UNIVERSITY OF CALIFORNIA

Los Angeles

Bioinformatic Strategies for Population Precision Health

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

by

Christa Caggiano

2023

ABSTRACT OF THE DISSERTATION

Bioinformatic Strategies for Population Precision Health

by

Christa Caggiano
Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2023
Professor Noah A. Zaitlen, Chair

Population precision health represents a paradigm shift in healthcare, emphasizing the need for tailored and personalized approaches to improve health outcomes at a population level. Population precision health recognizes the heterogeneity within populations and leverages advances in genomics, epigenomics, and clinical data repositories to deliver targeted interventions and preventive strategies. By integrating genomic and clinical data, population precision health aims to identify individuals at increased risk for specific diseases and tailor interventions based on their unique genetic and environmental profiles. In this work, I present strategies to address three key challenges of implementing population precision health. I develop algorithms to non-invasively detect tissue death, which can be used for disease diagnosis and prevention. I then use these algorithms as the foundation of a scalable cell-free DNA platform to monitor disease at the population level. Lastly, I employ machine learning algorithms in a large genetic biobank to identify population-specific genetic and health risks. Together, this work represents a step toward implementing non-invasive disease screening and monitoring in diverse groups, which will be a crucial element of deploying population precision health.

The dissertation of Christa Caggiano is approved.

Matteo Pellegrini

Bogdan Pasaniuc

Paivi Elisabeth Pajukanta

Noah A. Zaitlen, Committee Chair

University of California, Los Angeles

2023

For my brother, Michael, and all those suffering from genetic diseases.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Firstly, I would like to thank my PhD advisor, Noah Zaitlen. I started my PhD with Noah six years ago at UCSF, and moved to Los Angeles when he moved his lab to UCLA. The choice was a difficult one, but one that I am so glad I made. Noah has been my biggest supporter and believed in me when I was ready to quit. He has pushed me to pursue new challenges and to become a better, more precise, scientist. Most of all, Noah has irrevocably shaped my scientific career, giving me the freedom to explore what I was interested in and follow new research directions. I admire Noah so much and am extremely grateful to have worked together during my PhD, and hopefully, for many years to come.

I would also like to thank my many wonderful collaborators that have made science fun and worth doing. Andy Dahl has been an amazing mentor, friend, and collaborator and helped me get through numerous crises. Andy has given me some of the best advice and without him, I am not sure I would have made it through my first paper. Barbara Celona is an amazing experimental scientist, and I am in awe of her technical skills and her perseverance. She has been a key contributor to a number of projects during my PhD, and we would not have the beautiful data we have without her. Marco Morselli, was so patient, willing, and open to collaborating with computational scientists who often had no idea what they were doing. The cfDNA capture project would be nowhere without him. Fleur Garton has been a friend and mentor since the beginning of my PhD, and over the past six years, we have discussed and dreamed of ALS biomarkers across time zones and continents. She has taught me so much about ALS biology. Lastly, I would like to thank Gillian Belbin. She is the most fun and kind person to work with, and she taught me, quite literally, everything I know about identity-by-descent. Our paper would be much worse without her.

Over the past 6 years I have made so many friends from both my PhD programs at UCSF, UCLA, and elsewhere. Snow Zun, Kodi Taraszka, Ella Petter, Mike Thompson, and Roohy Shemirani have all been incredible people to discuss science and life with.

I would also like to thank my thesis committee, Bogdan Pasaniuc, Matteo Pellegrini, and

Paivi Pajukanta. I have had many fruitful conversations about science with them, and it has shaped the direction of this work tremendously. I would also like to thank Bruna Balliu and Valerie Arboleda who have both also been incredible role models for me and have taught me a lot about science.

My family has been pivotal in every phase of my scientific career, and I could not have done it without them. They encouraged me to pursue science and always encouraged me. From shuttling me back and forth to my first science job in high school to reading and editing my writing, they have always been a huge support system for which I am forever grateful.

I would like to thank my partner, Arya Boudaie. Arya has endlessly supported me. He is the reason that I am pursuing a computational biology degree. On a personal level, he has moved cross country to support my scientific dreams, (and is doing it again for my postdoc). Throughout my PhD, he has been my emotional rock, and I am confident I could not have done it without his constant reassurance and guidance. He's learned so much about genetics just to understand my work, and in turn, has taught me so much about computer science and software engineering. With my most recent identity-by-descent project, he has even become a collaborator and co-author, and it has been such a joy to experience science and life with him.

Lastly, and most importantly, I would like to thank the patients who contributed their DNA to make this work possible. They have donated the most deeply personal part of themselves, their unique genetic code, for the hope of a better world. For many, their genomes that live on my computer are the last thing that remains of them. I am so grateful for their contribution to science in general, but especially to my science, and aspire to live up to their expectations.

VITA

2012 – 2014	Research Assistant, Laboratory for Laser Energetics, Rochester, New York, USA.
2013 – 2017	B.S., Biological Physics and Art History, Brandeis University, Waltham, Massachusetts, USA.
2014 – 2017	Student Scholar, Women's Studies Research Center Brandeis University, Waltham, Massachusetts, USA.
2017 – 2017	Bioinformatics Intern, Asuragen, Austin, Texas, USA
2017 – 2023	Graduate Research Assistant, University of California, Los Angeles, California, USA

PUBLICATIONS

C. Caggiano*, M. Morselli*, B. Celona, X. Qian *et al.* Non-invasive cell-free DNA biomarker discovery in amyotrophic lateral sclerosis. *in preparation* (2023) (* Authors contributed equally

C. Caggiano, A. Boudaie, R. Shemirani, J. Mefford, E., *et al.* Disease risk and healthcare utilization among ancestrally-diverse groups in the Los Angeles region. *Nature Medicine*, *in press* (2023)

R. Johnson, Y. Ding, V. Venkateswaran, A. Bhattacharya, K. Boulier, A. Chiu, S. Knyazev, T. Schwarz, M. Freund, L. Zhan, K. S. Burch, **C. Caggiano**, *et al.* Leveraging genomic diversity for discovery in an electronic health record linked biobank: the UCLA ATLAS Community Health Initiative. *Genome Medicine*, **14**, 1 (2022).

R. Van Paemel, A. De Koker, **C. Caggiano**, A. Morlion, *et al.* Genome-wide study of the effect of blood collection tubes on the cell-free DNA methylome. *Epigenetics*, **16**, 7 (2021).

C. Caggiano, B. Celona, F. Garton, J. Mefford,, *et al.* Comprehensive cell type decomposition of circulating cell-free DNA with CelFiE. *Nature Communications*, **12**, 1 (2021).

R. W. Newberry, T. Arhar, J. Costello, G. C. Hartoularos, *et al* [including **C. Caggiano**]. Robust Sequence Determinants of -Synuclein Toxicity in Yeast Implicate Membrane Binding. *ACS Chemical Biology*, **15**, 8 (2020).

CHAPTER 1

Introduction

All individuals have their own distinct genetic and environmental histories, that throughout their lives, will interact to influence their risk for disease. Consequently, a single treatment or care plan may only be suited to some individuals. For example, patients with breast cancer that have a mutation in the BRCA gene respond to a drug called a PARP inhibitor [172]. Breast cancer patients with a mutation in the HER2 gene, on the other hand, do not

To characterize an individual's risk for disease, clinicians and researchers need to capture multi-modal information about that individual. This includes bio-molecular measurements, such as genomic and metabolomic tests, along with data on that individual's family history, cultural background, and lifestyle [5]. Innovations in the 21st century have facilitated obtaining this data. For example, since the completion of the Human Genome Project in 2003 [66], the cost of sequencing a genome has dramatically decreased, making it increasingly feasible to capture the entire genome of an individual. Furthermore, with the widespread adoption of electronic health records (EHRs), healthcare data is rapidly expanding. EHRs can be used to learn about patient longitudinal trajectories, be linked to pedigrees to learn how disease risk is transmitted in families, or be used to identify clusters of patients with similar diseases.

However, there are still significant challenges in the implementation of precision medicine. To meet the threshold for integration into clinical care, precision treatments must have robust evidence. While some scenarios exist for which this threshold is met, many diseases still need more investigation into their genetic and environmental causes. Medicine has long been biased toward studying European ancestry populations, which can limit this study in several ways. Historically, people tended to have children with those in a similar population or geographic location, meaning that if genetic mutations arose in a population, those mutations may become more common in that population. Cultural and social practices, such as practices of endogamy and consanguinity [148][167], also affect the rate of genetic mutations in a population. The study of only European ancestry populations limits the identification of genetic loci impacting health in diverse populations. Furthermore, populations often have different environments that modulate disease risk. These differences could be structural. For example, it is well-documented that racism negatively affects health [7]. Environmental differences could also relate to lifestyle factors, like diet [75] and exercise, or a person's ability and desire to seek healthcare. Thus, to truly enact personalized medicine, more inclusive research is needed into population-specific disease risks.

While many precision medicine initiatives have focused on genomics data due to its value in understanding biological mechanisms, this approach has limitations. An individual's genome is static, meaning it does not give information on disease risk over time. Genomics assays also generally cannot learn about lifestyle or structural disease risks. Epigenetics, or the study of how the environment alters gene function, can be used as an alternative data source [170]. An individual's epigenome is dynamic, which means that it can offer insight into the health of an individual at a moment in time or be probed longitudinally to monitor changes in a person's biological state [84]. Crucially, the epigenome can be used to learn about lifestyle or environmental risk factors, such as air pollution or diet [140]. The epigenome is also correlated with the populations that an individual belongs to, meaning that it can be used to interrogate the complex interaction of population, environment, and genetic risk.

Epigenetics has been used to time disease onset [44], assess survival probability [114][49], and predict patient responses to medications [95]. However, many current epigenetic technologies are invasive and expensive [187][100], which reduces their relevance to clinical care. Additionally, many epigenetic studies only focus on one tissue, which can be limiting in characterizing how tissues interact with each other to produce disease. Lastly, many epigenetic studies have been done with relatively small sample sizes. Larger studies in diverse cohorts would facilitate a greater understanding of tissue and environmental-specific disease risks. Therefore, new technologies are needed to implement epigenetic precision medicine on a large scale.

These limitations highlight the need not just for precision medicine, but also for initiatives that further *precision population health.* A population precision health framework will address population-specific disease risk at scale, with an emphasis on prevention [61][81]. Genomic and epigenomic tools can be used to learn about the mechanism behind population-specific disease risk, the impact of environmental and sociocultural forces on disease, and identify opportunities to reduce disease burden. Importantly, population in this context can mean not only the typical populations characterized by race, ethnicity, or nationality, but also any community an individual belongs to that impacts their risk for disease [92].

In this thesis, I introduce novel strategies to empower precision population health. First, I develop statistical algorithms that focus on the use of non-invasive epigenetic data to detect and monitor tissue death, which is a hallmark of disease. This addresses the goal of developing tests that can monitor the health of many tissues at once, and to identify opportunities for intervention and prevention. I applied these algorithms to data from pregnant and amyotrophic lateral sclerosis patients. Next, to deploy these algorithms to a large cohort, I developed a scalable and inexpensive sequencing technology that can be used to measure disease-relevant regions of the epigenome. This work aims to meet to the goal of applying epigenetic screening measures to a large population and to further the study of epigenetic correlates of disease. Lastly, I performed a study that examined population-specific disease risks in a large biobank. This study is aligned with the goals of precision medicine by identifying disease risks affected by both environmental and genetic factors. Together, these studies advance population precision health initiatives, creating a foundation for future work that aspires towards personalized solutions to disease.

This work is organized into three chapters. Chapter 5 focuses on algorithms to non-invasively detect tissue death. Chapter 6 extends this work to develop a scalable cell-free

DNA platform to monitor disease. Lastly, chapter 7 uses large genetic biobanks for the identification of population-specific health risks.

CHAPTER 2

Comprehensive cell type decomposition of circulating cell-free DNA with CelFiE

2.1 Introduction

Cells die at different rates as a function of disease state, age, environmental exposure, and behavior [124][116]. A quantifiable indication of cell death could facilitate disease diagnosis and prognosis, prioritize patients for admission into clinical trials, and improve evaluation of treatment efficacy and disease progression [76][179][173][20]. Circulating cell-free DNA (cfDNA) is a promising candidate biomarker as it is released into the blood stream after cell death [161][88][86]. In healthy individuals, cfDNA in the blood arises from normal cell turnover, but in individuals with a disease, cfDNA can come from illness-specific cell death [70]. As a result, cfDNA levels have been shown to be elevated in individuals with cancer, autoimmune diseases, transplantation responses, and trauma [62][171][181][59]. CfDNA has also become the clinical standard for noninvasive prenatal testing [122], and many companies and research groups are sequencing cfDNA to identify the presence of somatic mutations related to tumors [182][166][153].

To understand what drives the increased presence of cfDNA in people with disease, this work focuses on the decomposition of cfDNA in blood into its cell types of origin. While each germline cell has nearly the same DNA sequence, DNA methylation is cell type specific [99], and there is a rich literature of complex tissue decomposition approaches using DNA methylation [68][69][138][139]. Recent work has attempted to use cfDNA methylation patterns to decompose tissues of origin for cfDNA [89][120][97][156]. These approaches, however, do not address some of the unique challenges of cfDNA. Previous work was designed for reference and input data from methylation chips, which are high coverage and have relatively low noise. Since cfDNA is only present in the blood in small amounts, an onerous amount of blood must be extracted from a patient to get the required amount of input DNA for methylation chips, which may not be practical for clinical use [134]. Other technologies and methods focus on sensitive detection of specific tissues or cancer sites [77][94] [151]. While increasingly powerful, these approaches can not provide biomarker discovery or comprehensive decomposition of constitutive cell types. In this work, we used whole genome bisulfite sequencing (WGBS) to assess the methylation of cfDNA. Unlike methylation arrays that target specific genomic locations, WGBS covers the entire genome, typically resulting in lower coverage per site, and increased noise relative to array data. Thus, WGBS presents computational challenges for decomposition of methylation data as current computational methods are ill-equipped to handle such noise in either the reference or input. Previous methods are also limited by which DNA methylation sites (CpGs) are chosen. Methylation arrays survey a limited number of CpGs, which may not be maximally informative of cell type. Some approaches also rely on selecting a set of CpGs designed for a particular dataset [32][89][97]. While curated site selection is useful for specific biological queries, it can cause bias when generalized to other settings or diseases. Choosing which sites to include in a decomposition can substantially influence which cell types are predicted because different sites are informative for different cell types. Another important limitation of previous cfDNA decomposition methods is that the results are restricted to the cell types included in the reference panel. However, as there are many thousands of cell types throughout the body, it is currently impossible to incorporate them into a reference panel. Thus, the specific choice of reference cell types will lead to biases in the decomposition results.

In this work, we develop an efficient expectation maximization (EM) algorithm, CelFiE (CELl Free DNA Estimation via expectation-maximization) for cfDNA decomposition that allows for low coverage and noisy data and apply it in a range of challenging real world scenarios. CelFiE can estimate unknown cell types not included in a reference panel and is not dependent on curated input methylation sites. We show in realistic simulations that CelFiE can accurately estimate known and unknown cell types, even at low coverage and with relatively few sites, and can detect rare cell types that contribute to only a small fraction of the total cfDNA. Decomposition of real WGBS complex mixtures demonstrates that CelFiE is robust to several violations of the model assumptions. Specifically, the real data contain correlations across regions and between cell types, read counts with heavy-tailed distributions, and reference samples that are heterogeneous mixtures of many cell types. Additionally, we develop an approach for unbiased CpG methylation site selection for use in the decomposition algorithm.

We apply CelFiE to two cfDNA data sets. First, we examined the positive control of cfDNA extracted from pregnant and nonpregnant women. We observe a significant placental component in the decomposition estimates only from pregnant women, providing validation for CelFiE. We then applied CelFiE to cfDNA from amyotrophic lateral sclerosis (ALS) patients and age matched controls. Currently, there are no established circulating biomarkers for ALS. As a result, it is difficult to monitor disease progression and efficiently evaluate treatment response [176]. cfDNA provides an opportunity to measure cell death in ALS that could fill these gaps. We find a significantly elevated skeletal muscle component in ALS patients. This novel observation, along with the successful decomposition of cfDNA from pregnant women, demonstrates that CelFiE has the potential for broad translational utility in understanding the biology of cell death, and in applications such as quantitative biomarker discovery, or in the noninvasive monitoring and diagnosis of disease.

2.2 Results

2.2.1 CelFiE Overview

CelFiE estimates the contribution of various cell types to the cfDNA of an individual via an EM optimization algorithm. The input to CelFiE is WGBS reference data consisting of T total cell types and WGBS cfDNA samples for N total individuals. Its output is the proportion of the reference cell types that make up each individual's cfDNA, such that the proportion of all T cell types sums to one for each individual. Notably, an arbitrary number of cell types can be missing, which addresses potential biases arising from estimating the proportions of cell types from a restricted reference panel. CelFiE also estimates the methylation values for each of the cell types included in the reference, which accommodates the currently noisy and low-coverage reference data sets. These developments are facilitated by CelFiE's EM algorithm, which is a flexible framework for parameter estimation, even when there is missing data. Complete details on CelFiE can be found in Section 2.4.

2.2.2 Evaluation using simulated cfDNA mixtures

We began by simulating cfDNA mixtures informed by realistic sequencing conditions and comparing the results of CelFiE and other decomposition tools. First, we compared CelFiE to a least-squares regression optimization method. Least-squares regression is a popular choice for decomposition problems, but is not guaranteed to produce an estimate of cell type proportions that sums to one. To compare CelFiE to a constrained optimization method, we implemented a second optimization method referred to here as the ”projection method.” In this approach, we computed the projection of the cell-type proportion estimates onto the L1-ball [45], which constrained the estimates of cell-type proportions to lie on the probability simplex and thus, sum to one. Furthermore, in our projection approach, we optimize a binomial log-likelihood that is parameterized by the number of methylated and unmethylated reads. By accounting for read data, this method is a more direct comparison to CelFiE (see section 2.4.5 for implementation details).

We also compared CelFiE to a previously published cfDNA decomposition tool, MethAtas [120]. Unlike CelFiE, which explicitly models WGBS reads, MethAtlas is designed to decompose methylation array data. MethAtlas also does not model missing data or estimate the methylation values for the reference cell types. Briefly, it optimizes $\|Y\alpha - \beta\|$ using non-negative least squares constrained by $\alpha \geq 0$, where Y is a reference matrix of array data, β is the observed cfDNA methylation measured on an array, and α is the cell type proportions vector that is being solved for. While MethAtlas is not designed for low read count data and thus not directly analogous to CelFiE, it is, to the best of our knowledge, the only other cfDNA decomposition algorithm that allows the inclusion of arbitrary input sites and does not restrict to specific cell types in the reference data.

MethAtlas provides a comprehensive reference matrix, composed of 25 tissues and cell types, over ∼6,000 CpG loci [101]. To ensure a fair comparison, we simulated data that matched the size of this reference data with 25 cell types and 6000 CpGs. The true methylation proportion of each CpG was drawn independently from a uniform distribution, so that the methylation of each CpG was between 0% and 100%. The choice of a uniform distribution allowed for variability across cell types for a given CpG. To characterize the decomposition performance of CelFiE across both rare and abundant cell types, we defined the true cell type proportion vector as $(1, \ldots, T)/\binom{T+1}{2}$, where $T = 25$ is the number of cell types truly in the mixture.

For CelFiE and projection method, the input data were the number of methylated reads and read depth at each site. The reference read depths were drawn independently from a Poisson distribution centered at 10, a relatively low sequencing depth for a WGBS experiment [194]. The number of methylated reads for a given CpG in each of the 25 cell types was drawn from a binomial distribution, where the probability of success was the true methylation value in that cell type, and the number of trials was the read depth at that locus. cfDNA read depths for each CpG were simulated from a Poisson distribution centered at 10, and then the reads for each CpG were assigned to originate from a cell type based on the cell type proportion vector for the cfDNA mixture. A read was determined to be either methylated or unmethylated given the true methylation proportion in that read's cell type of origin at that CpG. Since MethAtlas and least-squares regression do not take read counts as input, we calculated the methylation proportion for a CpG by dividing the methylated reads by the depth at that locus. While these methods were not designed for read count data, by doing this we were able to compare MethAtlas, least squares regression, and CelFiE on the same data. Additionally, to compare the least-squares regression estimates to the proportions produced by the other methods, we divided the vector of estimates produced by least-squares regression by its sum. In total, we performed 50 independent simulations for CelFiE and all comparison methods. Below, we consider additional simulations from real data, which are free from the distributional assumptions above.

CelFiE performed better than MethAtlas at these low read depths (Figure 2.1). Per each simulation, we calculated the Pearson's correlation between the true cell type proportion vector and the estimated proportions vector. For MethAtlas, the mean r^2 across replicates was 0.59±0.17, while CelFiE's mean r^2 was 0.96±0.01. As expected, CelFiE also performed better than linear least-squares regression, which had an mean r^2 of 0.73±0.11 (Figure S1A). CelFiE and the projection optimization method (mean r^2=0.95±0.02) performed similarly under these conditions (Figure S1B). However, a major limitation of our projection optimization method is that, unlike CelFiE, it is unable to estimate missing cell types, which we discuss further below.

To further characterize the properties of CelFiE, we varied the number of CpGs (100, 1,000 and 10,000), which represented conditions with varying amounts of information about cell type. We then focused on a single cell type and varied its proportion between 0% and 100%. In total we simulated 10 cell types, where one cell type was fixed. The remaining 9 additional cell type proportions were drawn from an independent uniform distribution and then normalized so that all proportions sum to one. Data was simulated for 1 individual with 50 independent simulations.

Performance was assessed by calculating the Pearson's correlation between the estimated cell type proportions and the true proportions for 50 replicates. We found that as the number of sites increased, the ability of CelFiE to accurately decompose the cfDNA mixtures improved (Figure 2.2A), especially for less abundant cell types. We further characterized the performance of CelFiE by calculating the correlation between the estimated methylation proportions of the fixed cell type with the true methylation proportions when the reference and input data were at 1x, 5x, 10x, or 100x coverage (Figure 2.2B). At the very low depth of 1x, the mean Pearson's correlation was $r^2 = 0.45 \pm 0.09$, which increased substantially at 5x coverage to $r^2 = 0.83 \pm 0.03$. As the sequencing depth increased, the correlation continued to increase.

Next, we examined the performance of CelFiE when two cell types with highly correlated methylation values were included in the reference panel, since many real cell types share substantial architecture with each other. We generated simulated methylation proportions for the two cell types with a Pearson's correlation between 0 and 1 at 100x depth and ran CelFiE for mixtures of 1,000 CpG sites. When the cell types are very correlated, we found that CelFiE is unable to distinguish between the two cell types. As the cell types become less related, CelFiE improved in its ability to disambiguate the two cell types (Figure S2). We note, however, that CelFiE accurately estimates the sum of the two cell types, even when they are perfectly correlated.

2.2.3 Detection of Differences Between Groups

Previous work suggests that a large portion of cfDNA originates from white blood cells [89]. This implies that a non-haematopoietic cell type of clinical significance may only be present in a population of interest at a low proportion in the mixture. To assess the ability of CelFiE to estimate rare cell types, we simulated data to resemble a small case-control study of 10 total individuals. Five individuals with a low proportion of a single cell-type (0.1%, 0.5%, 1%, or 5%) were simulated to represent the cases. The remaining 5 individuals were simulated to have 0% of that cell type, representing the controls. To understand how CelFiE's ability to estimate rare cell types changes as a function of sequencing depth, we simulated input and reference reads at 5x, 10x, 100x, and 1000x coverage for 1000 CpGs. We ran CelFiE jointly on all 10 individuals to prevent bias and assessed whether CelFiE can meaningfully discriminate between the two groups.

We plotted the CelFiE estimates for individuals whose cfDNA mixtures do and do not have that rare cell type (Figure 2.3A-D). We found that as both the depth and the cell-type proportion increased, CelFiE's ability to distinguish between the two groups improved. A grouped t-test was used to assess whether CelFiE is estimating a significant difference between the groups. At a depth of 5x, CelFiE was only able to distinguish between the groups of the most abundant fixed cell type, 5%, with an average estimate of 0.041 ± 0.018 in the group with the cell type and $2.51 \times 10^{-3} \pm 4.71 \times 10^{-3}$ in the group without. Despite the estimates being slightly underestimated, this difference was significantly different between the groups ($p = 4.8 \times 10^{-8}$), suggesting that CelFiE may have utility in detecting differences between groups even at extremely low depths. As the depth increased to 1000x, CelFiE significantly differentiated between all four fixed percentages ($p < 0.001$) and the estimates became more confident. We found that as we continued to increase the depth, CelFiE was able to detect arbitrarily small differences between the groups (at 10,000X and 0.01%, $p = 8.32 \times 10^{-9}$). In practice, however, the ability of CelFiE to detect these minute differences is limited by biological and technical constraints, such as the amount of cfDNA in blood or DNA degraded by bisulfite conversion. Nonetheless, these results demonstrate that CelFiE can accurately estimate cell types of relatively rare abundance when the read depth is high.

2.2.4 Unknown Cell Types

We then turned to understanding the behavior of CelFiE when estimating unknown cell-types. To accomplish this, we simulated data with low read counts, creating reference and cfDNA reads for 1000 CpGs at 10x depth, as in previous simulations. We simulated $t = 10$ cell types with one unknown cell type excluded from the reference data. We began by simulating a missing component that was relatively large. Its proportion, $\alpha_{unknown}$, was drawn as $\alpha_{unknown} \sim \mathcal{N}(0.2, 0.1)$ and truncated to be between 0 and 1. The remaining cell type proportions of the known cell types were drawn from a uniform distribution and all proportions were normalized to sum to 1. We then simulated cfDNA reads for 10, 50, 100, 500, and 1000 individuals. Note that the problem is not identifiable when the number of individuals is smaller than the number of unknowns. The mean squared error (MSE) was calculated between the estimated unknown proportion and the true simulated proportion. As the number of people included in the decomposition was increased, the performance of CelFiE improved (Figure 2.4A).

We next considered mixtures with two unknown cell types, one that was relatively large and one that was relatively small. For each person, the first unknown proportion, $\alpha_{unknown1}$, was drawn from $\alpha_{unknown1} \sim \mathcal{N}(0.2, 0.1)$, and the second unknown was drawn from $\alpha_{unknown2} \sim \mathcal{N}(0.1, 0.1)$. The proportions of the remaining cell types were simulated as above. Since the inferred CelFiE labels are not identified (i.e., CelFiE's estimated $\alpha_{unknown1}$ can correspond to either missing reference cell type 1 or 2), we assigned the unknowns by examining the estimated methylation fractions of each CpG. We estimated the correlation between the true and unknown methylation fractions and assigned the unknown to the true cell type with the highest correlation. After assigning the unknowns, we calculated the MSE between the true proportion and the estimated proportion. Furthermore, we calculated the Pearson's correlation between the true and estimated methylation fractions for each unknown (Figure S3). We observed that more individuals are needed to accurately estimate the unknown components when an additional unknown was added (Figure 2.4B). We also noted the presence of outliers in the estimates, which was likely due to differences in the simulated data that were randomly drawn in each replicate of our experiment.

We next examined how decomposition estimates are biased when there is a missing cell type, but no unknown is estimated. We generated simulated mixtures as above, for 1000 CpGs and 10 cell types truly in the reference, and for 100 people at 100x depth. CelFiE was ran twice: once when the missing cell type was the highest tissue in the mixture (∼20%) and secondly, when the missing cell type was approximately the average of all cell types contained in the mixture (∼10%). (Figure S5). To measure the bias of the estimates, we calculated the percent difference, defined as the true cell type proportion minus the estimate, divided by the true proportion. When the missing cell type was high, the average percent difference across all tissues was 0.32 ± 0.86. This meant that on average, without estimating the unknown, CelFiE produced cell type proportion estimates that were 32% higher than the truth. Likewise, when the missing cell type was lower, the average percent difference decreased to 0.21 ± 0.69, likely because there was less missing signal to be distributed across the cell types actually estimated. When there was an unknown included in CelFiE, the overestimate on average, decreased to -0.02 ± 0.62 and -0.11 ± 0.40, respectively. This result indicated that the larger the missing cell type, the more biased the cfDNA decomposition estimates will be without an unknown component, which may demonstrate the utility of CelFiE.

CelFiE's ability to accurately estimate unknowns contrasts with previous cfDNA decomposition methods, which can only estimate proportions of cell types in the reference. This creates a bias in the decomposition that can be addressed with CelFiE. Specifically, if we simulate cfDNA mixtures with a cell type excluded from the reference as above and run MethAtlas, it will produce biased estimates. On average, these estimates had an average percent difference that was $29 \pm 68\%$ larger than the true proportions (excluding the missing cell type, which was not estimated). We found similar biases in our least squares regression method, which on average, overestimated by $29 \pm 20\%$, and in our projection method, which on average, overestimated by $17 \pm 57\%$ (Figure S4). The difference in performance between CelFiE and comparison methods is more similar at high read depths and when all cell types are known (Figure S6).

2.2.5 Performance on WGBS cfDNA mixtures

We next considered simulated mixtures made from real WGBS data, which are substantially more complex and violate several assumptions of the CelFiE algorithm. In particular, the reference data contain tissues composed of multiple cell types, CpGs are correlated locally across genomic regions and between cell types, and read counts have heavy-tailed distributions reflecting true biological and technical heterogeneity across sites. Therefore, to examine how robust CelFiE is to these complications, we used biological replicates for 10 WGBS data sets (small intestine, pancreas, monocytes, stomach, tibial nerve, macrophages, memory B cells, adipose, neutrophils, and CD4+ T cells), downloaded from the ENCODE and BLUEPRINT projects [**?**][41][50]. In all experiments, we chose to include tissues to see if their complex cell type mixtures might contribute to decomposition errors. One set of WGBS biological replicates was assigned to make up the cfDNA mixtures; the other was assigned to the reference matrix.

Since roughly 80% of CpG sites in the human genome do not vary between cell types [193], randomly selected CpGs will contain mostly uninformative loci for cell-type decomposition. A reference panel that contains too many uninformative CpGs will reduce the performance of a decomposition algorithm. To demonstrate this, we simulated data for 100, 1000, and 10000 CpGs, where the true methylation values for 10 cell types were drawn from a normal distribution centered on 0.5. The variance across cell types was chosen to be between 0.01 and 1. The lower the variance, the less informative a CpG would be for cell type status. A cfDNA mixture for one individual and no missing cell types was simulated. The results of this experiment indicated that as the variance increased, CelFiE's ability to decompose the mixtures also increased (Figure S7). Therefore, to limit uninformative CpGs included in our analysis, we developed a method for choosing a set of unbiased informative CpGs in real data, which we called tissue informative markers (TIMs) (Section 2.4.11). We selected 100 TIMs per WGBS sample for use in these simulations, excluding common variants with a minor allele frequency greater than 1% [6]. Selecting TIMs improved performance in CelFiE decomposition (Figure S8). Furthermore, because DNA methylation of nearby CpGs are correlated [103], we combined information from proximal CpG sites 250bp upstream and downstream of each TIM (Section 2.4.11). These combined TIM regions improved the decomposition over single CpGs (Figure S9). We simulated cell type proportions 50 times for 100 people, as in Section 2.2.2. The proportion of CD4+ T cells was drawn from a normal distribution centered around 20% and the proportion of small intestine was centered around 10%. The remaining cell types proportions were drawn per person from a random uniform distribution.

We first assessed CelFiE's performance on WGBS samples without any cell type missing from the reference panel (Figure 2.5A). Despite the complexity of the data, we found that CelFiE still performed well. The average Pearson's correlation between the estimated cell type proportions and true cell type proportions was $r^2 = 0.83 \pm 0.16$. The average Pearson's correlation of the estimated methylation values and the true methylation values was similarly high, with an average r^2 of 0.96±0.01 (Figure S10A). For comparison, we adapted MethAtlas for whole-genome data. We used our selected TIMs and converted the read counts to proportions. The Pearson's correlation between the estimated methylation proportions and true proportions for MethAtlas was lower than that of CelFiE, 0.43 ± 0.24, which further illustrated that MethAtlas is not suitable for noisy read count data.

Next, we investigated CelFiE's ability to estimate mixtures with a substantial unknown component. We first masked only the most abundant cell type from the reference, the CD4+ T cell sample. Using the same true cell type proportions as in the simulations with no missing samples, we performed 50 simulations with 100 people (Figure 2.5B). The correlation between the estimated and true cell type proportions decreased only slightly in the case of no missing data, $r^2 = 0.8 \pm 0.16$, and we found that the correlation to the true methylation values was still high, with an average Pearson's $r^2 = 0.96 \pm 0.01$ across all cell types (Figure S10B). Subsequently, we masked two reference samples, CD4+ T cell and small intestine, from the reference panel. The true CD4+ T cell proportion was still centered around 20%, while the small intestine was centered around 10%. We found that CelFiE's ability to successfully decompose a complex mixture decreased when there are two missing cell types (Figure 2.5C). However, the estimated correlation to the true WGBS methylation values remained high, with an average Pearson's $r^2 = 0.95 \pm 0.04$ (Figure 2.5C and Figure S10C).

To further validate CelFiE's ability to estimate missing cell types, we assessed how similar the learned methylation proportions for the missing cell types are to the true methylation proportions for CD4+ T cells and small intestine. To do this, we appended the methylation proportions learned by CelFiE for the two unknown cell types to the matrix of true reference methylation proportions, including the values for T cells and small intestine that were originally masked. We calculated a distance matrix for the reference matrix plus unknowns and used this to perform hierarchical clustering. Figure 2.6 shows that the unknown cell types were segregated with their true cell type. For the case of one unknown, the unknown that was truly T cell clusters with the reference T cell sample. Furthermore, the average Pearson's correlation between the learned unknown cell type methylation proportions and the reference T cell methylation proportions was higher than all other cell types, $r^2 = 0.95$, suggesting that CelFiE learned the correct cell type for one unknown. For the two unknown cell types, unknown 1 remained clustered with the reference CD4+ T cell sample and had a high correlation with the reference CD4+ T cell methylation patterns, $r^2 = 0.94$. Unknown 2 clustered with the reference small intestine sample along with other gastrointestinal tissues. The correlation between the estimated and true small intestine methylation values was the highest of all pairings, $r^2 = 0.87$. Together with the data presented in Figure 2.5B, these observations suggest that even with an incomplete reference, CelFiE estimates both the correct cell type proportion and cell type methylation values.

2.2.6 Application to Pregnancy

To validate CelFiE, we first choose to analyze cfDNA from pregnant and non-pregnant females since these populations provide a robust example of a verifiable positive and a control group [168]. Unlike the decomposition of cell types in blood, there is no FACS or similar existing standard for cfDNA. Nonetheless, we know a priori that non-pregnant women will not have placenta cfDNA in their bloodstream.

To test CelFiE in pregnant and non-pregnant women, we downloaded publicly available WGBS cfDNA of 7 pregnant and 8 non-pregnant women [71]. All women were between 11- and 25- weeks gestation at the time of cfDNA extraction. Next, we subset the WGBS sites to the same TIMs we use in Section 2.2.5 and summed all reads +/- 250 bp around each TIM (See Methods). Twenty WGBS datasets from the ENCODE and BLUEPRINT projects were chosen for the reference panel, representing tissues and cell types throughout the body and blood, along with one unknown category. The decomposition result is the random restart with the highest log likelihood of 10 total restarts.

CelFiE estimated a high proportion of white blood cells (dendritic cells, eosinophils, monocytes, neutrophils, etc.), consistent with previous estimates based on cfDNA and our expectation that blood cells have high rates of cell turnover [120] [127]. CelFiE detected a small proportion of cfDNA coming from gastrointestinal tissues, such as the small intestine or stomach, which may also be due to the relatively high cell shedding in these tissues [185]. We used a single unknown cell type component and we estimated that it is large, with a mean of 0.31 ± 0.04 in non-pregnant women and a mean of 0.25 ± 0.06 in pregnant women (Figure 2.7A). To better understand which tissues and cell types are driving the unknown component, we performed hierarchical clustering on the estimated methylation values for the unknown component with the methylation values for the known cell types contained in the reference panel (Figure S11). We found that it clustered most closely with endothelial cells. This suggests that as reference panels improve, there is additional biological insight that may be gained by using CelFiE.

To evaluate which cell types differ the most between pregnancy states, we performed grouped two-sample t-tests of inferred cell type proportions. As expected, placenta showed the greatest difference, ranging from 9.3% to 29.7% (median 11.9%), and 2.9×10^{-16} to 2.1×10^{-2} (median 2.3×10^{-12}) in pregnant and non-pregnant women, respectively (two-sided grouped t-test, $p = 4.5 \times 10^{-5}$). We also found that CelFiE estimated a higher placental component in the second trimester (median 11.2% in trimester 1 and median 15.3% in trimester 2), concordant with the growth of the placenta throughout pregnancy (Figure 2.7B). This is also consistent with previous estimates of the proportion of placental DNA in the cfDNA of pregnant women (median 15.3% in trimester 1/2) [164]. We restricted statistical tests to the relevant tissue, in this case the placenta, but estimates are provided for all tissues and cell types in Figure 2.7.

To further validate our method, we compared CelFiE predictions with those from our WGBS adaption of MethAtlas, least squares regression, and our projection method (Figure S12A-C). While these methods are not explicitly designed to be ran on WGBS data, all three methods estimated a higher proportion of placental cfDNA in pregnant women than in non-pregnant women, as we expected (Table **??**). Least squares regression, however, produced negative estimates, suggesting that this method is unsuitable for real data applications. Furthermore, all three methods estimated proportions of blood cell types that may be inconsistent with known cell-type proportions in whole blood. For example, all methods estimated a large erythroblast component, on average about 24%. This was higher than expected since nucleated red blood cells are generally rarer than white blood cells in the blood [36]. Furthermore, white blood cells, such as neutrophils, have a much higher turnover rate, making them more likely to appear in cfDNA [115]. While the high proportion of erythroblasts may indicate the presence of a red blood cell precursor not captured by the current reference panel, it may also be a consequence of a bias introduced by missing tissues in the reference panel. For instance, CelFiE ran with an unknown component on the same data estimated an erythroblast proportion of 0.073 ± 0.052. When CelFiE was ran without an unknown component (Figure S12D), the erythroblast proportion increased to 0.29 ± 0.11. This could suggest that, as seen in Figure S5, decomposition estimates without unknown components may cause overestimation of other cell types in the mixture.

2.2.7 Application to ALS

Lastly, we examined cfDNA in ALS patients and age-matched controls (Sections 2.4.6 and 2.4.7). ALS patients represented a range of disease severity and onset sites. We first examined the overall abundance of cfDNA in cases (n=28, mean 297.72 ± 110.57pg/ul) and controls (n=25, mean 218.78 ± 139.17pg/ul). We observed a significant excess in cases (Figure 2.8A, $p = 5.00 \times 10^{-3}$), but it was unknown what tissue or tissues are responsible for this increase. To explore possible overrepresented tissues in ALS cfDNA, we applied CelFiE first to a discovery cohort composed of 8 controls and 8 cases from both the University of Queensland and UCSF (Figure S13A). As with the pregnancy cfDNA, we confined the WGBS data to TIM sites and then summed +/- 250 bp around the TIMs. We decomposed all mixtures using the same reference tissues as Section 2.2.6 and one unknown. We restricted statistical tests to two biologically relevant tissues for ALS: skeletal muscle and tibial nerve. Notably, we found a difference in the estimated skeletal muscle proportions, specifically finding an excess in cases relative to controls ($p = 5.02 \times 10^{-2}$) (Figure S14A).

We validated this difference with an independent replication of 8 cases and 8 controls from University of California San Francisco (UCSF) for which WGBS was performed. As expected, we found that the mixture was composed largely of blood cells (Figure S13B), with the top 5 tissues by proportion being neutrophils, monocytes, macrophages, eosinophils, and erythroblasts. In addition, CelFiE estimated a large unknown component, with a mean proportion of 0.42±0.11 for ALS cases and 0.30±0.19 for control samples. This large unknown component did not cluster closely with any cell type or tissue contained in our reference panel when we applied hierarchical clustering on the CelFiE estimate of the unknown methylation values (Figure S15), which could indicate that CelFiE captured a substantial signal not captured by other methods. Furthermore, we replicated the significantly higher skeletal muscle component in ALS cases, with a mean muscle proportion of 0.057 ± 0.06, while CelFiE estimated an average proportion of $8.9 \times 10^{-4} \pm 1.3 \times 10^{-3}$ in cases (grouped t-test $p = 7.8 \times 10^{-3}$) (Figure S14B). CelFie ran on the combined data (Figure S13C), estimated a mean proportion of 0.038 ± 0.020 in ALS samples and $1.7 \times 10^{-3} \pm 2.6 \times 10^{-3}$ in controls (grouped t-test $p = 2.4 \times 10^{-3}$) (Figure 2.8B).

Finally, we ran least squares regression, our projection method, and MethAtlas on our combined ALS cfDNA data (Figure S16). As in Figure S12, we found that these methods estimated higher proportions of erythroblasts than CelFiE, and that least squares regression produced negative estimates. We did find, however, that all three methods recapitulated our finding of a higher proportion of skeletal muscle in ALS patients (Table **??**). While these differences are similar in magnitude to those from CelFiE, they are less significant (least squares: $p = 0.019$; projection method: $p = 0.026$; MethAtlas: $p = 0.012$), possibly due to the higher error in these methods.

Together, these results suggest that cfDNA is a promising direction to identify the first quantitative biomarker for muscle atrophy and death that is a hallmark of ALS.

2.3 Discussion

During disease or increased cell turnover, elevated levels of cfDNA can be detected in the blood. For example, increases in the amount of cfDNA have been detected in patients with multiple types of cancer, autoimmune diseases, as well as acute episodes of myocardial infarction, trauma, transplantation response, and exercise [165][174][144]. Correspondingly, the utility of cfDNA as a diagnostic biomarker has been demonstrated in an increasing number of settings, including prenatal testing [72] and the detection tumor specific mutations [93][132]. Of great interest, however, is that assessments of cfDNA can now also provide information about cfDNA cellular origin [156][89][120][97]. This type of qualitative and quantitative assessment presents an individualized, unbiased approach to understanding cellular turnover over time. However, these technologies are nascent, noisy, and expensive.

In this work, we presented an algorithm, CelFiE, to decompose complex cfDNA mixtures into their cell types of origin. CelFiE can accurately decompose cfDNA mixtures with low sequencing coverage in both the reference cell types and the patient cfDNA samples. We also showed that CelFiE could estimate cell type proportions using relatively few sites, and that its performance improves as more tissue informative sites are selected. This could indicate CelFiE's utility in methylation capture panel development, where highly informative sites are selected and sequenced to high depth [96]. Furthermore, as cohort sizes are expanded, it can accurately estimate multiple unknown cell types, which reduces bias and increases confidence in the decomposition. Finally, the EM algorithm underlying CelFiE is computationally efficient, with iteration cost scaling linearly with the number of samples, CpG sites, and cell types.

We began by validating CelFiE extensively in simulations. In the context of simulated low read-count methylation data, CelFiE outperformed linear least squares regression, our novel L1-projection method, and MethAtlas, another cfDNA decomposition method. Since these methods are not explicitly designed for this data regime, CelFiE's improvements may make it a useful addition to the tools available to cfDNA researchers. To further demonstrate the accuracy of CelFiE, we applied it to real data from pregnant women. Decomposition estimates of placenta from pregnant women were significantly different from non-pregnant women. This provided a natural validation for CelFie, illustrating that it can correctly learn differences in cfDNA cell type of origin, even in real data sets.

In our study of ALS patients, we found that cfDNA levels are increased in ALS cases compared to controls. To understand what cell types are driving this difference, we applied CelFiE to the cfDNA samples, finding significantly higher skeletal muscle in patients with ALS. Future work will expand on this result by expanding the cohort size, and by testing for associations between cell type of origin and disease progression or severity. We may also test for associations between decomposition estimates and disease onset site. Furthermore, as cohort sizes expand, we will have the power to estimate multiple unknown categories. These multiple unknown categories could be used to further subtype ALS cases. We consider the current results, a promising step forward, especially as ALS currently has no reliable biomarker. These results also suggest that CelFiE might prove useful for quantifying cell death in other complex diseases.

The accuracy of CelFiE depends on several factors including read depth, the cell-type specificity of the sites considered, the abundance of key cell types, and the quantity and quality of reference data and cfDNA patient samples. Recent technologies for digesting or capturing specific regions of cfDNA [83], may allow deeper sequencing of informative CpGs. Selecting such TIM CpGs demonstrated marked improvement in accuracy and could be used to select sites for capture.

There are a number of areas for improvement. Many of the reference samples used here were complex mixtures of cell types and could be modeled as such, similar to the recent approach, FEAST [152], which modeled reference mixtures of microbial communities. Moreover, WGBS simulation results showed a high degree of correlation between replicates, but we believe modeling inter-person heterogeneity will likely improve the results further in real cfDNA samples. We currently account for the local correlation of CpG methylation by summing proximal CpG methylation states, but nearby CpGs may not always convey identical cell type information. Future work could also focus on modelling the relationship between cell types and tissues. For example, since cell types are correlated in their methylation profiles, it could be interesting to consider a hierarchical model in which the composition can be considered at different levels of cell type phylogeny [154]. This may help us gain additional power to identify samples, particularly highly similar cell types or tissues. Finally, the addition of non-CpG methylation and cfDNA fragment length may provide additional sources of information about cell types of origin.

In summary, we present CelFiE, an efficient EM algorithm for decomposing cfDNA mixtures into their cell type of origin, even when the data are low count or noisy. CelFiE can additionally robustly estimate both known and unknown cell types in cfDNA. Overall, our work demonstrates that CelFiE could be a useful tool for quantifying cell death, applicable to biomarker discovery and disease monitoring.

2.4 Methods

2.4.1 CelFiE Overview

We assume that we are provided with a bisulfite sequenced reference data set, composed of T cell types indexed by t, at M CpG sites indexed by m. Bisulfite sequencing produces read counts from specific cell types that we collect in two $T \times M$ matrices: Y and D^Y, where, Y_{tm} and D^Y_{tm} are the number of methylated and total reads at CpG m, respectively, in reference cell type t. Together, these two matrices represent the reference cfDNA data.

We are also provided with cfDNA extracted from N individuals indexed by n. The bisulfite sequencing read counts of the cfDNA are given in two $N \times M$ matrices X and D^X, with X_{nm} and D^X_{nm} giving the number of methylated and total reads at CpG m in the cfDNA from individual n, respectively. These two matrices represent the sample cfDNA data.

CelFiE takes as input the matrices Y, D^Y, X_{nm}, and D^X_{nm}, and then outputs a matrix α, where α_{nt} is the fraction of the cfDNA in person n that originated from cell type t.

2.4.2 Model

We model the cfDNA as a mixture of DNA from cell types in the reference panel and, potentially, unknown cell types absent from the reference panel. We assume that the individuals are independent given the true, unknown methylation proportions of each cell type, and the individual-specific cell type proportions.

We assume that reference data are drawn from a binomial distribution:

$$Y_{tm}|D^Y_{tm}, \beta_{tm} \overset{iid}{\sim} \text{Binomial}(D^Y_{tm}, \beta_{tm}) \tag{2.1}$$

where $\beta_{tm} \in [0, 1]$ is the true, unknown proportion of DNA in a cell type that is methylated at position m. This model assumes no intra-cell type heterogeneity, in the sense that each cell in a cell type has identical methylation probability.

Next, we model the samples in the cfDNA data. We assume each cfDNA read is drawn from some cell type t at some marker m, and in turn that its methylation value is drawn from a Bernoulli distribution governed solely by the methylation proportion in the cell type of origin:

$$x_{nmc}|\beta, Z_{nmc} = t \overset{iid}{\sim} \text{Bernoulli}(\beta_{tm}) \tag{2.2}$$

where x_{nmc} is the methylation status of the c-th read from sample n at position m, and $Z_{nmc} = t$ indicates that t is the cell type of origin for this read. For each person and methylation site, we define the total number of methylated reads as $X_{nm} := \sum_{c=1}^{D^X_{nm}} x_{nmc}$. This simply sums the methylation status over all reads for each person at each site. In the special case where $D^X_{nm} = 0$, we define $X_{nm} = 0$.

Finally, we assume that the cell type of origin of each cfDNA molecule is drawn independently from some individual-specific multinomial distribution:

$$Z_{nmc}|\alpha_n \overset{iid}{\sim} \text{Multinomial}(\alpha_{n1}, \ldots, \alpha_{nT}) \tag{2.3}$$

where α_{nt} is the probability that a read from person n comes from cell type t.

2.4.3 EM algorithm for one cfDNA sample

For simplicity, we first describe CelFiE in the case where the cfDNA data set contains only a single person, meaning the decomposition relies almost exclusively on the reference panel. We then explain how CelFiE can jointly model multiple individuals in the cfDNA data, as well as how and why this enables the estimation of unknown cell types. Full details of both algorithm derivations are given in Section **??**.

Formally, assume there is only one sample in the cfDNA data (i.e. $N = 1$). We define z_{tmc} as a binary indicator for whether read c at CpG m for the single cfDNA individual originates from cell type t. In relation to Z above, $z_{tmc} = 1$ if $Z_{1mc} = t$, and otherwise 0. That is, Z_{1mc} is a categorical variable, and z_{tmc} indicates which value Z_{1mc} takes.

To calculate the full data likelihood, $P(x, z, Y|\alpha, \beta)$, we first factorize it into $P(x, Y|z, \alpha, \beta)\cdot$ $P(z|\alpha, \beta)$. This then simplifies into three components:

$$P(x, z, Y|\alpha, \beta) = P(x|z, \beta)P(z|\alpha)P(Y|\beta) \tag{2.4}$$

The first component defines the probability of the cfDNA reads, given which cell type they come from and the methylation proportions of those cell types. The third component analogously defines the probability of drawing the reference reads. The second component describes the probability of observing a specific cell type in the cfDNA, which is determined by the proportion of each cell type in the individual's cfDNA.

We show in **??** that the resulting log-likelihood is equivalent to:

$$\sum_{t,m,c} z_{tmc} \left[x_{mc} \log\left(\beta_{tm}\right) + (1 - x_{mc}) \log\left(1 - \beta_{tm}\right)\right] + \sum_{t,m,c} z_{tmc} \log \alpha_t$$
$$+ \sum_{t,m} \big(Y_{tm} \log \beta_{tm} + (D^Y_{tm} - Y_{tm}) \log(1 - \beta_{tm})\big)$$

For this one-sample section, we drop an index on x and write x_{mc} instead of x_{1mc}. Analogously, we write $X_{nm} = X_m$ as the total number of methylated reads at position m (and D^X_{nm} as D^X_m).

To calculate the expected log-likelihood, i.e., the Q function, we must integrate over the conditional distribution for the missing data, i.e. $P(z|x, \beta, \alpha)$. Since z_{tmc} is binary and each read and site is assumed independent, this distribution is the probability that each z_{tmc} is 1. In other words, the probabilities that each read comes from each cell type are sufficient statistics, and are given by:

$$P(z_{tmc} = 1|x_{mc}, \beta, \alpha) = \frac{\beta_{tm}^{x_{mc}}(1-\beta_{tm})^{1-x_{mc}}\alpha_t}{\sum_k \beta_{kt}^{x_{mc}}(1-\beta_{kt})^{1-x_{mc}}\alpha_k} =: \tilde{p}_{tmc}(\alpha, \beta) \tag{2.5}$$

Conceptually, if read c is methylated, this indicates the read is more likely to come from cell types with high methylation proportion, as β_{tm} is larger (and vice versa if the read is unmethylated). Regardless the methylation state, however, this equation also says that the read is likelier to come from more common cell types, as α_t is larger.

This final term $\tilde{p}_{tmc}(\alpha, \beta)$, seems complex. However, it actually only depends on the specific read c through its methylation status, and takes only two values. We can redefine it in simpler terms, which represents the probability of each cell type for each read depending on its methylation:

$$\begin{aligned} \frac{\beta_{tm}\alpha_t}{\sum_k \beta_{kt}\alpha_k} &=: p_{tm1}(\alpha, \beta) = \tilde{p}_{tmc}(\alpha, \beta) \ \text{ if } x_{mc} = 1 \\ \frac{(1-\beta_{tm})\alpha_t}{\sum_i (1-\beta_{kt})\alpha_k} &=: p_{tm0}(\alpha, \beta) = \tilde{p}_{tmc}(\alpha, \beta) \ \text{ if } x_{mc} = 0 \end{aligned} \tag{2.6}$$

2.4.3.1 E step:

The Q function is defined at iteration i by:

$$Q_i(\beta, \alpha) := z|x, \alpha^{(i)}, \beta^{(i)} \log P(x, z, y|\alpha, \beta) \tag{2.7}$$

where $\alpha^{(i)}$ and $\beta^{(i)}$ are the parameter estimates of the cell type proportions and methylation proportions from the last EM step. Let $p_{tm}^{(i)} := p_{tm1}(\alpha^{(i)}, \beta^{(i)})$, which is the probability that a methylated read at site m comes from cell type t given the previously estimated parameters from iteration i. Then Q_i is:

$$Q_i(\beta, \alpha) = \sum_{t,m} \left[\left(Y_{tm} + p_{tm1}^{(i)} X_m \right) \log\left(\beta_{tm}\right) + \left(D_{tm}^{Y} - Y_{tm} + p_{tm0}^{(i)} (D_m^X - X_m) \right) \log\left(1 - \beta_{tm}\right) \right] \tag{2.8}$$

$$+ \sum_{t,m} \left(X_m p_{tm1}^{(i)} + (D_m^X - X_m) p_{tm0}^{(i)} \right) \log \alpha_t \tag{2.9}$$

The first line in this equation captures the expected total number of methylated reads (first term in the sum) and the total number of expected unmethylated reads (second term) for each cell type and site. Each of these terms combines both the reference and cfDNA contribution, e.g. the first term combines the total methylated reads from the relevant reference cell type (Y_{tm}) with the expected number of methylated reads from that cell type in the cfDNA mixture ($p_{tm1}^{(i)} X_m$).

Complementary to the first line, the second line determines the likelihood of α and does not depend on β. It captures the likelihood of observing the expected cell type frequencies. This is given by the sum of the expected methylated and the expected unmethylated reads over all loci.

2.4.3.2 M step:

To update the estimated cell type proportions, α, we maximize Q_i under the constraint that α is a probability vector, i.e., its entries are non-negative and sum to one. The maximizer is:

$$\alpha_t = \frac{\sum_m \left(x_m p^{(i)}_{tm1} + (D^X_m - x_m) p^{(i)}_{tm0} \right)}{\sum_{k,m} \left(x_m p^{(i)}_{km1} + (D^X_m - x_m) p^{(i)}_{km0} \right)} \tag{2.10}$$

The numerator is simply the number of reads expected to originate from each cell type, which is calculated by adding the expected contributions from the methylated and the unmethylated reads. The proportions are then obtained by normalizing these numerators to sum to 1.

The other M step update is for β, the proportion of reads that are methylated at each site and in each cell type:

$$\beta_{tm} = \frac{p^{(i)}_{tm1} x_m + Y_{tm}}{p^{(i)}_{tm0}(D^X_m - x_m) + D^Y_{tm} - Y_{tm} + p^{(i)}_{tm1} x_m + Y_{tm}} \tag{2.11}$$

Intuitively, this is the ratio of the expected number of methylated vs total reads from cell type t at site m. This update is conceptually similar to the α update in the sense that it matches an estimated proportion to an expected proportion. For α_t, this is the expected proportion of reads deriving from cell type t; for β_{tm}, this is the expected proportion of reads from cell type t that are methylated at site m.

2.4.4 EM algorithm for multiple cfDNA samples

We now return to allowing $N > 1$ cfDNA samples. In this setting, α is a matrix, because each cfDNA sample may have different proportions of each cell type in their cfDNA mixture. Further, x_{nmc} and Z_{nmc} are now 3-dimensional arrays indexed by cfDNA individual n, methylation site m, and sequencing read c, and the binary indicators z_{nmtc} are now 4-dimensional, as they additionally index each cell type.

The conditional distribution for z at each step of the EM algorithm now becomes:

$$P(z_{ntmc} = 1|x_{nmc}, \beta, \alpha) = \frac{\beta_{tm}^{x_{nmc}}(1-\beta_{tm})^{1-x_{nmc}}\alpha_t}{\sum_k \beta_{tk}^{x_{nmc}}(1-\beta_{tk})^{1-x_{nmc}}\alpha_k} =: \tilde{p}_{ntmc}(\alpha_n, \beta) \tag{2.12}$$

As before, this $\tilde{p}$ term depends on c only through x_{nmc}, and so we simplify terms by defining $\tilde{p}_{ntmc}(\alpha_n, \beta) = p_{ntmj}(\alpha_n, \beta)$ if $x_{nmc} = j$ for $j = 0, 1$.

To simplify the E step, we define the responsibilities by $p_{ntmj}^{(i)} := p_{ntmj}(\alpha_n^{(i)}, \beta^{(i)})$. For $j = 0$, this gives the conditional probability that an unmethylated read from individual n as site m comes from cell type t given the current parameter estimates; $j = 1$ gives the analogous probability for methylated reads. Since we assume cfDNA individuals are independent given α and β, the E step is a simple generalization of the one-sample E step that sums over samples and can be written:

$$Q_i(\alpha, \beta) = \sum_{n,t,m} \left[\left(Y_{tm} + p_{ntm1}^{(i)} X_{nm} \right) \log(\beta_{tm}) + \left(D_{tm}^{Y} - Y_{tm} + p_{ntm0}^{(i)}(D_{nm}^{X} - X_{nm}) \right) \log(1-\beta_{tm}) \right] \tag{2.13}$$

$$+ \sum_{n,t,m} \left(X_{nm} p_{ntm1}^{(i)} + (D_{nm}^{X} - X_{nm}) p_{ntm0}^{(i)} \right) \log \alpha_{nt} \tag{2.14}$$

This Q function can be interpreted identically to the single-sample Q function. The only difference is that now reference reads are added with expected cfDNA reads for multiple individuals, and the expectations $(p_{ntmj}^{(i)})$ depend on cfDNA individual n as well as cell type t, CpG site m, and methylation status j.

Q_i additively splits over row of α, therefore, the updates for each $\alpha_{n,}$ are identical to the single-sample α updates, where α_{nt} replaces α_t, X_{nm} replaces X_m, D_{nm}^X replaces D_m^X, and $p_{ntmj}^{(i)}$ replaces $p_{tmj}^{(i)}$. This means that if we condition on the number of reads coming from each cell type in person n, the estimates of that person's cell type proportion do not depend on anything else.

For β_{tm}, the M-step again compares the expected number of methylated and unmethylated reads at CpG m from cell type t, where the expectation combines reads from reference cell type t with the expected number of cfDNA reads from cell type t. The only difference is that now the expectation combines the expected contributions from multiple cfDNA samples:

$$\beta_{tm} = \frac{\sum_n p_{ntm1}^{(i)} X_{nm} + Y_{tm}}{\sum_n p_{ntm0}^{(i)}(D_{nm}^X - X_{nm}) + D_{tm}^Y - Y_{tm} + \sum_n p_{ntm1}^{(i)} X_{nm} + Y_{tm}} \tag{2.15}$$

2.4.4.1 Unknown sources:

It is likely that there are cell types in the cfDNA mixture not contained in the reference data. To estimate the proportion of an unknown cell type with CelFiE, we append a zero row to D^Y and Y, and then run CelFiE as usual. This produces an EM that is mathematically similar to the STRUCTURE model of mixtures of human populations [137]. Essentially, CelFiE estimates methylation patterns and abundances for the unknown cell type(s) that maximize the overall likelihood. To model more than one unknown cell types, additional rows of zeros are added to D^Y and Y. Note that if the number of unknown cell types is greater than the number of individuals, the problem is not identified.

2.4.4.2 Regularization and Missing Data:

Missing observations are allowed in both the reference and the input. It is represented as a 0 entered in both X/D^X or Y/D^Y. In practice, we add a methylated and unmethylated pseudocount to every entry of X and Y/D^X and D^Y to stabilize the algorithm and likelihood in case of cell type/site combinations with very low coverage.

2.4.4.3 Computational cost

Each iteration of the EM algorithm in CelFiE involves three calculations. First, $p^{(i)}_{ntmj}$ is evaluated for each sample n, cell type t, CpG site m, and methylation status $j = 0, 1$; each calculation is independent of the input data dimensions, hence evaluating $p^{(i)}$ is $O(NTM)$. Second, α_{nt} must be evaluated, which involves summing over M sites for each n and t, giving overall complexity $O(NTM)$. Finally, updating β_{tm} requires summing over all cfDNA individuals and the reference cell type data, again giving overall complexity $O(NTM)$. Overall, this means that CelFiE scales linearly in sample size, number of CpGs, and number of cell types.

We also note that if multiple references were included, the cost would not multiply–rather, the cost would increase to $O((N + N_{ref})TM)$, where N_{ref} is the (maximum) number of reference samples per cell type.

2.4.5 Other Decomposition Methods

Linear least-squares regression was implemented using the linregress package from SciPy (v 1.5.2) in Python [180]. We minimized $min||X\alpha - Y||_2^2$ where X was the methylation proportions of the cfDNA input and Y was the methylation proportions of the reference matrix. We estimated α, which was the cell-type proportions of the cfDNA mixture. Since least-squares regression does not return estimates that sum to one, we divided α by its sum.

Projection onto the L1 ball was a implemented in a custom Python script available at https://github.com/christacaggiano/celfie. There, we optimize a binomial log-likelihood, where the number of successes is the number of methylated cfDNA reads, the number of trials is the cfDNA read depths, and the probability of success is the reference methylation values multiplied by the estimate of cell type proportions for a given iteration. Maximum likelihood optimization was performed using the L-BFGS algorithm in the SciPy Minimize package.

MethAtlas was run using code available at https://github.com/nloyfer/meth_atlas commit #0223493. It was run using the following command: deconvolve.py -a <reference path> <ouput directory> <samples path>.

2.4.6 ALS Subjects

ALS patients were recruited jointly from the University of California San Francisco ALS Center and the University of Queensland ALS clinics under clinician supervision. All participants provided informed consent and the study was approved both by the Human Research Ethics Committee at the University of Queensland (IRB 2018002470) and by the UCSF Committee on Human Research (IRB 10-05027).

12 cases and 12 controls from San Francisco and 4 cases and 4 controls from Queensland were included in this study. Controls were from non-related family members or caregivers. cfDNA was extracted after subjects were at rest for more than 30 minutes to prevent possible confounding from exercise. We collected 20 mL of whole blood from controls and 10 mL from cases, to allow for further analyses.

2.4.7 ALS cfDNA Sequencing

Whole blood was collected in PAXgene Blood ccfDNA tubes (Qiagen, Cat. No. 768115) and centrifuged at 1,900 x g for 10 min at RT to isolate plasma. Plasma was centrifuged twice at 16,000 x g for 10 min and stored at -80degrees C until cfDNA extraction. Circulating cfDNA was extracted from 4 ml (ALS patients) or 8 ml (controls) of plasma using the QIAamp Circulating Nucleic Acid kit (Qiagen, Cat. No. 55114). Larger volumes of control blood were collected to ensure equal amounts of total cell-free DNA (compared to patients) were analyzed. cfDNA quality and concentration were assessed with an Agilent 2100 Bioanalyzer, using the Agilent High Sensitivity DNA kit (Agilent, Cat. No. 5067-4626). 10 ng of cfDNA were bisulfite-treated and purified using the EZ DNA Methylation-Direct Kit (Zymo Research Cat. No D5020). Libraries for whole genome bisulfite-sequencing were generated using Accel-NGS® Methyl-Seq DNA Library Kit (Swift Biosciences, Cat. No. 30024) and Accel-NGS Methyl-Seq Dual Indexing kit (Swift Biosciences, Cat. No. 38096), with 8 cycles of indexing PCR. Libraries were quantified by qPCR with the Hyper Library Quantification kit (Kapa, Cat. No. KR0405) and paired-end sequenced on a NovaSeq 6000 System (Illumina).

ALS cfDNA Data Processing

Our ALS case-control WGBS data (including both the UCSF and UQ data) were processed according to the ENCODE consortium guidelines [**?**]. Quality of the fastq files was assessed using FastQC (v 0.11.9) [3]. All samples had average phred scores ≥ 28. Adapters were trimmed from the paired end fastq files using TrimGalore (v 0.6.6). Four basepairs were trimmed from the 5' direction and 12 base pairs were trimmed from the 3' direction. Trimmed fastq files were mapped to a bisulfite converted hg38 genome using the Bismark (v 0.23.0) implementation of Bowtie2 (v 2.3.5.1). CpG methylation was from a Samtools (v 1.7) sorted Bismark generated bam file using MethylDackel (v 0.5.0). For this study we were only interested in CpG methylation, which is largely symmetric. Thus, we combined reads on each strand, using the MethylDackel "–mergeContext." option. To standardize methylation calls across all WGBS data sources, hg38 coordinates were reported as 0-indexed. All packages were installed using Anaconda (v 4.9.2). For more details, see `https://github.com/christacaggiano/ENCODE_WGBS`.

2.4.8 Pregnancy cfDNA Data Processing

Data from pregnant women and non-pregnant controls were taken from Jensen et al. at. Raw fastq files from were retrieved from dbGaP identifier phs000846. To ensure consistency across cfDNA samples, data was processed identically to 2.4.7. In the original Jensen et al. study design, multiple fastq files mapped to one sample. Thus, after methylation calling, we combined the appropriate methylation bed files into one per individual, for a total of 15 bed files.

2.4.9 WGBS Simulation Data

Ten adult (small intestine, pancreas, monocyte, stomach, tibial nerve, macrophage, memory B cell, adipose, neutrophil and T cell) WGBS bedMethyl files were obtained from the ENCODE and BLUEPRINT project [**?**][50] (Data identifiers described in Supplementary Table 3). BLUEPRINT data was downloaded as two bigWig files, a methylation signal bigWig and a coverage of methylation signal bigWig. These files were combined into one bedgraph-format file using the UCSC bigWigToBedGraph utility.

Each WGBS file had two biological replicates coming from distinct people. All bed file coordinates were harmonized to hg38 using hgLiftOver [80]. For each tissue or cell type, the file was restructured to report the number of methylated reads and read depth for each CpG locus. Coordinates were standardized to be zero-indexed.

2.4.10 WGBS Reference Data

Reference data for the real cfDNA decomposition experiments in 2.2.6 and 2.2.7 were retrieved from ENCODE and BLUEPRINT. Twenty tissues and cell types were chosen to be representative of the many tissues possible in cfDNA. To decrease noise, we combined two replicates of the tissue when available (see Supplementary Table 3 for individual accession numbers). As described previously, we mapped all data to hg38, and converted the coordinates to be 0-indexed.

2.4.11 Site Selection and Summing

2.4.11.1 Tissue informative markers

Only about 20% of autosomal CpGs vary by cell type [193]. Selecting sites that do vary enriches for information on tissue of origin and reduces the EM computational burden, which scales linearly in the number of sites. We propose selecting tissue informative markers (TIMs) without curation, an approach inspired by ancestry informative markers in population genetics [85] [146].

After processing (Section 2.4.9) the WGBS files, one replicate per tissue was segregated into a reference matrix. This reference matrix was used to calculate TIMs. We assess whether a CpG is a TIM one locus at a time. For each CpG, the distance between the percent methylation of that cell type and the median percent methylation for that CpG was calculated. Only CpGs where the median depth was greater than 15 and had no missing data were considered. The top N (default=100) CpGs with the greatest distance per cell type were selected. TIMs provide increased accuracy in decomposition, and vastly improve computation time. We reference cell types to have overlapping TIMs (i.e., one CpG may be a TIM for both pancreas and liver). We combine proximal CpGs (+/-250bp) around TIMs to increase confidence in the methylation state for a particular CpG (see Site Combination). To test the performance of TIMs, we create a complex mixture of ten WGBS samples and calculate 100 TIMs per sample (for a total of 1,000 CpGS). We compared CelFiE decomposition estimates using 1,000 random summed 500bp regions, 1,080 500bp regions published in Sun et al [164], and our TIM regions. For the data set of WGBS mixtures, TIMs perform better than random and better than the Sun et al regions (Figure S8). We believe that TIMs will be especially desirable for downstream applications, where permuting random WGBS CpG sites is not feasible, or in the development of a capture panel (see Section 2.3).

2.4.11.2 Site combination

To demonstrate whether summing sites improves CelFiE's ability to discriminate tissues, we create complex mixtures of WGBS samples, as in the previous section. We either use single TIMs, or add all methylated and unmethylated counts for all CpGs +/-250bp around a TIM. Summing CpGs +/-250 improves the performance of CelFiE (Figure S9).

2.5 Figures

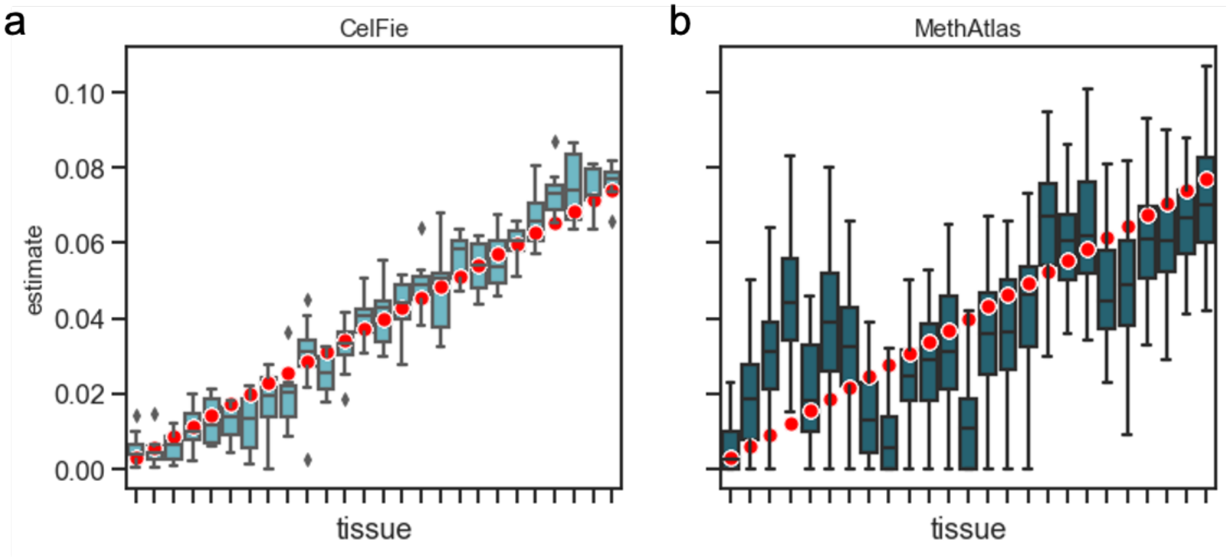


Figure 2.1: **Decomposition of simulated cfDNA mixtures**. Decomposition results by CelFiE (A) and MethAtlas (B). 50 replications for a single simulated individual were performed, and the estimated mixing proportions were plotted (light blue and dark blue boxes, respectively). The red dots indicate the true cell type proportion for each simulated tissue. The center line of the box indicates the mean, the outer edges of the box indicate the upper and lower quartiles, and the whiskers indicate the maxima and minima of the distribution.

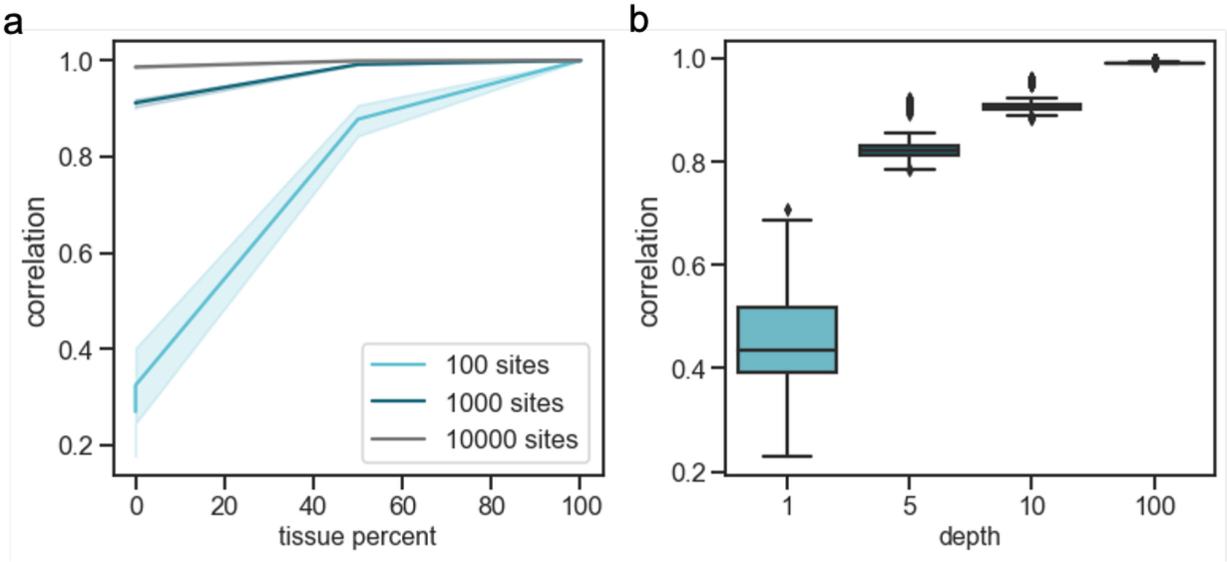


Figure 2.2: **The performance of CelFiE on simulated mixtures.** First, a cell type is fixed at a proportion between 0% and 100%, and reads are simulated for 100 (light blue line), 1000 (dark blue line), and 10000 (black line) CpG sites at 10x depth (A). The Pearson's correlation between the true and estimated cell type proportion is plotted. Solid lines indicate the mean and the shading around the line indicates a 95% confidence interval. On (B) the average Pearson's correlation between the true methylation values for the fixed tissues and the CelFiE estimated methylation values for 1000 sites simulated with 1x, 5x, 10x, and 100x depths (light blue boxes). The center of the boxplot indicates the mean of the distribution, the edges of the box indicate the upper and and lower quartiles, and edge of the whiskers indicate the maxima and minima of the distribution. Data is shown for 50 independent simulations of one individual.

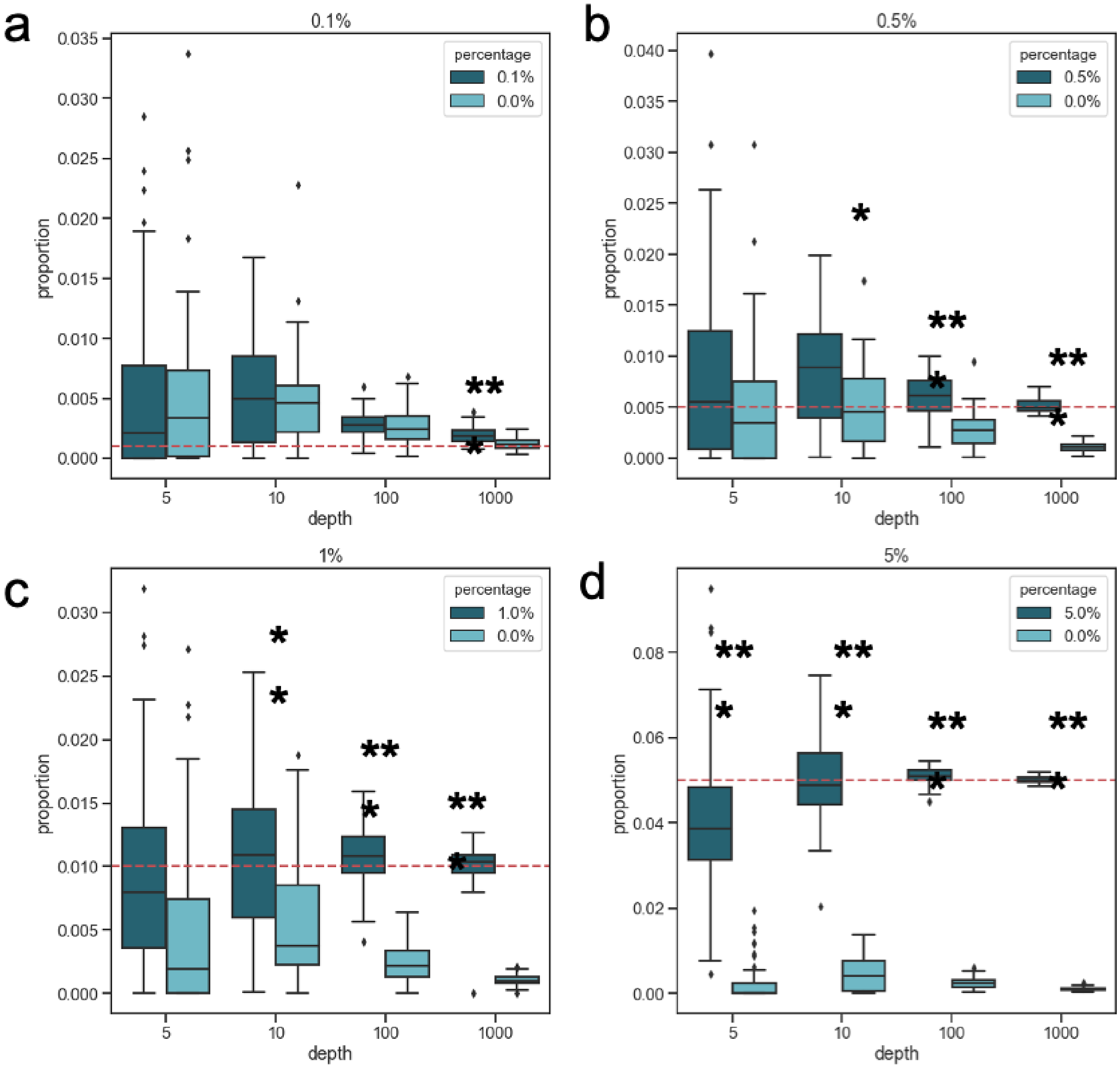


Figure 2.3: **Decomposition sensitivity at low cell type proportions**. Cell type proportion estimates for n=5 simulated individuals (dark blue boxes) with a cell type of interest and n=5 individuals without that cell type (light blue boxes). Cell type proportions are simulated at (A) 0.1% (two-sided grouped t-test; 5x: n.s., 10x: n.s, 100x: n.s., 1000x: p=3.5 × 10−5), (B) 0.5% (two-sided grouped t-test; 5x: n.s., 10x: p=0.013, 100x: 2.1 × 10−6, 1000x: p=5.7 × 10−11), (C) 1% (two-sided grouped t-test; 5x: n.s., 10x: p=1.5 × 10−3, 100x: 2.8 × 10−9, 1000x: p=4.3 × 10−12), or (D) 5% (two-sided grouped t-test; 5x: 4.8 × 10−8, 10x: p=5.4 × 10−9, 100x: 1.8 × 10−14, 1000x: p < 2.0 × 10−16). The true fixed percentage of the cases is indicated by a red dotted line. Significant differences between the groups are indicated by * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$). The center line of the box indicates the mean, the outer edges of the box indicate the upper and lower quartiles, and the whiskers indicate the maxima and minima of the distribution. Data is shown for 50 independent simulations.

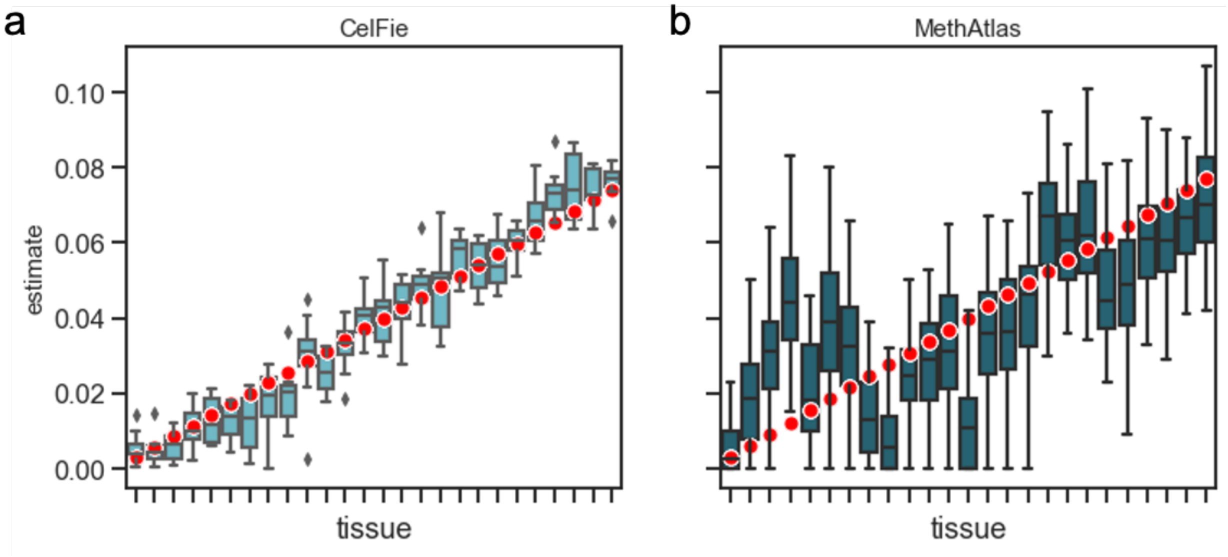


Figure 2.4: **Missing data decomposition simulations.** Decomposition results for 50 independent simulations of cfDNA mixtures with missing cell types in the reference. We simulate cfDNA for 10, 50, 100, 500 and 1000 people, and exclude one cell type truly in the mixture at 20% (light blue) (A) or two cell types (B), one in the mixture at a mean proportion of 20% (light blue), and the other at 10% (dark blue). We calculate the MSE between the true unknown proportion and the CelFiE estimate for 50 simulation experiments. The 95% confidence interval is indicated by the light and dark blue shading.

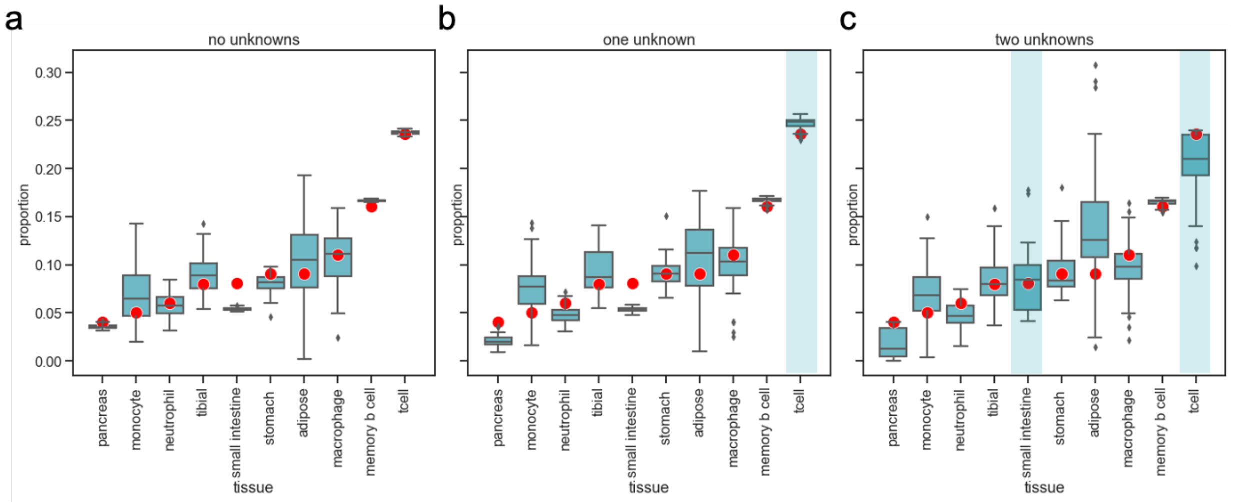


Figure 2.5: **WGBS simulation decomposition**. CelFiE cell type proportion estimates or a randomly selected individual's real WGBS cfDNA over 50 simulation experiments. The blue boxes represent estimates of the true cell type composition (red dots) for 100 individuals in 50 simulation experiments in the scenario where there are no missing cell types (A), when CD4+ T cells are a missing cell type (indicated by blue shading) (B) and when CD4+ T cell and small intestine are both missing (C). The center line of the boxplot indicates the mean, the outer edges of the box indicate the upper and lower quartiles, and the whiskers indicate the maxima and minima of the distribution.

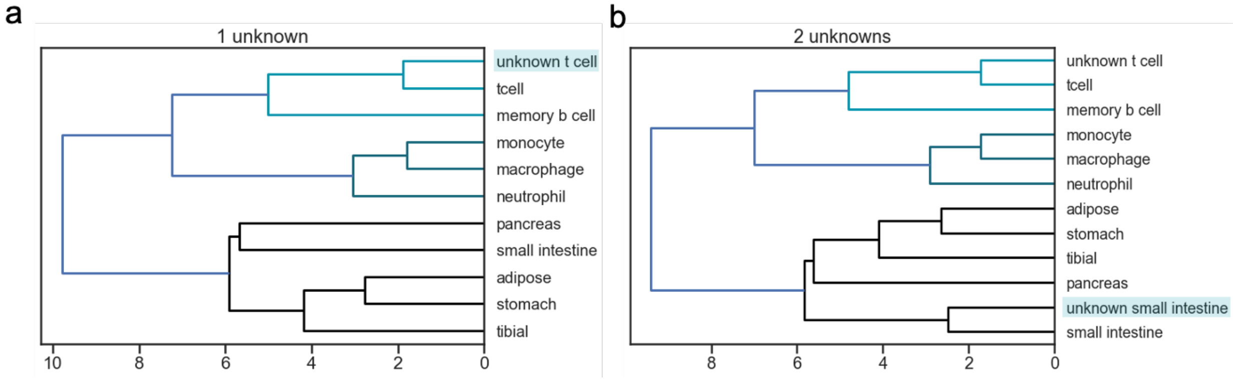


Figure 2.6: **CelFiE unknown hierarchical clustering.** Hierarchical clustering of the CelFiE methylation proportion estimates for (A) one unknown and (B) 2 unknowns with the true WGBS methylation proportions. The shaded blue box indicates the unknown tissue. The light blue, dark blue, and black colors indicate clusters of tissues detected by the hierarchical clustering algorithm.

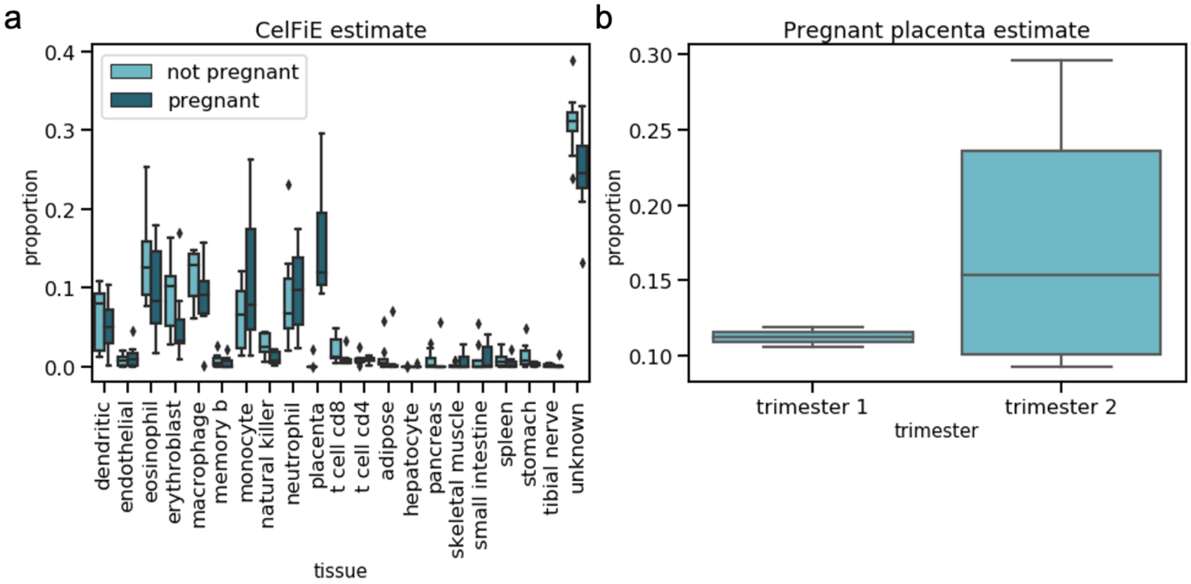


Figure 2.7: **CelFiE estimates for pregnant women**. Decomposition estimates for cfDNA derived from pregnant women and non-pregnant controls. (A) CelFiE decomposition estimates for independent samples of n=8 non-pregnant (light blue) and n=7 pregnant women (dark blue). (B) CelFiE placenta estimates for n=3 pregnant women in the first trimester and n=4 women in the second trimester. In all cases, the center line of the boxplot indicates the mean, the outer edges of the box indicate the upper and lower quartiles, and the whiskers indicate the maxima and minima of the distribution.

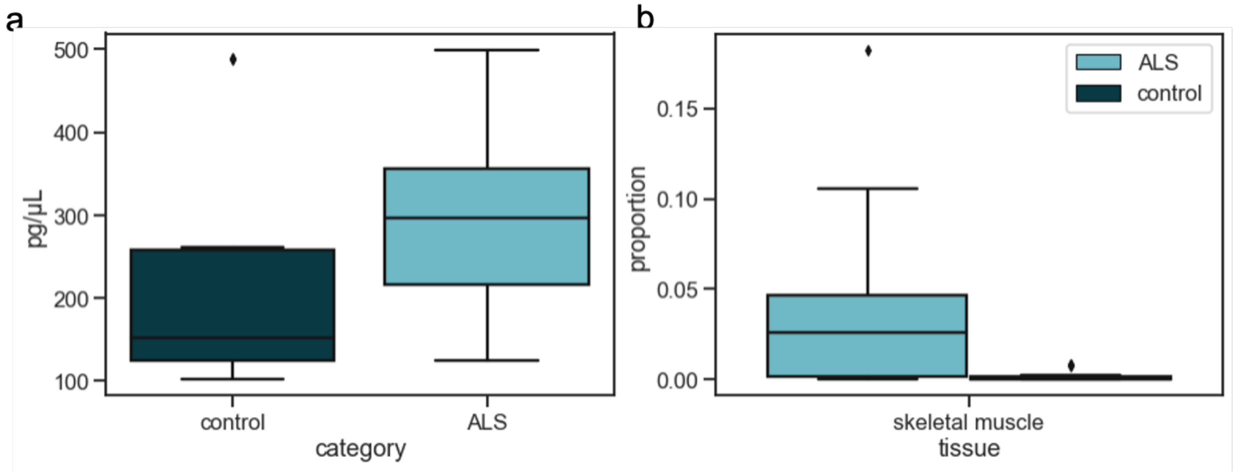


Figure 2.8: **CfDNA concentration and decomposition estimates for ALS patients and age matched controls.** (A) CfDNA concentrations for n=28 independent cases and n=25 independent controls and (B) CelFiE skeletal muscle estimates for n=16 ALS patients (light blue) and n=16 controls (dark blue) from both UCSF and University of Queensland. In both panels, the center line of the boxplot indicates the mean, the outer edges of the box indicate the upper and lower quartiles, and the whiskers indicate the maxima and minima of the distribution.

CHAPTER 3

Non-invasive cell-free DNA biomarker discovery in amyotrophic lateral sclerosis

3.1 Introduction

Amyotrophic lateral sclerosis (ALS) is a devastating neurodegenerative disease characterized by the progressive death of motor neurons [63]. There is currently no cure for ALS and it has an average life expectancy of only 2 to 5 years after diagnosis [162]. A significant challenge in the study of ALS is its heterogeneity [133]. Patients can vary in their disease onset site, overall symptoms, and survival time, making the diagnosis and treatment difficult [82]. Significant global research effort has focused on the development of a quantitative biomarker for ALS, which could reduce the time to diagnosis [159], assist in longitudinal monitoring [33], and facilitate clinical trials seeking to develop effective ALS treatments [149][175].

While there is currently no biomarker used regularly in the clinic, several candidates have been proposed. Recent efforts have especially focused on minimally invasive biomarkers, which can be easily integrated into existing clinical workflows [159]. Neurofilament proteins, detected in both cerebral spinal fluid and serum, have been shown to be helpful in both detecting the presence of ALS and in predicting patient prognosis [16][102][54][177][191]. Circulating microRNAs, which can be detected in the blood, are another class of potential biomarkers [129][111]. In particular, a microRNA enriched in neurons, miR-181, has been demonstrated to be enriched in ALS patients relative to controls and can predict patient survival [104].

Our recent work has highlighted plasma cell-free DNA (cfDNA) as an alternative biomarker for ALS [24]. CfDNA fragments are released into the blood and originate from dying cells throughout the body [161][37]. Fragments can be deconvolved into the cell or tissue type of origin via their DNA methylation patterns [101][119][160], which in the context of an illness, can be used to learn about cell-specific death in disease [89]. Previously, we observed increased cfDNA originating from skeletal muscle in ALS patients relative to healthy controls, consistent with muscle atrophy that is characteristic of ALS [24]. However, since cfDNA originates from both diseased and non-diseased tissues, it has the advantage of learning about cellular health throughout the body. Recent studies have highlighted the role of inflammation [113], [67], and the microbiome [18] in ALS, both of which can be effectively captured by cfDNA [28]. Therefore, cfDNA can be used to learn a multidimensional picture of disease, beyond a specific neurological context.

Since DNA methylation is cell-type specific [193], it is an ideal modality for understanding the cellular contributions to cfDNA. In this work, we develop a scalable targeted sequencing approach, designed to target regions of DNA methylation informative for tissue and disease status (Fig. 3.1). This capture approach has the benefit of enriching only for methylation sites that vary between tissues, thus reducing costs relative to whole-genome bisulfite sequencing (WGBS). Furthermore, since cfDNA is only present in the blood in low quantities, we designed a methodology that requires a minimum input of 10ng of DNA, which is less than other low-cost methylation screening assays like methylation chips.

We apply this technology to two independent cohorts of cfDNA from ALS patients and age-matched controls, one set of 96 patients from the University of Queensland, Brisbane, Australia (UQ), and another set of 96 patients from the University of California San Francisco, San Francisco, United States (UCSF).

3.2 Results

3.2.1 Cohort characteristics

The two cohorts from UCSF and UQ had substantial phenotypic heterogeneity in the characteristics of the patients, both within a cohort and between the cohorts (Fig. 3.2a-b). The median age at the time of cfDNA collection was slightly older for the UCSF cohort, although the age range of patients enrolled in this cohort was larger. There was a greater proportion of male patients at UQ. While the majority of patients in both cohorts were primarily of European genetic ancestry, there were also patients with Asian, African, and American continental genetic ancestry in the UCSF cohort.

There was also variation in the clinical characteristics of the ALS patients in the two cohorts (Fig. 3.2d-f). The median age of diagnosis of ALS patients at UCSF was 65, which was slightly older than the UQ median age of 58. However, UCSF patients were diagnosed as young as 26 and old as 83. A main difference between the two cohorts was disease severity. In general, patients at UCSF had lower ALS Functional Rating Scale-Revised (ALS-FRS-R) scores at the time of cfDNA collection. The ALS-FRS-R scale qualitatively measures physical functioning, such that lower scores mean a patient is less able to complete normal daily tasks [29]. Furthermore, the time since diagnosis was generally longer for UCSF patients. This indicated that, on average, patients recruited at UCSF had more severe degeneration than patients from UQ.

Our previous work introduced the concept of tissue informative markers (TIMs) as a method to identify methylation sites that vary between tissues and cell types. Briefly, a TIM is a site that is either hyper- or hypo- methylated relative to the average methylation value of all other tissues at that site (Fig. 3.3a) (See Methods for more details). Based on tissues identified as present in the cfDNA in our previous work and other recent work, we selected 19 tissues to identify TIMs (Table 1). Tissue methylomes were obtained from two public reference consortiums, ENCODE and Blueprint, and included several white blood cells, large organs, epithelium, and brain. In this work, we use TIMs to prioritize regions of the epigenome for capture. We focused on CpG sites, as most non-CpG methylation sites are not methylated in adult tissues.

TIMs could be any CpG in the genome, however, we applied several filtering criteria to enrich for sites that would most likely be captured in cfDNA. To ensure that potential TIMs were commonly observed in cfDNA, we used our previously published WGBS cfDNA data from ALS patients and controls, along with WGBS cfDNA from pregnant women. We kept only the TIMs that had at least an average read depth of 10x across both WGBS datasets. Furthermore, we removed TIMs that overlapped a common SNP (minor allele frequency ¿5%) and TIMs that were less than 500bp from another TIM.

An important property of cfDNA is that their fragmentation patterns are non-random. cfDNA observed in blood generally are fragments approximately 160bp long, corresponding to the length of DNA wrapped around nucleosomes. This suggests that cfDNA fragments are protected from degradation in the blood by the presence of tightly-associated histone proteins. Since DNA from compacted chromatin is more likely to be protected and methylated, we would expect hypermethylated cfDNA fragments to be likely to be observed in the blood. As such, we chose to select a greater proportion of TIMs that were hypermethylated relative to other tissues, with an average of 300 TIMs per tissue (Table 1). Of those 300 TIMs, an average of 150 were hypermethylated and 50 were hypomethylated relative to the other tissues (Fig. 3.3b).

After quality control (Methods), 5,666 TIMs were selected. TIM sites were distributed throughout the genome, excluding the Y chromosome. Hypermethylated TIMs were closer to transcription start sites and CpG Islands than hypomethylated TIMs (Fig. 3.2c). This is consistent with the requirement that at a hypermethylated TIM, all other tissues at that site are unmethylated. Unmethylated sites are more likely to be involved in genome regulation. Likewise, hypomethylated TIMs were more likely to be in intergenic and intronic regions (Fig. 3.2d). Together, this suggests that hypermethylated and hypomethylated TIMs offer complementary types of genomic information.

3.2.2 Capture panel validation

For each of the 5,666 TIMs, both a methylated and unmethylated probe was designed to bind to and capture both possible states of the targeted CpG. To increase the efficiency of the capture, probes were designed to target a window of 120bp around the TIM. During bisulfite conversion, any cytosine base not protected by a methyl group in position 5 is converted into thymine. Since methylation in humans primarily occurs at CpG sites, this means that all cytosines on the forward strand would be converted to thymine. Thus, to capture the unmethylated CpG state, the unmethylated probe was designed with all guanine bases converted to adenine. For the methylated state, where only cytosines in a CpG dinucleotide would be protected from the bisulfite treatment, only non-CpG guanine bases were converted to an adenine.

To ensure that the methylation capture panel protocol could accurately profile the methylation state of the chosen CpGs, we performed several validation experiments. First, we used universal methylated DNA standards to create mixtures where the CpG sites were methylated 0, 25, 50, and 100% of the time. We then assessed the methylation proportion measured after capture. We found that the observed methylation was highly concordant with the true methylation proportion (Fig. 3.3e). Next, to examine how the capture panel might perform in real-world cfDNA scenarios, we validated the capture panel using sheared genomic DNA from blood, along with healthy cfDNA samples. The methylation proportions for these samples were significantly correlated with methylation from white blood cells, as expected given that healthy cfDNA arises primarily from blood cell turnover. We performed deconvolution on these samples and confirmed that the majority of cfDNA was originating from neutrophils and lymphocytes, consistent with published research. Together, these experiments demonstrate that our approach for targeting TIMs can correctly capture the methylation state of cfDNA.

3.2.3 cfDNA capture

cfDNA was extracted from both cohorts. We confirmed our previous finding of elevated cfDNA in ALS patients relative to controls (Fig. 3.4a) (UQ logistic p-value=2.0x10-2, UCLA logistic p-value=2.0x10-3). As expected due to the relatively low input cfDNA quantity, we noted a relatively high deduplication rate after first sequencing UQ samples (average=X). However, since standard position-based deduplication protocols might be overly conservative for short cfDNA fragments, to better quantify the true methylation state, we added a unique molecular identifier (UMI) to UCLA samples for sequencing. We found that UMIs recovered approximately 20% more reads on average relative to position-based deduplication.

In total, after sequencing, the average on-target coverage of UQ samples was 55.29 reads per CpG. The average on-target coverage of UCLA samples was 208.75 reads per CpG. For both cohorts, this was higher per-CpG read coverage than WGBS at an analogous number of reads (Fig. 4b-c). The average methylation proportion at TIM sites was highly correlated (Pearson's R2=0.98) between the two cohorts (Fig. 3.4d), suggesting that even with the different deduplication strategies, similar methylation profiles were obtained. Similar to the validation experiments, we observed that hypermethylated TIMs were less methylated in both cohorts (Fig. S).

3.2.4 Correlation with disease status

To examine the differences between the ALS case and control methylation patterns, we first performed dimensionality reduction over the captured TIMs using principal component analysis (PCA). In both the cohorts, cases and controls separated in PC space and PC2 was significantly associated with case/control status in both cohorts (Fig. 3.55a-b) (UQ logistic p-value: 2.00x10-2, UCLA logistic p-value: 7.40x10-5). This suggested that there were widespread differences in the cfDNA methylation patterns of ALS cases and controls.

The separation difference between cases and controls in PC space motivated the development of a prediction algorithm that integrated information from all targeted CpGs to classify

ALS cases. A prediction algorithm may have value in assisting in physician diagnosis but also, the probabilities associated with a binary prediction task can be used to identify ALS patients who more closely resemble controls, which may have value in understanding disease severity or progression. To do this prediction, we use Lasso regression, implemented in the BigStatsR package. Lasso regression is a regularization technique that incorporates feature selection and regularization to enhance predictive accuracy and model interpretability.

We used methylation proportions as input to the Lasso algorithm, with missing data imputed using SoftImpute. Non-penalized covariates included age, genetic sex, genetic ancestry, sequencing depth, and initial cfDNA concentration. First, we trained the model on UQ samples only and implemented 10-fold cross-validation to select model parameters. We then repeated this process for the UCSF data. The within-cohort area under the precision-recall curve (AUC) for the UQ samples was high, with AUC=0.82. For the UCSF within-cohort model, the AUC was even higher, AUC=0.98. The higher AUC in the UCSF model is likely attributed both to the introduction of UMIs to the sequencing for these samples, which likely decreased noise, and also that this cohort had more severe disease, which could facilitate differentiation between cases and controls.

The UQ-trained model was then tested on held-out UCSF samples. The model had a high AUC of 0.88, which was even higher than the accuracy of the within-UQ model. For the UCSF trained model applied to UQ data, the AUC was lower, 0.74 (Figure 5c). This difference might be again attributed to the less severe disease present in the UQ ALS patients. We also calculated area under the precision recall curve (AUPRC) and found that AUPRC was high in both cohorts, but slightly lower in the UCLA trained model. Despite differences in performance, overall, the relativley high transferability suggested that the cfDNA methylation patterns could be used to learn about ALS disease in independent cohorts.

Lastly, we performed cfDNA deconvolution with CelFiE. CelFiE is a supervised cell-type decomposition algorithm that is designed to work with methylation read count data and missing or noisy reference data. As input, CelFiE takes cfDNA read count data, and estimates the proportion of the cfDNA mixture originating from the tissues in the reference dataset, along with a specified number of unknown tissues. We ran CelFiE using the targeted TIM sites as input, and the set of reference tissues that the TIMs were designed for as the reference sample. In the UCSF cohort, we confirmed our previous finding of elevated skeletal muscle cfDNA in ALS patients (logistic p=0.03), with ALS patients having 2.0% of their cfDNA originating from skeletal muscle, and controls having only 0.8%. In the UQ cohort, the difference between cases and controls was not significant (p=0.38), although cases, on average, had slightly more of their cfDNA originating from skeletal muscle, 1.3% versus 1.0%. As UQ skeletal muscle estimates from ALS patients were closer to the average from UCSF controls, the difference might be explained by patients with less muscle degeneration in the UQ cohort.

3.3 Discussion

Here, we present a scalable cfDNA capture protocol that measures the methylation status of disease and tissue-relevant CpG sites. We applied this capture technology to two independent cohorts of ALS patients and age-matched controls and examined the correlation with ALS disease status and progression. We found that cfDNA can significantly predict ALS disease status in both cohorts and the prediction models are transferable between cohorts. Furthermore, cfDNA tissue of origin deconvolution confirmed skeletal muscle as being elevated in the cfDNA of ALS patients, which indicated that cfDNA can learn about disease-specific degeneration. We conclude that cfDNA has the potential to be a clinically relevant biomarker for ALS, with value in disease diagnosis and quantitative measurement of progression.

One of the key strengths of this work is that cfDNA can provide a comprehensive picture of a patient's biological state and is not limited to a specific tissue or context. For example, many biomarker candidates for ALS focus on biomolecules obtained from neurological tissues, like neurofilaments. However, many proposed biomarkers can be found in other neurodegenerative disorders or even non-neurological conditions. This lack of specificity makes it difficult to differentiate ALS from other diseases, leading to potential misdiagnosis or limited accuracy in disease monitoring. In this work, we show that cfDNA from skeletal muscle is a specific predictor of ALS disease status, which may not be elevated in other neurological disorders. However, by capturing and quantifying methylation levels at multiple tissue-informative CpG sites simultaneously, the panel has the potential to also learn about biological processes occurring in ALS outside of neurodegeneration. In particular, cfDNA is well-suited to measuring inflammation, which has been of recent interest in ALS pathophysiology. Future work with deeper clinical phenotyping could provide additional insight into how cfDNA relates to inflammatory markers in ALS, providing a complementary avenue for investigation into disease mechanisms.

cfDNA is also a valuable biomarker candidate because it is non-invasive. Many biomarker candidates require invasive procedures, such as cerebrospinal fluid (CSF) or muscle biopsies, which can be burdensome, costly, and carry associated risks. Furthermore, the protocol presented here was designed for low DNA input, which also has the advantage of not requiring copious amounts of blood for patients. The assay could be easily integrated into existing clinical workflows, using discarded patient samples already collected during care. Furthermore, the low cost of capture relative to whole-genome sequencing makes it a viable candidate for implementation in the clinic, especially for longitudinal monitoring.

Several considerations should be taken into account when interpreting the results of this study. Firstly, the sample size used for panel evaluation might limit the generalizability of the findings. ALS is an extremely heterogeneous disease and it is likely that all potential subtypes of the disease were not well represented. Furthermore, both cohorts were of primarily European ancestry. To ensure the robustness and reliability of the panel's performance, further validation on larger and more diverse cohorts is warranted.

Next, while methylation capture arrays allow for a more cost-effective and focused analysis over relevant CpG sites, targeted capture also limits the coverage of the genome. This has the potential to miss important methylation changes occurring outside the targeted regions. Additionally, since we relied on published tissue methylation data sets that are low coverage and inherently noisy, TIM selection might be affected. Marker selection and overall algorithm performance might be improved by better, high-coverage reference data. Reference panel design for cfDNA applications is a robust area of current research, and incorporating new samples or biobanks into ALS disease prediction could be an area for future research.

Overall, the design of the cell-free DNA methylation capture panel presented in this study represents a significant advancement in the field of ALS research. The panel demonstrates promising potential as a non-invasive and diagnostic tool for ALS, which could facilitate timely intervention and personalized treatment strategies. Further research and validation are necessary to refine the panel's performance, assess its generalizability, and address practical considerations. Nonetheless, this study paves the way for the integration of DNA methylation biomarkers into the clinical management of ALS, bringing us closer to improved patient outcomes.

3.4 Methods

3.4.1 Patient Recruitment and Clinical Data

ALS patients were recruited from the UCSF ALS Clinic in San Francisco, California, USA and the Royal Brisbane and Women's Hospital in Brisbane, Australia under neurologist supervision. All participants provided informed consent and the study was approved both by the Human Research Ethics Committee at the University of Queensland (IRB 2018002470) and by the UCSF Committee on Human Research (IRB 10-05027).

Age-matched healthy controls were recruited from non-related family members or caregivers. In total, 46 ALS cases of varying disease stages were obtained from each site along with 46 controls. For cases and controls, age, sex, and self-reported race/ethnicity were recorded. At UCSF, for ALS cases at the time of visit, FVC and ALSFRS-R were taken, and ALSFRS-R slope and FVC slope relative to the previous visit were calculated. The symptom onset site and date of first symtoms were also recorded. For the UQ samples, ALSFRS-R was recorded and progression was calculated as ((48 enrollment ALSFRS-R)/time (in months) from symptom onset to enrollment).

cfDNA was extracted from patients after a 30min rest period to prevent cfDNA originating from exercise. At each site, 20mL of whole blood from controls and 10mL of whole blood from cases, were extracted.

4.2 Library Preparation and Sequencing cfDNA from 2-8 ml of plasma was extracted using the QIAGEN Circulating Nucleic Acid kit according to the manufacturer's recommendations. Extracted cfDNA was quantified using Qubit dsDNA HS Assay and visualized using the cfDNA assay (Agilent - TapeStation 4200). cfDNA was bisulfite converted using the Zymo Lightning kit (Zymo Research) and underwent library preparation using the Accel-NGS Methyl-Seq (Swift Biosciences) according to the manufacturer's instructions with a major modification. Briefly, the denatured BS-converted cfDNA was subject to the adaptase, extension, and ligation reaction. Following the ligation purification, the DNA underwent primer extension (98C for 1 minute; 70C for 2 minutes; 65C for 5 minutes; 4C hold) using oligos containing random UMI and i5 barcodes. The extension using a UMI-containing primer allows the tagging of each individual molecule in order to be able to remove PCR duplicates and correctly estimate DNA methylation levels.

Following exonuclease I treatment and subsequent purification, the libraries were then amplified using a universal custom P5 primer and custom i7-barcoded P7 primers (initial denaturation: 98C for 30 seconds; 15 cycles of: 98C for 10 seconds, 60C for 30 seconds, 68C for 60 seconds; final extension: 68C for 5 minutes; 4C hold). The resulting unique-dual indexed libraries were then purified, quantified using the Qubit HS-dsDNA assay, the quality checked using the D1000-HS assay (Agilent - TapeStation 4200), and grouped as 12-plex pools. Each pool was then subject to hybridization capture using the xGen Hybridization Capture Kit (IDT) using custom probes designed on approximately 5000 pre-selected regions. See section 4.4 on how regions were selected for methylation capture.

For each top and bottom strands of the regions of interest, two probes were designed: one "unmethyl" probe with all G bases converted to A, and one "methyl" probe with all non-CpG G bases converted to A.

Following the hybridization capture, a final amplification PCR (initial denaturation: 98C for 30 seconds; 10 cycles of: 98C for 10 seconds, 60C for 30 seconds, 68C for 60 seconds; final extension: 68C for 5 minutes; 4C hold) has been performed, followed by SPRI beads purification and quantification as QC as previously described.

The final pool of libraries has been submitted for sequencing on an Illumina NovaSeq6000 (S4 lane - 150 PE, 8bases for i7, 17 bases for i5). not sure how they sequenced in Australia

3.4.2 Bioinformatics processing

For data generated at UCLA, UMIs were first extracted from the index read and added to the header of the corresponding R1 and R2 fastq file using umitools. This step was skipped for UQ samples since UMIs were not sequenced. For samples from both institutions, adapters were trimmed using trimgalore. Read alignment, processing, and methylation calling were performed using BsBolt v 1.6.1. Reads were aligned to an hg38 bisulfite converted genome, which was generated using the BsBolt Index over an hg38 fasta file obtained from the UCSC genome browser. Reads were aligned using BsBolt Align in paired end mode with default parameters.

To prepare for duplicate removal, aligned reads were subject to samtools fixmate and sorted. At UCLA, UMIs were used to remove duplicates using umi$_toolsdedupinpairedendmode.AtUQ, dupli$

For both cohorts, CpG methylation was called using the command BsBolt CallMethylation -BG -CG -remove-ccgg. The CG parameter restricted to only CpG sites (ignoring non-CpG methylation), the the BG parameter sent the output to a bedgraph file and the -remove-ccgg parameter removed methylation calls in ccgg regions.

3.4.3 Tissue informative marker selection

TIMs were selected for 19 tissues and cell types: dendritic cells, endothelial cells, eosinophils, erythroblasts, macrophages, monocytes, neutrophils, T-cells, adipose, brain, fibroblast, heart, hepatocytes, lung, megakaryocytes, skeletal muscle, small intestine, placenta, and mammary epithelial cells. These tissues were determined based on our previous work to be relevant to

ALS, or selected based on previous publications to be the primary contributors to cfDNA. At least two WGBS samples per reference dataset were obtained. The average methylation per CpG for the reference tissue replicates was calculated.

Per CpG, for one tissue at a time, the distance between the methylation proportion at that tissue and the mean methylation of all other tissues was calculated. The N sites per tissue with the greatest difference were kept as TIMs. If two tissues had the same CpG classified as a TIM, it was removed from both lists.

To begin, we selected 500 potential TIM sites and then performed quality control checks. To ensure that TIMs were sites that would be covered in cfDNA data, we used two WGBS cfDNA datasets and removed any CpG site that had less than an average of 10X coverage in both datasets. We also removed TIMs that overlapped a common SNP (minor allele frequency $> 5\%$). Since we wanted to have the greatest diversity of regions targeted in the capture, if there were multiple TIM sites within 500bp of each other, we kept only the first site.

3.4.4 Deconvolution

cfDNA deconvolution was performed using CelFiE, which is a supervised deconvolution algorithm that is designed for noisy read count data and missing reference tissues. Input sites for CelFiE were the on-target TIMs selected for capture, As demonstrated in the CelFiE publication, summing reads from adjacent CpGs can improve deconvolution performance by decreasing sampling noise. As such, reads were summed +/-250bp around the target CpG. Sites with no reads covering the CpG were set to have a read depth of zero.

Deconvolution was performed using the same tissues used for capture. CelFiE can estimate an arbitrary number of unknown tissues. Since CelFiE learns from both the input and reference data, the number of samples influences the accuracy of unknown estimation. Based on simulation experiments published in the original CelFiE paper, 2 unknowns were chosen for the sample size of 96 total cfDNA input samples.

The reference panel for CelFiE consisted of 19 tissues over the same on-target TIMs as the input matrix. References samples were WGBS samples obtained from ENCODE and Blueprint. Reference samples were also summed in 500bp regions around the target CpG.

CelFiE was run over the samtools deduplicated UQ samples, the UMI deduplicated samples, and both cohorts combined. The CelFIE default of 10 random restarts was used.

After running deconvolution, differences in cell-type proportion between cases and controls were tested for one tissue at a time using the Python StatsModels package. A logistic regression model was run where the outcome was the binary case/control status and the input variable was the estimated tissue of origin proportion for a given tissue. Self-reported age, genetic sex, and genetic ancestry were used as covariates.

3.4.5 Principal component analysis

PCA was performed on the methylation proportions for on-target sites. Reads from CpGs +/-250bp around a TIM were summed. Methylated and unmethylated reads were converted in a matrix of the proportion of methylated reads. Missing values were dropped. Before performing PCA, the matrix was standardized using the StandardScaler() function in the Python SciPy package. PC's were calculated for each cohort independently and 10 principal components were calculated for each cohort. 4.6 Machine learning preprocessing

Samples with more than 10% of targeted CpGs missing, meaning that no reads were covering a CpG, were removed. Any site that had a median read coverage of 1 read or less was also removed. For the remaining sites and samples, the input matrix was made by dividing the number of methylated reads by the total number of reads. Imputation was performed over the methylation proportion matrix using SoftImpute, implemented in the Python package fancyImpute Sex and race/ethnicity were one-hot encoded and added as columns in the input matrix. Age, cfDNA starting concentration, and total sample read depth were also added. Two separate matrices were kept, one for the ALS case/control status, and one for the methylation proportion and covariates.

3.4.6 Lasso regression

Lasso regression was performed in R using the BigStatsR package. We performed logistic regression separately for each cohort. ALS disease status served as the binary outcome variable, while the DNA methylation proportion at targeted CpGs and clinical variables served as predictors. We incorporated relevant covariates, such as age, biological sex, genetic ancestry, cfDNA concentration, and total sequencing depth, into the regression models as non-penalized variables.

Models were first trained on each cohort separately and then applied to the second cohort. The alpha parameter which controls model sparsity, was selected by performing ten-fold cross-validation on the training cohort.

Models were evaluated using area under the receiver operating characteristic curve (AUC) and the area under the precision-recall curve (AUPRC).

3.5 Figures

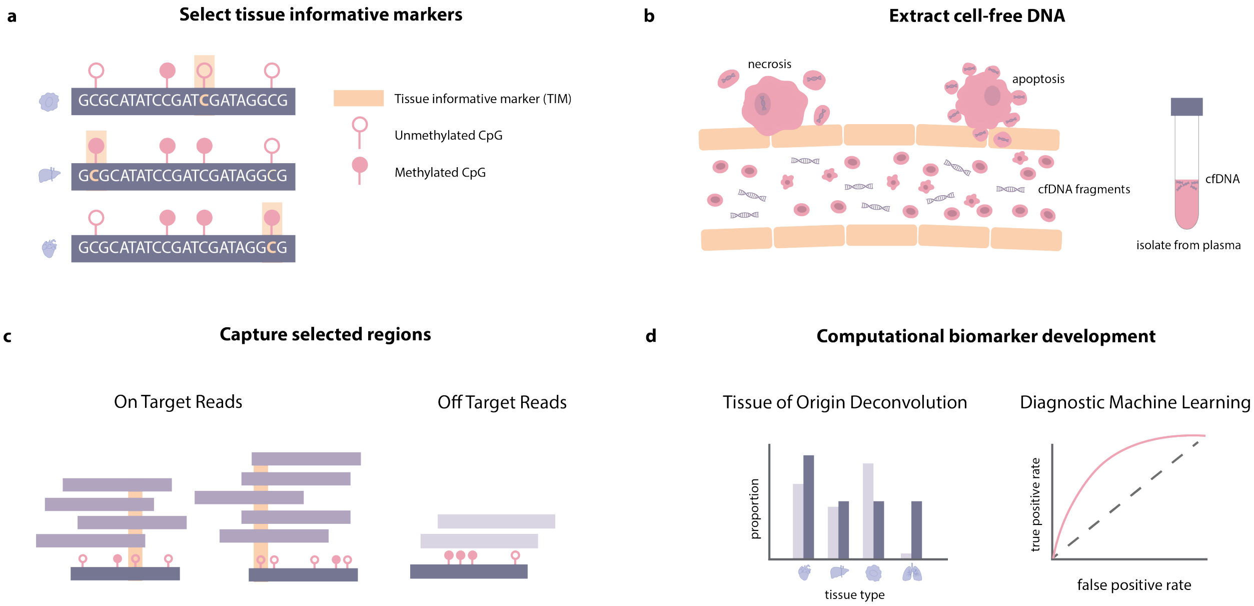


Figure 3.1: **Schematic of cfDNA methylation capture.** (a) First, tissue informative markers (TIMs) are selected to capture CpG sites that are hypermethylated or hypomethylated in a tissue of interest. (b) Next, cfDNA is extracted from the blood plasma of ALS cases and controls. (c) The cfDNA is bisulfite sequenced and amplified at the on-target region around a TIM. Some off-target reads will also be captured. (d) Using bioinformatics, we analyze the tissue of origin of the cfDNA samples and perform machine learning to identify features of disease.

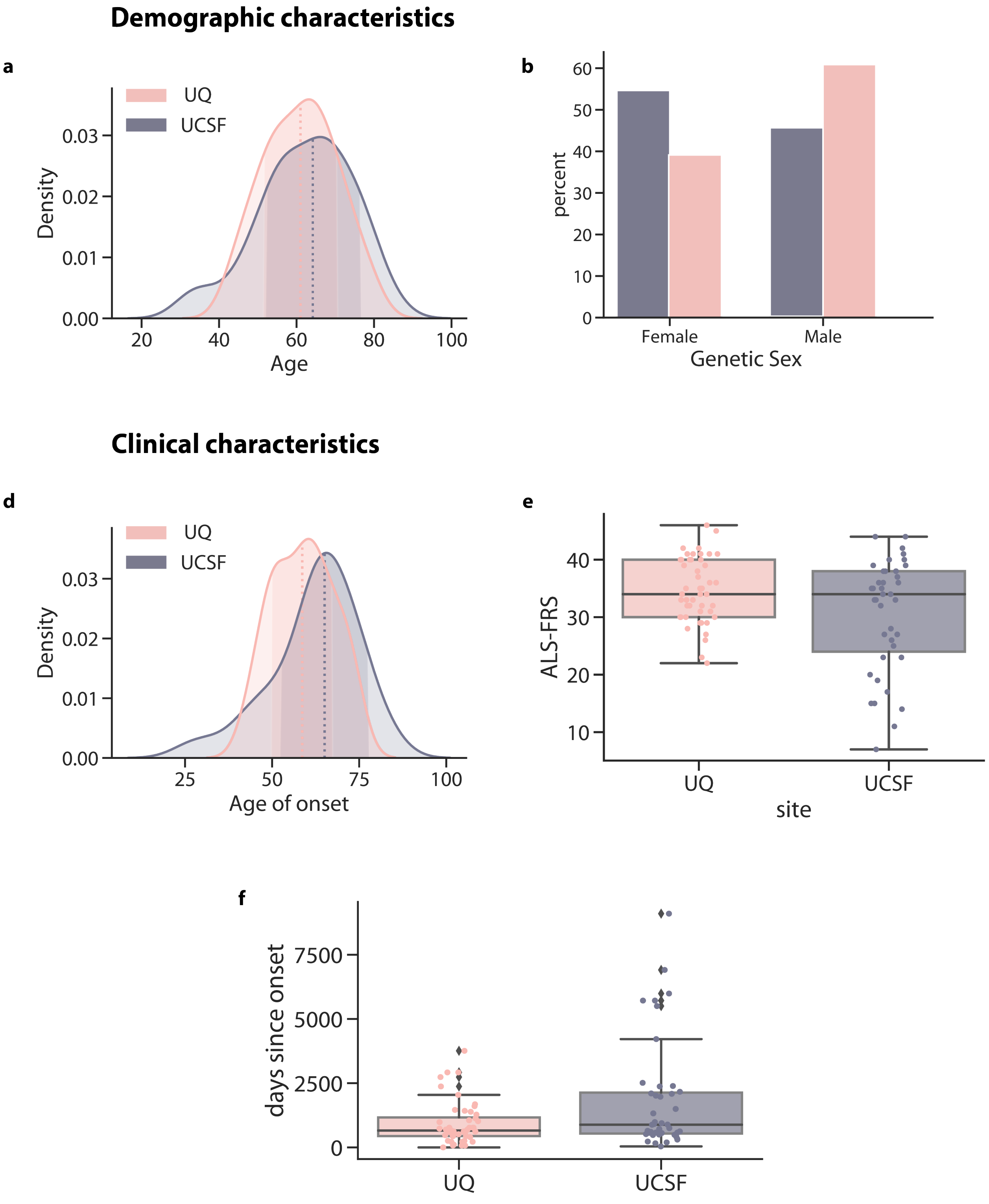


Figure 3.2: **Cohort demographic and clinical characteristics.** For both the cases and controls in the UQ and UCSF cohorts, the distribution of sample (a) ages (b) the number of female and male participants, (c) the sample genetic ancestry. For the ALS samples in each cohort, (d) the age of onset, (e) the ALS-FRS-R at time of cfDNA extraction, and (f) the days of since symptom onset.

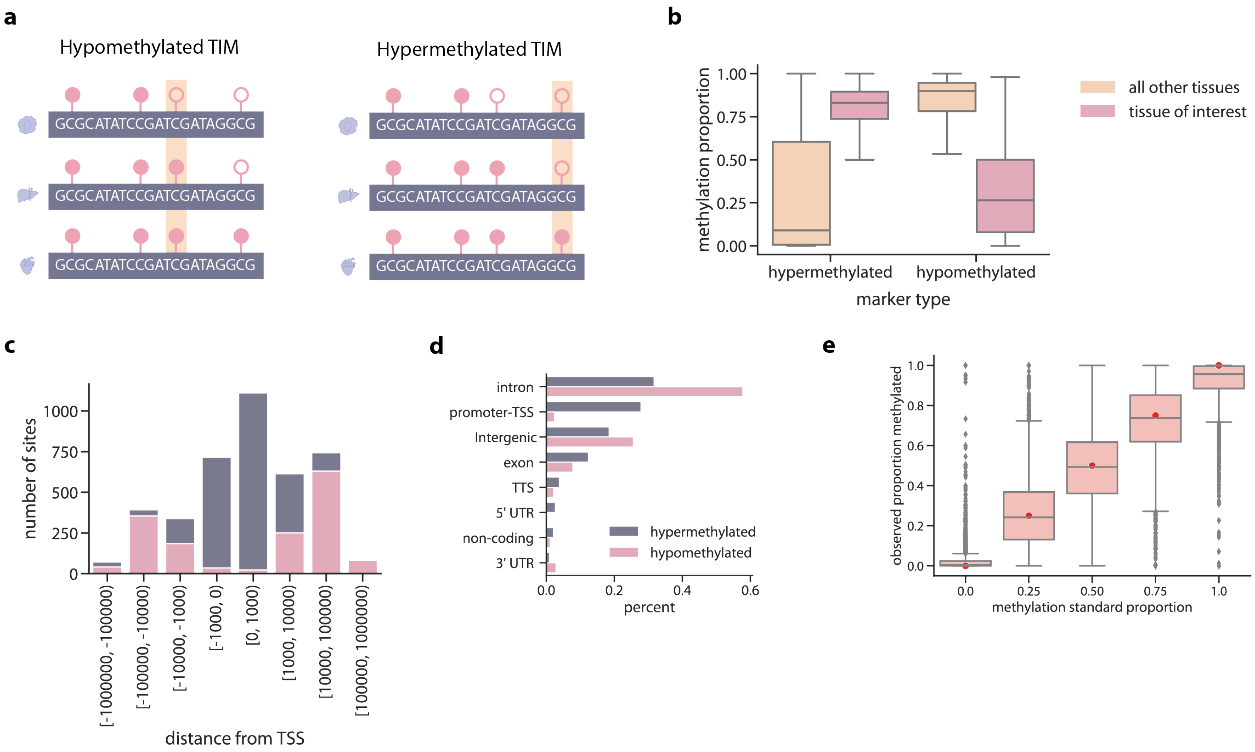


Figure 3.3: **Capture panel design.** (a) The panel was designed to capture both hypomethylated TIMs, which were CpG sites who were less methylated relative to other tissues, and hypermethylated TIMs, which were designed to capture sites more methylated than other tissues. (b) The methylation proportion of reference tissues at either the site the TIM was selected for, or all other tissues. (c) The distance a TIM was from the transcription start site of a gene. (d) The number of hyper- and hypomethylated TIMs in different genomic regions. (e) For a validation experiment where mixtures of DNA with a true methylation proportion between 0 and 1 were captured, the observed methylation proportion.

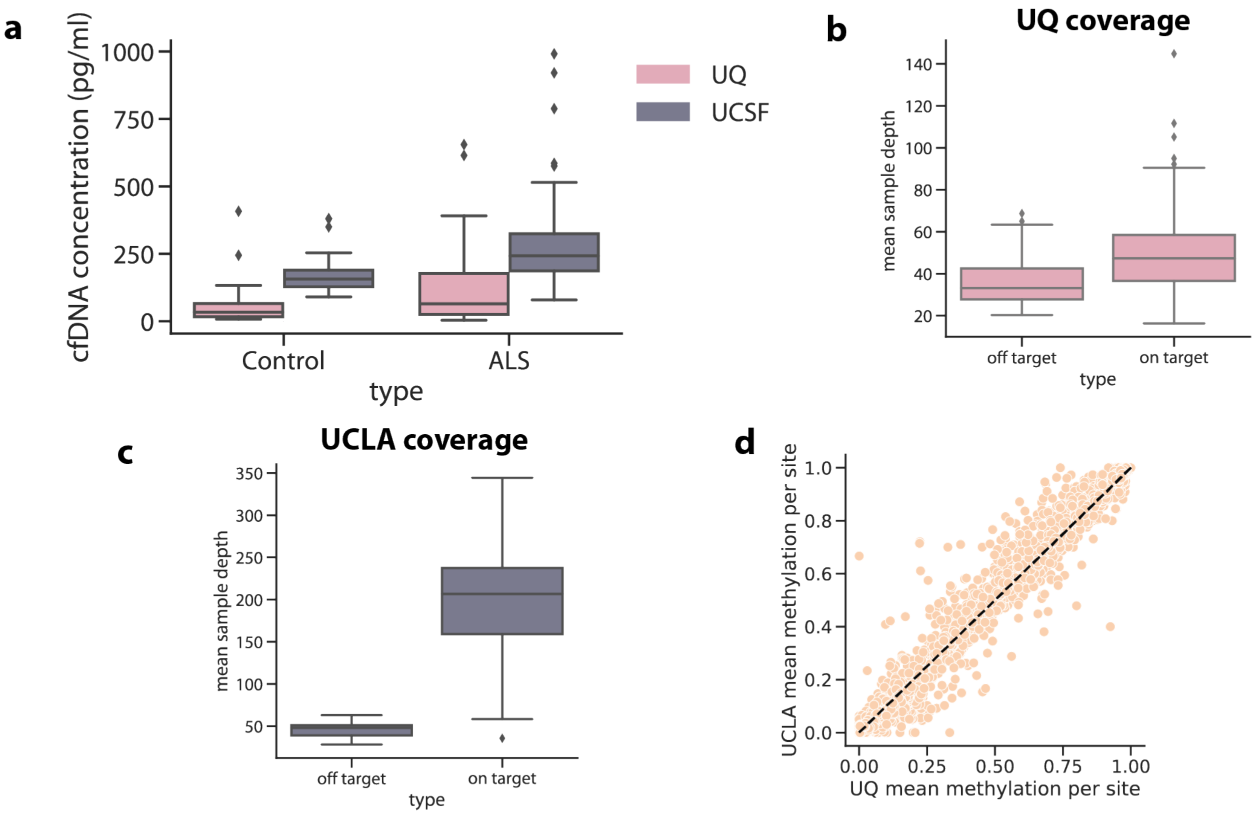


Figure 3.4: **Capture panel performance on cfDNA data.** (a) The starting cfDNA concentration of ALS patients and controls for each cohort. (b) Coverage of the on-target and off-target CpG sites of UQ cohort samples and (c) UCLA cohort samples. (d) Correlation between the UQ and UCLA methylation proportions at on-target sites.

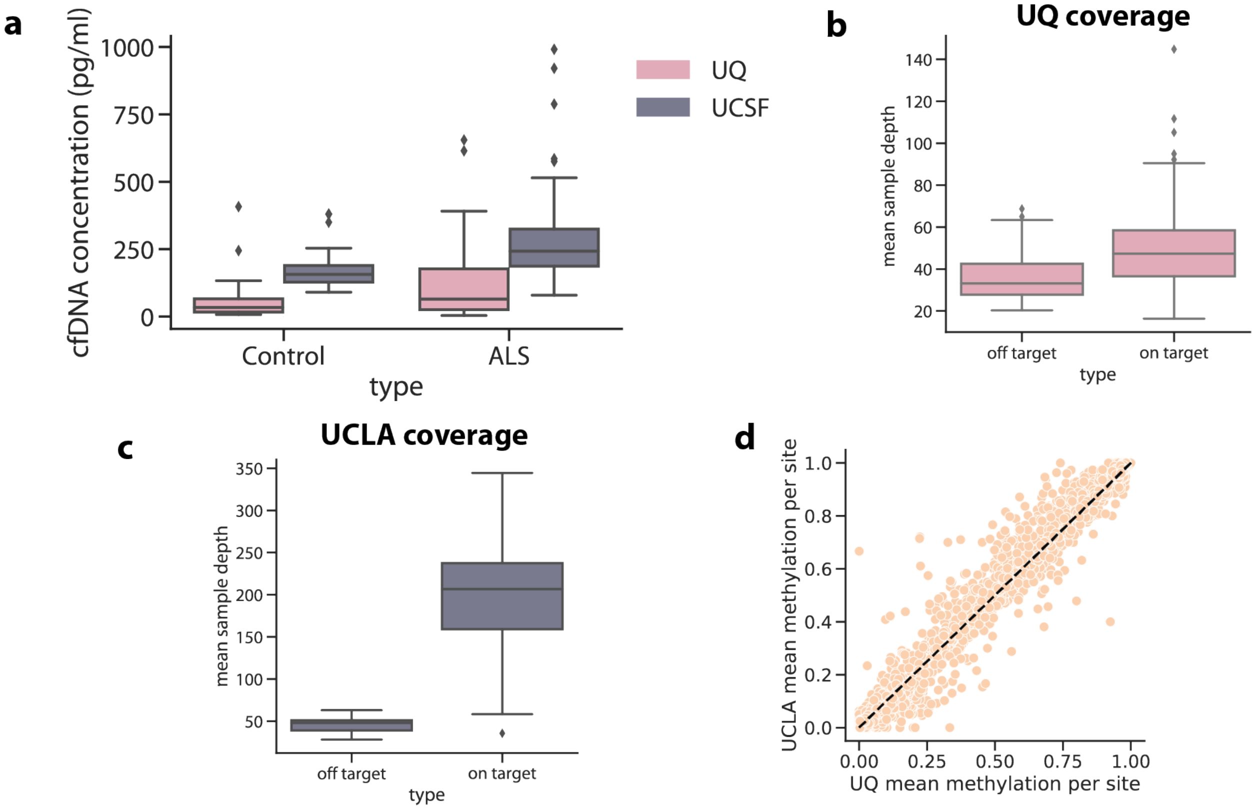


Figure 3.5: **cfDNA methylation disease classification.** (a) Principal component analysis on UQ samples. (b) Principal component analysis over UCSF samples. (c) For the UQ-trained model that was tested on UCSF data, and the UCSF model tested on UQ data, the area under the curve and (d) area under the precision-recall curve.

3.6 Tables

Tissue	Hypermethylated	Hypomethylated	Total
T-cell	274	59	333
Adipose	239	43	282
Brain	313	53	366
Dendritic cell	210	84	294
Endothelial cell	291	60	351
Eosinophil	89	172	261
Erythroblast	136	128	264
Fibroblast	301	51	352
Heart left ventricle	210	48	258
Hepatocyte	283	51	334
Lung	234	46	280
Macrophage	88	167	255
Mammary epithelial cell	232	50	282
Megakaryocyte	208	75	283
Monocyte	74	166	240
Neutrophil	72	190	262
Placenta	251	55	306
Skeletal muscle	308	53	361
Small intestine	249	53	302

Table 3.1: **TIM selection design.** Per tissue selected for capture, the number of hypermethylated TIMs selected, the number of hypomethylated TIMs selected, and the total number of final TIMs selected for capture.

CHAPTER 4

Disease risk and healthcare utilization among ancestrally-diverse groups in the Los Angeles region

4.1 Introduction

Individuals belong to many populations (Box 4.2), each with unique health risks. This can be a consequence of a population's shared cultural or physical environment, genetics, or a combination of both. Structural factors, including racism and socioeconomic status, also shape the health of populations, particularly in the United States [184][51][56]. Therefore, understanding population-level differences in disease risk is important for reducing health disparities and developing personalized interventions [109][105]. New large-scale biobanks tied to electronic health records (EHR) present an ideal opportunity to study population health [126]. Previous biobank studies have identified new genetic associations to complex traits [73] examined how diseases track through families [64], and produced polygenic risk scores for multiple ancestries9.

Our work, and other previous work [10][142][39][57] have used identity-by-descent segments (Box 1) to find fine-scale populations who share genetic ancestry in biobanks. Identity-by-descent segments are identical stretches of DNA inherited from a shared ancestor. People whose ancestors lived in the same geographic location or who were part of the same ethno-linguistic group tend to have a greater proportion of their genome identical-by-descent [65] These clusters of people may also share an environment, including structural factors like discrimination, which can be relevant for understanding why or how patients visit the hospital. We have previously shown that individuals within identity-by-descent-based clusters often share clinical diagnoses[10].

Here, we use identity-by-descent sharing to define fine-scale population clusters and to analyze their health system utilization within the ATLAS Community Health Initiative16 (ATLAS). ATLAS is part of the University of Los Angeles (UCLA) health system located in Los Angeles, a city with a rich history of recent and past immigration [74]. We note that identity-by-descent clusters offer one lens into the study of health outcomes alongside others including socially determined concepts of race and ethnicity (Table 1). We examined the relationship between identity-by-descent clusters and healthcare system utilization inferred from electronic medical records. We identified thousands of cluster-specific health associations and cluster-specific enrichments of clinically actionable genetic variants. To facilitate the use of the large set of associations, we developed a web framework allowing interactive access to the results presented.

4.2 Results

4.2.1 ATLAS Community Health Initiative

The ATLAS Community Health Initiative [73] includes 35,968 patients with genotyping and de-identified EHR data (see Methods)Patients are diverse both genetically, and in terms of EHR reported demographic characteristics [73]. ATLAS demographics are consistent with the overall patient population of UCLA Health, but the demographics of UCLA differ from that of Los Angeles. Socioeconomic factors and racial discrimination strongly influence where people live in Los Angeles, especially as West Los Angeles. contains some of the wealthiest zip codes in the nation according to Census and IRS income data [22]. Despite this, 40% of ATLAS patients identify as a race other than White, making it substantially more diverse than many other biobanks that have participants with predominantly European ancestry [26]. Some groups, including Middle Eastern and North African (MENA) populations like Iranians or Armenians, are not well represented in current biobanks. Thus, ATLAS offers opportunities to study health in diverse communities16.

Identifying fine-scale identity-by-descent clusters To identify fine-scale clusters we first inferred patient relationships via identity-by-descent sharing (Fig. 4.1). Studying identity-by-descent clusters offer advantages over clustering patients through EHR-reported measures alone10. In ATLAS, a large proportion of patients have missing or "other race" specified in their EHR. Other demographic characteristics may be missing for complex and non-random reasons, and when included, they are not guaranteed to be accurate [11]. Therefore, for this study, we focus on groups identified via genetic ancestry. Genetic ancestry is a distinct concept from race, which is a social construct [92]

To define the identity-by-descent clusters, we called pairwise identity-by-descent between all ATLAS participants and reference individuals from the 1000 Genomes Project [6], the Simons Genome Diversity Project [106] and the Human Genome Diversity Project [13]. Identity-by-descent segments were estimated using iLASH [150] and clusters were identified with the Louvain community detection algorithm [17]. Sensitivity analyses were performed with additional phasing and identity-by-descent calling algorithms; pairwise identity-by-descent segments were highly concordant between the methods (Pearson's R^2=0.91) and alternative clustering algorithms over the alternatively phased data produced similar clusters.

We detected 367 identity-by-descent clusters, each of which was given an identifier determined by three iterations of Louvain clustering (e.g., "cluster-1-0-2"). There was substantial variation in cluster size, ranging from 2 to 2,030 individuals. Differences in cluster size, historic population size, and complex patterns of genetic relatedness resulted in differential cluster densities. In some clusters, like cluster 3-8-2, nearly every pair of individuals share identity-by-descent segments, while in other clusters, like cluster 1-6-10 or 5-7-0, individuals share fewer connections. Admixture analysis39 revealed substantial genetic diversity between the clusters, with continental ancestry sources from the Americas, West Asia, Europe, Africa, East Asia, and South Asia (Fig. 4.3a).

To further refine the clusters, we used the approach of Dai et al [39]. and merged subclusters with low genetic differentiation, measured as Hudson's fixation index (F_{st})(F_{st} <0.001). This produced clusters differentiated enough to represent the diversity of ATLAS, while still powered for statistical analyses. Finer-scale clusters might be relevant for specific medical or population genetics questions. For example, the subclusters that were merged together to make the predominantly European ancestry cluster each had a different distribution of identity-by-descent sharing Computing F_{st} to UK BioBank [23] participants born outside the UK suggested that these subclusters represent individuals with Northern, Southern, and Eastern European ancestry (Fig. 4.3b).

After F_{st} merging, 24 clusters with at least 30 ATLAS participants representing 97.8% of ATLAS remained for downstream analysis. These 24 clusters were assigned a name. The ATLAS biobank does not contain the country of origin of participants, which was used in our previous studies to annotate cluster identity [11]. Instead, we annotated clusters by using reference data in the clustering algorithm. For clusters without reference data, de-identified EHR demographic information, such as EHR reported race and ethnicity, preferred language, and religion, were used to refine and determine cluster annotations. Importantly, the label given to a cluster serves as a broad interpretation of the cluster's demographic and ancestral ties and does not necessarily reflect the self-identity of members (see Discussion). Furthermore, the clusters discussed here are specific to Los Angeles, especially those who visit UCLA Health, and may not be representative of the global population.

Using external reference data (Fig. 4.3d, global genetic ancestry, principal component analysis (PCA), and EHR-reported demographics, we identified identity-by-descent clusters reflecting the demography of Los Angeles. There was a large cluster of Mexican and Central American patients. Further Louvain clustering of this cluster with additional indigenous reference samples from Mexico [55] revealed subclusters with ancestry from northern Mexico and Baja California, central Mexico and Oaxaca, and Guatemala (Supplementary Table 2). We also identified three distinct Black and African American identity-by-descent clusters, containing patients with African American, Afro-Caribbean, and West African ancestries respectively (Fig. 4.3c). Several clusters had MENA global genetic ancestry (Fig. 4.3a), consistent with Los Angeles County having the largest population of people from the Greater Middle East in the United States [131]. Two distinct clusters contained patients of Iranian descent, one with patients with EHR-reported Jewish religion while the other contained patients who reported other religions. One cluster was enriched for patients of Armenian descent, consistent with Los Angeles having the largest population of diaspora Armenians in the US [123]. Lastly, we identified several Asian identity-by-descent clusters. These included clusters with patients that have predominantly East Asian global genetic ancestry (Fig. 2a), and also clusters with South Asian ancestry.

Our previous work [11] found that clustering using identity-by-descent offered enhanced resolution relative to PCA. Similarly, we found that many of the clusters overlapped in PC space. This was especially true for the Middle Eastern and South Asian identity-by-descent clusters, which were tightly clustered with the European cluster.

4.2.2 Health system utilization of identity-by-descent clusters

We next sought to understand how individuals in the identity-by-descent clusters accessed the hospital system using EHR data. Patients in clusters varied substantially by age, sex, and BMI, as well as the fraction carrying private health insurance. However, the proportion patients with private insurance coverage was high for all clusters, likely driven by the fact that not having quality insurance coverage is a primary obstacle to obtaining healthcare in the United States [52].

We used logistic regression to test for associations between EHR-phecode [183] based diagnoses and cluster membership. To account for differences in diagnosis frequencies between medical contexts, we separately assessed the code assignments both for outpatient encounters and emergency room (E.R.) visits and controlled for age, sex, and BMI. More complex combinations of ICD10 codes are often used in place of phecodes for improving phenotypic specificity. To explore this, we used additional phenotype definitions for Alzheimer's disease and related dementias [38].

We began by comparing outpatient phecode assignments in the Ashkenazi Jewish identity-by-descent cluster (n=5309) to all other participants. We tested n=1131 phecodes assigned to at least 30 patients in outpatient encounters. 236 phecodes were significantly associated with cluster membership at Benjamini-Hochberg false discovery rate of 5% (Fig. 4.4a). Consistent with previous studies of Ashkenazi Jewish individuals [11][143][141]patients in the cluster were more likely diagnosed with ulcerative colitis (OR=2.24, 95% CI: [1.83, 2.75], q-value=5.34x10-13) and regional enteritis (OR=2.93, 95% CI: [2.41, 3.56], q-value=2.39x10-24). We further identified less well-characterized associations, particularly for several mental health disorders, including eating disorders (OR=3.37, 95% CI: [2.45, 4.64], q-value=6.79x10^{-}12), anxiety disorder (OR=1.7, 95% CI: [1.59, 1.82], q-value=9.90x10-52), and major depressive disorder (OR=1.62, 95% CI: [1.47, 1.78], q-value=2.55x10-20). All these associations remained significant at FDR 5% when restricting the analysis to only compare the Ashkenazi Jewish cluster with the European cluster.

In E.R. visits, membership in the Ashkenazi Jewish identity-by-descent cluster was significantly associated with major depression as the primary diagnosis (OR=2.29, 95% CI: [1.32, 3.98], q-value=4.86x10-2). While these results were consistent with previous reports of mental health conditions in European Jewish communities [91][135], we emphasize that this association does not indicate a causal relationship between identity-by-descent cluster membership and these disorders [189].

We next examined associations in the African American and Mexican and Central American identity-by-descent clusters. This analysis revealed several associations in both outpatient (Fig. 4.4) and emergency room contexts. Consistent with previous literature[158], patients in the African American cluster were more likely diagnosed with sickle cell anemia (OR=50.29, 95% CI: [29.08, 86.97], q-value=1.33x10-42)(Fig. 4.4b). We also identified a significant increase in uterine leiomyomas in the African American identity-by-descent cluster (OR=2.92, 95% CI: [2.4, 3.55], q-value=2.16x10-24), consistent with the increased burden of uterine fibroids in African American women and representing a substantial health disparity [47]. In the Mexican and Central American cluster, there was a strong enrichment of type2 diabetes (OR=2.37, 95% CI: [2.2, 2.56], q-value=3.27x10-104) and chronic liver disease (OR=5.52, 95% CI: [4.65, 6.56], q-value=3.47x10-81)(Fig. 4.4).

To further characterize the disease risk of Latino patients, we examined how phecode associations differ between the three Mexican and Central American subclusters, the Afro-Caribbean cluster, and the Puerto Rican identity-by-descent cluster. 106 phecodes showed effect size heterogeneity54 across these five clusters. For example, while phecodes relating to lung disease (i.e. pulmonary fibrosis and lung transplants), were associated identifying as Latino in the EHR, the association was most primarily driven by patients in the the Afro-Caribbean cluster. Even within the three Mexican and Central American subclusters, there was heterogeneity. The Guatemalan and Central American subcluster was the only subcluster associated with several pregnancy phecodes, including anemia during pregnancy (OR=2.57, 95% CI: [0.94, 1.48], q-value=4.84×10-5) and short gestation period (OR=5.04, 95% CI: [2.73, 5.95], q-value=4.86×10-5 The Central Mexican subcluster was the only subcluster associated with the coccidioidomycosis fungal infection(OR=3.98, 95% CI: [1.92, 3.71], q-value=3.86×10-5). Overall, these differences offer further evidence that grouping patients only by Hispanic and Latino ethnicity is too coarse.

We further examined disease associations in MENA and Asian clusters (Fig. 3d). We began with the Iranian (n=315) and Iranian Jewish (n=264) identity-by-descent clusters. These two clusters shared several associations in outpatient diagnoses. Individuals from both clusters were less likely to be diagnosed with skin cancer (Iranian Jewish: OR=0.1, 95% CI: [0.03, 0.28], q-value=3.09x10-3, Iranian: OR=0.26, 95% CI: [0.13, 0.51], q-value=4.07x10-2) However, the phecode with the smallest p-value for each cluster, non-toxic multinodular goiter in the Iranian cluster (OR=2.58, 95% CI: [1.63, 4.08], q-value=4.07x10-2) and adjustment disorder in the Iranian Jewish cluster (OR=2.89, 95% CI: [2.04, 4.09], q-value=2.31x10-6), were not the same. Other associations included an enrichment of phecodes relating to bacterial enteritis in the Egyptian Christian identity-by-descent cluster (n=92) (OR=7.42, 95% CI: [3.56, 15.47], q-value=1.04x10-4) and phecodes relating to bronchus cancer in the Korean identity-by-descent (cluster (n=546)(OR=2.82, 95% CI: [1.84, 4.32], q-value=2.56x10-4).

We also observed an increased number of diagnoses relating to viral hepatitis B in identity-by-descent clusters with Asian ancestry patients. Asian ancestry as a risk factor for viral hepatitis B is widely documented [27]. However, we noted that there were differences between the fine-scale Asian ancestry clusters. For example, individuals in the Chinese identity-by-descent cluster (n=1547) (OR=19.12, 95% CI: [14.92, 24.5], q-value=1.88x10-117) were more likely to receive a diagnosis of hepatitis B, while diagnoses of hepatitis B were not elevated in the Japanese cluster (n=596) (OR=1.15, 95% CI: [0.47, 2.8], q-value=1.00x10-1). We performed a mixed-effects meta-regression using the odds ratios estimated for each Asian ancestry cluster [178]. The effect sizes significantly differed between the clusters for this phecode and others (meta-regression p=2.23x10-15), showing the value of fine-scale information.

To explore whether the associations reported here were specific to UCLA or could be generalizable to other settings, we used BioMe summary statistic data published in Belbin et. al10. For six BioMe identity-by-descent clusters found in ATLAS (Supplementary Table 3), the correlation of effect sizes was high, R2=0.69 (IQR=[0.63, 0.84]) (Extended Fig. 10). Many associations in BioMe were found in ATLAS, including elevated rates of gout in the Filipino cluster (OR=4.91, 95% CI: [3.77, 6.4], q-value=2.24x10-29), chronic lymphocytic thyroiditis in the Ashkenazi Jewish cluster (OR=1.51, 95% CI: [1.3, 1.76], q-value=3.07x10-6), and peripheral vascular disease in the African American cluster (OR=2.0, 95% CI: [1.58, 2.53], q-value=3.21x10–7) (Supplementary Table 4). Unlike BioMe, the ATLAS European cluster did not have an elevated rate of multiple sclerosis (OR=1.2, 95% CI: [0.93, 1.54], q-value=3.55x10-1). Associations were calculated relative to a background population and differences between ATLAS and BioMe might be driven by differences in comparator clusters, environment, or the underlying fine-scale populations.

While phecodes assigned to an identity-by-descent cluster can be relative to the entire biobank, we also explored enrichments between closely related clusters. Phecode association tests for the Armenian cluster were performed against four comparator clusters- against the entire biobank, against the European cluster, against the two Iranian clusters, and against all MENA ancestry identity-by-descent clusters. We restricted to phecodes with more than 30 patients in all four groups and examined phecodes significant in all four comparisons (Fig. 4.4a). Phecodes relating to heart disease were more likely to be associated with the

Armenian cluster relative to all comparison groups. This result is consistent with previous reports of Armenian ancestry as a risk factor for cardiovascular disease56. Next, we examined whether there were phecodes associated with the Armenian identity-by-descent cluster that had significantly different effect sizes across the comparison groups (Fig. 4.4b). Seven phecodes had a nominally significant meta-regression p-value ($p<0.05$), e.g. non-toxic uninodular goiters. The Armenian cluster was more likely than the biobank and the European cluster to be associated with this phecode, but less likely to be diagnosed with this phecode relative to the Iranian and MENA clusters. This example illustrates the importance of holistically evaluating cluster-disease associations, as they are likely determined by context and environment.

We next sought to evaluate how individuals in identity-by-descent interface with the health system. We found that many clusters were significantly less likely to visit a routine care provider than the European cluster. For example, individuals who belonged to the European cluster were significantly more likely to visit a primary care physician (OR=1.33, 95% CI: [1.27, 1.4], q-value=7.19x10-29) than other biobank participants (Extended Fig. 9b). We observed differential utilization of the emergency room by clusters. Patients in the African American and the Mexican and Central American, identity-by-descent clusters were more likely to visit the emergency room, a well-documented health inequity that is associated with worse outcomes [35][147][9]. However, we also identified other clusters that were more likely to visit the emergency room, including the Iranian Jewish (OR=1.78, 95% CI: [1.41, 2.25], q-value=7.64x10-6) and Armenian (OR=2.34, 95% CI: [1.53, 3.57], q-value=3.98x10-2), identity-by-descent clusters both of primarily MENA ancestry. Emergency room use for these populations is not widely documented.

We next examined how individuals from different identity-by-descent clusters interact with the health system over time, which can give insights into the dynamic nature of disease. We plotted two typical phecodes (Methods), kidney transplants and major depressive disorder for the 6 largest clusters. The proportion of patients assigned a phecode relating to kidney transplants significantly increased between 2016 and 2019 for the Filipino (p=4.42x10-

5), Mexican and Central American (p=1.77x 10-31), and African American (p=5.30x10-7) identity-by-descent clusters, but not in the Ashkenazi Jewish, European, or Chinese clusters. Diagnoses generally increased but dropped sharply in 2020, which might be attributed to the decrease in procedures performed during Covid-19 shelter-in-place orders.

Phecodes relating to mental health conditions were heterogeneous between clusters. The Ashkenazi Jewish identity-by-descent cluster had the highest proportion of patients diagnosed with major depressive disorder. By 2020, this cluster had five times as many diagnoses as the Chinese identity-by-descent cluster. This cluster had a consistently low proportion receiving the phecode, and while most other clusters had an increasing number of diagnoses with time, the Chinese cluster had a slow or even decreasing proportion. For any of these diagnoses, it is not necessarily true that the rates of diagnosis indicate the actual prevalence of the health conditions in the cluster. Instead, these results indicate the complex dynamics between how clusters interact with the health system, which could be a function of doctor choice, insurance coverage, practitioner perceptions, or other forces.

Identity-by-descent clusters can facilitate the study of pathogenic alleles in diverse groups, which are often underrepresented in genetic screening efforts [1]. To do this we examined the minor allele frequency (MAF) of pathogenic mutations that have been previously reported to be enriched within particular groups. One example is Familial Mediterranean Fever, which is caused by mutations in the MEFV gene [157]. We restricted to pathogenic MEFV SNPs and performed a Fisher's exact test comparing cluster allele frequencies to the rest of ATLAS. One pathogenic SNP genotyped in MEFV (rs28940579) was significant at FDR 5% in several MENA ancestry clusters. These included the Ashkenazi Jewish (MAF: 2.9×10-2, p=2.6×10-159) Armenian, (MAF: 4.2×10-2, p=1.7×10-21), and the Lebanese Christian identity-by-descent clusters (MAF: 3.7×10-2, p=1×10-8), which all had elevated frequencies compared to the remaining biobank excluding these clusters (biobank MAF: 9.55×10-4). Of all ATLAS clusters, diagnosis with FMF was strongly associated with membership in the Armenian cluster (OR=17.36, 95% CI: [6.99, 46.95], p=1.0×10-8, consistent with literature finding of high FMF burden in individuals of Armenian descent [117]. However, the high carrier rate in other clusters motivates disease screening in other populations.

We also analyzed pathogenic variants in the HBB gene, which is implicated in thalassemia and sickle cell disease [25]. Sickle cell disease is known to be associated with African ancestry [158] and in the phecode analysis, it was significantly associated with membership in the African American identity-by-descent cluster. Consistent with that observation, we found a pathogenic HBB allele, rs34598529, that was significantly more common in this cluster (biobank MAF: 3.02x10-5 cluster MAF: 2.20x10-3, p=1.52x10-9). Furthermore, we found two pathogenic alleles in HBB associated with membership in the Chinese identity-by-descent cluster. Both alleles, rs34451549 (biobank MAF: 0.00, cluster MAF: 3.54x10-3, p=1.12x10-13) and rs33931746 (biobank MAF: 1.15x10-5, cluster MAF: 1.18x10-3, p=1.89x10-4), are documented to be associated with beta-thalassemia in East Asian populations [79][188] and are at elevated frequencies in these populations in gnomAD, a large database of allele frequency data [78]. Furthermore, patients in this cluster were also more likely to receive diagnoses of hemoglobinopathies (OR=2.81, 95% CI: [1.87, 4.21], p=3.93×10-5) than the remaining biobank participants. This result illustrates that patients of many different ancestry backgrounds could experience elevated genetic risk in the HBB gene.

Lastly, we broadly studied genetic risk variants associated with each identity-by-descent cluster and found over 100 loci that were at elevated frequencies in a specific cluster. Examples included elevated MAF of a pathogenic allele associated with transthyretin cardiac amyloidosis in the African American cluster (biobank MAF: 2.14x10-4, cluster MAF: 1.78x10-2, p=4.76x10-66), and an allele associated with Lynch Syndrome in the Mexican and Central American cluster (biobank MAF: 0.0, cluster MAF: 6.95x10-4, p=5.59x10-7)[58][40]. We further identified several lesser-known associations. One finding was rs28937594, which was significantly higher in the Iranian Jewish identity-by-descent cluster (biobank MAF: 5.80x10-5 cluster MAF: 0.024, p=5.58x10-28). Rs28937594 is in the GNE gene and is implicated in hereditary inclusion-body myopathy, an ultra-rare recessive disease [136]. While no ATLAS participants were homogenous for the SNP or diagnosed with the disease, this SNP has been reported to be a founder mutation in Iranian Jewish populations [46]. Interestingly, in the

Iranian identity-by-descent cluster, the MAF for this SNP was also high, but not significant (cluster MAF: 0.0017, p=0.1512). Overall, this supports the idea that identity-by-descent clusters can confirm and refine variants included in genetic screening programs [2].

4.2.3 Genetics of identity-by-descent clusters

Identity-by-descent clusters also present opportunities for learning about historical or demographic factors, which can have implications for personalizing care or developing precision treatments [167][48]. First, we analyzed the distribution of total identity-by-descent shared between pairs of individuals in a cluster (Fig. 4.5a)(Supplementary Table 5). The Iranian Jewish cluster had the highest level of total identity-by-descent sharing (mean = 57.43 cM, 95% CI: [(56.80 - 58.06]). This is higher than other clusters that contained populations expected to have founder effects. The Iranian cluster also had relatively high identity-by-descent sharing (total pairwise identity-by-descent mean=15.70 cM, 95% CI: [14.54 - 16.86]), but not as high as the Iranian Jewish cluster, highlighting the role of cultural factors.

Additionally, we examined cluster runs of homozygosity (ROH) (Fig 4.6b), which occur when an individual inherits identical copies of a haplotype from each parent [28]. ROH can reflect the demographic processes, such as consanguinity, and is implicated in risk for complex diseases [90][118]. We found elevated amounts of ROH in several MENA clusters and South Asian ancestry clusters. The amount of within-cluster identity-by-descent sharing did not always correlate with the rate of ROH. This observation may be attributed to differences in the historical and modern demographic processes, like the practice of endogamy or historical population bottlenecks.

We used the IBDNe program [21] to estimate cluster-specific historical effective population size (Fig. 4.6c). Consistent with previous reports [12], we observed a large bottleneck in the Puerto Rican cluster, with a minimum population size occurring around 15 generations ago. We also observed historic population size reduction in several other clusters, especially in MENA ancestry clusters. The bottleneck timing in these clusters is similar, approximately 13-15 generations ago. Despite the similarity in the timing of the bottleneck, the estimates of the max population size differed. For example, the population size of the Iranian Jewish cluster was estimated to be less than 10,000 for the last 30 generations, which is very small, and could be relevant for understanding the genetic disease burden in this group.

Patterns of identity-by-descent sharing between clusters can further reveal modern and historical relationships. We first computed pairwise Hudson's F_{st} in the largest identity-by-descent clusters (Fig. 4.6d), which revealed complex within-continent sharing patterns. While there was low differentiation between the Iranian and Iranian Jewish clusters, (F_{st}=0.0055), the Iranian cluster exhibited a smaller F_{st} with the Armenian, Egyptian Christian, and Lebanese Christian clusters. It is important to note, however, that the F_{st} estimates used here do not capture the effect of rare variants [15].

Lastly, we created a network representation of identity-by-descent sharing, where the nodes of the network were a cluster and the edges were the median identity-by-descent shared between clusters (Fig. 4.6e). From this representation, we observed that geography affected cluster relationships. For example, clusters with MENA ancestry were close in network space, with the Pakistani cluster acting as a bridge between them and the South Asian identity-by-descent clusters. We also observed some unexpected relationships. The Mexican and Central American cluster shared more identity-by-descent on average with the Ashkenazi Jewish cluster (mean=0.243 cM, 95% CI:[0.243, 0.244]) than European cluster (mean=0.0372 cM, 95% CI: [0.0371, 0.0373). A similar trend was observed for the Puerto Rican identity-by-descent cluster. Other reports have found a contribution of Jewish ancestry to Latin American populations [30].

4.3 Discussion

To ensure that precision medicine initiatives are applicable to all people, it is important to understand the diverse determinants of health. In this study, we analyzed clusters of people who share genetic ancestry. Identifying these fine-scale ancestry clusters is useful in the study of health disparities, especially with respect to the coarse race and ethnicity information usually recorded in biobanks. While people who share ancestry may share genetic risk for disease, they may also share an environment, which is particularly important for understanding disease risk. Race, ethnicity, and religion are social constructs and are not determined by genetics, although they may be correlated [19]. It is simultaneously true, however, that identity by race, ethnicity, and religion can affect access to and quality of healthcare in the United States81. Thus, this approach provides a complementary lens for identifying potential health differences between people living in Los Angeles.

These findings can inform provision of care at UCLA Health and similar health systems. We identified pathogenic loci that segregated at higher frequencies in the Chinese, Iranian Jewish, Armenian, and African American clusters. Historically, in the United States, carrier screening guidelines are based on self-reported race and ethnicity [34][4]. Many of the associations we identified would be missed by these guidelines. Furthermore, allele frequency data is often only available for limited ancestry groups66, and pathogenicity or penetrance may differ across ancestries [108]. This work supports calls to expand genetic screening efforts to more people [2][1] regardless of race or ethnicity. We make allele frequencies available for all clusters to facilitate studies on genetic disease in diverse groups.

These results occur within, and support the existence of, an unequal healthcare system. For example, the African American and Mexican and Central American clusters were both associated with severe diseases, like chronic renal failure and liver transplants. This could be a consequence of the burden of systematic racism, which adversely affects the health of minority groups in America [7], and reduced access to quality insurance, which affects care and varies by race and ethnicity. These results may be further compounded by the fact that the main UCLA Health facilities are in west Los Angeles, which includes some of the wealthiest neighborhoods in Los Angeles County. Thus, clusters with from economically disadvantaged households might be traveling further to access specialty care at UCLA and thus have greater health needs motivating the longer trip.

There are several limitations to this work. Although we used genetics to identify clusters, genetics is likely not the only causal factor for these results. The reported associations are strictly correlative and may be specific to UCLA Health. Additionally, defining a population or cluster is not straightforward [130][92], and the definition of ancestry itself is subject to disagreement [112]. We followed previous studies and chose a genetic similarity criterion, but any number of criteria or algorithms could have been used. Additionally, the clusters are not necessarily equivalent. Some were tightly related in network space, while others had more diffuse patterns of connection. While every participant in ATLAS is placed into a cluster, this approach may have limitations for individuals with multiple ancestries.

The individuals whose data comprise ATLAS are not representative of a random sample of the general Los Angeles population. The ATLAS biobank is opt-in, which means that an individual's participation can be influenced by their level of comfort and trust with health research. Since medical research has a long history of unethical experimentation on people of color, these groups may be less willing to participate [125]. Another source of participation bias is that individuals who come to a hospital are usually unwell. The severity of ill health may vary with geographic distance from UCLA. Other socioeconomic factors, such as age, education, and household income are also associated with when and if patients receive diagnoses [8][169][186]. These differences may also be exacerbated by biases from health practitioners, which systematically affect care [43].

Lastly, we focused on population-level analyses in this work. When translating results to individuals, the limitations of genetic ancestry must be considered. Genetic ancestry is continuous, and many individuals have multiple ancestries. Identity-by-descent clusters as a biomarker must be inclusive and tailored to individuals for clinical use [110]. Furthermore, access to genetic information will inevitably have intrinsic biases. Health systems will have to evaluate the impact of genomic medicine initiatives on the populations they serve [2] as well as provide education to their patients and practitioners [163]. In particular, evidence-based recommendations on when to use ancestry, race, and ethnicity tailored to specific diseases and treatment options are needed [19].

Overall, we identified and characterized the health profiles of diverse Los Angeles identity-by-descent clusters. This represents an advance toward equitable health research and, along with our website, can empower future studies on health outcomes in Los Angeles.

4.4 Methods

4.4.1 Ethics

Patient Recruitment and Sample Collection for Precision Health Activities at UCLA is an approved study by the UCLA Institutional Review Board (IRB17-001013). All necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived.

4.4.2 Patients and Recruitment

The UCLA ATLAS Community Health Initiative aims to create a genomic resource to enable translational and precision medicine [73]. In ATLAS, genotyping data is tied to de-identified EHRs as part of the UCLA Health IT Discovery Data Repository and Dashboard (DDR) [31]. UCLA primarily serves patients on the west side of Los Angeles, but also has more than 200 clinics throughout the area, making it one of the largest health systems in Los Angeles.

Enrollment in ATLAS is elective and patients enroll in ATLAS when they visit a UCLA site for a blood draw. ATLAS has a 65% opt-in rate (see Lajonchere et al. [87] for more details on participation). As of 2021, there were approximately 35968 participants with full genotyping and DDR data available [73]. No statistical method was used to predetermine sample size. The experiments were not randomized and the Investigators were not blinded to allocation during experiments and outcome assessment. A complete description of the ATLAS project and data is available in Johnson et al. [73].

4.4.3 EHR Data

Each patient's genotype data was tied to Electronic Health Records (EHR) collected during patients visits on EPIC systems using a de-identified ID. Patient EHR was pulled for 2016-

2020 and included visit information, diagnosis information, and demographics. For the normal outpatient data, we restricted to visits that were labeled as scheduled appointments and that did not have a code associated with an inpatient, ICU, or trauma stay. Emergency room data was any visit that happened within an emergency room department. Diagnoses assigned in emergency rooms were restricted to the primary reason for the visit. Each visit contained information on patient weight, height, and BMI measured at the visit. We calculated the median BMI for a patient across all encounters and used this as the BMI for that patient in our association testing. The EHR was queried using Microsoft SQL Server 2014.

4.4.4 Demographic information

Demographic information was restricted to race/ethnicity, preferred religion, preferred language, sex, and birth date. Sex was indicated as binary. To calculate patient age, we calculated the patient age at the time of each visit and took the maximum age overall for each patient. For EHR-reported race/ethnicity, patients were designated (by themselves or a healthcare staff member) as "White," "Black", "Asian", "Native American", or "Pacific Islander." Asian patients could be further designated as Chinese, Japanese, Korean, Thai, Filipino, Vietnamese, Taiwanese, Pakistani, Indian, or Indonesian, although not all Asian patients had one of these identifiers. Hispanic/Latino patients were designated as "Hispanic", which was further subdivided into several other sub-identifiers, such as "Spanish origin", "Chicano/a" or "Cuban". For visualization, we considered the main race/ethnic categories and not the sub-designations. There were numerous preferred languages and religions. For simplicity, we examined the languages that had more than 5 individuals who indicated that they preferred that language. Furthermore, the preferred religion was restricted to consider major religions: Christianity, Islam, Judaism, Hinduism, Sikhism, and Buddhism. Christianity was further subdivided into Protestant and Catholic. Other religions were condensed into an "Other Religion" category.

4.4.5 Diagnoses and Phecodes

We used phecodes to study disease associations. Diagnoses are coded in the DDR as ICD10 codes. For all encounters that occurred at UCLA between January 2016 (the start of the DDR) and January 2021, we found all unique diagnoses assigned to a patient in an outpatient setting, which included in-person doctor visits and video calls. In outpatient visits, we included all diagnoses given in a visit. For emergency room visits, we restricted to diagnoses given as the primary reason for the visit, which was coded by the diagnosing clinician (i.e., if a person showed up to the emergency room for a heart attack who also had diabetes, the primary reason for the visit was the heart attack).

ICD10 codes were then merged into phecodes using the mappings provided at Phecode Map 1.2 with the first 5 characters of an ICD10 code (i.e., if the ICD10 code was V80.720S, only V80.720 would be used for mapping).

Alternative phenotype definitions were defined with ICD 10 codes or with using the procedure orders. Specialties

Specialty utilization was determined by the specialty of the primary provider for a patient encounter. Providers with multiple specialties were only counted for their primary specialty. We grouped subspecialities into one specialty. For example, “Neurology, sleep medicine” and “Neurology, movement disorders’ were both counted as a visit to a neurologist. Changes in phecodes over time

We calculated the proportion of a cluster assigned a phecode in a given year. We then calculated the inter-year difference in the proportion of people diagnosed in 2020 and 2016. Since we were interested in phecodes that might have different trajectories between clusters, we identified the phecodes that had the greatest variance in the inter-year difference between the 6 largest clusters.

4.4.6 Other phenotype definitions

We focused on definitions of phenotypes defined via phecodes because they have been shown to work well in the context of EHRs [183]. However, phecodes tend to be broad and are optimized for generalizability across health systems and for genetic association studies. Depending on the application, other phenotype definitions might be more relevant.

To explore this, we utilized two different phenotype definitions. One phenotype definition was a curated list of ICD codes relevant to Alzheimer's and related dementias (see below), and that is used by physicians for defining clinical cohorts. The other phenotype was brain MRI imaging orders. We performed a logistic regression to assess the relationship between cluster membership and ever having the phenotype, controlling for age, sex, and BMI.

4.4.7 Genomics pre-processing and quality control

Genotyping for ATLAS was performed on a custom genotyping chip, with sites from the global screening array. Data was mapped to hg38 and all SNPs were mapped to the 147 build of dbSNP96. All preprocessing and quality control steps were performed using PLINK 1.997 and bcftools v1.998.

For ATLAS samples, we removed any individuals whose genotyped sex mismatched their EHR-reported sex. We did this by using the PLINK –update-sex command to update the PLINK fam files to contain the EHR sex and the PLINK –check-sex to identify samples with discrepancies between the estimated genotype sex and EHR sex.

ATLAS data was merged with genotyping data from the 1000 Genome Project (1000GP), the Simons Genome Diversity Project (SGDP), and the Human Genome Diversity Project (HGDP). All reference data were converted to hg38 for merging using CrossMap [190]. Samples that overlapped between the different projects were removed using PLINK –keep. Rsids were harmonized across projects using bcftools annotate. Data were then standardized using bcftools norm and a hg38 genome reference. After merging, sites or individuals with more than 1% missing were removed using plink –mind and –geno. For identity-by-descent analysis, only SNPs with MAF ¿ 5% were kept.

4.4.8 Phasing

Before identity-by-descent calling, data was statistically phased using Shapeit4 [42] using default parameters and the hg38 map files distributed with the software. To speed up computation, one chromosome was phased at a time.

4.4.9 PCA

To prevent the large sample size of ATLAS from distorting the relationship populations in PC space, PCA was performed first on only the reference samples. ATLAS samples were then projected onto the reference PCs. To enable visualization, the reference data, and the ATLAS sample PCA results were plotted separately on adjoining axes.

4.4.10 Identity-by-descent calling and processing

For identity-by-descent calling, the genotype data were converted from PLINK bed files into PLINK ped/map files using a custom Python script that preserves phasing. Centimorgan information for the map files was pulled from the same genetic maps used in Shapeit4.

Identity-by-descent segments were called using iLASH37 with the following parameters: slice size 350, step size 350, perm count 20, shingle size 15, shingle overlap 0, bucket count 5, max thread 20, match threshold 0.99, interest threshold 0.70, min length 2.9, auto slice 1, slice length 2.9, cm overlap 1, minhash threshold 55. Identity-by-descent was called for one chromosome at a time.

4.4.11 Identity-by-descent quality control

After identity-by-descent segments were called, we removed outliers as in Belbin et al [10]. Firstly, any identity-by-descent segments overlapping centromeres or telomeres were removed. Identity-by-descent tracts intersecting the HLA region were also removed. To find other regions of the genome that may have erroneously high identity-by-descent, we calculated the total amount of identity-by-descent contained at each SNP in our input file by summing all segments that overlapped that SNP. SNPs that had total identity-by-descent greater or less than 3 standard deviations from the genome-wide mean were removed. In total, 6696 were removed.

For downstream analysis, identity-by-descent segment lengths were summed between individuals, meaning that for a given pair of individuals, all the identity-by-descent segments that they shared across all chromosomes were added together to create one summary number.

We removed pairs of individuals who were immediate family members using two methods. Firstly, we used the PLINK 2.0 implementation of KING [107] to identify relatives of third degree or closer, using the parameter of –king-cutoff with a value of 0.0884. KING was run on all SNPs with MAF > 0.05 and after linkage pruning, using PLINK and –indep-pairwise 50 10 0.1. As KING may underestimate the relatedness of individuals, especially in the case of individuals with high levels of autozygosity [150], we also filtered pairs based on the total amount of identity-by-descent shared. Using empirical data reported to DNA Painter [14], we determined a conservative threshold of second-degree relatedness was a threshold of 1000cM. We removed any pairs with identity-by-descent higher than this threshold.

4.4.12 Sensitivity analyses

To characterize the robustness of our results to the choice of phasing and identity-by-descent calling algorithms, we performed additional sensitivity analyses with different algorithm choices. Statistical phasing was performed with Eagle v2.4.1 [98] and identity-by-descent calling was performed using hap-ibd [192]. As with iLASH, identity-by-descent was called for segments > 3.0cM long and on individuals who were unrelated (more than third-degree relatives). After calling all identity-by-descent segments across ATLAS and the reference data, we summed the total amount of identity-by-descent shared between a pair of individuals. We then calculated the Pearson's correlation between the total identity-by-descent shared between a pair detected with shapeit4 + iLASH and the total amount detected with

Eagle + hap-ibd.

We further characterized the robustness of the clusters initially identified with iLASH. We re-performed Louvain clustering as we did previously, using three iterations of clustering and merging any clusters with F_{st} ¡ 0.001. To assess the consistency of the clustering, we randomly sampled 10,000 ATLAS pairs and asked if they were in the same cluster originally, was the pair still in the same cluster with the new algorithm, or vice-versa.

4.4.13 Cluster identification

To infer clusters, we followed the approach of Dai et al. and used the Louvain method for cluster detection [17]. This method finds structure in large networks and has been shown to work well on genetic data12. We applied this algorithm to an undirected network constructed from identity-by-descent sharing, where each node represented an individual and edge weights were defined as the genome-wide sum of identity-by-descent sharing between the nodes. An advantage of the Louvain algorithm is that it can be run iteratively, meaning that an initial run over the entirety of the graph can be used to define broad substructure, which can be further resolved into more fine-scale clusters upon subsequent iterations.

For cluster detection, we used the Python package NetworkX [60]. We created an undirected graph representation of the identity-by-descent matches, where each node was an individual and an edge between individuals was weighted by the total amount of identity-by-descent matches shared between the two people.

Louvain clustering implemented in NetworkX, was used iteratively to detect fine-scale populations. It was first run to detect a primary set of clusters. Each cluster was then subject to Louvain clustering again, and these subclusters were clustered once more, for a total of three runs of Louvain clustering.

After generating clusters with the Louvain algorithms, the clusters were merged using F_{st}, as in Dai et al12. We used the implementation of Hudson's F_{st} from PLINK 2.0. It was run on all pairs of clusters from the third level of the Louvain clustering and clusters that had $F_{st} < 0.001$ were merged. Since F_{st} may perform poorly in small populations, clusters with less than 10 people were removed [15]. This threshold was selected because it gave good separation of clusters on a subcontinental level.

4.4.14 Cluster identity and demographics

We primarily used external reference data to characterize what populations may be contributing to a cluster. Some clusters did not contain any reference data, or the reference data did not capture important aspects of the cluster. For example, there was no Ashkenazi Jewish reference data, only reference data labeled by European countries. To address this problem, we used the de-identified EHR demographic table as an additional source of information. This included EHR-reported race and ethnicity, preferred language, and religion. We emphasize that race, ethnicity, and religion are not determined by identity-by-descent segments but represent sociocultural characteristics that may be related to characterizing the cluster. We chose to use religion when it was relevant to identifying a historically persecuted group (i.e. “Lebanese Christian” instead of just “Lebanese”). These groups often have distinct histories and cultural practices, which can affect demography, environment, and disease risk. For example, it is well known that Ashkenazi Jews have distinct genetic risks relative to other Europeans [155]. Thus, including religion in this study may offer opportunities to improve the health of understudied ethnoreligious groups.

The majority of ATLAS patients are not Latino, have no religious preference, and indicated that they prefer to speak English. We, therefore, explored cluster identity using individuals who preferred a different language or religion or were identified as Hispanic/Latino in the EHR (note that the actual number of English speakers may be lower, as some patients may not, for societal or practical reasons, have this information included in their medical records).

For downstream analysis, we focused on identity-by-descent clusters that had more than 40 members to ensure a large enough sample size for our EHR and genetic analyses.

Additional summary statistic reference data were used to compute Hudson’s F_{st} between

ATLAS identity-by-descent clusters and external populations, including identity-by-descent clusters identified in the BioMe biobank. This enabled additional refinement, along with the use of EHR demographic information and cluster-level admixture analyses.

4.4.15 Latino subclusters

We obtained an additional reference dataset that focused on fine-scale indigenous populations of Mexico [55]. Importantly, some of these indigenous groups also live in neighboring Guatemala, Belize, Honduras, and El Salvador, which were all part of the historic Mesoamerica region that was broken up by Spanish colonization [128]. We merged the genetic data from the indigenous populations and that of patients from the Mexican and Central American identity-by-descent cluster and performed an additional level of Louvain clustering. As above, we merged clustered with low differentiation ($F_{st} < 0.001$). One set of four subclusters were merged for subsequent analyses and was referred to the Central American identity-by-descent cluster.

EHR demographic characteristics were explored for each subcluster. Phecode associations for each subcluster were also compared using the and heterogeneity in effect size was analyzed for the three largest subclusters along with the Puerto Rican and Afro-Caribbean clusters

4.4.16 Identity-by-descent distribution

To find the distribution of identity-by-descent in a cluster, we considered segments of individuals assigned to the same cluster. We summed the identity-by-descent segments to get the total identity-by-descent shared between the pair and calculated the distribution of total identity-by-descent between members of the cluster. ROH distribution

For ROH, we first performed linkage pruning and MAF filtering using PLINK and the parameters –maf 0.01 –indep-pairwise 50 10 0.1. ROH calling was also performed using PLINK and the parameters -homozyg –homozyg-density 200 –homozyg-gap 500 –homozyg-kb 3000 –homozyg-snp 65 –homozyg-window-het 0 –homozyg-window-missing 3 –homozyg-window-snp 65. Detected ROH were summed within an individual. We then calculated the distribution of detected ROH of all individuals within a cluster. IBDNe

IBDNe was run using the identity-by-descent haplotypes estimated using iLASH . We filtered the iLASH output for each chromosome to individuals from a single cluster. The haplotypes were combined into one file for IBDNe input. IBDNe was run with default parameters and the hg38 genetic map provided on the IBDNe website.

4.4.17 F_{st}

For the heatmap of F_{st}, we calculated the pairwise Hudson's F_{st}, as described in the Louvain clustering section. We calculated F_{st} between the largest final clusters (after Louvain clustering and merging). Data was visualized using Python Seaborn clustermap with default parameters.

4.4.18 Genetic relatedness network

The network visualization between clusters was developed using NetworkX. The input was a matrix where each row and columns represented one of the largest clusters, and each entry was the mean identity-by-descent shared between the two clusters. To find this mean, we found all possible pairs of individuals between the two clusters. If the pair did not have any identity-by-descent detected, we set their sharing to 0 and then calculated the mean over all possible pairs. This was to prevent biasing the mean identity-by-descent by limiting it to only pairs that had identity-by-descent detected. This square matrix was then used to create a weighted undirected graph, where the nodes were the clusters and the edges were the mean identity-by-descent between the clusters. We visualized the graph using 1000 iterations of the Fruchterman-Reingold force-directed algorithm [53].

4.4.19 Association testing

Statistical testing was done using the Python StatsModel [145] package. For each phecode, we determined whether an individual has ever had been assigned that phecode in an outpatient context, making the outcome binary. Cluster status was binary and could either be a particular cluster vs all other biobank participants, or a particular cluster compared against another cluster. We tested whether binary cluster status was associated with phecode assignment using the StatsModel GLM command with the family set to binomial. We corrected for sex, age, and BMI in these analyses. S

The same statistical framework was used to test for emergency room diagnoses and specialty visits, where instead of phecode assignment, the outcome was whether an individual had visited a doctor with a given specialty reported in the EHR. In all cases, we restricted to specialties, diagnoses, or zip codes with at least 30 visits.

An association was considered significant after controlling for false discovery rate at 5% using the Benjamini and Hochberg procedure. Multiple test correction was performed across phecodes each time a regression analysis was performed, i.e. for each cluster-background comparison.

4.4.20 Heterogeneity test

To calculate whether there was a significant difference in the effect sizes between clusters for a given phecode, we performed a mixed-effects meta-regression test for heterogeneity, implemented in the R package metafor [178]. Specifically, we used the function rma.uni.

4.4.21 Reproducibility

To assess the reproducibility of the results presented in this work, we obtained published association statistics taken from the BioMe biobank at Mt. Sinai10. For 6 related identity-by-descent clusters comprised of similar populations in ATLAS and Biobank (see Supplementary Table 3), we computed odds ratios for phecodes tested in both biobanks. We compared the effect size of the estimates using a Pearson's correlation.

4.4.22 Website

The website hosting the data visualization is implemented as a single-page application [121]. The application is developed in the JavaScript framework React, where each graph page is implemented as a separate component. The map plot is powered by the deck.gl library developed by Mapbox, which provides maps for data overlays. The other graphs are powered by the react-plotly.js library developed by Plotly, which provides a React interface to create interactive plots. The application has no backend, as the data is relatively small, requires no modification or manipulation per request, and is not subject to any privacy concerns due to its approval for release. All the data is stored in static JSON files that the application directly references to generate data visualizations. The website code and underlying data are publicly available on Github with an MIT license, which will allow others to contribute to the application as well as use the code to build visualizations for their own organizations.

4.4.23 Data Visualization

Data analysis was done in Python 3.7 using Jupyter Notebooks. Visualization was done using Seaborn and Matplotlib.

4.5 Figures

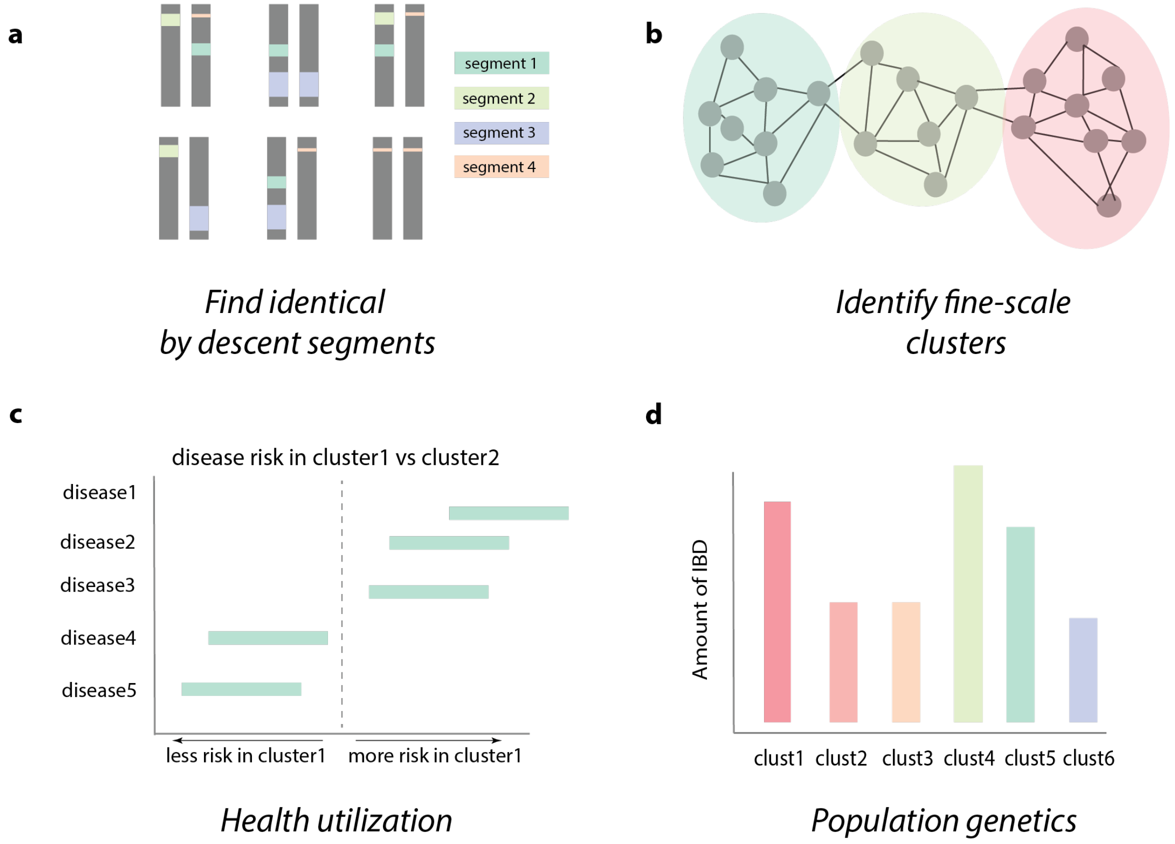
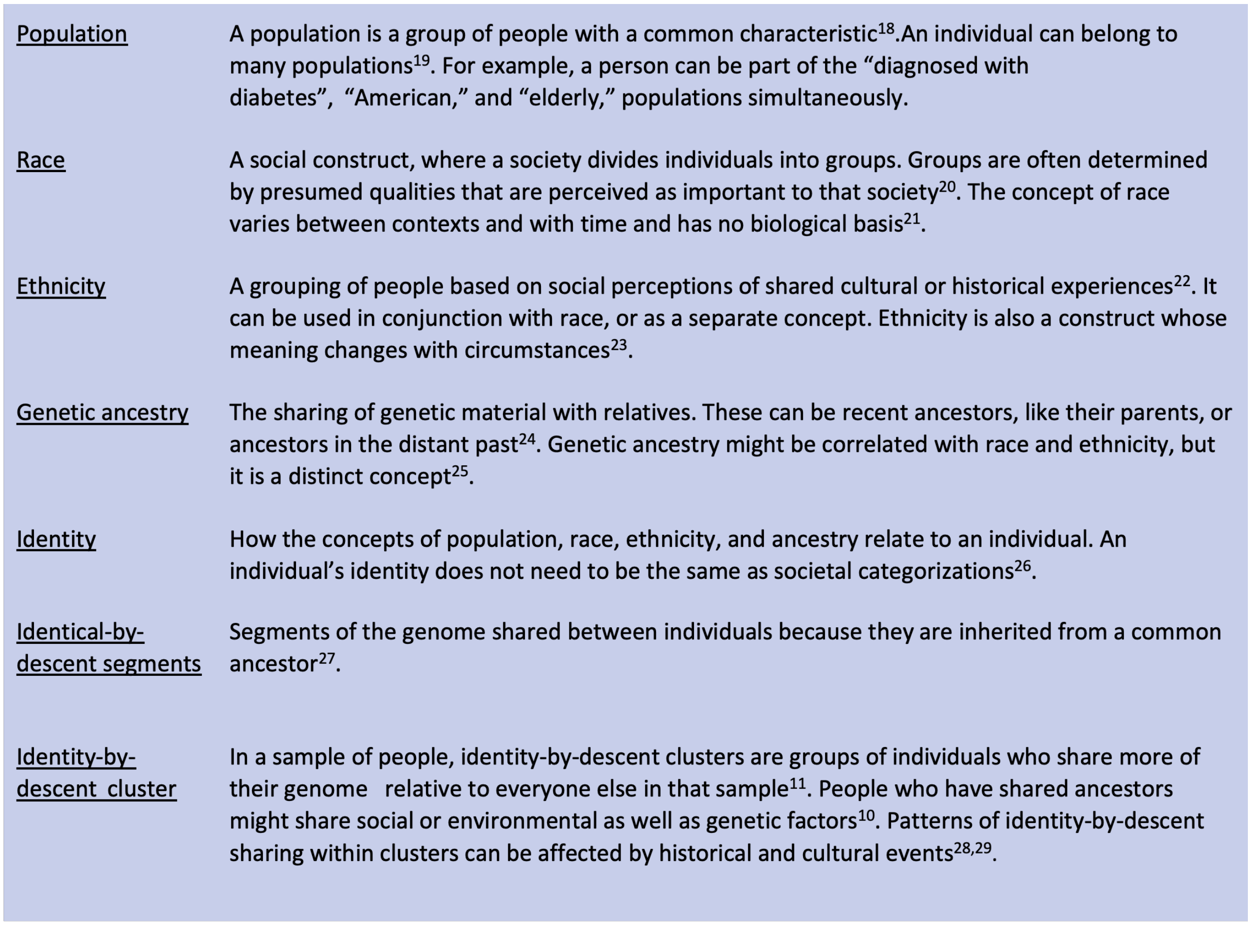


Figure 4.1: **An overview of the fine-scale cluster detection approach.** A schematic of identity-by-descent calling and cluster annotation. (a) We first infer identity-by-descent segments for all biobank participants and reference samples. (b) We then identify fine-scale clusters using Louvain clustering (c) and we explore patterns of enrichment for cluster-specific health utilization. (d) Finally, we measure patterns of genetic relatedness both within and between clusters.



Population

A population is a group of people with a common characteristic[18].An individual can belong to many populations[19]. For example, a person can be part of the “diagnosed with diabetes”, “American,” and “elderly,” populations simultaneously.

Race

A social construct, where a society divides individuals into groups. Groups are often determined by presumed qualities that are perceived as important to that society[20]. The concept of race varies between contexts and with time and has no biological basis[21].

Ethnicity

A grouping of people based on social perceptions of shared cultural or historical experiences[22]. It can be used in conjunction with race, or as a separate concept. Ethnicity is also a construct whose meaning changes with circumstances[23].

Genetic ancestry

The sharing of genetic material with relatives. These can be recent ancestors, like their parents, or ancestors in the distant past[24]. Genetic ancestry might be correlated with race and ethnicity, but it is a distinct concept[25].

Identity

How the concepts of population, race, ethnicity, and ancestry relate to an individual. An individual’s identity does not need to be the same as societal categorizations[26].

Identical-by-descent segments

Segments of the genome shared between individuals because they are inherited from a common ancestor[27].

Identity-by-descent cluster

In a sample of people, identity-by-descent clusters are groups of individuals who share more of their genome relative to everyone else in that sample[11]. People who have shared ancestors might share social or environmental as well as genetic factors[10]. Patterns of identity-by-descent sharing within clusters can be affected by historical and cultural events[28,29].

Figure 4.2: **Definitions of frequently used words relating to ancestry.**

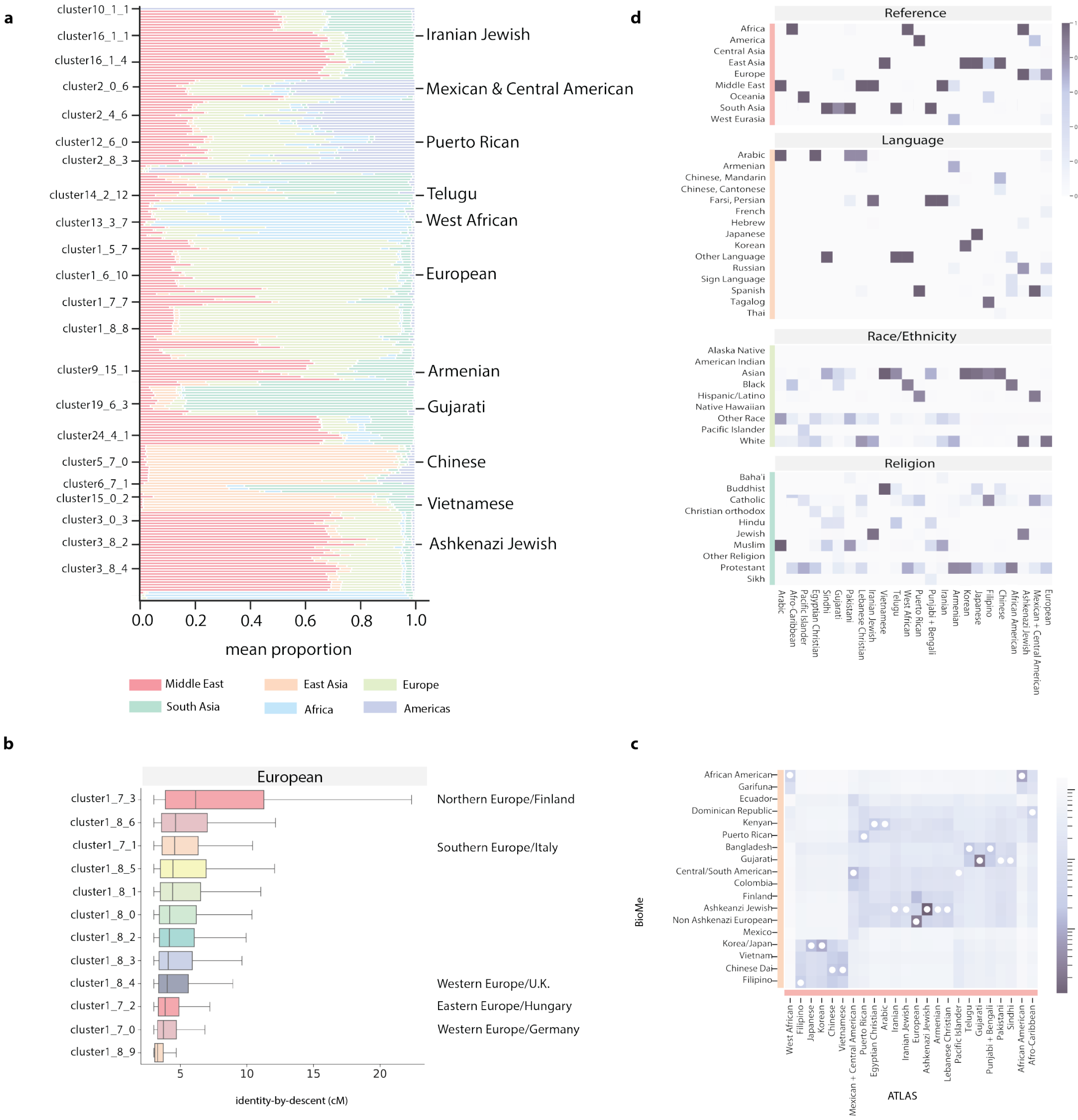


Figure 4.3: **Genetic and demographic properties of clusters** (a) The mean admixture fractions for each of the identity-by-descent clusters. Each line corresponds to one ATLAS cluster. The components refer to genetic ancestry from the Middle East, East Asia, Europe, South or Central Asia, Africa, and the Americas. The left column indicates the identity-by-decent cluster number, and the right column gives examples of names given to the largest clusters. (b) The distribution of identity-by-descent within subclusters that were merged to make one European cluster (n= 17017). The names on the left indicate the identity-by-descent cluster number, and the name on the right indicate relatedness from comparison with the UK BioBank. The center line of the box indicates the mean, the outer edges of the box indicate the upper and lower quartiles, and the whiskers indicate the maxima and minima of the distribution.(c) The Hudson's fixation index (FST) value between identity-by-descent clusters identified in BioMe at Mount Sinai and ATLAS identity-by-descent clusters demonstrates the relationship between ATLAS and populations outside of UCLA Health. The darker the color, the smaller the FST value. The smallest FST value for each of the ATLAS clusters is indicated by a white dot. (d) For each of the largest clusters, (from top to bottom) the proportion of reference data by continent in each cluster, the proportion that

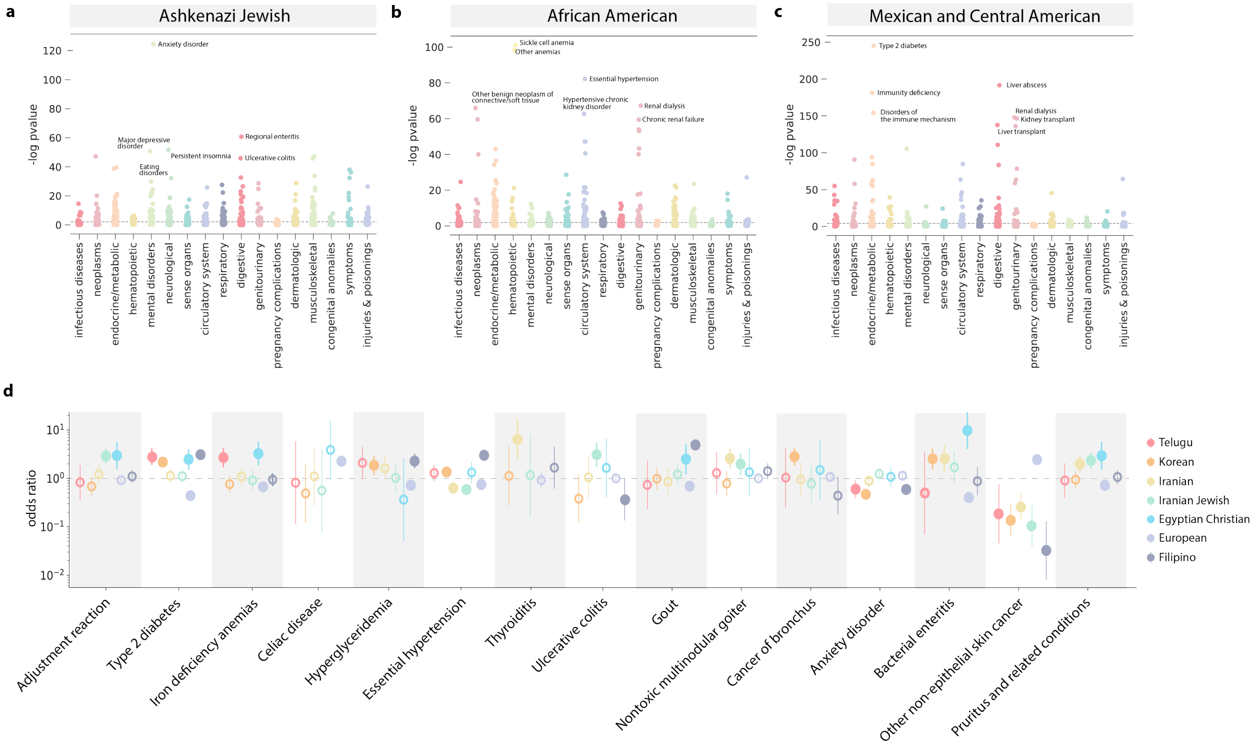


Figure 4.4: **Phecode associations for selected clusters.** Phecodes associations for (n=1131) identity-by-descent clusters relative to the remaining biobank participants. Results are shown for the (a) Ashkenazi Jewish (n=5309) (b) African American (n=1877) and (c) Mexican and Central American (n=6075) identity-by-descent clusters. Phecodes are grouped by phenotypic category. Top significant (Benjamini-Hochberg false discovery rate (FDR) at 5%) associations for each cluster are labeled, Bonferroni significance is indicated by a grey dotted line. (d) Odds ratios of association between identity-by-descent clusters and phecodes for the Telugu (n=276), Korean (n=546), Iranian (n=350), Iranian Jewish (n=264), Egyptian Christian (n=92), European (n=17017), and Filipino (n=796) clusters. Vertical bars indicate the standard error. Dots represent the odds ratio and a solid indicates significance at FDR 5%. Open dots indicate a non-significant association.

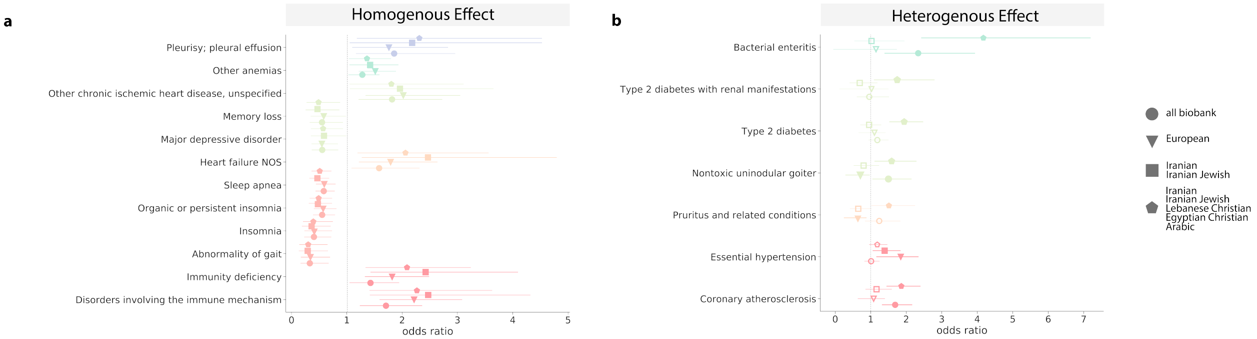


Figure 4.5: **Phecodes associated with the Armenian identity-by-descent cluster.** For each phecode, the odds ratio that membership in the Armenian cluster (n=491) was associated with that phecode, compared to the rest of the biobank, the European cluster (n=17017), the Iranian and Iranian Jewish clusters (n=614) and MENA ancestry clusters(n=960). In (a), phecodes that are FDR significant at 5% (where logistic regression q<0.05) in all comparison groups and had the same direction of effect ("homogenous effect"), are shown. In (b), phecodes that have a "heterogeneous effect," (mixed-effects meta-regression test where p<0.05) are shown. Phecodes of the same color are from the same phecode category. In each plot, the dot represents the odds ratio and the lines represent the standard error.

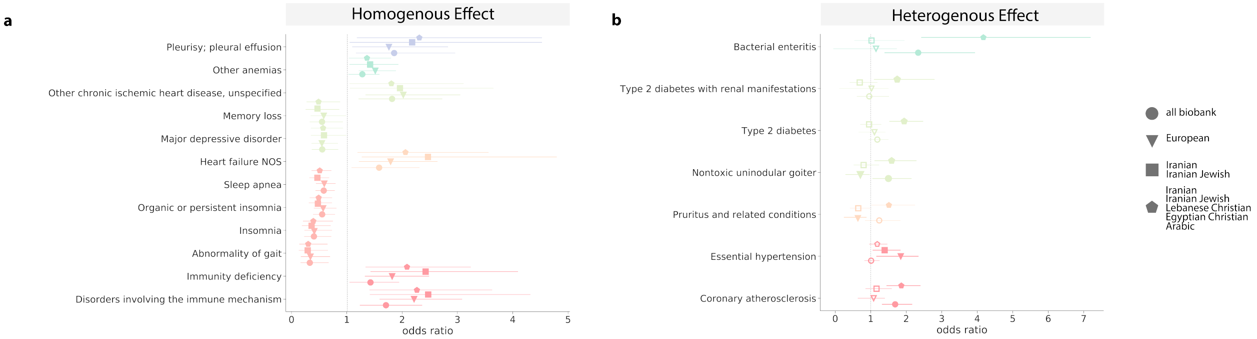
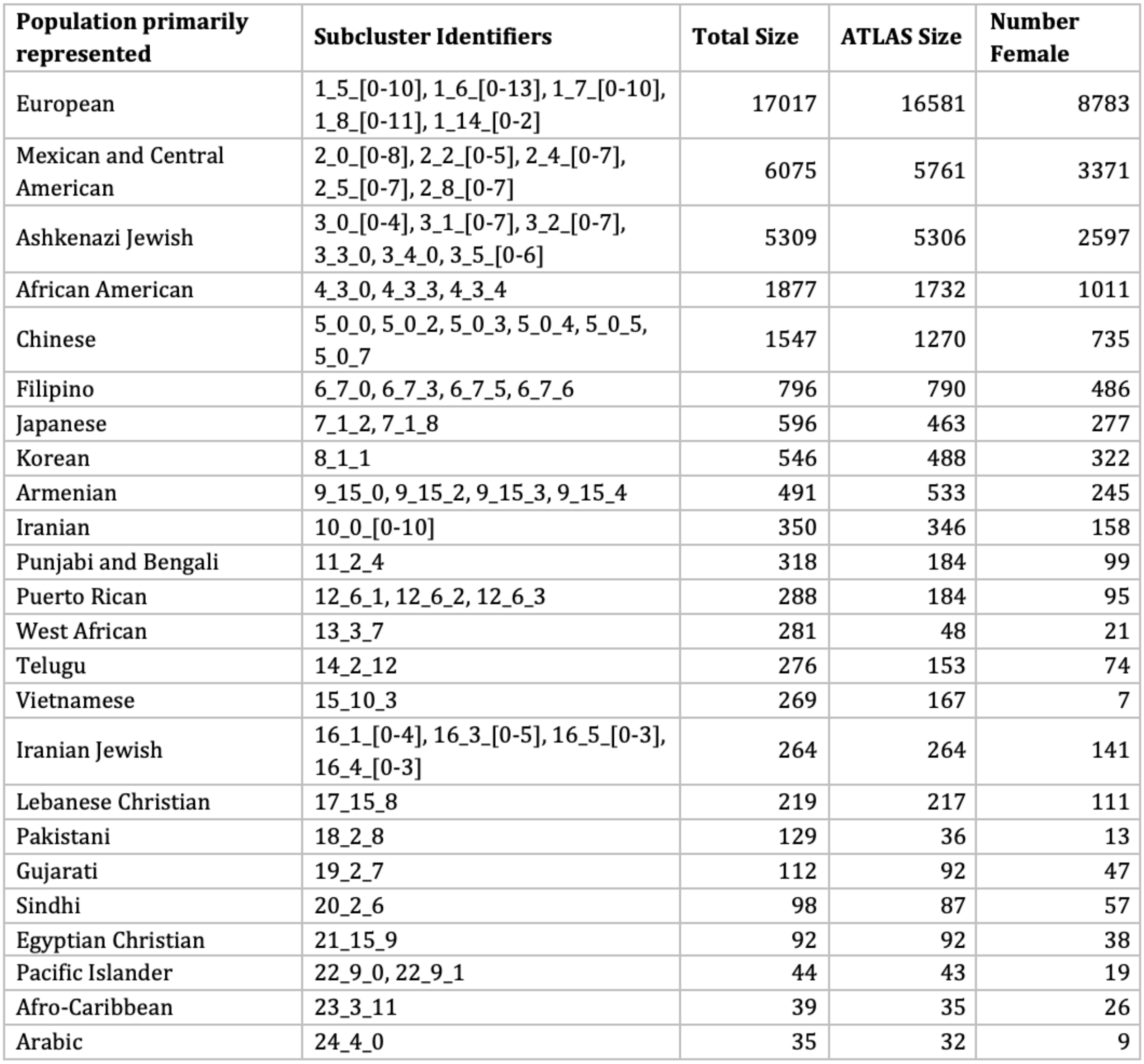


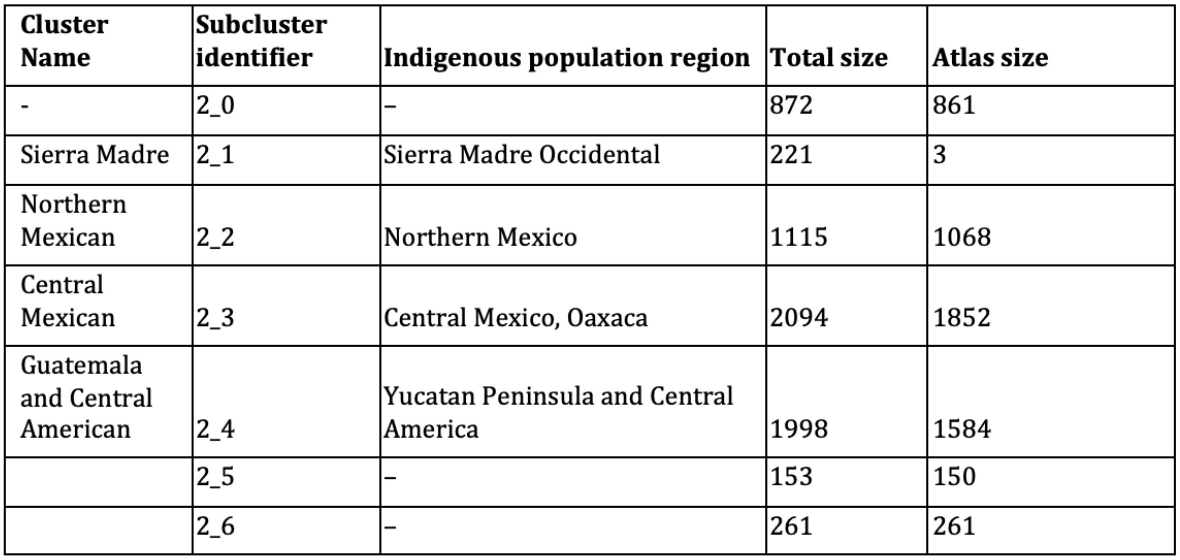
Figure 4.6: **The genetic properties of the largest identity-by-descent clusters.**(a) The distribution of total pairwise identity-by-descent (cM) and (b) total amount of ROH detected shared between individuals of a given cluster. The center line of the box indicates the mean, the outer edges of the box indicate the upper and lower quartiles, and the whiskers indicate the maxima and minima of the distribution. (c) IBDNe estimates of historic population size for 9 selected clusters, where the line is the mean estimate of the population size for each generation from present, and the shaded region indicates the 95% CI of the estimate. Dips in the population size can suggest founder effects. (d) Pairwise Hudson's FST estimates between UCLA ATLAS identity-by-descent clusters, where the darker color indicates lower FST, suggesting less differentiation between the pair of clusters. (e) A network diagram of identity-by-descent sharing between clusters, where each node is a cluster and each edge is weighted by the amount of identity-by-descent shared between the clusters. The graph was visualized using 1000 iterations of the Fruchterman-Reingold algorithm. For clarity, the 3 edges with the largest amount of identity-by-descent shared per cluster are displayed.

4.6 Tables



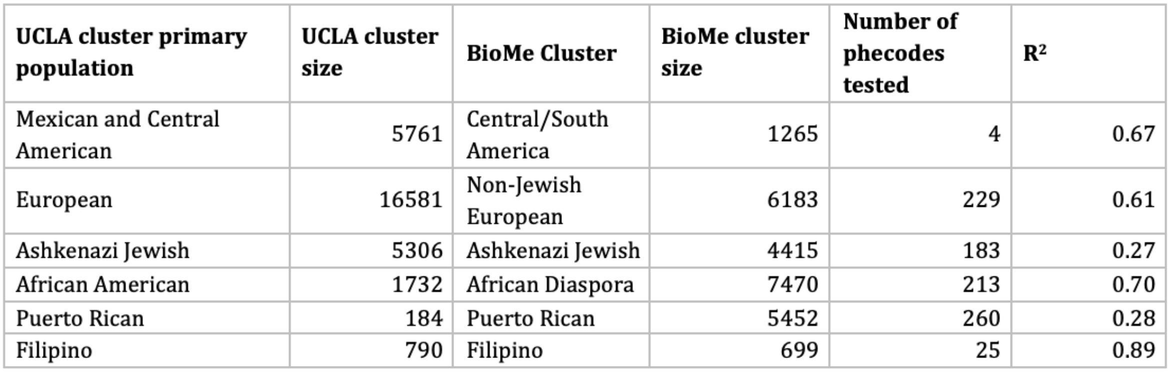
Population primarily represented	Subcluster Identifiers	Total Size	ATLAS Size	Number Female
European	1_5_[0-10], 1_6_[0-13], 1_7_[0-10], 1_8_[0-11], 1_14_[0-2]	17017	16581	8783
Mexican and Central American	2_0_[0-8], 2_2_[0-5], 2_4_[0-7], 2_5_[0-7], 2_8_[0-7]	6075	5761	3371
Ashkenazi Jewish	3_0_[0-4], 3_1_[0-7], 3_2_[0-7], 3_3_0, 3_4_0, 3_5_[0-6]	5309	5306	2597
African American	4_3_0, 4_3_3, 4_3_4	1877	1732	1011
Chinese	5_0_0, 5_0_2, 5_0_3, 5_0_4, 5_0_5, 5_0_7	1547	1270	735
Filipino	6_7_0, 6_7_3, 6_7_5, 6_7_6	796	790	486
Japanese	7_1_2, 7_1_8	596	463	277
Korean	8_1_1	546	488	322
Armenian	9_15_0, 9_15_2, 9_15_3, 9_15_4	491	533	245
Iranian	10_0_[0-10]	350	346	158
Punjabi and Bengali	11_2_4	318	184	99
Puerto Rican	12_6_1, 12_6_2, 12_6_3	288	184	95
West African	13_3_7	281	48	21
Telugu	14_2_12	276	153	74
Vietnamese	15_10_3	269	167	7
Iranian Jewish	16_1_[0-4], 16_3_[0-5], 16_5_[0-3], 16_4_[0-3]	264	264	141
Lebanese Christian	17_15_8	219	217	111
Pakistani	18_2_8	129	36	13
Gujarati	19_2_7	112	92	47
Sindhi	20_2_6	98	87	57
Egyptian Christian	21_15_9	92	92	38
Pacific Islander	22_9_0, 22_9_1	44	43	19
Afro-Caribbean	23_3_11	39	35	26
Arabic	24_4_0	35	32	9

Table 4.1: **Largest ATLAS identity-by-descent clusters.** For the 24 largest ATLAS identity-by-descent clusters, the cluster number, the population that the cluster primarily represents, the subclusters identified from the Louvain algorithm, and the total size of the cluster.



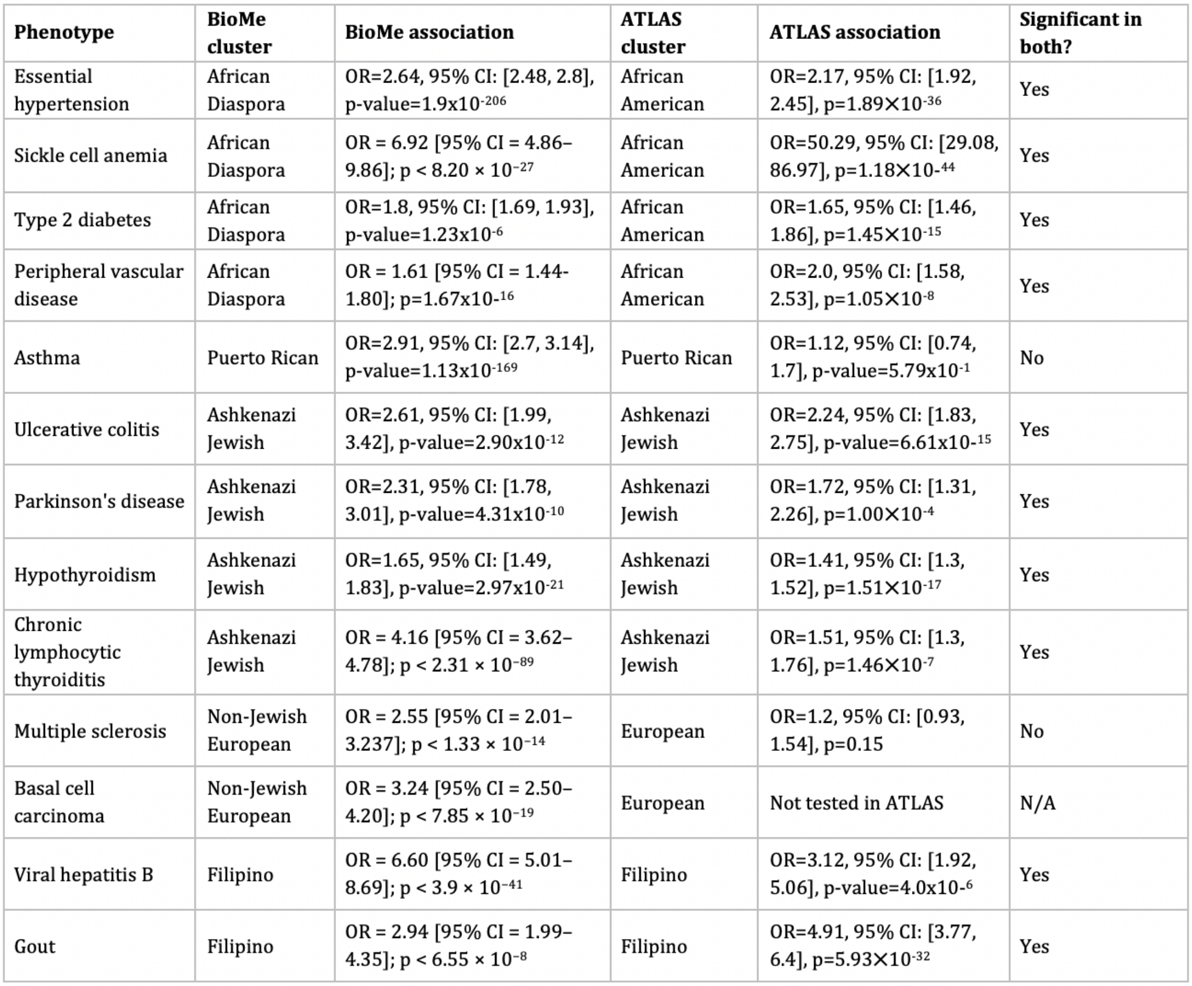
Cluster Name	Subcluster identifier	Indigenous population region	Total size	Atlas size
-	2_0	–	872	861
Sierra Madre	2_1	Sierra Madre Occidental	221	3
Northern Mexican	2_2	Northern Mexico	1115	1068
Central Mexican	2_3	Central Mexico, Oaxaca	2094	1852
Guatemala and Central American	2_4	Yucatan Peninsula and Central America	1998	1584
	2_5	–	153	150
	2_6	–	261	261

Table 4.2: **Mexican and Central American subclusters.** For the Mexican and Central American identity-by-descent cluster, the six subclusters identified. The four largest were given representative names.



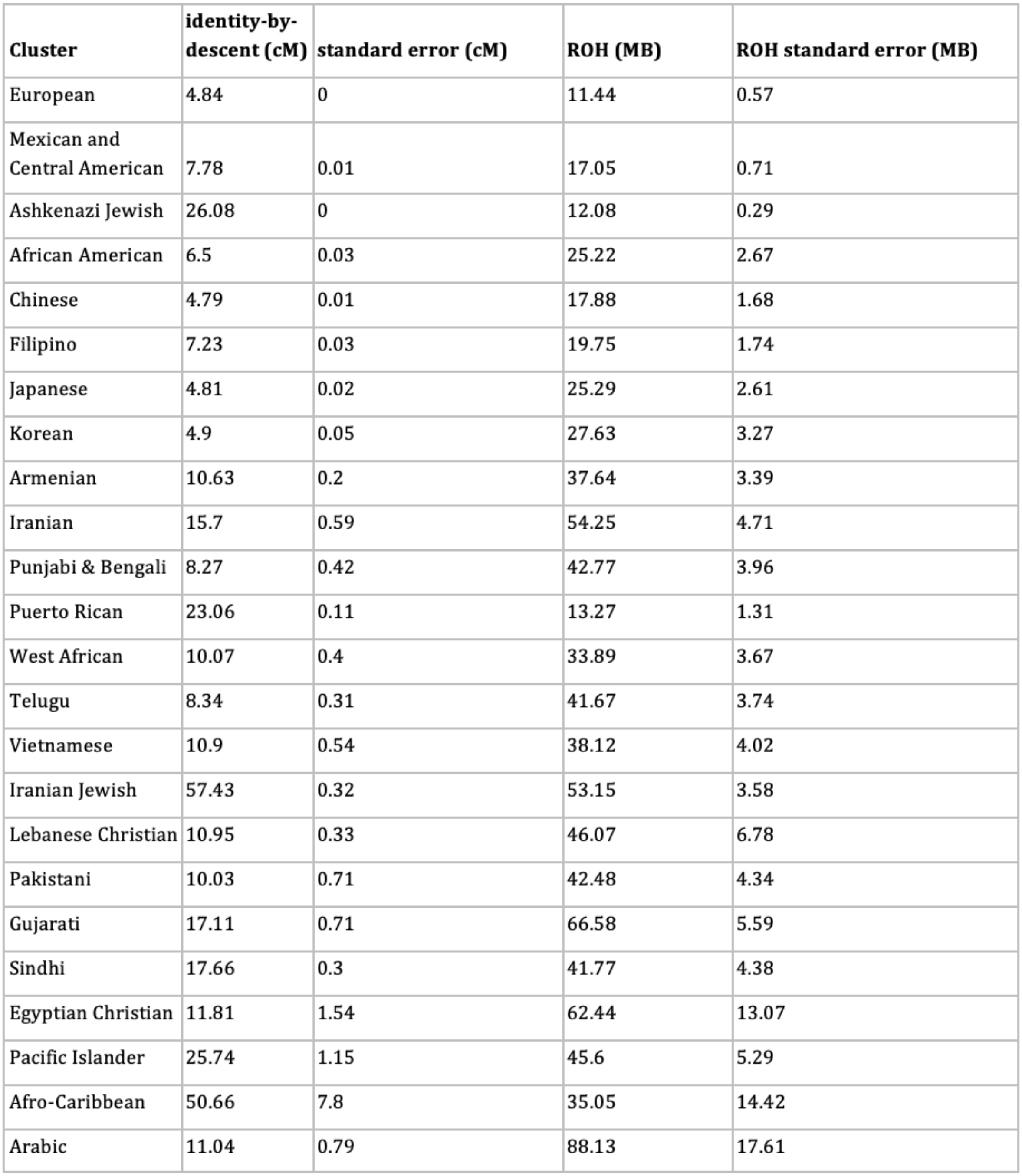
UCLA cluster primary population	UCLA cluster size	BioMe Cluster	BioMe cluster size	Number of phecodes tested	R^2
Mexican and Central American	5761	Central/South America	1265	4	0.67
European	16581	Non-Jewish European	6183	229	0.61
Ashkenazi Jewish	5306	Ashkenazi Jewish	4415	183	0.27
African American	1732	African Diaspora	7470	213	0.70
Puerto Rican	184	Puerto Rican	5452	260	0.28
Filipino	790	Filipino	699	25	0.89

Table 4.3: **Replication in BioMe.** Pearson's correlation between effect sizes of cluster-phecode associations calculated in BioMe and ATLAS for six clusters that are enriched for similar populations.



Phenotype	BioMe cluster	BioMe association	ATLAS cluster	ATLAS association	Significant in both?
Essential hypertension	African Diaspora	OR=2.64, 95% CI: [2.48, 2.8], p-value=1.9x10^{-206}	African American	OR=2.17, 95% CI: [1.92, 2.45], p=1.89×10^{-36}	Yes
Sickle cell anemia	African Diaspora	OR = 6.92 [95% CI = 4.86–9.86]; p < 8.20 × 10^{-27}	African American	OR=50.29, 95% CI: [29.08, 86.97], p=1.18×10^{-44}	Yes
Type 2 diabetes	African Diaspora	OR=1.8, 95% CI: [1.69, 1.93], p-value=1.23x10^{-6}	African American	OR=1.65, 95% CI: [1.46, 1.86], p=1.45×10^{-15}	Yes
Peripheral vascular disease	African Diaspora	OR = 1.61 [95% CI = 1.44-1.80]; p=1.67x10-^{16}	African American	OR=2.0, 95% CI: [1.58, 2.53], p=1.05×10^{-8}	Yes
Asthma	Puerto Rican	OR=2.91, 95% CI: [2.7, 3.14], p-value=1.13x10^{-169}	Puerto Rican	OR=1.12, 95% CI: [0.74, 1.7], p-value=5.79x10^{-1}	No
Ulcerative colitis	Ashkenazi Jewish	OR=2.61, 95% CI: [1.99, 3.42], p-value=2.90x10^{-12}	Ashkenazi Jewish	OR=2.24, 95% CI: [1.83, 2.75], p-value=6.61x10-^{15}	Yes
Parkinson's disease	Ashkenazi Jewish	OR=2.31, 95% CI: [1.78, 3.01], p-value=4.31x10^{-10}	Ashkenazi Jewish	OR=1.72, 95% CI: [1.31, 2.26], p=1.00×10^{-4}	Yes
Hypothyroidism	Ashkenazi Jewish	OR=1.65, 95% CI: [1.49, 1.83], p-value=2.97x10^{-21}	Ashkenazi Jewish	OR=1.41, 95% CI: [1.3, 1.52], p=1.51×10^{-17}	Yes
Chronic lymphocytic thyroiditis	Ashkenazi Jewish	OR = 4.16 [95% CI = 3.62–4.78]; p < 2.31 × 10^{-89}	Ashkenazi Jewish	OR=1.51, 95% CI: [1.3, 1.76], p=1.46×10^{-7}	Yes
Multiple sclerosis	Non-Jewish European	OR = 2.55 [95% CI = 2.01–3.237]; p < 1.33 × 10^{-14}	European	OR=1.2, 95% CI: [0.93, 1.54], p=0.15	No
Basal cell carcinoma	Non-Jewish European	OR = 3.24 [95% CI = 2.50–4.20]; p < 7.85 × 10^{-19}	European	Not tested in ATLAS	N/A
Viral hepatitis B	Filipino	OR = 6.60 [95% CI = 5.01–8.69]; p < 3.9 × 10^{-41}	Filipino	OR=3.12, 95% CI: [1.92, 5.06], p-value=4.0x10-^{6}	Yes
Gout	Filipino	OR = 2.94 [95% CI = 1.99–4.35]; p < 6.55 × 10^{-8}	Filipino	OR=4.91, 95% CI: [3.77, 6.4], p=5.93×10^{-32}	Yes

Table 4.4: **BioMe phenotypes in ATLAS.** For cluster-phecode associations published in BioMe, the odds ratios and p-values are obtained from a logistic regression analysis for 6 clusters in ATLAS and in BioME that are enriched for similar populations.



Cluster	identity-by-descent (cM)	standard error (cM)	ROH (MB)	ROH standard error (MB)
European	4.84	0	11.44	0.57
Mexican and Central American	7.78	0.01	17.05	0.71
Ashkenazi Jewish	26.08	0	12.08	0.29
African American	6.5	0.03	25.22	2.67
Chinese	4.79	0.01	17.88	1.68
Filipino	7.23	0.03	19.75	1.74
Japanese	4.81	0.02	25.29	2.61
Korean	4.9	0.05	27.63	3.27
Armenian	10.63	0.2	37.64	3.39
Iranian	15.7	0.59	54.25	4.71
Punjabi & Bengali	8.27	0.42	42.77	3.96
Puerto Rican	23.06	0.11	13.27	1.31
West African	10.07	0.4	33.89	3.67
Telugu	8.34	0.31	41.67	3.74
Vietnamese	10.9	0.54	38.12	4.02
Iranian Jewish	57.43	0.32	53.15	3.58
Lebanese Christian	10.95	0.33	46.07	6.78
Pakistani	10.03	0.71	42.48	4.34
Gujarati	17.11	0.71	66.58	5.59
Sindhi	17.66	0.3	41.77	4.38
Egyptian Christian	11.81	1.54	62.44	13.07
Pacific Islander	25.74	1.15	45.6	5.29
Afro-Caribbean	50.66	7.8	35.05	14.42
Arabic	11.04	0.79	88.13	17.61

Table 4.5: **Identity-by-descent and ROH within clusters.** For the 24 largest ATLAS clusters, the mean total pairwise identity-by-descent detected between individuals in the cluster and the mean ROH detected within individuals of the cluster.

CHAPTER 5

Bibliography

BIBLIOGRAPHY

[1] N. S. Abul-Husn, E. R. Soper, J. A. Odgis, S. Cullina, D. Bobo, A. Moscati, J. E. Rodriguez, R. J. F. Loos, J. H. Cho, G. M. Belbin, S. A. Suckiel, E. E. Kenny, CBIPM Genomics Team, and Regeneron Genetics Center, "Exome sequencing reveals a high prevalence of BRCA1 and BRCA2 founder variants in a diverse population-based biobank," *Genome Medicine*, vol. 12, no. 1, p. 2, Dec. 2019. [Online]. Available: https://doi.org/10.1186/s13073-019-0691-1 73, 77

[2] N. S. Abul-Husn, E. R. Soper, G. T. Braganza, J. E. Rodriguez, N. Zeid, S. Cullina, D. Bobo, A. Moscati, A. Merkelson, R. J. F. Loos, J. H. Cho, G. M. Belbin, S. A. Suckiel, and E. E. Kenny, "Implementing genomic screening in diverse populations," *Genome Medicine*, vol. 13, no. 1, p. 17, Feb. 2021. [Online]. Available: https://doi.org/10.1186/s13073-021-00832-y 75, 77, 78

[3] S. Andrews, "s-andrews/FastQC," Jan. 2021, original-date: 2017-12-21T11:48:51Z. [Online]. Available: https://github.com/s-andrews/FastQC 33

[4] A. Arjunan, D. R. Darnes, K. G. Sagaser, and A. B. Svenson, "Addressing Reproductive Healthcare Disparities through Equitable Carrier Screening: Medical Racism and Genetic Discrimination in United States' History Highlights the Needs for Change in Obstetrical Genetics Care," *Societies*, vol. 12, no. 2, p. 33, Apr. 2022, number: 2 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2075-4698/12/2/33 77

[5] E. A. Ashley, "Towards precision medicine," *Nature Reviews Genetics*, vol. 17, no. 9, pp. 507–522, Sep. 2016, number: 9 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nrg.2016.86 1

[6] A. Auton, G. R. Abecasis, D. M. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, E. E. Eichler, P. Flicek, S. B. Gabriel, R. A. Gibbs, E. D. Green, M. E. Hurles, B. M. Knoppers, J. O. Korbel,

E. S. Lander, C. Lee, H. Lehrach, E. R. Mardis, G. T. Marth, G. A. McVean, D. A. Nickerson, J. P. Schmidt, S. T. Sherry, J. Wang, R. K. Wilson, R. A. Gibbs, E. Boerwinkle, H. Doddapaneni, Y. Han, V. Korchina, C. Kovar, S. Lee, D. Muzny, J. G. Reid, Y. Zhu, J. Wang, Y. Chang, Q. Feng, X. Fang, X. Guo, M. Jian, H. Jiang, X. Jin, T. Lan, G. Li, J. Li, Y. Li, S. Liu, X. Liu, Y. Lu, X. Ma, M. Tang, B. Wang, G. Wang, H. Wu, R. Wu, X. Xu, Y. Yin, D. Zhang, W. Zhang, J. Zhao, M. Zhao, X. Zheng, E. S. Lander, D. M. Altshuler, S. B. Gabriel, N. Gupta, N. Gharani, L. H. Toji, N. P. Gerry, A. M. Resch, P. Flicek, J. Barker, L. Clarke, L. Gil, S. E. Hunt, G. Kelman, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, A. Roa, D. Smirnov, R. E. Smith, I. Streeter, A. Thormann, I. Toneva, B. Vaughan, X. Zheng-Bradley, D. R. Bentley, R. Grocock, S. Humphray, T. James, Z. Kingsbury, H. Lehrach, R. Sudbrak, M. W. Albrecht, V. S. Amstislavskiy, T. A. Borodina, M. Lienhard, F. Mertes, M. Sultan, B. Timmermann, M.-L. Yaspo, E. R. Mardis, R. K. Wilson, L. Fulton, R. Fulton, S. T. Sherry, V. Ananiev, Z. Belaia, D. Beloslyudtsev, N. Bouk, C. Chen, D. Church, R. Cohen, C. Cook, J. Garner, T. Hefferon, M. Kimelman, C. Liu, J. Lopez, P. Meric, C. O'Sullivan, Y. Ostapchuk, L. Phan, S. Ponomarov, V. Schneider, E. Shekhtman, K. Sirotkin, D. Slotta, H. Zhang, G. A. McVean, R. M. Durbin, S. Balasubramaniam, J. Burton, P. Danecek, T. M. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, J. P. Schmidt, C. J. Davies, J. Gollub, T. Webster, B. Wong, Y. Zhan, A. Auton, C. L. Campbell, Y. Kong, A. Marcketta, R. A. Gibbs, F. Yu, L. Antunes, M. Bainbridge, D. Muzny, A. Sabo, Z. Huang, J. Wang, L. J. M. Coin, L. Fang, X. Guo, X. Jin, G. Li, Q. Li, Y. Li, Z. Li, H. Lin, B. Liu, R. Luo, H. Shao, Y. Xie, C. Ye, C. Yu, F. Zhang, H. Zheng, H. Zhu, C. Alkan, E. Dal, F. Kahveci, G. T. Marth, E. P. Garrison, D. Kural, W.-P. Lee, W. Fung Leong, M. Stromberg, A. N. Ward, J. Wu, M. Zhang, M. J. Daly, M. A. DePristo, R. E. Handsaker, D. M. Altshuler, E. Banks, G. Bhatia, G. del Angel, S. B. Gabriel, G. Genovese, N. Gupta, H. Li, S. Kashin, E. S. Lander, S. A. McCarroll, J. C. Nemesh, R. E. Poplin, S. C. Yoon, J. Lihm, V. Makarov, A. G. Clark, S. Gottipati, A. Keinan,

J. L. Rodriguez-Flores, J. O. Korbel, T. Rausch, M. H. Fritz, A. M. Stütz, P. Flicek, K. Beal, L. Clarke, A. Datta, J. Herrero, W. M. McLaren, G. R. S. Ritchie, R. E. Smith, D. Zerbino, X. Zheng-Bradley, P. C. Sabeti, I. Shlyakhter, S. F. Schaffner, J. Vitti, D. N. Cooper, E. V. Ball, P. D. Stenson, D. R. Bentley, B. Barnes, M. Bauer, R. Keira Cheetham, A. Cox, M. Eberle, S. Humphray, S. Kahn, L. Murray, J. Peden, R. Shaw, E. E. Kenny, M. A. Batzer, M. K. Konkel, J. A. Walker, D. G. MacArthur, M. Lek, R. Sudbrak, V. S. Amstislavskiy, R. Herwig, E. R. Mardis, L. Ding, D. C. Koboldt, D. Larson, K. Ye, S. Gravel, The 1000 Genomes Project Consortium, Corresponding authors, Steering committee, Production group, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, Coriell Institute for Medical Research, E. B. I. 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 N. National Eye Institute, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, Oct. 2015, number: 7571 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nature15393 15, 66

[7] Z. D. Bailey, J. M. Feldman, and M. T. Bassett, "How Structural Racism Works — Racist Policies as a Root Cause of U.S. Racial Health Inequities," *New England Journal of Medicine*, vol. 384, no. 8, pp. 768–773, Feb. 2021, publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMms2025396. [Online]. Available: https://doi.org/10.1056/NEJMms2025396 2, 77

[8] A. C. Bax, D. E. Bard, S. P. Cuffe, R. E. McKeown, and M. L. Wolraich, "The Association Between Race/Ethnicity and Socioeconomic Factors and

the Diagnosis and Treatment of Children with Attention-Deficit Hyperactivity Disorder," *Journal of Developmental & Behavioral Pediatrics*, vol. 40, no. 2, pp. 81–91, Mar. 2019. [Online]. Available: https://journals.lww.com/jrnldbp/Fulltext/2019/02000/The_Association_Between_Race_Ethnicity_and.1.aspx?casa_token=12BlZTejgIEAAAAA:7KiH1onD6fUCFADiKV8wMxTlyScDzfNk5JUp_jkSpuotE4v-_BlJt3iLAYKabyPtRAlB-y1euUKkdpohz1EF3gub 78

[9] M. Bazargan, J. L. Smith, S. Cobb, L. Barkley, C. Wisseh, E. Ngula, R. J. Thomas, and S. Assari, "Emergency Department Utilization among Underserved African American Older Adults in South Los Angeles," *International Journal of Environmental Research and Public Health*, vol. 16, no. 7, p. 1175, Jan. 2019, number: 7 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1660-4601/16/7/1175 72

[10] G. M. Belbin, S. Cullina, S. Wenric, E. R. Soper, B. S. Glicksberg, D. Torre, A. Moscati, G. L. Wojcik, R. Shemirani, N. D. Beckmann, A. Cohain, E. P. Sorokin, D. S. Park, J.-L. Ambite, S. Ellis, A. Auton, E. P. Bottinger, J. H. Cho, R. J. F. Loos, N. S. Abul-Husn, N. A. Zaitlen, C. R. Gignoux, and E. E. Kenny, "Toward a fine-scale population health monitoring system," *Cell*, vol. 184, no. 8, pp. 2068–2083.e11, Apr. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092867421003652 64, 65, 83

[11] G. M. Belbin, S. Rutledge, T. Dodatko, S. Cullina, M. C. Turchin, S. Kohli, D. Torre, M.-C. Yee, C. R. Gignoux, N. S. Abul-Husn, S. M. Houten, and E. E. Kenny, "Leveraging Health Systems Data to Characterize a Large Effect Variant Conferring Risk for Liver Disease in Puerto Ricans," *medRxiv*, p. 2021.03.31.21254662, Apr. 2021, publisher: Cold Spring Harbor Laboratory Press. [Online]. Available: https://www.medrxiv.org/content/10.1101/2021.03.31.21254662v1 66, 67, 68, 69

[12] G. M. Belbin, J. Odgis, E. P. Sorokin, M.-C. Yee, S. Kohli, B. S. Glicksberg, C. R. Gignoux, G. L. Wojcik, T. Van Vleck, J. M. Jeff, M. Linderman, C. Schurmann,

D. Ruderfer, X. Cai, A. Merkelson, A. E. Justice, K. L. Young, M. Graff, K. E. North, U. Peters, R. James, L. Hindorff, R. Kornreich, L. Edelmann, O. Gottesman, E. E. Stahl, J. H. Cho, R. J. Loos, E. P. Bottinger, G. N. Nadkarni, N. S. Abul-Husn, and E. E. Kenny, "Genetic identification of a common collagen disease in Puerto Ricans via identity-by-descent mapping in a health system," *eLife*, vol. 6, p. e25060, Sep. 2017, publisher: eLife Sciences Publications, Ltd. [Online]. Available: https://doi.org/10.7554/eLife.25060 75

[13] A. Bergström, C. Stringer, M. Hajdinjak, E. M. L. Scerri, and P. Skoglund, "Origins of modern human ancestry," *Nature*, vol. 590, no. 7845, pp. 229–237, Feb. 2021, number: 7845 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41586-021-03244-5 66

[14] B. Bettinger, "Shared cM Project," 2020. [Online]. Available: https://dnapainter.com/tools/sharedcmv4 84

[15] G. Bhatia, N. J. Patterson, S. Sankararaman, and A. L. Price, "Estimating and interpreting Fst: the impact of rare variants," *Genome Research*, p. gr.154831.113, Jul. 2013, company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. [Online]. Available: https://genome.cshlp.org/content/early/2013/07/16/gr.154831.113 76, 86

[16] K. Bjornevik, M. Cortese, B. C. Healy, J. Kuhle, M. J. Mina, Y. Leng, S. J. Elledge, D. W. Niebuhr, A. I. Scher, K. L. Munger, and A. Ascherio, "Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis," *Science*, Jan. 2022, publisher: American Association for the Advancement of Science. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.abj8222 43

[17] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and*

Experiment, vol. 2008, no. 10, p. P10008, Oct. 2008, arXiv: 0803.0476. [Online]. Available: http://arxiv.org/abs/0803.0476 66, 85

[18] S. L. Boddy, I. Giovannelli, M. Sassani, J. Cooper-Knock, M. P. Snyder, E. Segal, E. Elinav, L. A. Barker, P. J. Shaw, and C. J. McDermott, "The gut microbiome: a key player in the complexity of amyotrophic lateral sclerosis (ALS)," *BMC Medicine*, vol. 19, no. 1, p. 13, Jan. 2021. [Online]. Available: https://doi.org/10.1186/s12916-020-01885-3 44

[19] L. N. Borrell, J. R. Elhawary, E. Fuentes-Afflick, J. Witonsky, N. Bhakta, A. H. Wu, K. Bibbins-Domingo, J. R. Rodríguez-Santana, M. A. Lenoir, J. R. Gavin, R. A. Kittles, N. A. Zaitlen, D. S. Wilkes, N. R. Powe, E. Ziv, and E. G. Burchard, "Race and Genetic Ancestry in Medicine — A Time for Reckoning with Racism," *New England Journal of Medicine*, vol. 384, no. 5, pp. 474–480, Feb. 2021, publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMms2029562. [Online]. Available: https://doi.org/10.1056/NEJMms2029562 77, 78

[20] R. Bowser, M. R. Turner, and J. Shefner, "Biomarkers in amyotrophic lateral sclerosis: opportunities and limitations," *Nature Reviews Neurology*, vol. 7, no. 11, pp. 631–638, Nov. 2011. [Online]. Available: https://www.nature.com/articles/nrneurol.2011.151 5

[21] S. R. Browning and B. L. Browning, "Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent," *American Journal of Human Genetics*, vol. 97, no. 3, pp. 404–418, Sep. 2015. 75

[22] U. C. Bureau, "U.S. Census Bureau QuickFacts: Los Angeles city, California," 2021. [Online]. Available: https://www.census.gov/quickfacts/losangelescitycalifornia 65

[23] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, and J. Marchini, "The UK Biobank resource with deep phenotyping and genomic data," *Nature*,

vol. 562, no. 7726, pp. 203–209, Oct. 2018, bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 7726 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genome;Genome-wide association studies;Genotype;Haplotypes;Population genetics Subject_term_id: genome;genome-wide-association-studies;genotype;haplotypes;population-genetics. [Online]. Available: https://www.nature.com/articles/s41586-018-0579-z. 67

[24] C. Caggiano, B. Celona, F. Garton, J. Mefford, B. L. Black, R. Henderson, C. Lomen-Hoerth, A. Dahl, and N. Zaitlen, "Comprehensive cell type decomposition of circulating cell-free DNA with CelFiE," *Nature Communications*, vol. 12, no. 1, p. 2717, May 2021, bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Diagnostic markers;DNA methylation;Epigenomics;Statistical methods Subject_term_id: diagnostic-markers;dna-methylation;epigenomics;statistical-methods. [Online]. Available: https://www.nature.com/articles/s41467-021-22901-x 44

[25] T. Carlice-dos Reis, J. Viana, F. C. Moreira, G. d. L. Cardoso, J. Guerreiro, S. Santos, and Ribeiro-dos Santos, "Investigation of mutations in the HBB gene using the 1,000 genomes database," *PLoS ONE*, vol. 12, no. 4, p. e0174637, Apr. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5381778/ 74

[26] H. Carress, D. J. Lawson, and E. Elhaik, "Population genetic considerations for using biobanks as international resources in the pandemic era and beyond," *BMC Genomics*, vol. 22, no. 1, p. 351, May 2021. [Online]. Available: https://doi.org/10.1186/s12864-021-07618-x 65

[27] CDC, "People Born Outside of the United States and Viral Hepatitis," Sep. 2020. [Online]. Available: https://www.cdc.gov/hepatitis/populations/Born-Outside-United-States.htm 70

[28] F. C. Ceballos, P. K. Joshi, D. W. Clark, M. Ramsay, and J. F. Wilson, "Runs of homozygosity: windows into population history and trait architecture," *Nature*

Reviews Genetics, vol. 19, no. 4, pp. 220–234, Apr. 2018, number: 4 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nrg.2017.109 75

[29] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, and A. Nakanishi, "The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function," *Journal of the Neurological Sciences*, vol. 169, no. 1, pp. 13–21, Oct. 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022510X99002105 45

[30] J.-C. Chacón-Duque, K. Adhikari, M. Fuentes-Guajardo, J. Mendoza-Revilla, V. Acuña-Alonzo, R. Barquera, M. Quinto-Sánchez, J. Gómez-Valdés, P. Everardo Martínez, H. Villamil-Ramírez, T. Hünemeier, V. Ramallo, C. C. Silva de Cerqueira, M. Hurtado, V. Villegas, V. Granja, M. Villena, R. Vásquez, E. Llop, J. R. Sandoval, A. A. Salazar-Granara, M.-L. Parolin, K. Sandoval, R. I. Peñaloza-Espinosa, H. Rangel-Villalobos, C. A. Winkler, W. Klitz, C. Bravi, J. Molina, D. Corach, R. Barrantes, V. Gomes, C. Resende, L. Gusmão, A. Amorim, Y. Xue, J.-M. Dugoujon, P. Moral, R. González-José, L. Schuler-Faccini, F. M. Salzano, M.-C. Bortolini, S. Canizales-Quinteros, G. Poletti, C. Gallo, G. Bedoya, F. Rothhammer, D. Balding, G. Hellenthal, and A. Ruiz-Linares, "Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance," *Nature Communications*, vol. 9, no. 1, p. 5388, Dec. 2018, bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genetic variation;Quantitative trait loci Subject_term_id: genetic-variation;quantitative-trait-loci. [Online]. Available: https://www.nature.com/articles/s41467-018-07748-z 76

[31] T. S. Chang, Y. Ding, M. K. Freund, R. Johnson, T. Schwarz, J. M. Yabu, C. Hazlett, J. N. Chiang, D. A. Wulf, A. L. Antonio, M. Ariannejad, A. M. Badillo, B. Balliu, Y. Berkovich, M. Broudy, T. Dang, C. Denny, E. Eskin, E. Halperin, B. L. Hill, A. Jain, V. Katakwar, C. Lajonchere, C. Magyar, S. Minton,

G. Mohammed, A. Muhamed, P. Pavan, M. A. Pfeffer, N. Rakocz, A. Rudas, R. Salonga, T. J. Sanders, P. Tung, V. Vu, A. Zheng, D. H. Geschwind, M. J. Butte, and B. Pasaniuc, “Pre-existing conditions in Hispanics/Latinxs that are COVID-19 risk factors,” *iScience*, vol. 24, no. 3, p. 102188, Mar. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2589004221001565 79

[32] Z. Chatterton, N. Mendelev, S. Chen, T. Raj, R. Walker, W. Carr, G. Kamimori, M. Beeri, Y. Ge, A. Dwork, and F. Haghighi, “Brain-derived circulating cell-free DNA defines the brain region and cell specific origins associated with neuronal atrophy,” *bioRxiv*, p. 538827, Feb. 2019, publisher: Cold Spring Harbor Laboratory Section: New Results. [Online]. Available: https://www.biorxiv.org/content/10.1101/538827v1 6

[33] R. H. Chipika, E. Finegan, S. Li Hi Shing, O. Hardiman, and P. Bede, “Tracking a Fast-Moving Disease: Longitudinal Markers, Monitoring, and Clinical Trial Endpoints in ALS,” *Frontiers in Neurology*, vol. 10, p. 229, Mar. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6433752/ 43

[34] J. L. Clarke, “Impact of Pan-Ethnic Expanded Carrier Screening in Improving Population Health Outcomes: Proceedings from a Multi-Stakeholder Virtual Roundtable Summit, June 25, 2020,” *Population Health Management*, vol. 24, no. 5, pp. 622–630, Oct. 2021, publisher: Mary Ann Liebert, Inc., publishers. [Online]. Available: https://www.liebertpub.com/doi/10.1089/pop.2021.0073 77

[35] S. Cobb, M. Bazargan, S. Assari, L. Barkley, and S. Bazargan-Hejazi, “Emergency Department Utilization, Hospital Admissions, and Office-Based Physician Visits Among Under-Resourced African American and Latino Older Adults,” *Journal of Racial and Ethnic Health Disparities*, Jan. 2022. [Online]. Available: https://doi.org/10.1007/s40615-021-01211-4 72

[36] B. T. Constantino and B. Cogionis, “Nucleated RBCs—Significance in the Peripheral

Blood Film," *Laboratory Medicine*, vol. 31, no. 4, pp. 223–229, Apr. 2000. [Online]. Available: https://doi.org/10.1309/D70F-HCC1-XX1T-4ETE 19

[37] R. B. Corcoran and B. A. Chabner, "Application of Cell-free DNA Analysis to Cancer Treatment," *New England Journal of Medicine*, vol. 379, no. 18, pp. 1754–1765, Nov. 2018, publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMra1706174. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMra1706174 44

[38] R. A. Corriveau, W. J. Koroshetz, J. T. Gladman, S. Jeon, D. Babcock, D. A. Bennett, S. T. Carmichael, S. L.-J. Dickinson, D. W. Dickson, M. Emr, H. Fillit, S. M. Greenberg, M. L. Hutton, D. S. Knopman, J. J. Manly, K. S. Marder, C. S. Moy, C. H. Phelps, P. A. Scott, W. W. Seeley, B.-A. Sieber, N. B. Silverberg, M. L. Sutherland, A. Taylor, C. L. Torborg, S. P. Waddy, A. K. Gubitz, and D. M. Holtzman, "Alzheimer's Disease–Related Dementias Summit 2016: National research priorities," *Neurology*, vol. 89, no. 23, pp. 2381–2391, Dec. 2017, publisher: Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology Section: Views &amp; Reviews. [Online]. Available: https://n.neurology.org/content/89/23/2381 68

[39] C. L. Dai, M. M. Vazifeh, C.-H. Yeang, R. Tachet, R. S. Wells, M. G. Vilar, M. J. Daly, C. Ratti, and A. R. Martin, "Population Histories of the United States Revealed through Fine-Scale Migration and Haplotype Analysis," *The American Journal of Human Genetics*, vol. 106, no. 3, pp. 371–388, Mar. 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0002929720300446 64, 66

[40] S. M. Damrauer, K. Chaudhary, J. H. Cho, L. W. Liang, E. Argulian, L. Chan, A. Dobbyn, M. A. Guerraty, R. Judy, J. Kay, R. L. Kember, M. G. Levin, A. Saha, T. Van Vleck, S. S. Verma, J. Weaver, N. S. Abul-Husn, A. Baras, J. A. Chirinos, B. Drachman, E. E. Kenny, R. J. F. Loos, J. Narula, J. Overton, J. Reid, M. Ritchie, G. Sirugo, G. Nadkarni, D. J. Rader, and R. Do, "Association of the V122I Hereditary Transthyretin Amyloidosis Genetic Variant With Heart Failure Among Individuals of

African or Hispanic/Latino Ancestry," *JAMA*, vol. 322, no. 22, pp. 2191–2202, Dec. 2019. 74

[41] C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, K. C. Onate, K. Graham, S. R. Miyasato, T. R. Dreszer, J. S. Strattan, O. Jolanki, F. Y. Tanaka, and J. M. Cherry, "The Encyclopedia of DNA elements (ENCODE): data portal update," *Nucleic Acids Research*, vol. 46, no. D1, pp. D794–D801, 2018. 14

[42] O. Delaneau, J.-F. Zagury, M. R. Robinson, J. L. Marchini, and E. T. Dermitzakis, "Accurate, scalable and integrative haplotype estimation," *Nature Communications*, vol. 10, no. 1, p. 5436, Nov. 2019, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-019-13225-y 83

[43] A. Deyrup and J. L. Graves, "Racial Biology and Medical Misconceptions," *New England Journal of Medicine*, vol. 386, no. 6, pp. 501–503, Feb. 2022, publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMp2116224. [Online]. Available: https://doi.org/10.1056/NEJMp2116224 78

[44] M. V. Dogan, I. M. Grumbach, J. J. Michaelson, and R. A. Philibert, "Integrated genetic and epigenetic prediction of coronary heart disease in the Framingham Heart Study," *PLOS ONE*, vol. 13, no. 1, p. e0190549, Jan. 2018, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0190549 2

[45] J. Duchi, Y. Singer, and T. Chandra, *Efficient Projections onto the 1-Ball for Learning in High Dimensions*. 8

[46] I. Eisenberg, N. Avidan, T. Potikha, H. Hochner, M. Chen, T. Olender, M. Barash, M. Shemesh, M. Sadeh, G. Grabov-Nardini, I. Shmilevich, A. Friedmann, G. Karpati, W. G. Bradley, L. Baumbach, D. Lancet, E. B. Asher, J. S. Beckmann, Z. Argov, and S. Mitrani-Rosenbaum, "The UDP-N-acetylglucosamine 2-epimerase/N-

acetylmannosamine kinase gene is mutated in recessive hereditary inclusion body myopathy," *Nature Genetics*, vol. 29, no. 1, pp. 83–87, Sep. 2001. 74

[47] H. M. Eltoukhi, M. N. Modi, M. Weston, A. Y. Armstrong, and E. A. Stewart, "The health disparities of uterine fibroid tumors for African American women: a public health issue," *American Journal of Obstetrics and Gynecology*, vol. 210, no. 3, pp. 194–199, Mar. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000293781300834X 69

[48] J. Fallahi, Z. Anvar, V. Razban, M. Momtahan, B. Namavar-Jahromi, and M. Fardaei, "Founder Effect of KHDC3L, p.M1V Mutation, on Iranian Patients with Recurrent Hydatidiform Moles," *Iranian Journal of Medical Sciences*, vol. 45, no. 2, pp. 118–124, Mar. 2020. 75

[49] J. D. Faul, J. K. Kim, M. E. Levine, B. Thyagarajan, D. R. Weir, and E. M. Crimmins, "Epigenetic-based age acceleration in a representative sample of older Americans: Associations with aging-related morbidity and mortality," *Proceedings of the National Academy of Sciences*, vol. 120, no. 9, p. e2215840120, Feb. 2023, publisher: Proceedings of the National Academy of Sciences. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2215840120 2

[50] J. M. Fernández, V. de la Torre, D. Richardson, R. Royo, M. Puiggròs, V. Moncunill, S. Fragkogianni, L. Clarke, P. Flicek, D. Rico, D. Torrents, E. C. de Santa Pau, and A. Valencia, "The BLUEPRINT Data Analysis Portal," *Cell systems*, vol. 3, no. 5, pp. 491–495.e5, Nov. 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5919098/ 14, 34

[51] K. Fiscella and D. R. Williams, "Health Disparities Based on Socioeconomic Inequities: Implications for Urban Health Care," *Academic Medicine*, vol. 79, no. 12, pp. 1139–1147, Dec. 2004. [Online]. Available: https://journals.lww.com/academicmedicine/Fulltext/2004/12000/Health_Disparities_Based_on_Socioeconomic.4.aspx 64

[52] J. D. Freeman, S. Kadiyala, J. F. Bell, and D. P. Martin, “The Causal Effect of Health Insurance on Utilization and Outcomes in Adults: A Systematic Review of US Studies,” *Medical Care*, vol. 46, no. 10, pp. 1023–1032, 2008, publisher: Lippincott Williams & Wilkins. [Online]. Available: https://www.jstor.org/stable/40221801 68

[53] T. M. J. Fruchterman and E. M. Reingold, “Graph drawing by force-directed placement,” *Software: Practice and Experience*, vol. 21, no. 11, pp. 1129–1164, 1991, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/spe.4380211102. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.4380211102 88

[54] A. Gaiani, I. Martinelli, L. Bello, G. Querin, M. Puthenparampil, S. Ruggero, E. Toffanin, A. Cagnin, C. Briani, E. Pegoraro, and G. Sorarù, “Diagnostic and Prognostic Biomarkers in Amyotrophic Lateral Sclerosis: Neurofilament Light Chain Levels in Definite Subtypes of Disease,” *JAMA Neurology*, vol. 74, no. 5, pp. 525–532, May 2017. [Online]. Available: https://doi.org/10.1001/jamaneurol.2016.5398 43

[55] H. García-Ortiz, F. Barajas-Olmos, C. Contreras-Cubas, M. Cid-Soto, E. J. Córdova, F. Centeno-Cruz, E. Mendoza-Caamal, I. Cicerón-Arellano, M. Flores-Huacuja, P. Baca, D. A. Bolnick, M. Snow, S. E. Flores-Martínez, R. Ortiz-Lopez, A. W. Reynolds, A. Blanchet, M. Morales-Marín, R. Velázquez-Cruz, A. D. Kostic, C. Galaviz-Hernández, A. G. García-Zapién, J. C. Jiménez-López, G. León-Reyes, E. G. Salas-Bautista, B. P. Lazalde-Ramos, J. L. Jiménez-Ruíz, G. Salas-Martínez, J. Ramos-Madrigal, E. Mirzaeicheshmeh, Y. Saldaña-Alvarez, M. del Carmen Abrahantes-Pérez, F. Loeza-Becerra, R. Mojica-Espinosa, F. Sánchez-Quinto, H. Rangel-Villalobos, M. Sosa-Macías, J. Sánchez-Corona, A. Rojas-Martinez, A. Martínez-Hernández, and L. Orozco, “The genomic landscape of Mexican Indigenous populations brings insights into the peopling of the Americas,” *Nature Communications*, vol. 12, no. 1, p. 5942, Oct. 2021, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-021-26188-w 67, 87

[56] L. D. Geneviève, A. Martani, D. Shaw, B. S. Elger, and T. Wangmo, “Structural racism in precision medicine: leaving no one behind,” *BMC Medical Ethics*, vol. 21, no. 1, p. 17, Feb. 2020. [Online]. Available: https://doi.org/10.1186/s12910-020-0457-8 64

[57] E. Gilbert, A. Shanmugam, and G. L. Cavalleri, “Revealing the recent demographic history of Europe via haplotype sharing in the UK Biobank,” *Proceedings of the National Academy of Sciences*, vol. 119, no. 25, p. e2119281119, Jun. 2022, publisher: Proceedings of the National Academy of Sciences. [Online]. Available: https://www.pnas.org/doi/10.1073/pnas.2119281119 64

[58] J. J. Grzymski, G. Elhanan, J. A. Morales Rosado, E. Smith, K. A. Schlauch, R. Read, C. Rowan, N. Slotnick, S. Dabe, W. J. Metcalf, B. Lipp, H. Reed, L. Sharma, E. Levin, J. Kao, M. Rashkin, J. Bowes, K. Dunaway, A. Slonim, N. Washington, M. Ferber, A. Bolze, and J. T. Lu, “Population genetic screening efficiently identifies carriers of autosomal dominant diseases,” *Nature Medicine*, vol. 26, no. 8, pp. 1235–1239, Aug. 2020. 74

[59] M. Gögenur, J. Burcharth, and I. Gögenur, “The role of total cell-free DNA in predicting outcomes among trauma patients in the intensive care unit: a systematic review,” *Critical Care*, vol. 21, Jan. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5260039/ 5

[60] A. Hagberg, P. Swart, and D. S Chult, “Exploring network structure, dynamics, and function using networkx,” Jan. 2008. [Online]. Available: https://www.osti.gov/biblio/960616-exploring-network-structure-dynamics-function-using-networkx 85

[61] A. Han, A. Isaacson, and P. Muennig, “The promise of big data for precision population health management in the US,” *Public Health*, vol. 185, pp. 110–116, Aug. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0033350620301451 3

[62] X. Han, J. Wang, and Y. Sun, "Circulating Tumor DNA as Biomarkers for Cancer Detection," *Genomics, Proteomics & Bioinformatics*, vol. 15, no. 2, pp. 59–72, Apr. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1672022917300487 5

[63] O. Hardiman, A. Al-Chalabi, A. Chio, E. M. Corr, G. Logroscino, W. Robberecht, P. J. Shaw, Z. Simmons, and L. H. van den Berg, "Amyotrophic lateral sclerosis," *Nature Reviews Disease Primers*, vol. 3, no. 1, pp. 1–19, Oct. 2017, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nrdp201771 43

[64] S. Hateley, A. Lopez-Izquierdo, C. J. Jou, S. Cho, J. G. Schraiber, S. Song, C. T. Maguire, N. Torres, M. Riedel, N. E. Bowles, C. B. Arrington, B. J. Kennedy, S. P. Etheridge, S. Lai, C. Pribble, L. Meyers, D. Lundahl, J. Byrnes, J. M. Granka, C. A. Kauffman, G. Lemmon, S. Boyden, W. Scott Watkins, M. A. Karren, S. Knight, J. Brent Muhlestein, J. F. Carlquist, J. L. Anderson, K. G. Chahine, K. U. Shah, C. A. Ball, I. J. Benjamin, M. Yandell, and M. Tristani-Firouzi, "The history and geographic distribution of a KCNQ1 atrial fibrillation risk allele," *Nature Communications*, vol. 12, no. 1, p. 6442, Nov. 2021, bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Cardiovascular genetics;Genetics research;Population genetics Subject_term_id: cardiovascular-genetics;genetics-research;population-genetics. [Online]. Available: https://www.nature.com/articles/s41467-021-26741-7 64

[65] B. M. Henn, L. Hon, J. M. Macpherson, N. Eriksson, S. Saxonov, I. Pe'er, and J. L. Mountain, "Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples," *PLOS ONE*, vol. 7, no. 4, p. e34267, Apr. 2012, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0034267 64

[66] L. Hood and L. Rowen, “The Human Genome Project: big science transforms biology and medicine,” *Genome Medicine*, vol. 5, no. 9, p. 79, Sep. 2013. [Online]. Available: https://doi.org/10.1186/gm483 1

[67] P. J. Hop, R. A. Zwamborn, E. Hannon, G. L. Shireby, M. F. Nabais, E. M. Walker, W. van Rheenen, J. J. van Vugt, A. M. Dekker, H.-J. Westeneng, G. H. Tazelaar, K. R. van Eijk, M. Moisse, D. Baird, A. Al Khleifat, A. Iacoangeli, N. Ticozzi, A. Ratti, J. Cooper-Knock, K. E. Morrison, P. J. Shaw, A. N. Basak, A. Chiò, A. Calvo, C. Moglia, A. Canosa, M. Brunetti, M. Grassano, M. Gotkine, Y. Lerner, M. Zabari, P. Vourc’h, P. Corcia, P. Couratier, J. S. Mora Pardina, T. Salas, P. Dion, J. P. Ross, R. D. Henderson, S. Mathers, P. A. McCombe, M. Needham, G. Nicholson, D. B. Rowe, R. Pamphlett, K. A. Mather, P. S. Sachdev, S. Furlong, F. C. Garton, A. K. Henders, T. Lin, S. T. Ngo, F. J. Steyn, L. Wallace, K. L. Williams, BIOS Consortium, Brain MEND Consortium, M. M. Neto, R. J. Cauchi, I. P. Blair, M. C. Kiernan, V. Drory, M. Povedano, M. de Carvalho, S. Pinto, M. Weber, G. A. Rouleau, V. Silani, J. E. Landers, C. E. Shaw, P. M. Andersen, A. F. McRae, M. A. van Es, R. J. Pasterkamp, N. R. Wray, R. L. McLaughlin, O. Hardiman, K. P. Kenna, E. Tsai, H. Runz, A. Al-Chalabi, L. H. van den Berg, P. Van Damme, J. Mill, and J. H. Veldink, “Genome-wide study of DNA methylation shows alterations in metabolic, inflammatory, and cholesterol pathways in ALS,” *Science Translational Medicine*, vol. 14, no. 633, p. eabj0264, Feb. 2022, publisher: American Association for the Advancement of Science. [Online]. Available: https://www.science.org/doi/abs/10.1126/scitranslmed.abj0264 44

[68] E. A. Houseman, W. P. Accomando, D. C. Koestler, B. C. Christensen, C. J. Marsit, H. H. Nelson, J. K. Wiencke, and K. T. Kelsey, “DNA methylation arrays as surrogate measures of cell mixture distribution,” *BMC Bioinformatics*, vol. 13, no. 1, p. 86, May 2012. [Online]. Available: https://doi.org/10.1186/1471-2105-13-86 5

[69] E. A. Houseman, J. Molitor, and C. J. Marsit, “Reference-free cell mixture adjustments in analysis of DNA methylation data,” *Bioinformatics*, vol. 30,

no. 10, pp. 1431–1439, May 2014. [Online]. Available: https://academic.oup.com/bioinformatics/article/30/10/1431/266465 5

[70] S. Jahr, H. Hentze, S. Englisch, D. Hardt, F. O. Fackelmayer, R.-D. Hesch, and R. Knippers, "DNA Fragments in the Blood Plasma of Cancer Patients: Quantitations and Evidence for Their Origin from Apoptotic and Necrotic Cells," *Cancer Research*, vol. 61, no. 4, pp. 1659–1665, Feb. 2001. [Online]. Available: http://cancerres.aacrjournals.org/content/61/4/1659 5

[71] T. J. Jensen, S. K. Kim, Z. Zhu, C. Chin, C. Gebhard, T. Lu, C. Deciu, D. van den Boom, and M. Ehrich, "Whole genome bisulfite sequencing of cell-free DNA and its cellular contributors uncovers placenta hypomethylated domains," *Genome Biology*, vol. 16, p. 78, Apr. 2015. [Online]. Available: https://doi.org/10.1186/s13059-015-0645-x 17

[72] P. Jiang and Y. M. D. Lo, "The Long and Short of Circulating Cell-Free DNA and the Ins and Outs of Molecular Diagnostics," *Trends in genetics: TIG*, vol. 32, no. 6, pp. 360–371, 2016. 21

[73] R. Johnson, Y. Ding, V. Venkateswaran, A. Bhattacharya, A. Chiu, T. Schwarz, M. Freund, L. Zhan, K. S. Burch, C. Caggiano, B. Hill, N. Rakocz, B. Balliu, J. H. Sul, N. Zaitlen, V. A. Arboleda, E. Halperin, S. Sankararaman, M. J. Butte, UCLA Precision Health Data Discovery Repository Working Group, UCLA Precision Health ATLAS Working Group, C. Lajonchere, D. H. Geschwind, and B. Pasaniuc, "Leveraging genomic diversity for discovery in an EHR-linked biobank: the UCLA ATLAS Community Health Initiative," Genetic and Genomic Medicine, preprint, Sep. 2021. [Online]. Available: http://medrxiv.org/lookup/doi/10.1101/2021.09.22.21263987 64, 65, 79

[74] R. Johnson, Y. Ding, A. Bhattacharya, A. Chiu, C. Lajonchere, D. H. Geschwind, and B. Pasaniuc, "The UCLA ATLAS Community Health Initiative: promoting precision health research in a diverse biobank," Genetic and Genomic Medicine, preprint,

Feb. 2022. [Online]. Available: http://medrxiv.org/lookup/doi/10.1101/2022.02.12.22270895 65

[75] J. L. Johnston, J. C. Fanzo, and B. Cogill, "Understanding Sustainable Diets: A Descriptive Analysis of the Determinants and Processes That Influence Diets and Their Impact on Health, Food Security, and Environmental Sustainability," *Advances in Nutrition*, vol. 5, no. 4, pp. 418–429, Jul. 2014. [Online]. Available: https://doi.org/10.3945/an.113.005553 2

[76] D. Joka, K. Wahl, S. Moeller, J. Schlue, B. Vaske, M. J. Bahr, M. P. Manns, K. Schulze-Osthoff, and H. Bantel, "Prospective biopsy-controlled evaluation of cell death biomarkers for prediction of liver fibrosis and nonalcoholic steatohepatitis," *Hepatology*, vol. 55, no. 2, pp. 455–464, 2012. [Online]. Available: https://aasldpubs.onlinelibrary.wiley.com/doi/abs/10.1002/hep.24734 5

[77] S. Kang, Q. Li, Q. Chen, Y. Zhou, S. Park, G. Lee, B. Grimes, K. Krysan, M. Yu, W. Wang, F. Alber, F. Sun, S. M. Dubinett, W. Li, and X. J. Zhou, "CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA," *Genome Biology*, vol. 18, no. 1, p. 53, Mar. 2017. [Online]. Available: https://doi.org/10.1186/s13059-017-1191-5 6

[78] K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E.

Talkowski, B. M. Neale, M. J. Daly, and D. G. MacArthur, "The mutational constraint spectrum quantified from variation in 141,456 humans," *Nature*, vol. 581, no. 7809, pp. 434–443, May 2020, number: 7809 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41586-020-2308-7 74

[79] H. H. Kazazian, C. E. Dowling, P. G. Waber, S. Huang, and W. H. Lo, "The spectrum of beta-thalassemia genes in China and Southeast Asia," *Blood*, vol. 68, no. 4, pp. 964–966, Oct. 1986. 74

[80] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and a. D. Haussler, "The Human Genome Browser at UCSC," *Genome Research*, vol. 12, no. 6, pp. 996–1006, Jun. 2002. [Online]. Available: http://genome.cshlp.org/content/12/6/996 34

[81] M. J. Khoury and S. Galea, "Will Precision Medicine Improve Population Health?" *JAMA*, vol. 316, no. 13, pp. 1357–1358, Oct. 2016. [Online]. Available: https://doi.org/10.1001/jama.2016.12260 3

[82] M. C. Kiernan, S. Vucic, K. Talbot, C. J. McDermott, O. Hardiman, J. M. Shefner, A. Al-Chalabi, W. Huynh, M. Cudkowicz, P. Talman, L. H. Van den Berg, T. Dharmadasa, P. Wicks, C. Reilly, and M. R. Turner, "Improving clinical trial outcomes in amyotrophic lateral sclerosis," *Nature Reviews Neurology*, vol. 17, no. 2, pp. 104–118, Feb. 2021, number: 2 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41582-020-00434-z 43

[83] A. D. Koker, R. V. Paemel, B. D. Wilde, K. D. Preter, and N. Callewaert, "A versatile method for circulating cell-free DNA methylome profiling by reduced representation bisulfite sequencing," *bioRxiv*, p. 663195, Jun. 2019. [Online]. Available: https://www.biorxiv.org/content/10.1101/663195v2 22

[84] S. Komaki, H. Ohmomo, T. Hachiya, Y. Sutoh, K. Ono, R. Furukawa, S. Umekage, Y. Otsuka-Yamasaki, K. Tanno, M. Sasaki, and A. Shimizu, "Longitudinal

DNA methylation dynamics as a practical indicator in clinical epigenetics," *Clinical Epigenetics*, vol. 13, no. 1, p. 219, Dec. 2021. [Online]. Available: https://doi.org/10.1186/s13148-021-01202-6 2

[85] R. Kosoy, R. Nassir, C. Tian, P. A. White, L. M. Butler, G. Silva, R. Kittles, M. E. Alarcon-Riquelme, P. K. Gregersen, J. W. Belmont, F. M. De La Vega, and M. F. Seldin, "Ancestry Informative Marker Sets for Determining Continental Origin and Admixture Proportions in Common Populations in America," *Human mutation*, vol. 30, no. 1, pp. 69–78, Jan. 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3073397/ 35

[86] A. Kustanovich, R. Schwartz, T. Peretz, and A. Grinshpun, "Life and death of circulating cell-free DNA," *Cancer Biology & Therapy*, vol. 20, no. 8, pp. 1057–1067, Aug. 2019. [Online]. Available: https://doi.org/10.1080/15384047.2019.1598759 5

[87] C. Lajonchere, A. Naeim, S. Dry, N. Wenger, D. Elashoff, S. Vangala, A. Petruse, M. Ariannejad, C. Magyar, L. Johansen, G. Werre, M. Kroloff, and D. Geschwind, "An Integrated, Scalable, Electronic Video Consent Process to Power Precision Health Research: Large, Population-Based, Cohort Implementation and Scalability Study," *Journal of Medical Internet Research*, vol. 23, no. 12, p. e31121, Dec. 2021, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. [Online]. Available: https://www.jmir.org/2021/12/e31121 79

[88] D. Laurent, F. Semple, P. J. S. Lewis, E. Rose, H. A. Black, S. J. Forbes, M. J. Arends, J. W. Dear, and T. J. Aitman, "Absolute measurement of the tissue origins of cell-free DNA in the healthy state and following paracetamol overdose," *bioRxiv*, p. 715888, Jul. 2019. [Online]. Available: https://www.biorxiv.org/content/10.1101/715888v1 5

[89] R. Lehmann-Werman, D. Neiman, H. Zemmour, J. Moss, J. Magenheim, A. Vaknin-Dembinsky, S. Rubertsson, B. Nellgård, K. Blennow, H. Zetterberg,

K. Spalding, M. J. Haller, C. H. Wasserfall, D. A. Schatz, C. J. Greenbaum, C. Dorrell, M. Grompe, A. Zick, A. Hubert, M. Maoz, V. Fendrich, D. K. Bartsch, T. Golan, S. A. B. Sasson, G. Zamir, A. Razin, H. Cedar, A. M. J. Shapiro, B. Glaser, R. Shemer, and Y. Dor, "Identification of tissue-specific cell death using methylation patterns of circulating DNA," *Proceedings of the National Academy of Sciences*, vol. 113, no. 13, pp. E1826–E1834, Mar. 2016. [Online]. Available: http://www.pnas.org/content/113/13/E1826 5, 6, 11, 21, 44

[90] T. Lencz, J. Yu, R. R. Khan, E. Flaherty, S. Carmi, M. Lam, D. Ben-Avraham, N. Barzilai, S. Bressman, A. Darvasi, J. H. Cho, L. N. Clark, Z. H. Gümüş, J. Vijai, R. J. Klein, S. Lipkin, K. Offit, H. Ostrer, L. J. Ozelius, I. Peter, A. K. Malhotra, T. Maniatis, G. Atzmon, and I. Pe'er, "Novel ultra-rare exonic variants identified in a founder population implicate cadherins in schizophrenia," *Neuron*, vol. 109, no. 9, pp. 1465–1478.e4, May 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0896627321001525 75

[91] I. Levav, R. Kohn, J. M. Golding, and M. M. Weissman, "Vulnerability of Jews to affective disorders," *The American Journal of Psychiatry*, vol. 154, no. 7, pp. 941–947, Jul. 1997. 69

[92] A. C. F. Lewis, S. J. Molina, P. S. Appelbaum, B. Dauda, A. Di Rienzo, A. Fuentes, S. M. Fullerton, N. A. Garrison, N. Ghosh, E. M. Hammonds, D. S. Jones, E. E. Kenny, P. Kraft, S. S.-J. Lee, M. Mauro, J. Novembre, A. Panofsky, M. Sohail, B. M. Neale, and D. S. Allen, "Getting genetic ancestry right for science and society," *Science*, vol. 376, no. 6590, pp. 250–252, Apr. 2022, publisher: American Association for the Advancement of Science. [Online]. Available: https://www.science.org/doi/10.1126/science.abm7530 3, 66, 78

[93] B. T. Li, A. Drilon, M. L. Johnson, M. Hsu, C. S. Sima, C. McGinn, H. Sugita, M. G. Kris, and C. G. Azzoli, "A prospective study of total plasma cell-free DNA as a predictive biomarker for response to systemic therapy in patients with advanced non-

small-cell lung cancers," *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, vol. 27, no. 1, pp. 154–159, Jan. 2016. 21

[94] W. Li, Q. Li, S. Kang, M. Same, Y. Zhou, C. Sun, C.-C. Liu, L. Matsuoka, L. Sher, W. H. Wong, F. Alber, and X. Zhou, "CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data," *Nucleic Acids Research*, vol. 46, no. 15, p. e89, Sep. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6125664/ 6

[95] A. J. Lisoway, C. C. Zai, A. K. Tiwari, and J. L. Kennedy, "DNA methylation and clinical response to antidepressant medication in major depressive disorder: A review and recommendations," *Neuroscience Letters*, vol. 669, pp. 14–23, Mar. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304394016310199 2

[96] M. C. Liu, G. R. Oxnard, E. A. Klein, C. Swanton, M. V. Seiden, M. C. Liu, G. R. Oxnard, E. A. Klein, D. Smith, D. Richards, T. J. Yeatman, A. L. Cohn, R. Lapham, J. Clement, A. S. Parker, M. K. Tummala, K. McIntyre, M. A. Sekeres, A. H. Bryce, R. Siegel, X. Wang, D. P. Cosgrove, N. R. Abu-Rustum, J. Trent, D. D. Thiel, C. Becerra, M. Agrawal, L. E. Garbo, J. K. Giguere, R. M. Michels, R. P. Harris, S. L. Richey, T. A. McCarthy, D. M. Waterhouse, F. J. Couch, S. T. Wilks, A. K. Krie, R. Balaraman, A. Restrepo, M. W. Meshad, K. Rieger-Christ, T. Sullivan, C. M. Lee, D. R. Greenwald, W. Oh, C.-K. Tsao, N. Fleshner, H. F. Kennecke, M. F. Khalil, D. R. Spigel, A. P. Manhas, B. K. Ulrich, P. A. Kovoor, C. Stokoe, J. G. Courtright, H. A. Yimer, T. G. Larson, C. Swanton, M. V. Seiden, S. R. Cummings, F. Absalan, G. Alexander, B. Allen, H. Amini, A. M. Aravanis, S. Bagaria, L. Bazargan, J. F. Beausang, J. Berman, C. Betts, A. Blocker, J. Bredno, R. Calef, G. Cann, J. Carter, C. Chang, H. Chawla, X. Chen, T. C. Chien, D. Civello, K. Davydov, V. Demas, M. Desai, Z. Dong, S. Fayzullina, A. P. Fields, D. Filippova, P. Freese, E. T. Fung, S. Gnerre, S. Gross, M. Halks-Miller, M. P. Hall, A.-R. Hartman, C. Hou, E. Hubbell, N. Hunkapiller, K. Jagadeesh,

A. Jamshidi, R. Jiang, B. Jung, T. Kim, R. D. Klausner, K. N. Kurtzman, M. Lee, W. Lin, J. Lipson, H. Liu, Q. Liu, M. Lopatin, T. Maddala, M. C. Maher, C. Melton, A. Mich, S. Nautiyal, J. Newman, J. Newman, V. Nicula, C. Nicolaou, O. Nikolic, W. Pan, S. Patel, S. A. Prins, R. Rava, N. Ronaghi, O. Sakarya, R. V. Satya, J. Schellenberger, E. Scott, A. J. Sehnert, R. Shaknovich, A. Shanmugam, K. C. Shashidhar, L. Shen, A. Shenoy, S. Shojaee, P. Singh, K. K. Steffen, S. Tang, J. M. Toung, A. Valouev, O. Venn, R. T. Williams, T. Wu, H. H. Xu, C. Yakym, X. Yang, J. Yecies, A. S. Yip, J. Youngren, J. Yue, J. Zhang, L. Zhang, L. Q. Zhang, N. Zhang, C. Curtis, and D. A. Berry, "Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA," *Annals of Oncology*, vol. 31, no. 6, pp. 745–759, Jun. 2020, publisher: Elsevier. [Online]. Available: https://www.annalsofoncology.org/article/S0923-7534(20)36058-0/abstract 21

[97] X. Liu, J. Ren, N. Luo, H. Guo, Y. Zheng, J. Li, F. Tang, L. Wen, and J. Peng, "Comprehensive DNA methylation analysis of tissue of origin of plasma cell-free DNA by methylated CpG tandem amplification and sequencing (MCTA-Seq)," *Clinical Epigenetics*, vol. 11, no. 1, p. 93, Jun. 2019. [Online]. Available: https://doi.org/10.1186/s13148-019-0689-y 5, 6, 21

[98] P.-R. Loh, P. Danecek, P. F. Palamara, C. Fuchsberger, Y. A Reshef, H. K Finucane, S. Schoenherr, L. Forer, S. McCarthy, G. R. Abecasis, R. Durbin, and A. L Price, "Reference-based phasing using the Haplotype Reference Consortium panel," *Nature Genetics*, vol. 48, no. 11, pp. 1443–1448, Nov. 2016, number: 11 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/ng.3679 84

[99] K. Lokk, V. Modhukur, B. Rajashekar, K. Märtens, R. Mägi, R. Kolde, M. Koltšina, T. K. Nilsson, J. Vilo, A. Salumets, and N. Tõnisson, "DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns," *Genome Biology*, vol. 15, no. 4, p. r54, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4053947/ 5

[100] A. T. Lorincz, "The Promise and the Problems of Epigenetics Biomarkers in Cancer," *Expert opinion on medical diagnostics*, vol. 5, no. 5, pp. 375–379, Sep. 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3191528/ 2

[101] N. Loyfer, "Comprehensive Human Cell-Type Methylation Atlas Reveals Origins of Circulating Cell-Free Dna in Health and Disease," Sep. 2019, original-date: 2018-08-23T14:29:09Z. [Online]. Available: https://github.com/nloyfer/meth_atlas 9, 44

[102] C.-H. Lu, C. Macdonald-Wallis, E. Gray, N. Pearce, A. Petzold, N. Norgren, G. Giovannoni, P. Fratta, K. Sidle, M. Fish, R. Orrell, R. Howard, K. Talbot, L. Greensmith, J. Kuhle, M. R. Turner, and A. Malaspina, "Neurofilament light chain: A prognostic biomarker in amyotrophic lateral sclerosis," *Neurology*, vol. 84, no. 22, pp. 2247–2257, Jun. 2015, publisher: Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology Section: Article. [Online]. Available: https://n.neurology.org/content/84/22/2247 43

[103] C. Lövkvist, I. B. Dodd, K. Sneppen, and J. O. Haerter, "DNA methylation in human epigenomes depends on local topology of CpG sites," *Nucleic Acids Research*, vol. 44, no. 11, pp. 5123–5132, Jun. 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4914085/ 15

[104] I. Magen, N. S. Yacovzada, E. Yanowski, A. Coenen-Stass, J. Grosskreutz, C.-H. Lu, L. Greensmith, A. Malaspina, P. Fratta, and E. Hornstein, "Circulating miR-181 is a prognostic biomarker for amyotrophic lateral sclerosis," *Nature Neuroscience*, vol. 24, no. 11, pp. 1534–1541, Nov. 2021, number: 11 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41593-021-00936-z 43

[105] L. Majara, A. Kalungi, N. Koen, H. Zar, D. J. Stein, E. Kinyanda, E. G. Atkinson, and A. R. Martin, "Low generalizability of polygenic scores in African populations due to genetic and environmental diversity," Genetics, preprint, Jan. 2021. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2021.01.12.426453 64

[106] S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. Di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, and D. Reich, "The Simons Genome Diversity Project: 300 genomes from 142 diverse populations," *Nature*, vol. 538, no. 7624, pp. 201–206, Oct. 2016, bandiera_abtest: a Cg_type: Nature Research Journals Number: 7624 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genetic variation;Genetics Subject_term_id: genetic-variation;genetics. [Online]. Available: https://www.nature.com/articles/nature18964 66

[107] A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W.-M. Chen, "Robust relationship inference in genome-wide association studies," *Bioinformatics*, vol. 26, no. 22, pp. 2867–2873, Nov. 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3025716/ 84

[108] A. K. Manrai, B. H. Funke, H. L. Rehm, M. S. Olesen, B. A. Maron, P. Szolovits, D. M. Margulies, J. Loscalzo, and I. S. Kohane, "Genetic Misdiagnoses and the Potential for Health Disparities," *New England Journal of Medicine*, vol. 375, no. 7, pp. 655–665, Aug. 2016. [Online]. Available: http://www.nejm.org/doi/10.1056/NEJMsa1507092 77

[109] A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, and M. J.

Daly, “Clinical use of current polygenic risk scores may exacerbate health disparities,” *Nature Genetics*, vol. 51, no. 4, p. 584, Apr. 2019. [Online]. Available: https://www.nature.com/articles/s41588-019-0379-x 64

[110] D. O. Martschenko and J. L. Young, “Precision Medicine Needs to Think Outside the Box,” *Frontiers in Genetics*, vol. 13, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fgene.2022.795992 78

[111] J. M. Matamala, R. Arias-Carrasco, C. Sanchez, M. Uhrig, L. Bargsted, S. Matus, V. Maracaja-Coutinho, S. Abarzua, B. van Zundert, R. Verdugo, P. Manque, and C. Hetz, “Genome-wide circulating microRNA expression profiling reveals potential biomarkers for amyotrophic lateral sclerosis,” *Neurobiology of Aging*, vol. 64, pp. 123–138, Apr. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0197458017304177 43

[112] M. Mauro, D. S. Allen, B. Dauda, S. J. Molina, B. M. Neale, and A. C. F. Lewis, “A scoping review of guidelines for the use of race, ethnicity, and ancestry reveals widespread consensus but also points of ongoing disagreement,” *The American Journal of Human Genetics*, vol. 109, no. 12, pp. 2110–2125, Dec. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000292972200492X 78

[113] P. McCombe and R. Henderson, “The Role of Immune and Inflammatory Mechanisms in ALS,” *Current Molecular Medicine*, vol. 11, no. 3, pp. 246–254, Apr. 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3182412/ 44

[114] C. McCrory, G. Fiorito, B. Hernandez, S. Polidoro, A. M. O’Halloran, A. Hever, C. Ni Cheallaigh, A. T. Lu, S. Horvath, P. Vineis, and R. A. Kenny, “GrimAge Outperforms Other Epigenetic Clocks in the Prediction of Age-Related Clinical Phenotypes and All-Cause Mortality,” *The Journals of Gerontology: Series A*, vol. 76, no. 5, pp. 741–749, May 2021. [Online]. Available: https://doi.org/10.1093/gerona/glaa286 2

[115] C. R. McGrath, D. C. Hitchcock, and O. W. van Assendelft, “Total white blood cell counts for persons ages 1-74 years with differential leukocyte counts for adults ages 25-74 years: United States, 1971-75,” *Vital and Health Statistics. Series 11, Data from the National Health Survey*, no. 220, pp. 1–36, Jan. 1982. 19

[116] P. Meier, A. Finch, and G. Evan, “Apoptosis in development,” *Nature*, vol. 407, no. 6805, pp. 796–801, Oct. 2000. [Online]. Available: https://www.nature.com/articles/35037734 5

[117] M. M. Moradian, T. Sarkisian, H. Ajrapetyan, and N. Avanesian, “Genotype–phenotype studies in a large cohort of Armenian patients with familial Mediterranean fever suggest clinical disease with heterozygous MEFV mutations,” *Journal of Human Genetics*, vol. 55, no. 6, pp. 389–393, Jun. 2010, number: 6 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/jhg201052 73

[118] S. Moreno-Grau, M. V. Fernández, I. de Rojas, P. Garcia-González, I. Hernández, F. Farias, J. P. Budde, I. Quintela, L. Madrid, A. González-Pérez, L. Montrreal, E. Alarcón-Martín, M. Alegret, O. Maroñas, J. A. Pineda, J. Macías, M. Marquié, S. Valero, A. Benaque, J. Clarimón, M. J. Bullido, G. García-Ribas, P. Pástor, P. Sánchez-Juan, V. Álvarez, G. Piñol-Ripoll, J. M. García-Alberca, J. L. Royo, E. Franco-Macías, P. Mir, M. Calero, M. Medina, A. Rábano, J. Ávila, C. Antúnez, L. M. Real, A. Orellana, Carracedo, M. E. Sáez, L. Tárraga, M. Boada, C. Cruchaga, and A. Ruiz, “Long runs of homozygosity are associated with Alzheimer’s disease,” *Translational Psychiatry*, vol. 11, no. 1, pp. 1–12, Feb. 2021, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41398-020-01145-1 75

[119] J. Moss, A. Zick, A. Grinshpun, E. Carmon, M. Maoz, B. L. Ochana, O. Abraham, O. Arieli, L. Germansky, K. Meir, B. Glaser, R. Shemer, B. Uziely, and Y. Dor, “Circulating breast-derived DNA allows universal detection and monitoring of localized

breast cancer," *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, vol. 31, no. 3, pp. 395–403, Mar. 2020. 44

[120] J. Moss, J. Magenheim, D. Neiman, H. Zemmour, N. Loyfer, A. Korach, Y. Samet, M. Maoz, H. Druid, P. Arner, K.-Y. Fu, E. Kiss, K. L. Spalding, G. Landesberg, A. Zick, A. Grinshpun, A. M. J. Shapiro, M. Grompe, A. D. Wittenberg, B. Glaser, R. Shemer, T. Kaplan, and Y. Dor, "Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease," *Nature Communications*, vol. 9, no. 1, pp. 1–12, Nov. 2018. [Online]. Available: https://www.nature.com/articles/s41467-018-07466-6 5, 8, 17, 21

[121] Mozilla, "SPA (Single-page application) - MDN Web Docs Glossary: Definitions of Web-related terms | MDN." [Online]. Available: https://developer.mozilla.org/en-US/docs/Glossary/SPA 90

[122] T. J. Musci, G. Fairbrother, A. Batey, J. Bruursema, C. Struble, and K. Song, "Non-invasive prenatal testing with cell-free DNA: US physician attitudes toward implementation in clinical practice," *Prenatal Diagnosis*, vol. 33, no. 5, pp. 424–428, 2013. [Online]. Available: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1002/pd.4091 5

[123] Z. Naccashian, M. Hattar-Pollara, C. A. Ho, and S. P. Ayvazian, "Prevalence and Predictors of Diabetes Mellitus and Hypertension in Armenian Americans in Los Angeles," *The Diabetes Educator*, vol. 44, no. 2, pp. 130–143, Apr. 2018, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.1177/0145721718759981 68

[124] S. Nagata, "Apoptosis by Death Factor," *Cell*, vol. 88, no. 3, pp. 355–365, Feb. 1997. [Online]. Available: https://www.cell.com/cell/abstract/S0092-8674(00)81874-7 5

[125] A. Nuriddin, G. Mooney, and A. I. R. White, "Reckoning with histories of medical racism and violence in the USA," *The Lancet*, vol. 396, no.

10256, pp. 949–951, Oct. 2020, publisher: Elsevier. [Online]. Available: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)32032-8/fulltext 78

[126] A. of Us Research Program Investigators, "The "All of Us" Research Program," *New England Journal of Medicine*, vol. 381, no. 7, pp. 668–676, Aug. 2019, publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMsr1809937. [Online]. Available: https://doi.org/10.1056/NEJMsr1809937 64

[127] M. Ogawa, "Differentiation and proliferation of hematopoietic stem cells," *Blood*, vol. 81, no. 11, pp. 2844–2853, Jun. 1993. 17

[128] L. Ongaro, M. O. Scliar, R. Flores, A. Raveane, D. Marnetto, S. Sarno, G. A. Gnecchi-Ruscone, M. E. Alarcón-Riquelme, E. Patin, P. Wangkumhang, G. Hellenthal, M. Gonzalez-Santos, R. J. King, A. Kouvatsi, O. Balanovsky, E. Balanovska, L. Atramentova, S. Turdikulova, S. Mastana, D. Marjanovic, L. Mulahasanovic, A. Leskovac, M. F. Lima-Costa, A. C. Pereira, M. L. Barreto, B. L. Horta, N. Mabunda, C. A. May, A. Moreno-Estrada, A. Achilli, A. Olivieri, O. Semino, K. Tambets, T. Kivisild, D. Luiselli, A. Torroni, C. Capelli, E. Tarazona-Santos, M. Metspalu, L. Pagani, and F. Montinaro, "The Genomic Impact of European Colonization of the Americas," *Current Biology*, vol. 29, no. 23, pp. 3974–3986.e4, Dec. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982219313065 87

[129] A. Panio, C. Cava, S. D'Antona, G. Bertoli, and D. Porro, "Diagnostic Circulating miRNAs in Sporadic Amyotrophic Lateral Sclerosis," *Frontiers in Medicine*, vol. 9, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmed.2022.861960 43

[130] A. Panofsky and C. Bliss, "Ambiguity and Scientific Authority: Population Classification in Genomic Science," *American Sociological Review*, vol. 82, no. 1, pp. 59–87, Feb. 2017, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.1177/0003122416685812 78

[131] S. Parvini and E. Simani, “Are Arabs and Iranians white? Census says yes, but many disagree.” [Online]. Available: https://www.latimes.com/projects/la-me-census-middle-east-north-africa-race/ 67

[132] E. I. Pentsova, R. H. Shah, J. Tang, A. Boire, D. You, S. Briggs, A. Omuro, X. Lin, M. Fleisher, C. Grommes, K. S. Panageas, F. Meng, S. D. Selcuklu, S. Ogilvie, N. Distefano, L. Shagabayeva, M. Rosenblum, L. M. DeAngelis, A. Viale, I. K. Mellinghoff, and M. F. Berger, “Evaluating Cancer of the Central Nervous System Through Next-Generation Sequencing of Cerebrospinal Fluid,” *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, vol. 34, no. 20, pp. 2404–2415, 2016. 21

[133] D. Petrov, C. Mansfield, A. Moussy, and O. Hermine, “ALS Clinical Trials Review: 20 Years of Failure. Are We Any Closer to Registering a New Treatment?” *Frontiers in Aging Neuroscience*, vol. 9, p. 68, Mar. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5360725/ 43

[134] R. Pidsley, E. Zotenko, T. J. Peters, M. G. Lawrence, G. P. Risbridger, P. Molloy, S. Van Djik, B. Muhlhausler, C. Stirzaker, and S. J. Clark, “Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling,” *Genome Biology*, vol. 17, no. 1, p. 208, Oct. 2016. [Online]. Available: https://doi.org/10.1186/s13059-016-1066-1 6

[135] L. Pinhas, M. Heinmaa, P. Bryden, S. Bradley, and B. Toner, “Disordered Eating in Jewish Adolescent Girls,” *The Canadian Journal of Psychiatry*, vol. 53, no. 9, pp. 601–608, Sep. 2008, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.1177/070674370805300907 69

[136] O. Pogoryelova, J. A. González Coraspe, N. Nikolenko, H. Lochmüller, and A. Roos, “GNE myopathy: from clinics and genetics to pathology and research strategies,” *Orphanet Journal of Rare Diseases*, vol. 13, p. 70, May 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5930817/ 74

[137] J. K. Pritchard, M. Stephens, and P. Donnelly, “Inference of Population Structure Using Multilocus Genotype Data,” *Genetics*, vol. 155, no. 2, pp. 945–959, Jun. 2000. [Online]. Available: https://www.genetics.org/content/155/2/945 30

[138] E. Rahmani, N. Zaitlen, Y. Baran, C. Eng, D. Hu, J. Galanter, S. Oh, E. G. Burchard, E. Eskin, J. Zou, and E. Halperin, “Sparse PCA Corrects for Cell-Type Heterogeneity in Epigenome-Wide Association Studies,” *Nature methods*, vol. 13, no. 5, pp. 443–445, May 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5548182/ 5

[139] E. Rahmani, R. Schweiger, L. Shenhav, E. Eskin, and E. Halperin, “A Bayesian Framework for Estimating Cell Type Composition from DNA Methylation Without the Need for Methylation Reference,” in *Research in Computational Molecular Biology*, ser. Lecture Notes in Computer Science, S. C. Sahinalp, Ed. Cham: Springer International Publishing, 2017, pp. 207–223. 5

[140] C. F. Rider and C. Carlsten, “Air pollution and DNA methylation: effects of exposure in humans,” *Clinical Epigenetics*, vol. 11, no. 1, p. 131, Sep. 2019. [Online]. Available: https://doi.org/10.1186/s13148-019-0713-2 2

[141] M. P. Roth, G. M. Petersen, C. McElree, E. Feldman, and J. I. Rotter, “Geographic origins of Jewish patients with inflammatory bowel disease,” *Gastroenterology*, vol. 97, no. 4, pp. 900–904, Oct. 1989. 69

[142] J. N. Saada, G. Kalantzis, D. Shyr, F. Cooper, M. Robinson, A. Gusev, and P. F. Palamara, “Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations,” *Nature Communications*, vol. 11, no. 1, pp. 1–15, Nov. 2020, cc_license_type: cc_by Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genome-wide association studies;Haplotypes;Heritable quantitative trait;Population genetics Subject_term_id: genome-wide-association-studies;haplotypes;heritable-quantitative-

trait;population-genetics. [Online]. Available: https://www.nature.com/articles/s41467-020-19588-x 64

[143] E. R. Schiff, M. Frampton, F. Semplici, S. L. Bloom, S. A. McCartney, R. Vega, L. B. Lovat, E. Wood, A. L. Hart, D. Crespi, M. A. Furman, S. Mann, C. D. Murray, A. W. Segal, and A. P. Levine, "A New Look at Familial Risk of Inflammatory Bowel Disease in the Ashkenazi Jewish Population," *Digestive Diseases and Sciences*, vol. 63, no. 11, pp. 3049–3057, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6182437/ 69

[144] H. Schwarzenbach, D. S. B. Hoon, and K. Pantel, "Cell-free nucleic acids as biomarkers in cancer patients," *Nature Reviews. Cancer*, vol. 11, no. 6, pp. 426–437, Jun. 2011. 21

[145] S. Seabold and J. Perktold, "Statsmodels: Econometric and Statistical Modeling with Python," Austin, Texas, 2010, pp. 92–96. [Online]. Available: https://conference.scipy.org/proceedings/scipy2010/seabold.html 89

[146] M. F. Seldin and A. L. Price, "Application of Ancestry Informative Markers to Association Studies in European Americans," *PLOS Genetics*, vol. 4, no. 1, p. e5, Jan. 2008. [Online]. Available: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0040005 35

[147] T. H. Self, C. R. Chrisman, D. L. Mason, and M. J. Rumbak, "Reducing emergency department visits and hospitalizations in African American and Hispanic patients with asthma: a 15-year review," *The Journal of Asthma: Official Journal of the Association for the Care of Asthma*, vol. 42, no. 10, pp. 807–812, Dec. 2005. 72

[148] A. L. Severson, S. Carmi, and N. A. Rosenberg, "The Effect of Consanguinity on Between-Individual Identity-by-Descent Sharing," *Genetics*, vol. 212, no. 1, pp. 305–316, May 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6499533/ 2

[149] J. M. Shefner, R. Bedlack, J. A. Andrews, J. D. Berry, R. Bowser, R. Brown, J. D. Glass, N. J. Maragakis, T. M. Miller, J. D. Rothstein, and M. E. Cudkowicz, "Amyotrophic Lateral Sclerosis Clinical Trials and Interpretation of Functional End Points and Fluid Biomarkers: A Review," *JAMA Neurology*, vol. 79, no. 12, pp. 1312–1318, Dec. 2022. [Online]. Available: https://doi.org/10.1001/jamaneurol.2022.3282 43

[150] R. Shemirani, G. M. Belbin, C. L. Avery, E. E. Kenny, C. R. Gignoux, and J. L. Ambite, "Rapid detection of identity-by-descent tracts for mega-scale datasets," *Nature Communications*, vol. 12, no. 1, p. 3546, Jun. 2021, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-021-22910-w 66, 84

[151] S. Y. Shen, R. Singhania, G. Fehringer, A. Chakravarthy, M. H. A. Roehrl, D. Chadwick, P. C. Zuzarte, A. Borgida, T. T. Wang, T. Li, O. Kis, Z. Zhao, A. Spreafico, T. d. S. Medina, Y. Wang, D. Roulois, I. Ettayebi, Z. Chen, S. Chow, T. Murphy, A. Arruda, G. M. O'Kane, J. Liu, M. Mansour, J. D. McPherson, C. O'Brien, N. Leighl, P. L. Bedard, N. Fleshner, G. Liu, M. D. Minden, S. Gallinger, A. Goldenberg, T. J. Pugh, M. M. Hoffman, S. V. Bratman, R. J. Hung, and D. D. D. Carvalho, "Sensitive tumour detection and classification using plasma cell-free DNA methylomes," *Nature*, vol. 563, no. 7732, pp. 579–583, Nov. 2018. [Online]. Available: https://www.nature.com/articles/s41586-018-0703-0 6

[152] L. Shenhav, M. Thompson, T. A. Joseph, L. Briscoe, O. Furman, D. Bogumil, I. Mizrahi, I. Pe'er, and E. Halperin, "FEAST: fast expectation-maximization for microbial source tracking," *Nature Methods*, vol. 16, no. 7, pp. 627–632, Jul. 2019. [Online]. Available: http://www.nature.com/articles/s41592-019-0431-x 22

[153] C. Sheridan, "Investors keep the faith in cancer liquid biopsies," *Nature Biotechnology*, vol. 37, pp. 972–974, Aug. 2019. [Online]. Available: http://www.nature.com/articles/d41587-019-00022-7 5

[154] W. J. Shim, E. Sinniah, J. Xu, B. Vitrinel, M. Alexanian, G. Andreoletti, S. Shen, B. Balderson, G. Peng, N. Jing, Y. Sun, Y. Wang, P. P. L. Tam, A. Smith, M. Piper, L. Christiaen, Q. Nguyen, M. Bodén, and N. J. Palpant, "Conserved epigenetic regulatory logic infers genes governing cell identity," *bioRxiv*, p. 635516, Mar. 2020, publisher: Cold Spring Harbor Laboratory Section: New Results. [Online]. Available: https://www.biorxiv.org/content/10.1101/635516v5 22

[155] M. Slatkin, "A Population-Genetic Test of Founder Effects and Implications for Ashkenazi Jewish Diseases," *The American Journal of Human Genetics*, vol. 75, no. 2, pp. 282–293, Aug. 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0002929707624100 86

[156] M. W. Snyder, M. Kircher, A. J. Hill, R. M. Daza, and J. Shendure, "Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin," *Cell*, vol. 164, no. 1-2, pp. 57–68, Jan. 2016. 5, 21

[157] E. Sohar, M. Prass, J. Heller, and H. Heller, "Genetics of Familial Mediterranean Fever (FMF): A Disorder with Recessive Inheritance in Non-Ashkenazi Jews and Armenians," *Archives of Internal Medicine*, vol. 107, no. 4, pp. 529–538, Apr. 1961. [Online]. Available: https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/565719 73

[158] N. Solovieff, S. W. Hartley, C. T. Baldwin, E. S. Klings, M. T. Gladwin, J. G. Taylor, G. J. Kato, L. A. Farrer, M. H. Steinberg, and P. Sebastiani, "Ancestry of African Americans with Sickle Cell Disease," *Blood cells, molecules & diseases*, vol. 47, no. 1, pp. 41–45, Jun. 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3116635/ 69, 74

[159] K. A. Staats, D. R. Borchelt, M. G. Tansey, and J. Wymer, "Blood-based biomarkers of inflammation in amyotrophic lateral sclerosis," *Molecular Neurodegeneration*, vol. 17, no. 1, p. 11, Jan. 2022. [Online]. Available: https://doi.org/10.1186/s13024-022-00515-1 43

[160] M. L. Stackpole, W. Zeng, S. Li, C.-C. Liu, Y. Zhou, S. He, A. Yeh, Z. Wang, F. Sun, Q. Li, Z. Yuan, A. Yildirim, P.-J. Chen, P. Winograd, B. Tran, Y.-T. Lee, P. S. Li, Z. Noor, M. Yokomizo, P. Ahuja, Y. Zhu, H.-R. Tseng, J. S. Tomlinson, E. Garon, S. French, C. E. Magyar, S. Dry, C. Lajonchere, D. Geschwind, G. Choi, S. Saab, F. Alber, W. H. Wong, S. M. Dubinett, D. R. Aberle, V. Agopian, S.-H. B. Han, X. Ni, W. Li, and X. J. Zhou, "Cost-effective methylome sequencing of cell-free DNA for accurately detecting and locating cancer," *Nature Communications*, vol. 13, no. 1, p. 5566, Sep. 2022, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-022-32995-6 44

[161] M. Stroun, P. Maurice, V. Vasioukhin, J. Lyautey, C. Lederrey, F. Lefort, A. Rossier, X. Qi Chen, and P. Anker, "The Origin and Mechanism of Circulating DNA," *Annals of the New York Academy of Sciences*, vol. 906, pp. 161–8, May 2000. 5, 44

[162] W.-M. Su, Y.-F. Cheng, Z. Jiang, Q.-Q. Duan, T.-M. Yang, H.-F. Shang, and Y.-P. Chen, "Predictors of survival in patients with amyotrophic lateral sclerosis: A large meta-analysis," *eBioMedicine*, vol. 74, Dec. 2021, publisher: Elsevier. [Online]. Available: https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(21)00526-0/fulltext 43

[163] S. A. Suckiel, J. A. Odgis, K. M. Gallagher, J. E. Rodriguez, D. Watnick, G. Bertier, M. Sebastin, N. Yelton, E. Maria, J. Lopez, M. Ramos, N. Kelly, N. Teitelman, F. Beren, T. Kaszemacher, K. Davis, I. Laguerre, L. D. Richardson, G. A. Diaz, N. M. Pearson, S. B. Ellis, C. Stolte, M. Robinson, P. Kovatch, C. R. Horowitz, B. D. Gelb, J. M. Greally, L. J. Bauman, R. E. Zinberg, N. S. Abul-Husn, M. P. Wasserstein, and E. E. Kenny, "GUÍA: a digital platform to facilitate result disclosure in genetic counseling," *Genetics in Medicine*, vol. 23, no. 5, pp. 942–949, May 2021, number: 5 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41436-020-01063-z 78

[164] K. Sun, P. Jiang, K. C. A. Chan, J. Wong, Y. K. Y. Cheng, R. H. S. Liang,

W.-k. Chan, E. S. K. Ma, S. L. Chan, S. H. Cheng, R. W. Y. Chan, Y. K. Tong, S. S. M. Ng, R. S. M. Wong, D. S. C. Hui, T. N. Leung, T. Y. Leung, P. B. S. Lai, R. W. K. Chiu, and Y. M. D. Lo, "Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 40, pp. E5503–E5512, Oct. 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4603482/ 18, 35

[165] V. Swarup and M. R. Rajeswari, "Circulating (cell-free) nucleic acids–a promising, non-invasive tool for early detection of several human diseases," *FEBS letters*, vol. 581, no. 5, pp. 795–799, Mar. 2007. 21

[166] S. Taavitsainen, M. Annala, E. Ledet, K. Beja, P. J. Miller, M. Moses, M. Nykter, K. N. Chi, O. Sartor, and A. W. Wyatt, "Evaluation of Commercial Circulating Tumor DNA Test in Metastatic Prostate Cancer," *JCO Precision Oncology*, no. 3, pp. 1–9, Jun. 2019. [Online]. Available: https://ascopubs.org/doi/10.1200/PO.19.00014 5

[167] G. O. Tadmouri, P. Nair, T. Obeid, M. T. Al Ali, N. Al Khaja, and H. A. Hamamy, "Consanguinity and reproductive health among Arabs," *Reproductive Health*, vol. 6, no. 1, p. 17, Oct. 2009. [Online]. Available: https://doi.org/10.1186/1742-4755-6-17 2, 75

[168] E. Taglauer, L. Wilkins-Haug, and D. Bianchi, "Review: Cell-free fetal DNA in the maternal circulation as an indication of placental health and disease," *Placenta*, vol. 35, no. Suppl, pp. S64–S68, Feb. 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4886648/ 17

[169] P. Thomas, W. Zahorodny, B. Peng, S. Kim, N. Jani, W. Halperin, and M. Brimacombe, "The association of autism diagnosis with socioeconomic status," *Autism*, vol. 16, no. 2, pp. 201–213, Mar. 2012, publisher: SAGE Publications Ltd. [Online]. Available: https://doi.org/10.1177/1362361311413397 78

[170] E. G. Toraño, M. G. García, J. L. Fernández-Morera, P. Niño-García, and A. F. Fernández, "The Impact of External Factors on the Epigenome: In Utero and over Lifetime," *BioMed Research International*, vol. 2016, p. 2568635, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4887632/ 2

[171] S. Tug, S. Helmig, J. Menke, D. Zahn, T. Kubiak, A. Schwarting, and P. Simon, "Correlation between cell free DNA levels and medical evaluation of disease progression in systemic lupus erythematosus patients," *Cellular Immunology*, vol. 292, no. 1, pp. 32–39, Nov. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0008874914001270 5

[172] N. Tung and J. E. Garber, "PARP inhibition in breast cancer: progress made and future hopes," *npj Breast Cancer*, vol. 8, no. 1, pp. 1–5, Apr. 2022, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41523-022-00411-3 1

[173] M. R. Turner, R. Bowser, L. Bruijn, L. Dupuis, A. Ludolph, M. Mcgrath, G. Manfredi, N. Maragakis, R. G. Miller, S. L. Pullman, S. B. Rutkove, P. J. Shaw, J. Shefner, and K. H. Fischbeck, "Mechanisms, models and biomarkers in amyotrophic lateral sclerosis," *Amyotrophic lateral sclerosis & frontotemporal degeneration*, vol. 14, no. 0 1, pp. 19–32, May 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4284067/ 5

[174] M. Velders, G. Treff, K. Machus, E. Bosnyák, J. Steinacker, and U. Schumann, "Exercise is a potent stimulus for enhancing circulating DNase activity," *Clinical Biochemistry*, vol. 47, no. 6, pp. 471–474, Apr. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0009912013006024 21

[175] N. Verber and P. J. Shaw, "Biomarkers in amyotrophic lateral sclerosis: a review of new developments," *Current Opinion in Neurology*, vol. 33, no. 5, p. 662, Oct. 2020. [Online]. Available: https://journals.lww.com/co-neurology/Fulltext/2020/10000/Biomarkers_

in_amyotrophic_lateral_sclerosis__a.18.aspx?casa_token=_Qi-ftDIiCsAAAAA:15oWDZm-nTI3ytIcFg-e4irDZ5ZSRWpo1JILS5b4c4ibW3ZptVxojO-e_Cm0ZvHIkOCXwFVSNn8aP_OcPo5tiZeH 43

[176] N. S. Verber, S. R. Shepheard, M. Sassani, H. E. McDonough, S. A. Moore, J. J. P. Alix, I. D. Wilkinson, T. M. Jenkins, and P. J. Shaw, "Biomarkers in Motor Neuron Disease: A State of the Art Review," *Frontiers in Neurology*, vol. 10, Apr. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6456669/ 7

[177] F. Verde, P. Steinacker, J. H. Weishaupt, J. Kassubek, P. Oeckl, S. Halbgebauer, H. Tumani, C. A. F. v. Arnim, J. Dorst, E. Feneberg, B. Mayer, H.-P. Müller, M. Gorges, A. Rosenbohm, A. E. Volk, V. Silani, A. C. Ludolph, and M. Otto, "Neurofilament light chain in serum for the diagnosis of amyotrophic lateral sclerosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 90, no. 2, pp. 157–164, Feb. 2019, publisher: BMJ Publishing Group Ltd Section: Neurodegeneration. [Online]. Available: https://jnnp.bmj.com/content/90/2/157 43

[178] W. Viechtbauer, "Conducting Meta-Analyses in R with the metafor Package," *Journal of Statistical Software*, vol. 36, pp. 1–48, Aug. 2010. [Online]. Available: https://doi.org/10.18637/jss.v036.i03 71, 89

[179] M. Vila and S. Przedborski, "Targeting programmed cell death in neurodegenerative diseases," *Nature Reviews Neuroscience*, vol. 4, no. 5, pp. 365–375, May 2003. [Online]. Available: https://www.nature.com/articles/nrn1100 5

[180] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and P. van Mulbregt, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272,

Mar. 2020, number: 3 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41592-019-0686-2 31

[181] I. D. Vlaminck, L. Martin, M. Kertesz, K. Patel, M. Kowarsky, C. Strehl, G. Cohen, H. Luikart, N. F. Neff, J. Okamoto, M. R. Nicolls, D. Cornfield, D. Weill, H. Valantine, K. K. Khush, and S. R. Quake, "Noninvasive monitoring of infection and rejection after lung transplantation," *Proceedings of the National Academy of Sciences*, vol. 112, no. 43, pp. 13 336–13 341, Oct. 2015. [Online]. Available: https://www.pnas.org/content/112/43/13336 5

[182] S. Volik, M. Alcaide, R. D. Morin, and C. Collins, "Cell-free DNA (cfDNA): Clinical Significance and Utility in Cancer Shaped By Emerging Technologies," *Molecular Cancer Research*, vol. 14, no. 10, pp. 898–908, Oct. 2016. [Online]. Available: https://mcr.aacrjournals.org/content/14/10/898 5

[183] W.-Q. Wei, L. A. Bastarache, R. J. Carroll, J. E. Marlo, T. J. Osterman, E. R. Gamazon, N. J. Cox, D. M. Roden, and J. C. Denny, "Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record," *PLOS ONE*, vol. 12, no. 7, p. e0175508, Jul. 2017, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0175508 68, 82

[184] D. R. Williams, S. A. Mohammed, J. Leavell, and C. Collins, "Race, socioeconomic status, and health: Complexities, ongoing challenges, and research opportunities," *Annals of the New York Academy of Sciences*, vol. 1186, no. 1, pp. 69–101, 2010, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-6632.2009.05339.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.2009.05339.x 64

[185] J. M. Williams, C. A. Duckworth, M. D. Burkitt, A. J. M. Watson, B. J. Campbell, and D. M. Pritchard, "Epithelial Cell Shedding and Barrier Function,"

Veterinary Pathology, vol. 52, no. 3, pp. 445–455, May 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4441880/ 18

[186] S. K. Wise, M. D. Ghegan, E. Gorham, and R. J. Schlosser, "Socioeconomic factors in the diagnosis of allergic fungal rhinosinusitis," *Otolaryngology–Head and Neck Surgery*, vol. 138, no. 1, pp. 38–42, Jan. 2008, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.1016/j.otohns.2007.10.020 78

[187] Y. Xi and W. Li, "BSMAP: whole genome bisulfite sequence MAPping program," *BMC Bioinformatics*, vol. 10, no. 1, p. 232, Jul. 2009. [Online]. Available: https://doi.org/10.1186/1471-2105-10-232 2

[188] F. Xiong, M. Sun, X. Zhang, R. Cai, Y. Zhou, J. Lou, L. Zeng, Q. Sun, Q. Xiao, X. Shang, X. Wei, T. Zhang, P. Chen, and X. Xu, "Molecular epidemiological survey of haemoglobinopathies in the Guangxi Zhuang Autonomous Region of southern China," *Clinical Genetics*, vol. 78, no. 2, pp. 139–148, Aug. 2010. 74

[189] P. P. Yeung and S. Greenwald, "Jewish Americans and mental health: results of the NIMH Epidemiologic Catchment Area Study," *Social Psychiatry and Psychiatric Epidemiology*, vol. 27, no. 6, pp. 292–297, Nov. 1992. [Online]. Available: https://doi.org/10.1007/BF00788901 69

[190] H. Zhao, Z. Sun, J. Wang, H. Huang, J.-P. Kocher, and L. Wang, "CrossMap: a versatile tool for coordinate conversion between genome assemblies," *Bioinformatics (Oxford, England)*, vol. 30, no. 7, pp. 1006–1007, Apr. 2014. 82

[191] Y.-n. Zhou, Y.-h. Chen, S.-q. Dong, W.-b. Yang, T. Qian, X.-n. Liu, Q. Cheng, J.-c. Wang, and X.-j. Chen, "Role of Blood Neurofilaments in the Prognosis of Amyotrophic Lateral Sclerosis: A Meta-Analysis," *Frontiers in Neurology*, vol. 12, p. 712245, Oct. 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8526968/ 43

[192] Y. Zhou, S. R. Browning, and B. L. Browning, “A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data,” *The American Journal of Human Genetics*, vol. 106, no. 4, pp. 426–437, Apr. 2020, publisher: Elsevier. [Online]. Available: https://www.cell.com/ajhg/abstract/S0002-9297(20)30052-5 84

[193] M. J. Ziller, H. Gu, F. Müller, J. Donaghey, L. T.-Y. Tsai, O. Kohlbacher, P. L. De Jager, E. D. Rosen, D. A. Bennett, B. E. Bernstein, A. Gnirke, and A. Meissner, “Charting a dynamic DNA methylation landscape of the human genome,” *Nature*, vol. 500, no. 7463, pp. 477–481, Aug. 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3821869/ 15, 34, 44

[194] M. J. Ziller, K. D. Hansen, A. Meissner, and M. J. Aryee, “Coverage recommendations for methylation analysis by whole genome bisulfite sequencing,” *Nature methods*, vol. 12, no. 3, pp. 230–232, Mar. 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4344394/ 9