Title
Leveraging replicable sources of variability to increase power and interpretability in analyses of genomic datasets

Author
Thompson, Michael

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Leveraging replicable sources of variability to increase power and interpretability in analyses of genomic datasets

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

by

Michael James Thompson

2022

ABSTRACT OF THE DISSERTATION

Leveraging replicable sources of variability
to increase power and interpretability
in analyses of genomic datasets

by

Michael James Thompson
Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2022
Professor Eran Halperin, Chair

Many types of genomic datasets—including RNA sequencing (RNAseq) and DNA methylation—are influenced by innumerable sources of variability. Frequently, analyses of such variability focus on local effects due to genetics, often overlooking the components of variability related to context-level, individual-level, or environmental effects. Here, we leverage the idea that sources of variability are often conserved across genomic datasets to propose two approaches to partition variability: first into distinct biological and technical components, and second into orthogonal context-specific and context-shared genetic components. Using our methods, we perform more powerful and interpretable genomic association studies (such as transcriptome- or epigenome-wide association studies), and we uncover that heritability is more context-specific at the level of single-cell RNAseq, whereas it is more context-shared at the level of bulk (tissue) RNAseq. Subsequently, we perform an analysis of medical records to elucidate the informativeness and impacts of multiple genomics data types on phenotype imputation tasks. We show that risk scores derived from one's methylation are more informative than risk scores derived from one's genotypes in imputation tasks. The work presented here shows lasting impact on the design of multiple classes of genomic association studies as well as studies of the utility of genomic biomarkers in electronic medical records.

The dissertation of Michael James Thompson is approved.

Bogdan Pasaniuc

Matteo Pelligrini

Noah A. Zaitlen

Eran Halperin, Committee Chair

University of California, Los Angeles

2022

To my family,

for their cromulent encouragement and support,

and to Cara,

for reminding me life ain't chess.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

There are innumerable people who have helped me make it to this point. For the sake of brevity, this will not be exhaustive, and will more or less be written in chronological order of my academic career.

First, I need to thank my family for pushing me to follow my interests. I appreciate them advising me to pursue my passion toward computer science. Consequently, I must thank my undergraduate roommate, Justin Miller, for telling me that the Bioinformatics minor at UCLA existed, and that it enabled me to take the *fun* computer science courses that were otherwise unavailable to us in MIMG.

Of course, this minor would not have existed without Eleazar Eskin. Eleazar was a massive influence on my undergraduate (and therefore graduate) experience. Thank you for organizing this fantastic minor and collection of professors and resources. The minor not only helped me find science that I was genuinely interested in, but helped me find a group of individuals (mentioned below) who would become my life support throughout the degree.

Through the undergraduate bioinformatics research collective, I came to learn about the work of Tom Graeber. To Tom's surprise (and I assume many professor's worst nightmares) I showed up to his office unannounced, wearing business-causal clothes, holding a copy of my CV, and asking about his research and to join his lab. Despite this first impression, Tom was extremely welcoming and took time out of his day to explain the goals and projects in his lab as well as the people—including Niko Balanis—who were working on them. Niko, Tom, thank you for being fanastic, patient, and kind mentors throughout my undergraduate experience. You were very encouraging, and were the reasons I wanted to continue pursuing research. I am immensely grateful for your mentorship.

In terms of training me in statistics, I owe a shoutout to my favorite instructors at UCLA—Jessica Li, Sriram Sankararaman, Ying Nian Wu, and Mark Handcock. While I generally tried to take classes I had a reason to take, I was always inclined to take *any* topic taught by any of you. You are excellent instructors and do a fantastic job of explaining

otherwise very complex, hard-to-teach topics. While my teaching experience was limited to TAing, I am already practicing my teaching to hopefully reach a level as strong as yours in the future.

Nevertheless, my success in the classes taught by the above was due to my friendship with (dependence on) Alec Chiu, Brandon Jew, and Leah Briscoe. What began as small discussions in Chris Lee's course developed into great friendships, sleepless nights, and excellent food runs. Thank you for helping me throughout this. It's been amazing growing and working with you all, and I could not imagine a better group with whom to have gone through this all. Alec, your generosity in particular has been unmatched by anyone else I've come across throughout my time at UCLA. There is not a doubt in my mind that I wouldn't have made it this far without your friendship and support throughout these years, and I know that that's true of many others to whom you have graciously lent your brain-power.

Next, I'd like to thank Noah Zaitlen. You have been a fantastic mentor these last few years, and I look up to you as both a scientist and human being. You've not only taught me how to conduct extremely thorough, well thought-out science, but how to build and prioritize a culture of kindness when conducting science. I had almost enrolled at UCSF to join your lab, and it was a pleasant surprise when you moved to UCLA. Thank you for giving me the chance to work with you and Bruna on an extremely fulfilling project, and for continuing to challenge me with exciting problems. Seeing you two on Hollywood Boulevard after a concert at the bowl was quite possibly one of the most influential moments of my career—the point at which I learned that amazing scientists could also be cool, fun-loving human beings without having to choose one over the other.

Finally, this section could not be complete without a thank-you to Eran. Eran, your mentorship over the course of my PhD has been absolutely fantastic. When I first joined your lab, I was completely naive of the caliber of science that you were conducting, as well as the caliber of scientists you were training. I am very grateful that despite the fact that my initial abilities in statistics and science were quite limited, you not only had the patience to accept me into your lab and to teach me, but to surround me with the amazing guid-

ance and support that I needed (thank you to Liat Shenhav, Elior Rahmani, and Johnson Chen). I am still in awe over the way you are able to transition the relationship between you and your mentees from a basic student-teacher partnership to a competent, enriching collaborator-collaborator/mentor companionship. I appreciate you teaching me how to think about science and to complete projects (as well as dozens of life lessons along the way), but I will always think very fondly of working with you as a TA. Going from you teaching me every small step of how to conduct research when working on CONFINED, to working together, writing lectures and homeworks together and even to just chatting together during the odd time of running a pandemic class has been indescribable. Thank you for our time together at UCLA, and for putting CRG on the map—I am hopeful that this is simply a *nos vemos*.

Chapter Two of this dissertation is an abridged version of Mike Thompson, Zeyuan Johnson Chen, Elior Rahmani, Eran Halperin. "CONFINED: distinguishing biological from technical sources of variation by leveraging multiple methylation datasets". Genome Biology, 20, 138 2019.

Chapter Three is a version of a manuscript currently in revision at Nature Communications: Mike Thompson, Mary Grace Gordon, Andrew Lu, Anchit Tandon, Eran Halperin, Alexander Gusev, Chun Jimmie ye, Brunilda Balliu, Noah Zaitlen. "Multi-context genetic modeling of transcriptional regulation resolves novel disease loci".

Chapter Four is a version of Mike Thompson*, Brian Hill*, Nadav Rakocz, Jeffrey Chiang, IPH, Sriram Sankararaman, Ira Hofer, Maxime Cannesson, Noah Zaitlen, Eran Halperin "Methylation risk scores are associated with a collection of phenotypes within electronic health record systems". Nature Genome Medicine, accepted 2022.

This work was supported in part by the NIH Training Grant in Genomic Analysis and Interpretation T32HG002536.

VITA

2017-2022	PhD Candidate, Bioinformatics, University of California, Los Angeles, CA, USA
2013-2017	BS, Microbiology, Immunology, and Molecular Genetics (minor in Bioinformatics), University of California, Los Angeles, CA, USA

SELECTED PUBLICATIONS

* Denotes equal contribution

Mike Thompson*, Brian Hill*, Nadav Rakocz, Jeffrey Chiang, IPH, Sriram Sankararaman, Ira Hofer, Maxime Cannesson, Noah Zaitlen, Eran Halperin. Methylation risk scores are associated with a collection of phenotypes within electronic health record systems. *Nature Genomic Medicine*. 2022.

Richard K Perez, M Grace Gordon, Meena Subramaniam, Min Cheol Kim, George Hartoularos, Sasha Targ, Yang Sun, Anton Ogorodnikov, Andrew Lu, **Mike Thompson**, Nadav Rappoport, Andrew Dahl, Christina Lanata, Mehrdad Matloubian, Lenka Maliskova, Serena Kwek, Tony Li, Michal Slyper, Julia Waldman, Danielle Dionne, Orit Rozenblatt-Rosen, Lawrence Fond, Maria Dall'Era, Brunilda Balliu, Aviv Regev, Jinoos Yazdany, Lindsey Criswell, Noah Zaitlen, Chun Jimmie Ye. Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science*. 2022.

Nadav Rakocz, Jeffrey Chiang, Muneeswar Nittala, Giulia Corradetti, Liran Tiosano, Swetha Velaga, **Michael Thompson**, Brian Hill, Sriram Sankararaman, Jonathan Haines, Margaret

Pericak-Vance, Dwight Stambolian, Srinivas Sadda, Eran Halperin. Automated identification of clinical biomarkers from sparsely annotated 3-dimensional medical imaging. *Nature Digital Medicine*. 2021.

Mike Thompson, Zeyuan Johnson Chen, Elior Rahmani, Eran Halperin. CONFINED: distinguishing biological from technical sources of variation by leveraging multiple methylation datasets. *Genome Biology*. 2019.

Liat Shenhav, **Mike Thompson**, Tyler A. Joseph, Leah Briscoe, Ori Furman, David Bogumil, Itzhak Mizrahi, Itsik Pe'er, Eran Halperin. FEAST: Fast Expectation-Maximization for Microbial Source Tracking. *Nature Methods*. 2019.

Liat Shenhav, Ori Furman, Leah Briscoe, **Mike Thompson**, Itzhak Mizrahi, Eran Halperin. Modeling the temporal dynamics of the gut microbial community in adults and infants. *PLOS Computational Biology*. 2019.

Monica M. Olcina, Nikolas G. Balanis, Ryan K. Kim, B. Arman Aksoy, Julia Kodysh, **Michael J. Thompson**, Jeff Hammerbacher, Thomas G. Graeber, Amato J. Giaccia. Mutations in an Innate Immunity Pathway are Associated with Poor Overall Survival Outcomes and Hypoxic Signaling in Cancer. *Cell Reports*. 2018.

Bryan A. Smith, Nikolas G. Balanis, Avinash Nanjundiah, Katherine M. Sheu, Brandon L. Tsai, Qingfu Zhang, Jung Wook Park, **Michael Thompson**, Jiaoti Huang, Owen N. Witte, Thomas G. Graeber. A Human Adult Stem Cell Signature Marks Aggressive Variants across Epithelial Cancers. *Cell Reports* 2018.

CHAPTER 1

Introduction

1.1 Scope of Research

Innovations in sequencing technologies have led to a massive expansion of genomics datasets available to researchers[1, 2, 3, 4]. Commonly, such datasets are used to discover associations between genetic variability and the variability of a given phenotype or collection of phenotypes (including common traits and complex diseases[5], or even other genomics or biological measurements such as RNA expression[6] and CpG methylation[7]). Association studies are typically straight-forward analyses and enable researchers to discover regions of the genome that are related or causal to a phenotype, potentially elucidating mechanisms or pathways that may be informative for medicine, therapeutics, or basic science[8]. While large and densely phenotyped genomic datasets have enabled researchers to discover a substantial number of associations, the findings from these studies must be replicated across additional datasets before they can be further studied and considered valid[9].

Though replication is a powerful means to instill further confidence in a purported association, genomics datasets are affected by innumerable sources of variability that may hinder validation of discoveries or lead to spurious findings[10, 11, 12]. For example, epigenome-wide associations (EWAS), which aim to implicate associations between methylation levels at various loci and phenotypic variance, are at risk for confounding by age because age is correlated with many phenotypes and methylation sites[13, 14, 15]. Nevertheless, age is only one of many sources of variability in single context association analyses. Other sources of variability can include batch effects and population structure for genotypes[16, 17], as well as batch effects, population structure, smoking status, age, sex, BMI, and cell-type compo-

sition for DNA methylation and RNA sequencing datasets[13, 18, 19, 20, 21, 22, 23]. While the above technical sources of variability are not of interest and should undoubtedly be accounted for in analyses, the biological sources of variability may provide utility in achieving a study-specific aim, such as maximizing prediction power or conditioning in order to interpret an association. Accordingly, it is imperative to partition the biological variability from the technical variability in order to mitigate spurious conclusions[24].

Moreover, genomic analyses may be further complicated in the case of multi-level studies, or studies in which the same individual is measured across multiple contexts or datasets. Unlike single-context studies that contain independent samples or measurements, multi-level studies introduce further complexities at the level of variability[25]. For example, the Genotype-Tissue Expression project (GTEx) has collected the genotypes of roughly 1,000 individuals as well as their RNA-sequencing profiles in multiple tissues[1, 2, 6]. In addition to the aforementioned sources of variability, studies like GTEx include genomic effects that are shared across multiple contexts, genomic effects that are specific to each context, and individual-level effects that are shared across all context measurements of a given individual[25, 26, 27]. Since studies like GTEx not only aim to maximize their power by modeling all individuals at once but to understand the genomic architecture and specificity of expression and disease, they must model the individual-level and genomic effects that are replicated across contexts[27].

Finally, it is essential to evaluate the utility of various data types and their associated variability in downstream (e.g., medical) tasks[28]. Electronic health records (EHR)—which are often heterogeneous and sparse due to the fact that there may exist preference toward, for example, ordering more diagnostic tests and lab panels for individuals who are perceived to be at greater risk for an outcome than an otherwise healthy individual—present an excellent opportunity for highlighting the utility of external biomarkers[29]. Primarily, the sparsity present in EHR can lead to bias when performing imputation tasks, as the collection of individuals from which the estimator is generated is unlikely to be representative of the individuals for whom the imputation is performed[29]. Consequently, researchers have turned

to using external measurements of risk derived from patients' genomics measurements, such as polygenic risk scores (PRS). PRS, though associated with many outcomes, are often population-specific and do not replicate across groups of individuals with varying ancestry[28, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42]. Owing to these constraints, it is crucial to use a biomarker that leverages sources of variability that are replicable across populations.

1.2 Contributions and Overview

In this dissertation, we propose two computational approaches to model the replicable sources of variability present in genomics datasets: the first focuses on disentangling biological from technical sources of variability, and the second on modeling intra-individual effects to partition context-shared from context-specific genetic effects. We next leverage the fact that variability in DNA methylation datasets is often comprised of a wider range of biologically replicable sources than variability in genotype datasets to perform biomarker-informed imputation tasks in electronic medical records.

Chapter 2 begins by describing and classifying sources of variability in genomics datasets to motivate our introduction of CONFINED—an approach to disentangle technical from biological sources of variability. In brief, genomics datasets are affected by measurable and unmeasurable confounders, both of which can be of biological (e.g., cell-type composition or age) or technical (e.g., batch effects) origin. We developed CONFINED, an approach based on sparse canconical correlation analysis (CCA), to model the fact that technical variability is often dataset-specific, whereas biological effects are largely conserved across data. CONFINED finds replicable sources of variability that are conserved across datasets and improves over previous reference-free methods in the estimation of confounders such as cell-type composition, age, and sex. Moreover, we use simulations and real data to show that CONFINED is robust to batch effects and consistently generates components that reflect shared biology (even across multiple tissue types).

In Chapter 3, we present a decomposition and model—termed CONTENT—to capture

context-shared and context-specific genetic effects while leveraging the intra-individual effect present in multi-level studies. We apply CONTENT to GTEx and to CLUES (an in-house single-cell RNA sequencing dataset of peripheral blood mononuclear cells) and show that CONTENT is substantially more powerful than previous approaches when building genetic predictors of expression. Subsequently, we perform transcriptome-wide association studies (TWAS) on a collection of phenotypes and show that the models built by CONTENT not only discover more associations than models built by previous approaches, but that they are more interpretable, as they properly attribute genetic variabilty to its context-shared or context-specific component. Finally, we use CONTENT to show that with bulk, tissue-level RNA-sequencing, genetic effects are largely context-shared, whereas with single-cell-level RNA-sequencing, genetic effects are mostly context-specific.

Though polygenic risk scores (PRS) are associated with a variety of outcomes, their use in risk prediction is often coupled with covariates to improve power[28, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40]. Since methylation is influenced by many replicable sources of variability—including genetics, age, environment, diet, smoking status, exercise and lifetyle choices—we hypothesized that it would capture multi-factorial signal about predispositions to clinical conditions and therefore complement one's genetics as a clinical prediction tool[10, 13, 18, 19, 20, 21, 22, 23]. In chapter 4, we develop methylation risk scores (MRS) for over 600 outcomes in UCLA's EHR. We compare the MRS to in-house and external (UK Biobank) PRS and show that MRS are substantially more accurate than PRS in terms of R^2 and AUC on a variety of outcomes. Moreover, MRS replicated across multiple ancestral populations and in several external datasets. Lastly, we show that existing state-of-the-art EHR imputation approaches can be improved by adding MRS to their input. Our work provides one of the most extensive comparisons of MRS to PRS and demonstrates the potential utility of MRS as a clinical biomarker.

CHAPTER 2

Distinguishing biological from technical sources of variation by leveraging multiple methylation datasets

2.1 Background

While technological advances have provided a surplus of methylation datasets, analyses of these datasets are often complicated by innumerable possible sources of variability [11, 12]. In particular, epigenome-wide association studies (EWAS) and studies that aim to implicate observed methylation signal to phenotypic variance are particularly at risk for false associations due to unknown drivers of the observed signal that globally affect the epigenome [43, 44, 45]. For example, age is correlated with a large number of methylation sites and phenotypes[13, 14, 15], and thus if not corrected for, association between a specific methylation site and a phenotype may be primarily driven by a confounder such as age. In order to mitigate spurious associations in such association studies, it is crucial to elucidate and account for the sources of variation that globally affect the methylation patterns in the genome.

Sources of global methylation effects can be either technical or biological, and may also be measured or unmeasured. In the case of technical sources, most typical are batch effects, or variation resulting from different technicians or conditions during the data-preparing steps [46]. These sources should undoubtedly be identified and accounted for in analyses, for example by balancing cases, controls, and samples from different datasets, including measured potential confounders as covariates, regressing out the sources of confounding signals if they are measured, or otherwise estimating these potential sources of technical

effects and accounting for their estimates [24].

The case of biological sources is more complex; biological sources of variation such as age, sex, cell-type composition, genetics, ethnicity, co-morbidities, or responses to environmental factors like medication intake or smoking status indeed affect the global methylation patterns in the genome, and they are also often correlated to the phenotype of interest[13, 19, 20, 21, 22, 23]. However, due to logistical limitations, often only a few of these sources of biological variation are measured in a given study; moreover, it is often the case that some of the sources of variation that are correlated with the phenotype are unknown and hence unmeasured.

Unlike technical effects, there is much debate over the best practice of using these biological sources of variation in a model (e.g., [21, 43, 47, 48]) since one can argue that identifying these sources is an important ingredient in understanding the disease mechanism. Moreover, identifying these biological sources of variation may be useful in prediction algorithms related to the studied phenotype. In other words, it is context-specific whether one should include biological sources of variation in their model—considering the additional sources as confounders—or simply derive a model considering only the observed signal and accounting for the technical effects[49].

To capture signal corresponding to specific biological sources of variation, reference-based methods have been proposed. In the case of methylation, one commonly researched source of biological variability is cell-type composition. Houseman et al. developed an approach to estimate the true cell-type proportions in methylation datasets using "methylation signatures" (estimates of cell-type-specific methylation levels across a population)[50]. Reference-based methods and methods that leverage prior statistics, however, are limited to known sources of variability for which such reference data exists. In many cases, either the sources of variability are unknown, or there is no reference data that can be utilized for these methods (e.g., factors such as diet and exposure to air pollution[51, 52, 53], and tissues such as solid tumors or adipose[54]). In such cases, reference-based methods cannot be used.

In an attempt to overcome the above limitations, many reference-free methods [55, 56, 54, 57, 58, 59, 60], have been proposed. Though these methods can correct for cell-type

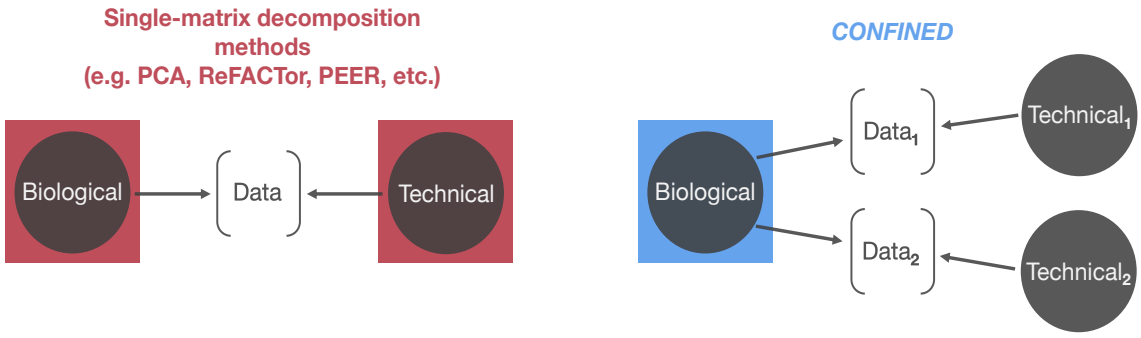

Figure 2.1: CONFINED compared to previous factorization approaches. Previous reference-free methods based on single-matrix decompositions (e.g. principal component analysis, non-negative matrix factorization) capture the dominant sources of variability which may be composed of both biological and technical effects (left). Here, we propose a method to capture solely biological variability (right).

composition in EWAS [61, 58] and may also capture other sources of variability, they are limited by the fact that it is impossible to know whether their components reflect biological or technical signal (Figure 2.1). While technical signal is not of interest and should be accounted for in the analysis, the biological signal can provide useful insights about underlying biological phenomena, for instance by being used to model the interaction with the methylation signal.

In this chapter, we propose a reference-free method that disentangles the technical sources of variation from the biological sources of variation. Our method is based on the observation that the same biological sources of variation typically affect different studies that are performed under the same conditions (e.g., on the same tissue type), while technical variability is study-specific. Thus, unlike previous unsupervised methods that utilize single-matrix decomposition techniques to account for covariates in methylation data, we propose the use of canonical correlation analysis (CCA), which captures shared signal across multiple datasets. In brief, CCA finds shared structure between two datasets by finding maximally-correlated linear transformations of the datasets and is used across many fields including cognitive science[62], psychology[63], and imaging[64]. CCA has been used in the context of genomics to capture genome-wide similarities between different genomic measurements (e.g., gene expression and genetics[65, 66], gene expression and copy number alterations[67, 68]) for the same set of individuals. As opposed to this traditional use of CCA, our method,

named CONFINED (CCA ON Features for INter-dataset Effect Detection), searches for genome-wide similarities between one methylation profile across two sets of individuals. By instead searching across a single genomic profile, we capture shared structure inherent to the underlying biology of the datasets.

The key discrepancy between CONFINED and previous reference-free methods is that CONFINED will only capture shared sources of variability. There are two notable examples in which a method like CONFINED can be leveraged over unsupervised methods that capture dataset-specific variability. First, when capturing unmeasurable and unknown sources of variability CONFINED will distinguish between the technical and biological components of such sources, as technical variability tends to be dataset-specific. Second, if the goal of a study were to elucidate the effects of a dataset-specific effect such as a treatment effect, and one wished to capture and control for covariates, single-matrix methods would fail and adjust away the effect of interest.

We evaluated the performance of CONFINED through both simulated and real data. Our evaluations demonstrate that CONFINED captures signal from only biologically replicable sources of variability. We show, as examples, CONFINED's improvement over previous methods by comparing their performance in capturing methylation signal due to known, measurable sources of variability such as cell-type composition, age, and sex in several whole-blood datasets. We also demonstrate that by inducing sparsity, CONFINED prioritizes features that recapitulate biological functionality inherent to both datasets. For example, when pairing two whole-blood datasets together, the sites best ranked by CONFINED were significantly enriched for immune cell function.

2.2 Methods

2.2.1 A brief introduction to canonical correlation analysis

We first explain the general idea of canonical correlation analysis (CCA) [69]. In the simplest terms, CCA attempts to maximize the correlation of two matrices via linear transformations.

CCA takes as input two matrices X_1 of dimension $n \times m_1$ and X_2 of dimension $n \times m_2$ where $n > m_1$ and m_2. In other words, both matrices have the same number of rows but not necessarily the same number of columns. CCA then attempts to find m_1- and m_2-length vectors a_1 and a_2, such that the correlation of $X_1 a_1$ and $X_2 a_2$ is maximized:

$$\max_{a_1, a_2} \ \text{corr}(X_1 a_1, X_2 a_2) \tag{2.1}$$

To produce a_1 and a_2, we first obtain vectors b_1 and b_2, the eigenvectors corresponding to the largest eigenvalues of the following matrices (where X_1 and X_2 are column-centered):

$$M_1 = \frac{1}{n}^{1/2} (X_1^T X_1)^{-1/2} (X_1^T X_2)(X_2^T X_2)^{-1/2} (X_2^T X_1)(X_1^T X_1)^{-1/2}$$

$$M_2 = \frac{1}{n}^{1/2} (X_2^T X_2)^{-1/2} (X_2^T X_1)(X_1^T X_1)^{-1/2} (X_1^T X_2)(X_2^T X_2)^{-1/2}$$

The vectors a_1 and a_2 are then obtained from a simple change of basis of b_1 and b_2 respectively:

$$a_1 = (\frac{1}{n} X_1^T X_1)^{-1/2} b_1$$

$$a_2 = (\frac{1}{n} X_2^T X_2)^{-1/2} b_2$$

The products $X_1 a_1$ and $X_2 a_2$ are referred to as the first canonical variables of the input matrices, and we let $u_1 = X_1 a_1$ and $u_2 = X_2 a_2$. CCA can produce up to $\min\{m_1, m_2\}$ pairs of canonical variables from the remaining eigenvectors, however, the first pair of canonical variables (corresponding to the largest eigenvalue) has the greatest correlation.

When seeking the second and subsequent pairs of canonical variables, one additional restriction is introduced—the new canonical variables must be orthogonal to all the previous ones:

$$\text{corr}(u_1^{(i)}, u_1^{(j)}) = \text{corr}(u_2^{(i)}, u_2^{(j)}) = 0 \quad i < j$$

Given this constraint, the solution for the i^{th} pair of canonical variables conveniently fol-

lows the same formula as the first pair, only that we substitute the eigenvector corresponding to the i^{th} largest eigenvalue for the eigenvector corresponding to the largest eigenvalue. We then column-wise concatenate all $u_i^{(j)}$ for each dataset to obtain two matrices (U_1 and U_2) of canonical variables of size $n \times \min\{m_1, m_2\}$. Simply put, the collection of canonical variables for each dataset is defined as follows:

$$U_1 = X_1 A_1 \;\; U_2 = X_2 A_2 \tag{2.2}$$

Where A_1 and A_2 are the eigenvectors of M_1 and M_2 respectively. The canonical variables are ordered such that their correlation (which is proportional to their corresponding eigenvalue) is in decreasing order:

$$\text{corr}(u_1^{(i)}, u_2^{(i)}) > \text{corr}(u_1^{(j)}, u_2^{(j)}) \quad i < j$$

Additionally, the canonical variables have the properties that each of their variances equal 1, and the covariance of $u_1^{(i)}$ and $u_1^{(j)}$ (and $u_2^{(i)}$ and $u_2^{(j)}$) is equal to 0 when $i \neq j$:

$$\frac{1}{n} U_1^T U_1 = I, \; \frac{1}{n} U_2^T U_2 = I$$

To reiterate, the basic goal of CCA is to find a_1 and a_2 such that corr($X_1 a_1$,$X_2 a_2$) is maximized. There are $\min\{m_1, m_2\}$ such vectors for each pair of datasets, yielding $\min\{m_1, m_2\}$ pairs of canonical variables.

2.2.2 A formal description of CONFINED

CCA has been used in genomics in many instances [70, 71, 72]. In these cases the rows correspond to individuals, while the columns correspond to features of genomic measurements. For example, each feature could be the expression of a specific gene in one matrix, and in the other matrix it could be the genotype allele, i.e., in this case X_1 corresponds to a gene expression matrix, and X_2 corresponds to a genotype matrix, but both measurements have been taken on the same set of individuals. In CONFINED, we transpose the problem.

Rather than searching for shared directions between two sets of genomic measurements, we instead search for shared directions of the same type of genomic measurement (in our case, methylation), but across two sets of individuals. Moreover, since we find that in practice many sources of variability in methylation only act on a fraction of the methylation sites in the genome[55, 22], CONFINED uses sparsity by limiting the analysis to a fraction of the methylation sites in the genome. We note that our method shares similarities with a recent application of CCA to single-cell expression datasets [73]. However, unlike this method, we search for shared structure across two sets of individuals rather than two sets of cells, and we assume the number of genomic features is larger than the number of individuals (or cells).

Formally, CONFINED takes as input two matrices, X_1 with dimension $m \times n_1$ and X_2 with dimension $m \times n_2$, of m measured methylation sites for n_1 and n_2 individuals respectively. In addition, it takes as input a sparsity parameter t, a dimensionality parameter l, and an output parameter specifying the number of components to generate k. To generate its components, CONFINED first selects the t most informative features then runs CCA on these t features:

1. Obtain U_1 and U_2 both of size $m \times \min\{n_1, n_2\}$ following Equations (1) and (2).
2. Construct $\tilde{U}_1$ and $\tilde{U}_2$ both of dimension $m \times l$ from the first l columns of U_1 and U_2 respectively.
3. Generate a low-rank approximation of each dataset:

$$\tilde{X}_1 = \tilde{U}_1 \tilde{U}_1^T X_1 \qquad \tilde{X}_2 = \tilde{U}_2 \tilde{U}_2^T X_2$$

4. For each site j in dataset i compute a score based on its correlation between itself and its low-rank approximation:

$$S_i^{(j)} = \text{corr}(X_i^{(j)}, \tilde{X}_i^{(j)})$$

5. Rank the sites with the highest inter-dataset score:

$$S_1^{(j)} + S_2^{(j)}$$

6. Perform CCA using the sites with the top t scores, returning CONFINED components $X_1^{[t]T}U_1^{[t]}$ of size $n_1 \times k$ for X_1 and $X_2^{[t]T}U_2^{[t]}$ of size $n_2 \times k$ for X_2.

We set l as the number of pairs of canonical variables with correlation greater than a threshold λ, or 1 in the case that no pairs have this correlation. In practice, we set λ to .95 and found this threshold using cross-validation. By finding the sites that are best approximated by a low-rank, correlated transformation, we therefore assume that the sites with the highest scores will be representative of features that are functionally shared (i.e. correlated) between the datasets. This step is analogous to one taken by ReFACTor [55], only that we leverage the *correlated* subspace of the two datasets rather than a *variable* subspace of one dataset. Though we emphasize that CONFINED can be used for general sources of global biological variation, for the purpose of comparing a single use-case of CONFINED to other methods, we evaluated the effect of t for estimation of cell-type composition in whole-blood datasets and found that CONFINED was robust when using a relatively small number of sites (< 10000) and we therefore recommend a default use of 2000 CpGs.

CONFINED is available as an R package at `https://github.com/cozygene/CONFINED`. The calculations in the R package were optimized with C++ code using `Rcpp` and `RcppArmadillo`. Also included with the package is an ultra-fast function for performing CCA.

2.2.3 Simulation of low-rank structure

We evaluated the performance of CONFINED using a simulated study. For the simulations, we generated $\widehat{X_i}$ for every dataset X_i:

$$\widehat{X_i} = X_i + Z_i W_i^T$$

Where Z_i is a random matrix of "scores" of size $m \times r$ with every entry z_{jk} drawn from the standard normal distribution and W_i is a matrix of "weights" of size $n_i \times r$ where every entry w_{jk} is drawn from the standard uniform distribution and each column $w_i^{(k)}$ is standardized to have norm 1.

In doing so, we add some structured, normally distributed noise that is specific to each dataset. By varying the number and length of the weight vectors $w_i^{(k)}$, we can also control the rank and magnitude of the structured noise. Intuitively, this noise emulates technical variation, as each dataset will have its own unique set of weight vectors.

2.2.4 Permutation testing

To validate the enrichment results reported by `missMethyl`[74], we performed permutation testing. `missMethyl` takes as input a set (i.e. sample) of CpG sites used to test for enrichment of gene ontology pathways, along with the population from which the sample of CpG sites was chosen. For the purpose of the permutation tests, our sample of CpG sites consisted of the top t sites reported by CONFINED, and the population of CpG sites was made up of the m sites in the input matrices. For each number of sites t, we ran `missMethyl` 1000 times, using a random selection of t sites from the m sites of the input datasets at each iteration. We then compared the permutation p-values to the p-values from using the top t CONFINED sites.

2.2.5 Usage of other methods

We compared CONFINED against several previous reference-free methods that were developed to capture cell-type composition. Notably, each method has several parameters the user is left to select, and we wished to provide a fair comparison across methods. In the case of PMA[67], we followed the authors' code and used their cross-validation function to estimate optimal parameters, which, as the reviewer mentions balances the fit of the model by optimizing the sparsity. In the case of PEER[18] we simply used the code in the authors' example in their github wiki. We also followed the authors' recommendations for optimizing

the sparsity parameter and feature-selection steps of ReFACTor[55]. In addition to the above we also tried each of the methods using the top 1,000 to 10,000 most variable sites (with a step size of 1,000) for a more fair comparison (similarly to how was done by Houseman et al. [54]). When we induced sparsity in PMA, PEER and NNMF, the methods' performance were generally lower than when using no sparsity. In terms of R^2, we describe the results when using 10,000 sites and no sparsity respectively: $R^2_{\text{PMA}} = .47$ as opposed to .54, $R^2_{\text{PEER}} = .49$ compared to .52, $R^2_{\text{NNMF}} = .49$ instead of .54. ReFACTor benefited most from sparsity and had the highest performance when using 2,000 sites $R^2_{\text{ReF}} = .79$.

2.2.6 Datasets analyzed

Throughout our main experiments, we used publicly available data generated from the Illumina Infinium Human Methylation 450k chip. Our analyses focused on four whole-blood datasets and one brain-tissue dataset: (1) an analysis of Rheumatoid arthritis patients and controls with 659 individuals from Liu et al. (GSE42861) [75] (2) a study of aging with 656 individuals from Hannum et al. (GSE40279) [76] (3-4) analysis and re-analysis of schizophrenia with 847 and 675 samples from Hannon et al. (GSE80417, GSE84727) [77] and (5) a dataset from Lunnon et al. with brain tissue from 122 individuals that was used to study Alzheimer's disease (GSE59685) [78].

The whole-blood datasets were preprocessed following guidelines suggested by Lehne et al. [79]. Using the R package `minfi` [80], we obtained and subsequently preprocessed the raw IDAT methylation files from the Liu et al. and Hannon et al. datasets. As there was no supplied IDAT file for the dataset of Hannum et al., we simply used their published intensity values. Following the guidelines of Lehne et al., we first removed single nucleotide polymorphism markers (total of 65) then applied the Illumina background correction to the obtained intensity values treating autosomal and sex chromosomes separately. We set our p-value detection threshold to 10^{-16} and set the probes whose p-values did not fall below this threshold as having missing values.

Further, we normalized the whole-blood data using quantile normalization of the intensity

values, subdivided by probe type, probe sub-type, and color channel. After finalizing the intensity levels, we calculated beta-normalized methylation levels for each probe. Probes that had more than 10% of their values missing were discarded from the datasets, and the remainder of missing values were imputed using R package `impute`. Additionally, following [58], we used GLINT [81] to remove polymorphic and cross-reactive sites [82] as well as sites from non-autosomal chromosomes.

The brain dataset from Lunnon et al. was already preprocessed using the function *dasen* from R package `wateRmelon`[83]. Notably, this function also operates on the raw intensity to generate normalized beta values and uses similar preprocessing steps, including quantile normalization and the removal of single nucleotide polymorphisms. As CONFINED takes as input matrices with the intersection of CpG sites in two datasets, the brain dataset was also analyzed with the removal of polymorphic and cross-reactive sites as well as sites from non-autosomal chromosomes.

Additionally, we removed from our analyses outliers and samples with missing information about their sources of variability. Samples whose principal components scores were over four standard deviations away from the mean were excluded, which led to us removing six samples from the Hannum et al. dataset and two samples from the Liu et al. dataset.

We also followed filtering procedures from other works that also used the same datasets, including the removal of consistently methylated or unmethylated sites [55, 58]. Prior to running any analyses, we filtered out methylation sites with standard deviation less than .02. After all preprocessing steps the dataset from (1) Liu et al. had 376021 sites and 658 individuals, (2) Hannum et al. had 382158 sites and 650 individuals, (3) Hannon et al. 381338 sites and 638 individuals, (4) Hannon et al. 382158 sites and 665 individuals, and (5) Lunnon et al. 485577 sites and 451 individuals.

In the analysis across tissue types as well as the brain and adipose analyses in the supplementary sections, we used the respective authors' preprocessed datasets. Notably, in many datasets, there were multiple studied phenotypes. When available, we used only the healthy individuals for the clustering experiment. We also removed sites with low standard

deviation ($< .02$) as well as sites with missing values. In the Huang et al. stomach dataset [84], the authors processed the raw signal intensities to functionally normalized beta values using minfi, and after filtering missing and low variables CpG sites, there were 304163 sites for 61 individuals. [85] et al. used minfi to generate functionally normalized M-values from stomach mucosa which we transformed to beta values for 42 individuals and 267858 sites. The normalized beta values of the lung dataset from Wielscher et al. [86] were generated using packages from Bioconductor and after our filtering contained 302023 sites measured for 33 individuals. Shi et al. [87] generated their beta values using the R package methylumi to perform exponential background correction and control-probed-based normalization, and after our filtering we were left with 316992 sites for 244 individuals. The brain [88] and liver [89] datasets of Horvath et al. contained Beta MIxture Quantile dilation (BMIQ) normalized [90] beta values for 260 individuals at 315050 sites and 79 individuals at 346808 sites respectively. The adipose and liver datasets from Bonder et al. [91] consisted of Subset-quantile Within Array Normalization (SWAN)-normalized beta values that were preprocessed using the minfi package, and after our filtering, the first adipose dataset had 287438 for 71 individuals, the second adipose dataset had 293425 sites for 71 individuals, and the liver dataset had 265523 for 110 individuals. The kidney dataset of Wei et al. [92] was processed by the R package RnBeads to conduct BMIQ normalization and background correction on their beta values, and after filtering out unhealthy individuals and sites with missing values and low standard deviation, we were left with 89763 sites for 46 individuals. The beta values for the kidney dataset of Ko et al. [93] were processed using Illumina GenomeStudio Software 2011.1 Methylation Module 1.8, and after filtering contained 338312 sites measured at 85 individuals. Teschendorff et al. [94] generated their breast dataset beta values using the minfi R package as well as their BMIQ normalization, and after our filtering, it contained 353644 for 92 individuals. The breast dataset of Song et al. [95] contained after filterting beta values for 121 individuals at 324431 sites and was generated using Partek Genomics Suite and SWAN normalization.

2.3 Results

2.3.1 A brief summary of CONFINED

We developed CONFINED to capture biological sources of variability in methylation datasets. As input, CONFINED takes two matrices with the same number of rows (methylation sites) but not necessarily the same number of columns (individuals), k the number of components to produce, and t the number of CpG sites to use, or in other words, a sparsity parameter. As output, CONFINED produces k components that can be used to model biological sources of variability for each input dataset.

Notably, CONFINED is based on CCA which considers two datasets simultaneously. Intuitively, CCA performs a decomposition of two matrices simultaneously, and hence finds linear combinations of features that define biological variation present in both datasets. Conversely, previous methods that decompose one matrix at a time essentially look for linear or non-linear (kernel-based) combinations of features that preserve dominant structure in a single dataset, and this structure may be a combination of both biological and technical signal. Thus, leveraging the shared structure of two datasets through CCA is crucial. Nonetheless, there are two substantial differences between CONFINED and traditional uses of CCA in genomic studies. First, CONFINED looks for shared structure of one methylation profile across two sets of individuals rather than looking for shared structure in one set of individuals across two sets of genomic measurements. Second, CONFINED performs a feature selection procedure that is critical to detect the shared sources of variability across the different datasets.

2.3.2 CONFINED finds biological sources of variability with high accuracy: Analysis across datasets of the same tissue type

We first evaluated CONFINED using a pair of whole-blood methylation datasets from Hannum et al.[76] and Liu et al.[75]. Along with their methylation data were measured sources of biological variation including patients' disease status, age, and sex. In addition to evaluating

CONFINED's ability to capture the measured biological factors, we also evaluated its performance on an unmeasured source of variation, cell-type composition. While in this section we focused on using two datasets corresponding to the same tissue type, we note that the studied phenotypes in the datasets were different (e.g., Hannum et al. studied aging whereas Liu et al. studied Rheumatoid arthritis). As CONFINED looks for only *shared* biological sources of variation, we excluded from our evaluations sources of variation that may only appear in one of the datasets, e.g. patient status. As we show below, using CONFINED we were able to produce components that correlated with both the measured and unmeasured sources of biological signal across both datasets.

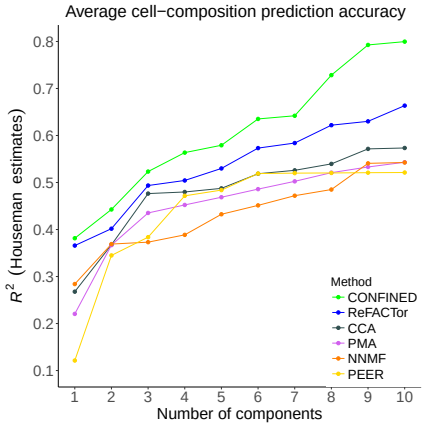

Figure 2.2: A comparison of CONFINED and previous reference-free methods in capturing leukocyte composition. We used each methods' components to capture cell-type proportions as estimated by the reference-based method of Houseman et al. across CD4 T cells, CD8 T cells, monocytes, B cells, natural killer cells, and granulocytes in whole-blood data from an aging study (Hannum et al.) as well as in whole-blood from a study of Rheumatoid arthritis (Liu et al., results omitted for brevity).

First, we evaluated CONFINED against other reference-free methods when capturing unmeasured biological sources of variability in two whole-blood datasets. Here, we used CONFINED to capture cell-type composition, which was unmeasured in both studies. We treated cell-type proportion estimates from the reference-based algorithm of Houseman et al. [54] as the ground-truth. Houseman et al. proposed a reference-based method for estimating proportions of immune cells in whole-blood methylation data by leveraging differentially methylated regions of DNA to form methylation signatures for individual cell-types. They

then use these signatures to obtain cell proportion estimates for several immune cells (CD4 T cells, CD8 T cells, B cells, natural killer cells, monocytes and granulocytes). In our experiments, we fit a linear model of each Houseman-estimated cell-type proportion using several components from each of the methods. CONFINED outperformed all of the previous methods we tested, with pronounced differences in its estimation of the composition of monocytes and natural killer cells (Figure 2.2). To clarify if the gain in performance was a result of CONFINED using more individuals or a more informative feature selection, we considered the situation in which two datasets are concatenated and supplied to a single-matrix-decomposition method as a single dataset, as well as the situation in which a single-matrix decomposition method leverages the features selected by CONFINED. In both procedures, however, the components of the single-matrix method were less correlated to cell-type composition than the components of CONFINED.

We next considered the ability of CONFINED when searching for known, measured sources of variability. For the same pair of blood datasets CONFINED's components captured age and sex with accuracy $R^2_{\text{age}} > .74$ and $R^2_{\text{sex}} > .70$ respectively (Figure 2.3). In the case of other methods, PMA [67] had the highest performance among previous methods, but was greatly outperformed by CONFINED ($R^2_{\text{age}} > .41$ and $R^2_{\text{sex}} > .37$). Notably, using relatively less sparsity to capture age and sex achieved greater accuracy, however this trend was not necessarily observed when using lower sparsity for capturing cell-type composition.

To better understand the implications of CONFINED's sparsity parameter, we evaluated the biological significance of the features selected by CONFINED using the R package `missMethyl` [74]. For a given set of methylation sites, `missMethyl` tests for enrichment in gene ontology (GO) pathways by first mapping the sites to genes (weighing the genes based on the number of sites that map to them), then performing a test built off of Wallenius' noncentral hypergeometric distribution. In order to avoid potential biases resulting from the parametric assumptions in the model of `missMethyl`, we performed permutation testing using its reported p-values. Our test yielded significant enrichment for various ontologies across multiple pairs of datasets. When we paired two whole-blood datasets, the highest ranked

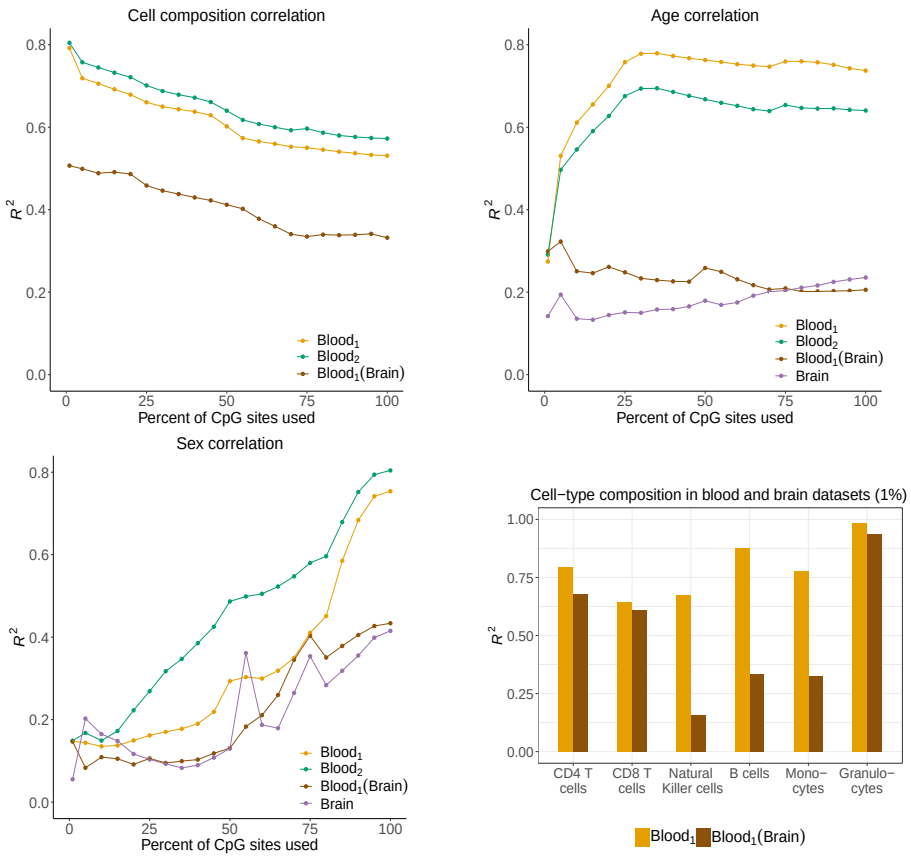

Figure 2.3: Biological drivers of variability captured by across a range of sparsity. We paired a whole-blood dataset (Liu et al.) with another whole blood dataset (Hannum et al.) and with a brain dataset (Lunnon et al.) to capture sources of variability in each dataset. We fit a linear model for each source of variability was using 10 components to obtain an R^2 value. We varied the percentage of CpG sites used from 1% (nearly entirely sparse) to 100% (no sparsity).

features by CONFINED were enriched for pathways generally involved with the immune response, leukocyte activation, and defense response. Notably, most of the significantly enriched pathways were related to the immune system or signaling (Table 2.1). When looking at the enrichment for adipose and brain tissues, we saw pathways concerning vascularization and sheathing respectively. These results underscore the importance of CONFINED's sparsity and provide support for CONFINED's ability to capture biologically meaningful signal, such as tissue-specific cell-type functions.

Table 2.1: Gene Ontology Enrichment of sites ranked by CONFINED. We tested enrichment of the highest-ranked sites by CONFINED in a blood-blood pair of datasets. Here, we set the sparsity parameter based on a rule learned through cross-validation ($t = 2072$), however we observed qualitatively similar results across a range of sparsity parameters, with increasing significance when we included a relatively larger number of CpG sites.

Ontology term	p-value (permutation)	p-value (missMethyl)
Immune system process	.001	6.9e−18
Immune response	.001	1.0e−15
Regulation of immune response	.026	3.0e−11
Defense response	.038	7.18e−11
Regulation of immune system response	.039	7.18e−11
Response to external biotic stimulus	.059	2.58e−10
Response to other organism	.059	2.58e−10
Leukocyte activation	.069	4.68e−10
Regulation of immune effector process	.090	1.86e−09
Response to biotic stimulus	.095	2.46e−09
Positive regulation of immune system process	.100	2.89e−09
Response to bacterium	.103	3.65e−09
Cell activation	.104	3.77e−09
Immune effector process	.104	3.77e−09
Response to stress	.136	1.77e−08
Lymphocyte activation	.139	1.25e−08
Positive regulation of immune response	.143	1.49e−08
Regulation of leukocyte activation	.145	1.59e−08
Regulation of cell activation	.185	2.91e−08
Protein binding	.190	3.10e−08

2.3.3 CONFINED distinguishes between dataset-specific and shared signal: Real data analysis with simulated dataset-specific effects

In the context of capturing biological signal, one of the main limitations of single-matrix decomposition methods (e.g., PCA, ReFACTor [55], PEER [18], non-negative matrix factorization (NNMF) [96]), is that each of their components may consist of a mixture of signal reflective of technical noise specific to a dataset, such as batch effects, and the biological signal. For instance, PCA and methods based on PCA, such as ReFACTor [55] and penalized matrix decomposition (PMA) [67], consider directions in the data that explain the most variability, but this variability is not limited to strictly global biological or replicable effects in the individual datasets. This issue may also be present in PEER [18], which includes a probabilistic version of factor analysis, as the latent factors driving the data may also include some effect from technical variability. Similarly, in NNMF [96] a data matrix is decomposed as a linear combination of different components, and some of the signal of the data matrix may be deconstructed by a component that captures technical variation. Intuitively, CONFINED should be robust to dataset-specific technical effects as it only looks for shared structure across datasets.

To illustrate that CONFINED captures only replicable biological signal, we simulated batch effects for two whole-blood methylation datasets from Hannum et al.[76] and Liu et al.[75] and compared our method to several earlier methods based on single-matrix decomposition. In this setting, we generated dataset-specific noise with low-rank structure and added it to each of the datasets prior to running any feature selection or method. Naturally, simulated batch effects induce technical variation in the datasets, and thus may interfere with methods' abilities to capture biological variation. We used the datasets with added noise to capture cell-proportion estimates of the original datasets as reported by the method proposed by Houseman et al. [50] (Figure 2.4).

We evaluated the performance of each method while varying the strength of simulated, dataset-specific technical effects and found that the components of CONFINED best captured the biological signal and that they were the only components that were robust to

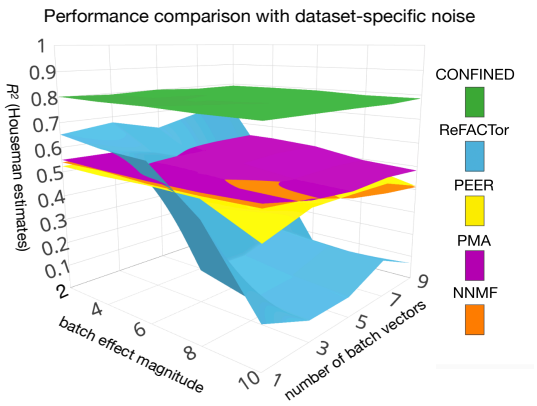

Figure 2.4: Capturing cell-composition in the presence of simulated technical noise. We added simulated batch effects to the whole-blood datasets of Liu et al. and Hannum et al. and compared the ability of, ReFACTor, PEER, PMA, and NNMF to capture cell-type composition in whole-blood. Here, we show the results of the Hannum et al. dataset, however the results of each method were quantitatively similar across both datasets.

technical variation across all levels of noise (Figure 2.4). In addition to the biological signal, the components of the previous methods captured signal pertaining to the simulated batch effects (average R^2 ranging from .131 to .984 depending on the strength of the batch effect).

We also considered the scenario in which a preprocessing step is taken prior to running each method in order to remove technical variation or noise. Here, we used Remove Unwanted Variation (RUV) [46, 12] to generate components which we regressed out from the datasets with added noise prior to running any of the previous methods. Using RUV as a preprocessing step helped improve the single-matrix methods in the presence of simulated technical noise, however the components generated by CONFINED in the presence of the technical noise (and without any such preprocessing) were still more correlated with cell-type composition than those produced by the single-matrix methods (average difference in R^2 between CONFINED and ReFACTor $> .10$).

In the case where one wishes to elucidate the effects of a treatment that has been administered to a set of individuals in one dataset, CONFINED may also be of use. In a second simulation experiment, we simulated a rank-one treatment effect following a similar strategy used in the batch effects simulations, only that we used the absolute value of the batch effect scores (i.e. we assumed that the treatment effect had the same directionality

across samples). We then added this positive treatment effect to a subset of individuals in one of the whole-blood datasets prior to any analysis. We paired the dataset with added treatment effects with one of the raw datasets and obtained the CONFINED components for each dataset. Afterward, we regressed out the top 10 CONFINED components from the treatment dataset. Comparing the PCA plots of the treatment dataset before and after preprocessing (i.e. removing the shared signal) shows how CONFINED can be leveraged to highlight a dataset-specific treatment effect (Figure 2.5). In the scenario where the treatment effect was a dominant source of variability, using CONFINED as a preprocessing step did not diminish the ability to distinguish between those who received treatment and those who did not.

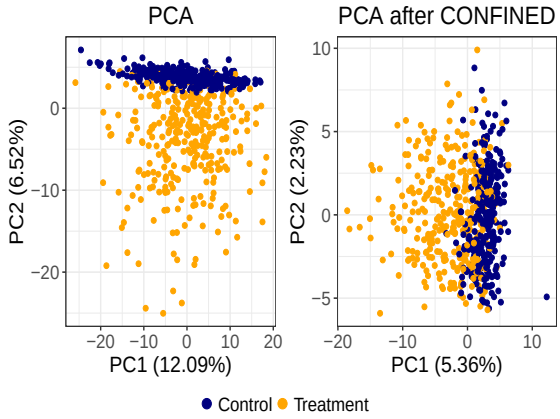

Figure 2.5: Highlighting treatment effect. We removed from a dataset with simulated treatment effect the components generated by CONFINED. Notably, this simulated treatment effect was not shared across datasets. On the left, PCA performed on the dataset prior to removing the CONFINED components, and on the right the PCA of the dataset after regressing out the CONFINED components.

2.3.4 CONFINED finds the shared biology across datasets: Analysis of datasets of different tissue types

We also used CONFINED's components to capture measured sources of biological variation across tissue-types (Figure 2.3). In one experiment, we paired a whole-blood dataset [75] with a dataset from Lunnon et al.[78] composed from brain tissue. Notably, the accuracy of CONFINED to capture each source of signal varied depending on the pairing of the

tissue-type (i.e blood-blood vs. blood-brain) and the sparsity parameter used.

When pairing the blood dataset with the brain dataset, CONFINED's components were correlated with some of the whole-blood dataset's measured biological factors with slightly less strength than when pairing it with a dataset of the same tissue type ($R^2_{\text{age}} > .27, R^2_{\text{sex}} > .39$) (Figure 2.3), possibly suggesting a different architecture for genome-wide variation across the different tissue types. Nonetheless, the cell-type composition accuracy for the blood dataset when paired with the brain dataset was still relatively high (average $R^2_{\text{cell}} = .54$). This is likely due to the fact that several types of immune cells are known to populate or have immune-related functions in the brain (e.g. resident T cells [97, 98], glia [99] and neutrophils (granulocytes)[100]). Therefore, the immune function of cells in the brain and immune cells in the blood may follow similar pathways that could be reflected in the epigenome. The biological sources of variability in the brain dataset were captured with overall less accuracy than the whole-blood biological sources of variability ($R^2_{\text{age}} > .21, R^2_{\text{sex}} > .33$).

When pairing the blood and brain datasets, we observed enrichment results somewhat similar to when using the blood-blood pair, but with less significance. The most enriched pathways in the blood-brain pair included several immune system or hematopoietic processes, but the less enriched pathways were primarily different than when pairing the two blood datasets. The pathways in the blood-brain pair were generally not significantly enriched using permutation testing, unless we used a relatively lower level of sparsity.

Considering CONFINED's ability to find the biological signal shared across two datasets, we performed an additional experiment in which we included datasets corresponding to tissues from the following types: adipose, blood, brain, breast, kidney, liver, lung, and stomach. For each tissue type, we gathered two datasets. Here, we wished to elucidate the shared structure across tissue-types, e.g. if it were possible to use CONFINED to cluster datasets based on their tissue type. For each pair of datasets, we saved the correlations output by CONFINED (i.e. the correlations between the canonical variables as defined in the Methods section), and used a statistic of these correlations to construct a distance matrix for use in hierarchical clustering. We took the mean of the top 10 correlations

between each pair of datasets, i, j, and populated each entry of the matrix$_{ij}$ with this mean correlation. Intuitively, this acts a metric of similarity between each dataset. After running hierarchical clustering, we found that tissues of the same type clustered together for each of the datasets. We believe that this presents evidence that CONFINED is in fact finding signal that recapitulates the underlying biology shared between two datasets.

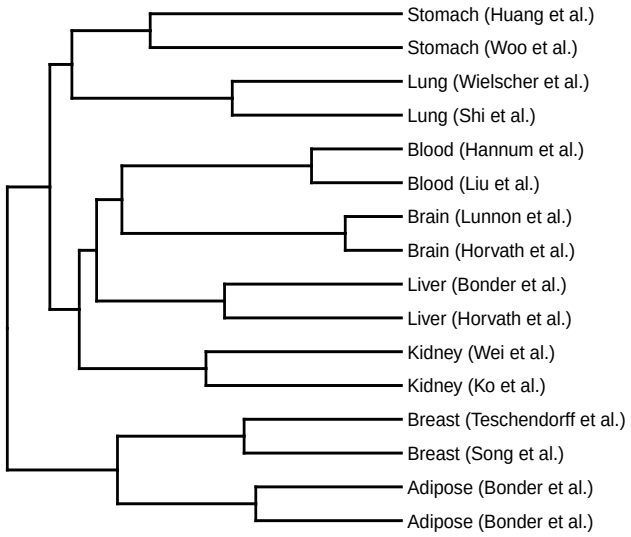

Figure 2.6: Capturing shared biology across datasets. To validate that CONFINED finds biology shared across datasets, we gathered 2 datasets for 9 tissue types, then considered their CCA-based correlations as a metric of similarity. Here, we perform hierarchical clustering, using as a metric of similarity the mean correlation of the top 10 CCA correlations.

2.4 Discussion

Here, we propose CONFINED, a sparse-CCA-based method to capture biologically replicable signal by leveraging shared structure between datasets. Though CONFINED captures the shared variability between two datasets, there may be sources of variability that are unknown or unmeasurable present in the datasets, and we cannot evalaute CONFINED's performance for these sources of variability. Therefore, we have highlighted the strength of CONFINED through examples of known measured and unmeasured sources of variability. Specifically, we showed its use and improved accuracy over other methods in the context of capturing

cell-type composition between datasets of the same tissue type. We also showed how it can be used to capture other sources of biological signal shared across datasets. Moreover, we provide evidence that CONFINED can be used as a feature selection mechanism, prioritizing features that are functionally shared between datasets.

Across several datasets we demonstrated that CONFINED accurately captured global biological sources of variability. In the case of cell-composition, the components produced by CONFINED better captured cell-type composition across all cell-types in methylation datasets (of the same tissue-type) than previous reference-free methods that were designed for capturing signal from cell-type composition. Additionally, CONFINED's components captured other replicable sources of variability such as age and sex. While cell-type composition was better captured when using a pair of datasets of the same tissue-type, we note that other biological factors may be better captured when pairing two datasets of different tissue types. Our results provide grounds for CONFINED as a means to capture replicable signal from biological sources across datasets.

Additionally, CONFINED is robust to technical variability. Through simulations, we demonstrated that CONFINED accurately captures biological signal in the presence of strong, dataset-specific technical noise. Other methods that leverage decompositions of single matrices produced components corresponding to the simulated technical noise, but the components produced by CONFINED were unaffected by the simulated noise. Therefore, leveraging *multiple* datasets through CONFINED can provide researchers a way to robustly account for signal arising from technical variation. Though the premise of CONFINED is to leverage the shared structure across two datasets to distinguish technical noise, we performed an experiment in which CONFINED uses a single dataset split into halves as input instead of two separate datasets. In this experiment CONFINED suffered from issues similar to single-matrix methods, and its performance was negatively affected by the presence of dataset-specific variability (average R^2 from $> .73$ to $> .55$ without and with batch effects respectively).

Though we develop a cross-validation routine and suggest a default setting for the sparsity

parameter (i.e. the number of features) in the specific case of capturing cell-type composition in methylation whole-blood datasets, we emphasize that the selection of the sparsity parameter in other cases may be non-trivial. Evaluating CONFINED on multiple datasets and sources of biological variability aside from cell-type composition, we found that the optimal sparsity parameter for cell-type composition may not be optimal for other covariates of interest. For instance, with a pair of blood datasets where the sex chromosomes were removed, sex was better captured as the number of features increased. This may be due to the fact that specific biological functions—such as the immune response—may be confined to several thousand methylation sites, whereas autosomal changes in methylation patterns due to more broad characteristics—such as age or sex—are more minute, and thus require more information or sites to capture. (Of course, when the sex chromosomes are included in the analysis, the accuracy of CONFINED can improve dramatically ($R^2_{\text{sex}} > .9$).) We suggest future investigations take place and considerations about underlying biology be taken into account for selecting the optimal sparsity parameter for biological signal aside from cell-type composition.

We also showed the utility of CONFINED as an unbiased way of selecting informative and potentially biologically relevant methylation sites. Intuitively, as CCA finds shared structure between datasets, this structure should be reflective of biological mechanisms that are common to a pair of datasets. In our experiments, CONFINED found methylation sites that capture the shared variability across different blood tissues, and this set of sites was significantly enriched for immune function. Similarly, for the brain-blood pair, we observed enrichment for some immune and hematopoietic function, but the enrichment was generally not significant. Thus, our results suggest that our feature-selection method may be useful in highlighting pathways that are similar across two datasets.

A similar concept to CONFINED has been previously introduced in the context of single-cell RNA-sequencing by Butler et al.[73]. However, mathematically, the problem Butler et al. solve is different as the number of "individuals" (in their case, cells) in single-cell RNA is much larger than the number of features (genes), whereas in our setting, the number

of individuals is much smaller than the number of features (methylation sites). Moreover, we show that a simple application of CCA does not suffice in the case of methylation, and thus CONFINED performs feature selection prior to performing CCA. In other words, CONFINED utilizes sparsity.

Importantly, determining the input and usage of the output of CONFINED is goal-specific. As the assumption of CONFINED is that the biological variability in two datasets is shared, we suggest pairing two datasets with similar characteristics, e.g. design protocol or sample collection. In such cases, for any pair of datasets, CONFINED can be used to capture variability or model biological factors that are present in both datasets for use in downstream analyses. On the other hand, CONFINED can be used as a preprocessing step to make dataset-specific effects more prevalent. In Figure 2.5, we show how CONFINED can be used to highlight a treatment effect that was present in a subset of individuals in one of the input datasets. Thus, CONFINED enables researchers to decide how they wish to model the shared or unshared variability in their datasets.

The parameters of CONFINED can be fine-tuned for downstream analyses. In general, we recommend inducing sparsity to capture variability due to specific functions, such as cell-type composition. For more broad characteristics, such as age and sex, we recommend less sparsity is induced. There may be tradeoffs when attempting to optimize the correlation of the CONFINED components and specific sources of variability, and we suggest from our empirical results using around fifty percent sparsity. We found the correlation threshold to be robust across a large range of values, but suggest using a relatively higher correlation such as .95. Lastly, we suggest using a low number (e.g. 6 or 10) of CONFINED components as people often do in EWAS with principal components [55, 101].

In summary, our results suggest that CONFINED will be a useful tool in capturing effects of biological variability as well as highlighting shared cellular mechanisms across multiple datasets. The components from CONFINED can be used in downstream analyses that wish to model only the biological signal of a methylation dataset or to include certain biological signals as confounders in statistical analyses. We suggest future research into the selection

of t, the number of informative sites to use for recovering signal for specific biological factors, as well as research into which pairs of phenotypes or datasets may be useful in extracting signal for specific biological drivers of variability. We posit that using extensions of CCA which include more than two datasets [67] may be a promising future direction.

CHAPTER 3

Multi-context genetic modeling of transcriptional regulation resolves novel disease loci

3.1 Background

A large portion of the signal discovered in genome-wide associations studies (GWAS) has been localized to non-coding regions [102]. In light of this, researchers have developed post-GWAS approaches to elucidate the functional consequences of variants and their impact on the etiology of traits [6]. One notable approach has been to generate genetic predictors of gene expression and leverage these predictors with GWAS data to associate genes with traits of interest[103, 104]. These transcriptome-wide association studies (TWAS) have not only shown great promise in terms of discovery and interpretation of association signals but have also helped prioritize potentially causal genes for complex diseases [105]. Nonetheless, methods like TWAS are limited by the accuracy and power of the genetic predictors generated in training datasets [106, 107, 108, 109, 110, 1].

The original TWAS methodology builds genetic predictors of expression on a context-by-context basis. For example, in a study with RNA-seq and genotypes collected across multiple tissues, the expression of each tissue would be modeled independently [103, 104]. More recent methods model multiple contexts simultaneously and leverage the sharing of genetic effects across contexts [109, 108, 110, 111]. However, these approaches do not maximize predictive power because they ignore the intra-individual correlation of gene expression across contexts inherent to studies with repeated sampling, e.g., the Genotype-Tissue Expression (GTEx) project [2] (Figure 3.1; [112, 113]). Moreover, they build predictors which are mixtures of

both context-specific and context-shared (pleiotropic) genetic effects, making it difficult to distinguish the relevant contexts for a disease gene, and are often computationally inefficient [109]. A recent approach by Wheeler et al. [114] does model correlated intra-individual noise with a linear-mixed model, but does not produce combined predictions of expression, reducing overall power. Finally, existing methods with the goal of maximizing the number of discoveries made may employ multiple testing strategies that either fail to control for all tests performed, (e.g., by controlling the false discovery rate (FDR) within each context separately [104, 115]), or limit their discoveries as they are based on conservative FWER control (e.g., by using Bonferroni adjustment across all contexts [115]). Together, these shortcomings reduce power and interpretability of TWAS.

Here, we introduce CONTENT—CONtexT spEcific geNeTics— a novel method that leverages the correlation structure of multi-context studies to efficiently and powerfully generate genetic predictors of gene expression. Briefly, CONTENT decomposes the gene expression of each individual across contexts into context-shared and context-specific components [26], builds genetic predictors for each component separately, and creates a final predictor using both components. To identify genes with significant disease associations, CONTENT employs a hierarchical testing procedure [116, 117]. CONTENT has several advantages over existing methods. First, it explicitly accounts for intra-individual correlation across contexts, boosting prediction performance. Second, by building specific and shared predictors, it can distinguish context-shared from context-specific genetic components of gene expression and disease. Third, it employs a recently developed hierarchical testing procedure [117] to not only adequately control the FDR across and within contexts, but boost power in cases where a gene has a significant association to disease in multiple contexts. Fourth, this adjustment procedure allows for inclusion of other TWAS predictors [109, 103, 111, 104, 110, 108], enabling approaches to be complementary in discovering associations. Finally, CONTENT is orders of magnitude more computationally efficient than several previous approaches.

We evaluated the performance of CONTENT over simulated data sets, GTEx[1, 6, 2], and a single-cell RNA-Seq data set[118, 119]. We show in simulations that CONTENT captures

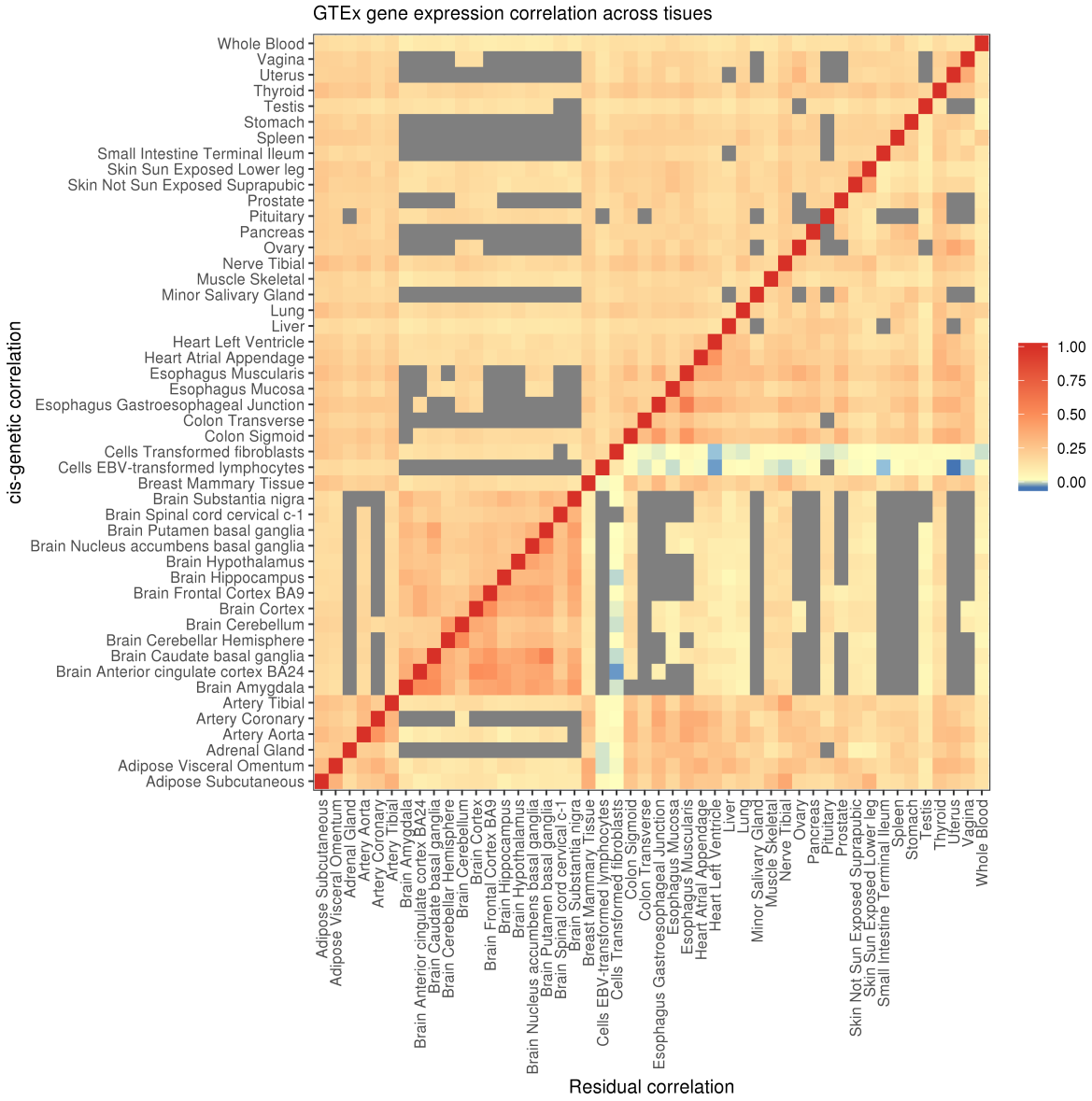

Figure 3.1: Gene expression correlation across tissues in the GTEx study. Using a linear mixed model with bivariate REML, we calculated cis-genetic and residual (which captures variance due to both trans-genetic effects as well as residual effects) variance and covariance components for each gene-tissue pair across GTEx. The gray units indicate tissue pairs with less than 10% sample overlap. In both the genetic (upper) and residual (lower) components, there was widespread cis-genetic and residual correlation, with the brain tissues showing higher correlations compared to other tissues.

a greater proportion of the heritable component of expression than previous methods (at minimum over 22% more), and that CONTENT successfully distinguishes the specific and shared components of genetic variability on expression. In applications to GTEx, CONTENT

improved over previous context-by-context methods both in the number of genes with a significant heritable component (average 42% increase in significant gene-tissue pairs discovered) as well as the proportion of variability explained by the heritable component (average increase of 28%) [104, 103]. Consistent with complex cell type heterogeneity within tissues [120, 121, 122, 123], we find that in applications to the single-cell data, genetic predictors at the cell type level have substantially more context-specific heritability than the tissue-level models. We then performed TWAS across 22 phenotypes using weights trained on GTEx and scRNA and found that CONTENT discovered over 51% independent, significantly associated loci. We provide CONTENT gene expression weights for both GTEx and the single-cell dataset at the TWAS/FUSION repository (`http://gusevlab.org/projects/fusion/`).

3.2 Methods

3.2.1 An overview of the CONTENT model

In this section, we detail the assumed generative model and objectives of CONTENT. CONTENT is based on the methodology and decomposition of a previous work by Lu et al., FastGxC [26]. In brief, like FastGxC, we assume that the expression of an individual in a given gene and context is a combination of a context-shared genetic component that is shared across different contexts and a context-specific genetic component that is specific to a context, that is

$$
\begin{aligned}
E_c &= E_G^{\text{Shared}} + E_{G,c}^{\text{Specific}} + \varepsilon_c \\
E_G^{\text{Shared}} &= g\beta \\
E_{G,c}^{\text{Specific}} &= g\gamma_c
\end{aligned}
$$

where E_c denotes the expression of the individual at the gene in context c, E_G^{Shared} and $E_{G,c}^{\text{Shared}}$ denote the components of the expression due to context-shared and context-specific

genetic effects respectively, β and γ_c represent the context-shared and context-specific cis-genetic effects respectively, g the individual's cis-genotypes and $\varepsilon_c \sim N(0, \sigma_c^2)$ represents the environmental effects (and non-cis-genetic effects) on the individual's gene expression.

The objective of CONTENT is to build a genetic predictor of context-specific phenotypes. While previous work has focused on building powerful genetic models for E_c, we aim to build unbiased models that partition and estimate the context-shared $g\beta$ and context-specific terms $g\gamma_t$. Specifically, we aim to maximize the power to detect the context-specific terms, allowing some leniency in the accuracy of context-shared terms, as we are interested in context-specific effects. Moreover, as a context-specific predictor can be used in downstream analyses to identify the specific context(s) through which genetic variation manifests its effect on the phenotype and disease risk, we also aim to minimize the correlation between the predicted context-specific component and the true context-shared component. Finally, our method must account for the correlated intra-individual noise across contexts, and do so in a computationally efficient manner.

3.2.2 Decomposing multilevel data

Many genomic datasets, such as those of GTEx, have a multilevel nature; first the individuals are sampled, and second an individual is measured in each context. To take the multilevel structure of the data into account, the observed expression on gene j can be decomposed into an offset term, a between-individual component and a within-individual component [25]. That is, if E_{ijc} denotes the observed expression level for individual i ($i = 1, \ldots, I$) on gene j ($j = 1, \ldots, J$) and context c ($c = 1, \ldots, C$), E_{ijc} can be decomposed as

$$E_{ijc} = E_{.j.} + (E_{ij.} - E_{.j.}) + (E_{ijc} - E_{ij.}) \tag{3.1}$$

where $E_{.j.} = \frac{1}{I \times C} \sum_{i=1}^{I} \sum_{c=1}^{C} E_{ijc}$ the mean expression of gene j computed over all (I) individuals and all (C) contexts, and $E_{ij.} = \frac{1}{C} \sum_{c=1}^{C} E_{ijc}$ the mean expression of individual i on gene j, computed over all contexts. In (1), $E_{.j.}$ is a term that is constant across individuals

and contexts for each gene, $(E_{ij.} - E_{.j.})$ is the between-individuals deviation, and $(E_{ijc} - E_{ij.})$ is the within-individual deviation of the expression on gene j in context c.

Variables that differ between but not within individuals, e.g. sex and genotype, will have an effect on $(E_{ij.} - E_{.j.})$ but not on $(E_{ijc} - E_{ij.})$. On the other hand, variables that change within individuals but are the same between individuals, e.g. the genetic effect on a specific context, will have an effect on $(E_{ijc} - E_{ij.})$ but not on $(E_{ij.} - E_{.j.})$.

In the context of estimation, we first center and scale the expression of gene j in each context c, i.e. $\frac{1}{I}\sum_{i=1}^{I} E_{ijc} = 0$ and $\frac{1}{I}\sum_{i=1}^{I} E_{ijc}^2 = 1$. Therefore, $E_{.j.} = \frac{1}{I \times C}\sum_{i=1}^{I}\sum_{c=1}^{C} E_{ijc} = 0$, and equation (3.1) simplifies to:

$$E_{ijc} = \underbrace{E_{ij.}}_{E_{ij}^{\text{Shared}}} + \underbrace{(E_{ijc} - E_{ij.})}_{E_{ijc}^{\text{Specific}}} \tag{3.2}$$

3.2.3 A formal description of CONTENT

We use the simplified decomposition in equation (3.2) to build genetic predictors of context-specific effects while accounting for the correlated intra-individual noise across contexts. Intuitively, the between-individuals variability serves as the component of expression that is shared across contexts, E^{Shared}, and the deviance from this shared component (i.e. the within-individual variability) serves as the context-specific component of expression, E^{Specific}. Moreover, treating the context-specific component as a deviance from the context-shared component leads the decomposition to have the property that as the correlation of intra-individual noise across contexts increases, the power to detect context-specificity also increases. In addition, the decomposition generates context-shared and context-specific components of expression that are orthogonal to each other. Further rationale for using the decomposed expression is included in the text by Lu et al. [26]. Lu et al. also include a description of the decomposition's equivalence to a linear mixed model.

For a single gene j, CONTENT takes as input centered, scaled, and residualized (over a set of covariates) expression measured across I individuals in C contexts and an $I \times m$

genotype matrix G_j with m measured cis-SNPs for gene j. CONTENT then decomposes the expression vectors into C context-specific components and a single context-shared component by simply calculating the mean of expression for each individual across contexts, and setting the context-specific expression for context c as the difference between the observed expression of context c and the calculated context-shared expression. As it has been observed that cis-genetic effects may be sparse and that the elastic net may perform best relative to other penalized linear models in the context of genetically regulated gene-expression [104, 114], CONTENT fits $C+1$ penalized linear models for the $C+1$ expression components using an elastic net. Lastly, CONTENT generates a final genetic predictor of expression by combining the context-shared and context-specific components. Importantly, as the context-specific component is a deviance from the context-shared component, the sign of the context-specific component must be properly realigned when combining both components of expression to make a final predictor. We refer to this linear combination of expression components as the "full" model of CONTENT and fit it using a simple linear regression:

1. Obtain E_j^{Shared} and E_{jc}^{Specific} from the decomposition.

2. Generate cis-genetic predictors of each component using cross-validated elastic net:

   (a) Fit cross-validated elastic net regressions for the shared and specific components:

$$E_j^{\text{Shared}} = \alpha^{\text{Shared}} + G_j\beta + \varepsilon^{\text{Shared}} \tag{3.3}$$

$$E_{jc}^{\text{Specific}} = \alpha_c^{\text{Specific}} + G_j\gamma_c + \varepsilon_c^{\text{Specific}} \tag{3.4}$$

   (b) Use the estimates to generate genetic predictors of each component:

$$\hat{E}_{jG}^{\text{Shared}} = \hat{\alpha}^{\text{Shared}} + G_j\hat{\beta} \tag{3.5}$$

$$\hat{E}_{jcG}^{\text{Specific}} = \hat{\alpha}_c^{\text{Specific}} + G_j\hat{\gamma}_c \tag{3.6}$$

3. Regress the expression of context c onto the context-shared and context-specific com-

ponents:

$$E_{jc} = \alpha_c^{\text{Full}} + \hat{E}_{jG}^{\text{Sh.}} w_{jc}^{\text{Sh.}} + \hat{E}_{jcG}^{\text{Sp.}} w_{jc}^{\text{Sp.}} + \varepsilon_{jc} \tag{3.7}$$

Within each regression, α represents the offset and we assume that all ε are from a normal distribution with mean 0 and standard deviation that is a function of the given outcome.

We save for each gene the set of estimated regression weights $\hat{w}_{jc}^{\text{Shared}}$ and $\hat{w}_{jc}^{\text{Specific}}$ from equation (4) for use in downstream analyses. Namely, in TWAS, each context receives a single vector of weights, and to test the association of a gene-context's full model to a trait, we simply use a weighted sum of the predictors learned from equation (3), $\hat{w}_{jc}^{\text{Sh.}} \hat{\beta} + \hat{w}_{jc}^{\text{Sp.}} \hat{\gamma}_c$. We also use the same procedure for the context-specific weight to ensure the correct directionality. To test for significance of genetic effects (i.e. to call an eGene or eAssociation), we correlate each component of expression—the context-shared, context-specific, and full—to its corresponding genetically predicted value.

3.2.4 Controlling the false discovery rate across contexts

Generally, methods for building genetic predictors of expression or TWAS predictors leverage either Bonferroni correction or false discovery rate (FDR). Nonetheless, using a Bonferroni correction may be too stringent (for example, as tests across contexts may be correlated), and using FDR within each context or across all contexts simultaneously may lead to an inflation or deflation to the false disovery proportion within certain contexts [116]. To simultaneously control the FDR across all contexts at once, a hierarchical false discovery correction—treeQTL—was developed [116]. The treeQTL procedure leverages the hierarchical structure of a collection of tests (e.g. gene level and gene-context level) to properly control the FDR across an arbitrary number of contexts and levels in the hierarchy as well as boost power in cases where a gene has a significant association in multiple contexts [116, 117, 106].

Notably, using CONTENT, our testing hierarchy contains 3 levels; (1) at the level of the gene, (2) at the level of the context, and (3) at the level of the method or model (Figure 3.2). Intuitively, a gene may contain a genetic component that is shared across all contexts, or a

given context may have its own genetic architecture. In CONTENT, a given context may have its own genetic predictor from either the context-specific component or the full model. Using treeQTL with this structure is robust across multiple contexts, and since the tree is structured such that a specific method/model is at the final level of testing for a context, it enables incorporation of additional models trained from other approaches (such as those fit on a context-by-context basis or by UTMOST). Moreover, we can add to the shared leaf an additional level of tests to account for additional components of effects-sharing, such as a brain tissue-shared component.

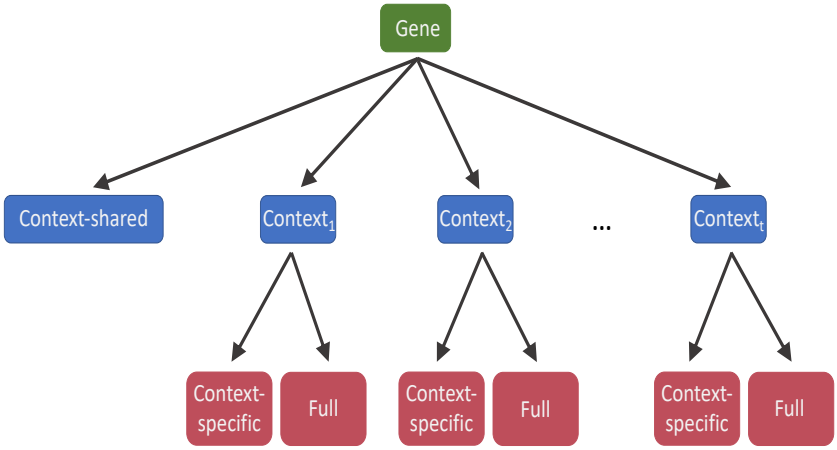

Figure 3.2: Hierarchical false discovery correction. Here, we show the structure of the hypothesis tests for determining whether a gene has a heritable component. A gene (green, top level) is considered heritable if it has a heritable context-shared component or if it was heritable for a specific context (blue, second level). A given gene-context may be heritable due to either the full or context-specific model of CONTENT (red, third level).

3.2.5 Comparison to other methods

We compared the prediction accuracy of CONTENT to a context-by-context TWAS model [103, 104] in which the expression of each context is modeled separately, and to UTMOST [109], a method that jointly learns the genetic effects on all contexts simultaneously. Specifically the model based on TWAS fits a penalized linear model for each context. UTMOST, on the other hand, employs a group LASSO penalty across all contexts simultaneously, allowing it to gain power over the context-by-context approach by considering all individuals

and contexts in a study at once. As we were we able to use a fast R package for penalized regression[124], we used 10-fold cross-validation to fit the context-by-context model. Owing to UTMOST's computational intensity, we used its default value of 5 folds for cross-validation.

We also compared CONTENT to a previous approach by Wheeler et al., orthogonal tissue decomposition (OTD)[114]. OTD is a direct correlate of CONTENT(Shared) and CONTENT(Specific), and is generated by fitting a mixed effects model across all contexts for a given individual. Namely, a mixed effects model is fit as follows: an individual's expression across all tissues is set as the outcome, the shared expression is modeled as a random individual-level intercept and is estimated using the posterior mean, and the specific expression is treated as the residuals from the fit model (after adjusting for covariates). Under infinite sample sizes, the components of OTD are equivalent to CONTENT(Shared) and CONTENT(Specific).

Evaluations on GTEx and CLUES We residualized the expression of each gene in each context over their corresponding covariates (e.g. PEER factors, age, sex, batch information) prior to fitting UTMOST and an elastic-net model for each context in the context-by-context approach. We did the same residualization before decomposing and then fitting the context-shared and context-specific components with an elastic net for CONTENT. After generating cross-validated predictors for each method, we examined the number of significantly predicted genes as well as the prediction accuracy (in terms of adjusted R^2) between the cross-validation-predicted and true gene expression per gene-context pair.

To properly control the false discovery proportion at .05 across-contexts and within-methods, we employed a hierarchical FDR correction [116, 117] separately for CONTENT, UTMOST, and the context-by-context approaches. Notably, using this correction for all methods provides a generous comparison to previous methods, as when there exists at least one significantly heritable gene-context association for a given gene, there is a relative gain in power over the context-by-context FDR for other contexts tested within this gene [116, 117].

Application to TWAS We performed transcription-wide association studies across 24 phenotypes (Table 3.1) using FUSION-TWAS[103]. FUSION-TWAS uses GWAS summary statistics and user-specified gene expression weights with an LD reference panel to perform the test of association between genetically predicted gene expression and a phenotype of interest. We tested a gene-context pair for association if the pair's expression was predicted at a nominal p-value of .1, and note that this threshold does not substantially alter the number of TWAS discoveries. Notably, previous methods may use their own test of gene-context-trait association or leverage set tests (e.g. Berk Jones[109]) to combine their associations across all contexts for a given gene and therefore increase power. In this comparison, we report the association as output by FUSION (a single gene-context-trait association) and corrected by hierarchical false discovery without any sort of set test for the sake of equality in the comparison. We ran FUSION-TWAS using the default recommended settings, with reference data from the 1000 genomes project [125]. TWAS weights were trained on the GTEx v7 dataset[6] as well as the CLUES[119] single-cell RNAseq dataset of PBMCs. For a given gene-context-trio, we ran (assuming each model built a weight for the gene-context under our nominal p-value threshold of 0.10) 5 TWAS—1) context-by-context, 2) UTMOST, 3) CONTENT(Shared), 4) CONTENT(Specific), and 5) CONTENT(Full). Notably, we re-trained each methods' predictors on genetic variants that are present in the LDREF cohort as well as GTEx or CLUES to ensure selected expression weights had overlap with the reference panel (LDREF).

Table 3.1: GWAS summary statistics used as input for TWAS. Abbreviation used for each trait as well as and its respective study and sample size. The collection of traits from the UKBiobank were self-reported and measured on the same set of individuals across traits.

Symbol	Trait	Study	Sample Size
AD	Alzheimer's disease	Lambert et al. Nat Genet. 2013	74,046
Asthma	Asthma (self-reported)	UKBB Loh et al. 2018 Nat Genet	361141.00
Bipolar	Bipolar Disorder	PGC Cell 2018	73,684
CAD	Coronary Artery Disease	CARDIoGRAM Nat Genet. 2011	86,995
CKD	Chronic Kidney Disease	Wuttke et al. Nat Genet. 2019	1,046,070
Crohn's	Crohn's Disease	IIBDGC Europeans Nat Genet. 2015	13,974
Eczema	Eczema (self-reported)	UKBB Loh et al. 2018 Nat Genet	361,141
FastGlu	Fasting Glucose	MAGIC Nat Genet. 2012	96,496
HDL	High-density Lipoprotein	Teslovich et al. Nature 2010	99,900
IBS	Irritible bowel syndrome (self-reported)	UKBB Loh et al. 2018 Nat Genet	361,141
LDL	Low-density lipoprotein	Global lipids genetics consotrium Nat Genet 2013	188,577
Lupus	Systemic Lupus Erythromous	Bentham et al. Nat Genet 2015	23,210
MDD	Major Depression Disorder	PGC; Howard et al. Nat Neuro 2019	807,553
MS	Multiple Sclerosis (self-reported)	UKBB Loh et al. 2018 Nat Genet	361,141
PBC	Primary biliary cirrhosis	Cordell et all. Nat Comm 2015	13,239
Psoriasis	Psoriasis (self-reported)	UKBB Loh et al. 2018 Nat Genet	361,141
RA	Rheumatoid Arthritis	Okada et al. Nature 2013	103,638
Sarcoidosis	Sarcoidosis (self-reported)	UKBB Loh et al. 2018 Nat Genet	361,141
Sjogren	Sjogren's Syndrome (self-reported)	UKBB Loh et al. 2018 Nat Genet	361,141
T1D	Type 1 Diabetes	Inshaw et al. Diabetologia 2021	17,685
T2D	Type 2 Diabetes	DIAGRAM Nat Genet 2018	898,130
Ulc colitis	Ulcerative Colitis (self-reported)	UKBB Loh et al. 2018 Nat Genet	361,141

Simulations to evaluate prediction accuracy To evaluate the properties of our method relative to other methods we perform a series of simulation experiments. We first simulate genotypes for each individual, where each individual i and each locus $m\,(m = 1 : M)$ is independent, and there are no rare SNPs:

$$G_{im} \sim \text{Bin}(2, \text{Unif}[.05, .50])$$

We then draw both context-shared ($\beta_{j.}$) and context-specific (β_{jc}) effect sizes for each SNP from a normal distribution with a Bernoulli random variable I_m controlling the probability

that the m^{th} SNP is causal (i.e. induce sparsity of genetic effects).:

$$I^m \sim \text{Bernoulli}(.05), \quad \beta_{j\cdot}^m \sim \text{N}\left(0, \frac{h^2}{M * \pi}\right) \times I^m, \text{ and } \quad \beta_{jc}^m \sim \text{N}\left(0, \frac{h_c^2}{\lambda * M * \pi}\right) \times I_\lambda^m$$

Here, h^2 and h_c^2 are the context-shared and context-specific heritabilities of expression on gene j. In general, the SNPs with nonzero context-specific effect sizes were subsampled from SNPs with nonzero context-shared effect sizes. We additionally simulate for a subset of contexts some number of truly context-specific eQTLs drawn from Poisson($\lambda = 1$) for randomly selected SNPs that were not eQTLs for the context-shared effects. Finally, we simulate the expression of gene j as follows:

$$E_{jc} = G_j\beta_{j\cdot} + G_j\beta_{jc} + \varepsilon_{jc} \tag{3.8}$$

$$\varepsilon \sim \mathcal{N}(0, \Sigma), \quad \Sigma \in \mathbb{R}^{C\times C} = \begin{bmatrix} \sigma_1^2 & \dots & \sigma_{1,C} \\ \vdots & \ddots & \vdots \\ \sigma_{C,1} & \dots & \sigma_C^2 \end{bmatrix} \tag{3.9}$$

where $\varepsilon \in \mathbb{R}^I$, represents the correlation of environment or intra-individual noise across contexts, $\sigma_c^2 = 1 - h^2 - h_c^2$ is the variances of each context c, and $\sigma_{c_1,c_2} = \rho_{c_1,c_2}\sigma_{c_1}\sigma_{c_2}$ is the covariance of context c_1 and c_2. We generated data under varying levels of context-specific heritability, truly context-specific eQTLs, causal SNPs, and correlation of intra-individual noise across contexts. The number of contexts was set to 20, and to replicate a setting similar to GTEx, the corresponding sample sizes of each ranged from 75 to 410 where individuals were not necessarily measured in every context. In our simulations, we generated one train and one test data set using the above framework. We evaluated the performance of each method by comparing the true and predicted expression in the test data set, using the predictor learned from the training data set.

To assess the effect of additional sharing on a subset of contexts, we also set up a simulation framework using the same generative process as above, only that a subset of contexts also received additional genetic effects. More rigorously, for this subset of contexts (acting as brain contexts in GTEx, for example), expression was generated as in equation (6) with an additional term:

$$E_{jc} = G_j\beta_{j\cdot} + G_j\beta_{jc} + G_j\beta_{j\dot{b}} + \varepsilon_{jc}, \quad \beta^m_{j\dot{b}} \sim \mathrm{N}\left(0, \frac{h^2_{\dot{b}}}{\lambda * M * \pi}\right) \times I^m_\lambda \tag{3.10}$$

where each variable is simulated as before, $\beta^m_{j\dot{b}}$ corresponds to additional genetic effects that are subsampled from SNPs that have a context-shared effect, and $h^2_{\dot{b}}$ is the brain-shared heritability.

Simulations of TWAS performance Using the above generated genotypes and gene expression, we simulated phenotypes to evaluate the performance of each method under the assumed model in TWAS. For a given phenotype, we randomly selected 300 gene-context pairs (100 genes, 3 contexts each) whose expression would comprise a portion of a phenotype. Explicitly, we generated a phenotype as follows:

$$y_i = E_i\delta + \varepsilon \quad \delta \sim N(0, \frac{\sigma^2_{ge}}{300}), \;\; \varepsilon_i \sim N(0, 1 - \frac{\sigma^2_{ge}}{300})$$

Where E_i is the standardized genetic expression of the 300 gene-context pairs for individual i, δ is the length-300 vector of effect sizes for each gene-contexts' expression, σ^2_{ge} is the variance in the phenotype y_i due to cis-genetic gene expression, and ε_i corresponds to environmental effects (or noise) as well as trans-genetic effects for individual i. In our simulations, we varied the heritability of gene expression and fixed variability in the phenotype due to genetic gene expression to .2. To simulate a wide range of genetic architectures, the proportion of heritability of gene expression due to the context-shared effects was sampled from a standard uniform distribution, and the proportion of heritability due to context-

specific effects was (1- the context-shared proportion). Once we generated a phenotype, we performed a TWAS using weights output from each method by imputing expression into a simulated external, independent set of 10000 genotypes that followed the same generation process as in the previous subsection.

3.3 Results

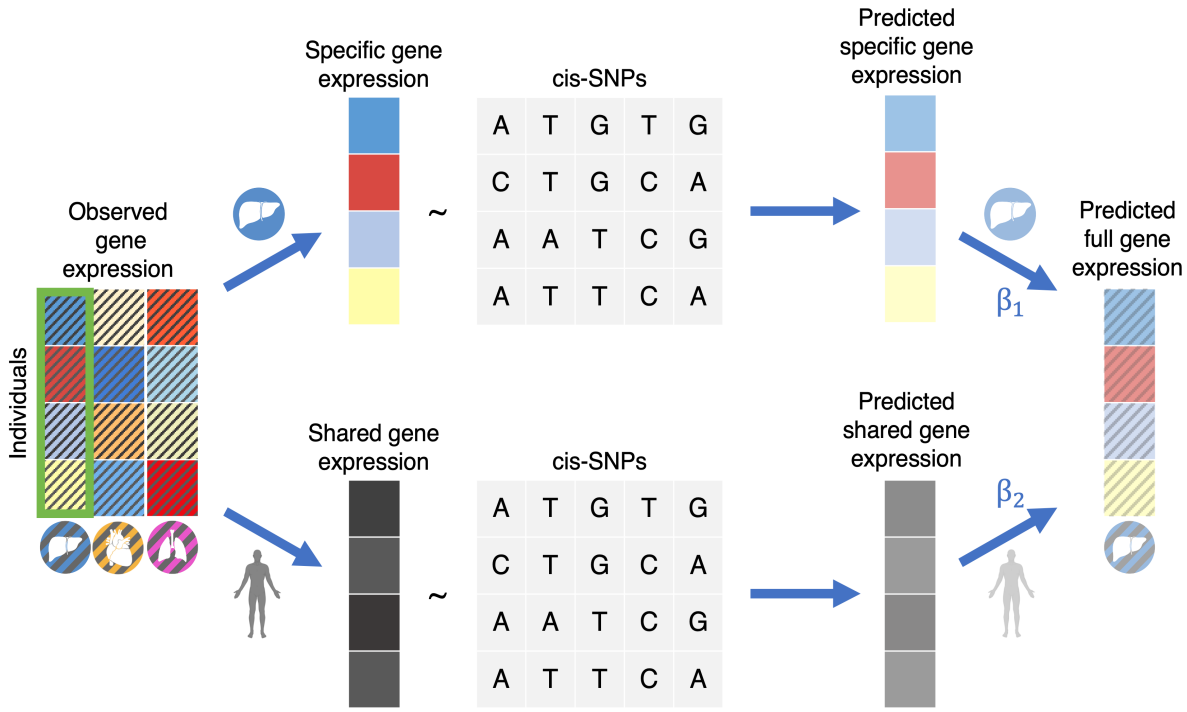

Figure 3.3: An overview of the CONTENT approach. CONTENT first decomposes the observed expression for each individual into context-specific and context-shared components following Lu et al. Then, CONTENT fits predictors for the context-shared component of expression as well as each context-specific component of expression (e.g., liver). Finally, for a given context, CONTENT combines the genetically predicted components into the full model using a simple regression.

3.3.1 Methods overview

We developed CONTENT, a method for generating genetic predictors of gene expression across contexts for use in downstream applications such as TWAS. Briefly, for each indi-

vidual, CONTENT leverages our recently developed FastGxC method [26] to decompose the gene expression across C contexts into one context-shared component and C context-specific components. Next, CONTENT builds genetic predictors for the shared component and each of the C context-specific components of expression using penalized regression. We refer to these predictors as the CONTENT(Shared) and CONTENT(Specific) models. In addition, CONTENT generates genetic predictors of the total expression in each context by combining the context-shared and context-specific genetic predictors with linear regression. We refer to these predictors as the CONTENT(Full) models. A given gene may have CONTENT(Specific), CONTENT(Shared), and/or CONTENT(Full) models depending on the architecture of genetic effects.

We residualized the expression of each gene in each context over their corresponding covariates (e.g. PEER factors, age, sex, batch information) prior to decomposing and then fitting an elastic net with double ten-fold cross-validation for both CONTENT(Shared) and CONTENT(Specific). We examined the number of significantly predicted genes as well as the prediction accuracy (in terms of adjusted R^2) between the cross-validation-predicted and true gene expression per gene-context pair. To properly control the FDR for each method across contexts and genes, we employed a hierarchical FDR correction [116, 117] (Figure 3.2 and Methods). We note that groups of contexts may comprise additional sources of pleiotropy (e.g. in GTEx the group of brain tissues may have their own shared effects in addition to the overall tissue-shared effects). The decomposition of CONTENT is flexible and can account for both levels of pleiotropy among contexts (see Supplementary Methods).

3.3.2 CONTENT is powerful and well-calibrated in simulated data.

We evaluate the prediction accuracy of CONTENT in a series of simulations and compare its performance to the context-by-context approach[103, 104], which builds predictors by fitting an elastic net in each context separately, as well as UTMOST[109], which builds predictors over all contexts simultaneously using a group LASSO penalty. Implicitly, we compare to the method from [114] which decomposes expression into orthogonal context-shared and

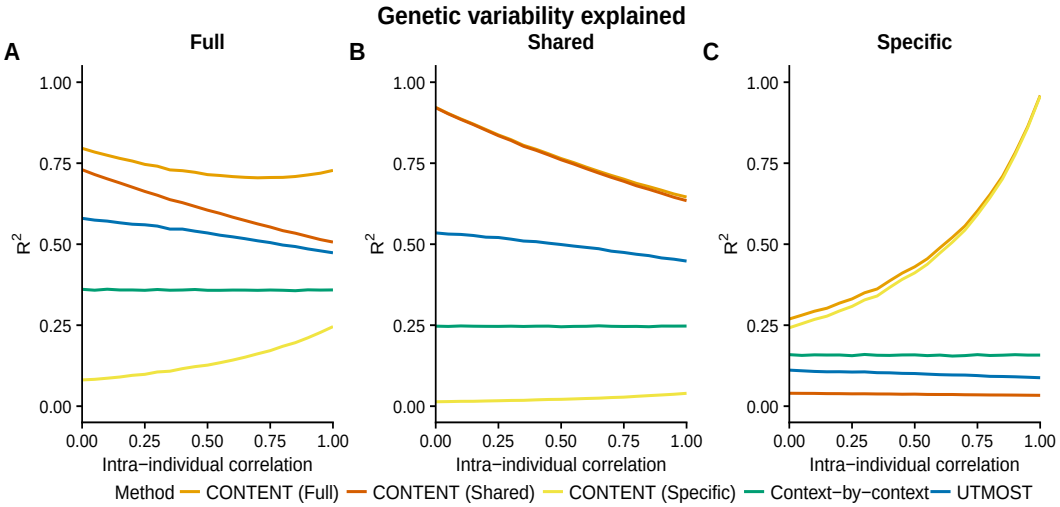

Figure 3.4: CONTENT is powerful and well-calibrated in simulated data. Accuracy of each method to predict the genetically regulated gene expression of each gene-context pair for different correlations of intra-individual noise across contexts. Mean adjusted R^2 across contexts between the true (A) full (context-specific + context-shared), (B) shared, and (C) specific genetic components of expression and the predicted component for each method and for different levels of intra individual correlation. The context-by-context approach and UTMOST output only a single predictor, and we show the variability captured by this predictor for each component of expression. CONTENT, however, generates predictors for all three components of expression, and notably, CONTENT(Specific) and CONTENT(Shared) capture their intended component of expression without capturing the opposite (i.e. the predictor for CONTENT(Specific) is uncorrelated with the true shared component of expression and vice versa). We show here the accuracy for each component and method on gene-contexts with both context-shared and context-specific effects, but show in Figure **??** the accuracy for all gene-contexts pairs.

context-specific components, as the CONTENT(Shared) and CONTENT(Specific) models are related to these components (See Methods). We omit comparison to other TWAS methods as many of them are built on the same framework as the context-by-context approach, or require external data, such as curated DNase I hypersensitivity measurements [110, 111, 108].

We used simulation parameters from GTEx, the largest multi-context eQTL study to-date, as a guideline. Specifically, we generated gene expression and genotype data such that context-specific genetic effects mostly lie on the same loci as context-shared eQTLs, and context-specific eQTLs without context-shared effects are rare [6, 26]. Intuitively, this framework assumes that, most often, SNPs affect expression of a gene in all contexts, but to a

different extent in each context (rather than, for example, acting as an eQTL in only a single context). We varied the proportion of contexts with context-specific heritability, the number of context-specific eQTLs without a context-shared effect, the number of causal SNPs, and the intra-individual residual correlation while keeping the number of genes (1000), contexts (20), *cis*-SNPs (500) and the proportion of context-shared and context-specific heritability constant (.3 and .1 respectively).

Throughout our simulations, CONTENT significantly outperformed the context-by-context and UTMOST approaches in terms of prediction accuracy of the total genetic contribution to expression variability (Figure 3.4A). The average increase in adjusted R^2 between the true genetic component of expression and the CONTENT(Full) predictor was .22 over UTMOST ($p<2e\text{-}16$ paired two-way t-test) and .48 over the context-by-context approach ($p<2e\text{-}16$ paired two-way t-test). Across nearly the entirety of parameter settings, CONTENT generated context-specific components that were uncorrelated with the true context-shared components (mean adjusted R^2=.023, and vice versa .026; Figure 3.4B,C). This property is central to the objective as it reduces confounding from pleiotropy in downstream applications such as context fine-mapping. As expected, the previous methods failed to disentangle the context-specific and context-shared components (Figure 3.4B,C), since they were not developed with this property in mind. Our results were consistent under different values of the simulation parameters (figures omitted for brevity).

3.3.3 CONTENT improves prediction accuracy over previous methods in the GTEx and CLUES datasets

We next evaluated CONTENT, the context-by-context approach, and UTMOST in terms of prediction accuracy and power across 22,447 genes measured in 48 tissues of 519 European individuals in the bulk RNA-seq GTEx data set [1, 6, 2]. Due to computational issues, UTMOST was examined only on 22,307 genes rather than the entire data set of 22,447 genes. (On this smaller subset of genes, the results were nearly identical to those presented here.) We also examined, for the first time in a large-scale TWAS context, a single-cell

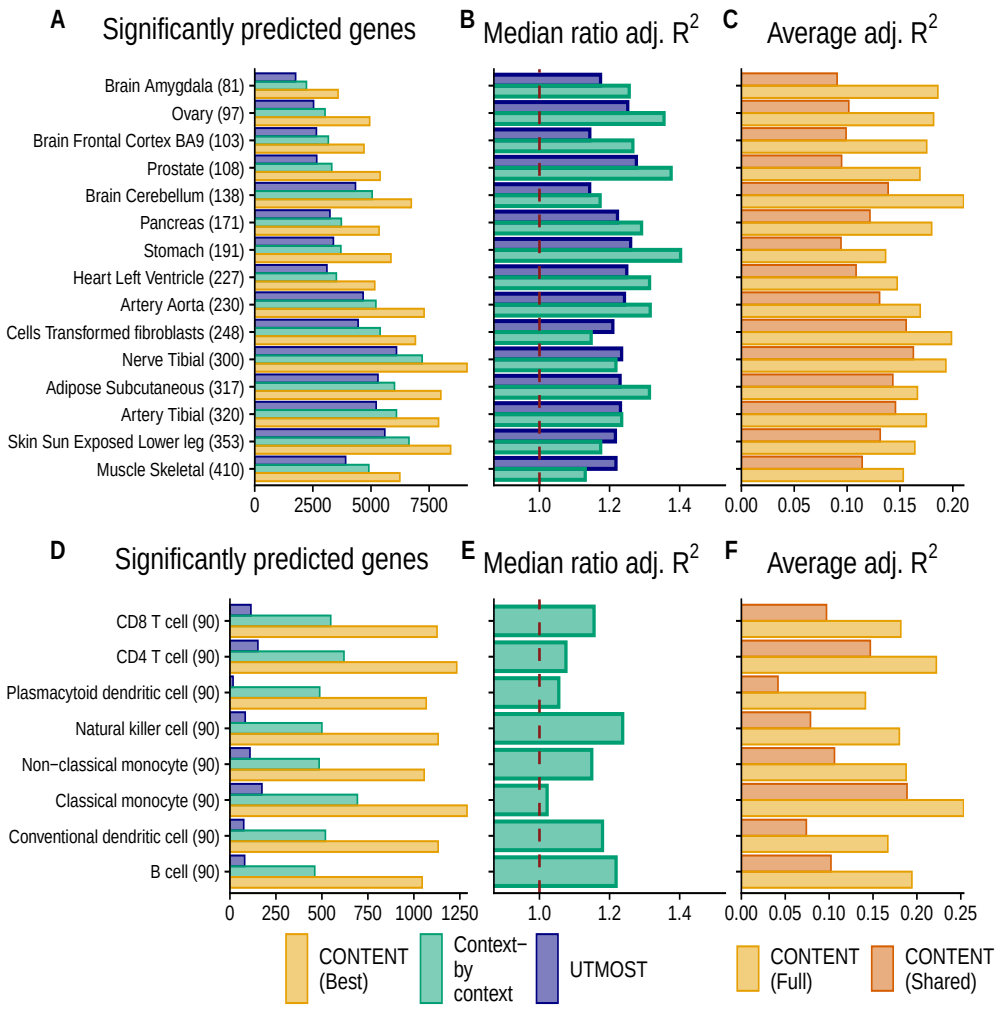

Figure 3.5: CONTENT outperforms existing approaches in the GTEx and scRNA-seq CLUES datasets. (A,D) Number of genes with a significantly predictable component (hFDR $\leq 5\%$) in GTEx (A) and CLUES (D); the sample sizes for each context are included in parentheses. (B,E) Ratio of expression prediction accuracy (adjusted R^2) of the best-performing cross-validated CONTENT model over the context-by-context (green) and UTMOST (blue) approaches (median across all genes significantly predicted by at least either method). Numbers above one indicate higher adjusted R^2 and thus prediction accuracy for CONTENT. (C,F) Prediction accuracy of CONTENT(Full) and CONTENT(Shared) when a gene-tissue has a significant shared, specific, and full model.

RNAseq data set from the California Lupus Epidemiology Study (CLUES) [118, 119]. The CLUES data set contained 9,592 genes measured in 9 cell types in peripheral blood from 90 individuals.

In GTEx, CONTENT identified more gene-tissue pairs with a significantly predictable

genetic component of expression (278,101 over 20,506 genes) than the context-by-context approach (195,607 over 17,723 genes) and UTMOST (167,865 over 11,442 genes) at an hFDR of 5% for all approaches. These results also held when using the traditional FDR approach within each context separately for all approaches. We also compared the performance of each method on the union of genes that were significantly predicted (hFDR $\leq$ 5%) by at least one method. As CONTENT can generate up to three models (Shared, Specific, Full) for a given gene-tissue pair, and because each gene may have its own unique architecture (i.e. different proportions of specific or shared heritability), we selected the model that achieved the greatest cross-validated adjusted R^2. CONTENT greatly outperformed the context-by-context and UTMOST approaches across all tissues (average 28% and 22% increase in adjusted R^2 across tissues and genes; Figure 3.5). Further, for genes with significant CONTENT(Shared), CONTENT(Specific), and CONTENT(Full) predictors, prediction accuracy increases substantially with the addition of the context-specific component to the context-shared component (average gain of CONTENT(Full) over CONTENT(Shared) adj. R^2 of 55.92%), emphasizing the need to extend previous approaches[114] with CONTENT(Full) to build a powerful predictor.

Within the single-cell CLUES data set, CONTENT again outperformed the context-by-context (in this case, cell type-by-cell type) and UTMOST approaches, discovering 9,080 heritable gene-cell type pairs (5,067 genes) whereas the context-by-context model and UTMOST found 4,314 (2,355 genes) and 804 (288 genes) respectively. The average improvement in adjusted R^2 of CONTENT over the context-by-context model was 13.6%. In gene-cell type pairs with significant CONTENT(Full), CONTENT(Specific), and CONTENT(Shared) models, CONTENT(Full) improved the adjusted R^2 over CONTENT(Shared) by 104.09%. Once more, the improvement in variability explained when including both the cell type-specific and cell type-shared components highlights the need to consider both components simultaneously when building a predictor.

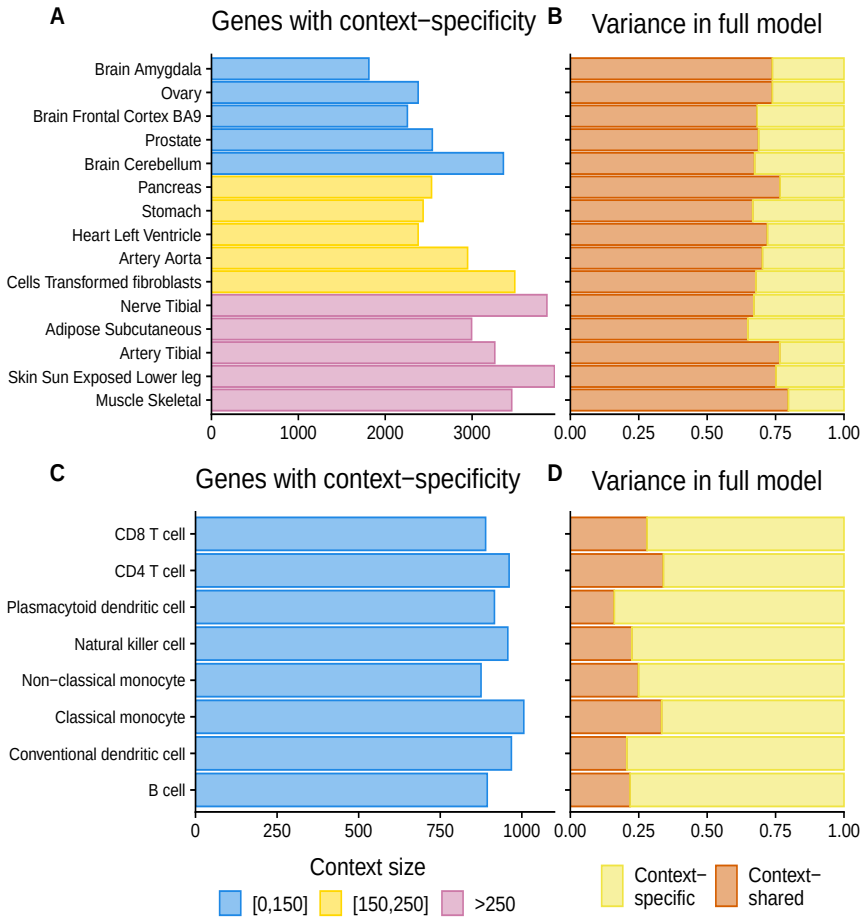

Figure 3.6: Contribution of context-specific genetic regulation in GTEx and CLUES. (A,C) Number of genes with a significant (FDR$\leq$ 5%) CONTENT(Specific) model of expression in GTEx (A) and CLUES (C). Color indicates sample size of context. (B,D) Proportion of expression variance of CONTENT(Full) explained by CONTENT(Specific) and CONTENT(Shared) for genes with a significant CONTENT(Full) model.

3.3.4 CONTENT discovers significant context-specific components of expression in bulk multi-tissue and single-cell datasets.

Given the ability of CONTENT to disentangle context-shared and context-specific variability, we examined the context-specific components of expression in GTEx and CLUES. In GTEx, CONTENT discovered 128,985 gene-tissue pairs (19,765 genes) with a significant context-specific genetic component of expression (Figures 3.6). As with previous reports [126, 26], we found that testis was the tissue with the greatest number of tissue-specific genetic components. Nonetheless, we observe that the tissues with larger sample sizes more

frequently had significant context-specific components. Consistent with previous works that have discovered extensive eQTL sharing across tissues [127, 6, 126], we found that in gene-tissue pairs with a CONTENT(Full) model, the variability explained was dominated by CONTENT(Shared) model—across tissues, the context-shared component explained on average 70% of the variability explained by CONTENT(Full).

In the CLUES data set, CONTENT discovered 7,466 gene-cell type pairs (4,658 genes) with a significant cell type-specific component of expression (hFDR $\leq$ 5%). We found that all cell types had a similar number of cell type-specific components, and emphasize that the sample size across all cell types was equivalent. Interestingly, in genes with a CONTENT(Full) model, the variability was often dominated by the cell type-specific variability (average 75% of the explained variability), unlike GTEx, in which the average tissue-specific variability explained only 30% of total variance. Consequently, we found that within the 20,433 genes in GTEx with any genetic component, 51.50% (10,522) had a significant shared component, whereas of the 5,067 genes in CLUES with a genetic component, only 14.25% (722) had a shared component. This is consistent with complex cell type heterogeneity in bulk tissues [128] since there is more power to discover eQTLs with pleiotropy across the underlying cell types.

3.3.5 CONTENT more accurately distinguishes disease-relevant genes than traditional TWAS approaches in simulated data.

We performed a simulation study in which we evaluated the sensitivity, specificity, and power of CONTENT, UTMOST, and context-by-context to implicate the correct gene in TWAS. In our experiments, we simulated a phenotype in which 20% of the variability was composed of the genetically regulated expression of 300 randomly selected gene-context pairs (100 genes and 3 contexts each). We simulated gene expression for 1,000 genes across 20 contexts as before, however, to capture a range of genetic architectures in the simulation, for each gene, we sampled from a standard uniform distribution to determine the proportion of shared variability. We varied the heritability of gene expression and considered power as a method's

ability to discover the correct genes associated with a phenotype. To compare power, we calculated the area under receiver-operating curve (AUC) using the maximum association statistic for a given gene across contexts.

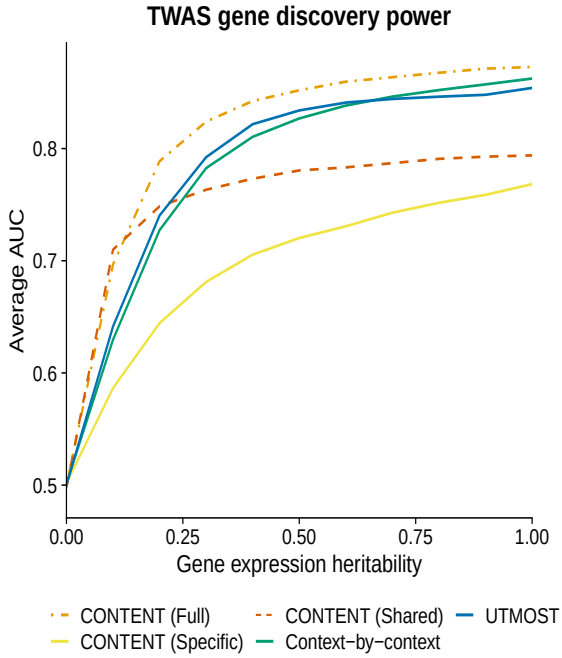

Figure 3.7: CONTENT(Full) is powerful, sensitive, and specific in simulated TWAS data. Average AUC from 1,000 TWAS simulations while varying the overall heritability of gene expression. Each phenotype (1,000 per proportion of heritability) was generated from 300 (100 genes and 3 contexts each) randomly selected gene-context pairs' genetically regulated gene expression, and the 300 gene-context pairs' genetically regulated expression accounted for 20% of the variability in the phenotype. In genes with low heritability, CONTENT(Shared) performed similarly to CONTENT (Full), however CONTENT(Full) was the most powerful method in discovering the correct genes for TWAS across the range of heritability. CONTENT(Full) was significantly more powerful than UTMOST and the context-by-context approach at all levels of heritability.

Across simulations, CONTENT(Full) was the highest powered in terms of gene discovery (Figure 3.7). CONTENT(Shared) performed very similarly to CONTENT(Full) in the setting with the lowest heritability, however, our simulations show the necessity for CONTENT(Full) as it substantially outperforms both CONTENT(Specific) and CONTENT(Shared) across a range of heritabilities. Moreover, CONTENT(Full) significantly outperformed both the context-by-context approach and UTMOST. Specifically, the range

of percent change in AUC of CONTENT(Full) over previous methods was as follows: CONTENT(Shared) 1.9%-9.9%; CONTENT(Specific) 13.6%-22.4%; UTMOST 2.2%-8.6%; context-by-context 1.2%-10.6%. Generally, we observed that CONTENT(Full) was its most powerful for genes in which there was both shared and specific effects, UTMOST was its most powerful in settings with high sharing, and the context-by-context approach was its most powerful in settings with low sharing and high specificity of genetic effects within contexts. As with previous methods [109], we performed simulations in which the causal context(s) has been observed. In real data applications, this may not occur, and in such cases, further complexities may arise due to genetic correlation. As they are issues of association fine-mapping, the complexities posed by missing tissues and cell types are beyond the scope of this manuscript, and we therefore leave the development of relevant methodology as future work.

3.3.6 Application of CONTENT to TWAS yields novel discoveries over previous methods.

We performed TWAS across 22 complex traits and diseases collected from a variety of GWAS [129, 130, 131, 132, 133, 134, 5, 135, 136, 137, 138, 139, 140, 141] using weights trained by CONTENT, UTMOST and the context-by-context approach on GTEx and CLUES. We passed forward weights to FUSION-TWAS[103]—a software that performs TWAS using GWAS summary statistics, user-specified gene expression weights, and an LD reference panel—for a gene-context pair if the pair's expression was predicted at a nominal p-value less than .1.

Across all traits at an hFDR of 5%, CONTENT discovered a median of 51% (range of 5 to 178%) and 135% (51-400%) more associations (unique TWAS loci) than the context-by-context approach and UTMOST respectively with GTEx weights, and 62% (0-289%) and 101% (47-600%) more loci than the context-by-context approach and UTMOST respectively with weights built from the CLUES dataset (Table 3.2). We find that, with GTEx weights, the associations implicated by the context-by-context approach had more overlap with the associations implicated by CONTENT(Specific) (median Jaccard similarity (JS)

across traits =.419) than CONTENT(Shared) (JS=.234). This is consistent with our simulation results in which the context-by-context approach was most powerful in cases of high context-specificity and low context-sharing. Conversely, the associations discovered by UTMOST, which leverages pleiotropy, had slightly higher overlap with CONTENT(Shared) (JS=.221) than CONTENT(Specific) (JS=.177). With CLUES weights, the context-by-context approach again had greater similarity with CONTENT(Specific) (JS=.291) than CONTENT(Shared) (JS=.098), however UTMOST discovered TWAS genes that had similar overlap between CONTENT(Shared) (JS=.119) and CONTENT(Specific) (JS=.135). As UTMOST, CONTENT, and the context-by-context approach discovered both overlapping and unique associations, we suggest that the approaches complement—rather than replace—one another.

We next compared the different CONTENT models to understand their properties in real data. With GTEx weights, CONTENT(Full) replicated an average of 98.3% and 67.3% of the associations discovered by CONTENT(Shared) and CONTENT(Specific) respectively (hFDR $\leq$ 5%). CONTENT(Full) replicated an average of 81.2% and 61.6% of the associations discovered by CONTENT(Shared) and CONTENT(Specific) respectively with the CLUES weights. Notably, CONTENT(Full) is the best predictor out of all the CONTENT models on average, and particularly when there exist both shared and specific effects. Consequently, across all traits, the inclusion of CONTENT(Full) with CONTENT(Shared) and CONTENT(Specific) led to an average increase of 12% and 21% in the number of genes with significant TWAS associations with GTEx weights and CLUES weights respectively.

We investigated the genes implicated by CONTENT(Full) that were not significant in CONTENT(Shared) or CONTENT(Specific) and found that many of the discoveries replicated known gene-trait associations. For example, CONTENT(Full) discovered a significant association of fasting glucose levels and CAMK2 (p=2.44e-23, brain cortex), a gene responsible for calcium signaling and regulation of hepatic glucose production [142], as well as BLVRA (4.21e-06, CD8 T cell), a gene involved in insulin signaling and likely metabolic syndrome [143]. Furthermore, CCL2, which is thought to be involved in HDL internalization and choles-

Table 3.2: CONTENT outperforms existing methods in TWAS across 22 complex traits and diseases. TWAS results (unique loci, merging genes within 1MB) across 22 complex traits and diseases using weights output by CONTENT, UTMOST, and the context-by-context method. CONTENT(All) refers to the collection of all loci output by at least one CONTENT model. CONTENT(Full) added an average of 15% and 19% of gene-trait discoveries over the CONTENT(Shared) and CONTENT(Specific) approaches together at an hFDR of 5% in GTEx and CLUES respectively. See Supplementary Table 3.1 for GWAS trait information.

	GTEx						CLUES					
Trait	Tissue-by-tissue	UTMOST	CONTENT (All)	CONTENT (Full)	CONTENT (Specific)	CONTENT (Shared)	Tissue-by-tissue	UTMOST	CONTENT (All)	CONTENT (Full)	CONTENT (Specific)	CONTENT (Shared)
AD	17	9	**24**	20	20	11	7	5	**15**	9	13	3
Asthma	155	90	**237**	181	181	67	74	63	**127**	101	104	34
Bipolar	42	45	**83**	63	63	39	9	14	**35**	20	25	5
CAD	10	11	**23**	18	18	8	6	6	**10**	7	6	0
CKD	26	19	**42**	31	31	18	2	4	**6**	5	5	1
Crohn's	77	63	**95**	73	73	47	27	22	**44**	30	37	9
Eczema	32	13	**57**	44	44	10	8	5	**11**	9	9	3
FastGlu	16	8	**19**	12	12	7	3	3	6	6	6	0
HDL	58	29	**79**	60	60	36	21	14	**28**	23	25	6
IBS	9	5	**25**	20	20	3	3	1	**7**	5	6	1
LDL	89	57	**132**	107	107	58	47	29	**51**	40	44	14
Lupus	93	54	**129**	94	94	51	36	27	**58**	42	48	11
MDD	99	79	**169**	132	132	62	20	29	**47**	32	39	3
MS	20	10	**42**	32	32	9	9	7	**11**	8	10	5
PBC	62	42	**65**	55	55	33	21	14	**30**	24	26	6
Psoriasis	47	22	**58**	46	46	16	13	10	**21**	17	16	6
RA	73	56	**99**	79	79	46	40	20	**51**	33	45	9
Sarcoidosis	19	13	**30**	27	27	8	6	4	6	6	4	2
Sjogren	17	9	**31**	25	25	6	4	2	**7**	6	6	1
T1D	77	64	**109**	88	88	49	26	23	**41**	36	29	13
T2D	193	115	**246**	208	208	112	76	76	**112**	77	98	17
Ulc colitis	16	10	**40**	30	30	7	5	4	**11**	9	7	2

terol efflux [144], was not implicated by either CONTENT(Shared) or CONTENT(Specific), but was implicated in the TWAS of HDL with CONTENT(Full) (p=2.30e-08, small intestine terminal ileum). CONTENT(Full) also discovered a significant association of F2 (prothrombin) and primary biliary cirrhosis (PBC) (1.47e-07, liver), whereas CONTENT(Shared) and CONTENT(Specific) did not; PBC patients have been shown to have higher prothrombin times than controls [145]. Moreover, CONTENT(Full) discovered an association of GIT1—a gene involved with synaptic transmission and plasticity[146, 147]—with bipolar disorder (BIP; B cell, p=3.20e-06) as well as an association of GSDMB—a gene involved with airway remodeling and airway-hyperresponsiveness[148]—and asthma (CD4 T cell, p=1.25e-20).

Moreover, the genes implicated by CONTENT but neither UTMOST nor the context-by-

context approach (at an hFDR of 5%) replicated previously associated genes-trait pairs, several of which with known biological relationships to the trait of interest. Within Alzheimer's disease, these genes included VGF[149], FZD4[150], and TRPV6 (a transient receptor potential channel) [151, 152] with the GTEx weights, as well as IRF7[153] and GANC[154] with CLUES weights. Additionally, in Crohn's disease, CONTENT implicated the following genes, whereas previous methods did not: STAT3[155] and CTBP2[156] with GTEx weights, as well as ATG16L[157] and PKAR2A[158] using CLUES weights. For major depression disorder (MDD), CONTENT implicated SYN2M[159] and CYB56AD1[160] using GTEx weights, and GAB1 [161], TLR4[160] and ARL3[162] using CLUES weights.

As the individuals comprising the GTEx and CLUES datasets are disjoint, we also investigated whether using both datasets could highlight relevant biological genes (akin to a replication study). We first examined LDL genes and found SORT1, which alters plasma LDL levels (GTEx min. p=2.15e-251, CLUES min. p=2.41e-19) [163, 164, 165]. We next found an association between S100A4, S100A8, S100A10, S100A11 as well as S100A12 (part of the epidermal differentiation complex) and Eczema using both datasets (S100A10 p=2.78e-41, p=2.90e-11)[166, 167]. Additionally, when we looked at discoveries made with GTEx and CLUES weights for Alzheimer's disease, we found MARK4 (p=8.72e-20, p=6.39e-63), a gene associated with tau phosphyrlation in granulovacuolar degeneration bodies [168]. Finally, both sets of weights produced a significant association of immune checkpoint gene CTLA4 (p=1.71e-11, p=2.28e-21) with Rheumatoid Arthritis[169].

While CONTENT discovered substantially more loci and genes than previous approaches, we also wished to verify that it does not enrich for false positives. To do so, we performed an analysis similar to one carried out by Ndungu et al. [170]. Briefly, Ndungu et al. evaluated the extent to which TWAS associations may be driven by horizontal pleiotropy or linkage disequilibrium by examining TWAS associations for a set of genes with a known causal relationship to a set of metabolites. In our analyses, we examined the within-locus ($\pm$ 1Mb) rank of the causal TWAS gene with its suspected metabolite when using weights built by CONTENT and the context-by-context approach on the GTEx dataset. To order

genes within a method, we first filtered for statistically significant gene-context-metabolite associations, then sorted genes by their maximum absolute TWAS association statistic between a given metabolite across contexts (and models for CONTENT). In line with our applications of TWAS to GTEx and CLUES, CONTENT discovered additional loci that were not discovered by the context-by-context approach (39 compared to 36 of 58 known gene-metabolite pairs). Moreover, despite having more models built per locus, CONTENT ranked the known causal gene similarly to the context-by-context approach on the intersection of gene-metabolite pairs discovered by both methods (CONTENT average rank of 2.257 compared to context-by-context rank of 2.371, where a ranking of 1 is ideal).

3.4 Discussion

We introduced CONTENT, a computationally efficient and powerful method to estimate the genetic contribution to expression in multi-context studies. CONTENT can distinguish the context-shared and context-specific components of genetic variability and can account for correlated intra-individual noise across contexts. Using a range of simulation and real studies, we showed that CONTENT outperforms previous methods in terms of prediction accuracy of the total genetic contribution to expression variability in each context. Interestingly, we also found that when there exists a gene with a genetic component of expression, the heritability is often dominated by the context-specific effects at the single-cell level, but at the tissue level, the heritability is dominated by the context-shared effects. Finally, CONTENT was more powerful, specific, and sensitive than previous approaches in applications to TWAS.

Using weights trained by CONTENT, UTMOST and the context-by-context approach, we discovered 12,150 unique gene-trait associations through TWAS. To our knowledge, we present the first application of TWAS trained on a single-cell RNAseq dataset for a collection of 90 individuals' PBMCs. For both the weights generated by GTEx and CLUES, CONTENT was largely more powerful than UTMOST and the context-by-context approach in TWAS. However, we emphasize that the approaches often capture genes unique to each approach. Each method may therefore complement each other and may be combined in

TWAS to maximize the number of discoveries made as different methods are likely favorable under different genetic architectures. Though we show that CONTENT may be useful in fine-mapping the specific tissue relevant for a TWAS association in simulations, we note that fine-mapping to the correct tissue in real data is a particularly difficult task. For example, throughout this manuscript, we assume that the causal tissue is included in the measured tissues, however, when this is not the case, CONTENT and all TWAS approaches may associate an incorrect, correlated tissue. For example, in the case of chronic kidney disease, CONTENT implicated GATM–a gene thought to be involved with kidney disease and GFR levels [171, 172, 173]–however, the significant association was within the thyroid. This may be due to the fact that kidney expression is not measured in this version of the GTEx dataset. Future work may explore using the CONTENT-trained weights and jointly fitting all TWAS Z scores, or otherwise accounting for missingness.

We also leveraged recently developed methodology for controlling the false discovery rate when summarizing significantly predicted genes, gene-contexts, and TWAS associations [116, 117]. This approach has been shown to effectively control the FDR across contexts in eQTL studies, and to our knowledge, it is the first time such an approach has been used to effectively control the FDR when predicting expression values and when making discoveries using TWAS. While our analyses focused on the comparison of CONTENT, UTMOST, and the context-by-context approach, we emphasize that by using this type of false discovery correction, all methods can be used in combination with one another, rather than in replacement of one another. For downstream analysis, combining all prediction methods is crucial, as certain genes or gene-context pairs may be (better) predicted by one method and not others. In the GTEx data for example, when we included models built by UTMOST and the context-by-context approach to the correction scheme for CONTENT, the number of genes for which there was a significant model for a given tissue increased on average by 7.56%.

Importantly, neither UTMOST nor the context-by-context method distinguishes the context-specific and context-shared components of genetic effects on expression. Implic-

itly, by modeling all contexts independently, the context-by-context fit is best-suited for cases in which there is no effect-sharing across contexts. As UTMOST considers all contexts simultaneously, its power is maximized in cases where the genetic effects are mostly shared. Additionally, these methods do not account for the shared correlated residuals between samples, thus they do not maximize their predictive power.

While a previous approach proposed by Wheeler et al. [114] does model the correlated intra-individual noise, CONTENT offers several advantages. The previous decomposition does not include an option to leverage both the context-shared and context-specific components of expression to form a final predictor of the observed expression for a given context. We show that this is especially crucial in the context of single-cell data wherein the prediction accuracy for a given gene-context increases drastically when using both components (Figure 3.5). Further, without properly combining both components (e.g. via regression), the context-specific genotype-expression weights produced by the previous decomposition may have the incorrect sign, as they are considered residuals of the context-shared component and are not properly re-calibrated to the observed expression. Unlike the novel decomposition proposed by CONTENT, this previous approach also does not intuitively allow for additional sources of pleiotropy or effects-sharing (see Supplementary Text for discussion of brain level sharing in GTEx). Finally, the decomposition used in the previous method is based on a linear mixed model fit on a per-gene basis, and is therefore much less computationally efficient.

Notably, a limitation of TWAS methods in general is interpretability, as associations may be confounded by linkage disequilibrium or horizontal pleiotropy [174, 170]. We emphasize that CONTENT discovered substantially more independent loci than previous methods, however, since CONTENT is more powerful than previous methods, it may build more models within a given locus relative to previous approaches. We performed a brief set of analyses in line with Ndungu et al. [170], in which we evaluated the ability of TWAS approaches to associate the suspected causal gene to a collection of metabolites. Despite CONTENT building more models than the context-by-context approach, it prioritized suspected genes

the same as or better than the context-by-context approach in addition to discovering several more loci than did the context-by-context approach (Supplementary Table **??**). We therefore conclude that, similarly to GWAS fine-mapping studies, resolution of downstream TWAS fine-mapping methods (e.g., FOCUS [174] should increase with the use of our models, as our gain in performance is akin to that expected from an increase in sample size. Moreover, since CONTENT discovers additional loci over previous approaches, it undoubtedly will present additional useful information for such studies.

In this manuscript we focused on prediction of the total genetic contribution to expression as well as the context-shared and context-specific components of expression. Nonetheless, future work using the methodology presented here can be extended to a wide variety of problems. Primarily, the decomposition can be used to efficiently estimate Gene×Context heritability using existing software for heritability estimation, e.g. *GCTA* [113], on the decomposed components offering computational speed up over existing methods for cross-context heritability estimation [127]. Additionally, the decomposed components from CONTENT may also be included in previous approaches, e.g. UTMOST, to gain further power. Further, by training each method on the single-cell level data, we offer researchers the means to pursue their own association analyses at a lower level of granularity than was previously available.

Notably, we found that single-cell data may have lower levels of effects-sharing than tissue-level data. While this may be due to genuine biological differences in genetic regulation, this finding is also consistent with a large degree of sharing of cell types across contexts. For example, endothelial cells can be found in tissues such as breast, endometrium, esophagus, eye, heart muscle, liver, lung, ovary, pancreas, placenta, prostate, skeletal muscle, and skin and often make up a substantial fraction of the collected tissue [175, 176]. We believe our work is consistent with this observation: primarily, the proportion of genes with a heritable component of expression that also have a shared component is substantially lower at the single cell level. What's more is that the ability to discover context-specific components of expression is indeed related to sample size in the GTEx dataset. Despite the above, and

having a lower number of individuals in the single-cell data, we discover a greater proportion of genes with a context-specific component than in GTEx. Further, when there exists a CONTENT(Full) model, it is dominated by the specific variability at the single-cell level, whereas it is dominated by the shared variability at the tissue level. Nonetheless, as this finding, to our knowledge, was previously unappreciated, it warrants further investigation.

In summary, we present a novel approach for generating context-shared and context-specific predictors that is much simpler than previous approaches [114, 26]. Moreover, unlike previous methods, we offer a way to combine both predictors, as well as extend the decomposition to additional levels of pleiotropy. Finally, we show utility of existing hierarchical FDR correction methods to properly adjust for analyses that take advantage of multiple methods as well as investigate genes in the space of multiple contexts. The increased prediction accuracy, specificity, computational speed, and hierarchical testing framework of CONTENT will be paramount to unveiling context-specific effects on disease as well as uncovering the mechanisms of context-specific genetic regulation.

CHAPTER 4

Methylation risk scores are associated with a collection of phenotypes within electronic health record systems

4.1 Background

Widespread adoption of electronic health record systems coupled with an increasing interest in hospital biobanking systems has spurred research efforts spanning machine-learning and genomics communities[30, 31, 32, 177, 178, 179, 180]. These efforts have produced increasingly accurate imputation (current state) and prediction (future state) of patient phenotypes from medical records [181, 182] and polygenic risk scores [30, 31, 32, 33, 34, 35, 28, 36], and are already being investigated in translational contexts [37, 38, 39, 40]. For example, recent work has shown that machine learning can leverage high-dimensional data to aid in the prediction of a multitude of clinical phenotypes including cardiac function and arrhythmia [183, 184, 185], post-operative complications [181, 182], sepsis [186], breast cancer [187, 34], and prostate cancer [188]. Nonetheless, a genetics-based predictor such as the polygenic risk score may be limited in predictive utility as it does not account for changes in disease risk—for example, due to age, or changes in environment—throughout one's lifespan [28].

In this work we examine the potential for epigenetic information to improve phenotype inference in combined biobank-EHR systems. As DNA methylation, henceforth referred to as simply "methylation", is affected by both genetics and environment—such as lifestyle choices, diet, exercise, and smoking status—it captures multi-factorial information about predispositions to clinical conditions [189, 19, 23, 190, 191, 192, 193]. Moreover, methylation is readily available for use in existing biobanks that collect DNA samples, and recent advance-

ments in methylation profiling technologies have enabled an abundance of large-scale studies of methylation and its role as a biomarker for a variety of phenotypes and health-related outcomes [189, 194, 195, 196, 197, 198, 199, 193]. It is therefore a natural candidate for an extension of PRS, and we hypothesized that methylation can be used to complement genetics as a clinical prediction tool. To that end, we have generated and evaluated methylation risk scores (MRS), which are linear combinations of CpG methylation states[189].

To comprehensively investigate the utility of MRS and characterize its properties, we conducted a study of 607 EHR-derived phenotypes spanning medications (e.g. vasopressers, glucocorticosteroids, fluoroquinolones), labs (e.g. creatinine, glucose, prothrombin time), and diagnoses (e.g. T2D, bacterial pneumonia, anemia) that were available for a sufficient number of patients in the cohort. The cohort contained 831 patients—to the best of our knowledge, the largest epigenetic biobank dataset to date (including genetics, methylation, and EHR)—from the UCLA Health ATLAS cohort across a wide range of ages (18-90), racial and ethnic groups, and overall health (including patients ascertained on kidney and heart disease, with matched controls), with corresponding genetic and EHR data. This provides the opportunity to study the potential contribution of methylation to larger biobanks and in multiple clinical contexts. We find that the MRS-based imputations were more informative compared to PRS in 84 (92%) medications, 32 (94%) labs, and 123 (82%) diagnoses, more than doubling the imputation accuracy in over half of the outcomes considered. We also show that the MRS improves the imputation accuracy over PRS for cases in which the PRS is trained on very large external biobanks (roughly 3 orders of magnitude larger), as opposed to 831 samples that are available in this study. We observe that MRS improves over PRS learned from large biobanks in 40% of the tested phenotypes. Further, as our cohort was ethnically diverse, we performed replicability analyses within each racial and ethnic subset of our data. We broadly showed the replicability of the five best-imputed (by MRS) medications, labs, and diagnoses—46% and 100% of which replicated in (n=118) non-white Hispanic-Latino- and (n=543) white non-Hispanic-Latino-identifying individuals respectively. Finally, we demonstrate the ability of MRS to transfer between methylation

arrays and cohorts by conducting an association study of kidney-related MRS in an external diabetic nephropathy EWAS [200], where the minimum replication p-value was 2.72×10^{-7}.

These results provide evidence for the utility of methylation in phenotype imputation in general, and in biobank settings in particular. However, the promise of clinical translation of genomic risk scores, including PRS or MRS, is highly dependent on the clinical context of the patient. There is a large body of work investigating phenotype imputation and prediction in clinical settings using EHR data alone, typically with machine learning techniques, without any genomic data. To the best of our knowledge, the question of whether genomic data can be used to complement such algorithms has not been studied. Since the application of MRS or PRS to clinical data without taking into account the EHR data provides a limited clinical utility, this is a natural question.

Here, we demonstrate that MRS can be used in conjunction with EHR data to improve the imputation of clinical data of patients. Critically, most machine learning approaches rely on imputation because of the inability of such algorithms to process missing data, making accurate imputation a crucial step. We found that the combination of MRS with a gold standard imputation approach—SoftImpute [201]—for clinical data imputation, provides improved accuracy (R^2) in 37.3% of the examined phenotypes with a median increase of 47.6%. This result provides the potential to improve machine learning algorithms that use the EHR data, by complementing the data with methylation information for the patients.

In summary, our results quantify the contribution of methylation information in clinical settings, both in isolation and in conjunction with the EHR data, and they demonstrate the potential utility of epigenetic biobanks in clinical settings.

4.2 Methods

4.2.1 Electronic Health Record Data

De-identified electronic health record data for this study was extracted from the perioperative data warehouse (PDW), a custom-built, robust data warehouse containing all patients who

have undergone surgery at UCLA Health since the implementation of UCLA's EMR (EPIC Systems, Madison, WI, USA) in March 2013. The PDW, which has been described previously [202], has a two-stage design. First, data are extracted from EPIC's Clarity database into 29 tables organised around three distinct concepts: patients, surgical procedures, and health system encounters. Then, these data are used to populate a series of 4000 distinct measures and metrics such as procedure duration, admission ICD codes, lab results, and medication orders.

4.2.2 Patient Ascertainment

Methylation and genotype samples were collected using blood from 831 patients as part of the UCLA ATLAS precision health initiative between October 26, 2016 and December 10, 2018 [203]. We include the following statements from [203] detailing IRB approval. Retrospective data collection and analysis was approved by the UCLA IRB. Patient Recruitment and Sample Collection for Precision Health Activities at UCLA is an approved study by the UCLA Institutional Review Board (UCLA IRB). IRB17-001013. All necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived.

The samples were collected from patients before undergoing surgery with general anesthesia at UCLA Health, and the patients had not undergone surgery in the 30 days prior to blood sample collection. Of these patients, 302 were selected for inclusion based on the presence of acute kidney injury (AKI), defined as an Acute Kidney Injury Network (AKIN) classification of one or greater, after undergoing surgery. An additional 348 patients were risk-matched controls, with either glomerular filtration rate (GFR) less than or equal to 38 (210 patients), or GFR greater than 38 and a propensity risk score that matched case patients (348 patients). The propensity score was created using available EHR features such as age, weight, BMI, and other preoperative features that were measured in the hospital. Within the control group, we also performed a similar procedure ascertained on whether individuals were a heart attack case. Controls for heart attack patients were also selected using propensity scoring. Demographics of the patient population are further described in

Table 4.1 below.

Table 4.1: Cohort patient demographics. AKIN is the Acute Kidney Injury Network Classification, BMI is Body Mass Index, GFR is glomerular filtration rate.

		Missing	Overall
n			831
Age, mean (SD)		0	61.0 (15.8)
Sex, n (%)	F	0	352 (42.4)
	M		479 (57.6)
BMI, mean (SD)		1	27.2 (6.6)
AKIN Classification, n (%)	0.0	0	537 (64.6)
	1.0		189 (22.7)
	2.0		27 (3.2)
	3.0		78 (9.4)
GFR > 38, n (%)	False	0	375 (45.1)
	True		456 (54.9)
Heart Attack, n (%)	False	601	146 (63.5)
	True		84 (36.5)
Self-Reported Ethnicity, n (%)	Cuban	0	2 (0.2)
	Hispanic or Latino		116 (14.0)
	Hispanic/Spanish origin Other		14 (1.7)
	Mexican, Mexican American, Chicano/a		37 (4.5)
	Not Hispanic or Latino		655 (78.8)
	Patient Refused		5 (0.6)
	Puerto Rican		2 (0.2)
Self-Reported Race, n (%)	American Indian	0	2 (0.2)
	Asian		73 (8.8)
	Black		72 (8.7)
	Declined to Specify		6 (0.7)
	Other Race		132 (15.9)
	Pacific Islander		3 (0.4)
	Unknown		1 (0.1)
	White or Caucasian		542 (65.2)

4.2.3 Medication Usage

For each medication, a patient was labeled as using a medication if the electronic health record contained a medication order that occurred before the methylation sample collection date. Medications were grouped by pharmaceutical subclass using the Generic Product Identifier (GPI) hierarchical classification system codes. Any medications that were ordered

in fewer than 5% of the patients were excluded from the analysis. In total, 168 pharmaceutical subclasses were considered in our analysis. The number of patients using medications from each subclass is shown in Supplementary Table A.1. In Supplementary Table A.2, we show for each pharmaceutical subclass the specific medication that patients in our cohort received.

4.2.4 Lab Results

The most recent lab result prior to the methylation sample collection was extracted from the PDW for each patient. Any labs with a result date that occurred more than 365 days before the methylation sample collection date were excluded from the analyses. Additionally, labs for which there were less than 50 patients with valid results were excluded. We were left with a total of 69 lab values on which to run our models.

4.2.5 Diagnosis Codes

International Classification of Diseases, Ninth Revision (ICD-9) and International Classification of Diseases, Tenth Revision (ICD-10) codes are a standard set of diagnosis codes, primarily used for billing purposes. While these codes provide a standardized methodology for describing a diagnosis, they are very specific. To map these specific diagnosis codes into meaningful, distinct diseases/traits, Denny et al. aggregated the ICD codes into phenotype codes (Phecodes) [204, 205]. Specifically, for each patient, we queried all diagnoses prior to the methylation sample collection date, and used the Phecode (version 1.2) mapping to aggregate ICD-9 and ICD-10 codes to unique, meaningful phenotypes. If a patient's diagnosis record had both ICD-9 and ICD-10 labels, the ICD-10 to Phecode mapping was used instead of the ICD-9 to Phecode mapping. Each Phecode was treated as a binary variable, indicating the presence or absence of a relevant diagnosis code at any point in time before sample collection. We excluded rare Phecodes (occurrence less than 5% of the patients) and, in total, our cohort contained 370 unique Phecode phenotypes.

4.2.6 Preprocessing of genotype data for cross-validation

We measured the genotypes for 831 individuals based on their DNA sampled from whole blood using the ATLAS genotype array. We preprocessed the genotype data using Beagle (d20) [206], PLINK (1.07) [207], and GCTA (1.93.2) [208]. We restricted the genotypes to autosomal variants, removed rare variants (MAF $< .05$), and filtered for variants that met Hardy-Weinberg equilibrium with p-value threshold 10^{-6}. We also removed individuals and variants with more than 1% missing values. For the purpose of running cross-validation, we used Beagle to impute only any remaining missing values, but did not impute to an external dataset. With our sample size and phenotypes evaluated, using genotypes imputed to an external reference did not significantly improve our results. In total we were left with 292,808 SNPs. To obtain principal components, we ran PCA using plink on the chipped genotypes.

4.2.7 Preprocessing and imputation of genotype data for comparison to external models

We used a version of the ATLAS genotype data that was imputed to an external dataset, as detailed in [203]. Briefly, after performing quality control, genotypes were uploaded the Michigan Imputation Server [209]. The server phases the genotype data using Eagle v2.4 [210] and performs imputation using the TOPMed Freeze5 imputation panel [211] using minimac4[212]. We applied the same quality control and filters to the imputed genotypes as we did the chipped genotypes, and we were left with a total of 5,574,956 SNPs.

4.2.8 Preprocessing of methylation array data

We measured methylation data for 831 individuals based on their DNA sampled from whole blood using the EPIC Illumina array. To generate beta-normalized methylation levels at each CpG, we ran the default pipeline of ENmix (1.22.0) [213] on the the raw probe data (IDAT files), which performs background correction, RELIC dye bias correction, and RCP probe-type bias adjustment. We removed from our analysis CpGs that coincided with SNP

loci as well as CpGs on the sex chromosome. We also filtered out outlier samples, defined as having a PC score more than 4 standard deviations away from the average PC score in the first two principal components. In the imputation tasks, we removed sites with low variability (standard deviation < 0.02) leading to a total of 269,471 sites.

4.2.9 Imputation using baseline medical features

To establish a baseline level of imputation performance, we constructed a set of features derived from basic patient information. We trained a simple linear (or logistic) model with 10-fold cross validation using an intercept and patients' age, sex, BMI, methylation-based cell-type proportions (from the reference-based method of Houseman et al. [214]), self-reported ancestry, first ten genetic principal components, and smoking status (never, former or current). Importantly, we wished to establish how well an outcome (medication, Phecode, or lab value) could be imputed by using covariates (e.g. ancestry, age, smoking status) that are known to be captured by genomics.

4.2.10 Imputation using a single penalized linear model

After establishing a baseline level of imputation performance, we performed penalized logistic and linear regression using either individuals' methylation, genotypes, or both. More concretely, we fit 10-fold cross-validation using LASSO, elastic net and ridge regularization under the following two models:

$$y = \alpha_G + G\beta_G + C\beta_C + \varepsilon_G \tag{4.1}$$

$$y = \alpha_M + M\beta_M + C\beta_C + \varepsilon_M \tag{4.2}$$

where y corresponds to the outcome, α the model-specific intercept, G the $n \times s$ genotypes, M the $n \times c$ methylation data, β the vector of length-s or -c effect sizes for the given explanatory variable, C and β_C the covariates from the baseline model and their cor-

responding effect sizes, and ε the length n noise vector. We refer to models (2) and (3) as the PRS and MRS respectively, and note that they also include the baseline features. After fitting all three penalized linear models for a given datatype and outcome, we selected a final model as determined by the model with the highest cross-validated metric (AUC or R^2 if the outcome was binary or continuous, respectively). We fit all penalized models using package *bigstatsr*[124]. We share MRS weights for outcomes that were significantly imputed at https://github.com/cozygene/EHR_MRS_UCLA.

4.2.11 Imputing lab results using EHR data and MRS values with softImpute

Imputing a partially-observed matrix of values is often formulated as a matrix-completion problem. In a matrix completion problem, the observed values of the matrix are used to estimate the values of the unobserved values by assuming that there is some underlying structure that is responsible for generating the data. For example, in the popular SoftImpute method [201], the data is assumed to be well-approximated by a low-rank representation, and the error between the observed values and the reconstructed values is minimized through a convex optimization procedure. However, since the unobserved values are, by definition, not observed, and therefore cannot be used to assess the imputation performance, the primary method for measuring the performance involves masking (removing) observed values and comparing the imputed values to these held-out, true values.

The EHR data used in the imputation procedure included demographic information, diagnosis codes, medication usage, and lab results, which were extracted from the EHR database using the previously described criteria. In addition to the EHR data, we also ran the imputation procedure while including relevant MRS values. Specifically, we included the MRS values for demographics, diagnosis codes, medication usage, and lab results that were imputed at a statistically significant level. These MRS values were added as additional observed features to the EHR matrix.

To estimate the imputation performance, we randomly masked 10% of the observed lab result values, and performed the imputation procedure (SoftImpute matrix completion) to

generate estimates of the missing values. However, since labs are most often ordered in panels, for example a metabolic panel, if a lab is missing then typically other labs that are part of the same panel are also missing. We simulated a more realistic missingness scenario by, instead of masking out values only from a specific lab l, masking out all labs that are ordered as a panel that include lab l. This masking procedure was done per lab, using 10-fold cross-validation, such that 10% of the non-missing values of a particular lab result (and its associated lab panels) were masked (removed), and the remaining 90% of the observed values were used to complete the matrix. Matrix completion was performed using the SoftImpute algorithm, as implemented in the *fancyimpute* [215] python package (version 0.5.5). The proportion of variance explained (R^2) of the true lab values by the imputed lab values was used to measure the imputation performance. Confidence intervals were derived using bootstrapping.

4.2.12 Hypothesis testing

To determine whether an imputation was significant or whether one predictor offered significant additional explanatory signal, we conducted our hypothesis tests using a linear (logistic) regression framework. Primarily, after running cross-validation or generating a single predictor $\hat{y}$ for an outcome y, we would test whether the imputation was significant by comparing it to y:

$$y = \alpha + \hat{y}\beta + \varepsilon \tag{4.3}$$

Where Equation (4) corresponded to linear regression when the outcome was continuous, and logistic regression when the outcome was binary, α was the intercept, and β was an effect size indicating association of the predictor with the outcome. Notably, by building our testing framework as a linear model, we can easily extend it to include additional predictors in order to test whether the additional predictors significantly improve the fit of the regression—or more simply, whether predictor $\hat{y_j}$ offers additional predictive power over $\hat{y_i}$ by conducting a likelihood ratio test of the following nested models:

$$y = \alpha_i + \hat{y}_i\beta_i + \varepsilon_i \quad (4.4)$$

$$y = \alpha_{ij} + \hat{y}_i\beta_i + \hat{y}_j\beta_j + \varepsilon_{ij} \quad (4.5)$$

Where i and j index either the baseline, MRS, or PRS models. We corrected for multiple hypothesis tests within each outcome and method by using a Bonferroni adjustment at α level .05.

4.2.13 Imputing external polygenic risk scores into the ATLAS cohort

We compared our in-house built risk scores to risk scores learned in the UKBiobank dataset[216, 217]. In both [216, 217] the authors construct their PRS using penalized regression akin to as we have done in our analyses. Notably, using penalized regression on individual-level genotypes allows one to automatically, optimally control for shrinkage and variable selection at the step of model generation[124, 218]. This is in contrast with many commonly used polygenic risk score tools such as LDPred[219] or PRSCS[220], that attempt to perform shrinkage or variable selection post-hoc on the level of summary statistics. After downloading the PRS from the PGS catalog[221] listed in Table 4.2, we imputed PRS into our cohort using our imputed genotypes using the score function of Plink. To account for population structure, we limited our analysis to individuals who self-identified as white, and passed filtering using manual inspection of principal components (Figure 4.1).

Ethical Approval and Patient Consent Retrospective data collection and analysis was approved by the UCLA IRB. All research was conducted in accordance with the tenets set forth in the Declaration of Helsinki. We include the following statements from [203] detailing IRB approval. Patient Recruitment and Sample Collection for Precision Health Activities at UCLA is an approved study by the UCLA Institutional Review Board (UCLA IRB). IRB17-001013. All necessary patient/participant consent has been obtained and the appropriate

Table 4.2: Polygenic scores used for the imputed genotypes. We list below the weights used for computing the polygenic risk scores. We downloaded the weights from the Polygenic Score Catalogue (PGS) from two studies of the UKBiobank (Methods).

Lab	PGS accession	Study	Number of variants in weight	Number of variants present in our data
Albumin	PGS000669	Sinnott-Armstrong et al.	11,912	9,172
Cholesterol	PGS000677	Sinnott-Armstrong et al.	17,204	13,401
Creatinine	PGS000679	Sinnott-Armstrong et al.	5,469	4,242
HGBA1C	PGS000685	Sinnott-Armstrong et al.	14,658	11,208
HDL	PGS000686	Sinnott-Armstrong et al.	25,070	19,123
Hematocrit	PGS001225	Tanigawa et al.	15,721	11,898
Hemoglobin	PGS001400	Tanigawa et al.	15,602	11,770
Mean corpuscular hemoglobin	PGS001219	Tanigawa et al.	13,003	9,853
Mean corpuscular volume	PGS001220	Tanigawa et al.	17,311	13,181
Urea nitrogen	PGS000701	Sinnott-Armstrong et al.	12,351	9,473

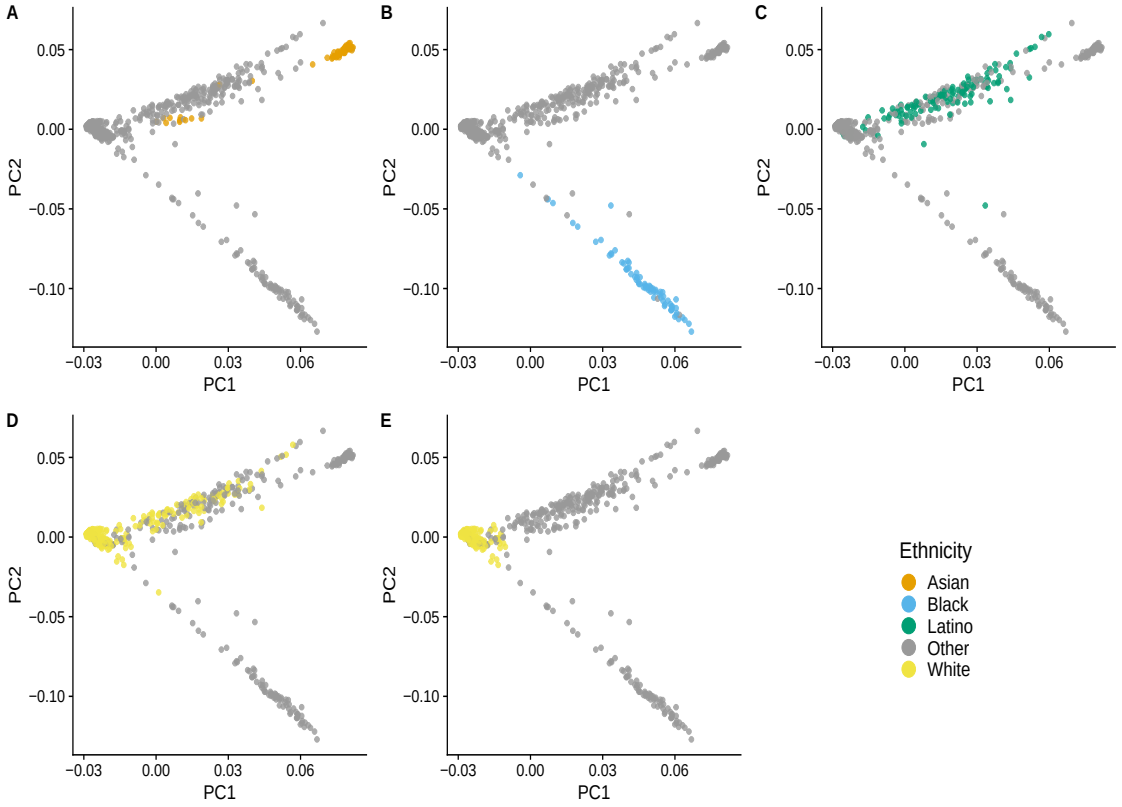

Figure 4.1: Self-reported ancestry along genetic PCs We show the primary self-identified ethnicity in each plot individually. For the analysis using external PRS we limited the set of white-identifying individuals to those who additionally had a PC1 score of $- < .01$. We show the individuals used in our analysis in plot E.

institutional forms have been archived.

4.3 Results

4.3.1 Risk model description

Analogous to the PRS[222, 124], we defined the MRS by a linear combination of m CpG site beta values c and weights w:

$$\text{MRS} = \sum_{i=1}^{m} w_i c_i \tag{4.6}$$

To ensure the methylation risk score added predictive value over commonly captured features (e.g. age and sex), we created a baseline predictive model that included patients' age, sex, reference-based methylation cell-type composition estimates [214], self-reported race-ethnicity, self-reported smoking status, and the first ten genetic principal components [23]. We fit the baseline model using a linear or logistic regression model depending on whether the outcome was continuous or binary. We compared the baseline model to models that included the baseline features as well as either methylation or genotype data. For both the MRS and PRS, we used regression with LASSO, elastic net, and ridge regularization over the genomic features while treating the baseline features as fixed covariates. We fit all models using 10-fold double cross-validation, wherein each training set an additional cross-validation was performed for hyperparameter selection, then this training-set cross-validated model was used to predict the held-out test set. We tested for significance using an association test (via linear regression) between the cross-validated predicted outcome (i.e. the concatenated predictor across all folds) and the true outcome. For full details see Methods.

4.3.2 Methylation risk scores significantly outperform the baseline and PRS models

From our EHR database, we extracted diagnosis codes , medication orders, and the most recent lab results, all of which occurred before the methylation samples were collected. We

aggregated the ICD codes into higher-level phenotypes according to the phenotype code (Phecode) mapping proposed by Denny et al. [204, 205] and grouped individual medications by pharmaceutical subclass to increase generalizability and power.

We trained penalized linear models to predict clinical phenotypes for which there was a sufficient number of patient data available, which included 168 medication subclasses, 69 lab values, and 370 Phecodes. Using a Bonferroni-adjusted association test, the baseline and MRS models significantly imputed the usage of 69 and 88 medications, 18 and 33 labs, and 106 and 139 Phecodes respectively. We compared the performance of the MRS to a model that used both the PRS and baseline features on the same set of individuals, which significantly imputed the usage of 53 medications, 20 lab results, and 93 Phecodes. Notably, the baseline model imputed a greater number of medications and Phecodes than models that leveraged a PRS, which suggests that including genomic features may either add noise or our sample size may not have been sufficient to discover their effects for certain outcomes. We also found that the baseline model gains some of its predictive power from genomics-derived features like ancestry PCs or estimated cell counts, and therefore a PRS or MRS may not offer a substantial improvement over these features for certain outcomes under the current sample sizes.

Next, we investigated outcomes for which genomics-based predictors add predictive power to the baseline features and, in such cases, the extent to which their inclusion improves predictive accuracy. On the outcomes for which the genomics-based predictors produced statistically significant imputations, we conducted a likelihood ratio test comparing an association test of the true outcome using the cross-validated baseline predictor alone, to a model that included the cross-validated baseline predictor as well as the cross-validated predictor that included both baseline and genomic features (Methods). The methylation significantly improved the baseline predictor for 54 medications, 29 labs, and 56 Phecodes, and led to a median increase of 10.74%, 141.52%, and 15.46% over the baseline predictor's accuracy (AUC, R^2) in each outcome, respectively (Figure 4.2). The genotypes significantly improved the baseline predictor for 8 medications, 3 labs, and 11 Phecodes, and led to a median in-

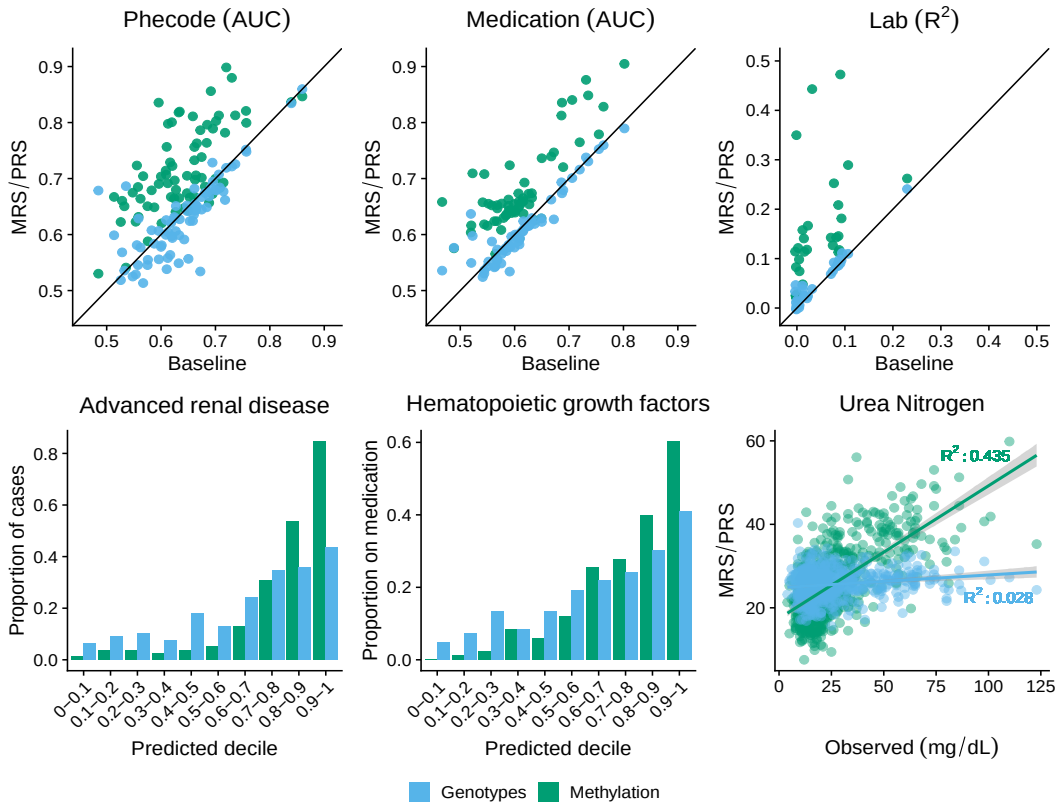

Figure 4.2: MRS increases imputation accuracy on a variety of outcomes (Top) The performance of the PRS (blue) and MRS (green) imputations on the y-axis with the baseline model performance on the x-axis. The performance of binary phenotypes (Phecodes, medications) is measured using area under the ROC curve (AUC) and the performance of continuous phenotypes (lab results) is measured using proportion of variance explained (R^2). Shown is the performance on the union of outcomes that were significantly improved over the baseline model by either the MRS or PRS and that were significantly imputed their corresonponding predictor (72 Phecodes, 59 medications, and 31 labs). (Bottom) The disease incidence as a function of the PRS (blue) and MRS (green) binned by deciles (left, middle); and the observed Urea Nitrogen lab result value plotted against its imputed value (right).

crease of 18.42% over the baseline in the R^2 of the labs, but a median decrease of 1.75% and 0.94% in AUC of the medications and Phecodes respectively (Figure 4.2). We note that our internal sample size is likely underpowered to discover small genetic effects and therefore suggest the contributions made by the genotypes may be due to additional ancestry signal that was not captured by the first few genetic PCs.

The medications that improved the greatest using methylation corresponded to drugs often prescribed to individuals with neutropenia (hematopoeitic growth factors, AUC base-

line .706 95% CI [.661,.748] to AUC methylation .840 [.807,.871]) or chronic kidney disease (phosphate binder agents AUC from .731 [.683, .777] to .876 [.842, .907]). The lab panels best improved with the addition of the methylation-based predictor included those related to kidney function as well as cell counts (Urea nitrogen baseline adjusted R^2 .032 [.007,.057] compared to .443 [.377,.509] with methylation, hemoglobin .107 [.063,.151] to .289 [.232,.346]). The addition of the genotype-based predictor improved the imputation of hematocrit (adjusted R^2 from .077 [.041,.114] to .092 [.052,.132]) and total protein (adjusted R^2 .094 [.047,.141] to .111 [.060,.162]), both of which are influenced by ancestry [223, 224]. In the context of Phecodes, methylation greatly increased the imputation of advanced renal disease over the baseline and genotype models (for example, AUC baseline .720 [.673,.762] to 0.898 [.867,.927] with methylation), and the genotype model increased the imputation of actinic keratosis (AUC from .694 [.631,.747] to .728 [.672,.784]).

Overall, when looking at the intersection of medications significantly imputed by either the methylation and genotypes or methylation and baseline, 92% were better imputed by methylation sites than genotypes (median 9.13% increase) and 78% were better imputed by methylation compared to the baseline (median 6.81% increase). Methylation improved the baseline imputation accuracy by over 15% for 14 medications. In the context of significantly imputed lab values, methylation explained more variability than the baseline (median 398% increase) and genotype (median 274% increase) predictors in 97% and 94% of the respective union of significantly imputed labs. For 22 labs, the percent increase of imputation accuracy was greater than 15% over the baseline model. Methylation was more accurate than the baseline (median 3.48% increase) or genotypes (median 6.58% increase) for 70% and 83% of each respective union of Phecodes. For 29 Phecodes, the methylation offered over a 15% increase in predictive accuracy compared to the baseline model. For a substantial proportion of outcomes, the MRS predictor added statistically significant predictive value over the PRS predictor. This was generally not true when comparing whether the PRS added predictive value over the MRS. For the imputation performance on the full list of phenotypes, see Supplemental Tables A.3, A.4, and **??**.

Importantly, cell-type composition, age, sex, BMI, smoking status and ancestry provide sufficient power for the imputation of many EHR outcomes. Moreover, it is likely that genomics derived features such as cell-type composition and ancestry PCs likely contribute to accurate imputation of several outcomes. In our analyses, we directly compared the power gained by methylation over the aforementioned set of baseline features. However, we note that obtaining these baseline features may be unnecessary as the methylation alone may capture their signal [225, 194, 23, 192, 226]. Further, previous reports have suggested that approaches that fit all methylation probes simultaneously with regularization may perform better when excluding latent confounders, such as cell type composition [227]. We therefore suggest that using the methylation alone is sufficient to replicate a substantial proportion of the associations generated from the baseline features.

4.3.3 Using methylation risk scores improves imputation approaches

Due to significant heterogeneity in patient populations, the diagnosis and treatment process can vary widely between patients, causing many variables to be left unobserved. This sparse structure in the data must be reconciled before performing many downstream analyses, and the imputation accuracy of these unobserved variables is therefore crucial to subsequent steps. A commonly-used approach for imputation is matrix completion, for example, SoftImpute [201], where the data matrix is reconstructed from a low-rank representation. Often, one would jointly use demographic information, diagnosis codes, lab results, and medications to generate an estimate of the unobserved EHR values using an imputation method such as SoftImpute, and therefore we used this as our baseline imputation estimate [228].

To investigate whether methylation can add additional useful information to the imputation, we included the MRS values as part of imputation procedure and compared the performance to the estimates that do not take methylation data into account (see Methods). Specifically, we included cross-validated MRS values for diagnosis codes, lab results, medications, and demographics that were significantly imputed as 261 additional features (i.e. columns of the input matrix) in the imputation procedure. We randomly removed a subset

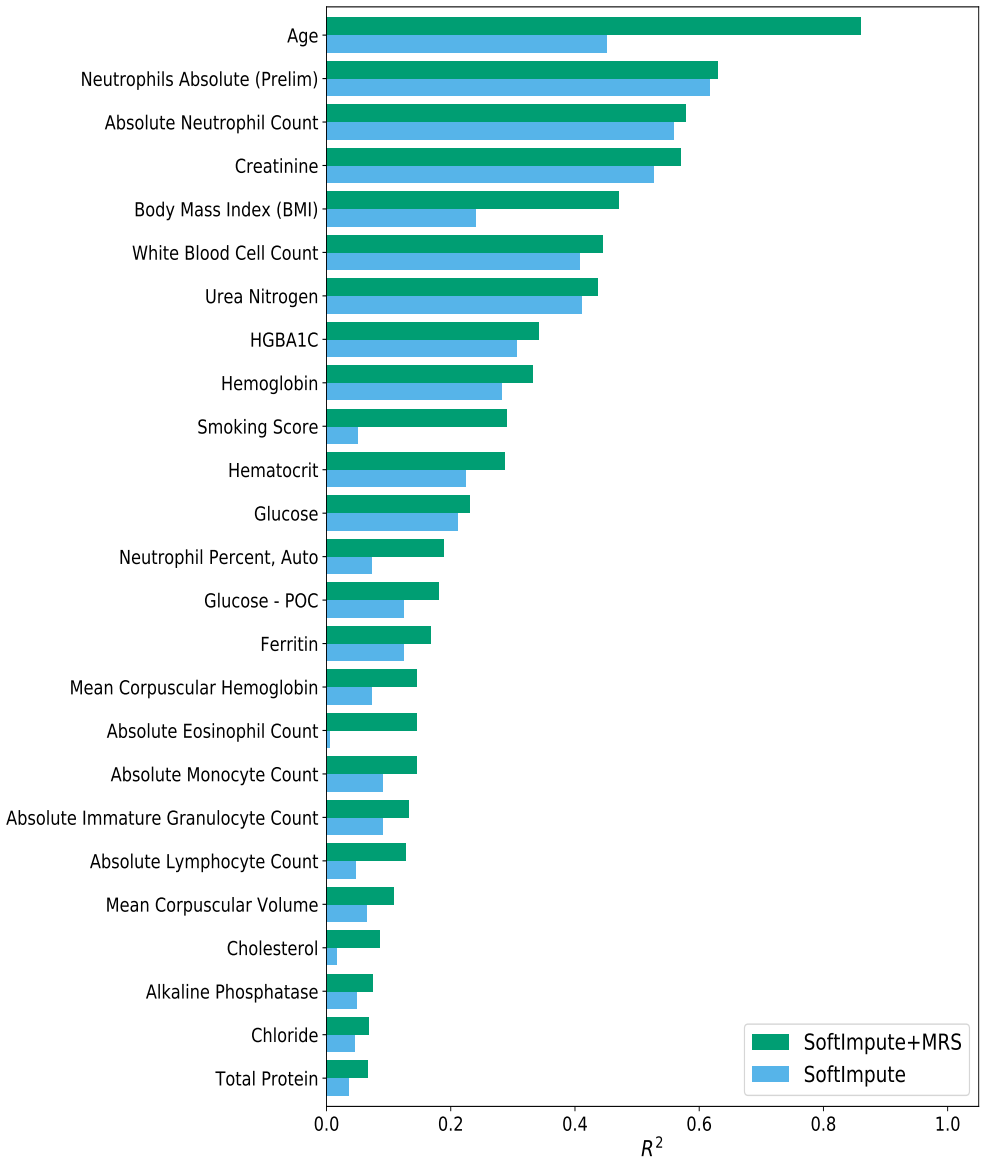

Figure 4.3: Improvement in lab result imputation performance by including MRS For lab results that were significantly better imputed using a matrix completion imputation procedure that included the MRS values, we compare the quality of the imputed values (R^2) using only the EHR data (SoftImpute) to the values generated when including the MRS values in addition to the EHR data (SoftImpute+MRS).

of the observed lab results, including other labs that are ordered as part of the same lab panel(s), and imputed the masked values using the remaining observed values. The imputed values were then compared to the held-out, masked values to assess the quality of the imputation. In Figure 4.3, we show the imputation accuracy (R^2 between the masked true and

imputed values) for labs where the addition of cross-validated MRS to the baseline SoftImpute procedure explained significantly more variability. Of the 67 lab results considered, 25 (37.3%) were significantly better imputed by including the MRS values. Including the MRS values led to a median increase of 47.6% (95% CI 17.3%-90.9%) in the imputation R^2 values.

4.3.4 Methylation risk scores will improve with larger sample sizes

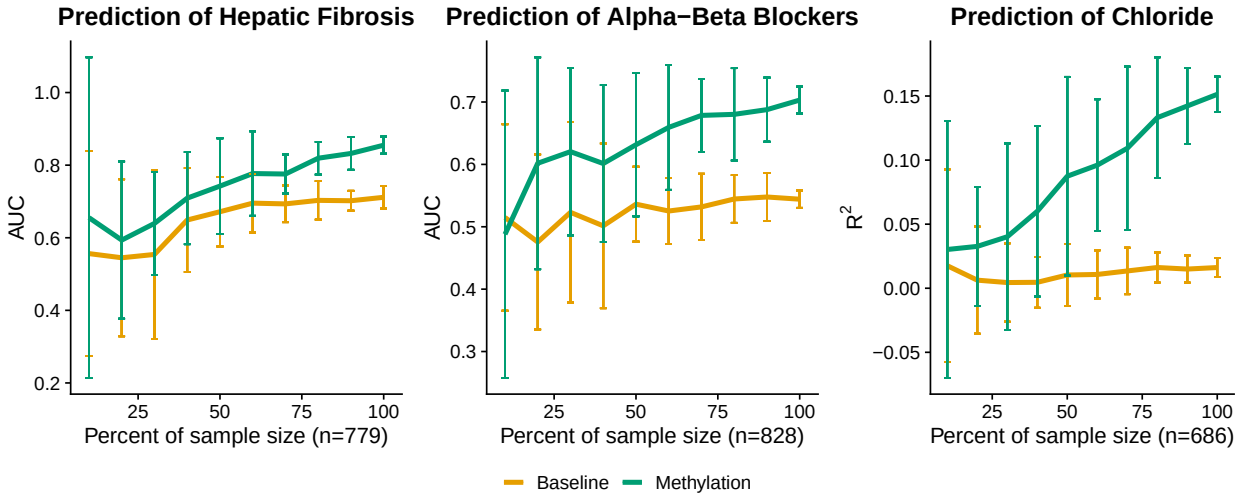

Figure 4.4: Imputation accuracy may improve with additional samples We downsampled the number of individuals to evaluate the imputation performance as a function of sample size using a well-imputed medication, lab value, and Phecode. The performance is significantly affected by the number of individuals, suggesting that there is additional power to be gained with the addition of more methylation samples.

In this study, our analyses of imputation accuracy were performed on 831 individuals' methylation and genetic features. For many phenotypes, the genetic effects are relatively small and require large sample sizes to identify associations between genomic features and the outcome of interest. Consequently, in many biobanks the number of individuals with measured genomic features is several orders of magnitude larger than our sample size [30, 31, 32]. While the methylation data provided sufficient power to significantly impute numerous outcomes, there may remain much power to be gained by increasing the number of methylation samples to numbers approaching biobank-scale.

To determine the role of sample size in our imputation accuracy, we performed an exper-

iment in which we downsampled the number of individuals in our data and trained models on the subsampled data. From the set of outcomes most accurately imputed by methylation and that also significantly improved the baseline's imputation, we chose 10 medications, labs, and Phecodes on which to perform 10-fold cross-validation. For each sample size, we repeated the procedure 20 times to attempt to mitigate variance due to ascertainment effect. Though we selected features that had high accuracy using the full set of data, our results suggest that our models may become more accurate as the sample size increases (Figure 4.4). We further posit that there may be additional outcomes that will be significantly imputed as the number of methylation samples increases.

4.3.5 Comparing MRS to UKBiobank PRS

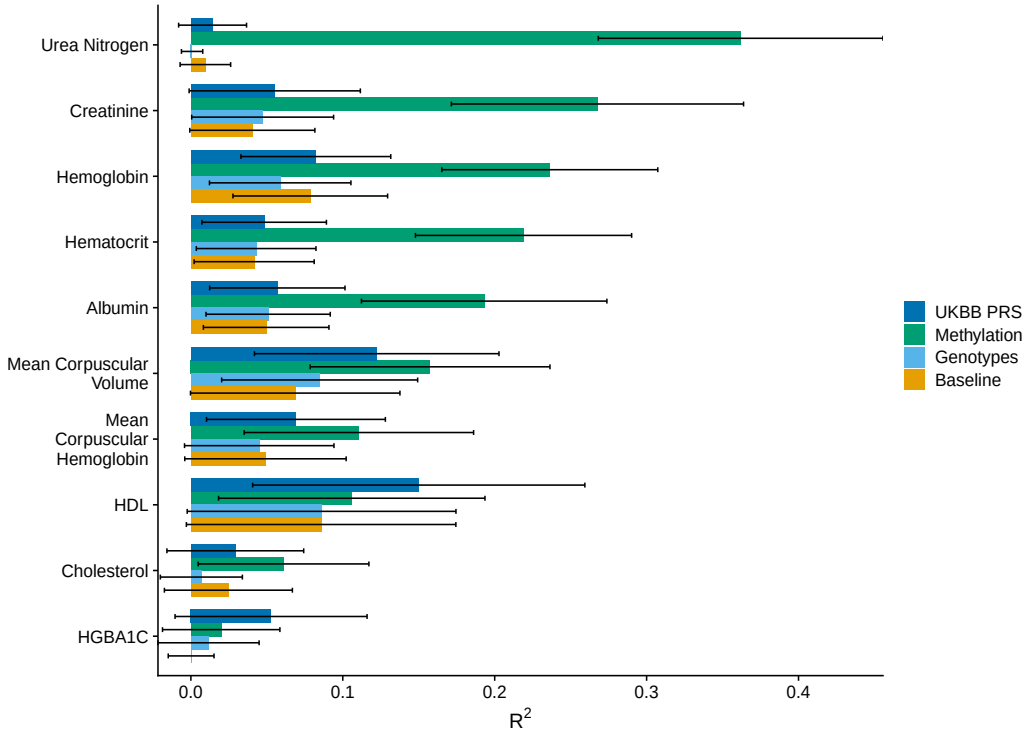

Figure 4.5: Labs as imputed by methylation, genotypes, and an externally-trained polygenic risk score The cross-validated R^2 between the true and imputed lab value on 541 unrelated patients of non-Hispanic-Latino white-identifying individuals using a baseline predictor as well as a baseline predictor with methylation, genotypes, and a PRS externally-trained from UKBiobank summary statistics. HDL corresponds to high-density lipoprotein cholesterol and HGBA1C to glycated hemoglobin.

As expected, due to a small sample size and the likely small effects of SNPs on phenotypes, the PRS developed using the UCLA cohort did not add substantial predictive power over the baseline features. Studies leveraging biobanks with sample sizes several magnitudes larger than the cohort at UCLA however, have shown non-zero heritability for a variety of phenotypes [30, 216, 229, 217]. Therefore, we sought to compare the MRS and PRS generated with the UCLA data to a polygenic risk score created using the UKBiobank data [30]. To do so, we obtained the genotype weights corresponding to 10 polygenic risk scores trained on the UKBiobank (Table 4.2) [30, 217, 229, 221] data and imputed the external risk scores into our health record system using PLINK [207]. We included in the comparison labs that were significantly imputed by the baseline model and excluded labs that corresponded to cell counts or labs for which the internal PRS outperformed the external PRS (indicating a mismatch in the phenotypes or cryptic population structure that was unaccounted for by principal components). While the external polygenic risk score improved substantially the imputation performance relative to the internal polygenic risk score, it did not significantly outperform the methylation for any of the tested phenotypes (Figure 4.5). The methylation remained the best predictor in general—even when trained on fewer than 1000 samples—significantly outperforming the other models in the imputation of urea nitrogen, creatinine, hemoglobin, hematocrit, and albumin. The externally-derived polygenic risk score greatly outperformed both the internally-derived PRS and the MRS when predicting glycated hemoglobin (HGBA1C) and HDL levels, however the improvement was not significant.

4.3.6 Evaluation of methylation risk scores across ancestral populations

Previous reports have suggested that a significant confounder to the application and versatility of polygenic risk scores is population structure, where a population-specific bias is induced that affects generalizability of PRS to different ancestries [42, 230, 41]. The collection of samples analyzed throughout this study is ethnically heterogeneous—individuals were self-identified as non-Hispanic/Latino European, Hispanic/Latino, Black, or Asian. Methy-

lation data is also influenced by differences in population [231], and in particular the first several methylation principal components sufficiently capture population structure in European and African groups [232, 233]). Consequently, we examined the performance of the methylation risk scores within and across ancestral populations.

Primarily, after training the models on the entire heterogeneous set of samples, we examined the predictive performance within each ancestral population. When we examined the top 10 best-imputed (by MRS across the entire set of individuals) lab panels, medications, and Phecodes, only 10 of the entire 180 possible comparisons ($\binom{4}{2}$ comparisons across 30 outcomes) displayed significant differences between the predictive performance within each population separately (7 of which are displayed in Figure 4.6).

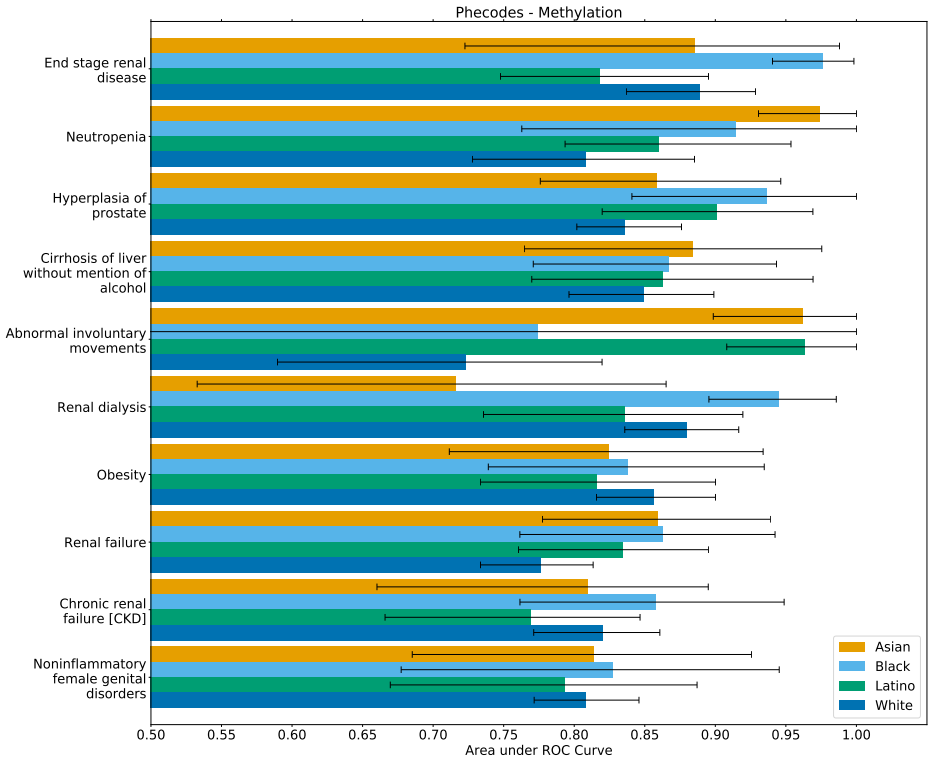

Figure 4.6: Best methylation-imputed Phecodes within ancestral populations. After training a model on the entire heterogeneous population of individuals, we evaluated the predictive performance within each population separately. We observed only 6 (of 60) significant differences between self-reported ancestral groupings.

In a second replication analysis we trained predictive models within ancestral groupings separately. As the individuals self-identified as either Black or Asian comprised less than 100

individuals in both groupings, we focused our analyses on Hispanic/Latino- and white-non-Hispanic/Latino-identifying individuals. We retrained models for the top 5 best-imputed (by MRS) medications, lab panels, and unique Phecodes on the Hispanic/Latino individuals and white non-Hispanic/Latino individuals alone and treated a prediction as significant if its association p-value was lower than .01. Creatinine, hemoglobin, and urea nitrogen replicated across both groupings, however, hematocrit and mean corpuscular hemoglobin did not replicate in the Latino/Hispanic grouping (Table 4.3). In the context of medications, CMV agents, osmotic diuretics, phosphate binder agents, hematopoietic growth factors, and immunosuppressive agents replicated within the white non-Hispanic/Latino population but only CMV and immunosuppressive agents replicated within the Hispanic/Latino population (Table 4.3). Finally, Phecodes corresponding to immunity deficiency, hypertensive renal disease and end-stage renal failure replicated within both groupings, however, neutropenia and anemia replicated only within the white non-Hispanic/Latino set of individuals (Table 4.3).

4.3.7 Replication of methylation risk scores across external datasets

To evaluate the transferability of the MRS to a different population, we performed several experiments in which we imputed the MRS into external datasets. Primarily, we focused on imputation of kidney-related outcomes as they were the most accurately imputed in our own cohort. To do so, we leveraged a dataset that used the HumanMethylation27k array to measure the methylation of 194 individuals who had Type 1 Diabetes, 49.7% of whom had nephropathy (cases) [234]. We re-trained the models for a Phecode corresponding chronic renal disease as well as labs corresponding to creatinine and urea nitrogen on our in-house data, limiting our analysis to the 27,000 sites that belonged to the external dataset. The imputed chronic renal disease was significantly associated with nephropathy in the external dataset (p=8.32e-05, AUC=.684 [.615,.758]. Further, both of the imputed values for creatinine and urea nitrogen were significantly associated with nephropathy (p=5.11e-07, AUC=.739 [.670,.808] and p=3.71e-05 AUC=.693 [.619,.767], respectively). Importantly,

Table 4.3: Replication statistics within ethnic groupings Predictive accuracy (R^2 and AUC) for MRS trained within only Latino/Hispanic- or white-non-Latino/Hispanic-identifying individuals compared to the accuracy trained on the entire, cross-ethnic cohort.

Outcome	Metric	Accuracy, p-value Hispanic/Latino (n=118)	Accuracy, p-value white, non-Hispanic/Latino (n=543)	Accuracy, p-value all ethnicities (n=833)
Creatinine	R^2	.217, 4.63e-07	.356, 7.47e-46	.457, 1.27e-95
Hematocrit	R^2	.045, 2.91e-02	.188, 1.87e-21	.246, 1.14e-42
Hemoglobin	R^2	.096, 1.21e-03	.204, 2.54e-23	.283, 3.02e-50
Mean corpuscular hemoglobin	R^2	.050, 2.12e-02	.122, 9.70e-14	.208, 7.04e-35
Urea nitrogen	R^2	.289, 2.97e-09	.349, 7.61e-44	.435, 2.50e-87
CMV Agents	AUC	.874, 9.27e-07	.875, 3.47e-16	.905, 1.72e-38
Osmotic Diuretics	AUC	.530, 0.841	.842, 2.27e-12	848, 6.37e-34
Phosphate binder agents	AUC	.608, 0.321	.819, 7.76e-17	.876, 1.11e-50
Hematopoietic growth factors	AUC	.567, 0.476	.780, 1.51e-19	.840, 1.75e-45
Immunosuppressive agents	AUC	.721, 1.43e-04	.823, 6.36e-22	.828, 9.44e-41
Neutropenia	AUC	.689, 5.60e-02	.800, 7.68e-10	.836, 1.11e-19
Immunity deficiency	AUC	.889, 4.06e-09	.818, 3.26e-19	.821,9.74e-33
Anemia	AUC	.637, 9.75e-02	.698, 3.13e-08	.789, 1.40e-32
Hypertensive renal disease	AUC	.715, 1.35e-04	.688, 6.74e-10	.801,1.45e-42
End-stage renal failure	AUC	.677, 1.80e-03	.868, 2.51e-29	.898, 5.46e-72

when limiting our internal analysis to sites only on the 27k array, the association signal decreased (for chronic renal disease from p=6.81e-51 to p=3.13e-29, creatinine p=1.27e-95 to p=3.14e-62, and urea nitrogen p=2.50e-87 to p=8.44e-34). However, likely due to correlation between CpGs, the association tests for outcomes trained on the smaller set of sites were still significant.

Second, we expanded our replication analyses to include phenotypes that were unrelated to kidney function. In these analyses, we revisited epigenome-wide association studies (EWAS) of Schizophrenia [235] and Rheumatoid Arthritis [236] and imputed commonly prescribed medications for each dataset—for Schizophrenia we used phenothiazines, and for Rheumatoid Arthritis we used glucocorticosteroids. To ensure our MRS captured medication intake status and were not merely serving as proxies for the disease, we re-trained our models while conditioning on the trait of interest. The imputed phenothaizine intake was significantly associated with Schizophrenia case-control status (p=8.71e-04, AUC=.568

[.527,.611]) and the imputed glucocorticosteroids usage was significantly associated with Rheumatoid Arthritis case-control status (p=2.72e-07, AUC=.626 [.584,.669]. Weights for both medications were trained on CpGs corresponding to those present on the HumanMethylation450k array and also included their corresponding disease in the baseline set of covariates. Accordingly, the association signal of phenothiazines dropped from 1.14e-07 to 3.99e-05 and the performance of glucocorticoids dropped from 1.35e-16 to 1.82e-15 when compared to the MRS trained on the set of EPIC array CpGs and with the baseline features as covariates.

4.4 Discussion

In this study, we provide a comprehensive investigation of the utility of methylation risk scores in a clinical setting. We used (to our knowledge) the largest methylation biobank cohort produced to date, which includes methylation, genotype, and comprehensive EHR data for all patients. We find that the MRS improved imputation performance over a baseline model by 10.65%, 156.31%, and 14.59% when predicting medication usage, lab panel values, and diagnosis codes respectively. These contributions are significantly more substantial than those obtained by PRS.

The vision of genomic biobanks is that the genomic data will be translated into improved clinical diagnosis and treatment decisions [28, 237, 35]. In practice, clinical decisions are not expected to be based solely on genomic information, but rather on the combination of the genomic, medical, and demographic information of the patient. While previous studies have used a limited number of key features as a baseline for imputation of a phenotype (e.g., age, sex, and major comorbidities) [228, 238, 239, 240], to the best of our knowledge, these studies did not take into account the entire familial-genetic or environmental history of the patients. Thus, the question of whether genomic data (methylation or genetics) can be used to improve imputation over the EHR data is critical in order to claim clinical relevance. Our results demonstrate that adding MRS to existing EHR-based imputation frameworks improve imputation accuracy by over 29% in a clinical context.

It is well appreciated that PRS are sensitive to the studied population, and it is often the case that a PRS developed for one ethnic group performs poorly on others [41, 42]. It is therefore important to evaluate the population effect on MRS performance. For this reason, we measured the transferability of our results across different populations, and we observe that the accuracy of the MRS was robust to population structure. This is likely driven by the diversity of the training cohort used, but also due to the fact the we are under-powered to discover subtle differences in imputation accuracy due to our sample sizes. Nonetheless, since we observed very few large differences in accuracy across populations, we are hopeful that our results will inspire future investigations to continue to recruit diverse cohorts and to examine these differences at length with greater sample sizes.

While our study was focused on methylation, there are many other possibilities for the introduction of genomic data in clinical settings. First and foremost, genetic data has been heavily studied by others and large biobanks including genetic data of patients already exists. However, other measurements such as RNA, microbiome, metabolomics, or proteomics may also be relevant. Some of these have logistic and cost considerations at scale. One of the advantages of methylation is that DNA biobanks already exist in large numbers, and the cost of measuring methylation is close to that of measuring genetic data. Moreover, different genomic measurements may provide different snapshots of the patient's data, risk, or health status. Methylation, for example, is known to capture one's smoking status[19], and may therefore be particularly useful for cases in which researchers intend to use self-reported features that may suffer from patient recall bias or honesty. Tangentially, while polygenic risk scores provide a lifetime risk for a patient, methylation risk scores may provide the current risk of the patient over the last few months [241, 242, 243], and other genomic information may provide risk with the resolution of days or hours (e.g., RNA or certain metabolomics [244, 245, 246, 247]). Nonetheless, owing to the dynamic nature of methylation, it is currently unclear what the range or duration of the methylation risk score are. Furthermore, while methylation patterns are associated with outcomes, it is generally unknown if they cause a disease or are a response to a disease [248].

To assist the research community in investigating methylation in the context of disease, we provide the MRS predictors for all significantly predicted outcomes at https://github.com/cozygene/EHR_MRS_UCLA. While our samples were ascertained on kidney and heart disease, we show that our weights successfully replicated across three internal datasets, including studies of Rheumatoid Arthritis and Schizophrenia. Consequently, our weights may be used by researchers and clinicians in different ways. For example, in many epigenome-wide association studies (EWAS), in which associations between specific methylation CpG sites and a phenotype are studied, one may wish to account for patients' comorbidities and medications, which are often not available to the study. Using the MRS database, the researchers leveraging EWAS will be able to incorporate such covariates into their model.

There are multiple potential next steps for the examination of methylation in clinical contexts. First, in this work we focused our attention on the imputation of the phenotypes, or in other words, the inference as to whether the patient is currently diagnosed with a disease. We hope that our findings will be able to be translated to the inference of future clinical events, i.e., prediction of future deterioration or disease occurrence. Second, our analyses did not focus on generating models for a specific patient demographic (e.g. only senior patients) and we were limited to methylation collected from blood samples. As methylation is known to vary across age and tissue type, models may be improved by focusing on individuals of a specific demographic, or by assaying a tissue relevant for a given phenotype (e.g. liver tissue for metabolic disorders. Third, although our evaluation is across the largest dataset which includes both EHR, methylation, and genotype data, the sample size of our study is still moderate compared to genetic studies that are performed on biobanks. Indeed, we demonstrate that for some of the phenotypes, an increase in sample size will likely lead to a substantially improved imputation accuracy (Figure 4.4). Moreover, larger sample size data may be able to reveal the quantity or contribution of genetics verses methylation to the MRS imputation accuracy [227]. In light of our results, as well as the fact that many biobanks have already obtained blood or DNA samples, we recommend that future biobanks consider

measuring methylation in addition to the genotypes across a large number of patients.

APPENDIX A

Supplementary Material - Methylation risk scores are associated with a collection of phenotypes within electronic health record systems

Table A.1: Number of samples with reported usage of medications in the pharmaceutical subclasses. Pharmaceutical subclasses are sorted by number of samples.

Pharmaceutical Subclass	Number of Samples (Percent)
Sodium	699 (80.9%)
Opioid Agonists	639 (74.0%)
Local Anesthetics - Amides	589 (68.2%)
Non-Barbiturate Hypnotics	584 (67.6%)
5-HT3 Receptor Antagonists	549 (63.5%)
Analgesics Other	544 (63.0%)
Radiographic Contrast Media	535 (61.9%)
Anesthetics - Misc.	507 (58.7%)
Glucocorticosteroids	499 (57.8%)
Salicylates	459 (53.1%)
Heparins And Heparinoid-Like Agents	459 (53.1%)
Opioid Combinations	458 (53.0%)
HMG CoA Reductase Inhibitors	456 (52.8%)
Proton Pump Inhibitors	443 (51.3%)
Oil Soluble Vitamins	434 (50.2%)

Continued on next page

Table A.1: Number of samples with reported usage of medications in the pharmaceutical subclasses. Pharmaceutical subclasses are sorted by number of samples.

Pharmaceutical Subclass	Number of Samples (Percent)
Vasopressors	421 (48.7%)
Surfactant Laxatives	398 (46.1%)
Electrolyte Mixtures	390 (45.1%)
Antiarrhythmics Type I-B	383 (44.3%)
Beta Blockers Cardio-Selective	383 (44.3%)
Cephalosporins - 1st Generation	369 (42.7%)
Calcium Channel Blockers	367 (42.5%)
Loop Diuretics	346 (40.0%)
Miscellaneous Contrast Media	341 (39.5%)
Nondepolarizing Muscle Relaxants	336 (38.9%)
Fluoroquinolones	327 (37.8%)
Stimulant Laxatives	326 (37.7%)
Nonsteroidal Anti-inflammatory Agents (NSAIDs)	313 (36.2%)
Sympathomimetics	308 (35.6%)
Antihistamines - Ethanolamines	301 (34.8%)
Laxatives - Miscellaneous	293 (33.9%)
Magnesium	290 (33.6%)
Local Anesthetics - Topical	280 (32.4%)
Potassium	277 (32.1%)
Insulin	269 (31.1%)
Benzodiazepines	265 (30.7%)
Diagnostic Radiopharmaceuticals	264 (30.6%)
Anticonvulsants - Misc.	260 (30.1%)
Carbohydrates	252 (29.2%)
Saline Laxatives	250 (28.9%)

Continued on next page

Table A.1: Number of samples with reported usage of medications in the pharmaceutical subclasses. Pharmaceutical subclasses are sorted by number of samples.

Pharmaceutical Subclass	Number of Samples (Percent)
Antispasmodics	250 (28.9%)
H-2 Antagonists	232 (26.9%)
Angiotensin II Receptor Antagonists	231 (26.7%)
ACE Inhibitors	225 (26.0%)
Penicillin Combinations	219 (25.3%)
Cephalosporins - 3rd Generation	217 (25.1%)
Nitrates	215 (24.9%)
Glycopeptides	213 (24.7%)
Alpha-Beta Blockers	210 (24.3%)
Calcium	207 (24.0%)
Multivitamins	207 (24.0%)
Local Anesthetic Combinations	200 (23.1%)
Anti-infective Misc. - Combinations	198 (22.9%)
Anti-infective Agents - Misc.	193 (22.3%)
Plasma Proteins	190 (22.0%)
Diagnostic Drugs	189 (21.9%)
Water Soluble Vitamins	188 (21.8%)
Phenothiazines	184 (21.3%)
Gastrointestinal Stimulants	182 (21.1%)
Corticosteroids - Topical	182 (21.1%)
Central Muscle Relaxants	181 (20.9%)
Viral Vaccines	175 (20.3%)
Iron	170 (19.7%)
Vasodilators	168 (19.4%)
Antibiotics - Topical	166 (19.2%)

Continued on next page

Table A.1: Number of samples with reported usage of medications in the pharmaceutical subclasses. Pharmaceutical subclasses are sorted by number of samples.

Pharmaceutical Subclass	Number of Samples (Percent)
Hematopoietic Growth Factors	165 (19.1%)
Azithromycin	164 (19.0%)
Antacids - Calcium Salts	164 (19.0%)
Antimyasthenic/Cholinergic Agents	158 (18.3%)
Nasal Steroids	158 (18.3%)
Selective Serotonin Reuptake Inhibitors (SSRIs)	157 (18.2%)
Thiazides and Thiazide-Like Diuretics	157 (18.2%)
Misc. Nutritional Substances	155 (17.9%)
Opioid Antagonists	155 (17.9%)
Platelet Aggregation Inhibitors	154 (17.8%)
Thyroid Hormones	149 (17.2%)
Antifungals - Topical	149 (17.2%)
Bacterial Vaccines	144 (16.7%)
Immunosuppressive Agents	142 (16.4%)
Phosphate Binder Agents	140 (16.2%)
Serotonin Modulators	136 (15.7%)
Laxative Combinations	136 (15.7%)
Biguanides	135 (15.6%)
Depolarizing Muscle Relaxants	135 (15.6%)
Genitourinary Irrigants	134 (15.5%)
Prostatic Hypertrophy Agents	134 (15.5%)
Bronchodilators - Anticholinergics	131 (15.2%)
Antiflatulents	130 (15.0%)
Antacid Combinations	127 (14.7%)
Aminopenicillins	126 (14.6%)

Continued on next page

Table A.1: Number of samples with reported usage of medications in the pharmaceutical subclasses. Pharmaceutical subclasses are sorted by number of samples.

Pharmaceutical Subclass	Number of Samples (Percent)
Imidazole-Related Antifungals	125 (14.5%)
Diagnostic Tests	120 (13.9%)
Cobalamins	118 (13.7%)
Folic Acid/Folates	116 (13.4%)
B-Complex w/ Folic Acid	116 (13.4%)
Antihistamines - Non-Sedating	113 (13.1%)
Anesthetics Topical Oral	108 (12.5%)
Diabetic Supplies	107 (12.4%)
Osmotic Diuretics	106 (12.3%)
Tetracyclines	105 (12.2%)
Multiple Vitamins w/ Minerals	105 (12.2%)
Ophthalmic Anti-infectives	104 (12.0%)
Metabolic Modifiers	102 (11.8%)
Potassium Removing Agents	102 (11.8%)
Potassium Sparing Diuretics	101 (11.7%)
Hemostatics - Topical	101 (11.7%)
Ophthalmics - Misc.	101 (11.7%)
Gout Agents	100 (11.6%)
Alternative Medicine - M's	99 (11.5%)
Parenteral Therapy Supplies	99 (11.5%)
Cough/Cold/Allergy Combinations	99 (11.5%)
Antiseptics - Mouth/Throat	98 (11.3%)
Direct Factor Xa Inhibitors	97 (11.2%)
Anti-infectives - Throat	94 (10.9%)
Anti-inflammatory Agents - Topical	93 (10.8%)

Continued on next page

Table A.1: Number of samples with reported usage of medications in the pharmaceutical subclasses. Pharmaceutical subclasses are sorted by number of samples.

Pharmaceutical Subclass	Number of Samples (Percent)
Coumarin Anticoagulants	92 (10.6%)
Posterior Pituitary Hormones	91 (10.5%)
Antidotes and Specific Antagonists	90 (10.4%)
Antiadrenergic Antihypertensives	90 (10.4%)
Ophthalmic Steroids	90 (10.4%)
Antitussives	88 (10.2%)
Lincosamides	84 (9.7%)
Dibenzapines	83 (9.6%)
Bone Density Regulators	81 (9.4%)
Antianxiety Agents - Misc.	80 (9.3%)
Phosphate	78 (9.0%)
Antiemetics - Anticholinergic	77 (8.9%)
Antiperistaltic Agents	76 (8.8%)
Herpes Agents	76 (8.8%)
Bicarbonates	75 (8.7%)
Liquid Vehicles	72 (8.3%)
Antiarrhythmics Type III	72 (8.3%)
Artificial Tears and Lubricants	71 (8.2%)
Antidiarrheal/Probiotic Agents - Misc.	71 (8.2%)
Toxoid Combinations	70 (8.1%)
Urinary Antispasmodic - Antimuscarinics (Antich...	67 (7.8%)
Lozenges	67 (7.8%)
CMV Agents	66 (7.6%)
Thrombolytic Enzymes	66 (7.6%)
Impotence Agents	65 (7.5%)

Continued on next page

Table A.1: Number of samples with reported usage of medications in the pharmaceutical subclasses. Pharmaceutical subclasses are sorted by number of samples.

Pharmaceutical Subclass	Number of Samples (Percent)
Alternative Medicine - C's	64 (7.4%)
Sulfonylureas	63 (7.3%)
Antihypertensive Combinations	63 (7.3%)
Specialty Vitamins Products	63 (7.3%)
Aminoglycosides	61 (7.1%)
Cephalosporins - 2nd Generation	60 (6.9%)
Alkalinizers	59 (6.8%)
Opioid Partial Agonists	73 (6.8%)
Urinary Anti-infectives	58 (6.7%)
Irrigation Solutions	58 (6.7%)
Influenza Agents	57 (6.6%)
Expectorants	57 (6.6%)
Beta Blockers Non-Selective	56 (6.5%)
Tricyclic Agents	56 (6.5%)
Serotonin-Norepinephrine Reuptake Inhibitors (S...	56 (6.5%)
Cephalosporins - 4th Generation	55 (6.4%)
Antihistamines-Topical	55 (6.4%)
Antacids - Bicarbonate	54 (6.2%)
Bulk Laxatives	53 (6.1%)
Alpha-2 Receptor Antagonists (Tetracyclics)	52 (6.0%)
Ophthalmic Local Anesthetics	49 (5.7%)
Hemostatics - Systemic	49 (5.7%)
Zinc	48 (5.6%)
Dipeptidyl Peptidase-4 (DPP-4) Inhibitors	47 (5.4%)
Gallstone Solubilizing Agents	47 (5.4%)

Continued on next page

Table A.1: Number of samples with reported usage of medications in the pharmaceutical subclasses. Pharmaceutical subclasses are sorted by number of samples.

Pharmaceutical Subclass	Number of Samples (Percent)
Cycloplegic Mydriatics	47 (5.4%)
Protamine	58 (5.4%)
Butyrophenones	46 (5.3%)
Antidepressants - Misc.	45 (5.2%)
Mucolytics	45 (5.2%)
Leukotriene Modulators	44 (5.1%)
B-Complex Vitamins	44 (5.1%)
Acne Products	44 (5.1%)

Table A.2: Medications used in each pharmaceutical subclass

Pharmaceutical Subclass	Drug Name
ALKALINIZERS	BICITRA
ALKALINIZERS	CITRIC
ALKALINIZERS	CYTRA-2
ALKALINIZERS	CYTRA-3
ALKALINIZERS	POT
ALKALINIZERS	POTASSIUM
ANTI-INFECTIVES - THROAT	CLOTRIMAZOLE
ANTI-INFECTIVES - THROAT	MICONAZOLE
ANTI-INFECTIVES - THROAT	NYSTATIN
B-COMPLEX W/ FOLIC ACID	B
B-COMPLEX W/ FOLIC ACID	B-COMPLEX
B-COMPLEX W/ FOLIC ACID	DIALYVITE
B-COMPLEX W/ FOLIC ACID	FULL
B-COMPLEX W/ FOLIC ACID	NEPHRO-VITE
B-COMPLEX W/ FOLIC ACID	NEPHROCAPS
B-COMPLEX W/ FOLIC ACID	RENA-VITE
B-COMPLEX W/ FOLIC ACID	RENAL
B-COMPLEX W/ FOLIC ACID	RENAL-VITE
B-COMPLEX W/ FOLIC ACID	VOL-CARE
B-COMPLEX W/ FOLIC ACID	VP-VITE
BIGUANIDES	METFORMIN
CALCIUM CHANNEL BLOCKERS	ADALAT
CALCIUM CHANNEL BLOCKERS	AFEDITAB
CALCIUM CHANNEL BLOCKERS	AMLODIPINE
CALCIUM CHANNEL BLOCKERS	CARTIA

Continued on next page

Table A.2: Medications used in each pharmaceutical subclass

Pharmaceutical Subclass	Drug Name
CALCIUM CHANNEL BLOCKERS	DILT-XR
CALCIUM CHANNEL BLOCKERS	DILTIAZEM
CALCIUM CHANNEL BLOCKERS	FELODIPINE
CALCIUM CHANNEL BLOCKERS	ISRADIPINE
CALCIUM CHANNEL BLOCKERS	NICARDIPINE
CALCIUM CHANNEL BLOCKERS	NIFEDICAL
CALCIUM CHANNEL BLOCKERS	NIFEDIPINE
CALCIUM CHANNEL BLOCKERS	NIMODIPINE
CALCIUM CHANNEL BLOCKERS	NORVASC
CALCIUM CHANNEL BLOCKERS	VERAPAMIL
CMV AGENTS	VALCYTE
CMV AGENTS	VALGANCICLOVIR
DIBENZAPINES	OLANZAPINE
DIBENZAPINES	QUETIAPINE
DIBENZAPINES	ZYPREXA
HEMATOPOIETIC GROWTH FACTORS	ARANESP
HEMATOPOIETIC GROWTH FACTORS	DARBEPOETIN
HEMATOPOIETIC GROWTH FACTORS	EPOETIN
HEMATOPOIETIC GROWTH FACTORS	EPOGEN
HEMATOPOIETIC GROWTH FACTORS	FILGRASTIM
HEMATOPOIETIC GROWTH FACTORS	FILGRASTIM-SNDZ
HEMATOPOIETIC GROWTH FACTORS	MIRCERA
HEMATOPOIETIC GROWTH FACTORS	NEULASTA
HEMATOPOIETIC GROWTH FACTORS	NEUPOGEN
HEMATOPOIETIC GROWTH FACTORS	PEGFILGRASTIM

Continued on next page

Table A.2: Medications used in each pharmaceutical subclass

Pharmaceutical Subclass	Drug Name
HEMATOPOIETIC GROWTH FACTORS	PROCRIT
HEMATOPOIETIC GROWTH FACTORS	ROMIPLOSTIM
HEMATOPOIETIC GROWTH FACTORS	ZARXIO
IMMUNOSUPPRESSIVE AGENTS	ANTI-THYMOCYTE
IMMUNOSUPPRESSIVE AGENTS	AZATHIOPRINE
IMMUNOSUPPRESSIVE AGENTS	BASILIXIMAB
IMMUNOSUPPRESSIVE AGENTS	BELATACEPT
IMMUNOSUPPRESSIVE AGENTS	CELLCEPT
IMMUNOSUPPRESSIVE AGENTS	CYCLOSPORINE
IMMUNOSUPPRESSIVE AGENTS	EVEROLIMUS
IMMUNOSUPPRESSIVE AGENTS	IDS
IMMUNOSUPPRESSIVE AGENTS	MYCOPHENOLATE
IMMUNOSUPPRESSIVE AGENTS	MYCOPHENOLIC
IMMUNOSUPPRESSIVE AGENTS	MYFORTIC
IMMUNOSUPPRESSIVE AGENTS	NEORAL
IMMUNOSUPPRESSIVE AGENTS	PROGRAF
IMMUNOSUPPRESSIVE AGENTS	RAPAMUNE
IMMUNOSUPPRESSIVE AGENTS	SIROLIMUS
IMMUNOSUPPRESSIVE AGENTS	TACROLIMUS
METABOLIC MODIFIERS	CALCITRIOL
METABOLIC MODIFIERS	CINACALCET
METABOLIC MODIFIERS	DOXERCALCIFEROL
METABOLIC MODIFIERS	HECTOROL
METABOLIC MODIFIERS	PARICALCITOL
METABOLIC MODIFIERS	ROCALTROL

Continued on next page

Table A.2: Medications used in each pharmaceutical subclass

Pharmaceutical Subclass	Drug Name
METABOLIC MODIFIERS	SENSIPAR
METABOLIC MODIFIERS	ZEMPLAR
OSMOTIC DIURETICS	MANNITOL
PHOSPHATE BINDER AGENTS	AURYXIA
PHOSPHATE BINDER AGENTS	CALCIUM
PHOSPHATE BINDER AGENTS	FERRIC
PHOSPHATE BINDER AGENTS	FOSRENOL
PHOSPHATE BINDER AGENTS	LANTHANUM
PHOSPHATE BINDER AGENTS	PHOSLO
PHOSPHATE BINDER AGENTS	RENAGEL
PHOSPHATE BINDER AGENTS	RENVELA
PHOSPHATE BINDER AGENTS	SEVELAMER
PHOSPHATE BINDER AGENTS	SUCROFERRIC
PHOSPHATE BINDER AGENTS	VELPHORO
POTASSIUM REMOVING RESINS	KALEXATE
POTASSIUM REMOVING RESINS	KAYEXALATE
POTASSIUM REMOVING RESINS	KIONEX
POTASSIUM REMOVING RESINS	PATIROMER
POTASSIUM REMOVING RESINS	SODIUM
POTASSIUM REMOVING RESINS	VELTASSA
SELECTIVE SEROTONIN REUPTAKE INHIBITORS (SSRIS)	CITALOPRAM
SELECTIVE SEROTONIN REUPTAKE INHIBITORS (SSRIS)	ESCITALOPRAM
SELECTIVE SEROTONIN REUPTAKE INHIBITORS (SSRIS)	FLUOXETINE
SELECTIVE SEROTONIN REUPTAKE INHIBITORS (SSRIS)	FLUVOXAMINE
SELECTIVE SEROTONIN REUPTAKE INHIBITORS (SSRIS)	LEXAPRO

Continued on next page

Table A.2: Medications used in each pharmaceutical subclass

Pharmaceutical Subclass	Drug Name
SELECTIVE SEROTONIN REUPTAKE INHIBITORS (SSRIS)	PAROXETINE
SELECTIVE SEROTONIN REUPTAKE INHIBITORS (SSRIS)	SERTRALINE
SELECTIVE SEROTONIN REUPTAKE INHIBITORS (SSRIS)	ZOLOFT
SPECIALTY VITAMINS PRODUCTS	MG-PLUS
SPECIALTY VITAMINS PRODUCTS	ONE-A-DAY
SPECIALTY VITAMINS PRODUCTS	PROSTATE
SULFONYLUREAS	GLIMEPIRIDE
SULFONYLUREAS	GLIPIZIDE
SULFONYLUREAS	GLYBURIDE
THROMBOLYTIC ENZYMES	ALTEPLASE
VASODILATORS	HYDRALAZINE
VASODILATORS	MINOXIDIL
VASODILATORS	NITROPRUSSIDE

Table A.3: Mean (95% confidence interval) area under the ROC curve for predicting medication usage, grouped by pharmaceutical subclass, using the baseline, methylation data, and genotype data. Confidence intervals determined using bootstrapping.

Pharmaceutical Subclass	Baseline	Methylation	Genotypes
CMV Agents	0.80 (0.75-0.85)	0.90 (0.86-0.94)	0.79 (0.74-0.84)
Phosphate Binder Agents	0.73 (0.68-0.77)	0.88 (0.84-0.91)	0.74 (0.70-0.78)
Osmotic Diuretics	0.74 (0.69-0.78)	0.85 (0.81-0.88)	0.73 (0.68-0.78)
Hematopoietic Growth Factors	0.70 (0.66-0.75)	0.84 (0.81-0.87)	0.70 (0.66-0.75)
B-Complex w/ Folic Acid	0.69 (0.64-0.74)	0.84 (0.79-0.87)	0.67 (0.62-0.72)
Immunosuppressive Agents	0.77 (0.72-0.81)	0.83 (0.79-0.86)	0.76 (0.70-0.80)
Metabolic Modifiers	0.69 (0.63-0.74)	0.81 (0.77-0.86)	0.70 (0.63-0.75)
Prostatic Hypertrophy Agents	0.76 (0.72-0.79)	0.78 (0.74-0.81)	0.76 (0.71-0.79)
Antacids - Bicarbonate	0.78 (0.71-0.85)	0.78 (0.71-0.83)	0.75 (0.69-0.81)
Anti-infectives - Throat	0.72 (0.67-0.77)	0.77 (0.71-0.82)	0.72 (0.66-0.78)
Cycloplegic Mydriatics	0.76 (0.71-0.82)	0.75 (0.68-0.81)	0.76 (0.69-0.82)
Thrombolytic Enzymes	0.67 (0.60-0.74)	0.75 (0.68-0.81)	0.63 (0.56-0.70)
Plasma Proteins	0.67 (0.62-0.72)	0.74 (0.69-0.78)	0.66 (0.61-0.71)
Potassium Removing Agents	0.64 (0.59-0.70)	0.74 (0.69-0.79)	0.62 (0.56-0.67)
Cephalosporins - 4th Generation	0.67 (0.59-0.74)	0.73 (0.64-0.80)	0.65 (0.57-0.74)
Gallstone Solubilizing Agents	0.59 (0.50-0.69)	0.72 (0.64-0.79)	0.50 (0.41-0.59)
Imidazole-Related Antifungals	0.69 (0.64-0.73)	0.72 (0.67-0.77)	0.68 (0.63-0.73)
HMG CoA Reductase Inhibitors	0.72 (0.67-0.75)	0.72 (0.68-0.75)	0.71 (0.68-0.74)
Alkalinizers	0.70 (0.62-0.76)	0.71 (0.63-0.78)	0.68 (0.59-0.75)
Bone Density Regulators	0.70 (0.63-0.75)	0.71 (0.65-0.75)	0.70 (0.64-0.76)
Parenteral Therapy Supplies	0.52 (0.45-0.59)	0.71 (0.65-0.76)	0.60 (0.54-0.66)
Vasodilators	0.54 (0.49-0.59)	0.70 (0.66-0.75)	0.48 (0.43-0.52)
Salicylates	0.69 (0.65-0.72)	0.70 (0.66-0.73)	0.68 (0.64-0.72)
Ophthalmic Local Anesthetics	0.69 (0.63-0.76)	0.69 (0.63-0.76)	0.68 (0.61-0.76)

Continued on next page

Table A.3: Mean (95% confidence interval) area under the ROC curve for predicting medication usage, grouped by pharmaceutical subclass, using the baseline, methylation data, and genotype data. Confidence intervals determined using bootstrapping.

Pharmaceutical Subclass	Baseline	Methylation	Genotypes
Proton Pump Inhibitors	0.62 (0.58-0.66)	0.69 (0.66-0.73)	0.61 (0.57-0.66)
Impotence Agents	0.70 (0.64-0.75)	0.68 (0.62-0.74)	0.67 (0.60-0.73)
Ophthalmic Steroids	0.70 (0.64-0.75)	0.68 (0.62-0.73)	0.68 (0.62-0.74)
Diabetic Supplies	0.64 (0.58-0.69)	0.67 (0.62-0.72)	0.62 (0.56-0.67)
Phosphate	0.65 (0.57-0.72)	0.67 (0.60-0.74)	0.60 (0.52-0.66)
Loop Diuretics	0.63 (0.60-0.67)	0.67 (0.63-0.71)	0.63 (0.59-0.67)
Antiseptics - Mouth/Throat	0.63 (0.57-0.68)	0.67 (0.62-0.73)	0.62 (0.54-0.67)
Anti-infective Agents - Misc.	0.61 (0.56-0.65)	0.67 (0.62-0.72)	0.60 (0.55-0.65)
Specialty Vitamins Products	0.60 (0.52-0.68)	0.67 (0.58-0.75)	0.61 (0.52-0.69)
Anti-infective Misc. - Combinations	0.64 (0.59-0.68)	0.67 (0.62-0.72)	0.62 (0.57-0.66)
Iron	0.66 (0.62-0.71)	0.67 (0.62-0.71)	0.66 (0.61-0.71)
Cephalosporins - 3rd Generation	0.62 (0.57-0.66)	0.66 (0.63-0.71)	0.61 (0.57-0.65)
Glucocorticosteroids	0.59 (0.55-0.63)	0.66 (0.63-0.70)	0.58 (0.54-0.62)
Antihistamines - Ethanolamines	0.62 (0.58-0.65)	0.66 (0.62-0.70)	0.62 (0.58-0.66)
Fluoroquinolones	0.54 (0.50-0.58)	0.66 (0.62-0.70)	0.52 (0.48-0.57)
Calcium	0.61 (0.57-0.65)	0.66 (0.62-0.71)	0.60 (0.56-0.64)
Benzodiazepines	0.61 (0.57-0.66)	0.66 (0.62-0.70)	0.61 (0.56-0.64)
Biguanides	0.67 (0.62-0.71)	0.66 (0.61-0.71)	0.65 (0.60-0.70)
Local Anesthetic Combinations	0.57 (0.53-0.62)	0.66 (0.61-0.70)	0.56 (0.52-0.61)
Ophthalmics - Misc.	0.66 (0.60-0.72)	0.66 (0.60-0.71)	0.65 (0.60-0.70)
Antacid Combinations	0.63 (0.58-0.68)	0.66 (0.61-0.70)	0.62 (0.56-0.67)
Dibenzapines	0.62 (0.56-0.68)	0.66 (0.58-0.72)	0.58 (0.52-0.63)
Nitrates	0.64 (0.61-0.68)	0.66 (0.61-0.70)	0.63 (0.59-0.67)
Insulin	0.63 (0.58-0.67)	0.66 (0.61-0.69)	0.62 (0.57-0.66)

Continued on next page

Table A.3: Mean (95% confidence interval) area under the ROC curve for predicting medication usage, grouped by pharmaceutical subclass, using the baseline, methylation data, and genotype data. Confidence intervals determined using bootstrapping.

Pharmaceutical Subclass	Baseline	Methylation	Genotypes
Liquid Vehicles	0.59 (0.51-0.67)	0.65 (0.58-0.73)	0.58 (0.51-0.66)
Alternative Medicine - M's	0.55 (0.49-0.61)	0.65 (0.59-0.71)	0.59 (0.52-0.64)
5-HT3 Receptor Antagonists	0.57 (0.53-0.60)	0.65 (0.61-0.69)	0.55 (0.51-0.60)
Antiperistaltic Agents	0.55 (0.48-0.62)	0.65 (0.57-0.73)	0.51 (0.44-0.58)
Analgesics Other	0.59 (0.55-0.63)	0.65 (0.61-0.68)	0.58 (0.54-0.62)
Carbohydrates	0.61 (0.57-0.65)	0.65 (0.61-0.69)	0.59 (0.54-0.64)
Laxatives - Miscellaneous	0.61 (0.57-0.65)	0.65 (0.61-0.68)	0.59 (0.55-0.64)
Alpha-Beta Blockers	0.57 (0.52-0.61)	0.65 (0.60-0.69)	0.57 (0.53-0.61)
Saline Laxatives	0.59 (0.54-0.63)	0.64 (0.60-0.70)	0.58 (0.54-0.62)
Potassium	0.59 (0.55-0.63)	0.64 (0.60-0.68)	0.58 (0.54-0.62)
Stimulant Laxatives	0.60 (0.56-0.64)	0.64 (0.60-0.68)	0.60 (0.56-0.64)
Urinary Anti-infectives	0.60 (0.53-0.67)	0.64 (0.55-0.71)	0.51 (0.42-0.58)
Calcium Channel Blockers	0.59 (0.56-0.63)	0.64 (0.60-0.67)	0.60 (0.56-0.63)
Glycopeptides	0.60 (0.56-0.65)	0.64 (0.59-0.68)	0.59 (0.54-0.63)
Magnesium	0.58 (0.54-0.62)	0.64 (0.60-0.67)	0.57 (0.53-0.61)
Heparins And Heparinoid-Like Agents	0.62 (0.58-0.65)	0.64 (0.60-0.67)	0.61 (0.57-0.64)
Bicarbonates	0.60 (0.53-0.67)	0.63 (0.56-0.70)	0.61 (0.53-0.68)
Electrolyte Mixtures	0.58 (0.55-0.62)	0.63 (0.60-0.67)	0.57 (0.54-0.61)
Thyroid Hormones	0.64 (0.58-0.69)	0.63 (0.58-0.68)	0.63 (0.58-0.68)
Antihypertensive Combinations	0.64 (0.56-0.70)	0.63 (0.57-0.71)	0.59 (0.50-0.67)
Folic Acid/Folates	0.61 (0.54-0.67)	0.63 (0.58-0.69)	0.60 (0.55-0.66)
Diagnostic Radiopharmaceuticals	0.58 (0.54-0.62)	0.63 (0.60-0.67)	0.57 (0.52-0.61)
Surfactant Laxatives	0.60 (0.57-0.64)	0.63 (0.58-0.66)	0.60 (0.56-0.64)
Thiazides and Thiazide-Like Diuretics	0.65 (0.61-0.71)	0.63 (0.58-0.68)	0.63 (0.58-0.67)

Continued on next page

Table A.3: Mean (95% confidence interval) area under the ROC curve for predicting medication usage, grouped by pharmaceutical subclass, using the baseline, methylation data, and genotype data. Confidence intervals determined using bootstrapping.

Pharmaceutical Subclass	Baseline	Methylation	Genotypes
Expectorants	0.57 (0.47-0.65)	0.63 (0.54-0.72)	0.44 (0.36-0.52)
Platelet Aggregation Inhibitors	0.64 (0.59-0.68)	0.63 (0.58-0.67)	0.62 (0.58-0.67)
Antiflatulents	0.56 (0.50-0.61)	0.63 (0.58-0.68)	0.54 (0.48-0.60)
Vasopressors	0.57 (0.53-0.61)	0.62 (0.59-0.65)	0.55 (0.52-0.60)
Opioid Antagonists	0.60 (0.54-0.65)	0.62 (0.58-0.67)	0.60 (0.54-0.65)
Antibiotics - Topical	0.55 (0.50-0.60)	0.62 (0.58-0.67)	0.50 (0.45-0.56)
Antidiarrheal/Probiotic Agents - Misc.	0.56 (0.49-0.63)	0.62 (0.55-0.69)	0.51 (0.43-0.59)
Serotonin-Norepinephrine Reuptake Inhibitors (S...	0.68 (0.61-0.74)	0.62 (0.54-0.70)	0.66 (0.58-0.73)
Phenothiazines	0.52 (0.46-0.56)	0.62 (0.56-0.66)	0.46 (0.41-0.50)
Beta Blockers Cardio-Selective	0.54 (0.51-0.58)	0.61 (0.58-0.65)	0.53 (0.49-0.57)
Misc. Nutritional Substances	0.55 (0.49-0.59)	0.61 (0.56-0.67)	0.54 (0.49-0.59)
Alpha-2 Receptor Antagonists (Tetracyclics)	0.66 (0.56-0.75)	0.61 (0.51-0.70)	0.67 (0.58-0.75)
Bronchodilators - Anticholinergics	0.63 (0.58-0.67)	0.61 (0.56-0.67)	0.61 (0.56-0.66)
Tetracyclines	0.51 (0.45-0.59)	0.61 (0.56-0.67)	0.40 (0.35-0.46)
Antacids - Calcium Salts	0.58 (0.53-0.63)	0.61 (0.56-0.66)	0.55 (0.48-0.60)
Penicillin Combinations	0.57 (0.52-0.61)	0.61 (0.56-0.65)	0.55 (0.51-0.60)
Gout Agents	0.65 (0.60-0.69)	0.61 (0.54-0.66)	0.63 (0.57-0.68)
Radiographic Contrast Media	0.61 (0.57-0.64)	0.61 (0.57-0.64)	0.60 (0.56-0.63)
Sodium	0.47 (0.43-0.52)	0.61 (0.56-0.65)	0.54 (0.48-0.59)
Diagnostic Tests	0.60 (0.55-0.65)	0.61 (0.55-0.66)	0.57 (0.52-0.63)
Sympathomimetics	0.59 (0.55-0.62)	0.61 (0.57-0.65)	0.57 (0.53-0.61)
Antiarrhythmics Type III	0.56 (0.48-0.62)	0.60 (0.53-0.67)	0.53 (0.46-0.61)
Antihistamines-Topical	0.48 (0.41-0.57)	0.60 (0.53-0.68)	0.39 (0.31-0.47)
Antidotes and Specific Antagonists	0.39 (0.33-0.46)	0.60 (0.54-0.67)	0.59 (0.52-0.65)

Continued on next page

Table A.3: Mean (95% confidence interval) area under the ROC curve for predicting medication usage, grouped by pharmaceutical subclass, using the baseline, methylation data, and genotype data. Confidence intervals determined using bootstrapping.

Pharmaceutical Subclass	Baseline	Methylation	Genotypes
Coumarin Anticoagulants	0.62 (0.57-0.68)	0.60 (0.55-0.66)	0.58 (0.52-0.62)
Bacterial Vaccines	0.56 (0.51-0.62)	0.60 (0.55-0.65)	0.56 (0.52-0.61)
Genitourinary Irrigants	0.41 (0.36-0.46)	0.60 (0.55-0.65)	0.54 (0.49-0.59)
Anesthetics Topical Oral	0.63 (0.57-0.69)	0.60 (0.53-0.66)	0.60 (0.54-0.67)
Cobalamins	0.60 (0.54-0.65)	0.60 (0.55-0.66)	0.57 (0.52-0.63)
Posterior Pituitary Hormones	0.56 (0.50-0.62)	0.59 (0.53-0.66)	0.50 (0.43-0.56)
Gastrointestinal Stimulants	0.52 (0.47-0.57)	0.59 (0.55-0.65)	0.53 (0.48-0.57)
Antiadrenergic Antihypertensives	0.55 (0.48-0.60)	0.59 (0.54-0.65)	0.49 (0.43-0.54)
Angiotensin II Receptor Antagonists	0.61 (0.56-0.66)	0.59 (0.55-0.63)	0.58 (0.54-0.62)
Antitussives	0.50 (0.43-0.56)	0.59 (0.53-0.65)	0.50 (0.44-0.55)
B-Complex Vitamins	0.49 (0.41-0.57)	0.59 (0.47-0.67)	0.48 (0.40-0.56)
Cephalosporins - 1st Generation	0.55 (0.51-0.59)	0.59 (0.55-0.63)	0.53 (0.49-0.58)
Diagnostic Drugs	0.59 (0.54-0.63)	0.59 (0.54-0.63)	0.59 (0.54-0.63)
Potassium Sparing Diuretics	0.60 (0.53-0.66)	0.58 (0.52-0.66)	0.57 (0.50-0.64)
Selective Serotonin Reuptake Inhibitors (SSRIs)	0.59 (0.55-0.64)	0.58 (0.53-0.64)	0.57 (0.52-0.62)
Anesthetics - Misc.	0.53 (0.49-0.56)	0.58 (0.55-0.62)	0.52 (0.49-0.55)
Irrigation Solutions	0.58 (0.50-0.64)	0.58 (0.51-0.64)	0.55 (0.47-0.62)
Lozenges	0.58 (0.50-0.65)	0.58 (0.49-0.66)	0.51 (0.43-0.60)
Aminoglycosides	0.55 (0.48-0.62)	0.58 (0.50-0.64)	0.45 (0.38-0.54)
Non-Barbiturate Hypnotics	0.55 (0.51-0.59)	0.57 (0.53-0.61)	0.53 (0.49-0.57)
Ophthalmic Anti-infectives	0.61 (0.56-0.66)	0.57 (0.53-0.63)	0.61 (0.54-0.66)
Antiarrhythmics Type I-B	0.54 (0.50-0.58)	0.57 (0.54-0.62)	0.51 (0.48-0.55)
Oil Soluble Vitamins	0.57 (0.53-0.61)	0.57 (0.53-0.61)	0.59 (0.56-0.64)
Direct Factor Xa Inhibitors	0.58 (0.52-0.64)	0.57 (0.51-0.63)	0.57 (0.51-0.63)

Continued on next page

Table A.3: Mean (95% confidence interval) area under the ROC curve for predicting medication usage, grouped by pharmaceutical subclass, using the baseline, methylation data, and genotype data. Confidence intervals determined using bootstrapping.

Pharmaceutical Subclass	Baseline	Methylation	Genotypes
H-2 Antagonists	0.52 (0.48-0.57)	0.57 (0.53-0.62)	0.50 (0.47-0.55)
Multivitamins	0.51 (0.46-0.55)	0.57 (0.52-0.62)	0.52 (0.47-0.57)
Hemostatics - Topical	0.57 (0.50-0.63)	0.57 (0.50-0.64)	0.53 (0.47-0.58)
Artificial Tears and Lubricants	0.56 (0.48-0.63)	0.57 (0.51-0.64)	0.58 (0.51-0.65)
Anti-inflammatory Agents - Topical	0.58 (0.52-0.65)	0.57 (0.51-0.64)	0.58 (0.51-0.64)
Opioid Agonists	0.58 (0.53-0.62)	0.57 (0.53-0.61)	0.55 (0.51-0.60)
Leukotriene Modulators	0.55 (0.46-0.63)	0.57 (0.48-0.66)	0.50 (0.40-0.57)
Antianxiety Agents - Misc.	0.58 (0.52-0.65)	0.57 (0.50-0.64)	0.53 (0.46-0.60)
Local Anesthetics - Amides	0.58 (0.53-0.62)	0.57 (0.53-0.61)	0.56 (0.51-0.60)
Water Soluble Vitamins	0.57 (0.52-0.62)	0.57 (0.52-0.62)	0.56 (0.51-0.61)
Nondepolarizing Muscle Relaxants	0.56 (0.52-0.60)	0.57 (0.52-0.61)	0.54 (0.51-0.58)
Urinary Antispasmodic - Antimuscarinics (Antich...	0.59 (0.51-0.66)	0.57 (0.49-0.64)	0.59 (0.51-0.66)
Bulk Laxatives	0.53 (0.46-0.61)	0.56 (0.50-0.64)	0.53 (0.45-0.61)
Antiemetics - Anticholinergic	0.53 (0.46-0.60)	0.56 (0.50-0.63)	0.45 (0.39-0.51)
Aminopenicillins	0.55 (0.49-0.60)	0.56 (0.51-0.62)	0.54 (0.48-0.59)
Serotonin Modulators	0.53 (0.48-0.58)	0.56 (0.51-0.61)	0.53 (0.48-0.59)
Viral Vaccines	0.56 (0.51-0.61)	0.56 (0.51-0.60)	0.52 (0.47-0.57)
Laxative Combinations	0.57 (0.52-0.63)	0.56 (0.50-0.61)	0.55 (0.50-0.61)
Antimyasthenic/Cholinergic Agents	0.54 (0.50-0.59)	0.56 (0.50-0.60)	0.52 (0.46-0.58)
Azithromycin	0.59 (0.54-0.63)	0.55 (0.51-0.60)	0.57 (0.53-0.63)
Local Anesthetics - Topical	0.54 (0.49-0.58)	0.55 (0.51-0.60)	0.52 (0.47-0.56)
ACE Inhibitors	0.51 (0.47-0.55)	0.55 (0.50-0.60)	0.48 (0.43-0.52)
Herpes Agents	0.53 (0.45-0.59)	0.55 (0.48-0.62)	0.48 (0.41-0.55)
Influenza Agents	0.51 (0.44-0.58)	0.54 (0.48-0.61)	0.45 (0.38-0.53)

Continued on next page

Table A.3: Mean (95% confidence interval) area under the ROC curve for predicting medication usage, grouped by pharmaceutical subclass, using the baseline, methylation data, and genotype data. Confidence intervals determined using bootstrapping.

Pharmaceutical Subclass	Baseline	Methylation	Genotypes
Beta Blockers Non-Selective	0.52 (0.44-0.59)	0.54 (0.45-0.64)	0.59 (0.50-0.69)
Sulfonylureas	0.59 (0.51-0.66)	0.54 (0.47-0.62)	0.61 (0.55-0.69)
Nasal Steroids	0.56 (0.51-0.60)	0.54 (0.50-0.59)	0.51 (0.46-0.57)
Antispasmodics	0.56 (0.52-0.60)	0.54 (0.50-0.58)	0.56 (0.52-0.59)
Cough/Cold/Allergy Combinations	0.58 (0.52-0.65)	0.54 (0.48-0.61)	0.55 (0.49-0.63)
Cephalosporins - 2nd Generation	0.52 (0.44-0.59)	0.54 (0.47-0.62)	0.49 (0.42-0.56)
Opioid Combinations	0.53 (0.49-0.57)	0.54 (0.50-0.58)	0.50 (0.46-0.53)
Anticonvulsants - Misc.	0.54 (0.49-0.58)	0.54 (0.50-0.58)	0.51 (0.47-0.55)
Multiple Vitamins w/ Minerals	0.58 (0.51-0.65)	0.54 (0.47-0.59)	0.54 (0.49-0.60)
Protamine	0.53 (0.46-0.61)	0.54 (0.44-0.64)	0.50 (0.39-0.59)
Lincosamides	0.52 (0.46-0.59)	0.53 (0.47-0.59)	0.52 (0.44-0.59)
Hemostatics - Systemic	0.46 (0.37-0.55)	0.53 (0.44-0.62)	0.54 (0.47-0.62)
Antihistamines - Non-Sedating	0.54 (0.47-0.60)	0.52 (0.46-0.57)	0.53 (0.46-0.58)
Acne Products	0.59 (0.49-0.66)	0.50 (0.41-0.60)	0.41 (0.33-0.50)
Miscellaneous Contrast Media	0.49 (0.46-0.53)	0.50 (0.46-0.54)	0.53 (0.49-0.57)
Nonsteroidal Anti-inflammatory Agents (NSAIDs)	0.52 (0.49-0.57)	0.49 (0.45-0.53)	0.53 (0.48-0.57)
Toxoid Combinations	0.55 (0.49-0.63)	0.49 (0.42-0.57)	0.55 (0.48-0.63)
Corticosteroids - Topical	0.51 (0.46-0.56)	0.49 (0.43-0.53)	0.43 (0.38-0.47)
Tricyclic Agents	0.53 (0.44-0.61)	0.48 (0.39-0.56)	0.58 (0.51-0.65)
Central Muscle Relaxants	0.52 (0.48-0.57)	0.48 (0.43-0.52)	0.48 (0.43-0.53)
Depolarizing Muscle Relaxants	0.53 (0.47-0.58)	0.46 (0.40-0.51)	0.49 (0.43-0.54)
Antifungals - Topical	0.50 (0.44-0.55)	0.45 (0.40-0.50)	0.48 (0.43-0.54)
Cardiac Glycosides	0.52 (0.42-0.60)	0.43 (0.33-0.52)	0.54 (0.46-0.63)
Alternative Medicine - C's	0.53 (0.47-0.60)	0.37 (0.30-0.44)	0.51 (0.45-0.60)

Table A.4: Mean (95% confidence interval) R^2 for predicting the most recent lab result using the baseline, methylation data, and genotype data. Confidence intervals determined using bootstrapping. Activated Partial Thromboplastin Time (APTT); Point of care (POC); Pulmonary function test (PFT); Forced expiratory volume in 1 second (FEV1)

Lab Test	Baseline	Methylation	Genotypes
Troponin	0.65 (0.53-0.74)	0.62 (0.51-0.72)	0.60 (0.50-0.70)
Creatinine	0.08 (0.01-0.13)	0.43 (0.38-0.47)	0.09 (0.04-0.13)
Troponin interpretation	0.46 (0.31-0.62)	0.41 (0.26-0.54)	0.33 (0.20-0.48)
Urea nitrogen	0.01 (-0.04-0.05)	0.40 (0.35-0.45)	0.03 (-0.00-0.05)
Absolute eosinophil count	-0.06 (-0.10–0.03)	0.33 (0.27-0.40)	-0.01 (-0.02-0.00)
Hemoglobin	0.10 (0.04-0.15)	0.28 (0.23-0.33)	0.10 (0.05-0.14)
Neutrophil percent (auto)	0.23 (0.12-0.32)	0.26 (0.18-0.33)	0.22 (0.13-0.30)
PFT FEV1 (pre)	0.12 (-0.18-0.36)	0.26 (0.07-0.38)	0.18 (-0.02-0.32)
Hematocrit	0.07 (0.02-0.11)	0.24 (0.20-0.28)	0.07 (0.04-0.11)
Mean corpuscular hemoglobin	0.07 (0.01-0.13)	0.20 (0.15-0.26)	0.08 (0.04-0.12)
Mean corpuscular volume	0.09 (0.03-0.14)	0.18 (0.12-0.24)	0.09 (0.04-0.13)
Absolute lymphocyte count	0.06 (-0.04-0.17)	0.17 (0.08-0.33)	0.10 (0.03-0.23)
Platelet count (auto)	-0.00 (-0.05-0.05)	0.16 (0.12-0.21)	0.02 (-0.00-0.04)
Absolute neutrophil count	0.08 (0.01-0.13)	0.15 (0.10-0.20)	0.08 (0.04-0.11)
Albumin	0.07 (0.01-0.13)	0.14 (0.08-0.18)	0.08 (0.04-0.12)
Chloride	-0.01 (-0.05-0.03)	0.13 (0.09-0.18)	0.01 (-0.01-0.03)
Absolute immature granulocyte count	-0.09 (-0.25-0.01)	0.13 (0.05-0.20)	-0.00 (-0.04-0.02)
Absolute monocyte count	0.08 (0.00-0.14)	0.12 (0.06-0.17)	0.07 (0.01-0.13)
White blood cell count	-0.01 (-0.07-0.04)	0.11 (0.05-0.19)	0.02 (-0.01-0.05)
Neutrophils absolute (prelim).	0.07 (0.01-0.13)	0.11 (0.06-0.18)	0.08 (0.03-0.13)
HgbA1C	-0.07 (-0.15-0.01)	0.11 (0.06-0.17)	-0.01 (-0.04-0.01)
Total protein	0.09 (0.03-0.14)	0.11 (0.06-0.16)	0.08 (0.04-0.12)
Sodium	-0.00 (-0.04-0.04)	0.10 (0.07-0.14)	0.03 (0.00-0.05)

Continued on next page

Table A.4: Mean (95% confidence interval) R^2 for predicting the most recent lab result using the baseline, methylation data, and genotype data. Confidence intervals determined using bootstrapping. Activated Partial Thromboplastin Time (APTT); Point of care (POC); Pulmonary function test (PFT); Forced expiratory volume in 1 second (FEV1)

Lab Test	Baseline	Methylation	Genotypes
Ferritin	-0.21 (-0.39–0.07)	0.10 (0.04-0.15)	-0.02 (-0.05–0.00)
Sedimentation rate erythrocyte	-0.10 (-0.38-0.12)	0.10 (-0.01-0.18)	0.08 (-0.01-0.17)
Iron binding capacity	-0.10 (-0.19–0.00)	0.09 (0.03-0.14)	-0.00 (-0.03-0.03)
Absolute basophil count	-0.10 (-0.25–0.02)	0.07 (0.05-0.11)	-0.01 (-0.02–0.00)
Glucose	-0.02 (-0.07-0.02)	0.07 (0.04-0.10)	0.01 (-0.01-0.02)
Qrs.duration	-0.04 (-0.11-0.02)	0.05 (0.00-0.08)	0.01 (-0.01-0.02)
Cholesterol HDL	0.03 (-0.07-0.12)	0.04 (-0.03-0.10)	0.06 (0.00-0.12)
Hematocrit OSL	-0.31 (-0.81-0.09)	0.04 (-0.13-0.19)	-0.05 (-0.22-0.08)
Cholesterol	-0.05 (-0.11-0.02)	0.03 (-0.02-0.08)	0.01 (-0.02-0.04)
Ventricular rate	-0.06 (-0.14-0.01)	0.03 (-0.00-0.06)	0.02 (-0.01-0.04)
Anion gap	-0.04 (-0.07–0.01)	0.02 (-0.00-0.05)	-0.00 (-0.02-0.01)
Alanine aminotransferase	-0.05 (-0.09–0.02)	0.02 (0.01-0.04)	-0.01 (-0.01-0.00)
R axis	-0.03 (-0.10-0.03)	0.01 (-0.01-0.04)	0.00 (-0.02-0.03)
Magnesium	-0.11 (-0.19–0.04)	0.01 (-0.02-0.05)	-0.00 (-0.03-0.02)
Alkaline phosphatase	-0.02 (-0.07-0.02)	0.01 (-0.01-0.03)	0.00 (-0.01-0.01)
T.axis	-0.08 (-0.19–0.01)	0.01 (-0.02-0.04)	-0.00 (-0.01-0.01)
Cholesterol LDL (calculated)	-0.10 (-0.22–0.02)	0.01 (-0.02-0.04)	-0.01 (-0.03-0.01)
Potassium	-0.03 (-0.07-0.00)	0.01 (-0.01-0.02)	-0.01 (-0.02-0.01)
Aspartate aminotransferase	-0.02 (-0.07-0.03)	0.01 (-0.01-0.03)	0.00 (-0.02-0.02)
International normalized ratio (INR)	-0.02 (-0.06-0.02)	0.01 (-0.01-0.03)	0.00 (-0.02-0.02)
Bilirubin total	-0.05 (-0.16–0.01)	0.01 (-0.02-0.02)	-0.00 (-0.02-0.00)
Prothrombin time	-0.02 (-0.10-0.02)	0.01 (-0.01-0.03)	0.00 (-0.01-0.02)
QT interval	-0.07 (-0.12–0.01)	0.00 (-0.02-0.02)	0.01 (-0.01-0.02)

Continued on next page

Table A.4: Mean (95% confidence interval) R^2 for predicting the most recent lab result using the baseline, methylation data, and genotype data. Confidence intervals determined using bootstrapping. Activated Partial Thromboplastin Time (APTT); Point of care (POC); Pulmonary function test (PFT); Forced expiratory volume in 1 second (FEV1)

Lab Test	Baseline	Methylation	Genotypes
Glucose (POC)	-0.08 (-0.21-0.02)	0.00 (-0.03-0.03)	0.00 (-0.04-0.04)
X saturation	-0.18 (-0.30–0.06)	0.00 (-0.04-0.04)	-0.01 (-0.03–0.00)
PR interval	-0.05 (-0.15-0.03)	0.00 (-0.04-0.03)	-0.00 (-0.03-0.02)
Brain natriuretic peptide (BNP)	-0.51 (-1.38–0.03)	0.00 (-0.08-0.05)	0.03 (-0.07-0.11)
Glucose whole blood	-0.24 (-0.49–0.09)	-0.00 (-0.03-0.03)	-0.03 (-0.08-0.00)
Atrial rate	-0.05 (-0.13-0.01)	-0.00 (-0.02-0.01)	-0.00 (-0.02-0.02)
P axis	-0.04 (-0.10-0.01)	-0.00 (-0.02-0.02)	-0.00 (-0.02-0.02)
QtC calculation (bezet)	-0.07 (-0.14–0.01)	-0.00 (-0.02-0.02)	-0.00 (-0.02-0.01)
PFT FEV1 (pre) (percent ref)	-0.31 (-0.58–0.06)	-0.00 (-0.06-0.04)	-0.03 (-0.07–0.01)
Blood lactate	-0.29 (-0.64-0.02)	-0.01 (-0.08-0.05)	0.00 (-0.06-0.05)
APTT	-0.05 (-0.12–0.01)	-0.01 (-0.02-0.00)	-0.01 (-0.03-0.00)
Triglycerides	-0.13 (-0.23–0.05)	-0.01 (-0.03-0.01)	-0.01 (-0.03–0.00)
Thyroid stimulating hormone (TSH)	-0.13 (-0.37–0.05)	-0.01 (-0.04–0.01)	-0.02 (-0.04–0.00)
Calcium	-0.05 (-0.08–0.02)	-0.01 (-0.02–0.00)	-0.02 (-0.03–0.00)
Iron	-0.13 (-0.26–0.00)	-0.02 (-0.06-0.03)	-0.02 (-0.05-0.01)
Urea nitrogen (OSL)	-0.61 (-1.93-0.00)	-0.04 (-0.17-0.05)	-0.04 (-0.22-0.03)
Bilirubin conjugated	-0.57 (-1.25–0.17)	-0.06 (-0.18–0.00)	-0.06 (-0.17–0.03)
C reactive protein (CRP)	-0.44 (-1.14–0.13)	-0.06 (-0.24–0.01)	-0.01 (-0.08-0.02)
Left ventricular ejection fraction	-1.80 (-3.29–0.91)	-0.06 (-0.21-0.01)	-0.07 (-0.28-0.03)
Hemoglobin (OSL)	-2.66 (-12.80–0.23)	-0.07 (-0.37–0.02)	-0.09 (-0.45–0.02)
Chloride (OSL)	-1.96 (-4.43–0.46)	-0.19 (-0.70-0.03)	-0.17 (-0.65-0.04)
Sodium (OSL)	-304.86 (-2851.16–0.13)	-2.03 (-16.23–0.02)	-1.57 (-16.42–0.01)
Potassium (OSL).	-178.77 (-828.09–0.14)	-2.33 (-8.10–0.02)	-2.42 (-8.14–0.02)

[1] GTEx Consortium. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[2] The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

[3] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[4] The "all of us"research program. *New England Journal of Medicine*, 381(7):668–676, 2022/05/06 2019.

[5] Po-Ru Loh, Gleb Kichaev, Steven Gazal, Armin P. Schoech, and Alkes L. Price. Mixed-model association for biobank-scale datasets. *Nature Genetics*, 50(7):906–908, 2018.

[6] GTEx Consortium, François Aguet, Andrew A. Brown, Stephane E. Castel, Joe R. Davis, Yuan He, Brian Jo, Pejman Mohammadi, YoSon Park, Princy Parsana, Ayellet V. Segrè, Benjamin J. Strober, Zachary Zappala, Beryl B. Cummings, Ellen T. Gelfand, Kane Hadley, Katherine H. Huang, Monkol Lek, Xiao Li, Jared L. Nedzel, Duyen Y. Nguyen, Michael S. Noble, Timothy J. Sullivan, Taru Tukiainen, Daniel G. MacArthur, Gad Getz, Anjene Addington, Ping Guan, Susan Koester, A. Roger Little, Nicole C. Lockhart, Helen M. Moore, Abhi Rao, Jeffery P. Struewing, Simona Volpi, Lori E. Brigham, Richard Hasz, Marcus Hunter, Christopher Johns, Mark Johnson, Gene Kopen, William F. Leinweber, John T. Lonsdale, Alisa McDonald, Bernadette Mestichelli, Kevin Myer, Bryan Roe, Michael Salvatore, Saboor Shad, Jeffrey A. Thomas, Gary Walters, Michael Washington, Joseph Wheeler, Jason Bridge, Barbara A. Foster, Bryan M. Gillard, Ellen Karasik, Rachna Kumar, Mark Miklos, Michael T. Moser, Scott D. Jewell, Robert G. Montroy, Daniel C. Rohrer, Dana Valley, Deborah C. Mash, David A. Davis, Leslie Sobin, Mary E. Barcus, Philip A. Branton, Nathan S. Abell, Brunilda Balliu, Olivier Delaneau, Laure Frésard, Eric R. Gamazon, Diego Garrido-Martín, Ariel D. H. Gewirtz, Genna Gliner, Michael J. Gloudemans, Buhm Han, Amy Z. He, Farhad Hormozdiari, Xin Li, Boxiang Liu, Eun Yong Kang, Ian C. McDowell, Halit Ongen, John J. Palowitch, Christine B. Peterson, Gerald Quon, Stephan Ripke, Ashis Saha, Andrey A. Shabalin, Tyler C. Shimko, Jae Hoon Sul, Nicole A. Teran, Emily K. Tsang, Hailei Zhang, Yi-Hui Zhou, Carlos D. Bustamante, Nancy J. Cox, Roderic Guigó, Manolis Kellis, Mark I. McCarthy, Donald F. Conrad, Eleazar Eskin, Gen Li, Andrew B. Nobel, Chiara Sabatti, Barbara E. Stranger, Xiaoquan Wen, Fred A. Wright, Kristin G. Ardlie, Emmanouil T. Dermitzakis, Tuuli Lappalainen, Kristin G. Ardlie, Beryl B. Cummings, Ellen T. Gelfand, Robert E. Handsaker, Katherine H. Huang, Seva Kashin, Konrad J. Karczewski, Daniel G. MacArthur,

Jared L. Nedzel, Duyen T. Nguyen, Michael S. Noble, Ayellet V. Segrè, Casandra A. Trowbridge, Nathan S. Abell, Ruth Barshir, Omer Basha, Alexis Battle, Gireesh K. Bogu, Andrew Brown, Christopher D. Brown, Stephane E. Castel, Lin S. Chen, Colby Chiang, Donald F. Conrad, Nancy J. Cox, Farhan N. Damani, Joe R. Davis, Emmanouil T. Dermitzakis, Barbara E. Engelhardt, Pedro G. Ferreira, Eric R. Gamazon, Ariel D. H. Gewirtz, Michael J. Gloudemans, Roderic Guigo, Ira M. Hall, Cedric Howald, Hae Kyung Im, Eun Yong Kang, Yungil Kim, Sarah Kim-Hellmuth, Serghei Mangul, Mark I. McCarthy, Ian C. McDowell, Jean Monlong, Stephen B. Montgomery, Manuel Muñoz-Aguirre, Anne W. Ndungu, Dan L. Nicolae, Andrew B. Nobel, Meritxell Oliva, John J. Palowitch, Nikolaos Panousis, Panagiotis Papasaikas, Anthony J. Payne, Christine B. Peterson, Jie Quan, Ferran Reverter, Michael Sammeth, Alexandra J. Scott, Andrey A. Shabalin, Reza Sodaei, Matthew Stephens, Barbara E. Stranger, Benjamin J. Strober, Emily K. Tsang, Sarah Urbut, Martijn van de Bunt, Gao Wang, Fred A. Wright, Hualin S. Xi, Esti Yeger-Lotem, Judith B. Zaugg, Joshua M. Akey, Daniel Bates, Joanne Chan, Lin S. Chen, Melina Claussnitzer, Kathryn Demanelis, Morgan Diegel, Jennifer A. Doherty, Andrew P. Feinberg, Marian S. Fernando, Jessica Halow, Kasper D. Hansen, Eric Haugen, Peter F. Hickey, Lei Hou, Farzana Jasmine, Ruiqi Jian, Lihua Jiang, Audra Johnson, Rajinder Kaul, Muhammad G. Kibriya, Kristen Lee, Jin Billy Li, Qin Li, Jessica Lin, Shin Lin, Sandra Linder, Caroline Linke, Yaping Liu, Matthew T. Maurano, Benoit Molinie, Stephen B. Montgomery, Jemma Nelson, Fidencio J. Neri, Yongjin Park, Brandon L. Pierce, Nicola J. Rinaldi, Lindsay F. Rizzardi, Richard Sandstrom, Andrew Skol, Kevin S. Smith, Michael P. Snyder, John Stamatoyannopoulos, Barbara E. Stranger, Hua Tang, Emily K. Tsang, Li Wang, Meng Wang, Nicholas Van Wittenberghe, Fan Wu, Rui Zhang, Concepcion R. Nierras, Philip A. Branton, Latarsha J. Carithers, Helen M. Moore, Jimmie B. Vaught, Sarah E. Gould, Nicole C. Lockart, Casey Martin, Jeffery P. Struewing, Anjene M. Addington, Susan E. Koester, A. Roger Little, Lori E. Brigham, William F. Leinweber, John T. Lonsdale, Brian Roe, Jeffrey A. Thomas, Barbara A. Foster, Bryan M. Gillard, Michael T. Moser, Scott D. Jewell, Robert G. Montroy, Daniel C. Rohrer, Dana R. Valley, David A. Davis, Deborah C. Mash, Anita H. Undale, Anna M. Smith, David E. Tabor, Nancy V. Roche, Jeffrey A. McLean, Negin Vatanian, Karna L. Robinson, Mary E. Barcus, Kimberly M. Valentino, Liqun Qi, Steven Hunter, Pushpa Hariharan, Shilpi Singh, Ki Sung Um, Takunda Matose, Maria M. Tomaszewski, Laura K. Barker, Maghboeba Mosavel, Laura A. Siminoff, Heather M. Traino, Paul Flicek, Thomas Juettemann, Magali Ruffier, Dan Sheppard, Kieron Taylor, Stephen J. Trevanion, Daniel R. Zerbino, Brian Craft, Mary Goldman, Maximilian Haeussler, W. James Kent, Christopher M. Lee, Benedict Paten, Kate R. Rosenbloom, John Vivian, and Jingchun Zhu. Genetic effects on gene expression across human tissues. *Nature*, 550:204 EP –, 10 2017.

[7] Tom R. Gaunt, Hashem A. Shihab, Gibran Hemani, Josine L. Min, Geoff Woodward, Oliver Lyttleton, Jie Zheng, Aparna Duggirala, Wendy L. McArdle, Karen Ho, Susan M. Ring, David M. Evans, George Davey Smith, and Caroline L. Relton. Systematic identification of genetic influences on methylation across the human life course.

Genome Biology, 17(1):61, 2016.

[8] William S. Bush and Jason H. Moore. Chapter 11: Genome-wide association studies. *PLOS Computational Biology*, 8(12):1–11, 12 2012.

[9] Peter Kraft, Eleftheria Zeggini, and John P. A. Ioannidis. Replication in Genome-Wide Association Studies. *Statistical Science*, 24(4):561 – 573, 2009.

[10] Mike Thompson, Zeyuan Johnson Chen, Elior Rahmani, and Eran Halperin. Confined: distinguishing biological from technical sources of variation by leveraging multiple methylation datasets. *Genome Biology*, 20(1):138, 2019.

[11] Florian Schmidt, Markus List, Engin Cukuroglu, Sebastian Köhler, Jonathan Göke, and Marcel H Schulz. An ontology-based method for assessing batch effect adjustment approaches in heterogeneous datasets. *Bioinformatics*, 34(17):i908–i916, 2018.

[12] Jovana Maksimovic, Johann A Gagnon-Bartsch, Terence P Speed, and Alicia Oshlack. Removing unwanted variation in a differential methylation analysis of illumina humanmethylation450 array data. *Nucleic acids research*, 43(16):e106–e106, 09 2015.

[13] Steve Horvath. Dna methylation age of human tissues and cell types. *Genome Biology*, 14(10):3156, 2013.

[14] Michael J Rae, Robert N Butler, Judith Campisi, Aubrey D N J de Grey, Caleb E Finch, Michael Gough, George M Martin, Jan Vijg, Kevin M Perrott, and Barbara J Logan. The demographic and biomedical case for late-life interventions in aging. *Science translational medicine*, 2(40):40cm21–40cm21, 07 2010.

[15] Luigi Ferrucci, Charles Hesdorffer, Stefania Bandinelli, and Eleanor M. Simonsick. Frailty as a nexus between the biology of aging, environmental conditions and clinical geriatrics. *Public Health Reviews*, 32(2):475–488, 2010.

[16] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

[17] Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

[18] Oliver Stegle, Leopold Parts, Richard Durbin, and John Winn. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLOS Computational Biology*, 6(5):1–11, 05 2010.

[19] Ken Lee and Zdenka Pausova. Cigarette smoking and dna methylation. *Frontiers in Genetics*, 4:132, 2013.

[20] Paula Singmann, Doron Shem-Tov, Simone Wahl, Harald Grallert, Giovanni Fiorito, So-Youn Shin, Katharina Schramm, Petra Wolf, Sonja Kunze, Yael Baran, Simonetta Guarrera, Paolo Vineis, Vittorio Krogh, Salvatore Panico, Rosario Tumino, Anja Kretschmer, Christian Gieger, Annette Peters, Holger Prokisch, Caroline L. Relton, Giuseppe Matullo, Thomas Illig, Melanie Waldenberger, and Eran Halperin. Characterization of whole-genome autosomal differences of dna methylation between men and women. *Epigenetics & Chromatin*, 8(1):43, Oct 2015.

[21] James Flanagan. *Epigenome-Wide Association Studies (EWAS): Past, present, and future*, volume 1238. 11 2015.

[22] Elior Rahmani, Liat Shenhav, Regev Schweiger, Paul Yousefi, Karen Huen, Brenda Eskenazi, Celeste Eng, Scott Huntsman, Donglei Hu, Joshua Galanter, Sam S. Oh, Melanie Waldenberger, Konstantin Strauch, Harald Grallert, Thomas Meitinger, Christian Gieger, Nina Holland, Esteban G. Burchard, Noah Zaitlen, and Eran Halperin. Genome-wide methylation data mirror ancestry information. *Epigenetics & Chromatin*, 10(1):1, 2017.

[23] Joshua M Galanter, Christopher R Gignoux, Sam S Oh, Dara Torgerson, Maria Pino-Yanes, Neeta Thakur, Celeste Eng, Donglei Hu, Scott Huntsman, Harold J Farber, Pedro C Avila, Emerita Brigino-Buenaventura, Michael A LeNoir, Kelly Meade, Denise Serebrisky, William Rodríguez-Cintrón, Rajesh Kumar, Jose R Rodríguez-Santana, Max A Seibold, Luisa N Borrell, Esteban G Burchard, and Noah Zaitlen. Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. *eLife*, 6:e20532, jan 2017.

[24] Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11:733 EP –, 09 2010.

[25] Lee Joseph Cronbach and Noreen M. Webb. Between-class and within-class effects in a reported aptitude x treatment interaction: Reanalysis of a study by g. l. anderson. 1975.

[26] Andrew Lu, Mike Thompson, M Grace Gordon, Andy Dahl, Chun Jimmie Ye, Noah Zaitlen, and Brunilda Balliu. Fast and powerful statistical method for context-specific qtl mapping in multi-context genomic studies. *bioRxiv*, 2021.

[27] Mike Thompson, Mary Grace Gordon, Andrew Lu, Anchit Tandon, Eran Halperin, Alexander Gusev, Chun Jimmie Ye, Brunilda Balliu, and Noah Zaitlen. Multi-context genetic modeling of transcriptional regulation resolves novel disease loci. *bioRxiv*, 2021.

[28] Cathryn M. Lewis and Evangelos Vassos. Polygenic risk scores: from research tools to clinical instruments. *Genome Medicine*, 12(1):44, 2020.

[29] Sebastien Haneuse, David Arterburn, and Michael J. Daniels. Assessing Missing Data Assumptions in EHR-Based Studies: A Complex and Underappreciated Task. *JAMA Network Open*, 4(2):e210184–e210184, 02 2021.

[30] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, 12(3):e1001779, March 2015. Publisher: Public Library of Science.

[31] Catherine A McCarty, Russell A Wilke, Philip F Giampietro, Steve D Wesbrook, and Michael D Caldwell. Marshfield clinic personalized medicine research project (pmrp): design, methods and recruitment for a large population-based biobank. *Personalized Medicine*, 2(1):49–79, 2021/01/04 2005.

[32] D M Roden, J M Pulley, M A Basford, G R Bernard, E W Clayton, J R Balser, and D R Masys. Development of a large-scale de-identified dna biobank to enable personalized medicine. *Clin Pharmacol Ther*, 84(3):362–369, Sep 2008.

[33] Amit V Khera, Mark Chaffin, Krishna G Aragam, Mary E Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S Lander, Steven A Lubitz, Patrick T Ellinor, and Sekar Kathiresan. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*, 50(9):1219–1224, Sep 2018.

[34] Nasim Mavaddat, Kyriaki Michailidou, Joe Dennis, Michael Lush, Laura Fachal, Andrew Lee, Jonathan P Tyrer, Ting-Huei Chen, Qin Wang, Manjeet K Bolla, Xin Yang, Muriel A Adank, Thomas Ahearn, Kristiina Aittomäki, Jamie Allen, Irene L Andrulis, Hoda Anton-Culver, Natalia N Antonenkova, Volker Arndt, Kristan J Aronson, Paul L Auer, Päivi Auvinen, Myrto Barrdahl, Laura E Beane Freeman, Matthias W Beckmann, Sabine Behrens, Javier Benitez, Marina Bermisheva, Leslie Bernstein, Carl Blomqvist, Natalia V Bogdanova, Stig E Bojesen, Bernardo Bonanni, Anne-Lise Børresen-Dale, Hiltrud Brauch, Michael Bremer, Hermann Brenner, Adam Brentnall, Ian W Brock, Angela Brooks-Wilson, Sara Y Brucker, Thomas Brüning, Barbara Burwinkel, Daniele Campa, Brian D Carter, Jose E Castelao, Stephen J Chanock, Rowan Chlebowski, Hans Christiansen, Christine L Clarke, J Margriet Collée, Emilie Cordina-Duverger, Sten Cornelissen, Fergus J Couch, Angela Cox, Simon S Cross, Kamila Czene, Mary B Daly, Peter Devilee, Thilo Dörk, Isabel Dos-Santos-Silva, Martine Dumont, Lorraine Durcan, Miriam Dwek, Diana M Eccles, Arif B Ekici, A Heather Eliassen, Carolina Ellberg, Christoph Engel, Mikael Eriksson, D Gareth Evans, Peter A Fasching, Jonine Figueroa, Olivia Fletcher, Henrik Flyger, Asta Försti, Lin Fritschi, Marike Gabrielson, Manuela Gago-Dominguez, Susan M Gapstur, JoséA García-Sáenz, Mia M Gaudet, Vassilios Georgoulias, Graham G Giles, Irina R Gilyazova, Gord Glendon, Mark S Goldberg, David E Goldgar, Anna González-Neira, Grethe I Grenaker Alnæs, Mervi Grip, Jacek Gronwald, Anne Grundy, Pascal Guénel, Lothar Haeberle, Eric

Hahnen, Christopher A Haiman, Niclas Håkansson, Ute Hamann, Susan E Hankinson, Elaine F Harkness, Steven N Hart, Wei He, Alexander Hein, Jane Heyworth, Peter Hillemanns, Antoinette Hollestelle, Maartje J Hooning, Robert N Hoover, John L Hopper, Anthony Howell, Guanmengqian Huang, Keith Humphreys, David J Hunter, Milena Jakimovska, Anna Jakubowska, Wolfgang Janni, Esther M John, Nichola Johnson, Michael E Jones, Arja Jukkola-Vuorinen, Audrey Jung, Rudolf Kaaks, Katarzyna Kaczmarek, Vesa Kataja, Renske Keeman, Michael J Kerin, Elza Khusnutdinova, Johanna I Kiiski, Julia A Knight, Yon-Dschun Ko, Veli-Matti Kosma, Stella Koutros, Vessela N Kristensen, Ute Krüger, Tabea Kühl, Diether Lambrechts, Loic Le Marchand, Eunjung Lee, Flavio Lejbkowicz, Jenna Lilyquist, Annika Lindblom, Sara Lindström, Jolanta Lissowska, Wing-Yee Lo, Sibylle Loibl, Jirong Long, Jan Lubiński, Michael P Lux, Robert J MacInnis, Tom Maishman, Enes Makalic, Ivana Maleva Kostovska, Arto Mannermaa, Siranoush Manoukian, Sara Margolin, John W M Martens, Maria Elena Martinez, Dimitrios Mavroudis, Catriona McLean, Alfons Meindl, Usha Menon, Pooja Middha, Nicola Miller, Fernando Moreno, Anna Marie Mulligan, Claire Mulot, Victor M Muñoz-Garzon, Susan L Neuhausen, Heli Nevanlinna, Patrick Neven, William G Newman, Sune F Nielsen, Børge G Nordestgaard, Aaron Norman, Kenneth Offit, Janet E Olson, Håkan Olsson, Nick Orr, V Shane Pankratz, Tjoung-Won Park-Simon, Jose I A Perez, Clara Pérez-Barrios, Paolo Peterlongo, Julian Peto, Mila Pinchev, Dijana Plaseska-Karanfilska, Eric C Polley, Ross Prentice, Nadege Presneau, Darya Prokofyeva, Kristen Purrington, Katri Pylkäs, Brigitte Rack, Paolo Radice, Rohini Rau-Murthy, Gad Rennert, Hedy S Rennert, Valerie Rhenius, Mark Robson, Atocha Romero, Kathryn J Ruddy, Matthias Ruebner, Emmanouil Saloustros, Dale P Sandler, Elinor J Sawyer, Daniel F Schmidt, Rita K Schmutzler, Andreas Schneeweiss, Minouk J Schoemaker, Fredrick Schumacher, Peter Schürmann, Lukas Schwentner, Christopher Scott, Rodney J Scott, Caroline Seynaeve, Mitul Shah, Mark E Sherman, Martha J Shrubsole, Xiao-Ou Shu, Susan Slager, Ann Smeets, Christof Sohn, Penny Soucy, Melissa C Southey, John J Spinelli, Christa Stegmaier, Jennifer Stone, Anthony J Swerdlow, Rulla M Tamimi, William J Tapper, Jack A Taylor, Mary Beth Terry, Kathrin Thöne, Rob A E M Tollenaar, Ian Tomlinson, Thérèse Truong, Maria Tzardi, Hans-Ulrich Ulmer, Michael Untch, Celine M Vachon, Elke M van Veen, Joseph Vijai, Clarice R Weinberg, Camilla Wendt, Alice S Whittemore, Hans Wildiers, Walter Willett, Robert Winqvist, Alicja Wolk, Xiaohong R Yang, Drakoulis Yannoukakos, Yan Zhang, Wei Zheng, Argyrios Ziogas, Alison M Dunning, Deborah J Thompson, Georgia Chenevix-Trench, Jenny Chang-Claude, Marjanka K Schmidt, Per Hall, Roger L Milne, Paul D P Pharoah, Antonis C Antoniou, Nilanjan Chatterjee, Peter Kraft, Montserrat García-Closas, Jacques Simard, and Douglas F Easton. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am J Hum Genet*, 104(1):21–34, Jan 2019.

[35] Cathryn M Lewis and Saskia P Hagenaars. Progressing polygenic medicine in psychiatry through electronic health records. *JAMA Psychiatry*, 76(5):470–472, May 2019.

[36] Miklos D. Kertai, Jonathan D. Mosley, Jing He, Abinaya Ramakrishnan, Mark J. Abdelmalak, Yurim Hong, M. Benjamin Shoemaker, Dan M. Roden, and Lisa Bastarache.

Predictive accuracy of a polygenic risk score for postoperative atrial fibrillation after cardiac surgery. *Circulation: Genomic and Precision Medicine*, 14(2):e003269, 2021/07/01 2021.

[37] Feras Hatib, Zhongping Jian, Sai Buddi, Christine Lee, Jos Settels, Karen Sibert, Joseph Rinehart, and Maxime Cannesson. Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis. *Anesthesiology*, 129(4):663–674, October 2018.

[38] Marije Wijnberge, Bart F. Geerts, Liselotte Hol, Nikki Lemmers, Marijn P. Mulder, Patrick Berge, Jimmy Schenk, Lotte E. Terwindt, Markus W. Hollmann, Alexander P. Vlaar, and Denise P. Veelo. Effect of a Machine Learning–Derived Early Warning System for Intraoperative Hypotension vs Standard Care on Depth and Duration of Intraoperative Hypotension During Elective Noncardiac Surgery: The HYPE Randomized Clinical Trial. *JAMA*, February 2020.

[39] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22):2402–2410, December 2016.

[40] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–12, New York, NY, USA, April 2020. Association for Computing Machinery.

[41] Alicia R. Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4):584–591, April 2019. Number: 4 Publisher: Nature Publishing Group.

[42] L. Duncan, H. Shen, B. Gelaye, J. Meijsen, K. Ressler, M. Feldman, R. Peterson, and B. Domingue. Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*, 10(1):3328, July 2019. Number: 1 Publisher: Nature Publishing Group.

[43] Karin B Michels, Alexandra M Binder, Sarah Dedeurwaerder, Charles B Epstein, John M Greally, Ivo Gut, E Andres Houseman, Benedetta Izzi, Karl T Kelsey, Alexander Meissner, Aleksandar Milosavljevic, Kimberly D Siegmund, Christoph Bock, and Rafael A Irizarry. Recommendations for the design and analysis of epigenome-wide association studies. *Nature Methods*, 10:949 EP –, 09 2013.

[44] Ino D. Karemaker and Michiel Vermeulen. Single-cell dna methylation profiling: Technologies and biological applications. *Trends in Biotechnology*, 36(9):952 – 965, 2018.

[45] Andrew E. Jaffe and Rafael A. Irizarry. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, 15(2):R31, Feb 2014.

[46] Johann A. Gagnon-Bartsch and Terence P. Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012.

[47] Elior Rahmani, Noah Zaitlen, Yael Baran, Celeste Eng, Donglei Hu, Joshua Galanter, Sam Oh, Esteban G Burchard, Eleazar Eskin, James Zou, and Eran Halperin. Correcting for cell-type heterogeneity in dna methylation: a comprehensive evaluation. *Nature Methods*, 14:218 EP –, 02 2017.

[48] Shijie C Zheng, Stephan Beck, Andrew E Jaffe, Devin C Koestler, Kasper D Hansen, Andres E Houseman, Rafael A Irizarry, and Andrew E Teschendorff. Correcting for cell-type heterogeneity in epigenome-wide association studies: revisiting previous analyses. *Nature Methods*, 14:216 EP –, 02 2017.

[49] Joanna D Holbrook, Rae-Chi Huang, Sheila J Barton, Richard Saffery, and Karen A Lillycrop. Is cellular heterogeneity merely a confounder to be removed from epigenome-wide association studies? *Epigenomics*, 9(8):1143–1150, 2019/01/13 2017.

[50] Eugene Andres Houseman, William P. Accomando, Devin C. Koestler, Brock C. Christensen, Carmen J. Marsit, Heather H. Nelson, John K. Wiencke, and Karl T. Kelsey. Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13(1):86, May 2012.

[51] Andrew E. Teschendorff and Caroline L. Relton. Statistical and integrative system-level analysis of dna methylation data. *Nature Reviews Genetics*, 19:129 EP –, 11 2017.

[52] Stephanie Lovinsky-Desir and Rachel L Miller. Epigenetics, asthma, and allergic diseases: a review of the latest advancements. *Current allergy and asthma reports*, 12(3):211–220, 06 2012.

[53] Andrea Baccarelli, Robert O. Wright, Valentina Bollati, Letizia Tarantini, Augusto A. Litonjua, Helen H. Suh, Antonella Zanobetti, David Sparrow, Pantel S. Vokonas, and Joel Schwartz. Rapid dna methylation changes after exposure to traffic particles. *American Journal of Respiratory and Critical Care Medicine*, 179(7):572–578, 2009. PMID: 19136372.

[54] E. Andres Houseman, Molly L. Kile, David C. Christiani, Tan A. Ince, Karl T. Kelsey, and Carmen J. Marsit. Reference-free deconvolution of dna methylation data and mediation by cell composition effects. *BMC Bioinformatics*, 17(1):259, Jun 2016.

[55] Elior Rahmani, Noah Zaitlen, Yael Baran, Celeste Eng, Donglei Hu, Joshua Galanter, Sam Oh, Esteban G Burchard, Eleazar Eskin, James Zou, and Eran Halperin. Sparse pca corrects for cell type heterogeneity in epigenome-wide association studies. *Nature Methods*, 13:443 EP –, 03 2016.

[56] James Zou, Christoph Lippert, David Heckerman, Martin Aryee, and Jennifer Listgarten. Epigenome-wide association studies without the need for cell-type composition. *Nature Methods*, 11:309 EP –, 01 2014.

[57] Pavlo Lutsik, Martin Slawski, Gilles Gasparoni, Nikita Vedeneev, Matthias Hein, and Jörn Walter. Medecom: discovery and quantification of latent components of heterogeneous methylomes. *Genome Biology*, 18(1):55, Mar 2017.

[58] Elior Rahmani, Regev Schweiger, Liat Shenhav, Theodora Wingert, Ira Hofer, Eilon Gabel, Eleazar Eskin, and Eran Halperin. Bayescce: a bayesian framework for estimating cell-type composition from dna methylation without the need for methylation reference. *Genome Biology*, 19(1):141, Sep 2018.

[59] Eugene Andres Houseman, John Molitor, and Carmen J. Marsit. Reference-free cell mixture adjustments in analysis of dna methylation data. *Bioinformatics*, 30(10):1431–1439, 2014.

[60] Elior Rahmani, Noah Zaitlen, Yael Baran, Celeste Eng, Donglei Hu, Joshua Galanter, Sam Oh, Esteban G Burchard, Eleazar Eskin, James Zou, and Eran Halperin. Correcting for cell-type heterogeneity in dna methylation: a comprehensive evaluation. *Nature Methods*, 14:218 EP –, 02 2017.

[61] James M. Flanagan. *Epigenome-Wide Association Studies (EWAS): Past, Present, and Future*, pages 51–63. Springer New York, New York, NY, 2015.

[62] Pedro Silva Moreira, Nadine Correia Santos, Nuno Sousa, and Patrício Soares Costa. The use of canonical correlation analysis to assess the relationship between executive functioning and verbal memory in older adults. *Gerontology & geriatric medicine*, 1:2333721415602820; 2333721415602820–2333721415602820, 08 2015.

[63] Alissa Sherry and Robin K. Henson. Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer. *Journal of Personality Assessment*, 84(1):37–48, 02 2005.

[64] Sami Sieranoja, Md Sahidullah, and Tomi Kinnunen. Audiovisual synchrony detection with optimized audio features. 2018.

[65] Brielin C. Brown, Nicolas L. Bray, and Lior Pachter. Expression reflects population structure. *PLOS Genetics*, 14(12):e1007841–, 12 2018.

[66] Elena Parkhomenko, David Tritchler, and Joseph Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 9, 2009.

[67] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics (Oxford, England)*, 10(3):515–534, 07 2009.

[68] Charlotte Soneson, Henrik Lilljebjörn, Thoas Fioretos, and Magnus Fontes. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. *BMC Bioinformatics*, 11(1):191, 2010.

[69] HAROLD HOTELLING. Relations between two sets of variates*. *Biometrika*, 28(3-4):321–377, 12 1936.

[70] Tessel E Galesloot, Kristel van Steen, Lambertus A L M Kiemeney, Luc L Janss, and Sita H Vermeulen. A comparison of multivariate genome-wide association methods. *PloS one*, 9(4):e95923; e95923–e95923, 04 2014.

[71] Michael Inouye, Samuli Ripatti, Johannes Kettunen, Leo-Pekka Lyytikäinen, Niku Oksala, Pirkka-Pekka Laurila, Antti J Kangas, Pasi Soininen, Markku J Savolainen, Jorma Viikari, Mika Kähönen, Markus Perola, Veikko Salomaa, Olli Raitakari, Terho Lehtimäki, Marja-Riitta Taskinen, Marjo-Riitta Järvelin, Mika Ala-Korpela, Aarno Palotie, and Paul I W de Bakker. Novel loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. *PLoS genetics*, 8(8):e1002907; e1002907–e1002907, 08 2012.

[72] Anna Cichonska, Juho Rousu, Pekka Marttinen, Antti J Kangas, Pasi Soininen, Terho Lehtimäki, Olli T Raitakari, Marjo-Riitta Järvelin, Veikko Salomaa, Mika Ala-Korpela, Samuli Ripatti, and Matti Pirinen. metacca: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics (Oxford, England)*, 32(13):1981–1989, 07 2016.

[73] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36:411 EP –, 04 2018.

[74] Belinda Phipson, Jovana Maksimovic, and Alicia Oshlack. missmethyl: an r package for analyzing data from illumina's humanmethylation450 platform. *Bioinformatics*, 32(2):286–288, 2016.

[75] Yun Liu, Martin J Aryee, Leonid Padyukov, M Daniele Fallin, Espen Hesselberg, Arni Runarsson, Lovisa Reinius, Nathalie Acevedo, Margaret Taub, Marcus Ronninger, Klementy Shchetynsky, Annika Scheynius, Juha Kere, Lars Alfredsson, Lars Klareskog, Tomas J Ekström, and Andrew P Feinberg. Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology*, 31:142 EP –, 01 2013.

[76] Gregory Hannum, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, SriniVas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, Rob Deconde, Menzies Chen, Indika Rajapakse, Stephen Friend, Trey Ideker, and Kang Zhang. Genome-wide

methylation profiles reveal quantitative views of human aging rates. *Molecular cell*, 49(2):359–367, 01 2013.

[77] Eilis Hannon, Emma Dempster, Joana Viana, Joe Burrage, Adam R. Smith, Ruby Macdonald, David St Clair, Colette Mustard, Gerome Breen, Sebastian Therman, Jaakko Kaprio, Timothea Toulopoulou, Hilleke E. Hulshoff Pol, Marc M. Bohlken, Rene S. Kahn, Igor Nenadic, Christina M. Hultman, Robin M. Murray, David A. Collier, Nick Bass, Hugh Gurling, Andrew McQuillin, Leonard Schalkwyk, and Jonathan Mill. An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential dna methylation. *Genome Biology*, 17(1):176, Aug 2016.

[78] Katie Lunnon, Rebecca Smith, Eilis Hannon, Philip L De Jager, Gyan Srivastava, Manuela Volta, Claire Troakes, Safa Al-Sarraj, Joe Burrage, Ruby Macdonald, Daniel Condliffe, Lorna W Harries, Pavel Katsel, Vahram Haroutunian, Zachary Kaminsky, Catharine Joachim, John Powell, Simon Lovestone, David A Bennett, Leonard C Schalkwyk, and Jonathan Mill. Methylomic profiling implicates cortical deregulation of ank1 in alzheimer's disease. *Nature Neuroscience*, 17:1164 EP –, 08 2014.

[79] Benjamin Lehne, Alexander W. Drong, Marie Loh, Weihua Zhang, William R. Scott, Sian-Tsung Tan, Uzma Afzal, James Scott, Marjo-Riitta Jarvelin, Paul Elliott, Mark I. McCarthy, Jaspal S. Kooner, and John C. Chambers. A coherent approach for analysis of the illumina humanmethylation450 beadchip improves data quality and performance in epigenome-wide association studies. *Genome Biology*, 16(1):37, Feb 2015.

[80] Martin J Aryee, Andrew E Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P Feinberg, Kasper D Hansen, and Rafael A Irizarry. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics (Oxford, England)*, 30(10):1363–1369, 05 2014.

[81] Regev Schweiger, Reut Yedidim, Elior Rahmani, Liat Shenhav, Omer Weissbrod, Noah Zaitlen, and Eran Halperin. GLINT: a user-friendly toolset for the analysis of high-throughput DNA-methylation array data. *Bioinformatics*, 33(12):1870–1872, 02 2017.

[82] Yi-an Chen, Mathieu Lemire, Sanaa Choufani, Darci T Butcher, Daria Grafodatskaya, Brent W Zanke, Steven Gallinger, Thomas J Hudson, and Rosanna Weksberg. Discovery of cross-reactive probes and polymorphic cpgs in the illumina infinium humanmethylation450 microarray. *Epigenetics*, 8(2):203–209, 02 2013.

[83] Ruth Pidsley, Chloe C. Y Wong, Manuela Volta, Katie Lunnon, Jonathan Mill, and Leonard C. Schalkwyk. A data-driven approach to preprocessing illumina 450k methylation array data. *BMC Genomics*, 14(1):293, May 2013.

[84] Kie Kyon Huang, Kalpana Ramnarayanan, Feng Zhu, Supriya Srivastava, Chang Xu, Angie Lay Keng Tan, Minghui Lee, Suting Tay, Kakoli Das, Manjie Xing, Aliya Fatehullah, Syed Muhammad Fahmy Alkaff, Tony Kiat Hon Lim, Jonathan Lee, Khek Yu Ho, Steven George Rozen, Bin Tean Teh, Nick Barker, Chung King Chia, Christopher

Khor, Choon Jin Ooi, Kwong Ming Fock, Jimmy So, Wee Chian Lim, Khoon Lin Ling, Tiing Leong Ang, Andrew Wong, Jaideepraj Rao, Andrea Rajnakova, Lee Guan Lim, Wai Ming Yap, Ming Teh, Khay Guan Yeoh, and Patrick Tan. Genomic and epigenomic profiling of high-risk intestinal metaplasia reveals molecular determinants of progression to gastric cancer. *Cancer Cell*, 33(1):137–150, Jan 2018.

[85] Hae Dong Woo, Nora Fernandez-Jimenez, Akram Ghantous, Davide Degli Esposti, Cyrille Cuenin, Vincent Cahais, Il Ju Choi, Young-Il Kim, Jeongseon Kim, and Zdenko Herceg. Genome-wide profiling of normal gastric mucosa identifies helicobacter pylori- and cancer-associated dna methylome changes. *Int J Cancer*, 143(3):597–609, Aug 2018.

[86] Matthias Wielscher, Klemens Vierlinger, Ulrike Kegler, Rolf Ziesche, Andrea Gsur, and Andreas Weinhausel. Diagnostic performance of plasma dna methylation profiles in lung cancer, pulmonary fibrosis and copd. *EBioMedicine*, 2(8):929–936, Aug 2015.

[87] Jianxin Shi, Crystal N Marconett, Jubao Duan, Paula L Hyland, Peng Li, Zhaoming Wang, William Wheeler, Beiyun Zhou, Mihaela Campan, Diane S Lee, Jing Huang, Weiyin Zhou, Tim Triche, Laufey Amundadottir, Andrew Warner, Amy Hutchinson, Po-Han Chen, Brian S I Chung, Angela C Pesatori, Dario Consonni, Pier Alberto Bertazzi, Andrew W Bergen, Mathew Freedman, Kimberly D Siegmund, Benjamin P Berman, Zea Borok, Nilanjan Chatterjee, Margaret A Tucker, Neil E Caporaso, Stephen J Chanock, Ite A Laird-Offringa, and Maria Teresa Landi. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. *Nat Commun*, 5:3365, Feb 2014.

[88] Steve Horvath, Vei Mah, Ake T. Lu, Jennifer S. Woo, Oi-Wa Choi, Anna J. Jasinska, José A. Riancho, Spencer Tung, Natalie S. Coles, Jonathan Braun, Harry V. Vinters, and L. Stephen Coles. The cerebellum ages slowly according to the epigenetic clock. *Aging*, 7(5):294–306, 2015.

[89] Steve Horvath, Wiebke Erhart, Mario Brosch, Ole Ammerpohl, Witigo von Schonfels, Markus Ahrens, Nils Heits, Jordana T Bell, Pei-Chien Tsai, Tim D Spector, Panos Deloukas, Reiner Siebert, Bence Sipos, Thomas Becker, Christoph Rocken, Clemens Schafmayer, and Jochen Hampe. Obesity accelerates epigenetic aging of human liver. *Proc Natl Acad Sci U S A*, 111(43):15538–15543, Oct 2014.

[90] Andrew E Teschendorff, Francesco Marabita, Matthias Lechner, Thomas Bartlett, Jesper Tegner, David Gomez-Cabrero, and Stephan Beck. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k dna methylation data. *Bioinformatics*, 29(2):189–196, Jan 2013.

[91] Marc Jan Bonder, Silva Kasela, Mart Kals, Riin Tamm, Kaie Lokk, Isabel Barragan, Wim A. Buurman, Patrick Deelen, Jan-Willem Greve, Maxim Ivanov, Sander S. Rensen, Jana V. van Vliet-Ostaptchouk, Marcel G. Wolfs, Jingyuan Fu, Marten H. Hofker, Cisca Wijmenga, Alexandra Zhernakova, Magnus Ingelman-Sundberg, Lude

Franke, and Lili Milani. Genetic and epigenetic regulation of gene expression in fetal and adult human livers. *BMC Genomics*, 15(1):860, Oct 2014.

[92] Jin-Huan Wei, Ahmed Haddad, Kai-Jie Wu, Hong-Wei Zhao, Payal Kapur, Zhi-Ling Zhang, Liang-Yun Zhao, Zhen-Hua Chen, Yun-Yun Zhou, Jian-Cheng Zhou, Bin Wang, Yan-Hong Yu, Mu-Yan Cai, Dan Xie, Bing Liao, Cai-Xia Li, Pei-Xing Li, Zong-Ren Wang, Fang-Jian Zhou, Lei Shi, Qing-Zuo Liu, Zhen-Li Gao, Da-Lin He, Wei Chen, Jer-Tsong Hsieh, Quan-Zhen Li, Vitaly Margulis, and Jun-Hang Luo. A cpg-methylation-based assay to predict survival in clear cell renal cell carcinoma. *Nat Commun*, 6:8699, Oct 2015.

[93] Yi-An Ko, Davoud Mohtat, Masako Suzuki, Ae Seo Deok Park, Maria Concepcion Izquierdo, Sang Youb Han, Hyun Mi Kang, Han Si, Thomas Hostetter, James M Pullman, Melissa Fazzari, Amit Verma, Deyou Zheng, John M Greally, and Katalin Susztak. Cytosine methylation changes in enhancer regions of core pro-fibrotic genes characterize kidney fibrosis development. *Genome Biol*, 14(10):R108, 2013.

[94] Andrew E Teschendorff, Yang Gao, Allison Jones, Matthias Ruebner, Matthias W Beckmann, David L Wachter, Peter A Fasching, and Martin Widschwendter. Dna methylation outliers in normal breast tissue identify field defects that are enriched in cancer. *Nat Commun*, 7:10478, Jan 2016.

[95] Min-Ae Song, Theodore M Brasky, Daniel Y Weng, Joseph P McElroy, Catalin Marian, Michael J Higgins, Christine Ambrosone, Scott L Spear, Adana A Llanos, Bhaskar V S Kallakury, Jo L Freudenheim, and Peter G Shields. Landscape of genome-wide age-related dna methylation in breast tissue. *Oncotarget*, 8(70):114648–114662, Dec 2017.

[96] E Andres Houseman, Molly L Kile, David C Christiani, Tan A Ince, Karl T Kelsey, and Carmen J Marsit. Reference-free deconvolution of dna methylation data and mediation by cell composition effects. *BMC bioinformatics*, 17:259; 259–259, 06 2016.

[97] Joost Smolders, Kirstin M. Heutinck, Nina L. Fransen, Ester B. M. Remmerswaal, Pleun Hombrink, Ineke J. M. ten Berge, RenéA. W. van Lier, Inge Huitinga, and Jörg Hamann. Tissue-resident memory t cells populate the human brain. *Nature Communications*, 9(1):4593, 2018.

[98] Joost Smolders, Ester B. M. Remmerswaal, Karianne G. Schuurman, Jeroen Melief, Corbert G. van Eden, René A. W. van Lier, Inge Huitinga, and Jörg Hamann. Characteristics of differentiated cd8+ and cd4+ t cells present in the human brain. *Acta Neuropathologica*, 126(4):525–535, Oct 2013.

[99] Sae-Bom Jeon, Hee Jung Yoon, Se-Ho Park, In-Hoo Kim, and Eun Jung Park. Sulfatide, a major lipid component of myelin sheath, activates inflammatory responses as an endogenous stimulator in brain-resident immune cells. *The Journal of Immunology*, 181(11):8077–8087, 2008.

[100] Jan-Kolja Strecker, Antje Schmidt, Wolf-Rüdiger Schäbitz, and Jens Minnerup. Neutrophil granulocytes in cerebral ischemia – evolution from killers to key players. *Neurochemistry International*, 107:117 – 126, 2017.

[101] Richard T Barfield, Lynn M Almli, Varun Kilaru, Alicia K Smith, Kristina B Mercer, Richard Duncan, Torsten Klengel, Divya Mehta, Elisabeth B Binder, Michael P Epstein, Kerry J Ressler, and Karen N Conneely. Accounting for population stratification in dna methylation studies. *Genetic epidemiology*, 38(3):231–241, 04 2014.

[102] Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K. Johnson, Theresa K. Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R. Scott Hansen, Shane Neph, Peter J. Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R. Sunyaev, Rajinder Kaul, and John A. Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195, 2012.

[103] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W J H Penninx, Rick Jansen, Eco J C de Geus, Dorret I Boomsma, Fred A Wright, Patrick F Sullivan, Elina Nikkola, Marcus Alvarez, Mete Civelek, Aldons J Lusis, Terho Lehtimäki, Emma Raitoharju, Mika Kähönen, Ilkka Seppälä, Olli T Raitakari, Johanna Kuusisto, Markku Laakso, Alkes L Price, Päivi Pajukanta, and Bogdan Pasaniuc. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48:245 EP –, 02 2016.

[104] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, GTEx Consortium, Dan L Nicolae, Nancy J Cox, and Hae Kyung Im. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47:1091 EP –, 08 2015.

[105] Michael Wainberg, Nasa Sinnott-Armstrong, Nicholas Mancuso, Alvaro N. Barbeira, David A. Knowles, David Golan, Raili Ermel, Arno Ruusalepp, Thomas Quertermous, Ke Hao, Johan L. M. Björkegren, Hae Kyung Im, Bogdan Pasaniuc, Manuel A. Rivas, and Anshul Kundaje. Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics*, 51(4):592–599, 2019.

[106] Timothée Flutre, Xiaoquan Wen, Jonathan Pritchard, and Matthew Stephens. A statistical framework for joint eqtl analysis in multiple tissues. *PLOS Genetics*, 9(5):1–13, 05 2013.

[107] Nicholas Mancuso, Huwenbo Shi, Pagé Goddard, Gleb Kichaev, Alexander Gusev, and Bogdan Pasaniuc. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *The American Journal of Human Genetics*, 100(3):473 – 487, 2017.

[108] Dan Zhou, Yi Jiang, Xue Zhong, Nancy J. Cox, Chunyu Liu, and Eric R. Gamazon. A unified framework for joint-tissue transcriptome-wide association and mendelian randomization analysis. *Nature Genetics*, 52(11):1239–1246, 2020.

[109] Yiming Hu, Mo Li, Qiongshi Lu, Haoyi Weng, Jiawei Wang, Seyedeh M. Zekavat, Zhaolong Yu, Boyang Li, Jianlei Gu, Sydney Muchnik, Yu Shi, Brian W. Kunkle, Shubhabrata Mukherjee, Pradeep Natarajan, Adam Naj, Amanda Kuzma, Yi Zhao, Paul K. Crane, Hui Lu, Hongyu Zhao, and Alzheimer's Disease Genetics Consortium. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature Genetics*, 51(3):568–576, 2019.

[110] Alvaro N. Barbeira, Milton Pividori, Jiamao Zheng, Heather E. Wheeler, Dan L. Nicolae, and Hae Kyung Im. Integrating predicted transcriptome from multiple tissues improves association detection. *PLOS Genetics*, 15(1):1–20, 01 2019.

[111] Helian Feng, Nicholas Mancuso, Alexander Gusev, Arunabha Majumdar, Megan Major, Bogdan Pasaniuc, and Peter Kraft. Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improve the power of transcriptome-wide association studies. *bioRxiv*, 2020.

[112] S H Lee, J Yang, M E Goddard, P M Visscher, and N R Wray. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, 28(19):2540–2542, Oct 2012.

[113] Jian Yang, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. Gcta: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2019/11/04 2011.

[114] Heather E. Wheeler, Kaanan P. Shah, Jonathon Brenner, Tzintzuni Garcia, Keston Aquino-Michaels, GTEx Consortium, Nancy J. Cox, Dan L. Nicolae, and Hae Kyung Im. Survey of the heritability and sparse architecture of gene expression traits across human tissues. *PLOS Genetics*, 12(11):1–23, 11 2016.

[115] Alvaro N. Barbeira, Scott P. Dickinson, Rodrigo Bonazzola, Jiamao Zheng, Heather E. Wheeler, Jason M. Torres, Eric S. Torstenson, Kaanan P. Shah, Tzintzuni Garcia, Todd L. Edwards, Eli A. Stahl, Laura M. Huckins, François Aguet, Kristin G. Ardlie, Beryl B. Cummings, Ellen T. Gelfand, Gad Getz, Kane Hadley, Robert E. Handsaker, Katherine H. Huang, Seva Kashin, Konrad J. Karczewski, Monkol Lek, Xiao Li, Daniel G. MacArthur, Jared L. Nedzel, Duyen T. Nguyen, Michael S. Noble, Ayellet V. Segrè, Casandra A. Trowbridge, Taru Tukiainen, Nathan S. Abell, Brunilda Balliu, Ruth Barshir, Omer Basha, Alexis Battle, Gireesh K. Bogu, Andrew Brown, Christopher D. Brown, Stephane E. Castel, Lin S. Chen, Colby Chiang, Donald F. Conrad, Farhan N. Damani, Joe R. Davis, Olivier Delaneau, Emmanouil T. Dermitzakis, Barbara E. Engelhardt, Eleazar Eskin, Pedro G. Ferreira, Laure Frésard, Eric R. Gamazon, Diego Garrido-Martín, Ariel D. H. Gewirtz, Genna Gliner, Michael J. Gloudemans,

Roderic Guigo, Ira M. Hall, Buhm Han, Yuan He, Farhad Hormozdiari, Cedric Howald, Brian Jo, Eun Yong Kang, Yungil Kim, Sarah Kim-Hellmuth, Tuuli Lappalainen, Gen Li, Xin Li, Boxiang Liu, Serghei Mangul, Mark I. McCarthy, Ian C. McDowell, Pejman Mohammadi, Jean Monlong, Stephen B. Montgomery, Manuel Muñoz-Aguirre, Anne W. Ndungu, Andrew B. Nobel, Meritxell Oliva, Halit Ongen, John J. Palowitch, Nikolaos Panousis, Panagiotis Papasaikas, YoSon Park, Princy Parsana, Anthony J. Payne, Christine B. Peterson, Jie Quan, Ferran Reverter, Chiara Sabatti, Ashis Saha, Michael Sammeth, Alexandra J. Scott, Andrey A. Shabalin, Reza Sodaei, Matthew Stephens, Barbara E. Stranger, Benjamin J. Strober, Jae Hoon Sul, Emily K. Tsang, Sarah Urbut, Martijn van de Bunt, Gao Wang, Xiaoquan Wen, Fred A. Wright, Hualin S. Xi, Esti Yeger-Lotem, Zachary Zappala, Judith B. Zaugg, Yi-Hui Zhou, Joshua M. Akey, Daniel Bates, Joanne Chan, Lin S. Chen, Melina Claussnitzer, Kathryn Demanelis, Morgan Diegel, Jennifer A. Doherty, Andrew P. Feinberg, Marian S. Fernando, Jessica Halow, Kasper D. Hansen, Eric Haugen, Peter F. Hickey, Lei Hou, Farzana Jasmine, Ruiqi Jian, Lihua Jiang, Audra Johnson, Rajinder Kaul, Manolis Kellis, Muhammad G. Kibriya, Kristen Lee, Jin Billy Li, Qin Li, Jessica Lin, Shin Lin, Sandra Linder, Caroline Linke, Yaping Liu, Matthew T. Maurano, Benoit Molinie, Stephen B. Montgomery, Jemma Nelson, Fidencio J. Neri, Yongjin Park, Brandon L. Pierce, Nicola J. Rinaldi, Lindsay F. Rizzardi, Richard Sandstrom, Andrew Skol, Kevin S. Smith, Michael P. Snyder, John Stamatoyannopoulos, Barbara E. Stranger, Hua Tang, Emily K. Tsang, Li Wang, Meng Wang, Nicholas Van Wittenberghe, Fan Wu, Rui Zhang, Concepcion R. Nierras, Philip A. Branton, Latarsha J. Carithers, Ping Guan, Helen M. Moore, Abhi Rao, Jimmie B. Vaught, Sarah E. Gould, Nicole C. Lockart, Casey Martin, Jeffery P. Struewing, Simona Volpi, Anjene M. Addington, Susan E. Koester, A. Roger Little, Lori E. Brigham, Richard Hasz, Marcus Hunter, Christopher Johns, Mark Johnson, Gene Kopen, William F. Leinweber, John T. Lonsdale, Alisa McDonald, Bernadette Mestichelli, Kevin Myer, Brian Roe, Michael Salvatore, Saboor Shad, Jeffrey A. Thomas, Gary Walters, Michael Washington, Joseph Wheeler, Jason Bridge, Barbara A. Foster, Bryan M. Gillard, Ellen Karasik, Rachna Kumar, Mark Miklos, Michael T. Moser, Scott D. Jewell, Robert G. Montroy, Daniel C. Rohrer, Dana R. Valley, David A. Davis, Deborah C. Mash, Anita H. Undale, Anna M. Smith, David E. Tabor, Nancy V. Roche, Jeffrey A. McLean, Negin Vatanian, Karna L. Robinson, Leslie Sobin, Mary E. Barcus, Kimberly M. Valentino, Liqun Qi, Steven Hunter, Pushpa Hariharan, Shilpi Singh, Ki Sung Um, Takunda Matose, Maria M. Tomaszewski, Laura K. Barker, Maghboeba Mosavel, Laura A. Siminoff, Heather M. Traino, Paul Flicek, Thomas Juettemann, Magali Ruffier, Dan Sheppard, Kieron Taylor, Stephen J. Trevanion, Daniel R. Zerbino, Brian Craft, Mary Goldman, Maximilian Haeussler, W. James Kent, Christopher M. Lee, Benedict Paten, Kate R. Rosenbloom, John Vivian, Jingchun Zhu, Dan L. Nicolae, Nancy J. Cox, Hae Kyung Im, GTEx Consortium, Data Analysis & Coordinating Center (LDACC)-Analysis Working Group Laboratory, Statistical Methods groups Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGrI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site-NDrI, Biospecimen Collection Source Site-rPCI, Biospecimen Core resource VArI, Brain Bank repository-University of Mi-

ami Brain Endowment Bank, Leidos Biomedical-Project Management, ELSI Study, Genome Browser Data Integration & Visualization-EBI, and University of California Santa Cruz Genome Browser Data Integration & Visualization-UCSC Genomics Institute. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from gwas summary statistics. *Nature Communications*, 9(1):1825, 2018.

[116] C B Peterson, M Bogomolov, Y Benjamini, and C Sabatti. Treeqtl: hierarchical error control for eqtl findings. *Bioinformatics*, 32(16):2556–2558, Aug 2016.

[117] Christine B Peterson, Marina Bogomolov, Yoav Benjamini, and Chiara Sabatti. Many phenotypes without many false discoveries: Error controlling strategies for multitrait association studies. *Genet Epidemiol*, 40(1):45–56, Jan 2016.

[118] Cristina M. Lanata, Ishan Paranjpe, Joanne Nititham, Kimberly E. Taylor, Milena Gianfrancesco, Manish Paranjpe, Shan Andrews, Sharon A. Chung, Brooke Rhead, Lisa F. Barcellos, Laura Trupin, Patricia Katz, Maria Dall'Era, Jinoos Yazdany, Marina Sirota, and Lindsey A. Criswell. A phenotypic and genomics approach in a multi-ethnic cohort to subtype systemic lupus erythematosus. *Nature Communications*, 10(1):3902, 2019.

[119] Gaia Andreoletti, Cristina M. Lanata, Laura Trupin, Ishan Paranjpe, Tia S. Jain, Joanne Nititham, Kimberly E. Taylor, Alexis J. Combes, Lenka Maliskova, Chun Jimmie Ye, Patricia Katz, Maria Dall'Era, Jinoos Yazdany, Lindsey A. Criswell, and Marina Sirota. Transcriptomic analysis of immune cells in a multi-ethnic cohort of systemic lupus erythematosus patients identifies ethnicity- and disease-specific expression signatures. *Communications Biology*, 4(1):488, 2021.

[120] Monique G. P. van der Wijst, Harm Brugge, Dylan H. de Vries, Patrick Deelen, Morris A. Swertz, Lude Franke, LifeLines Cohort Study, and BIOS Consortium. Single-cell rna sequencing identifies celltype-specific cis-eqtls and co-expression qtls. *Nature Genetics*, 50(4):493–497, 2018.

[121] Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, Rachel E Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A Criswell, and Chun Jimmie Ye. Multiplexed droplet single-cell rna-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1):89–94, 2018.

[122] Maayan Baron, Adrian Veres, Samuel L. Wolock, Aubrey L. Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K. Wagner, Shai S. Shen-Orr, Allon M. Klein, Douglas A. Melton, and Itai Yanai. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Systems*, 3(4):346–360.e4, 2016.

[123] Brandon Jew, Marcus Alvarez, Elior Rahmani, Zong Miao, Arthur Ko, Kristina M. Garske, Jae Hoon Sul, Kirsi H. Pietiläinen, Päivi Pajukanta, and Eran Halperin. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature Communications*, 11(1):1971, 2020.

[124] Florian Prive, Hugues Aschard, Andrey Ziyatdinov, and Michael G B Blum. Efficient analysis of large-scale genome-wide data with two r packages: bigstatsr and bigsnpr. *Bioinformatics*, 34(16):2781–2787, Aug 2018.

[125] Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Eric D. Green, Matthew E. Hurles, Bartha M. Knoppers, Jan O. Korbel, Eric S. Lander, Charles Lee, Hans Lehrach, Elaine R. Mardis, Gabor T. Marth, Gil A. McVean, Deborah A. Nickerson, Jeanette P. Schmidt, Stephen T. Sherry, Jun Wang, Richard K. Wilson, Richard A. Gibbs, Eric Boerwinkle, Harsha Doddapaneni, Yi Han, Viktoriya Korchina, Christie Kovar, Sandra Lee, Donna Muzny, Jeffrey G. Reid, Yiming Zhu, Yuqi Chang, Qiang Feng, Xiaodong Fang, Xiaosen Guo, Min Jian, Hui Jiang, Xin Jin, Tianming Lan, Guoqing Li, Jingxiang Li, Yingrui Li, Shengmao Liu, Xiao Liu, Yao Lu, Xuedi Ma, Meifang Tang, Bo Wang, Guangbiao Wang, Honglong Wu, Renhua Wu, Xun Xu, Ye Yin, Dandan Zhang, Wenwei Zhang, Jiao Zhao, Meiru Zhao, Xiaole Zheng, Eric S. Lander, David M. Altshuler, Stacey B. Gabriel, Namrata Gupta, Neda Gharani, Lorraine H. Toji, Norman P. Gerry, Alissa M. Resch, Jonathan Barker, Laura Clarke, Laurent Gil, Sarah E. Hunt, Gavin Kelman, Eugene Kulesha, Rasko Leinonen, William M. McLaren, Rajesh Radhakrishnan, Asier Roa, Dmitriy Smirnov, Richard E. Smith, Ian Streeter, Anja Thormann, Iliana Toneva, Brendan Vaughan, Xiangqun Zheng-Bradley, David R. Bentley, Russell Grocock, Sean Humphray, Terena James, Zoya Kingsbury, Ralf Sudbrak, Marcus W. Albrecht, Vyacheslav S. Amstislavskiy, Tatiana A. Borodina, Matthias Lienhard, Florian Mertes, Marc Sultan, Bernd Timmermann, Marie-Laure Yaspo, Elaine R. Mardis, Richard K. Wilson, Lucinda Fulton, Robert Fulton, Stephen T. Sherry, Victor Ananiev, Zinaida Belaia, Dimitriy Beloslyudtsev, Nathan Bouk, Chao Chen, Deanna Church, Robert Cohen, Charles Cook, John Garner, Timothy Hefferon, Mikhail Kimelman, Chunlei Liu, John Lopez, Peter Meric, Chris O'Sullivan, Yuri Ostapchuk, Lon Phan, Sergiy Ponomarov, Valerie Schneider, Eugene Shekhtman, Karl Sirotkin, Douglas Slotta, Hua Zhang, Gil A. McVean, Richard M. Durbin, Senduran Balasubramaniam, John Burton, Petr Danecek, Thomas M. Keane, Anja Kolb-Kokocinski, Shane McCarthy, James Stalker, Michael Quail, Jeanette P. Schmidt, Christopher J. Davies, Jeremy Gollub, Teresa Webster, Brant Wong, Yiping Zhan, Christopher L. Campbell, Yu Kong, Anthony Marcketta, Richard A. Gibbs, Fuli Yu, Lilian Antunes, Matthew Bainbridge, Aniko Sabo, Zhuoyi Huang, Lachlan J. M. Coin, Lin Fang, Qibin Li, Zhenyu Li, Haoxiang Lin, Binghang Liu, Ruibang Luo, Haojing Shao, Yinlong Xie, Chen Ye, Chang Yu, Fan Zhang, Hancheng Zheng, Hongmei Zhu, Can Alkan, Elif Dal, Fatma Kahveci, Gabor T. Marth, Erik P. Garrison, Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Michael Stromberg, Alistair N. Ward, Jiantao Wu, Mengyao Zhang, Mark J. Daly, Mark A. DePristo, Robert E. Handsaker, David M. Altshuler, Eric Banks, Gaurav Bhatia, Guillermo del Angel, Stacey B. Gabriel, Giulio Genovese, Heng Li, Seva Kashin, Eric S. Lander, Steven A. McCarroll, James C. Nemesh, Ryan E. Poplin, Seungtai C. Yoon, Jayon Lihm, Vladimir Makarov, Andrew G. Clark, Srikanth Gottipati, Alon Keinan,

Juan L. Rodriguez-Flores, Jan O. Korbel, Tobias Rausch, Markus H. Fritz, Adrian M. Stütz, Kathryn Beal, Avik Datta, Javier Herrero, William M. McLaren, Graham R. S. Ritchie, Richard E. Smith, Daniel Zerbino, Pardis C. Sabeti, Ilya Shlyakhter, Stephen F. Schaffner, Joseph Vitti, David N. Cooper, Edward V. Ball, Peter D. Stenson, David R. Bentley, Bret Barnes, Markus Bauer, R. Keira Cheetham, Anthony Cox, Michael Eberle, Scott Kahn, Lisa Murray, John Peden, Richard Shaw, Eimear E. Kenny, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Daniel G. MacArthur, Monkol Lek, Vyacheslav S. Amstislavskiy, Ralf Herwig, Elaine R. Mardis, Li Ding, Daniel C. Koboldt, David Larson, Kai Ye, Simon Gravel, The 1000 Genomes Project Consortium, Corresponding authors, Steering committee, Production group, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT, Harvard, Coriell Institute for Medical Research, European Bioinformatics Institute European Molecular Biology Laboratory, Illumina, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, US National Institutes of Health, University of Oxford, Wellcome Trust Sanger Institute, Analysis group, Affymetrix, Albert Einstein College of Medicine, Bilkent University, Boston College, Cold Spring Harbor Laboratory, Cornell University, European Molecular Biology Laboratory, Harvard University, Human Gene Mutation Database, Icahn School of Medicine at Mount Sinai, Louisiana State University, Massachusetts General Hospital, McGill University, and NIH National Eye Institute. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[126] Sarah M. Urbut, Gao Wang, Peter Carbonetto, and Matthew Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics*, 51(1):187–195, 2019.

[127] Rachel Moore, Francesco Paolo Casale, Marc Jan Bonder, Danilo Horta, Bastiaan T. Heijmans, Peter A. C. 't Hoen, Joyce van Meurs, Aaron Isaacs, Rick Jansen, Lude Franke, Dorret I. Boomsma, René Pool, Jenny van Dongen, Jouke J. Hottenga, Marleen M. J. van Greevenbroek, Coen D. A. Stehouwer, Carla J. H. van der Kallen, Casper G. Schalkwijk, Cisca Wijmenga, Alexandra Zhernakova, Ettje F. Tigchelaar, P. Eline Slagboom, Marian Beekman, Joris Deelen, Diana van Heemst, Jan H. Veldink, Leonard H. van den Berg, Cornelia M. van Duijn, Bert A. Hofman, AndréG. Uitterlinden, P. Mila Jhamai, Michael Verbiest, H. Eka D. Suchiman, Marijn Verkerk, Ruud van der Breggen, Jeroen van Rooij, Nico Lakenberg, Hailiang Mei, Maarten van Iterson, Michiel van Galen, Jan Bot, Peter van't Hof, Patrick Deelen, Irene Nooren, Matthijs Moed, Martijn Vermaat, Dasha V. Zhernakova, René Luijk, Freerk van Dijk, Wibowo Arindrarto, Szymon M. Kielbasa, Morris A. Swertz, Erik W. van Zwet, Inês Barroso, Oliver Stegle, and BIOS Consortium. A linear mixed-model approach to study multivariate gene–environment interactions. *Nature Genetics*, 51(1):180–186, 2019.

[128] Sarah Kim-Hellmuth, François Aguet, Meritxell Oliva, Manuel Muñoz-Aguirre, Silva Kasela, Valentin Wucher, Stephane E. Castel, Andrew R. Hamel, Ana Viñuela, Amy L. Roberts, Serghei Mangul, Xiaoquan Wen, Gao Wang, Alvaro N. Barbeira, Diego Garrido-Martín, Brian B. Nadel, Yuxin Zou, Rodrigo Bonazzola, Jie Quan, Andrew

Brown, Angel Martinez-Perez, José Manuel Soria, GTEx Consortium§, Gad Getz, Emmanouil T. Dermitzakis, Kerrin S. Small, Matthew Stephens, Hualin S. Xi, Hae Kyung Im, Roderic Guigó, Ayellet V. Segrè, Barbara E. Stranger, Kristin G. Ardlie, and Tuuli Lappalainen. Cell type–specific genetic regulation of gene expression across human tissues. 369(6509).

[129] Jean-Charles Lambert, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, Gary W Beecham, Benjamin Grenier-Boley, Giancarlo Russo, Tricia A Thornton-Wells, Nicola Jones, Albert V Smith, Vincent Chouraki, Charlene Thomas, M Arfan Ikram, Diana Zelenika, Badri N Vardarajan, Yoichiro Kamatani, Chiao-Feng Lin, Amy Gerrish, Helena Schmidt, Brian Kunkle, Melanie L Dunstan, Agustin Ruiz, Marie-Thérèse Bihoreau, Seung-Hoan Choi, Christiane Reitz, Florence Pasquier, Paul Hollingworth, Alfredo Ramirez, Olivier Hanon, Annette L Fitzpatrick, Joseph D Buxbaum, Dominique Campion, Paul K Crane, Clinton Baldwin, Tim Becker, Vilmundur Gudnason, Carlos Cruchaga, David Craig, Najaf Amin, Claudine Berr, Oscar L Lopez, Philip L De Jager, Vincent Deramecourt, Janet A Johnston, Denis Evans, Simon Lovestone, Luc Letenneur, Francisco J Morón, David C Rubinsztein, Gudny Eiriksdottir, Kristel Sleegers, Alison M Goate, Nathalie Fiévet, Matthew J Huentelman, Michael Gill, Kristelle Brown, M Ilyas Kamboh, Lina Keller, Pascale Barberger-Gateau, Bernadette McGuinness, Eric B Larson, Robert Green, Amanda J Myers, Carole Dufouil, Stephen Todd, David Wallon, Seth Love, Ekaterina Rogaeva, John Gallacher, Peter St George-Hyslop, Jordi Clarimon, Alberto Lleo, Anthony Bayer, Debby W Tsuang, Lei Yu, Magda Tsolaki, Paola Bossù, Gianfranco Spalletta, Petroula Proitsi, John Collinge, Sandro Sorbi, Florentino Sanchez-Garcia, Nick C Fox, John Hardy, Maria Candida Deniz Naranjo, Paolo Bosco, Robert Clarke, Carol Brayne, Daniela Galimberti, Michelangelo Mancuso, Fiona Matthews, Susanne Moebus, Patrizia Mecocci, Maria Del Zompo, Wolfgang Maier, Harald Hampel, Alberto Pilotto, Maria Bullido, Francesco Panza, Paolo Caffarra, Benedetta Nacmias, John R Gilbert, Manuel Mayhaus, Lars Lannfelt, Hakon Hakonarson, Sabrina Pichler, Minerva M Carrasquillo, Martin Ingelsson, Duane Beekly, Victoria Alvarez, Fanggeng Zou, Otto Valladares, Steven G Younkin, Eliecer Coto, Kara L Hamilton-Nelson, Wei Gu, Cristina Razquin, Pau Pastor, Ignacio Mateo, Michael J Owen, Kelley M Faber, Palmi V Jonsson, Onofre Combarros, Michael C O'Donovan, Laura B Cantwell, Hilkka Soininen, Deborah Blacker, Simon Mead, Thomas H Mosley, David A Bennett, Tamara B Harris, Laura Fratiglioni, Clive Holmes, Renee F A G de Bruijn, Peter Passmore, Thomas J Montine, Karolien Bettens, Jerome I Rotter, Alexis Brice, Kevin Morgan, Tatiana M Foroud, Walter A Kukull, Didier Hannequin, John F Powell, Michael A Nalls, Karen Ritchie, Kathryn L Lunetta, John S K Kauwe, Eric Boerwinkle, Matthias Riemenschneider, Mercè Boada, Mikko Hiltunen, Eden R Martin, Reinhold Schmidt, Dan Rujescu, Li-San Wang, Jean-François Dartigues, Richard Mayeux, Christophe Tzourio, Albert Hofman, Markus M Nöthen, Caroline Graff, Bruce M Psaty, Lesley Jones, Jonathan L Haines, Peter A Holmans, Mark Lathrop, Margaret A Pericak-Vance, Lenore J Launer, Lindsay A Farrer, Cornelia M van Duijn, Christine Van Broeck-

hoven, Valentina Moskvina, Sudha Seshadri, Julie Williams, Gerard D Schellenberg, Philippe Amouyel, European Alzheimer's Disease Initiative (EADI), Genetic, Environmental Risk in Alzheimer's Disease (GERAD), Alzheimer's Disease Genetic Consortium (ADGC), Cohorts for Heart, and Aging Research in Genomic Epidemiology (CHARGE). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nature Genetics*, 45(12):1452–1458, 2013.

[130] Douglas M. Ruderfer, Stephan Ripke, Andrew McQuillin, James Boocock, Eli A. Stahl, Jennifer M. Whitehead Pavlides, Niamh Mullins, Alexander W. Charney, Anil P. S. Ori, Loes M. Olde Loohuis, Enrico Domenici, Arianna Di Florio, Sergi Papiol, Janos L. Kalman, Vassily Trubetskoy, Rolf Adolfsson, Ingrid Agartz, Esben Agerbo, Huda Akil, Diego Albani, Margot Albus, Martin Alda, Madeline Alexander, Ney Alliey-Rodriguez, Thomas D. Als, Farooq Amin, Adebayo Anjorin, Maria J. Arranz, Swapnil Awasthi, Silviu A. Bacanu, Judith A. Badner, Marie Baekvad-Hansen, Steven Bakker, Gavin Band, Jack D. Barchas, Ines Barroso, Nicholas Bass, Michael Bauer, Bernhard T. Baune, Martin Begemann, Celine Bellenguez, Jr. Belliveau, Richard A., Frank Bellivier, Stephan Bender, Judit Bene, Sarah E. Bergen, Wade H. Berrettini, Elizabeth Bevilacqua, Joanna M. Biernacka, Tim B. Bigdeli, Donald W. Black, Hannah Blackburn, Jenefer M. Blackwell, Douglas H. R. Blackwood, Carsten Bocker Pedersen, Michael Boehnke, Marco Boks, Anders D. Borglum, Elvira Bramon, Gerome Breen, Matthew A. Brown, Richard Bruggeman, Nancy G. Buccola, Randy L. Buckner, Monika Budde, Brendan Bulik-Sullivan, Suzannah J. Bumpstead, William Bunney, Margit Burmeister, Joseph D. Buxbaum, Jonas Bybjerg-Grauholm, William Byerley, Wiepke Cahn, Guiqing Cai, Murray J. Cairns, Dominique Campion, Rita M. Cantor, Vaughan J. Carr, Noa Carrera, Juan P. Casas, Miquel Casas, Stanley V. Catts, Pablo Cervantes, Kimberley D. Chambert, Raymond C. K. Chan, Eric Y. H. Chen, Ronald Y. L. Chen, Wei Cheng, Eric F. C. Cheung, Siow Ann Chong, Toni-Kim Clarke, C. Robert Cloninger, David Cohen, Nadine Cohen, Jonathan R. I. Coleman, David A. Collier, Paul Cormican, William Coryell, Nicholas Craddock, David W. Craig, Benedicto Crespo-Facorro, James J. Crowley, Cristiana Cruceanu, David Curtis, Piotr M. Czerski, Anders M. Dale, Mark J. Daly, Udo Dannlowski, Ariel Darvasi, Michael Davidson, Kenneth L. Davis, Christiaan A. de Leeuw, Franziska Degenhardt, Jurgen Del Favero, Lynn E. DeLisi, Panos Deloukas, Ditte Demontis, J. Raymond DePaulo, Marta di Forti, Dimitris Dikeos, Timothy Dinan, Srdjan Djurovic, Amanda L. Dobbyn, Peter Donnelly, Gary Donohoe, Elodie Drapeau, Serge Dronov, Jubao Duan, Frank Dudbridge, Audrey Duncanson, Howard Edenberg, Sarah Edkins, Hannelore Ehrenreich, Peter Eichhammer, Torbjorn Elvsashagen, Johan Eriksson, Valentina Escott-Price, Tonu Esko, Laurent Essioux, Bruno Etain, Chun Chieh Fan, Kai-How Farh, Martilias S. Farrell, Matthew Flickinger, Tatiana M. Foroud, Liz Forty, Josef Frank, Lude Franke, Christine Fraser, Robert Freedman, Colin Freeman, Nelson B. Freimer, Joseph I. Friedman, Menachem Fromer, Mark A. Frye, Janice M. Fullerton, Katrin Gade, Julie Garnham, Helena A. Gaspar, Pablo V. Gejman, Giulio Genovese, Lyudmila Georgieva, Claudia Giambartolomei, Eleni Giannoulatou, Ina Giegling, Michael Gill, Matthew Gillman, Marianne Giortz Pedersen, Paola Giusti-Rodriguez, Stephanie Go-

dard, Fernando Goes, Jacqueline I. Goldstein, Srihari Gopal, Scott D. Gordon, Katherine Gordon-Smith, Jacob Gratten, Emma Gray, Elaine K. Green, Melissa J. Green, Tiffany A. Greenwood, Maria Grigoroiu-Serbanescu, Jakob Grove, Weihua Guan, Hugh Gurling, Jose Guzman Parra, Rhian Gwilliam, Lieuwe de Haan, Jeremy Hall, Mei-Hua Hall, Christian Hammer, Naomi Hammond, Marian L. Hamshere, Mark Hansen, Thomas Hansen, Vahram Haroutunian, Annette M. Hartmann, Joanna Hauser, Martin Hautzinger, Urs Heilbronner, Garrett Hellenthal, Frans A. Henskens, Stefan Herms, Maria Hipolito, Joel N. Hirschhorn, Per Hoffmann, Mads V. Hollegaard, David M. Hougaard, Hailiang Huang, Laura Huckins, Christina M. Hultman, Sarah E. Hunt, Masashi Ikeda, Nakao Iwata, Conrad Iyegbe, Assen V. Jablensky, Stephane Jamain, Janusz Jankowski, Alagurevathi Jayakumar, Inge Joa, Ian Jones, Lisa A. Jones, Erik G. Jonsson, Antonio Julia, Anders Jureus, Anna K. Kahler, Rene S. Kahn, Luba Kalaydjieva, Radhika Kandaswamy, Sena Karachanak-Yankova, Juha Karjalainen, Robert Karlsson, David Kavanagh, Matthew C. Keller, Brian J. Kelly, John Kelsoe, James L. Kennedy, Andrey Khrunin, Yunjung Kim, George Kirov, Sarah Kittel-Schneider, Janis Klovins, Jo Knight, Sarah V. Knott, James A. Knowles, Manolis Kogevinas, Bettina Konte, Eugenia Kravariti, Vaidutis Kucinskas, Zita Ausrele Kucinskiene, Ralph Kupka, Hana Kuzelova-Ptackova, Mikael Landen, Cordelia Langford, Claudine Laurent, Jacob Lawrence, Stephen Lawrie, William B. Lawson, Markus Leber, Marion Leboyer, Phil H. Lee, Jimmy Lee Chee Keong, Sophie E. Legge, Todd Lencz, Bernard Lerer, Douglas F. Levinson, Shawn E. Levy, Cathryn M. Lewis, Jun Z. Li, Miaoxin Li, Qingqin S. Li, Tao Li, Kung-Yee Liang, Jennifer Liddle, Jeffrey Lieberman, Svetlana Limborska, Kuang Lin, Don H. Linszen, Jolanta Lissowska, Chunyu Liu, Jianjun Liu, Jouko Lonnqvist, Carmel M. Loughland, Jan Lubinski, Susanne Lucae, Jr. Macek, Milan, Donald J. MacIntyre, Patrik K. E. Magnusson, Brion S. Maher, Pamela B. Mahon, Wolfgang Maier, Anil K. Malhotra, Jacques Mallet, Ulrik F. Malt, Hugh S. Markus, Sara Marsal, Nicholas G. Martin, Ignacio Mata, Christopher G. Mathew, Manuel Mattheisen, Morten Mattingsdal, Fermin Mayoral, Owen T. McCann, Robert W. McCarley, Steven A. McCarroll, Mark I. McCarthy, Colm McDonald, Susan L. McElroy, Peter McGuffin, Melvin G. McInnis, Andrew M. McIntosh, James D. McKay, Francis J. McMahon, Helena Medeiros, Sarah E. Medland, Sandra Meier, Carin J. Meijer, Bela Melegh, Ingrid Melle, Fan Meng, Raquelle I. Mesholam-Gately, Andres Metspalu, Patricia T. Michie, Lili Milani, Vihra Milanova, Philip B. Mitchell, Younes Mokrab, Grant W. Montgomery, Jennifer L. Moran, Gunnar Morken, Derek W. Morris, Ole Mors, Preben B. Mortensen, Bryan J. Mowry, Thomas W. Mühleisen, Bertram Müller-Myhsok, Kieran C. Murphy, Robin M. Murray, Richard M. Myers, Inez Myin-Germeys, Benjamin M. Neale, Mari Nelis, Igor Nenadic, Deborah A. Nertney, Gerald Nestadt, Kristin K. Nicodemus, Caroline M. Nievergelt, Liene Nikitina-Zake, Vishwajit Nimgaonkar, Laura Nisenbaum, Merete Nordentoft, Annelie Nordin, Markus M. Nöthen, Evaristus A. Nwulia, Eadbhard O'Callaghan, Claire O'Donovan, Colm O'Dushlaine, F. Anthony O'Neill, Ketil J. Oedegaard, Sang-Yun Oh, Ann Olincy, Line Olsen, Lilijana Oruc, Jim Van Os, Michael J. Owen, Sara A. Paciga, Colin N. A. Palmer, Aarno Palotie, Christos Pantelis, George N. Papadimitriou, Elena Parkhomenko, Carlos Pato, Michele T. Pato, Tiina Paunio, Richard Pearson, Diana O. Perkins, Roy H. Perlis,

Amy Perry, Tune H. Pers, Tracey L. Petryshen, Andrea Pfennig, Marco Picchioni, Olli Pietilainen, Jonathan Pimm, Matti Pirinen, Robert Plomin, Andrew J. Pocklington, Danielle Posthuma, James B. Potash, Simon C. Potter, John Powell, Alkes Price, Ann E. Pulver, Shaun M. Purcell, Digby Quested, Josep Antoni Ramos-Quiroga, Henrik B. Rasmussen, Anna Rautanen, Radhi Ravindrarajah, Eline J. Regeer, Abraham Reichenberg, Andreas Reif, Mark A. Reimers, Marta Ribases, John P. Rice, Alexander L. Richards, Michelle Ricketts, Brien P. Riley, Fabio Rivas, Margarita Rivera, Joshua L. Roffman, Guy A. Rouleau, Panos Roussos, Dan Rujescu, Veikko Salomaa, Cristina Sanchez-Mora, Alan R. Sanders, Stephen J. Sawcer, Ulrich Schall, Alan F. Schatzberg, William A. Scheftner, Peter R. Schofield, Nicholas J. Schork, Sibylle G. Schwab, Edward M. Scolnick, Laura J. Scott, Rodney J. Scott, Larry J. Seidman, Alessandro Serretti, Pak C. Sham, Cynthia Shannon Weickert, Tatyana Shehktman, Jianxin Shi, Paul D. Shilling, Engilbert Sigurdsson, Jeremy M. Silverman, Kang Sim, Claire Slaney, Petr Slominsky, Olav B. Smeland, Jordan W. Smoller, Hon-Cheong So, Janet L. Sobell, Erik Soderman, Christine Soholm Hansen, Chris C. A. Spencer, Anne T. Spijker, David St Clair, Hreinn Stefansson, Kari Stefansson, Stacy Steinberg, Elisabeth Stogmann, Eystein Stordal, Amy Strange, Richard E. Straub, John S. Strauss, Fabian Streit, Eric Strengman, Jana Strohmaier, T. Scott Stroup, Zhan Su, Mythily Subramaniam, Jaana Suvisaari, Dragan M. Svrakic, Jin P. Szatkiewicz, Szabolcs Szelinger, Avazeh Tashakkori-Ghanbaria, Srinivas Thirumalai, Robert C. Thompson, Thorgeir E. Thorgeirsson, Draga Toncheva, Paul A. Tooney, Sarah Tosato, Timothea Toulopoulou, Richard C. Trembath, Jens Treutlein, Gustavo Turecki, Arne E. Vaaler, Helmut Vedder, Eduard Vieta, John Vincent, Peter M. Visscher, Ananth C. Viswanathan, Damjan Vukcevic, John Waddington, Matthew Waller, Dermot Walsh, Muriel Walshe, James T. R. Walters, Dai Wang, Qiang Wang, Weiqing Wang, Yunpeng Wang, Stanley J. Watson, Bradley T. Webb, Thomas W. Weickert, Daniel R. Weinberger, Matthias Weisbrod, Mark Weiser, Thomas Werge, Paul Weston, Pamela Whittaker, Sara Widaa, Durk Wiersma, Dieter B. Wildenauer, Nigel M. Williams, Stephanie Williams, Stephanie H. Witt, Aaron R. Wolen, Emily H. M. Wong, Nicholas W. Wood, Brandon K. Wormley, Jing Qin Wu, Simon Xi, Wei Xu, Allan H. Young, Clement C. Zai, Peter Zandi, Peng Zhang, Xuebin Zheng, Fritz Zimprich, Sebastian Zollner, Aiden Corvin, Ayman H. Fanous, Sven Cichon, Marcella Rietschel, Elliot S. Gershon, Thomas G. Schulze, Alfredo B. Cuellar-Barboza, Andreas J. Forstner, Peter A. Holmans, John I. Nurnberger, Ole A. Andreassen, S. Hong Lee, Michael C. O'Donovan, Patrick F. Sullivan, Roel A. Ophoff, Naomi R. Wray, Pamela Sklar, and Kenneth S. Kendler. Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. *Cell*, 173(7):1705–1715.e16, 2021/05/13 2018.

[131] Heribert Schunkert, Inke R König, Sekar Kathiresan, Muredach P Reilly, Themistocles L Assimes, Hilma Holm, Michael Preuss, Alexandre F R Stewart, Maja Barbalic, Christian Gieger, Devin Absher, Zouhair Aherrahrou, Hooman Allayee, David Altshuler, Sonia S Anand, Karl Andersen, Jeffrey L Anderson, Diego Ardissino, Stephen G Ball, Anthony J Balmforth, Timothy A Barnes, Diane M Becker, Lewis C Becker,

Klaus Berger, Joshua C Bis, S Matthijs Boekholdt, Eric Boerwinkle, Peter S Braund, Morris J Brown, Mary Susan Burnett, Ian Buysschaert, John F Carlquist, Li Chen, Sven Cichon, Veryan Codd, Robert W Davies, George Dedoussis, Abbas Dehghan, Serkalem Demissie, Joseph M Devaney, Patrick Diemert, Ron Do, Angela Doering, Sandra Eifert, Nour Eddine El Mokhtari, Stephen G Ellis, Roberto Elosua, James C Engert, Stephen E Epstein, Ulf de Faire, Marcus Fischer, Aaron R Folsom, Jennifer Freyer, Bruna Gigante, Domenico Girelli, Solveig Gretarsdottir, Vilmundur Gudnason, Jeffrey R Gulcher, Eran Halperin, Naomi Hammond, Stanley L Hazen, Albert Hofman, Benjamin D Horne, Thomas Illig, Carlos Iribarren, Gregory T Jones, J Wouter Jukema, Michael A Kaiser, Lee M Kaplan, John J P Kastelein, Kay-Tee Khaw, Joshua W Knowles, Genovefa Kolovou, Augustine Kong, Reijo Laaksonen, Diether Lambrechts, Karin Leander, Guillaume Lettre, Mingyao Li, Wolfgang Lieb, Christina Loley, Andrew J Lotery, Pier M Mannucci, Seraya Maouche, Nicola Martinelli, Pascal P McKeown, Christa Meisinger, Thomas Meitinger, Olle Melander, Pier Angelica Merlini, Vincent Mooser, Thomas Morgan, Thomas W Mühleisen, Joseph B Muhlestein, Thomas Münzel, Kiran Musunuru, Janja Nahrstaedt, Christopher P Nelson, Markus M Nöthen, Oliviero Olivieri, Riyaz S Patel, Chris C Patterson, Annette Peters, Flora Peyvandi, Liming Qu, Arshed A Quyyumi, Daniel J Rader, Loukianos S Rallidis, Catherine Rice, Frits R Rosendaal, Diana Rubin, Veikko Salomaa, M Lourdes Sampietro, Manj S Sandhu, Eric Schadt, Arne Schäfer, Arne Schillert, Stefan Schreiber, Jürgen Schrezenmeir, Stephen M Schwartz, David S Siscovick, Mohan Sivananthan, Suthesh Sivapalaratnam, Albert Smith, Tamara B Smith, Jaapjan D Snoep, Nicole Soranzo, John A Spertus, Klaus Stark, Kathy Stirrups, Monika Stoll, W H Wilson Tang, Stephanie Tennstedt, Gudmundur Thorgeirsson, Gudmar Thorleifsson, Maciej Tomaszewski, Andre G Uitterlinden, Andre M van Rij, Benjamin F Voight, Nick J Wareham, George A Wells, H-Erich Wichmann, Philipp S Wild, Christina Willenborg, Jaqueline C M Witteman, Benjamin J Wright, Shu Ye, Tanja Zeller, Andreas Ziegler, Francois Cambien, Alison H Goodall, L Adrienne Cupples, Thomas Quertermous, Winfried März, Christian Hengstenberg, Stefan Blankenberg, Willem H Ouwehand, Alistair S Hall, Panos Deloukas, John R Thompson, Kari Stefansson, Robert Roberts, Unnur Thorsteinsdottir, Christopher J O'Donnell, Ruth McPherson, Jeanette Erdmann, Nilesh J Samani, Cardiogenics, and the CARDIoGRAM Consortium. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics*, 43(4):333–338, 2011.

[132] Matthias Wuttke, Yong Li, Man Li, Karsten B. Sieber, Mary F. Feitosa, Mathias Gorski, Adrienne Tin, Lihua Wang, Audrey Y. Chu, Anselm Hoppmann, Holger Kirsten, Ayush Giri, Jin-Fang Chai, Gardar Sveinbjornsson, Bamidele O. Tayo, Teresa Nutile, Christian Fuchsberger, Jonathan Marten, Massimiliano Cocca, Sahar Ghasemi, Yizhe Xu, Katrin Horn, Damia Noce, Peter J. van der Most, Sanaz Sedaghat, Zhi Yu, Masato Akiyama, Saima Afaq, Tarunveer S. Ahluwalia, Peter Almgren, Najaf Amin, Johan Ärnlöv, Stephan J. L. Bakker, Nisha Bansal, Daniela Baptista, Sven Bergmann, Mary L. Biggs, Ginevra Biino, Michael Boehnke, Eric Boerwinkle, Mathilde Boissel, Erwin P. Bottinger, Thibaud S. Boutin, Hermann Brenner, Marco Brumat, Ralph

Burkhardt, Adam S. Butterworth, Eric Campana, Archie Campbell, Harry Campbell, Mickaël Canouil, Robert J. Carroll, Eulalia Catamo, John C. Chambers, Miao-Ling Chee, Miao-Li Chee, Xu Chen, Ching-Yu Cheng, Yurong Cheng, Kaare Christensen, Renata Cifkova, Marina Ciullo, Maria Pina Concas, James P. Cook, Josef Coresh, Tanguy Corre, Cinzia Felicita Sala, Daniele Cusi, John Danesh, E. Warwick Daw, Martin H. de Borst, Alessandro De Grandi, Renée de Mutsert, Aiko P. J. de Vries, Frauke Degenhardt, Graciela Delgado, Ayse Demirkan, Emanuele Di Angelantonio, Katalin Dittrich, Jasmin Divers, Rajkumar Dorajoo, Kai-Uwe Eckardt, Georg Ehret, Paul Elliott, Karlhans Endlich, Michele K. Evans, Janine F. Felix, Valencia Hui Xian Foo, Oscar H. Franco, Andre Franke, Barry I. Freedman, Sandra Freitag-Wolf, Yechiel Friedlander, Philippe Froguel, Ron T. Gansevoort, He Gao, Paolo Gasparini, J. Michael Gaziano, Vilmantas Giedraitis, Christian Gieger, Giorgia Girotto, Franco Giulianini, Martin Gögele, Scott D. Gordon, Daniel F. Gudbjartsson, Vilmundur Gudnason, Toomas Haller, Pavel Hamet, Tamara B. Harris, Catharina A. Hartman, Caroline Hayward, Jacklyn N. Hellwege, Chew-Kiat Heng, Andrew A. Hicks, Edith Hofer, Wei Huang, Nina Hutri-Kähönen, Shih-Jen Hwang, M. Arfan Ikram, Olafur S. Indridason, Erik Ingelsson, Marcus Ising, Vincent W. V. Jaddoe, Johanna Jakobsdottir, Jost B. Jonas, Peter K. Joshi, Navya Shilpa Josyula, Bettina Jung, Mika Kähönen, Yoichiro Kamatani, Candace M. Kammerer, Masahiro Kanai, Mika Kastarinen, Shona M. Kerr, Chiea-Chuen Khor, Wieland Kiess, Marcus E. Kleber, Wolfgang Koenig, Jaspal S. Kooner, Antje Körner, Peter Kovacs, Aldi T. Kraja, Alena Krajcoviechova, Holly Kramer, Bernhard K. Krämer, Florian Kronenberg, Michiaki Kubo, Brigitte Kühnel, Mikko Kuokkanen, Johanna Kuusisto, Martina La Bianca, Markku Laakso, Leslie A. Lange, Carl D. Langefeld, Jeannette Jen-Mai Lee, Benjamin Lehne, Terho Lehtimäki, Wolfgang Lieb, Su-Chi Lim, Lars Lind, Cecilia M. Lindgren, Jun Liu, Jianjun Liu, Markus Loeffler, Ruth J. F. Loos, Susanne Lucae, Mary Ann Lukas, Leo-Pekka Lyytikäinen, Reedik Mägi, Patrik K. E. Magnusson, Anubha Mahajan, Nicholas G. Martin, Jade Martins, Winfried März, Deborah Mascalzoni, Koichi Matsuda, Christa Meisinger, Thomas Meitinger, Olle Melander, Andres Metspalu, Evgenia K. Mikaelsdottir, Yuri Milaneschi, Kozeta Miliku, Pashupati P. Mishra, Karen L. Mohlke, Nina Mononen, Grant W. Montgomery, Dennis O. Mook-Kanamori, Josyf C. Mychaleckyj, Girish N. Nadkarni, Mike A. Nalls, Matthias Nauck, Kjell Nikus, Boting Ning, Ilja M. Nolte, Raymond Noordam, Jeffrey O'Connell, Michelle L. O'Donoghue, Isleifur Olafsson, Albertine J. Oldehinkel, Marju Orho-Melander, Willem H. Ouwehand, Sandosh Padmanabhan, Nicholette D. Palmer, Runolfur Palsson, Brenda W. J. H. Penninx, Thomas Perls, Markus Perola, Mario Pirastu, Nicola Pirastu, Giorgio Pistis, Anna I. Podgornaia, Ozren Polasek, Belen Ponte, David J. Porteous, Tanja Poulain, Peter P. Pramstaller, Michael H. Preuss, Bram P. Prins, Michael A. Province, Ton J. Rabelink, Laura M. Raffield, Olli T. Raitakari, Dermot F. Reilly, Rainer Rettig, Myriam Rheinberger, Kenneth M. Rice, Paul M. Ridker, Fernando Rivadeneira, Federica Rizzi, David J. Roberts, Antonietta Robino, Peter Rossing, Igor Rudan, Rico Rueedi, Daniela Ruggiero, Kathleen A. Ryan, Yasaman Saba, Charumathi Sabanayagam, Veikko Salomaa, Erika Salvi, Kai-Uwe Saum, Helena Schmidt, Reinhold Schmidt, Ben Schöttker, Christina-Alexandra Schulz, Nicole Schupf, Christian M. Shaffer, Yuan Shi, Albert V.

Smith, Blair H. Smith, Nicole Soranzo, Cassandra N. Spracklen, Konstantin Strauch, Heather M. Stringham, Michael Stumvoll, Per O. Svensson, Silke Szymczak, E-Shyong Tai, Salman M. Tajuddin, Nicholas Y. Q. Tan, Kent D. Taylor, Andrej Teren, Yih-Chung Tham, Joachim Thiery, Chris H. L. Thio, Hauke Thomsen, Gudmar Thorleifsson, Daniela Toniolo, Anke Tönjes, Johanne Tremblay, Ioanna Tzoulaki, AndréG. Uitterlinden, Simona Vaccargiu, Rob M. van Dam, Pim van der Harst, Cornelia M. van Duijn, Digna R. Velez Edward, Niek Verweij, Suzanne Vogelezang, Uwe Völker, Peter Vollenweider, Gerard Waeber, Melanie Waldenberger, Lars Wallentin, Ya Xing Wang, Chaolong Wang, Dawn M. Waterworth, Wen Bin Wei, Harvey White, John B. Whitfield, Sarah H. Wild, James F. Wilson, Mary K. Wojczynski, Charlene Wong, Tien-Yin Wong, Liang Xu, Qiong Yang, Lifelines Cohort Study, and V. A. Million Veteran Program. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nature Genetics*, 51(6):957–972, 2019.

[133] Alisa K Manning, Marie-France Hivert, Robert A Scott, Jonna L Grimsby, Nabila Bouatia-Naji, Han Chen, Denis Rybin, Ching-Ti Liu, Lawrence F Bielak, Inga Prokopenko, Najaf Amin, Daniel Barnes, Gemma Cadby, Jouke-Jan Hottenga, Erik Ingelsson, Anne U Jackson, Toby Johnson, Stavroula Kanoni, Claes Ladenvall, Vasiliki Lagou, Jari Lahti, Cecile Lecoeur, Yongmei Liu, Maria Teresa Martinez-Larrad, May E Montasser, Pau Navarro, John R B Perry, Laura J Rasmussen-Torvik, Perttu Salo, Naveed Sattar, Dmitry Shungin, Rona J Strawbridge, Toshiko Tanaka, Cornelia M van Duijn, Ping An, Mariza de Andrade, Jeanette S Andrews, Thor Aspelund, Mustafa Atalay, Yurii Aulchenko, Beverley Balkau, Stefania Bandinelli, Jacques S Beckmann, John P Beilby, Claire Bellis, Richard N Bergman, John Blangero, Mladen Boban, Michael Boehnke, Eric Boerwinkle, Lori L Bonnycastle, Dorret I Boomsma, Ingrid B Borecki, Yvonne Böttcher, Claude Bouchard, Eric Brunner, Danijela Budimir, Harry Campbell, Olga Carlson, Peter S Chines, Robert Clarke, Francis S Collins, Arturo Corbatón-Anchuelo, David Couper, Ulf de Faire, George V Dedoussis, Panos Deloukas, Maria Dimitriou, Josephine M Egan, Gudny Eiriksdottir, Michael R Erdos, Johan G Eriksson, Elodie Eury, Luigi Ferrucci, Ian Ford, Nita G Forouhi, Caroline S Fox, Maria Grazia Franzosi, Paul W Franks, Timothy M Frayling, Philippe Froguel, Pilar Galan, Eco de Geus, Bruna Gigante, Nicole L Glazer, Anuj Goel, Leif Groop, Vilmundur Gudnason, Göran Hallmans, Anders Hamsten, Ola Hansson, Tamara B Harris, Caroline Hayward, Simon Heath, Serge Hercberg, Andrew A Hicks, Aroon Hingorani, Albert Hofman, Jennie Hui, Joseph Hung, Marjo-Riitta Jarvelin, Min A Jhun, Paul C D Johnson, J Wouter Jukema, Antti Jula, W H Kao, Jaakko Kaprio, Sharon L R Kardia, Sirkka Keinanen-Kiukaanniemi, Mika Kivimaki, Ivana Kolcic, Peter Kovacs, Meena Kumari, Johanna Kuusisto, Kirsten Ohm Kyvik, Markku Laakso, Timo Lakka, Lars Lannfelt, G Mark Lathrop, Lenore J Launer, Karin Leander, Guo Li, Lars Lind, Jaana Lindstrom, Stéphane Lobbens, Ruth J F Loos, Jian'an Luan, Valeriya Lyssenko, Reedik Mägi, Patrik K E Magnusson, Michael Marmot, Pierre Meneton, Karen L Mohlke, Vincent Mooser, Mario A Morken, Iva Miljkovic, Narisu Narisu, Jeff O'Connell, Ken K Ong, Ben A Oostra, Lyle J Palmer, Aarno Palotie, James S Pankow, John F Peden, Nancy L Pedersen, Marina Pehlic, Leena Peltonen,

Brenda Penninx, Marijana Pericic, Markus Perola, Louis Perusse, Patricia A Peyser, Ozren Polasek, Peter P Pramstaller, Michael A Province, Katri Räikkönen, Rainer Rauramaa, Emil Rehnberg, Ken Rice, Jerome I Rotter, Igor Rudan, Aimo Ruokonen, Timo Saaristo, Maria Sabater-Lleal, Veikko Salomaa, David B Savage, Richa Saxena, Peter Schwarz, Udo Seedorf, Bengt Sennblad, Manuel Serrano-Rios, Alan R Shuldiner, Eric J G Sijbrands, David S Siscovick, Johannes H Smit, Kerrin S Small, Nicholas L Smith, Albert Vernon Smith, Alena Stančáková, Kathleen Stirrups, Michael Stumvoll, Yan V Sun, Amy J Swift, Anke Tönjes, Jaakko Tuomilehto, Stella Trompet, Andre G Uitterlinden, Matti Uusitupa, Max Vikström, Veronique Vitart, Marie-Claude Vohl, Benjamin F Voight, Peter Vollenweider, Gerard Waeber, Dawn M Waterworth, Hugh Watkins, Eleanor Wheeler, Elisabeth Widen, Sarah H Wild, Sara M Willems, Gonneke Willemsen, James F Wilson, Jacqueline C M Witteman, Alan F Wright, Hanieh Yaghootkar, Diana Zelenika, Tatijana Zemunik, Lina Zgaga, Nicholas J Wareham, Mark I McCarthy, Ines Barroso, Richard M Watanabe, Jose C Florez, Josée Dupuis, James B Meigs, Claudia Langenberg, DIAbetes Genetics Replication, Meta analysis (DIAGRAM) Consortium, and The Multiple Tissue Human Expression Resource (MUTHER) Consortium. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature Genetics*, 44(6):659–669, 2012.

[134] Tanya M. Teslovich, Kiran Musunuru, Albert V. Smith, Andrew C. Edmondson, Ioannis M. Stylianou, Masahiro Koseki, James P. Pirruccello, Samuli Ripatti, Daniel I. Chasman, Cristen J. Willer, Christopher T. Johansen, Sigrid W. Fouchier, Aaron Isaacs, Gina M. Peloso, Maja Barbalic, Sally L. Ricketts, Joshua C. Bis, Yurii S. Aulchenko, Gudmar Thorleifsson, Mary F. Feitosa, John Chambers, Marju Orho-Melander, Olle Melander, Toby Johnson, Xiaohui Li, Xiuqing Guo, Mingyao Li, Yoon Shin Cho, Min Jin Go, Young Jin Kim, Jong-Young Lee, Taesung Park, Kyunga Kim, Xueling Sim, Rick Twee-Hee Ong, Damien C. Croteau-Chonka, Leslie A. Lange, Joshua D. Smith, Kijoung Song, Jing Hua Zhao, Xin Yuan, Jian'an Luan, Claudia Lamina, Andreas Ziegler, Weihua Zhang, Robert Y. L. Zee, Alan F. Wright, Jacqueline C. M. Witteman, James F. Wilson, Gonneke Willemsen, H. Erich Wichmann, John B. Whitfield, Dawn M. Waterworth, Nicholas J. Wareham, Gérard Waeber, Peter Vollenweider, Benjamin F. Voight, Veronique Vitart, Andre G. Uitterlinden, Manuela Uda, Jaakko Tuomilehto, John R. Thompson, Toshiko Tanaka, Ida Surakka, Heather M. Stringham, Tim D. Spector, Nicole Soranzo, Johannes H. Smit, Juha Sinisalo, Kaisa Silander, Eric J. G. Sijbrands, Angelo Scuteri, James Scott, David Schlessinger, Serena Sanna, Veikko Salomaa, Juha Saharinen, Chiara Sabatti, Aimo Ruokonen, Igor Rudan, Lynda M. Rose, Robert Roberts, Mark Rieder, Bruce M. Psaty, Peter P. Pramstaller, Irene Pichler, Markus Perola, Brenda W. J. H. Penninx, Nancy L. Pedersen, Cristian Pattaro, Alex N. Parker, Guillaume Pare, Ben A. Oostra, Christopher J. O'Donnell, Markku S. Nieminen, Deborah A. Nickerson, Grant W. Montgomery, Thomas Meitinger, Ruth McPherson, Mark I. McCarthy, Wendy McArdle, David Masson, Nicholas G. Martin, Fabio Marroni, Massimo Mangino, Patrik K. E. Magnusson, Gavin Lucas, Robert Luben, Ruth J. F. Loos, Marja-Liisa Lokki,

Guillaume Lettre, Claudia Langenberg, Lenore J. Launer, Edward G. Lakatta, Reijo Laaksonen, Kirsten O. Kyvik, Florian Kronenberg, Inke R. König, Kay-Tee Khaw, Jaakko Kaprio, Lee M. Kaplan, Åsa Johansson, Marjo-Riitta Jarvelin, A. Cecile J. W. Janssens, Erik Ingelsson, Wilmar Igl, G. Kees Hovingh, Jouke-Jan Hottenga, Albert Hofman, Andrew A. Hicks, Christian Hengstenberg, Iris M. Heid, Caroline Hayward, Aki S. Havulinna, Nicholas D. Hastie, Tamara B. Harris, Talin Haritunians, Alistair S. Hall, Ulf Gyllensten, Candace Guiducci, Leif C. Groop, Elena Gonzalez, Christian Gieger, Nelson B. Freimer, Luigi Ferrucci, Jeanette Erdmann, Paul Elliott, Kenechi G. Ejebe, Angela Döring, Anna F. Dominiczak, Serkalem Demissie, Panagiotis Deloukas, Eco J. C. de Geus, Ulf de Faire, Gabriel Crawford, Francis S. Collins, Yiider I. Chen, Mark J. Caulfield, Harry Campbell, Noel P. Burtt, Lori L. Bonnycastle, Dorret I. Boomsma, S. Matthijs Boekholdt, Richard N. Bergman, Inês Barroso, Stefania Bandinelli, Christie M. Ballantyne, Themistocles L. Assimes, Thomas Quertermous, David Altshuler, Mark Seielstad, Tien Y. Wong, E-Shyong Tai, Alan B. Feranil, Christopher W. Kuzawa, Linda S. Adair, Herman A. Taylor Jr, Ingrid B. Borecki, Stacey B. Gabriel, James G. Wilson, Hilma Holm, Unnur Thorsteinsdottir, Vilmundur Gudnason, Ronald M. Krauss, Karen L. Mohlke, Jose M. Ordovas, Patricia B. Munroe, Jaspal S. Kooner, Alan R. Tall, Robert A. Hegele, John J. P. Kastelein, Eric E. Schadt, Jerome I. Rotter, Eric Boerwinkle, David P. Strachan, Vincent Mooser, Kari Stefansson, Muredach P. Reilly, Nilesh J Samani, Heribert Schunkert, L. Adrienne Cupples, Manjinder S. Sandhu, Paul M Ridker, Daniel J. Rader, Cornelia M. van Duijn, Leena Peltonen, Gonçalo R. Abecasis, Michael Boehnke, and Sekar Kathiresan. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–713, 2010.

[135] Cristen J Willer, Ellen M Schmidt, Sebanti Sengupta, Gina M Peloso, Stefan Gustafsson, Stavroula Kanoni, Andrea Ganna, Jin Chen, Martin L Buchkovich, Samia Mora, Jacques S Beckmann, Jennifer L Bragg-Gresham, Hsing-Yi Chang, Ayşe Demirkan, Heleen M Den Hertog, Ron Do, Louise A Donnelly, Georg B Ehret, Tõnu Esko, Mary F Feitosa, Teresa Ferreira, Krista Fischer, Pierre Fontanillas, Ross M Fraser, Daniel F Freitag, Deepti Gurdasani, Kauko Heikkilä, Elina Hyppönen, Aaron Isaacs, Anne U Jackson, Åsa Johansson, Toby Johnson, Marika Kaakinen, Johannes Kettunen, Marcus E Kleber, Xiaohui Li, Jian'an Luan, Leo-Pekka Lyytikäinen, Patrik K E Magnusson, Massimo Mangino, Evelin Mihailov, May E Montasser, Martina Müller-Nurasyid, Ilja M Nolte, Jeffrey R O'Connell, Cameron D Palmer, Markus Perola, Ann-Kristin Petersen, Serena Sanna, Richa Saxena, Susan K Service, Sonia Shah, Dmitry Shungin, Carlo Sidore, Ci Song, Rona J Strawbridge, Ida Surakka, Toshiko Tanaka, Tanya M Teslovich, Gudmar Thorleifsson, Evita G Van den Herik, Benjamin F Voight, Kelly A Volcik, Lindsay L Waite, Andrew Wong, Ying Wu, Weihua Zhang, Devin Absher, Gershim Asiki, Inês Barroso, Latonya F Been, Jennifer L Bolton, Lori L Bonnycastle, Paolo Brambilla, Mary S Burnett, Giancarlo Cesana, Maria Dimitriou, Alex S F Doney, Angela Döring, Paul Elliott, Stephen E Epstein, Gudmundur Ingi Eyjolfsson, Bruna Gigante, Mark O Goodarzi, Harald Grallert, Martha L Gravito, Christopher J Groves, Göran Hallmans, Anna-Liisa Hartikainen, Caroline Hayward, Dena Hernandez, An-

drew A Hicks, Hilma Holm, Yi-Jen Hung, Thomas Illig, Michelle R Jones, Pontiano Kaleebu, John J P Kastelein, Kay-Tee Khaw, Eric Kim, Norman Klopp, Pirjo Komulainen, Meena Kumari, Claudia Langenberg, Terho Lehtimäki, Shih-Yi Lin, Jaana Lindström, Ruth J F Loos, François Mach, Wendy L McArdle, Christa Meisinger, Braxton D Mitchell, Gabrielle Müller, Ramaiah Nagaraja, Narisu Narisu, Tuomo V M Nieminen, Rebecca N Nsubuga, Isleifur Olafsson, Ken K Ong, Aarno Palotie, Theodore Papamarkou, Cristina Pomilla, Anneli Pouta, Daniel J Rader, Muredach P Reilly, Paul M Ridker, Fernando Rivadeneira, Igor Rudan, Aimo Ruokonen, Nilesh Samani, Hubert Scharnagl, Janet Seeley, Kaisa Silander, Alena Stancáková, Kathleen Stirrups, Amy J Swift, Laurence Tiret, Andre G Uitterlinden, L Joost van Pelt, Sailaja Vedantam, Nicholas Wainwright, Cisca Wijmenga, Sarah H Wild, Gonneke Willemsen, Tom Wilsgaard, James F Wilson, Elizabeth H Young, Jing Hua Zhao, Linda S Adair, Dominique Arveiler, Themistocles L Assimes, Stefania Bandinelli, Franklyn Bennett, Murielle Bochud, Bernhard O Boehm, Dorret I Boomsma, Ingrid B Borecki, Stefan R Bornstein, Pascal Bovet, Michel Burnier, Harry Campbell, Aravinda Chakravarti, John C Chambers, Yii-Der Ida Chen, Francis S Collins, Richard S Cooper, John Danesh, George Dedoussis, Ulf de Faire, Alan B Feranil, Jean Ferrières, Luigi Ferrucci, Nelson B Freimer, Christian Gieger, Leif C Groop, Vilmundur Gudnason, Ulf Gyllensten, Anders Hamsten, Tamara B Harris, Aroon Hingorani, Joel N Hirschhorn, Albert Hofman, G Kees Hovingh, Chao Agnes Hsiung, Steve E Humphries, Steven C Hunt, Kristian Hveem, Carlos Iribarren, Marjo-Riitta Järvelin, Antti Jula, Mika Kähönen, Jaakko Kaprio, Antero Kesäniemi, Mika Kivimaki, Jaspal S Kooner, Peter J Koudstaal, Ronald M Krauss, Diana Kuh, Johanna Kuusisto, Kirsten O Kyvik, Markku Laakso, Timo A Lakka, Lars Lind, Cecilia M Lindgren, Nicholas G Martin, Winfried März, Mark I McCarthy, Colin A McKenzie, Pierre Meneton, Andres Metspalu, Leena Moilanen, Andrew D Morris, Patricia B Munroe, Inger Njølstad, Nancy L Pedersen, Chris Power, Peter P Pramstaller, Jackie F Price, Bruce M Psaty, Thomas Quertermous, Rainer Rauramaa, Danish Saleheen, Veikko Salomaa, Dharambir K Sanghera, Jouko Saramies, Peter E H Schwarz, Wayne H-H Sheu, Alan R Shuldiner, Agneta Siegbahn, Tim D Spector, Kari Stefansson, David P Strachan, Bamidele O Tayo, Elena Tremoli, Jaakko Tuomilehto, Matti Uusitupa, Cornelia M van Duijn, Peter Vollenweider, Lars Wallentin, Nicholas J Wareham, John B Whitfield, Bruce H R Wolffenbuttel, Jose M Ordovas, Eric Boerwinkle, Colin N A Palmer, Unnur Thorsteinsdottir, Daniel I Chasman, Jerome I Rotter, Paul W Franks, Samuli Ripatti, L Adrienne Cupples, Manjinder S Sandhu, Stephen S Rich, Michael Boehnke, Panos Deloukas, Sekar Kathiresan, Karen L Mohlke, Erik Ingelsson, Gonçalo R Abecasis, and Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, 45(11):1274–1283, 2013.

[136] James Bentham, David L Morris, Deborah S Cunninghame Graham, Christopher L Pinder, Philip Tombleson, Timothy W Behrens, Javier Martín, Benjamin P Fairfax, Julian C Knight, Lingyan Chen, Joseph Replogle, Ann-Christine Syvänen, Lars Rönnblom, Robert R Graham, Joan E Wither, John D Rioux, Marta E Alarcón-Riquelme, and Timothy J Vyse. Genetic association analyses implicate aberrant reg-

ulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nature Genetics*, 47(12):1457–1464, 2015.

[137] The International Multiple Sclerosis Genetics Consortium (IMSGC). Evidence for polygenic susceptibility to multiple sclerosis—the shape of things to come. *The American Journal of Human Genetics*, 86(4):621–625, 2021/05/13 2010.

[138] Heather J. Cordell, Younghun Han, George F. Mells, Yafang Li, Gideon M. Hirschfield, Casey S. Greene, Gang Xie, Brian D. Juran, Dakai Zhu, David C. Qian, James A. B. Floyd, Katherine I. Morley, Daniele Prati, Ana Lleo, Daniele Cusi, Erik M Schlicht, Craig Lammert, Elizabeth J Atkinson, Landon L Chan, Mariza de Andrade, Tobias Balschun, Andrew L Mason, Robert P Myers, Jinyi Zhang, Piotr Milkiewicz, Jia Qu, Joseph A Odin, Velimir A Luketic, Bruce R Bacon, Henry C Bodenheimer Jr, Valentina Liakina, Catherine Vincent, Cynthia Levy, Peter K Gregersen, Piero L Almasio, Domenico Alvaro, Pietro Andreone, Angelo Andriulli, Cristina Barlassina, Pier Maria Battezzati, Antonio Benedetti, Francesca Bernuzzi, Ilaria Bianchi, Maria Consiglia Bragazzi, Maurizia Brunetto, Savino Bruno, Giovanni Casella, Barbara Coco, Agostino Colli, Massimo Colombo, Silvia Colombo, Carmela Cursaro, Lory Saveria Crocè, Andrea Crosignani, Maria Francesca Donato, Gianfranco Elia, Luca Fabris, Carlo Ferrari, Annarosa Floreani, Barbara Foglieni, Rosanna Fontana, Andrea Galli, Roberta Lazzari, Fabio Macaluso, Federica Malinverno, Fabio Marra, Marco Marzioni, Alberto Mattalia, Renzo Montanari, Lorenzo Morini, Filomena Morisco, Mousa Hani S, Luigi Muratori, Paolo Muratori, Grazia A Niro, Vincenzo O Palmieri, Antonio Picciotto, Mauro Podda, Piero Portincasa, Vincenzo Ronca, Floriano Rosina, Sonia Rossi, Ilaria Sogno, Giancarlo Spinzi, Marta Spreafico, Mario Strazzabosco, Sonia Tarallo, Mirko Tarocchi, Claudio Tiribelli, Pierluigi Toniutto, Maria Vinci, Massimo Zuin, Chin Lye Ch'ng, Mesbah Rahman, Tom Yapp, Richard Sturgess, Christopher Healey, Marek Czajkowski, Anton Gunasekera, Pranab Gyawali, Purushothaman Premchand, Kapil Kapur, Richard Marley, Graham Foster, Alan Watson, Aruna Dias, Javaid Subhani, Rory Harvey, Roger McCorry, David Ramanaden, Jaber Gasem, Richard Evans, Thiriloganathan Mathialahan, Christopher Shorrock, George Lipscomb, Paul Southern, Jeremy Tibble, David Gorard, Altaf Palegwala, Susan Jones, Marco Carbone, Mohamed Dawwas, Graeme Alexander, Sunil Dolwani, Martin Prince, Matthew Foxton, David Elphick, Harriet Mitchison, Ian Gooding, Mazn Karmo, Sushma Saksena, Mike Mendall, Minesh Patel, Roland Ede, Andrew Austin, Joanna Sayer, Lorraine Hankey, Christopher Hovell, Neil Fisher, Martyn Carter, Konrad Koss, Andrzej Piotrowicz, Charles Grimley, David Neal, Guan Lim, Sass Levi, Aftab Ala, Andrea Broad, Athar Saeed, Gordon Wood, Jonathan Brown, Mark Wilkinson, Harriet Gordon, John Ramage, Jo Ridpath, Theodore Ngatchu, Bob Grover, Syed Shaukat, Ray Shidrawi, George Abouda, Faiz Ali, Ian Rees, Imroz Salam, Mark Narain, Ashley Brown, Simon Taylor-Robinson, Simon Williams, Leonie Grellier, Paul Banim, Debashis Das, Andrew Chilton, Michael Heneghan, Howard Curtis, Markus Gess, Ian Drake, Mark Aldersley, Mervyn Davies, Rebecca Jones, Alastair McNair, Raj Srirajaskanthan, Maxton Pitcher, Sambit Sen, George Bird, Adrian Barnardo, Paul Kitchen, Kevin Yoong, Oza Chirag, Nurani Sivaramakrishnan, George MacFaul, David Jones, Amir Shah,

Chris Evans, Subrata Saha, Katharine Pollock, Peter Bramley, Ashis Mukhopadhya, Andrew Fraser, Peter Mills, Christopher Shallcross, Stewart Campbell, Andrew Bathgate, Alan Shepherd, John Dillon, Simon Rushbrook, Robert Przemioslo, Christopher Macdonald, Jane Metcalf, Udi Shmueli, Andrew Davis, Asifabbas Naqvi, Tom Lee, Stephen D Ryder, Jane Collier, Howard Klass, Mary Ninkovic, Matthew Cramp, Nicholas Sharer, Richard Aspinall, Patrick Goggin, Deb Ghosh, Andrew Douds, Barbara Hoeroldt, Jonathan Booth, Earl Williams, Hyder Hussaini, William Stableforth, Reuben Ayres, Douglas Thorburn, Eileen Marshall, Andrew Burroughs, Steven Mann, Martin Lombard, Paul Richardson, Imran Patanwala, Julia Maltby, Matthew Brookes, Ray Mathew, Samir Vyas, Saket Singhal, Dermot Gleeson, Sharat Misra, Jeff Butterworth, Keith George, Tim Harding, Andrew Douglass, Simon Panter, Jeremy Shearman, Gary Bray, Graham Butcher, Daniel Forton, John Mclindon, Matthew Cowan, Gregory Whatley, Aditya Mandal, Hemant Gupta, Pradeep Sanghi, Sanjiv Jain, Steve Pereira, Geeta Prasad, Gill Watts, Mark Wright, James Neuberger, Fiona Gordon, Esther Unitt, Allister Grant, Toby Delahooke, Andrew Higham, Alison Brind, Mark Cox, Subramaniam Ramakrishnan, Alistair King, Carole Collins, Simon Whalley, Andy Li, Jocelyn Fraser, Andrew Bell, Voi Shim Wong, Amit Singhal, Ian Gee, Yeng Ang, Rupert Ransford, James Gotto, Charles Millson, Jane Bowles, Caradog Thomas, Melanie Harrison, Roman Galaska, Jennie Kendall, Jessica Whiteman, Caroline Lawlor, Catherine Gray, Keith Elliott, Caroline Mulvaney-Jones, Lucie Hobson, Greta Van Duyvenvoorde, Alison Loftus, Katie Seward, Canadian-US PBC Consortium, Italian PBC Genetics Study Group, and UK-PBC Consortium. International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nature Communications*, 6(1):8019, 2015.

[139] Yukinori Okada, Di Wu, Gosia Trynka, Towfique Raj, Chikashi Terao, Katsunori Ikari, Yuta Kochi, Koichiro Ohmura, Akari Suzuki, Shinji Yoshida, Robert R. Graham, Arun Manoharan, Ward Ortmann, Tushar Bhangale, Joshua C. Denny, Robert J. Carroll, Anne E. Eyler, Jeffrey D. Greenberg, Joel M. Kremer, Dimitrios A. Pappas, Lei Jiang, Jian Yin, Lingying Ye, Ding-Feng Su, Jian Yang, Gang Xie, Ed Keystone, Harm-Jan Westra, Tõnu Esko, Andres Metspalu, Xuezhong Zhou, Namrata Gupta, Daniel Mirel, Eli A. Stahl, Dorothée Diogo, Jing Cui, Katherine Liao, Michael H. Guo, Keiko Myouzen, Takahisa Kawaguchi, Marieke J. H. Coenen, Piet L. C. M. van Riel, Mart A. F. J. van de Laar, Henk-Jan Guchelaar, Tom W. J. Huizinga, Philippe Dieudé, Xavier Mariette, S. Louis Bridges Jr, Alexandra Zhernakova, Rene E. M. Toes, Paul P. Tak, Corinne Miceli-Richard, So-Young Bang, Hye-Soon Lee, Javier Martin, Miguel A. Gonzalez-Gay, Luis Rodriguez-Rodriguez, Solbritt Rantapää-Dahlqvist, Lisbeth Ärlestig, Hyon K. Choi, Yoichiro Kamatani, Pilar Galan, Mark Lathrop, Steve Eyre, John Bowes, Anne Barton, Niek de Vries, Larry W. Moreland, Lindsey A. Criswell, Elizabeth W. Karlson, Atsuo Taniguchi, Ryo Yamada, Michiaki Kubo, Jun S. Liu, Sang-Cheol Bae, Jane Worthington, Leonid Padyukov, Lars Klareskog, Peter K. Gregersen, Soumya Raychaudhuri, Barbara E. Stranger, Philip L. De Jager, Lude Franke, Peter M. Visscher, Matthew A. Brown, Hisashi Yamanaka, Tsuneyo Mimori, Atsushi Takahashi, Huji Xu, Timothy W. Behrens, Katherine A.

Siminovitch, Shigeki Momohara, Fumihiko Matsuda, Kazuhiko Yamamoto, Robert M. Plenge, the RACI consortium, and the GARNET consortium. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488):376–381, 2014.

[140] Jamie R. J. Inshaw, Carlo Sidore, Francesco Cucca, M. Irina Stefana, Daniel J. M. Crouch, Mark I. McCarthy, Anubha Mahajan, and John A. Todd. Analysis of overlapping genetic association in type 1 and type 2 diabetes. *Diabetologia*, 64(6):1342–1347, 2021.

[141] Anubha Mahajan, Daniel Taliun, Matthias Thurner, Neil R. Robertson, Jason M. Torres, N. William Rayner, Anthony J. Payne, Valgerdur Steinthorsdottir, Robert A. Scott, Niels Grarup, James P. Cook, Ellen M. Schmidt, Matthias Wuttke, Chloé Sarnowski, Reedik Mägi, Jana Nano, Christian Gieger, Stella Trompet, Cécile Lecoeur, Michael H. Preuss, Bram Peter Prins, Xiuqing Guo, Lawrence F. Bielak, Jennifer E. Below, Donald W. Bowden, John Campbell Chambers, Young Jin Kim, Maggie C. Y. Ng, Lauren E. Petty, Xueling Sim, Weihua Zhang, Amanda J. Bennett, Jette Bork-Jensen, Chad M. Brummett, Mickaël Canouil, Kai-Uwe Ec kardt, Krista Fischer, Sharon L. R. Kardia, Florian Kronenberg, Kristi Läll, Ching-Ti Liu, Adam E. Locke, Jian'an Luan, Ioanna Ntalla, Vibe Nylander, Sebastian Schönherr, Claudia Schurmann, Loïc Yengo, Erwin P. Bottinger, Ivan Brandslund, Cramer Christensen, George Dedoussis, Jose C. Florez, Ian Ford, Oscar H. Franco, Timothy M. Frayling, Vilmantas Giedraitis, Sophie Hackinger, Andrew T. Hattersley, Christian Herder, M. Arfan Ikram, Martin Ingelsson, Marit E. Jørgensen, Torben Jørgensen, Jennifer Kriebel, Johanna Kuusisto, Symen Ligthart, Cecilia M. Lindgren, Allan Linneberg, Valeriya Lyssenko, Vasiliki Mamakou, Thomas Meitinger, Karen L. Mohlke, Andrew D. Morris, Girish Nadkarni, James S. Pankow, Annette Peters, Naveed Sattar, Alena Stančáková, Konstantin Strauch, Kent D. Taylor, Barbara Thorand, Gudmar Thorleifsson, Unnur Thorsteinsdottir, Jaakko Tuomilehto, Daniel R. Witte, Josée Dupuis, Patricia A. Peyser, Eleftheria Zeggini, Ruth J. F. Loos, Philippe Froguel, Erik Ingelsson, Lars Lind, Leif Groop, Markku Laakso, Francis S. Collins, J. Wouter Jukema, Colin N. A. Palmer, Harald Grallert, Andres Metspalu, Abbas Dehghan, Anna Köttgen, Goncalo R. Abecasis, James B. Meigs, Jerome I. Rotter, Jonathan Marchini, Oluf Pedersen, Torben Hansen, Claudia Langenberg, Nicholas J. Wareham, Kari Stefansson, Anna L. Gloyn, Andrew P. Morris, Michael Boehnke, and Mark I. McCarthy. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics*, 50(11):1505–1513, 2018.

[142] Lale Ozcan, Catherine C. L. Wong, Gang Li, Tao Xu, Utpal Pajvani, Sung Kyu Robin Park, Anetta Wronska, Bi-Xing Chen, Andrew R. Marks, Akiyoshi Fukamizu, Johannes Backs, Harold A. Singer, III Yates, John R., Domenico Accili, and Ira Tabas. Calcium signaling through camkii regulates hepatic glucose production in fasting and obesity. *Cell Metabolism*, 15(5):739–751, 2022/04/07 2012.

[143] Flavia Agata Cimini, Andrea Arena, Ilaria Barchetta, Antonella Tramutola, Valentina Ceccarelli, Chiara Lanzillotta, Mario Fontana, Laura Bertoccini, Frida Leonetti, Danila

Capoccia, Gianfranco Silecchia, Claudio Di Cristofano, Caterina Chiappetta, Fabio Di Domenico, Marco Giorgio Baroni, Marzia Perluigi, Maria Gisella Cavallo, and Eugenio Barone. Reduced biliverdin reductase-a levels are associated with early alterations of insulin signaling in obesity. *Biochim Biophys Acta Mol Basis Dis*, 1865(6):1490–1501, Jun 2019.

[144] Run-Lu Sun, Can-Xia Huang, Jin-Lan Bao, Jie-Yu Jiang, Bo Zhang, Shu-Xian Zhou, Wei-Bin Cai, Hong Wang, Jing-Feng Wang, and Yu-Ling Zhang. Cc-chemokine ligand 2 (ccl2) suppresses high density lipoprotein (hdl) internalization and cholesterol efflux via cc-chemokine receptor 2 (ccr2) induction and p42/44 mitogen-activated protein kinase (mapk) activation in human endothelial cells. *J Biol Chem*, 291(37):19532–19544, Sep 2016.

[145] D M Ritter, C A Jr Owen, E J Bowie, S R Rettke, T L Cole, H F Taswell, D M Ilstrup, R H Wiesner, and R A Krom. Evaluation of preoperative hematology-coagulation screening in liver transplantation. *Mayo Clin Proc*, 64(2):216–223, Feb 1989.

[146] M J Kim, J. Biag, D M Fass, M C Lewis, Q. Zhang, M. Fleishman, S P Gangwar, M. Machius, M. Fromer, S M Purcell, S A McCarroll, G. Rudenko, R T Premont, E M Scolnick, and S J Haggarty. Functional analysis of rare variants found in schizophrenia implicates a critical role for git1–pak3 signaling in neuroplasticity. *Molecular Psychiatry*, 22(3):417–429, 2017.

[147] Katharine R. Smith, Elizabeth C. Davenport, Jing Wei, Xiangning Li, Manavendra Pathania, Victoria Vaccaro, Zhen Yan, and Josef T. Kittler. Git1 and pix are essential for gabaa receptor synaptic stability and inhibitory neurotransmission. *Cell Reports*, 9(1):298–310, 2014.

[148] Sudipta Das, Marina Miller, Andrew K. Beppu, James Mueller, Matthew D. McGeough, Christine Vuong, Maya R. Karta, Peter Rosenthal, Fazila Chouiali, Taylor A. Doherty, Richard C. Kurten, Qutayba Hamid, Hal M. Hoffman, and David H. Broide. Gsdmb induces an asthma phenotype characterized by increased airway responsiveness and remodeling without lung inflammation. *Proceedings of the National Academy of Sciences*, 113(46):13132–13137, 2016.

[149] Noam D. Beckmann, Wei-Jye Lin, Minghui Wang, Ariella T. Cohain, Alexander W. Charney, Pei Wang, Weiping Ma, Ying-Chih Wang, Cheng Jiang, Mickael Audrain, Phillip H. Comella, Amanda K. Fakira, Siddharth P. Hariharan, Gillian M. Belbin, Kiran Girdhar, Allan I. Levey, Nicholas T. Seyfried, Eric B. Dammer, Duc Duong, James J. Lah, Jean-Vianney Haure-Mirande, Ben Shackleton, Tomas Fanutza, Robert Blitzer, Eimear Kenny, Jun Zhu, Vahram Haroutunian, Pavel Katsel, Sam Gandy, Zhidong Tu, Michelle E. Ehrlich, Bin Zhang, Stephen R. Salton, and Eric E. Schadt. Multiscale causal networks identify vgf as a key regulator of alzheimer's disease. *Nature Communications*, 11(1):3942, 2020.

[150] Lin Jia, Juan Piña-Crespo, and Yonghe Li. Restoring wnt/-catenin signaling is a promising therapeutic strategy for alzheimer's disease. *Molecular Brain*, 12(1):104, 2019.

[151] Milena Duitama, Viviana Vargas-López, Zulma Casas, Sonia L Albarracin, Jhon-Jairo Sutachan, and Yolima P Torres. Trp channels role in pain associated with neurodegenerative diseases. *Front Neurosci*, 14:782, 2020.

[152] Laura Thei, Jennifer Imm, Eleni Kaisis, Mark L Dallas, and Talitha L Kerrigan. Microglia in alzheimer's disease: A role for ion channels. *Frontiers in neuroscience*, 12:676–676, 09 2018.

[153] Ethan R Roy, Baiping Wang, Ying-Wooi Wan, Gabriel Chiu, Allysa Cole, Zhuoran Yin, Nicholas E Propson, Yin Xu, Joanna L Jankowsky, Zhandong Liu, Virginia M-Y Lee, John Q Trojanowski, Stephen D Ginsberg, Oleg Butovsky, Hui Zheng, and Wei Cao. Type i interferon response drives neuroinflammation and synapse loss in alzheimer disease. *J Clin Invest*, 130(4):1912–1930, Apr 2020.

[154] Lei Meng, Zhe Wang, Hong-Fang Ji, and Liang Shen. Causal association evaluation of diabetes with alzheimer's disease and genetic analysis of antidiabetic drugs against alzheimer's disease. *Cell & Bioscience*, 12(1):28, 2022.

[155] Ken Sugimoto. Role of stat3 in inflammatory bowel disease. *World journal of gastroenterology*, 14(33):5110–5114, 09 2008.

[156] Yiguo Shen, David Kapfhamer, Angela M. Minnella, Ji-Eun Kim, Seok Joon Won, Yanting Chen, Yong Huang, Ley Hian Low, Stephen M. Massa, and Raymond A. Swanson. Bioenergetic state regulates innate inflammatory responses through the transcriptional co-repressor ctbp. *Nature Communications*, 8(1):624, 2017.

[157] Saleh A Naser, Melissa Arce, Anam Khaja, Marlene Fernandez, Najih Naser, Sammer Elwasila, and Saisathya Thanigachalam. Role of atg16l, nod2 and il23r in crohn's disease pathogenesis. *World journal of gastroenterology*, 18(5):412–424, 02 2012.

[158] Lumin Wei, Rongjing Zhang, Jinzhao Zhang, Juanjuan Li, Deping Kong, Qi Wang, Jing Fang, and Lifu Wang. Prkar2a deficiency protects mice from experimental colitis by increasing ifn-stimulated gene expression and modulating the intestinal microbiota. *Mucosal Immunology*, 14(6):1282–1294, 2021.

[159] Cristiana Cruceanu, Elena Kutsarova, Elizabeth S Chen, David R Checknita, Corina Nagy, Juan Pablo Lopez, Martin Alda, Guy A Rouleau, and Gustavo Turecki. Dna hypomethylation of synapsin ii cpg islands associates with increased gene expression in bipolar disorder and major depression. *BMC Psychiatry*, 16(1):286, Aug 2016.

[160] Saveen Sall, Willie Thompson, Aurianna Santos, and Donard S Dwyer. Analysis of major depression risk genes reveals evolutionary conservation, shared phenotypes, and extensive genetic interactions. *Frontiers in psychiatry*, 12:698029–698029, 07 2021.

[161] Cheng Jiang and Stephen R Salton. The role of neurotrophins in major depressive disorder. *Translational neuroscience*, 4(1):46–58, 03 2013.

[162] Falk W. Lohoff. Overview of the genetics of major depressive disorder. *Current Psychiatry Reports*, 12(6):539–546, 2010.

[163] Alanna Strong, Kevin Patel, and Daniel J Rader. Sortilin and lipoprotein metabolism: making sense out of complexity. *Current opinion in lipidology*, 25(5):350–357, 10 2014.

[164] Kiran Musunuru, Alanna Strong, Maria Frank-Kamenetsky, Noemi E Lee, Tim Ahfeldt, Katherine V Sachs, Xiaoyu Li, Hui Li, Nicolas Kuperwasser, Vera M Ruda, James P Pirruccello, Brian Muchmore, Ludmila Prokunina-Olsson, Jennifer L Hall, Eric E Schadt, Carlos R Morales, Sissel Lund-Katz, Michael C Phillips, Jamie Wong, William Cantley, Timothy Racie, Kenechi G Ejebe, Marju Orho-Melander, Olle Melander, Victor Koteliansky, Kevin Fitzgerald, Ronald M Krauss, Chad A Cowan, Sekar Kathiresan, and Daniel J Rader. From noncoding variant to phenotype via sort1 at the 1p13 cholesterol locus. *Nature*, 466(7307):714–719, Aug 2010.

[165] Claudia Goettsch, Mads Kjolby, and Elena Aikawa. Sortilin and its multiple roles in cardiovascular and metabolic diseases. *Arterioscler Thromb Vasc Biol*, 38(1):19–25, Jan 2018.

[166] Richard L. Eckert, Ann-Marie Broome, Monica Ruse, Nancy Robinson, David Ryan, and Kathleen Lee. S100 proteins in the epidermis. *Journal of Investigative Dermatology*, 123(1):23–33, 2004.

[167] Regine Gläser, Ulf Meyer-Hoffert, Jürgen Harder, Jesko Cordes, Maike Wittersheim, Julia Kobliakova, Regina Fölster-Holst, Ehrhardt Proksch, Jens-Michael Schröder, and Thomas Schwarz. The antimicrobial protein psoriasin (s100a7) is upregulated in atopic dermatitis and after experimental skin barrier disruption. *J Invest Dermatol*, 129(3):641–649, Mar 2009.

[168] Harald Lund, Elin Gustafsson, Anne Svensson, Maria Nilsson, Margareta Berg, Dan Sunnemark, and Gabriel von Euler. Mark4 and mark3 associate with early tau phosphorylation in alzheimer's disease granulovacuolar degeneration bodies. *Acta Neuropathologica Communications*, 2(1):22, 2014.

[169] Maurizio Cutolo, Alberto Sulli, Sabrina Paolino, and Carmen Pizzorni. Ctla-4 blockade in the treatment of rheumatoid arthritis: an update. *Expert Rev Clin Immunol*, 12(4):417–425, 2016.

[170] Anne Ndungu, Anthony Payne, Jason M. Torres, Martijn van de Bunt, and Mark I. McCarthy. A multi-tissue transcriptome analysis of human metabolites guides interpretability of associations based on multi-snp models for gene expression. *The American Journal of Human Genetics*, 106(2):188–201, 2020.

[171] Andrew S. Levey, Kai-Uwe Eckardt, Yusuke Tsukamoto, Adeera Levin, Josef Coresh, Jerome Rossert, Dick D.E. Zeeuw, Thomas H. Hostetter, Norbert Lameire, and Garabed Eknoyan. Definition and classification of chronic kidney disease: A position statement from kidney disease: Improving global outcomes (kdigo). *Kidney International*, 67(6):2089–2100, 2005.

[172] Anna Köttgen, Nicole L Glazer, Abbas Dehghan, Shih-Jen Hwang, Ronit Katz, Man Li, Qiong Yang, Vilmundur Gudnason, Lenore J Launer, Tamara B Harris, Albert V Smith, Dan E Arking, Brad C Astor, Eric Boerwinkle, Georg B Ehret, Ingo Ruczinski, Robert B Scharpf, Yii-Der Ida Chen, Ian H de Boer, Talin Haritunians, Thomas Lumley, Mark Sarnak, David Siscovick, Emelia J Benjamin, Daniel Levy, Ashish Upadhyay, Yurii S Aulchenko, Albert Hofman, Fernando Rivadeneira, AndréG Uitterlinden, Cornelia M van Duijn, Daniel I Chasman, Guillaume Paré, Paul M Ridker, W H Linda Kao, Jacqueline C Witteman, Josef Coresh, Michael G Shlipak, and Caroline S Fox. Multiple loci associated with indices of renal function and chronic kidney disease. *Nat Genet*, 41(6):712–717, Jun 2009.

[173] Matthias Wuttke, Yong Li, Man Li, Karsten B. Sieber, Mary F. Feitosa, Mathias Gorski, Adrienne Tin, Lihua Wang, Audrey Y. Chu, Anselm Hoppmann, Holger Kirsten, Ayush Giri, Jin-Fang Chai, Gardar Sveinbjornsson, Bamidele O. Tayo, Teresa Nutile, Christian Fuchsberger, Jonathan Marten, Massimiliano Cocca, Sahar Ghasemi, Yizhe Xu, Katrin Horn, Damia Noce, Peter J. van der Most, Sanaz Sedaghat, Zhi Yu, Masato Akiyama, Saima Afaq, Tarunveer S. Ahluwalia, Peter Almgren, Najaf Amin, Johan Ärnlöv, Stephan J. L. Bakker, Nisha Bansal, Daniela Baptista, Sven Bergmann, Mary L. Biggs, Ginevra Biino, Michael Boehnke, Eric Boerwinkle, Mathilde Boissel, Erwin P. Bottinger, Thibaud S. Boutin, Hermann Brenner, Marco Brumat, Ralph Burkhardt, Adam S. Butterworth, Eric Campana, Archie Campbell, Harry Campbell, Mickaël Canouil, Robert J. Carroll, Eulalia Catamo, John C. Chambers, Miao-Ling Chee, Miao-Li Chee, Xu Chen, Ching-Yu Cheng, Yurong Cheng, Kaare Christensen, Renata Cifkova, Marina Ciullo, Maria Pina Concas, James P. Cook, Josef Coresh, Tanguy Corre, Cinzia Felicita Sala, Daniele Cusi, John Danesh, E. Warwick Daw, Martin H. de Borst, Alessandro De Grandi, Renée de Mutsert, Aiko P. J. de Vries, Frauke Degenhardt, Graciela Delgado, Ayse Demirkan, Emanuele Di Angelantonio, Katalin Dittrich, Jasmin Divers, Rajkumar Dorajoo, Kai-Uwe Eckardt, Georg Ehret, Paul Elliott, Karlhans Endlich, Michele K. Evans, Janine F. Felix, Valencia Hui Xian Foo, Oscar H. Franco, Andre Franke, Barry I. Freedman, Sandra Freitag-Wolf, Yechiel Friedlander, Philippe Froguel, Ron T. Gansevoort, He Gao, Paolo Gasparini, J. Michael Gaziano, Vilmantas Giedraitis, Christian Gieger, Giorgia Girotto, Franco Giulianini, Martin Gögele, Scott D. Gordon, Daniel F. Gudbjartsson, Vilmundur Gudnason, Toomas Haller, Pavel Hamet, Tamara B. Harris, Catharina A. Hartman, Caroline Hayward, Jacklyn N. Hellwege, Chew-Kiat Heng, Andrew A. Hicks, Edith Hofer, Wei Huang, Nina Hutri-Kähönen, Shih-Jen Hwang, M. Arfan Ikram, Olafur S. Indridason, Erik Ingelsson, Marcus Ising, Vincent W. V. Jaddoe, Johanna Jakobsdottir, Jost B. Jonas, Peter K. Joshi, Navya Shilpa Josyula, Bettina Jung, Mika Kähönen, Yoichiro Kamatani, Candace M. Kammerer, Masahiro Kanai, Mika Kastarinen, Shona M.

Kerr, Chiea-Chuen Khor, Wieland Kiess, Marcus E. Kleber, Wolfgang Koenig, Jaspal S. Kooner, Antje Körner, Peter Kovacs, Aldi T. Kraja, Alena Krajcoviechova, Holly Kramer, Bernhard K. Krämer, Florian Kronenberg, Michiaki Kubo, Brigitte Kühnel, Mikko Kuokkanen, Johanna Kuusisto, Martina La Bianca, Markku Laakso, Leslie A. Lange, Carl D. Langefeld, Jeannette Jen-Mai Lee, Benjamin Lehne, Terho Lehtimäki, Wolfgang Lieb, Su-Chi Lim, Lars Lind, Cecilia M. Lindgren, Jun Liu, Jianjun Liu, Markus Loeffler, Ruth J. F. Loos, Susanne Lucae, Mary Ann Lukas, Leo-Pekka Lyytikäinen, Reedik Mägi, Patrik K. E. Magnusson, Anubha Mahajan, Nicholas G. Martin, Jade Martins, Winfried März, Deborah Mascalzoni, Koichi Matsuda, Christa Meisinger, Thomas Meitinger, Olle Melander, Andres Metspalu, Evgenia K. Mikaelsdottir, Yuri Milaneschi, Kozeta Miliku, Pashupati P. Mishra, Karen L. Mohlke, Nina Mononen, Grant W. Montgomery, Dennis O. Mook-Kanamori, Josyf C. Mychaleckyj, Girish N. Nadkarni, Mike A. Nalls, Matthias Nauck, Kjell Nikus, Boting Ning, Ilja M. Nolte, Raymond Noordam, Jeffrey O'Connell, Michelle L. O'Donoghue, Isleifur Olafsson, Albertine J. Oldehinkel, Marju Orho-Melander, Willem H. Ouwehand, Sandosh Padmanabhan, Nicholette D. Palmer, Runolfur Palsson, Brenda W. J. H. Penninx, Thomas Perls, Markus Perola, Mario Pirastu, Nicola Pirastu, Giorgio Pistis, Anna I. Podgornaia, Ozren Polasek, Belen Ponte, David J. Porteous, Tanja Poulain, Peter P. Pramstaller, Michael H. Preuss, Bram P. Prins, Michael A. Province, Ton J. Rabelink, Laura M. Raffield, Olli T. Raitakari, Dermot F. Reilly, Rainer Rettig, Myriam Rheinberger, Kenneth M. Rice, Paul M. Ridker, Fernando Rivadeneira, Federica Rizzi, David J. Roberts, Antonietta Robino, Peter Rossing, Igor Rudan, Rico Rueedi, Daniela Ruggiero, Kathleen A. Ryan, Yasaman Saba, Charumathi Sabanayagam, Veikko Salomaa, Erika Salvi, Kai-Uwe Saum, Helena Schmidt, Reinhold Schmidt, Ben Schöttker, Christina-Alexandra Schulz, Nicole Schupf, Christian M. Shaffer, Yuan Shi, Albert V. Smith, Blair H. Smith, Nicole Soranzo, Cassandra N. Spracklen, Konstantin Strauch, Heather M. Stringham, Michael Stumvoll, Per O. Svensson, Silke Szymczak, E-Shyong Tai, Salman M. Tajuddin, Nicholas Y. Q. Tan, Kent D. Taylor, Andrej Teren, Yih-Chung Tham, Joachim Thiery, Chris H. L. Thio, Hauke Thomsen, Gudmar Thorleifsson, Daniela Toniolo, Anke Tönjes, Johanne Tremblay, Ioanna Tzoulaki, AndréG. Uitterlinden, Simona Vaccargiu, Rob M. van Dam, Pim van der Harst, Cornelia M. van Duijn, Digna R. Velez Edward, Niek Verweij, Suzanne Vogelezang, Uwe Völker, Peter Vollenweider, Gerard Waeber, Melanie Waldenberger, Lars Wallentin, Ya Xing Wang, Chaolong Wang, Dawn M. Waterworth, Wen Bin Wei, Harvey White, John B. Whitfield, Sarah H. Wild, James F. Wilson, Mary K. Wojczynski, Charlene Wong, Tien-Yin Wong, Liang Xu, Qiong Yang, Lifelines Cohort Study, and V. A. Million Veteran Program. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nature Genetics*, 51(6):957–972, 2019.

[174] Nicholas Mancuso, Malika K. Freund, Ruth Johnson, Huwenbo Shi, Gleb Kichaev, Alexander Gusev, and Bogdan Pasaniuc. Probabilistic fine-mapping of transcriptome-wide association studies. *Nature Genetics*, 51(4):675–682, 2019.

[175] Mathias Uhlén, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, In-

gMarie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina Al-Khalili Szigyarto, Jacob Odeberg, Dijana Djureinovic, Jenny Ottosson Takanen, Sophia Hober, Tove Alm, Per-Henrik Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan Rockberg, Peter Nilsson, Jochen M. Schwenk, Marica Hamsten, Kalle von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, Gunnar von Heijne, Jens Nielsen, and Fredrik Pontén. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.

[176] Max Karlsson, Cheng Zhang, Loren Méar, Wen Zhong, Andreas Digre, Borbala Katona, Evelina Sjöstedt, Lynn Butler, Jacob Odeberg, Philip Dusart, Fredrik Edfors, Per Oksvold, Kalle von Feilitzen, Martin Zwahlen, Muhammad Arif, Ozlem Altay, Xiangyu Li, Mehmet Ozcan, Adil Mardinoglu, Linn Fagerberg, Jan Mulder, Yonglun Luo, Fredrik Ponten, Mathias Uhlén, and Cecilia Lindskog. A single–cell type transcriptomics map of human tissues. *Science Advances*, 7(31):eabh2169, 2021.

[177] Lisa Bastarache, Jacob J. Hughey, Scott Hebbring, Joy Marlo, Wanke Zhao, Wanting T. Ho, Sara L. Van Driest, Tracy L. McGregor, Jonathan D. Mosley, Quinn S. Wells, Michael Temple, Andrea H. Ramirez, Robert Carroll, Travis Osterman, Todd Edwards, Douglas Ruderfer, Digna R. Velez Edwards, Rizwan Hamid, Joy Cogan, Andrew Glazer, Wei-Qi Wei, QiPing Feng, Murray Brilliant, Zhizhuang J. Zhao, Nancy J. Cox, Dan M. Roden, and Joshua C. Denny. Phenotype risk scores identify patients with unrecognized mendelian disease patterns. *Science*, 359(6381):1233–1239, 2018.

[178] Tim Hulsen, Saumya S. Jamuar, Alan R. Moody, Jason H. Karnes, Orsolya Varga, Stine Hedensted, Roberto Spreafico, David A. Hafler, and Eoin F. McKinney. From big data to precision medicine. *Frontiers in Medicine*, 6:34, 2019.

[179] Huiying Liang, Brian Y. Tsui, Hao Ni, Carolina C. S. Valentim, Sally L. Baxter, Guangjian Liu, Wenjia Cai, Daniel S. Kermany, Xin Sun, Jiancong Chen, Liya He, Jie Zhu, Pin Tian, Hua Shao, Lianghong Zheng, Rui Hou, Sierra Hewett, Gen Li, Ping Liang, Xuan Zang, Zhiqi Zhang, Liyan Pan, Huimin Cai, Rujuan Ling, Shuhua Li, Yongwang Cui, Shusheng Tang, Hong Ye, Xiaoyan Huang, Waner He, Wenqing Liang, Qing Zhang, Jianmin Jiang, Wei Yu, Jianqun Gao, Wanxing Ou, Yingmin Deng, Qiaozhen Hou, Bei Wang, Cuichan Yao, Yan Liang, Shu Zhang, Yaou Duan, Runze Zhang, Sarah Gibson, Charlotte L. Zhang, Oulan Li, Edward D. Zhang, Gabriel Karin, Nathan Nguyen, Xiaokang Wu, Cindy Wen, Jie Xu, Wenqin Xu, Bochu Wang, Winston Wang, Jing Li, Bianca Pizzato, Caroline Bao, Daoman Xiang, Wanting He, Suiqin He, Yugui Zhou, Weldon Haw, Michael Goldbaum, Adriana Tremoulet, Chun-Nan Hsu, Hannah Carter, Long Zhu, Kang Zhang, and Huimin Xia. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature Medicine*, 25(3):433–438, 2019.

[180] Michelle M. Clark, Amber Hildreth, Sergey Batalov, Yan Ding, Shimul Chowdhury, Kelly Watkins, Katarzyna Ellsworth, Brandon Camp, Cyrielle I. Kint, Calum Yacoubian, Lauge Farnaes, Matthew N. Bainbridge, Curtis Beebe, Joshua J. A. Braun, Margaret Bray, Jeanne Carroll, Julie A. Cakici, Sara A. Caylor, Christina Clarke,

Mitchell P. Creed, Jennifer Friedman, Alison Frith, Richard Gain, Mary Gaughran, Shauna George, Sheldon Gilmer, Joseph Gleeson, Jeremy Gore, Haiying Grunenwald, Raymond L. Hovey, Marie L. Janes, Kejia Lin, Paul D. McDonagh, Kyle McBride, Patrick Mulrooney, Shareef Nahas, Daeheon Oh, Albert Oriol, Laura Puckett, Zia Rady, Martin G. Reese, Julie Ryu, Lisa Salz, Erica Sanford, Lawrence Stewart, Nathaly Sweeney, Mari Tokita, Luca Van Der Kraan, Sarah White, Kristen Wigby, Brett Williams, Terence Wong, Meredith S. Wright, Catherine Yamada, Peter Schols, John Reynders, Kevin Hall, David Dimmock, Narayanan Veeraraghavan, Thomas Defay, and Stephen F. Kingsmore. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Science Translational Medicine*, 11(489), 2019.

[181] Kristin M Corey, Sehj Kashyap, Elizabeth Lorenzi, Sandhya A Lagoo-Deenadayalan, Katherine Heller, Krista Whalen, Suresh Balu, Mitchell T Heflin, Shelley R McDonald, Madhav Swaminathan, and Mark Sendak. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): A retrospective, single-site study. *PLOS Medicine*, 15(11):e1002701–e1002701, November 2018.

[182] Brian L. Hill, Robert Brown, Eilon Gabel, Nadav Rakocz, Christine Lee, Maxime Cannesson, Pierre Baldi, Loes Olde Loohuis, Ruth Johnson, Brandon Jew, Uri Maoz, Aman Mahajan, Sriram Sankararaman, Ira Hofer, and Eran Halperin. An automated machine learning-based model predicts postoperative mortality using readily-extractable preoperative electronic health record data. *British Journal of Anaesthesia*, 123(6):877–886, December 2019. Publisher: Elsevier.

[183] Amirata Ghorbani, David Ouyang, Abubakar Abid, Bryan He, Jonathan H. Chen, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. Deep learning interpretation of echocardiograms. *npj Digital Medicine*, 3(1):1–10, January 2020. Number: 1 Publisher: Nature Publishing Group.

[184] Antônio H. Ribeiro, Manoel Horta Ribeiro, Gabriela M. M. Paixão, Derick M. Oliveira, Paulo R. Gomes, Jéssica A. Canazart, Milton P. S. Ferreira, Carl R. Andersson, Peter W. Macfarlane, Wagner Meira Jr, Thomas B. Schön, and Antonio Luiz P. Ribeiro. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11(1):1760, April 2020. Number: 1 Publisher: Nature Publishing Group.

[185] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1):65–69, 2019.

[186] Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 2018.

[187] Paige Maas, Myrto Barrdahl, Amit D. Joshi, Paul L. Auer, Mia M. Gaudet, Roger L. Milne, Fredrick R. Schumacher, William F. Anderson, David Check, Subham Chattopadhyay, Laura Baglietto, Christine D. Berg, Stephen J. Chanock, David G. Cox, Jonine D. Figueroa, Mitchell H. Gail, Barry I. Graubard, Christopher A. Haiman, Susan E. Hankinson, Robert N. Hoover, Claudine Isaacs, Laurence N. Kolonel, Loic Le Marchand, I-Min Lee, Sara Lindström, Kim Overvad, Isabelle Romieu, Maria-Jose Sanchez, Melissa C. Southey, Daniel O. Stram, Rosario Tumino, Tyler J. VanderWeele, Walter C. Willett, Shumin Zhang, Julie E. Buring, Federico Canzian, Susan M. Gapstur, Brian E. Henderson, David J. Hunter, Graham G Giles, Ross L. Prentice, Regina G. Ziegler, Peter Kraft, Montse Garcia-Closas, and Nilanjan Chatterjee. Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. *JAMA Oncology*, 2(10):1295–1302, 10 2016.

[188] Fredrick R Schumacher, Ali Amin Al Olama, Sonja I Berndt, Sara Benlloch, Mahbubl Ahmed, Edward J Saunders, Tokhir Dadaev, Daniel Leongamornlert, Ezequiel Anokian, Clara Cieza-Borrella, Chee Goh, Mark N Brook, Xin Sheng, Laura Fachal, Joe Dennis, Jonathan Tyrer, Kenneth Muir, Artitaya Lophatananon, Victoria L Stevens, Susan M Gapstur, Brian D Carter, Catherine M Tangen, Phyllis J Goodman, Ian M Jr Thompson, Jyotsna Batra, Suzanne Chambers, Leire Moya, Judith Clements, Lisa Horvath, Wayne Tilley, Gail P Risbridger, Henrik Gronberg, Markus Aly, Tobias Nordström, Paul Pharoah, Nora Pashayan, Johanna Schleutker, Teuvo L J Tammela, Csilla Sipeky, Anssi Auvinen, Demetrius Albanes, Stephanie Weinstein, Alicja Wolk, Niclas Håkansson, Catharine M L West, Alison M Dunning, Neil Burnet, Lorelei A Mucci, Edward Giovannucci, Gerald L Andriole, Olivier Cussenot, Géraldine Cancel-Tassin, Stella Koutros, Laura E Beane Freeman, Karina Dalsgaard Sorensen, Torben Falck Orntoft, Michael Borre, Lovise Maehle, Eli Marie Grindedal, David E Neal, Jenny L Donovan, Freddie C Hamdy, Richard M Martin, Ruth C Travis, Tim J Key, Robert J Hamilton, Neil E Fleshner, Antonio Finelli, Sue Ann Ingles, Mariana C Stern, Barry S Rosenstein, Sarah L Kerns, Harry Ostrer, Yong-Jie Lu, Hong-Wei Zhang, Ninghan Feng, Xueying Mao, Xin Guo, Guomin Wang, Zan Sun, Graham G Giles, Melissa C Southey, Robert J MacInnis, Liesel M FitzGerald, Adam S Kibel, Bettina F Drake, Ana Vega, Antonio Gómez-Caamaño, Robert Szulkin, Martin Eklund, Manolis Kogevinas, Javier Llorca, Gemma Castaño-Vinyals, Kathryn L Penney, Meir Stampfer, Jong Y Park, Thomas A Sellers, Hui-Yi Lin, Janet L Stanford, Cezary Cybulski, Dominika Wokolorczyk, Jan Lubinski, Elaine A Ostrander, Milan S Geybels, Børge G Nordestgaard, Sune F Nielsen, Maren Weischer, Rasmus Bisbjerg, Martin Andreas Røder, Peter Iversen, Hermann Brenner, Katarina Cuk, Bernd Holleczek, Christiane Maier, Manuel Luedeke, Thomas Schnoeller, Jeri Kim, Christopher J Logothetis, Esther M John, Manuel R Teixeira, Paula Paulo, Marta Cardoso, Susan L Neuhausen, Linda Steele, Yuan Chun Ding, Kim De Ruyck, Gert De Meerleer, Piet Ost, Azad Razack, Jasmine Lim, Soo-Hwang Teo, Daniel W Lin, Lisa F Newcomb, Davor Lessel, Marija Gamulin, Tomislav Kulis, Radka Kaneva, Nawaid Usmani, Sandeep Singhal, Chavdar Slavov, Vanio Mitev, Matthew Parliament, Frank Claessens, Steven Joniau, Thomas Van den Broeck, Samantha Larkin, Paul A Townsend, Claire Aukim-Hastie,

Manuela Gago-Dominguez, Jose Esteban Castelao, Maria Elena Martinez, Monique J Roobol, Guido Jenster, Ron H N van Schaik, Florence Menegaux, Thérèse Truong, Yves Akoli Koudou, Jianfeng Xu, Kay-Tee Khaw, Lisa Cannon-Albright, Hardev Pandha, Agnieszka Michael, Stephen N Thibodeau, Shannon K McDonnell, Daniel J Schaid, Sara Lindstrom, Constance Turman, Jing Ma, David J Hunter, Elio Riboli, Afshan Siddiq, Federico Canzian, Laurence N Kolonel, Loic Le Marchand, Robert N Hoover, Mitchell J Machiela, Zuxi Cui, Peter Kraft, Christopher I Amos, David V Conti, Douglas F Easton, Fredrik Wiklund, Stephen J Chanock, Brian E Henderson, Zsofia Kote-Jarai, Christopher A Haiman, and Rosalind A Eeles. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat Genet*, 50(7):928–936, Jul 2018.

[189] Anke Hüls and Darina Czamara. Methodological challenges in constructing dna methylation risk scores. *Epigenetics*, 15(1-2):1–11, Jan-Feb 2020.

[190] Elizabeth Hibler, Lei Huang, Jorge Andrade, and Bonnie Spring. Impact of a diet and activity health promotion intervention on regional patterns of dna methylation. *Clinical Epigenetics*, 11(1):133, 2019.

[191] Alexandra J White, Dale P Sandler, Sophia C E Bolick, Zongli Xu, Jack A Taylor, and Lisa A DeRoo. Recreational and household physical activity at different time points and dna global methylation. *Eur J Cancer*, 49(9):2199–2206, Jun 2013.

[192] Fang Fang Zhang, Alfredo Morabia, Joan Carroll, Karina Gonzalez, Kimberly Fulda, Manleen Kaur, Jamboor K. Vishwanatha, Regina M. Santella, and Roberto Cardarelli. Dietary Patterns Are Associated with Levels of Global Genomic DNA Methylation in a Cancer-Free Population. *The Journal of Nutrition*, 141(6):1165–1171, 04 2011.

[193] Katherine J Dick, Christopher P Nelson, Loukia Tsaprouni, Johanna K Sandling, Dylan Aïssi, Simone Wahl, Eshwar Meduri, Pierre-Emmanuel Morange, France Gagnon, Harald Grallert, Melanie Waldenberger, Annette Peters, Jeanette Erdmann, Christian Hengstenberg, Francois Cambien, Alison H Goodall, Willem H Ouwehand, Heribert Schunkert, John R Thompson, Tim D Spector, Christian Gieger, David-Alexandre Trégouët, Panos Deloukas, and Nilesh J Samani. Dna methylation and body-mass index: a genome-wide analysis. *The Lancet*, 383(9933):1990–1998, 2014.

[194] Victor V Levenson. Dna methylation as a universal biomarker. *Expert Review of Molecular Diagnostics*, 10(4):481–488, 05 2010.

[195] Katarzyna Kamińska, Ewelina Nalejska, Marta Kubiak, Joanna Wojtysiak, Łukasz Żołna, Janusz Kowalewski, and Marzena Anna Lewandowska. Prognostic and predictive epigenetic biomarkers in oncology. *Molecular Diagnosis & Therapy*, 23(1):83–95, 2019.

[196] Audrey Y. Chu, Adrienne Tin, Pascal Schlosser, Yi-An Ko, Chengxiang Qiu, Chen Yao, Roby Joehanes, Morgan E. Grams, Liming Liang, Caroline A. Gluck, Chunyu Liu, Josef Coresh, Shih-Jen Hwang, Daniel Levy, Eric Boerwinkle, James S. Pankow,

Qiong Yang, Myriam Fornage, Caroline S. Fox, Katalin Susztak, and Anna Köttgen. Epigenome-wide association studies identify dna methylation associated with kidney function. *Nature Communications*, 8(1):1286, 2017.

[197] Yun Liu, Martin J Aryee, Leonid Padyukov, M Daniele Fallin, Espen Hesselberg, Arni Runarsson, Lovisa Reinius, Nathalie Acevedo, Margaret Taub, Marcus Ronninger, Klementy Shchetynsky, Annika Scheynius, Juha Kere, Lars Alfredsson, Lars Klareskog, Tomas J Ekström, and Andrew P Feinberg. Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology*, 31:142 EP –, 01 2013.

[198] Vardhman K. Rakyan, Huriya Beyan, Thomas A. Down, Mohammed I. Hawa, Siarhei Maslau, Deeqo Aden, Antoine Daunay, Florence Busato, Charles A. Mein, Burkhard Manfras, Kerith-Rae M. Dias, Christopher G. Bell, Jörg Tost, Bernhard O. Boehm, Stephan Beck, and R. David Leslie. Identification of type 1 diabetes–associated dna methylation variable positions that precede disease diagnosis. *PLOS Genetics*, 7(9):1–9, 09 2011.

[199] Jimmy L Huynh, Paras Garg, Tin Htwe Thin, Seungyeul Yoo, Ranjan Dutta, Bruce D Trapp, Vahram Haroutunian, Jun Zhu, Michael J Donovan, Andrew J Sharp, and Patrizia Casaccia. Epigenome-wide differences in pathology-free regions of multiple sclerosis–affected brains. *Nature Neuroscience*, 17(1):121–130, 2014.

[200] Christopher G. Bell, Andrew E. Teschendorff, Vardhman K. Rakyan, Alexander P. Maxwell, Stephan Beck, and David A. Savage. Genome-wide DNA methylation analysis for diabetic nephropathy in type 1 diabetes mellitus. *BMC Medical Genomics*, 3(1):33, August 2010.

[201] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *Journal of machine learning research : JMLR*, 11:2287–2322, March 2010.

[202] Ira S Hofer, Eilon Gabel, Michael Pfeffer, Mohammed Mahbouba, and Aman Mahajan. A Systematic Approach to Creation of a Perioperative Data Warehouse. *Anesthesia & Analgesia*, 122(6), 2016.

[203] Ruth Johnson, Yi Ding, Vidhya Venkateswaran, Arjun Bhattacharya, Alec Chiu, Tommer Schwarz, Malika Freund, Lingyu Zhan, Kathryn S. Burch, Christa Caggiano, Brian Hill, Nadav Rakocz, Brunilda Balliu, Jae Hoon Sul, Noah Zaitlen, Valerie A. Arboleda, Eran Halperin, Sriram Sankararaman, Manish J. Butte, UCLA Precision Health Data Discovery Repository Working Group, UCLA Precision Health ATLAS Working Group, Clara Lajonchere, Daniel H. Geschwind, and Bogdan Pasaniuc. Leveraging genomic diversity for discovery in an ehr-linked biobank: the ucla atlas community health initiative. *medRxiv*, 2021.

[204] Joshua C. Denny, Marylyn D. Ritchie, Melissa A. Basford, Jill M. Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R. Masys, Dan M. Roden, and

Dana C. Crawford. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210, May 2010. Publisher: Oxford Academic.

[205] Joshua C. Denny, Lisa Bastarache, Marylyn D. Ritchie, Robert J. Carroll, Raquel Zink, Jonathan D. Mosley, Julie R. Field, Jill M. Pulley, Andrea H. Ramirez, Erica Bowton, Melissa A. Basford, David S. Carrell, Peggy L. Peissig, Abel N. Kho, Jennifer A. Pacheco, Luke V. Rasmussen, David R. Crosslin, Paul K. Crane, Jyotishman Pathak, Suzette J. Bielinski, Sarah A. Pendergrass, Hua Xu, Lucia A. Hindorff, Rongling Li, Teri A. Manolio, Christopher G. Chute, Rex L. Chisholm, Eric B. Larson, Gail P. Jarvik, Murray H. Brilliant, Catherine A. McCarty, Iftikhar J. Kullo, Jonathan L. Haines, Dana C. Crawford, Daniel R. Masys, and Dan M. Roden. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, 31(12):1102–1111, December 2013. Number: 12 Publisher: Nature Publishing Group.

[206] Sharon R. Browning and Brian L. Browning. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics*, 81(5):1084–1097, November 2007. Publisher: Elsevier.

[207] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics*, 81(3):559–575, September 2007.

[208] Jian Yang, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. GCTA: A Tool for Genome-wide Complex Trait Analysis. *American Journal of Human Genetics*, 88(1):76–82, January 2011.

[209] Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, Emily Y Chew, Shawn Levy, Matt McGue, David Schlessinger, Dwight Stambolian, Po-Ru Loh, William G Iacono, Anand Swaroop, Laura J Scott, Francesco Cucca, Florian Kronenberg, Michael Boehnke, Gonçalo R Abecasis, and Christian Fuchsberger. Next-generation genotype imputation service and methods. *Nat Genet*, 48(10):1284–1287, Oct 2016.

[210] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R Abecasis, Richard Durbin, and Alkes L Price. Reference-based phasing using the haplotype reference consortium panel. *Nature Genetics*, 48(11):1443–1448, 2016.

[211] Daniel Taliun, Daniel N. Harris, Michael D. Kessler, Jedidiah Carlson, Zachary A. Szpiech, Raul Torres, Sarah A. Gagliano Taliun, André Corvelo, Stephanie M. Gogarten,

Hyun Min Kang, Achilleas N. Pitsillides, Jonathon LeFaive, Seung-been Lee, Xiaowen Tian, Brian L. Browning, Sayantan Das, Anne-Katrin Emde, Wayne E. Clarke, Douglas P. Loesch, Amol C. Shetty, Thomas W. Blackwell, Albert V. Smith, Quenna Wong, Xiaoming Liu, Matthew P. Conomos, Dean M. Bobo, François Aguet, Christine Albert, Alvaro Alonso, Kristin G. Ardlie, Dan E. Arking, Stella Aslibekyan, Paul L. Auer, John Barnard, R. Graham Barr, Lucas Barwick, Lewis C. Becker, Rebecca L. Beer, Emelia J. Benjamin, Lawrence F. Bielak, John Blangero, Michael Boehnke, Donald W. Bowden, Jennifer A. Brody, Esteban G. Burchard, Brian E. Cade, James F. Casella, Brandon Chalazan, Daniel I. Chasman, Yii-Der Ida Chen, Michael H. Cho, Seung Hoan Choi, Mina K. Chung, Clary B. Clish, Adolfo Correa, Joanne E. Curran, Brian Custer, Dawood Darbar, Michelle Daya, Mariza de Andrade, Dawn L. DeMeo, Susan K. Dutcher, Patrick T. Ellinor, Leslie S. Emery, Celeste Eng, Diane Fatkin, Tasha Fingerlin, Lukas Forer, Myriam Fornage, Nora Franceschini, Christian Fuchsberger, Stephanie M. Fullerton, Soren Germer, Mark T. Gladwin, Daniel J. Gottlieb, Xiuqing Guo, Michael E. Hall, Jiang He, Nancy L. Heard-Costa, Susan R. Heckbert, Marguerite R. Irvin, Jill M. Johnsen, Andrew D. Johnson, Robert Kaplan, Sharon L. R. Kardia, Tanika Kelly, Shannon Kelly, Eimear E. Kenny, Douglas P. Kiel, Robert Klemmer, Barbara A. Konkle, Charles Kooperberg, Anna Köttgen, Leslie A. Lange, Jessica Lasky-Su, Daniel Levy, Xihong Lin, Keng-Han Lin, Chunyu Liu, Ruth J. F. Loos, Lori Garman, Robert Gerszten, Steven A. Lubitz, Kathryn L. Lunetta, Angel C. Y. Mak, Ani Manichaikul, Alisa K. Manning, Rasika A. Mathias, David D. McManus, Stephen T. McGarvey, James B. Meigs, Deborah A. Meyers, Julie L. Mikulla, Mollie A. Minear, Braxton D. Mitchell, Sanghamitra Mohanty, May E. Montasser, Courtney Montgomery, Alanna C. Morrison, Joanne M. Murabito, Andrea Natale, Pradeep Natarajan, Sarah C. Nelson, Kari E. North, Jeffrey R. O'Connell, Nicholette D. Palmer, Nathan Pankratz, Gina M. Peloso, Patricia A. Peyser, Jacob Pleiness, Wendy S. Post, Bruce M. Psaty, D. C. Rao, Susan Redline, Alexander P. Reiner, Dan Roden, Jerome I. Rotter, Ingo Ruczinski, Chloé Sarnowski, Sebastian Schoenherr, David A. Schwartz, Jeong-Sun Seo, Sudha Seshadri, Vivien A. Sheehan, Wayne H. Sheu, M. Benjamin Shoemaker, Nicholas L. Smith, Jennifer A. Smith, Nona Sotoodehnia, Adrienne M. Stilp, Weihong Tang, Kent D. Taylor, Marilyn Telen, Timothy A. Thornton, Russell P. Tracy, David J. Van Den Berg, Ramachandran S. Vasan, Karine A. Viaud-Martinez, Scott Vrieze, Daniel E. Weeks, Bruce S. Weir, Scott T. Weiss, Lu-Chen Weng, Cristen J. Willer, Yingze Zhang, Xutong Zhao, Donna K. Arnett, Allison E. Ashley-Koch, Kathleen C. Barnes, Eric Boerwinkle, Stacey Gabriel, Richard Gibbs, Kenneth M. Rice, Stephen S. Rich, Edwin K. Silverman, Pankaj Qasba, Weiniu Gan, Namiko Abe, Laura Almasy, Seth Ament, Peter Anderson, Pramod Anugu, Deborah Applebaum-Bowden, Tim Assimes, Dimitrios Avramopoulos, Emily Barron-Casella, Terri Beaty, Gerald Beck, Diane Becker, Amber Beitelshees, Takis Benos, Marcos Bezerra, Joshua Bis, Russell Bowler, Ulrich Broeckel, Jai Broome, Karen Bunting, Carlos Bustamante, Erin Buth, Jonathan Cardwell, Vincent Carey, Cara Carty, Richard Casaburi, Peter Castaldi, Mark Chaffin, Christy Chang, Yi-Cheng Chang, Sameer Chavan, Bo-Juen Chen, Wei-Min Chen, Lee-Ming Chuang, Ren-Hua Chung, Suzy Comhair, Elaine Cornell, Carolyn Crandall, James Crapo, Jeffrey Curtis, Coleen Dam-

cott, Sean David, Colleen Davis, Lisa de las Fuentes, Michael DeBaun, Ranjan Deka, Scott Devine, Qing Duan, Ravi Duggirala, Jon Peter Durda, Charles Eaton, Lynette Ekunwe, Adel El Boueiz, Serpil Erzurum, Charles Farber, Matthew Flickinger, Chris Frazar, Mao Fu, Lucinda Fulton, Shanshan Gao, Yan Gao, Margery Gass, Bruce Gelb, Xiaoqi Priscilla Geng, Mark Geraci, Auyon Ghosh, Chris Gignoux, David Glahn, Da-Wei Gong, Harald Goring, Sharon Graw, Daniel Grine, C. Charles Gu, Yue Guan, Namrata Gupta, Jeff Haessler, Nicola L. Hawley, Ben Heavner, David Herrington, Craig Hersh, Bertha Hidalgo, James Hixson, Brian Hobbs, John Hokanson, Elliott Hong, Karin Hoth, Chao Agnes Hsiung, Yi-Jen Hung, Haley Huston, Chii Min Hwu, Rebecca Jackson, Deepti Jain, Min A. Jhun, Craig Johnson, Rich Johnston, Kimberly Jones, Sekar Kathiresan, Alyna Khan, Wonji Kim, Greg Kinney, Holly Kramer, Christoph Lange, Ethan Lange, Leslie Lange, Cecelia Laurie, Meryl LeBoff, Jiwon Lee, Seunggeun Shawn Lee, Wen-Jane Lee, David Levine, Joshua Lewis, Xiaohui Li, Yun Li, Henry Lin, Honghuang Lin, Keng Han Lin, Simin Liu, Yongmei Liu, Yu Liu, James Luo, Michael Mahaney, and NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium. Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *Nature*, 590(7845):290–299, 2021.

[212] Christian Fuchsberger, Gonçalo R Abecasis, and David A Hinds. minimac2: faster genotype imputation. *Bioinformatics*, 31(5):782–784, Mar 2015.

[213] Zongli Xu, Liang Niu, Leping Li, and Jack A. Taylor. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Research*, 44(3):e20, February 2016.

[214] Eugene Andres Houseman, William P. Accomando, Devin C. Koestler, Brock C. Christensen, Carmen J. Marsit, Heather H. Nelson, John K. Wiencke, and Karl T. Kelsey. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13(1):86, May 2012.

[215] Alex Rubinsteyn, Sergey Feldman, Tim O'Donnell, and Brett Beaulieu-Jones. hammerlab/fancyimpute: Version 0.2.0. September 2017.

[216] Yosuke Tanigawa, Junyang Qian, Guhan Venkataraman, Johanne Marie Justesen, Ruilin Li, Robert Tibshirani, Trevor Hastie, and Manuel A. Rivas. Significant sparse polygenic risk scores across 813 traits in uk biobank. *medRxiv*, 2021.

[217] Nasa Sinnott-Armstrong, Yosuke Tanigawa, David Amar, Nina Mars, Christian Benner, Matthew Aguirre, Guhan Ram Venkataraman, Michael Wainberg, Hanna M. Ollila, Tuomo Kiiskinen, Aki S. Havulinna, James P. Pirruccello, Junyang Qian, Anna Shcherbina, Fatima Rodriguez, Themistocles L. Assimes, Vineeta Agarwala, Robert Tibshirani, Trevor Hastie, Samuli Ripatti, Jonathan K. Pritchard, Mark J. Daly, Manuel A. Rivas, and FinnGen. Genetics of 35 blood and urine biomarkers in the uk biobank. *Nature Genetics*, 53(2):185–194, 2021.

[218] Gad Abraham, Adam Kowalczyk, Justin Zobel, and Michael Inouye. Sparsnp: fast and memory-efficient analysis of all snps for phenotype prediction. *BMC Bioinformatics*, 13:88, May 2012.

[219] Bjarni J Vilhjálmsson, Jian Yang, Hilary K Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, Tristan Hayeck, Hong-Hee Won, Biology Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Risk of Inherited Variants in Breast Cancer (DRIVE) study, Sekar Kathiresan, Michele Pato, Carlos Pato, Rulla Tamimi, Eli Stahl, Noah Zaitlen, Bogdan Pasaniuc, Gillian Belbin, Eimear E Kenny, Mikkel H Schierup, Philip De Jager, Nikolaos A Patsopoulos, Steve McCarroll, Mark Daly, Shaun Purcell, Daniel Chasman, Benjamin Neale, Michael Goddard, Peter M Visscher, Peter Kraft, Nick Patterson, and Alkes L Price. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *American journal of human genetics*, 97(4):576–592, 10 2015.

[220] Tian Ge, Chia-Yen Chen, Yang Ni, Yen-Chen Anne Feng, and Jordan W. Smoller. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nature Communications*, 10(1):1776, 2019.

[221] Samuel A. Lambert, Laurent Gil, Simon Jupp, Scott C. Ritchie, Yu Xu, Annalisa Buniello, Aoife McMahon, Gad Abraham, Michael Chapman, Helen Parkinson, John Danesh, Jacqueline A. L. MacArthur, and Michael Inouye. The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nature Genetics*, 53(4):420–425, 2021.

[222] Frank Dudbridge. Power and Predictive Accuracy of Polygenic Risk Scores. *PLOS Genetics*, 9(3):e1003348, March 2013. Publisher: Public Library of Science.

[223] Ernest Beutler and Carol West. Hematologic differences between african-americans and whites: the roles of iron deficiency and alpha-thalassemia on hemoglobin levels and mean corpuscular volume. *Blood*, 106(2):740–745, Jul 2005.

[224] Eunjung Lim, Jill Miyamura, and John J Chen. Racial/ethnic-specific reference intervals for common laboratory tests: A comparison among asians, blacks, hispanics, and white. *Hawaii J Med Public Health*, 74(9):302–310, Sep 2015.

[225] Steve Horvath. Dna methylation age of human tissues and cell types. *Genome Biology*, 14(10):3156, 2013.

[226] Paula Singmann, Doron Shem-Tov, Simone Wahl, Harald Grallert, Giovanni Fiorito, So-Youn Shin, Katharina Schramm, Petra Wolf, Sonja Kunze, Yael Baran, Simonetta Guarrera, Paolo Vineis, Vittorio Krogh, Salvatore Panico, Rosario Tumino, Anja Kretschmer, Christian Gieger, Annette Peters, Holger Prokisch, Caroline L. Relton, Giuseppe Matullo, Thomas Illig, Melanie Waldenberger, and Eran Halperin. Characterization of whole-genome autosomal differences of dna methylation between men and women. *Epigenetics & Chromatin*, 8(1):43, Oct 2015.

[227] Daniel Trejo Banos, Daniel L. McCartney, Marion Patxot, Lucas Anchieri, Thomas Battram, Colette Christiansen, Ricardo Costeira, Rosie M. Walker, Stewart W. Morris, Archie Campbell, Qian Zhang, David J. Porteous, Allan F. McRae, Naomi R. Wray, Peter M. Visscher, Chris S. Haley, Kathryn L. Evans, Ian J. Deary, Andrew M. McIntosh, Gibran Hemani, Jordana T. Bell, Riccardo E. Marioni, and Matthew R. Robinson. Bayesian reassessment of the epigenetic architecture of complex traits. *Nature Communications*, 11(1):2865, 2020.

[228] Brett K. Beaulieu-Jones, Daniel R. Lavage, John W. Snyder, Jason H. Moore, Sarah A. Pendergrass, and Christopher R. Bauer. Characterizing and Managing Missing Structured Data in Electronic Health Records: Data Analysis. *JMIR Medical Informatics*, 6(1):e8960, February 2018. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.

[229] Dragana Vuckovic, Erik L. Bao, Parsa Akbari, Caleb A. Lareau, Abdou Mousas, Tao Jiang, Ming-Huei Chen, Laura M. Raffield, Manuel Tardaguila, Jennifer E. Huffman, Scott C. Ritchie, Karyn Megy, Hannes Ponstingl, Christopher J. Penkett, Patrick K. Albers, Emilie M. Wigdor, Saori Sakaue, Arden Moscati, Regina Manansala, Ken Sin Lo, Huijun Qian, Masato Akiyama, Traci M. Bartz, Yoav Ben-Shlomo, Andrew Beswick, Jette Bork-Jensen, Erwin P. Bottinger, Jennifer A. Brody, Frank J.A. van Rooij, Kumaraswamy N. Chitrala, Peter W.F. Wilson, Hélène Choquet, John Danesh, Emanuele Di Angelantonio, Niki Dimou, Jingzhong Ding, Paul Elliott, Tõnu Esko, Michele K. Evans, Stephan B. Felix, James S. Floyd, Linda Broer, Niels Grarup, Michael H. Guo, Qi Guo, Andreas Greinacher, Jeff Haessler, Torben Hansen, Joanna M.M. Howson, Wei Huang, Eric Jorgenson, Tim Kacprowski, Mika Kähönen, Yoichiro Kamatani, Masahiro Kanai, Savita Karthikeyan, Fotios Koskeridis, Leslie A. Lange, Terho Lehtimäki, Allan Linneberg, Yongmei Liu, Leo-Pekka Lyytikäinen, Ani Manichaikul, Koichi Matsuda, Karen L. Mohlke, Nina Mononen, Yoshinori Murakami, Girish N. Nadkarni, Kjell Nikus, Nathan Pankratz, Oluf Pedersen, Michael Preuss, Bruce M. Psaty, Olli T. Raitakari, Stephen S. Rich, Benjamin A.T. Rodriguez, Jonathan D. Rosen, Jerome I. Rotter, Petra Schubert, Cassandra N. Spracklen, Praveen Surendran, Hua Tang, Jean-Claude Tardif, Mohsen Ghanbari, Uwe Völker, Henry Völzke, Nicholas A. Watkins, Stefan Weiss, Na Cai, Kousik Kundu, Stephen B. Watt, Klaudia Walter, Alan B. Zonderman, Kelly Cho, Yun Li, Ruth J.F. Loos, Julian C. Knight, Michel Georges, Oliver Stegle, Evangelos Evangelou, Yukinori Okada, David J. Roberts, Michael Inouye, Andrew D. Johnson, Paul L. Auer, William J. Astle, Alexander P. Reiner, Adam S. Butterworth, Willem H. Ouwehand, Guillaume Lettre, Vijay G. Sankaran, and Nicole Soranzo. The polygenic and monogenic basis of blood traits and diseases. *Cell*, 182(5):1214–1231.e11, 2020.

[230] Sini Kerminen, Alicia R. Martin, Jukka Koskela, Sanni E. Ruotsalainen, Aki S. Havulinna, Ida Surakka, Aarno Palotie, Markus Perola, Veikko Salomaa, Mark J. Daly, Samuli Ripatti, and Matti Pirinen. Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland. *The American Journal of Human*

Genetics, 104(6):1169–1181, June 2019.

[231] Elior Rahmani, Liat Shenhav, Regev Schweiger, Paul Yousefi, Karen Huen, Brenda Eskenazi, Celeste Eng, Scott Huntsman, Donglei Hu, Joshua Galanter, Sam S. Oh, Melanie Waldenberger, Konstantin Strauch, Harald Grallert, Thomas Meitinger, Christian Gieger, Nina Holland, Esteban G. Burchard, Noah Zaitlen, and Eran Halperin. Genome-wide methylation data mirror ancestry information. *Epigenetics & Chromatin*, 10(1):1, January 2017.

[232] Richard T. Barfield, Lynn M. Almli, Varun Kilaru, Alicia K. Smith, Kristina B. Mercer, Richard Duncan, Torsten Klengel, Divya Mehta, Elisabeth B. Binder, Michael P. Epstein, Kerry J. Ressler, and Karen N. Conneely. Accounting for Population Stratification in DNA Methylation Studies. *Genetic Epidemiology*, 38(3):231–241, 2014. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gepi.21789.

[233] Erika L. Moen, Xu Zhang, Wenbo Mu, Shannon M. Delaney, Claudia Wing, Jennifer McQuade, Jamie Myers, Lucy A. Godley, M. Eileen Dolan, and Wei Zhang. Genome-Wide Variation of Cytosine Modifications Between European and African Populations and the Implications for Complex Traits. *Genetics*, 194(4):987–996, August 2013. Publisher: Genetics Section: Investigations.

[234] Christopher G Bell, Andrew E Teschendorff, Vardhman K Rakyan, Alexander P Maxwell, Stephan Beck, and David A Savage. Genome-wide dna methylation analysis for diabetic nephropathy in type 1 diabetes mellitus. *BMC Med Genomics*, 3:33, Aug 2010.

[235] Eilis Hannon, Emma Dempster, Joana Viana, Joe Burrage, Adam R Smith, Ruby Macdonald, David St Clair, Colette Mustard, Gerome Breen, Sebastian Therman, Jaakko Kaprio, Timothea Toulopoulou, Hilleke E Hulshoff Pol, Marc M Bohlken, Rene S Kahn, Igor Nenadic, Christina M Hultman, Robin M Murray, David A Collier, Nick Bass, Hugh Gurling, Andrew McQuillin, Leonard Schalkwyk, and Jonathan Mill. An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential dna methylation. *Genome Biol*, 17(1):176, Aug 2016.

[236] Yun Liu, Martin J Aryee, Leonid Padyukov, M Daniele Fallin, Espen Hesselberg, Arni Runarsson, Lovisa Reinius, Nathalie Acevedo, Margaret Taub, Marcus Ronninger, Klementy Shchetynsky, Annika Scheynius, Juha Kere, Lars Alfredsson, Lars Klareskog, Tomas J Ekström, and Andrew P Feinberg. Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol*, 31(2):142–147, Feb 2013.

[237] Daniel W Belsky. Translating polygenic analysis for prevention: From who to how. *Circulation. Cardiovascular genetics*, 10(3):e001798, 06 2017.

[238] Ali Jazayeri, Ou Stella Liang, and Christopher C. Yang. Imputation of missing data in electronic health records based on patients'similarities. *Journal of Healthcare Informatics Research*, 4(3):295–307, 2020.

[239] Riccardo Miotto, Li Li, Brian A. Kidd, and Joel T. Dudley. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6(1):26094, 2016.

[240] Tao Zheng, Wei Xie, Liling Xu, Xiaoying He, Ya Zhang, Mingrong You, Gong Yang, and You Chen. A machine learning-based framework to identify type 2 diabetes through electronic health records. *International Journal of Medical Informatics*, 97:120–127, 2017.

[241] Elior Rahmani, Noah Zaitlen, Yael Baran, Celeste Eng, Donglei Hu, Joshua Galanter, Sam Oh, Esteban G Burchard, Eleazar Eskin, James Zou, and Eran Halperin. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nature Methods*, 13:443, March 2016. Publisher: Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.

[242] Steve Horvath. DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10):3156, December 2013.

[243] Kara N Fitzgerald, Romilly Hodges, Douglas Hanes, Emily Stack, David Cheishvili, Moshe Szyf, Janine Henkel, Melissa W Twedt, Despina Giannopoulou, Josette Herdell, Sally Logan, and Ryan Bradley. Potential reversal of epigenetic age using a diet and lifestyle intervention: a pilot randomized clinical trial. *Aging*, 13(7):9419–9432, 04 2021.

[244] Jiajia Li, Gregory R. Grant, John B. Hogenesch, and Michael E. Hughes. Chapter sixteen - considerations for rna-seq analysis of circadian rhythms. In Amita Sehgal, editor, *Circadian Rhythms and Biological Clocks, Part A*, volume 551 of *Methods in Enzymology*, pages 349–367. Academic Press, 2015.

[245] Alexessander Couto Alves, Craig A. Glastonbury, Julia S. El-Sayed Moustafa, and Kerrin S. Small. Fasting and time of day independently modulate circadian rhythm relevant gene expression in adipose and skin tissue. *BMC Genomics*, 19(1):659, 2018.

[246] Romanas Chaleckis, Itsuo Murakami, Junko Takada, Hiroshi Kondoh, and Mitsuhiro Yanagida. Individual variability in human blood metabolites identifies age-related differences. *Proceedings of the National Academy of Sciences*, 113(16):4252–4259, 2016.

[247] Gad Asher and Paolo Sassone-Corsi. Time for food: the intimate interplay between nutrition, metabolism, and the circadian clock. *Cell*, 161(1):84–92, Mar 2015.

[248] Caroline L Relton and George Davey Smith. Two-step epigenetic mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *Int J Epidemiol*, 41(1):161–176, Feb 2012.