sequence may suggest that the drug or compound in question may also exhibit activity against the gene harboring the associated variant. It may also indicate that, if the homologous gene has a similar function to the gene harboring the associated variant, the modulation of that gene may affect the longevity-related phenotype.

We emphasize eQTLs as relevant molecular phenotypes in much of our analyses because they have received the most attention in the literature, have the most information about their influence on different tissues, and have the most resources cataloging them. In addition, by focusing on eQTLs in potential target genes and their LD relationships to other variants, we expose the potential that a variant associated with longevity could reveal a molecular mechanism worthy of scrutiny as a drug target. We admit that there may be other variants in the genes of interest that are not themselves eQTLs, nor in LD with eQTLs, that may actually induce or contribute to an as-yet uncharacterized molecular function that could be pharmacologically modulated. We also emphasize that the definition of longevity is widely debated and a crucially important topic for putting aspects of our findings into perspective. We make no attempt to resolve this debate but rather use the published data based on the authors' definitions of longevity to make broader claims about genetic information and putative longevity-enhancing drugs and drug targets. Figure 1 provides a schematic summarizing our sources of information for longevity-associated variants as well as putative longevity-enhancing drugs. Figure 1 also provides the main databases and query tools used in our analyses, which are discussed in greater detail in the Methods section.

Methods

Longevity Drug and Compound Data Sources

There are many drugs and compounds that have been hypothesized to influence human longevity based on a wide variety of studies (see, eg, the DrugAge database (47)). We limited the number of drugs we considered in the present analysis to those receiving the most attention based on our internet and literature searchers, although we are pursuing more complete analyses of a larger collection of drugs. Note that many "drugs" we list are actually experimental compounds not yet approved for use but are rather in various stages of development. We first considered the drugs and compounds found to significantly increase longevity in mice from the NIA-sponsored Interventions Testing Program (ITP) (20). The ITP follows a rigorous standard protocol to test drugs and compounds for their effects on mouse longevity. We also considered drugs and compounds that have been proposed to be evaluated in human clinical trials based on the credibility of the published science behind them. URLs and relevant references with information describing these efforts are provided in Supplementary Tables where appropriate. Finally, we considered the drugs and compounds ranking highly as likely to affect human longevity based on the systems biology and cross-species analysis of Fuentealba and colleagues (48) because the authors did *not* include more comprehensive human genetic association study result information in their otherwise very thorough analysis of candidacy and properties of the drug targets.

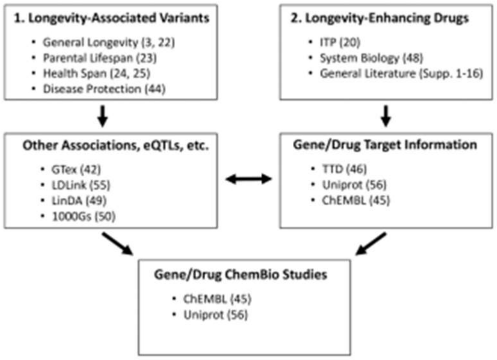

1. Longevity-Associated Variants
• General Longevity (3, 22)
• Parental Lifespan (23)
• Health Span (24, 25)
• Disease Protection (44)
2. Longevity-Enhancing Drugs
• ITP (20)
• System Biology (48)
• General Literature (Supp. 1-16)
Other Associations, eQTLs, etc.
• GTex (42)
• LDLink (55)
• LinDA (49)
• 1000Gs (50)
Gene/Drug Target Information
• TTD (46)
• Uniprot (56)
• ChEMBL (45)
Gene/Drug ChemBio Studies
• ChEMBL (45)
• Uniprot (56)

Figure 1. Schematic of the resources used for both determining: (i) if a gene harboring a longevity-associated variant is a reasonable drug target and (ii) if there is genetic evidence supporting the targets of potential longevity-enhancing drugs. Note that numbers in parentheses denote references.

Identifying Variants and Their Associations in the Gene Targets of Longevity-Enhancing Drugs

For each putative longevity-enhancing drug and compound we considered, we identified the primary gene targets of those drugs and compounds using the Therapeutic Target Database (TTD) (45). We emphasize that TTD, although well curated, does not contain exhaustive information about drug targets that could be obtained from an analysis of, for example, the downstream effects of a drug on genes in a particular pathway. For each gene target, we used the LinDA web-based query tools (49) to determine the number of eQTL variants within them that could reflect compelling genetically mediated molecular phenotypes for drug development purposes (eg, pharmacologic modulation studies of the expression of a gene that varies naturally between long-lived and short-lived individuals or between carriers and noncarriers of specific genetic variants). Note that we used conventional statistical significance criteria also used on the website to determine eQTL status, though of course it would be important to explore how the use of different criteria would change our findings. For each of the eQTL variants we also used the LinDA query tools to identify variants that were in LD with these eQTLs (LD > 0.8) that were associated with (i) longevity; (ii) diseases and other clinical phenotypes; and/or (iii) other molecular phenotypes (eQTLs; sQTLs [splicing QTLS]; aseQTLs [allele-specific expression QTLs]; polyAQTLs [alternative polyA QTLs]; repeatQTLs [repeats expansion expression-level QTLs]; pQTLs: dhsQTL [DnaseI hypersensitive sites QTLs]; hQTLs [Histone modifications QTLs]; mQTLs [DNA methylation QTLs]). Variants in LD with eQTLs were based on the use of the European cohort from the 1000 Genomes Project (50). Note that we chose an LD cutoff of 0.8 because the effect sizes of most variants associated with longevity are weak. If a true functional variant is in weak LD with a variant with a weak effect on longevity, it is difficult to argue that the influence of that variant on longevity is due to the molecular phenotype induced by the variant for which it is in weak LD. Of course, further studies assuming different LD strengths could be revealing and should be pursued. We did not weigh the evidence for reported phenotypic associations, but merely point to published studies claiming an association with a particular phenotype. Further exploration is needed to accurately assess the strength of the evidence for each association and how it may support the belief that the gene harboring that variant is a reasonable drug target. For eQTL information, we summarized studies involving different tissues using the GTex database (42). To summarize genetic association information, we summed the number of associated variants (for longevity and other phenotypes, including the molecular phenotypes) falling into various categories and report these categories here.

Associated Variants With Longevity, Healthspan, and Disease Protection Sources

We gathered information about genetic variants associated with human longevity, healthspan, and protection against disease from multiple publications. For variants associated with longevity, we used the recent review by Partridge and colleagues focusing on variants with replication studies (3,22,51–54), the meta-analysis of GWAS studies by Sebastiani and colleagues (22), and the GWAS study of parental life span using the UK Biobank and LifeGen study data by Timmers and colleagues (23). For variants associated with healthspan, we used the study involving the UK Biobank by Zenin and colleagues (24) as well as the study on *TTR* gene variants by Hornstrup and colleagues (25). For genes harboring rare variants that appear to protect individuals from getting certain diseases, we used the list in the review by Harper and colleagues (44).

Characteristics and Drug Information for Genes Harboring Variants Associated With Longevity-Related Phenotypes

For variants associated with longevity, healthspan, and protection against diseases, we first identified the genes reported to harbor, be near, or modulated by, the variant from the publications cited. We then determined these genes' druggability based on information in Chembl (45) and TTD (46). We determined if the associated variants were themselves eQTLs using the LinDA resources (49). We also used the LinDA resources to determine whether the associated variants were in LD (>0.8) with variants associated with longevity, other age-related phenotypes, or various molecular phenotypes (eg, eQTLs or pQTLs). To capture additional information about the various tissues affected by the eQTLs, we used the LDLink resources and query tools using the European cohort information of the 1000 Genomes Project (50,55). We queried each of the single-nucleotide polymorphisms (SNPs) in LD with other variants associated with various phenotypes, and summed the count of disease/trait associations across different disease categories to get a total number of SNPs. When eQTLS were in LD with a longevity phenotype-associated variant, we also catalogued the various tissues that these eQTLs affected—based on the GTex database (42). For each protein-coding gene harboring the variants, we identified its protein equivalent in UniProt (56) using the BioConductor package "Uniprot.ws." Known and experimental drugs targeting these proteins were identified with custom functions querying a downloaded SQLite instance of the ChEMBL database (Version 24) (45). The drug-target annotation functions developed for this step have been implemented as an R software package (Girke and colleagues, manuscript in review). Representative drugs were provided for any gene encoding druggable proteins identified from both TTD and ChEMBL. Because many genes in the human genome are part of gene families, we included in an extra panel of our drug-target annotation routine all nearest neighbor proteins. They had to share with each protein, encoded by a variant harboring gene, a sequence identity of at least 90% based on the UniRef90 entries in UniProt. Including nearest neighbor protein sequences is important because closely related proteins are usually targeted by the same drugs. Yet, drug development and screening efforts can only focus on one or a few targets within a protein family. Thus, incorporating these family relationships reduces the false-negative rate of our approach.

Results

Candidate Longevity-Enhancing Drugs and Compounds

ITP drugs

We first considered the drugs and compounds shown to have an effect on longevity in mice from the ITP (20). Table 1 summarizes the results. For some drugs and compounds, multiple target proteins are listed in the TTD (46). We note that experimental evidence may not suggest that a drug exhibits activity against all of these targets. In addition, information in the TTD (and other databases) may simply be wrong. Note that despite its having a positive effect on longevity based on ITP studies, we did not include the dietary supplement Protandim (a mixture of milk thistle, bacopa extract, ashwagandha, green tea extract, and turmeric extract) because it is not a defined active small molecule with a specific target as indicated in the TTD. Protandim may affect a number of genes possibly relevant to longevity based on association studies; however, from Table 1, it can be seen that none of the drugs and compounds with an effect on mice interact with human gene targets that harbor variants associated with longevity. However, some drugs target genes that harbor variants associated with age-related diseases (eg, 17-alpha-estradiol and target gene *ESR1*). In addition, it is important to point out that 17-alpha-estradiol and acarbose exhibited sex-specific effects in mice based on the ITP studies, complicating their relationships to a target gene or protein. Many other target genes harbor eQTLs and other variants that are in LD with many other phenotypes and eQTLs that affect longevity-related tissues (eg, acarbose targets gene *MGAM* with variants in LD with eQTLs affecting the expression levels of genes in the brain). These findings do not suggest that the drugs found to affect longevity, as defined by the ITP, in mice will not affect human longevity, but rather that variation in the genes they target are not overtly associated with human longevity.

Proposed longevity-enhancing clinical trial drugs

We identified multiple drugs that have been proposed as potential longevity-enhancing compounds and for which some claim about them being evaluated in a clinical trial has been made (see Supplementary Tables for references). Supplementary Table 1 describes the results. None of the reported drugs targets a gene that harbors a variant associated with longevity. A few genes (eg, *AlkiSi*, alternatively named *TGFBR1*) harbor variants that have been shown to be associated with age-related diseases, but the reproducibility of these associations needs to be considered and explored. Many of the drugs listed in Supplementary Table 1 target gene products that harbor variants that are themselves eQTLs, are in LD with variants that affect gene function, or are associated with age-related diseases and phenotypes (eg, Fisetin and the *FABG* gene; J147 and the *ATP5A1* gene). This provides evidence that a rich genetically mediated set of phenomena exists that could make these genes even more compelling longevity-enhancing drug targets, either through their ability to stave off age-related diseases, slow the aging rate, or both, if explored in greater depth.

Highly ranked drugs by Fuentealba and colleagues

Fuentealba and colleagues (48) conducted a series of analyses to evaluate the evidence that certain drugs target genes that, if modulated, are likely to affect fundamental processes implicated in aging and longevity. These analyses leveraged state-of-the-art systems biology analyses and databases and resulted in two lists of prioritized drugs. The first list (Table 1 in Fuentealba and colleagues (48))

Table 1. Human Genetic Information Related to Drugs That Exhibited Statistically Significant Positive Effects on Mouse Longevity From the NIA-Sponsored Interventions Testing Program (20)

Drug/Compound Information				Variants		Associations Involving Variants in LD With Target Gene eQTLs							eQTL Tissues							
Drug	Gene Target	TTD Mechanism	Indication	eQTLs	# LD	Long	Age Rel	Other	LD eQTLs	LD pQTLs	LD mQTLs	LD Other	Adipose	Artery	Brain	Heart	Muscle	Skin	WB	Total
Rapamycin	*OPRK1*	A	C	8	185	0	0	0	27	0	8	12	0	0	59	0	0	0	0	59
Rapamycin	*MTOR*	I	C	24	1,753	0	22	7	548	0	143	217	0	0	2	0	9	17	212	437
Rapamycin	*FKBP1A*	B	C	33	524	0	0	0	178	0	16	82	15	32	1	0	9	27	1	132
17Aalpha estradiol	*ESR1*	A	M	15	890	0	42	8	48	0	206	47	0	53	0	0	1	3	0	204
17-alpha-estradiol	*ESR2*	A	M	28	881	0	10	25	360	0	81	284	0	0	58	9	226	486	76	1,348
Acarbose	*MGAM*	M	D	41	3,167	0	1	2	4,789	26	143	1,256	0	0	1,599	0	0	0	9	2,134
NDGA	*ERBB2*	M	P	12	531	0	0	25	569	0	159	134	7	0	0	110	0	161	0	736

Notes: TTD mechanism = mechanism of action per the Therapeutic Target Database, where A = agonist, I = inhibitor, B = binder, and M = modulator (TTD) (46). Indication = indication of drug on disease where C = CAS, multiple myeloma, M = menopause, D = diabetes, cardiovascular disease, and P = prostate cancer. eQTLs = number of eQTLs in the gene target based on the LinDA eGENE query tool (49). # LD = number of variants in linkage disequilibrium (LD > 0.8 among the 100 genomes European cohort) with the eQTL variants in the gene target per the LinDA eGENE query. Long = number LD variants associated Longevity from the literature based on the LinDA eGENE GWAS summary. Age Rel = comparable number LD variants with GWAS associations to age-related diseases (cancer, cardiovascular disease, metabolic diseases such as diabetes, osteoporosis, and other age-related bone diseases, and Alzheimer's and Parkinson's diseases). Other = number LD variants with GWAS associations associated other phenotypes; LD eQ, LD pQ, LD mQ, LD Other = number of variants eQTL, pQTL and mQTLs themselves in LD (>0.8) with variants with eQTLs in the gene target based on the LinDA molecular QTL summary. eQTL Tissues = number of variants associated with the Gene Target in GTex that affect certain tissues where WB = whole blood and Total = sum of eQTLs including all other tissues.

considers drugs whose gene targets contribute to processes and networks of relevance to longevity and aging. Of the drugs in this list, six drugs had been shown to influence longevity in nonhuman species (resveratrol, genistein, simvastatin, epigallocatechin gallate, celecoxib, and sirolimus). The second list of drugs (Table 2 in Fuentealba and colleagues (48)) was based on multiple criteria including their reported biological activity. Of the drugs in this list, three drugs had been shown to influence longevity in nonhuman species: trichostatin, geldanamycin, and celecoxib. Despite the sophistication of the approach taken to identify candidate drugs for enhancing human longevity, Fuentealba and colleagues (48) did not consider human genetic support in the form of GWAS, eQTL, and other association studies. We note that we could not identify information necessary to conduct our assessments for a few of the drugs listed by Fuentealba and colleagues (48) including cAMP, epigallocatechin gallate, dorsomorphin, doxorubicin, selenium, indole-3 carbinol, cisplatin, and etoposide. Also, the potential side effects of many of these drugs in humans need further attention given their use as chemotherapeutic agents.

Supplementary Table 2 (reflecting drugs in Table 1 of Feuntealba and colleagues (48)) and Table 2 (reflecting drugs in Table 2 of Feuntealba and colleagues (48)) provide the results of our assessments of the genetic support for the two lists of drugs. It is interesting that none of the drugs/compounds target a gene that harbors variants associated with longevity. However, all of the target genes harbor eQTLs, suggesting that functional variants affect those genes. In addition, several of the drugs and compounds target genes, which contain variants in LD with eQTLs that are associated with many age-related diseases and additional functional variants such as eQTLs, pQTLs, and mQTLs. For those target genes enriched for eQTLs, many of the tissues affected by these eQTLs are important in aging (eg, the *NOS2* gene targeted by resveratrol; the *HSP90AA1* gene targeted by tanespimycin and geldanamycin).

Variants Associated With Longevity-Related Phenotypes

Variants associated with longevity

The review on aging research by Partridge and colleagues (3) discusses a number of variants in specific genes that have been shown to be strongly associated with human longevity. Table 3 provides our assessment of those variants and genes. We find that at least two of the genes harboring longevity-associated variants are not considered druggable because they are noncoding genes and hence not considered in the TTD (*LINC02227* and *USP2-AS1*). Two of the variants are themselves eQTLs, suggesting that they could reveal mechanisms for their association with longevity that could motivate the genes they reside in as potential drug targets (*FOXO3A* and *RAD50/IL13*). Many of the variants were in LD with eQTLs, and other variants were associated with a wide variety of phenotypes, with the exception of the rs28926173 variant in the *MC2R* gene and the rs139137459 variant in *USP2-AS1*. These SNPs do not appear to be in strong LD with other variants of functional significance, raising questions about the biology behind their associations with longevity. Interestingly, two of the genes harboring longevity-associated variants have been the focus of pharmacologic studies (eg, the #Ch columns in Table 3): *FOXO3A* and *MC2R*. Further exploration of the studies focused on *FOXO3A* suggests that resveratrol (which has been extensively studied and considered a candidate longevity-enhancing drug despite not exhibiting effects on longevity in mice) has an effect on that gene (57) and that the efficacy of

Table 2. Human Genetic Information Related to Drugs Identified by Fuentealba and Colleagues as Being Good Candidate for Promoting Healthy Aging Based onThese Drugs' Multiple Levels of Biological Action (48)

Drug/Compound Information						Variants		Associations Involving Variants in LD With Target Gene eQTLs							eQTL Tissues							
Drug	Extend	Toxic	Status	Gene Target	Mechanism	eQTLs	# LD	Long	Age Rel	Other	LD eQTLs	LD pQTLs	LD mQTLs	LD Other	Adipose	Artery	Brain	Heart	Muscle	Skin	WB	Sum
Tanespimycin	N	N	I	*HSP90AA1*	I	19	401	0	0	4	103	0	18	57	35	4	3	0	92	250	0	393
Imatinib	N	N	A	*KIT*	I	8	511	0	0	11	133	0	133	42	0	0	0	0	0	0	0	114
Imatinib	N	N	A	*PDGFRB*	I	22	425	0	1	7	160	0	404	106	0	32	0	22	1	0	0	107
Imatinib	N	N	A	*ABL1*	I	18	734	0	0	3	32	0	7	9	0	3	0	4	0	0	0	19
Sunitinib	N	N	A	*KDR*	M	13	309	0	0	0	16	0	4	0	0	3	5	0	0	0	0	812
Trichostatin	Y	N	E	*HDAC1*	I	15	1,054	0	0	1	50	0	3	11	4	4	5	4	2	5	0	32
Geldanamycin	Y	N	I	*HSP90AA1*	I	19	401	0	0	4	103	0	18	57	0	4	3	0	92	250	0	358
Sorafenib	N	N	A	*KDR*	M	13	309	0	0	0	16	0	4	0	0	3	5	0	0	0	0	812
Sorafenib	N	N	A	*KIT*	M	8	511	0	0	11	133	0	133	42	0	0	0	0	0	0	0	114
Sorafenib	N	N	A	*PDGFRB*	M	22	425	0	1	7	160	0	404	106	0	32	0	22	1	0	0	107
Dasatinib	N	N	A	*SRC*	I	20	216	0	0	2	273	0	46	25	0	0	0	0	0	0	0	6
Dasatinib	N	N	A	*ABL1*	I	18	734	0	0	3	32	0	7	9	0	3	0	4	0	0	0	19
Dasatinib	N	N	A	*LCK*	I	15	1,792	0	0	0	52	2	28	19	0	0	10	0	0	0	0	36
Dasatinib	N	N	A	*FYN*	I	34	618	0	1	5	193	0	56	46	0	184	0	1	0	0	0	373
Erlotinib	N	N	A	*EGFR*	I	17	87	0	1	0	23	0	6	0	0	0	8	9	7	146	0	245
Celecoxib	Y	N	A	*PTGS2*	I	17	1,154	0	0	10	194	5	21	130	0	0	0	0	0	0	0	9

Notes: See Table 1. Extend = evidence that the drug can increase life span in model species per Fuentealba and colleagues (48); toxic = evidence exists that the drug is toxic per Fuentealba and colleagues (48). Extend: Y = yes, N = no. Toxic: Y = yes, N = no. Status: I = Investigational, A = approved, E = experimental. Mechanism: I = inhibitor, M = modulator.

Table 3. Variant Effect Annotations and Drug-Target Information on Variants Found to Be Associated With Human Longevity as Reviewed by Partridge and Colleagues (3)

Associated Variant Information				Druggable?		Annotations		Associations Involving Variants in LD With Target SNP							Chem Studies on Gene				
SNP	Gene	Refs	Chrom	PCh	TTD	eQTL?	# LD	Long	Age Rel	Other	LD eQ	LD pQ	LD mQ	LD O	#Ch	#Ch A	#TTD	#I/M	#Ant
rs6857	*APOE*	22	19	N	Y	N	1	0	18	2	0	0	0	0	0	0	2	2	0
rs2149954	*LINC02227*	51	5	N	N	Y	34	2	10	0	1	0	1	3	0	0	0	0	0
rs10457180	*FOXO3A*	52	6	Y	N	Y	26	0	17	8	2	0	1	0	9	2	0	0	0
rs2706372	*RAD50/IL13*	53	5	Y	N	Y	106	0	0	17	4	0	18	5	0	6	8	5	1
rs2892613	*MC2R*	54	18	Y	Y	N	1	0	0	0	0	0	0	0	34	4	1	0	0
rs139137459	*USP2-AS1*	54	11	N	N	N	1	0	0	0	0	0	0	0	0	0	0	0	0

Notes: See Table 1. SNP = SNP found to exhibit a statistically significant association with human longevity. Druggable PCh, TTD = the gene harboring the associated variants status as "druggable" according to ChEMBL (45) and the TTD (46), respectively, where Y = yes and N = no. eQTL = whether or not the associated variant is an eQTL based on GTEX query, where Y = yes and N = no (42). Chem Studies on Genes: #Ch = number of ChEMBL (45) entries with reported activity against gene sequences homologous to the sequence of the gene harboring the longevity-associated variant. #Ch A = number of ChEMBL entries with in-depth annotation information without leveraging paralogs, #TTD = number of entries for the gene in the TTD (46). #I/M = number of drugs in the TTD that modulate or agonize the gene. #Ant = number of drugs in the TTD antagonize the gene.

corticotropin administration is influenced by variants in the *MC2R* gene (58). The eQTL effects of the longevity-associated *FOXO3A* variant rs10457180 are on tibial nerve and artery tissue, making their relevance to pharmacologic modulation and potential connection to resveratrol in need of further investigation.

The meta-analysis of four GWAS studies described by Sebastiani and colleagues (22) led to the identification of 8 longevity-associated variants, including an *APOE* variant. Supplementary Table 3 provides our assessment of those variants. We note that Sebastiani and colleagues (22) did compile some information about eQTLs in LD with those variants but did not have access to the most recently developed tools and databases. Unfortunately, only one of the genes harboring longevity-associated variants is thought to be druggable, though a few of the genes have variants, which are in LD with other variants exhibiting functional effects (eg, the rs7185375 variant in the *SIAH1* gene and the rs72834698 variant in the *HIST1H2BD* gene).

Timmers and colleagues (23) conducted a very large GWAS on parental life spans as a proxy for an individual life span among participants in the UK Biobank Study and LifeGen study data (23). This study focused on natural variation in life span and not exceptional longevity as a unique phenotype. They identified 12 associated variants using standard GWAS (reviewed in our Supplementary Table 4a) as well as 7 additional variants using a Bayesian analysis that accommodated mortality risk factors in the association test with longevity (reviewed in our Supplementary Table 4b). Our assessment of these variants again suggests that many are within genes that are not thought to be druggable, or at least within gene products for which no known or experimental drugs are available, despite many being in LD with a variant associated with a wide variety of age-related diseases, phenotypes, and functional effects. A few of the genes harboring longevity-associated variants have been the focus of a large number of pharmacologic studies (eg, the *HTT* gene harboring the rs61348208 variants and the *ATXN2* gene harboring the rs11065979 variant).

Timmers and colleagues (23) also pursued an in-depth set of analyses seeking to identify genes whose expression levels are in likely variant-mediated causal pathways involving longevity based on Mendelian randomization tests (43). Table 4 provides the results of our assessments of these variants. Note that the authors identified multiple genes whose expression levels passed statistical criteria for being in a causal pathway from an associated variant to longevity. We present information about the reported associated variants only, but note that other variants in each of the implicated genes that might be of interest. Unfortunately, only a few of these variants implicate genes in a causal pathway involving longevity that are thought to be druggable, although all of them, being eQTLs themselves, are in LD with a number of variants associated with other phenotypes and a wide variety of functional variants. Many of the tissues affected by the longevity-associated eQTL variants are relevant to aging (eg, skeletal muscle, associated variant rs4970836 within gene *CELSR2*).

Variants associated with healthspan

Zenin and colleagues recently pursued a GWAS exploring the age at which an individual likely succumbed to disease and was in a healthy state prior to that age over their life and identified 12 "healthspan"-associated variants (24). Along with these health-enhancing variants, we also gathered information about a variant in the *TTR* gene that has been shown to influence healthspan and longevity as discussed by Hornstrup and colleagues (25). Supplementary Table 5 provides the results, with the first 12 rows corresponding to the variants identified by Zenin and colleagues (24) and the last row corresponding

Table 4. Variant Effect Annotations and Drug-Target Information on eQTL Variants With Statistically Significant Evidence That They Are in a Causal Pathway Associated With Human Longevity Based on Analyses by Timmers and Colleagues (23)

Associated Variant Information					Drug?		Annot.	Associations Involving Variants in LD With Target SNP								Chem Studies on Gene				
SNP	Gene	Chr	Example Tissue	Number of Tissues	PCh	TTD	eQTL?	# LD	Long	Age Rel	Other	LD eQ	LD pQ	LD mQ	LD O	#Ch	#Ch A	#TTD	#I/M	#Ant
rs429358	*AC006126.4*	19	Testis	1	N	N	Y	1	2	14	32	0	0	1	0	0	0	0	0	—
rs429358	*CEACAM19*	19	Thyroid	2	N	N	Y	1	2	14	32	0	0	1	0	0	0	0	—	—
rs429358	*APOC1*	19	Esophagus mucosa	1	N	N	Y	1	2	14	32	0	0	1	0	0	0	0	—	—
rs11065979	*ALDH2*	12	Sun-exposed skin	1	Y	Y	Y	24	0	95	93	22	0	39	1	0	1	3	2	6
rs11065979	*CUX2*	12	Muscle skeletal	1	N	N	Y	24	0	95	93	22	0	39	1	0	0	0	—	—
rs1230666	*AP4B1-AS1*	1	Transformed fibroblasts	1	N	N	Y	1	0	1	2	0	0	0	0	353	0	0	—	—
rs3130507	*CCHCR1*	6	Sun-exposed skin	2	N	N	Y	18	0	0	0	2	0	5	2	0	0	0	—	—
rs3130507	*PSORS1C1*	6	Artery aorta	4	N	N	Y	18	0	0	0	2	0	5	2	0	0	0	—	—
rs4970836	*CELSR2*	1	Muscle skeletal	3	N	N	Y	11	0	16	56	52	0	2	11	0	0	0	—	—
rs4970836	*PSRC1*	1	Liver	3	N	N	Y	11	0	16	56	52	0	2	11	0	0	0	—	—
rs6224	*FES*	15	Transformed fibroblasts	9	N	Y	Y	10	0	6	0	11	0	3	0	1,835	0	2	0	2
rs6224	*FURIN*	15	Artery aorta	1	N	Y	Y	10	0	6	0	11	0	3	0	511	0	3	0	3
rs6224	*RCCD1*	15	Brain cerebellum	2	N	N	Y	10	0	6	0	11	0	3	0	0	0	0	—	—
rs113160991	*PMS2P3*	7	Esophagus muscularis	6	N	N	Y	13	0	0	0	30	0	1	15	0	0	0	—	—
rs8042849	*RP11-650L12.2*	15	Lung	2	N	N	Y	41	2	13	149	13	0	10	16	0	0	0	—	—
rs111333005	*SLC22A1*	6	Pituitary	6	N	N	Y	33	0	0	0	6	0	0	0	437	0	0	—	—

Notes: See Table 3. A dash (—) in a cell indicates that we could not find information about the gene or variants in the gene with the resources we used. Drug = Druggable? Annot. = Annotations. PCh: Y = yes, N = no. TTD: Y = yes, N = no. eQTL: Y = yes, N = no.

to the *TTR* variant discussed by Hornstrup and colleagues (25). As with the longevity-associated variants, many of the variants found to be associated with healthspan are not located in genes coding for proteins for which drug-target information is available, although many themselves are eQTLs or in LD with eQTLs and variants associated with nonlongevity phenotypes.

Variants shown to be protective against diseases

We finally considered genes that harbor multiple rare variants that have been associated with protection against diseases (ie, they seem to confer health benefits to those that possess them) as reviewed Harper and colleagues (44). These protective variants may or may not protect against all or most age-related diseases, however. Given the number and rarity of the variants exhibiting protective effects, we considered the properties of the genes they were identified in, rather than the individual variants themselves. Supplementary Table 6 provides the results. All but one (*SLC30A8*) is considered to be druggable, which makes sense since these genes have gathered a great deal of interest as drug targets. There are many eQTLs and variants associated with phenotypes other than longevity, but whether these variants are in LD with the rare variants thought to have functional effects, and induce the positive phenotypes they are associated with, needs further exploration.

Discussion

The genomics era has resulted in a number of major initiatives focusing on naturally occurring human genetic variation, such as the Human Genome Project (59), the International Hapmap Project (60), the 1000 Genomes Project (50), and The Cancer Genome Atlas (TCGA) project (61), among many others. We considered the utility of genetic information in prioritizing or validating drug targets for longevity-enhancing interventions. We identified drugs and compounds thought to have potential to enhance human longevity, collated naturally occurring variants in the gene and protein targets for these drugs, and looked to see whether these variants have been associated with longevity, or if they influence the molecular functions of those targets. We also brought together, from published literature, lists of variants allegedly associated with longevity, healthspan, or protection from disease, and asked if the genes they reside in are reasonable targets for drug development.

The fact that we found that no drug hypothesized to modify human longevity targets a gene that harbors variants found to be associated with longevity to date and that no associated variants have led to a longevity-enhancing drug development campaign, suggests, among other things, that (i) many proposed drugs are not targeting relevant biology related to human longevity (at least as revealed by GWAS); (ii) the genetic associations from GWAS are too weak and ambiguous to reveal compelling targets; and/or (iii) more comprehensive data sets and studies are needed to make genetic association data "actionable" at some level. We believe that the third explanation is likely the best because we did find that there is an incredibly rich biology uncovered by the effects of genetic variants on mechanisms, like gene expression levels, that could be exploited in drug-target identification studies with more systematic analysis. In this light, our work can be seen as a starting point for more comprehensive assessments of genetically mediated biological targets for longevity-enhancing drugs. For example, we believe our work can motivate more sophisticated consideration of genetic association and, for example, eQTL and pQTL studies in work like that of Partridge and colleagues (48) and Cardoso and colleagues (62), which seek to integrate different sources of information in analyses designed to prioritize drugs and biomarkers for further study (27). In fact, a very recent study exploring the utility of genetic association studies in drug target analysis for immune-related diseases provides an excellent example of the type of integration that we feel is necessary (63). Unfortunately, the data sets and information that the authors exploited, including study results using assays on humans interrogating processes known to be of fundamental importance to immune diseases, are lacking with respect to human longevity. It is noteworthy, however, that the National Institutes on Aging (NIA) of the U.S. National Institutes of Health (NIH) have recently funded initiatives designed to generate more sophisticated data and methods that could enable longevity-enhancing drug-target identification and validation (eg, the Longevity Consortium (https://www.longevityconsortium.org) and the Longevity Genomics (https://www.longevitygenomics.org) initiatives.

Given the hype surrounding genetic studies and the somewhat humbling results of our studies, which suggest no obvious connections between genetic associations and drugs currently hypothesized to enhance human longevity exist, we believe our work exposes a number of serious shortcomings with the use of genetic data for identifying, prioritizing, or validating drug targets for human longevity that are also touched on in the study by Fang and colleagues (63). We describe a few of these shortcomings below—many of which are relevant to our very specific analyses—but feel these descriptions are less of a focused critique of what we have produced and more of an indication of what needs to be done going forward, so that better integration of genetic information into bioinformatics analyses can be pursued (27).

Exploitation of Results of GWAS Involving Other Ancestral Groups

We used variant and LD information obtained from individuals of European descent, though many variants are population specific and/or exhibit different LD relationships in individuals of non-European descent.

Consideration of Different Levels of LD

We chose to only consider variants with an LD strength > 0.8 for target variants or those in LD with eQTLs within a gene. Different LD strengths could provide a different picture of the functional landscape of a gene.

Consideration of the Direction of Effect of a Variant's Associations

A variant could increase or decrease, for example, the expression level of a gene. If this variant is associated with a relevant phenotype as well, then the direction of effect on gene expression level could indicate whether a drug should enhance or antagonize the expression of the gene to achieve the same phenotypic effect.

Leveraging Pleiotropy and Unpacking Diseases Associated With Variants

We cataloged variants associated with many age-related diseases, but if many variants are associated with the same disease, this provides a different picture of the pleiotropic effects of the gene than if many variants are associated with different diseases.

Unpacking the Number of Associations for a Gene

We summed up the number of variants associated with different phenotypes, but the resulting sum may involve different variants in varying degrees of LD or variants in very strong LD. These two scenarios have different biological consequences, wherein variation induced by a gene's functional differences attributable to individual variants is due to a single haplotype that deviates from the others functionally or whether there are multiple haplotypes (alleles) that each differs from the others. In addition, the mere assignment of a variant to an individual gene can be problematic if the variant resides in DNA sequence that does not encode a particular gene or if the sequence does encode a gene but that gene is alternatively spliced such that the variant may not affect all forms of protein translated from that gene's sequence.

Exploring the Effects of Multiple eQTLs Within a Gene

A gene that harbors many eQTLs, pQTLs, etc. is likely to regulate a wider range of molecular phenomena. This could indicate that the gene participates in a network filled with feedback and redundancy mechanisms, which could affect its candidacy as a drug target.

Making Better Use of Orthology Information

Many drugs and compounds have been tested in model organisms for their effects on longevity, such as those pursued by the ITP (20), but the relevance of the effects observed in model species to humans is an open question. Exploring the degree of homology between nonhuman and human genes and incorporating this information into cross-species analyses may be useful in this context.

Better Phenotyping and Indices of Health

Individual life span is a very crude phenotype and does not capture the underlying "subclinical" aspects of health that might exist in people who die early of nongenetic causes (eg, accidents, malnutrition, war, etc.) nor what might be possessed by people who would have died earlier without extensive health care or a favorable but rare environment. Therefore, better measures of underlying robustness, vitality, and functional enhancements (eg, muscle strength, excellent vision, etc.) are needed for GWAS and related studies.

Better Molecular Phenotyping of Longevity-Related Processes

Disease-focused research communities often exploit extensive molecular phenotyping (eg, lipid biology in cardiovascular disease, inflammation in rheumatic disease, etc.) to help put drug targets and potential interventions into biological perspective. Researchers investigating longevity need better phenotyping of aging-related processes, such as "rate of aging measures" or measurable facets of the hallmarks of aging, that could be subjected to GWAS (64,65). This activity could lead to better biomarkers of aging to be considered in causal analyses of longevity (see below).

Incorporating Biomarker Data

eQTLs, pQTLs, and so forth capture the effects of variants on measurable molecular phenotypes. These molecular phenotypes could themselves be tested for association with longevity (eg, a gene's expression level in whole blood or skin may correlate with longevity). Many molecular phenotypes have been treated as biomarkers and tested for association with longevity and aging-related phenotypes (62). Information about whether such biomarkers are associated with longevity-related phenotypes could help solidify causal chains leading from a variant to a longevity-related phenotype, but the tissue in which that biomarker has been measured is important to consider. Note that systematically testing such causal chains for prioritization is crucial if there are many such potential causal connections (43).

Exploiting the Power of Network Biology

The role of a gene within a broader network of genes is important for placing drug candidates into context. For example, a gene may harbor a variant associated with a longevity phenotype, but its druggability has not been demonstrated yet. However, if that gene is known to modulate another gene in an extended causal chain leading from the variant to the longevity phenotype, then the gene that is modulated by the other could be thought of as a drug target. Thus, including network module and pathway information into studies like ours may be of crucial importance.

As we have emphasized, although our efforts to compile and process relevant information on the genetic support for longevity-enhancing drug targets are hardly exhaustive, we pursued it to show the potential, and limitations, of the use of such information. We ultimately find that there is a great deal of potential in using genetic association information for longevity-enhancing drug-target studies, particularly with respect to prioritization and lead development. However, we also believe that more detailed and focused mining of the information, along with relevant query tool and resource development, will be necessary to have a broader impact. We hope that our efforts will motivate the pursuit of appropriate studies and tool development.

Supplementary Material

Supplementary data are available at *The Journals of Gerontology, Series A: Biological Sciences and Medical Sciences* online.

Funding

This paper was published as part of a supplement sponsored and funded by AARP. The statements and opinions expressed herein by the authors are for information, debate, and discussion, and do not necessarily represent official policies of AARP.

Acknowledgments

The authors want to thank the members of the Longevity Consortium, including associated NIA staff, for comments and suggestions as this work was being pursued. Again, we emphasize, however, that the content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest

None reported.

References

1. Vaiserman A, Lushchak O. Implementation of longevity-promoting supplements and medications in public health practice: achievements, challenges and future perspectives. *J Transl Med*. 2017;15:160. doi:10.1186/s12967-017-1259-8

2. Kaeberlein M, Rabinovitch PS, Martin GM. Healthy aging: the ultimate preventative medicine. *Science*. 2015;350:1191–1193. doi:10.1126/science.aad3267
3. Partridge L, Deelen J, Slagboom PE. Facing up to the global challenges of ageing. *Nature*. 2018;561:45–56. doi:10.1038/s41586-018-0457-8
4. Mahmoudi S, Xu L, Brunet A. Turning back time with emerging rejuvenation strategies. *Nat Cell Biol*. 2019;21:32–43. doi:10.1038/s41556-018-0206-0
5. Longo VD, Antebi A, Bartke A, et al. Interventions to slow aging in humans: are we ready? *Aging Cell*. 2015;14:497–510. doi:10.1111/acel.12338
6. Broer L, van Duijn CM. GWAS and meta-analysis in aging/longevity. *Adv Exp Med Biol*. 2015;847:107–125. doi:10.1007/978-1-4939-2404-2_5
7. Brooks-Wilson AR. Genetics of healthy aging and longevity. *Hum Genet*. 2013;132:1323–1338. doi:10.1007/s00439-013-1342-z
8. Crocco P, Montesanto A, Dato S, et al. Inter-individual variability in xenobiotic-metabolizing enzymes: implications for human aging and longevity. *Genes (Basel)*. 2019;10. doi:10.3390/genes10050403
9. Dato S, Rose G, Crocco P, et al. The genetics of human longevity: an intricacy of genes, environment, culture and microbiome. *Mech Ageing Dev*. 2017;165(Pt B):147–155. doi:10.1016/j.mad.2017.03.011
10. Ma S, Gladyshev VN. Molecular signatures of longevity: insights from cross-species comparative studies. *Semin Cell Dev Biol*. 2017;70:190–203. doi:10.1016/j.semcdb.2017.08.007
11. Cohen AA. Aging across the tree of life: the importance of a comparative perspective for the use of animal models in aging. *Biochim Biophys Acta Mol Basis Dis*. 2018;1864:2680–2689. doi:10.1016/j.bbadis.2017.05.028
12. Tian X, Seluanov A, Gorbunova V. Molecular mechanisms determining lifespan in short- and long-lived species. *Trends Endocrinol Metab*. 2017;28:722–734. doi:10.1016/j.tem.2017.07.004
13. Singh PP, Demmitt BA, Nath RD, Brunet A. The genetics of aging: a vertebrate perspective. *Cell*. 2019;177:200–220. doi:10.1016/j.cell.2019.02.038
14. Hook M, Roy S, Williams EG, et al. Genetic cartography of longevity in humans and mice: current landscape and horizons. *Biochim Biophys Acta Mol Basis Dis*. 2018;1864:2718–2732. doi:10.1016/j.bbadis.2018.01.026
15. Bogue MA, Peters LL, Paigen B, et al. Accessing data resources in the mouse phenome database for genetic analysis of murine life span and health span. *J Gerontol A Biol Sci Med Sci*. 2016;71:170–177. doi:10.1093/gerona/glu223
16. Ackert-Bicknell CL, Anderson LC, Sheehan S, et al. Aging research using mouse models. *Curr Protoc Mouse Biol*. 2015;5:95–133. doi:10.1002/9780470942390.mo140195
17. Zhang P, Judy M, Lee SJ, Kenyon C. Direct and indirect gene regulation by a life-extending *FOXO* protein in *C. elegans*: roles for GATA factors and lipid gene regulators. *Cell Metab*. 2013;17:85–100. doi:10.1016/j.cmet.2012.12.013
18. Ocampo A, Reddy P, Martinez-Redondo P, et al. In vivo amelioration of age-associated hallmarks by partial reprogramming. *Cell*. 2016;167:1719–1733.e12. doi:10.1016/j.cell.2016.11.052
19. Ye X, Linton JM, Schork NJ, Buck LB, Petrascheck M. A pharmacological network for lifespan extension in *Caenorhabditis elegans*. *Aging Cell*. 2014;13:206–215. doi:10.1111/acel.12163
20. Nadon NL, Strong R, Miller RA, Harrison DE. NIA Interventions Testing Program: investigating putative aging intervention agents in a genetically heterogeneous mouse model. *EBioMed*. 2017;21:3–4. doi:10.1016/j.ebiom.2016.11.038
21. Hackam DG, Redelmeier DA. Translation of research evidence from animals to humans. *JAMA*. 2006;296:1731–1732. doi:10.1001/jama.296.14.1731
22. Sebastiani P, Gurinovich A, Bae H, et al. Four genome-wide association studies identify new extreme longevity variants. *J Gerontol A Biol Sci Med Sci*. 2017;72:1453–1464. doi:10.1093/gerona/glx027
23. Timmers PR, Mounier N, Lall K, et al. Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *Elife*. 2019;8. doi:10.7554/eLife.39856
24. Zenin A, Tsepilov Y, Sharapov S, et al. Identification of 12 genetic loci associated with human healthspan. *Commun Biol*. 2019;2:41. doi:10.1038/s42003-019-0290-0
25. Hornstrup LS, Frikke-Schmidt R, Nordestgaard BG, Tybjærg-Hansen A. Genetic stabilization of transthyretin, cerebrovascular disease, and life expectancy. *Arterioscler Thromb Vasc Biol*. 2013;33:1441–1447. doi:10.1161/ATVBAHA.113.301273
26. Sebastiani P, Solovieff N, Dewan AT, et al. Genetic signatures of exceptional longevity in humans. *PLoS One*. 2012;7:e29848. doi:10.1371/journal.pone.0029848
27. Schork NJ, Raghavachari N; Workshop Speakers and Participants. Report: NIA workshop on translating genetic variants associated with longevity into drug targets. *Geroscience*. 2018;40:523–538. doi:10.1007/s11357-018-0046-7
28. Sanseau P, Agarwal P, Barnes MR, et al. Use of genome-wide association studies for drug repositioning. *Nat Biotechnol*. 2012;30:317–320. doi:10.1038/nbt.2151
29. Nelson MR, Tipney H, Painter JL, et al. The support of human genetic evidence for approved drug indications. *Nat Genet*. 2015;47:856–860. doi:10.1038/ng.3314
30. Hurle MR, Nelson MR, Agarwal P, Cardon LR. Impact of genetically supported target selection on R&D productivity. *Nat Rev Drug Discov*. 2016;15:596–597. doi:10.1038/nrd.2016.187
31. King EA, Davis JW, Degner JF. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *bioRxiv*. 2019. doi:10.1101/513945
32. Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. *Nat Rev Drug Discov*. 2013;12:581–594. doi:10.1038/nrd4051
33. Feng LB, Grosse SD, Green RF, Fink AK, Sawicki GS. Precision medicine in action: the impact of ivacaftor on cystic fibrosis-related hospitalizations. *Health Aff (Millwood)*. 2018;37:773–779. doi:10.1377/hlthaff.2017.1554
34. Paton DM. *PCSK9* inhibitors: monoclonal antibodies for the treatment of hypercholesterolemia. *Drugs Today (Barc)*. 2016;52:183–192. doi:10.1358/dot.2016.52.3.2440527
35. Aronson SJ, Rehm HL. Building the foundation for genomics in precision medicine. *Nature*. 2015;526:336–342. doi:10.1038/nature15816
36. Zahn LM. Unleashing the power of precision medicine. *Science*. 2016;354:1546–1548. doi:10.1126/science.354.6319.1546-j
37. Schork AJ, Thompson WK, Pham P, et al.; Tobacco and Genetics Consortium; Bipolar Disorder Psychiatric Genomics Consortium; Schizophrenia Psychiatric Genomics Consortium. All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS Genet*. 2013;9:e1003449. doi:10.1371/journal.pgen.1003449
38. Finan C, Gaulton A, Kruger FA, *et al*. The druggable genome and support for target identification and validation in drug development. *Sci Transl Med*. 2017;9. doi:10.1126/scitranslmed.aag1166
39. Floris M, Olla S, Schlessinger D, Cucca F. Genetic-driven druggable target identification and validation. *Trends Genet*. 2018;34:558–570. doi:10.1016/j.tig.2018.04.004
40. Oprea TI, Bologa CG, Brunak S, et al. Unexplored therapeutic opportunities in the human genome. *Nat Rev Drug Discov*. 2018;17:377. doi:10.1038/nrd.2018.52
41. Brown AA, Viñuela A, Delaneau O, Spector TD, Small KS, Dermitzakis ET. Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nat Genet*. 2017;49:1747–1751. doi:10.1038/ng.3979
42. eGTEx_Project. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nat Genet*. 2017;49:1664–1670. doi:10.1038/ng.3969
43. Zheng J, Baird D, Borges MC, et al. Recent developments in Mendelian randomization studies. *Curr Epidemiol Rep*. 2017;4:330–345. doi:10.1007/s40471-017-0128-6
44. Harper AR, Nayee S, Topol EJ. Protective alleles and modifier variants in human health and disease. *Nat Rev Genet*. 2015;16:689–701. doi:10.1038/nrg4017

45. Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017. *Nucleic Acids Res*. 2017;45:D945–D954. doi:10.1093/nar/gkw1074

46. Li YH, Yu CY, Li XX, et al. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res*. 2018;46:D1121–D1127. doi:10.1093/nar/gkx1076

47. Tacutu R, Thornton D, Johnson E, et al. Human ageing genomic resources: new and updated databases. *Nucleic Acids Res*. 2018;46:D1083–D1090. doi:10.1093/nar/gkx1042

48. Fuentealba M, Dönertaş HM, Williams R, Labbadia J, Thornton JM, Partridge L. Using the drug-protein interactome to identify anti-ageing compounds for humans. *PLoS Comput Biol*. 2019;15:e1006639. doi:10.1371/journal.pcbi.1006639

49. IRGB. LinDA (LINkage Disequilibrium-based Annotation) Browser. Cagliari, Province of Cagliari, Italy: Università degli Studi di Cagliari; 2019. http://linda.irgb.cnr.it/.

50. 1000_Genomes_Project_Consortium, Auton A, Brooks LD, Durbin RM, et al. A global reference for human genetic variation. *Nature*. 2015;526:68–74. doi:10.1038/nature15393

51. Deelen J, Beekman M, Uh HW, et al. Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Hum Mol Genet*. 2014;23:4420–4432. doi:10.1093/hmg/ddu139

52. Broer L, Buchman AS, Deelen J, et al. GWAS of longevity in CHARGE consortium confirms *APOE* and *FOXO3* candidacy. *J Gerontol A Biol Sci Med Sci*. 2015;70:110–118. doi:10.1093/gerona/glu166

53. Flachsbart F, Ellinghaus D, Gentschew L, et al. Immunochip analysis identifies association of the *RAD50/IL13* region with human longevity. *Aging Cell*. 2016;15:585–588. doi:10.1111/acel.12471

54. Pilling LC, Kuo CL, Sicinski K, et al. Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging (Albany, NY)* 2017;9:2504–2520. doi:10.18632/aging.101334

55. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*. 2015;31:3555–3557. doi:10.1093/bioinformatics/btv402

56. The_UniProt_Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 2017;45:D158–D169. doi:10.1093/nar/gkh131

57. Kwon SH, Choi HR, Kang YA, Park KC. Depigmenting effect of resveratrol is dependent on *FOXO3a* activation without *SIRT1* activation. *Int J Mol Sci*. 2017;18. doi:10.3390/ijms18061213

58. Ding YX, Zou LP, He B, Yue WH, Liu ZL, Zhang D. *ACTH* receptor (*MC2R*) promoter variants associated with infantile spasms modulate *MC2R* expression and responsiveness to *ACTH*. *Pharmacogenet Genomics*. 2010;20:71–76. doi:10.1097/FPC.0b013e328333a172

59. Green ED, Watson JD, Collins FS. Human genome project: twenty-five years of big biology. *Nature*. 2015;526:29–31. doi:10.1038/526029a

60. International HapMap Consortium. The international hapMap project. *Nature*. 2003;426:789–796. doi:10.1038/nature02168

61. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, *et al*. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;45:1113–1120. doi:10.1038/ng.2764

62. Cardoso AL, Fernandes A, Aguilar-Pimentel JA, et al. Towards frailty biomarkers: candidates from genes and pathways regulated in aging and age-related diseases. *Ageing Res Rev*. 2018;47:214–277. doi:10.1016/j.arr.2018.07.004

63. Fang H, De Wolf H, Knezevic B, et al.; ULTRA-DD Consortium. A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nat Genet*. 2019;51:1082–1091. doi:10.1038/s41588-019-0456-1

64. Marioni RE, Harris SE, Shah S, et al. The epigenetic clock and telomere length are independently associated with chronological age and mortality. *Int J Epidemiol*. 2016. doi:10.1093/ije/dyw041

65. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. *Cell*. 2013;153:1194–1217. doi:10.1016/j.cell.2013.05.039

Chapter 4.2, in full, is a reprint of the material as it appears in Genetic Support for Longevity-Enhancing Drug Targets: Issues, Preliminary Data, and Future Directions. 2019. McCorrison J, Girke T, Goetz LH, Miller R, Schork NJ. The dissertation author was a primary investigator and lead author of this paper.

CHAPTER 5: CONCLUSION

In any experiment or study of naturally-occurring phenomena, it is not unusual for a researcher to attempt to identify some core truths about a phenomenon of interest by gathering together different data or analysis results and determining if they are consistent or reveal something collectively that they cannot individually. Alternatively, one may generate, and then test, a specific hypothesis by using an initial ‘training’ set followed by an analysis of a ‘test’ set using insights obtained from the analysis of the training set. Finally, a researcher may simply aggregate massive amounts of individual data points thought to be complementary and seek to identify patterns in those data that may reveal some new biological insight. All these efforts assume that the different sources of data are not without so much error as to be useless and have been generated in ways that defy their harmonization. Information about how the data were generated and in what context – i.e., metadata – can be used to guide analyses of those data as well as interpretations of the results of those analyses.

In each experiment I considered, the results of many of the analyses were informed by the subject-specific metadata and helped shape the hypotheses that were generated and tested. The metadata shed light on why the data, in some contexts, would not have been ‘harmonizable’ or appropriate to analyze together if appropriate analytical methods were not used that made specific use of the metadata. Some of the analysis results may have been foreshadowed or obvious (e.g. RNA sequencing quality score across single neuron sequencing samples are clear indicators of the reliability of the results), but what was unknown was the degree to which the data needed to be analyzed with the metadata taken into account or the need for novel methods to accommodate an analysis that takes into account the metadata in a practical setting. Other analyses I pursued were less obvious, until placed under more sophisticated statistical scrutiny (e.g. the available data about the mechanisms of actions of drugs and their gene targets thought to extend lifespan and data about the genes harboring variants that are associated with human lifespan do not overlap). By leveraging statistical methods developed in a wide variety of contexts with very different scientific orientations (e.g., traditional microbiome sequencing analysis methods used in evaluating the effect on mood of a digital therapeutic), I show that a greater focus on metadata-aware data science as a methodology can be used to expand the scope of research in many different fields.

My analyses incorporating metadata in single cell RNA sequencing studies was particularly interesting because of the depth of complications that arise from the very negative consequences of having poor quality control measures in those studies. This has been well-documented; for example, in 2012 the first methodologies were being developed to account for the ability to isolate and amplify RNA from an individual bacteria.[1] Now, only 8 years later, FACS laser sorting and automated high-throughput sequencing has enabled the rapid isolation and sequencing of much smaller, and much 'cleaner' representations of individual cell types, but come with a number of guidelines as to how to ensure that the data have been generated reliably so that data aggregation efforts can be pursued.[2,3]

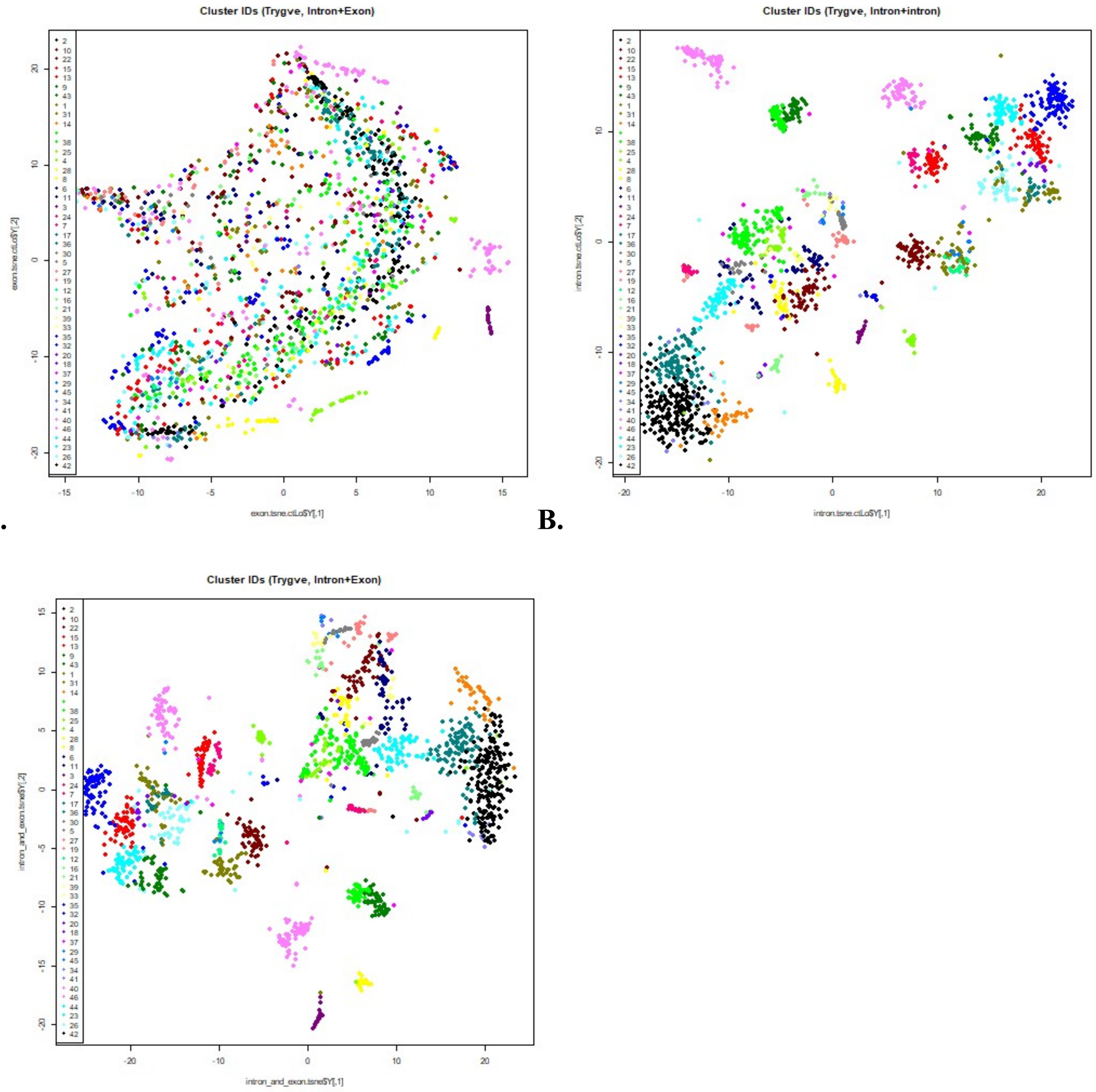
As I have shown, ignoring the manner in which data have been generated and yet combining data sets comes at the cost of biases introduced at many stages of an analysis of those data. Basically, the underlying complexities introduced by varying experimental protocols that are cobbled together complicate downstream analyses, but could also be the secret to motivating methods to improve aggregated analyses and improve interpretations of aggregated analyses. For example, through a feedback loop of early project planning for metadata capture, the recording of actual sequencing protocols used, designing appropriate information capture strategies through each step in an assay and data generation process, and the use of rigorous statistical analysis techniques, I have shown the value of leveraging metadata to improve and develop one of the first high-throughput analysis pipelines for aggregated analysis of multiple data sets (Chapter 2.2), which many teams have now leveraged outside of our group to identify new target cell types in a variety of human tissue including the brain and lung [4,5].

Furthermore, I showed that the use of methods that leverage metadata can also be used to model and measure the degree to which a signal is lost due to a poor appreciation of meta-data. For example, without a well-informed choice of which reference genomes should be used in a cross-species transcriptomics study, it is unclear if the resulting transcript abundances and ortholog determinations reflect true transcript abundances. I suggested that the ability to leverage information which may already be a part of the broader knowledge-base, i.e., bird phylogeny, can be used to improve the confidence with which inferences about the relationships between transcripts identified across the species can be made. By using insights into the amount of variation in relevant phenotypes (e.g., transcript abundance) explained by metadata about the generation of the transcript profiles, my colleagues and I were able to make more advanced decisions about how samples inform an aggregated data set.

All of the results of my studies must be considered in light of efforts within the community at large to generate data, store those data, and make those data available to the scientific community for aggregated and integrated analysis. The efforts to make data available must come with an understanding of how important it is to also make available as much metadata about how those data were generated as possible. Ultimately, there is no doubt that as the scientific community becomes more comfortable in the digital era in which large-scale data collections become routine, the role of metadata in relevant data analyses will become more pronounced.

APPENDIX

APPENDIX A (2.4)

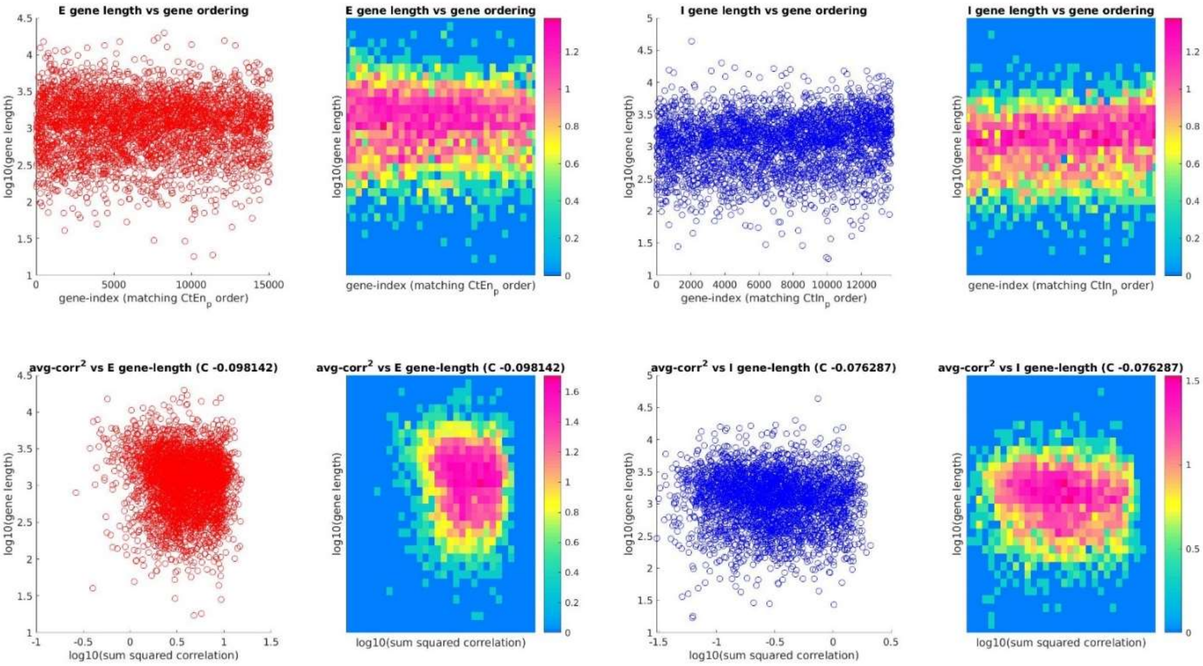


A.

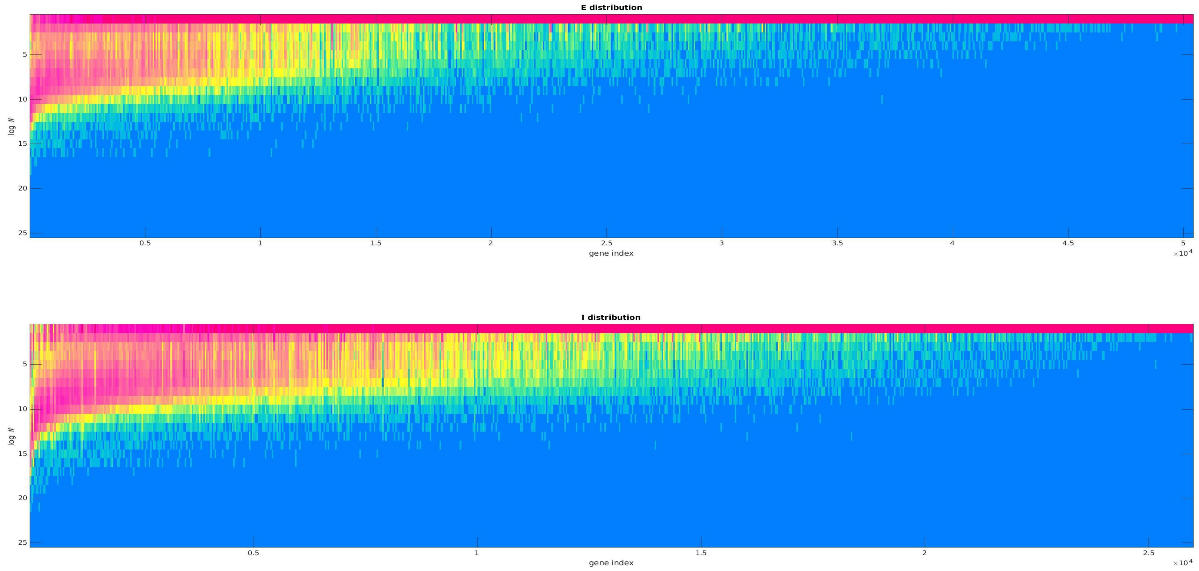
B.

C.

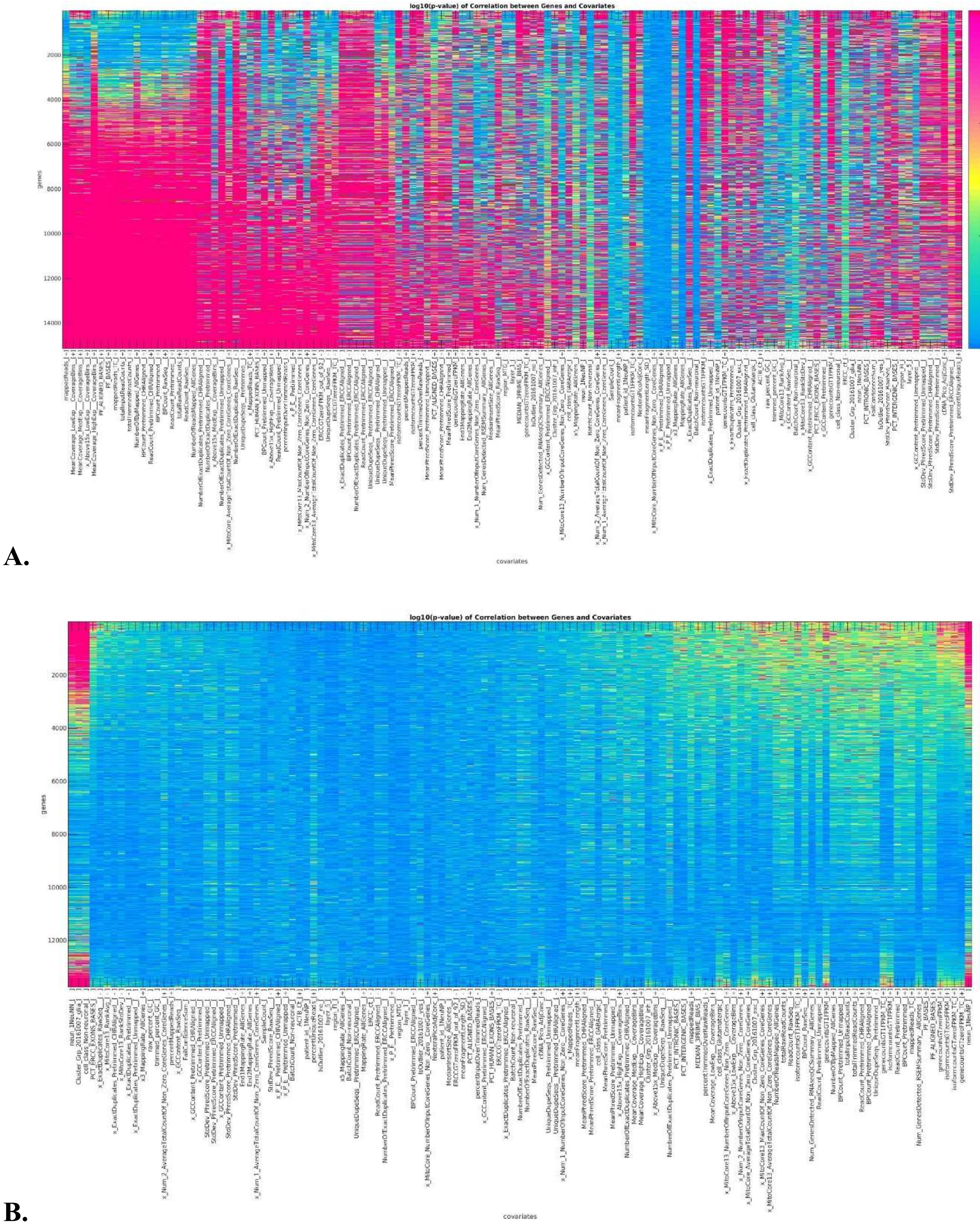
Supplemental Figure 1. Recomposition of example data in intron, exon, and intron-and-exon abundance space.



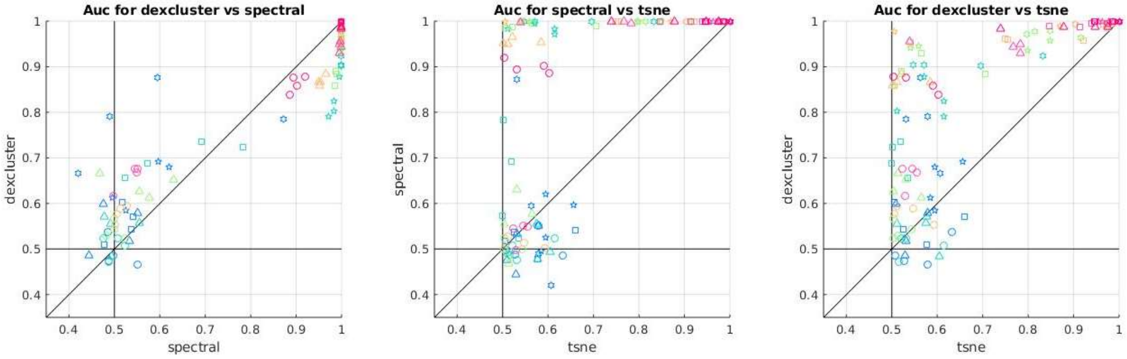
Supplemental Figure 2. Length-associated bias observations. A lack of correlation between gene-length and the average-squared-correlation between that gene and the covariates.



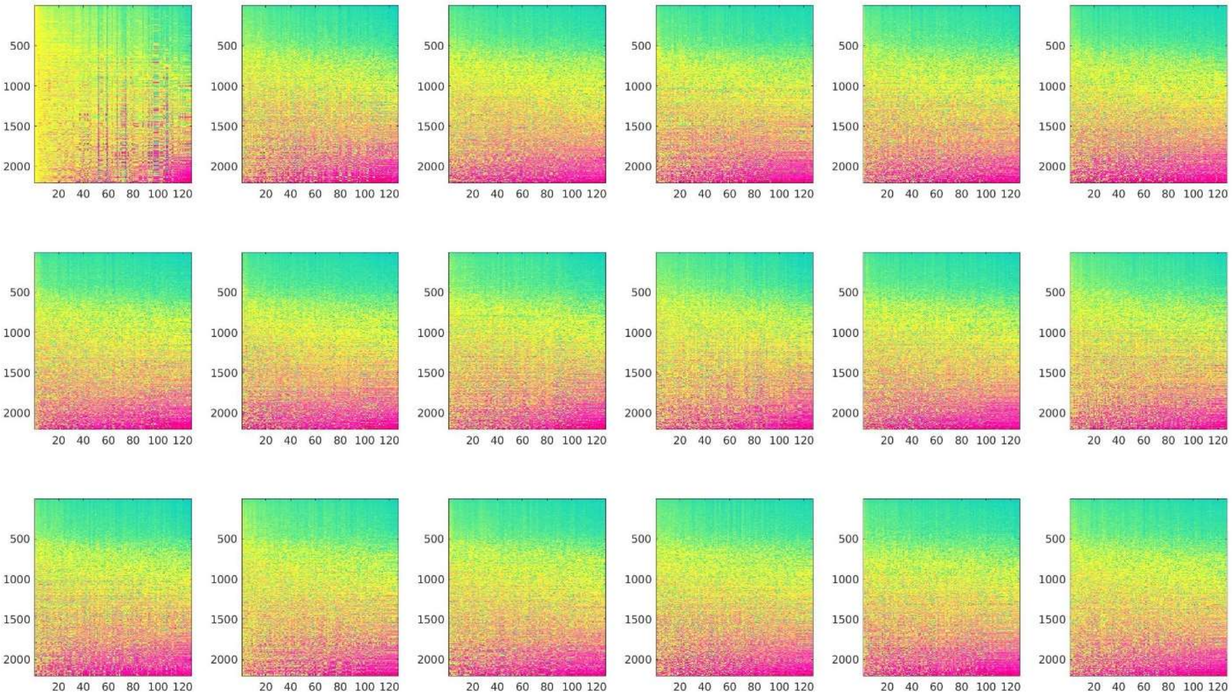
Supplemental Figure 3. Population-wide Bimodal Distribution-derived Methods do not Identify QC groups defining variation between cell types (other than ones like ‘layer type’, etc. of biological relevance) Vertical axis is log(abundance); each vertical slice is a histogram (taken across samples). Horizontal axis lists genes. Top subplot shows Exons, bottom subplot shows Introns.



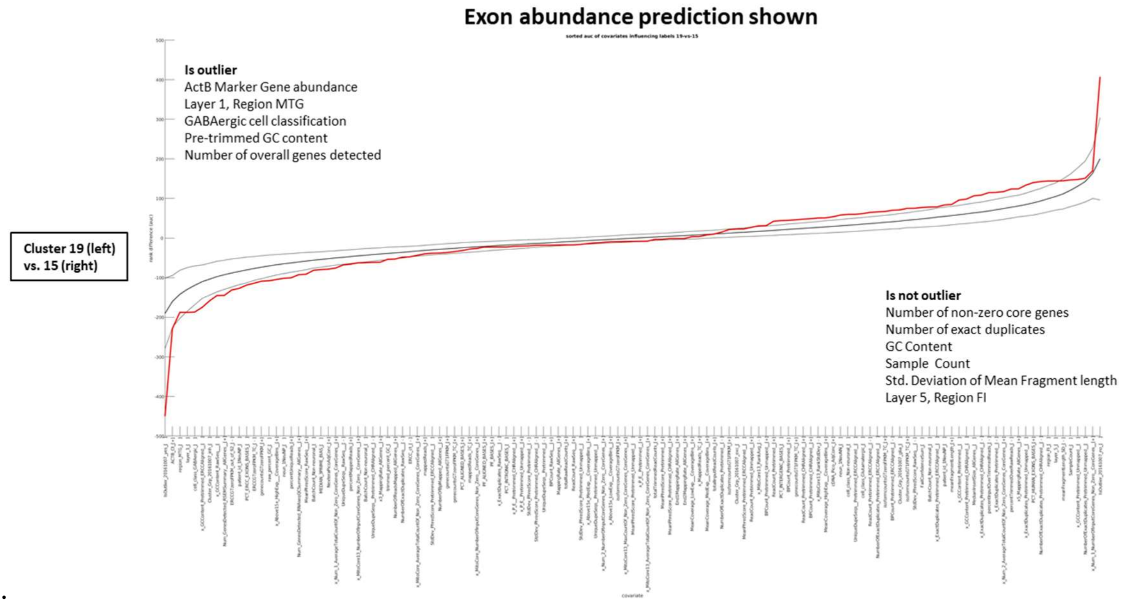
Supplemental Figure 4. Unsupervised Covariate vs. Sample Z-score matrix. **A)** Exons. B) Introns.



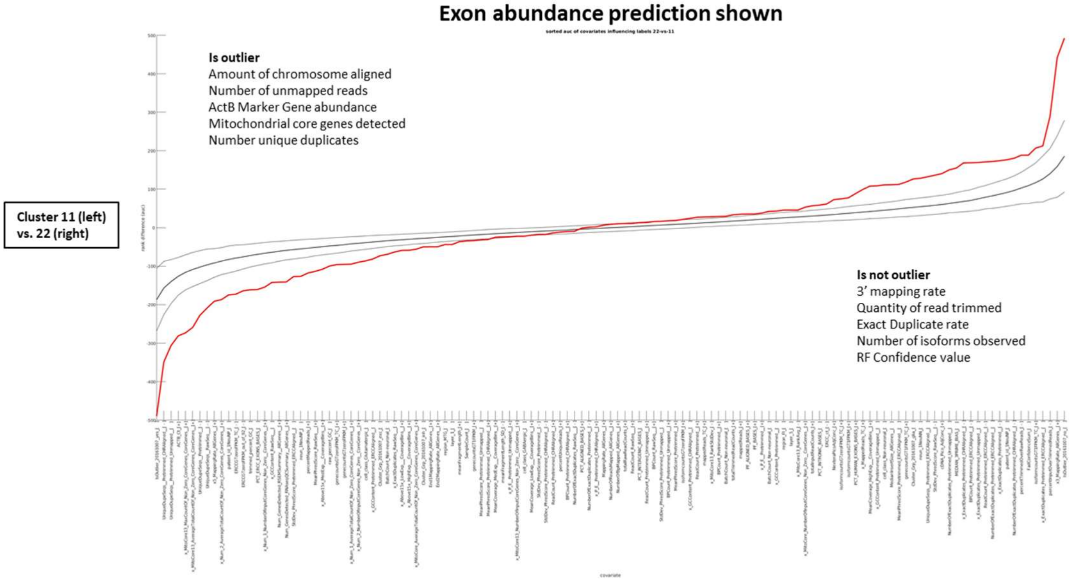
Supplemental Figure 5. The comparison of accuracy between spectral, nearest neighbor (tsne) and density (dexcluster) based interpretation of simulated data (e.g. the 'planted bicluster problem'). Variations in the color of nodes indicate X and the type of node indicates X.



Supplemental Figure 6. AIBS clusters' z-score matrix across all cluster pairs in the supervised context were checked to see if they were more or less strongly correlated than the z-score matrix we'd see if we scrambled the covariates at random. Heatmap for the Z-score matrix (sorted by simple spectral clustering).

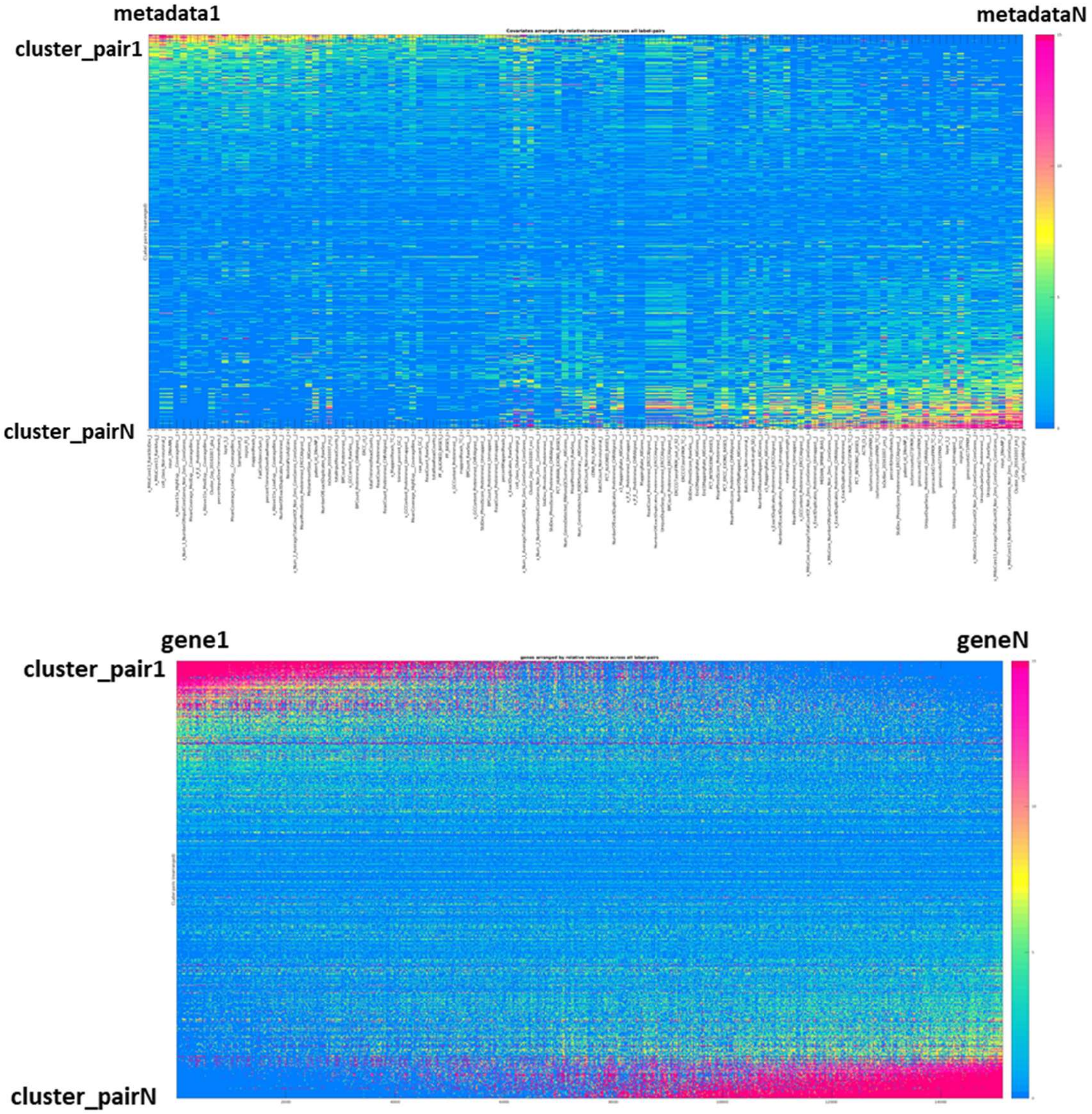


A.



B.

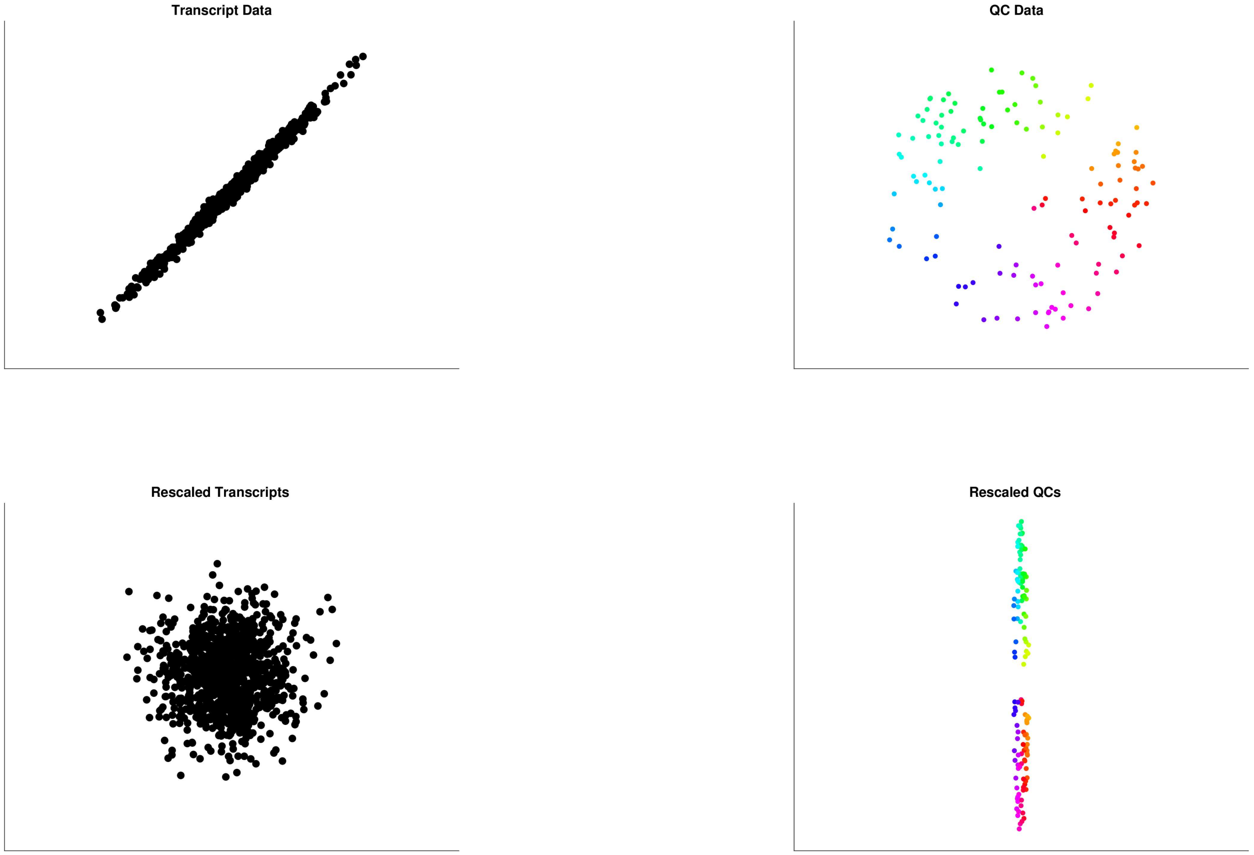
Supplemental Figure 7. Exon abundance. Outlier clusters adjacent to "validated" cell types have variance strongly predicted by sets of quality metrics associated with outlier terms associated with **A)** tissue-specific sampling and relative gene abundance observed, and **B)** common quality control terms associated with fail cases in prior literature. [2]



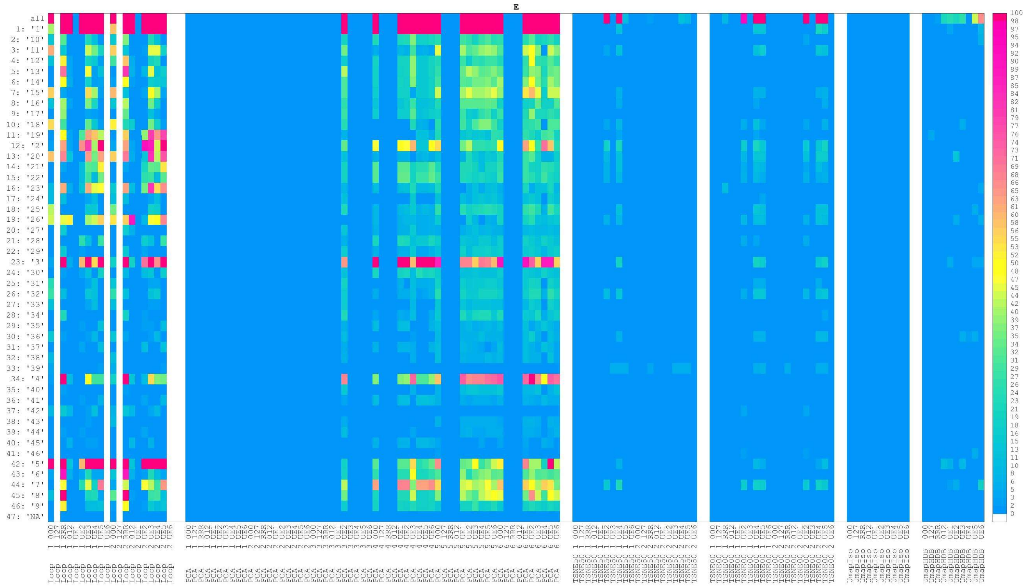
A.

B.

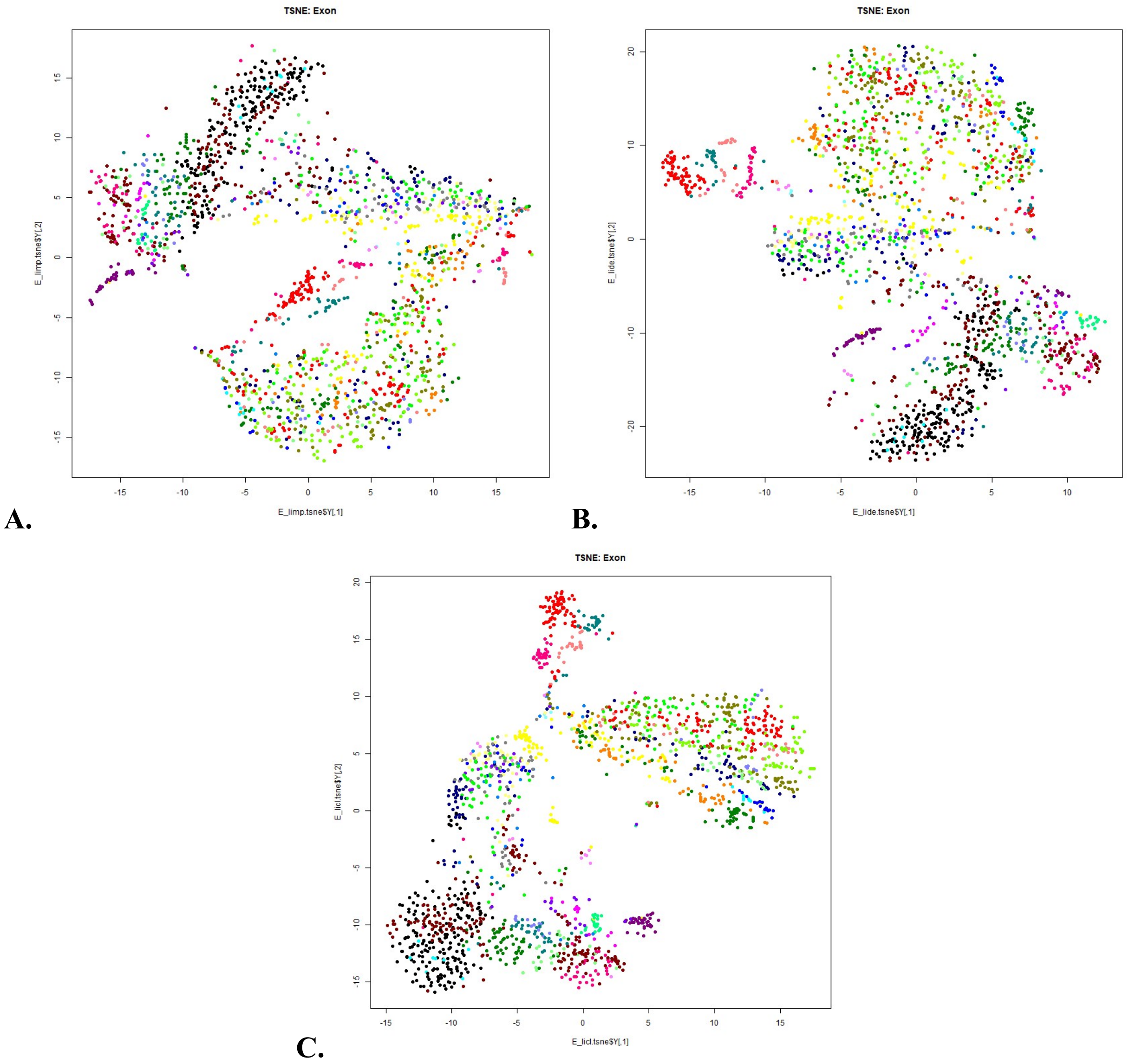
Supplemental Figure 8. Exons. A) The absolute-value of the z-score associated with the (raw) differential-expression of that particular covariate-rank over that cluster pair. **B)** The absolute-value of the z-score associated with the (raw) differential-expression of that particular gene-rank over that cluster pair.



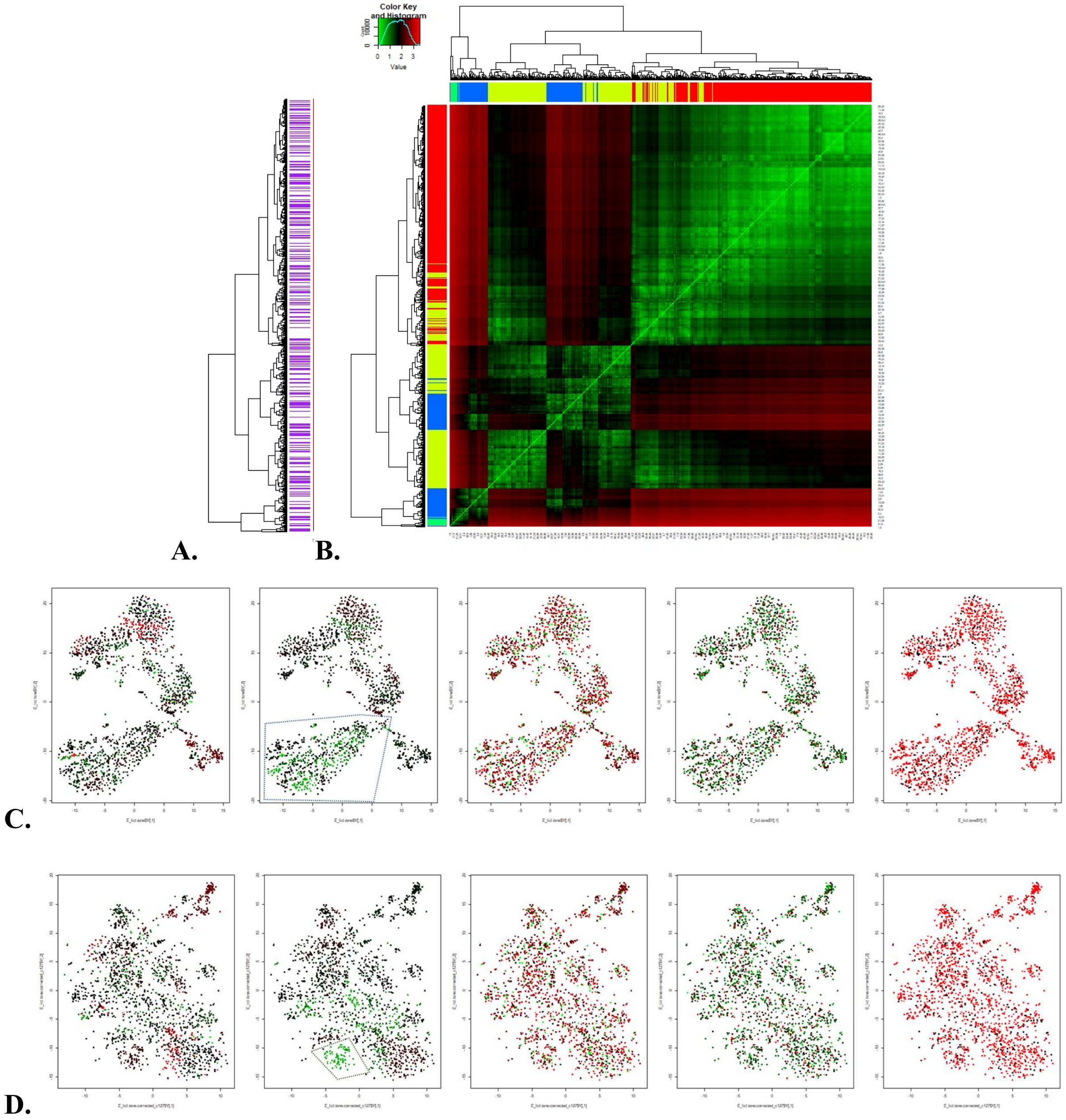
Supplemental Figure 9. In this cartoon we imagine a simple data-set with only 2 samples (shown here as the horizontal- and vertical-coordinates, respectively. On the top-left we show 1000 transcripts (black dots), each represented as a point in sample-space. On the top-right we show 127 QCs (colored to guide the eye), also represented as points in sample-space. On the bottom we show the data after decorrelating the transcripts.



Supplemental Figure 10. Compare to Table 2. The AIBS clusters to the unsupervised clustering of the Exon data (both pre- and post-covariate-correction) by X and gathered the negative of the log of the p-value. Higher means better (i.e., less likely to happen by chance). *(Note: Some columns are missing in this draft.)*

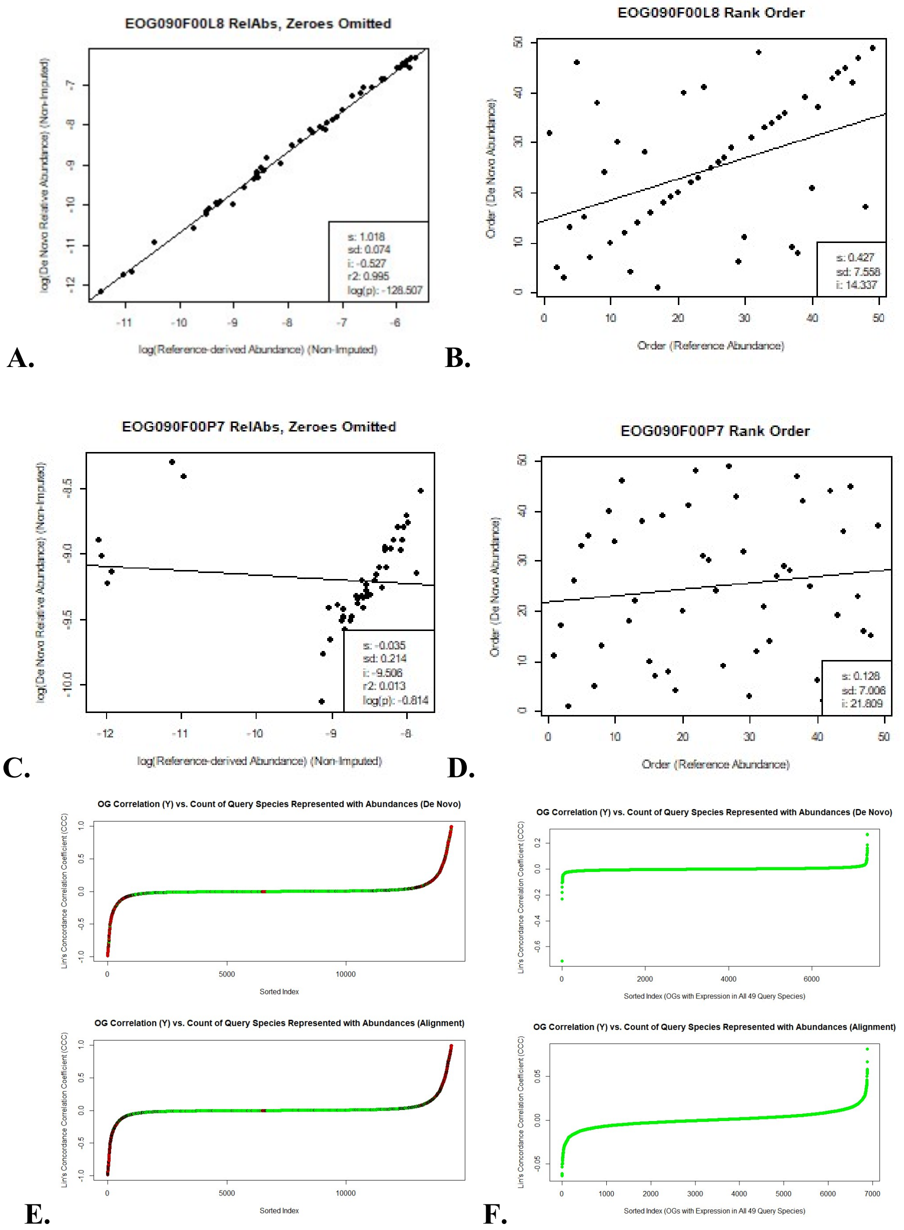


Supplemental Figure 11. Spectral ordination and clustering of samples in transcriptomic space. A-C) Exonic data, TSNE ordinated, and colored by original clusters defined by Bakken et. al. [cite]. **A)** Log normalized and imputed. **B)** The former method, then converted to relative abundance. **C)** The former method, followed by centering by log ratio.

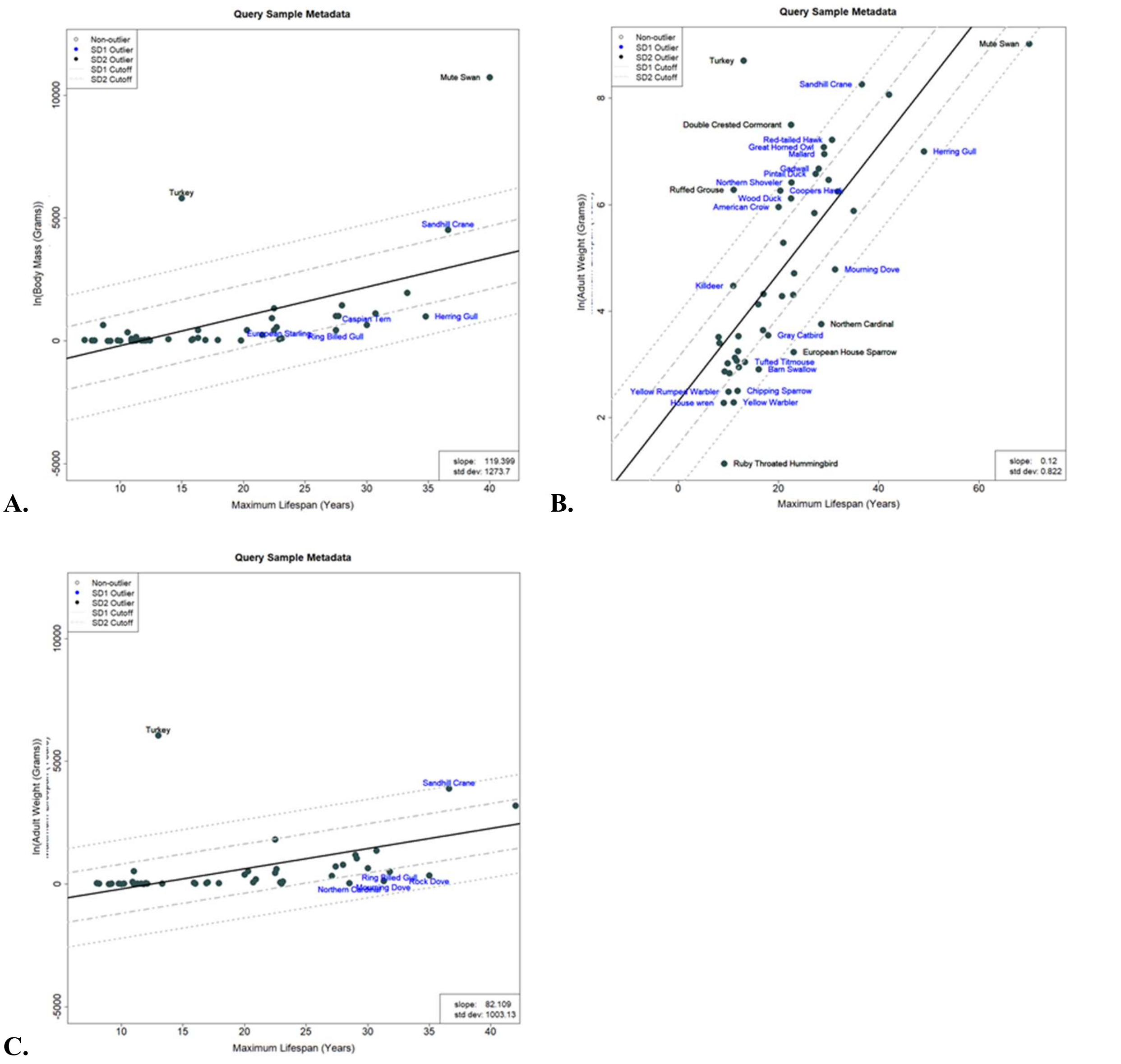


Supplemental Figure 12. Experimentation with experimental QC terms using linear modelling. A) Outlier (purple) vs. non-outlier labels for each row in “E”. **B)** The bias mode specific cluster pair distance matrix for cluster representing “Comparable BP After Trimming (Sequencer Bias)”, including the following terms: 1) Total Read Counts {Input before Trimming, Input after Trimming, Pre-trimmed and aligned to Human Reference, Pretrimmed and Unmapped, and 2) Total Base Pair count { Input before Trimming, Pre-trimmed and aligned to Huamn Reference, Pre-trimmed and unmapped.}.**C-D)** Compare to licl TSNE ordination. For each hierarchical cluster (5 shown, right), count all instances of a cluster in the grouping of cluster pair associations with the “bias mode”. **C)** Before correction **D)** After bias-mode-specific covariate correction (using only the covariates in the bias mode described above, under header “B”.

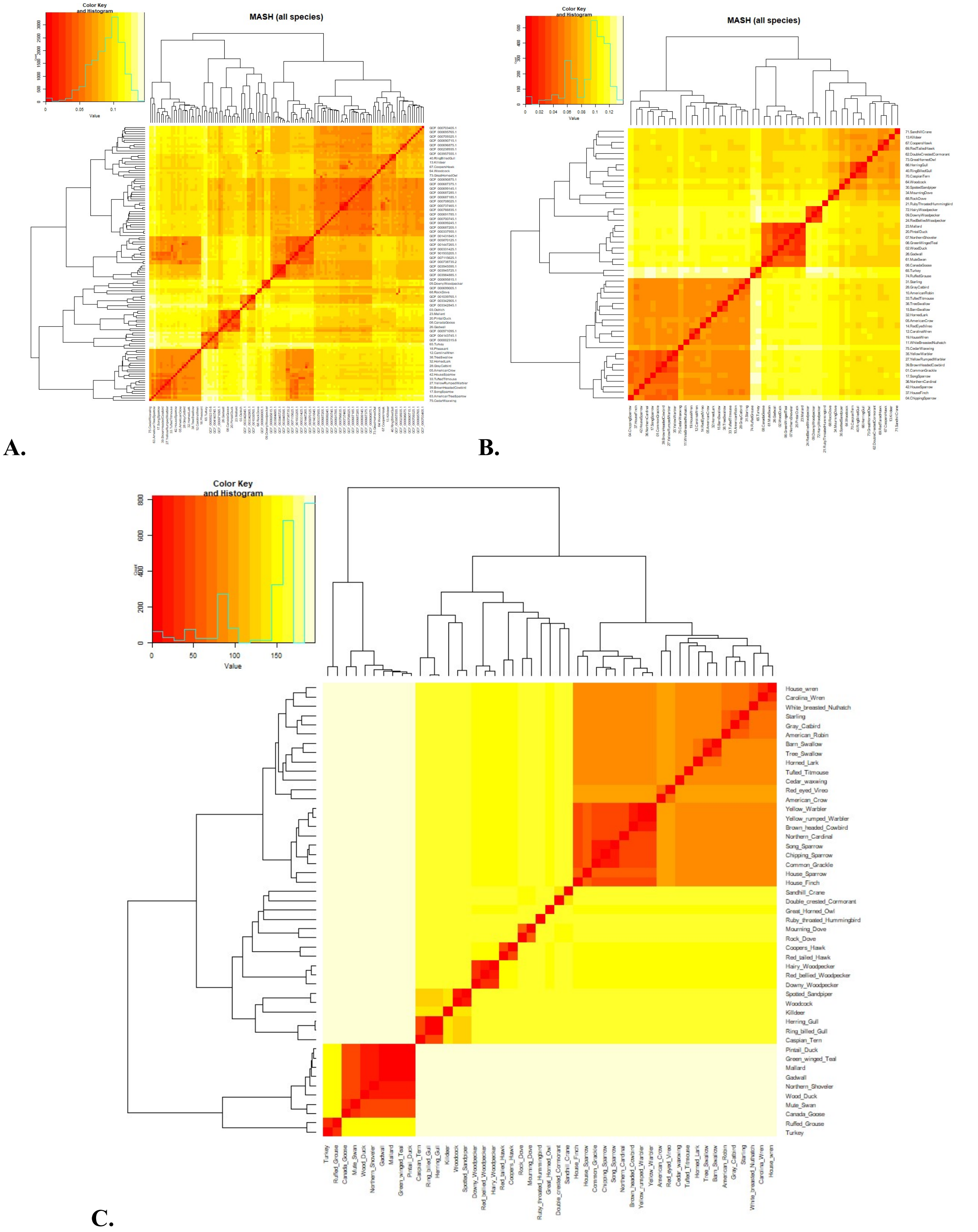
APPENDIX B (3.2)



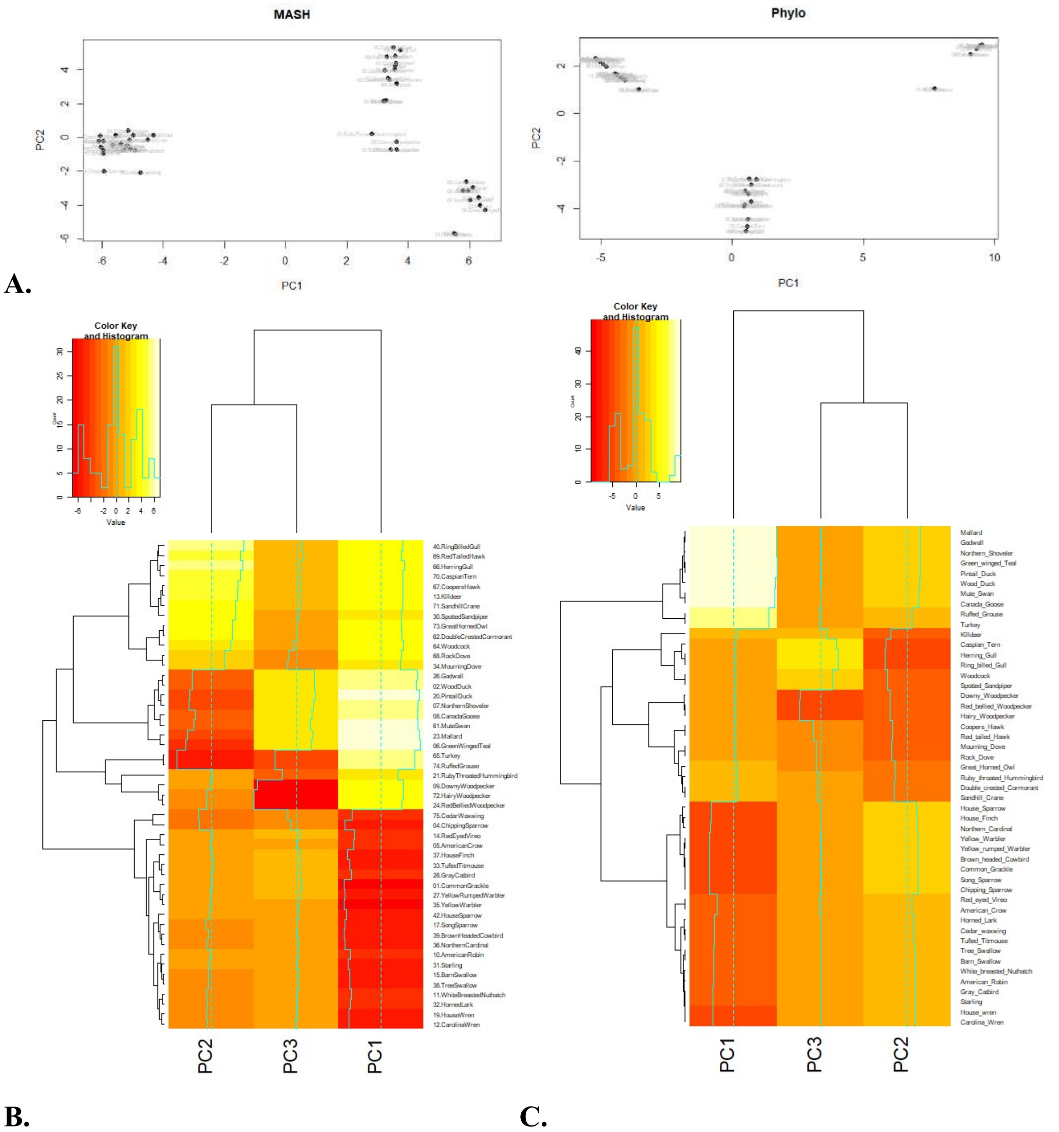
Supplemental Figure 1. OG Correlation Analysis. **A-D)** Correlation between relative natural log of abundance in alignment (x) vs. de novo (y) abundance context. **A-B)** High correlation example. **C-D)** Low correlation example. **E-F)** Lin's CCC value rho.q value for all OGs (expressed over 1-49 query species, where red is low (1), black is approx. the median (245), and green is is high (49) count of query species expressing the OG. **E)** All OGs. **F)** All OGs expressed over all 49 query species (those included in further study.)



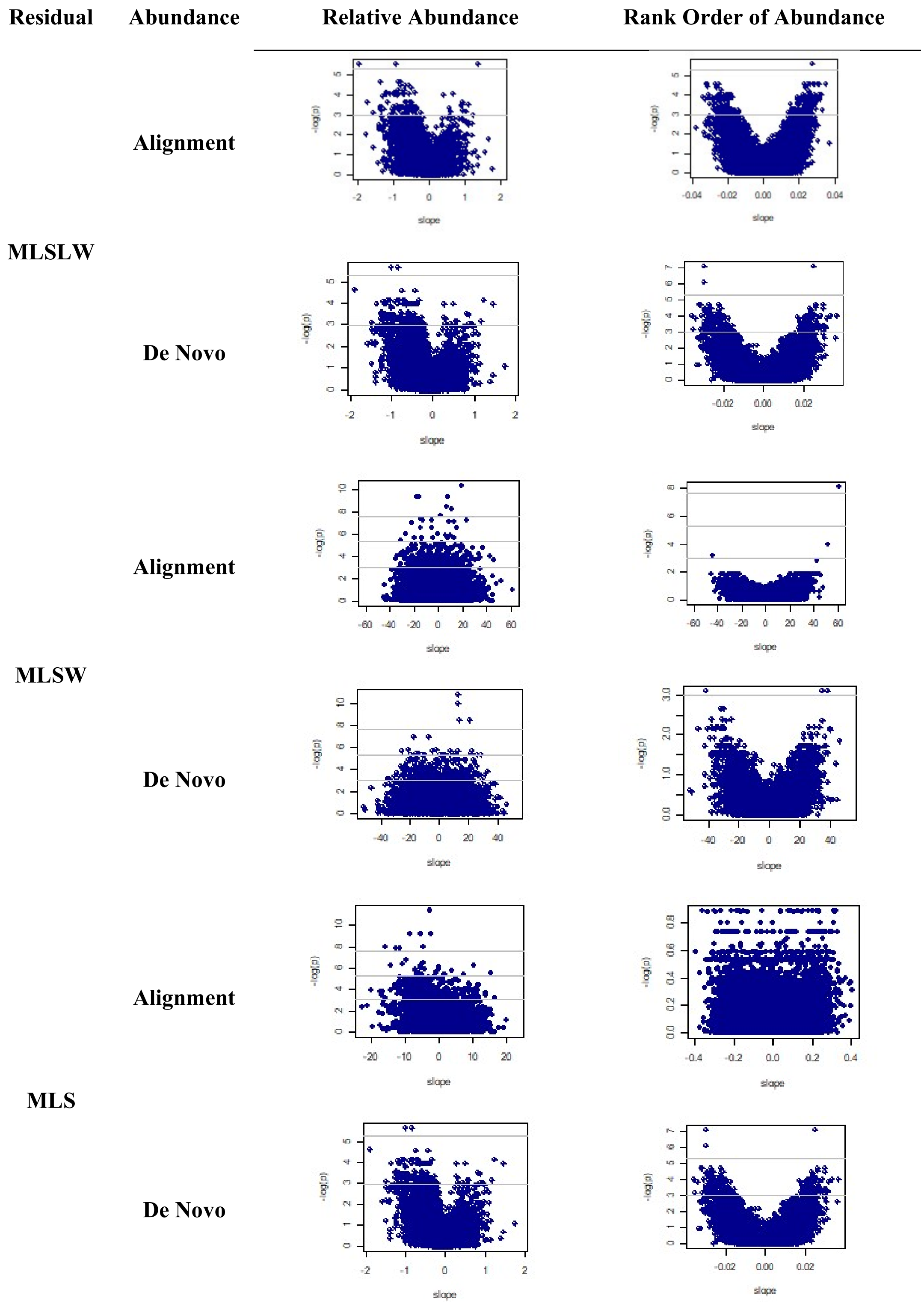
Supplemental Figure 2: Standard deviation comparisons calculated using the residual of the linear fit of the raw **(A,C)** and log-base-e-normalized body mass **(B)** of each query species, in grams, against the maximum lifespan of each species. Metadata provided in Supplemental Table 5a. **A)** Columns = "Dependent Variable: BMG (A)", "Independent Variable: MLS (A)") **B-C)** Columns = "Dependent Variable: AW (A)", "Independent Variable: MLS (A)")   These plots are comparable to the in alternative metadata context in Figure 2.



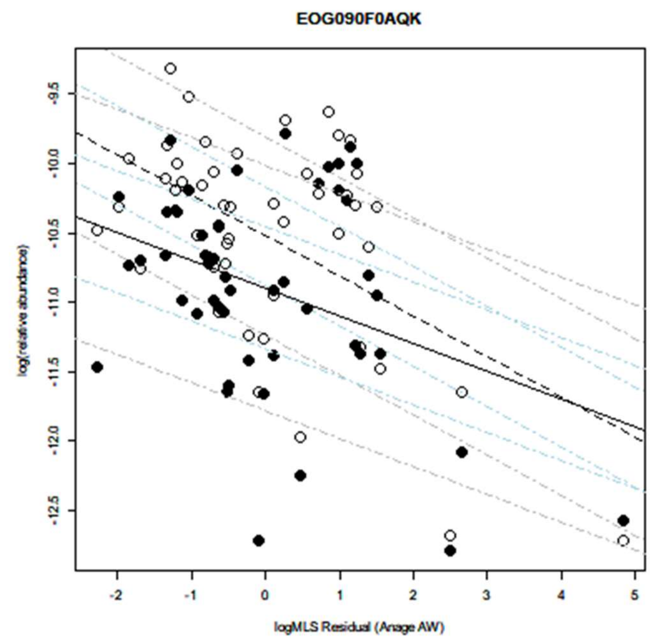
Supplemental Figure 3. Query-query sample distance matrices. **A)** All Avian GigaDB reference genome kmer similarity distance matrix. **B)** Subset to denote the distances amongst the 49 query species of interest. **C)** The comparable distance matrix from the phylogeny tree in Figure 1.



Supplemental Figure 4. A, Top) Principal component analysis showing the ordination of query-query sample distance In the context of the MASH distance matrix (Supplemental Figure 3B), over 2 principal components describing 37.55% of variance across the population.. **A, Bottom)** Repeated in the context of the distance matrix derived from the phylogenetic tree (Supplemental Figure 3C). **B-C)** Query sample ANOVA residuals for the linear fit of the multivariate model *MLS ~ log(BMG) + PC1 + PC2 + PC3*, where the PC's 1, 2, and 3 represent 26.83%, 7..99%, and 2.74% of population variance, each. (Supplemental Table 7) **B)** Using the MASH distance matrix. C) Using the Phylogenetic Tree distance matrix. See



Supplementary Figure 5. Number of unique ortholog groups (Y, negative natural log of raw p-value) and their associated phylogenetically contrast-adjusted slope (X, calculated using the caper model package), A) CAPER, MLSLW, log(Rel Abs), (BMG (A)), B) CAPER, MLSW log(Rel Abs) , (BMG (A)), C) CAPER, MLS log(Rel Abs) , (BMG (A)), D) CAPER, MLSLW, Rank(Abs), (BMG (A)) , E) CAPER, MLSW Rank (Abs) , (BMG (A)) , F) CAPER, MLS Rank.

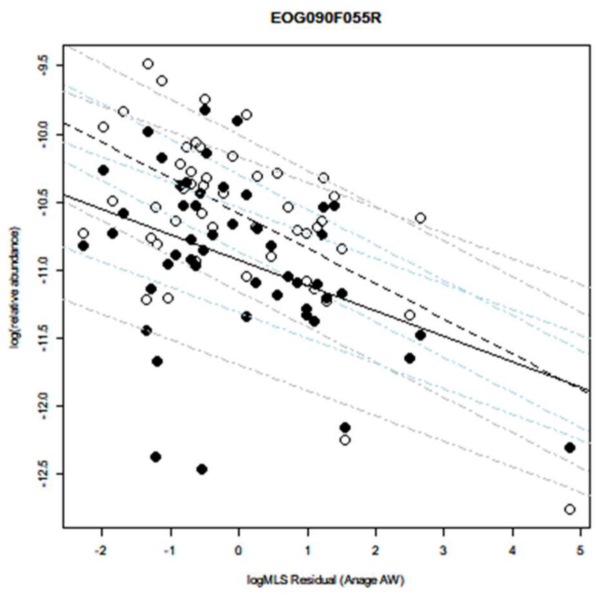


Query-specific Symbols	ARMC10 (armadillo repeat containing 10)	
HuRef Equivalents	ARMC10 (Decreases activity of tumor suppressor p53, role in cell growth and survival)	
Biological Processes	NA	NA
Molecular Functions	GO:0005488	binding
Cellular Components	NA	NA
Interpro Domains	Many	IPR006911, IPR011989, IPR016024
Unclassified GO Terms	GO:0003674, GO:0005515	

	BMG (O)	**BMG (A)**	**AW (A)**
LM Residuals (CCC rho.q)	0.9541201	0.951162	0.9509247
Phylogeny Contrasts (CCC rho.q)	0.4988913	0.6059601	0.6630369

	LM		**CAPER**		**MDMR (Mash)**		**MDMR (Phylo)**	
BMG (A)	**Align**	**De novo**	**Align**	**De novo**	**Align**	**De novo**	**Align**	**De novo**
Slope	-0.80999	-0.83848	-0.80999	-0.83848	NA	NA	NA	NA
P	0.012611	9.83E-05	1.78E-05	2.33E-07	0.008	0.02	0.016	0.04
FDR(P)	0.06711	0.004268	0.010292	0.001345	0.016722	0.026614	0.025684	0.048562
AW (A)	**Align**	**De novo**	**Align**	**De novo**	**Align**	**De novo**	**Align**	**De novo**
Slope	-0.80999	-0.83848	-0.80999	-0.83848	NA	NA	NA	NA
P	0.010376	8.51E-05	1.75E-05	7.37E-07	0.008	0.024	0.018	0.044
FDR(P)	0.058336	0.00387	0.011225	0.003574	0.015906	0.031317	0.02842	0.052906

Supplementary Figure 6a. See Figure 4.

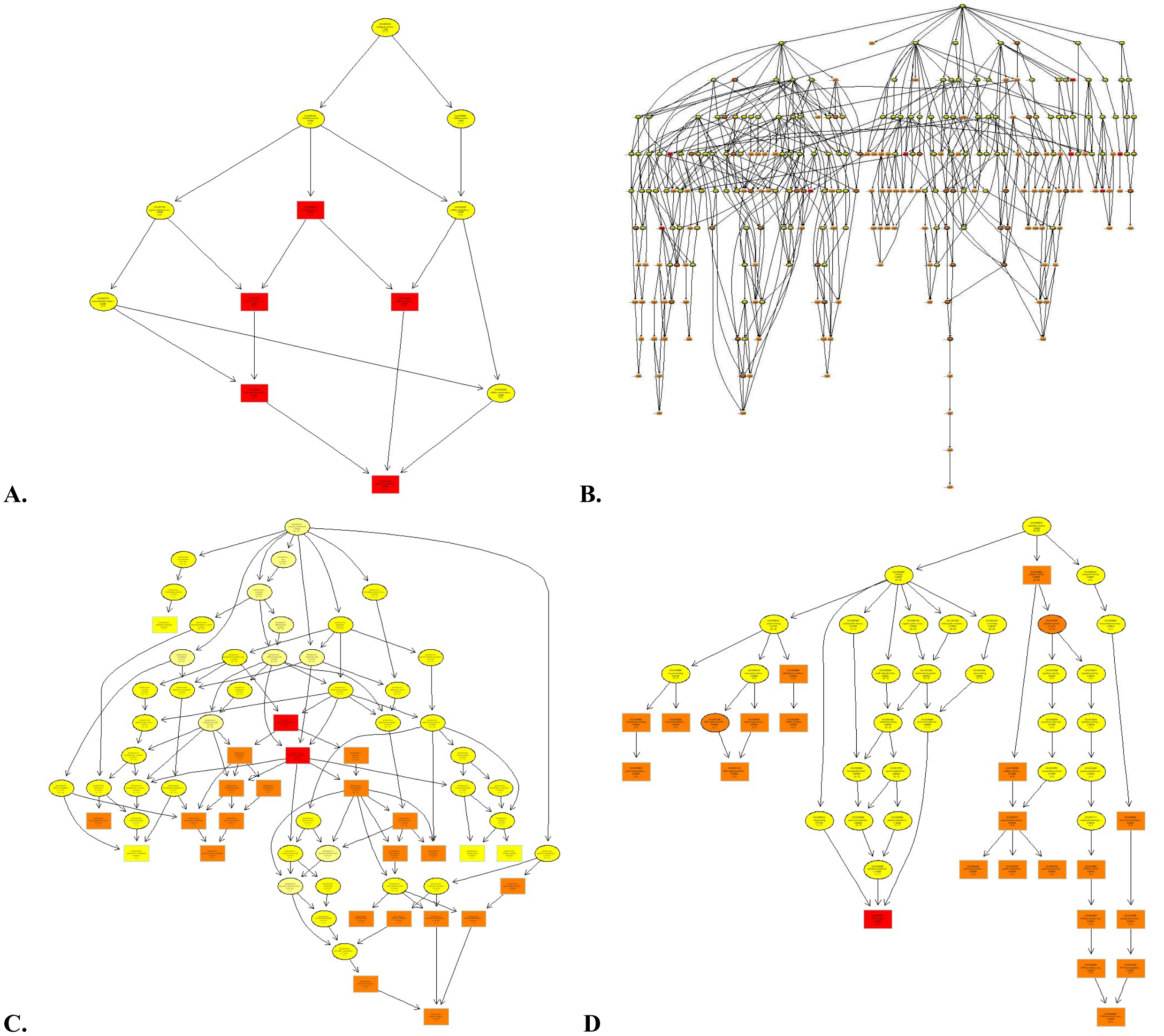


Query-specific Symbols	KLHDC1 (Kelch domain containing 1)	
HuRef Equivalents	KLHDC1/MST025 (Inhibits transcription factor LZIP) – Causes Retinitis Pigmentosa	
Biological Processes	NA	NA
Molecular Functions	NA	NA
Cellular Components	GO:0005737	cytoplasm
Interpro Domains	IPR015915	Kelch-type beta propeller
Unclassified GO Terms	GO:0003674, GO:0005515	

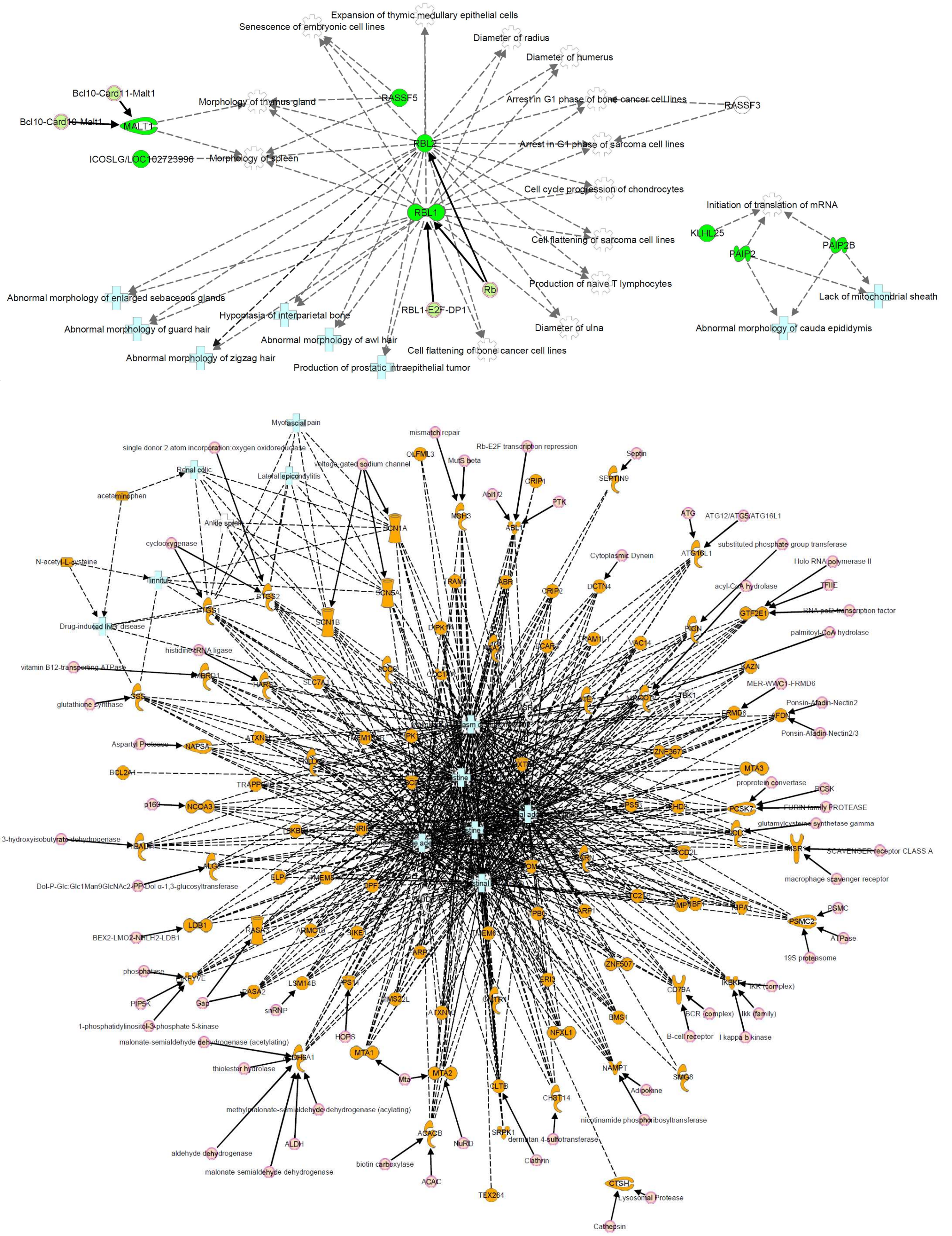
	BMG (O)	**BMG (A)**	**AW (A)**
LM Residuals (CCC rho.q)	0.9472192	0.9010957	0.8988462
Phylogeny Contrasts (CCC rho.q)	0.4444235	**0.2668113**	**0.2363392**

	LM		**CAPER**		**MDMR (Mash)**		**MDMR (Phylo)**	
BMG (A)	**Align**	**De novo**	**Align**	**De novo**	**Align**	**De novo**	**Align**	**De novo**
Slope	-0.23285	-1.00897	-0.23285	-1.00897	NA	NA	NA	NA
P	0.003972	1.16E-05	0.028606	1.74E-06	0.018	0.026	0.022	0.026
FDR(P)	0.032253	0.001334	0.223021	0.003233	0.024512	0.0328	0.031028	0.034896
AW (A)	**Align**	**De novo**	**Align**	**De novo**	**Align**	**De novo**	**Align**	**De novo**
Slope	-0.23285	-1.00897	-0.23285	-1.00897	NA	NA	NA	NA
P	0.003135	9.42E-06	0.037584	1.86E-06	0.016	0.028	0.026	0.028
FDR(P)	0.027657	0.001006	0.243555	0.003574	0.023235	0.035369	0.035532	0.037511

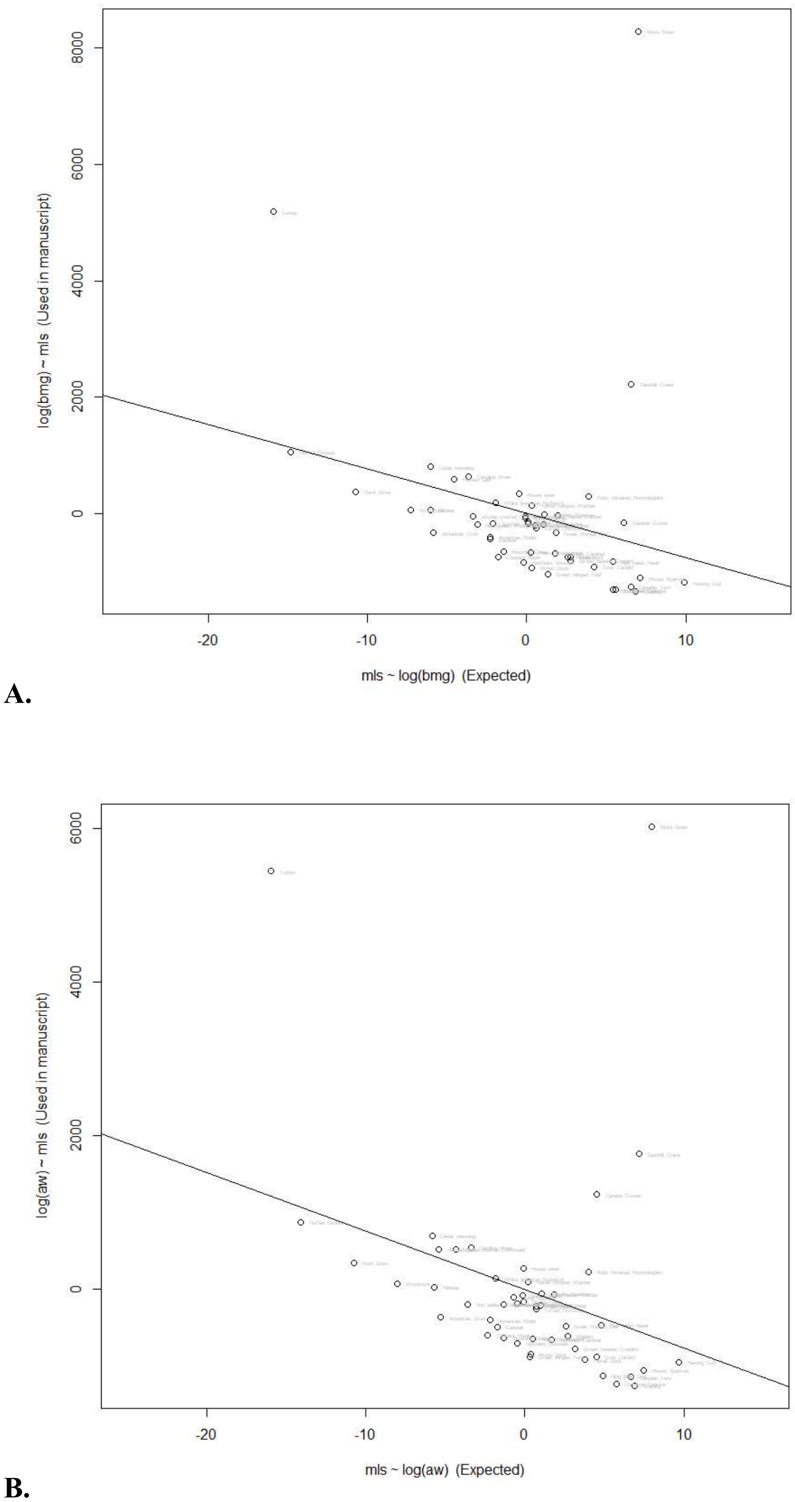
Supplementary Figure 6b. See Figure 4.



Supplemental Figure 7. TopGO Gene Ontology Results. Metadata = BMG (A). GO terms associated with significant OGs from both Alignment and De Novo abundance contexts. **A,C,D.)** Positive association with residual on lifespan. **B)** Negative association with residual on lifespan. **A-B)** Biological processes, cutoff = KS 0.05. **C-D.)** Cellular components, cutoff = KS 0.05. **E-F.)** Molecular functions, cutoff = KS 0.05.


Expansion of thymic medullary epithelial cells
Senescence of embryonic cell lines
Diameter of radius
Diameter of humerus
Bcl10-Card11-Malt1
RASSF5
Morphology of thymus gland
Arrest in G1 phase of bone cancer cell lines
RASSF3
Bcl10-Card10-Malt1
MALT1
RBL2
Arrest in G1 phase of sarcoma cell lines
ICOSLG/LOC102723996
Morphology of spleen
Cell cycle progression of chondrocytes
Initiation of translation of mRNA
RBL1
KLHL25
Cell flattening of sarcoma cell lines
PAIP2B
PAIP2
Rb
Production of naive T lymphocytes
Abnormal morphology of enlarged sebaceous glands
Lack of mitochondrial sheath
Hypoplasia of interparietal bone
RBL1-E2F-DP1
Abnormal morphology of guard hair
Diameter of ulna
Abnormal morphology of cauda epididymis
Abnormal morphology of awl hair
Cell flattening of bone cancer cell lines
Abnormal morphology of zigzag hair
Production of prostatic intraepithelial tumor
A.
Myofascial pain
mismatch repair
single donor 2 atom incorporation:oxygen oxidoreductase
Rb-E2F transcription repression
Septin
Renal colic
voltage-gated sodium channel
OLFML3
MutS beta
Lateral epicondylitis
SEPTIN9
acetaminophen
Abl1/2
CRIP1
PTK
MSH3
ABL1
Ankle sprain
SCN1A
ATG
ATG12/ATG5/ATG16L1
cyclooxygenase
substituted phosphate group transferase
ATG16L1
N-acetyl-L-cysteine
Tinnitus
Cytoplasmic Dynein
Holo RNA polymerase II
PTGS2
CRIP2
DCTN4
acyl-CoA hydrolase
TFIIE
PTGS1
SCN5A
GTF2E1
RNA pol2 transcription factor
Drug-induced liver disease
SCN1B
PIGN
palmitoyl-CoA hydrolase
histidine-tRNA ligase
KAZN
vitamin B12-transporting ATPase
MER-WWC1-FRMD6
TBK1
Ponsin-Afadin-Nectin2
glutathione synthase
GSS
FRMD6
AFDN
Ponsin-Afadin-Nectin2/3
Aspartyl Protease
NAPSA
ATXN1
ZNF507
BCL2A1
MTA3
proprotein convertase
TRAPPC
PCSK
p160
NCOA3
PSS
PCSK7
FURIN family PROTEASE
glutamylcysteine synthetase gamma
3-hydroxyisobutyrate dehydrogenase
MSR1
SCAVENGER receptor CLASS A
ALG8
macrophage scavenger receptor
Dol-P-Glc:Glc1Man9GlcNAc2-PP-Dol α-1,3-glucosyltransferase
PSMC
PSMC2
ATPase
LDB1
BEX2-LMO2-NHLH2-LDB1
RASA2
19S proteasome
phosphatase
PIKFYVE
PIP5K
Gas
RASA2
LSM14B
ZNF507
CD79A
IKBKB
IKK (complex)
snRNP
BCR (complex)
Ikk (family)
1-phosphatidylinositol-3-phosphate 5-kinase
HOPS
NFXL1
B-cell receptor
I kappa b kinase
malonate-semialdehyde dehydrogenase (acetylating)
ALDH6A1
MTA1
thiolester hydrolase
Mta
MTA2
NAMPT
SMG8
Adipokine
CLTB
CHST14
methylmalonate-semialdehyde dehydrogenase (acylating)
nicotinamide phosphoribosyltransferase
ACACB
NuRD
SRPK1
dermatan 4-sulfotransferase
aldehyde dehydrogenase
ALDH
biotin carboxylase
Clathrin
malonate-semialdehyde dehydrogenase
ACAC
CTSH
TEX264
Lysosomal Protease
Cathepsin
B.

Supplemental Figure 8. Ingenuity Pathway Figure (hand-curated). Metadata = A (AW) Significant OGs shared in Alignment and De Novo abundance contexts with Ingenuity Pathway equivalents (Human, Mouse, Rat) Data used for Figure examples: Model = CAPER, Dependent = MLSLW, Independent = LG, Metadata = BMG (O+A), FDR p-value cutoff = 0.5. Direction of association = **A) IP pathway selection cutoff = 5e-6 (none significant at 5e-8)** Positive (Dark Green=Positive Association with Residual on Lifespan, Light Green = IP Linked Associated Target). **B) IP pathway selection cutoff = 5e-8:** Negative (Dark Orange=Positive Association with Residual on Lifespan, Light Orange = IP Linked Associated Target). Some of the most significant Disease and Function terms, those with greater than 100 connections to genes were omitted for visualization (see Supplemental Table). Associations between Human reference genes co-defined for significant OGs (dark green, dark orange) and diseases (light blue) and functions (white) are indicated with a dashed line. Connections between these genes and their linked targets from the IP database are provided with a solid line (Table 1, Table 2) Compare to Table 5.



Supplemental Figure 9. The relationship between LM residuals for each of the 49 query species involved in our study in the context of regressing X = log(Body Mass in Grams) on maximum lifespan and Y = maximum lifespan on log(Body Mass in Grams). A) Metadata = BMG (A). B) A) Metadata = AW (A).

Supplemental Table 1. 49 Query Species, their associated assignment to their best-defined reference, and associated metadata. [citations] MLSLW Residuals and related metadata provided in the BMG (A) context. See supplemental table 3 for additional metadata in alternative contexts.* = Curate (BMG (O)), ** = Curated+AnAge (BMG (A))

Common Name	Scientific Name	Family	Body Mass Grams*	MLS Years*	Source	MASH Best Reference (2020)	MASH score for Best Reference (query > ref)	STAMPY Best Reference (2012)	MLSW Residual**	MLSW Outlier**	MLSLW Residual**	MLSLW Outlier**
American Crow	Corvus brachyrhynchos	Corvidae	448	16.3	US BBL	American Crow	0.033	American Crow	-192.855	+	1.401475234	+
Ruby Throated Hummingbird	Archilochus colubris	Trochilidae	3.1	9.2	US BBL	Annas Hummingbird	0.032	Annas Hummingbird	415.677	- -	-2.232521418	- -
House Finch	Carpodacus mexicanus	Fringillidae	21.4	11.7	US BBL	Common Canary	0.039	Zebra Finch	191.287		-0.680056082	
Red-tailed Hawk	Buteo jamaicensis	Accipitridae	1126	30.7	US BBL	Bald Eagle	0.036	Peregrine Falcon	-548.860	+	1.03945654	
Coopers Hawk	Accipiter cooperii	Accipitridae	439	20.3	US BBL	Bald Eagle	0.040	Peregrine Falcon	-230.846	+	1.345191164	
Great Horned Owl	Bubo virginianus	Strigidae	1309	28	US BBL	Burrowing Owl	0.055	Downy Woodpecker	-60.579	+	1.49629332	
Common Grackle	Quiscalus quiscula	Icteridae	114	23.1	US BBL	Small Tree Finch	0.042	Zebra Finch	-826.427		-0.339018806	-
Northern Cardinal	Cardinalis cardinalis	Cardinalidae	44.7	15.8	US BBL	Small Tree Finch	0.045	Zebra Finch	-1421.261	- -	-2.009469698	
Yellow Warbler	Dendroica petechia	Parulidae	9.5	11	US BBL	Small Tree Finch	0.046	Zebra Finch	238.369	-	-1.372318683	-
Brown Headed Cowbird	Molothrus ater	Icteridae	43.9	16.9	US BBL	Small Tree Finch	0.046	Zebra Finch	-298.784		-0.581914667	-
Yellow Rumped Warbler	Zenaida macroura	Parulidae	12.6	10	US BBL	Small Tree Finch	0.047	Zebra Finch	337.005	-	-1.061295945	
European House Sparrow	Passer domesticus	Passeridae	27.7	19.8	EURING	Small Tree Finch	0.054	Zebra Finch	-903.063	- -	-1.741788091	- -
Rock Dove	Columba livia	Columbidae	350	10.6	EURING	Rock Dove	0.030	Rock Dove	-1740.395		-0.644897373	+ +
Mourning Dove	Zenaida macroura	Columbidae	119	16.3	US BBL	Rock Dove	0.054	Rock Dove	-1609.842	-	-1.246766415	
Canada Goose	Branta canadensis	Anatidae	1952	33.3	US BBL	Chinese Goose	0.034	Mallard	-814.847		0.234006727	
Mute Swan	Cygnus olor	Anatidae	10735	40	Science Daily	Chinese Goose	0.039	Mallard	5262.346		-1.420428383	
Double Crested Cormorant	Phalacrocorax auritus	Phalacrocoracidae	1674	22.5	US BBL	Great Cormorant	0.038	Crested Ibis	447.555	+ +	2.189697468	+
Ruffed Grouse	Bonasa umbellus	Phasianidae	577	8.6	Decker Major et al. 1981	Ring-Nicked Pheasant	0.054	Chicken	872.869	+ +	2.844088244	+ +
European Starling	Sturnus vulgaris	Sturnidae	82.3	22.9	EURING	Starling	0.029	Zebra Finch	-846.099		-0.733735641	-
Gray Catbird	Dumetella carolinensis	Mimidae	36.9	17.9	US BBL	Starling	0.049	Zebra Finch	-401.020	-	-0.843174694	-
American Robin	Turdus migratorius	Turdidae	77.3	13.9	US BBL	Starling	0.057	Zebra Finch	-273.647		0.004278462	
Turkey	Meleagris gallopavo	Phasianidae	5811	15	Cardozo, 1995	Turkey	0.035	Turkey	5846.597	+ +	4.803962466	+ +
Downy Woodpecker	Picoides pubescens	Picidae	27	11.9	US BBL	Downy Woodpecker	0.029	Downy Woodpecker	163.596		-0.654268973	
Hairy Woodpecker	Picoides villosus	Picidae	66.3	15.9	US BBL	Downy Woodpecker	0.034	Downy Woodpecker	-178.348		-0.017261344	
Red-bellied Woodpecker	Melanerpes carolinus	Picidae	61.7	12.3	US BBL	Downy Woodpecker	0.045	Downy Woodpecker	-646.800		-0.665011319	
Pintail Duck	Anas acuta	Anatidae	1011	27.5	EURING	Mallard	0.032	Mallard	-344.961	+	1.327617706	
Mallard	Anas platyrhynchos	Anatidae	1082	27.7	US BBL	Mallard	0.034	Mallard	-500.242	+	1.13253564	
Gadwall	Anas strepera	Anatidae	920	22.3	EURING	Mallard	0.034	Mallard	-493.943	+	1.161315656	+
Green Winged Teal	Anas crecca	Anatidae	341	21.5	EURING	Mallard	0.036	Mallard	-1076.971		-0.033626344	
Northern Shoveler	Anas clypeata	Anatidae	613	22.7	BTO	Mallard	0.039	Mallard	-338.109	+	1.301931183	
Wood Duck	Aix sponsa	Anatidae	658	22.5	US BBL	Mallard	0.039	Mallard	-434.445	+	1.101556479	
Sandhill Crane	Grus canadensis	Gruidae	4513	36.6	US BBL	South African Crowned Crane	0.044	Killdeer	2267.988	+	1.71993887	
Red Eyed Vireo	Vireo olivaceus	Vireonidae	16.7	12	US BBL	Hooded Crow	0.056	American Crow	322.877		-0.712227761	
Killdeer	Charadrius vociferus	Charadriidae	97	10.9	US BBL	Killdeer	0.030	Killdeer	335.532	+	0.963097247	+
Ring Billed Gull	Larus delawarensis	Laridae	518.5	27.5	US BBL	Killdeer	0.061	Killdeer	-1342.160		-0.034435108	
Herring Gull	Larus argentatus	Laridae	1135	34.8	EURING	Killdeer	0.064	Killdeer	-2443.298	-	-1.274620274	-
Caspian Tern	Hydroprogne caspia	Laridae	656	30	EURING	Killdeer	0.065	Killdeer	-951.215		0.583167771	
Spotted Sandpiper	Actitis macularius	Scolopacidae	42.5	12	US BBL	Ruff	0.050	Killdeer	174.733		0.005926092	
American Woodcock	Scolopax minor	Icteridae	197.5	11.3	US BBL	Ruff	0.055	Zebra Finch	-571.128		0.243044399	+
Tufted Titmouse	Baeolophus bicolor	Paridae	21.6	12.4	US BBL	Tibetan Ground-Tit	0.036	Zebra Finch	28.206	-	-0.826842421	
White Breasted Nuthatch	Sitta carolinensis	Sittidae	21.1	9.8	US BBL	Tibetan Ground-Tit	0.062	Zebra Finch	365.932		-0.430376439	
Barn Swallow	Hirundo rustica	Hirundinidae	16	11.1	EURING	Tibetan Ground-Tit	0.063	Zebra Finch	-236.311	-	-1.333076233	
Tree Swallow	Tachycineta bicolor	Hirundinidae	20.1	12.1	US BBL	Tibetan Ground-Tit	0.063	Zebra Finch	142.669		-0.754854919	
Carolina Wren	Thryothorus ludovicianus	Troglodytidae	18.7	7.7	US BBL	Tibetan Ground-Tit	0.063	Zebra Finch	421.513		-0.479145455	
House wren	Troglodytes aedon	Troglodytidae	10.9	9	US BBL	Tibetan Ground-Tit	0.064	Zebra Finch	433.041	-	-0.994912689	
Horned Lark	Eremophila alpestris	Alaudidae	31.3	7.9	US BBL	Tibetan Ground-Tit	0.065	Zebra Finch	544.777		-0.005611438	
Cedar waxwing	Bombycilla cedrorum	Bombycillidae	32	7.1	US BBL	Tibetan Ground-Tit	0.071	Zebra Finch	531.449		0.178034426	
Song Sparrow	Melospiza melodia	Emberizidae	20.8	11.3	US BBL	White-Throated Sparrow	0.043	Zebra Finch	218.978		-0.709912397	
Chipping Sparrow	Spizella passerina	Emberizidae	12.3	10.9	US BBL	White-Throated Sparrow	0.051	Zebra Finch	163.460	-	-1.243046083	

Supplemental Table 2a. Best reference – metadata table.

Reference Metadata	Num Query Assigned		Reference-specific IDs				
Common Name	Assigned (MASH)	Prev Assigned (STAMPY)	Protein IDs	CDS Reference IDs	Feature Table Defined Symbols	Feature Table Defined Symbols (No LOC)	Feature Table Defined Symbols (Huref Overlaps, of 54,597)
Atlantic Canary	1	0	29,660	29,717	17,261	12,394	12,202
Bald Eagle	2	0	27,572	27,583	15,655	12,918	12,425
Burrowing Owl	1	0	32,506	32,534	15,675	11,713	11,446
Small Tree Finch	6	0	26,242	26,250	16,664	13,141	12,956
Chinese Goose	2	0	29,421	29,441	20,704	11,828	11,423
Common Cormorant	1	0	15,826	15,830	14,687	9,085	8,712
Common Pheasant	1	0	16,428	16,448	19,182	13,082	12,885
Common Starling	3	0	15,339	15,348	16,715	12,672	12,283
Grey Crowned Crane	1	0	16,163	16,179	15,279	9,834	9,438
Hooded Crow	1	0	25,265	25,280	17,118	11,552	11,131
Ruff	2	0	29,601	29,603	18,471	12,397	12,018
Tibetan Ground-Jay	8	0	31,811	31,821	16,415	13,051	12,620
White-Throated Sparrow	2	0	30,889	30,937	17,249	11,491	11,235
Anna's Hummingbird	1	1	26,611	26,619	16,080	12,454	12,268
Common Turkey	1	1	27,746	27,752	24,107	11,708	11,512
American Crow	1	2	39,904	39,930	19,188	12,180	11,682
Carrier Pigeon	2	2	29,214	29,231	26,483	12,373	11,933
Downy Woodpecker	3	4	29,007	29,046	14,283	11,755	11,298
Killdeer	4	6	30,353	30,374	14,649	11,959	11,492
Duck	6	8	28,897	28,914	23,951	12,063	11,769
Crested Ibis	0	1	Data not provided.				
Chicken	0	1					
Peregrine Falcon	0	2					
Zebra Finch	0	21					
Sum			**538,455**	**538,837**	**135,805**	**15,639**	**14,454**

Supplemental Table 2b. Best reference - shared defined common symbol counts.

Common Name	Canary	Bald Eagle	Burrowing Owl	Small Tree Finch	Chinese Goose	Cormorant	Pheasant	Starling	Grey Crowned Crane	Hooded Crow	Ruff	Tibetan Ground Jay	White Throated Sparrow	Annas Hummingbird	Turkey	Crow	Pigeon	Downy Woodpecker	Killdeer	Mallard
Symbol Overlaps (Defined)																				
Atlantic Canary	17261	11227	10873	11937	10711	8271	11903	11402	8885	10916	11099	11485	11020	11547	10679	11250	11316	10638	10737	11020
Bald Eagle	11227	15655	10789	11703	11341	8813	11683	11971	9523	10784	11813	12223	10566	11267	10511	11406	11472	11410	11606	10914
Burrowing Owl	10873	10789	15675	11089	10579	8352	11041	10717	8926	10537	10693	10789	10526	10988	10229	10687	10827	10320	10587	10775
Camarhynchus Parvulus	11937	11703	11089	16664	10875	8342	12281	11891	8962	10844	11453	12129	11002	11949	11010	11368	11464	10864	10936	11299
Chinese Goose	10711	11341	10579	10875	20704	8638	10967	11207	9259	10582	11163	11357	10268	10741	10046	10957	10970	10813	11077	10594
Common Cormorant	8271	8813	8352	8342	8638	14687	8368	8594	8136	8331	8537	8682	8114	8303	7789	8482	8473	8563	8800	8149
Common Pheasant	11903	11683	11041	12281	10967	8368	19182	11695	9024	10888	11470	11876	10990	11847	11220	11355	11610	10836	10949	11325
Common Starling	11402	11971	10717	11891	11207	8594	11695	16715	9234	10894	11774	12386	10711	11341	10498	11508	11416	11162	11238	10916
Grey Crowned Crane	8885	9523	8926	8962	9259	8136	9024	9234	15279	8871	9206	9354	8645	8892	8400	9093	9086	9172	9465	8728
Hooded Crow	10916	10784	10537	10844	10582	8331	10888	10894	8871	17118	10714	10948	10592	10665	9920	11171	11015	10412	10573	10409
Ruff	11099	11813	10693	11453	11163	8537	11470	11774	9206	10714	18471	11886	10467	11169	10408	11257	11336	11015	11207	10812
Tibetan Ground-Jay	11485	12223	10789	12129	11357	8682	11876	12386	9354	10948	11886	16415	10751	11420	10620	11601	11545	11305	11392	11004
White-Throated Sparrow	11020	10566	10526	11002	10268	8114	10990	10711	8645	10592	10467	10751	17249	10749	9949	10714	10745	10170	10287	10485
Anna's Hummingbird	11547	11267	10988	11949	10741	8303	11847	11341	8892	10665	11169	11420	10749	16080	10747	11008	11215	10675	10726	11186
Common Turkey	10679	10511	10229	11010	10046	7789	11220	10498	8400	9920	10408	10620	9949	10747	24107	10248	10486	9883	10013	10332
American Crow	11250	11406	10687	11368	10957	8482	11355	11508	9093	11171	11257	11601	10714	11008	10248	19188	11418	10832	10977	10662
Carrier Pigeon	11316	11472	10827	11464	10970	8473	11610	11416	9086	11015	11336	11545	10745	11215	10486	11418	26483	10842	10981	10890
Downy Woodpecker	10638	11410	10320	10864	10813	8563	10836	11162	9172	10412	11015	11305	10170	10675	9883	10832	10842	14283	11045	10311
Killdeer	10737	11606	10587	10936	11077	8800	10949	11238	9465	10573	11207	11392	10287	10726	10013	10977	10981	11045	14649	10507
Duck	11020	10914	10775	11299	10594	8149	11325	10916	8728	10409	10812	11004	10485	11186	10332	10662	10890	10311	10507	23951
Symbol Overlaps (No LOC)																				
Atlantic Canary	12394	11227	10873	11937	10711	8271	11903	11402	8885	10916	11099	11485	11020	11547	10679	11250	11316	10638	10737	11020
Bald Eagle	11227	12918	10789	11703	11341	8813	11683	11971	9523	10784	11813	12223	10566	11267	10511	11406	11472	11410	11606	10914
Burrowing Owl	10873	10789	11713	11089	10579	8352	11041	10717	8926	10537	10693	10789	10526	10988	10229	10687	10827	10320	10587	10775
Camarhynchus Parvulus	11937	11703	11089	13141	10875	8342	12281	11891	8962	10844	11453	12129	11002	11949	11010	11368	11464	10864	10936	11299
Chinese Goose	10711	11341	10579	10875	11828	8638	10967	11207	9259	10582	11163	11357	10268	10741	10046	10957	10970	10813	11077	10594
Common Cormorant	8271	8813	8352	8342	8638	9085	8368	8594	8136	8331	8537	8682	8114	8303	7789	8482	8473	8563	8800	8149
Common Pheasant	11903	11683	11041	12281	10967	8368	13082	11695	9024	10888	11470	11876	10990	11847	11220	11355	11610	10836	10949	11325
Common Starling	11402	11971	10717	11891	11207	8594	11695	12672	9234	10894	11774	12386	10711	11341	10498	11508	11416	11162	11238	10916
Grey Crowned Crane	8885	9523	8926	8962	9259	8136	9024	9234	9834	8871	9206	9354	8645	8892	8400	9093	9086	9172	9465	8728
Hooded Crow	10916	10784	10537	10844	10582	8331	10888	10894	8871	11552	10714	10948	10592	10665	9920	11171	11015	10412	10573	10409
Ruff	11099	11813	10693	11453	11163	8537	11470	11774	9206	10714	12397	11886	10467	11169	10408	11257	11336	11015	11207	10812
Tibetan Ground-Jay	11485	12223	10789	12129	11357	8682	11876	12386	9354	10948	11886	13051	10751	11420	10620	11601	11545	11305	11392	11004
White-Throated Sparrow	11020	10566	10526	11002	10268	8114	10990	10711	8645	10592	10467	10751	11491	10749	9949	10714	10745	10170	10287	10485
Anna's Hummingbird	11547	11267	10988	11949	10741	8303	11847	11341	8892	10665	11169	11420	10749	12454	10747	11008	11215	10675	10726	11186
Common Turkey	10679	10511	10229	11010	10046	7789	11220	10498	8400	9920	10408	10620	9949	10747	11708	10248	10486	9883	10013	10332
American Crow	11250	11406	10687	11368	10957	8482	11355	11508	9093	11171	11257	11601	10714	11008	10248	12180	11418	10832	10977	10662
Carrier Pigeon	11316	11472	10827	11464	10970	8473	11610	11416	9086	11015	11336	11545	10745	11215	10486	11418	12373	10842	10981	10890
Downy Woodpecker	10638	11410	10320	10864	10813	8563	10836	11162	9172	10412	11015	11305	10170	10675	9883	10832	10842	11755	11045	10311
Killdeer	10737	11606	10587	10936	11077	8800	10949	11238	9465	10573	11207	11392	10287	10726	10013	10977	10981	11045	11959	10507
Duck	11020	10914	10775	11299	10594	8149	11325	10916	8728	10409	10812	11004	10485	11186	10332	10662	10890	10311	10507	12063
Symbol Overlaps (HuRef)																				
Atlantic Canary	12202	11180	10827	11896	10665	8233	11737	11354	8849	10761	11051	11437	10857	11501	10641	11093	11153	10592	10694	10972
Bald Eagle	11180	12425	10672	11662	10952	8451	11645	11596	9141	10547	11447	11813	10447	11221	10475	11111	11250	10966	11155	10795
Burrowing Owl	10827	10672	11446	10948	10463	8258	11005	10600	8836	10421	10578	10670	10409	10845	10151	10570	10706	10209	10472	10579
Camarhynchus Parvulus	11896	11662	10948	12956	10836	8312	12241	11851	8930	10806	11413	12088	10964	11810	10940	11330	11423	10824	10899	11190
Chinese Goose	10665	10952	10463	10836	11423	8318	10931	10857	8928	10351	10818	10977	10153	10697	10012	10673	10757	10435	10689	10480
Common Cormorant	8233	8451	8258	8312	8318	8712	8339	8306	7801	8138	8253	8368	8016	8267	7761	8243	8297	8209	8437	8057
Common Pheasant	11737	11645	11005	12241	10931	8339	12885	11657	8995	10746	11432	11838	10838	11809	11179	11210	11450	10799	10915	11287
Common Starling	11354	11596	10600	11851	10857	8306	11657	12283	8935	10656	11409	11998	10591	11296	10462	11213	11196	10813	10885	10799
Grey Crowned Crane	8849	9141	8836	8930	8928	7801	8995	8935	9438	8676	8911	9029	8553	8857	8372	8851	8905	8800	9086	8639
Hooded Crow	10761	10547	10421	10806	10351	8138	10746	10656	8676	11131	10479	10706	10364	10622	9885	10768	10641	10184	10339	10296
Ruff	11051	11447	10578	11413	10818	8253	11432	11409	8911	10479	12018	11514	10351	11123	10372	10969	11114	10672	10857	10696
Tibetan Ground-Jay	11437	11813	10670	12088	10977	8368	11838	11998	9029	10706	11514	12620	10630	11374	10584	11300	11320	10925	11008	10885
White-Throated Sparrow	10857	10447	10409	10964	10153	8016	10838	10591	8553	10364	10351	10630	11235	10706	9913	10484	10509	10059	10171	10373
Anna's Hummingbird	11501	11221	10845	11810	10697	8267	11809	11296	8857	10622	11123	11374	10706	12268	10681	10965	11172	10631	10684	11076
Common Turkey	10641	10475	10151	10940	10012	7761	11179	10462	8372	9885	10372	10584	9913	10681	11512	10211	10447	9847	9980	10250
American Crow	11093	11111	10570	11330	10673	8243	11210	11213	8851	10768	10969	11300	10484	10965	10211	11682	11037	10549	10690	10551
Carrier Pigeon	11153	11250	10706	11423	10757	8297	11450	11196	8905	10641	11114	11320	10509	11172	10447	11037	11933	10631	10762	10772
Downy Woodpecker	10592	10966	10209	10824	10435	8209	10799	10813	8800	10184	10672	10925	10059	10631	9847	10549	10631	11298	10607	10199
Killdeer	10694	11155	10472	10899	10689	8437	10915	10885	9086	10339	10857	11008	10171	10684	9980	10690	10762	10607	11492	10392
Duck	10972	10795	10579	11190	10480	8057	11287	10799	8639	10296	10696	10885	10373	11076	10250	10551	10772	10199	10392	11769

Supplemental Table 3. SwitssProt results. The percentage of query proteins observed in the corresponding AvianDB reference genome for which an ID match of N% identity (rows) was identified in the Human reference genome.

% Identity to Human Reference Protein	01.CommonGrackle	02.WoodDuck	03.Ostrich	04.ChippingSparrow	05.AmericanCrow	06.GreenWingedTeal	07.NorthernShoveler	08.CanadaGoose	09.DownyWoodpecker	10.AmericanRobin	11.WhiteBreastedNuthatch	12.CarolinaWren	13.Killdeer	14.RedEyedVireo	15.BarnSwallow	17.SongSparrow	18.Pheasant	19.HouseWren	20.PintailDuck	21.RubyThroatedHummingbird	23.Mallard	24.RedBelliedWoodpecker	26.Gadwall	27.YellowRumpedWarbler	28.GrayCatbird	30.SpottedSandpiper	31.Starling	32.HornedLark	33.TuftedTitmouse	34.MourningDove	35.YellowWarbler	36.NorthernCardinal	37.HouseFinch	38.TreeSwallow	39.BrownHeadedCowbird	40.RingBilledGull	41.YellowThroatedWarbler	42.HouseSparrow	61.MuteSwan	62.DoubleCrestedCormorant	63.AmericanTreeSparrow	64.Woodcock	65.Turkey	66.HerringGull	67.CoopersHawk	68.RockDove	69.RedTailedHawk	70.CaspianTern	71.SandhillCrane	72.HairyWoodpecker	73.GreatHornedOwl	74.RuffedGrouse	75.CedarWaxwing
100	0.79	0.83	0.74	0.97	0.81	1.08	0.94	0.71	0.76	0.73	0.75	0.87	0.89	0.55	0.73	0.80	0.78	1.00	0.84	0.81	1.03	0.67	0.96	0.81	0.89	1.02	0.75	0.78	0.67	0.84	0.74	0.98	0.97	0.71	0.82	0.88	0.80	0.64	0.84	0.87	8.17	0.81	1.05	0.90	0.82	0.74	0.70	0.75	0.91	0.89	0.89	0.96	0.99
99	0.82	0.76	0.77	0.66	0.69	0.73	0.66	0.76	0.76	0.89	0.59	0.65	0.61	0.70	0.67	0.74	0.75	0.71	0.71	0.68	0.81	0.63	0.81	0.83	0.65	0.81	0.74	0.64	0.75	0.69	0.72	0.72	0.93	0.64	0.65	0.88	0.74	0.58	0.67	0.64	10.13	0.78	0.70	0.80	0.85	0.77	0.74	0.83	0.70	0.66	0.67	0.64	0.61
98	1.39	1.63	1.53	1.47	1.48	1.48	1.79	1.50	1.54	1.59	1.41	1.50	1.55	1.50	1.44	1.59	1.55	1.46	1.52	1.50	1.71	1.43	1.90	1.62	1.99	1.87	1.39	1.67	1.60	1.56	1.43	1.61	1.79	1.43	1.32	1.97	1.48	1.38	1.58	1.48	12.04	1.83	1.56	1.73	1.66	1.52	1.52	1.65	1.62	1.55	1.56	1.39	1.46
97	1.78	1.90	1.90	1.67	1.74	1.59	2.13	1.76	1.57	1.80	1.72	1.81	1.71	1.77	1.53	1.82	1.72	1.80	1.73	1.71	1.94	1.65	2.33	1.82	2.01	2.00	1.97	1.84	1.67	1.89	1.79	1.78	2.14	1.49	1.81	2.12	1.78	1.53	1.91	1.76	8.30	1.93	2.06	1.71	1.85	1.75	1.59	1.59	1.79	1.68	1.78	1.82	1.66
96	2.01	2.01	2.34	1.65	1.89	2.23	2.33	1.95	1.86	2.10	1.93	1.95	2.17	2.03	1.99	1.98	2.00	2.16	2.33	1.94	2.12	1.98	2.15	2.04	2.22	2.42	2.02	2.10	2.06	1.95	1.88	2.08	2.08	1.91	1.83	2.31	1.81	1.99	2.23	2.21	6.44	2.26	2.25	2.00	2.05	2.11	1.77	1.90	2.16	2.24	2.06	2.00	1.92
95	2.03	2.26	2.22	1.93	2.07	2.12	2.37	2.17	1.99	1.90	2.20	1.95	2.29	1.91	2.27	2.16	2.11	1.90	2.27	2.19	2.55	1.85	2.45	2.40	2.43	2.44	2.09	2.10	2.06	2.20	2.11	2.09	2.50	1.88	2.13	2.57	2.21	2.05	2.19	2.14	5.03	2.27	2.23	2.16	2.11	2.23	1.97	2.03	2.47	2.05	2.10	1.95	1.96
94	2.66	2.48	2.30	2.26	2.15	2.34	2.40	2.19	2.07	2.20	2.33	2.07	2.09	2.31	2.29	2.43	2.34	2.14	2.47	2.42	2.42	2.10	2.86	2.48	2.35	2.42	2.12	2.12	2.34	2.32	2.29	2.29	2.75	2.17	2.61	2.79	2.23	2.21	2.37	2.38	3.79	2.33	2.36	2.40	2.31	2.20	2.35	2.23	2.25	2.20	2.12	2.12	2.18
93	2.38	2.44	2.51	2.27	2.28	2.30	2.33	2.37	2.38	2.40	2.27	2.49	2.63	2.51	2.18	2.49	2.49	2.45	2.51	2.49	2.52	2.41	2.66	2.52	2.48	2.56	2.24	2.62	2.52	2.34	2.39	2.61	2.37	2.41	2.23	2.68	2.25	2.32	2.54	2.57	3.08	2.55	2.31	2.57	2.42	2.47	2.45	2.37	2.27	2.33	2.28	2.29	2.18
92	2.34	2.54	2.43	2.57	2.20	2.51	2.69	2.43	2.35	2.27	2.33	2.29	2.34	2.46	2.50	2.60	2.62	2.45	2.55	2.59	2.47	2.18	2.61	2.45	2.50	2.69	2.45	2.38	2.55	2.33	2.68	2.49	2.66	2.53	2.71	2.63	2.52	2.47	2.54	2.44	2.86	2.37	2.52	2.57	2.22	2.49	2.43	2.53	2.68	2.44	2.42	2.39	2.35
91	2.50	2.89	2.43	2.55	2.55	2.50	2.82	2.63	2.33	2.51	2.50	2.53	2.64	2.71	2.44	2.64	2.58	2.49	2.86	2.43	2.58	2.13	3.08	2.61	2.62	2.77	2.54	2.56	2.50	2.67	2.72	2.68	2.70	2.24	2.54	2.59	2.34	2.49	2.66	2.63	2.25	2.75	2.75	2.50	2.71	2.53	2.49	2.51	2.56	2.30	2.42	2.39	2.46
90	2.39	2.81	2.51	2.55	2.41	2.50	2.78	2.58	2.35	2.51	2.52	2.66	2.72	2.49	2.61	2.52	2.48	2.55	2.63	2.42	2.68	2.36	2.64	2.76	2.44	2.57	2.43	2.51	2.52	2.87	2.41	2.61	2.69	2.55	2.40	2.68	2.57	2.44	2.67	2.60	1.97	2.57	2.71	2.68	2.55	2.78	2.64	2.54	2.55	2.24	2.47	2.38	2.35
89	2.38	2.61	2.48	2.30	2.26	2.43	2.51	2.30	2.55	2.57	2.33	2.39	2.33	2.34	2.54	2.52	2.47	2.20	2.55	2.33	2.46	2.42	2.62	2.62	2.26	2.71	2.42	2.40	2.44	2.41	2.65	2.31	2.63	2.41	2.39	2.66	2.31	2.51	2.59	2.52	1.63	2.43	2.44	2.38	2.42	2.23	2.46	2.31	2.66	2.30	2.44	2.24	2.40
88	2.42	2.35	2.56	2.30	2.58	2.45	2.75	2.36	2.36	2.48	2.58	2.50	2.57	2.35	2.69	2.52	2.47	2.32	2.55	2.39	2.50	2.35	2.43	2.70	2.50	2.41	2.47	2.61	2.36	2.42	2.33	2.62	2.32	2.45	2.50	2.95	2.51	2.41	2.48	2.59	1.57	2.57	2.68	2.79	2.57	2.26	2.39	2.45	2.51	2.56	2.55	2.43	2.62
87	2.12	2.51	2.65	2.32	2.42	2.34	2.52	2.59	2.21	2.29	2.36	2.49	2.39	2.24	2.32	2.40	2.30	2.25	2.30	2.25	2.59	2.48	2.38	2.52	2.45	2.62	2.52	2.40	2.25	2.40	2.39	2.15	2.55	2.40	2.53	2.61	2.19	2.54	2.62	2.48	1.42	2.71	2.60	2.55	2.38	2.48	2.42	2.53	2.13	2.29	2.46	2.24	2.38
86	2.06	2.73	2.35	2.31	2.22	2.48	2.67	2.49	2.44	2.55	2.41	2.54	2.59	2.41	2.41	2.55	2.40	2.56	2.38	2.71	2.46	2.34	2.50	2.38	2.32	2.28	2.35	2.18	2.80	2.34	2.67	2.62	2.49	2.18	2.47	2.59	2.61	2.16	2.53	2.37	1.48	2.52	2.51	2.42	2.50	2.42	2.63	2.56	2.60	2.36	2.35	2.53	2.54
85	2.50	2.41	2.41	2.54	2.25	2.44	2.59	2.55	2.61	2.50	2.39	2.43	2.41	2.47	2.42	2.35	2.35	2.24	2.32	2.44	2.34	2.53	2.65	2.44	2.53	2.30	2.03	2.41	2.42	2.39	2.42	2.38	2.85	2.36	2.53	2.25	2.41	2.28	2.54	2.33	1.24	2.51	2.54	2.63	2.49	2.67	2.55	2.32	2.51	2.53	2.50	2.21	2.28
84	2.32	2.31	2.37	2.30	2.20	2.21	2.43	2.40	2.47	2.28	2.37	2.49	2.41	2.34	2.30	2.31	2.25	2.40	2.27	2.30	2.47	2.39	2.66	2.43	2.29	2.63	2.30	2.45	2.31	2.35	2.30	2.32	2.29	2.15	2.32	2.55	2.35	2.36	2.27	2.47	1.22	2.37	2.37	2.13	2.42	2.30	2.36	2.34	2.49	2.24	2.50	2.20	2.44
83	2.23	2.48	2.43	2.15	2.53	2.30	2.44	2.41	2.30	2.15	2.37	2.18	2.44	2.23	2.25	1.97	2.51	2.15	2.32	2.41	2.36	2.36	2.25	2.26	2.29	2.59	2.27	2.25	2.15	2.28	2.16	2.09	2.35	2.34	2.24	2.25	2.27	2.34	2.80	2.53	1.07	2.43	2.35	2.25	2.48	2.35	2.34	2.32	2.27	2.43	2.35	2.26	2.04
82	2.14	2.38	2.42	2.35	2.20	2.30	2.29	2.42	2.48	2.21	2.50	2.50	2.21	2.26	2.57	2.42	2.37	2.36	2.19	2.18	2.28	2.24	2.06	2.36	2.39	2.48	2.34	2.17	2.25	2.58	2.42	2.24	2.21	2.30	2.04	2.43	2.30	2.33	2.54	2.20	1.09	2.45	2.24	2.19	2.07	2.62	2.17	2.20	2.43	2.43	2.20	2.27	2.19
81	2.23	2.03	2.30	2.32	2.25	2.21	2.26	2.14	2.38	2.17	2.32	2.19	2.29	2.31	2.04	2.30	2.10	2.45	2.30	2.22	2.10	2.31	2.08	2.19	2.16	2.28	2.40	2.38	2.27	2.42	2.01	2.32	2.26	2.34	2.23	2.08	2.22	2.27	2.07	2.45	1.07	2.36	2.27	2.18	2.37	2.28	2.31	2.23	2.17	2.29	2.19	2.26	2.24
80	2.44	2.62	2.88	2.22	2.45	2.29	2.45	2.34	2.27	2.15	2.36	2.35	2.27	2.27	2.31	2.38	2.42	2.30	2.38	2.30	2.87	2.30	2.34	2.26	2.38	2.56	2.29	2.18	2.40	2.50	2.13	2.33	2.48	2.30	2.18	2.57	2.29	2.30	2.58	2.49	1.18	2.33	2.40	2.42	2.42	2.26	2.35	2.49	2.43	2.34	2.44	2.21	2.27
79	2.09	2.12	1.97	1.83	2.12	2.15	2.25	2.34	2.24	2.02	2.14	2.15	2.31	2.22	2.00	2.14	2.08	1.90	2.35	2.09	2.26	2.25	1.95	1.85	2.00	1.79	2.15	2.19	2.31	1.99	1.97	1.85	1.95	2.07	2.17	2.16	2.16	2.09	2.20	2.05	0.95	1.90	1.89	2.15	2.20	2.16	2.24	2.05	2.04	2.03	2.04	1.97	2.01
78	2.33	2.30	2.27	1.93	2.20	2.29	2.28	2.14	2.27	2.23	2.00	2.07	2.11	2.02	2.18	2.40	2.39	2.45	2.30	2.18	2.23	2.17	2.18	2.10	2.13	2.17	2.35	2.36	2.15	2.21	2.14	2.23	2.34	2.15	2.21	1.93	2.13	2.14	2.16	2.25	0.98	2.20	2.13	2.33	2.42	2.29	2.39	2.10	2.20	2.14	2.23	2.00	2.08
77	1.98	2.53	2.20	1.99	2.17	2.06	2.06	2.25	2.01	2.07	2.15	1.97	2.31	2.11	2.25	2.13	2.13	2.02	2.02	2.11	2.06	2.27	1.97	2.30	2.09	2.23	2.32	2.06	2.32	2.06	2.11	2.17	1.95	2.22	2.11	2.23	2.09	2.20	2.00	2.07	0.83	2.12	1.98	2.28	2.12	1.98	2.28	2.08	2.14	2.29	2.26	2.03	2.10
76	2.03	2.27	2.19	2.03	2.25	2.08	2.13	2.11	2.24	2.09	2.14	2.19	2.25	2.01	1.99	2.03	2.09	2.16	2.17	2.24	2.10	2.18	2.25	2.07	2.37	2.13	2.42	2.08	2.21	2.00	2.26	2.08	2.03	2.14	2.09	2.32	2.24	2.21	2.19	2.24	1.06	2.11	2.11	2.08	2.21	2.28	2.14	2.23	2.25	2.23	2.11	2.20	2.21
75	2.21	2.33	2.39	2.11	2.35	2.41	2.26	2.41	2.48	2.19	2.35	2.19	2.33	1.94	2.15	2.24	2.26	2.27	2.30	2.38	2.29	2.20	2.33	2.29	2.38	2.15	2.14	2.09	2.20	2.10	2.49	2.28	2.22	2.18	2.37	2.19	2.49	2.41	2.34	2.10	1.07	2.32	2.22	2.09	2.12	2.28	2.27	2.40	2.16	2.25	2.14	2.38	2.20
74	2.11	1.80	2.02	1.90	1.96	1.79	1.71	1.85	2.07	2.06	1.79	1.85	1.95	1.75	1.93	1.94	2.11	1.88	1.95	1.79	2.07	2.03	1.71	1.92	1.94	1.72	1.93	1.97	1.96	1.81	2.03	1.91	1.74	1.95	1.98	2.00	1.98	2.21	1.91	2.11	0.89	1.87	1.80	2.11	2.02	2.09	2.07	1.95	1.99	2.01	2.05	2.02	2.03
73	2.00	1.86	1.90	1.88	2.19	1.90	1.79	2.17	1.94	1.87	1.77	1.87	1.97	2.10	1.96	1.90	1.93	1.98	1.89	1.85	1.91	2.11	1.99	2.02	2.11	1.86	2.02	1.95	1.90	2.09	2.03	1.96	2.10	1.90	1.76	2.01	1.71	1.92	1.83	2.04	0.86	1.90	2.08	1.95	2.07	1.95	1.98	2.10	2.02	1.87	1.98	2.22	1.99
72	1.88	1.87	2.21	1.93	1.97	1.95	1.78	2.14	1.83	1.97	2.11	2.20	1.82	2.06	2.02	1.95	2.20	1.94	2.17	1.97	1.95	1.98	1.87	1.83	1.93	1.99	1.92	1.89	1.78	1.81	1.80	1.93	1.86	2.03	1.88	1.80	1.86	2.14	1.91	1.90	0.84	2.14	2.01	2.06	2.29	2.10	2.37	1.99	1.91	2.12	2.20	2.07	1.98
71	1.97	1.95	1.81	1.83	1.75	1.85	1.94	1.96	1.77	1.96	1.86	1.93	1.70	1.92	1.85	1.78	2.02	1.97	2.03	1.94	1.92	1.92	1.94	1.90	1.61	1.81	1.90	1.80	1.76	1.88	1.95	2.05	1.78	2.16	1.76	1.78	1.81	1.93	1.87	1.90	0.84	1.76	1.78	1.84	1.78	1.82	1.91	1.84	1.91	1.75	1.89	1.96	1.95
70	1.80	1.92	2.01	1.93	2.06	1.84	1.76	1.73	1.79	1.60	1.90	1.81	1.94	1.91	1.69	1.78	2.07	1.94	1.76	2.08	1.98	1.91	1.80	1.79	1.79	1.95	2.07	1.95	1.89	1.84	1.75	1.90	1.81	2.08	1.75	1.87	1.81	1.84	1.72	1.96	0.78	1.88	1.93	1.94	2.02	1.88	1.71	1.95	1.90	1.85	1.81	1.94	1.92
69	1.66	1.58	1.80	1.84	1.74	1.89	1.52	1.80	1.76	1.80	1.74	1.62	1.62	1.94	1.87	1.82	1.90	1.80	1.71	1.71	1.83	1.69	1.71	1.83	1.75	1.59	1.74	1.54	1.75	1.85	1.72	1.69	1.75	1.57	1.85	1.62	1.86	1.95	1.86	1.80	0.76	1.51	1.75	1.64	1.78	1.82	1.88	1.72	1.84	1.71	1.82	1.78	1.72
68	1.80	1.76	1.73	1.83	1.97	1.75	1.74	1.66	1.86	1.72	1.71	1.71	1.81	1.81	1.93	1.83	1.71	1.82	1.81	1.81	1.55	1.76	1.55	1.72	1.72	1.69	1.65	1.64	1.48	1.74	1.69	1.87	1.52	1.74	1.88	1.69	1.85	1.72	1.74	1.70	0.76	1.57	1.59	1.93	1.86	1.73	1.70	1.75	1.71	1.70	1.73	1.85	1.72
67	1.66	1.58	1.66	1.81	1.62	1.66	1.63	1.69	1.69	1.62	1.69	1.69	1.73	1.66	1.66	1.59	1.71	1.71	1.49	1.80	1.53	1.80	1.57	1.52	1.73	1.58	1.85	1.80	1.54	1.66	1.79	1.62	1.59	1.95	1.75	1.42	1.82	1.68	1.60	1.59	0.71	1.55	1.75	1.52	1.80	1.58	1.79	1.74	1.63	1.69	1.65	1.62	1.68
66	1.70	1.62	1.63	1.51	1.80	1.65	1.32	1.74	1.73	1.64	1.73	1.59	1.67	1.87	1.57	1.78	1.71	1.59	1.49	1.70	1.50	1.78	1.74	1.52	1.57	1.63	1.63	1.71	1.74	1.66	1.49	1.53	1.57	1.64	1.70	1.49	1.81	1.51	1.65	1.72	0.75	1.65	1.50	1.64	1.64	1.67	1.73	1.61	1.66	1.47	1.60	1.70	1.63
65	1.72	1.70	1.44	1.60	1.59	1.60	1.57	1.47	1.69	1.46	1.67	1.65	1.47	1.76	1.88	1.67	1.73	1.54	1.54	1.64	1.38	1.52	1.49	1.63	1.76	1.51	1.52	1.78	2.00	1.52	1.54	1.57	1.57	1.73	1.66	1.30	1.52	1.69	1.34	1.38	0.70	1.59	1.54	1.44	1.67	1.56	1.55	1.43	1.47	1.69	1.56	1.73	1.66
64	1.58	1.42	1.46	1.58	1.71	1.54	1.42	1.35	1.62	1.64	1.66	1.49	1.30	1.59	1.81	1.49	1.59	1.68	1.46	1.53	1.36	1.41	1.26	1.33	1.42	1.68	1.62	1.49	1.67	1.44	1.50	1.58	1.48	1.58	1.50	1.33	1.40	1.55	1.37	1.53	0.58	1.39	1.44	1.43	1.41	1.47	1.57	1.54	1.45	1.59	1.40	1.45	1.68
63	1.51	1.45	1.45	1.49	1.55	1.41	1.40	1.38	1.41	1.63	1.44	1.48	1.51	1.53	1.57	1.46	1.40	1.60	1.26	1.45	1.41	1.35	1.26	1.48	1.39	1.35	1.36	1.44	1.45	1.52	1.40	1.40	1.34	1.51	1.46	1.31	1.60	1.48	1.34	1.51	0.61	1.42	1.44	1.35	1.24	1.54	1.41	1.50	1.46	1.49	1.46	1.43	1.54
62	1.38	1.49	1.45	1.58	1.43	1.47	1.35	1.24	1.38	1.62	1.39	1.63	1.32	1.45	1.48	1.43	1.43	1.54	1.43	1.35	1.36	1.51	1.17	1.67	1.35	1.19	1.55	1.56	1.61	1.49	1.49	1.30	1.44	1.53	1.49	1.31	1.50	1.34	1.37	1.52	0.55	1.44	1.42	1.32	1.49	1.42	1.48	1.24	1.44	1.39	1.35	1.33	1.41
61	1.59	1.20	1.28	1.38	1.38	1.31	1.23	1.40	1.29	1.45	1.32	1.40	1.42	1.31	1.41	1.52	1.27	1.28	1.15	1.37	1.18	1.47	1.31	1.29	1.30	1.23	1.32	1.34	1.45	1.33	1.49	1.53	1.31	1.49	1.38	1.24	1.33	1.22	1.25	1.37	0.52	1.38	1.23	1.41	1.35	1.46	1.52	1.33	1.29	1.24	1.37	1.36	1.40
60	1.52	1.28	1.44	1.44	1.53	1.31	1.20	1.37	1.34	1.28	1.40	1.38	1.46	1.38	1.34	1.40	1.34	1.41	1.51	1.31	1.48	1.48	1.30	1.14	1.38	1.43	1.27	1.52	1.34	1.37	1.61	1.35	1.37	1.43	1.50	1.26	1.28	1.34	1.17	1.14	0.62	1.32	1.41	1.29	1.33	1.35	1.22	1.42	1.39	1.38	1.66	1.45	1.55
59	1.36	1.06	1.10	1.35	1.24	1.09	1.46	1.02	1.30	1.13	1.14	1.04	1.11	1.26	1.16	1.01	1.21	1.20	1.33	1.24	1.16	1.22	1.21	1.04	1.15	1.14	1.06	1.17	1.14	1.08	1.27	1.33	1.07	1.09	1.21	1.09	1.23	1.18	1.12	1.07	0.46	1.11	1.23	1.15	1.10	1.05	1.20	1.12	1.06	1.24	1.08	1.28	1.21
58	1.23	1.12	1.11	1.51	1.29	1.15	1.05	1.22	1.28	1.17	1.21	1.38	1.34	1.22	1.29	1.24	1.16	1.35	1.11	1.26	1.01	1.21	1.08	1.15	1.10	1.19	1.39	1.34	1.30	1.13	1.22	1.17	1.06	1.39	1.23	0.88	1.21	1.27	1.22	1.09	0.49	0.94	1.16	1.00	0.97	1.17	1.14	1.22	1.01	1.19	1.08	1.25	1.21
57	1.28	0.98	1.05	1.27	1.08	1.13	0.95	1.00	1.13	1.18	1.17	1.25	1.05	1.24	1.43	1.13	1.09	1.13	0.99	1.17	0.97	1.30	1.12	1.14	0.92	1.05	1.19	1.10	1.15	1.08	1.23	1.27	0.94	1.29	1.19	1.20	1.24	1.16	1.07	1.11	0.50	1.10	1.12	1.08	1.02	1.22	1.25	1.11	1.17	1.40	1.12	1.15	1.18
56	1.19	0.98	0.97	1.25	1.12	1.11	0.97	1.13	1.03	1.25	1.26	1.17	0.85	1.40	1.16	1.21	1.25	1.21	1.10	1.33	1.20	1.26	0.99	1.09	1.02	1.01	1.02	1.13	1.31	1.15	1.18	1.17	1.15	1.24	1.20	0.93	1.21	1.16	1.03	1.05	0.49	1.09	1.05	1.18	0.91	1.00	0.95	1.17	1.04	1.00	1.10	1.17	1.21
55	1.01	1.01	0.82	1.03	1.18	1.01	0.95	1.00	1.12	1.05	1.20	0.92	1.03	1.04	1.09	1.10	1.02	0.98	0.90	1.20	0.96	1.25	0.95	1.18	0.98	1.01	1.02	1.10	0.91	1.07	0.91	1.00	0.97	1.00	1.17	0.95	1.06	1.11	0.98	1.07	0.46	0.99	1.06	1.18	0.95	1.13	1.03	1.17	1.11	1.15	0.98	1.06	1.04
54	0.94	0.83	0.89	1.20	1.02	1.07	0.89	1.09	1.00	1.08	1.02	1.00	0.96	0.87	0.83	0.92	0.83	1.05	1.08	1.04	0.86	1.10	0.88	0.90	0.97	0.94	1.10	1.05	0.87	0.90	0.88	0.94	0.88	1.01	0.95	0.84	0.97	1.04	0.98	0.96	0.37	0.85	0.95	0.95	0.98	0.97	0.95	1.02	0.98	1.09	0.93	1.01	0.99
53	0.96	0.84	0.82	1.05	0.91	1.01	0.92	0.98	1.06	1.03	0.94	0.92	0.87	0.95	0.99	0.93	0.85	1.03	0.82	0.90	0.87	1.10	0.77	1.01	1.01	0.84	0.99	1.16	0.97	1.00	0.95	0.99	1.04	1.09	0.87	0.89	1.07	0.90	0.95	1.06	0.43	1.03	0.91	0.96	0.89	0.97	0.90	0.94	0.88	1.05	0.96	0.98	0.94
52	0.82	0.72	0.93	0.99	0.95	0.89	0.86	0.89	0.91	1.02	0.98	0.86	0.82	0.89	0.95	0.89	0.87	0.82	0.92	0.90	0.89	0.96	0.93	0.93	0.73	0.81	1.00	0.91	1.06	0.95	0.84	0.82	0.71	1.00	0.99	0.83	1.04	1.01	0.90	0.92	0.46	0.91	0.86	0.83	0.87	0.92	0.89	0.93	0.83	1.05	0.94	1.07	1.04
51	0.96	0.82	0.84	0.89	0.84	0.86	0.87	0.76	1.00	0.85	0.93	0.93	0.75	0.98	0.82	0.77	0.84	0.86	0.81	0.92	0.92	0.90	0.83	0.76	0.78	0.67	0.80	0.86	0.89	0.76	0.94	0.86	0.97	1.05	0.83	0.61	0.91	0.99	0.85	0.87	0.35	0.94	0.83	0.73	0.81	0.76	0.85	0.95	0.74	0.98	0.79	0.90	0.98
50	0.90	0.96	0.97	1.07	0.92	0.97	0.81	0.89	1.09	0.90	1.06	0.87	0.86	0.84	0.86	0.80	0.87	0.89	0.88	1.03	0.85	0.96	0.90	0.99	1.03	1.01	0.86	1.01	0.94	0.91	1.05	0.84	0.85	0.84	0.95	0.81	0.98	0.94	0.77	0.85	0.39	0.89	0.96	0.88	0.89	0.83	0.98	0.92	0.74	0.98	1.05	1.07	0.97
49	0.62	0.68	0.55	0.79	0.70	0.70	0.69	0.74	0.70	0.59	0.62	0.69	0.58	0.62	0.61	0.61	0.53	0.65	0.52	0.59	0.62	0.68	0.59	0.55	0.58	0.65	0.71	0.61	0.55	0.63	0.63	0.58	0.68	0.62	0.63	0.53	0.54	0.74	0.48	0.63	0.20	0.53	0.64	0.74	0.71	0.58	0.61	0.60	0.66	0.68	0.68	0.77	0.60
48	0.75	0.69	0.71	0.68	0.75	0.78	0.71	0.74	0.72	0.80	0.71	0.64	0.85	0.80	0.68	0.67	0.63	0.82	0.82	0.67	0.73	0.85	0.61	0.67	0.79	0.58	0.75	0.57	0.61	0.69	0.68	0.67	0.55	0.72	0.72	0.66	0.73	0.73	0.73	0.63	0.26	0.84	0.70	0.66	0.63	0.71	0.68	0.77	0.67	0.81	0.86	0.77	0.81
47	0.78	0.60	0.65	0.82	0.64	0.65	0.53	0.64	0.77	0.77	0.74	0.69	0.64	0.67	0.58	0.64	0.64	0.62	0.68	0.67	0.71	0.69	0.49	0.62	0.67	0.64	0.65	0.63	0.68	0.65	0.67	0.80	0.59	0.69	0.69	0.63	0.69	0.77	0.72	0.75	0.24	0.73	0.58	0.76	0.64	0.70	0.63	0.66	0.67	0.63	0.69	0.78	0.70
46	0.70	0.57	0.48	0.77	0.61	0.65	0.56	0.79	0.72	0.59	0.64	0.76	0.73	0.62	0.72	0.64	0.60	0.63	0.63	0.68	0.48	0.69	0.61	0.70	0.66	0.57	0.69	0.72	0.65	0.60	0.74	0.66	0.56	0.66	0.78	0.61	0.74	0.64	0.63	0.53	0.23	0.64	0.64	0.60	0.58	0.45	0.53	0.64	0.59	0.55	0.70	0.70	0.65
45	0.62	0.57	0.56	0.58	0.61	0.65	0.56	0.64	0.70	0.68	0.68	0.66	0.58	0.58	0.69	0.57	0.57	0.59	0.69	0.60	0.60	0.63	0.51	0.52	0.76	0.58	0.61	0.58	0.60	0.78	0.68	0.54	0.48	0.69	0.48	0.73	0.52	0.69	0.60	0.56	0.21	0.64	0.46	0.56	0.68	0.71	0.67	0.63	0.69	0.65	0.73	0.70	0.68
44	0.62	0.55	0.44	0.57	0.61	0.62	0.40	0.61	0.63	0.62	0.56	0.57	0.62	0.50	0.52	0.49	0.56	0.63	0.58	0.56	0.49	0.61	0.49	0.53	0.59	0.56	0.49	0.58	0.53	0.53	0.53	0.68	0.55	0.55	0.58	0.49	0.52	0.64	0.56	0.47	0.19	0.53	0.57	0.56	0.52	0.55	0.50	0.58	0.60	0.52	0.59	0.71	0.61
43	0.51	0.44	0.53	0.67	0.60	0.52	0.56	0.50	0.53	0.52	0.60	0.51	0.48	0.55	0.45	0.61	0.45	0.52	0.57	0.48	0.40	0.47	0.47	0.40	0.51	0.45	0.47	0.51	0.41	0.51	0.53	0.64	0.42	0.56	0.52	0.61	0.56	0.62	0.49	0.47	0.19	0.58	0.37	0.57	0.47	0.54	0.53	0.60	0.53	0.58	0.46	0.54	0.55
42	0.55	0.47	0.49	0.48	0.41	0.51	0.48	0.53	0.49	0.65	0.48	0.60	0.55	0.50	0.52	0.42	0.42	0.58	0.40	0.41	0.35	0.51	0.47	0.49	0.39	0.50	0.52	0.48	0.48	0.47	0.43	0.42	0.42	0.49	0.53	0.33	0.57	0.46	0.50	0.49	0.17	0.43	0.42	0.55	0.46	0.47	0.49	0.54	0.45	0.45	0.44	0.51	0.52
41	0.42	0.46	0.38	0.43	0.42	0.46	0.52	0.47	0.44	0.48	0.47	0.43	0.42	0.47	0.44	0.43	0.46	0.38	0.53	0.42	0.43	0.50	0.43	0.36	0.40	0.49	0.42	0.41	0.41	0.50	0.39	0.38	0.56	0.45	0.49	0.44	0.41	0.46	0.43	0.44	0.16	0.39	0.44	0.42	0.38	0.38	0.37	0.38	0.41	0.43	0.49	0.47	0.44
40	0.50	0.38	0.37	0.39	0.38	0.46	0.30	0.44	0.40	0.46	0.41	0.43	0.47	0.41	0.39	0.47	0.45	0.47	0.42	0.45	0.43	0.49	0.43	0.32	0.39	0.25	0.39	0.38	0.37	0.45	0.45	0.40	0.41	0.38	0.38	0.37	0.34	0.34	0.46	0.39	0.15	0.33	0.33	0.42	0.43	0.39	0.40	0.41	0.34	0.41	0.47	0.54	0.45
39	0.31	0.31	0.28	0.40	0.31	0.38	0.38	0.40	0.30	0.46	0.32	0.37	0.45	0.46	0.38	0.33	0.35	0.41	0.28	0.34	0.28	0.39	0.34	0.37	0.46	0.29	0.31	0.33	0.29	0.43	0.37	0.32	0.30	0.31	0.39	0.27	0.31	0.41	0.40	0.33	0.14	0.32	0.40	0.36	0.38	0.34	0.33	0.42	0.33	0.33	0.32	0.42	0.50
38	0.39	0.33	0.34	0.35	0.29	0.36	0.33	0.31	0.36	0.33	0.32	0.37	0.39	0.35	0.32	0.34	0.36	0.29	0.36	0.32	0.29	0.26	0.34	0.41	0.27	0.31	0.42	0.33	0.38	0.34	0.32	0.34	0.22	0.37	0.34	0.26	0.38	0.29	0.34	0.33	0.11	0.35	0.25	0.25	0.28	0.29	0.39	0.31	0.43	0.37	0.39	0.30	0.35
37	0.36	0.27	0.31	0.37	0.31	0.34	0.21	0.33	0.28	0.42	0.31	0.28	0.32	0.33	0.29	0.32	0.22	0.32	0.29	0.27	0.26	0.30	0.25	0.26	0.32	0.18	0.31	0.25	0.33	0.32	0.40	0.28	0.27	0.36	0.36	0.19	0.36	0.28	0.17	0.29	0.12	0.29	0.31	0.32	0.31	0.29	0.30	0.40	0.36	0.29	0.29	0.32	0.34
36	0.31	0.23	0.18	0.28	0.32	0.23	0.22	0.22	0.38	0.28	0.25	0.36	0.22	0.30	0.32	0.26	0.31	0.31	0.20	0.26	0.24	0.35	0.26	0.33	0.33	0.23	0.39	0.33	0.25	0.30	0.24	0.26	0.19	0.30	0.32	0.28	0.30	0.36	0.25	0.29	0.08	0.26	0.32	0.21	0.32	0.27	0.23	0.25	0.23	0.33	0.24	0.32	0.30
35	0.25	0.21	0.28	0.22	0.24	0.35	0.27	0.27	0.21	0.24	0.21	0.26	0.23	0.24	0.20	0.25	0.27	0.22	0.22	0.26	0.23	0.27	0.24	0.22	0.19	0.15	0.20	0.27	0.22	0.33	0.23	0.36	0.25	0.36	0.24	0.29	0.32	0.27	0.18	0.20	0.09	0.21	0.28	0.25	0.23	0.26	0.24	0.26	0.31	0.36	0.28	0.31	0.27
34	0.23	0.17	0.25	0.27	0.31	0.14	0.27	0.17	0.22	0.21	0.28	0.19	0.22	0.22	0.21	0.19	0.25	0.17	0.22	0.19	0.21	0.22	0.24	0.27	0.18	0.11	0.29	0.19	0.26	0.27	0.32	0.20	0.24	0.32	0.21	0.24	0.27	0.19	0.11	0.21	0.06	0.13	0.22	0.19	0.18	0.26	0.21	0.20	0.19	0.26	0.19	0.22	0.29
33	0.23	0.12	0.17	0.19	0.18	0.19	0.19	0.21	0.18	0.19	0.15	0.20	0.21	0.23	0.25	0.17	0.10	0.16	0.17	0.17	0.18	0.20	0.13	0.22	0.27	0.12	0.24	0.26	0.22	0.22	0.20	0.21	0.22	0.20	0.23	0.12	0.26	0.11	0.22	0.18	0.05	0.26	0.22	0.17	0.20	0.18	0.26	0.19	0.18	0.22	0.30	0.24	0.29
32	0.19	0.10	0.16	0.23	0.17	0.17	0.14	0.17	0.19	0.31	0.14	0.21	0.16	0.17	0.18	0.22	0.16	0.20	0.14	0.15	0.19	0.16	0.14	0.21	0.13	0.11	0.25	0.21	0.15	0.16	0.19	0.18	0.17	0.22	0.20	0.14	0.28	0.22	0.17	0.19	0.05	0.08	0.22	0.15	0.18	0.18	0.14	0.20	0.19	0.18	0.20	0.17	0.20
31	0.18	0.15	0.09	0.24	0.15	0.20	0.12	0.19	0.13	0.17	0.15	0.19	0.20	0.18	0.17	0.19	0.12	0.11	0.20	0.15	0.14	0.18	0.12	0.16	0.09	0.18	0.16	0.16	0.18	0.16	0.22	0.18	0.21	0.16	0.21	0.11	0.16	0.22	0.21	0.16	0.06	0.10	0.15	0.10	0.12	0.07	0.13	0.20	0.13	0.17	0.16	0.15	0.11
30	0.07	0.08	0.14	0.15	0.15	0.16	0.09	0.12	0.13	0.12	0.16	0.15	0.11	0.14	0.11	0.13	0.13	0.16	0.13	0.09	0.11	0.11	0.06	0.14	0.11	0.12	0.13	0.11	0.11	0.10	0.14	0.17	0.12	0.07	0.08	0.14	0.10	0.19	0.09	0.10	0.04	0.13	0.10	0.13	0.15	0.13	0.10	0.16	0.13	0.16	0.14	0.12	0.16
29	0.08	0.05	0.06	0.08	0.09	0.10	0.08	0.09	0.05	0.07	0.13	0.13	0.09	0.14	0.10	0.10	0.09	0.11	0.06	0.05	0.12	0.06	0.08	0.11	0.10	0.07	0.14	0.10	0.11	0.10	0.07	0.04	0.06	0.14	0.09	0.08	0.08	0.11	0.11	0.05	0.04	0.10	0.04	0.10	0.05	0.10	0.06	0.14	0.11	0.04	0.07	0.12	0.14
28	0.11	0.10	0.09	0.12	0.08	0.09	0.05	0.05	0.01	0.05	0.06	0.05	0.04	0.10	0.11	0.06	0.06	0.07	0.02	0.09	0.07	0.03	0.07	0.09	0.05	0.07	0.04	0.08	0.09	0.09	0.06	0.04	0.04	0.04	0.09	0.06	0.08	0.10	0.08	0.05	0.02	0.06	0.09	0.06	0.09	0.07	0.05	0.06	0.07	0.09	0.06	0.06	0.10
27	0.04	0.07	0.06	0.10	0.06	0.09	0.05	0.06	0.04	0.09	0.05	0.09	0.04	0.08	0.05	0.06	0.06	0.05	0.09	0.03	0.04	0.05	0.07	0.09	0.03	0.04	0.03	0.09	0.05	0.05	0.04	0.03	0.07	0.03	0.04	0.03	0.04	0.07	0.06	0.05	0.01	0.04	0.08	0.03	0.11	0.05	0.07	0.06	0.04	0.06	0.07	0.06	0.08
26	0.02	0.04	0.03	0.06	0.06	0.07	0.07	0.06	0.03	0.04	0.02	0.07	0.06	0.09	0.04	0.03	0.03	0.10	0.04	0.02	0.05	0.02	0.02	0.03	0.08	0.01	0.05	0.04	0.03	0.05	0.04	0.06	0.03	0.03	0.03	0.04	0.03	0.05	0.05	0.05	0.02	0.01	0.05	0.04	0.03	0.06	0.06	0.08	0.04	0.03	0.07	0.05	0.03
25	0.02	0.01	0.03	0.02	0.02	0.03	0.01	0.01	0.02	0.03	0.06	0.04	0.04	0.01	0.06	0.01	0.03	0.04	0.01	0.01	0.00	0.04	0.04	0.04	0.03	0.03	0.01	0.03	0.05	0.02	0.04	0.03	0.02	0.03	0.01	0.02	0.03	0.04	0.02	0.02	0.01	0.02	0.02	0.03	0.02	0.04	0.01	0.02	0.02	0.05	0.01	0.01	0.05
24	0.01	0.00	0.03	0.02	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.02	0.03	0.02	0.00	0.02	0.01	0.01	0.02	0.01	0.03	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.01	0.02	0.00	0.05	0.02	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.03	0.02	0.00	0.02	0.01	0.01	0.02	0.00	0.02	0.02
23	0.01	0.00	0.00	0.02	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.00	0.02	0.01	0.01	0.01	0.03	0.01	0.02	0.01	0.01	0.02	0.02	0.01	0.02	0.01	0.01	0.00	0.00	0.02	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.04	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.01	0.02	0.00	0.01	0.00	0.00	0.00
22	0.01	0.01	0.00	0.01	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.02	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.01
21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
20	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Supplemental Table 4. Histogram of OG counts (alignment vs. de novo) with OG representation (0 -> 49)

(Omitted in this draft.)

Supplemental Table 5a. Metadata Contexts and Linear Model Residuals. MLSW = Residual of independent variable on dependent variable using raw mass in grams. MLSLW = Equivalent residual using the natural log of mass in grams.

	Independent Variable			Dependent Variable		Linear Residual							
	BMG		AW	MLS		MLSW				MLSLW			
	O	A	A	O	A	BMG (O)	BMG (A)	AW (A)	Std. Dev.	BMG (O)	BMG (A)	AW (A)	Std. Dev.
American Crow	448	448	384.8	7.1	8	-128.36	-192.86	-246.02	48.1061	1.50659	1.40148	1.23811	0.11047
Ruby Throated Hummingbird	3.1	3.2	3.1	7.7	8.2	286.102	415.677	267.268	65.971	-2.1806	-2.2325	-2.2735	0.03801
House Finch	21.4	20.4	21.4	7.9	9	1.80803	191.287	80.2963	77.7329	-0.7015	-0.6801	-0.6419	0.02465
Red-tailed Hawk	1126	1126	1362	8.6	9.1	-1193.3	-548.86	-147.38	430.818	-0.1804	1.03946	1.21641	0.62096
Coopers Hawk	439	439	526.64	9	9.2	-621.51	-230.85	-128.81	212.339	0.76169	1.34519	1.51585	0.3229
Great Horned Owl	1309	1450	1191.2	9.2	9.8	-683.5	-60.579	-178.6	270.162	0.45933	1.49629	1.28669	0.44768
Common Grackle	114	114	111	9.8	10	-1285.4	-826.43	-774.36	229.63	-1.0938	-0.339	-0.3776	0.3471
Northern Cardinal	44.7	41	42.6	10	10.2	-471.15	-1421.3	-1286.1	419.682	-0.7077	-2.0095	-1.9841	0.60779
Yellow Warbler	9.5	9.5	9.8	10.6	10.9	74.6342	238.369	117.962	69.2692	-1.3868	-1.3723	-1.3508	0.01478
Brown Headed Cowbird	43.9	42.5	38.1	10.9	11	-605.09	-298.78	-338.18	136.061	-0.925	-0.5819	-0.7019	0.14215
Yellow Rumped Warbler	12.6	11.5	12	10.9	11	198.772	337.005	202.27	64.3548	-0.9232	-1.0613	-1.0281	0.05884
European House Sparrow	27.7	27.7	25.3	11	11.3	-972.3	-903.06	-851.84	49.3567	-1.9108	-1.7418	-1.8443	0.06952
Rock Dove	350	350	358.7	11.1	11.6	463.549	-1740.4	-1503.8	987.907	2.2923	-0.6449	-0.6345	1.38216
Mourning Dove	119	123	119	11.3	11.8	-457.36	-1609.8	-1439.6	507.943	0.18092	-1.2468	-1.2933	0.68424
Canada Goose	1952	1952	3200	11.3	11.9	-682	-814.85	762.788	714.453	-0.1012	0.23401	0.71284	0.33404
Mute Swan	10735	10735	8300	11.7	12	7290.05	5262.35	3563.74	1523.24	0.38975	-1.4204	-1.6984	0.92584
Double Crested Cormorant	1674	1330	1817	11.9	12.1	347.204	447.555	980.91	278.113	1.70162	2.1897	2.48993	0.32486
Ruffed Grouse	577	644	532	12	13	932.624	872.869	640.162	126.165	3.15451	2.84409	2.64344	0.21024
European Starling	82.3	75	74	12	13.3	-1292.9	-846.1	-794.93	223.666	-1.3834	-0.7337	-0.759	0.30049
Gray Catbird	36.9	36.9	34.5	12.1	15.9	-733.12	-401.02	-423.89	151.453	-1.2798	-0.8432	-0.9213	0.19012
American Robin	77.3	77.3	75.5	12.3	16	-208.57	-273.65	-308.99	41.5897	0.18426	0.00428	-0.03	0.09397
Turkey	5811	5811	6050	12.4	16.9	5391.98	5846.6	5993.94	256.198	4.3048	4.80396	4.83429	0.24277
Downy Woodpecker	27	21.7	25.6	13.9	17	-16.799	163.596	59.8636	73.9221	-0.5053	-0.6543	-0.4988	0.07182
Hairy Woodpecker	66.3	66.3	62	15	17.9	-461.65	-178.35	-232.17	122.844	-0.3316	-0.0173	-0.0948	0.13368
Red-bellied Woodpecker	61.7	61.7	72.5	15.8	20	-30.514	-646.8	-615.79	283.494	0.24869	-0.665	-0.5151	0.4001
Pintail Duck	1011	1011	721	15.9	20.3	-920.98	-344.96	-517.42	241.386	0.29159	1.32762	0.97686	0.43024
Mallard	1082	1020	1048.1	16.3	20.7	-874.19	-500.24	-329.91	227.326	0.32323	1.13254	1.14669	0.38489
Gadwall	920	920	791	16.3	20.9	-382.59	-493.94	-496.69	53.1519	1.13926	1.16132	0.99743	0.07262
Green Winged Teal	341	250	343.8	16.9	22.5	-864.76	-1077	-869.99	98.8275	0.29169	-0.0336	0.27233	0.149
Northern Shoveler	613	554	613	17.9	22.5	-738	-338.11	-231.3	218.091	0.66079	1.30193	1.39134	0.32537
Wood Duck	658	448	452.8	19.8	22.6	-668.8	-434.45	-383.29	124.298	0.76786	1.10156	1.10044	0.15705
Sandhill Crane	4513	4513	3890	20.3	22.9	1479.58	2267.99	1896.18	322.041	0.13912	1.71994	1.55694	0.70991
Red Eyed Vireo	16.7	16.7	17	21.5	23	-39.203	322.877	190.849	149.613	-1.0038	-0.7122	-0.7039	0.13948
Killdeer	97	97	88	22.3	23.1	174.238	335.532	204.372	70.0213	0.95473	0.9631	0.85615	0.04857
Ring Billed Gull	518.5	439	518.5	22.5	27.1	-1413.5	-1342.2	-1081.2	142.827	-0.3762	-0.0344	0.11847	0.20678
Herring Gull	1135	1000	1094	22.5	27.4	-1680.6	-2443.3	-1918	318.697	-0.9151	-1.2746	-1.2016	0.15514
Caspian Tern	656	656	644	22.7	28	-1578.6	-951.21	-807.91	334.673	-0.5938	0.58317	0.55151	0.54753
Spotted Sandpiper	42.5	42.5	34	22.9	28.5	-13.403	174.733	60.0527	77.4183	-0.0697	0.00593	-0.227	0.09702
American Woodcock	197.5	156.7	197.5	23.1	29	226.323	-571.13	-507.22	361.8	1.5933	0.24304	0.46298	0.59153
Tufted Titmouse	21.6	21.6	21	27.5	29.1	-82.718	28.2059	-59.689	47.7958	-0.819	-0.8268	-0.865	0.02011
White Breasted Nuthatch	21.1	21.1	20.5	27.5	30	231.479	365.932	227.192	64.4158	-0.3714	-0.4304	-0.4686	0.03996
Barn Swallow	16	18	18.3	27.7	30.7	69.0305	-236.31	-284.08	156.42	-0.8836	-1.3331	-1.3271	0.21048
Tree Swallow	20.1	20.1	19	28	31.3	-47.907	142.669	36.8419	77.9608	-0.8366	-0.7549	-0.8209	0.03544
Carolina Wren	18.7	18.7	17.5	30	31.8	483.258	421.513	273.457	88.0335	-0.1118	-0.4791	-0.5547	0.19347
House wren	10.9	10.9	9.7	30.7	35	318.109	433.041	282.079	64.3745	-0.887	-0.9949	-1.1208	0.09552
Horned Lark	31.3	26	33.5	33.3	36.6	471.65	544.777	387.988	64.0568	0.36709	-0.0056	0.23881	0.1546
Cedar waxwing	32	32	30	34.8	42	569.18	531.449	368.066	87.2827	0.53413	0.17803	0.10443	0.18764
Song Sparrow	20.8	19.1	22.7	36.6	49	49.623	218.978	106.229	70.3938	-0.6575	-0.7099	-0.5469	0.06795
Chipping Sparrow	12.3	11.9	12.2	40	70	89.538	163.46	54.6745	45.3556	-1.1104	-1.243	-1.2279	0.05929

Supplemental Table 5b. Metadata Contexts and Caper Phylogeny Contrasts. MLSW = Residual of independent variable on dependent variable using raw mass in grams. MLSLW = Equivalent residual using the natural log of mass in grams.

	MLSLW				MLSW			
	BMG		AW	Std, Dev	BMG		AW	Std, Dev
Node	O	A	A		O	A	A	
AY	0.696216	0.391707	0.449764	0.132008	0.143320	-0.261353	-0.073953	0.165357
AX	-0.292371	-0.593349	-0.613852	0.146954	0.222999	-0.050793	0.068113	0.112096
AW	4.625498	2.959913	-1.596262	2.629812	0.597961	-1.211682	-0.487078	0.743652
AV	0.856952	1.428176	-0.208675	0.678326	-0.534295	-0.331060	-0.159369	0.153243
AT	0.414617	0.454868	-1.974793	1.135985	-0.074383	-0.057803	-0.054099	0.008820
AU	-0.410716	-0.051436	0.011897	0.186099	-0.308548	-0.163431	-0.016015	0.119427
AS	1.837134	0.943825	0.716865	0.483565	0.851552	-0.006329	0.358180	0.351543
AR	0.617445	1.145357	-0.092638	0.507230	-0.198825	0.072637	0.091111	0.132537
AQ	0.095351	0.057975	-0.630648	0.333778	-0.258526	-0.391349	-0.371062	0.058422
AP	-0.301419	-1.396420	-3.080070	1.142833	-0.667152	-0.462599	-0.454171	0.098474
AO	0.239351	0.307686	-0.114260	0.184917	-0.132048	-0.102551	-0.097612	0.015203
AN	0.064297	1.199443	1.439575	0.599779	-1.519437	-1.130998	-1.080636	0.196063
AM	0.359222	-0.279870	0.556637	0.357018	0.531729	-0.048689	0.124660	0.243274
AL	1.091643	0.640965	-0.172959	0.523323	0.534835	0.176512	0.319237	0.147290
AK	-0.151461	0.480033	-0.053907	0.277567	0.391818	0.829108	1.040364	0.270075
AJ	-1.006339	0.328000	1.100666	0.870309	-0.387877	-0.067378	-0.033968	0.159544
AI	-0.820623	0.169217	-1.428719	0.658530	-2.554283	2.271657	1.116575	2.057482
AH	-1.186759	-3.312595	0.255959	1.465727	5.858837	4.632178	3.012065	1.165884
AG	-0.585751	-0.694011	-1.119650	0.230443	-0.149730	-0.303195	-0.184246	0.065737
AF	-1.777695	-1.336231	-0.188739	0.669694	-0.763589	-0.587568	-0.569844	0.087455
AE	-1.616557	-1.701540	3.360503	2.366494	-2.972969	-3.943577	-4.730551	0.718834
AD	-1.256273	-1.068833	-0.015200	0.546254	-0.921427	-0.953403	-1.176473	0.113447
AC	-0.801399	-0.952589	0.683074	0.738009	-0.997792	-1.305704	-1.633655	0.259633
AB	2.831414	3.388990	0.182290	1.398875	2.205325	2.744232	3.749619	0.639971
AA	0.312376	0.700064	0.338199	0.176986	-0.075181	0.157721	0.130901	0.104047
T	0.487343	0.628287	1.809114	0.592667	-0.173501	-0.189597	-0.133265	0.023691
S	0.188746	1.148671	-0.079174	0.527135	-0.491746	0.073302	0.118034	0.277511
R	0.016291	0.448969	1.321815	0.542982	0.629547	1.033167	1.140074	0.219843
Q	-0.158606	-0.336982	0.138005	0.195905	-0.142100	-0.171059	-0.029926	0.060864
P	-0.302714	0.517707	-0.191086	0.363308	-0.271081	0.188120	0.301942	0.247695
O	0.883563	1.298231	-0.778922	0.897550	0.342649	0.490308	0.391877	0.061387
M	-0.155090	0.218078	-1.303691	0.647587	0.601311	0.876789	0.895752	0.134554
L	-0.430073	0.681986	-0.102043	0.466545	0.952256	1.632358	1.792193	0.364170
Z	0.564285	0.498688	0.834360	0.145266	-0.307026	-0.505616	-0.428390	0.081738
Y	-0.443200	-0.063572	1.361830	0.777033	-0.612013	-0.588380	-0.475839	0.059411
V	-0.726665	-0.519487	0.671252	0.615986	-0.848391	-1.018202	-0.859521	0.077559
U	-0.314009	-0.589462	0.058185	0.265382	-0.232173	-0.671757	-0.554671	0.185876
X	-0.744876	0.416681	-0.732507	0.544671	-0.991241	-0.425237	-0.352542	0.285498
W	-0.838217	0.208024	0.151180	0.480365	-0.096229	-0.350882	-0.298391	0.109784
J	0.010192	-0.063598	0.080924	0.059005	-0.025004	-0.014060	-0.013403	0.005320
H	-0.280016	-0.217624	-0.808313	0.264975	-0.188260	-0.200372	-0.192280	0.005037
G	-0.342287	-0.263976	0.564319	0.410168	-0.202921	-0.267922	-0.258105	0.028610
F	0.033182	-0.330124	0.450688	0.319021	-0.069972	-0.223066	-0.213526	0.070029
E	-0.270945	-0.121100	0.153854	0.175912	-0.168563	-0.055916	-0.051223	0.054242
D	0.583679	0.514862	-0.850932	0.660660	-0.079299	-0.109901	-0.098961	0.012661
C	-1.144205	-0.830109	-0.219315	0.384006	-0.327649	-0.312256	-0.309757	0.007911
B	-2.091727	-1.628563	1.648800	1.664905	-0.663042	-0.443557	-0.425319	0.108022
A	0.535895	-0.631730	0.512193	0.544923	-0.045598	-0.059298	-0.062805	0.007424

Supplemental Table 6: Leveraging MDMR to validate Metadata Contexts for Highlight in Manuscript. Mass: “Original” metadata by Rich M. and Steve S, imputed with updates available from AnAge is vest fit in both distance contexts. Lifespan: “AnAge” MLS is best fit to distances oserved in both contexts, but less well-fit when used to calculate residuals. Residuals: Performance is most consistent using the phylogenetic distance matrix. The lowest residual uses the “original” metadata updated with AnAge information wherever defined. It was more highly correlated to the phylogenetic distance matrix. Significant: * = < 0.01, ** = < 0.001, *** = < 0.0001.

Metric	MASH			Phylogeny		
	Stat	R2	P	Stat	R2	P
Phylogenetic Correlations to Mass, in each Context						
BMG (O)	0.0844	0.0778	0.002**	0.0926	0.0848	0.002**
log(BMG (O))	0.267	0.211	<0.002***	0.245	0.197	<0.002***
BMG (A)	0.0816	0.0754	0.002**	0.0896	0.0822	0.002**
log(BMG (A))	0.262	0.208	<0.002***	0.239	0.193	<0.002***
AW (A)	0.0998	0.0908	<0.002***	0.107	0.0967	<0.002***
log(AW (A))	0.265	0.21	<0.002***	0.241	0.194	<0.002***
Phylogenetic Correlations to Lifespan						
MLS (O)	0.135	0.119	<0.002***	0.127	0.113	<0.002***
MLS (A)	0.0998	0.0908	<0.002***	0.107	0.0967	<0.002***
Phylogenetic Correlations to Residuals, in each Context						
res(BMG ~ MLS) (O)	0.0298	0.0289	0.194	0.0341	0.033	0.098
res(log(BMG) ~ MLS) (O)	0.142	0.124	<0.002***	0.131	0.116	<0.002***
res(BMG ~ MLS) (A)	0.0293	0.0285	0.2	0.0333	0.0322	0.11
res(log(BMG) ~ MLS) (A)	0.139	0.122	<0.002***	0.127	0.112	<0.002***
res(AW (A) ~ MLS) (A)	0.129	0.114	<0.002***	0.129	0.114	<0.002***
res(log(AW (A)) ~ MLS) (A)	0.16	0.138	<0.002***	0.129	0.114	<0.002***

Supplemental Table 7. Variance explained by each principal component in each distance matrix context.

	Variance Explained (MASH)		Variance Explained (Phylo)	
PC	%	Sum	%	Sum
PC1	26.83	26.83	25.13	25.13
PC2	7.99	34.82	8.09	33.22
PC3	2.74	37.55	5.11	38.33
PC4	1.86	39.41	2.32	40.65
PC5	1.51	40.92	1.27	41.92
PC6	1.4	42.32	0.92	42.84
PC7	1.07	43.39	0.87	43.71
PC8	0.89	44.28	0.72	44.43
PC9	0.81	45.09	0.56	44.99

Supplemental Table 8: Principle Components provide additional insight for simple linear model comparisons: Sample Distance Correlation to Transcriptomic Distance. Use the principal components as additional independent variables/covariates in a regression model with MLS as the dependent variable and body size as an independent variable. See if the regression coefficient is significant for body size along with the principal components.

		MASH DISTANCE					PHYLO DISTANCE				
		LM : Whole	LM > Anova : p				LM : Whole	LM > Anova : p			
Variables tested:	Metadata	model p	MLS	PC1	PC2	PC3	model p	MLS	PC1	PC2	PC3
MLS ~ PC1 + PC2	BMG (O)	0.000373813	NA	0.0001651	0.1641677	NA	4.49E-09	NA	1.38E-09	0.08437	NA
	BMG (A)	0.00012516	NA	5.81E-05	0.1301	NA	7.71E-09	NA	2.39E-09	0.08756	NA
	AW (A)	1.30E-08	NA	2.04E-09	0.8403	NA	6.00E-09	NA	2.01E-09	0.07025	NA
MLS ~ log(BMG) + PC1	BMG (O)	4.59E-05	0.0001686	0.0061606	NA	NA	2.59E-15	1.54E-14	1.74E-06	NA	NA
	BMG (A)	2.38E-05	0.0001762	0.0023029	NA	NA	2.37E-13	3.10E-12	2.37E-06	NA	NA
	AW (A)	1.25E-13	6.47E-12	1.76E-07	NA	NA	3.01E-12	4.89E-11	4.61E-06	NA	NA
MLS ~ log(BMG) + PC1 + PC2	BMG (O)	2.80E-05	0.0001119	0.00479	0.0476216	NA	8.56E-15	1.32E-14	1.43E-06	0.1385	NA
	BMG (A)	1.20E-05	0.0001093	0.0016265	0.0363013	NA	9.88E-13	3.24E-12	2.27E-06	0.2242	NA
	AW (A)	4.66E-13	6.15E-12	1.59E-07	0.1867	NA	1.28E-12	1.20E-11	1.80E-06	0.01538	NA
MLS ~ log(BMG) + PC1 + PC2 + PC3	BMG (O)	1.75E-07	1.48E-05	0.0013861	0.0237509	0.0002859	6.66E-14	2.54E-14	1.85E-06	0.1428	0.8164
	BMG (A)	1.21E-06	3.91E-05	0.0007682	0.0246713	0.0061258	6.28E-12	5.32E-12	2.84E-06	0.2285	0.6713
	AW (A)	8.21E-14	8.26E-13	3.44E-08	0.155847	0.006466	3.48E-12	1.09E-11	1.58E-06	0.0145	0.1709

ST9a. Tabular version of phylogenetic tree distance matrix

	American_Crow	Ruby_throated_Hummingbird	House_Finch	Red_tailed_Hawk	Coopers_Hawk	Great_Horned_Owl	Common_Grackle	Northern_Cardinal	Yellow_Warbler	Brown_headed_Cowbird	Yellow_rumped_Warbler	House_Sparrow	Rock_Dove	Mourning_Dove	Canada_Goose	Mute_Swan	Double_crested_Cormorant	Ruffed_Grouse	Starling	Gray_Catbird	American_Robin	Turkey	Downy_Woodpecker	Hairy_Woodpecker	Red_bellied_Woodpecker	Pintail_Duck	Mallard	Gadwall	Green_winged_Teal	Northern_Shoveler	Wood_Duck	Sandhill_Crane	Red_eyed_Vireo	Killdeer	Ring_billed_Gull	Herring_Gull	Caspian_Tern	Spotted_Sandpiper	Woodcock	Tufted_Titmouse	White_breasted_Nuthatch	Barn_Swallow	Tree_Swallow	Carolina_Wren	House_wren	Horned_Lark	Cedar_waxwing	Song_Sparrow	Chipping_Sparrow
American_Crow	0	159.2	100.2	156.8	156.8	156.8	100.2	100.2	100.2	100.2	100.1	100.2	158	158	195.2	195.2	158	195.2	100.2	100.2	100.2	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	75.6	158.8	158.8	158.8	158.8	158.8	158.8	100.2	100.2	100.2	100.2	100.2	100.2	100.2	100.2	100.2	100.2
Ruby_throated_Hummingbird	159.2	0	159.2	159.2	159.2	159.2	159.2	159.2	159.2	159.2	159.1	159.2	158.2	158.2	195.2	195.2	158.2	195.2	159.2	159.2	159.2	195.2	159.2	159.1	159.1	196.2	195.2	195.2	196.2	195.17	195.2	158.2	159.2	159.2	159.2	159.2	159.2	159.2	159.2	159.2	159.2	159.2	159.2	159.2	159.2	159.2	159.2	159.2	159.2
House_Finch	100.2	159.2	0	156.8	156.8	156.8	52	52	52	52	51.9	58.6	158	158	195.2	195.2	158	195.2	90.2	90.2	90.2	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	91.2	90.2	91.2	91.2	90.2	90.2	91.2	90.2	52	52
Red_tailed_Hawk	156.8	159.2	156.8	0	48.4	155.4	156.8	156.8	156.8	156.8	156.7	156.8	158	158	195.2	195.2	158	195.2	156.8	156.8	156.8	195.2	151.8	151.7	151.7	196.2	195.2	195.2	196.2	195.17	195.2	158	156.8	158.8	158.8	158.8	158.8	158.8	158.8	156.8	156.8	156.8	156.8	156.8	156.8	156.8	156.8	156.8	156.8
Coopers_Hawk	156.8	159.2	156.8	48.4	0	155.4	156.8	156.8	156.8	156.8	156.7	156.8	158	158	195.2	195.2	158	195.2	156.8	156.8	156.8	195.2	151.8	151.7	151.7	196.2	195.2	195.2	196.2	195.17	195.2	158	156.8	158.8	158.8	158.8	158.8	158.8	158.8	156.8	156.8	156.8	156.8	156.8	156.8	156.8	156.8	156.8	156.8
Great_Horned_Owl	156.8	159.2	156.8	155.4	155.4	0	156.8	156.8	156.8	156.8	156.7	156.8	158	158	195.2	195.2	158	195.2	156.8	156.8	156.8	195.2	155.4	155.3	155.3	196.2	195.2	195.2	196.2	195.17	195.2	158	156.8	158.8	158.8	158.8	158.8	158.8	158.8	156.8	156.8	156.8	156.8	156.8	156.8	156.8	156.8	156.8	156.8
Common_Grackle	100.2	159.2	52	156.8	156.8	156.8	0	45.8	45.8	45.8	45.7	58.6	158	158	195.2	195.2	158	195.2	90.2	90.2	90.2	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	91.2	90.2	91.2	91.2	90.2	90.2	91.2	90.2	24	24
Northern_Cardinal	100.2	159.2	52	156.8	156.8	156.8	45.8	0	43	43	42.9	58.6	158	158	195.2	195.2	158	195.2	90.2	90.2	90.2	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	91.2	90.2	91.2	91.2	90.2	90.2	91.2	90.2	45.8	45.8
Yellow_Warbler	100.2	159.2	52	156.8	156.8	156.8	45.8	43	0	21	10.4	58.6	158	158	195.2	195.2	158	195.2	90.2	90.2	90.2	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	91.2	90.2	91.2	91.2	90.2	90.2	91.2	90.2	45.8	45.8
Brown_headed_Cowbird	100.2	159.2	52	156.8	156.8	156.8	45.8	43	21	0	20.9	58.6	158	158	195.2	195.2	158	195.2	90.2	90.2	90.2	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	91.2	90.2	91.2	91.2	90.2	90.2	91.2	90.2	45.8	45.8
Yellow_rumped_Warbler	100.1	159.1	51.9	156.7	156.7	156.7	45.7	42.9	10.4	20.9	0	58.5	157.9	157.9	195.1	195.1	157.9	195.1	90.1	90.1	90.1	195.1	156.7	156.6	156.6	196.1	195.1	195.1	196.1	195.07	195.1	157.9	100.1	158.7	158.7	158.7	158.7	158.7	158.7	91.1	90.1	91.1	91.1	90.1	90.1	91.1	90.1	45.7	45.7
House_Sparrow	100.2	159.2	58.6	156.8	156.8	156.8	58.6	58.6	58.6	58.6	58.5	0	158	158	195.2	195.2	158	195.2	90.2	90.2	90.2	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	91.2	90.2	91.2	91.2	90.2	90.2	91.2	90.2	58.6	58.6
Rock_Dove	158	158.2	158	158	158	158	158	158	158	158	157.9	158	0	52.6	194.2	194.2	149	194.2	158	158	158	194.2	158	157.9	157.9	195.2	194.2	194.2	195.2	194.17	194.2	149	158	158	158	158	158	158	158	158	158	158	158	158	158	158	158	158	158
Mourning_Dove	158	158.2	158	158	158	158	158	158	158	158	157.9	158	52.6	0	194.2	194.2	149	194.2	158	158	158	194.2	158	157.9	157.9	195.2	194.2	194.2	195.2	194.17	194.2	149	158	158	158	158	158	158	158	158	158	158	158	158	158	158	158	158	158
Canada_Goose	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.1	195.2	194.2	194.2	0	27	194.2	144.8	195.2	195.2	195.2	144.8	195.2	195.1	195.1	49.2	48.2	48.2	49.2	48.17	48.2	194.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2
Mute_Swan	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.1	195.2	194.2	194.2	27	0	194.2	144.8	195.2	195.2	195.2	144.8	195.2	195.1	195.1	49.2	48.2	48.2	49.2	48.17	48.2	194.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2
Double_crested_Cormorant	158	158.2	158	158	158	158	158	158	158	158	157.9	158	149	149	194.2	194.2	0	194.2	158	158	158	194.2	158	157.9	157.9	195.2	194.2	194.2	195.2	194.17	194.2	143	158	158	158	158	158	158	158	158	158	158	158	158	158	158	158	158	158
Ruffed_Grouse	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.1	195.2	194.2	194.2	144.8	144.8	194.2	0	195.2	195.2	195.2	26.6	195.2	195.1	195.1	145.8	144.8	144.8	145.8	144.77	144.8	194.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2
Starling	100.2	159.2	90.2	156.8	156.8	156.8	90.2	90.2	90.2	90.2	90.1	90.2	158	158	195.2	195.2	158	195.2	0	43.8	56.2	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	91.2	77.6	91.2	91.2	77.6	77.6	91.2	89.2	90.2	90.2
Gray_Catbird	100.2	159.2	90.2	156.8	156.8	156.8	90.2	90.2	90.2	90.2	90.1	90.2	158	158	195.2	195.2	158	195.2	43.8	0	56.2	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	91.2	77.6	91.2	91.2	77.6	77.6	91.2	89.2	90.2	90.2
American_Robin	100.2	159.2	90.2	156.8	156.8	156.8	90.2	90.2	90.2	90.2	90.1	90.2	158	158	195.2	195.2	158	195.2	56.2	56.2	0	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	91.2	77.6	91.2	91.2	77.6	77.6	91.2	89.2	90.2	90.2
Turkey	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.1	195.2	194.2	194.2	144.8	144.8	194.2	26.6	195.2	195.2	195.2	0	195.2	195.1	195.1	145.8	144.8	144.8	145.8	144.77	144.8	194.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2
Downy_Woodpecker	156.8	159.2	156.8	151.8	151.8	155.4	156.8	156.8	156.8	156.8	156.7	156.8	158	158	195.2	195.2	158	195.2	156.8	156.8	156.8	195.2	0	28.5	28.5	196.2	195.2	195.2	196.2	195.17	195.2	158	156.8	158.8	158.8	158.8	158.8	158.8	158.8	156.8	156.8	156.8	156.8	156.8	156.8	156.8	156.8	156.8	156.8
Hairy_Woodpecker	156.7	159.1	156.7	151.7	151.7	155.3	156.7	156.7	156.7	156.7	156.6	156.7	157.9	157.9	195.1	195.1	157.9	195.1	156.7	156.7	156.7	195.1	28.5	0	14.1	196.1	195.1	195.1	196.1	195.07	195.1	157.9	156.7	158.7	158.7	158.7	158.7	158.7	158.7	156.7	156.7	156.7	156.7	156.7	156.7	156.7	156.7	156.7	156.7
Red_bellied_Woodpecker	156.7	159.1	156.7	151.7	151.7	155.3	156.7	156.7	156.7	156.7	156.6	156.7	157.9	157.9	195.1	195.1	157.9	195.1	156.7	156.7	156.7	195.1	28.5	14.1	0	196.1	195.1	195.1	196.1	195.07	195.1	157.9	156.7	158.7	158.7	158.7	158.7	158.7	158.7	156.7	156.7	156.7	156.7	156.7	156.7	156.7	156.7	156.7	156.7
Pintail_Duck	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.1	196.2	195.2	195.2	49.2	49.2	195.2	145.8	196.2	196.2	196.2	145.8	196.2	196.1	196.1	0	9.48	12.54	7.98	15.71	25.6	195.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2
Mallard	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.1	195.2	194.2	194.2	48.2	48.2	194.2	144.8	195.2	195.2	195.2	144.8	195.2	195.1	195.1	9.48	0	11.54	9.48	14.71	24.6	194.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2
Gadwall	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.1	195.2	194.2	194.2	48.2	48.2	194.2	144.8	195.2	195.2	195.2	144.8	195.2	195.1	195.1	12.54	11.54	0	12.54	14.71	24.6	194.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2
Green_winged_Teal	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.1	196.2	195.2	195.2	49.2	49.2	195.2	145.8	196.2	196.2	196.2	145.8	196.2	196.1	196.1	7.98	9.48	12.54	0	15.71	25.6	195.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2	196.2
Northern_Shoveler	195.17	195.17	195.17	195.17	195.17	195.17	195.17	195.17	195.17	195.17	195.07	195.17	194.17	194.17	48.17	48.17	194.17	144.77	195.17	195.17	195.17	144.77	195.17	195.07	195.07	15.71	14.71	14.71	15.71	0	24.57	194.17	195.17	195.17	195.17	195.17	195.17	195.17	195.17	195.17	195.17	195.17	195.17	195.17	195.17	195.17	195.17	195.17	195.17
Wood_Duck	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.1	195.2	194.2	194.2	48.2	48.2	194.2	144.8	195.2	195.2	195.2	144.8	195.2	195.1	195.1	25.6	24.6	24.6	25.6	24.57	0	194.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2	195.2
Sandhill_Crane	158	158.2	158	158	158	158	158	158	158	158	157.9	158	149	149	194.2	194.2	143	194.2	158	158	158	194.2	158	157.9	157.9	195.2	194.2	194.2	195.2	194.17	194.2	0	158	158	158	158	158	158	158	158	158	158	158	158	158	158	158	158	158
Red_eyed_Vireo	75.6	159.2	100.2	156.8	156.8	156.8	100.2	100.2	100.2	100.2	100.1	100.2	158	158	195.2	195.2	158	195.2	100.2	100.2	100.2	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	0	158.8	158.8	158.8	158.8	158.8	158.8	100.2	100.2	100.2	100.2	100.2	100.2	100.2	100.2	100.2	100.2
Killdeer	158.8	159.2	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.7	158.8	158	158	195.2	195.2	158	195.2	158.8	158.8	158.8	195.2	158.8	158.7	158.7	196.2	195.2	195.2	196.2	195.17	195.2	158	158.8	0	135.8	135.8	135.8	135.8	135.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8
Ring_billed_Gull	158.8	159.2	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.7	158.8	158	158	195.2	195.2	158	195.2	158.8	158.8	158.8	195.2	158.8	158.7	158.7	196.2	195.2	195.2	196.2	195.17	195.2	158	158.8	135.8	0	3.16	45	126.8	126.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8
Herring_Gull	158.8	159.2	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.7	158.8	158	158	195.2	195.2	158	195.2	158.8	158.8	158.8	195.2	158.8	158.7	158.7	196.2	195.2	195.2	196.2	195.17	195.2	158	158.8	135.8	3.16	0	45	126.8	126.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8
Caspian_Tern	158.8	159.2	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.7	158.8	158	158	195.2	195.2	158	195.2	158.8	158.8	158.8	195.2	158.8	158.7	158.7	196.2	195.2	195.2	196.2	195.17	195.2	158	158.8	135.8	45	45	0	126.8	126.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8
Spotted_Sandpiper	158.8	159.2	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.7	158.8	158	158	195.2	195.2	158	195.2	158.8	158.8	158.8	195.2	158.8	158.7	158.7	196.2	195.2	195.2	196.2	195.17	195.2	158	158.8	135.8	126.8	126.8	126.8	0	15.2	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8
Woodcock	158.8	159.2	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.7	158.8	158	158	195.2	195.2	158	195.2	158.8	158.8	158.8	195.2	158.8	158.7	158.7	196.2	195.2	195.2	196.2	195.17	195.2	158	158.8	135.8	126.8	126.8	126.8	15.2	0	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8	158.8
Tufted_Titmouse	100.2	159.2	91.2	156.8	156.8	156.8	91.2	91.2	91.2	91.2	91.1	91.2	158	158	195.2	195.2	158	195.2	91.2	91.2	91.2	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	0	91.2	84.4	84.4	91.2	91.2	84.4	91.2	91.2	91.2
White_breasted_Nuthatch	100.2	159.2	90.2	156.8	156.8	156.8	90.2	90.2	90.2	90.2	90.1	90.2	158	158	195.2	195.2	158	195.2	77.6	77.6	77.6	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	91.2	0	91.2	91.2	64.4	64.4	91.2	89.2	90.2	90.2
Barn_Swallow	100.2	159.2	91.2	156.8	156.8	156.8	91.2	91.2	91.2	91.2	91.1	91.2	158	158	195.2	195.2	158	195.2	91.2	91.2	91.2	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	84.4	91.2	0	37.6	91.2	91.2	69.8	91.2	91.2	91.2
Tree_Swallow	100.2	159.2	91.2	156.8	156.8	156.8	91.2	91.2	91.2	91.2	91.1	91.2	158	158	195.2	195.2	158	195.2	91.2	91.2	91.2	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	84.4	91.2	37.6	0	91.2	91.2	69.8	91.2	91.2	91.2
Carolina_Wren	100.2	159.2	90.2	156.8	156.8	156.8	90.2	90.2	90.2	90.2	90.1	90.2	158	158	195.2	195.2	158	195.2	77.6	77.6	77.6	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	91.2	64.4	91.2	91.2	0	28.6	91.2	89.2	90.2	90.2
House_wren	100.2	159.2	90.2	156.8	156.8	156.8	90.2	90.2	90.2	90.2	90.1	90.2	158	158	195.2	195.2	158	195.2	77.6	77.6	77.6	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	91.2	64.4	91.2	91.2	28.6	0	91.2	89.2	90.2	90.2
Horned_Lark	100.2	159.2	91.2	156.8	156.8	156.8	91.2	91.2	91.2	91.2	91.1	91.2	158	158	195.2	195.2	158	195.2	91.2	91.2	91.2	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	84.4	91.2	69.8	69.8	91.2	91.2	0	91.2	91.2	91.2
Cedar_waxwing	100.2	159.2	90.2	156.8	156.8	156.8	90.2	90.2	90.2	90.2	90.1	90.2	158	158	195.2	195.2	158	195.2	89.2	89.2	89.2	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	91.2	89.2	91.2	91.2	89.2	89.2	91.2	0	90.2	90.2
Song_Sparrow	100.2	159.2	52	156.8	156.8	156.8	24	45.8	45.8	45.8	45.7	58.6	158	158	195.2	195.2	158	195.2	90.2	90.2	90.2	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	91.2	90.2	91.2	91.2	90.2	90.2	91.2	90.2	0	19.94
Chipping_Sparrow	100.2	159.2	52	156.8	156.8	156.8	24	45.8	45.8	45.8	45.7	58.6	158	158	195.2	195.2	158	195.2	90.2	90.2	90.2	195.2	156.8	156.7	156.7	196.2	195.2	195.2	196.2	195.17	195.2	158	100.2	158.8	158.8	158.8	158.8	158.8	158.8	91.2	90.2	91.2	91.2	90.2	90.2	91.2	90.2	19.94	0

ST9b. Tabular version of MASH tree distance matrix

	05.AmericanCrow	21.RubyThroatedHummingbird	37.HouseFinch	69.RedTailedHawk	67.CoopersHawk	73.GreatHornedOwl	01.CommonGrackle	36.NorthernCardinal	35.YellowWarbler	39.BrownHeadedCowbird	27.YellowRumpedWarbler	42.HouseSparrow	68.RockDove	34.MourningDove	08.CanadaGoose	61.MuteSwan	62.DoubleCrestedCormorant	74.RuffedGrouse	31.Starling	28.GrayCatbird	10.AmericanRobin	65.Turkey	09.DownyWoodpecker	72.HairyWoodpecker	24.RedBelliedWoodpecker	20.PintailDuck	23.Mallard	26.Gadwall	06.GreenWingedTeal	07.NorthernShoveler	02.WoodDuck	71.SandhillCrane	14.RedEyedVireo	13.Killdeer	40.RingBilledGull	66.HerringGull	70.CaspianTern	30.SpottedSandpiper	64.Woodcock	33.TuftedTitmouse	11.WhiteBreastedNuthatch	15.BarnSwallow	38.TreeSwallow	12.CarolinaWren	19.HouseWren	32.HornedLark	75.CedarWaxwing	17.SongSparrow	04.ChippingSparrow
05.AmericanCrow	0.000000	0.107802	0.065740	0.077263	0.078409	0.079467	0.054708	0.065394	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
21.RubyThroatedHummingbird	0.107802	0.000000	0.109889	0.089629	0.091597	0.094678	0.105576	0.111340	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
37.HouseFinch	0.065740	0.109889	0.000000	0.079618	0.078028	0.101880	0.041158	0.047055	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
69.RedTailedHawk	0.077263	0.089629	0.079618	0.000000	0.028243	0.097005	0.083080	0.097101	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
67.CoopersHawk	0.078409	0.091597	0.078028	0.028243	0.000000	0.096530	0.086153	0.107886	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
73.GreatHornedOwl	0.079467	0.094678	0.101880	0.097005	0.096530	0.000000	0.089741	0.102486	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
01.CommonGrackle	0.054708	0.105576	0.041158	0.083080	0.086153	0.089741	0.000000	0.028359	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
36.NorthernCardinal	0.065394	0.111340	0.047055	0.097101	0.107886	0.102486	0.028359	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
35.YellowWarbler	0.065943	0.108153	0.044671	[illegible]	[illegible]	0.103846	[illegible]	0.041327	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
39.BrownHeadedCowbird	0.065837	0.110793	0.047585	[illegible]	[illegible]	[illegible]	0.022945	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
27.YellowRumpedWarbler	0.066442	0.107822	[illegible]	[illegible]	[illegible]	0.102851	0.054293	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
42.HouseSparrow	[illegible]	0.109760	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
68.RockDove	[illegible]	0.103516	[illegible]	[illegible]	[illegible]	[illegible]	0.104495	0.107822	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
34.MourningDove	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
08.CanadaGoose	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
61.MuteSwan	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
62.DoubleCrestedCormorant	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
74.RuffedGrouse	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
31.Starling	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
28.GrayCatbird	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
10.AmericanRobin	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
65.Turkey	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
09.DownyWoodpecker	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
72.HairyWoodpecker	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
24.RedBelliedWoodpecker	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
20.PintailDuck	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
23.Mallard	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
26.Gadwall	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
06.GreenWingedTeal	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
07.NorthernShoveler	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
02.WoodDuck	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
71.SandhillCrane	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
14.RedEyedVireo	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
13.Killdeer	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
40.RingBilledGull	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
66.HerringGull	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
70.CaspianTern	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
30.SpottedSandpiper	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
64.Woodcock	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
33.TuftedTitmouse	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
11.WhiteBreastedNuthatch	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
15.BarnSwallow	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
38.TreeSwallow	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
12.CarolinaWren	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
19.HouseWren	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]	[illegible]
32.HornedLark	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]	[illegible]
75.CedarWaxwing	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]	[illegible]
17.SongSparrow	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000	[illegible]
04.ChippingSparrow	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.000000

Supplemental Table 10. Repeat Table 2 for Alternative Metadata Context : AW (A) (Alignment-sorted)

Most Significant Associations with Lifespan : Model = res(MLS ~ log(BMG)) ~ log(rel(gene_abs))										Alternative Model Performance (FDR-adjusted)								Alternative Model Performance (FDR-adjusted)							
Top OGs with Shared Huref Common Symbol								CCC rho.q (Alignment vs. De Novo Correlation)		LM		CAPER		MDMR (Mash)		MDMR (Phylo)		LM		CAPER		MDMR (Mash)		MDMR (Phylo)	
Rank	Directionality	Pvalue Cutoff	Gene Symbols	Ortholog Group	Slope of Assoc.	P-value (Raw)	P-value (FDR)	LM Residuals	Phylogeny Contrasts	Align	De novo	Align	De novo	Align	De novo	Align	De novo	Align	De novo	Align	De novo	Align	De novo	Align	De novo
Alignment-Derived Abundances																									
Metadata Context = AW (A)																		Metadata Context = BMG (A)							
1	Positive	0.05	PAIP2	EOG090F0C08	1.3623709	1.27407E-06	0.003886	0.92544	0.8322029	0.009514	0.000492	NA	0.015741	0.015906	0.159081	0.026981	0.203471	0.009794	0.000532	0.005292	0.009905	0.018536	0.146522	0.024231	0.204554
2			KLHL25	EOG090F03CE	0.6584221	8.80433E-05	0.017149	9.88E-01	2.60E-01	0.121253	0.153574	NA	0.276747	0.017319	0.01595	0.031813	0.026314	0.09648	0.131638	0.012095	0.249758	0.016722	0.016511	0.027257	0.023624
3			ICOSLG, LICOS	EOG090F07AA	0.3939026	9.86428E-05	0.018373	9.99E-01	6.94E-01	0.689745	0.815603	NA	0.019301	0.015906	0.01595	0.026135	0.026314	0.65656	0.769182	0.029544	0.025903	0.016722	0.016511	0.023909	0.023624
4			RASSF3	EOG090F099I	0.9227489	0.000165918	0.025932	9.43E-01	3.76E-01	0.002022	0.026424	NA	0.425608	0.543789	0.159081	0.440578	0.158019	0.001773	0.022977	0.010684	0.348687	0.53777	0.162325	0.431437	0.149319
5			CNT6L	EOG090F03WS	1.0807058	0.000218491	0.029339	0.928319	0.8035534	0.000944	0.006152	NA	0.167358	0.145377	0.023125	0.106966	0.028358	0.001045	0.005931	0.025244	0.125604	0.140548	0.022342	0.094386	0.027222
6			WDR5	EOG090F075P	0.9173344	0.000233736	0.029991	9.69E-01	1.53E-01	0.181114	0.277951	NA	0.541132	0.023235	0.025183	0.041234	0.033659	0.170155	0.307939	0.023259	0.566226	0.020486	0.020275	0.034913	0.031051
7			MALT1	EOG090F02FF	0.8800394	0.000256363	0.031495	9.45E-01	5.50E-01	0.026107	0.009733	NA	0.053098	0.051954	0.019162	0.071114	0.033659	0.023029	0.010875	0.022543	0.055937	0.051522	0.018393	0.060828	0.031051
8			None defined.	EOG090F09MR	0.5462945	0.0004174	0.035761	0.986819	0.5260895	0.919617	0.528718	NA	0.062731	0.015906	0.023125	0.026135	0.02706	0.912809	0.539747	0.03861	0.066854	0.016722	0.020275	0.023909	0.027222
9			RBL1, RBL2	EOG090F017T	0.9150194	0.000595868	0.040959	9.75E-01	2.47E-01	0.000897	0.003543	NA	0.325717	0.091998	0.079687	0.069177	0.058774	0.001195	0.004653	0.068836	0.425949	0.07539	0.071229	0.070553	0.050344
10			BAHD1	EOG090F022R	1.0245268	0.000609705	0.041417	0.966906	0.1559621	0.023074	0.014582	NA	0.509027	0.027281	0.04355	0.031813	0.062761	0.017122	0.012352	0.025244	0.447106	0.022522	0.043098	0.032947	0.056378
11			UGP2	EOG090F04H0	0.5966175	0.000654647	0.042954	0.998339	0.3675871	0.281849	0.198298	NA	0.053748	0.023235	0.033316	0.039324	0.047104	0.297229	0.205555	0.025739	0.037433	0.020486	0.028674	0.031028	0.040725
1	Negative	0.025	TMEM87A	EOG090F03QS	-1.9802229	2.01923E-06	0.003886	9.47E-01	8.20E-01	0.000607	0.000478	NA	0.00986	0.636551	0.332258	0.548852	0.27032	0.000493	0.000404	0.003083	0.003233	0.668231	0.339588	0.577701	0.2518
2			GCLC	EOG090F02SE	-0.9376417	1.90249E-06	0.003886	9.90E-01	9.53E-01	0.007393	0.011461	NA	0.003574	0.023235	0.017226	0.026135	0.026314	0.00937	0.015077	0.007507	0.004985	0.024512	0.018393	0.023909	0.023624
3			AFDN, MLLTT4	EOG090F00EM	-1.381762	8.54259E-06	0.00959	0.98725	0.8778568	0.000991	0.000934	NA	0.018887	0.015906	0.01595	0.026135	0.026314	0.001277	0.001352	0.006807	0.019382	0.016722	0.016511	0.023909	0.023624
4			MTA1, MTA3	EOG090F02S8	-0.8890178	9.96497E-06	0.00959	0.858234	0.6582843	0.000944	0.450547	NA	0.190444	0.26927	0.031317	0.406719	0.043118	0.001059	0.434967	0.007507	0.175725	0.240924	0.026614	0.417908	0.044635
5			CHST14	EOG090F07EP	-0.8070168	9.94359E-06	0.00959	0.994734	0.6185213	0.579044	0.45064	NA	0.214262	0.015906	0.01595	0.026135	0.026314	0.641044	0.467231	0.006807	0.17615	0.016722	0.016511	0.023909	0.023624
6			ARMC10	EOG090F0AQK	-0.8099887	1.74963E-05	0.011225	0.951162	6.06E-01	0.058336	0.00387	NA	0.003574	0.015906	0.031317	0.02842	0.052906	0.06711	0.004268	0.010292	0.001345	0.016722	0.026614	0.025684	0.048562
7			GMDS	EOG090F06BO	-0.5686311	1.70399E-05	0.011225	0.987097	0.6644349	0.413945	0.97385	NA	0.752456	0.015906	0.01595	0.026135	0.026314	0.513611	0.942479	0.020564	0.904536	0.016722	0.016511	0.023909	0.023624
8			KAZN	EOG090F07E1	-0.6816273	1.69285E-05	0.011225	0.90265	6.99E-01	0.000488	0.036946	NA	0.178642	0.073954	0.035369	0.069177	0.035583	0.000647	0.038267	0.006807	0.132233	0.065511	0.030802	0.05684	0.032971
9			MSRA	EOG090F0CAU	-0.5022559	2.22872E-05	0.012869	0.958972	0.2047721	0.005973	0.06957	NA	0.203446	0.085834	0.033316	0.10882	0.037511	0.007783	0.064358	0.020564	0.193265	0.087319	0.028674	0.10994	0.032971
10			ACCS	EOG090F040J	-0.7057513	2.81174E-05	0.014493	0.989623	8.57E-01	0.136943	0.378506	NA	0.108376	0	0.01595	0	0.026314	0.139567	0.339257	0.022287	0.145611	0	0.016511	0	0.023624
11			MRPL13	EOG090F0A3O	-0.8601564	3.01211E-05	0.014493	0.938577	0.6558909	0.031195	0.142267	NA	0.167549	0.066067	0.073682	0.059318	0.078803	0.043438	0.169437	0.020564	0.275027	0.055575	0.061341	0.058822	0.076298
12			TRIM25	EOG090F039M	-0.3927903	0.000035996	0.015988	0.984802	0.6447667	0.147415	0.136313	NA	0.016142	0.025251	0.023125	0.045129	0.043118	0.173777	0.181291	0.022543	0.022443	0.022522	0.020275	0.042733	0.04265
13			TMEM63A	EOG090F022V	-0.9622118	4.30734E-05	0.01658	0.928913	0.635143	4.81E-05	0.0079	NA	0.100638	0	0	0	0	6.86E-05	0.009566	0.02138	0.112558	0	0	0	0
14			SMG8	EOG090F01MM	-1.210637	4.1002E-05	0.01658	0.987727	0.629342	0.194763	0.065853	NA	0.016783	0.019183	0.025183	0.026135	0.035583	0.201493	0.064469	0.025244	0.019396	0.018536	0.030802	0.023909	0.036794
15			CMTR1	EOG090F01WE	-0.7020964	7.50584E-05	0.017149	0.980052	0.4838038	0.010398	0.01723	NA	0.018887	0.073954	0.06781	0.039324	0.051076	0.011369	0.019618	0.035424	0.026688	0.071353	0.065294	0.040689	0.048562
16			ABCD2	EOG090F02LH	-0.8570946	8.14913E-05	0.017149	0.976111	0.4617398	0.062894	0.111671	NA	0.029299	0.02935	0.025183	0.033717	0.026314	0.056525	0.114699	0.022543	0.025903	0.028713	0.020275	0.038734	0.024334
17			C2CD2L	EOG090F04AY	-0.7479158	8.81058E-05	0.017149	0.958907	0.3140142	0.02266	0.078837	NA	0.033436	0.055934	0.037485	0.059318	0.031829	0.01767	0.06329	0.02138	0.019382	0.057556	0.041139	0.064785	0.032971
18			TOM1	EOG090F052R	-0.7489444	6.14343E-05	0.017149	0.956129	0.4496638	0.005864	0.00685	NA	0.045106	0.055934	0.083633	0.077077	0.110663	0.006359	0.010848	0.023188	0.094785	0.055575	0.069284	0.072572	0.105965
19			TRAM1, TRAM1L1	EOG090F072D	-0.9811094	8.91006E-05	0.017149	0.959578	0.3436689	0.003533	0.013823	NA	0.197399	0.073954	0.07571	0.051411	0.051076	0.00411	0.015077	0.025244	0.199825	0.07734	0.075073	0.050979	0.05234
20			MARCH1, MARCH8	EOG090F07QP	-0.9662476	7.73219E-05	0.017149	0.971575	0.772529	0.072529	0.087999	NA	0.016783	0.015906	0.029316	0.026135	0.030073	0.06943	0.072509	0.010292	0.00748	0.016722	0.028674	0.023909	0.024334
21			LSM14B	EOG090F0889	-0.9293703	6.12433E-05	0.017149	0.944112	0.7357913	0.000277	0.000252	NA	0.016538	0.019183	0.049542	0.031813	0.114679	0.000632	0.000418	0.048636	0.027301	0.016722	0.047067	0.027257	0.105965
22			ERI3	EOG090F089Q	-1.0938566	7.30288E-05	0.017149	0.961957	0.7557582	0.046192	0.36681	NA	0.116022	0.015906	0.01595	0.026981	0.026314	0.052588	0.419028	0.020564	0.126848	0.016722	0.016511	0.025684	0.023624
23			PIKFYVE	EOG090F007L	-1.2027791	5.85901E-05	0.017149	0.993314	0.5419014	0.203142	0.380614	NA	0.016142	0	0.01595	0	0.026314	0.195718	0.37385	0.022287	0.019382	0	0	0	0.023624
24			NFXL1	EOG090F01W5	-0.8843063	5.97702E-05	0.017149	0.963058	0.7182771	0.008705	0.05777	NA	0.018887	0.520162	0.085622	0.458046	0.08465	0.010655	0.068518	0.022543	0.020584	0.486105	0.075073	0.447488	0.072345
25			IKBKE, TBK1	EOG090F0359	-0.568745	8.17434E-05	0.017149	0.995462	0.861342	0.404491	0.417006	NA	0.030417	0.015906	0.01595	0.026135	0.026314	0.407292	0.402012	0.013451	0.019382	0.016722	0.016511	0.023909	0.023624
26			ATXN1L	EOG090F03JS	-1.209321	4.83325E-05	0.017149	0.988396	-0.1013906	0.391556	0.64098	NA	0.249616	0.015906	0.01595	0.026135	0.026314	0.374123	0.689765	0.010292	0.339117	0.016722	0.016511	0.023909	0.023624
27			DCTN4	EOG090F04K0	-0.8718973	6.85492E-05	0.017149	0.991569	0.5533904	0.368439	0.747909	NA	0.015741	0.015906	0.01595	0.026135	0.026314	0.465849	0.87447	0.025244	0.020661	0.016722	0.016511	0.023909	0.023624
28			None defined.	EOG090F04KC	-0.9901861	8.48052E-05	0.017149	0.959296	0.4621886	0.163471	0.956796	NA	0.427618	0.015906	0.01595	0.026135	0.026314	0.149477	0.951298	0.025244	0.509514	0.016722	0.016511	0.023909	0.023624
29			VMP1	EOG090F069J	-0.9130043	0.000103705	0.018712	0.957585	0.380357	0.100521	0.808093	NA	0.904267	0.015906	0.01595	0.026135	0.026314	0.120419	0.766499	0.025244	0.925876	0.016722	0.016511	0.023909	0.024334
30			HIBADH	EOG090F07PA	-1.3114934	0.000109233	0.019112	0.996509	0.6466836	0.221879	0.276726	NA	0.030417	0	0	0.026135	0.026314	0.258565	0.307128	0.025244	0.03093	0	0	0.023909	0.023624

Supplemental Table 11. AW (A): Number significant GO terms for each CAPER model at each cutoff. See Table 3.

Tool: Database	Dependent variable	Independent variable	De Novo			Alignment			Shared		
			Pos	Neg	All	Pos	Neg	All	Pos	Neg	All
Ingenuity Pathways: Diseases and Functions	MLSLW	LG	36	174	NA	44	162	NA	44	174	NA
TopGO: Biological Processes	MLS	LG	NA	NA	NA	50	155	192	57	162	240
	MLSW	LG	0	96	0	94	85	129	38	26	37
	MLSLW	LG	5	107	98	5	124	98	5	107	86
TopGO: Cellular Components	MLS	LG	NA	NA	NA	9	26	22	10	12	12
	MLSW	LG	6	1	5	7	2	5	7	2	3
	MLSLW	LG	1	25	20	0	23	20	0	25	19
TopGO: Molecular Processes	MLS	LG	NA	NA	NA	12	50	40	13	26	38
	MLSW	LG	18	29	43	26	29	43	18	10	29
	MLSLW	LG	3	26	28	0	30	28	0	26	27

Supplemental Table 12a. Significant Biological Processes. CAPER, Dependent = MLSLW, Independent = LG, Metadata = AW (A)

Shared Order	Direction of Association	Align-Only Significant	De Novo-Only Significant	GO.ID	Term	Annotated	Significant	Expected	Rank in classicFisher	classicFisher	classicKS	elimKS
1	**All**	1	1	GO:0001818	negative regulation of cytokine producti...	5	5	5	1	1	0.0032	0.0032
2		2	2	GO:0048518	positive regulation of biological proces...	38	38	38	2	1	0.0033	0.0033
3		3	3	GO:0006511	ubiquitin-dependent protein catabolic pr...	6	6	6	3	1	0.0057	0.0057
4		4	4	GO:0043412	macromolecule modification	37	37	37	4	1	0.0074	0.0074
5		NA	NA	GO:0044281	small molecule metabolic process	16	16	16	5	1	0.0075	0.0075
6		8	8	GO:0006464	cellular protein modification process	35	35	35	6	1	0.0116	0.0116
7		9	9	GO:0036211	protein modification process	35	35	35	7	1	0.0116	0.0116
8		15	15	GO:0140014	mitotic nuclear division	3	3	3	8	1	0.0116	0.0116
9		18	18	GO:0009893	positive regulation of metabolic process	25	25	25	9	1	0.0117	0.0117
10		11	11	GO:0030031	cell projection assembly	6	6	6	10	1	0.0125	0.0125
11		12	12	GO:0044782	cilium organization	6	6	6	11	1	0.0125	0.0125
12		13	13	GO:0060271	cilium assembly	6	6	6	12	1	0.0125	0.0125
13		14	14	GO:0120031	plasma membrane bounded cell projection ...	6	6	6	13	1	0.0125	0.0125
14		16	16	GO:0010498	proteasomal protein catabolic process	5	5	5	14	1	0.0141	0.0141
15		10	10	GO:0009605	response to external stimulus	11	11	11	15	1	0.0143	0.0143
16		6	6	GO:0009894	regulation of catabolic process	10	10	10	16	1	0.0145	0.0145
17		7	7	GO:0031329	regulation of cellular catabolic process	10	10	10	17	1	0.0145	0.0145
18		19	19	GO:0045787	positive regulation of cell cycle	5	5	5	18	1	0.0179	0.0179
19		20	20	GO:0090068	positive regulation of cell cycle proces...	5	5	5	19	1	0.0179	0.0179
20		17	17	GO:0071310	cellular response to organic substance	14	14	14	20	1	0.0181	0.0181
21		21	21	GO:0009896	positive regulation of catabolic process	7	7	7	21	1	0.0192	0.0192
22		22	22	GO:0031331	positive regulation of cellular cataboli...	7	7	7	22	1	0.0192	0.0192
NA		5	NA	GO:0055086	nucleobase-containing small molecule met...	See Suplemental Data						
NA		23	NA	GO:0016567	protein ubiquitination							
NA		24	NA	GO:0032446	protein modification by small protein co...							
NA		25	NA	GO:0034097	response to cytokine							
NA		26	NA	GO:0070647	protein modification by small protein co...							
NA	**Pos**			None								
1	**Neg**	1	1	GO:0001818	negative regulation of cytokine producti...	5	5	5	1	1	0.00342	0.0034
2		3	3	GO:0048522	positive regulation of cellular process	28	28	28	2	1	0.00516	0.0052
3		2	2	GO:0009605	response to external stimulus	10	10	10	3	1	0.00531	0.0053
4		8	8	GO:0009893	positive regulation of metabolic process	21	21	21	4	1	0.00782	0.0078
5		11	11	GO:0016032	viral process	8	8	8	5	1	0.00816	0.0082
6		4	4	GO:0071310	cellular response to organic substance	13	13	13	6	1	0.00873	0.0087
7		7	7	GO:0006464	cellular protein modification process	32	32	32	7	1	0.0092	0.0092
8		6	6	GO:0009890	negative regulation of biosynthetic proc...	11	11	11	8	1	0.0092	0.0092
9		9	9	GO:0031396	regulation of protein ubiquitination	3	3	3	9	1	0.01098	0.011
10		10	10	GO:1903320	regulation of protein modification by sm...	3	3	3	10	1	0.01098	0.011
11		14	14	GO:0045787	positive regulation of cell cycle	3	3	3	11	1	0.01237	0.0124
12		15	15	GO:0090068	positive regulation of cell cycle proces...	3	3	3	12	1	0.01237	0.0124
13		16	16	GO:0140014	mitotic nuclear division	3	3	3	13	1	0.01237	0.0124
14		17	17	GO:1903047	mitotic cell cycle process	5	5	5	14	1	0.01284	0.0128
15		27	**27**	GO:0030031	cell projection assembly	6	6	6	15	1	0.01565	0.0156
16		28	28	GO:0044782	cilium organization	6	6	6	16	1	0.01565	0.0156
17		29	NA	GO:0060271	cilium assembly	6	6	6	17	1	0.01565	0.0156
18		30	NA	GO:0120031	plasma membrane bounded cell projection ...	6	6	6	18	1	0.01565	0.0156
19		18	18	GO:0006511	ubiquitin-dependent protein catabolic pr...	5	5	5	19	1	0.0159	0.0159
20		19	19	GO:0010498	proteasomal protein catabolic process	5	5	5	20	1	0.0159	0.0159
21		20	20	GO:0016567	protein ubiquitination	5	5	5	21	1	0.0159	0.0159
22		21	21	GO:0019941	modification-dependent protein catabolic...	5	5	5	22	1	0.0159	0.0159
23		22	22	GO:0032446	protein modification by small protein co...	5	5	5	23	1	0.0159	0.0159
24		23	23	GO:0043632	modification-dependent macromolecule cat...	5	5	5	24	1	0.0159	0.0159
25		24	24	GO:0045087	innate immune response	5	5	5	25	1	0.0159	0.0159
26		25	25	GO:0051603	proteolysis involved in cellular protein...	5	5	5	26	1	0.0159	0.0159
27		26	**26**	GO:0070647	protein modification by small protein co...	5	5	5	27	1	0.0159	0.0159
28		NA	NA	GO:0006952	defense response	10	10	10	28	1	0.01698	0.017
NA		5	NA	GO:0044281	small molecule metabolic process	See Suplemental Data						
NA		12	NA	GO:0044403	symbiont process							
NA		13	NA	GO:0044419	interspecies interaction between organis...							
NA		31	NA	GO:0051704	multi-organism process							

Supplemental Table 12b. Significant Cellular Components. CAPER, Dependent = MLSLW, Independent = LG, Metadata = AW (A)

Shared Order	Direction of Association	Align-Only Significant	De Novo-Only Significant	GO.ID	Term	Annotated	Significant	Expected	Rank in classicFisher	classicFisher	classicKS	elimKS
1	**All**	1	1	GO:0043227	membrane-bounded organelle	83	83	83	1	1	0.0037	0.0037
2		2	2	GO:0043229	intracellular organelle	85	85	85	2	1	0.0044	0.0044
3		5	5	GO:0043231	intracellular membrane-bounded organelle	79	79	79	3	1	0.0118	0.0118
4		3	3	GO:0055037	recycling endosome	3	3	3	4	1	0.0278	0.0278
5		6	6	GO:0000118	histone deacetylase complex	2	2	2	5	1	0.0323	0.0323
6		7	7	GO:0016581	NuRD complex	2	2	2	6	1	0.0323	0.0323
7		8	8	GO:0090545	CHD-type complex	2	2	2	7	1	0.0323	0.0323
8		NA	NA	GO:0005634	nucleus	50	50	50	8	1	0.0344	0.0344
9		NA	NA	GO:0044446	intracellular organelle part	74	74	74	9	1	0.0378	0.0378
10		9	9	GO:0016604	nuclear body	8	8	8	10	1	0.0414	0.0414
11		10	10	GO:1904949	ATPase complex	4	4	4	11	1	0.0427	0.0427
12		11	11	GO:0005829	cytosol	49	49	49	12	1	0.0443	0.0443
13		12	12	GO:0005777	peroxisome	2	2	2	13	1	0.0472	0.0472
14		13	13	GO:0005778	peroxisomal membrane	2	2	2	14	1	0.0472	0.0472
15		14	14	GO:0031903	microbody membrane	2	2	2	15	1	0.0472	0.0472
16		15	15	GO:0042579	microbody	2	2	2	16	1	0.0472	0.0472
17		16	16	GO:0044438	microbody part	2	2	2	17	1	0.0472	0.0472
18		17	17	GO:0044439	peroxisomal part	2	2	2	18	1	0.0472	0.0472
NA		4	NA	GO:0005737	cytoplasm	See Suplemental Data						
NA		18	NA	GO:0000922	spindle pole							
NA		19	NA	GO:0005912	adherens junction							
NA	**Pos**			None.								
1	**Neg**	1	1	GO:0043231	intracellular membrane-bounded organelle	72	72	72	1	1	0.00164	0.0016
2		4	4	GO:0044446	intracellular organelle part	68	68	68	2	1	0.00502	0.005
3		3	3	GO:0005634	nucleus	43	43	43	3	1	0.01451	0.0145
4		11	11	GO:0005730	nucleolus	7	7	7	4	1	0.01768	0.0177
5		NA	NA	GO:0044428	nuclear part	35	35	35	5	1	0.01843	0.0184
6		NA	NA	GO:0031981	nuclear lumen	33	33	33	6	1	0.02384	0.0238
7		2	2	GO:0055037	recycling endosome	3	3	3	7	1	0.02942	0.0294
8		NA	21	GO:0005654	nucleoplasm	30	30	30	8	1	0.03259	0.0326
9		5	5	GO:0000118	histone deacetylase complex	2	2	2	9	1	0.03468	0.0347
10		6	6	GO:0016581	NuRD complex	2	2	2	10	1	0.03468	0.0347
11		7	7	GO:0017053	transcriptional repressor complex	2	2	2	11	1	0.03468	0.0347
12		8	8	GO:0090545	CHD-type complex	2	2	2	12	1	0.03468	0.0347
13		9	9	GO:0090568	nuclear transcriptional repressor comple...	2	2	2	13	1	0.03468	0.0347
14		10	10	GO:0016604	nuclear body	8	8	8	14	1	0.04396	0.044
15		12	**12**	GO:1904949	ATPase complex	4	4	4	15	1	0.04452	0.0445
16		13	13	GO:0005777	peroxisome	2	2	2	16	1	0.04873	0.0487
17		14	14	GO:0005778	peroxisomal membrane	2	2	2	17	1	0.04873	0.0487
18		15	15	GO:0031903	microbody membrane	2	2	2	18	1	0.04873	0.0487
19		16	16	GO:0042579	microbody	2	2	2	19	1	0.04873	0.0487
20		17	17	GO:0044438	microbody part	2	2	2	20	1	0.04873	0.0487
21		18	18	GO:0044439	peroxisomal part	2	2	2	21	1	0.04873	0.0487
22		19	NA	GO:0000922	spindle pole							
23		20	NA	GO:0005912	adherens junction	See Suplemental Data						

Supplemental Table 12c. C. Significant Molecular Functions. CAPER, Dependent = MLSLW, Independent = LG, Metadata = AW (A)

Shared Order	Direction of Association	Align-Only Significant	De Novo-Only Significant	GO.ID	Term	Annotated	Significant	Expected	Rank in classicFisher	classicFisher	classicKS	elimKS
1	**All**	1	1	GO:0005524	ATP binding	13	13	13	1	1	0.0021	0.0021
2		2	2	GO:0003824	catalytic activity	39	39	39	2	1	0.0188	0.0188
3		3	3	GO:0019902	phosphatase binding	2	2	2	3	1	0.0216	0.0216
4		4	4	GO:0019903	protein phosphatase binding	2	2	2	4	1	0.0216	0.0216
5		5	5	GO:0001103	RNA polymerase II repressing transcripti...	2	2	2	5	1	0.0323	0.0323
6		6	6	GO:0042826	histone deacetylase binding	2	2	2	6	1	0.0323	0.0323
7		7	7	GO:0070491	repressing transcription factor binding	2	2	2	7	1	0.0323	0.0323
8		8	8	GO:0140096	catalytic activity, acting on a protein	13	13	13	8	1	0.0348	0.0348
9		9	9	GO:0016879	ligase activity, forming carbon-nitrogen...	2	2	2	9	1	0.0349	0.0349
10		10	10	GO:0016874	ligase activity	3	3	3	10	1	0.0408	0.0408
11		11	11	GO:0045296	cadherin binding	3	3	3	11	1	0.0408	0.0408
12		12	12	GO:0050839	cell adhesion molecule binding	3	3	3	12	1	0.0408	0.0408
13		13	13	GO:0000287	magnesium ion binding	2	2	2	13	1	0.0438	0.0438
14		14	14	GO:0005488	binding	90	90	90	14	1	0.031	0.0496
NA		16	16	GO:0016757	transferase activity, transferring glyco...	See Suplemental Data						
NA	**Pos**	1	NA	GO:0043167	ion binding	3	3	3	1	1	0.016	0.016
NA		2	NA	GO:0043169	cation binding	2	2	2	2	1	0.049	0.049
NA		3	NA	GO:0046872	metal ion binding	2	2	2	3	1	0.049	0.049
1	**Neg**	1	1	GO:0005524	ATP binding	13	13	13	1	1	0.0023	0.0023
2		2	2	GO:0140096	catalytic activity, acting on a protein	11	11	11	2	1	0.0067	0.0067
3		3	3	GO:0019902	phosphatase binding	2	2	2	3	1	0.0221	0.0221
4		4	4	GO:0019903	protein phosphatase binding	2	2	2	4	1	0.0221	0.0221
5		5	5	GO:0005102	signaling receptor binding	4	4	4	5	1	0.0298	0.0298
6		6	7	GO:0001103	RNA polymerase II repressing transcripti...	2	2	2	6	1	0.0347	0.0347
7		7	8	GO:0042826	histone deacetylase binding	2	2	2	7	1	0.0347	0.0347
8		8	9	GO:0070491	repressing transcription factor binding	2	2	2	8	1	0.0347	0.0347
9		9	14	GO:0016879	ligase activity, forming carbon-nitrogen...	2	2	2	9	1	0.0378	0.0378
10		10	11	GO:0016874	ligase activity	3	3	3	10	1	0.0406	0.0406
11		11	12	GO:0045296	cadherin binding	3	3	3	11	1	0.0406	0.0406
12		12	13	GO:0050839	cell adhesion molecule binding	3	3	3	12	1	0.0406	0.0406
13		13	NA	GO:0042802	identical protein binding	9	9	9	13	1	0.0434	0.0434
14		14	15	GO:0000287	magnesium ion binding	2	2	2	14	1	0.0448	0.0448
15		15	**10**	GO:0005488	binding	79	79	79	15	1	0.0268	0.0453
16		16	NA	GO:0005515	protein binding	64	64	64	16	1	0.0498	0.0498
NA		17	18	GO:0016757	transferase activity, transferring glyco...	See Suplemental Data						

Supplemental Table 13. Significant Ingenuity Pathway Diseases or Functions. CAPER, Dependent = MLSLW, Independent = LG, Metadata = AW (A)

Shared Order	Direction	P-value Cutoff	Rank Order (Align)	Rank Order (De Novo)	Categories	Diseases or Functions Annotation	P-value (Shared)	P-value (Align)	P-value (De novo)	# Molecules (Shared)	# Molecules (Align)	# Molecules (De Novo)
1	**Pos**	**5.00E-06**	1	1	Cancer, Organismal Injury and Abnormalities, Reproductive System Disease, Tumor Morphology	Production of prostatic intraepithelial tumor	1.68E-07	1.68E-07	4.06E-07	2	2	2
2			2	3	Cell Cycle, Connective Tissue Development and Function	Cell cycle progression of chondrocytes	5.05E-07	5.05E-07	1.22E-06	2	2	2
3			3	4	Cellular Growth and Proliferation, Hematological System Development and Function, Lymphoid Tissue Structure and Development, Organ Development, Organ Morphology, Tissue Development, Tissue Morphology	Expansion of thymic medullary epithelial cells	5.05E-07	5.05E-07	1.22E-06	2	2	2
4			4	5	Dermatological Diseases and Conditions, Hair and Skin Development and Function, Organ Morphology, Organismal Injury and Abnormalities	Abnormal morphology of guard hair	1.01E-06	1.01E-06	2.43E-06	2	2	2
5			5	6	Connective Tissue Development and Function, Organ Morphology, Organismal Development, Skeletal and Muscular System Development and Function, Tissue Development, Tissue Morphology	Diameter of humerus	1.01E-06	1.01E-06	2.43E-06	2	2	2
6			6	7	Developmental Disorder	Hypoplasia of interparietal bone	1.01E-06	1.01E-06	2.43E-06	2	2	2
7			7	8	Cell-mediated Immune Response, Cellular Development, Cellular Function and Maintenance, Cellular Growth and Proliferation, Embryonic Development,	Production of naive T lymphocytes	1.01E-06	1.01E-06	2.43E-06	2	2	2
8			8	16	Cell Cycle	Arrest in G1 phase of bone cancer cell lines	1.13E-06	1.13E-06	4.36E-06	3	3	3
9			9	17	Cell Cycle	Arrest in G1 phase of sarcoma cell lines	0.0000012	1.20E-06	4.64E-06	3	3	3
10			10	9	Dermatological Diseases and Conditions, Hair and Skin Development and Function, Organ Morphology, Organismal Injury and Abnormalities	Abnormal morphology of zigzag hair	1.68E-06	1.68E-06	4.05E-06	2	2	2
11			11	10	Cell Morphology	Cell flattening of bone cancer cell lines	1.68E-06	1.68E-06	4.05E-06	2	2	2
12			12	11	Cell Morphology	Cell flattening of sarcoma cell lines	1.68E-06	1.68E-06	4.05E-06	2	2	2
13			13	12	Connective Tissue Development and Function, Organ Morphology, Organismal Development, Skeletal and Muscular System Development and Function, Tissue Development, Tissue Morphology	Diameter of radius	1.68E-06	1.68E-06	4.05E-06	2	2	2
14			14	13	Connective Tissue Development and Function, Organ Morphology, Organismal Development, Skeletal and Muscular System Development and Function, Tissue Development, Tissue Morphology	Diameter of ulna	1.68E-06	1.68E-06	4.05E-06	2	2	2
15			15	14	Cell Morphology, Endocrine System Disorders, Organ Morphology, Organismal Injury and Abnormalities, Reproductive System Development and Function, Reproductive System Disease	Lack of mitochondrial sheath	1.68E-06	1.68E-06	4.05E-06	2	2	2
16			16	20	Gene Expression, Protein Synthesis	Initiation of translation of mRNA	1.79E-06	1.79E-06	6.90E-06	3	3	3
17			17	18	Dermatological Diseases and Conditions, Hair and Skin Development and Function, Organ Morphology, Organismal Injury and Abnormalities	Abnormal morphology of awl hair	2.52E-06	2.52E-06	6.08E-06	2	2	2
18			18	19	Dermatological Diseases and Conditions, Organ Morphology, Organismal Injury and Abnormalities	Abnormal morphology of enlarged sebaceous glands	2.52E-06	2.52E-06	6.08E-06	2	2	2
19			19	29	Hematological System Development and Function, Lymphoid Tissue Structure and Development, Organ Morphology, Organismal Development, Tissue Morphology	Morphology of spleen	2.93E-06	2.93E-06	2.95E-05	5	5	5
20			20	25	Hematological System Development and Function, Lymphoid Tissue Structure and Development, Organ Morphology, Organismal Development, Tissue Morphology	Morphology of thymus gland	3.35E-06	3.35E-06	2.08E-05	4	4	4
21			21	22	Cell Cycle, Embryonic Development	Senescence of embryonic cell lines	3.95E-06	3.95E-06	1.52E-05	3	3	3
22			22	21	Endocrine System Disorders, Organ Morphology, Organismal Injury and Abnormalities, Reproductive System Development and Function, Reproductive	Abnormal morphology of cauda epididymis	4.71E-06	4.71E-06	1.13E-05	2	2	2
69			69	2	Cancer, Organismal Injury and Abnormalities	Non-melanoma solid tumor	1.38E-04	1.38E-04	1.20E-06	15	15	23
87			87	15	Cancer, Organismal Injury and Abnormalities	Malignant solid tumor	3.09E-04	3.09E-04	4.15E-06	15	15	23
1	**Neg**	**5.00E-08**	1	1	Cancer, Organismal Injury and Abnormalities	Tumorigenesis of tissue	3.31E-16	2.56E-17	3.31E-16	131	134	131
2			2	2	Cancer, Organismal Injury and Abnormalities	Head and neck tumor	2.08E-15	6.46E-16	2.08E-15	110	112	110
3			4	3	Cancer, Organismal Injury and Abnormalities	Head and neck carcinoma	9.77E-15	3.02E-15	9.77E-15	107	109	107
4			5	4	Cancer, Organismal Injury and Abnormalities	Cancer of secretory structure	1.34E-14	4.44E-15	1.34E-14	111	113	111
5			3	5	Cancer, Organismal Injury and Abnormalities	Cancer	3.21E-14	2.88E-15	3.21E-14	138	141	138
6			6	6	Cancer, Organismal Injury and Abnormalities	Neck neoplasm	7.3E-14	6.16E-15	7.30E-14	103	106	103
7			NA	7	Cancer, Endocrine System Disorders, Organismal Injury and Abnormalities	Nonpituitary endocrine tumor	8.24E-14	NA	8.24E-14	103	NA	103
8			1	8	Cancer, Endocrine System Disorders, Organismal Injury and Abnormalities	Thyroid gland tumor	9.52E-14	8.04E-15	9.52E-14	102	105	102
9			2	9	Cancer, Endocrine System Disorders, Organismal Injury and Abnormalities	Thyroid carcinoma	2.18E-13	1.89E-14	2.18E-13	101	104	101
10			3	10	Cancer, Organismal Injury and Abnormalities	Nonhematologic malignant neoplasm	6.45E-13	6.94E-14	6.45E-13	126	129	126
11			4	11	Cancer, Organismal Injury and Abnormalities	Carcinoma	7.3E-13	7.89E-14	7.30E-13	124	127	124
12			12	12	Cancer, Organismal Injury and Abnormalities	Adenocarcinoma	9.44E-13	1.17E-12	9.44E-13	112	113	112
13			11	13	Cancer, Organismal Injury and Abnormalities	Extracranial solid tumor	4.11E-12	4.79E-13	4.11E-12	129	132	129
14			17	14	Cancer, Organismal Injury and Abnormalities	Abdominal neoplasm	6.96E-12	1.03E-10	6.96E-12	119	118	119
15			3	15	Cancer, Organismal Injury and Abnormalities	Abdominal adenocarcinoma	7E-12	7.66E-11	7.00E-12	108	107	108
16			7	16	Cancer, Organismal Injury and Abnormalities	Abdominal carcinoma	1.05E-11	1.72E-10	1.05E-11	113	116	113
17			13	17	Cancer, Organismal Injury and Abnormalities	Non-melanoma solid tumor	1.06E-11	1.29E-12	1.06E-11	125	128	125
18			4	18	Cancer, Organismal Injury and Abnormalities	Abdominal cancer	1.28E-11	1.28E-10	1.28E-11	117	112	117
19			2	19	Cancer, Gastrointestinal Disease, Organismal Injury and Abnormalities	Gastrointestinal tumor	2.02E-11	6.77E-11	2.02E-11	103	103	103
20			3	20	Cancer, Gastrointestinal Disease, Organismal Injury and Abnormalities	Gastrointestinal tract cancer	3.3E-11	1.08E-10	3.30E-11	102	102	102
21			1	21	Cancer, Gastrointestinal Disease, Organismal Injury and Abnormalities	Gastrointestinal carcinoma	3.46E-11	1.05E-10	3.46E-11	98	98	98
22			5	22	Cancer, Gastrointestinal Disease, Organismal Injury and Abnormalities	Large intestine carcinoma	4.54E-11	1.29E-10	4.54E-11	94	94	94
23			6	23	Cancer, Gastrointestinal Disease, Organismal Injury and Abnormalities	Gastrointestinal adenocarcinoma	5.5E-11	1.58E-10	5.50E-11	95	95	95
24			8	24	Cancer, Gastrointestinal Disease, Organismal Injury and Abnormalities	Large intestine neoplasm	7.01E-11	2.10E-10	7.01E-11	98	98	98
25			2	25	Auditory Disease	Tinnitus	9.84E-11	1.07E-10	9.84E-11	6	6	6
26			1	26	Cancer, Organismal Injury and Abnormalities	Malignant solid tumor	1.09E-10	1.48E-11	1.09E-10	127	130	127
27			9	27	Cancer, Gastrointestinal Disease, Organismal Injury and Abnormalities	Malignant neoplasm of large intestine	1.32E-10	3.86E-10	1.32E-10	97	97	97
28			10	28	Cancer, Gastrointestinal Disease, Organismal Injury and Abnormalities	Large intestine adenocarcinoma	1.95E-10	5.24E-10	1.95E-10	92	92	92
29			11	29	Cancer, Gastrointestinal Disease, Organismal Injury and Abnormalities	Digestive organ tumor	1.95E-10	6.99E-10	1.95E-10	109	109	109
30			12	30	Cancer, Gastrointestinal Disease, Organismal Injury and Abnormalities	Digestive system cancer	2.73E-10	9.31E-10	2.73E-10	107	107	107
31			13	31	Organismal Injury and Abnormalities, Renal and Urological Disease	Renal colic	6.74E-09	7.29E-09	6.74E-09	6	6	6
32			14	32	Gastrointestinal Disease, Hepatic System Disease, Organismal Injury and Abnormalities	Drug-induced liver disease	1.36E-08	1.48E-08	1.36E-08	6	6	6
33			15	33	Connective Tissue Disorders, Organismal Injury and Abnormalities, Skeletal and Muscular Disorders	Ankle sprain	1.48E-08	1.58E-08	1.48E-08	5	5	5
34			16	34	Organismal Injury and Abnormalities	Myofascial pain	1.97E-08	2.10E-08	1.97E-08	5	5	5
35			18	35	Connective Tissue Disorders, Inflammatory Disease, Organismal Injury and Abnormalities, Skeletal and Muscular Disorders	Lateral epicondylitis	3.32E-08	3.55E-08	3.32E-08	5	5	5
39			17	34	Malignant genitourinary solid tumor	Malignant genitourinary solid tumor	9.38E-08	3.45E-08	3.45E-08	87	87	85

APPENDIX C (4.2)

SUPPLEMENTARY TABLES: REFERENCES

See the main text for additional context and references for the supplementary tables

Supplementary Table 1. The drugs being considered for human clinical trials focused on longevity were obtained from multiple sources (see Supplementary Internet References below) as well as published studies.[1-16]

Supplementary Table 2. The top-ranked drugs based on protein-protein interaction networks are from Fuentealba et al.[17]

Supplementary Table 3. The variants associated with lifespan are from the meta-analysis by Sebastiani et al.[18]

Supplementary Table 4a. The variants associated with parental lifespan are from Timmers et al.[19]

Supplementary Table 4b. The variants associated with parental lifespan after taking into account factors affecting mortality are from Timmers et al.[19]

Supplementary Table 5. The healthspan variants considered are from Zenin et al.[20] and Hornstrup et al.[21]

Supplementary Table 6. The genes considered are taken from the review by Harper et al. and associated references.[22-27]

Supplementary Internet References: URLs and links to information about drugs listed in Supplementary Table 1.

Supplementary Table 1. Human genetic information related to drugs that are being considered for study in human clinical trials on aspects of human aging and longevity obtained from internet and literature search sources (see Methods).

Drug/Compound Information				Variants		Associations Involving Variants in LD with Target Gene eQTLs							eQTL Tissues							
Drug	**Target**	**Refs**	**TTD Mechanism**	**eQTLs**	**# LD**	**Long**	**Age Rel**	**Other**	**LD eQTLs**	**LD pQTLs**	**LD mQTLs**	**LD Other**	**Adipose**	**Artery**	**Brain**	**Heart**	**Muscle**	**Skin**	**WB**	**Total**
NMN	NMNAT1	1	A	19	156	0	0	0	68	0	23	16	25	5	1	16	48	42	0	250
NMN	NMNAT3	1	A	43	539	0	1	0	833	0	133	103	27	108	82	25	23	15	0	474
NMN	BST1	1	A	31	541	0	11	10	328	0	48	184	0	0	0	74	12	0	24	383
TA-65	TERT	2	A	9	96	0	45	5	24	0	45	14	0	0	0	0	0	3	0	64
Doxycycline	SOX2	3	A	8	740	0	0	0	8	0	0	0	0	5	1	0	0	1	0	11
Doxycycline	KLF4	3	A	15	112	0	0	0	33	0	2	2	10	0	0	0	0	2	0	12
Doxycycline	c-Myc (MYC)	3	A	18	1554	0	4	6	31	0	3	21	0	0	19	0	0	3	1	27
CPHPC	APCS	4	M	3	8	0	0	0	6	0	0	24	0	0	0	0	0	0	0	1
Rosiglitazone	PPARG	5	G	4	98	0	16	40	259	0	17	216	0	1	*^{E}	27	0	7	0	548
J147 (circumin)	ATP5A1	6	I	36	1604	0	0	14	857	0	177	315	98	3	14	133	120	62	138	899
NPT088	α-synuclein (SNCA)	7	I	36	1475	0	9	11	527	0	158	170	0	307	2	217	0	1	24	909
UBX0101	MDM2	8	D	26	396	0	0	0	314	0	2	211	21	0	0	4	0	0	103	147
Quercetin	Lyso-PAF (LPCAT2)	9	I	46	2424	0	0	32	920	0	100	605	20	214	1	44	2	15	0	468
Dasatinib	SRC	9	I	20	216	0	0	2	273	0	46	25	0	0	0	0	0	0	0	6
Dasatinib	ABL1	9	I	18	734	0	0	3	32	0	7	9	0	3	0	4	0	0	0	19
Dasatinib	LCK	9	I	15	1792	0	0	0	52	2	28	19	24	0	10	0	0	0	0	60
Dasatinib	FYN	9	I	34	618	0	1	5	193	0	56	46	0	184	0	1	0	0	0	373
Fisetin	FABG (HSD17B8)	10	I	54	1175	0	3	16	*^{A}	0	*^{C}	338	*^{D}	852	*^{F}	313	*^{G}	*^{H}	60	5863
Fisetin	CDK6	10	I	12	156	0	6	9	49	0	14	9	0	4	3	0	1	0	0	43
Fisetin	FASN	10	I	25	2656	0	0	34	*^{B}	0	597	836	2	17	2	0	511	2	397	1889
UBX1967	BCL-2 (BCL2)	11	I	26	220	0	0	10	50	0	11	8	15	27	1	48	15	0	0	142
Alk5i	ALK5 (TGFBR1)	12	I	18	275	0	4	8	43	0	10	46	1	10	0	0	1	0	83	99
RTB101	TORC1 (CRTC1)	13	I	17	111	0	0	1	39	0	18	4	2	32	25	0	0	0	0	87
SRK-015	TGFβ (TGFB1)	14	I	21	728	0	12	1	239	5	48	67	12	0	6	9	32	0	0	268
SM04690	CLK2	15	I	12	240	0	0	1	158	0	68	54	0	2	23	0	0	0	0	26
SM04690	DYRK1A	15	I	23	870	0	1	0	62	0	6	14	3	0	1	3	0	10	0	234
MSI-1436	PTP1B (PTPN1)	16	I	28	883	0	1	8	165	0	7	34	0	0	0	1	0	0	21	66

Key: See Table 1. Refs = Relevant reference discussing the drug and/or its gene target. TTD Mechanism: A=Activator, M=Modulator, G=Agonist, I=Inhibitor, D=Disrupt interaction. Large outliers highlighted and emitted for printability: LD eQTLs (> 1000): *^{A}=1202, *^{B}=1029; LD mQTLs (>1000): *^{C}=1229; Adipose (>100): *^{D} = 239, Brain (> 100): *^{E}=494, *^{F}=481; Muscle (>1000): *^{G}=1123; Skin: (>100): *^{H}=760.

Supplementary Table 2. Human genetic information related to drugs identified by Fuentealba et al. as being good candidates for promoting healthy aging based on these drugs being significantly enriched for ageing-related targets [17].

Drug/Compound Information						Variants		Associations Involving Variants in LD with Target Gene eQTLs							eQTL Tissues							
Drug	**Extend**	**Toxic**	**Status**	**Gene Target**	**TTD Mechanism**	**eQTLs**	**# LD**	**Long**	**Age Rel**	**Other**	**LD eQTLs**	**LD pQTLs**	**LD mQTLs**	**LD Other**	**Adipose**	**Artery**	**Brain**	**Heart**	**Muscle**	**Skin**	**WB**	**Sum**
Resveratrol	Y	N	I	NOS2	I	29	516	0	0	36	*C	0	214	843	*D	*F	*G	*I	149	305	0	*K
Resveratrol	Y	N	I	PTGS1	I	41	484	0	0	27	725	3	28	133	2	64	1	24	2	18	0	181
Resveratrol	Y	N	I	PTGS2	I	17	*A	0	0	10	194	5	21	130	0	0	0	0	0	0	0	9
Sunitinib	N	N	A	KDR	M	13	309	0	0	0	16	0	4	0	0	3	5	0	0	0	0	812
Genistein	Y	N	I	ESR2	M	28	881	0	10	25	360	0	81	284	0	0	58	9	226	486	76	*L
Simvastatin	Y	N	A	HMGCR	I	14	533	1	1	23	47	1	27	125	0	0	1	0	34	0	0	35
Simvastatin	Y	N	P	PPARG	G	34	987	0	16	40	259	0	17	216	0	1	*H	27	0	7	0	548
Tanespimycin	N	N	I	HSP90AA1	I	19	401	0	0	4	103	0	18	57	35	4	3	0	92	250	0	393
Regorafenib	N	N	A	KDR	M	13	309	0	0	0	16	0	4	0	0	3	5	0	0	0	0	812
Regorafenib	N	N	A	RET	M	26	992	0	0	38	458	3	174	184	*E	47	1	0	0	80	0	668
Regorafenib	N	N	A	KIT	M	8	511	0	0	11	133	0	133	42	0	0	0	0	0	0	0	114
Celecoxib	Y	N	A	PTGS2	I	17	*A	0	0	10	194	5	21	130	0	0	0	0	0	0	0	9
Hydrogen peroxide	N	Y	I	PTPN1	I	28	883	0	1	8	165	0	7	34	0	0	0	1	0	0	21	66
GW-501516	N	N	I	PPARD	M	24	622	0	2	5	120	0	19	5	0	0	0	0	6	34	11	336
Bexarotene	N	N	A	RXRA	M	21	799	0	0	0	33	0	19	10	0	11	0	0	180	2	0	253
Sorafenib	N	N	A	PDGFRB	M	22	425	0	1	7	160	0	404	106	0	32	0	22	1	0	0	107
Sorafenib	N	N	A	KIT	M	8	511	0	0	11	133	0	133	42	0	0	0	0	0	0	0	114
Sorafenib	N	N	A	KDR	M	13	309	0	0	0	16	0	4	0	0	3	5	0	0	0	0	812
Sirolimus	Y	N	A	MTOR	I	24	*B	0	22	7	548	0	143	217	0	0	2	0	9	17	*J	437

Key: See Table 1; Extend = evidence that the drug can increase lifespan in model species per Fuentealba et al.[17]; Toxic = evidence exists that the drug is toxic per Fuentealba et al.[17] Extend: Y=Yes, N=No. Toxic: Y=Yes, N=No. Status: I=Investigational, A=Approved, P=Phase 3. TTD Mechanism: I=Inhibitor, M=Modulator, G=Agonist. Large outliers highlighted and emitted for printability: # LD: (>1000): *A=1154, *B=1753; LD eQTLs (>1000): C*=2381; Adipose (>100): *D=201, *E=135; Artery (>100): *F=364; Brain (>100): *G=772, *H=492; Heart (>100): *I=123; WB (>100) *J=212; Sum (>1000): *K=3374, *L=1348.

Supplementary Table 3. Human genetic information related to variants identified by Sebastiani et al. exhibit statistically significant evidence for association with longevity.[18]

Associated Variant Information			Druggable?		Annotations		Associations Involving Variants in LD with Target SNP							Chem Studies on Gene				
SNP	**Gene**	**Chrom**	**PCh**	**TTD**	**eQTL?**	**# LD**	**Long**	**Age Rel**	**Other**	**LD eQ**	**LD pQ**	**LD mQ**	**LD O**	**#Ch**	**#Ch A**	**#TTD**	**#I/M**	**#Ant**
rs6857	APOE	19	N	Y	N	1	0	18	2	0	0	0	0	0	0	2	2	0
rs769449	PVRL2	19	N	N	N	3	1	11	2	0	0	0	0	0	0	0	0	0
rs3764814	USP42	7	N	N	N	4	0	0	0	0	0	3	0	-	-	-	-	-
rs7976168	TMTC2	12	N	N	N	16	0	0	0	0	0	2	0	0	0	0	0	0
rs7185375	SIAH1	16	N	N	N	1	0	0	0	0	0	0	0	0	0	0	0	0
rs2008465	GRHL1	2	N	N	N	2	0	0	0	0	0	0	0	0	0	0	0	0
rs72834698	HIST1H2BD	6	N	N	N	1	0	0	0	0	0	1	0	0	0	0	0	0
rs28391193	ELOVL6	4	N	N	N	33	0	0	0	0	0	0	0	132	0	0	0	0

Key: See Table 3. Pch: Y=Yes, N=No. TTD: Y=Yes, N=No. eQTL: Y=Yes, N=No.

Supplementary Table 4a. Human genetic information related to variants identified by Timmers et al. that exhibit statistically significant evidence for association with parental lifespan based on a standard GWAS.[19]

Associated Variant Information			Drug?		Annot.		Associations Involving Variants in LD with Target SNP							Chem Studies on Gene				
SNP	**Gene**	**Chrom**	**PCh**	**TTD**	**eQTL?**	**# LD**	**Long**	**Age Rel**	**Other**	**LD eQ**	**LD pQ**	**LD mQ**	**LD O**	**#Ch**	**#Ch A**	**#TTD**	**#I/M**	**#Ant**
rs1230666	MAGI3	1	E	N	N	1	0	1	2	0	0	0	0	22	0	-	-	-
rs1275922	KCNK3	2	Y	N	N	5	0	5	0	0	0	0	0	750	6	-	-	-
rs61348208	HTT	4	E	Y	N	19	1	0	0	0	0	3	1	19125	0	2	0	0
rs34967069	HLA-DQA1	6	N	E	N	33	1	1	0	0	0	1	2	0	0	-	-	-
rs10455872	LPA	6	N	Y	N	4	2	9	12	0	0	2	0	0	0	1	0	0
rs1556516	CDKN2B-AS1	9	N	N	Y	60	0	34	12	1	0	0	0	0	0	-	-	-
rs11065979	ATXN2/BRAP	12	E	N	Y	24	0	94	94	22	0	39	1	54410	0	-	-	-
rs8042849	CHRNA3/5	15	Y	N	Y	41	2	13	149	13	0	10	16	2381	6	-	-	-
rs6224	FURIN/FES (FES)	15	E	Y	Y	10	0	6	0	11	0	3	0	1835	0	1	2	0
rs12924886	HP	16	N	Y	Y	7	0	0	14	65	0	2	75	0	0	1	0	0
rs142158911	LDLR	19	E	Y	N	52	0	7	24	0	0	0	0	185	0	1	1	0
rs429358	APOE	19	N	Y	N	1	2	9	37	0	0	1	0	0	0	2	0	0

Key: See Table 3. Drug?=Druggable. Annot.=Annotations. Pch: E=Experimental. TTD: Y=Yes, N=No. eQTL?: Y=Yes, N=No.

Supplementary Table 4b. Human genetic information related to variants identified by Timmers et al. that exhibit statistically significant evidence for association with parental lifespan based on a Bayesian GWAS using mortality risk factors.[19]

Associated Variant Information			Drug?		Annot.		Associations Involving Variants in LD with Target SNP							Chem Studies on Gene			
SNP	Gene	Chrom	PCh	TTD	eQTL?	# LD	Long	Age Rel	Other	LD eQ	LD pQ	LD mQ	LD O	#Ch	#Ch A	#TTD	#I/M
rs4970836	CELSR2/PSRC1	1	N	N	Y	11	0	16	56	52	0	2	11	0	0	-	-
rs6744653	TMEM18	2	N	N	Y	231	0	82	7	1	0	34	0	0	0	-	-
rs10211471	GBX2/ASB18	2	N	N	N	9	0	0	0	0	0	16	0	0	0	-	-
rs111333005	IGF2R	6	Y	Y	Y	33	0	0	0	6	0	0	0	20	0	3	1
rs113160991	POM121C	7	N	N	Y	13	0	0	0	30	0	1	15	0	0	-	-
rs56179563	ZC3HC1	7	N	N	N	2	0	14	3	0	0	2	0	0	0	-	-
rs2519093	ABO	9	N	N	Y	13	0	19	112	3	0	9	0	0	0	-	-

Key: See Table 3. Drug?=Druggable. Annot.=Annotations. Pch: E=Experimental. TTD: Y=Yes, N=No. eQTL?: Y=Yes, N=No.

Supplementary Table 5. Human genetic information related to variants that exhibit statistically significant evidence for association with healthspan based on an analysis of by Zenin et al.[20] and Hornstrup et al.[21]

Associated Variant Information			Drug?		Annot.		Associations Involving Variants in LD with Target SNP							Chem Studies on Gene				
SNP	**Gene**	**Chrom**	**PCh**	**TTD**	**eQTL?**	**# LD**	**Long**	**Age Rel**	**Other**	**LD eQ**	**LD pQ**	**LD mQ**	**LD O**	**# PCh**	**# PCh A**	**# TTD**	**# I/M**	**# Ant**
rs10197246	ALS2CR12	2	N	-	Y	20	0	4	2	22	0	0	15	0	0	-	-	-
rs12203592	IRF4	6	N	N	Y	1	0	4	22	13	0	2	9	0	0	0	0	0
rs1049053	HLA-DQB1	6	N	N	Y	73	0	2	3	42	6	15	22	0	0	0	0	0
rs10455872	LPA	6	N	Y	N	4	2	9	12	0	0	2	0	0	0	1	1	0
rs140570886	LPA	6	N	Y	N	9	0	2	2	0	0	0	0	0	0	1	1	0
rs7859727	CDKN2B-AS1	9	N	N	Y	60	0	35	10	1	0	0	0	0	0	0	0	0
rs34872471	TCF7L2	10	E	N	N	11	0	67	10	0	0	1	0	637	0	0	0	0
rs2860197	FGFR2	10	Y	Y	N	28	0	18	0	0	0	10	0	2953	16	9	11	1
rs1126809	TYR	11	Y	Y	N	5	0	6	9	0	0	0	0	352	2	8	105	0
rs4784227	CASC16	16	N	N	Y	12	0	8	0	1	0	0	0	0	0	0	0	0
rs4268748	DEF8	16	N	N	Y	1	0	1	3	1	0	0	0	0	0	0	0	0
rs159428	NOL4L	20	N	N	Y	1	0	0	0	1	0	0	2	0	0	0	0	0
rs28933981	TTR	18	Y	N	N	2	0	0	0	0	0	0	0	2	0	0	0	0

Key: See Table 3. Drug?=Druggable. Annot.=Annotations. Pch: E=Experimental. TTD: Y=Yes, N=No, “ -“ =Undefined in database. eQTL?: Y=Yes, N=No.

Supplementary Table 6. Variant effect annotations and drug target information on genes harboring multiple rare variants that exhibit statistically significant evidence that they contribute to protective effects against disease as reviewed by Harper et al.[22]

Associated Gene Information			Drug?		Annot.		Associations Involving Variants in LD with Target Gene eQTLs							Chem Studies on Gene				
Gene	Refs	Function	PCh	TTD	eQTL?	# LD	Long	Age Rel	Other	LD eQ	LD pQ	LD mQ	LD O	# PCh	# PCh A	# TTD	# I/M	# Ant
PCSK9	23	Decreased LDL	Y	Y	12	151	0	4	6	52	0	17	19	188	5	7	1	1
LPA	24	Reduced CVD risk	N	Y	10	105	0	7	2	42	0	5	21	0	0	1	1	0
APOC3	25	Reduced triglycerides	N	Y	6	53	0	0	0	6	0	2	0	0	0	1	1	0
NPC1L1	26	Reduced CAD risk	Y	Y	13	433	0	0	4	174	0	21	37	318	1	1	0	2
SLC30A8	27	Reduced T2D risk	N	N	9	146	0	4	3	23	0	3	1	0	0	0	0	0

Key: See Table 3; Function = favorable phenotype associated with rare variants in the gene. Drug?=Druggable. Annot.=Annotations. Pch: E=Experimental. TTD: Y=Yes, N=No

Supplementary Internet References

Drug	Internet link
NMN	https://www.leafscience.org/dna-repair/
TA-65	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5178008/
Doxycycline	https://www.salk.edu/news-release/turning-back-time-salk-scientists-reverse-signs-aging/
CPHPC	https://www.ncbi.nlm.nih.gov/pubmed/26225229
Rosiglitazone	https://www.ncbi.nlm.nih.gov/pubmed/?term=PMID%3A+16123266
J147 (circumin)	https://www.salk.edu/news-release/alzheimers-drug-turns-back-clock-powerhouse-cell/
NPT088	http://www.proclarabio.com/our-programs/
UBX0101	http://ir.unitybiotechnology.com/news-releases/news-release-details/unity-biotechnology-expands-ongoing-ubx0101-phase-1-study
Quercetin	https://www.scripps.edu/news-and-events/press-room/2015/20150309agingcell.html
Dasatinib	https://www.scripps.edu/news-and-events/press-room/2015/20150309agingcell.html
Fisetin	https://www.leafscience.org/fisetin-may-be-a-low-hanging-fruit-for-aging/
UBX1967	http://ir.unitybiotechnology.com/news-releases/news-release-details/unity-biotechnology-announces-completion-ubx1967-license-and
Alk5i	https://www.leafscience.org/conboy-interview/
RTB101	https://www.restorbio.com/about-torc1
SRK-015	https://scholarrock.com/pipeline/srk-015-for-sma/intro/
SM04690	https://www.globenewswire.com/news-release/2019/01/29/1706787/0/en/Samumed-to-Present-Novel-Biological-Targets-of-SM04690-for-Treatment-of-Knee-Osteoarthritis.html
MSI-1436	https://www.eurekalert.org/pub_releases/2017-09/mdib-nbr090717.php

REFERENCES (APPENDIX C)

1. Li, J., et al., *A conserved NAD(+) binding pocket that regulates protein-protein interactions during aging.* Science, 2017. **355**(6331): p. 1312-1317. doi: 10.1126/science.aad8242.
2. Salvador, L., et al., *A Natural Product Telomerase Activator Lengthens Telomeres in Humans: A Randomized, Double Blind, and Placebo Controlled Study.* Rejuvenation Res, 2016. **19**(6): p. 478-484. doi: 10.1089/rej.2015.1793.
3. Ocampo, A., et al., *In Vivo Amelioration of Age-Associated Hallmarks by Partial Reprogramming.* Cell, 2016. **167**(7): p. 1719-1733 e12. doi: 10.1016/j.cell.2016.11.052
4. Pepys, M.B., et al., *Targeted pharmacological depletion of serum amyloid P component for treatment of human amyloidosis.* Nature, 2002. **417**(6886): p. 254-9. doi: 10.1038/417254a.
5. Kurosu, H., et al., *Suppression of aging in mice by the hormone Klotho.* Science, 2005. **309**(5742): p. 1829-33. doi: 10.1126/science.1112766.
6. Goldberg, J., et al., *The mitochondrial ATP synthase is a shared drug target for aging and dementia.* Aging Cell, 2018. **17**(2). doi: 10.1111/acel.12715.
7. Levenson, J.M., et al., *NPT088 reduces both amyloid-beta and tau pathologies in transgenic mice.* Alzheimers Dement (N Y), 2016. **2**(3): p. 141-155. doi: 10.1016/j.trci.2016.06.004.
8. Baker, D.J., et al., *Clearance of p16Ink4a-positive senescent cells delays ageing-associated disorders.* Nature, 2011. **479**(7372): p. 232-6. doi: 10.1038/nature10600.
9. Zhu, Y., et al., *The Achilles' heel of senescent cells: from transcriptome to senolytic drugs.* Aging Cell, 2015. **14**(4): p. 644-58. doi: 10.1111/acel.12344.
10. Yousefzadeh, M.J., et al., *Fisetin is a senotherapeutic that extends health and lifespan.* EBioMedicine, 2018. **36**: p. 18-28. doi: 10.1016/j.ebiom.2018.09.015.
11. Wu, D. and C. Prives, *Relevance of the p53-MDM2 axis to aging.* Cell Death Differ, 2018. **25**(1): p. 169-179. doi: 10.1038/cdd.2017.187.
12. Yousef, H., et al., *Systemic attenuation of the TGF-beta pathway by a single drug simultaneously rejuvenates hippocampal neurogenesis and myogenesis in the same old mammal.* Oncotarget, 2015. **6**(14): p. 11959-78. doi: 10.18632/oncotarget.3851.
13. Mannick, J.B., et al., *TORC1 inhibition enhances immune function and reduces infections in the elderly.* Sci Transl Med, 2018. **10**(449). doi: 10.1126/scitranslmed.aaq1564.
14. Pirruccello-Straub, M., et al., *Blocking extracellular activation of myostatin as a strategy for treating muscle wasting.* Sci Rep, 2018. **8**(1): p. 2292. doi: 10.1038/s41598-018-20524-9.
15. Yazici, Y., et al., *A novel Wnt pathway inhibitor, SM04690, for the treatment of moderate to severe osteoarthritis of the knee: results of a 24-week, randomized, controlled, phase 1 study.* Osteoarthritis Cartilage, 2017. **25**(10): p. 1598-1606. doi: 10.1016/j.joca.2017.07.006.
16. Smith, A.M., et al., *The protein tyrosine phosphatase 1B inhibitor MSI-1436 stimulates regeneration of heart and multiple other tissues.* NPJ Regen Med, 2017. **2**: p. 4. doi: 10.1038/s41536-017-0008-1.
17. Fuentealba, M., et al., *Using the drug-protein interactome to identify anti-ageing compounds for humans.* PLoS Comput Biol, 2019. **15**(1): p. e1006639. doi: 10.1371/journal.pcbi.1006639.
18. Sebastiani, P., et al., *Four Genome-Wide Association Studies Identify New Extreme Longevity Variants.* J Gerontol A Biol Sci Med Sci, 2017. **72**(11): p. 1453-1464. doi: 10.1093/gerona/glx027.

19. Timmers, P.R., et al., *Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances.* Elife, 2019. **8**. doi: 10.7554/eLife.39856.

20. Zenin, A., et al., *Identification of 12 genetic loci associated with human healthspan.* Commun Biol, 2019. **2**: p. 41. doi: 10.1038/s42003-019-0290-0.

21. Hornstrup, L.S., et al., *Genetic stabilization of transthyretin, cerebrovascular disease, and life expectancy.* Arterioscler Thromb Vasc Biol, 2013. **33**(6): p. 1441-7. doi: 10.1161/ATVBAHA.113.301273.

22. Harper, A.R., S. Nayee, and E.J. Topol, *Protective alleles and modifier variants in human health and disease.* Nat Rev Genet, 2015. **16**(12): p. 689-701. doi: 10.1038/nrg4017.

23. Cohen, J.C., et al., *Sequence variations in PCSK9, low LDL, and protection against coronary heart disease.* N Engl J Med, 2006. **354**(12): p. 1264-72. doi: 10.1056/NEJMoa054013.

24. Lim, E.T., et al., *Distribution and medical impact of loss-of-function variants in the Finnish founder population.* PLoS Genet, 2014. **10**(7): p. e1004494. doi: 10.1371/journal.pgen.1004494.

25. Tg, et al., *Loss-of-function mutations in APOC3, triglycerides, and coronary disease.* N Engl J Med, 2014. **371**(1): p. 22-31. doi: 10.1056/NEJMoa1307095.

26. Myocardial Infarction Genetics Consortium, I., et al., *Inactivating mutations in NPC1L1 and protection from coronary heart disease.* N Engl J Med, 2014. **371**(22): p. 2072-82. doi: 10.1056/NEJMoa1405386.

27. Flannick, J., et al., *Loss-of-function mutations in SLC30A8 protect against type 2 diabetes.* Nat Genet, 2014. **46**(4): p. 357-63. doi: 10.1038/ng.2915.

REFERENCES

1.1.1 Makin S. The emerging world of digital therapeutics. *Nature.* 2019 Sep; 573(7775): S106-S109. DOI: 10.1038/d41586-019-02873-1.

1.1.2. Nathanson, L, Rivers SE, Flynn LM, Brackett MA. Creating emotionally intelligent schools with RULER. *Emotion Review*. 2016; 8(4): 1-6.

1.1.3. Thayer RE. Toward a psychological theory of multidimensional activation (arousal). *Motivation and Emotion.* 1978; 2(1): 1–34. DOI: 10.1007/BF00992729

1.1.4. Lövheim, H. A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical hypotheses.* 2012; 78(2): 341-8. DOI: 10.1016/j.mehy.2011.11.016.

1.1.5 Schnorr S, Candela, M, Rampelli S, Centanni M, Consolandi C, Basaglia G, Turroni S, Biagi E, Peano C, Severgnini M, Fiori J, Gotti R, De Bellis G, Luiselli D, Brigidi P, Mabulla A, Marlowe F, Henry AG, Crittenden AN. Gut microbiome of the Hadza hunter-gatherers. *Nature Communications.* 2014 Apr 15; 5: 3654. DOI: 10.1038/ncomms4654

1.1.6 Gomez A, Sharma AK, Mallott EK, Petrzelkova KJ, Robinson CAJ, Yeoman CJ, Carbonero F, Pafco B, Rothman JM, Ulanov A, Vlckova K, Amato KR, Schnorr SL, Dominy NJ, Modry D, Todd A, Torralba M, Nelson KE, Burns MB, Blekhman R, Remis M, Stumpf RM, Wilson BA, Gaskins HR, Garber PA, White BA, Leigh SR. Plasticity in the Human gut microbiome defies evolutionary constraints. *mSphere*. 2019 Jul; 4(4) e00271-19; DOI: 10.1128/mSphere.00271-19

1.1.7. Gaulke CA, Sharpton TJ. The influence of ethnicity and geography on human gut microbiome composition. *Nature Medicine.* 2018; 24(10): 1495–1496. DOI: 10.1038/s41591-018-0210-8

1.1.8. Guinane CM, Cotter PD. Role of the gut microbiota in health and chronic gastrointestinal disease: Understanding a hidden metabolic organ. *Therapeutic Advances in Gastroenterology.* 2013; 6(4): 295-308. DOI: 10.1177/1756283X13482996

1.1.9. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner F, Pedersen O, de Vos WM, Brunak S, Doré J; MetaHIT Consortium, Antolín M, Artiguenave F, Blottiere HM, Almeida M, Brechot C, Cara C, Chervaux C, Cultrone A, Delorme C, Denariaz G, Dervyn R, Foerstner KU, Friss C, van de Guchte M, Guedon E, Haimet F, Huber W, van Hylckama-Vlieg J, Jamet A, Juste C, Kaci G, Knol J, Lakhdari O, Layec S, Le Roux K, Maguin E, Mérieux A, Melo Minardi R, M'rini C, Muller J, Oozeer R, Parkhill J, Renault P, Rescigno M, Sanchez N, Sunagawa S, Torrejon A, Turner K, Vandemeulebrouck G, Varela E, Winogradsky Y, Zeller G, Weissenbach J, Ehrlich SD, Bork P. Enterotypes of the human gut microbiome. *Nature.* 2011 May 12; 473**:** 174–180. DOI: 10.1038/nature099441.1.8

1.1.10. Bray, JR, Curtis JT. An ordination of upland forest communities of southern Wisconsin. *Ecological Monographs.* 1957; 27(4): 325-349.

1.1.11. Borg I, Groenen P. Modern Multidimensional Scaling: Theory and applications (2nd ed.). New York: Springer-Verlag; 2005: 207–212. ISBN 978-0-387-94845-4.

1.1.12. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. Cluster: Cluster Analysis Basics and Extensions. *R package version 2.1.0.* 2019.

2.1.1. Lasken RS. Genomic sequencing of uncultured microorganisms from single cells. *Nature Reviews Microbiology.* 2012; 10(9): 631–640.

2.1.2. Binga, EK, Lasken RS. and Neufeld LD. Something from (almost) nothing: The impact of multiple displacement amplification on microbial ecology. *The ISME Journal.* 2008; 2(3): 233-241.

2.1.3. Lasken RS. Genomic DNA amplification by the multiple displacement amplification (MDA) method. *Biochemical Soc Trans*. 2009 Apr 3; 37: 450-453.

2.1.4. McCorrison JM, Venepally P, Singh I, Fouts DE, Lasken RS, Methé BA. NeatFreq: reference-free data reduction and coverage normalization for De Novosequence assembly. *BMC Bioinformatics*. 2014; 15(1): 357. DOI: 10.1186/s12859-014-0357-3.

2.1.5. Chitsaz H, Yee-Greenbaum JL, Tesler G, Lombardo MJ, Dupont CL, Badger JH, Novotny M, Rusch DB, Fraser LJ, Gormley NA, Schulz-Trieglaff O, Smith GP, Evers DJ, Pevzner PA, Lasken RS. Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nature Biotechnology*. 2011; 29(10): 915–921 DOI: 10.1038/nbt.1966.

2.1.6. Hutchison III CA, Smith HO, Pfannkoch C, Venter JC. Cell-free cloning using φ29 DNA polymerase. *Proceedings of the National Academy of Sciences*. 2005; 102(48), 17332–17336.

2.1.7. Marcy Y, Ishoey T, Lasken RS, Stockwell TB, Walenz BP, Halpern AL, Beeson KY, Goldberg SMD, Quake SR. Nanoliter reactors improve multiple displacement amplification of genomes from single cells. *PLoS Genetics.* 2007; 3(9), e155.

2.1.8. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*. 2014; 9(1), 171-181

2.1.9. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*. 2013; *10*(11), 1096-1098.

2.1.10. Hawrylycz M, Miller JA, Menon V, Feng D, Dolbeare T, Guillozet-Bongaarts AL, Jegga AG, Aronow BJ, Lee CK, Bernard A, Glasser MF, Dierker DL, Menche J, Szafer A, Collman F, Grange P, Berman KA, Mihalas S, Yao Z, Stewart L, Barabási AL, Schulkin J, Phillips J, Ng L, Dang C, Haynor DR, Jones A, Van Essen DC, Koch C, Lein E. Canonical genetic signatures of the adult human brain. *Nature Neuroscience.* 2015 Dec; 18(12): 1832-44. DOI: 10.1038/nn.4171.

2.1.11. Kuan L, Yang L, Lau C, Feng D, Bernard A, Sunkin SM, Zeng H, Dang C, Hawrylycz M, Ng L. Neuroinformatics of the Allen Mouse Brain Connectivity Atlas. *Methods*. 2015; 73: 4-17. DOI: 10.1016/j.ymeth.2014.12.013.

2.1.12. Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, Levi B, Gray LT, Sorensen SA, Dolbeare T, Bertagnolli D, Goldy J, Shapovalova N, Parry S, Lee C, Smith K, Bernard A, Madisen L, Sunkin SM, Hawrylycz M, Koch C, Zeng H. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience.* 2016; 9: 335–346.

2.1.13. Lacar B, Linker SB, Jaeger BN, Krishnaswami S, Barron JJ, Kelder MJE, Parylak S, Paquola ACM, Venepally P, Novotny M, O'Connor C, Fitzpatrick C, Erwin JA, Hsu JY, Husband D, McConnell MJ, Lasken R, Gage FH. Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nature Communications*. 2016; 7(1): 1-13. DOI: 10.1038/ncomms11022.

2.1.14. Španić E, Langer Horvat L, Hof PR, Šimić G. Role of microglial cells in alzheimer's disease tau propagation. *Frontiers in Aging Neuroscience.* 2019 Oct 4; 11: 271. DOI: 10.3389/fnagi.2019.00271

2.1.15. Molnár Z, Kaas JH, De Carlos JA, Hevner RF, Lein E, Němec P. Evolution and development of the mammalian cerebral cortex. *Brain, Behavior and Evolution*. 2014; 83(2): 126-139

2.1.16. Guillozet-Bongaarts AL, Hyde TM, Dalley RA, Hawrylycz MJ, Henry A, Hof PR, Hohmann J, Jones AR, Kuan CL, Royall J, Shen E, Swanson B, Zeng H, Kleinman JE. Altered gene expression in the dorsolateral prefrontal cortex of individuals with schizophrenia. *Molecular Psychiatry.* 2014 Apr; 19(4): 478–485.

2.1.17. Krishnaswami SR, Grindberg RV, Novotny M, Venepally P, Lacar B, Bhutani K, Linker SB, Pham S, Erwin JA, Miller JA, Hodge R, McCarthy JK, Kelder M, McCorrison JM, Aevermann BD, Diez Fuertes F, Scheuermann RH, Lee J, Lein ES, Schork N, McConnell MJ, Gage FH, Lasken RS. Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. *Nature Protocols*. 2016 Mar; 11(3): 499-524. DOI: 10.1038/nprot.2016.015

2.3.1. Duran RCD, Wei H, Wu JQ. Single-cell RNA-sequencing of the brain. *Clinical and Translational Medicine.* 2017; 6(1): 20. DOI: 10.1186/s40169-017-0150-9

2.3.2. Zhang Q, He Y, Luo N, Patel SJ, Han Y, Gao R, Modak M, Carotta S, Haslinger C, Kind D, Peet GW, Zhong G, Lu S, Zhu W, Mao Y, Xiao M, Bergmann M, Hu X, Kerkar SP, Vogt AB, Pflanz S, Liu K, Peng J, Ren X, Zhang Z. Landscape and Dynamics of Single Immune Cells in Hepatocellular Carcinoma. *Cell.* 2019 Oct 31; 179(4): 829-845. DOI: 10.1016/j.cell.2019.10.003.

2.4.1. Picelli S, Björklund AK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for senstitive full-length transcriptome profiling in single cells. *Nature Methods*. 2013; 10(11): 1096–1098.

2.4.2. Aevermann B, McCorrison JM, Venepally P, Hodge R, Bakken, T, Miller J, Novotny M, Tran DN, Diezfuertes F, Christiansen L, Zhang F, Steemers F, Lasken RS, Lein ED, Schork N, Scheuermann RH. Production of a preliminary quality control pipeline for single nuclei Rna-Seq and its application in the analysis of cell type diversity of post-mortem human brain neocortex. In *Pacific Symposium on Biocomputing.* 2017; 564-575.

2.4.3. Krishnaswami SR, Grindberg RV, Novotny M, Venepally P, Lacar B, Bhutani K, Linker SB, Pham S, Erwin JA, Miller JA, Hodge R, McCarthy JK, Kelder M, McCorrison JM, Aevermann BD, Diez Fuertes F, Scheuermann RH, Lee J, Lein ES, Schork N, McConnell MJ, Gage FH, Lasken RS. Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. *Nature Protocols*. 2016 Mar; 11(3): 499-524. DOI: 10.1038/nprot.2016.015

2.4.4. Haghverdi L, Lun AT, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 2018; 36(5): 421-427.

2.4.5. Choi J, Pacheco CM, Mosbergen R, Korn O, Chen T, Nagpal I, Englart S, Angel PW, Wells CA. Stemformatics: visualize and download curated stem cell data. *Nucleic Acids Research.* 2019 Jan 8; 47(D1): D841-D846. DOI: 10.1093/nar/gky1064. PMID: 30407577]

2.4.6. Tian L, Su S, Dong X, Amann-Zalcenstein D, Biben C, Seidi A, Hilton DJ, Naik SH, Ritchie ME. scPipe: A flexible R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data. *PLoS Computational Biology.* 2018 Aug 10; 14(8): e1006361. DOI: 10.1371/journal.pcbi.1006361.

2.4.7. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology.* 2018; 36(5): 411-420.

2.4.8. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, Linsley PS, Gottardo R. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology.* 2015; 16(1): 278.

2.4.9. Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, Frey B, Irimia M, Blencowe BJ. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome research*, 2014 Nov; 24(11), 1774-1786.

2.4.10. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck III WM, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive integration of single-cell data. *Cell.* 2019; 177(7): 1888-1902. DOI: 10.1016/j.cell.2019.05.031.

2.4.11. Lasken RS. Genomic sequencing of uncultured microorganisms from single cells. *Nature Reviews Microbiology.* 2012; 10(9): 631–640.

2.4.12. Lasken RS. Genomic DNA amplification by the multiple displacement amplification (MDA) method. *Biochem Soc Trans*. 2009; 37: 450–453.

2.4.13. McCorrison JM, Venepally P, Singh I, Fouts DE, Lasken RS, Methé BA. NeatFreq: reference-free data reduction and coverage normalization for De Novosequence assembly. *BMC Bioinformatics*. 2014; 15(1): 357. DOI: 10.1186/s12859-014-0357-3.

2.4.14. Chitsaz H, Yee-Greenbaum JL, Tesler G, Lombardo MJ, Dupont CL, Badger JH, Novotny M, Rusch DB, Fraser LJ, Gormley NA, Schulz-Trieglaff O, Smith GP, Evers DJ, Pevzner PA, Lasken RS. Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nature Biotechnology.* 2011; 29, 915–921. DOI: 10.1038/nbt.1966.

2.4.15. Hutchison III CA, Smith HO, Pfannkoch C, Venter JC. Cell-free cloning using φ29 DNA polymerase. *Proceedings of the National Academy of Sciences*. 2005; 102(48), 17332–17336.

2.4.16. Rangan AV, McGrouther CC, Kelsoe J, Schork N, Stahl E, et al. (2018) A loop-counting method for covariate-corrected low-rank biclustering of gene-expression and genome-wide association study data. *PLOS Computational Biology*. 2018; 14(5): e1006105. DOI: 10.1371/journal.pcbi.1006105

2.4.17. Deshpande Y, Montanari A. Finding Hidden Cliques of Size $$\sqrt{N/e} in Nearly Linear Time. *Foundations of Computational Mathematics.* 2013 Apr 26; 15(4): 1069-1128.

2.4.18. Alon N, Krivelevich M, and Sudakov B. Finding a large hidden clique in a random graph.

Random Structures and Algorithms. 1998 Dec 07; 13(3-4): 457-466.

2.4.19. Shabalin AA, Weigman VJ, Perou CM, Nobel AB. Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics.* 2009 May 11; 3(3): 985-1012.

2.4.20. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*. 2014; 30(15): 2114-2120.

2.4.21. FastQC. [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/].

2.4.22. DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics.* 2012; 28(11): 1530-1532.

2.4.23. Baik J, Arous GB, Péché S. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*. 2005; *33*(5): 1643-1697.

2.4.24. Chung JE, Magland JF, Barnett AH, Tolosa VM, Tooker AC, Lee KY, Shah KG, Felix SH, Frank LM, Greengard LF. A fully automated approach to spike sorting. *Neuron.* 2017; 95(6) 1381–1394.

2.4.25. Linderman GC, Rachh M, Hoskins JG, Steinerberger S, Kluger Y. Efficient Algorithms for t-distributed Stochastic Neighborhood Embedding. *arXiv preprint: 1712.09005.* 25 Dec 2017.

2.4.26. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, Ginhoux F, Newell. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology.* 2019; 37(1): 38–44. DOI: 10.1038/nbt.4314.

2.4.27. McInnes L, Healy J, Astels, S. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software.* 2017; 2(11): 205. DOI: 10.21105/joss.00205

2.4.28. Mahalanobis PC. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India.* 1936; 2(1): 49–55.

2.4.29. Suzuki R, Shimodaira H. Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics.* 2006 Jun 15; 22(12): 1540–1542. DOI: 10.1093/bioinformatics/btl117

2.4.30. Mishra A, Dey DK, Chen K. Sequential co-sparse factor regression. *Journal of Computational Graphical Statistics.* 2017; 26(4): 814–825.

3.2.1. Soto-Gamez A, Demaria M. Therapeutic interventions for aging: the case of cellular senescence. *Drug Discovery Today*. 2017; 22.5: 786-795.

3.2.2. Sebastiani P, Perls TT. The genetics of extreme longevity: Lessons from the new England centenarian study. *Frontiers in Genetics*. 2012; *3*: 277.

3.2.3. Kitani K, Minami C, Yamamoto T, Kanai S, Ivy GO, Carrillo MC. Pharmacological interventions in aging and age-associated disorders: potentials of propargylamines for human use. *Annals of the New York Academy of Sciences.* 2002 Apr; 959(1): 295-307.

3.2.4. Sebastiani P, Gurinovich A, Nygaard M, Sasaki T, Sweigart B, Bae H, Andersen SL, Villa F, Atzmon G, Kaare C, Yasumichi A, Barzilai N, Puca A, Christiansen L, Hirose N, Perls TT. APOE alleles and extreme human longevity. *The Journals of Gerontology: Series A.* 2019 Jan; 74(1): 44-51. DOI: 10.1093/gerona/gly174

3.2.5. Sebastiani P, Solovieff N, Puca A, Hartley SW, Melista E, Andersen S, Dworkis DA, Wilk JB, Myers RH, Steinberg, MH, Montano, M, Baldwin CT, Perls TT. Genetic signatures of exceptional longevity in humans. *Science.* 2010 Jul 01. DOI:10.1126/science.1190532.

3.2.6. Partridge L., Deelen J, Slagboom, PE. Facing up to the global challenges of ageing. *Nature.* 2018; 561(7721): 45-56. DOI: 10.1038/s41586-018-0457-8.

3.2.7. Kaya A, Ma S, Wasko B, Lee M, Kaeberlein M, Gladyshev VN. Defining molecular basis for longevity traits in natural yeast isolates. *NPJ Aging and Mechanisms of Disease.* 2015; 1(1): 1-9.

3.2.8. Uno M, Nishida E. Lifespan-regulating genes in C. elegans. *NPJ aging and mechanisms of disease.* 2016; 2(1): 1-8.

3.2.9. Spencer CC, Howell CE, Wright AR, Promislow DE. Testing an 'aging gene' in long-lived drosophila strains: increased longevity depends on sex and genetic background. *Aging Cell.* 2003; 2(2): 123-130.

3.2.10. Sahm A, Bens M, Szafranski K, Holtze S, Groth M, Görlach M, Calkhoven C, Muller C, Schwab M, Kraus J, Kestler HA, Cellerino A, Burda H, Hildebrandt T, Dammann P, Platzer M. Long-lived rodents reveal signatures of positive selection in genes associated with lifespan. *PLoS Genetics.* 2018; 14(3): e1007272.

3.2.11. Wirthlin M, Lima NCB, Guedes RLM, Soares AER, Almeida LGP, Cavaleiro NP, Morais GLD, Chaves AV, Howard JT, Teixeira MDM, Schneider PN, Santos FR, Schatz MC, Felipe MS, Miyaki CY, Alexio A, Schneider MPC, Jarvis ED, Mello CV. Parrot genomes and the evolution of heightened longevity and cognition. *Cell.* 2017 Dec 17; 28(24): 4001-4008. DOI: 10.1016/j.cub.2018.10.050

3.2.12. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003; 13(9): 2178-2189. doi:10.1101/gr.1224503

3.2.13. al. ASem (2020). DescTools: Tools for Descriptive Statistics. R package version 0.99.35, https://cran.r-project.org/package=DescTools.

3.2.12. Felsenstein J. Phylogenies and the Comparative Method. *The American Naturalist*. 1985; 125(1): 1-15.

3.2.13. Zapala MA, Schork NJ. Statistical properties of multivariate distance matrix regression for high-dimensional data analysis. *Frontiers in Genetics*. 2012; 3:190. DOI: 10.3389/fgene.2012.00190.

3.2.14. Zapala MA, Schork NJ. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proceedings of the National Academy of Sciences of the United States of America*. 2006 Dec; 103(51): 19430-19435. DOI: 10.1073/pnas.0609333103.

3.2.15. Nievergelt CM, Libiger O, Schork NJ. Generalized analysis of molecular variance. PLoS Genet. 2007; 3(4): e51. DOI: 10.1371/journal.pgen.0030051

3.2.16. Tacutu R, Thornton D, Johnson E, Budovsky A, Barardo D, Craig T, Diana E, Lehmann G, Toren D, Wang J, Fraifeld VE, de Magalhaes JP. Human ageing genomic resources: New and updated databases. *Nucleic Acids Research.* 2018; 46(D1): D1083-D1090.

3.2.17. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic acids research*. 2019 Jan 08; *47*(D1): D807-D811. DOI: 10.1093/nar/gky1053

3.2.18. Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, Pearse W. Caper: comparative analyses of phylogenetics and evolution in R. R package version 0.5. 2012.

3.2.19. Krämer A, Green J, Pollard Jr J, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics*. 2014; 30(4): 523-530. DOI:10.1093/bioinformatics/btt703

3.2.20. Sneddon TP, Li P, Edmunds, SC. GigaDB: announcing the GigaScience database. *GigaScience*. 2012; 1(1): 2047-217X. DOI: 10.1186/2047-217X-1-11.

3.2.21. McKinney C, Yu D, Mohr I. A new role for the cellular PABP repressor Paip2 as an innate restriction factor capable of limiting productive cytomegalovirus replication. *Genes & Development*. 2013; 27(16): 1809-1820. DOI:10.1101/gad.221341.113.

3.2.22. Hirata T, Fujita M, Nakamura S, Gotoh K, Motooka D, Murakami Y, Maeda Y, Kinoshita T. Post-Golgi anterograde transport requires GARP-dependent endosome-to-TGN retrograde transport. *Molecular Biology of the Cell.* 2015; 26(17): 3071-3084.

3.2.23. Serrat R, Mirra, S, Figueiro-Silva J, Navas-Perez E, Quevedo M, Lopez-Domenech G, Podlesniy P, Ulloa F, Garcia-Fernandez J, Trullas R, Soriana E. The Armc10/SVH gene: genome context, regulation of mitochondrial dynamics and protection against Aβ-induced mitochondrial fragmentation. Cell Death & Disease. 2014; 5(4): e1163. DOI: 10.1038/cddis.2014.121

3.2.24. Lai AY, Wade PA. Cancer biology and NuRD: a multifaceted chromatin remodeling complex. *Nature Reviews Cancer*. 2011 Jul 7; 11(8): 588-596. DOI:10.1038/nrc3091

3.2.25. Alver BH, Kim KH, Lu P, Wang X, Manchester HE, Wang W, Haswell JR, Park PJ. The SWI/SNF chromatin remodelling complex is required for maintenance of lineage specific enhancers. *Nature Communications.* 2017 Mar 06; 8(1): 1-10. DOI: 10.1038/ncomms14648

3.2.26. Längst G, Manelyte L. Chromatin remodelers: From function to dysfunction. *Genes.* 2015 Jun 12; 6(2): 299-324. DOI:10.3390/genes6020299

3.2.27. Hakim O, Resch W, Yamane A, Klein I, Kieffer-Kwon KR, Jankovic M, Oliveira T, Bothmer A, Voss TC, Ansarah-Sobrinho C, Mathe E, Liang G, Cobell J, Nakahashi H, Robbiani DF, Nussenzweig A, Hager GL, Nussenzweig MC, Casellas R. DNA damage defines sites of recurrent chromosomal translocations in B lymphocytes. *Nature.* 2012 Feb 07; 484(7392): 69-74. DOI: 10.1038/nature10909.

3.2.28. Garschall K, Flatt T. The interplay between immunity and aging in Drosophila. *F1000Research.* 2018 Feb 07; 7:160. DOI:10.12688/f1000research.13117.1

3.2.29. Bharath LP, Ip BC, Nikolajczyk BS. Adaptive immunity and metabolic health: Harmony becomes dissonant in obesity and aging. *Comprehensive Physiology*. 2017 Sep 12; 7(4): 1307-1337. DOI: 10.1002/cphy.c160042.

3.2.30. Harper JM, Min Wang, Galecki AT, Ro J, Williams JB, Miller RA. Fibroblasts from long-lived bird species are resistant to multiple forms of stress. *Journal of Experimental Biology.* 2011; 214(11): 1902-1910; DOI: 10.1242/jeb.054643.

3.2.31. Schuierer S, Carbone W, Knehr J, Petitjean V, Fernandez A, Sultan M, Roma, G. A comprehensive assessment of RNA-seq protocols for degraded and low-quantity samples. *BMC Genomics.* 2017 Jun 05; 18(1): 442. DOI:10.1186/s12864-017-3827-y

3.2.32. Pease J, Sooknanan R. A rapid, directional RNA-seq library preparation workflow for Illumina® sequencing. *Nature Methods.* 2012; 9(3): i–ii DOI: 10.1038/nmeth.f.355

3.2.33. Ondov, BD, Treangen TJ, Melsted P, Mallonee, AB, Bergman NH, Koren S, Phillippy

AM. Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology.* 2016; 17(1): 132. DOI: 10.1186/s13059-016-0997-x.

3.2.34. Rubolini D, Liker A, Garamszegi LZ, Møller AP, Saino N. Using the BirdTree.org website to obtain robust phylogenies for avian comparative studies: A primer. *Current Zoology*. 2015 Dec 01; 61(6): 959–965, DOI: 10.1093/czoolo/61.6.959

3.2.35. cophenetic.phylo [https://www.rdocumentation.org/packages/ape/versions/5.3/topics/cophenetic.phylo]

3.2.36. Jombart T, Dray S. Adephylo: Exploratory analyses for the phylogenetic comparative method. *Bioinformatics.* 2010; 26(15): 1-21.

3.2.37. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Research Notes.* 2016 Feb 12; 9(1): 88. DOI: 10.1186/s13104-016-1900-2.

3.2.38. Bolger, AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*. 2014; 30(15): 2114-2120.

3.2.39. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology.* 2019; 37(8): 907–915.

3.2.40. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*. 2019; 20(1): 1-13. DOI: 10.1186/s13059-019-1910-1.

3.2.41. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Palma FD, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev, A. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology.* 2011 May 15; 29(7): 644-652. DOI:10.1038/nbt.1883.

3.2.42. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods.* 2012; 9(4): 357-359.

3.2.43. Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011; 12(2): 323. DOI: 10.1186/1471-2105-12-323.

3.2.44. Bairoch A, Boeckmann B, Ferro S, Gasteiger E. Swiss-Prot: Juggling between evolution and stability. *Briefings in Bioinformatics.* 2004; 5(1): 39-55.

3.2.45. Altschul, SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of Molecular Biology*. 1990; 215(3): 403-410.

3.2.46. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics*. 2012; 28(23): 3150-3152. DOI: 10.1093/bioinformatics/bts565.

3.2.47. Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs and Helene Wagner (2019). vegan: Community Ecology Package. R package version 2.5-6. https://CRAN.R-project.org/package=vegan

3.2.48. Benjamini, Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1995; 57(1): 289-300.

3.2.49. Alexa A, Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. *R package version* 2. 32.0. 2016

3.2.50. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P, Flicek P. Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database*. 2011 Jul 23. DOI: 10.1093/database/bar030

3.2.51. Li X. ALL: A data package. R package version 1.29.0. 2019

3.2.52. Hansen KD, Gentry J, Long L, Gentleman R, Falcon S, Hahne F, Sarkar D (2020). Rgraphviz: Provides plotting capabilities for R graph objects. R package version 2.32.0.

5.1. Lasken RS. Genomic DNA amplification by the multiple displacement amplification (MDA) method. Biochem Soc Trans 1 April 2009; 37 (2): 450–453. DOI: 10.1042/BST0370450.

5.2. Okada S, Saiwai H, Kumamaru H, Kubota K, Harada A, Yamaguchi M, Iwamoto Y, Ohkawa Y. Flow cytometric sorting of neuronal and glial nuclei from central nervous system tissue. J Cell Physiol. 2011 Feb;226(2):552-8. DOI: 10.1002/jcp.22365.

5.3. Picelli S, Björklund AK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for senstitive full-length transcriptome profiling in single cells. Nature Methods. 2013; 10(11): 1096–1098.

5.4. van der Poel M, Ulas T, Mizee MR, Hsiao C, Miedema SSM, Adelia, Schuurman KG, Helder B, Tas SW, Schultze JL, Hamann J, Huitinga I. Transcriptional profiling of human microglia reveals grey–white matter heterogeneity and multiple sclerosis-associated changes. Nat Commun 10, 1139 (2019). DOI: 10.1038/s41467-019-08976-7.

5.5. Schiller HB, Montoro DT, Simon LM, Rawlins EL, Meyer KB, Strunz M, Vieira Braga FA, Timens W, Koppelman GH, Budinger GRS, Burgess JK, Waghray A, van den Berge M, Theis FJ, Regev A, Kaminski N, Rajagopal J, Teichmann SA, Misharin AV, Nawijn MC. The Human Lung Cell Atlas: A High-Resolution Reference Map of the Human Lung in Health and Disease. Am J Respir Cell Mol Biol. 2019 Jul;61(1):31-41. DOI: 10.1165/rcmb.2018-0416TR.