

Integrative Analyses of Genetic and Genomic Sequence Data to Improve GWAS Interpretation

by

Sarah C. Hanks

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2023

Doctoral Committee:

Professor Michael Boehnke, Co-Chair
Research Professor Laura J. Scott, Co-Chair
Dr. Christian Fuchsberger, Eurac Research
Professor Hyun Min Kang
Associate Professor Stephen C. J. Parker
Professor Xiaoquan Wen

Sarah C. Hanks

schanks@umich.edu

ORCID iD: [0000-0003-2978-5289](https://orcid.org/0000-0003-2978-5289)

© Sarah C. Hanks 2023

Dedication

To my parents, Dr. Kathleen Gould and Dr. Steven Hanks, and to my sister, soon-to-be Dr. Jessica Hanks.

Acknowledgements

This dissertation would not have been possible without the dedicated mentorship of each member of my dissertation committee. Thank you to Hyun Min Kang for years of insightful and helpful feedback on every project presented here. Thank you to Christian Fuchsberger for patiently guiding me through my first first-author paper and to William Wen for patiently guiding me through my first methods paper. Thank you to Steve Parker and the entire Parker lab for sharing your biological and computational expertise and for being a second academic home at Michigan. I am especially grateful for my Co-Chairs Laura Scott and Michael Boehnke. Laura, thank you for so generously sharing your time and approach to every step of the scientific process, from study design and hypothesis generation to writing and presentation. Your curiosity has continually inspired me, and any success I have in designing figures or slides is almost entirely attributable to your guidance. Mike, thank you for everything you have done to support my growth as a scientist and colleague over the years. Despite having “admit[ted] that [you are] ‘a busy person’” (thanks, Reviewer 2), your invaluable advice has only ever been a quickly responded email away. I look to you as an exemplary role model for supportive and effective scientific leadership. Finally, thank you to Mike, Laura, Hyun, and Steve for being such fantastic F31 Co-Sponsors and for treating me with the respect of a junior colleague throughout the process.

I have had the opportunity to learn from so many other incredible professors during graduate school. I would like to particularly thank Gonalo Abecasis, Bhramar Mukherjee,

Xiang Zhou, and Sebastian Zöellner for your art in planning and teaching courses that were informative, engaging, and motivating all at once. I would also like to thank Matt Zawistowski for your dedicated mentorship as I begin my own journey as a teacher.

I am also indebted to all past and present members of the Boehnke/Scott group. A special thank you to Anne Jackson and Heather Stringham for teaching me so much about study design, reproducible research, and collaboration, and for so patiently fielding a million questions about any and everything. Thank you to Ryan Welch for your computational support and kind analysis suggestions. I am also particularly grateful to past Boehnke/Scott post-docs and senior graduate students for your support and advice and for being my scientific and academic role models. Thank you also to all the members of the FUSION team for being helpful and inspiring members of my extended academic family.

My time in graduate school would not have been the same without the Genome Sciences Training Program. Thank you to Mike and to Dawn Keene and Peggy White for your leadership and dedication. I am grateful for all my fellow trainees and for the opportunity to learn alongside and from one another.

I am grateful for so many truly wonderful friends that I had the privilege to meet in graduate school. Thank you to Aubrey Annis, Elizabeth Chase, Abhay Hukku, Kevin Liao, Michelle McNulty, Jenny Nguyen, Pedro Orozco del Pino, Anita Pandit, Cathy Smith, Mike Sweeney, Nicky Wakim, Josh Weinstock, Cynthia Zajac, Christina Zhou, and many others. A huge thank you to Emily Roberts and Brooke Wolford for your inspiration and support. I am so grateful to have you both as lifelong friends and colleagues.

Finally, words cannot express my gratitude to my family for their encouragement, support, and love at every stage of my life. Thank you for sharing, through nature and nurture, your love of learning.

Table of Contents

Dedication	ii
Acknowledgements	iii
List of Tables	x
List of Figures	xii
Abstract	xv
Chapter 1 Introduction	1
Chapter 2 Extent to Which Array Genotyping and Imputation with Large Reference Panels Approximates Deep Whole Genome Sequencing.....	5
2.1 Introduction.....	5
2.2 Methods.....	6
2.2.1 Genetic data resources	6
2.2.2 Genotype imputation.....	8
2.2.3 Evaluation of imputation quality	9
2.2.4 Predicted variant consequences	10
2.2.5 Fine-scale ancestry estimation	10
2.2.6 Effect of regional genomic features on imputation quality	11
2.2.7 Effect of real vs. WGS-based array genotypes on evaluation of imputation quality.....	12
2.2.8 Effect of variant caller on evaluation of imputation quality	12
2.2.9 Imputability tool for the Michigan Imputation Server.....	12
2.3 Results.....	13

2.3.1 Whole genome sequencing studies of four ancestries	13
2.3.2 Impact of reference panel on genotype imputation quality	13
2.3.3 Less influence of genotype array size with TOPMed- compared to 1000G and HRC-based imputation.....	15
2.3.4 Individual-level imputation accuracy varies with finer-scale ancestry.....	15
2.3.5 Imputation quality varies across the genome.....	17
2.3.6 Local genomic features explain little variability in imputation quality	18
2.3.7 Impact of variant predicted function and type on imputation quality.....	19
2.4 Discussion	20
2.5 Figures and Tables	24
2.6 Acknowledgements and publication	28
2.7 Supplementary Material.....	30
Chapter 3 Statistical Methods for Genetic Colocalization in a Single Cohort Design	61
3.1 Introduction.....	61
3.2 Methods.....	63
3.2.1 Data resources and processing	63
3.2.2 Simulations	63
3.3 Results.....	67
3.3.1 One-sample colocalization design	67
3.3.2 Two-sample colocalization in one-sample design	68
3.3.3 Adjusting for true trait-shared confounder reduces Type I and Type II error rates. 68	
3.3.4 Estimating trait-shared confounder with probabilistic principal component analysis introduces collider bias	69
3.3.5 Regressing SNP effects on estimated confounder offers no improvements over two-sample methods for one-sample design.....	70
3.4 Discussion	70

3.5 Tables and Figures	73
3.6 Supplementary Material.....	76
Chapter 4 Extensive Differential Gene Expression and Regulation by Sex in Human Skeletal Muscle.....	82
4.1 Introduction.....	82
4.2 Methods.....	83
4.2.1 Data collection and processing	83
4.2.2 Statistical analysis of single nucleus data	86
4.2.3 Statistical analysis of bulk data.....	89
4.3 Results.....	91
4.3.1 Gene and miRNA expression and chromatin accessibility assayed in 281 vastus lateralis muscle biopsies	91
4.3.2 Clustering of snRNA-seq and snATAC-seq nuclei identifies 12 cell types	92
4.3.3 Single nucleus data show muscle cell type composition differs by sex	92
4.3.4 Differential gene expression by sex in muscle cell types	93
4.3.5 Differential gene expression by sex is concordant across muscle fiber types	94
4.3.6 LncRNAs and pseudogenes enriched for differential expression by sex in muscle cell types	94
4.3.7 Mitochondrial activity, signal transduction, and cell differentiation pathways enriched for sex-biased genes in muscle fiber types.....	95
4.3.8 Differential gene expression by sex in bulk skeletal muscle	96
4.3.9 Concordance of sex-biased expression between cell types and bulk skeletal muscle	97
4.3.10 Differential miRNA expression by sex in bulk skeletal muscle	97
4.3.11 Differential chromatin accessibility by sex in muscle cell types.....	99
4.3.12 Differential chromatin accessibility by sex is concordant across muscle fiber types	99
4.3.13 Sex-biased peaks enriched for gene regulatory function	100

4.3.14 Directional concordance of differential accessibility of promoter region peaks and gene expression	101
4.4 Discussion	101
4.5 Tables and Figures	106
4.6 Supplementary Material.....	110
Chapter 5 Discussion	134
5.1 Summary and immediate extensions	134
5.2 Emerging themes and future directions	138
5.2.1 Ancestral diversity in publicly available genetic and genomic resources	138
5.2.2 Modeling continuous nature of genetic ancestry	139
5.2.3 Promises and challenges of translational genetics	140
Bibliography	142

List of Tables

Table 2.1 Whole-genome sequencing (WGS) datasets.....	24
Supplementary Table 2.1 Whole genome sequencing (WGS)-based genotype arrays	52
Supplementary Table 2.2 Number of well-imputed biallelic single nucleotide variants (SNVs) in each whole genome sequencing (WGS) study by reference panel, genotype array, ancestry, and minor allele frequency (MAF) category	53
Supplementary Table 2.3 Proportion of biallelic single nucleotide variants (SNVs) in each whole genome sequencing (WGS) study that are well-imputed ($r^2 > 0.8$) by reference panel, genotype array, ancestry, and minor allele frequency (MAF) category	54
Supplementary Table 2.4 Minor allele frequency (MAF) threshold above which array genotyping and imputation can approximate whole genome sequencing (WGS) for biallelic single nucleotide variants (SNVs) by reference panel, genotype array, and ancestry	55
Supplementary Table 2.5 Mean heterozygous concordance rates by reference panel, genotype array, ancestry, and MAF category	56
Supplementary Table 2.6 25th, 50th, and 75th percentiles of the number of consecutive well-imputed (observed imputation $r^2 > 0.8$) biallelic single nucleotide variants (SNVs) by reference panel, genotype array, ancestry, and minor allele frequency (MAF) category	57
Supplementary Table 2.7 25th, 50th, and 75th percentiles of the length in kilobases (kb) of consecutively well-imputed (observed imputation $r^2 > 0.8$) variants by reference panel, genotype array, ancestry, and minor allele frequency (MAF) category	58
Supplementary Table 2.8 Proportion of biallelic single nucleotide variants (SNVs) in each whole genome sequencing (WGS) study that are well-imputed ($r^2 > 0.8$) by reference panel, genotype array, ancestry, and predicted impact on protein coding.....	59
Supplementary Table 2.9 Minor allele frequency (MAF) threshold above which array genotyping and imputation can approximate whole genome sequencing (WGS) with the TOPMed panel by genotype array, ancestry, and variant type	60
Supplementary Table 3.1 Unadjusted fastENLOC results	78
Supplementary Table 3.2 Unadjusted DAP-G results	79

Supplementary Table 3.3 Confounder-adjusted fastENLOC results	79
Supplementary Table 3.4 Confounder-adjusted DAP-G results	79
Supplementary Table 3.5 Probabilistic principal component analysis (PPCA)-adjusted fastENLOC results	79
Supplementary Table 3.6 Probabilistic principal component analysis (PPCA)-adjusted DAP-G results	80
Supplementary Table 3.7 Ridge regression residual-adjusted fastENLOC results	80
Supplementary Table 3.8 Ridge regression residual-adjusted DAP-G results	81
Supplementary Table 4.1 Sample characteristics and sample size by datatype of 118 female and 163 male vastus lateralis biopsy donors from the FUSION Tissue Biopsy Study.....	127
Supplementary Table 4.2 Mean and standard deviation of the proportion of nuclei by cell type and sex	128
Supplementary Table 4.3 Associations of sex with the number of nuclei in each cell type from negative binomial regressions.....	129
Supplementary Table 4.4 Number and mean UMI of genes tested for differential expression by sex by gene type and cell type. Gene types were defined by GENCODE.....	130
Supplementary Table 4.5 Association of gene type with differential expression by sex status .	131
Supplementary Table 4.6 Association of chromatin state with differential accessibility by sex status for autosomal peaks	132
Supplementary Table 4.7 Association of sex-biased promoter peaks with sex-biased gene expression in the fiber types, the single nucleus pseudobulk, and the bulk.....	133

List of Figures

Figure 2.1 Proportion of well-imputed ($r^2 > 0.8$) bi-allelic SNVs by reference panel, study ancestry, and genotyping array	25
Figure 2.2 Heterozygous genotype concordance rates for rare variants by ancestry with TOPMed panel imputation.....	26
Figure 2.3 Regional variability in TOPMed reference panel imputation quality	27
Figure 2.4 Genomic features associated with TOPMed imputation quality of bi-allelic SNVs by ancestry	28
Supplementary Figure 2.1 Effect of sample size on imputation quality metrics	30
Supplementary Figure 2.2 Mean observed imputation r^2 of biallelic SNVs by reference panel, study ancestry, and genotyping array.....	31
Supplementary Figure 2.3 Imputation quality of biallelic SNVs by reference panel using WGS-based and real Illumina OmniExpress arrays.....	32
Supplementary Figure 2.4 Proportion of well-imputed ($r^2 > 0.8$) biallelic SNVs by reference panel, genotyping array, and variant caller in Finnish study	33
Supplementary Figure 2.5 Heterozygous genotype concordance rates for low-frequency variants by ancestry with TOPMed panel imputation.....	34
Supplementary Figure 2.6 Heterozygous genotype concordance rates for common variants by ancestry with TOPMed panel imputation.	35
Supplementary Figure 2.7 Principal component analysis of WGS samples.....	36
Supplementary Figure 2.8 Regional variability in imputation quality of common variants with the TOPMed reference panel by genotyping array and ancestry across all chromosomes.....	47
Supplementary Figure 2.9 Repeat classes associated with TOPMed imputation quality of biallelic SNVs by ancestry	48
Supplementary Figure 2.10 Genomic features associated with TOPMed imputation quality of biallelic SNVs by ancestry	49

Supplementary Figure 2.11 Proportion of well-imputed ($r^2 > 0.8$) biallelic SNVs by predicted functional impact and ancestry	50
Supplementary Figure 2.12 Proportion of well-imputed ($r^2 > 0.8$) variants by variant type, genotyping array, and ancestry with the TOPMed panel.....	51
Supplementary Figure 2.13 Distribution of MAF for biallelic SNVs by ancestry	52
Figure 3.1 Directed acyclic graphs of colocalization in a two-sample and one-sample design ...	73
Figure 3.2 Two-sample colocalization in a single-cohort design	74
Figure 3.3 Colocalization from adjusted marginal association analyses in a single-cohort design	75
Supplementary Figure 3.1 Correlation of simulated continuous phenotypes	76
Supplementary Figure 3.2 Estimation of the trait-shared, non-genetic confounder with probabilistic principal component analysis (PPCA)	77
Supplementary Figure 3.3 Estimation of the trait-shared, non-genetic confounder with ridge regression residuals from probabilistic principal component analysis (PPCA).....	78
Figure 4.1 Sex differences in cell type composition of human skeletal muscle	106
Figure 4.2 Sex differences in cell type-specific gene expression in human skeletal muscle.....	107
Figure 4.3 Comparison of sex differences in gene expression from bulk vs. single nucleus RNA-seq	108
Figure 4.4 Sex differences in cell-type specific chromatin accessibility in human skeletal muscle	109
Supplementary Figure 4.1 Sample sizes across molecular data modalities.....	110
Supplementary Figure 4.2 Sex differences in cell type composition of human skeletal muscle using RNA nuclei.....	111
Supplementary Figure 4.3 Sex differences in cell type composition of human skeletal muscle using ATAC nuclei	112
Supplementary Figure 4.4 The cumulative distribution of the absolute fold change of sex-biased genes by cell type, chromosome, and direction of effect	113
Supplementary Figure 4.5 Comparison of differential expression by sex with and without adjusting for oral glucose tolerance test (OGTT) status	114
Supplementary Figure 4.6 Comparison of differential expression by sex across muscle cell types	116

Supplementary Figure 4.7 Association of gene type with sex-biased expression	117
Supplementary Figure 4.8 Gene expression levels in muscle fibers of top 20 autosomal sex-biased genes in oxidative phosphorylation and caveola GO terms.....	118
Supplementary Figure 4.9 Comparison of gene set enrichment test results for differential expression by sex with and without adjusting for oral glucose tolerance test (OGTT) status....	119
Supplementary Figure 4.10 Comparison of differential expression by sex between bulk and pseudobulk by expression level	120
Supplementary Figure 4.11 Comparison of differential expression by sex between FUSION bulk and GTEx bulk.....	121
Supplementary Figure 4.12 The cumulative distribution of the absolute fold change of sex-biased peaks by cell type, chromosome, and direction of effect	122
Supplementary Figure 4.13 Comparison of differential accessibility by sex with and without adjusting for oral glucose tolerance test (OGTT) status	123
Supplementary Figure 4.14 Comparison of differential accessibility by sex across muscle cell types	125
Supplementary Figure 4.15 Association of chromatin state with differential accessibility by sex	126

Abstract

Genome-wide association studies (GWAS) to date have identified hundreds of thousands of genetic variants associated with tens of thousands of complex human diseases and traits. However, fully understanding the biological mechanisms of these associations remains challenging and requires more complete identification of causal genetic variants and their functional consequences in human cells and tissues. Here, we propose novel methods and approaches for integrative analyses of genetic and genomic sequence data to further this understanding.

In the first project, we quantify the extent to which array genotyping and imputation can approximate deep whole genome sequencing (WGS) across a range of ancestries, reference panels, and genotyping arrays. Deep WGS, the gold standard technology for genetic variant identification and genotyping, remains very expensive for most large studies. In this chapter, we use WGS data from studies of individuals of African, Hispanic/Latino, and European ancestry in the US, and of Finnish ancestry in Finland (a population isolate) and perform genotype imputation using the genetic variants present on the Illumina Core, OmniExpress, MEGA, and Omni 2.5M arrays with the 1000G, HRC, and TOPMed imputation reference panels. We find that using the Omni 2.5M array and the TOPMed panel, $\geq 90\%$ of biallelic single nucleotide variants (SNVs) are well-imputed ($r^2 > 0.8$) down to minor allele frequencies (MAF) of 0.14% in African, 0.11% in Hispanic/Latino, 0.35% in European, and 0.85% in Finnish ancestries. We find that individual-level imputation quality varies widely between and within the three US

populations. Imputation quality also varies across genomic regions, producing regions where even common ($MAF > 5\%$) variants were not consistently well-imputed across ancestries.

In the second project, we investigate the consequences of violating the independent-cohorts assumption of genetic colocalization methods. Colocalization analysis aims to identify genetic variants that are causal for multiple association signals at a single locus. Existing colocalization methods explicitly assume that the phenotypes are measured in independent, non-overlapping samples. In this chapter, we present simulation analyses that demonstrate the consequences of applying these methods in a single cohort. We show that Type I error is well-controlled when the ratio of shared to trait-specific error variance is low but becomes problematic with increased sharing. For scenarios with well-controlled Type I error, we show that the one-sample design is more powerful than the two-sample design due to better linkage disequilibrium matching. Power can be further improved in the one-sample design when shared non-genetic factors are measured and controlled for in the marginal association analyses.

In the third project, we examine sex differences in gene expression and regulation in human skeletal muscle at the single nucleus resolution. We identify thousands of sex-biased genes across Type 1, 2A, and 2X muscle fibers and other, less abundant cell types. We find that sex-biased expression is highly concordant across the muscle fiber types and bulk muscle tissue and is enriched for genes in mitochondrial activity (males) and muscle regeneration (females) pathways. We also find that lncRNAs and miRNAs, two classes of genes with regulatory functions, show extensive sex-biased expression in the fiber-type and bulk data, respectively. We find widespread sex-biased chromatin accessibility enriched in regulatory chromatin states. Together, these results highlight nuclear and cytoplasmic mechanisms for sex-differential gene regulation in skeletal muscle.

Chapter 1 Introduction

An important goal of human genetics is to identify genes that influence heritable traits and diseases to improve strategies for prevention and treatment.¹ Genome wide association studies (GWAS) have identified hundreds of thousands of genetic variants associated with tens of thousands of human traits and diseases,² but the vast majority of these variants lie in non-coding regions.³ These variants are thought to regulate gene expression, but linking the genetic variants to genes and the biological contexts in which they influence the trait is challenging.⁴

Functional studies improve our understanding of the relationship between regulatory variants and genes by measuring gene expression and its regulation through related molecular phenotypes (e.g. DNA methylation, chromatin accessibility) in disease-relevant tissues.⁵⁻⁷ RNA sequencing (RNA-seq)⁸ quantifies levels of gene expression and can be performed in bulk tissue or at single-cell (scRNA-seq)⁹ or single-nucleus (snRNA-seq)¹⁰ resolution. Studies with both genotype and RNA-seq data on the same set of individuals can perform expression quantitative trait loci (eQTL) analyses to identify associations between genetic variants and gene expression levels.¹¹ Similarly, studies with both genotype and ATAC-seq data quantifying chromatin accessibility⁷ can perform chromatin accessibility quantitative trait loci (caQTL) analyses to identify associations between genetic variants and accessible regulatory regions.¹² These analyses provide some evidence for how regulatory variants may influence molecular phenotypes, but they alone do not reveal the mechanisms of most non-coding trait-associated genetic variants.¹³

One line of evidence that can further implicate an eQTL gene or a gene in close proximity to a disease-associated regulatory variant is the identification of rare coding variation in that gene. Compared to common variants, rare variants represent more recent variation and are more likely to have large effects on disease risk, but they require larger sample sizes both to detect and to identify their associations.¹⁴ Deep whole genome sequencing (WGS) is the current gold standard for accurately capturing most genetic variants across the genome and minor allele frequency (MAF) spectrum, especially for rare variants.¹⁵ But despite recent improvements in sequencing technologies and corresponding decreases in sequencing cost and increases in sequencing throughput, deep WGS remains prohibitively expensive for most large studies.^{15,16} In contrast, genotype arrays assay hundreds of thousands to millions of variants, representing only a small fraction of genetic variation, but at a much lower cost. Variants that are not array genotyped can be statistically inferred by comparing sample haplotypes to an external reference panel of sequenced haplotypes via genotype imputation.¹⁷ Recent increases in the size, sequencing depth, and diversity of imputation reference panels have improved imputation quality,^{18–20} suggesting that under some conditions the less expensive arrays may capture most genetic variants with similar accuracy to costly WGS.²⁰ In Chapter 2, we quantify the extent to which array genotyping and imputation can approximate deep WGS across a range of ancestries, reference panels, and genotyping arrays.

Another line of evidence that implicates genes in disease pathways comes from combining GWAS and functional analyses with genetic colocalization analysis. Colocalization methods integrate associations with multiple phenotypes, most often from GWAS and eQTL analyses, by seeking to identify variants that are causal for both associations.^{4,21} When successful, these methods provide evidence that a variant influences the GWAS trait by

regulating the expression of a particular gene in a particular tissue or cell type.²² Because tissue samples are usually more difficult to obtain than disease status, eQTL studies are often much smaller than GWAS, and the two types of analyses are often performed on non-overlapping samples.⁵ Many colocalization methods explicitly assume independence of eQTL and GWAS data.²³ However, an increasing number of studies have collected multiple molecular phenotypes measured on the same set of samples.^{6,24} One key advantage of this type of study design is that the pattern of linkage disequilibrium is necessarily the same for both association analyses, which satisfies another assumption of current methods.²² However, there are no methods designed for colocalization analysis with single-cohort designs. In Chapter 3, we evaluate the consequences of violating the non-overlapping cohort assumption and provide guidelines for researchers conducting colocalization analysis of two phenotypes in a single cohort.

A complicating factor in identifying disease-related genes is that the effects of some genes depend on other biological or environmental variables, such as age, sex, diet, infection, or stress. Gene by environment interactions have been identified for a wide range of traits (some examples include psychiatric,^{25,26} cardiometabolic,^{27,28} and neurodegenerative^{29,30} diseases). Chromosomal sex is one such interacting variable that has profound effects on human health at all stages of life, from puberty³¹ to menopause³² to life expectancy itself.³³ Throughout the lifespan, sex differences in disease risk have been observed for a wide range of conditions, including autoimmune disorders,³⁴ cardiovascular disease³⁵, infectious disease³⁶, and age-related disorders like osteoporosis³⁷ and dementia³⁸, among many others. Sex-stratified GWAS have identified genetic variants that are associated with anthropometric measures, arthritis, and gout in only one sex,³⁹ and sex is associated with expression levels of thousands of genes in tissues throughout the human body.⁴⁰ Characterizing the pervasive effects of sex on molecular

phenotypes and uncovering the regulatory mechanisms that drive these differences will further help contextualize the impact of genetic variation on health and disease. In Chapter 4, we examine sex differences in gene expression at the cell-type and whole-tissue levels in human skeletal muscle and highlight several potential nuclear and cytoplasmic regulatory mechanisms for these differences.

Together, these chapters improve our ability to perform tests of genetic associations and interpret their results with the goal of improving our understanding of the biological mechanisms that cause variability in human disease risks and complex traits.

Chapter 2 Extent to Which Array Genotyping and Imputation with Large Reference Panels Approximates Deep Whole Genome Sequencing

2.1 Introduction

Deep whole genome sequencing (WGS) accurately captures most genetic variants across the genome and minor allele frequency (MAF) spectrum.¹⁵ Advances in sequencing technologies, and corresponding decreases in sequencing cost, have enabled ever larger human sequencing studies.^{20,41–43} Such studies have identified rare alleles that cause Mendelian diseases^{44–46} and contribute to risk of common diseases⁴⁷ and variation in quantitative traits.^{41,43} However, deep WGS remains prohibitively expensive and computationally intensive for large studies.^{15,16}

In contrast to WGS, genotype arrays assay hundreds of thousands to millions of variants, representing only a small fraction of genetic variation, but at a much lower cost. Variants that are not array genotyped can be statistically inferred by comparing sample haplotypes to an external reference panel of sequenced haplotypes via genotype imputation.¹⁷ Most common ($\text{MAF} > 5\%$) variants are present in recent reference panels and can be imputed with high accuracy from genotype arrays.^{18–20} However, low-frequency ($0.5\% < \text{MAF} \leq 5\%$) and rare ($\text{MAF} \leq 0.5\%$) variants appear less often or may be absent from the reference panel, making their imputation less accurate or impossible.⁴⁸ Therefore, using inexpensive genotype arrays and imputation in place of costly deep WGS can result in lower coverage and less accurate genotyping of rare genetic variation.

Reference panel, genotype array, sample ancestry, and genomic location all influence imputation quality.^{18,19,48,49} Previous studies have evaluated imputation quality with the multiethnic 1000 Genomes Phase 3 (1000G),¹⁸ the predominantly European Haplotype Reference Consortium (HRC),¹⁹ and two releases of the multiethnic Trans-Omics for Precision Medicine (TOPMed)^{20,50} panels, finding that larger, more diverse, more deeply sequenced panels support more accurate imputation. Likewise, denser genotype arrays are associated with higher imputation quality,^{19,48,51} although the effect of array size on imputation quality has not been studied with the TOPMed panels. Regional variability in imputation quality with the 1000G panel is associated with genomic features including repeats and GC content,⁴⁹ but the degree to which imputation quality varies across the genome with the larger HRC or TOPMed panels is unknown. It is also unknown to what extent individual-level imputation quality varies within populations for any reference panel.

Here, we determine the extent to which genotyping with the Illumina Core, OmniExpress, MEGA, and Omni 2.5M arrays followed by imputation with the 1000G, HRC, and TOPMed reference panels can approximate deep WGS in studies with individuals of African, Hispanic/Latino, non-Finnish European, and Finnish ancestries. Depending on the MAF of variants relevant to the research question, study ancestry, and genomic location, we found that array genotyping and imputation can approximate WGS. Our findings, together with our new RsqBrowser tool for querying imputation quality, should help guide investigator decisions between these two technologies.

2.2 Methods

2.2.1 Genetic data resources

2.2.1.1 Whole genome sequencing data and processing

We used WGS data from the BioMe,⁵² InPSYght, METSIM,^{24,53} and MLOF⁵⁴ studies. Detailed descriptions of sample collection, sequencing, and data processing for BioMe and MLOF are provided by the TOPMed Informatics Research Center.²⁰ Corresponding information is available for the METSIM study.²⁴ The InPSYght study is a deep whole-genome sequencing US-based case-control study of individuals of admixed African-European or African genetic ancestry. Cases have either bipolar disorder or schizophrenia. The study is composed of samples from the Genomic Psychiatry Cohort (GPC),^{55,56} Consortium on the Genetics of Schizophrenia (COGS),⁵⁷ and from the NIMH repository from the Bipolar Genome Study (BIGS),⁵⁸ Lithium treatment moderate dose use study (LiTMUS)⁵⁹ and Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD) studies,⁶⁰ all obtained from the NIMH repository. Whole genome sequencing of the samples (mean depth 27 \pm 5.5 X) was performed at the Broad Institute. We excluded individuals with: sex mismatches (n=20), non-XX or XY sex karyotypes (n=17), duplicates (n=366), >5% DNA contamination (n=4), an excess of singletons (n=39), <25% global African ancestry as determined by ADMIXTURE⁶¹ analysis of array genotype data, or for whom <98% of sites were at a sequencing depth of ≥ 10 (n=14).

Participants from the BioMe biobank self-reported as Hispanic/Latino and were recruited at the Mount Sinai Health System in New York City (N=4,677; Table 2.1). Participants in the MLOF study self-reported as non-Hispanic white and were recruited throughout the US (N=2,987). Participants in the METSIM study were recruited in Kuopio, Finland (N=3,045). Based on recruiting location, self-reported and genetic ancestry, we designated the population groups Hispanic/Latino (BioMe), African (InPSYght), Finnish (METSIM), and European (MLOF) ancestry for the purposes of this study.

In all studies, we removed participants inferred to be related at a second degree or closer relationship using KING⁶² to any other individual genotyped in TOPMed Freeze 9 (n=157,675), including all participants in these four studies and all individuals in the TOPMed imputation reference panel. This filtering yielded 3,141 participants in BioMe, 7,169 in InPSYght, 2,703 in METSIM, and 2,429 in MLOF. We then randomly downsampled to 2,429 individuals in each study (Supplementary Figure 2.1).

WGS variant calling for all four studies was performed jointly with TOPMed Freeze 9 by the TOPMed Informatics Research Center (IRC) using the TOPMed Variant Calling/ GotCloud pipeline^{20,63}. We analyzed biallelic SNVs, multiallelic SNVs, biallelic indels, and multiallelic indels separately with n-allele variants recoded and analyzed as n-1 biallelic variants at the same position.

2.2.1.2 Array genotyping in METSIM

METSIM participants were genotyped with the Illumina Human OmniExpress array. Variants with poor mapping of probes to GRCh37, call rate <95%, or deviations from Hardy-Weinberg equilibrium ($p < 10^{-6}$) were removed.⁶⁴

2.2.2 Genotype imputation

For each study, we subsetted WGS variants to those present on the Illumina Infinium Core (0.3M markers), Illumina Omni Express (0.7M), Infinium Omni 2.5M (2.4M), and Multi-Ethnic Genotyping (MEGA; 1.8M) arrays (Supplementary Table 2.1). We refer to these WGS variant subsets as WGS-based arrays. For each study, we phased the selected variants with Eagle 2.4.1 and imputed genotypes using Minimac4 on the Michigan Imputation Server (pipeline version 1.2.4)⁶⁵ with the (1) 1000 Genomes Phase 3 (n=2,504), (2) Haplotype Reference Consortium (n=32,470), and (3) modified TOPMed (n=88,804) reference panels. MLOF and

BioMe are included in the full, publicly available TOPMed r2 (n=97,256) panel. To avoid overlap of participants and the presence of close relatives in the reference panel, we removed 4,694 BioMe (4,668 Hispanic/Latino and 26 missing ethnicity) and 3,758 MLOF (2,977 non-Hispanic white and 781 missing race/ethnicity) individuals from the full TOPMed r2 panel to create our modified TOPMed panel.

2.2.3 Evaluation of imputation quality

2.2.3.1 Observed imputation r^2

For each variant, we calculated the observed imputation r^2 as the squared Pearson correlation coefficient between the imputed genotype dosages and the sequence-based genotypes. We assigned $r^2=0$ for any variant present in the sequenced individuals but absent from the reference panels and so not imputed. For each variant category (biallelic SNVs, biallelic indels, multiallelic SNVs, and multiallelic indels) and each WGS-based array, we calculated the proportion of variants that were well-imputed (observed imputation $r^2>0.8$) within study-specific MAF bins of size 0.00025 for MAF between 0.0002 and 0.002 and of size 0.001 for MAF > 0.002. Each minor allele for multiallelic variants was analyzed independently of the other minor alleles of the same variant so that multiallelic variants had the same number of r^2 measurements as minor alleles.

2.2.3.2 Genotype concordance

Separately for common, low-frequency, and rare biallelic SNVs, we calculated the heterozygous concordance rate between the imputed best-guess genotypes and sequenced-based genotypes as the proportion of heterozygous variants in WGS that were present in the reference

panel that were also heterozygous in the imputed data using bed-diff.⁶⁶ We excluded biallelic SNVs that were absent from the reference panels in these calculations.

2.2.4 Predicted variant consequences

In each study, we used VEP⁶⁷ to predict the functional consequences of biallelic SNVs. We partitioned variants into four classes based on the predicted impact on protein coding: high, moderate, low, and modifier. While variants in the high and moderate classes are likely to change protein behavior, variants in the low impact class are unlikely to do so. Modifier variants are mostly non-coding with no evidence of impacting protein coding.

2.2.5 Fine-scale ancestry estimation

For InPSYght, we estimated the proportion of African ancestry present in each individual using RFMix⁶⁸ with two reference groups representing African and European ancestry from 1000G. For BioMe, participants had previously been grouped by continental origin and into identity-by-descent (IBD) communities representing groups with shared recent genetic ancestry.⁵² We labeled BioMe participants as from a Caribbean population if their continental origin was Caribbean or if they were members of the Puerto Rican or Dominican IBD communities. We labeled all other BioMe participants with non-missing continental origin as non-Caribbean. Participants not from Puerto Rican or Dominican IBD communities and with missing continental origin information were not included in comparisons between Caribbean and non-Caribbean populations.

We performed principal component analysis (PCA) to obtain fine-scale ancestry information for all four studies. We used 1000G-imputed genotypes on chromosome 1 to project

participants from each of the four studies onto the 938 reference samples from the Human Genome Diversity Project⁶⁹ using the LASER server.⁷⁰

2.2.6 Effect of regional genomic features on imputation quality

2.2.6.1 Genomic features datasets

We downloaded GC content over 5bp intervals, the genomic positions of segmental duplications, the genomic positions of structural variants annotated with the Database of Genomic Variants, and the genomic positions of repeats identified with RepeatMasker from the UCSC Genome Browser database.⁷¹ Recombination rate was calculated using the HapMap GrCh38 genetic map⁷² as centimorgans per megabase.

2.2.6.2 Relationship between genomic features and TOPMed imputation quality

In each study, we performed LD pruning to obtain a set of near-independent biallelic SNVs on chromosome 20, retaining variants with pairwise $r^2 < 0.2$ within a sliding 50kb window with a 5 variant step size with PLINK v2.0.^{73–75} For each retained variant, we defined five aggregate measures of genomic features over 10kb windows centered at the variant: mean GC content, number of repeats, number of structural variants, presence of ≥ 1 segmental duplication, and mean recombination rate. We defined the linear distance of the variant from the nearest array-genotyped variant. For each of the six genomic features, we performed a logistic regression to test the association between dichotomous imputation quality (observed imputation $r^2 > 0.8$ vs. ≤ 0.8) and the feature, adjusting for variant MAF as a categorical variable with 9 bins and breaks at 0, 0.0003, 0.0006, 0.0009, 0.001, 0.0032, 0.01, 0.032, 0.1, and 0.5. We also performed zero-one inflated beta regression to test the association between the continuous observed imputation r^2 and each feature with the same MAF adjustment. Zero-one inflated beta regression models the

association of the genomic features with the observed imputation r^2 in the open interval $0 < r^2 < 1$ (mean μ and variance σ^2) and the probabilities of observed imputation $r^2=0$ (ν) and observed imputation $r^2=1$ (τ) in a piecewise manner.⁷⁶ In both regression models, we centered and scaled continuous and count predictors for comparability.

2.2.7 Effect of real vs. WGS-based array genotypes on evaluation of imputation quality

To determine if the WGS-based imputation results were consistent with the genotype array-based-results, we imputed the real OmniExpress array with each of the three reference panels in METSIM. For each reference panel, we compared the observed imputation r^2 and genotype concordance metrics for the real array-based genotype imputation to WGS-based array genotype imputation results (from above).

2.2.8 Effect of variant caller on evaluation of imputation quality

Variants in TOPMed and the four study datasets were called with the TOPMed Variant Calling/ GotCloud pipeline.^{20,63} To assess the impact of variant calling tool, we recalled METSIM WGS variants using GATK version 3.5.⁷⁷ In the comparison, we excluded variants from the GATK callset deviating from Hardy-Weinberg equilibrium ($p < 10^{-6}$), with $>2\%$ missingness, or with allelic imbalance <0.3 or >0.7 . We also excluded variants in regions of low complexity, centromeres, segmental duplications, or satellite regions.⁷⁸ After filtering, 21.8M variants remained in the subset of 2,429 individuals used for imputation analysis in METSIM. We then created each of the four WGS-based arrays using both METSIM callsets and evaluated imputation performance by comparing the imputed variants to the respective sequenced variants.

2.2.9 Imputability tool for the Michigan Imputation Server

We developed RsqBrowser, a tool that allows researchers to query for the observed imputation r^2 for variants or regions of interest. Users specify the genomic position in build GRCh38 and select the genotype array, imputation reference panel, and sample ancestry. RsqBrowser returns a table with the position and observed imputation r^2 for all variants in the specified regions or genes. We have deployed this tool on the Michigan Imputation Server.

2.3 Results

2.3.1 Whole genome sequencing studies of four ancestries

We used WGS data in four studies as gold standard genotypes. These four represent three major US populations: African, Hispanic/Latino, and European ancestry, and a population isolate: Finnish ancestry from Finland (Table 2.1). We observed that our primary metric of imputation quality, the observed imputation r^2 , was upwardly biased in small samples for low-frequency and rare variants (Supplementary Figure 2.1). To avoid any biases comparing across datasets of different sample sizes, we randomly downsampled the African, Hispanic/Latino, and Finnish ancestry datasets to 2,429 individuals to match the smaller European ancestry dataset. After all sample- and variant-level filtering, we included in our analysis 79.3M, 68.8M, 62.2M, and 22.1M variants in the African, Hispanic/Latino, European, and Finnish ancestry studies, respectively (Table 2.1). In each study, >91% of these variants were biallelic SNVs. The others were multiallelic SNVs (0.3-1.6%), and biallelic (6.5-6.6%) and multiallelic indels (0.005-0.2%).

2.3.2 Impact of reference panel on genotype imputation quality

For each WGS study participant, we subsetting WGS genotypes to those present on the Illumina Core (0.3M markers), OmniExpress (0.7M), Multi-Ethnic Genotyping (MEGA) (1.8M), and Omni 2.5M (2.4M) arrays. We then carried out genotype imputation on these genotype array

subsets using the 1000G and HRC imputation reference panels, as well as a modified TOPMed panel. Because the Hispanic/Latino and European WGS datasets were included in the TOPMed panel, we restricted the TOPMed panel to a subset of 88,804 reference samples that did not overlap our WGS datasets for all analyses. To measure the imputation quality of each sequenced variant in the study, we calculated the squared Pearson correlation between sequenced genotypes and imputed genotype dosages (observed imputation r^2). We consider array genotyping followed by imputation to approximate WGS for the MAF bins for which >90% of variants are well-imputed (observed imputation $r^2 > 0.8$).

As expected, across all combinations of reference panels, ancestries, and MAF, the densest genotype array (Omni 2.5M) had both the highest mean observed imputation r^2 and highest number and proportion of well-imputed variants (Figure 2.1, Supplementary Figure 2.2, Supplementary Table 2.2, Supplementary Table 2.3). For the Omni 2.5M array and in all ancestries, TOPMed-based imputation approximated WGS for variants of lower MAF compared to the HRC or 1000G panels. TOPMED-based panel imputation approximated WGS at lower MAF thresholds in African and Hispanic/Latino ancestry (0.14 and 0.11%) than in European or Finnish ancestry (0.35 and 0.85%) (Figure 2.1A-B, Supplementary Table 2.4). Between the previously available 1000G and HRC panels, imputation quality was higher with 1000G for African and Hispanic/Latino ancestry studies and with HRC for European and Finnish ancestry studies. With the Omni 2.5M array, the fold change in variant MAF for which TOPMed-based imputation approximated WGS compared to the next best performing panel was highest in African and Hispanic/Latino ancestry studies (17.0X and 21.0X, comparing to 1000G) and lower in European and Finnish ancestry studies (8.0X and 1.4X, comparing to HRC). Although we used subsets of sequenced variants instead of actual genotyping arrays, we saw minimal

difference in results for the Finnish study for which we had Illumina OmniExpress array data (Supplementary Figure 2.3). Results for the Finnish study were also consistent using genotypes from a second variant caller (Supplementary Figure 2.4). These results show that with the TOPMed panel, it is possible for array genotyping and imputation to approximate WGS at a population level for common and low-frequency variants in these three US-based studies.

2.3.3 Less influence of genotype array size with TOPMed- compared to 1000G and HRC-based imputation

For all four ancestries and all three imputation reference panels, imputation quality increased with larger array size (Figure 2.1C). However, the difference in TOPMed imputation quality among the Omni 2.5M, MEGA, and OmniExpress arrays was minimal. For example, in the African ancestry study, TOPMed imputation approximated WGS for variants with $MAF \geq 0.14\%$ with the Omni 2.5M array, $\geq 0.17\%$ with the MEGA array, and $\geq 0.24\%$ with the OmniExpress array (Supplementary Table 2.4). This threshold was higher with the smaller Core array ($\geq 0.84\%$). In contrast, genotype array size had a larger effect on imputation quality with the HRC and 1000G panels in African and Hispanic/Latino ancestry studies. For African ancestry, 1000G imputation approximated WGS at a much lower MAF with the Omni 2.5M array ($\geq 2.5\%$) compared to the OmniExpress array ($\geq 14.0\%$). And with the HRC panel, imputation with the OmniExpress array could not approximate WGS at any MAF in African ancestry.

2.3.4 Individual-level imputation accuracy varies with finer-scale ancestry

Because imputation quality depends on the shared ancestry between reference panel and sample haplotypes,¹⁷ we hypothesized that imputation quality within the four WGS studies would vary with finer-scale ancestry. To measure individual-level imputation quality, we

calculated concordance rates between heterozygous sequenced and imputed genotypes separately for study-specific rare, low-frequency, and common biallelic SNVs in each individual. As expected, concordance rates varied across individuals more for rare variants than for low-frequency and common variants (TOPMed: Figure 2.2, Supplementary Figure 2.5, Supplementary Figure 2.6, all panels: Supplementary Table 2.5). With the TOPMed panel, mean heterozygous concordance rates for rare variants were higher in individuals of African and Hispanic/Latino ancestry (0.93 in both) compared to individuals of European and Finnish ancestry (0.86 and 0.82) (Figure 2.2A). Concordance rates varied most within Hispanic/Latino individuals (10th-90th percentile: 0.80-0.98).

We next stratified African and Hispanic/Latino study participants by finer-scale measures of ancestry. The African American population in the United States is primarily of African and European ancestries.⁷⁹ We therefore estimated the proportion of African ancestry for each individual in the African ancestry study assuming two populations, which ranged from 0.26 to 1.00 (mean 0.82). For the 2,307 individuals with an estimated African ancestry <0.95, individuals with higher proportions of African ancestry had higher genotype concordance rates with the TOPMed panel (Figure 2.2B). For instance, concordance rates for those with an estimated proportion between 0.86-0.95 were higher (mean 0.93) than for those between 0.26-0.35 (mean 0.89). In contrast, concordance rates for the 122 individuals with estimated proportion of African ancestry >0.95 were lower (mean 0.91) than for individuals with smaller estimated proportions of African ancestry.

Hispanic/Latino populations in the United States are admixed with primarily European, Native American, and African ancestry, with individuals of Caribbean origin usually having more African ancestry.⁸⁰ The concordance rates for individuals from Caribbean populations were

higher (mean 0.96) compared to those from non-Caribbean (mean 0.79) populations with the TOPMed panel (Figure 2.2C).

We also estimated finer-scale ancestry in all four studies with principal component analysis (PCA), projecting the study individuals onto 938 reference samples from the Human Genome Diversity Project (HGDP).⁶⁹ The first two PCs reflect clines of European (high PC2), African (high PC1 and low PC2), and Native American (low PC1 and low PC2) ancestries (**Error! Reference source not found.**). To see how TOPMed imputation quality varied with fine-scale ancestry, we divided individuals into concordance rate quintiles calculated jointly across all four studies. In all studies, individuals clustering closer to HGDP individuals of African ancestry were more likely to be in higher concordance rate quintiles, while those clustering closer to Native American or European populations were more likely to be in lower concordance rate quintiles for rare (**Error! Reference source not found.D**) and low-frequency variants (Supplementary Figure 2.5). As expected, there was little variability in common variant imputation quality (Supplementary Figure 2.6).

Taken together, these results demonstrate that TOPMed imputation quality varies across individuals with finer-scale ancestry. Among the populations studied here, population subsets with large proportions of African ancestry, including Hispanic/Latino ancestry individuals of Caribbean origin, were on average the most accurately imputed for rare variants. However, individuals with the greatest proportions of African ancestry in the African study were not the most accurately imputed. The heterozygote concordance rates from HRC and 1000G imputation also varied with finer-scale ancestry for rare variants (Supplementary Table 2.5).

2.3.5 Imputation quality varies across the genome

Sequence quality and genotype array density are not uniform across the genome. Because these factors influence imputation, we sought to quantify the regional variability in imputation quality. We first visualized the observed imputation r^2 for common variants (MAF>5%) across the chromosomes. Although the vast majority (>99.6%) of common variants are well-imputed (observed imputation r^2 >0.8) in all four ancestries with the Omni 2.5M array and TOPMed reference panel (Supplementary Table 2.3), we identified clusters of common variants that were not well-imputed at the same genomic positions across ancestries and genotype arrays (Figure 2.3A, Supplementary Figure 2.8). There are likewise regions with better-than-average imputation quality, including the *HLA* region on chromosome 6 (Supplementary Figure 2.8), that is characterized by high LD and dense genotype array coverage.^{51,81}

To assess regional variability in imputation quality, we calculated the lengths of runs of consecutively well-imputed variants separately for rare, low-frequency, and common biallelic SNVs across the genome (Figure 2.3B). We identified a large variability in the number of consecutively well-imputed common and low-frequency variants with the TOPMed panel (e.g. IQR in African ancestry is 41-750 common variants (10.4-253.2kb) and 9-287 low-frequency variants (2.3-84.3kb) with the Omni 2.5M array (Supplementary Table 2.6, Supplementary Table 2.7). As expected, the lengths of consecutive well-imputed rare variants were much shorter, with a maximum length of 34-45 variants depending on ancestry (Supplementary Table 2.6).

2.3.6 Local genomic features explain little variability in imputation quality

Genomic features including high GC content and the presence of large duplications or repeats have been associated with regions of poor imputation quality in Europeans using the 1000G panel.⁴⁹ To test the effects of genomic features on imputation quality with the TOPMed panel, we performed logistic regressions in each of our four studies with the imputed quality

status (observed imputation $r^2 > 0.8$) as the dichotomous outcome for independent variants on chromosome 20. In separate models, we tested the associations of distance to the nearest genotyped variant and the following features aggregated over a 10kb window centered at the variant: mean GC content, mean recombination rate, number of repeats, number of structural variants, and presence of ≥ 1 segmental duplication, adjusting for bins of MAF. We found that higher recombination rate, lower GC content, greater distance to genotype array variants, more structural variants, and the presence of segmental duplications were all associated with lower imputation quality (Figure 2.4A). The effect of nearby repeats was not consistent across ancestries or repeat class, although nearby simple repeats were associated with worse imputation quality in all ancestries (Supplementary Figure 2.9). However, none of the tested genomic features meaningfully impacted the proportion of variability in imputation quality beyond variant MAF (Figure 2.4B). Results were similar when modeling imputation quality as a continuous variable (Supplementary Figure 2.10) and were consistent across reference panels.

2.3.7 Impact of variant predicted function and type on imputation quality

Protein-coding variants are often of high clinical significance and easier to interpret compared to non-coding variants; they are also more likely to be rare and more difficult to impute.^{20,82} To determine the extent to which variants that impact protein coding are well-imputed, we classified sequenced biallelic SNVs by predicted impact on protein coding. With the TOPMed panel and Omni 2.5M array, we found that 50.7-66.7% of variants predicted to have high or moderate impact on protein coding were well-imputed (Supplementary Table 2.8). We found no meaningful difference in imputation quality between the protein coding classes when controlling for MAF (Supplementary Figure 2.11).

Multiallelic SNVs and all indels have been shown to have lower imputation quality than biallelic SNVs with the 1000G panel,^{18,49} and indels are absent from the HRC panel.¹⁹ To quantify the effect of variant type on imputation quality, we calculated the proportion of well-imputed indels, multiallelic SNVs, and multiallelic indels using the TOPMed panel and all four genotype arrays. We observed very similar MAF thresholds for which imputation could approximate WGS among biallelic SNVs, biallelic indels, and multiallelic SNVs (Supplementary Figure 2.12). Multiallelic indels were less well-imputed. For example, in African ancestry, TOPMed imputation with the Omni 2.5M array approximated WGS at similar MAF thresholds for biallelic SNVs and indels and multiallelic SNVs (0.14%, 0.24%, 0.16% respectively) compared to 0.55% for multiallelic indels (Supplementary Table 2.9).

2.4 Discussion

Here, we used deep WGS from studies of African, Hispanic/Latino, European, and Finnish ancestry to quantify the extent to which array genotyping followed by genotype imputation can approximate WGS. We performed imputation using genotypes present on the Illumina Core, OmniExpress, MEGA, and Omni 2.5M arrays with the 1000G, HRC, and TOPMed reference panels. We found that with the largest array (Omni 2.5M) and largest reference panel (TOPMed), array genotyping followed by imputation can approximate WGS at a population level for variants with $MAF \geq 0.14\%$ in African ancestry, $\geq 0.11\%$ in Hispanic/Latino ancestry, $\geq 0.35\%$ in European ancestry, and $\geq 0.84\%$ in Finnish ancestry. Particularly for the African and Hispanic/Latino ancestry studies, TOPMed imputation approximated WGS at much lower MAF than HRC or 1000G imputation, which is consistent with previous analyses showing improvements in these populations even with a smaller version of the TOPMed panel.⁵⁰ For analyses primarily investigating the genetic effects of common and low-frequency variants, such

as single-variant GWAS, in any of the four populations, array genotyping and imputation is sufficient to accurately capture genetic variants and given differences in cost allows for much larger sample sizes than WGS. Large proportions (~44-60%) of rare variants with MAF even lower than the reported thresholds were also well-imputed, highlighting the potential for well-powered rare-variant studies without WGS, although not all rare variants can be reliably imputed. Because we restricted the TOPMed panel to reference samples that did not overlap the WGS datasets, we expect that imputation quality with the full TOPMed-R2 panel to be even better than reported here. In particular, we would expect higher imputation quality for Hispanic/Latino studies as a large proportion of the Hispanic/Latino individuals in the TOPMed-R2 panel were excluded here.

Because genotype array size has been shown to be positively associated with imputation quality with the 1000G and HRC panels, we examined the impact of genotype arrays on the extent to which array genotyping and imputation can approximate WGS. As expected, imputation quality was higher when using larger arrays. However, the effect of genotype array choice on TOPMed imputation was much smaller than on HRC or 1000G imputation. The difference between the Omni 2.5M, MEGA, and OmniExpress arrays was minimal, suggesting that researchers imputing with the TOPMed panel in these populations may opt for the less expensive OmniExpress array with little loss of information. However, we did find lower imputation quality using the smaller Core array (~307k variants) and might expect even lower quality for arrays with fewer markers.

WGS is also used for clinical purposes including diagnosis, screening, and identifying therapeutic targets.⁴⁴ Variants predicted to alter protein function are often of high clinical significance.⁸² In the populations studied here, we found that only 50.7-66.7% of biallelic SNVs

with moderate or high predicted impact on protein coding were well-imputed, as might be expected given the generally low MAF of these variants. To quantify individual imputation quality in contrast to population-level imputation quality, we calculated the heterozygous concordance rates between sequenced biallelic SNVs and imputed best-guess genotypes. For all three reference panels, we found that the concordance rates for rare and low-frequency variants varied widely among individuals in the African, Hispanic/Latino, and European ancestry studies and were associated with finer-scale ancestry. Because of this variability and the large proportion of rare variants that are not accurately imputed with the available imputation reference panels, we believe that WGS cannot currently be reliably approximated in clinical settings with array genotyping and imputation.

Despite large numbers of African and Hispanic/Latino haplotypes in the TOPMed reference panel, more than half of the TOPMed haplotypes are European. Still, we found that TOPMed imputation quality was highest for the African and Hispanic/Latino ancestry studies and for individuals with large proportions of African ancestry among the populations studied here. A first possible explanation is that there are proportionally more rare variants in non-African populations that have undergone recent bottlenecks and subsequent population growth, as is true in the three US populations studied here (**Error! Reference source not found.**). In these populations, it can be more difficult to identify the haplotype background of the rare variation.^{18,83} A second possible explanation is that individuals with large proportions of African ancestry in these studies match more closely by chance with a subset of TOPMed haplotypes than do the individuals with large proportions of Native American or European ancestry. However, relatively higher imputation quality with the TOPMed panel in samples of African and Hispanic/Latino ancestry compared to European ancestry was previously reported in a separate

set of samples.²⁰ Third, admixture could impact the accuracy of haplotype phasing of the sample or reference haplotypes. Taken together, these results emphasize the importance of ancestrally diverse reference panels like TOPMed and suggest that reference panel composition is not the only factor explaining ancestry differences in imputation quality.

While nearly all common and low-frequency variants are well-imputed with the TOPMed panel in the populations studied, there was substantial variability by genomic region in imputation across the MAF spectrum. Some regions, such as the densely genotyped *HLA* locus, had higher imputation quality than what would be expected based on variant MAF alone. We found that lower recombination rate and higher GC content around a variant were associated with higher imputation quality in all four studies, but that none of these features except MAF explained a substantial proportion of variability in imputation quality. Given the difficulty of predicting hard-to-impute regions/variants, we developed RsqBrowser, a tool that allows researchers to query empirical imputation quality for specific variants or genomic regions of interest by ancestry, which is available on the Michigan Imputation Server.

The results presented here are limited by the use of high quality but imperfect WGS as a gold standard. We did not consider any variants that were imputed from the reference panels but not detected in the WGS. We also note that the results presented here cannot necessarily be extended to other populations or population isolates, particularly those such as East and South Asian populations, that are not represented or represented in smaller numbers in the TOPMed panel. Furthermore, we only used WGS from one study for each population that we analyzed. For some ancestries, particularly population isolates, other population-specific reference panels may perform better than the three commonly used imputation panels analyzed here.

While array genotyping and imputation cannot fully replace deep WGS, we found that it can approximate WGS for variants down to specific MAF thresholds depending on genotype array and reference panel choices as well as sample ancestry. Researchers' decision to invest in one technology over another will depend on these criteria, genomic location, and the MAF of variants relevant to their research questions.

2.5 Figures and Tables

Study	Ancestry	Mean depth	Sample size		Number of variants in 2,429 samples used in analyses				
			Total	Unrelated	Bi-allelic		Multi-allelic		Total
					SNV	Indel	SNV	Indel	
InPSYght	African	27	7,717	7,169	72.6M	1.3M	5.3M	0.2M	79.3M
BioMe	Hispanic/Latino	37	4,677	3,141	63.2M	0.9M	4.5M	0.1M	68.8M
MLOF	European	39	2,987	2,429	57.3M	0.8M	4.1M	0.1M	62.2M
METSIM	Finnish	24	3,045	2,703	20.5M	0.1M	1.4M	10K	22.0M

Table 2.1 Whole-genome sequencing (WGS) datasets

The study name, ancestry, mean sequencing depth, sample size (total and unrelated subset), and number of variants, including single-nucleotide variants (SNVs) and indels, for the four WGS datasets.

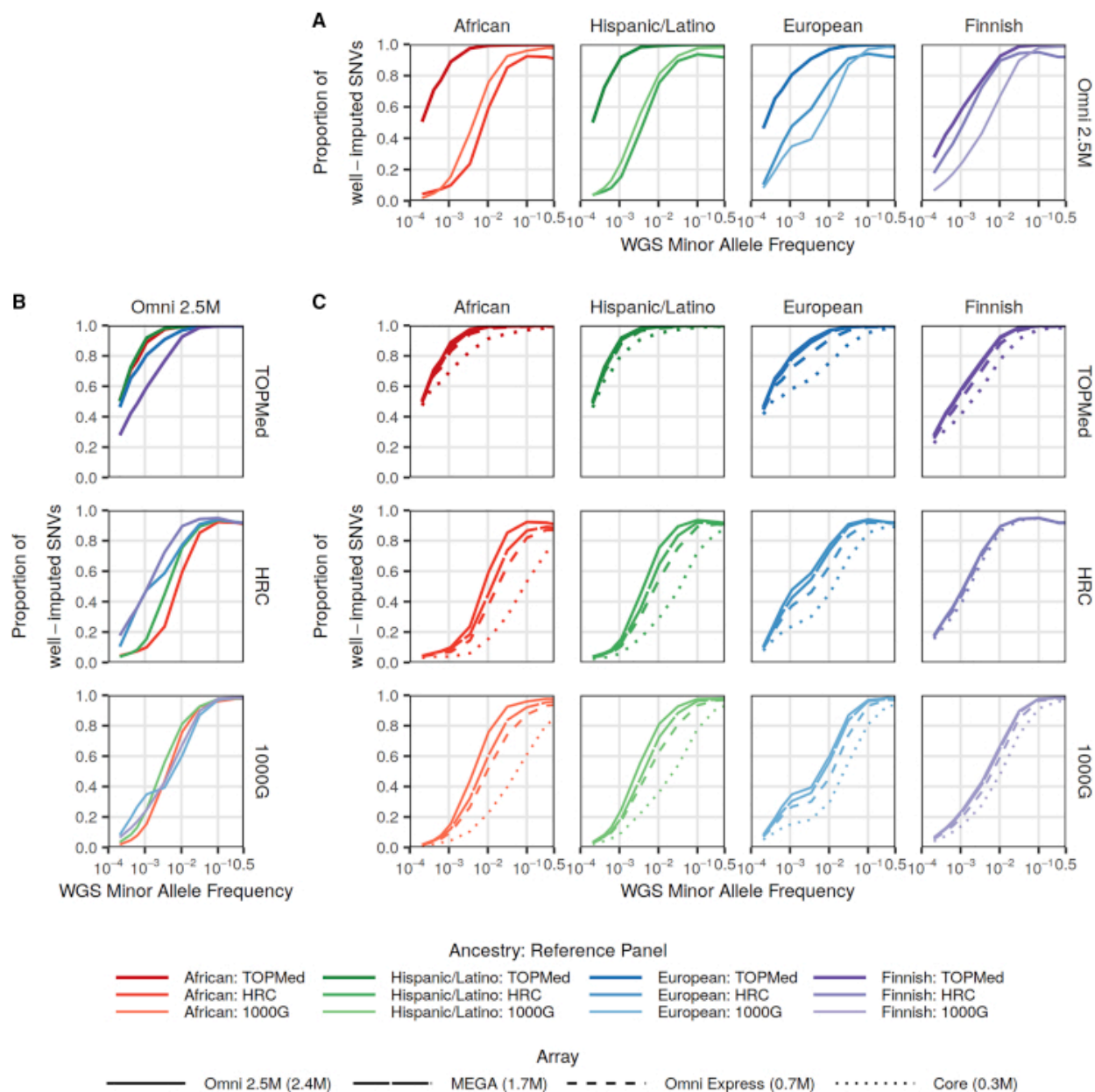


Figure 2.1 Proportion of well-imputed ($r^2 > 0.8$) bi-allelic SNVs by reference panel, study ancestry, and genotyping array

The proportion of sequenced variants that are well-imputed ($r^2 > 0.8$) with the TOPMed, HRC, and 1000G imputation reference panels. (A) Comparison across the reference panels using the Illumina Omni 2.5M array. (B) Comparison across the four studies using the Illumina Omni 2.5M array. (C) Comparison across four Illumina genotyping arrays: Omni 2.5M, MEGA, Omni Express, and Core by ancestry (columns) and imputation reference panels (rows). In all plots, the x axes show minor-allele frequency (MAF) calculated separately by study. Sequenced bi-allelic SNVs not present in reference panels were assigned $r^2 = 0$. Bi-allelic SNVs were then aggregated by MAF bins of width 0.00025 MAF for MAF between 0.0002 and 0.002 and of size 0.001 MAF for MAF > 0.002 ; those plotted here correspond to singletons, doubletons, and tripletons in each study, as well as those with mean MAF closest to the values 0.001, 0.0032, 0.01, 0.032, 0.1, 0.32, and 0.5.

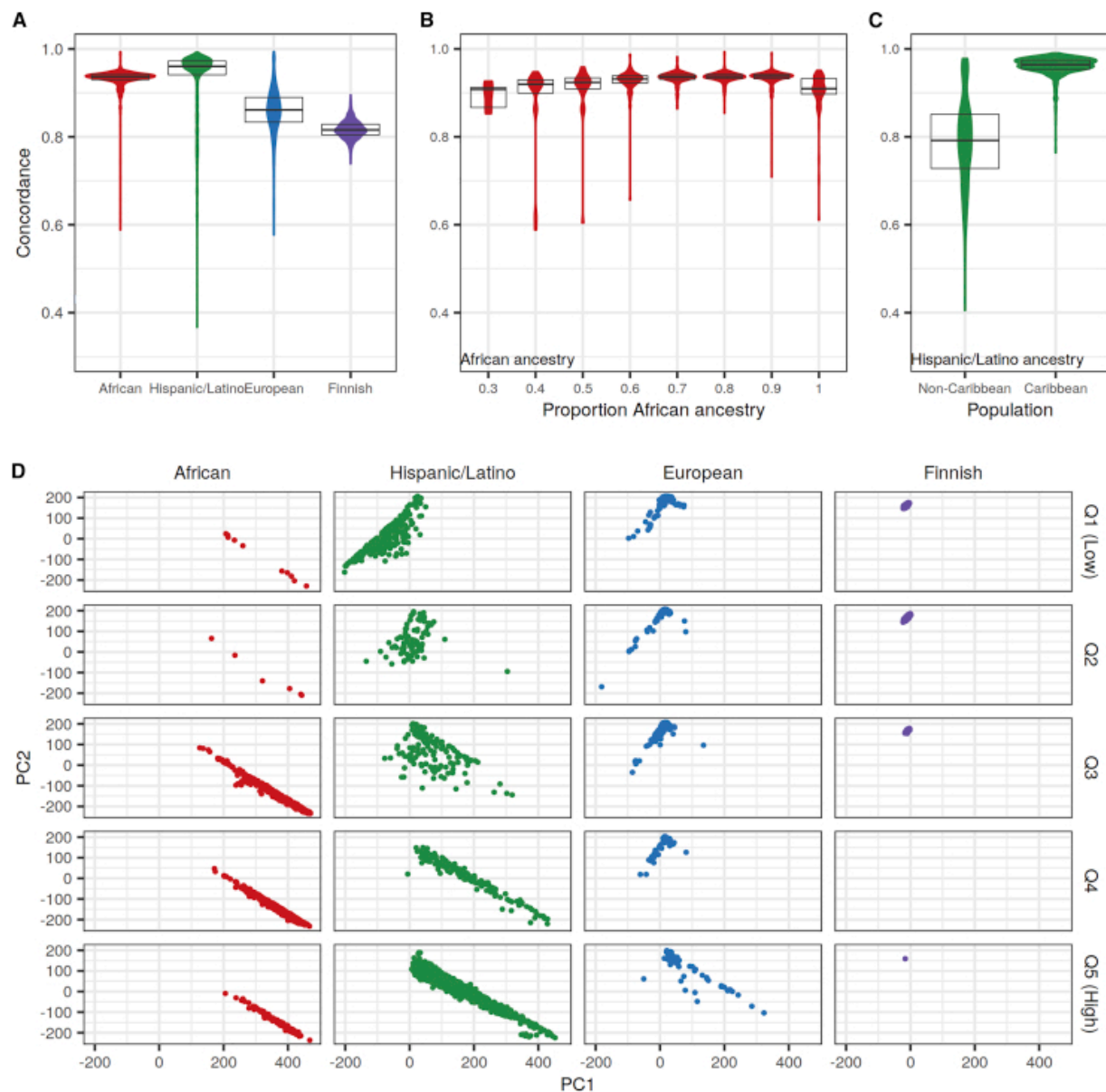


Figure 2.2 Heterozygous genotype concordance rates for rare variants by ancestry with TOPMed panel imputation

Heterozygous concordance rates were calculated between sequenced and TOPMed-imputed genotypes for rare ($MAF < 0.5\%$, calculated separately in each study) bi-allelic SNVs with the Omni 2.5M array. (A) Distribution of concordance rates in each of the four studies. Boxplots correspond to 25th, 50th, and 75th percentiles. (B) Distribution of concordance rates by bins of estimated proportion of African ancestry in the admixed African study. (C) Distribution of concordance rates in Caribbean and non-Caribbean populations in the Hispanic/Latino study. (D) Principal-component analysis (PCA) by genotype concordance quintile and ancestry. PCA was performed by projecting onto the Human Genome Diversity Project reference samples. Genotype concordance quintiles were calculated across all four studies and correspond to concordance rates of 0.37–0.82 (Q1), 0.82–0.86 (Q2), 0.86–0.93 (Q3), 0.93–0.95 (Q4), and 0.95–0.99 (Q5). Points are colored by ancestry.

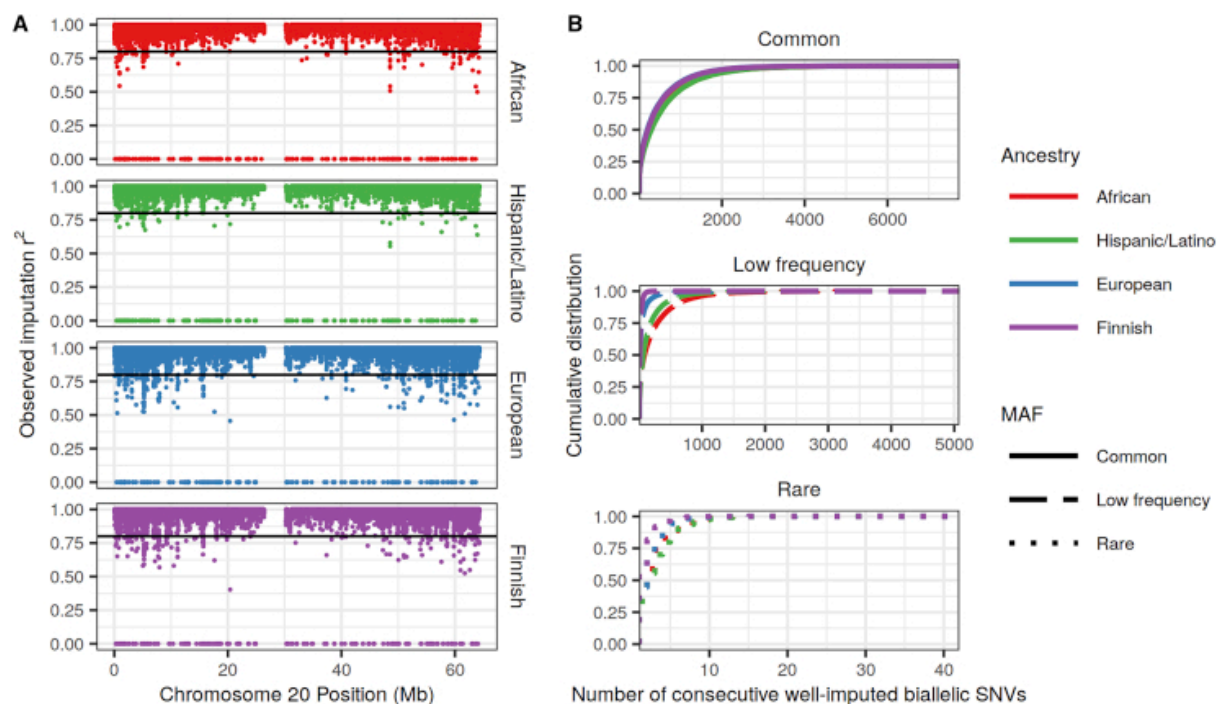


Figure 2.3 Regional variability in TOPMed reference panel imputation quality

(A) Observed imputation r^2 by genomic position (Mb) of common ($MAF > 0.05$) bi-allelic SNVs on chromosome 20. Sequenced bi-allelic SNVs not present in reference panels were assigned $r^2 = 0$. The horizontal line at $r^2 = 0.8$ corresponds to the threshold used to determine well-imputed variants. (B) Cumulative distribution of the number of consecutively well-imputed ($r^2 > 0.8$) bi-allelic SNVs in each MAF category: common ($MAF \geq 0.05$), low frequency ($0.005 \leq MAF < 0.05$), and rare ($MAF < 0.005$), as calculated separately in each study. For common variants, European and Finnish curves appear to overlap and African and Hispanic/Latino curves appear to overlap.

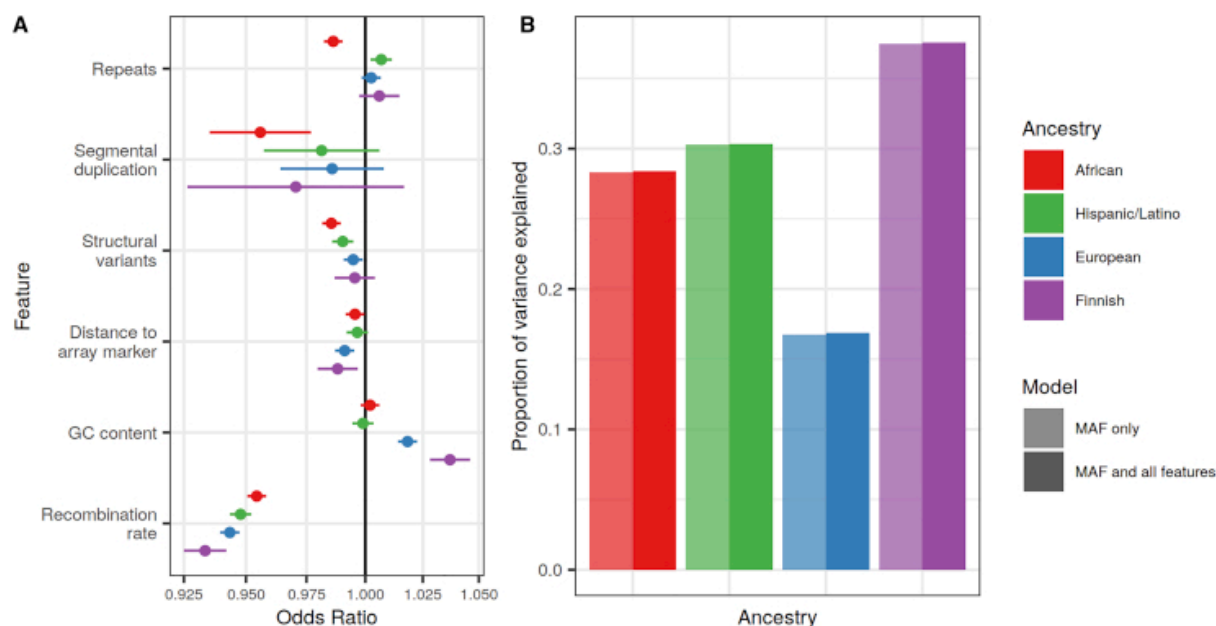


Figure 2.4 Genomic features associated with TOPMed imputation quality of bi-allelic SNVs by ancestry

(A) The odds ratios and corresponding unadjusted 95% confidence intervals from logistic regression models. Estimates are from separate models testing the associations between characteristics of regional genomic features and whether or not a variant is well imputed (observed imputation $r^2 > 0.8$), adjusting for variant MAF. (B) The proportion of variance explained (Nagelkerke R^2) for each logistic regression models with MAF only or with MAF and all six tested genomic features in one joint model.

2.6 Acknowledgements and publication

The results presented in this chapter have been peer-reviewed and published. Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Genome Sequencing for “NHLBI TOPMed: Whole Genome Sequencing in the BioMe Study” (phs phs001644) was performed at the Baylor College of Medicine Human Genome Sequencing Center and the McDonnell Genome Institute. Genome Sequencing for “NHLBI TOPMed: Whole Genome Sequencing in MLOF Study” (phs001515) was performed at the Baylor College of Medicine Human Genome Sequencing Center and New York Genome Center Genomics. Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were

provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Administrative Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

The Mount Sinai BioMe Biobank has been supported by The Andrea and Charles Bronfman Philanthropies and in part by Federal funds from the NHLBI and NHGRI (U01HG00638001; U01HG007417; X01HL134588). We thank all participants in the Mount Sinai Biobank. We also thank all our recruiters who have assisted and continue to assist in data collection and management and are grateful for the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

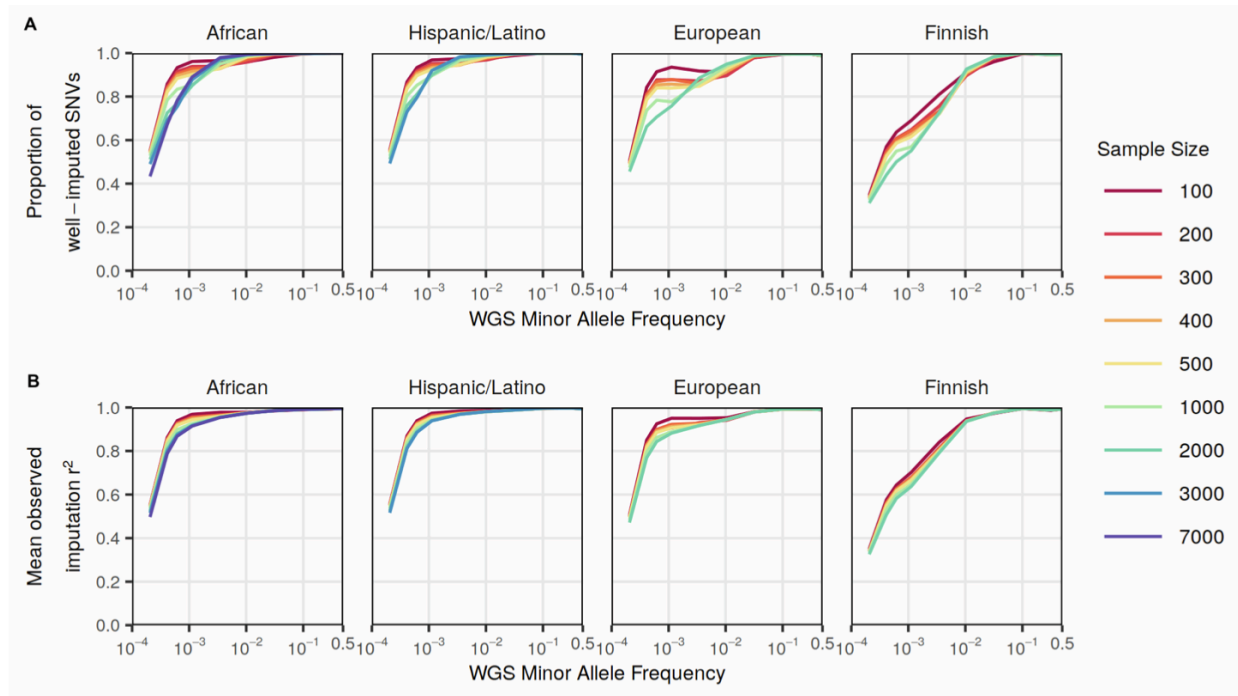
The My Life, Our Future samples and data are made possible through the partnership of Bloodworks Northwest, the American Thrombosis and Hemostasis Network, the National Hemophilia Foundation, and Bioverativ. We gratefully acknowledge the hemophilia treatment centers and their patients who provided biological samples and phenotypic data.

Bio-samples and data for the InPSYght study were obtained from NIMH Repository & Genomics Resource, a centralized national biorepository for genetic studies of psychiatric disorders. Contributing studies include the Genomic Psychiatry Cohort, the Consortium of the Genetics of Schizophrenia, the Bipolar Genome Study, the Lithium treatment moderate dose use study, and the Systematic Treatment Enhancement Program for Bipolar Disorder. We gratefully acknowledge the participants who provided biological samples and data for these studies.

We gratefully acknowledge the participants who provided biological samples and data for the METSIM study. We gratefully acknowledge Hyun Min Kang for his guidance in using beddiff to calculate genotype concordance metrics and other useful discussions. We gratefully acknowledge Corbin Quick for his script and guidance in calculating the observed imputation r^2 .

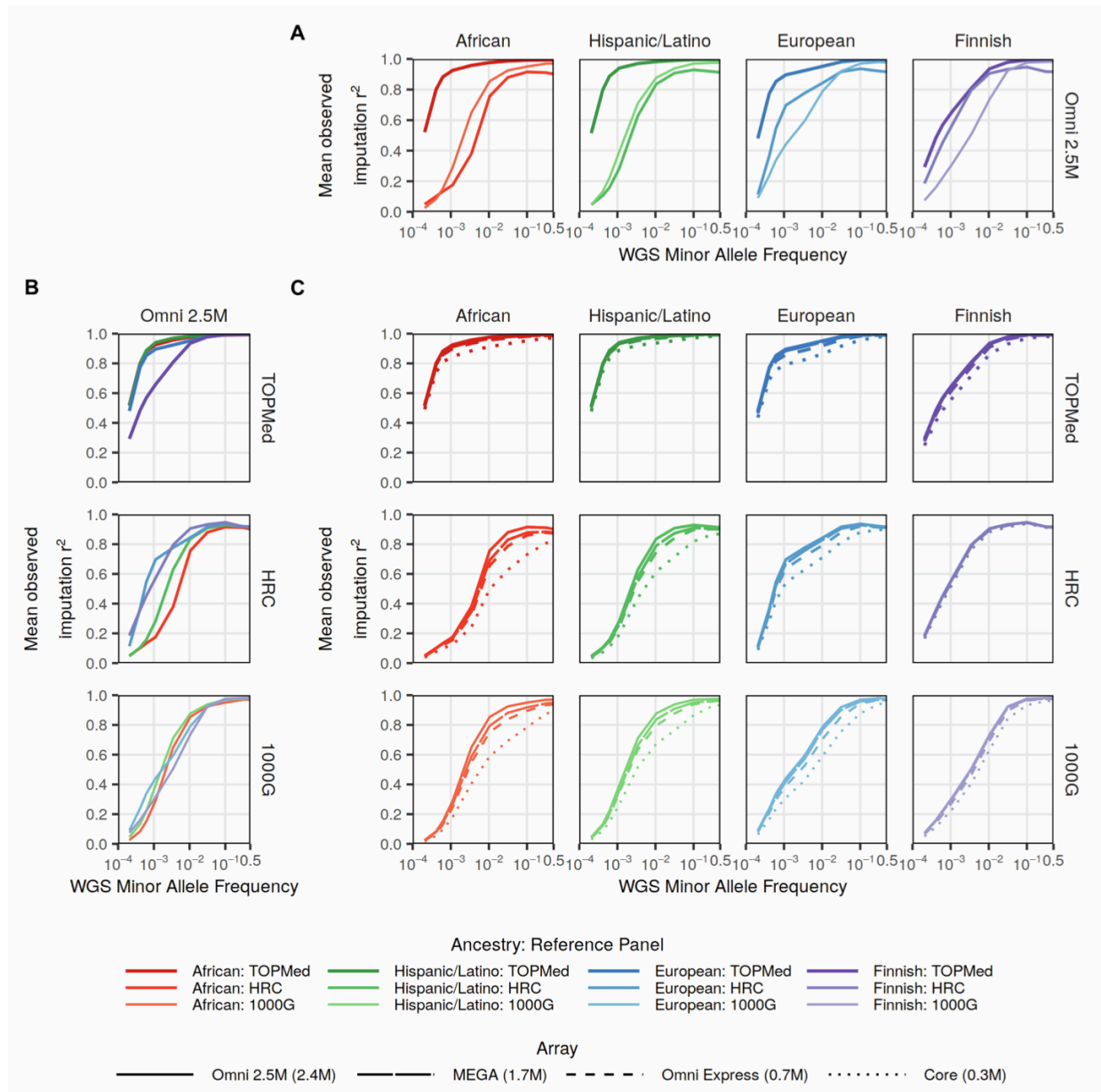
This work was supported by NHGRI grant R01 HG009976 (Boehnke). S.C.H was also supported by NHGRI grant F31 HG011186.

2.7 Supplementary Material



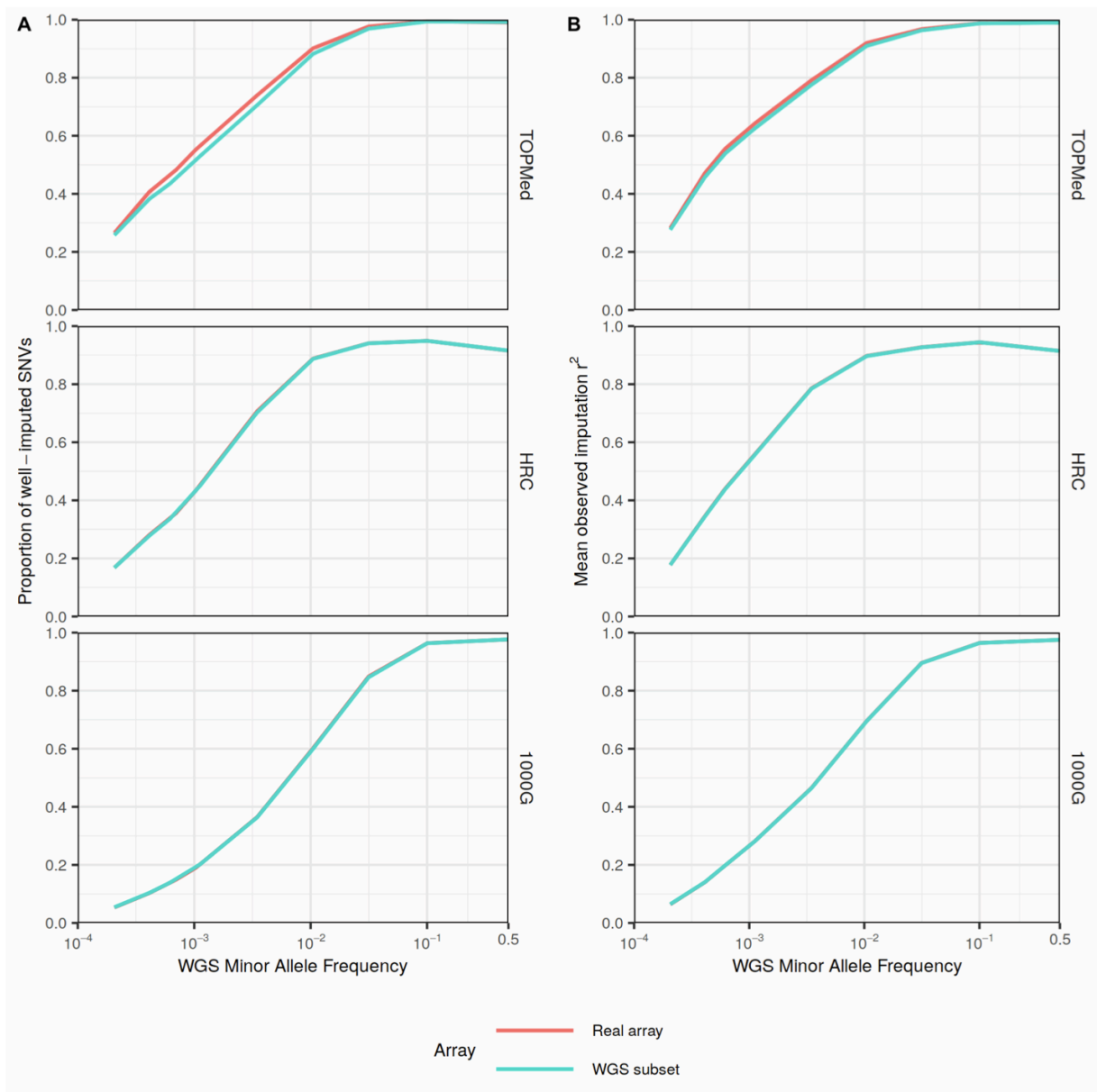
Supplementary Figure 2.1 Effect of sample size on imputation quality metrics

Random subsets of individuals were taken from each of the WGS studies as the total sample size of unrelated individuals allowed (up to 7,000 for African, 3,000 for Hispanic/Latino, and 2,000 for European and Finnish). Imputation was performed with the Omni 2.5M array and the TOPMed imputation reference panel. A. The proportion of sequenced biallelic SNVs that are well-imputed ($r^2 > 0.8$) by sample size. B. The mean r^2 by sample size. In both plots, the x-axes show minor allele frequency (MAF) calculated separately by study based on the 2,429 samples used in the main analyses. Sequenced biallelic SNVs not present in reference panels were assigned $r^2 = 0$. Biallelic SNVs were then aggregated by MAF bins of width 0.00025 MAF for MAF between 0.0002 and 0.002 and of size 0.001 MAF for MAF > 0.002 ; those plotted here correspond to singletons, doubletons, and tripletons in each study, as well as those with mean MAF closest to the values 0.001, 0.0032, 0.01, 0.032, 0.1, 0.32, and 0.5.



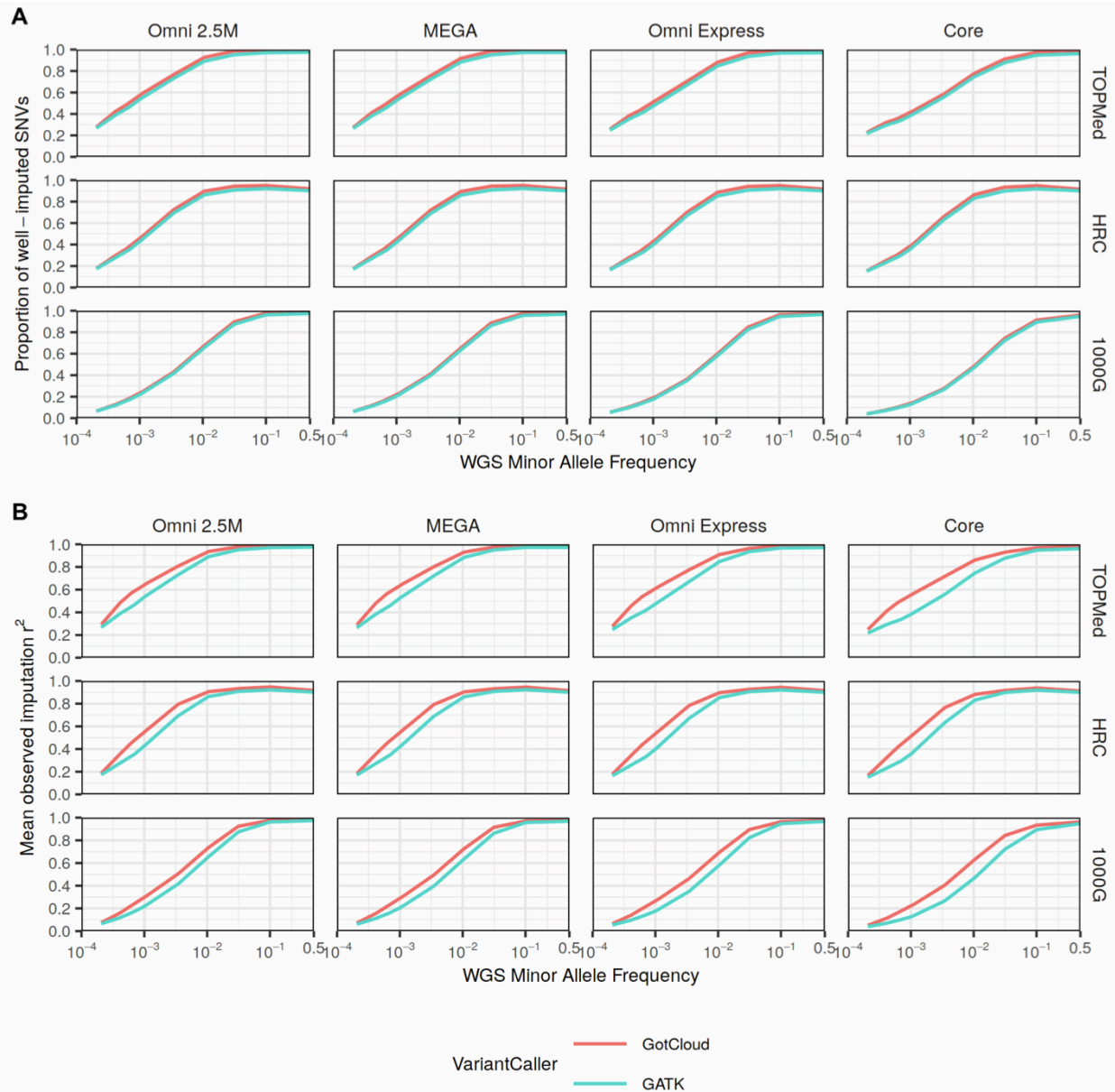
Supplementary Figure 2.2 Mean observed imputation r^2 of biallelic SNVs by reference panel, study ancestry, and genotyping array

The mean observed imputation r^2 with the TOPMed, HRC, and 1000G imputation reference panels. A. Comparison across the reference panels using the Illumina Omni 2.5M array. B. Comparison across the four studies using the Illumina Omni 2.5M array. C. Comparison across four Illumina genotyping arrays: Omni 2.5M, MEGA, Omni Express, and Core by ancestry (columns) and imputation reference panels (rows). In all plots, the x-axes show minor allele frequency (MAF) calculated separately by study. Sequenced biallelic SNVs not present in reference panels were assigned $r^2=0$. Biallelic SNVs were then aggregated by MAF bins of width 0.00025 MAF for MAF between 0.0002 and 0.002 and of size 0.001 MAF for MAF > 0.002; those plotted here correspond to singletons, doubletons, and tripletons in each study, as well as those with mean MAF closest to the values 0.001, 0.0032, 0.01, 0.032, 0.1, 0.32, and 0.5.



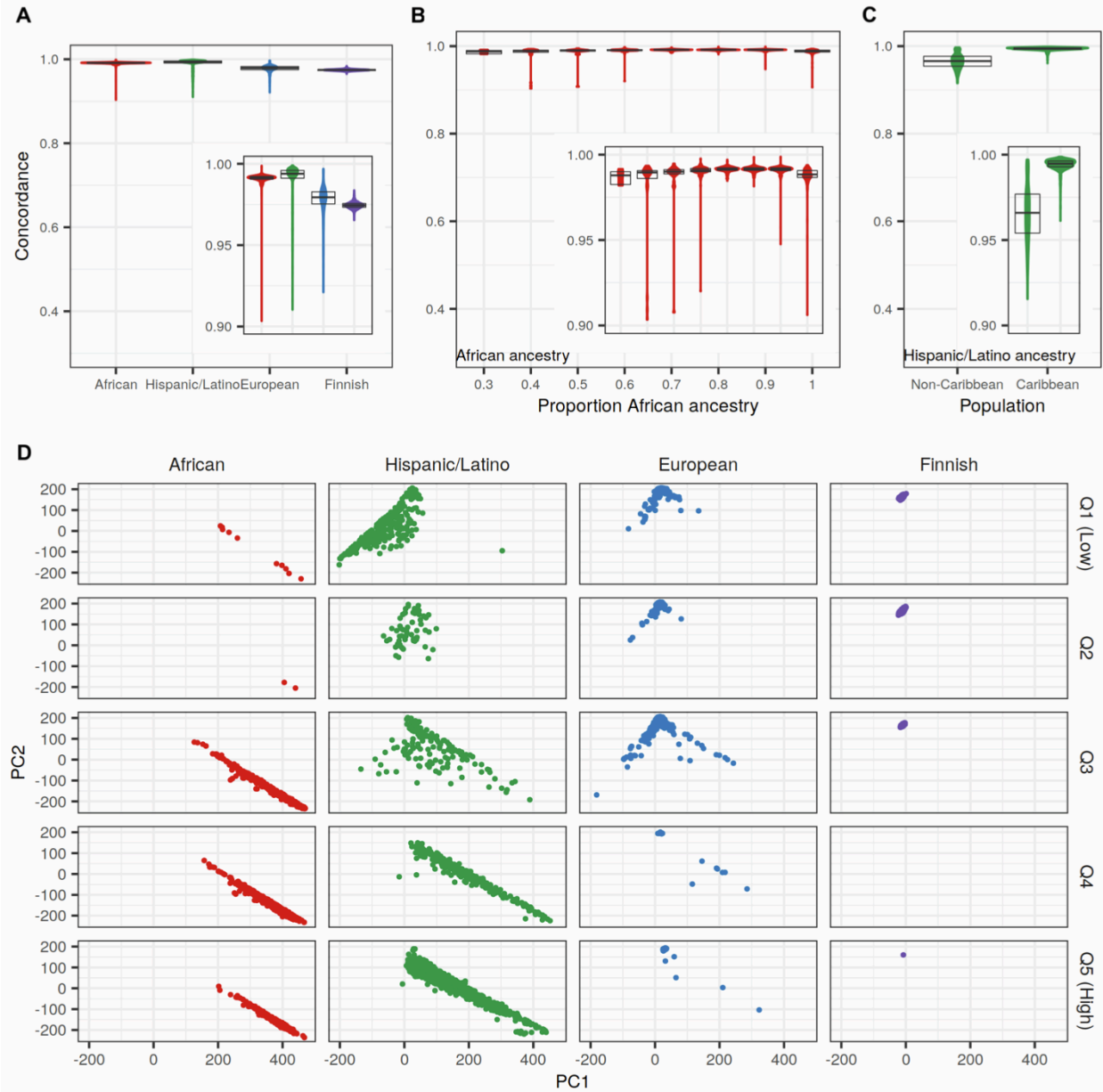
Supplementary Figure 2.3 Imputation quality of biallelic SNVs by reference panel using WGS- based and real Illumina OmniExpress arrays

A. The proportion of sequenced biallelic SNVs imputed from real array data (red line) or from WGS-based array (blue line) in the Finnish study that are well-imputed ($r^2 > 0.8$) by imputation reference panel. B. The mean observed imputation r^2 for the same variants. In all plots, the x-axes show minor allele frequency (MAF) calculated separately by study. Variants were aggregated by MAF bins of size 0.00025 MAF for MAF between 0.0002 and 0.002 and of size 0.001 MAF for MAF > 0.002 ; those plotted here correspond to singletons, doubletons, and tripletons in each study, as well as those with mean MAF closest to the values 0.001, 0.0032, 0.01, 0.316, 0.1, and 0.5. The lines appear entirely overlapping for the HRC and 1000G reference panels.



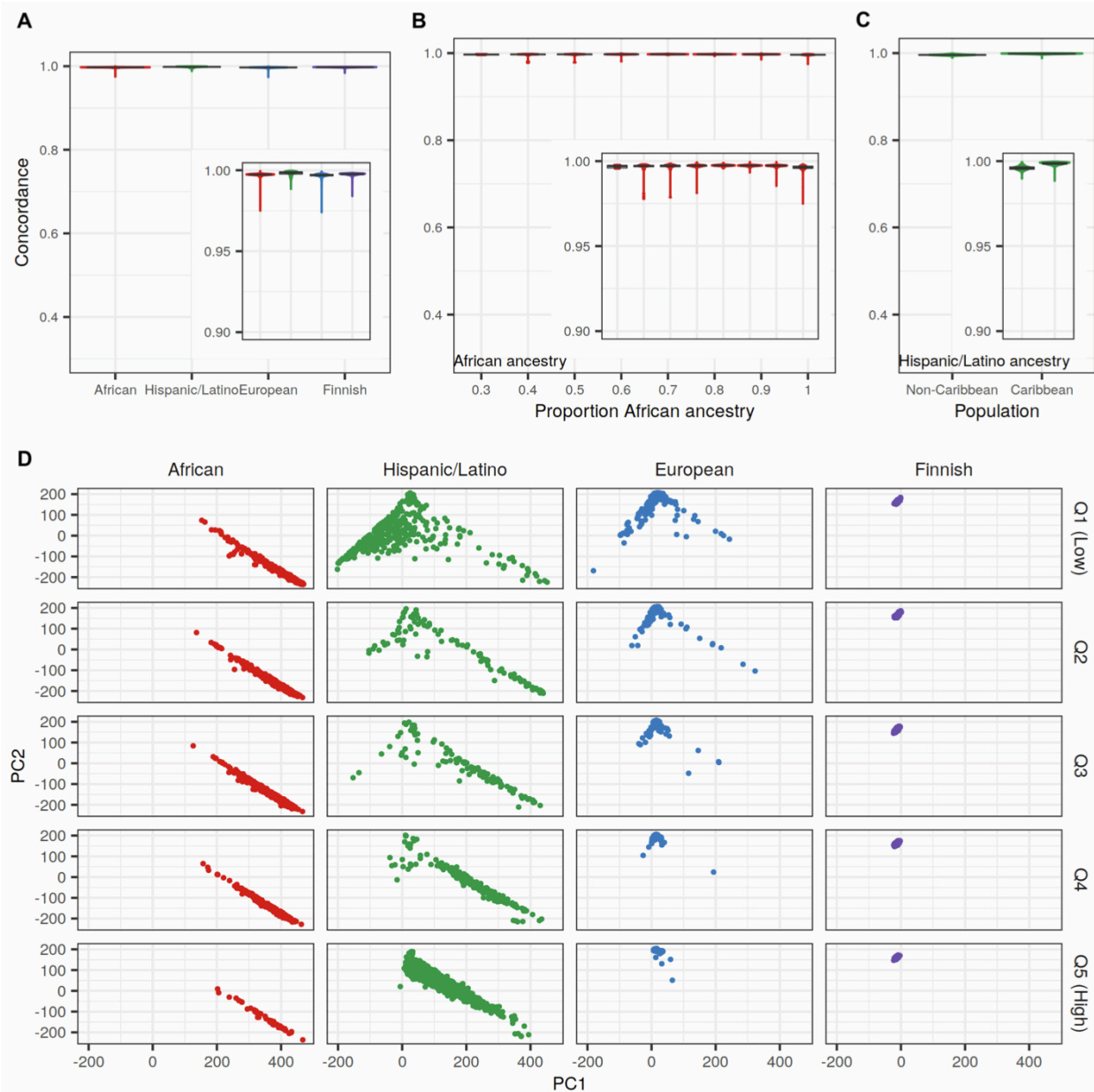
Supplementary Figure 2.4 Proportion of well-imputed ($r^2 > 0.8$) biallelic SNVs by reference panel, genotyping array, and variant caller in Finnish study

The proportion of sequenced biallelic SNVs called with the GotCloud pipeline (red line) or GATK pipeline (blue line) in the Finnish study that are well-imputed ($r^2 > 0.8$) by reference panel (rows) and genotyping array (columns). In all plots, the x-axes show minor allele frequency (MAF) calculated separately by study. Variants were aggregated by MAF bins of size 0.00025 MAF for MAF between 0.0002 and 0.002 and of size 0.001 MAF for MAF > 0.002 ; those plotted here correspond to singletons, doubletons, and tripletons in each study, as well as those with mean MAF closest to the values 0.001, 0.0032, 0.01, 0.316, 0.1, and 0.5.



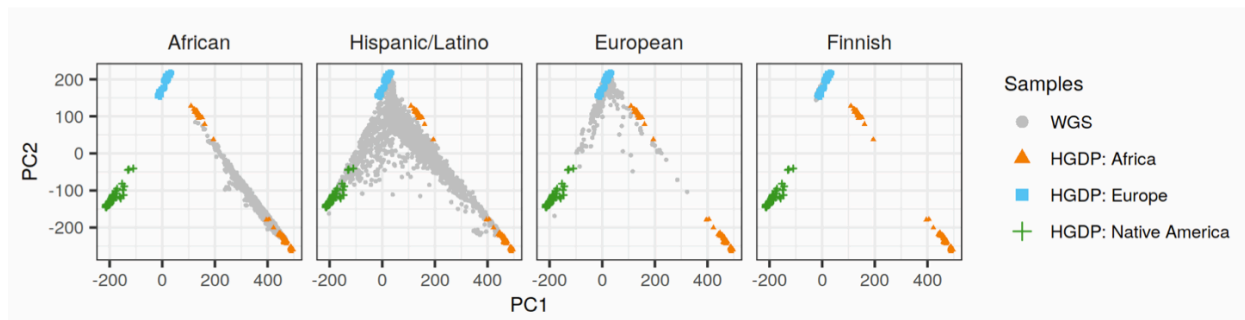
Supplementary Figure 2.5 Heterozygous genotype concordance rates for low-frequency variants by ancestry with TOPMed panel imputation

Heterozygous concordance rates were calculated between sequenced and TOPMed imputed genotypes for low-frequency ($0.5\% < \text{MAF} < 5\%$, calculated separately in each study) biallelic SNVs with the Omni2.5M array. A. Distribution of concordance rates in each of the four studies. Boxplots correspond to 25th, 50th, and 75th percentiles. B. Distribution of concordance rates by bins of estimated proportion of African ancestry in the admixed African study. C. Distribution of concordance rates in Caribbean and non-Caribbean populations in the Hispanic/Latino study. The inset figures in panels A-C show the same distributions with a restricted y-axis. D. Principal component analysis (PCA) by genotype concordance quintile and ancestry. PCA was performed by projecting onto the Human Genome Diversity Project reference samples. Genotype concordance quintiles were calculated across all four studies and correspond to concordance rates of 0.903-0.964 (Q1), 0.964-0.971 (Q2), 0.971-0.973 (Q3), 0.973-0.974 (Q4), and 0.974-0.974 (Q5). Points are colored by ancestry.



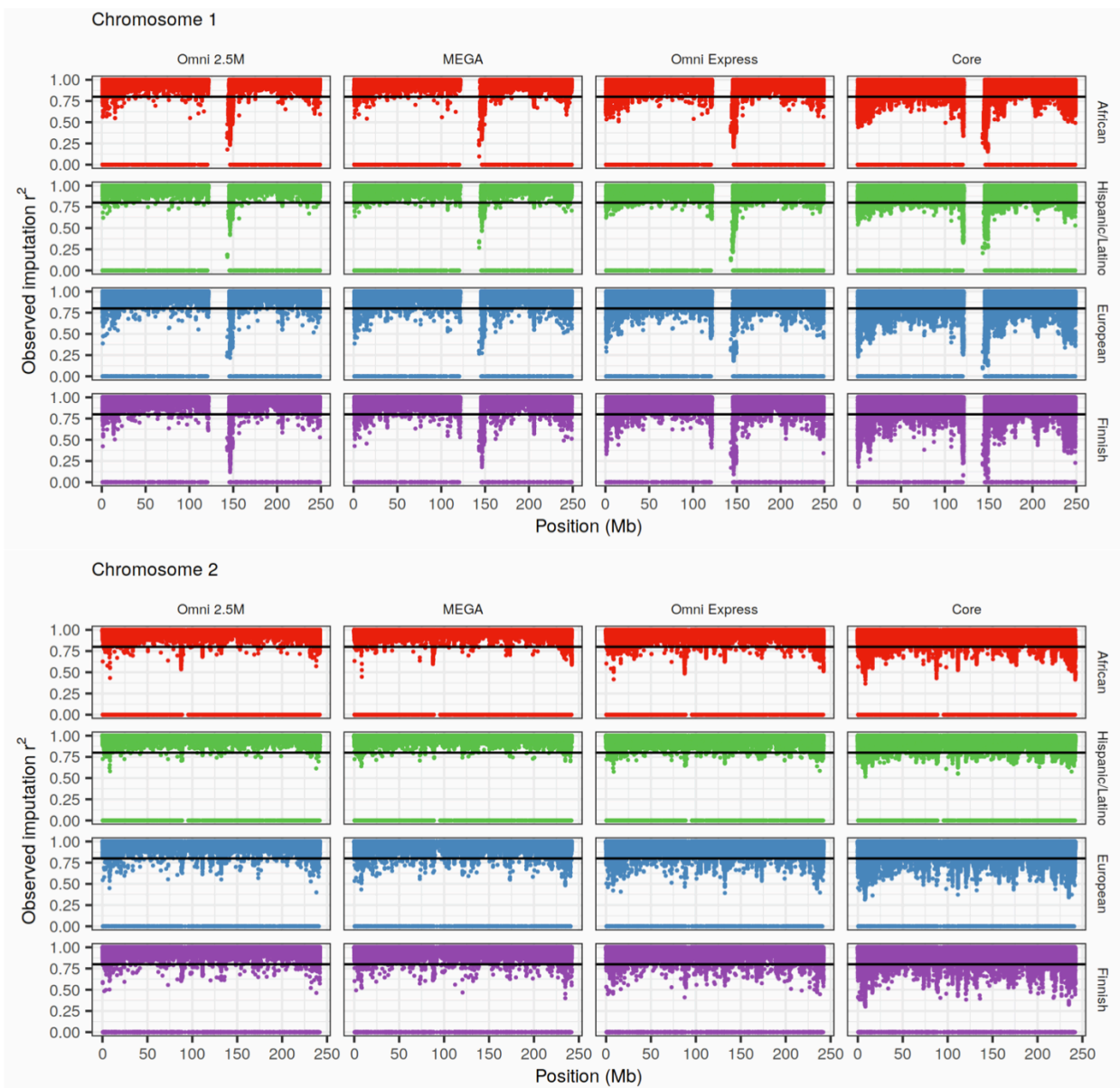
Supplementary Figure 2.6 Heterozygous genotype concordance rates for common variants by ancestry with TOPMed panel imputation.

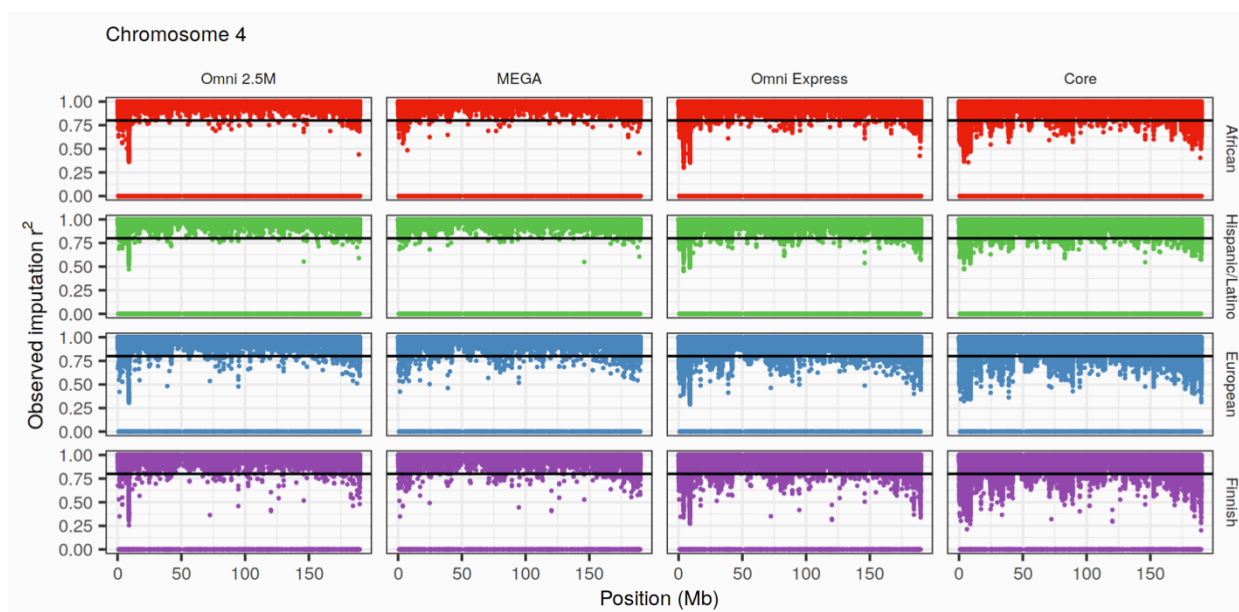
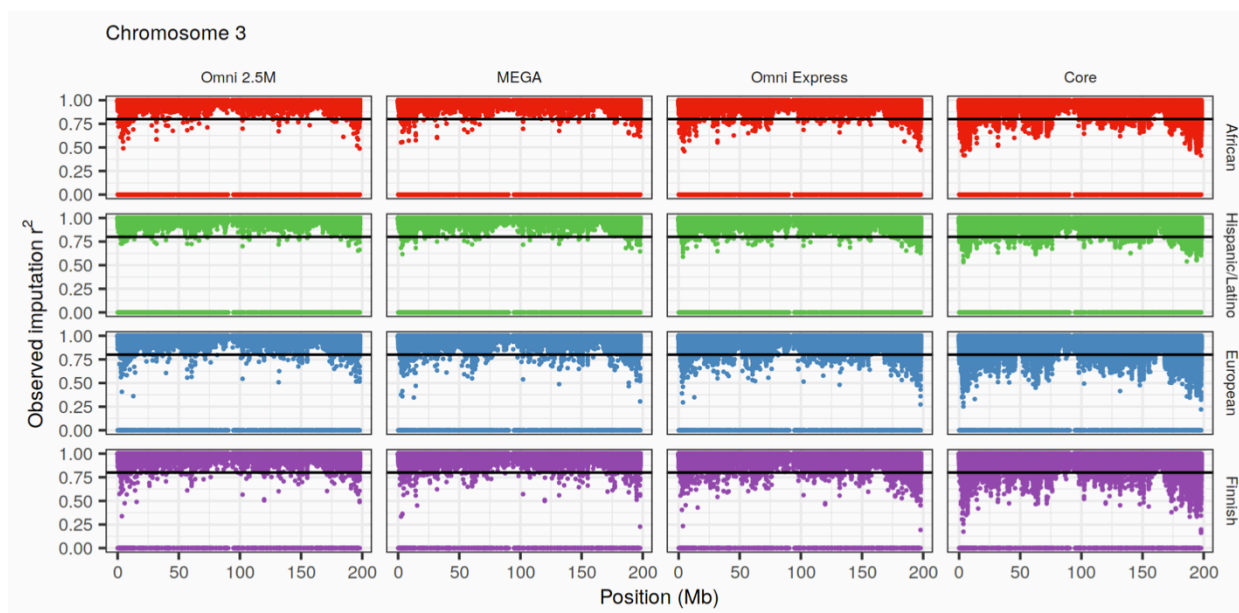
Heterozygous concordance rates were calculated between sequenced and TOPMed imputed genotypes for common (MAF>5%, calculated separately in each study) biallelic SNVs with the Omni2.5M array. A. Distribution of concordance rates in each of the four studies. Boxplots correspond to 25th, 50th, and 75th percentiles. B. Distribution of concordance rates by bins of estimated proportion of African ancestry in the admixed African study. C. Distribution of concordance rates in Caribbean and non-Caribbean populations in the Hispanic/Latino study. The inset figures in panels A-C show the same distributions with a restricted y-axis. D. Principal component analysis (PCA) by genotype concordance quintile and ancestry. PCA was performed by projecting onto the Human Genome Diversity Project reference samples. Genotype concordance quintiles were calculated across all four studies and correspond to concordance rates of 0.974-0.995 (Q1), 0.995-0.996 (Q2), 0.996-0.996 (Q3), 0.996-0.997 (Q4), and 0.997-0.997 (Q5). Points are colored by ancestry.

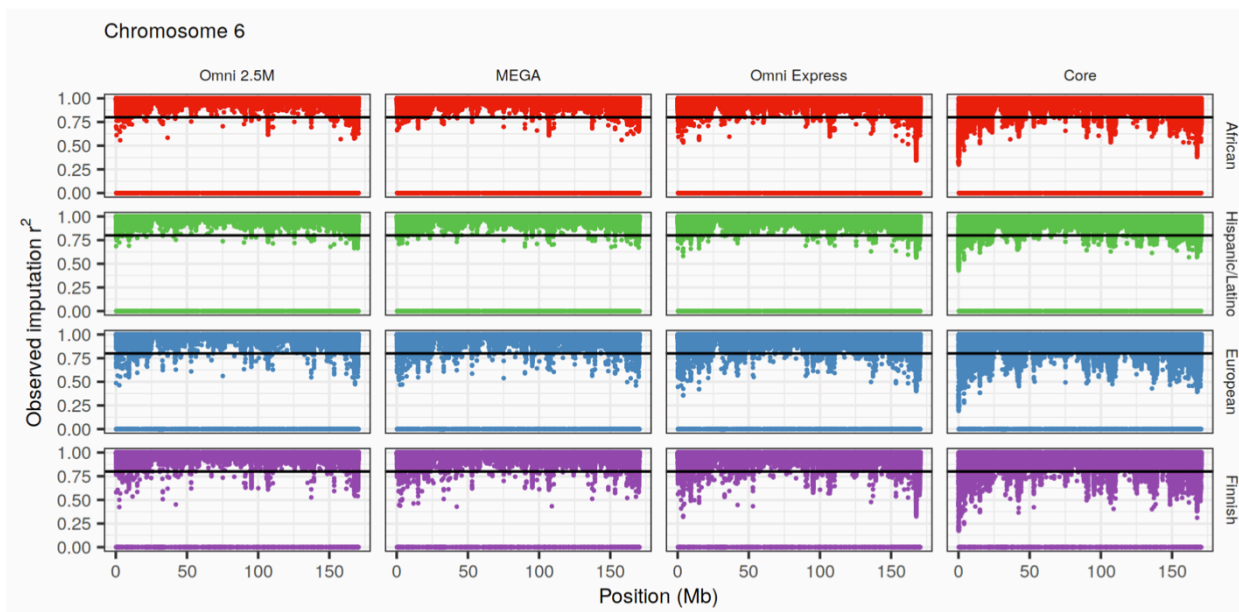
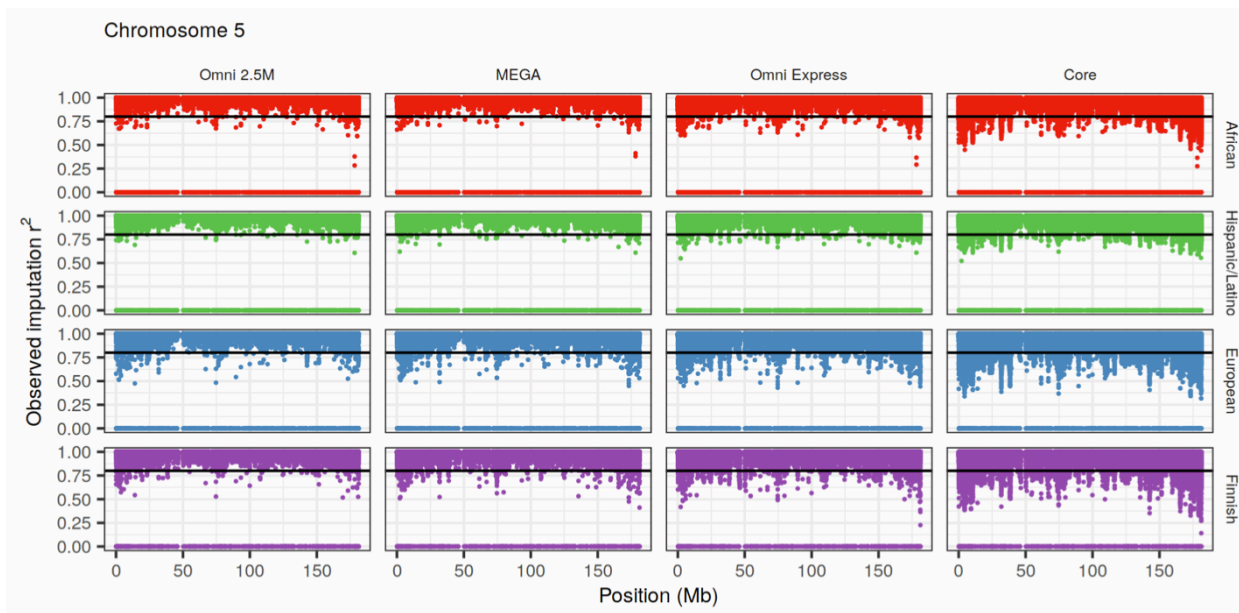


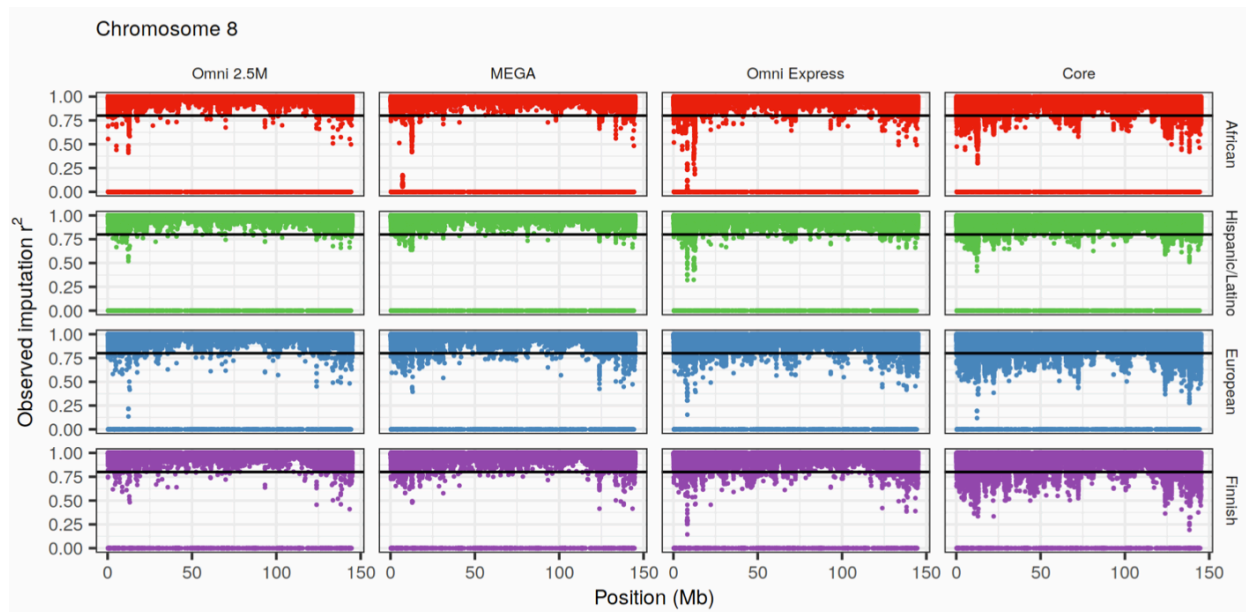
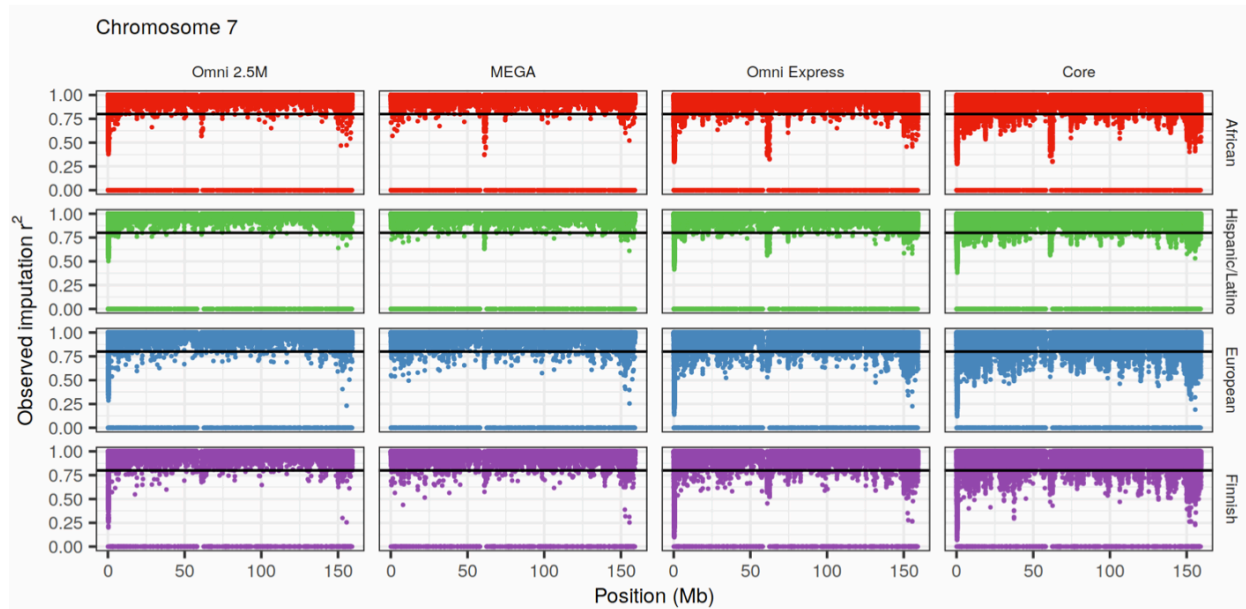
Supplementary Figure 2.7 Principal component analysis of WGS samples

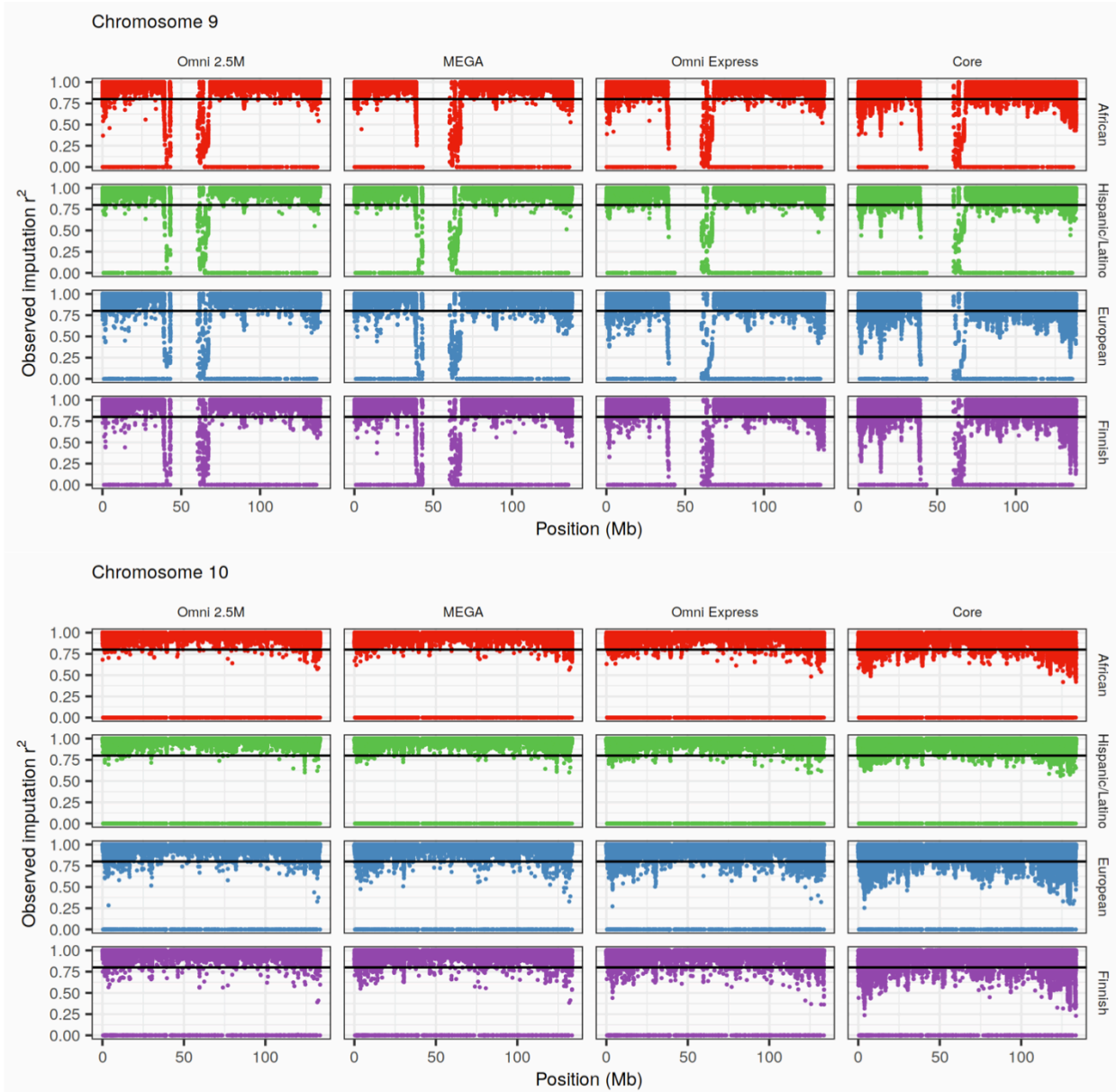
PC1 and PC2 for the four WGS studies and Human Genome Diversity Project (HGDP) reference samples from Africa (n=129), Europe (n=156), and Native America (n=63). PCA was performed by projecting onto all HGDP reference samples (n=938).

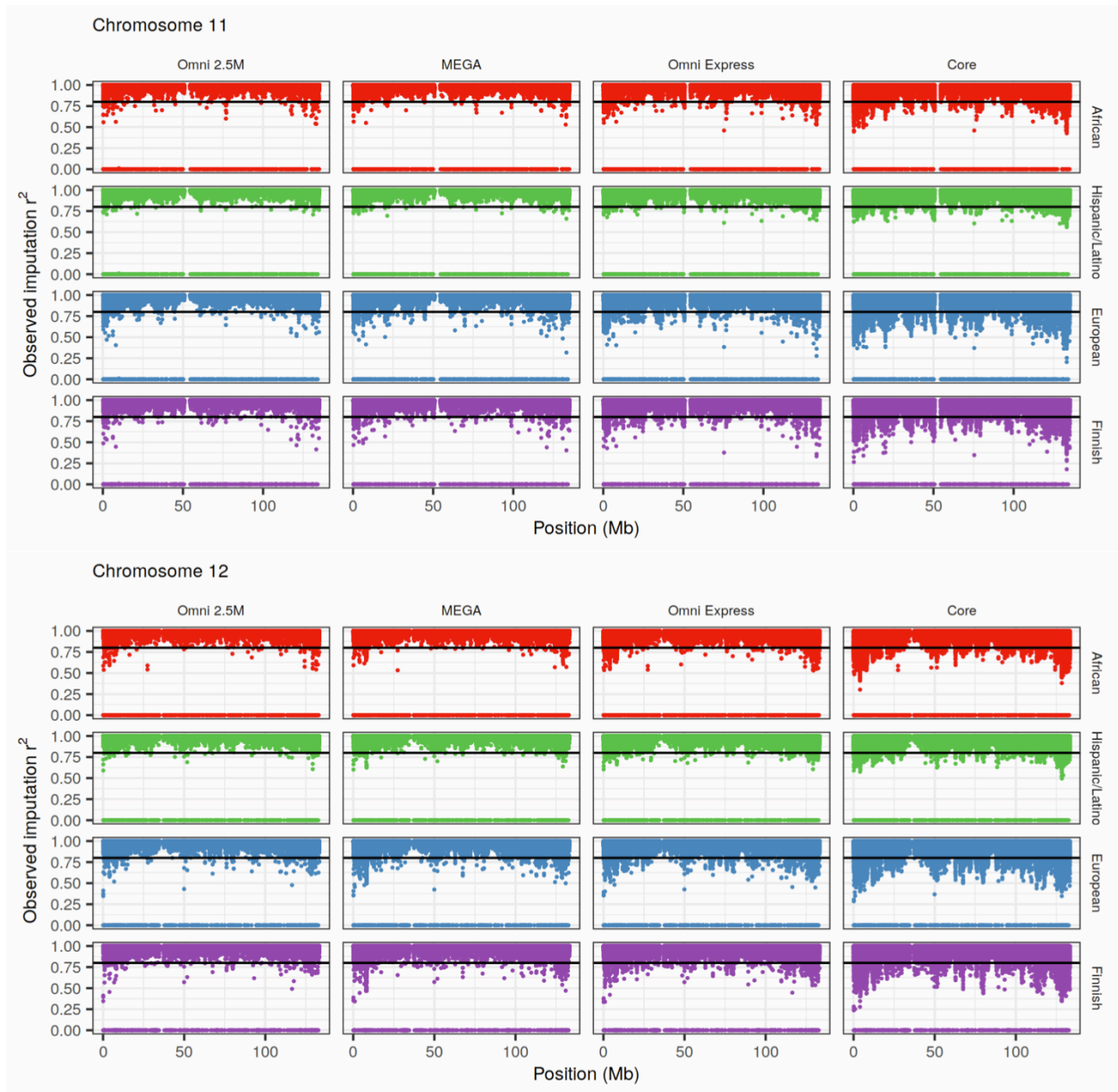


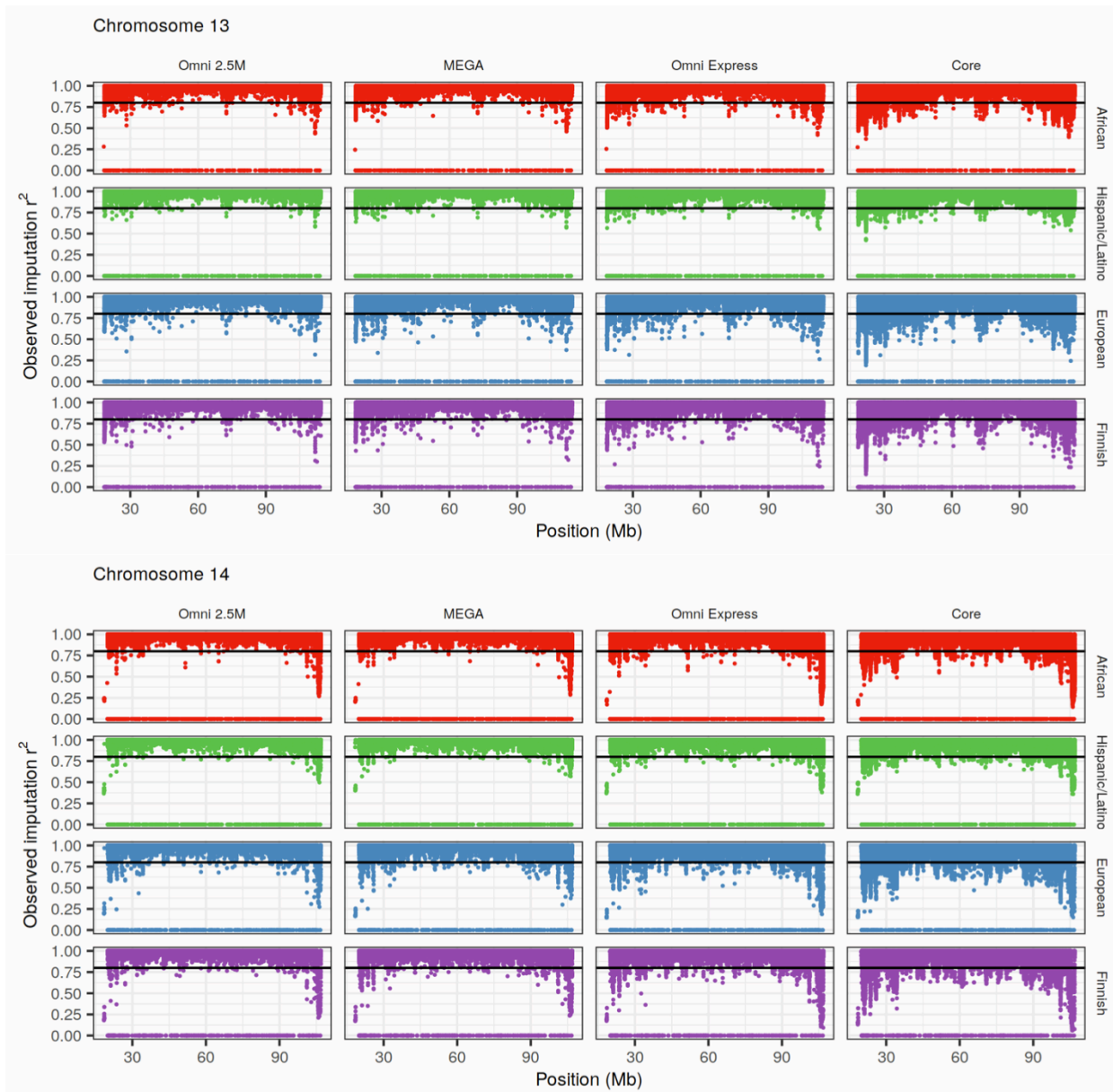


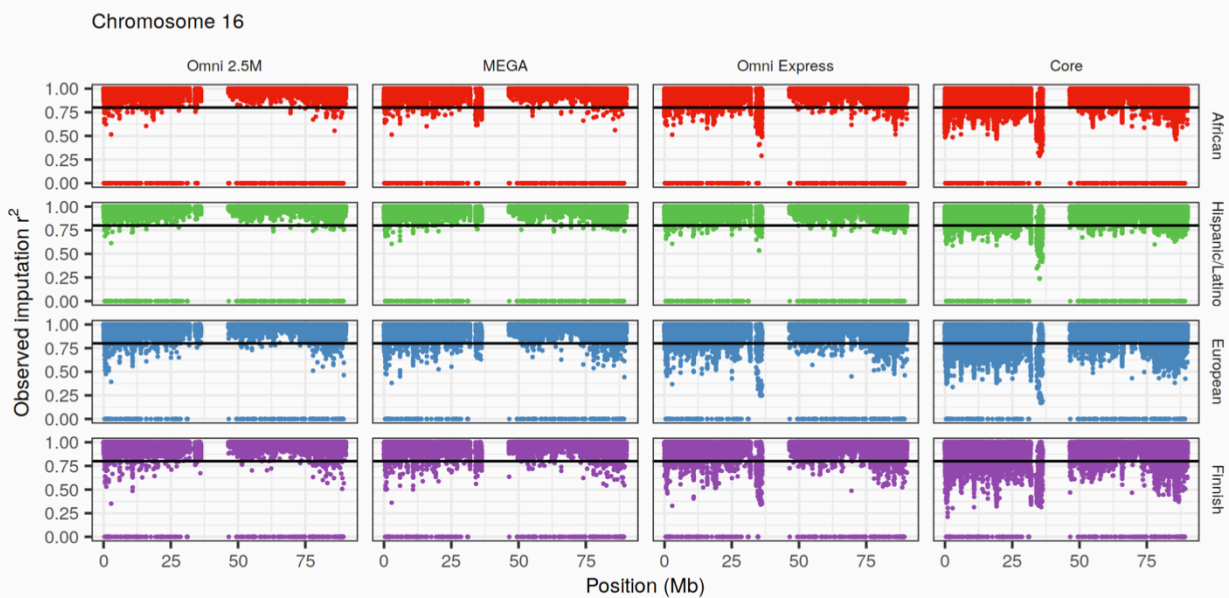
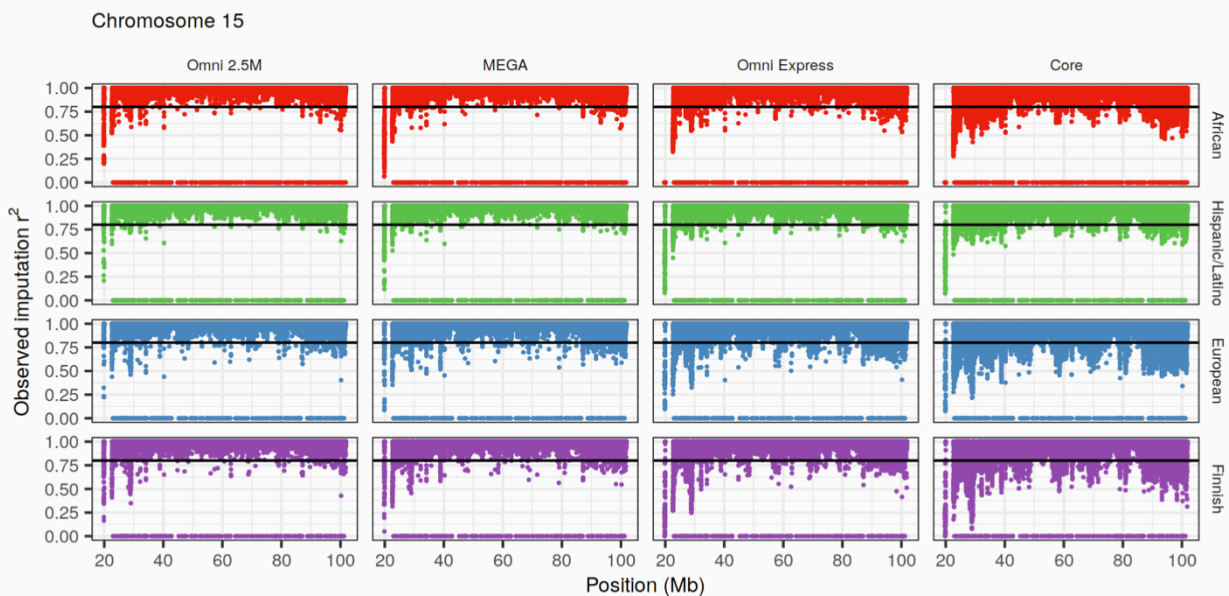


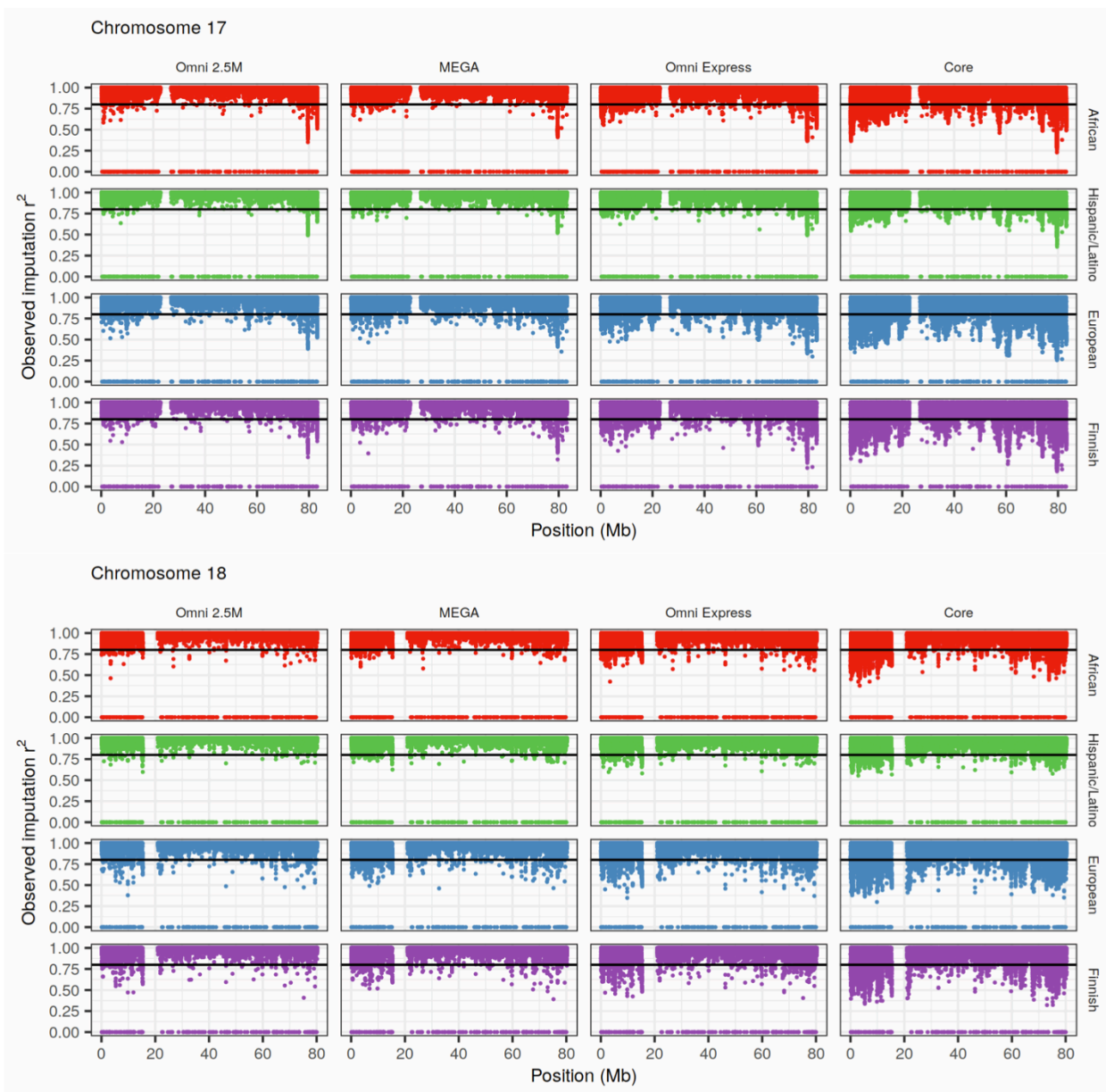


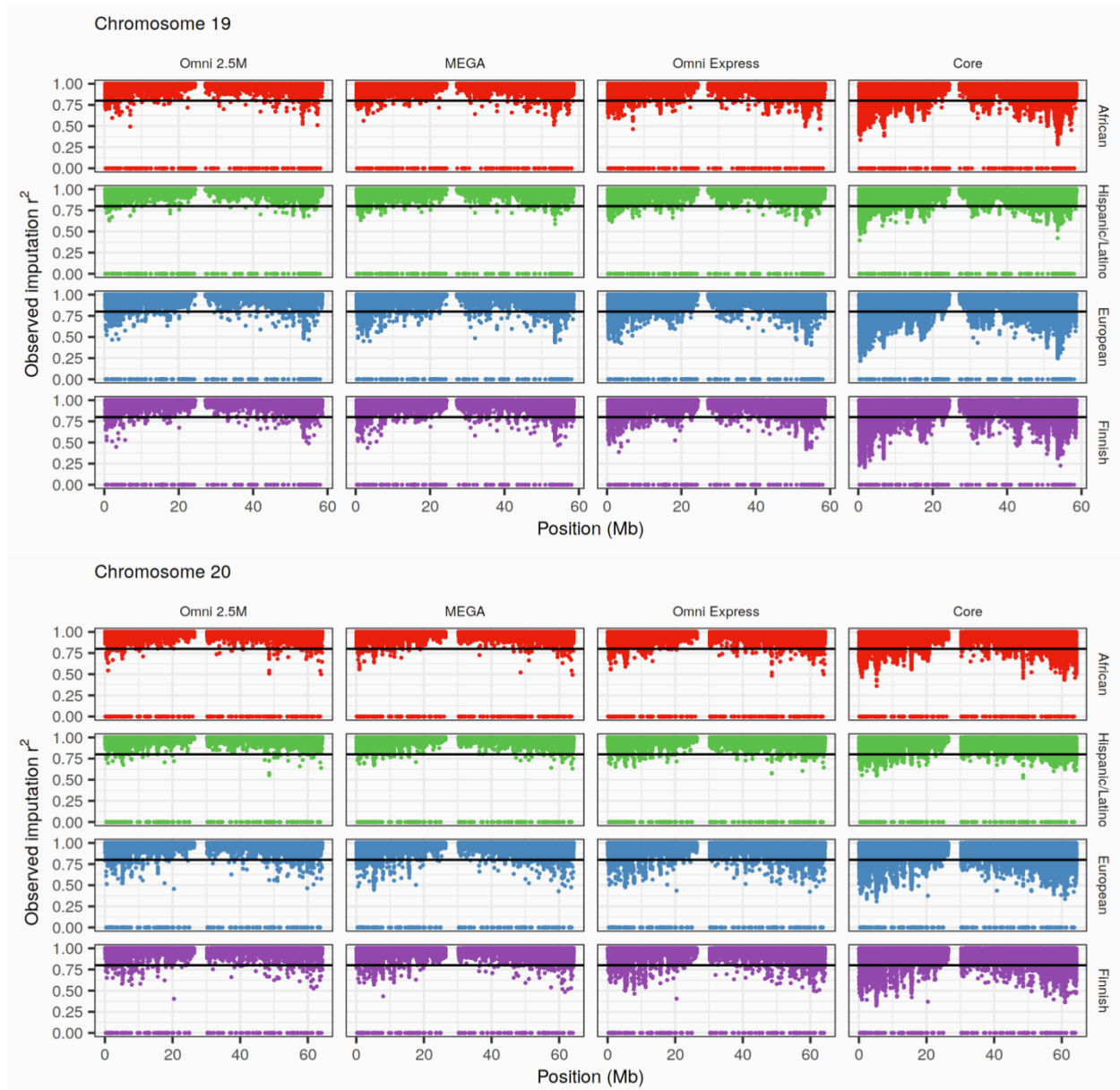


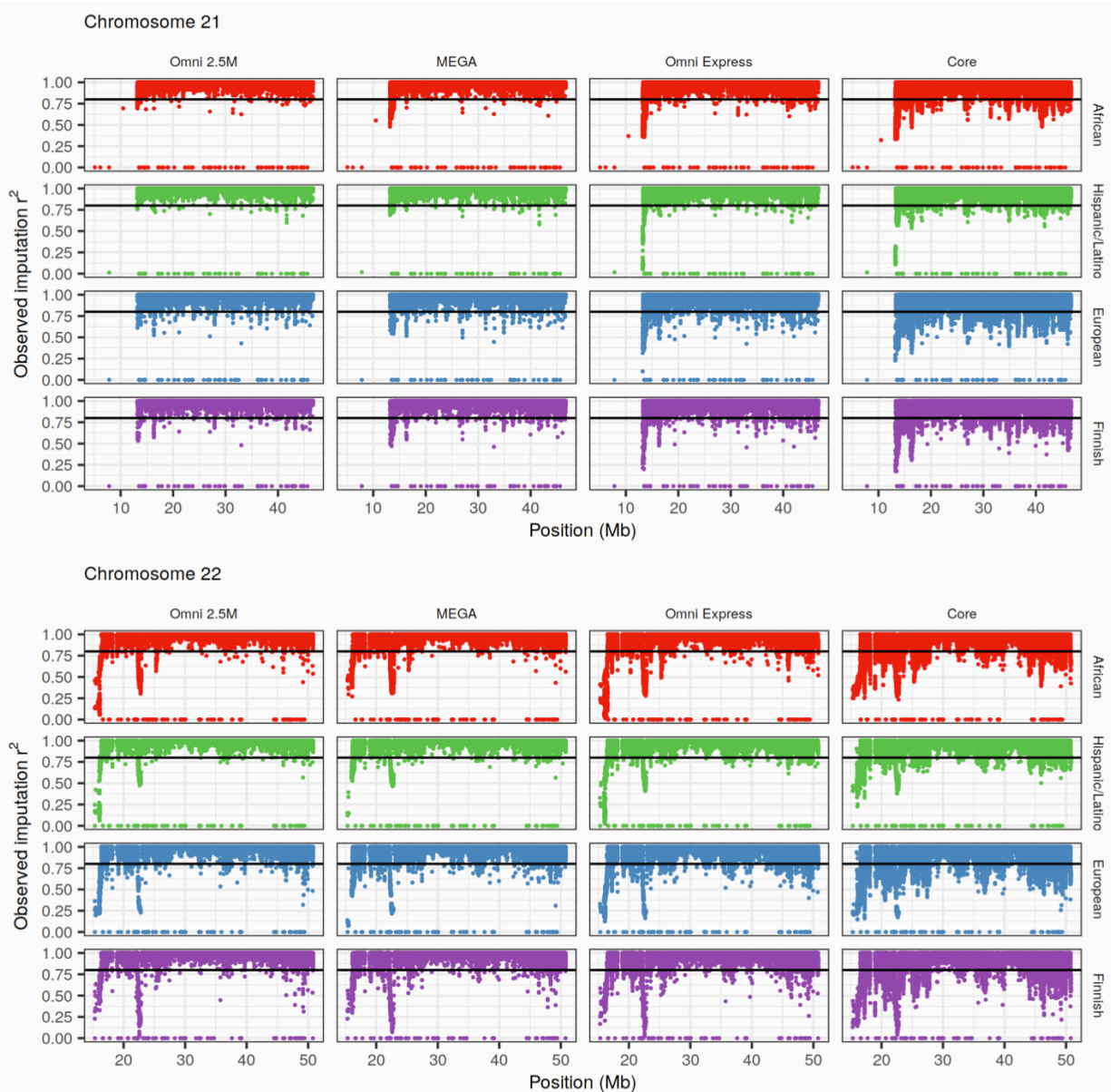






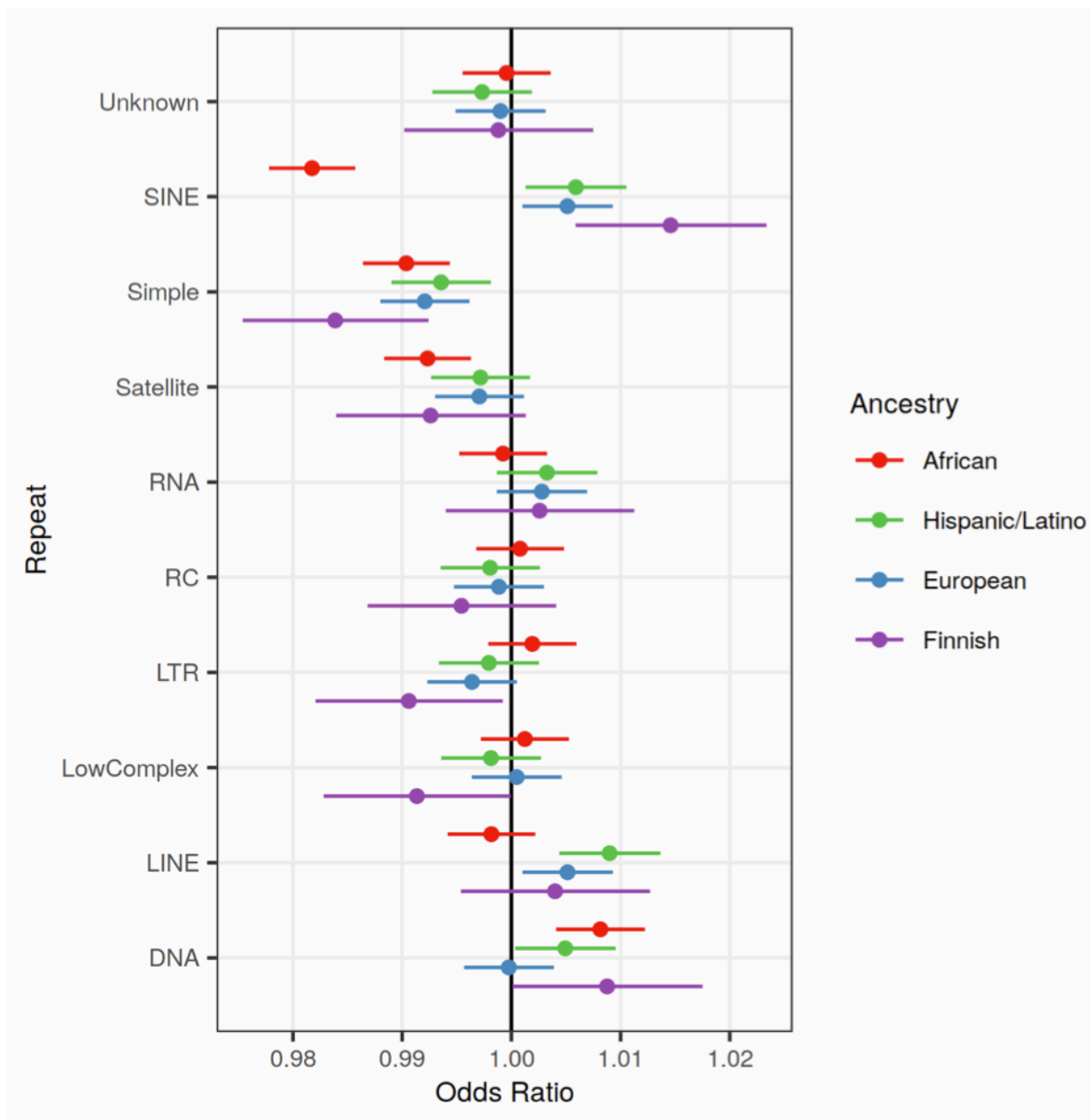






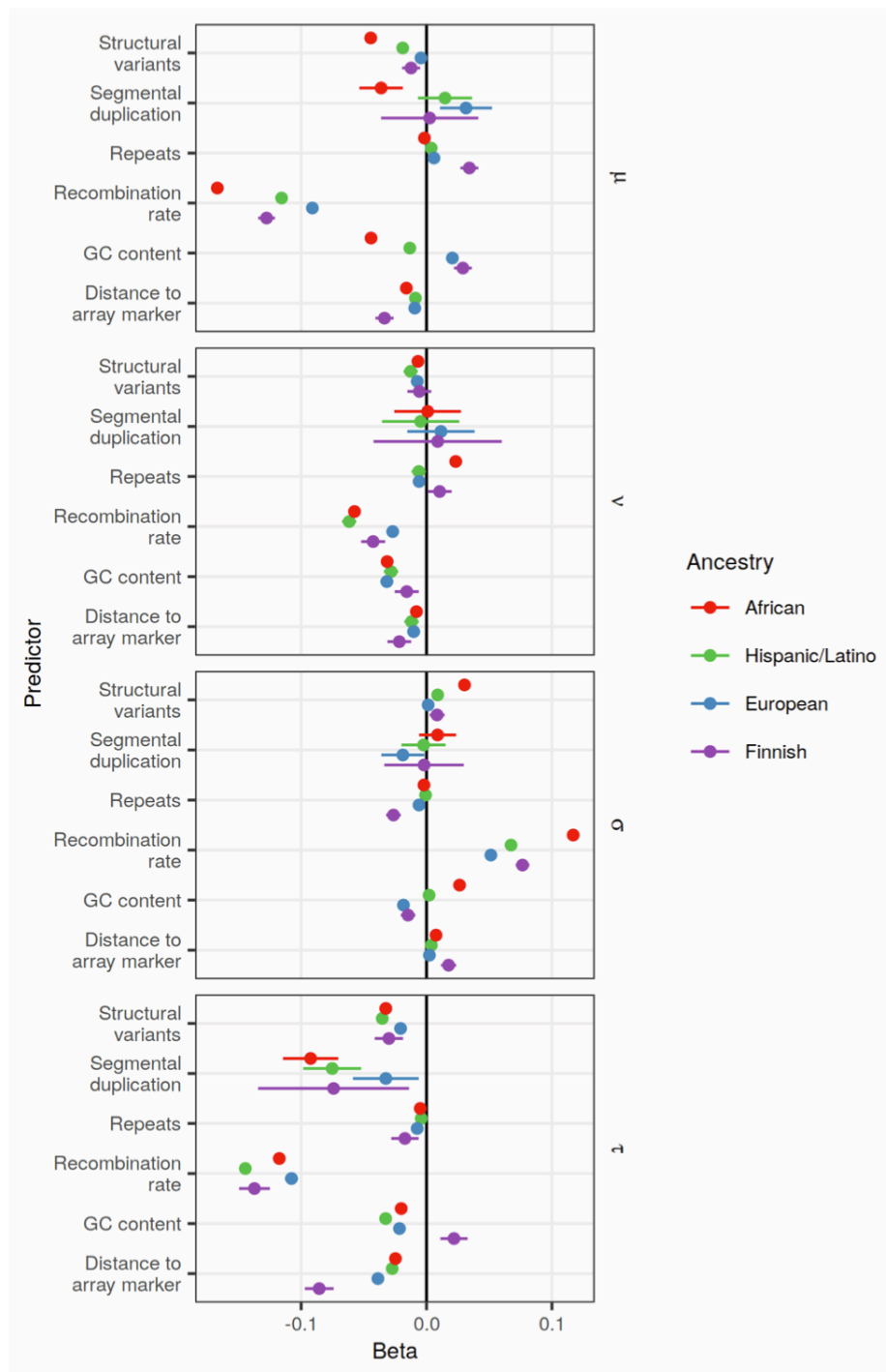
Supplementary Figure 2.8 Regional variability in imputation quality of common variants with the TOPMed reference panel by genotyping array and ancestry across all chromosomes

Observed imputation r^2 by genomic position (Mb) for common (MAF > 0.05) biallelic SNVs across all chromosomes by genotyping array (columns) and ancestry (rows). Variants above the horizontal black lines are well-imputed (observed imputation r^2 > 0.08).



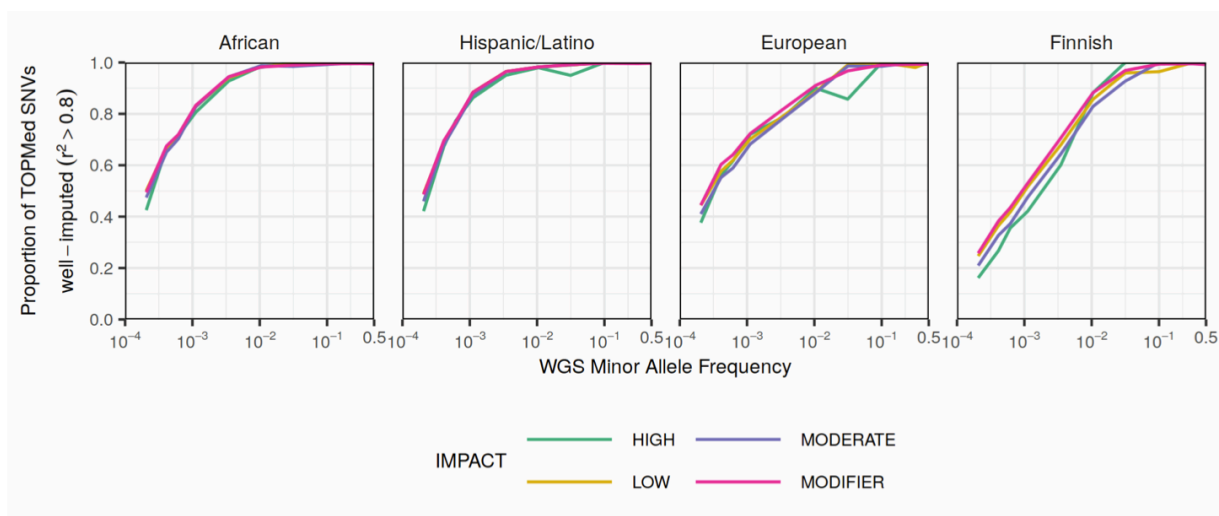
Supplementary Figure 2.9 Repeat classes associated with TOPMed imputation quality of biallelic SNVs by ancestry

The odds ratios and corresponding 95% confidence intervals from logistic regression models. Estimates are from separate models testing the associations between each repeat class and whether or not a variant is well-imputed (observed imputation $r^2 > 0.8$) adjusting for variant MAF. Repeat classes as defined by RepeatMasker include DNA repeat elements (DNA), long interspersed repeated elements (LINE), low complexity repeats (LowComplex), long terminal repeat elements including retrotransposons (LTR), rolling circle repeats (RC), RNA repeats (RNA), satellite repeats, microsatellites (Simple), short interspersed repeat elements including ALUs (SINE), and repeats of unknown class.



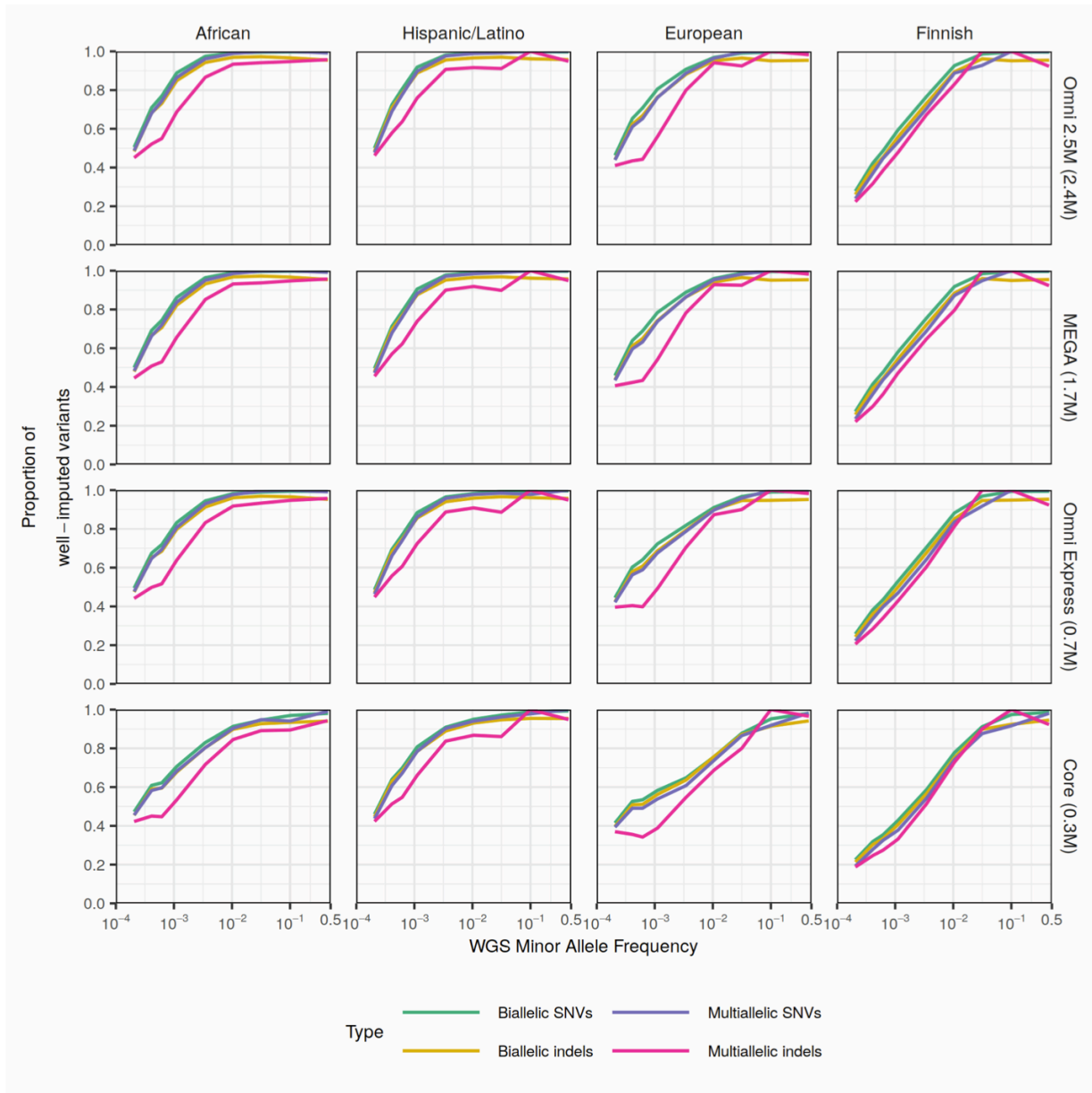
Supplementary Figure 2.10 Genomic features associated with TOPMed imputation quality of biallelic SNVs by ancestry

The odds ratios and corresponding 95% confidence intervals from zero-one inflated beta regression models testing the association of genomic features with the observed imputation r^2 in the open interval $0 < r^2 < 1$ (mean μ and variance-related parameter σ) and the probabilities of observed imputation $r^2=0$ (ν) or $r^2=1$ (τ). Estimates are from separate models testing the associations between characteristics of regional genomic features and imputation quality (observed imputation r^2) adjusting for variant MAF.



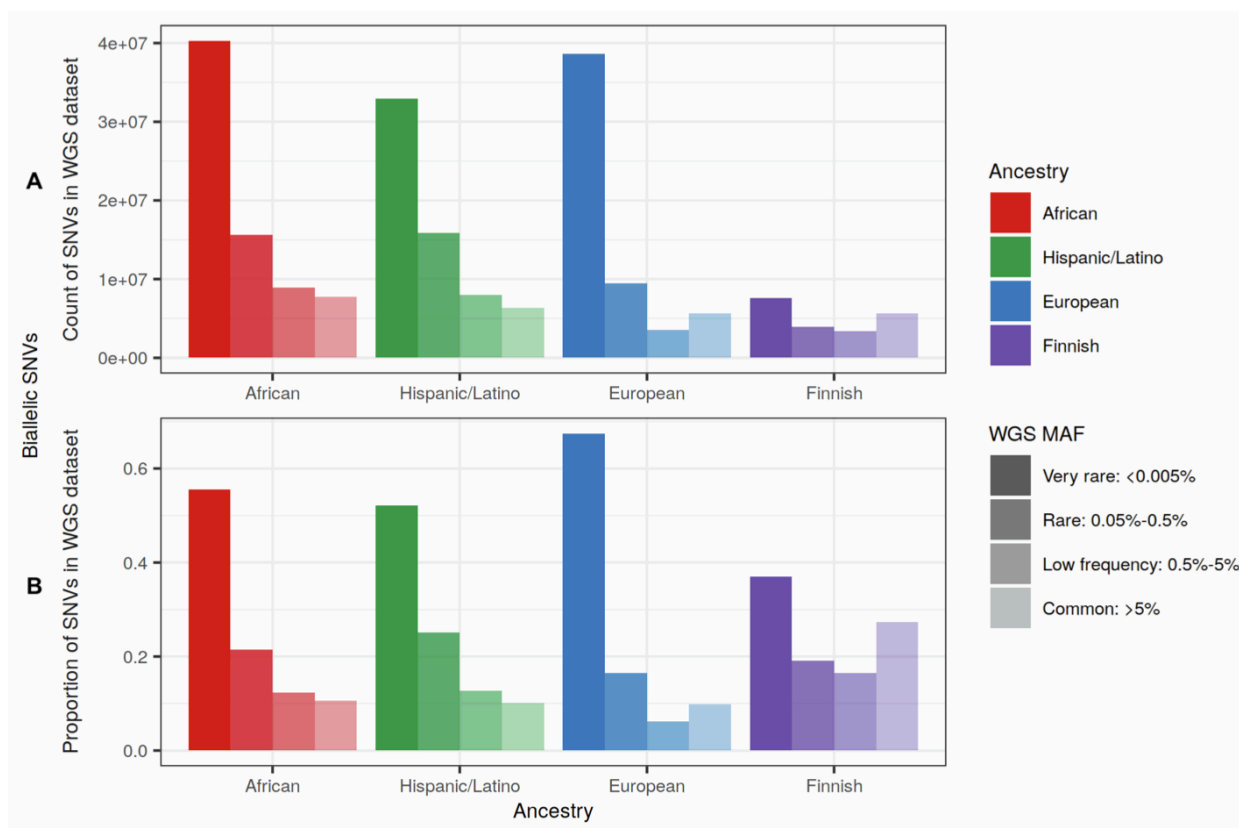
Supplementary Figure 2.11 Proportion of well-imputed ($r^2 > 0.8$) biallelic SNVs by predicted functional impact and ancestry

The predicted functional impact of all sequenced biallelic SNVs was determined with VEP. The x-axes show minor allele frequency (MAF) calculated separately by study. Biallelic SNVs were then aggregated by MAF bins of width 0.00025 MAF for MAF between 0.0002 and 0.002 and of size 0.001 MAF for MAF > 0.002 ; those plotted here correspond to singletons, doubletons, and tripletons in each study, as well as those with mean MAF closest to the values 0.001, 0.0032, 0.01, 0.032, 0.1, 0.32, and 0.5.



Supplementary Figure 2.12 Proportion of well-imputed ($r^2 > 0.8$) variants by variant type, genotyping array, and ancestry with the TOPMed panel

The proportion of sequenced variants that are well-imputed by genotyping array (rows) and ancestry (columns). X-axes show minor allele frequency (MAF) calculated separately in each study. Sequenced variants not present in reference panels were assigned $r^2 = 0$. Variants were then aggregated by MAF bins of width 0.00025 MAF for MAF between 0.0002 and 0.002, bins of width 0.001 MAF for MAF between 0.002 and 0.002, and one bin of width 0.1 MAF for MAF between 0.4 and 0.5. MAF bins plotted here correspond to singletons, doubletons, and tripletons in each study, as well as those with mean MAF closest to the values 0.01, 0.0032, 0.01, 0.316, 0.1, and 0.5.



Supplementary Figure 2.13 Distribution of MAF for biallelic SNVs by ancestry

A. Barplots of the number of biallelic SNVs in each MAF category for each WGS dataset. B. Barplots of the proportion of biallelic SNVs in each MAF category for each WGS dataset.

Array	Number of variants	African	Hispanic/Latino	European	Finnish
Omni 2.5M	2,381,000	2,132,501	2,330,998	2,330,998	2,264,709
MEGA	1,780,000	1,415,237	1,759,171	1,759,171	1,676,050
OmniExpress	710,000	680,234	706,652	706,652	698,865
Core	307,000	266,727	288,599	288,599	302,423

Supplementary Table 2.1 Whole genome sequencing (WGS)-based genotype arrays

The numbers of variants included on the Illumina arrays and the actual number of WGS variants in each study used to create the WGS-based arrays.

Reference panel	Array	MAF	African	Hispanic/Latino	European	Finnish
TOPMed	Omni 2.5M	Common	7.7M	6.3M	5.6M	5.6M
		Low frequency	8.9M	8.0M	3.4M	3.2M
		Rare	35.6M	32.4M	26.5M	4.8M
	MEGA	Common	7.7M	6.3M	5.6M	5.6M
		Low frequency	8.9M	8.0M	3.4M	3.2M
		Rare	35.0M	32.0M	26.1M	4.7M
	OmniExpress	Common	7.7M	6.3M	5.6M	5.6M
		Low frequency	8.8M	7.9M	3.3M	3.1M
		Rare	34.2M	31.4M	24.9M	4.4M
	Core	Common	7.5M	6.3M	5.5M	5.5M
		Low frequency	8.2M	7.7M	2.8M	2.8M
		Rare	31.2M	29.2M	22.2M	3.7M
HRC	Omni 2.5M	Common	7.1M	5.9M	5.2M	5.2M
		Low frequency	6.0M	6.2M	2.9M	3.1M
		Rare	4.0M	5.1M	9.4M	3.6M
	MEGA	Common	6.7M	5.8M	5.2M	5.2M
		Low frequency	4.9M	5.4M	2.8M	3.1M
		Rare	3.6M	4.3M	8.6M	3.6M
	OmniExpress	Common	6.5M	5.7M	5.2M	5.2M
		Low frequency	4.1M	4.7M	2.5M	3.1M
		Rare	3.1M	3.7M	7.8M	3.4M
	Core	Common	4.7M	5.0M	4.9M	5.2M
		Low frequency	1.9M	2.7M	1.9M	3.0M
		Rare	2.0M	2.3M	5.7M	3.1M
1000G	Omni 2.5M	Common	7.5M	6.2M	5.5M	5.5M
		Low frequency	7.2M	6.6M	2.4M	2.6M
		Rare	4.4M	6.5M	7.0M	1.7M
	MEGA	Common	7.2M	6.1M	5.4M	5.5M
		Low frequency	6.1M	6.0M	2.3M	2.5M
		Rare	3.5M	5.4M	6.3M	1.6M
	OmniExpress	Common	6.9M	6.0M	5.3M	5.4M
		Low frequency	5.3M	5.4M	2.0M	2.4M
		Rare	2.9M	4.8M	5.6M	1.4M
	Core	Common	5.4M	5.4M	5.0M	5.2M
		Low frequency	2.6M	3.3M	1.4M	2.0M
		Rare	1.4M	2.9M	3.7M	1.1M

Supplementary Table 2.2 Number of well-imputed biallelic single nucleotide variants (SNVs) in each whole genome sequencing (WGS) study by reference panel, genotype array, ancestry, and minor allele frequency (MAF) category

Reference panel	Array	MAF	African	Hispanic/Latino	European	Finnish
TOPMed	Omni 2.5M	Common	0.997	0.997	0.996	0.996
		Low frequency	0.993	0.992	0.974	0.945
		Rare	0.637	0.664	0.552	0.415
	MEGA	Common	0.997	0.997	0.996	0.996
		Low frequency	0.992	0.992	0.967	0.939
		Rare	0.626	0.656	0.543	0.408
	OmniExpress	Common	0.994	0.996	0.992	0.993
		Low frequency	0.984	0.985	0.927	0.913
		Rare	0.613	0.642	0.517	0.379
	Core	Common	0.973	0.990	0.969	0.978
		Low frequency	0.922	0.954	0.800	0.830
		Rare	0.559	0.598	0.461	0.318
HRC	Omni 2.5M	Common	0.921	0.926	0.929	0.933
		Low frequency	0.668	0.772	0.812	0.908
		Rare	0.071	0.104	0.195	0.314
	MEGA	Common	0.871	0.914	0.926	0.933
		Low frequency	0.546	0.679	0.784	0.907
		Rare	0.065	0.088	0.180	0.310
	OmniExpress	Common	0.834	0.894	0.917	0.932
		Low frequency	0.463	0.591	0.701	0.901
		Rare	0.055	0.076	0.162	0.298
	Core	Common	0.609	0.792	0.875	0.931
		Low frequency	0.208	0.338	0.539	0.886
		Rare	0.036	0.047	0.119	0.273
1000G	Omni 2.5M	Common	0.970	0.976	0.974	0.977
		Low frequency	0.801	0.828	0.692	0.760
		Rare	0.079	0.134	0.145	0.150
	MEGA	Common	0.936	0.965	0.965	0.974
		Low frequency	0.679	0.752	0.658	0.745
		Rare	0.063	0.110	0.131	0.142
	OmniExpress	Common	0.895	0.946	0.948	0.966
		Low frequency	0.590	0.667	0.559	0.697
		Rare	0.052	0.098	0.116	0.125
	Core	Common	0.691	0.851	0.880	0.930
		Low frequency	0.286	0.409	0.401	0.590
		Rare	0.025	0.058	0.077	0.091

Supplementary Table 2.3 Proportion of biallelic single nucleotide variants (SNVs) in each whole genome sequencing (WGS) study that are well-imputed ($r^2 > 0.8$) by reference panel, genotype array, ancestry, and minor allele frequency (MAF) category

Reference panel	Array	African	Hispanic/Latino	European	Finnish
TOPMed	Omni 2.5M	0.0014	0.0011	0.0035	0.0084
	MEGA	0.0016	0.0011	0.0045	0.0095
	OmniExpress	0.0024	0.0014	0.0095	0.0126
	Core	0.0084	0.0035	0.0395	0.0275
HRC	Omni 2.5M	0.0485	0.0364	0.0276	0.0115
	MEGA	0.3065	0.0565	0.0346	0.0115
	OmniExpress	NA	0.1055	0.0585	0.0135
	Core	NA	NA	0.2015	0.0154
1000G	Omni 2.5M	0.0245	0.0235	0.0385	0.0325
	MEGA	0.0665	0.0364	0.0455	0.0365
	OmniExpress	0.1395	0.0675	0.0705	0.0515
	Core	NA	0.2225	0.1704	0.0945

Supplementary Table 2.4 Minor allele frequency (MAF) threshold above which array genotyping and imputation can approximate whole genome sequencing (WGS) for biallelic single nucleotide variants (SNVs) by reference panel, genotype array, and ancestry

Threshold is the smallest MAF for which >90% of biallelic SNVs are well-imputed (observed imputation $r^2 > 0.8$).

	Array	MAF	African				Hispanic/ Latino			European	Finnish
			All	0.25-0.5	0.5-0.75	0.75-1.0	All	NC	C	All	All
TOPMed	Omni 2.5M	Common	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		Low frequency	0.99	0.98	0.99	0.99	0.99	0.99	0.97	0.98	0.97
		Rare	0.93	0.89	0.93	0.93	0.93	0.79	0.96	0.86	0.82
	MEGA	Common	0.92	0.92	0.92	0.92	0.96	0.97	0.96	0.99	1.00
		Low frequency	0.98	0.97	0.99	0.99	0.99	0.99	0.96	0.97	0.97
		Rare	0.91	0.87	0.91	0.91	0.92	0.78	0.95	0.84	0.81
	Omni Express	Common	0.92	0.92	0.92	0.92	0.96	0.97	0.96	0.99	1.00
		Low frequency	0.98	0.97	0.99	0.99	0.99	0.99	0.96	0.97	0.97
		Rare	0.91	0.86	0.91	0.91	0.91	0.75	0.95	0.82	0.78
	Core	Common	0.98	0.97	0.98	0.98	0.97	0.97	0.96	0.98	0.99
		Low frequency	0.98	0.97	0.99	0.99	0.99	0.99	0.96	0.97	0.97
		Rare	0.86	0.80	0.86	0.86	0.88	0.68	0.92	0.76	0.71
HRC	Omni 2.5M	Common	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.99	1.00
		Low frequency	0.91	0.91	0.91	0.91	0.92	0.92	0.90	0.94	0.97
		Rare	0.70	0.73	0.72	0.69	0.70	0.67	0.71	0.70	0.82
	MEGA	Common	0.96	0.97	0.96	0.96	0.98	0.98	0.98	0.99	1.00
		Low frequency	0.81	0.82	0.82	0.81	0.89	0.89	0.88	0.93	0.97
		Rare	0.65	0.71	0.69	0.64	0.65	0.63	0.66	0.68	0.81
	Omni Express	Common	0.95	0.96	0.96	0.95	0.98	0.98	0.98	0.98	1.00
		Low frequency	0.83	0.85	0.84	0.83	0.84	0.85	0.82	0.90	0.97
		Rare	0.62	0.67	0.66	0.61	0.60	0.57	0.61	0.65	0.80
	Core	Common	0.89	0.92	0.90	0.89	0.95	0.95	0.95	0.97	0.99
		Low frequency	0.68	0.73	0.70	0.67	0.73	0.73	0.70	0.84	0.95
		Rare	0.50	0.58	0.55	0.49	0.51	0.47	0.51	0.55	0.77
1000G	Omni 2.5M	Common	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.97	0.99
		Low frequency	0.93	0.92	0.93	0.93	0.91	0.92	0.89	0.88	0.90
		Rare	0.73	0.69	0.71	0.73	0.73	0.66	0.75	0.62	0.71
	MEGA	Common	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.97	0.99
		Low frequency	0.88	0.87	0.88	0.88	0.89	0.90	0.87	0.86	0.91
		Rare	0.65	0.64	0.65	0.65	0.68	0.61	0.69	0.58	0.68
	Omni Express	Common	0.96	0.96	0.96	0.96	0.98	0.98	0.98	0.96	0.99
		Low frequency	0.85	0.85	0.85	0.85	0.86	0.87	0.82	0.81	0.88
		Rare	0.61	0.58	0.61	0.62	0.64	0.56	0.66	0.54	0.63
	Core	Common	0.90	0.92	0.91	0.90	0.95	0.95	0.95	0.94	0.97
		Low frequency	0.70	0.72	0.71	0.70	0.74	0.75	0.69	0.69	0.78
		Rare	0.44	0.44	0.45	0.44	0.51	0.42	0.53	0.40	0.52

Supplementary Table 2.5 Mean heterozygous concordance rates by reference panel, genotype array, ancestry, and MAF category

Summary statistics are further broken down for the African ancestry study by estimated proportion of African ancestry (0.25-0.5, 0.5-0.75, 0.75-1.00) and for the Hispanic/Latino ancestry study by Caribbean (C) and non-Caribbean (NC) origin.

	Array	MAF	Number of consecutively well-imputed ($r^2>0.8$) biallelic SNVs											
			African			Hispanic/ Latino			European			Finnish		
			25 th	50 th	75 th	25 th	50 th	75 th	25 th	50 th	75 th	25 th	50 th	75 th
TOPMed	Omni 2.5M	Common	41	277	750	52	295	777	33	197	576	35	210	592
		Low frequency	9	85	287	18	66	186	4	12	41	4	11	25
		Rare	1	2	4	1	2	4	1	2	3	1	1	2
	MEGA	Common	21	243	715	41	276	753	16	139	473	17	157	505
		Low frequency	5	45	205	14	57	166	3	9	30	4	10	23
		Rare	1	2	4	1	2	4	1	2	3	1	1	2
	Omni Express	Common	4	106	512	17	193	616	7	56	267	8	72	328
		Low frequency	3	15	98	7	30	92	2	6	16	3	7	16
		Rare	1	2	3	1	2	4	1	2	3	1	1	2
	Core	Common	1	3	20	2	17	194	2	10	46	2	11	61
		Low frequency	1	4	13	2	8	28	1	3	7	2	4	9
		Rare	1	2	3	1	2	3	1	1	2	1	1	2
HRC	Omni 2.5M	Common	2	9	28	3	13	33	4	14	35	4	15	38
		Low frequency	1	2	5	1	3	7	2	4	8	3	9	19
		Rare	1	1	1	1	1	1	1	1	1	1	1	2
	MEGA	Common	2	4	15	3	10	28	4	14	33	4	15	38
		Low frequency	1	2	3	1	2	5	1	3	7	3	8	18
		Rare	1	1	1	1	1	1	1	1	1	1	1	2
	Omni Express	Common	1	4	12	2	8	22	3	12	29	4	15	38
		Low frequency	1	2	3	1	2	4	1	2	5	3	8	17
		Rare	1	1	1	1	1	1	1	1	1	1	1	2
	Core	Common	1	2	5	1	3	10	2	7	19	4	15	37
		Low frequency	1	1	2	1	1	3	1	2	4	3	6	15
		Rare	1	1	1	1	1	1	1	1	1	1	1	2
1000G	Omni 2.5M	Common	3	15	68	5	27	85	6	25	72	7	30	88
		Low frequency	2	4	8	2	4	9	1	2	5	1	3	6
		Rare	1	1	1	1	1	1	1	1	1	1	1	1
	MEGA	Common	2	4	20	3	12	54	4	16	53	5	24	74
		Low frequency	1	2	5	1	3	6	1	2	4	1	3	5
		Rare	1	1	1	1	1	1	1	1	1	1	1	1
	Omni Express	Common	1	4	14	2	9	37	3	13	36	4	18	55
		Low frequency	1	2	4	1	2	5	1	2	3	1	2	5
		Rare	1	1	1	1	1	1	1	1	1	1	1	1
	Core	Common	1	2	5	1	3	11	2	5	17	2	8	27
		Low frequency	1	1	2	1	2	3	1	1	3	1	2	4
		Rare	1	1	1	1	1	1	1	1	1	1	1	1

Supplementary Table 2.6 25th, 50th, and 75th percentiles of the number of consecutive well-imputed (observed imputation $r^2>0.8$) biallelic single nucleotide variants (SNVs) by reference panel, genotype array, ancestry, and minor allele frequency (MAF) category

	Array	MAF	Length in kb of consecutively well-imputed ($r^2>0.8$) biallelic SNVs											
			African			Hispanic/ Latino			European			Finnish		
			25 th	50 th	75 th	25 th	50 th	75 th	25 th	50 th	75 th	25 th	50 th	75 th
TOPMed	Omni 2.5M	Common	10.4	84.8	253.2	15.7	109.6	315.3	11.6	80.1	256.5	12.5	87.2	267.4
		Low frequency	2.3	23.9	84.3	5.2	20.7	60.1	1.6	7.8	29.8	2.3	8.0	19.2
		Rare	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.2
	MEGA	Common	5.1	73.3	241.6	11.4	101.3	302.2	5.3	55.7	210.5	6.3	64.9	224.4
		Low frequency	1.0	12.4	59.8	4.0	17.3	53.5	1.1	5.7	21.4	1.9	7.1	17.4
		Rare	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.2
	Omni Express	Common	0.7	31.3	162.8	4.6	70.4	245.3	2.5	23.9	113.6	3.2	31.1	144.6
		Low frequency	0.3	3.8	27.9	1.7	9.2	29.7	0.5	3.4	11.2	1.0	4.6	12.0
		Rare	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.2
	Core	Common	0.0	0.6	5.9	0.3	6.2	73.1	0.3	4.1	21.2	0.4	5.0	28.1
		Low frequency	0.0	0.7	3.4	0.3	2.1	8.9	0.0	1.2	4.5	0.1	1.9	6.0
		Rare	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1
HRC	Omni 2.5M	Common	0.2	2.3	8.5	0.6	4.0	12.3	0.8	5.0	14.9	0.8	5.3	16.2
		Low frequency	0.0	0.3	1.0	0.0	0.6	1.9	0.1	1.7	5.2	1.4	5.6	13.8
		Rare	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	MEGA	Common	0.0	0.9	4.3	0.3	2.9	10.1	0.7	4.7	14.2	0.8	5.3	16.2
		Low frequency	0.0	0.1	0.7	0.0	0.3	1.2	0.0	1.4	4.4	1.3	5.5	13.5
		Rare	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	Omni Express	Common	0.0	0.7	3.4	0.2	2.2	8.2	0.7	4.2	12.4	0.8	5.3	16.1
		Low frequency	0.0	0.0	0.5	0.0	0.2	1.0	0.0	0.8	2.9	1.1	5.1	12.7
		Rare	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	Core	Common	0.0	0.2	1.1	0.0	0.8	3.6	0.3	2.4	8.1	0.8	5.2	15.9
		Low frequency	0.0	0.0	0.2	0.0	0.0	0.5	0.0	0.3	1.7	0.8	4.0	10.6
		Rare	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1000G	Omni 2.5M	Common	0.3	4.0	21.2	1.2	9.7	33.4	1.9	10.6	31.6	2.1	12.9	40.0
		Low frequency	0.0	0.6	2.0	0.1	0.9	2.6	0.0	0.7	2.5	0.0	1.2	3.7
		Rare	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	MEGA	Common	0.0	1.0	6.0	0.4	4.2	20.4	0.9	6.7	23.2	1.5	9.9	33.1
		Low frequency	0.0	0.3	1.1	0.0	0.5	1.7	0.0	0.5	2.2	0.0	1.1	3.4
		Rare	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Omni Express	Common	0.0	0.7	4.0	0.3	3.0	14.3	0.9	5.4	16.5	1.3	8.1	25.0
		Low frequency	0.0	0.2	0.9	0.0	0.3	1.3	0.0	0.2	1.5	0.0	0.8	2.7
		Rare	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Core	Common	0.0	0.2	1.3	0.0	0.8	4.2	0.1	1.9	7.6	0.4	3.4	12.5
		Low frequency	0.0	0.0	0.4	0.0	0.0	0.6	0.0	0.0	1.0	0.0	0.3	1.9
		Rare	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Supplementary Table 2.7 25th, 50th, and 75th percentiles of the length in kilobases (kb) of consecutively well-imputed (observed imputation $r^2>0.8$) variants by reference panel, genotype array, ancestry, and minor allele frequency (MAF) category

Reference panel	Array	Impact	African	Hispanic/Latino	European	Finnish
TOPMed	Omni 2.5M	High	0.605	0.641	0.511	0.456
		Moderate	0.645	0.668	0.540	0.510
		Low	0.706	0.726	0.603	0.620
		Modifier	0.719	0.739	0.621	0.662
	MEGA	High	0.551	0.593	0.443	0.387
		Moderate	0.571	0.608	0.448	0.416
		Low	0.619	0.658	0.499	0.521
		Modifier	0.648	0.683	0.532	0.584
	OmniExpress	High	0.589	0.622	0.483	0.430
		Moderate	0.626	0.651	0.508	0.480
		Low	0.684	0.707	0.568	0.589
		Modifier	0.699	0.721	0.589	0.635
	Core	High	0.551	0.593	0.443	0.387
		Moderate	0.571	0.608	0.448	0.416
		Low	0.619	0.658	0.499	0.521
		Modifier	0.648	0.683	0.532	0.584
HRC	Omni 2.5M	High	0.151	0.181	0.191	0.392
		Moderate	0.163	0.191	0.204	0.442
		Low	0.222	0.259	0.280	0.555
		Modifier	0.235	0.271	0.305	0.582
	MEGA	High	0.173	0.210	0.229	0.420
		Moderate	0.209	0.251	0.268	0.485
		Low	0.202	0.238	0.271	0.555
		Modifier	0.210	0.246	0.290	0.579
	OmniExpress	High	0.123	0.147	0.166	0.382
		Moderate	0.129	0.154	0.175	0.430
		Low	0.174	0.207	0.243	0.542
		Modifier	0.188	0.224	0.270	0.571
	Core	High	0.081	0.107	0.139	0.368
		Moderate	0.074	0.099	0.133	0.406
		Low	0.099	0.135	0.189	0.518
		Modifier	0.119	0.159	0.220	0.554
1000G	Omni 2.5M	High	0.162	0.200	0.148	0.282
		Moderate	0.170	0.208	0.152	0.314
		Low	0.240	0.284	0.231	0.437
		Modifier	0.263	0.307	0.261	0.478
	MEGA	High	0.187	0.233	0.208	0.351
		Moderate	0.225	0.276	0.250	0.432
		Low	0.216	0.262	0.224	0.438
		Modifier	0.232	0.277	0.246	0.469
	OmniExpress	High	0.128	0.165	0.125	0.257
		Moderate	0.131	0.170	0.126	0.286
		Low	0.186	0.232	0.194	0.405
		Modifier	0.208	0.256	0.225	0.450
	Core	High	0.081	0.120	0.104	0.235
		Moderate	0.071	0.110	0.089	0.238
		Low	0.103	0.152	0.142	0.346
		Modifier	0.129	0.183	0.176	0.404

Supplementary Table 2.8 Proportion of biallelic single nucleotide variants (SNVs) in each whole genome sequencing (WGS) study that are well-imputed ($r^2 > 0.8$) by reference panel, genotype array, ancestry, and predicted impact on protein coding

Predicted impact was estimated with VEP.

Array	Variant type	African	Hispanic/ Latino	European	Finnish
Omni 2.5M	Biallelic SNV	0.0014	0.0011	0.0035	0.0084
	Biallelic indel	0.0024	0.0014	0.0045	0.0115
	Multiallelic SNV	0.0016	0.0014	0.0045	0.0115
	Multiallelic indel	0.0055	0.0035	0.0075	0.0144
MEGA	Biallelic SNV	0.0017	0.0011	0.0045	0.0095
	Biallelic indel	0.0024	0.0014	0.0055	0.0126
	Multiallelic SNV	0.0024	0.0014	0.0055	0.0126
	Multiallelic indel	0.0065	0.0045	0.0105	0.0165
OmniExpress	Biallelic SNV	0.0024	0.0014	0.0095	0.0126
	Biallelic indel	0.0035	0.0019	0.0115	0.0165
	Multiallelic SNV	0.0035	0.0016	0.0115	0.0164
	Multiallelic indel	0.0074	0.0045	0.0145	0.0165
Core	Biallelic SNV	0.0084	0.0035	0.0395	0.0275
	Biallelic indel	0.0115	0.0045	0.0425	0.0336
	Multiallelic SNV	0.0105	0.0035	0.0405	0.0224
	Multiallelic indel	0.0185	0.0075	0.0284	0.0384

Supplementary Table 2.9 Minor allele frequency (MAF) threshold above which array genotyping and imputation can approximate whole genome sequencing (WGS) with the TOPMed panel by genotype array, ancestry, and variant type

Threshold is the smallest MAF for which >90% of variants are well-imputed (observed imputation $r^2 > 0.8$).

Chapter 3 Statistical Methods for Genetic Colocalization in a Single Cohort Design

3.1 Introduction

Genome-wide association studies (GWAS) have successfully identified hundreds of thousands of genetic associations with complex traits and diseases.² However, determining the molecular mechanisms underlying most GWAS associations remains challenging.⁴ First, linkage disequilibrium (LD), or the nonrandom association of alleles at different loci,⁸⁴ limits our ability to pinpoint the causal variant(s) when they are inherited together with non-causal variants. Second, the vast majority of GWAS variants lie in noncoding regions of the genome.⁸⁵ These variants are thought to influence traits through the regulation of gene expression, but it is often unclear which gene(s) they regulate.⁴ Determining the genes causally related to complex traits and how they are impacted by genetic variation is essential for understanding diseases pathways and paves the way for downstream treatment and prevention efforts.

Genetic colocalization analyses seek to identify genetic variants or loci that are causal for multiple association signals.⁸⁶ The identification of genetic variants that are causally related to both a complex disease and an intermediate molecular phenotype (e.g. gene expression, DNA methylation, metabolite levels) can provide evidence for target genes or tissues to prioritize in follow-up studies.²² Colocalization analyses have improved our understanding of the molecular pathways of anthropometric traits,⁸⁷ circulating biomarker levels,⁸⁸ cardiometabolic diseases,^{6,89,90} and autoimmune disorders,^{91,92} among many others.

Colocalization analysis relies on the accurate estimation of LD between genetic variants. Existing colocalization methods implicitly assume that the patterns of LD match perfectly

between any datasets used in the marginal association analyses of each trait. When this assumption is violated, as is nearly always the case when the complex disease trait and the molecular phenotype are measured in separate cohorts, the power of colocalization analysis is diminished.²³ As the cost of high-throughput methods for assaying molecular data decreases, an increasing number of studies have measured multiple phenotypes on the same set of individuals. For example, the UK Biobank has collected plasma proteomic data on tens of thousands of individuals for whom disease information is readily available.⁹³ In addition, a number of smaller, disease-focused studies have collected multiple forms of molecular data on the same individuals. Examples include the METSIM Study, which has measured more than 1,300 metabolites on over 6,000 individuals⁸⁸ and the FUSION Study, which has measured single nucleus resolution skeletal muscle gene expression and chromatin accessibility on nearly 300 individuals (see Chapter 4). This one cohort, multi-phenotype design has the potential to facilitate more powerful colocalization analyses that further our understanding of the impact of genetic variation on many types of traits.

However, existing probabilistic tools (e.g. coloc,⁸⁶ eCAVIAR,²² enloc/fastENLOC^{94,95}) for colocalization also explicitly assume that the marginal association analyses are performed in separate cohorts with non-overlapping sets of individuals. For the corresponding analysis methods, this assumption appears in the estimation of the likelihood function, where GWAS and eQTL results (or association results of any two traits) are treated as independent.²³ The effects of this model misspecification were found to be largely ignorable in the case of colocalization analysis of ten traits with HyPrColoc,⁹⁶ but the extent to which these conclusions apply to the colocalization of other sets of traits or to scenarios with different correlation structure between traits has not been investigated.

Here, we present simulation analyses that demonstrate the consequences of using existing two-sample methods in colocalization analysis of two phenotypes from a one-sample design, in violation of the non-overlapping cohorts assumption. We show that Type I error is well-controlled when the ratio of trait-shared to trait-specific error variance is low but increases with increased sharing. For scenarios with well-controlled Type I error, we show that the one-sample design can be more powerful than the two-sample design due to better LD matching. Power can be further improved in the one-sample design when trait-shared non-genetic factors are measured and controlled for in the marginal association analyses. Our findings provide practical guidelines for researchers to appropriately perform colocalization analysis with existing tools in a single cohort and provide a performance benchmark for future methods that accommodate this more powerful design.

3.2 Methods

3.2.1 Data resources and processing

3.2.1.1 METSIM whole genome sequence data

A detailed description of sample collection, sequencing, and data processing can be found in Laakso et al 2017⁵³ and Yin et al 2022.²⁴ Briefly, the METSIM Study comprises 10,197 male Finish participants aged 45-74 who were recruited in Kuopio, Finland. Whole genome sequencing (wave 1) was performed in a subset of 3,074 participants to an average depth of 23X. Genetic variants with missingness >2%, HWE p-value <10⁻⁶, or allele imbalance <30% were removed.

3.2.2 Simulations

3.2.2.1 Simulation of continuous phenotype for one-sample design

We randomly sampled 1,000 individuals from the METSIM WGS dataset. To represent fine-mapping of a single locus, we used genotypes of 5,000 consecutive common (minor allele frequency >5%) SNPs (16.9 Mb) on chromosome 20. We simulated two continuous traits, Y_1 and Y_2 , which could represent gene expression and a quantitative GWAS trait. We first simulated Y_1 for the 1,000 individuals as follows:

Equation 3.1

$$Y_{1i} = \beta_1 g_{1i} + \beta_2 g_{2i} + \beta_3 g_{3i} + u_i + e_{1i}$$

where g_1, g_2, g_3 are the causal SNPs for Y_1 selected randomly without replacement from the 5,000 SNPs and $\beta_1, \beta_2, \beta_3 \sim N(0, V = 0.6)$ are the Y_1 -specific genetic effects. The non-genetic error term is partitioned into the Y_1 -specific error $e_1 \sim N(0, 1)$ and confounder $u \sim N(0, \phi^2)$; $\phi^2 = \{0, 0.5, 1, 2\}$ that affects both Y_1 and Y_2 (Equation 3.2, Equation 3.3, Equation 3.4). Larger ϕ^2 values represent larger non-genetic shared effects between Y_1 and Y_2 .

We then simulated Y_2 under three different colocalizations scenarios: null, pleiotropy, and mediator. Under the null scenario, we simulated Y_2 as follows:

Equation 3.2

$$Y_{2i} = \beta_4 g_{4i} + \beta_5 g_{5i} + \beta_6 g_{6i} + u_i + e_{2i}$$

where g_4, g_5, g_6 are the causal SNPs for Y_2 selected randomly without replacement from the 4,997 SNPs that are not causal for Y_1 , $\beta_4, \beta_5, \beta_6 \sim N(0, 0.6)$ are the Y_2 -specific genetic effects, and $e_2 \sim N(0, 1)$ is the Y_2 -specific error. We simulated Y_1 and Y_2 750 times under the null scenario. Under the pleiotropy scenario, we simulated Y_2 as follows:

Equation 3.3

$$Y_{2i} = \beta_4 g_{3i} + \beta_5 g_{5i} + \beta_6 g_{6i} + u_i + e_{2i}$$

where g_3 represents a colocalized SNP that is causal for both Y_1 and Y_2 , g_5, g_6 are causal SNPs for only Y_2 selected randomly without replacement from the 4,997 SNPs that are not causal for Y_1 , $\beta_4, \dots, \beta_6 \sim N(0, 0.6)$ are the Y_2 -specific genetic effects, and $e_2 \sim N(0, 1)$ is the Y_2 -specific error. We simulated Y_1 and Y_2 200 times under the pleiotropy scenario. Under the mediator scenario, we simulated Y_2 as follows:

Equation 3.4

$$Y_{2i} = \alpha Y_{1i} + u_i + e_{2i}$$

where $\alpha \sim N\left(0, \frac{1}{9}\right)$, so that all 3 causal SNPs for Y_1 are also causal for Y_2 , and $e_2 \sim N(0, 1)$ is the Y_2 -specific error. We simulated Y_1 and Y_2 50 times under the mediator scenario.

3.2.2.2 Multi-SNP association analysis and fine-mapping

We combined the 750 null, 200 pleiotropy, and 50 mediator simulations into one dataset to represent a true colocalization enrichment, defined as the log odds ratio quantifying the enrichment of eQTLs in GWAS signals, of 5.52. For each simulation, we used DAP-G⁹⁷ to perform multi-SNP association analysis and fine-mapping separately for Y_1 and Y_2 using an LD control setting of 0.5. We considered clusters with posterior inclusion probabilities greater than 0.95 to be causal for the trait. Clusters that did not contain a true causal variant were considered false positives. We calculated the false positive rate and power (number of true positive clusters divided by 3,000 truly causal SNPs) across all simulations separately for Y_1 and Y_2 . We repeated

the above association and fine-mapping analyses adjusting for the true simulated u in the marginal association analyses of both Y_1 and Y_2 (known confounder adjustment).

3.2.2.3 Probabilistic principal component analysis and adjustment

We first performed fine-mapping with DAP-G without adjusting for covariates as above and extracted the estimated transcriptome-wide association study (TWAS) weights $w_{1,...,5000}$ (Y_1) and $x_{1,...,5000}$ (Y_2) for all SNPs. To remove the estimated genetic effect from Y_1 and Y_2 , we calculated pseudo-residuals as follows:

Equation 3.5

$$r_1 = y_1 - \sum_{p=1}^{5,000} g_p w_p$$

Equation 3.6

$$r_2 = Y_2 - \sum_{p=1}^{5,000} g_p x_p$$

We then calculated the first principal component from r_1 and r_2 . We repeated the association and fine-mapping procedures for both Y_1 and Y_2 adjusting for the principal component in each.

3.2.2.4 Ridge regression on probabilistic principal components and adjustment

We performed ridge regression with all genotypes as covariates and the first principal component calculated above as the outcome. We then performed the association and fine-mapping procedures for Y_1 and Y_2 adjusting for the residuals from this ridge regression. We repeated the ridge regression three times using three parameter values (0.005, 0.01, and 0.05) for the lambda tuning parameter. We chose these tuning parameter values to maximize the amount of genetic information removed from Y_1 and Y_2 .

3.2.2.5 Colocalization

We performed colocalization analysis using fastENLOC.⁹⁵ We calculated Bayesian FDR from the cluster-level regional colocalization probability and considered clusters with $\text{FDR} < 5\%$ to be colocalized. Clusters with $\text{FDR} < 5\%$ that did not contain a variant that was causal for both traits in the simulation were considered false positives. We calculated the false positive rate and power (number of true positives divided by 350 truly colocalized SNPs) across all simulations.

3.3 Results

3.3.1 One-sample colocalization design

We simulated continuous phenotypes Y_1 and Y_2 from real sequenced-based genotype data from the METSIM Study.⁸⁸ Here, we will refer to Y_1 as gene expression level and Y_2 as a quantitative complex trait, but they could represent any two continuous phenotypes. In both one- and two-sample designs, we assume that SNP G influences Y_1 and Y_2 with effect sizes β_1 and β_2 , respectively, and we are interested in estimating the joint causal status d of G on Y_1 and γ of G on Y_2 (Figure 3.1). In the two-sample design, Y_1 and Y_2 are measured in non-overlapping samples such that the non-genetic error terms ϵ_1 and ϵ_2 represent both environmental effects and cohort-specific random error on Y_1 and Y_2 , respectively (Figure 3.1A). In the one-sample design, because Y_1 and Y_2 are measured on the same set of samples, the non-genetic error terms can be partitioned into a shared confounder u and trait-specific error terms e_1 and e_2 (Figure 3.1B). To assess the effects of the magnitude of u on colocalization error rates, we simulated 1,000 independent genes on 1,000 samples assuming a one-sample design for a range of confounder variance magnitudes denoted ϕ^2 (see Methods). In the simulations, we included genes with a range of shared causal SNPs from 0 (null) to 1 (pleiotropy) to 3 (mediator). These scenarios

represent our assumption that most genes will not have a colocalized variant (null), some genes will have a colocalized variant that affects both Y_1 and Y_2 but the gene expression level does not mediate the association between the colocalized variant and Y_2 (pleiotropy), and that the gene expression level of a few genes mediates the association between the genetic variants and Y_2 (mediator). The correlation of the simulated phenotypes increased with higher values of ϕ^2 and more shared causal SNPs between Y_1 and Y_2 (Supplementary Figure 3.1).

3.3.2 Two-sample colocalization in one-sample design

To evaluate the consequences of applying existing two-sample methods for colocalization of a one-sample design, we performed fine-mapping analysis with DAP-G⁹⁷ and colocalization analysis with fastENLOC⁹⁵ on the simulated data. The fine-mapping FDR was <0.007 for all values of ϕ^2 and the power similarly decreased with higher values of ϕ^2 , ranging from 0.701 (eQTL, $\phi^2 = 0$) to 0.474 (GWAS, $\phi^2 = 2$) (Supplementary Table 3.2). We found that the colocalization enrichment parameter was accurately estimated for all tested values of the confounder magnitude ϕ^2 (Figure 3.2). The colocalization FDR increased with higher values of ϕ^2 , such that the FDR was well-controlled at the 0.05 level only when $\phi^2 < 0.5$ (Figure 3.2B, Supplementary Table 3.1). The colocalization power decreased with higher values of ϕ^2 , ranging from 0.520 ($\phi^2 = 0$) to 0.266 ($\phi^2 = 2$) (Figure 3.2C).

3.3.3 Adjusting for true trait-shared confounder reduces Type I and Type II error rates

To establish an upper bound for single-cohort colocalization performance, we repeated the fine-mapping step adjusting for the true simulated confounder u . We found that the colocalization enrichment parameter was accurately estimated (Figure 3.3A) and that the colocalization FDR was well-controlled at the 0.05 level for all values of the confounder

magnitude ϕ^2 (Figure 3.3B, Supplementary Table 3.3). The power of the adjusted colocalization was higher than the standard two-sample colocalization analysis for all values of ϕ^2 , and this improvement was greatest for large values of ϕ^2 . For example, when $\phi^2 = 2$, the adjusted colocalization power was 0.440 compared to the unadjusted colocalization power 0.266. As expected, the adjusted fine-mapping FDR was similar to the unadjusted fine-mapping FDR but the adjusted fine-mapping power was higher than the unadjusted fine-mapping power because we were controlling for an additional associated variable in the association analysis (Supplementary Table 3.4).

3.3.4 Estimating trait-shared confounder with probabilistic principal component analysis introduces collider bias

Because adjusting for the true confounder u successfully reduced colocalization Type I and Type II error, we attempted to estimate u with principal probabilistic principal component analysis (PPCA) for scenarios in which u was unmeasured to first remove the genetic effect and then estimate the proportion of remaining variance shared between the two traits. The correlation between the estimates from PPCA and the true simulated u increased with higher ϕ^2 ; the mean squared Pearson correlation for $\phi^2 = 0.5$ was 0.48 and for $\phi^2 = 2$ was 0.79 (Supplementary Figure 3.2). Despite this high correlation, we found that performing colocalization from the PPCA-adjusted fine-mapping output resulted in an inflated enrichment parameter estimate (Figure 3.3A) and extremely high FDR, primarily due to trait-specific causal variants being inferred as causal variants for the other trait for which they were not causal (Figure 3.3B, Supplementary Table 3.5). The inflated false positive rate was also present at the fine-mapping stage (Supplementary Table 3.6). We hypothesize that our procedure for estimating u introduced collider bias because we conditioned on Y_1 and Y_2 , thus creating a path from the Y_1 -specific

genetic components to the Y_2 -specific genetic components in the directed acyclic graph (Figure 3.1B), which is supported by the high proportion of trait-specific causal variants falsely inferred as colocalized.

3.3.5 Regressing SNP effects on estimated confounder offers no improvements over two-sample methods for one-sample design

To reduce the effects of the collider bias introduced by the PPCA procedure, we used ridge regression in an attempt to regress out all genetic effects from the estimated principal component. We considered a range of shrinkage tuning parameters λ (0.05, 0.01, 0.005), all close to 0 to enable near-complete removal of the genetic-effects. We considered the residuals from this regression analysis as another potential estimator of u . This estimator was less correlated with u than the principal component, and the degree of correlation was lower for smaller values of ϕ^2 and λ (Supplementary Figure 3.3). For $\lambda=0.01$, the mean squared Pearson correlation for $\phi^2 = 0.5$ was 0.07 and for $\phi^2 = 2$ was 0.08. For all values of λ , the colocalization enrichment parameter was well-estimated (Figure 3.3A). The colocalization FDR increased with larger values of λ and was higher than the unadjusted colocalization FDR for all values of ϕ^2 (Figure 3.3B, Supplementary Table 3.7). The colocalization power was similar to the unadjusted colocalization power for all values of ϕ^2 (Figure 3.3C). The fine-mapping FDR also increased with larger values of λ and ϕ^2 and did not show improvement over unadjusted fine-mapping analysis (Supplementary Table 3.8).

3.4 Discussion

We presented simulation analyses to show the consequences of applying two-sample methods for colocalization in a single cohort, in violation of the non-overlapping cohorts

assumption. We found that Type I error is well-controlled when the phenotypes do not share large non-genetic effects or when these effects are measured and controlled for. Although we were able to accurately estimate large non-genetic effects, we found that adjusting for these estimated confounders introduced collider bias and led to inflated fine-mapping and colocalization Type I error rates.

Colocalization and fine-mapping analyses are often underpowered to detect most causal SNPs with modest effect sizes.⁹⁸ The single-cohort design for colocalization presents an opportunity to improve colocalization power through perfect LD matching and the opportunity to account for shared non-genetic factors between traits. Many such confounders with potentially large effects on both molecular phenotypes and complex traits (e.g. sex, age) are commonly measured in genetic studies and can be adjusted for in fine-mapping and colocalization analyses to achieve this reduction in Type II error.

We made several simplifying assumptions in our simulation study. First, we assumed a SNP heritability of 0.38-0.64 depending on the confounder variance magnitude for the eQTL analysis. This is higher than the heritability normally observed in cis-eQTL studies⁹⁹⁻¹⁰¹ but enabled more powerful fine-mapping analyses to evaluate changes in colocalization accuracy by degree of shared confounding. We additionally assumed the same heritability for the eQTL and GWAS trait in the null and pleiotropy simulations, which may be unlikely as gene expression levels are mechanistically closer to genetic variation and therefore are likely to be more heritable.⁹⁹ We also assumed that all error terms followed a normal distribution centered at 0, which is a common assumption of linear regression.

Here, we treated genes as independent simulations and did not allow for any non-genetic factors (u , e_1 , e_2) to be shared across genes. If we had instead assumed that the confounder u

impacts many genes in addition to the complex trait, it is possible that it could be estimated and adjusted for more accurately in the marginal eQTL analysis. Many methods, such as PEER¹⁰² and SVA,¹⁰³ perform this type of estimation. However, because PEER factors also have the potential to introduce collider bias,¹⁰⁴ future work is needed to evaluate the effects of this adjustment on colocalization error rates in a single-cohort design.

Because colocalization Type I error rates are positively associated with the degree of shared confounding, we recommend that researchers performing colocalization with overlapping samples attempt to detect such confounding following the probabilistic PCA procedure we outlined. If PCA explains a large proportion of the variance between the TWAS-adjusted residuals from Y_1 and Y_2 , caution should be used in applying two-sample colocalization methods. Conversely, if there is little shared non-genetic variance or if confounding variables have been measured and adjusted for, two-sample methods for colocalization can be accurately applied in a single-sample design.

3.5 Tables and Figures

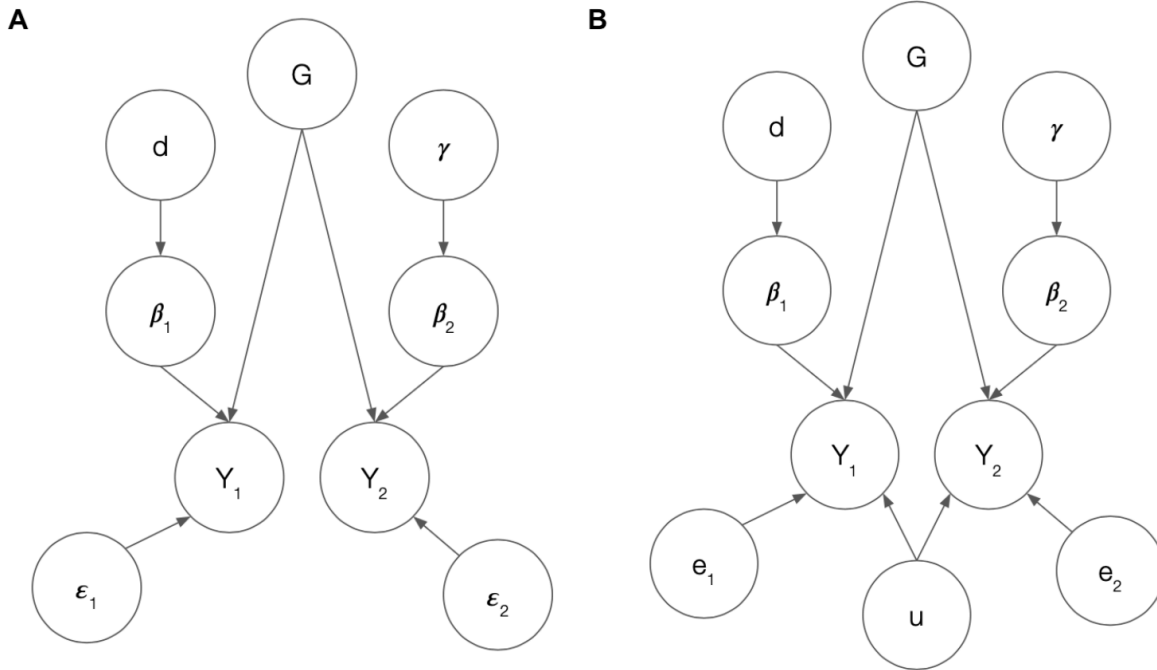


Figure 3.1 Directed acyclic graphs of colocalization in a two-sample and one-sample design

In both scenarios, we are interested in estimating the causal status of SNP G on the gene expression/ Y_1 (d) and GWAS trait/ Y_2 (γ). When $d = 1$, β_1 , the additive SNP effect on Y_1 , is nonzero. Similarly, when $\gamma = 1$, the additive SNP effect on Y_2 is nonzero. Y_1 and Y_2 are also affected by non-genetic error terms ϵ_1 and ϵ_2 . In the two-sample design (A), the degree of overlap between ϵ_1 and ϵ_2 is unknowable. In the single-sample design (B), ϵ_1 and ϵ_2 can be partitioned into a shared non-genetic confounder (u) and trait-specific error terms e_1 and e_2 .

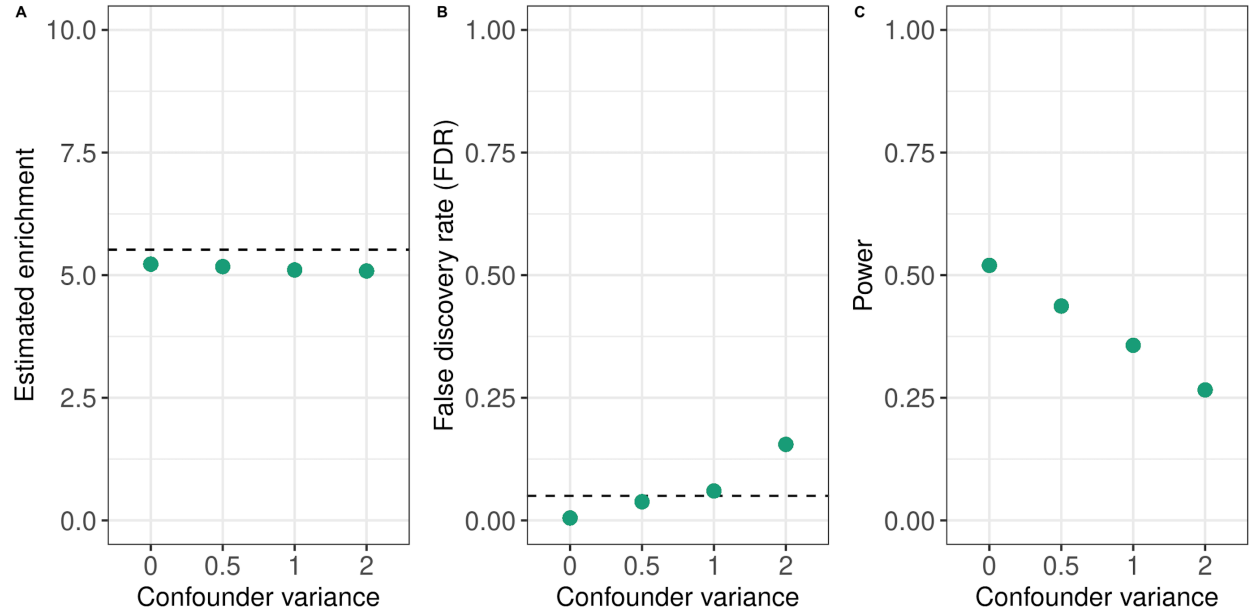


Figure 3.2 Two-sample colocalization in a single-cohort design

Colocalization results from fastENLOC applied to simulated data over a range of confounder (non-genetic error term shared across both traits) variance magnitudes ϕ^2 . We used a Bayesian FDR<5% threshold calculated on the regional colocalization probabilities (RCP) to determine significant clusters of variants across 1,000 simulations. We considered variants in significant clusters to be colocalized. A) The fastENLOC shrinkage-based enrichment parameter of eQTL (Y_1 -associated) variants in GWAS (Y_2 -associated) hits. The dotted line corresponds to the true simulated enrichment parameter (5.52) across the 1,000 simulations. B) The colocalization false discovery rate (FDR) calculated as the number of colocalized variants that were not simulated as causal for both traits (false positives) divided by the total number of colocalized variants across 1,000 simulations. The dotted line corresponds to 5% FDR. C) The colocalization power calculated as the number of colocalized variants that were simulated to be causal for both traits (true positives) divided by 250, the true number of colocalized variants across 1,000 simulations.

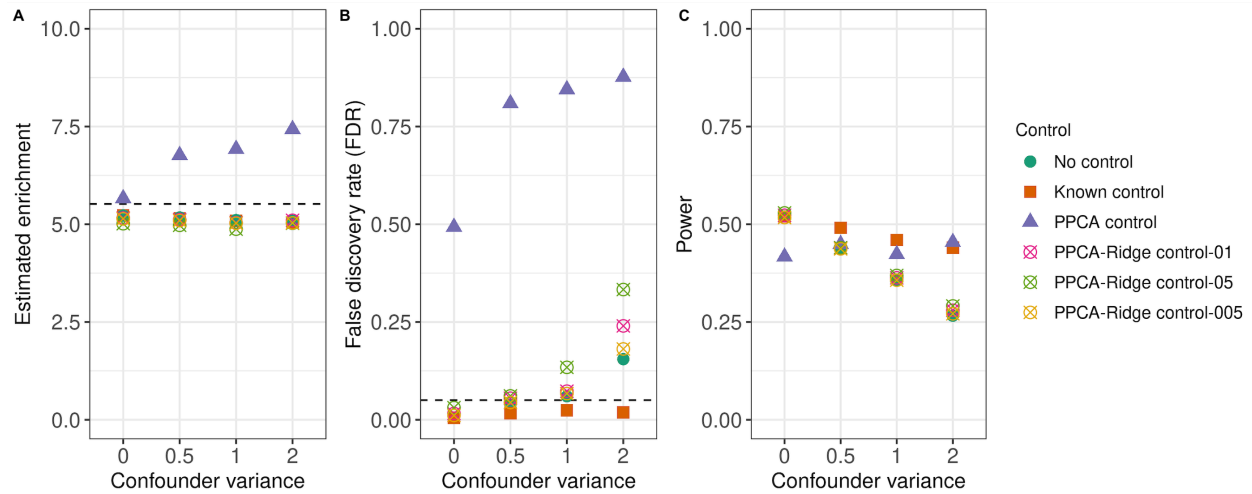
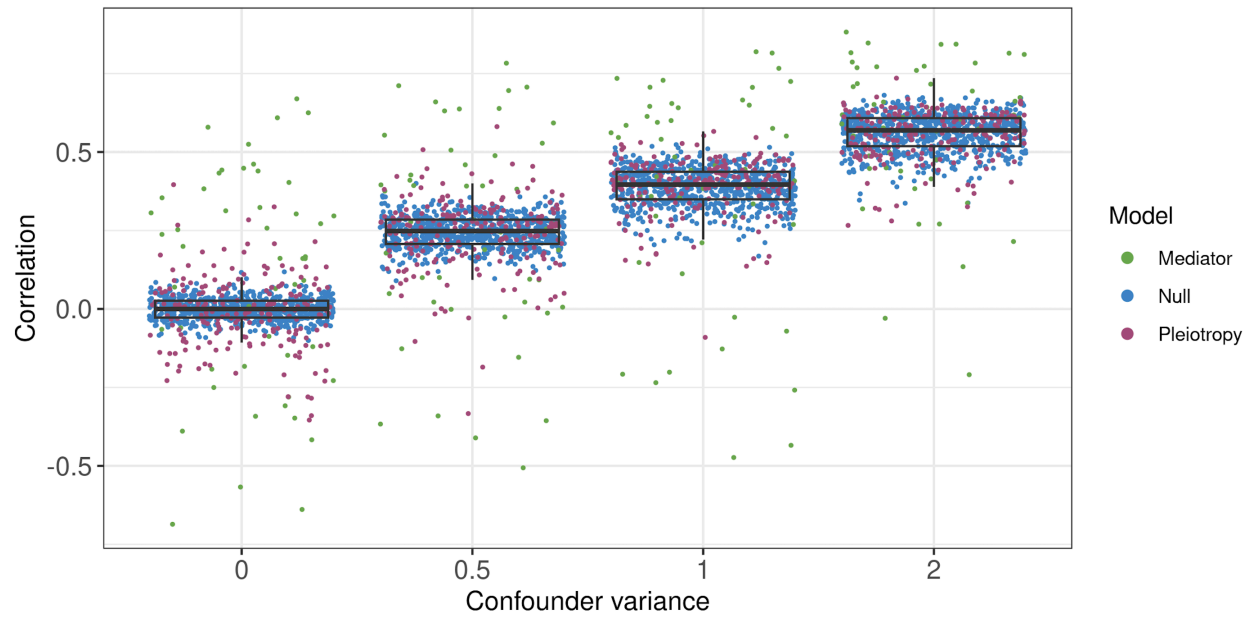


Figure 3.3 Colocalization from adjusted marginal association analyses in a single-cohort design

FastENLOC colocalization results from adjusted fine-mapping results applied to simulated data over a range of confounder (non-genetic error term shared across both traits) variance magnitudes ϕ^2 . The color and shape of points corresponds to the fine-mapping adjustment strategy: known control (adjusting for the true confounder u), PPCA control, and PPCA-Ridge control with lambda parameter values 0.005, 0.01, and 0.05. The unadjusted analysis (green solid circles) shown in Figure 3.2 is included for comparison. A) The fastENLOC shrinkage-based enrichment parameter of eQTL (Y_1 -associated) variants in GWAS (Y_2 -associated) hits. The dotted line corresponds to the true simulated enrichment parameter (5.52) across the 1,000 simulations. B) The colocalization false discovery rate (FDR) calculated as the number of colocalized variants that were not simulated as causal for both traits (false positives) divided by the total number of colocalized variants across 1,000 simulations. The dotted line corresponds to 5% FDR. C) The colocalization power calculated as the number of colocalized variants that were simulated to be causal for both traits (true positives) divided by 250, the true number of colocalized variants across 1,000 simulations.

3.6 Supplementary Material



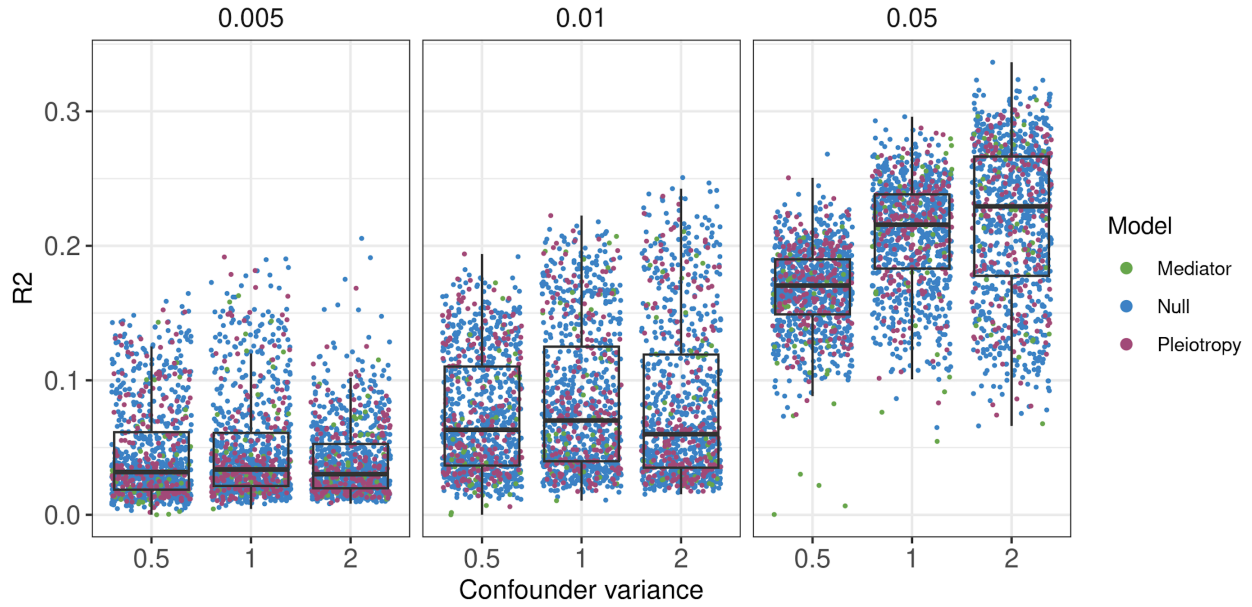
Supplementary Figure 3.1 Correlation of simulated continuous phenotypes

Boxplots showing the distributions across 1,000 simulations of the Pearson correlations between the simulated gene expression (Y_1) and complex trait (Y_2) phenotypes calculated across 1,000 individuals by the magnitude of the confounder (non-genetic error term shared across both traits) variance. This variance, denoted ϕ^2 , ranged from 0 (no shared non-genetic term) to 2 (twice the magnitude of the trait-specific error variance). Points are colored by simulated colocalization scenario: null ($n=750$), pleiotropy ($n=200$), and mediator ($n=50$).



Supplementary Figure 3.2 Estimation of the trait-shared, non-genetic confounder with probabilistic principal component analysis (PPCA)

Boxplots showing the distributions across 1,000 simulations of the squared Pearson correlation between the first principal component estimated from the TWAS-adjusted Y_1 and Y_2 pseudo-residuals and the true simulated confounder by confounder variance ϕ^2 . Points are colored by simulated colocalization scenario: null ($n=750$), pleiotropy ($n=200$), and mediator ($n=50$).



Supplementary Figure 3.3 Estimation of the trait-shared, non-genetic confounder with ridge regression residuals from probabilistic principal component analysis (PPCA)

Boxplots showing the distributions across 1,000 simulations of the squared Pearson correlation between the residuals from ridge regression performed by regressing all genotypes on the first principal component estimated from the TWAS-adjusted Y_1 and Y_2 pseudo-residuals and the true simulated confounder. Boxplots are shown by confounder variance ϕ^2 for a range of ridge regression lambda values (0.005, 0.01, and 0.05). Points are colored by simulated colocalization scenario: null ($n=750$), pleiotropy ($n=200$), and mediator ($n=50$).

Confounder variance ϕ^2	Number of colocalized variants		
	True positives	False positives	Total colocalized
0	182	1	183
0.5	153	6	159
1	125	8	133
2	93	17	110

Supplementary Table 3.1 Unadjusted fastENLOC results

The number of true and false positive colocalized variants from fastENLOC analysis applied to simulated data in a single cohort design by the magnitude of the simulated confounder variance

Confounder variance ϕ^2	eQTL (Y_1) fine-mapping			GWAS (Y_2) fine-mapping		
	True positives	False positives	Total fine-mapped	True positives	False positives	Total fine-mapped
0	2,103	3	2,106	2,001	3	2,004
0.5	1,919	5	1,924	1,876	7	1,883
1	1,820	7	1,827	1,653	11	1,664
2	1,519	4	1,523	1,421	9	1,430

Supplementary Table 3.2 Unadjusted DAP-G results

The number of true and false positive fine-mapped variants from DAP-G analysis applied to simulated data in a single cohort design. Fine-mapping results are shown separately for eQTL (Y_1 -associated) and GWAS (Y_2 -associated) variants for a range of confounder variance magnitude values.

Confounder variance ϕ^2	Number of colocated variants		
	True positives	False positives	Total colocated
0	182	1	183
0.5	172	3	175
1	161	4	165
2	154	3	157

Supplementary Table 3.3 Confounder-adjusted fastENLOC results

The number of true and false positive colocated variants from fastENLOC analysis of the true confounder adjusted marginal associations by the magnitude of the simulated confounder variance

Confounder variance ϕ^2	eQTL (Y_1) fine-mapping			GWAS (Y_2) fine-mapping		
	True positives	False positives	Total fine-mapped	True positives	False positives	Total fine-mapped
0	2,103	3	2,106	2,001	3	2,004
0.5	2,107	3	2,110	2,046	5	2,051
1	2,137	3	2,140	2,014	6	2,020
2	2,072	5	2,077	2,019	4	2,023

Supplementary Table 3.4 Confounder-adjusted DAP-G results

The number of true and false positive fine-mapped variants from confounder-adjusted DAP-G analysis applied to simulated data in a single cohort design. DAP-G analysis was performed adjusting for the true simulated confounder. Fine-mapping results are shown separately for eQTL (Y_1 -associated) and GWAS (Y_2 -associated) variants for a range of confounder variance magnitude values.

Confounder variance ϕ^2	Number of colocated variants			
	True positives	False positives-collider	False positives - other	Total colocated
0	146	93	49	288
0.5	157	509	155	821
1	148	647	160	955
2	159	909	388	1,297

Supplementary Table 3.5 Probabilistic principal component analysis (PPCA)-adjusted fastENLOC results

The number of true and false positive colocated variants from fastENLOC analysis of the PPCA-adjusted marginal associations by the magnitude of the simulated confounder variance. False positives were divided into those due to collider bias (causal for only one trait in the simulation) and others.

Confounder variance ϕ^2	eQTL (Y_1) fine-mapping			GWAS (Y_2) fine-mapping		
	True positives	False positives	Total fine-mapped	True positives	False positives	Total fine-mapped
0	2,134	95	2,229	1,736	71	1,807
0.5	2,048	357	2,405	1,889	31	1,920
1	2,014	343	2,357	1,792	50	1,842
2	1,860	433	2,293	1,703	111	1,814

Supplementary Table 3.6 Probabilistic principal component analysis (PPCA)-adjusted DAP-G results

The number of true and false positive fine-mapped variants from probabilistic principal component analysis (PPCA)-adjusted DAP-G analysis applied to simulated data in a single cohort design. DAP-G analysis was performed adjusting for the first principal component estimated from the TWAS-adjusted Y_1 and Y_2 pseudo-residuals. Fine-mapping results are shown separately for eQTL (Y_1 -associated) and GWAS (Y_2 -associated) variants for a range of confounder variance magnitude values.

Ridge regression lambda	Confounder variance ϕ^2	Number of colocized variants		
		True positives	False positives	Total colocized
0.005	0	181	2	183
	0.5	153	7	160
	1	125	9	134
	2	95	21	116
0.01	0	183	3	186
	0.5	154	9	163
	1	127	10	137
	2	98	31	129
0.05	0	187	6	191
	0.5	154	10	164
	1	129	20	149
	2	102	51	153

Supplementary Table 3.7 Ridge regression residual-adjusted fastENLOC results

The number of true and false positive colocized variants from fastENLOC analysis of the ridge regression-adjusted marginal associations by the lambda parameter from the ridge regression and the magnitude of the simulated confounder variance.

Ridge regression lambda	Confounder variance ϕ^2	eQTL (Y_1) fine-mapping			GWAS (Y_2) fine-mapping		
		True positives	False positives	Total fine-mapped	True positives	False positives	Total fine-mapped
0.005	0	2,120	5	2,125	2,025	10	2,035
	0.5	1,942	7	1,949	1,910	10	1,920
	1	1,841	8	1,849	1,686	14	1,700
	2	1,555	4	1,559	1,449	10	1,459
0.01	0	2,126	8	2,134	2,046	14	2,060
	0.5	1,961	12	1,973	1,926	11	1,937
	1	1,861	9	1,870	1,719	15	1,734
	2	1,581	5	1,586	1,476	16	1,493
0.05	0	2,150	19	2,169	2,100	43	2,143
	0.5	2,020	27	2,047	1,979	28	2,007
	1	1,939	24	1,963	1,795	33	1,828
	2	1,672	27	1,699	1,570	41	1,611

Supplementary Table 3.8 Ridge regression residual-adjusted DAP-G results

The number of true and false positive fine-mapped variants from ridge regression residual-adjusted DAP-G analysis applied to simulated data in a single cohort design. DAP-G analysis was performed adjusting for residuals from ridge regression performed by regressing all genotypes on the first principal component estimated from the TWAS-adjusted Y_1 and Y_2 pseudo-residuals. Fine-mapping results are shown separately for eQTL (Y_1 -associated) and GWAS (Y_2 -associated) variants by ridge regression lambda parameter values and confounder variance magnitudes.

Chapter 4 Extensive Differential Gene Expression and Regulation by Sex in Human Skeletal Muscle

4.1 Introduction

Human skeletal muscle exhibits sex differences in its size, composition, physiology, and disease susceptibility. On average, male muscle is larger than female muscle, both in the cross-sectional area of individual muscle fibers¹⁰⁵ and in proportion to total body mass.¹⁰⁶ Male muscle typically has larger proportions of fast-twitch, glycolytic (Type 2) fibers compared to female muscle, whereas female muscle typically has larger proportions of slow-twitch, oxidative (Type 1) fibers.^{107,108} Male muscle usually shows greater contractile strength whereas female muscle is less fatigable and shows greater endurance.¹⁰⁷ There are also differences by sex in disease prevalence or progression of diseases such as type 2 diabetes,¹⁰⁹ obesity,¹¹⁰ cardiovascular disease,¹¹¹ and osteoporosis¹¹² for which skeletal muscle is a relevant tissue.^{113–}

116

While the molecular mechanisms leading to these physiological differences remain largely unknown, studies measuring gene expression with microarrays^{117–120} or RNA-seq^{40,121–125} have identified hundreds to thousands of genes that are differentially expressed by sex in bulk muscle tissue. Most gene expression studies conducted to date in human skeletal muscle have small sample sizes of no more than 30 participants^{117–121} or use postmortem samples from the GTEx consortium;^{40,122,123,125} all have used bulk skeletal muscle tissue. A subset of these studies^{40,118,120,121,123} performed gene set enrichment analyses¹²⁶ to identify biologically

meaningful sets of genes overrepresented in the differentially expressed genes in muscle, but the same pathways have not been consistently identified.

Sex differences in transcriptional and post-transcriptional regulation and cell type composition can contribute to the observed sex differences in gene expression in bulk muscle tissue. At the transcriptional level, muscle sex-biased genes are enriched for differential DNA methylation¹²⁴ and differential targeting by transcription factors, including sex hormone receptors.^{40,123} At the post-transcriptional level, microRNAs (miRNAs) have been suggested to regulate sex differences in gene expression,^{39,127} and 80 sex-biased miRNAs have been identified in human muscle tissue.¹²⁸ Sex differences in cell type composition may also explain some of the observed sex-biased genes in any analysis of bulk tissue,^{40,124} but the extent to which differential gene and miRNA expression is associated with cell type composition differences in muscle is currently unknown. Single nucleus RNA-seq (snRNA-seq), which measures RNA abundance in single nuclei, can help resolve this question.¹⁰

Here, we combine single nucleus and bulk gene expression, single nucleus chromatin accessibility, and bulk miRNA expression data measured in up to 281 vastus lateralis samples from living Finnish donors. We characterize sex differences in cell type composition and identify widespread differential gene expression and regulation in individual muscle cell types and in muscle tissue as a whole. These results further our understanding of the molecular basis of sex differences in skeletal muscle tissue and may help uncover mechanisms of sex differences in muscle physiology and disease.

4.2 Methods

4.2.1 Data collection and processing

4.2.1.1 *FUSION Tissue Biopsy Study*

The Finland-United States Investigation of NIDDM Genetics (FUSION) Tissue Biopsy Study is described in Scott et al.¹²⁹ Briefly, we obtained ~250mg vastus lateralis skeletal muscle biopsies using a conchotome from 331 living participants at three study sites in Finland (Helsinki, Kuopio, and Savitaipale) between 2009-2013. We cleaned the biopsies of non-muscle tissue and froze them within 30 seconds of sampling. All physicians were trained to perform the biopsy in an identical manner following a standardized protocol.

4.2.1.2 *Single nucleus RNA- and ATAC-seq*

We profiled snRNA-seq and snATAC-seq in 10 batches from 287 frozen muscle tissue biopsy samples. Across all samples, there were 180,583 RNA-seq nuclei and 268,543 ATAC-seq nuclei that passed droplet-level QC thresholds. We performed ambient RNA decontamination for RNA-seq droplets with DecontX.¹³⁰ We jointly clustered RNA-seq and ATAC-seq nuclei with Liger and annotated the clusters with known marker gene expression as adipocytes, endothelial cells, macrophages, mesenchymal stem cells, mixed muscle fiber, neuromuscular junction, neuronal, satellite cells, smooth muscle, T cells, Type 1 muscle fiber, Type 2a muscle fiber, and Type 2x muscle fiber. We excluded all mixed muscle fiber nuclei from analyses due to the high proportion of exonic reads. We excluded three samples with less than 100 RNA-seq nuclei or less than 100 ATAC-seq nuclei. We further excluded one sample from each of two pairs of first degree relatives, and one non-Finnish participant. After cell-type and sample QC, we retained 177,350 RNA-seq nuclei from 279 individuals and 252,219 ATAC-seq nuclei from 281 individuals. The 279 snRNA individuals were a subset of the 281 snATAC individuals (Supplementary Figure 4.1).

4.2.1.3 Bulk RNA-seq

Data collection and processing methods for the bulk RNA-seq are described in detail in Scott et al.¹²⁹ Briefly, we sequenced mRNA in 301 frozen muscle tissue biopsy samples to a mean depth of 91.3M strand-specific paired-end reads. Here, we analyzed 268 of these samples that also had single nucleus data to adjust for cell type proportions (Supplementary Figure 4.1).

4.2.1.4 Bulk small RNA-seq (miRNA)

We measured miRNA expression levels for 296 skeletal muscle tissue samples. The total RNA isolated for mRNA-sequencing was also used for miRNA isolation and sequencing. miRNA libraries were prepared at the NIH Intramural Sequencing Core (NISC) from 1 µg total RNA using Illumina's TruSeq Small RNA Library Kit according to the manufacturer's guidelines, except a 10% acrylamide gel was used to better separate the library from adapters. Libraries were pooled in groups of four to eight for gel purification. Single-end 51-base sequencing was performed on Illumina HiSeq 2500 sequencers in Rapid Mode using version 2 chemistry. We mapped miRNA sequence reads using the exceRpt¹³¹ pipeline (v4.4.0) with default parameters. We counted reads mapped to each miRNA of miRBase¹³² (version 21).

Because the same RNA extracts were used for both mRNA-seq and miRNA-seq, we excluded one sample identified as contaminated in mRNA-seq from mi-RNAseq analysis as well. We assessed the quality of each miRNA-seq dataset through metrics generated by exceRpt, including read length and library size, and did not observe outliers. After QC, 290 skeletal muscle tissue miRNA-seq samples remained for analysis. Here, we analyzed 256 of these samples that also had single nucleus data to be able to adjust for cell type proportions (Supplementary Figure 4.1).

4.2.2 Statistical analysis of single nucleus data

4.2.2.1 Differential cell type composition by sex

We used 279 FUSION samples with at least 100 nuclei from both RNA and ATAC modalities. For each cell type, we used negative binomial models to test for the association between the number of ATAC-seq and RNA-seq nuclei and sex, adjusting for age, batch, and city of collection, with an offset of the log of total nuclei across cell types. We corrected for multiple testing across the cell types by using a threshold of false discovery rate (FDR) <5%.

4.2.2.2 Differential gene expression by sex in muscle cell types

For each cell type, for samples with at least 10 nuclei, we analyzed genes with ≥ 1 count for at least 25% of the samples. We tested for the association between gene expression counts in the cell type (rounded to the nearest integer value) and sex using a negative binomial model as implemented in DESeq2 version 1.36.¹³³ For the DESeq2 analysis, we used the recommended single cell settings, including using a likelihood ratio test, setting the *minmu* parameter to 1×10^{-6} , which is appropriate for datasets with many genes with expected counts <1, and calculating single-cell specific size factors. We included batch, sample collection site, age, median mitochondrial fraction (across nuclei), and total RNA nuclei counts to account for differences in ability to isolate nuclei across all cell types as sample covariates. Quantitative covariates were inverse-normalized to limit influence of outlying values. We used a threshold of FDR <5% across all tested genes within each cell type for this analysis and all subsequent analyses.

4.2.2.3 Downsampling Type 1 fiber gene counts

We downsampled the gene counts from the Type 1 fiber to approximate the power to detect sex-biased expression in each of the other cell types. For a given cell type, we subset the individuals analyzed in Type 1 fiber to the same individuals analyzed in the cell type. We multiplied the Type 1 fiber gene UMIs by the fraction of total UMIs in the cell type divided by the total UMIs in Type 1, rounding the gene UMIs to the nearest integer for analysis with DESeq2, thereby approximating the total UMIs in the cell type. We then tested for differential expression by sex in the downsampled Type 1 dataset as described above without further excluding individuals or genes.

4.2.2.4 Differential gene expression in total single nucleus pseudobulk

We created a total single nucleus pseudobulk dataset by summing the gene counts for all the cell types for each of the 279 samples. We tested for association between total single nucleus pseudobulk gene expression counts and sex using DESeq2 version 1.36 with the recommended bulk RNA settings. We used the same covariates for pseudobulk as for the muscle cell types, with the addition of cell type proportion covariates. We analyzed genes with at least 5 counts in at least 25% of the 279 samples, parallel to our analysis of bulk tissue (below).

4.2.2.5 Gene type enrichment in muscle cell types

For each cell type, we tested for association between differential expression status and gene type with logistic regression. We defined gene types from GENCODE annotations and grouped genes into protein-coding genes, lncRNAs (3' overlapping ncRNA, antisense, bidirectional promoter lncRNA, lincRNA, macro lncRNA, non coding, processed transcript, sense intronic, and sense overlapping), pseudogenes (pseudogene, processed pseudogene, polymorphic pseudogene, transcribed processed pseudogene, transcribed unprocessed

pseudogene, transcribed unitary pseudogene, unitary pseudogene, and unprocessed pseudogene), and others (immunoglobulin genes, rRNA, miRNA, scaRNA, snoRNA, and those without annotations in GENCODE). We repeated the analyses adjusting for bins of mean UMI across all samples with breaks at 0, 1, 2, 3, 4, 5, 10, 50, 100, 500, 1,000, and 5,000 UMI to account for the greater power to detect differences by sex in genes with higher expression levels.

4.2.2.6 Gene set enrichment analysis in muscle fiber types

Separately for each of the three muscle fiber types, we identified Gene Ontology (GO) terms enriched for genes expressed more highly in males or females using RNA-Enrich.¹³⁴ For each GO term, RNA-Enrich uses a logistic regression model to test for association between GO term membership as the outcome and the signed $-\log_{10}$ p-value from the differential expression analysis, accounting for gene expression level. We considered all GO Biological Processes, Cellular Components, and Molecular Functions gene sets that contain between 10 and 1,000 genes to focus on terms that contain more than one sex-biased gene and that do not represent non-specific pathways. To reduce the impact of outliers on the analysis, we inverse normalized p-values on the $-\log_{10}$ p-value scale prior to running RNA-Enrich. We used Revigo¹³⁵ to prune and select non-overlapping GO terms in Figure 2D.

4.2.2.7 Differential chromatin accessibility by sex in muscle cell types

We used 281 samples with at least 100 total ATAC nuclei. For each cell type, within samples with at least 10 nuclei, we analyzed the summed ATAC peak counts with mean peak count >1 . We tested for the association between peak counts and sex using negative binomial models as implemented in DESeq2 version 1.36 using the recommended single cell settings. We included batch, sample collection site, age, median mitochondrial fraction (across nuclei), TSS enrichment, and total ATAC nuclei across all cell types as sample covariates.

4.2.2.8 Chromatin state enrichment analysis in muscle fiber types

We obtained the genomic coordinates of chromatin states inferred by the 15 state ChromHMM model in two bulk skeletal muscle reference samples (E107: male, E108: female) from the Roadmap Epigenomics Consortium.¹³⁶ In each of the three muscle fiber types, we annotated the midpoint of each ATAC-seq peak with the chromatin state in the female and separately in the male reference sample. Across all cell types, 68.4% of peaks were annotated as the same state in the male and female reference samples (termed consensus state). For each fiber type and each annotation method (consensus, female, male) we tested for an association between differential accessibility status and the chromatin state with a logistic regression model. We repeated the analyses adjusting for bins of mean peak count across all samples with breaks at 0, 1, 2, 3, 4, 5, 10, 50, 100, 500, and 1,000, and to account for the greater power to detect differences by sex in peaks with higher counts.

4.2.2.9 Transcription factor binding site enrichment analysis in muscle fiber types

We obtained transcription factor binding site (TFBS) coordinates for 540 transcription factors (TF) described in D'Oliveira Albanus et al.¹³⁷ For each TF, we identified ATAC-seq peaks that overlapped the TFBS coordinates by ≥ 1 basepair. We used a logistic regression model to test for association between TFBS overlap and the inverse normalized signed $-\log_{10}$ p-value from the differential accessibility by sex analysis, adjusting for bins of mean peak count across all samples (as above).

4.2.3 Statistical analysis of bulk data

4.2.3.1 Differential mRNA expression in bulk skeletal muscle

We tested for association between bulk gene expression counts and sex using DESeq2 version 1.36 with the recommended bulk RNA settings. We adjusted for age, median insert size, mean RNA integrity number (RIN), median transcript integrity number (TIN), batch, sample collection site, mean GC content, and cell type proportion covariates (from the single nucleus data). We analyzed genes with at least 5 counts in at least 25% of the 268 samples. We used a significance threshold of $FDR < 5\%$ across all tested genes in the bulk.

4.2.3.2 Differential mRNA expression by sex in GTEx bulk skeletal muscle

We downloaded GTEx Analysis Freeze 8 bulk skeletal muscle RNA-seq counts and phenotype data files for 790 individuals from the GTEx portal.¹³⁸ We tested for association between gene expression counts and sex using DESeq2 version 1.36 with the recommended bulk RNA settings. We adjusted for age, Hardy Scale death circumstances, RNA integrity number, RNA isolation batch, and donor enrollment site. We analyzed genes with at least 5 counts in at least 25% of the 790 samples. We used a threshold of $FDR < 5\%$ across all tested genes in the GTEx dataset.

4.2.3.3 Differential miRNA expression by sex in bulk skeletal muscle

We tested for association between the miRNA counts and sex using DESeq2 version 1.36 with the recommended bulk RNA settings, using age, miRNA batch, sample collection site, and cell type proportion covariates. We analyzed genes with at least 5 counts in at least 25% of the 256 samples with both miRNA and cell type proportion data. We used a threshold of $FDR < 5\%$ across all tested miRNAs.

4.2.3.4 MiRNA targeting enrichment analysis in bulk skeletal muscle

We downloaded predicted gene targets from TargetScan version 8.0¹³⁹ for 742 mature miRNAs that we tested for differential expression by sex. Using target predictions with a cumulative weighted context++ score¹³⁹ of -0.6 or less, we counted the number of miRNA targeting each gene. For the intersection of genes tested for differential expression in bulk muscle tissue and genes scanned for target sites (n=15,199), we used logistic regression to test for association between differential expression status of genes scanned for target site and the number of differentially expressed miRNA predicted to target the gene, adjusting for the inverse-normalized 3' UTR length of the genes representative transcripts (from TargetScan) and inverse-normalized mean expression, quantified in counts per million (CPM), of the gene. For each of 605 miRNA families with at least 3 predicted gene targets, we tested for association between the predicted gene target status and differential expression status of the gene using a Firth logistic regression model, adjusting for inverse-normalized 3' UTR length and inverse-normalized average expression of the gene.

4.2.3.5 Concordance of miRNA 5p and 3p arms in bulk skeletal muscle

We examined concordance in the direction of differential expression for mature miRNAs where both the 5p and 3p arms were measured and tested for differential expression. (5p,3p) pairs that were differentially expressed in the same direction with both arms having FDR<5% were counted as concordant.

4.3 Results

4.3.1 Gene and miRNA expression and chromatin accessibility assayed in 281 vastus lateralis muscle biopsies

We analyzed data from vastus lateralis muscle biopsies of 281 (118 female and 163 male) living Finnish donors from the FUSION Tissue Biopsy Study.¹²⁹ Male and female donors were of similar age (mean 59.7 years for males and 60.9 years for females) and BMI (mean 27.9 for males and 27.4 for females) at the time of biopsy (Supplementary Table 4.1). A larger proportion of males (31.3%) than females (18.6%) had type 2 diabetes. We measured and analyzed bulk gene expression (n=268) and miRNA expression (n=256) as well as single nucleus gene expression (n=279) and chromatin accessibility (n=281) in individual muscle cell types (Supplementary Table 4.1).

4.3.2 Clustering of snRNA-seq and snATAC-seq nuclei identifies 12 cell types

We identified 12 cell types from the joint clustering of the 177,350 snRNA-seq and 252,219 single nucleus ATAC-seq (snATAC-seq) nuclei (Figure 4.1A). Across all individuals, the most abundant cell types were the three muscle fiber types: Type 1 (slow twitch oxidative fiber; mean proportion = 0.34), Type 2A (fast twitch oxidative fiber; 0.20), and Type 2X (fast twitch glycolytic fiber; 0.16), followed by endothelial cells (0.10) and mesenchymal stem cells (0.05). Less abundant cell types included smooth muscle cells, T cells, satellite cells, neuromuscular junction, neuronal cells, adipocytes, and macrophages (mean proportion <0.05 each) (Supplementary Table 4.2). There was substantial variability in the proportions of nuclei from each cell type across individuals in both sexes (Figure 4.1B), but the rank abundance of each cell type was the same in both sexes (Figure 4.1C).

4.3.3 Single nucleus data show muscle cell type composition differs by sex

To quantify sex differences in muscle cell type abundance, we tested the association of sex with the combined number of snRNA-seq and snATAC-seq nuclei for each of the 12 cell

types. On average, females had 25% more neuronal ($p=2 \times 10^{-6}$), 21% more Type 1 muscle fiber ($p=1 \times 10^{-5}$), and 11% more satellite cell ($p=0.039$) nuclei than males; males had 52% more Type 2X muscle fiber ($p=5 \times 10^{-12}$) nuclei than females (Figure 4.1D, Supplementary Table 4.3). There were no significant differences in the abundance of the remaining 8 cell types. There was no significant difference in the abundance of satellite cells when analyzing only snATAC-seq nuclei, but all other results were consistent when analyzing snRNA-seq and snATAC-seq nuclei separately (Supplementary Table 4.3, Supplementary Figure 4.2, Supplementary Figure 4.3).

4.3.4 Differential gene expression by sex in muscle cell types

We tested genes for differential expression by sex in the ten most abundant cell types. We found 3,349, 2,625, and 2,106 sex-biased genes (false discovery rate (FDR) <5%) in the Type 1, Type 2A, and Type 2X fibers, respectively, representing 12-14% of the tested genes (Figure 4.2A). We found 630 sex-biased genes across the other cell types, including 399 sex-biased genes in mesenchymal stem cells (2.7% of those tested) and 168 sex-biased genes in satellite cells (1.5% of those tested). Consistent with escape from X inactivation, in all cell types, sex-biased genes on the X chromosome were more likely to be female-biased. The median fold changes for both male- and female-biased expression of autosomal and X chromosome genes were similar (FC=1.2-1.3). There were, however, more female-biased X chromosome genes with larger effects (80th percentile FC=1.6-2.0) compared to male-biased chromosome X genes (80th percentile FC=1.5-1.6) and autosomal sex-biased genes (80th percentile FC=1.4-1.6) (Supplementary Figure 4.4). Adjusting the analysis for oral glucose tolerance test (OGTT) status had little impact on results (Supplementary Figure 4.5).

We detected a larger number of sex-biased genes in the fiber types compared to non-fiber cell types. This could be due to biological differences in the effects of sex on gene expression or

to lower power given the smaller non-fiber cell type sample sizes (Supplementary Table 4.1) coupled with fewer nuclei or counts per gene (UMIs) (Supplementary Table 4.4). To help equalize the power to detect sex-biased genes in the Type 1 fiber (most abundant cell type) compared to other less abundant cell types, we downsampled the Type 1 fiber individuals and UMIs (see Methods). Compared to the downsampled Type 1 data, we observed more sex-biased genes in the Type 2A fiber (1.1 fold) and Type 2X fiber (1.2 fold) and fewer sex-biased genes in endothelial cells (0.1 fold), smooth muscle (0.4 fold), and neuromuscular junction (0.3 fold) (Figure 4.2B).

4.3.5 Differential gene expression by sex is concordant across muscle fiber types

The direction and magnitude of sex-biased genes across the three muscle fiber types were highly concordant by sex. Between pairs of fiber types, 98-100% of sex-biased genes were more highly expressed in the same sex (Supplementary Figure 4.6). The direction of sex-biased expression between the fiber types and non-muscle cell types was less concordant. For example, only 88% of sex-biased genes (67% of autosomal sex-biased genes) in both Type 1 muscle fiber and endothelial cells were more highly expressed in the same sex (Supplementary Figure 4.6).

4.3.6 LncRNAs and pseudogenes enriched for differential expression by sex in muscle cell types

Although the roles of individual noncoding genes in establishing and maintaining sex differences have been extensively studied (i.e. *XIST*, *TSIX*),^{140,141} the extent to which classes of noncoding genes are enriched or depleted for differential expression by sex has not been studied. Because the level of gene expression varied by gene type (Supplementary Table 4.4) and was associated with power to detect differential expression (Figure 4.2C), we tested for enrichment

by gene type with and without adjusting for mean UMI of each gene. In the unadjusted analyses for the three fiber types, lncRNAs and pseudogenes were depleted for differential expression compared to protein-coding genes (Supplementary Table 4.5). However, in UMI-adjusted analysis, in Type 1 fiber, lncRNAs were 1.38 (95% CI: 1.25-1.53; $p=5.1 \times 10^{-10}$) times and pseudogenes 1.30 (95% CI: 1.09, 1.56, $p=0.0037$) times more likely to be differentially expressed by sex than protein-coding genes (Figure 4.2C, Supplementary Table 4.5, Supplementary Figure 4.7); this enrichment remained when restricting the analysis to autosomal genes (Supplementary Figure 4.7). The most significant autosomal sex-biased lncRNAs were *FAM230C* and *SNHG14*, which were both male-biased in all three muscle fiber types.

4.3.7 Mitochondrial activity, signal transduction, and cell differentiation pathways enriched for sex-biased genes in muscle fiber types

To identify biological pathways enriched for sex-biased genes, we performed GO term enrichment analysis with RNA-Enrich.¹³⁴ Expression was higher in males than females in genes in mitochondria-related and energy metabolism GO terms in all three muscle fiber types (Figure 4.2D). The top autosomal sex-biased genes of the oxidative phosphorylation GO term show concordant directions of effect across all three fiber types (Figure 4.2E). Most genes had a consistently increasing or decreasing gradient of expression from Type 1 to Type 2A to Type 2X fiber. Other genes, such as the most significant sex-biased gene in the pathway, *NDUFA10*, were expressed with a given sex at similar levels across the three fiber types in each sex (**Error! Reference source not found.**F, Supplementary Figure 4.8A). Expression was higher in females than males in genes in the caveola, signal transduction pathways, and cell differentiation-related GO terms (Figure 4.2D). The top autosomal sex-biased genes of the caveola GO term, including components of caveolae such as *CAVIN1* and *CAVIN4*, show higher expression levels in females

in one or more fiber types (Figure 4.2G, Supplementary Figure 4.8B). Many genes in the caveola GO term also showed a gradient of expression with Type 2A as an intermediate (Supplementary Figure 4.8B). The most significant sex-biased gene, *SMO*, showed highest expression in Type 2X fiber (**Error! Reference source not found.H**). Results changed little when adjusting for OGTT status (Supplementary Figure 4.9).

4.3.8 Differential gene expression by sex in bulk skeletal muscle

Bulk RNA-seq captures the nuclear and cytoplasmic mRNA of all cell types present in muscle. Single nucleus pseudobulk, formed by summing the UMIs across all nuclei for an individual, is a representation of the nuclear mRNA of all cell types with statistical removal of non-nuclear contamination. To infer cytoplasmic (by way of total) and nuclear patterns of differential expression by sex, we tested for differential gene expression in the bulk and single nucleus pseudobulk data, adjusting for estimates of cell type proportions from the snRNA-seq data. We identified 8,870 (39.7%) sex-biased genes in the bulk and 3,192 (17.1%) sex-biased genes in the pseudobulk. Among the 15,722 genes analyzed in both the bulk and pseudobulk, many more were found to be significantly differentially expressed by sex in bulk only (5,048 genes) compared to pseudobulk only (952 genes) (Figure 4.3A). Of the 1,886 sex-biased genes identified in bulk and pseudobulk, 96% were more highly expressed in the same sex (Figure 4.3A). We hypothesized that the greater number of sex-biased genes identified in the bulk may be due to deeper sequencing. However, even when comparing genes with similar mean expression levels in the bulk and pseudobulk, there were more sex-biased genes identified in the bulk (Supplementary Figure 4.10).

As an external comparison for bulk data, we tested sex-biased expression in the bulk skeletal muscle data from GTEx,¹³⁸ although we could not adjust for estimates of cell type

proportions in this analysis. We found that 1,925 (93.9%) of the 2,050 sex-biased genes identified in both the FUSION and GTEx bulk datasets were more highly expressed in the same sex (Supplementary Figure 4.11).

4.3.9 Concordance of sex-biased expression between cell types and bulk skeletal muscle

At the cell type level, we identified 5,128 sex-biased genes across one or more fiber types of which 2,463 were not identified in the bulk data. Of the 2,665 sex-biased genes identified in both the bulk and at least one of the fiber types, 2,464 (92.5%) were more highly expressed in the same sex (Figure 4.3B). This high concordance is reflected also in the GO terms enriched for sex-biased genes in the bulk; the most significant GO terms in the bulk, including those related to mitochondrial function, are significantly enriched in the pseudobulk and at least nominally enriched in the fiber types for higher expression in the same sex as bulk (Figure 4.3C). There were 132 sex-biased genes more highly expressed in one sex in the fiber types and in the other sex in bulk (Figure 4.3B). For example, BCLAF1 is female-biased in the bulk and male-biased in the fiber types (Figure 4.3D).

We identified 630 sex-biased genes in the non-fiber cell types of which 294 were not identified in the bulk data. Of the 336 sex-biased genes identified in both the bulk and at least one of the non-fiber cell types, 84% were more highly expressed in the same sex (Figure 4.3B). One counterexample is LPP, which is female-biased in the bulk and the fiber types but male-biased in mesenchymal stem cells and satellite cells (Figure 4.3E).

4.3.10 Differential miRNA expression by sex in bulk skeletal muscle

MiRNAs are short, noncoding genes that regulate gene expression and translation, usually by binding to the 3' UTR and promoting degradation of their target genes.¹⁴² Sex

differences in miRNA expression may contribute to sex differences in the post-transcriptional regulation of gene expression. To quantify sex differences in bulk miRNA expression, we tested for differential miRNA expression. We found 156 sex-biased miRNAs (20.7% of 755 tested) with a median absolute fold change of male to female counts of 1.3. MiRNAs derived from the same primary transcript are processed into two mature miRNAs (5p and 3p arms), both of which can be functional in a cell.^{143,144} In the absence of sex differences in miRNA degradation, we would expect pairs of 5p and 3p mature miRNAs derived from the same gene to have concordant directions of effect by sex. Among the 755 miRNAs tested for differential expression, 396 were part of 198 5p/3p pairs. Of these pairs, 20 were differentially expressed by sex for both the 5p and 3p arms; all had concordant directions of effect by sex (Figure 4.3F). Additionally, of the 36 5p/3p pairs that were sex-biased for only one arm, 89% had nominally concordant direction of effect by sex with the other arm, suggesting that sex-biased expression of these miRNAs was likely primarily due to sex differences in transcriptional regulation.

To investigate whether post-transcriptional regulation by sex-biased miRNAs may cause sex differences in gene expression in bulk muscle, we used TargetScan¹³⁹ to predict the mRNA targets of 742 miRNAs. The proportion of sex-biased mRNAs did not differ by the number of sex-biased miRNAs targeting the mRNA (Figure 4.3G). We hypothesized that miRNAs might play a stronger role in regulation of genes that did not show sex-biased expression in the single nucleus data, as these may represent genes with sex differences in regulatory processes in the cytoplasm. However, the proportion of sex-biased genes in this subset of genes also did not differ by the number of sex-biased targeting miRNAs (Figure 4.3G). Because specific miRNAs may have stronger effects in a given tissue,¹⁴⁵ we tested each miRNA family for enrichment of sex-biased genes among the miRNA family's predicted targets. We did not find enrichment of

sex-biased genes among the targets of any miRNA. This held true when considering only miRNA families with at least one miRNA among the 10% most highly expressed miRNAs or miRNA families containing at least one sex-biased miRNA.

4.3.11 Differential chromatin accessibility by sex in muscle cell types

Open or accessible chromatin in gene promoters allows the binding of transcriptional machinery and is associated with positive regulation of gene expression.¹⁴⁶ Differential chromatin accessibility by sex may contribute to sex differences in the transcriptional regulation of gene expression. We tested peaks on the autosomal and X chromosomes for differential chromatin accessibility by sex. We found 54,154, 62,680, and 38,197 sex-biased peaks (FDR<5%) in the Type 1, Type 2A, and Type 2X fibers, respectively (5.3-12.7% of tested peaks) (Figure 4.4A). Sex-biased peaks on the X chromosome were more likely to be female-biased. The median fold change of male- and female-biased peaks on the X chromosome (median FC=1.5-1.6 across the fiber types) were larger than male- and female-biased peaks on the autosomal chromosomes (median FC=1.3-1.4) (Supplementary Figure 4.12). To match the data available for the other cell types, we downsampled individuals and peak counts in the most abundant cell type, Type 1 fiber. Compared to the downsampled Type 1 fiber dataset, we observed more sex-biased peaks in the Type 2A fiber (2.9 fold), mesenchymal stem cell (6.2 fold), and satellite cell (18.6 fold) nuclei (Figure 4.4B). Results were little changed when adjusting for OGTT status (Supplementary Figure 4.13).

4.3.12 Differential chromatin accessibility by sex is concordant across muscle fiber types

The direction and magnitude of sex-biased peaks across the three muscle fiber types was highly concordant (Supplementary Figure 4.14). Between pairs of fiber types, over 99% of sex-biased peaks were more highly expressed in the same sex.

4.3.13 Sex-biased peaks enriched for gene regulatory function

To characterize the potential regulatory functions of sex-biased peaks in the three fiber types, we tested for enrichment of autosomal sex-biased peaks in chromatin states defined from bulk skeletal muscle reference samples.^{136,147} Compared to the quiescent state, in all fiber types, sex-biased peaks were significantly enriched in enhancers, transcription start sites (TSS), and flanking transcription states and depleted in the strong transcription state (Supplementary Figure 4.15, Supplementary Table 4.6). When adjusting for the peak read counts, which are lowest in the quiescent state, sex-biased peaks were significantly depleted in almost all states including enhancers, TSS, and flanking transcription states relative to the quiescent state (Figure 4.4C, Supplementary Table 4.6). These results were consistent when using chromatin states annotated with a male reference sample, a female reference sample, and when requiring the same annotation in both reference samples (Supplementary Figure 4.15).

To determine which transcription factors may be involved in sex differences in chromatin accessibility, we tested for enrichment of 540 transcription factor binding sites (TFBS) in sex-biased peaks in the three fiber types separately by chromatin state. We identified 335 TFBS that were significantly enriched in sex-biased peaks in at least one chromatin state in all three fiber types. TFBS enrichment was highly concordant across the fiber types. The most strongly enriched TFBS in male-biased peaks were androgen and glucocorticoid receptor binding sites (NR3C1 family) (Figure 4.4D). In Type 1 fiber, the most strongly enriched TFBS in female-biased peaks were ZNF35 and PITX2 (Figure 4.4D).

4.3.14 Directional concordance of differential accessibility of promoter region peaks and gene expression

Transcriptional regulation of gene expression can occur through the modulation of chromatin structure at gene promoters.¹⁴⁸ To assess the relationship of sex differences in chromatin accessibility with sex differences in gene expression, we counted the number of sex-biased peaks within 1kb upstream of the canonical gene TSS for each autosomal gene in the three fiber types, the single nucleus pseudobulk, and the bulk data (using sex-biased peaks in the Type 1 fiber for the pseudobulk and bulk). In these five datasets, 6-17% of sex-biased genes had one or more sex-biased peaks in the promoter region (Figure 4.4E). In each dataset, genes with at least one sex-biased peak were more likely to show sex-biased expression in the same direction as the peak. For example, in Type 1 fiber, genes with at least one male-biased promoter peak were 4.58 ($p=2.8 \times 10^{-51}$) times more likely to show male-biased expression and genes with at least one female-biased promoter peak were 5.00 ($p=3.5 \times 10^{-126}$) times more likely to show female-biased expression compared to genes without sex-biased promoter peaks (Supplementary Table 4.7).

4.4 Discussion

We found extensive sex differences in gene expression and regulation in human skeletal muscle, at both bulk and cell type (single-nucleus) resolution. Consistent with previous studies, we identified thousands of sex-biased genes in bulk muscle tissue. We showed, for the first time, widespread sex-biased expression in individual muscle cell types, identifying >2,100 sex-biased genes in Type 1, Type 2A, and Type 2X fiber nuclei. The high concordance of sex-biased expression across the fiber types and bulk tissue suggests that most sex-biased genes identified in

bulk represent sex-biased transcriptional control of gene expression and do not reflect confounding by sex differences in fiber type composition.

Muscle fiber types range from the mitochondria-rich oxidative, slow-twitch Type 1 fiber, to the intermediate Type 2A fiber, to the glycolytic, fast-twitch Type 2X fiber.^{149,150} Previous studies in bulk muscle tissue have identified inconsistent enrichments of male-^{118,123} or female-biased^{120,121} gene expression in pathways related to mitochondrial activity. We find that within each fiber type and in bulk muscle, biological pathways for mitochondrial components and oxidative energy metabolism are enriched for male-biased expression. Consistent with histological studies,¹⁰⁷ we find more Type 1 fiber nuclei in females and more Type 2X fiber nuclei in males. It is possible that the higher expression of genes in these mitochondria-related pathways in males may reflect a relatively higher level of mitochondria in each fiber type to offset the lower proportion of mitochondria-rich Type 1 fiber in males.

Skeletal muscle regeneration, characterized by the activation and differentiation of satellite cells, is essential for the preservation of muscle mass and function in response to injury.^{151,152} We find that females have a higher proportion of satellite cells than males. In addition, in muscle fibers, we find that the genes comprising caveolae, membrane organelles enriched in cholesterol with many cellular functions including a role in the repair and regeneration of muscle,¹⁵¹ are enriched for higher expression in females than in males. Similarly, we find genes in pathways related to morphogenesis, which in the context of muscle tissue may be related to regeneration,¹⁵³ to be more highly expressed in females. PITX2, a transcription factor involved in muscle homeostasis,¹⁵⁴ had one of the strongest enrichments for female-biased ATAC-seq peaks. Together, these results suggest significant sex differences in the muscle

regeneration pathway, potentially contributing to the greater endurance and recovery patterns¹⁰⁷ and slower loss of muscle strength with age¹⁵⁵ observed in females.

In addition to the sex-biased expression found in the fiber types, we identified 630 sex-biased genes in the less abundant cell types. Single nucleus resolution data are essential to uncover the cell-type specific effects of the subset of these genes that were expressed more highly in the opposite sex in bulk tissue, such as *LPP*. The smaller number of sex-biased genes identified in the non-fiber types compared to the fiber types is mostly due to lower power from smaller numbers of nuclei; as sample sizes increase, we would expect to identify more sex-biased genes in these cell types.

Our comparisons between the bulk and single nucleus datasets suggest that most sex-biased expression is likely due to sex differences in transcriptional regulation. We found that 96% of sex-biased genes identified in both the bulk and single nucleus pseudobulk tissues were more highly expressed in the same sex, suggesting that the sex-biased expression for these genes is driven mostly by processes occurring in the nucleus. We found extensive differences in chromatin accessibility, identifying tens of thousands of sex-biased peaks in the fiber types. For the small number of genes with sex-biased peaks in the promoter, sex-biased expression was positively associated with the accessibility of the promoter region, suggesting that these sex differences are due to sex differences in transcriptional regulation. This was true even for bulk sex-biased genes not identified in the pseudobulk, which could indicate that sex differences from these genes are also due to transcriptional regulation and were not detected in pseudobulk due to lower power. Exceptions to the high concordance of bulk and pseudobulk include genes such as *BCLAF1*, which encodes a protein involved in muscle regeneration.¹⁵⁶ Such examples could represent different regulatory processes in the nucleus and cytoplasm.

We also identified 156 sex-biased miRNAs, which can regulate gene expression in the cytoplasm. The concordant sex-biased expression of the 3p and 5p arms suggest that sex differences in the expression levels of miRNAs themselves are primarily transcriptionally regulated. We were unable to identify miRNA regulation of sex-biased expression. It is possible, however, that we were not able to identify miRNA targets accurately with computational tools alone because of the tissue-specific nature of miRNA targeting.

We reach strikingly different conclusions about the enrichments of differential expression by type of gene and of chromatin accessibility by chromatin state when adjusting or not adjusting for mean gene or peak read count in the analysis. The mean gene (UMI) or peak read count is positively (and non-linearly) associated with the power to detect sex differences (Figures 2C, 4C). In the unadjusted gene expression enrichment analysis, protein-coding genes, which have the highest mean counts of all gene types in the muscle fiber types, are strongly enriched for sex expression differences compared to other types of genes. However, when adjusting for read count, we found that lncRNAs are more highly enriched for expression differences by sex compared to protein-coding genes (**Error! Reference source not found.C**, Supplementary Table 4.5). LncRNAs may form local hubs of transcription¹⁵⁷ that could contribute to the known clustering of genes that are differentially expressed by sex.⁴⁰ In the unadjusted chromatin state enrichment analysis, we found that peaks in the quiescent state, which have lower counts than most other states, were depleted for sex differences compared to peaks in other chromatin states (Supplementary Table 4.6). However, when adjusting for peak read counts, we found that the quiescent state is enriched for sex differences (**Error! Reference source not found.C**) compared to other states. These peaks that show the enrichment are the most highly expressed peaks in the

quiescent chromatin regions, are particularly sensitive to regulation by sex, and could be in regions that are misclassified as quiescent.

The FUSION Tissue Biopsy Study provides unique advantages for studying sex differences in gene expression and regulation in human skeletal muscle. With 281 skeletal muscle biopsies, it is the largest study to date with snRNA-seq and snATAC-seq to date. All of the biopsies were taken from living donors, and thus do not have variability in gene expression introduced by severe illness or the death process.¹⁵⁸ The median age at the time of biopsy was 61 years, and over 88% of donors were older than 50. We were therefore unable to ascertain sex differences in gene expression at younger ages when circulating sex hormone levels show greater differences between males and females,¹⁵⁹ but showed instead that extensive sex differences in gene expression remain in older individuals. Although all FUSION donors are Finnish, the high degree of concordance in sex-biased expression between bulk FUSION and GTEx, particularly for genes with larger fold changes, suggests that our findings will generalize to non-Finnish populations.

Overall, our findings demonstrate the potential of integrating bulk and single nucleus data and provide transcriptome-level insights into sex differences in skeletal muscle biology. Future studies will be needed to show the mechanisms by which sex-biased gene expression contributes to sex differences in skeletal muscle physiology and disease susceptibility.

4.5 Tables and Figures

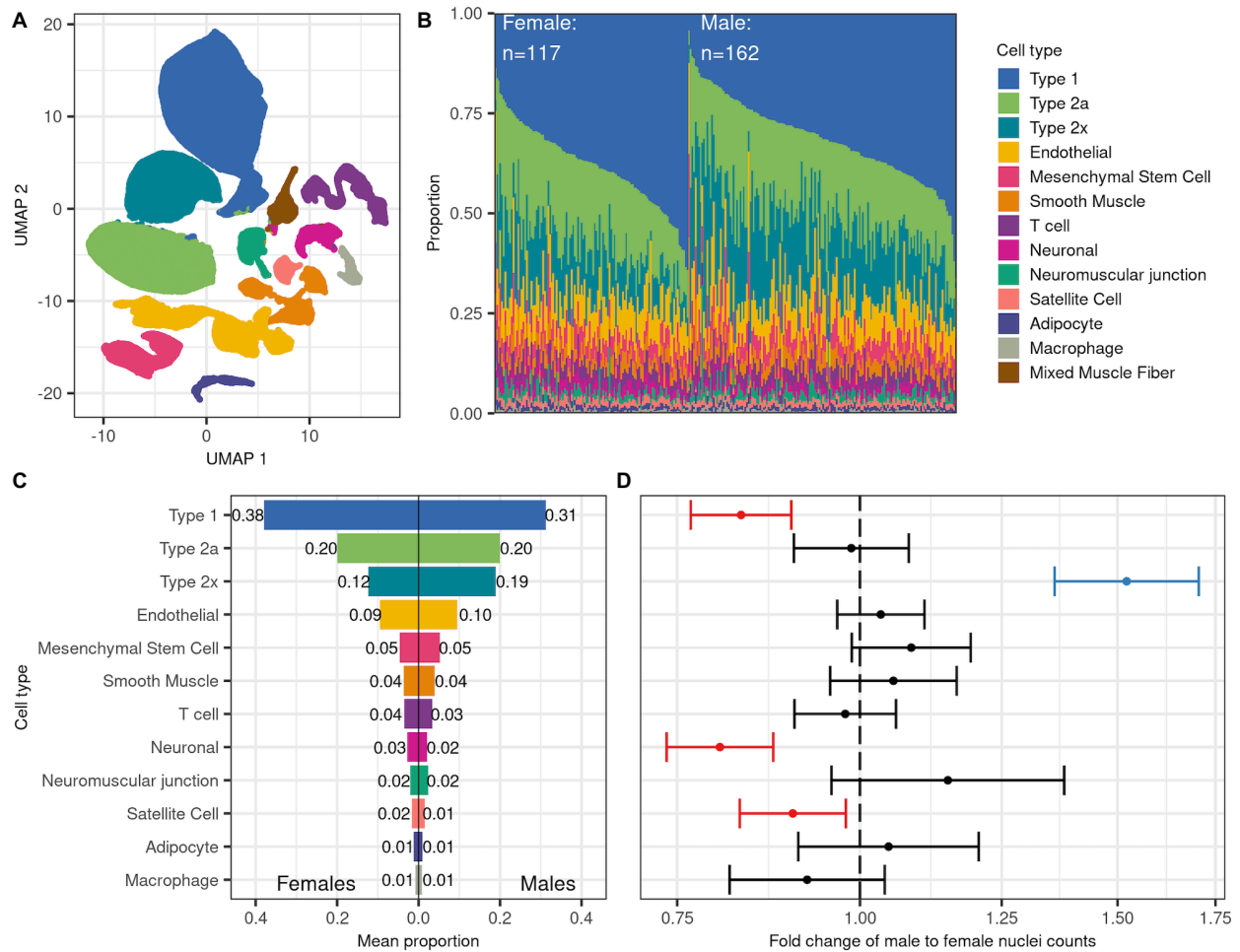


Figure 4.1 Sex differences in cell type composition of human skeletal muscle

A. UMAP projection of 429,569 RNA and ATAC nuclei across 279 individuals. B. Cell type proportions for each individual sorted by sex and proportion of Type 1 muscle fiber nuclei. Mixed muscle fiber was removed from analysis. C. Mean cell type proportions by sex. D. Fold change and 95% confidence intervals for the combined number of RNA and ATAC male nuclei compared to the number of female nuclei in each cell type. Cell types with significantly more nuclei at FDR < 0.05 in females are colored red and in males are colored blue.

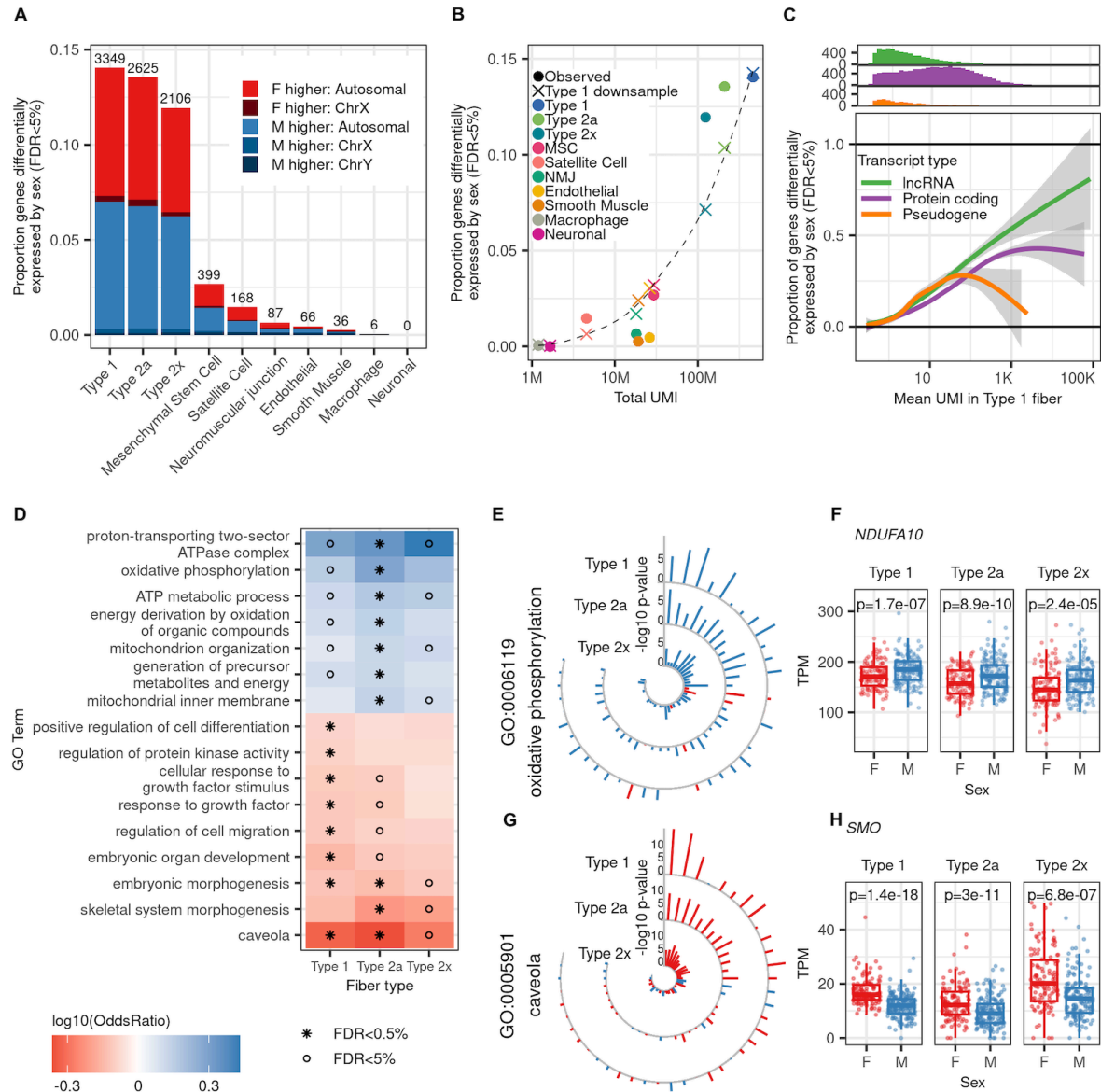


Figure 4.2 Sex differences in cell type-specific gene expression in human skeletal muscle.

A. The number and proportion of genes significantly differentially expressed by sex (FDR<5%) in 10 muscle cell types. B. The proportion of genes significantly differentially expressed by sex (FDR<5%) by the total UMI across all samples in each cell type in the observed data and in Type 1 muscle fiber data downsampled to match the sample size and total UMI of the other cell types. C. Histograms and smooth curves with 95% confidence intervals for the proportion of genes significantly differentially expressed by sex (FDR<5%) by mean UMI across all samples in Type 1 muscle fiber. D. The set of GO terms that are highly enriched (FDR<0.1%) for genes expressed higher in males (odds ratio >1) or females (odds ratio <1) in at least one muscle fiber type. E. The $-\log_{10}$ p-values across the three muscle fiber types colored by direction of effect for the top 40 autosomal genes in the biological process GO term oxidative phosphorylation. F. Boxplots of the TPMs of *NDUF10* in the female and male samples across the three muscle fiber types. One male and two female outlying samples with TPMs>430 in Type 2X and one outlying male with 0 TPMs in Type 2A are not shown. G. The $-\log_{10}$ p-values across the three muscle fiber types colored by direction of effect for the top 40 autosomal genes in the cellular component GO term caveola. H. Boxplots of the TPMs of *SMO* in the female and male samples across the three muscle fiber types. Two female outlying samples with TPMs>67 in Type 2X and one female outlying sample with TPM>60 in Type 2A are not shown.

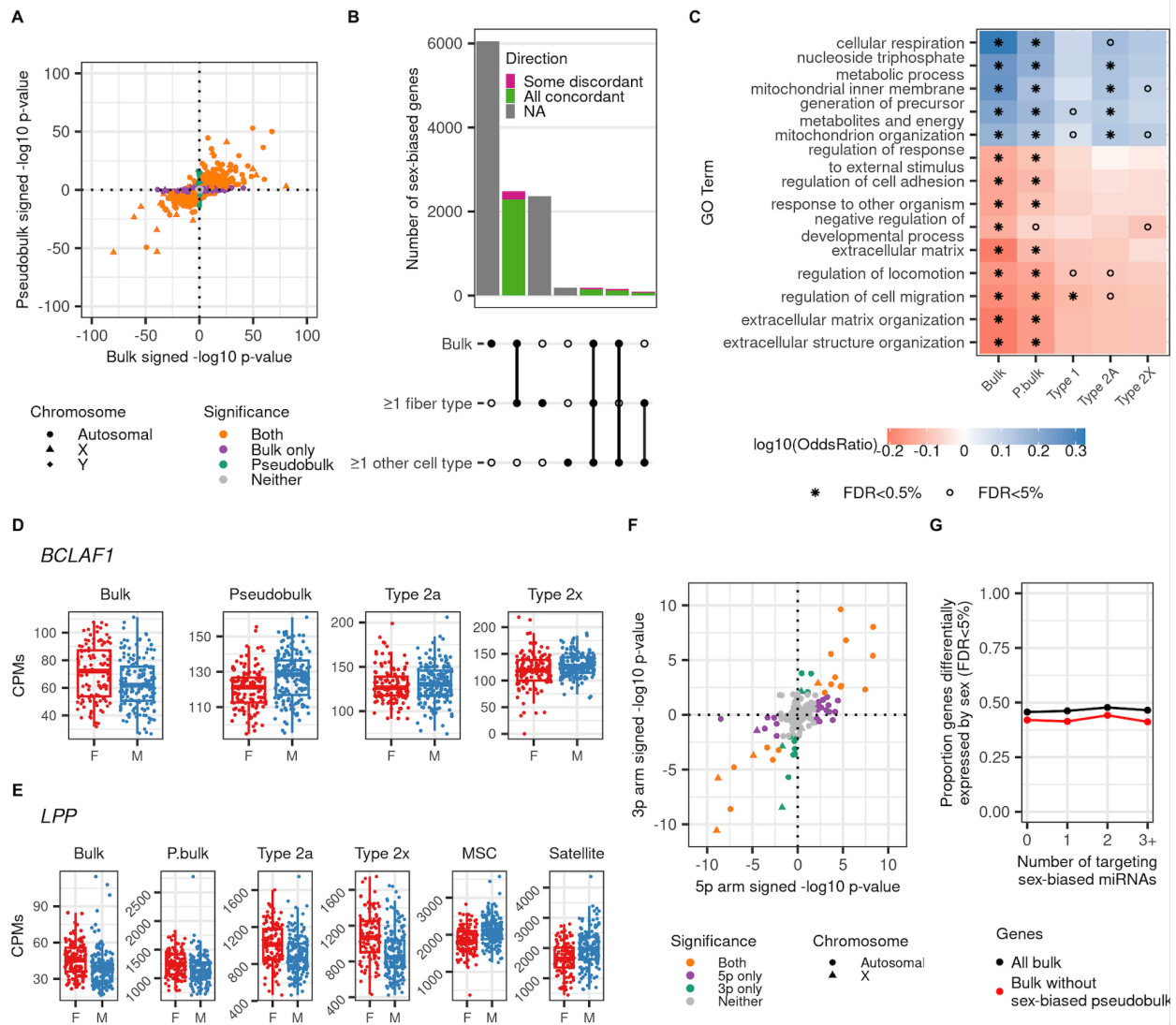


Figure 4.3 Comparison of sex differences in gene expression from bulk vs. single nucleus RNA-seq

A. Scatterplot of the signed $-\log_{10}$ p-values (>0 higher in males; <0 higher in females) of differential expression between the bulk and single nucleus pseudobulk. B. Directional upset plot of the number of significantly differentially expressed genes identified in the bulk RNA-seq, at least one fiber type, or at least one non-fiber cell type from snRNA-seq data. C. The set of GO terms that are most highly enriched for genes expressed higher in males (odds ratio >1) or females (odds ratio <1) in the bulk. D. Expression levels of *BCLAF1* in CPMs by sex in bulk, pseudobulk, and Type 2a and 2x fibers. In the Type 2X fiber, one outlying male had a CPM value of 356, which is not shown in the plot. E. Expression levels of *LPP* in CPMs by sex in bulk, pseudobulk, Type 2a and 2x fibers, mesenchymal stem cells, and satellite cells. F. Scatterplot of the signed $-\log_{10}$ p-values (>0 higher in males; <0 higher in females) of differential expression between miRNA 5p and 3p arms. G. The proportion of genes significantly differentially expressed by sex (FDR $<5\%$) in bulk muscle by the number of predicted miRNA targeting the gene that are significantly differentially expressed by sex.

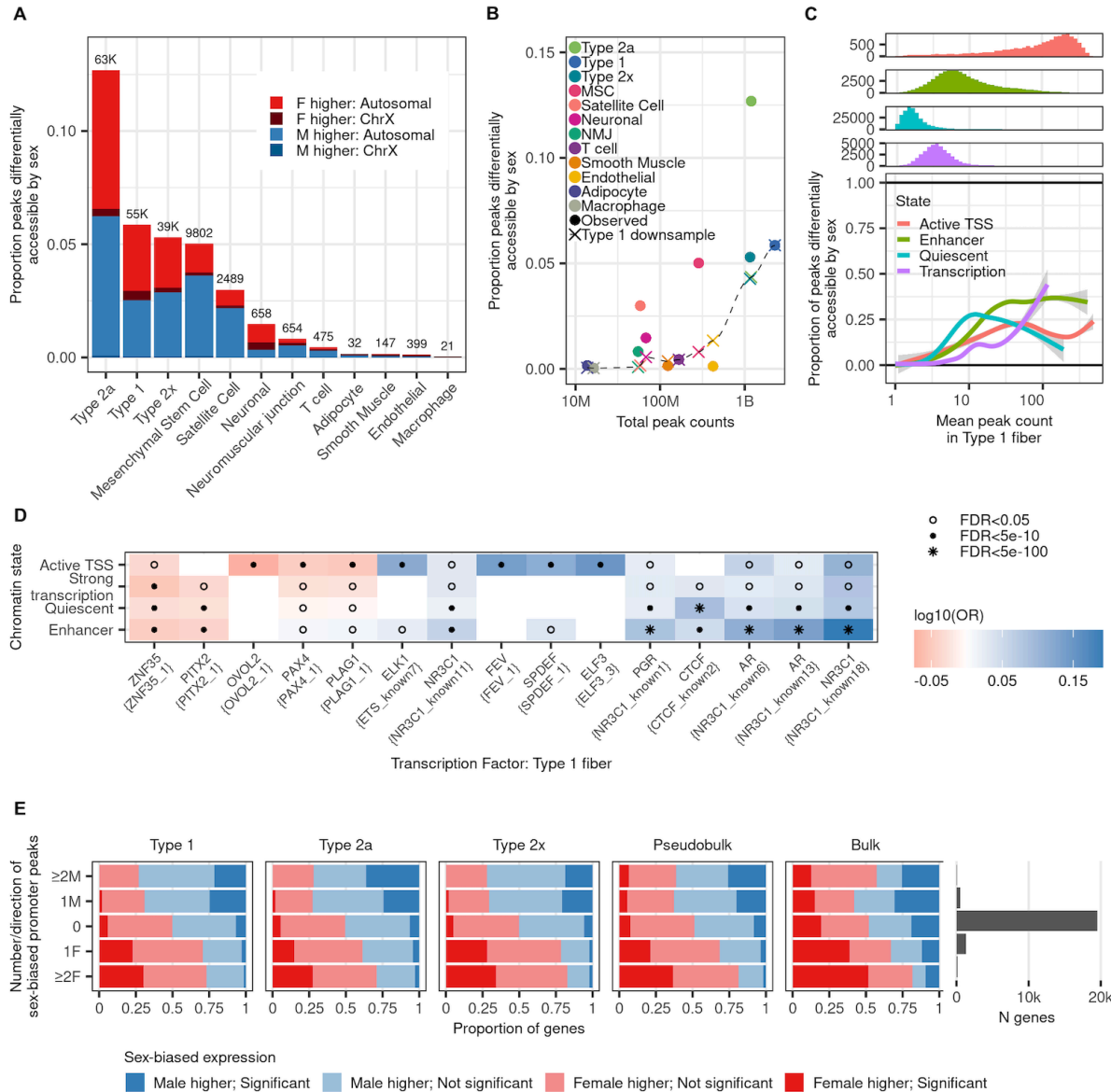
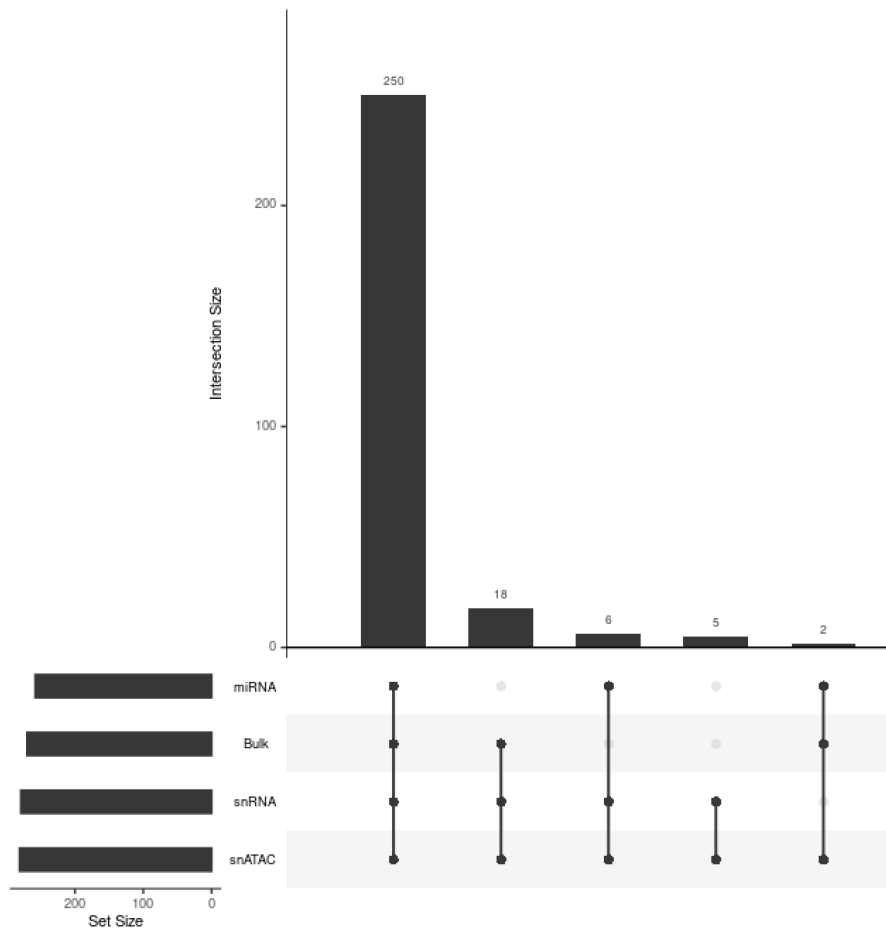


Figure 4.4 Sex differences in cell-type specific chromatin accessibility in human skeletal muscle

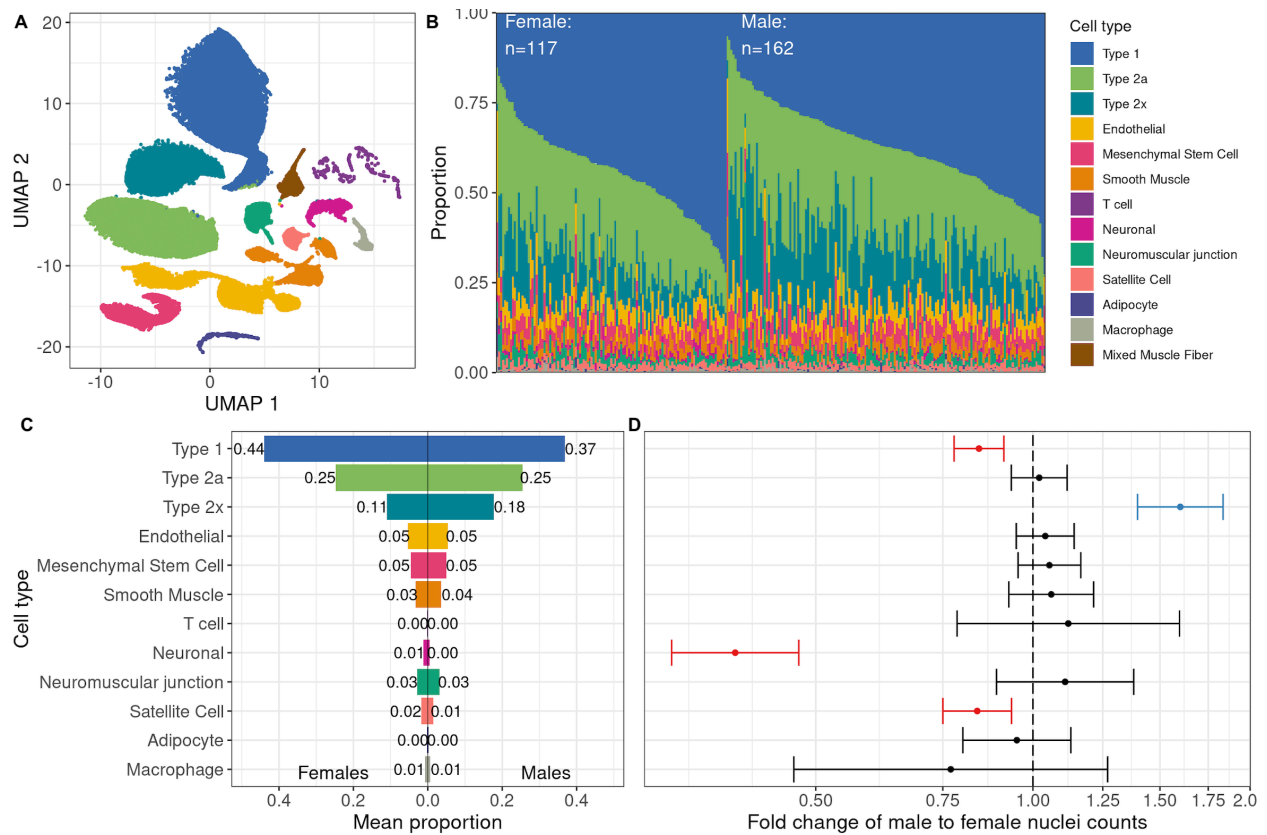
A. The number and proportion of peaks significantly differentially accessible by sex (FDR<5%) in 12 muscle cell types. B. The proportion of peaks significantly differentially accessible by sex (FDR<5%) by the total counts across all samples in each cell type in the observed data and in Type 1 muscle fiber data downsampled to match the sample size and total counts of the other cell types. C. The smooth curve and 95% confidence intervals for the proportion of peaks significantly differentially expressed by sex (FDR<5%) by chromatin state by mean peak count across all samples and histograms of the number of peaks by state and mean peak count in Type 1 muscle fiber. D. The set of transcription factor binding sites that are most highly enriched for peaks with higher counts in males (odds ratio >1) or females (odds ratio <1) in Type 1 muscle fiber in active TSS, strong transcription, quiescent, and enhancer consensus chromatin states. E. The proportion of autosomal genes with 0, 1, ≥2 sex-biased peaks <1kb upstream of gene TSS colored by their differential expression status and direction in the three fiber types, pseudobulk, and bulk. The single nucleus pseudobulk and bulk genes are annotated with sex-biased peaks from the Type 1 fiber. The histogram on the right shows the number of genes in bulk with each number/direction of DA peaks; the distribution is similar for the three fiber types and pseudobulk.

4.6 Supplementary Material



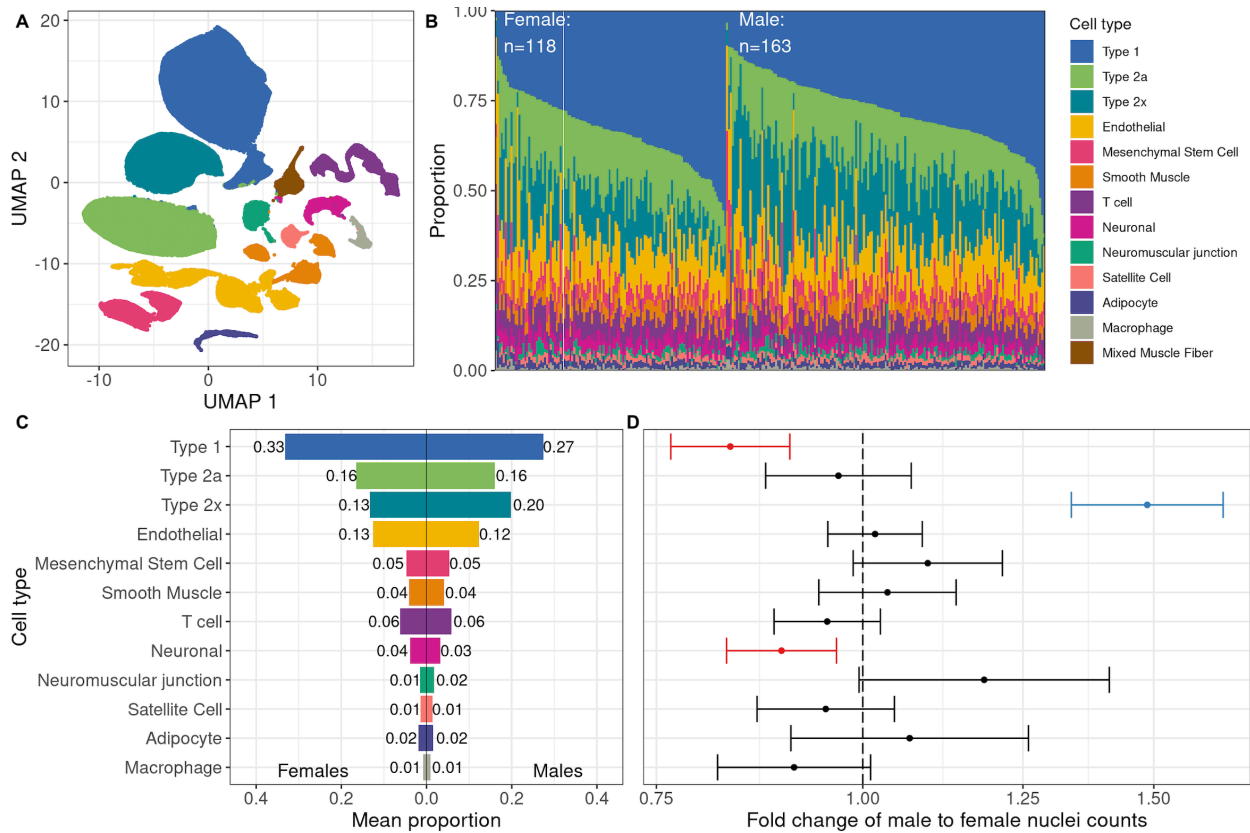
Supplementary Figure 4.1 Sample sizes across molecular data modalities

An upset plot describing the sample overlap between datasets for which bulk miRNA expression (miRNA), bulk mRNA expression (Bulk), single nucleus gene expression (snRNA), and single nucleus chromatin accessibility (snATAC) were collected.



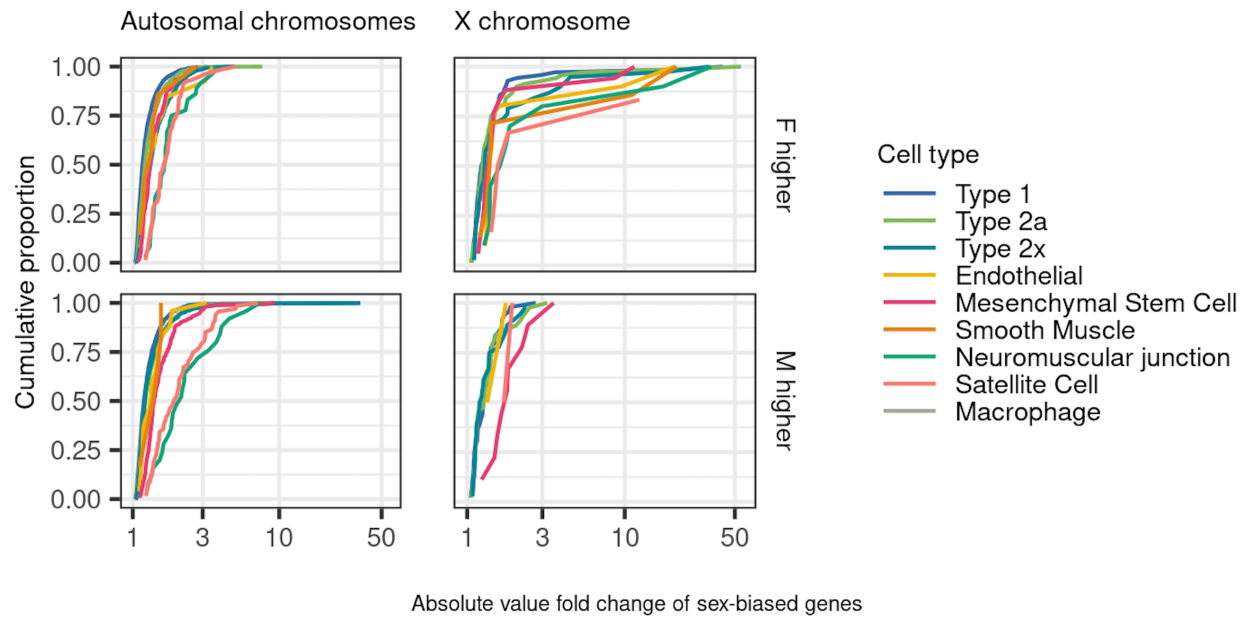
Supplementary Figure 4.2 Sex differences in cell type composition of human skeletal muscle using RNA nuclei

A. UMAP projection of 177,350 RNA nuclei across 279 individuals. B. Cell type proportions for each individual sorted by sex and proportion of Type 1 muscle fiber nuclei. C. Mean cell type proportions by sex based on RNA nuclei. D. Fold change and 95% confidence intervals for the number of male RNA nuclei compared to the number of female nuclei in each cell type. Cell types with significantly more nuclei at FDR < 0.05 in females are colored red and in males are colored blue.

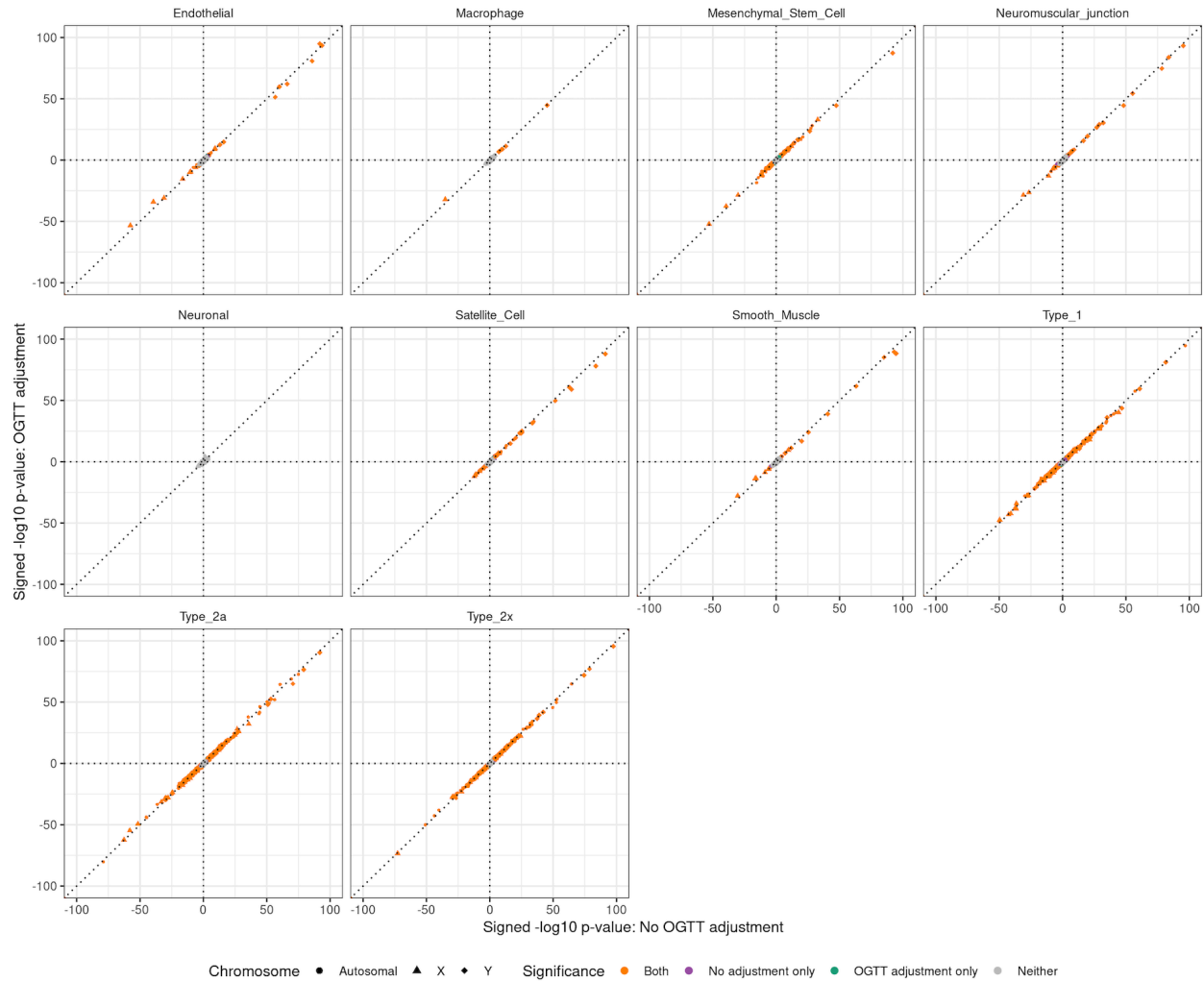


Supplementary Figure 4.3 Sex differences in cell type composition of human skeletal muscle using ATAC nuclei

A. UMAP projection of 252,219 ATAC nuclei across 281 individuals. B. Cell type proportions for each individual sorted by sex and proportion of Type 1 muscle fiber nuclei. C. Mean cell type proportions by sex based on ATAC nuclei. D. Fold change and 95% confidence intervals for the number of male ATAC nuclei compared to the number of female nuclei in each cell type. Cell types with significantly more nuclei at FDR < 0.05 in females are colored red and in males are colored blue.

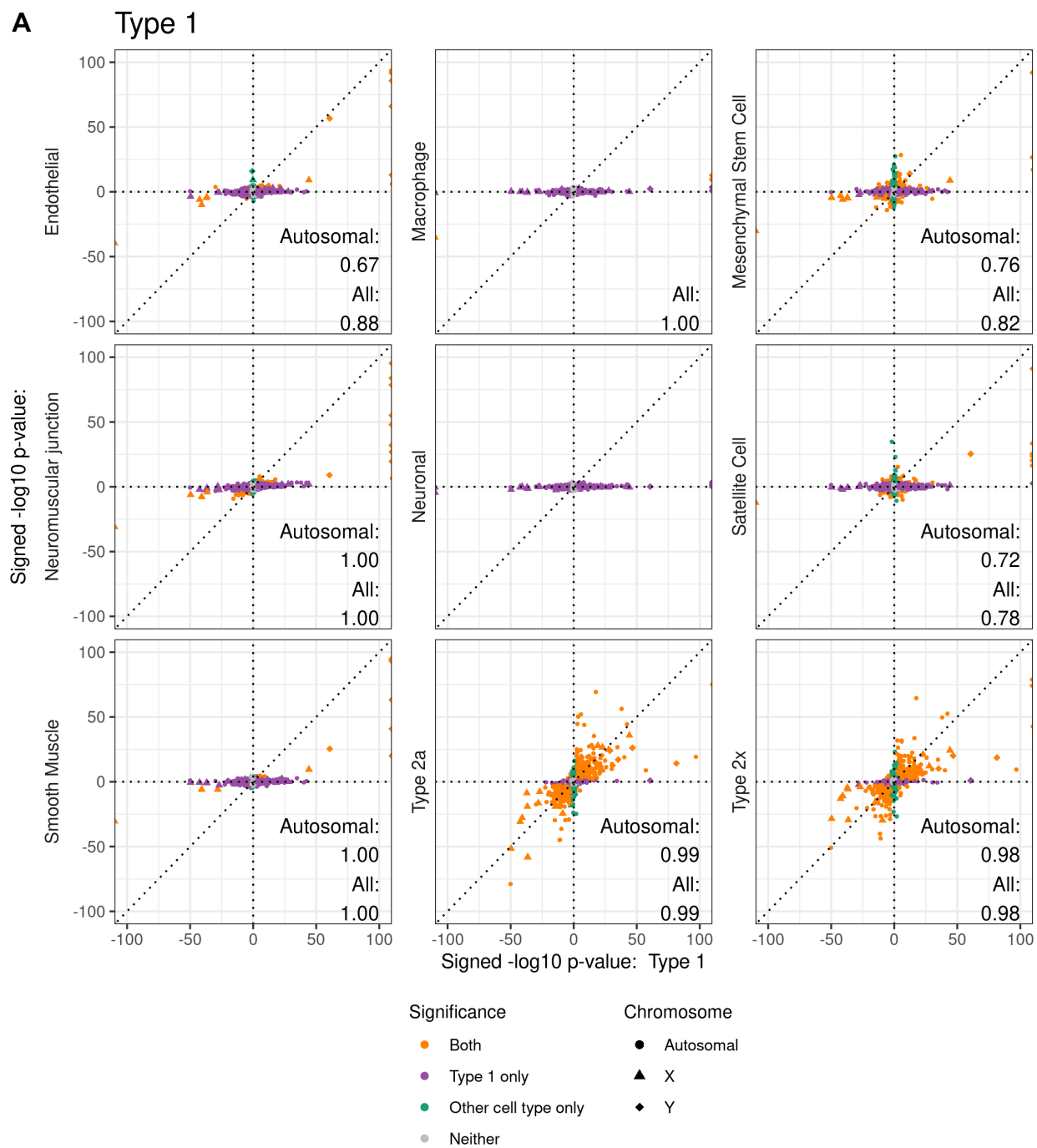


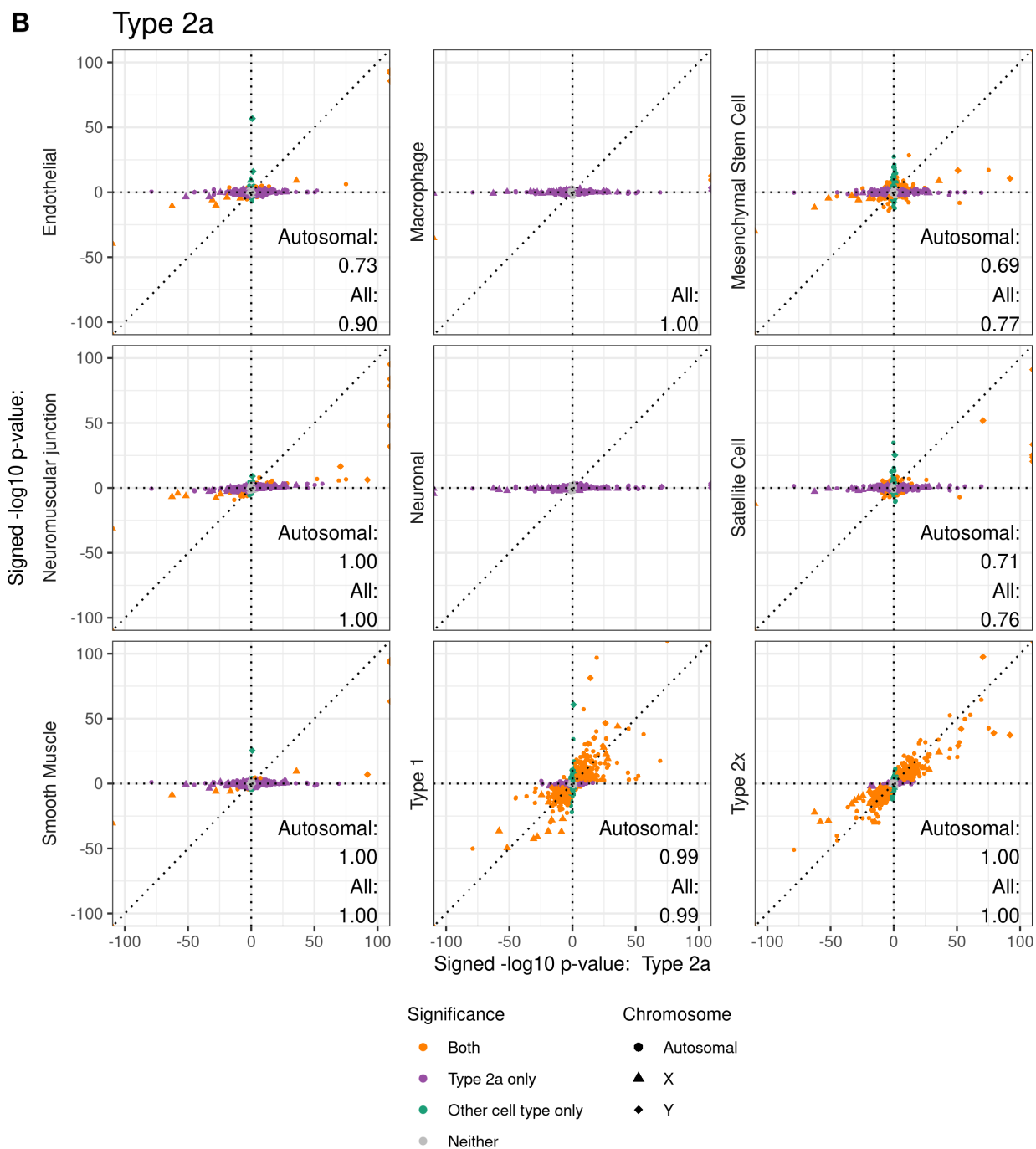
Supplementary Figure 4.4 The cumulative distribution of the absolute fold change of sex-biased genes by cell type, chromosome, and direction of effect



Supplementary Figure 4.5 Comparison of differential expression by sex with and without adjusting for oral glucose tolerance test (OGTT) status

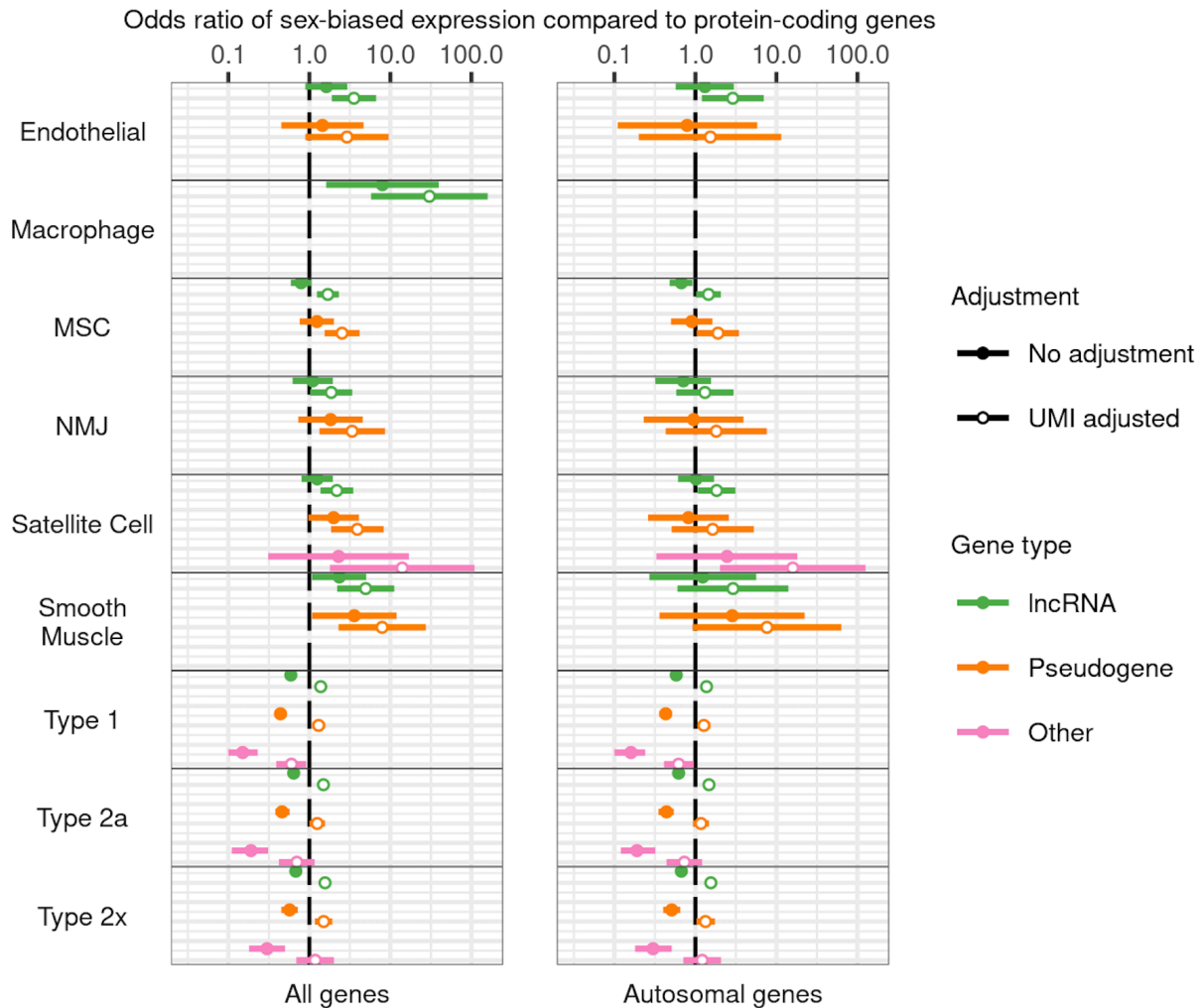
Scatterplots of the signed $-\log_{10}$ p-values (>0 higher in males; <0 higher in females) of differential expression for a model not adjusting for OGTT status (x-axis) and a model adjusting for OGTT status (y-axis). Signed $-\log_{10}$ p-values with an absolute magnitude >100 are not shown. Each gene is colored by the significance (FDR 0.05) from each model.





Supplementary Figure 4.6 Comparison of differential expression by sex across muscle cell types

Scatterplots of the signed $-\log_{10}$ p-values (>0 higher in males; <0 higher in females) of differential expression between A. Type 1 muscle fiber and all other cell types and B. Type 2A muscle fiber and all other cell types for genes in common between each cell type pair. Signed $-\log_{10}$ p-values with an absolute magnitude >100 are not shown. Each gene is colored by the significance (FDR 0.05) in each cell type. The proportion of genes that are significant in both cell types with concordant direction of effect is shown for each cell type pair in the lower left corner.

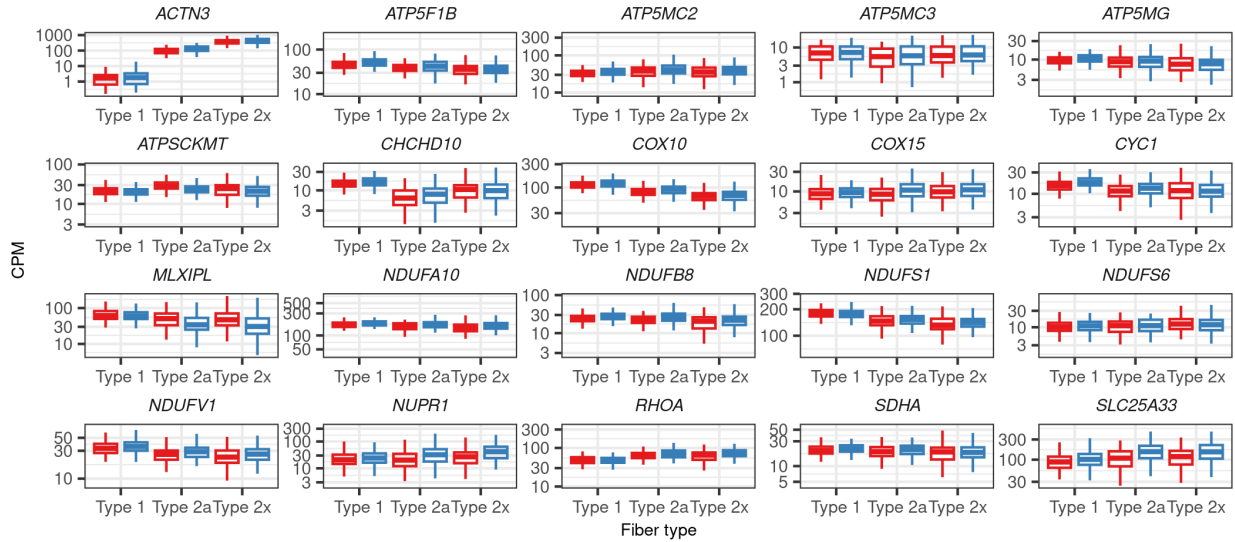


Supplementary Figure 4.7 Association of gene type with sex-biased expression

The odds ratios and corresponding 95% confidence intervals by cell type from logistic regression models testing the association between sex-biased expression (FDR<5%) and gene type (lncRNA, pseudogene, other) compared to protein-coding genes, with and without adjusting for mean UMI. The column row includes all genes and the second row includes only genes on autosomal chromosomes. There were 0 sex-biased genes for any combination of gene type and cell type that does not appear on this plot.

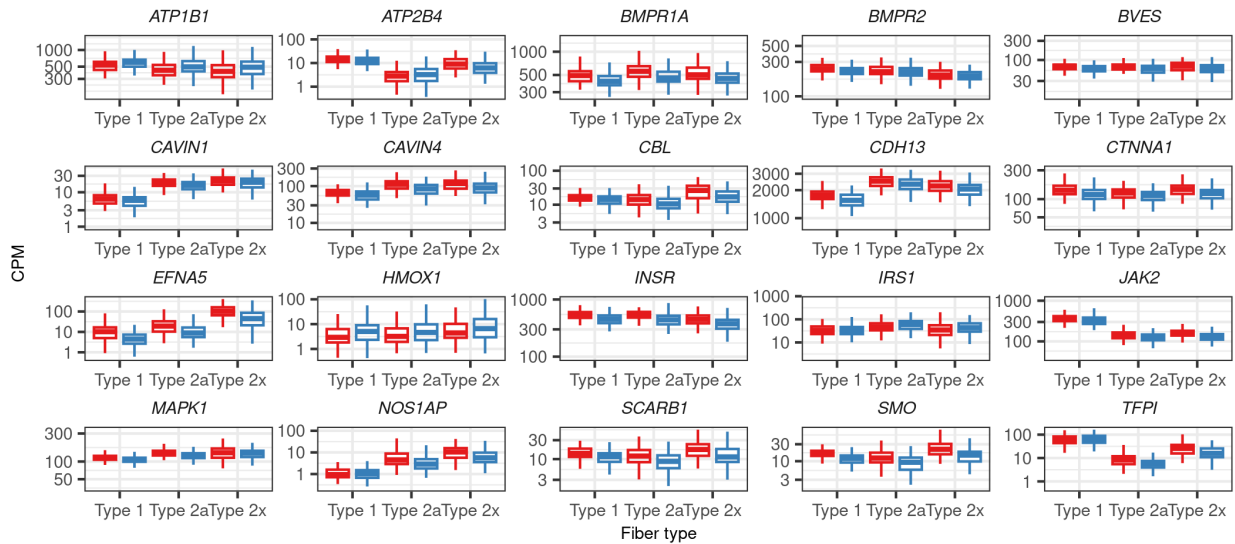
A

Oxidative phosphorylation



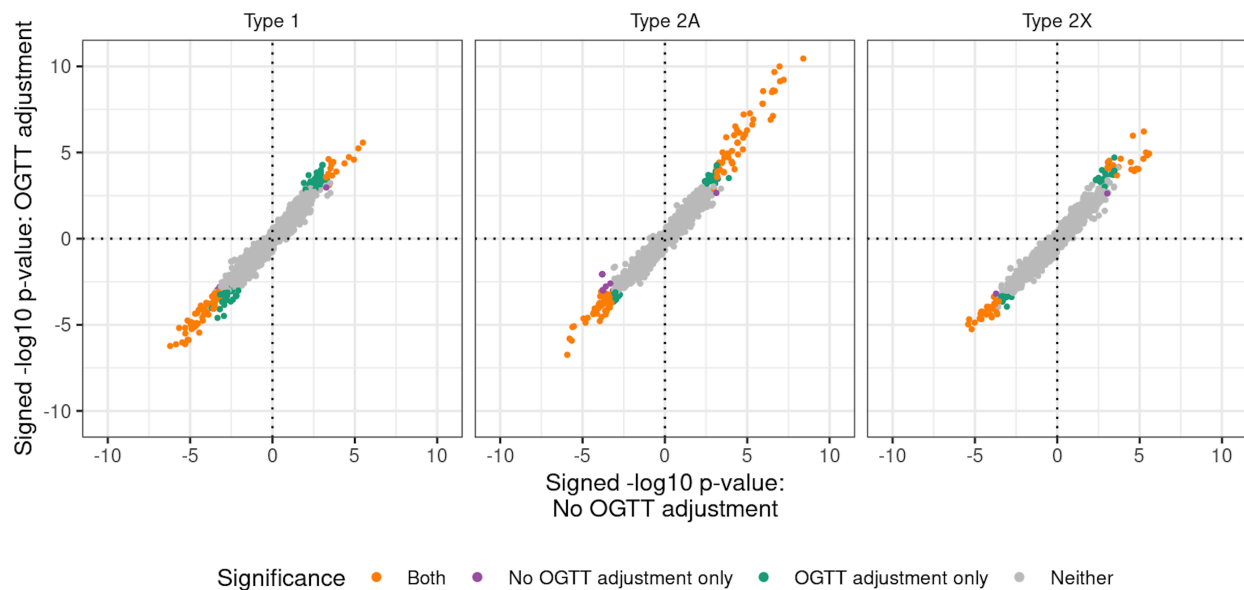
B

Caveola



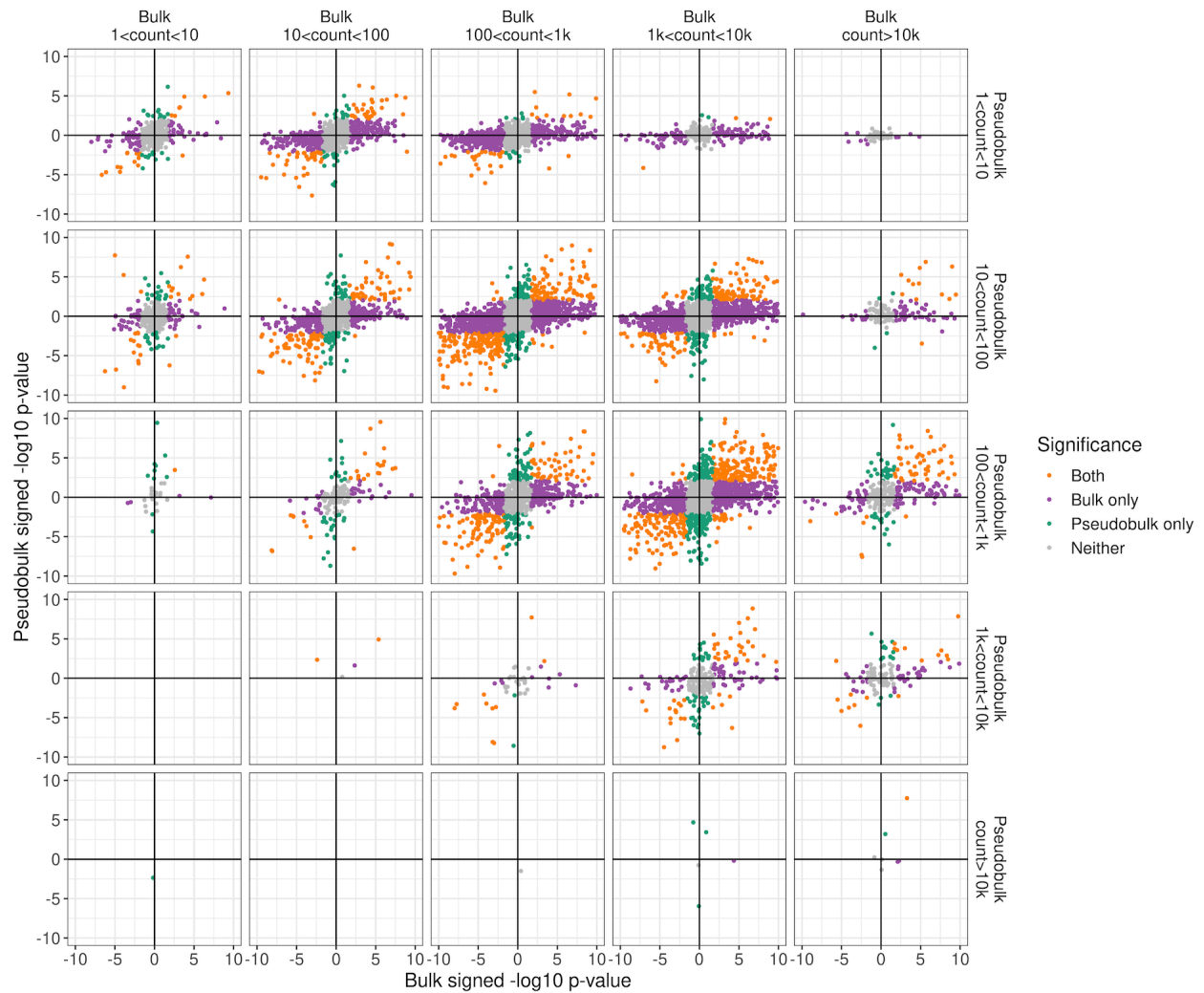
Supplementary Figure 4.8 Gene expression levels in muscle fibers of top 20 autosomal sex-biased genes in oxidative phosphorylation and caveola GO terms

Boxplots of the CPMs are shown for females (red) and males (blue) in Type 1, Type 2a, and Type 2x fibers for genes in A. the GO biological process oxidative phosphorylation and in B. the GO cellular component caveola.



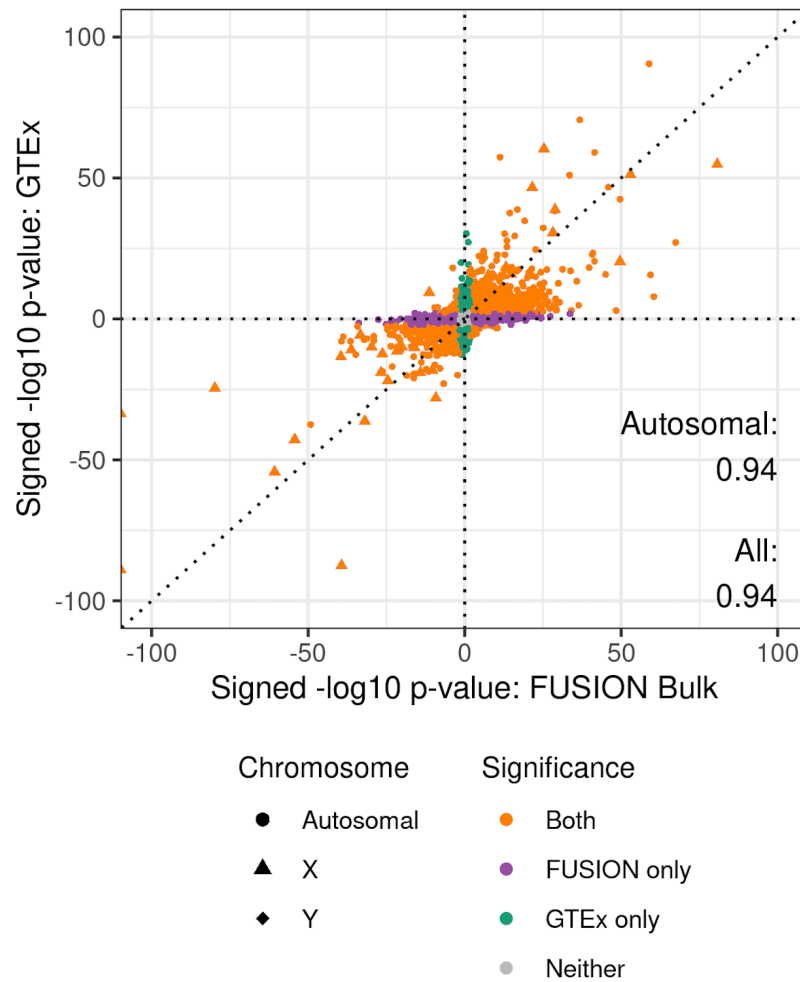
Supplementary Figure 4.9 Comparison of gene set enrichment test results for differential expression by sex with and without adjusting for oral glucose tolerance test (OGTT) status

Scatterplots of the signed $-\log_{10}$ p-values (>0 higher in males; <0 higher in females) of enrichment for GO terms calculated from the results of differential expression analyses from a model not adjusting for OGTT status (x-axis) and a model adjusting for OGTT status (y-axis). Each GO term is colored by the significance (FDR 0.05) from each model.



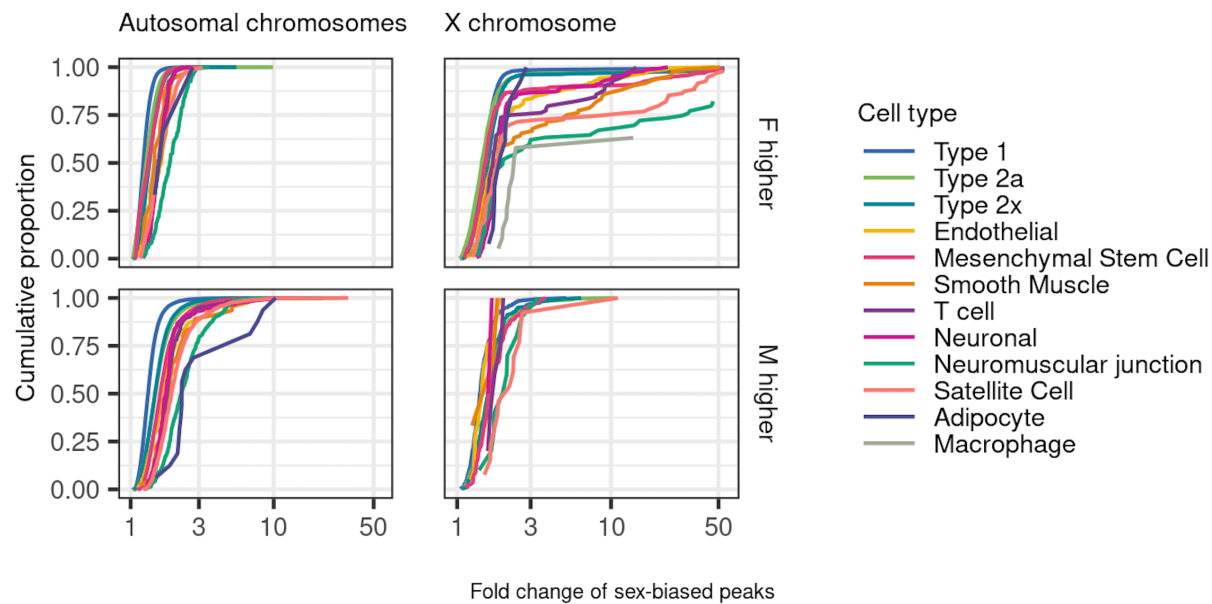
Supplementary Figure 4.10 Comparison of differential expression by sex between bulk and pseudobulk by expression level

Scatterplots of the signed $-\log_{10}$ p-values (>0 higher in males; <0 higher in females) of differential expression for autosomal genes tested in bulk and pseudobulk by bins of expression level measured in counts. Signed $-\log_{10}$ p-values with an absolute magnitude >10 are not shown. Each gene is colored by the significance (FDR 0.05) in bulk and pseudobulk.

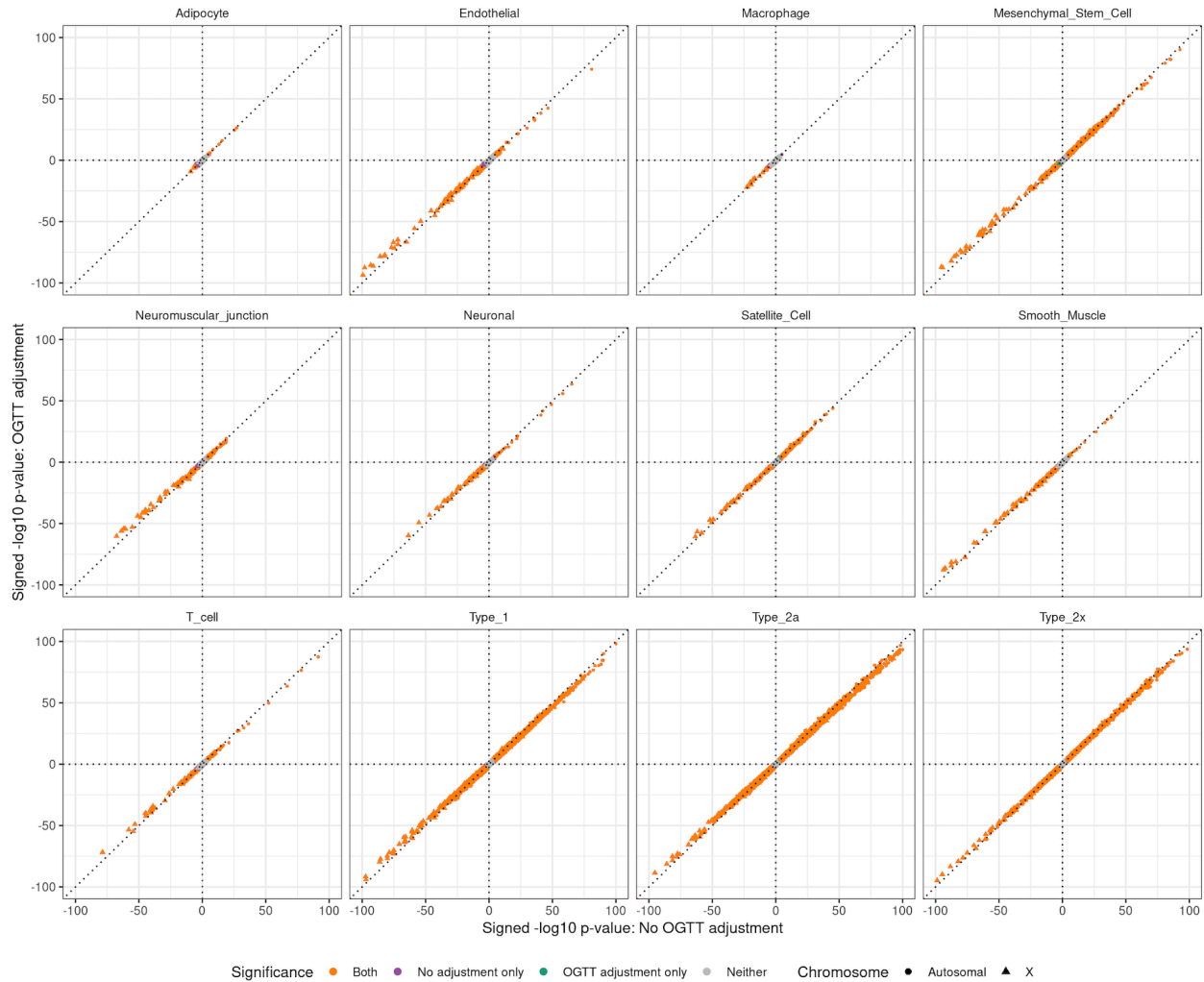


Supplementary Figure 4.11 Comparison of differential expression by sex between FUSION bulk and GTEx bulk

Scatterplot of the signed $-\log_{10}$ p-values (>0 higher in males; <0 higher in females) of differential expression for the 19,939 genes tested in both the FUSION and GTEx bulk skeletal muscle. Signed $-\log_{10}$ p-values with an absolute magnitude >100 are not shown. Each gene is colored by the significance (FDR 0.05) in the FUSION and GTEx datasets.

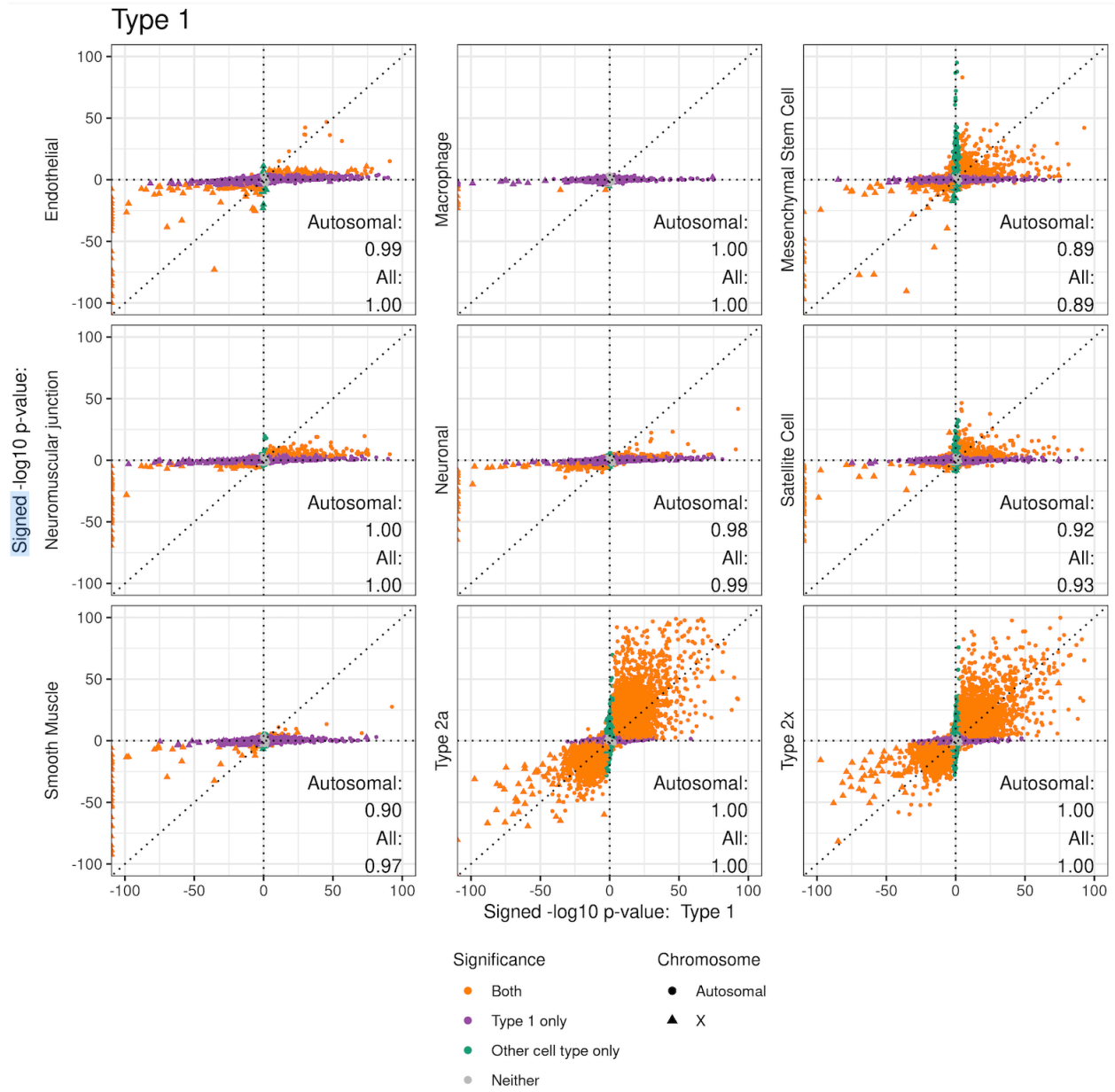


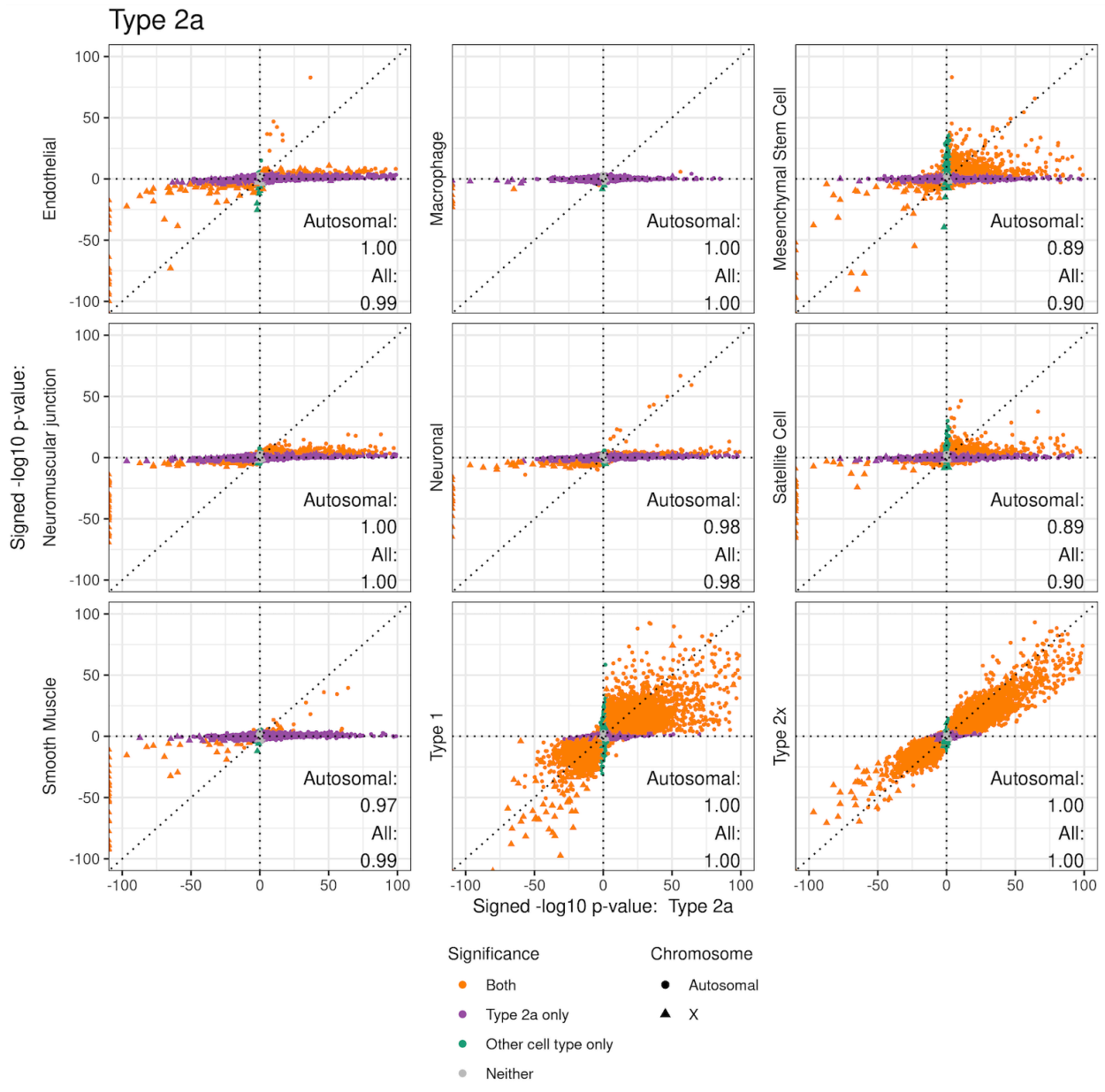
Supplementary Figure 4.12 The cumulative distribution of the absolute fold change of sex-biased peaks by cell type, chromosome, and direction of effect



Supplementary Figure 4.13 Comparison of differential accessibility by sex with and without adjusting for oral glucose tolerance test (OGTT) status

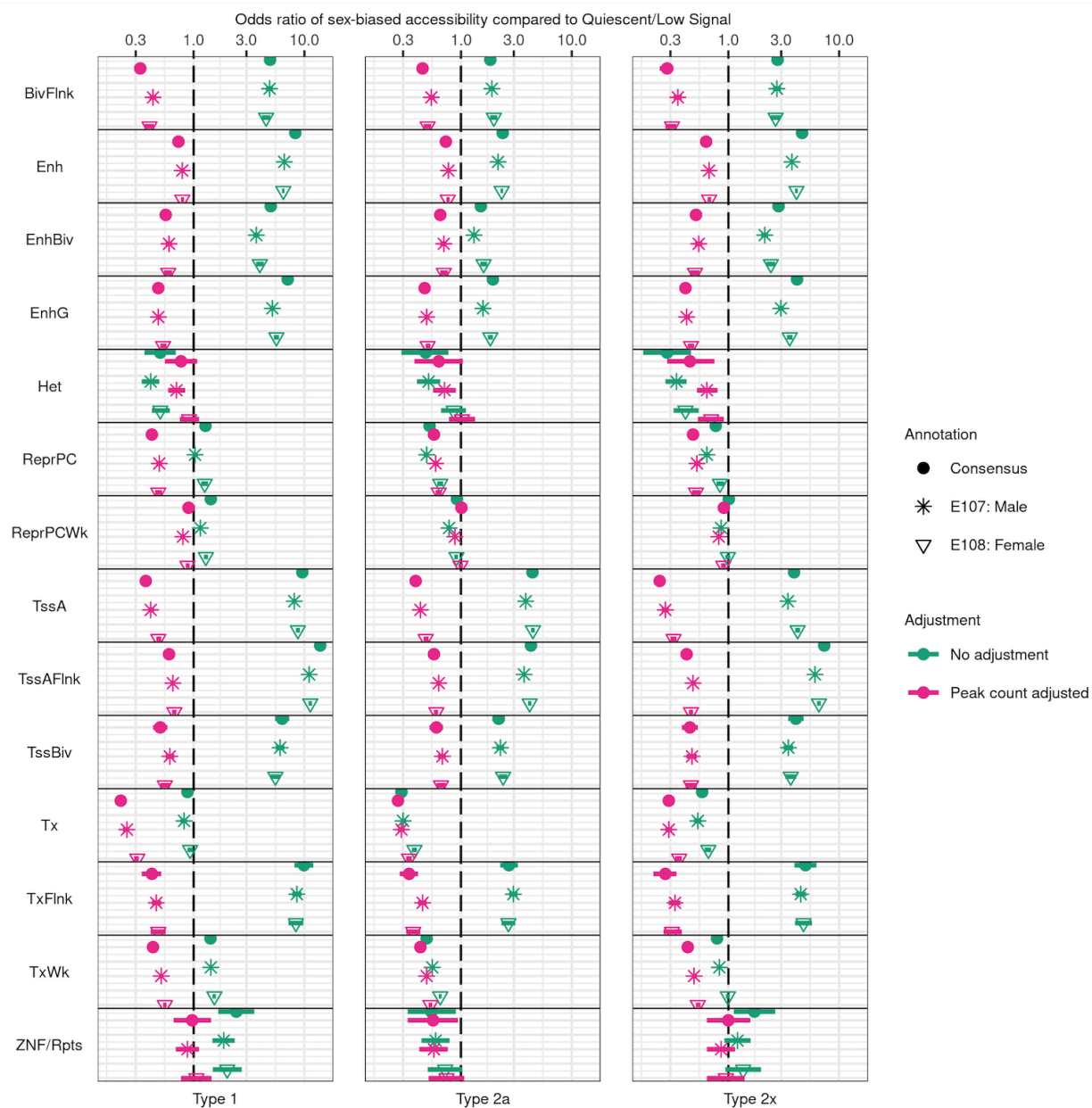
Scatterplots of the signed $-\log_{10}$ p-values (>0 higher in males; <0 higher in females) of differential accessibility for a model not adjusting for OGTT status (x-axis) and a model adjusting for OGTT status (y-axis). Signed $-\log_{10}$ p-values with an absolute magnitude >100 are not shown. Each peak is colored by the significance (FDR 0.05) from each model.





Supplementary Figure 4.14 Comparison of differential accessibility by sex across muscle cell types

Scatterplots of the signed $-\log_{10}$ p-values (>0 higher in males; <0 higher in females) of differential accessibility between A) Type 1 muscle fiber and all other cell types and B) Type 2A muscle fiber and all other cell types for peaks in common between each cell type pair. Signed $-\log_{10}$ p-values with an absolute magnitude >100 are not shown. Each peak is colored by the significance (FDR 0.05) in each cell type. The proportion of peaks that are significant in both cell types with concordant direction of effect is shown for each cell type pair in the lower left corner.



Supplementary Figure 4.15 Association of chromatin state with differential accessibility by sex

The odds ratios and corresponding 95% confidence intervals by fiber type from logistic regression models testing the association between differential accessibility by sex (FDR<5%) status and chromatin state (defined from female reference, male reference, or using only peaks with consensus calls) compared to the quiescent/low state, with and without adjusting for the mean peak count across all samples.

		Female (n=118)	Male (n=163)	Total (n=281)	
A. Sample characteristics (mean (SD))					
Age		60.9 (7.0)	59.7 (7.8)	60.2 (7.5)	
BMI		27.4 (4.1)	27.9 (4.4)	27.7 (4.3)	
OGTT (n (%))					
	NGT	52 (44.1%)	47 (28.8%)	99 (35.2%)	
	IFG	12 (10.2%)	27 (16.6%)	39 (13.9%)	
	IGT	32 (27.1%)	37 (22.7%)	69 (24.6%)	
	T2D	22 (18.6%)	51 (31.3%)	73 (26.0%)	
	Missing	0 (0%)	1 (0.6%)	1 (0.4%)	
B. Sample size by data type (n)					
Cell type(s)		Modality			
Bulk		RNA	110	158	268
		miRNA	108	148	256
Single nucleus	All	RNA	117	162	279
		ATAC	118	163	281
	Adipocyte	RNA	2	2	4
		ATAC	86	105	191
	Endothelial	RNA	113	155	268
		ATAC	118	163	281
	Macrophage	RNA	14	17	31
		ATAC	31	46	77
	Mesenchymal stem cell	RNA	110	154	264
		ATAC	111	161	272
	Neuromuscular junction	RNA	61	97	158
		ATAC	57	89	146
	Neuronal	RNA	33	4	37
		ATAC	111	140	251
	Satellite cell	RNA	65	64	129
		ATAC	74	101	175
	Smooth muscle	RNA	97	129	226
		ATAC	114	154	268
	T cell	RNA	1	1	2
		ATAC	115	163	278
	Type 1	RNA	117	162	279
		ATAC	117	163	280
	Type 2a	RNA	117	161	278
		ATAC	116	161	277
	Type 2x	RNA	114	160	274
		ATAC	116	163	279

Supplementary Table 4.1 Sample characteristics and sample size by datatype of 118 female and 163 male vastus lateralis biopsy donors from the FUSION Tissue Biopsy Study.

A. Mean and standard deviations of age and BMI at biopsy and the number and percent of samples with normal glucose tolerance (NGT), impaired fasting glucose (IFG), impaired glucose tolerance (IGT) and type 2 diabetes (T2D) from an oral glucose tolerance test (OGTT) by sex. B. Sample sizes by data type and sex. The sample size of individual cell types in the single nucleus data are the number of individuals with ≥ 10 nuclei per cell type and modality.

Cell type	Modality	Female	Male	Total
Adipocyte	RNA	0.00 (0.01)	0.00 (0.00)	0.00 (0.01)
	ATAC	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)
	Combined	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
Endothelial	RNA	0.05 (0.03)	0.05 (0.02)	0.05 (0.03)
	ATAC	0.13 (0.05)	0.12 (0.04)	0.12 (0.04)
	Combined	0.09 (0.03)	0.10 (0.03)	0.10 (0.03)
Macrophage	RNA	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
	ATAC	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
	Combined	0.01 (0.00)	0.01 (0.01)	0.01 (0.01)
Mesenchymal stem cell	RNA	0.05 (0.02)	0.05 (0.03)	0.05 (0.02)
	ATAC	0.05 (0.02)	0.05 (0.03)	0.05 (0.03)
	Combined	0.05 (0.02)	0.05 (0.03)	0.05 (0.02)
Neuromuscular junction	RNA	0.03 (0.03)	0.03 (0.04)	0.03 (0.04)
	ATAC	0.01 (0.01)	0.02 (0.02)	0.02 (0.02)
	Combined	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)
Neuronal	RNA	0.01 (0.01)	0.00 (0.00)	0.01 (0.01)
	ATAC	0.04 (0.01)	0.03 (0.01)	0.04 (0.02)
	Combined	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)
Satellite cell	RNA	0.02 (0.01)	0.01 (0.01)	0.02 (0.01)
	ATAC	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
	Combined	0.02 (0.01)	0.01 (0.01)	0.01 (0.01)
Smooth muscle	RNA	0.03 (0.02)	0.04 (0.0)	0.03 (0.03)
	ATAC	0.04 (0.02)	0.04 (0.02)	0.04 (0.02)
	Combined	0.04 (0.01)	0.04 (0.02)	0.04 (0.02)
T cell	RNA	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	ATAC	0.06 (0.03)	0.06 (0.02)	0.06 (0.02)
	Combined	0.04 (0.01)	0.03 (0.01)	0.03 (0.01)
Type 1	RNA	0.44 (0.12)	0.37 (0.12)	0.40 (0.12)
	ATAC	0.33 (0.10)	0.27 (0.09)	0.30 (0.10)
	Combined	0.38 (0.10)	0.31 (0.10)	0.34 (0.11)
Type 2a	RNA	0.25 (0.08)	0.25 (0.09)	0.25 (0.08)
	ATAC	0.16 (0.06)	0.16 (0.06)	0.16 (0.06)
	Combined	0.20 (0.07)	0.20 (0.07)	0.20 (0.07)
Type 2x	RNA	0.11 (0.06)	0.18 (0.09)	0.15 (0.09)
	ATAC	0.13 (0.06)	0.20 (0.08)	0.17 (0.08)
	Combined	0.12 (0.06)	0.19 (0.08)	0.16 (0.08)

Supplementary Table 4.2 Mean and standard deviation of the proportion of nuclei by cell type and sex

The statistics were calculated across 281 samples for snATAC-seq nuclei and 279 samples for snRNA-seq nuclei. The combined rows are calculated as the mean proportion of the sum of snATAC-seq and snRNA-seq nuclei across the 279 samples with both snATAC-seq and snRNA-seq data.

Cell type	Modality	Fold change (M to F) and 95% confidence interval	FDR adjusted p-value
Adipocyte	RNA	0.77 (0.47, 1.27)	0.54
	ATAC	0.91 (0.82, 1.01)	0.18
	Combined	0.92 (0.81, 1.04)	0.31
Endothelial	RNA	1.04 (0.94, 1.14)	0.54
	ATAC	1.02 (0.95, 1.09)	0.61
	Combined	1.03 (0.96, 1.11)	0.46
Macrophage	RNA	0.95 (0.80, 1.13)	0.61
	ATAC	1.07 (0.90, 1.26)	0.56
	Combined	1.05 (0.91, 1.21)	0.62
Mesenchymal stem cell	RNA	1.05 (0.95, 1.16)	0.54
	ATAC	1.09 (0.99, 1.21)	0.18
	Combined	1.08 (0.99, 1.19)	0.22
Neuromuscular junction	RNA	1.11 (0.89, 1.38)	0.54
	ATAC	1.18 (0.99, 1.41)	0.17
	Combined	1.15 (0.96, 1.38)	0.28
Neuronal	RNA	0.39 (0.32, 0.47)	5.0×10^{-19}
	ATAC	0.89 (0.83, 0.96)	0.015
	Combined	0.80 (0.74, 0.87)	1.6×10^{-6}
Satellite cell	RNA	0.84 (0.75, 0.93)	0.0044
	ATAC	0.95 (0.86, 1.04)	0.44
	Combined	0.90 (0.83, 0.98)	0.039
Smooth muscle	RNA	1.06 (0.93, 1.21)	0.54
	ATAC	1.03 (0.94, 1.14)	0.56
	Combined	1.05 (0.95, 1.16)	0.45
T cell	RNA	1.12 (0.78, 1.60)	0.61
	ATAC	0.95 (0.88, 1.02)	0.32
	Combined	0.98 (0.90, 1.06)	0.62
Type 1	RNA	0.84 (0.78, 0.91)	8.5×10^{-5}
	ATAC	0.83 (0.77, 0.90)	7.6×10^{-5}
	Combined	0.83 (0.77, 0.90)	1.4×10^{-5}
Type 2a	RNA	1.02 (0.93, 1.12)	0.66
	ATAC	0.97 (0.87, 1.07)	0.56
	Combined	0.99 (0.90, 1.08)	0.77
Type 2x	RNA	1.60 (1.40, 1.83)	8.6×10^{-11}
	ATAC	1.49 (1.34, 1.65)	2.4×10^{-12}
	Combined	1.52 (1.36, 1.70)	4.8×10^{-12}

Supplementary Table 4.3 Associations of sex with the number of nuclei in each cell type from negative binomial regressions

Fold changes >1 indicate more nuclei in males than females. FDR adjusted p-values were calculated separately for each modality.

Cell type	Gene type	Number of genes tested	Mean (SD) UMI per gene
Endothelial	Protein coding	11,670	7.63 (17.80)
	lncRNA	2,061	3.50 (42.23)
	Pseudogene	494	2.71 (5.78)
	Other	104	0.77 (0.87)
	All	14,329	6.82 (22.77)
Macrophage	Protein coding	10,144	3.98 (9.95)
	lncRNA	1,276	2.62 (27.75)
	Pseudogene	311	1.92 (7.32)
	Other	45	0.61 (0.59)
	All	11,776	3.76 (13.06)
Mesenchymal stem cell	Protein coding	11,844	8.36 (24.31)
	lncRNA	2,399	4.44 (66.28)
	Pseudogene	528	2.73 (4.81)
	Other	108	0.73 (0.60)
	All	14,879	7.48 (34.39)
Neuromuscular junction	Protein coding	10,847	8.97 (40.97)
	lncRNA	2,040	7.98 (150.86)
	Pseudogene	438	3.16 (9.51)
	Other	92	0.89 (0.74)
	All	13,417	8.58 (69.43)
Neuronal	Protein coding	10,089	3.75 (12.24)
	lncRNA	2,579	3.10 (12.54)
	Pseudogene	469	1.55 (2.92)
	Other	49	0.60 (0.40)
	All	13,186	3.53 (12.08)
Satellite cell	Protein coding	9,806	3.40 (7.38)
	lncRNA	1,330	3.05 (26.00)
	Pseudogene	294	1.65 (2.23)
	Other	32	0.59 (0.35)
	All	11,462	3.31 (11.19)
Smooth muscle	Protein coding	11,467	6.69 (17.23)
	lncRNA	1,846	4.51 (72.31)
	Pseudogene	402	2.22 (3.83)
	Other	85	0.88 (0.86)
	All	13,800	6.24 (30.78)
Type 1	Protein coding	14,502	97.86 (705.44)
	lncRNA	6,520	32.04 (1,069.12)
	Pseudogene	2,023	9.69 (68.85)
	Other	796	2.42 (4.84)
	All	23,841	69.20 (785.48)
Type 2a	Protein coding	12,879	50.42 (367.71)
	lncRNA	4,690	19.54 (561.85)
	Pseudogene	1,314	6.58 (44.63)
	Other	483	1.72 (2.87)
	All	19,366	38.75 (408.38)
Type 2x	Protein coding	12,359	31.63 (205.55)
	lncRNA	3,897	13.09 (277.75)
	Pseudogene	1,021	5.08 (27.43)
	Other	345	1.41 (2.03)
	All	17,622	25.40 (216.40)

Supplementary Table 4.4 Number and mean UMI of genes tested for differential expression by sex by gene type and cell type. Gene types were defined by GENCODE

Cell type	Gene type	Unadjusted for gene count		Adjusted for gene count	
		Odds Ratio and 95% confidence interval	P-value	Odds Ratio and 95% confidence interval	P-value
Endothelial	lncRNA	1.62 (0.89, 2.94)	0.11	3.55 (1.89, 6.67)	8.6x10 ⁻⁵
	Pseudogene	1.45 (0.45, 4.67)	0.53	2.91 (0.89, 9.53)	0.078
	Other	0 (0, Inf)	0.98	0 (0, Inf)	0.99
Macrophage	lncRNA	7.97 (1.61, 39.51)	0.011	30.27 (5.77, 158.84)	5.5x10 ⁻⁵
	Pseudogene	0 (0, Inf)	0.99	0 (0, Inf)	1.00
	Other	0 (0, Inf)	1.00	0 (0, Inf)	1.00
Mesenchymal stem cell	lncRNA	0.79 (0.59, 1.06)	0.12	1.69 (1.24, 2.31)	9.3x10 ⁻⁴
	Pseudogene	1.24 (0.76, 2.01)	0.38	2.53 (1.54, 4.17)	2.6x10 ⁻⁴
	Other	0 (0, Inf)	0.96	0 (0, Inf)	0.96
Neuromuscular junction	lncRNA	1.10 (0.62, 1.95)	0.76	1.86 (1.02, 3.38)	0.043
	Pseudogene	1.83 (0.73, 4.56)	0.19	3.37 (1.33, 8.59)	0.011
	Other	0 (0, Inf)	0.98	0 (0, Inf)	0.98
Satellite cell	lncRNA	1.25 (0.80, 1.95)	0.32	2.18 (1.37, 3.48)	0.0011
	Pseudogene	1.99 (0.97, 4.10)	0.062	3.91 (1.85, 8.26)	3.5x10 ⁻⁴
	Other	2.29 (0.31, 16.92)	0.42	13.99 (1.79, 109.35)	0.012
Smooth muscle	lncRNA	2.34 (1.08, 5.03)	0.03	4.95 (2.20, 11.17)	1.2x10 ⁻⁴
	Pseudogene	3.58 (1.08, 11.95)	0.038	7.91 (2.29, 27.36)	0.0011
	Other	0 (0, Inf)	0.99	0 (0, Inf)	0.99
Type 1	lncRNA	0.59 (0.54, 0.64)	3.8x10 ⁻³¹	1.38 (1.25, 1.53)	5.1x10 ⁻¹⁰
	Pseudogene	0.44 (0.37, 0.52)	1.3x10 ⁻²²	1.30 (1.09, 1.56)	0.0037
	Other	0.15 (0.10, 0.23)	1.7x10 ⁻¹⁹	0.60 (0.39, 0.91)	0.017
Type 2a	lncRNA	0.64 (0.57, 0.71)	3.8x10 ⁻¹⁷	1.49 (1.32, 1.68)	4.1x10 ⁻¹¹
	Pseudogene	0.46 (0.38, 0.57)	3.0x10 ⁻¹³	1.25 (1.00, 1.56)	0.049
	Other	0.19 (0.11, 0.31)	3.9x10 ⁻¹¹	0.70 (0.42, 1.16)	0.17
Type 2x	lncRNA	0.68 (0.60, 0.76)	1.9x10 ⁻¹⁰	1.56 (1.36, 1.78)	9.9x10 ⁻¹¹
	Pseudogene	0.57 (0.45, 0.72)	2.0x10 ⁻⁶	1.50 (1.17, 1.92)	0.0014
	Other	0.30 (0.18, 0.50)	4.6x10 ⁻⁶	1.18 (0.69, 2.01)	0.55

Supplementary Table 4.5 Association of gene type with differential expression by sex status

Summary statistics comparing the likelihood of differential gene expression by sex for lncRNAs, pseudogenes, and other non-protein coding genes compared to protein-coding genes from logistic regression models not adjusting and adjusting for categories of gene count. All gene type categories with an odds ratio of 0 had no significantly differentially expressed genes.

Chromatin state (symbol)	Fiber type	Unadjusted for peak count		Adjusted for peak count	
		Odds Ratio and 95% confidence interval	P-value	Odds Ratio and 95% confidence interval	P-value
Active TSS (TssA)	Type 1	9.58 (9.12, 10.07)	0.00	0.37 (0.34, 0.39)	7.0x10 ⁻¹⁸²
	Type 2a	4.42 (4.23, 4.63)	0.00	0.39 (0.37, 0.42)	1.2x10 ⁻¹⁸⁵
	Type 2x	3.92 (3.68, 4.17)	0.00	0.24 (0.22, 0.26)	1.0x10 ⁻²⁸¹
Flanking Active TSS (TssAFlnk)	Type 1	13.89 (13.27, 14.54)	0.00	0.60 (0.57, 0.64)	6.2x10 ⁻⁷⁰
	Type 2a	4.28 (4.09, 4.48)	0.00	0.57 (0.54, 0.60)	3.3x10 ⁻⁹⁶
	Type 2x	7.33 (6.95, 7.72)	0.00	0.42 (0.39, 0.45)	8.4x10 ⁻¹⁶⁶
Transcription at gene 5' and 3' (TxFlnk)	Type 1	9.90 (8.13, 12.05)	6.1x10 ⁻¹¹⁵	0.42 (0.34, 0.51)	2.2x10 ⁻¹⁷
	Type 2a	2.71 (2.26, 3.26)	6.9x10 ⁻²⁷	0.34 (0.28, 0.41)	2.8x10 ⁻²⁹
	Type 2x	4.97 (3.95, 6.24)	3.1x10 ⁻⁴³	0.27 (0.21, 0.34)	1.3x10 ⁻²⁸
Strong transcription (Tx)	Type 1	0.88 (0.82, 0.95)	0.0010	0.22 (0.21, 0.24)	1.8x10 ⁻³⁰¹
	Type 2a	0.29 (0.27, 0.31)	3.2x10 ⁻²⁶⁷	0.27 (0.26, 0.29)	7.5x10 ⁻²⁸²
	Type 2x	0.58 (0.53, 0.63)	8.9x10 ⁻³⁸	0.29 (0.26, 0.31)	2.5x10 ⁻¹⁷⁴
Weak transcription (TxWk)	Type 1	1.42 (1.35, 1.49)	6.2x10 ⁻⁴⁷	0.43 (0.41, 0.45)	9.8x10 ⁻²²⁶
	Type 2a	0.49 (0.47, 0.51)	1.7x10 ⁻¹⁹⁵	0.43 (0.41, 0.45)	1.7x10 ⁻²⁴⁷
	Type 2x	0.79 (0.75, 0.84)	2.2x10 ⁻¹⁵	0.43 (0.41, 0.46)	1.3x10 ⁻¹⁵⁵
Genic enhancers (EnhG)	Type 1	7.07 (6.60, 7.58)	0.00	0.48 (0.44, 0.51)	1.2x10 ⁻⁸³
	Type 2a	1.94 (1.82, 2.07)	1.2x10 ⁻⁸⁹	0.47 (0.44, 0.51)	5.3x10 ⁻⁹⁷
	Type 2x	4.17 (3.86, 4.50)	1.8x10 ⁻²⁸⁸	0.41 (0.38, 0.45)	1.1x10 ⁻⁹⁵
Enhancers (Enh)	Type 1	8.28 (8.02, 8.55)	0.00	0.73 (0.70, 0.76)	2.1x10 ⁻⁵¹
	Type 2a	2.38 (2.30, 2.46)	0.00	0.73 (0.70, 0.76)	1.0x10 ⁻⁵⁵
	Type 2x	4.63 (4.46, 4.80)	0.00	0.63 (0.60, 0.66)	1.6x10 ⁻⁸⁸
ZNF genes and repeats (ZNF/Rpts)	Type 1	2.43 (1.67, 3.53)	3.4x10 ⁻⁶	0.97 (0.66, 1.44)	0.89
	Type 2a	0.54 (0.33, 0.90)	0.018	0.56 (0.33, 0.94)	0.028
	Type 2x	1.72 (1.12, 2.65)	0.013	1.00 (0.64, 1.58)	0.99
Hetero-chromatin (Het)	Type 1	0.50 (0.36, 0.69)	2.4x10 ⁻⁵	0.77 (0.55, 1.08)	0.13
	Type 2a	0.48 (0.29, 0.77)	0.0028	0.63 (0.38, 1.04)	0.069
	Type 2x	0.28 (0.17, 0.46)	4.2x10 ⁻⁷	0.45 (0.28, 0.75)	0.0019
Bivalent/ Poised TSS (TssBiv)	Type 1	6.30 (5.43, 7.31)	1.9x10 ⁻¹³⁰	0.50 (0.43, 0.58)	1.7x10 ⁻¹⁸
	Type 2a	2.19 (1.92, 2.49)	1.1x10 ⁻³²	0.60 (0.52, 0.69)	2.8x10 ⁻¹³
	Type 2x	4.07 (3.47, 4.78)	7.6x10 ⁻⁶⁷	0.45 (0.38, 0.53)	5.5x10 ⁻²¹
Flanking Bivalent TSS/ Enhancer (BivFlnk)	Type 1	4.90 (4.39, 5.48)	4.1x10 ⁻¹⁷²	0.33 (0.29, 0.37)	4.1x10 ⁻⁷⁹
	Type 2a	1.84 (1.68, 2.02)	5.1x10 ⁻³⁹	0.45 (0.41, 0.50)	2.9x10 ⁻⁵⁸
	Type 2x	2.78 (2.45, 3.16)	4.7x10 ⁻⁵⁷	0.28 (0.24, 0.32)	1.4x10 ⁻⁸²
Bivalent Enhancer (EnhBiv)	Type 1	4.97 (4.46, 5.55)	2.0x10 ⁻¹⁸⁰	0.56 (0.50, 0.62)	2.0x10 ⁻²³
	Type 2a	1.51 (1.37, 1.67)	3.5x10 ⁻¹⁶	0.65 (0.59, 0.73)	2.3x10 ⁻¹⁵
	Type 2x	2.84 (2.51, 3.22)	2.5x10 ⁻⁶¹	0.51 (0.45, 0.58)	7.7x10 ⁻²⁴
Repressed PolyComb (ReprPC)	Type 1	1.28 (1.17, 1.39)	1.3x10 ⁻⁸	0.42 (0.39, 0.46)	9.8x10 ⁻⁸³
	Type 2a	0.52 (0.48, 0.56)	7.0x10 ⁻⁶⁴	0.57 (0.53, 0.61)	5.4x10 ⁻⁴⁶
	Type 2x	0.77 (0.70, 0.85)	8.5x10 ⁻⁸	0.48 (0.43, 0.53)	1.6x10 ⁻⁴⁷
Weak Repressed PolyComb (ReprPCWk)	Type 1	1.43 (1.35, 1.51)	9.8x10 ⁻³⁹	0.90 (0.85, 0.95)	1.7x10 ⁻⁴
	Type 2a	0.92 (0.87, 0.98)	0.0043	1.01 (0.95, 1.07)	0.77
	Type 2x	1.01 (0.95, 1.08)	0.69	0.91 (0.85, 0.97)	0.0056

Supplementary Table 4.6 Association of chromatin state with differential accessibility by sex status for autosomal peaks

Summary statistics comparing the likelihood of differential chromatin accessibility by sex for consensus chromatin states (same annotation in male and female reference samples) compared to the quiescent/low signal chromatin state from logistic regression models, with and without adjusting for categories of peak count.

Cell type (genes, promoter peaks)	a. Male-biased expression		b. Female-biased expression	
	Odds ratio and 95% confidence interval	P-value	Odds ratio and 95% confidence interval	P-value
Type 1, Type 1	4.58 (3.76, 5.58)	2.8×10^{-51}	5.00 (4.38, 5.71)	3.5×10^{-126}
Type 2a, Type 2a	4.95 (4.05, 6.06)	2.7×10^{-54}	3.45 (3.05, 3.91)	1.7×10^{-86}
Type 2x, Type 2x	4.22 (3.33, 5.35)	1.8×10^{-32}	7.92 (6.54, 9.61)	1.5×10^{-98}
Pseudobulk, Type 1	3.02 (2.42, 3.77)	2.5×10^{-22}	3.76 (3.27, 4.33)	7.9×10^{-76}
Bulk, Type 1	1.85 (1.55, 2.21)	5.8×10^{-12}	2.89 (2.60, 3.22)	4.0×10^{-83}

Supplementary Table 4.7 Association of sex-biased promoter peaks with sex-biased gene expression in the fiber types, the single nucleus pseudobulk, and the bulk

The odds ratios, 95% confidence intervals, and p-values from logistic regression models testing the association of (a) male-biased expression with having at least one male-biased peak ≥ 1 kb of gene TSS compared to having zero sex-biased peaks ≥ 1 kb of TSS and (b) female-biased expression with having at least one female-biased peak ≥ 1 kb of gene TSS compared to having zero sex-biased peaks.

Chapter 5 Discussion

5.1 Summary and immediate extensions

Translating GWAS signals into a more complete understanding of the molecular mechanisms contributing to disease risk remains challenging. In this dissertation, I have presented three projects that address distinct challenges in this pipeline, from variant ascertainment and association detection (Chapter 2) to the identification of causal genes (Chapter 3) to the characterization of diverse biological contexts in which genes function (Chapter 4). Here, I summarize the major findings from each chapter and propose potential extensions for future work.

In Chapter 2, we quantified the extent to which array genotyping and imputation can approximate whole genome sequencing. We found that for three major US populations (European American, African/African American, and Hispanic/Latino), imputation with the TOPMed reference panel can accurately estimate the genotypes of nearly all common and low-frequency SNVs. Researchers interested in studying the effects of these variants (e.g. with GWAS) can rely on TOPMed-based imputation for variant ascertainment and invest in larger sample sizes instead of more costly sequencing, with the caveat that TOPMed-based imputation quality varies substantially with genomic location. Researchers interested in studying rare variation in these populations can also use TOPMed-based imputation to investigate the effects of large numbers of rare variant genotypes. However, because ~50% of rare SNVs and even more indels and multiallelic variants are still not accurately imputed in these populations,

sequencing is still necessary when a more comprehensive set of rare variants is needed (e.g. for clinical use).

Because technologies that capture genetic variation continue to evolve and costs continue to decrease, it is essential to have frameworks and tools for comparing technologies to inform cost-effective study design decisions. The specific price differentials and minor allele frequency thresholds presented in Chapter 2 represent a snapshot of the field at the time of publication but will soon be outdated. For instance, due to a combination of updated sequencers, expiring patents, and a proliferation of start-up companies, short-read whole genome sequencing prices have fallen substantially over the last year to as low as \$200 per genome,¹⁶⁰ which is no longer an order of magnitude more expensive than most standard genotyping arrays. More studies may now be able to perform whole genome sequencing and directly study the effects of genetic variants across the minor allele frequency spectrum. Nonetheless, the framework for comparing two variant ascertainment strategies that we used in Chapter 2 is applicable to other, similar scenarios. For instance, long-read (>10kb read length) sequencing more accurately captures structural variants and resolves highly repetitive regions.¹⁶¹ However, it is even more expensive than short-read sequencing,¹⁶² which means that sample sizes are currently limited by price. A familiar strategy of imputing from a reference panel assayed with the more costly technology has been used to impute structural variants from a small number of samples with long-read sequencing to characterize their associations with cholesterol levels and height in an Icelandic population.¹⁶³ Our framework could be used here to evaluate the tradeoffs between accuracy/coverage and cost in order to make recommendations for researchers wishing to study the effects of many different types of genetic variation in diverse populations.

In Chapter 3, we evaluated the consequences of using existing methods for colocalization in a single cohort design in violation of the non-overlapping cohorts assumption. We showed that researchers can take advantage of the more powerful single-cohort design without inflated Type I error rates as long as non-genetic factors that affect both phenotypes are either small in magnitude or are measured and adjusted for in the marginal analyses. We showed that the existence of such factors can be estimated through probabilistic principal component analysis, but that adjusting for these estimates is not permissible due to inflated Type I error rates resulting from collider bias. The guidelines presented in this chapter can help researchers justify the use of existing colocalization methods or identify problematic scenarios where colocalization analysis alone is not likely to give reliable results

There are several avenues for extending colocalization methods, both in single- and multiple-cohort study designs. First, incorporating functional annotations into the priors of Bayesian colocalization methods may increase power to detect causal variants. For example, because open chromatin is associated with gene expression levels,¹⁴⁶ we would expect significant enrichment of caQTLs in eQTL sets; this enrichment has been observed in some cell types.^{164,165} Incorporating information about the variant's effects on chromatin accessibility or structure may therefore be useful in determining its causal eQTL status. Second, extending our analysis of single-cohort colocalization with two traits to a larger number of traits may be useful for cohorts for which large numbers of traits are measured on the same individuals (e.g. biobanks linked to electronic health record systems). HyPrColoc⁹⁶ and mvSUSIE¹⁶⁶ are methods that seek to perform multiple-trait colocalization and fine-mapping, respectively. However, HyPrColoc also assumes non-overlapping samples and provides only limited information about the consequences of violating this assumption. MvSUSIE, as it is not a colocalization method, does not provide

probabilities for causality for multiple traits. Finally, our strategy for measuring and accounting for confounding in the single-cohort design requires access to individual-level data, which is often impractical or impossible due to computational and privacy concerns. Methods based solely on adjusting summary statistics would be beneficial.

In Chapter 4, we characterized sex differences in gene expression and regulation at the cell-type and whole-tissue levels in skeletal muscle. We found highly concordant sex-biased expression of genes in mitochondrial activity (males) and muscle regeneration (females) across Type 1, Type 2A, and Type 2X muscle fibers and bulk muscle tissue, suggesting that the significant sex differences in fiber-type composition do not drive most sex differences seen in bulk data. LncRNAs and miRNAs, both classes of genes known for gene regulatory activity, showed extensive sex-biased expression in the fiber-type and bulk data, respectively. We found sex-biased chromatin accessibility to be ubiquitous but enriched in proximal and distal regulatory states; in gene promoters, sex-biased chromatin accessibility was positively associated with sex-biased gene expression. Binding sites for sex hormones and muscle-specific transcription factors were enriched in the sex-biased ATAC-seq peaks. Together, these results highlight nuclear and cytoplasmic mechanisms for sex-differential gene regulation in skeletal muscle.

Many of the findings raised in Chapter 4 pose new questions that could be answerable with additional data types. For example, many sex-biased genes were only observed in the bulk data and not in the single-nucleus data, which we hypothesized was due to a combination of greater power from higher counts in the bulk data and different regulatory processes in the cytoplasm compared to the nucleus of the cell. Spatial transcriptomics, which maps the location in addition to the expression levels of a gene,¹⁶⁷ could help resolve this question for individual genes. Additionally, our dataset was primarily comprised of individuals older than 50, which

meant that we were not well powered to ascertain sex differences in gene expression at younger ages or investigate interactions between sex and age. Because muscle tissue composition and gene expression are affected by age as well as sex,^{168–170} expanding our cohort to include individuals with younger ages would allow a more comprehensive assessment of sex differences in muscle across the lifespan. Finally, the associations that we identified between gene expression and sex are not sufficient to establish causal pathways that explain sex differences in muscle physiology or disease risk. Intermediate phenotypic data types, such as proteomics, could be useful for connecting sex-biased gene expression differences to higher-order phenotypes.

5.2 Emerging themes and future directions

5.2.1 Ancestral diversity in publicly available genetic and genomic resources

One emerging theme among the projects presented in this dissertation is the need for diverse (in terms of ancestry, sex, age, and disease/comorbidity status) cohorts with genetic and genomic data. To date, the majority of GWAS analyses have been conducted in populations of European ancestry, limiting genetic discovery and potentially exacerbating health disparities as genetic information is brought into clinical care.¹⁷¹ A necessary resource for conducting GWAS in non-European populations is an imputation reference panel (or panels) with substantial numbers of non-European haplotypes. Specifically, reference haplotypes should match the ancestry of the study sample as closely as possible. In Chapter 2, we illustrated that imputation with the TOPMed panel, which contains unprecedented numbers of African American and Hispanic/Latino samples, greatly improves imputation quality over the primarily European HRC panel in studies of African and Hispanic/Latino ancestry, particularly for individual with large proportions of (assumed) West African ancestry. TOPMed-based imputation in African American and Hispanic/Latino populations has led to the identification novel genetic

associations with type 2 diabetes,¹⁷² hematological traits,⁵⁰ and serum biomarkers,¹⁷³ demonstrating the utility of this resource.

In addition to genotype data, collecting functional genomic data on cohorts of diverse ancestry is essential to understanding the impact of genetic variation globally. The expression levels of some genes differ by ancestry due to genetic variation and environmental influences.^{174,175} The majority of samples in commonly-used, publicly-available resources with gene expression data like the GTEx Consortium and The Cancer Genome Atlas (TCGA) are of predominantly European ancestry.^{176,177} As discussed in Chapter 3, colocalization power decreases with the degree of LD mismatch between the eQTL and GWAS datasets. Therefore, these resources are less effective for colocalization with non-European GWAS datasets. As we showed that the single-cohort design has the potential for more powerful colocalization analyses, performing both eQTL and GWAS analyses in cohorts with African ancestry may be a particularly powerful strategy for uncovering the functional impact of noncoding genetic variants because African populations have less LD¹⁷⁸ and therefore higher fine-mapping power¹⁷⁹ than non-African populations.¹⁷⁸ Under the assumption that most true common causal variants and genes are shared across populations, this may be a powerful strategy for understanding of the functional impact in all ancestry groups of the genetic variants that are shared across populations.

5.2.2 Modeling continuous nature of genetic ancestry

As the size and diversity of genetic cohorts continue to increase, we will need new methods to appropriately analyze the data they generate. One area for methodological development is continuous ancestry modeling. Currently, many studies classify individuals into discrete population groups for convenience, counter to recommendations that the continuous nature of genetic ancestry be reflected in research whenever possible.¹⁸⁰ For example, in Chapter

2, we used four discrete population labels despite showing that imputation quality varies with fine-scale ancestry within these population groups. Evaluating the accuracy of polygenic risk scores (PRS) is one area in which modeling the continuous nature of genetic ancestry is particularly relevant. Because most GWAS are conducted primarily in European populations, PRS have lower predictive accuracy in non-European population groups.¹⁷¹ However, the individual-level accuracy of PRS has been shown to vary along a continuum of genetic ancestry.¹⁸¹ Future work in developing, training, and evaluating PRS while modeling continuous ancestry is necessary to fully characterize (and, ideally, optimize) PRS performance across individuals of diverse backgrounds.

5.2.3 Promises and challenges of translational genetics

In Chapter 1, we stated that one of the primary goals of human genetics was to identify genes that contribute to the risk of complex diseases to improve prevention and treatment strategies.¹ The work presented in Chapters 2-4 is focused on the first piece of that statement: identifying causal genes and their pathways. The second half of the statement, translating genetic information into useful clinical practices, is another active area of research. In the context of complex disease, PRS have shown promise in stratifying individuals based on their genetic risk for a wide range of diseases to inform intervention efforts.¹⁸² However, as previously discussed, PRS have higher predictive accuracy in European populations and therefore show a potential to exacerbate health disparities.¹⁷¹ Efforts to reduce these disparities include the methods to increase PRS portability from European to non-European populations¹⁸³ and the development of cross-ancestry PRS.¹⁸⁴ In the rare disease context, whole genome sequencing has led to much higher rates of diagnosis¹⁸⁵ and can inform treatment options through pharmacogenics.¹⁸⁶ However, using whole genome sequencing in the clinic also has the potential to exacerbate

health disparities because the European bias in reference datasets that are used for functional prediction has led to greater numbers of variants of unknown significance in non-European populations.¹⁸⁷

The emergence of electronic health record (EHR)-linked biobanks gives researchers access to clinical data on tens and hundreds of thousands of patients. Many biobanks have large numbers of non-European participants and, to varying degrees, reflect the ancestral diversity of their communities.^{52,188,189} EHR-linked biobanks are not only cost-effective resources for conducting genetic discovery analyses like GWAS on many diseases in diverse populations, but they also provide a platform for implementing and evaluating strategies for integrating genetics in clinical care.¹⁹⁰ Clinical trials using PRS and WGS data to inform clinical practices are ongoing,^{184,191} and much future work is needed to evaluate the efficacy of these strategies in a range of patient populations.

Fully realizing the potential of genetics to make broad and meaningful impacts on human health requires translating the biological knowledge we gain with GWAS and related analyses into clinical practice. In this endeavor, care must be taken to ensure the benefits of genetic research are shared as broadly as possible across people and populations.

Bibliography

1. Claussnitzer, M., Cho, J.H., Collins, R., Cox, N.J., Dermitzakis, E.T., Hurles, M.E., Kathiresan, S., Kenny, E.E., Lindgren, C.M., MacArthur, D.G., et al. (2020). A brief history of human disease genetics. *Nature* 577, 179–189. 10.1038/s41586-019-1879-7.
2. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47, D1005–D1012. 10.1093/nar/gky1120.
3. Zhang, F., and Lupski, J.R. (2015). Non-coding genetic variants in human disease. *Hum. Mol. Genet.* 24, R102-110. 10.1093/hmg/ddv259.
4. Cano-Gamez, E., and Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* 11.
5. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45, 580–585. 10.1038/ng.2653.
6. Taylor, D.L., Jackson, A.U., Narisu, N., Hemani, G., Erdos, M.R., Chines, P.S., Swift, A., Idol, J., Didion, J.P., Welch, R.P., et al. (2019). Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. *Proc. Natl. Acad. Sci.* 116, 10883–10888. 10.1073/pnas.1814263116.
7. Buenrostro, J., Wu, B., Chang, H., and Greenleaf, W. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* Ed. Frederick M Ausubel A1 109, 21.29.1-21.29.9. 10.1002/0471142727.mb2129s109.
8. Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. 10.1038/nrg2484.
9. Hwang, B., Lee, J.H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* 50, 1–14. 10.1038/s12276-018-0071-8.
10. Grindberg, R.V., Yee-Greenbaum, J.L., McConnell, M.J., Novotny, M., O'Shaughnessy, A.L., Lambert, G.M., Araúzo-Bravo, M.J., Lee, J., Fishman, M., Robbins, G.E., et al. (2013). RNA-sequencing from single nuclei. *Proc. Natl. Acad. Sci.* 110, 19802–19807. 10.1073/pnas.1319700110.

11. Nica, A.C., and Dermitzakis, E.T. (2013). Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. B Biol. Sci.* *368*, 20120362. 10.1098/rstb.2012.0362.
12. Kumasaka, N., Knights, A.J., and Gaffney, D.J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* *48*, 206–213. 10.1038/ng.3467.
13. Mountjoy, E., Schmidt, E.M., Carmona, M., Schwartzentruber, J., Peat, G., Miranda, A., Fumis, L., Hayhurst, J., Buniello, A., Karim, M.A., et al. (2021). An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet.* *53*, 1527–1533. 10.1038/s41588-021-00945-5.
14. Bomba, L., Walter, K., and Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* *18*, 77. 10.1186/s13059-017-1212-4.
15. Sazonovs, A., and Barrett, J.C. (2018). Rare-Variant Studies to Complement Genome-Wide Association Studies. *Annu. Rev. Genomics Hum. Genet.* *19*, 97–112. 10.1146/annurev-genom-083117-021641.
16. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* *20*, 467–484. 10.1038/s41576-019-0127-1.
17. Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype Imputation. *Annu. Rev. Genomics Hum. Genet.* *10*, 387–406. 10.1146/annurev.genom.9.081307.164242.
18. Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., et al. (2015). A global reference for human genetic variation. *Nature* *526*, 68–74. 10.1038/nature15393.
19. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* *48*, 1279–1283. 10.1038/ng.3643.
20. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* *590*, 290–299. 10.1038/s41586-021-03205-y.
21. Kanduri, C., Bock, C., Gundersen, S., Hovig, E., and Sandve, G.K. (2019). Colocalization analyses of genomic elements: approaches, recommendations and challenges. *Bioinformatics* *35*, 1615–1624. 10.1093/bioinformatics/bty835.
22. Hormozdiari, F., van de Bunt, M., Segrè, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* *99*, 1245–1260. 10.1016/j.ajhg.2016.10.003.

23. Hukku, A., Pividori, M., Luca, F., Pique-Regi, R., Im, H.K., and Wen, X. (2021). Probabilistic colocalization of genetic variants from complex and molecular traits: promise and limitations. *Am. J. Hum. Genet.* *108*, 25–35. 10.1016/j.ajhg.2020.11.012.
24. Yin, X., Chan, L.S., Bose, D., Jackson, A.U., VandeHaar, P., Locke, A.E., Fuchsberger, C., Stringham, H.M., Yu, K., Silva, L.F., et al. (2021). Genome-wide association study of 1,391 plasma metabolites in 6,136 Finnish men identifies 303 novel signals and provides biological insights into human diseases. *medRxiv*, 2021.10.19.21265094. 10.1101/2021.10.19.21265094.
25. Musci, R.J., Augustinavicius, J.L., and Volk, H. (2019). Gene-Environment Interactions in Psychiatry: Recent Evidence and Clinical Implications. *Curr. Psychiatry Rep.* *21*, 81. 10.1007/s11920-019-1065-5.
26. Chaste, P., and Leboyer, M. (2012). Autism risk factors: genes, environment, and gene-environment interactions. *Dialogues Clin. Neurosci.* *14*, 281–292. 10.31887/DCNS.2012.14.3/pchaste.
27. Kido, Y. (2016). Gene–environment interaction in type 2 diabetes. *Diabetol. Int.* *8*, 7–13. 10.1007/s13340-016-0299-2.
28. Westerman, K.E., Majarian, T.D., Giulianini, F., Jang, D.-K., Miao, J., Florez, J.C., Chen, H., Chasman, D.I., Udler, M.S., Manning, A.K., et al. (2022). Variance-quantitative trait loci enable systematic discovery of gene-environment interactions for cardiometabolic serum biomarkers. *Nat. Commun.* *13*, 3993. 10.1038/s41467-022-31625-5.
29. Waubant, E., Lucas, R., Mowry, E., Graves, J., Olsson, T., Alfredsson, L., and Langer-Gould, A. (2019). Environmental and genetic risk factors for MS: an integrated review. *Ann. Clin. Transl. Neurol.* *6*, 1905–1922. 10.1002/acn3.50862.
30. Georgiou, A., Demetriou, C.A., Christou, Y.P., Heraclides, A., Leonidou, E., Loukaides, P., Yiasoumi, E., Pantziaris, M., Kleopa, K.A., Papacostas, S.S., et al. (2019). Genetic and Environmental Factors Contributing to Parkinson’s Disease: A Case-Control Study in the Cypriot Population. *Front. Neurol.* *10*.
31. Marshall, W.A. (1970). Sex differences at puberty. *J. Biosoc. Sci.* *2*, 31–41. 10.1017/S0021932000023439.
32. Hart, D.A. (2022). Sex Differences in Biological Systems and the Conundrum of Menopause: Potential Commonalities in Post-Menopausal Disease Mechanisms. *Int. J. Mol. Sci.* *23*, 4119. 10.3390/ijms23084119.
33. Austad, S.N., and Fischer, K.E. (2016). Sex Differences in Lifespan. *Cell Metab.* *23*, 1022–1033. 10.1016/j.cmet.2016.05.019.
34. Voskuhl, R. (2011). Sex differences in autoimmune diseases. *Biol. Sex Differ.* *2*, 1. 10.1186/2042-6410-2-1.

35. Appelman, Y., van Rijn, B.B., ten Haaf, M.E., Boersma, E., and Peters, S.A.E. (2015). Sex differences in cardiovascular risk factors and disease prevention. *Atherosclerosis* 241, 211–218. 10.1016/j.atherosclerosis.2015.01.027.
36. Tadiri, C.P., Gisinger, T., Kautzky-Willer, A., Kublickiene, K., Herrero, M.T., Raparelli, V., Pilote, L., and Norris, C.M. (2020). The influence of sex and gender domains on COVID-19 cases and mortality. *CMAJ* 192, E1041–E1045. 10.1503/cmaj.200971.
37. Nieves, J.W. (2017). Sex-Differences in Skeletal Growth and Aging. *Curr. Osteoporos. Rep.* 15, 70–75. 10.1007/s11914-017-0349-0.
38. Mielke, M.M. (2018). Sex and Gender Differences in Alzheimer’s Disease Dementia. *Psychiatr. Times* 35, 14–17.
39. Bernabeu, E., Canela-Xandri, O., Rawlik, K., Talenti, A., Prendergast, J., and Tenesa, A. (2021). Sex differences in genetic architecture in the UK Biobank. *Nat. Genet.* 53, 1283–1289. 10.1038/s41588-021-00912-0.
40. Oliva, M., Muñoz-Aguirre, M., Kim-Hellmuth, S., Wucher, V., Gewirtz, A.D.H., Cotter, D.J., Parsana, P., Kasela, S., Balliu, B., Viñuela, A., et al. (2020). The impact of sex on gene expression across human tissues. *Science*.
41. Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R.B., Xu, C., Futema, M., Lawson, D., et al. (2015). The UK10K project identifies rare variants in health and disease. *Nature* 526, 82–90. 10.1038/nature14962.
42. Flannick, J., Mercader, J.M., Fuchsberger, C., Udler, M.S., Mahajan, A., Wessel, J., Teslovich, T.M., Caulkins, L., Koesterer, R., Barajas-Olmos, F., et al. (2019). Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* 570, 71–76. 10.1038/s41586-019-1231-2.
43. Van Hout, C.V., Tachmazidou, I., Backman, J.D., Hoffman, J.D., Liu, D., Pandey, A.K., Gonzaga-Jauregui, C., Khalid, S., Ye, B., Banerjee, N., et al. (2020). Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* 586, 749–756. 10.1038/s41586-020-2853-0.
44. Fernandez-Marmiesse, A., Gouveia, S., and Couce, M.L. (2018). NGS Technologies as a Turning Point in Rare Disease Research , Diagnosis and Treatment. *Curr. Med. Chem.* 25, 404–432. 10.2174/0929867324666170718101946.
45. Nishiguchi, K.M., Tearle, R.G., Liu, Y.P., Oh, E.C., Miyake, N., Benaglio, P., Harper, S., Koskiniemi-Kuendig, H., Venturini, G., Sharon, D., et al. (2013). Whole genome sequencing in patients with retinitis pigmentosa reveals pathogenic DNA structural changes and NEK2 as a new disease gene. *Proc. Natl. Acad. Sci. U. S. A.* 110, 16139–16144. 10.1073/pnas.1308243110.

46. Boycott, K.M., Vanstone, M.R., Bulman, D.E., and MacKenzie, A.E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet.* *14*, 681–691. 10.1038/nrg3555.
47. Cade, B.E., Lee, J., Sofer, T., Wang, H., Zhang, M., Chen, H., Gharib, S.A., Gottlieb, D.J., Guo, X., Lane, J.M., et al. (2021). Whole-genome association analyses of sleep-disordered breathing phenotypes in the NHLBI TOPMed program. *Genome Med.* *13*, 136. 10.1186/s13073-021-00917-8.
48. Das, S., Abecasis, G.R., and Browning, B.L. (2018). Genotype Imputation from Large Reference Panels. *Annu. Rev. Genomics Hum. Genet.* *19*, 73–96.
49. Liu, Q., Cirulli, E.T., Han, Y., Yao, S., Liu, S., and Zhu, Q. (2015). Systematic assessment of imputation performance using the 1000 Genomes reference panels. *Brief. Bioinform.* *16*, 549–562. 10.1093/bib/bbu035.
50. Kowalski, M.H., Qian, H., Hou, Z., Rosen, J.D., Tapia, A.L., Shan, Y., Jain, D., Argos, M., Arnett, D.K., Avery, C., et al. (2019). Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet.* *15*, e1008500. 10.1371/journal.pgen.1008500.
51. Verlouw, J.A.M., Clemens, E., de Vries, J.H., Zolk, O., Verkerk, A.J.M.H., am Zehnhoff-Dinnesen, A., Medina-Gomez, C., Lanvers-Kaminsky, C., Rivadeneira, F., Langer, T., et al. (2021). A comparison of genotyping arrays. *Eur. J. Hum. Genet.* *29*, 1611–1624. 10.1038/s41431-021-00917-7.
52. Belbin, G.M., Cullina, S., Wenric, S., Soper, E.R., Glicksberg, B.S., Torre, D., Moscati, A., Wojcik, G.L., Shemirani, R., Beckmann, N.D., et al. (2021). Toward a fine-scale population health monitoring system. *Cell* *184*, 2068–2083.e11. 10.1016/j.cell.2021.03.034.
53. Laakso, M., Kuusisto, J., Stančáková, A., Kuulasmaa, T., Pajukanta, P., Lusi, A.J., Collins, F.S., Mohlke, K.L., and Boehnke, M. (2017). The Metabolic Syndrome in Men study: a resource for studies of metabolic and cardiovascular diseases. *J. Lipid Res.* *58*, 481–493. 10.1194/jlr.O072629.
54. Johnsen, J.M., Fletcher, S.N., Dove, A., McCracken, H., Martin, B.K., Kircher, M., Josephson, N.C., Shendure, J., Ruuska, S., Valentino, L.A., et al. (2020). Results of Genetic Analysis of 11,341 Participants Enrolled in the My Life, Our Future (MLOF) Hemophilia Genotyping Initiative. *Blood* *136*, 19. 10.1182/blood-2020-140649.
55. Pato, M.T., Sobell, J.L., Medeiros, H., Abbott, C., Sklar, B.M., Buckley, P.F., Bromet, E.J., Escamilla, M.A., Fanous, A.H., Lehrer, D.S., et al. (2013). The genomic psychiatry cohort: partners in discovery. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet.* *162B*, 306–312. 10.1002/ajmg.b.32160.
56. Bigdeli, T.B., Genovese, G., Georgakopoulos, P., Meyers, J.L., Peterson, R.E., Iyegbe, C.O., Medeiros, H., Valderrama, J., Achtyes, E.D., Kotov, R., et al. (2020). Contributions of

- common genetic variants to risk of schizophrenia among individuals of African and Latino ancestry. *Mol. Psychiatry* 25, 2455–2467. 10.1038/s41380-019-0517-y.
57. Swerdlow, N.R., Gur, R.E., and Braff, D.L. (2015). Consortium on the Genetics of Schizophrenia (COGS) assessment of endophenotypes for schizophrenia: an introduction to this Special Issue of Schizophrenia Research. *Schizophr. Res.* 163, 9–16. 10.1016/j.schres.2014.09.047.
 58. Smith, E.N., Bloss, C.S., Badner, J.A., Barrett, T., Belmonte, P.L., Berrettini, W., Byerley, W., Coryell, W., Craig, D., Edenberg, H.J., et al. (2009). Genome-wide association study of bipolar disorder in European American and African American individuals. *Mol. Psychiatry* 14, 755–763. 10.1038/mp.2009.43.
 59. Nierenberg, A.A., Friedman, E.S., Bowden, C.L., Sylvia, L.G., Thase, M.E., Ketter, T., Ostacher, M.J., Leon, A.C., Reilly-Harrington, N., Iosifescu, D.V., et al. (2013). Lithium treatment moderate-dose use study (LiTMUS) for bipolar disorder: a randomized comparative effectiveness trial of optimized personalized treatment with and without lithium. *Am. J. Psychiatry* 170, 102–110. 10.1176/appi.ajp.2012.12060751.
 60. Sklar, P., Smoller, J.W., Fan, J., Ferreira, M. a. R., Perlis, R.H., Chambert, K., Nimgaonkar, V.L., McQueen, M.B., Faraone, S.V., Kirby, A., et al. (2008). Whole-genome association study of bipolar disorder. *Mol. Psychiatry* 13, 558–569. 10.1038/sj.mp.4002151.
 61. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. 10.1101/gr.094052.109.
 62. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics* 26, 2867–2873. 10.1093/bioinformatics/btq559.
 63. Jun, G., Wing, M.K., Abecasis, G.R., and Kang, H.M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res.* 25, 918–925. 10.1101/gr.176552.114.
 64. Teslovich, T.M., Kim, D.S., Yin, X., Stancáková, A., Jackson, A.U., Wielscher, M., Naj, A., Perry, J.R.B., Huyghe, J.R., Stringham, H.M., et al. (2018). Identification of seven novel loci associated with amino acid levels using single-variant and gene-based tests in 8545 Finnish men from the METSIM study. *Hum. Mol. Genet.* 27, 1664–1674. 10.1093/hmg/ddy067.
 65. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. *Nat. Genet.* 48, 1284–1287. 10.1038/ng.3656.
 66. Kang, H.M. (2020). APIGenome - Big data genomics analysis libraries & tools.
 67. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol.* 17, 122. 10.1186/s13059-016-0974-4.

68. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am. J. Hum. Genet.* *93*, 278–288. 10.1016/j.ajhg.2013.06.020.
69. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science* *319*, 1100–1104. 10.1126/science.1153717.
70. Taliun, D., Chothani, S.P., Schönherr, S., Forer, L., Boehnke, M., Abecasis, G.R., and Wang, C. (2017). LASER server: ancestry tracing with genotypes or sequence reads. *Bioinforma. Oxf. Engl.* *33*, 2056–2058. 10.1093/bioinformatics/btx075.
71. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* *32*, D493–496. 10.1093/nar/gkh103.
72. Browning, B.L. (2021). Beagle 5.3. Beagle 53. <http://faculty.washington.edu/browning/beagle/beagle.html>.
73. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* *81*, 559–575. 10.1086/519795.
74. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* *4*, s13742-015-0047–0048. 10.1186/s13742-015-0047-8.
75. Purcell, S.M., and Chang, C.C. (2022). PLINK 2.0. <http://www.cog-genomics.org/plink/2.0/>.
76. Stasinopoulos, D.M., and Rigby, R.A. (2008). Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *J. Stat. Softw.* *23*, 1–46. 10.18637/jss.v023.i07.
77. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytzky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* *20*, 1297–1303. 10.1101/gr.107524.110.
78. Ganel, L., Chen, L., Christ, R., Vangipurapu, J., Young, E., Das, I., Kanchi, K., Larson, D., Regier, A., Abel, H., et al. (2021). Mitochondrial genome copy number measured by DNA sequencing in human blood is strongly associated with metabolic traits via cell-type composition differences. *Hum. Genomics* *15*, 34. 10.1186/s40246-021-00335-2.
79. Parra, E.J., Marcini, A., Akey, J., Martinson, J., Batzer, M.A., Cooper, R., Forrester, T., Allison, D.B., Deka, R., Ferrell, R.E., et al. (1998). Estimating African American Admixture Proportions by Use of Population-Specific Alleles. *Am. J. Hum. Genet.* *63*, 1839–1851. 10.1086/302148.

80. Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M., Bustamante, C.D., and Ostrer, H. (2010). Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proc. Natl. Acad. Sci.* 107, 8954–8961. 10.1073/pnas.0914618107.
81. de Bakker, P.I.W., Mcvean, G., Sabeti, P.C., Miretti, M.M., Green, T., Marchini, J., Ke, X., Monsuur, A.J., Whittaker, P., Delgado, M., et al. (2006). A high resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* 38, 1166–1172. 10.1038/ng1885.
82. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337, 64–69. 10.1126/science.1219240.
83. Wojcik, G.L., Fuchsberger, C., Taliun, D., Welch, R., Martin, A.R., Shringarpure, S., Carlson, C.S., Abecasis, G., Kang, H.M., Boehnke, M., et al. (2018). Imputation-Aware Tag SNP Selection To Improve Power for Large-Scale, Multi-ethnic Association Studies. *G3 Bethesda Md* 8, 3255–3267. 10.1534/g3.118.200502.
84. Slatkin, M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* 9, 477–485. 10.1038/nrg2361.
85. Watanabe, K., Taskesen, E., van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* 8, 1826. 10.1038/s41467-017-01261-5.
86. Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., Pickrell, J., Jaffe, A.E., CommonMind Consortium, Pasaniuc, B., et al. (2018). A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinforma. Oxf. Engl.* 34, 2538–2545. 10.1093/bioinformatics/bty147.
87. Wu, Y., Broadaway, K.A., Raulerson, C.K., Scott, L.J., Pan, C., Ko, A., He, A., Tilford, C., Fuchsberger, C., Locke, A.E., et al. (2019). Colocalization of GWAS and eQTL signals at loci with multiple signals identifies additional candidate genes for body fat distribution. *Hum. Mol. Genet.* 28, 4161–4172. 10.1093/hmg/ddz263.
88. Yin, X., Bose, D., Kwon, A., Hanks, S.C., Jackson, A.U., Stringham, H.M., Welch, R., Oravilahti, A., Fernandes Silva, L., FinnGen, et al. (2022). Integrating transcriptomics, metabolomics, and GWAS helps reveal molecular mechanisms for metabolite levels and disease risk. *Am. J. Hum. Genet.* 109, 1727–1741. 10.1016/j.ajhg.2022.08.007.
89. Franceschini, N., Giambartolomei, C., de Vries, P.S., Finan, C., Bis, J.C., Huntley, R.P., Lovering, R.C., Tajuddin, S.M., Winkler, T.W., Graff, M., et al. (2018). GWAS and colocalization analyses implicate carotid intima-media thickness and carotid plaque loci in cardiovascular outcomes. *Nat. Commun.* 9, 5141. 10.1038/s41467-018-07340-5.

90. Gloudemans, M.J., Balliu, B., Nachun, D., Schnurr, T.M., Durrant, M.G., Ingelsson, E., Wabitsch, M., Quertermous, T., Montgomery, S.B., Knowles, J.W., et al. (2022). Integration of genetic colocalizations with physiological and pharmacological perturbations identifies cardiometabolic disease genes. *Genome Med.* *14*, 31. 10.1186/s13073-022-01036-8.
91. Fortune, M.D., Guo, H., Burren, O., Schofield, E., Walker, N.M., Ban, M., Sawcer, S.J., Bowes, J., Worthington, J., Barton, A., et al. (2015). Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nat. Genet.* *47*, 839–846. 10.1038/ng.3330.
92. McGowan, L.M., Davey Smith, G., Gaunt, T.R., and Richardson, T.G. (2019). Integrating Mendelian randomization and multiple-trait colocalization to uncover cell-specific inflammatory drivers of autoimmune and atopic disease. *Hum. Mol. Genet.* *28*, 3293–3300. 10.1093/hmg/ddz155.
93. Sun, B.B., Chiou, J., Traylor, M., Benner, C., Hsu, Y.-H., Richardson, T.G., Surendran, P., Mahajan, A., Robins, C., Vasquez-Grinnell, S.G., et al. (2022). Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. 2022.06.17.496443. 10.1101/2022.06.17.496443.
94. Wen, X., Pique-Regi, R., and Luca, F. (2017). Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLOS Genet.* *13*, e1006646. 10.1371/journal.pgen.1006646.
95. Pividori, M., Rajagopal, P.S., Barbeira, A., Liang, Y., Melia, O., Bastarache, L., Park, Y., Consortium, Gte., Wen, X., and Im, H.K. (2020). PhenomeXcan: Mapping the genome to the phenome through the transcriptome. *Sci. Adv.* *6*, eaba2083. 10.1126/sciadv.aba2083.
96. Foley, C.N., Staley, J.R., Breen, P.G., Sun, B.B., Kirk, P.D.W., Burgess, S., and Howson, J.M.M. (2021). A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nat. Commun.* *12*, 764. 10.1038/s41467-020-20885-8.
97. Wen, X., Lee, Y., Luca, F., and Pique-Regi, R. (2016). Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. *Am. J. Hum. Genet.* *98*, 1114–1129. 10.1016/j.ajhg.2016.03.029.
98. Yuan, K., Longchamps, R.J., Pardiñas, A.F., Yu, M., Chen, T.-T., Lin, S.-C., Chen, Y., Lam, M., Liu, R., Xia, Y., et al. (2023). Fine-mapping across diverse ancestries drives the discovery of putative causal variants underlying human complex traits and diseases. medRxiv, 2023.01.07.23284293. 10.1101/2023.01.07.23284293.
99. Yao, D.W., O'Connor, L.J., Price, A.L., and Gusev, A. (2020). Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* *52*, 626–633. 10.1038/s41588-020-0625-2.
100. Price, A.L., Helgason, A., Thorleifsson, G., McCarroll, S.A., Kong, A., and Stefansson, K. (2011). Single-Tissue and Cross-Tissue Heritability of Gene Expression Via Identity-by-

- Descent in Related or Unrelated Individuals. *PLOS Genet.* 7, e1001317. 10.1371/journal.pgen.1001317.
101. Hormozdiari, F., Gazal, S., van de Geijn, B., Finucane, H., Ju, C.J.-T., Loh, P.-R., Schoech, A., Reshef, Y., Liu, X., O'Connor, L., et al. (2018). Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* 50, 1041–1047. 10.1038/s41588-018-0148-2.
 102. Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* 7, 500–507. 10.1038/nprot.2011.457.
 103. Leek, J.T., and Storey, J.D. (2007). Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLOS Genet.* 3, e161. 10.1371/journal.pgen.0030161.
 104. Dahl, A., Guillemot, V., Mefford, J., Aschard, H., and Zaitlen, N. (2019). Adjusting for Principal Components of Molecular Phenotypes Induces Replicating False Positives. *Genetics* 211, 1179–1189. 10.1534/genetics.118.301768.
 105. Staron, R.S., Hagerman, F.C., Hikida, R.S., Murray, T.F., Hostler, D.P., Crill, M.T., Ragg, K.E., and Toma, K. (2000). Fiber type composition of the vastus lateralis muscle of young men and women. *J. Histochem. Cytochem. Off. J. Histochem. Soc.* 48, 623–629. 10.1177/002215540004800506.
 106. Janssen, I., Heymsfield, S.B., Wang, Z., and Ross, R. (2000). Skeletal muscle mass and distribution in 468 men and women aged 18–88 yr. *J. Appl. Physiol.* 89, 81–88. 10.1152/jap.2000.89.1.81.
 107. Haizlip, K.M., Harrison, B.C., and Leinwand, L.A. (2015). Sex-Based Differences in Skeletal Muscle Kinetics and Fiber-Type Composition. *Physiology* 30, 30–39. 10.1152/physiol.00024.2014.
 108. Staron, R.S., Hagerman, F.C., Hikida, R.S., Murray, T.F., Hostler, D.P., Crill, M.T., Ragg, K.E., and Toma, K. (2000). Fiber type composition of the vastus lateralis muscle of young men and women. *J. Histochem. Cytochem. Off. J. Histochem. Soc.* 48, 623–629. 10.1177/002215540004800506.
 109. Tramunt, B., Smati, S., Grandgeorge, N., Lenfant, F., Arnal, J.-F., Montagner, A., and Gourdy, P. (2020). Sex differences in metabolic regulation and diabetes susceptibility. *Diabetologia* 63, 453–461. 10.1007/s00125-019-05040-3.
 110. Lovejoy, J.C., Sainsbury, A., and Stock Conference 2008 Working Group (2009). Sex differences in obesity and the regulation of energy homeostasis. *Obes. Rev. Off. J. Int. Assoc. Study Obes.* 10, 154–167. 10.1111/j.1467-789X.2008.00529.x.
 111. Regitz-Zagrosek, V., and Kararigas, G. (2017). Mechanistic Pathways of Sex Differences in Cardiovascular Disease. *Physiol. Rev.* 97, 1–37. 10.1152/physrev.00021.2015.

112. Johnell, O., and Kanis, J.A. (2006). An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. *Osteoporos. Int.* *17*, 1726–1733. 10.1007/s00198-006-0172-4.
113. DeFronzo, R.A., and Tripathy, D. (2009). Skeletal muscle insulin resistance is the primary defect in type 2 diabetes. *Diabetes Care* *32 Suppl 2*, S157-163. 10.2337/dc09-S302.
114. Pedersen, B.K., and Febbraio, M.A. (2012). Muscles, exercise and obesity: skeletal muscle as a secretory organ. *Nat. Rev. Endocrinol.* *8*, 457–465. 10.1038/nrendo.2012.49.
115. Smith, A.G., and Muscat, G.E.O. (2005). Skeletal muscle and nuclear hormone receptors: Implications for cardiovascular and metabolic disease. *Int. J. Biochem. Cell Biol.* *37*, 2047–2063. 10.1016/j.biocel.2005.03.002.
116. Johannesdottir, F., Aspelund, T., Siggeirsdottir, K., Jonsson, B.Y., Mogensen, B., Sigurdsson, S., Harris, T.B., Gudnason, V.G., Lang, T.F., and Sigurdsson, G. (2012). Mid-thigh cortical bone structural parameters, muscle mass and strength and association with lower limb fractures in older men and women (AGES-Reykjavik Study). *Calcif. Tissue Int.* *90*, 354–364. 10.1007/s00223-012-9585-6.
117. Roth, S.M., Ferrell, R.E., Peters, D.G., Metter, E.J., Hurley, B.F., and Rogers, M.A. (2002). Influence of age, sex, and strength training on human muscle gene expression determined by microarray. *Physiol. Genomics* *10*, 181–190. 10.1152/physiolgenomics.00028.2002.
118. Welle, S., Tawil, R., and Thornton, C.A. (2008). Sex-Related Differences in Gene Expression in Human Skeletal Muscle. *PLoS ONE* *3*, e1385. 10.1371/journal.pone.0001385.
119. Maher, A.C., Fu, M.H., Isfort, R.J., Varbanov, A.R., Qu, X.A., and Tarnopolsky, M.A. (2009). Sex Differences in Global mRNA Content of Human Skeletal Muscle. *PLoS ONE* *4*, e6335. 10.1371/journal.pone.0006335.
120. Liu, D., Sartor, M.A., Nader, G.A., Gutmann, L., Treutelaar, M.K., Pistilli, E.E., Iglayreger, H.B., Burant, C.F., Hoffman, E.P., and Gordon, P.M. (2010). Skeletal muscle gene expression in response to resistance exercise: sex specific regulation. *BMC Genomics* *11*, 659. 10.1186/1471-2164-11-659.
121. Lindholm, M.E., Huss, M., Solnestam, B.W., Kjellqvist, S., Lundeberg, J., and Sundberg, C.J. (2014). The human skeletal muscle transcriptome: sex differences, alternative splicing, and tissue homogeneity assessed with RNA sequencing. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* *28*, 4571–4581. 10.1096/fj.14-255000.
122. Gershoni, M., and Petrokovski, S. (2017). The landscape of sex-differential transcriptome and its consequent selection in human adults. *BMC Biol.* *15*, 7. 10.1186/s12915-017-0352-z.
123. Lopes-Ramos, C.M., Chen, C.-Y., Kuijjer, M.L., Paulson, J.N., Sonawane, A.R., Fagny, M., Platig, J., Glass, K., Quackenbush, J., and DeMeo, D.L. (2020). Sex Differences in Gene

- Expression and Regulatory Networks across 29 Human Tissues. *Cell Rep.* 31, 107795. 10.1016/j.celrep.2020.107795.
124. Landen, S., Jacques, M., Hiam, D., Alvarez-Romero, J., Harvey, N.R., Haupt, L.M., Griffiths, L.R., Ashton, K.J., Lamon, S., Voisin, S., et al. (2021). Skeletal muscle methylome and transcriptome integration reveals profound sex differences related to muscle function and substrate metabolism. *Clin. Epigenetics* 13, 202. 10.1186/s13148-021-01188-1.
 125. Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., Young, T.R., Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J., et al. (2015). Human genomics. The human transcriptome across tissues and individuals. *Science* 348, 660–665. 10.1126/science.aaa0355.
 126. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* 102, 15545–15550. 10.1073/pnas.0506580102.
 127. Morgan, C.P., and Bale, T.L. (2012). Sex differences in microRNA regulation of gene expression: no smoke, just miRs. *Biol. Sex Differ.* 3, 22. 10.1186/2042-6410-3-22.
 128. Hiam, D., Landen, S., Jacques, M., Voisin, S., Lamon, S., and Eynon, N. (2023). The influence of Sex on microRNA expression in Human Skeletal Muscle (Physiology) 10.1101/2023.02.27.530361.
 129. Scott, L.J., Erdos, M.R., Huyghe, J.R., Welch, R.P., Beck, A.T., Wolford, B.N., Chines, P.S., Didion, J.P., Narisu, N., Stringham, H.M., et al. (2016). The genetic regulatory signature of type 2 diabetes in human skeletal muscle. *Nat. Commun.* 7, 11764. 10.1038/ncomms11764.
 130. Yang, S., Corbett, S.E., Koga, Y., Wang, Z., Johnson, W.E., Yajima, M., and Campbell, J.D. (2020). Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.* 21, 57. 10.1186/s13059-020-1950-6.
 131. Rozowsky, J., Kitchen, R.R., Park, J.J., Galeev, T.R., Diao, J., Warrell, J., Thistlethwaite, W., Subramanian, S.L., Milosavljevic, A., and Gerstein, M. (2019). exceRpt: A Comprehensive Analytic Platform for Extracellular RNA Profiling. *Cell Syst.* 8, 352-357.e3. 10.1016/j.cels.2019.03.004.
 132. Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Res.* 47, D155–D162. 10.1093/nar/gky1141.
 133. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. 10.1186/s13059-014-0550-8.

134. Lee, C., Patil, S., and Sartor, M.A. (2016). RNA-Enrich: a cut-off free functional enrichment testing method for RNA-seq with improved detection power. *Bioinforma. Oxf. Engl.* 32, 1100–1102. 10.1093/bioinformatics/btv694.
135. Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLOS ONE* 6, e21800. 10.1371/journal.pone.0021800.
136. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* 28, 1045–1048. 10.1038/nbt1010-1045.
137. D'Oliveira Albanus, R., Kyono, Y., Hensley, J., Varshney, A., Orchard, P., Kitzman, J.O., and Parker, S.C.J. (2021). Chromatin information content landscapes inform transcription factor and DNA interactions. *Nat. Commun.* 12, 1307. 10.1038/s41467-021-21534-4.
138. THE GTEx CONSORTIUM (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330. 10.1126/science.aaz1776.
139. Agarwal, V., Bell, G.W., Nam, J.-W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *eLife* 4, e05005. 10.7554/eLife.05005.
140. Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafrenière, R.G., Xing, Y., Lawrence, J., and Willard, H.F. (1992). The human XIST gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* 71, 527–542. 10.1016/0092-8674(92)90520-M.
141. Lee, J., Davidow, L.S., and Warshawsky, D. (1999). Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat. Genet.* 21, 400–404. 10.1038/7734.
142. O'Brien, J., Hayder, H., Zayed, Y., and Peng, C. (2018). Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Front. Endocrinol.* 9, 402. 10.3389/fendo.2018.00402.
143. Mitra, R., Sun, J., and Zhao, Z. (2015). microRNA regulation in cancer: one arm or two arms? *Int. J. Cancer* 137, 1516–1518. 10.1002/ijc.29512.
144. Mitra, R., Lin, C.-C., Eischen, C.M., Bandyopadhyay, S., and Zhao, Z. (2015). Concordant dysregulation of miR-5p and miR-3p arms of the same precursor microRNA may be a mechanism in inducing cell proliferation and tumorigenesis: a lung cancer study. *RNA N. Y. N* 21, 1055–1065. 10.1261/rna.048132.114.
145. Ludwig, N., Leidinger, P., Becker, K., Backes, C., Fehlmann, T., Pallasch, C., Rheinheimer, S., Meder, B., Stähler, C., Meese, E., et al. (2016). Distribution of miRNA expression across human tissues. *Nucleic Acids Res.* 44, 3865–3877. 10.1093/nar/gkw116.

146. Klemm, S.L., Shipony, Z., and Greenleaf, W.J. (2019). Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* 20, 207–220. 10.1038/s41576-018-0089-8.
147. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330. 10.1038/nature14248.
148. Zhao, Y., Zheng, D., and Cvekl, A. (2019). Profiling of chromatin accessibility and identification of general cis-regulatory mechanisms that control two ocular lens differentiation pathways. *Epigenetics Chromatin* 12, 27. 10.1186/s13072-019-0272-y.
149. Yasuda, T., Ishihara, T., Ichimura, A., and Ishihara, N. (2023). Mitochondrial dynamics define muscle fiber type by modulating cellular metabolic pathways. *Cell Rep.* 42, 112434. 10.1016/j.celrep.2023.112434.
150. Mishra, P., Varuzhanyan, G., Pham, A.H., and Chan, D.C. (2015). Mitochondrial dynamics is a distinguishing feature of skeletal muscle fiber types and regulates organellar compartmentalization. *Cell Metab.* 22, 1033–1044. 10.1016/j.cmet.2015.09.027.
151. Volonte, D., Peoples, A.J., and Galbiati, F. (2003). Modulation of myoblast fusion by caveolin-3 in dystrophic skeletal muscle cells: implications for Duchenne muscular dystrophy and limb-girdle muscular dystrophy-1C. *Mol. Biol. Cell* 14, 4075–4088. 10.1091/mbc.e03-03-0161.
152. LaBarge, S., McDonald, M., Smith-Powell, L., Auwerx, J., and Huss, J.M. (2014). Estrogen-related receptor- α (ERR α) deficiency in skeletal muscle impairs regeneration in response to injury. *FASEB J.* 28, 1082–1097. 10.1096/fj.13-229211.
153. Chang, C.-N., Singh, A.J., Gross, M.K., and Kiousi, C. (2019). Requirement of Pitx2 for Skeletal Muscle Homeostasis. *Dev. Biol.* 445, 90–102. 10.1016/j.ydbio.2018.11.001.
154. Li, L., Tao, G., Hill, M.C., Zhang, M., Morikawa, Y., and Martin, J.F. (2018). Pitx2 maintains mitochondrial function during regeneration to prevent myocardial fat deposition. *Dev. Camb. Engl.* 145, dev168609. 10.1242/dev.168609.
155. Tseng, L.A., Delmonico, M.J., Visser, M., Boudreau, R.M., Goodpaster, B.H., Schwartz, A.V., Simonsick, E.M., Satterfield, S., Harris, T., and Newman, A.B. (2014). Body composition explains sex differential in physical performance among older adults. *J. Gerontol. A. Biol. Sci. Med. Sci.* 69, 93–100. 10.1093/gerona/glt027.
156. Lowe, M., Lage, J., Paatela, E., Munson, D., Hostager, R., Yuan, C., Katoku-Kikyo, N., Ruiz-Estevez, M., Asakura, Y., Staats, J., et al. (2018). Cry2 Is Critical for Circadian Regulation of Myogenic Differentiation by Bclaf1-Mediated mRNA Stabilization of Cyclin D1 and Tmem176b. *Cell Rep.* 22, 2118–2132. 10.1016/j.celrep.2018.01.077.
157. Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* 22, 96–118. 10.1038/s41580-020-00315-9.

158. Ferreira, P.G., Muñoz-Aguirre, M., Reverter, F., Sá Godinho, C.P., Sousa, A., Amadoz, A., Sodaei, R., Hidalgo, M.R., Pervouchine, D., Carbonell-Caballero, J., et al. (2018). The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nat. Commun.* *9*, 490. 10.1038/s41467-017-02772-x.
159. Decaroli, M.C., and Rochira, V. (2016). Aging and sex hormones in males. *Virulence* *8*, 545–570. 10.1080/21505594.2016.1259053.
160. Pollie, R. (2023). Genomic Sequencing Costs Set to Head Down Again. *Engineering* *23*, 3–6. 10.1016/j.eng.2023.02.002.
161. Mantere, T., Kersten, S., and Hoischen, A. (2019). Long-Read Sequencing Emerging in Medical Genetics. *Front. Genet.* *10*.
162. Adewale, B.A. (2020). Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years? *Afr. J. Lab. Med.* *9*, 1340. 10.4102/ajlm.v9i1.1340.
163. Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H.P., Bjornsson, E., Jonsson, H., Atlason, B.A., Kristmundsdottir, S., Mehringer, S., Hardarson, M.T., et al. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* *53*, 779–786. 10.1038/s41588-021-00865-4.
164. Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A.J., Mann, A.L., Kundu, K., Hale, C., Dougan, G., and Gaffney, D.J. (2018). Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* *50*, 424–431. 10.1038/s41588-018-0046-7.
165. Wang, D., Wu, X., Jiang, G., Yang, J., Yu, Z., Yang, Y., Yang, W., Niu, X., Tang, K., and Gong, J. (2022). Systematic analysis of the effects of genetic variants on chromatin accessibility to decipher functional variants in non-coding regions. *Front. Oncol.* *12*.
166. Zou, Y., Carbonetto, P., Xie, D., Wang, G., and Stephens, M. (2023). Fast and flexible joint fine-mapping of multiple traits via the Sum of Single Effects model. *BioRxiv Prepr. Serv. Biol.*, 2023.04.14.536893. 10.1101/2023.04.14.536893.
167. Williams, C.G., Lee, H.J., Asatsuma, T., Vento-Tormo, R., and Haque, A. (2022). An introduction to spatial transcriptomics for biomedical research. *Genome Med.* *14*, 68. 10.1186/s13073-022-01075-1.
168. Lexell, J. (1995). Human aging, muscle mass, and fiber type composition. *J. Gerontol. A. Biol. Sci. Med. Sci.* *50 Spec No*, 11–16. 10.1093/gerona/50a.special_issue.11.
169. Tumasian, R.A., Harish, A., Kundu, G., Yang, J.-H., Ubaida-Mohien, C., Gonzalez-Freire, M., Kaileh, M., Zukley, L.M., Chia, C.W., Lyashkov, A., et al. (2021). Skeletal muscle transcriptome in healthy aging. *Nat. Commun.* *12*, 2014. 10.1038/s41467-021-22168-2.

170. St-Onge, M.-P., and Gallagher, D. (2010). Body composition changes with aging: The cause or the result of alterations in metabolic rate and macronutrient oxidation? *Nutr. Burbank Los Angel. Cty. Calif* 26, 152–155. 10.1016/j.nut.2009.07.004.
171. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* 51, 584–591. 10.1038/s41588-019-0379-x.
172. Huerta-Chagoya, A., Schroeder, P., Mandla, R., Deutsch, A.J., Zhu, W., Petty, L., Yi, X., Cole, J.B., Udler, M.S., Dornbos, P., et al. (2023). The power of TOPMed imputation for the discovery of Latino-enriched rare variants associated with type 2 diabetes. *Diabetologia* 66, 1273–1288. 10.1007/s00125-023-05912-9.
173. Sun, Q., Graff, M., Rowland, B., Wen, J., Huang, L., Miller-Fleming, T.W., Haessler, J., Preuss, M.H., Chai, J.-F., Lee, M.P., et al. (2022). Analyses of biomarker traits in diverse UK biobank participants identify associations missed by European-centric analysis strategies. *J. Hum. Genet.* 67, 87–93. 10.1038/s10038-021-00968-0.
174. Kachuri, L., Mak, A.C.Y., Hu, D., Eng, C., Huntsman, S., Elhawary, J.R., Gupta, N., Gabriel, S., Xiao, S., Keys, K.L., et al. (2023). Gene expression in African Americans, Puerto Ricans and Mexican Americans reveals ancestry-specific patterns of genetic architecture. *Nat. Genet.* 55, 952–963. 10.1038/s41588-023-01377-z.
175. Benjamin, K.J.M., Chen, Q., Eagles, N.J., Huuki-Myers, L.A., Collado-Torres, L., Stolz, J.M., Shin, J.H., Paquola, A.C.M., Hyde, T.M., Kleinman, J.E., et al. (2023). Genetic and environmental contributions to ancestry differences in gene expression in the human brain. *BioRxiv Prepr. Serv. Biol.*, 2023.03.28.534458. 10.1101/2023.03.28.534458.
176. Gay, N.R., Gloudemans, M., Antonio, M.L., Abell, N.S., Balliu, B., Park, Y., Martin, A.R., Musharoff, S., Rao, A.S., Aguet, F., et al. (2020). Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. *Genome Biol.* 21, 233. 10.1186/s13059-020-02113-0.
177. Carrot-Zhang, J., Chambwe, N., Damrauer, J.S., Knijnenburg, T.A., Robertson, A.G., Yau, C., Zhou, W., Berger, A.C., Huang, K.-L., Newberg, J.Y., et al. (2020). Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer. *Cancer Cell* 37, 639–654.e6. 10.1016/j.ccell.2020.04.012.
178. Campbell, M.C., and Tishkoff, S.A. (2008). AFRICAN GENETIC DIVERSITY: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. *Annu. Rev. Genomics Hum. Genet.* 9, 403–433. 10.1146/annurev.genom.9.081307.164258.
179. Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M.O., Choudhury, A., et al. (2015). The African Genome Variation Project shapes medical genetics in Africa. *Nature* 517, 327–332. 10.1038/nature13997.

180. Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field (2023). (National Academies Press) 10.17226/26902.
181. Ding, Y., Hou, K., Xu, Z., Pimplaskar, A., Petter, E., Boulier, K., Privé, F., Vilhjálms, B.J., Olde Loohuis, L.M., and Pasaniuc, B. (2023). Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* 618, 774–781. 10.1038/s41586-023-06079-4.
182. Lewis, C.M., and Vassos, E. (2020). Polygenic risk scores: from research tools to clinical instruments. *Genome Med.* 12, 44. 10.1186/s13073-020-00742-5.
183. Wang, Y., Tsuo, K., Kanai, M., Neale, B.M., and Martin, A.R. (2022). Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. *Annu. Rev. Biomed. Data Sci.* 5, 293–320. 10.1146/annurev-biodatasci-111721-074830.
184. Linder, J.E., Allworth, A., Bland, H.T., Caraballo, P.J., Chisholm, R.L., Clayton, E.W., Crosslin, D.R., Dikilitas, O., DiVietro, A., Esplin, E.D., et al. (2023). Returning integrated genomic risk and clinical recommendations: The eMERGE study. *Genet. Med.* 25, 100006. 10.1016/j.gim.2023.100006.
185. Stranneheim, H., Lagerstedt-Robinson, K., Magnusson, M., Kvarnung, M., Nilsson, D., Lesko, N., Engvall, M., Anderlid, B.-M., Arnell, H., Johansson, C.B., et al. (2021). Integration of whole genome sequencing into a healthcare setting: high diagnostic rates across multiple clinical entities in 3219 rare disease patients. *Genome Med.* 13, 40. 10.1186/s13073-021-00855-5.
186. Suwinski, P., Ong, C., Ling, M.H.T., Poh, Y.M., Khan, A.M., and Ong, H.S. (2019). Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics. *Front. Genet.* 10.
187. McInnes, G., Sharo, A.G., Koleske, M.L., Brown, J.E.H., Norstad, M., Adhikari, A.N., Wang, S., Brenner, S.E., Halpern, J., Koenig, B.A., et al. (2021). Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. *Am. J. Hum. Genet.* 108, 535–548. 10.1016/j.ajhg.2021.03.003.
188. Hall, J.B., Dumitrescu, L., Dilks, H.H., Crawford, D.C., and Bush, W.S. (2014). Accuracy of Administratively-Assigned Ancestry for Diverse Populations in an Electronic Medical Record-Linked Biobank. *PLoS ONE* 9, e99161. 10.1371/journal.pone.0099161.
189. Johnson, R., Ding, Y., Venkateswaran, V., Bhattacharya, A., Boulier, K., Chiu, A., Knyazev, S., Schwarz, T., Freund, M., Zhan, L., et al. (2022). Leveraging genomic diversity for discovery in an electronic health record linked biobank: the UCLA ATLAS Community Health Initiative. *Genome Med.* 14, 104. 10.1186/s13073-022-01106-x.
190. Abul-Husn, N.S., and Kenny, E.E. (2019). Personalized Medicine and the Power of Electronic Health Records. *Cell* 177, 58–69. 10.1016/j.cell.2019.02.039.
191. Odgis, J.A., Gallagher, K.M., Suckiel, S.A., Donohue, K.E., Ramos, M.A., Kelly, N.R., Bertier, G., Blackburn, C., Brown, K., Fielding, L., et al. (2021). The NYCKidSeq project:

study protocol for a randomized controlled trial incorporating genomics into the clinical care of diverse New York City children. *Trials* 22, 56. 10.1186/s13063-020-04953-4.