Title

Low-coverage transcriptomics for understanding genetic regulation of complex traits

Permalink

https://escholarship.org/uc/item/25g069kk

Author

Schwarz, Tommer Abraham

Publication Date

2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Low-coverage transcriptomics
for understanding genetic regulation of complex traits

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy in Bioinformatics

by

Tommer Abraham Schwarz

2022

ABSTRACT OF THE DISSERTATION

Low-coverage transcriptomics
for understanding genetic regulation of complex traits

by

Tommer Abraham Schwarz
Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2022
Professor Bogdan Pasaniuc, Chair

Mapping genetic variants that regulate gene expression (eQTL mapping) in large-scale RNA sequencing (RNA-seq) studies is often employed to understand functional consequences of regulatory variants. However, the high cost of RNA-Seq limits sample size, sequencing depth, and therefore, discovery power in eQTL studies. In this work, we demonstrate that, given a fixed budget, eQTL discovery power can be increased by lowering the sequencing depth per sample and increasing the number of individuals sequenced in the assay. We perform RNA-Seq of whole blood tissue across 1490 individuals at low-coverage (5.9 million reads/sample) and show that the effective power is higher than that of an RNA-Seq study of 570 individuals at moderate-coverage (13.9 million reads/sample). Next, we leverage synthetic datasets derived from real RNA-Seq data (50 million reads/sample) to explore the interplay of coverage and number individuals in eQTL studies, and show that a 10-fold reduction in coverage leads to only a 2.5-fold reduction in statistical power to identify eQTLs. Our work suggests that lowering coverage while increasing the number of individuals in RNA-Seq is an effective approach to increase discovery power in eQTL studies. We then build a pipeline using existing tools CIBERSORTx and bMIND to computationally deconvolute low-coverage bulk RNA-seq from a total of 1,996 individuals to estimate cell type expression. We show that cell type expression estimates are consistent with those from scRNA-seq and can be used as a powerful approach to finding ct-eQTLs. Next, we use medication history from this cohort to look for SNP x lithium interactions in ct-eQTLs, finding 110 examples of eGenes whose cell type expression is significantly associated with some SNP dependent of lithium usage.

The dissertation of Tommer Abraham Schwarz is approved.

Loes Marlein Olde Loohuis

Noah A. Zaitlen

Roel A. Ophoff

Valerie A. Arboleda

Bogdan Pasaniuc, Committee Chair

University of California, Los Angeles

2022

To my parents

TABLE OF CONTENTS

Chapter 1: Introduction

Chapter 2: Tradeoff between coverage and sample size in synthetic data

2.1 Abstract

2.2 Introduction

2.3 Results

2.3.1: Impact of coverage on discovery power

2.3.2 The importance of R^2 for estimating power in association studies

2.3.3 Using synthetic low-coverage RNA-seq to conduct cis-eQTL scans

2.3.4 Estimation accuracy in synthetic as a function of various gene characteristics

2.3.5 Optimal association power for eQTLs is attained at lower coverage with a larger number of samples

2.3.6 Optimal effective sample size under a fixed budget scenario

2.3.7 Impact of manipulating mapping rate on optimal effective sample size

2.3.8 Estimating costs of GTEx RNA-seq under our budget model

2.4 Discussion

2.5 Methods

2.6 Acknowledgments

2.7 Figures

Chapter 3: eQTL mapping using low-coverage RNA-seq in real data

3.1 Abstract

3.2 Introduction

3.3 Results

3.3.1: Whole blood RNA-seq datasets and background

3.3.2 Gene expression estimates using low-coverage RNA-seq are reliable

3.3.3 Low-coverage RNA-seq can be used for powerful eQTL mapping

3.3.4 Low-coverage RNA-seq successfully captures transcript-level expression

3.3.5 Comparison of total reads in low-coverage and high-coverage RNA-seq design

3.3.6 Low-coverage RNA-seq captures similar dynamic range of gene expression as moderate-coverage and high-coverage RNA-seq

3.3.7 Effect sizes of eGenes found using low-coverage and moderate-coverage RNA-seq remain highly concordant after accounting for lowly expressed genes

3.3.8 eQTLs found using low-coverage RNA-seq colocalize with GWAS loci

3.3.9 eQTLs found using low-coverage RNA-seq can be used for TWAS

3.3.10 Low-coverage and moderate-coverage RNA-seq have comparable computationally estimated cell type proportion values

3.4 Discussion

3.5 Methods

3.6 Acknowledgments

3.7 Figures

3.8 Supplementary Tables

3.9 Supplementary Note

Chapter 4: Computational deconvolution of bulk RNA-seq enables cell type biological insights

4.1 Abstract

4.2 Introduction

4.3 Results

4.3.1 Overview of cell type expression estimation pipeline

4.3.2 Cell type expression estimation using computational tools

4.3.3 Computationally estimated cell type proportions match sc-RNA-seq estimates

4.3.4: Cell type expression accounts for varying proportions of variance of BP status

4.3.5: Cell type eQTL analysis reveals more refined biological signal

4.3.6: Comparison of computationally estimated ct-eQTL effect sizes with ct-eQTL effect sizes derived from sc-RNA-seq

4.3.7: Lithium-SNP interaction models at the bulk and cell type expression levels

4.4 Discussion

4.5 Methods

4.6 Acknowledgments

4.7 Figures

4.8 Supplementary Tables

Chapter 5: Concluding remarks

References

LIST OF FIGURES

Figure 2.1: *Synthetic lower-coverage RNA-Seq captures expression signal*

Figure 2.2: *Effective sample size under various budget parameters*

Figure 2.3: V*ariability in correlations in synthetic data*.

Figure 2.4: *Variability of correlations as a function of average expression in a given gene*

Figure 2.5: *Variability of correlation as a function of various gene characteristics*.

Figure 2.6: *Concordance of gene expression as a function of presence of a predominantly expressed transcript*

Figure 3.1: *Concordance of eQTL discovery when using lower-coverage RNA-Seq vs moderate-coverage RNA-Seq*

Figure 3.2: eQTL analysis using *lower-coverage RNA-Seq is comparable to eQTL analysis from the GTEx Consortium*

Figure 3.3: *Distribution of ancestry among samples*

Figure 3.4: *Number of pseudoaligned reads per sample for low-coverage and high-coverage experiments*.

Figure 3.5: *Characteristics of eGenes that do/do not share the same eSNP between lower-coverage and moderate-coverage RNA-Seq.*

Figure 3.6: *Correlation of effect sizes between eQTL analyses low-coverage and GTEX with respect to LD-threshold between eSNPs.*

Figure 3.7: *Number of eGenes per mean expression quintile across datasets.*

Figure 3.8: *eGene concordance by quintile in moderate-coverage RNA-Seq and GTEx*

Figure 3.9: *Real data p-value comparison scatterplot with GTEX*.

Figure 3.10: *Estimation of cell-type proportions.*

Figure 3.11: *Overlap of significant eGenes using RNA-Seq from three different datasets*.

Figure 3.12: *Concordance of transcript expression estimation between lower-coverage RNA-Seq vs moderate-coverage RNA-Seq.*

Figure 4.1: *Visual schematic of deconvolution pipeline*

Figure 4.2: *Computationally estimated cell type expression*

Figure 4.3: *Computationally estimated ct-expression and ct-eQTLs*

Figure 4.4: *SNP-lithium interaction eQTLs*

Figure 4.5: *Computationally estimated cell type proportions compared to laboratory measured cell type proportion values (N=143)*

Figure 4.6: *Mean expression values of LM22 genes across all samples*

Figure 4.7: *Comparing computationally estimated cell type expression with scRNA-seq reference datasets*

Figure 4.8: *PCA of computationally deconvoluted expression*

Figure 4.9: *Number of cell types eGenes are significant in*

LIST OF TABLES

Table 2.1: Sequencing cost scenarios (corresponding to **Figure 2.2**)

Table 3.1: RNA-Seq datasets discussed in this section

Table 3.2: Coloc results for selected blood traits

Table 3.3: TWAS results for selected blood traits

Table S3.1: Number of typed and imputed SNPs after QC

Table S3.2: Whole blood RNA-Seq datasets and respective cost estimates.

Table S3.3: Transcript-level expression correlations by transcript type.

Table S4.1: Comparing computationally inferred ct-expression with scRNA-seq reference expression

Table S4.2: Number of ct-eGenes per cell type

Table S4.3: GREML estimates

Table S4.4: SNP-Lithium interaction results

Acknowledgements

There are so many people that have helped me get to this point and thanking them deserves its own dissertation. But for the sake of brevity, here are just a few of the people that have shaped my time in graduate school and Los Angeles.

First, I'd like to thank my family. To my parents, Katy and Avner, who have been relentlessly supportive for these past 5 years (and the 21 before, as well), thank you. And to my siblings, Maya and Oren, whose growth is my favorite thing to watch, thank you.

I owe a tremendous thank you to my advisor, Bogdan Pasaniuc. Bogdan, you have been a great mentor over these last five years and I feel so fortunate to have completed my PhD under your mentorship. You have cultivated an environment of learning and excitement about science, which has made all the difference in my graduate school experience. Your enthusiasm for research and openness is appreciated and has helped me become a stronger, more curious, more understanding scientist. Some PhD students may struggle to find time to sit down with their advisor but I'm proud to say I had lunch with mine most days over the last five years. Thank you for taking me on board way back when. It's been a pleasure.

To all of the members of Bogdan's lab that I have crossed paths with in my time at UCLA, you have really made my time at UCLA special. The lab has taken several different forms dependent on the state of the pandemic, but everyone has remained consistently supportive, insightful, friendly, and willing to help in any way they can. I am so glad to have met you all and you've all played a big role in my graduate school journey.

My committee has been instrumental along the way in providing feedback, new research directions, and encouragement. Thank you Val, Loes, Noah, and Roel.

There are many others within the UCLA community that have played significant roles in my graduate school experience. To Toni Boltz, much of this work doesn't get done without your efforts. It's going to take some adjusting to live in a world where we don't slack each other with research updates every day. To Benji Seitz and Amanda Johnson, I feel so fortunate to have met you both so early on in my time in LA. You have each been the PhD students that I have aspired to be.

I have been fortunate enough to find several communities that have made Los Angeles home during my time in graduate school. To everyone that I've lived with, played basketball with, ran with, and spent time with over these last five years – you have all greatly enriched my overall experience since I moved to LA and I am beyond thankful for that.

Chapter 2 consists of material covered in the open access article: **Schwarz T.**, Boltz T., Hou K., Bot M., Duan C., Olde Loohuis L.M., Boks M.P., Kahn R.S., Ophoff R.A. & Pasaniuc B. (2022) Powerful eQTL mapping through low coverage RNA Sequencing. *HGG Advances https://doi.org/10.1016/j.xhgg.2022.100103 .* T.S., B.P., and R.O. initialized the study. B.P. and R.O. directed and supervised the project. R.O., R.K., and M.P.B. collected samples. M.B. prepared samples for sequencing. Bioinformatics analysis was conducted by T.S., T.B., K.H., C.D., and L.O.L. . The first draft of the manuscript was drafted by T.S. and all authors contributed to editing, revisions, and approval.

Chapter 3 consists of more material covered in the same open access article: **Schwarz T.**, Boltz T., Hou K., Bot M., Duan C., Olde Loohuis L.M., Boks M.P., Kahn R.S., Ophoff R.A. & Pasaniuc B. (2022) Powerful eQTL mapping through low coverage RNA Sequencing. *HGG Advances https://doi.org/10.1016/j.xhgg.2022.100103 .*

Chapter 4 consists of material from the article currently *in prep*: **Schwarz T.*** & Boltz T.* *et al* (2022) Cell type decomposition of whole blood bulk RNA-Seq reveals cell type eQTLs . * denotes equal contributions. T.S., T.B., R.O., and B.P. designed the study and wrote the paper. Analysis was performed by T.S., T.B., K.H., and C.D. .

VITA

2015	Intern, Bioinformatics Gilead Sciences, Foster City
2016	Intern, Biostatistics/Bioinformatics Gilead Sciences, Foster City
2013-2017	B.S. Genetics & Genomics University of California, Davis
2020	Teaching Assistant, CM 121: Algorithms in Bioinformatics, University of California, Los Angeles
2017-2022	Graduate Student Researcher, Bioinformatics Interdepartmental Program University of California, Los Angeles

Awards

2020-2021	UCLA Training Program in Neurobehavioral Genetics (NIH-NIMH T32)
2018-2019	Modeling and Understanding Human Behavior (NIH NRT-MENTOR) Training fellowship

Publications

Schwarz T., Boltz T., Hou K., Bot M., Duan C., Olde Loohuis L.M., Boks M.P., Kahn R.S., Ophoff R.A. & Pasaniuc B. (2022) Powerful eQTL mapping through low coverage RNA Sequencing. *HGG Advances https://doi.org/10.1016/j.xhgg.2022.100103*

Chang T., Ding Y., Freund M.K., Johnson R.D., **Schwarz T.,** … (2021) Pre-existing conditions in Hispanics/Latinxs that are COVID-19 risk factors. *iScience* https://doi.org/10.1016/j.isci.2021.102188

Majumdar A., Giambartolomei C., Cai N., Haldar T., **Schwarz T.,** Gandal M., Flint J. & Pasaniuc B. (2021). Leveraging eQTLs to identify individual-level tissue of interest for a complex trait. *PLoS Computational Biology* https://doi.org/10.1371/journal.pcbi.1008915

Giambartolomei C., Seo J., **Schwarz T.**, Freund M.K., Johnson R.D., Spisak S., Baca S.C., Gusev A., Mancuso N., Pasaniuc B. & Freedman M.L. (2021) H3k27ac-HiChIP in prostate cell lines identifies risk genes for prostate cancer susceptibility. *American Journal of Human Genetics (In Press)* https://doi.org/10.1101/2020.10.23.352351

Johnson R.D., Ding Y., Venkateswaran V., Bhattacharya A., Chiu A., **Schwarz T.**, … , Zaitlen N., Arboleda V.A., Halperin E., Sankararaman S., Butte M.J., Lajonchere C., Geschwind D.H., Pasaniuc B., UCLA Precision Health Data Discovery Repository Working Group, UCLA Precision Health ATLAS Working Group. (2021) Leveraging genomic diversity for discovery in an EHR-linked biobankL the UCLA ATLAS Community Health Initiative. *medrxiv* https://doi.org/10.1101/2021.09.22.21263987

Baca S.C., Singler C., Zacharia, Seo J., Morova T, Hach F, Ding Y, **Schwarz T.**, Huang C.F., Kalita C., ... , Lack N.A., Pasaniuc B., Takeda D.Y., Gusev A. & Freedman M.L. (2021) Genetic determinants of chromatin reveal prostate cancer risk mediated by context-dependent gene regulation. *biorxiv* https://doi.org/10.1101/2021.05.10.443466

Mandric I., **Schwarz T.**, Majumdar A., Hou K., Briscoe L., Perez R., Subramaniam M., Hafemeister C., Satija R., Ye C.J., Pasaniuc B. & Halperin E. (2020) Optimized design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. *Nature Communications* https://doi.org/10.1038/s41467-020-19365-w

Chapter 1: Introduction

The human genome consists of approximately three billion pairs of nucleotides and a copy of is carried in each of the body's approximately three trillion cells, in the form of genetic material called DNA [60]. Of the three billion pairs, over 99% of them are identical across the human population [61]. However, at some positions, there exists variations, where some of the population may have one nucleotide, while others have another nucleotide. These positions are referred to as single nucleotide polymorphisms (SNPs) or genetic variants. Within the three billion pairs there are roughly 20,000 or so genes, which are regions of the genome that provide instructions for the bodies machinery to create RNA and proteins, through processes called transcription and translation, respectively [62]. RNA and proteins are the functional units of cells, and together affect different traits by different degrees.

An important note about the genome, as stated before, is that across each cell in the body, the genetic material is the same. This means that across many different tissue types (brain, blood, skin, etc.), cells carry the same genome. In order for each tissue to contain highly specialized cells, the quantities of RNA and protein produced from each gene must differ. For example, if a certain gene encodes instructions for production of a protein that performs critical brain functions, this gene might show high expression in brain cells, but lower expression in other tissue types.

When we think about ways to model risk for genetic traits, there are two components that we consider that lead to this risk [63-65]. The first being genetic factors that put people at risk for different traits and the second being environmental factors. This refers to where you live, local air quality, diet, other lifestyle characteristics, and everything that is not attributable to genetics. The amount that each of these categories affects risk for complex traits varies and we think about them in two major categories. First, rare or mendelian traits, that are characterized by having large genetic components, carrying genetic mutations inside gene regions that have large effects on the trait. Some examples of mendelian traits include cystic fibrosis, sickle cell anemia, and beta thalassemia. On the other hand, we have common or complex traits, characterized by substantial environmental components, varying levels of genetic contributions that are derived from many genetic variants with small effects, that lie outside coding regions. Some examples of complex traits include bipolar disorder, diabetes, and height. Because of the large genetic component and large effects of the variants implicated in mendelian traits, they are easier to identify and link to function [67-68]. However, in complex traits, many of the implicated variants lie outside of gene regions and have small effects on the overall outcome, making it difficult to draw conclusions about these variants [66].

Studying the genetic basis of complex traits is important for many reasons, including identifying therapeutic targets, early screening for diseases, and understanding how complex trait biology may differ across genetic ancestry groups. Roughly 20 years ago, the first genome-wide association studies (GWAS) were conducted in order to identify regions of the genome that are correlated with a trait [69]. This is done by gathering a large group of individuals and looking at all of the SNPs in the genome for where the presence or absence of one allele is associated with some trait. We can calculate the strength of association between each SNP and the disease by looking at the frequencies of each of them in a large group of cases and controls. We can measure the statistical association of all of the SNPs in the genome with the trait to find implicated regions. Using this approach, GWAS has been done for many traits that have implicated many different regions of the genome. In 2011, about eight years after scientists started doing these studies, there were 249 GWAS studies done that implicated 1,617 regions of the genome [71]. In the years that followed, measuring genotypes became much less expensive and more resources were allocated to these types of studies. By 2018, there were 5,687 studies that helped discover 71,673 region-trait associations [71].

GWAS has been tremendously successful in discovering genetic variants that are correlated with complex traits. However, the vast majority of risk loci identified in GWAS are difficult to interpret as they lie in noncoding regions of the genome [70]. It has been shown to lie in regulatory regions which leads to the hypothesis that these variants have some collective link to gene expression. Variants that regulate gene expression abundance, as measured through expression quantitative trait locus (eQTL) studies, provide insightful information about the functional interpretation of GWAS signals[1,2]. eQTL studies (or mapping) involve gathering a population of samples and measuring their genetics. Focusing on a SNP, we separate the population into three groups depending on their genotype at that position. For some gene nearby that SNP, we look at its measured expression level in each of the individuals, stratified by the three genotype groups. From this we can calculate the effect size, representing the magnitude of the effect that variant X has on the expression of gene A. We refer to the pair as an eQTL and the regulated gene as an eGene. By integrating eQTL associations with GWAS, we can hope to identify target genes that are driving the GWAS signal at a locus [3-6]. In order to use this approach, we must be able to measure gene expression in large groups of individuals.

RNA sequencing (RNA-Seq) is the state-of-the-art assay for measuring gene expression in bulk tissue and is therefore the assay of choice for eQTL mapping [7-8]. RNA-seq uses reads to sample the RNA being produced during transcription. In fact, the number of reads that goes into an RNA-seq experiment is a key factor that determines the quality of the experiment. Experiments with many reads will capture the underlying level of gene expression well, and we refer to them as having “high-coverage”. On the flip side, experiments with fewer reads will have noisier estimates of the underlying gene expression, and we refer to them as having “low-coverage”. Ideally, we would conduct all RNA-seq experiments with very high coverage, however, the high cost of RNA-Seq often limits the sample size and therefore reduces the discovery power of eQTL studies based on RNA-Seq [2,6,9]. Recent work from the eQTLGen consortium where they conducted a meta-cis-eQTL-analysis from 31,684 gene expression (combination of microarray and RNA-Seq) identified 16,987 eGenes. Consequent power analysis revealed that at a power of 0.80, 1,685 samples are needed to capture eGenes at an effect size of 0.124 (the median effect size observed among the 16,987 eGenes identified in the study) [10].

Traditional RNA-Seq study design prioritizes sequencing depth per individual (targeted levels of coverage in the range of 30-50 million reads) over the number of individuals (samples) included in the study [11-14, 72]. However, given that high levels of coverage per individual limits the sample size of a study, this results in a loss of statistical power in eQTL mapping. Previous studies have established that the low-coverage whole genome sequencing of a larger number of individuals attains increased power of association compared to higher-coverage studies of smaller sample sizes in GWAS [15-19]. This raises the hypothesis that, similarly as for whole genome sequencing and GWAS, lower coverage RNA-seq with a considerable increase in the number of individuals sequenced could increase power of discovery in eQTL studies.[20-24] Currently, there is no systematic approach for determining the optimal sample size (in terms of number of sequenced individuals) and coverage to maximize eQTL discovery power.

One application of eQTL discovery is integration with GWAS, using methods such as coloc [25], to better understand biological mechanisms driving these GWAS loci. Recent work from GTEx shows that just ~20% of GWAS loci colocalize with eQTLs in the most relevant tissue to the trait, and other work shows that an average of just ~11% of trait narrow-sense heritability is explained by cis-eQTLs measured in GTEx.[26-28] To better characterize GWAS loci, it is clear that large sample sizes are especially necessary for maximizing power in eQTL studies.[10] Looking back over the past decade since the inception of RNA-Seq, the size of RNA-Seq datasets has been steadily increasing as a result of decreasing sequencing costs and an emphasis on exploring the biological mechanisms behind GWAS hits.[29] Moving forward, as this trend continues, RNA-Seq experiment design is a critical part of maximizing data resources.[30]

In this work, we perform RNA-Seq in 1490 individuals at a lower coverage (average mapped read depth of 5.9 million reads/sample) and find that eQTL discovery power is better than that of an experiment with a similar budget, but with fewer individuals and higher coverage. Compared to moderate-coverage RNA-Seq[31] and GTEx, we find a high degree of consistency in both the gene expression as well as eQTL effects. We assess the interplay of coverage per sample and accuracy of expression estimates using synthetic RNA-Seq datasets generated by the down-sampling of real high-coverage data. Additionally, we generate synthetic data derived from an RNA-Seq experiment done at 50 million reads/sample to precisely show how decreasing coverage affects accuracy of gene quantification overall, and in different gene categories (by expression, numbers of transcripts, gene length, etc.). Our analyses show that a sequencing experiment conducted with a target coverage of 10 million reads/sample has an average correlation per-gene of 0.40, when compared to an experiment conducted with a target coverage of 50 million reads/sample. We provide evidence to show that under a fixed budget, sequencing at lower coverage levels (< 10 million reads/sample) and increased sample size can boost the effective sample size per unit of cost compared to standard approaches of eQTL study design.

Chapter 2: Trade off between coverage and sample size in synthetic data

2.1 Abstract

Expression quantitative trait studies (eQTL) mapping has proven to be a powerful approach to identify common genetic variants contributing to regulation of gene expression, and subsequently to complex traits and diseases. Here, we show via simulations, that under a fixed budget, low-coverage RNA-seq across an increased number of individuals attains more power for eQTL discovery than high-coverage RNA-seq with a limited number of individuals. This is quantified by effective sample size, or an estimate for the amount of individuals that would have needed to be sequenced at high-coverage to discover the same number of eQTLs. We find that with RNA-Seq data at 5-fold reduction in coverage, it is possible to capture upwards of 60% of the eGenes found in high-coverage data. Within the context of reducing experimental costs, our results suggest that low-coverage RNA-sequencing in many individuals can yield increased benefit for eQTL studies.

2.2 Introduction

Massive cost reductions in whole genome sequencing and genotyping has enabled researchers to collect genotypes at very large scales, leading to very high powered GWAS studies. While GWAS links genetic variants to risk for complex traits, eQTL mapping has emerged as an approach for linking genetic variants to gene expression, which can help provide some functional information about genetic variants. In the last decade, RNA-seq has become the method of choice for measuring gene expression, and while costs have decreased, it is still prohibitively expensive when compared to the costs of whole genome sequencing and genotyping. When designing an eQTL study with a fixed amount of resources, there exists a tradeoff between the number of samples included in the study and the level of coverage that the sequencing is performed at. These parameters directly influence how well the study will be powered to detect associations. Increasing sample size will reduce coverage, leading to more noisy gene expression estimates, while decreasing sample size will lead to more accurate gene expression estimates. Therefore, when designing eQTL studies with a fixed amount of resources, it is critically important to consider these parameters and their combined impact on association power.

In the past, eQTL studies have been conducted using RNA-seq experiments at very high levels of coverage (40-80 million reads/sample), and experimental guidelines suggest similarly high levels of coverage for the purpose of gaining a global view of gene expression. We hypothesize that it is still possible to capture much of the gene expression signal sequencing at lower levels of coverage, which would enable researchers to boost the number of samples included in their studies. In this work, we use actual high-coverage RNA-seq data to create many synthetic datasets at varying levels of coverage, and analyze how well we quantify gene expression and detect eQTLs at lower coverage levels.

2.3 Results

2.3.1: Impact of coverage on discovery power

We focus on exploring the interplay of number of individuals and coverage for optimizing power for discovery in this eQTL study. As simulating RNA-Seq data is challenging [34-35], we down sample reads from high-coverage RNA-Seq data to create synthetic datasets at various coverages (Methods). We observe that with just a fraction of the reads, it is still possible to estimate gene expression (**Figure 2.1A**). For example, we demonstrate using synthetic data that using just 10% of the data (5.0 million reads/sample) retains a per gene R^2 of 0.40, on average. In practice, increasing the number of samples in an RNA-Seq study leads to increased library preparation costs, making the increase in obtainable statistical association power less obvious.

	Cost per lane	Cost per sample
Scenario 1	$1790	$87
Scenario 2	$1790	$30
Scenario 3	$1790	$150
Scenario 4	$1000	$150

Table 2.1: *Sequencing cost scenarios (corresponding to **Figure 2.2**)*

The cost parameters corresponding to the effective sample size scenarios in **Figure 2.2**. Cost per sample reflects the cost of library prep to include an additional sample. Cost per lane reflects the cost per sequencing lane, which allows for 300 million reads.

2.3.2 The importance of R^2 for estimating power in association studies

It has been established that statistical power in association studies is a function of sample size, phenotype measurement accuracy, and genotype measurement accuracy [15,16,21,35]. This means that the power of a study with sample size N and estimated gene expression is approximately the same as the power of a study with sample size N, using the true gene expression measurements (Methods). In this scenario, R^2 is the correlation between the true expression and the expression estimates. We therefore report the squared correlation (R^2) between synthetic datasets at various coverages and the full data at an average of 50 million reads/sample (which is assumed to be the true gene expression). While these results show the mean R^2 for all genes obtained under one synthetic dataset (one draw) per coverage level, we find that the synthetic datasets are consistent across multiple draws at the same coverage level (**Figure 2.3A**) and each show similar correlations with the ground truth gene expression (**Figure 2.3B**).

2.3.3 Using synthetic low-coverage RNA-seq to conduct cis-eQTL scans

Next, we quantified how well lower-coverage RNA-Seq can be used to detect eGenes. We explore the number of genes with significant associations after FDR correction at 5% under various levels of simulated coverage (**Figure 2.1B**). Using synthetic data, as the number of reads per sample decreases, we find that many eGenes are still detectable. For example, at 10 million reads per sample, just 20% of the full coverage, 60% of the eGenes are still detected. In the context of eQTL studies, synthetic RNA-Seq supports the idea that sequencing at lower coverages over a higher number of individuals is a promising approach to boosting statistical power.

2.3.4 Estimation accuracy in synthetic as a function of various gene characteristics

Finally, we explore the estimation accuracy in the synthetic data as a function of relative gene expression abundance, since less abundant genes may not be captured altogether at lower sequencing coverages. We stratify genes into five groups based on their relative expression in the full dataset (M=50.3 million reads/sample) and report the R^2 for genes in each of these groups in synthetic data (**Figure 2.1C**). We observe that in the synthetic RNA-Seq dataset at 10 million reads/sample, we capture expression of highly expressed genes better than lower expressed genes. Specifically, for genes in the lowest through the highest quintiles of relative gene abundance, we find the average correlation (R^2) to the ground truth of expression to be 0.36, 0.44, 0.61, 0.73, 0.86, respectively. We observe the same effect for synthetic datasets at coverages of 1 million reads/sample and 25 million reads/sample (**Figure 2.4A** and **Figure 2.4B**). These results suggest that the ability to achieve similar power in eQTL analysis studies will differ per gene, and is a function of relative expression. We further investigate the properties of genes with quantification accuracy influenced by coverage levels of sequencing and find that that protein coding genes are more accurately quantified at lower coverage levels compared to non-protein coding genes (**Figure 2.5A**). Conversely, the number of transcripts per gene, gene length, and GC content do not appear to be factors that broadly influence the gene quantification accuracy when sequencing coverage is reduced (**Figure 2.5B**, **Figure 2.5C**, and **Figure 2.5D**). We also investigate in real data whether genes with a predominantly expressed transcript are better estimated in lower-coverage data compared to those genes that do not have a predominantly expressed transcript (**Figure 2.6**). We do not find that this is a factor that strongly impacts gene quantification accuracy in real data.

2.3.5 Optimal association power for eQTLs is attained at lower coverage with a larger number of samples

In the context of reducing experimental costs, we explored the trade-off between the number of samples sequenced and the average coverage per sample. To further evaluate the ability of lower-coverage sequencing to recapitulate expression signal observed in high-coverage data, we evaluated the expected effective sample size obtained with lower coverages per sample compared to a conventional approach of 50 million reads/sample. We down-sample reads (as described in Section 1 and Methods) from a high-coverage RNA-Seq experiment derived from Fibroblast tissue in order to create lower-coverage RNA-Seq synthetic data. This is done to match actual low coverage sequencing as closely as possible. To evaluate the relationship between cost, coverage, and sample size, we use the following equation to model the budget:

$B = N*e + N*g + N*a + \frac{N*b*c}{d} + f$ (Methods).

2.3.6 Optimal effective sample size under a fixed budget scenario

We compute the effective sample size of an eQTL study as a function of average coverage, which determines the number of samples sequenced under a fixed budget (Figure 2.2A). As an example, at a fixed budget of $300,000, the highest effective sample size is achieved by sequencing 1378 individuals using 13 million reads per sample, which leads to a corresponding effective sample size of 877. An experiment achieving the sample effective sample size, using 50 million reads per sample, would cost $384,418 (N = 877, R^2 = 1.0). Therefore, by lowering the coverage of each sample and increasing sample size, we achieve the same effective sample size at just 78.0% of the cost. In practice, it is common to observe a considerable discrepancy between the target number of reads in an experiment and the number of reads that successfully map to genes. This can be attributed to different library prep techniques, quality of samples, or tissue type. To show how mapping rate can influence the effective sample size of an experiment, we model effective sample size with varying levels of mapping rates (Methods). As expected, we observe that as the mapping rate increases, there is a corresponding increase in effective sample size (Figure 2.2C).

2.3.7 Impact of manipulating mapping rate on optimal effective sample size

With a budget of ~$300k and an expected mapping rate of 0.60 (chosen based on mapping rate of similar experiments using TruSeq Stranded plus rRNA and GlobinZero in whole blood tissue), we see the maximum effective sample size would be achieved at a target coverage of 16 million reads per sample, including 1274 individuals in the study. We estimate that achieving the same effective sample size using data with 50 million reads per sample would cost ~$320k (N = 723), or 1.06x the cost of sequencing 1274 individuals at a coverage of 16 million reads/sample. To explore other cost scenarios, we created a webtool where one can enter budget, costs, and other details about the experiment, in order to see how to achieve optimal effective sample size (https://tomschwarz.shinyapps.io/RNASeqCoverageCalculator/).

2.3.8 Estimating costs of GTEx RNA-seq under our budget model

We use this budget model to calculate the cost of the eQTL analysis performed by GTEx under standard cost assumptions (Methods). We find that the cost of this experiment (n = 668, 82 million reads/sample on average) would have been ~$620,000. The cost of the lower-coverage RNA-Seq (n = 1490, 5.9 million reads/sample, on average) under these assumptions is ~$293,000, just 47% of the cost of the GTEx experiment. The GTEx eQTL analysis reports 10544 eGenes with a significant association, while using the lower-coverage RNA-Seq leads to 7587 eGenes with a significant association, 72% of what GTEx reports. If we assume that genotypes have already been measured in the cohort (such that $g = 0$), the cost of the lower-coverage RNA-Seq experiment comes out to $215,000 , while the GTEX experiment comes out to ~$585,000. This means that using just ~36% of the cost, lower-coverage RNA-Seq has the power to detect ~72% of the eGenes with a significant association.

2.4 Discussion

Our study is, in part, motivated by previous findings of whole genome sequencing (WGS) studies benefiting from reduced coverage and increased sample sizes [15-16]. We note that though our application is similar, there remains some key differences. Primarily, there exists a high variance in the degree to which transcripts are expressed, which is not easily predictable [16]. While we generally refer to experiment-wide coverage of an experiment, coverage differs across transcripts due to factors such as gene length and number of transcripts per gene. Consequently, the nature of RNA-Seq data is such that lowering coverage of sequencing does not necessarily have a uniform effect on read sampling, which introduces an additional source of noise. It is important to explore the effects of reducing coverage in RNA-Seq as the necessary level of coverage in WGS studies are generally dictated by the structural variant (SNP, indel, CNV) of interest, with a fairly predictable change in detection with reduced coverage. On the other hand, the necessary level of coverage in RNA-Seq is related to its ability to detect lesser abundant transcripts, where the relationship between decreasing coverage and ability to quantify these transcripts is not understood as well.

2.5 Methods

Cohort Description

The samples included are from a study with individuals ascertained for bipolar disorder (BP). The cohort consists of 916 individuals with BP, 358 controls, and 216 relatives of the individuals with BP.

Connection between effect size and R^2

If *g* is the genotype at the SNP that we are testing for associations, and β is the effect size of that SNP when regressing on the true gene expression, y, and $\hat{\beta}$ is the effect size of that SNP when regressing on the estimated gene expression, $\tilde{y}$. The relationship between *y* and $\hat{y}$ is as follows that $R^2 = corr(y, \hat{y})$. It follows that the estimates of effect size for a SNP on the true gene expression, $\hat{\beta}$, are related to the estimate of effect size for a SNP on the estimated gene expression, $\hat{\tilde{\beta}}$ as $\hat{\tilde{\beta}} = cov(g,\ \tilde{y}) = cov\left(g,\ Ry + \varepsilon\right) = cov\left(g, Ry\right) + cov\left(g, \varepsilon\right) = R\hat{\beta}$ where ε is a random variable with mean 0 and variance 1. The association test statistics at low-coverage is $x_{ground} = Ncor^2(g, y)$ thus implying that the association statistic at low coverage is $x_{low-coverage} = Ncor^2\left(g, \tilde{y}\right) = N\hat{\tilde{\beta}}^2 = N(R\hat{\beta})^2 = R^2 {*} Ncor^2\left(g, y\right) = R^2 x_{ground}$, where N is the number of samples included in the association study.

Budget model

We modeled the cost of a large-scale bulk RNA-Seq experiment based on parameters from two different library prep techniques: (1) TruSeq Stranded plus rRNA and GlobinZero and (2) TruSeq Stranded polyA selected, both from the UCLA Neuroscience Genomics core. Cost, or *B*, is a function of the following: a, the library preparation cost per sample, *b*, which is the target coverage of each sample (in millions of reads). *c*, the cost per lane (which contains *d* million reads), *d* is the number of reads per sequencing lane (in millions), *g* is the cost of genotyping per sample, e is the cost of DNA and RNA extraction per sample, N is the number of samples in the association study, and f, any additional upfront or computational costs associated with analysis. Altogether, we model the budget as follows; $B = N^*e + N^*g + N^*a + \frac{N^*b^*c}{d} + f$

.

Synthetic low coverage RNA-Seq

We use high-coverage RNA-Seq (average of 50 million reads/sample, TruSeq Stranded polyA selected) from a set of 150 cell lines derived from human fibroblast cells. We assume this to be the ground truth of gene expression. We used seqtk (https://github.com/lh3/seqtk) to randomly down-sample reads at various coverages, uniformly. We performed five iterations of down-sampling at each level of coverage in order to account for potential variability in the sampling and sequencing errors.

RNA-Seq processing pipeline

We used FASTQC to visually inspect the read quality from the lower-coverage whole blood RNA-Seq (5.9M reads/sample), the moderate-coverage whole blood RNA-Seq (13.9M reads/ sample), and the high-coverage fibroblast RNA-Seq (50M reads/sample). We then used kallisto to pseudoalign reads to the GRCh37 gencode transcriptome (v33) and quantify estimates for transcript expression. We aggregated transcript counts to obtain gene level read counts using scripts from the GTEx consortium (https://github.com/broadinstitute/gtex-pipeline) [13].

cis-eQTL mapping

Excluding related individuals (pi_hat > 0.2) from the analysis, we perform cis-eQTL analysis mapping using FastQTL [37], using a defined window of 1 Mb both up and downstream of every gene's TSS, for sufficiently expressed genes (TPM > 0.1 in 20% of individuals). We run the eQTL analysis in permutation pass mode (1000 permutations, and perform multiple testing corrections using the q value FDR procedure, correcting at 5% unless otherwise specified. We then restrict our associations to the top (or leading) SNP per eGene.

R^2 adjustment

To account for the variability in mapping rate across different library prep techniques and different tissue types [46-47], we look at the mean R^2 at the expected coverage, which is calculated as *expected coverage = target coverage * estimated mapping rate.* Using mean R^2 values from comparing lower-coverage synthetic RNA-Seq to moderate-coverage RNA-Seq real data, we fit a log curve to estimate the adjusted mean R^2 (R^2_{adj}) at the expected coverage.

Effective Sample Size

Under a fixed-budget setting, we calculate effective sample size (N_{eff}) for a given coverage using the adjusted mean R^2 (R^2_{adj}) and the number of samples included at a given coverage level (N) $N_{eff} = R^2_{adj}*N$

2.6 Acknowledgments

We thank the study subjects for their willingness to provide specimens and clinical data. We thank Yi Ding, Kathryn Burch, Ruthie Johnson, Arjun Bhattacharya, and Malika Freund for meaningful discussion in helping make this work possible. T.S. was supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under award no. T32NS048004. T.B. was supported by the NIH (grant no. 5T32HG002536-19). This research was supported by the National Institute of Mental Health of the National Institutes of Health under award no. 5R01MH115676-04. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

2.7 Figures

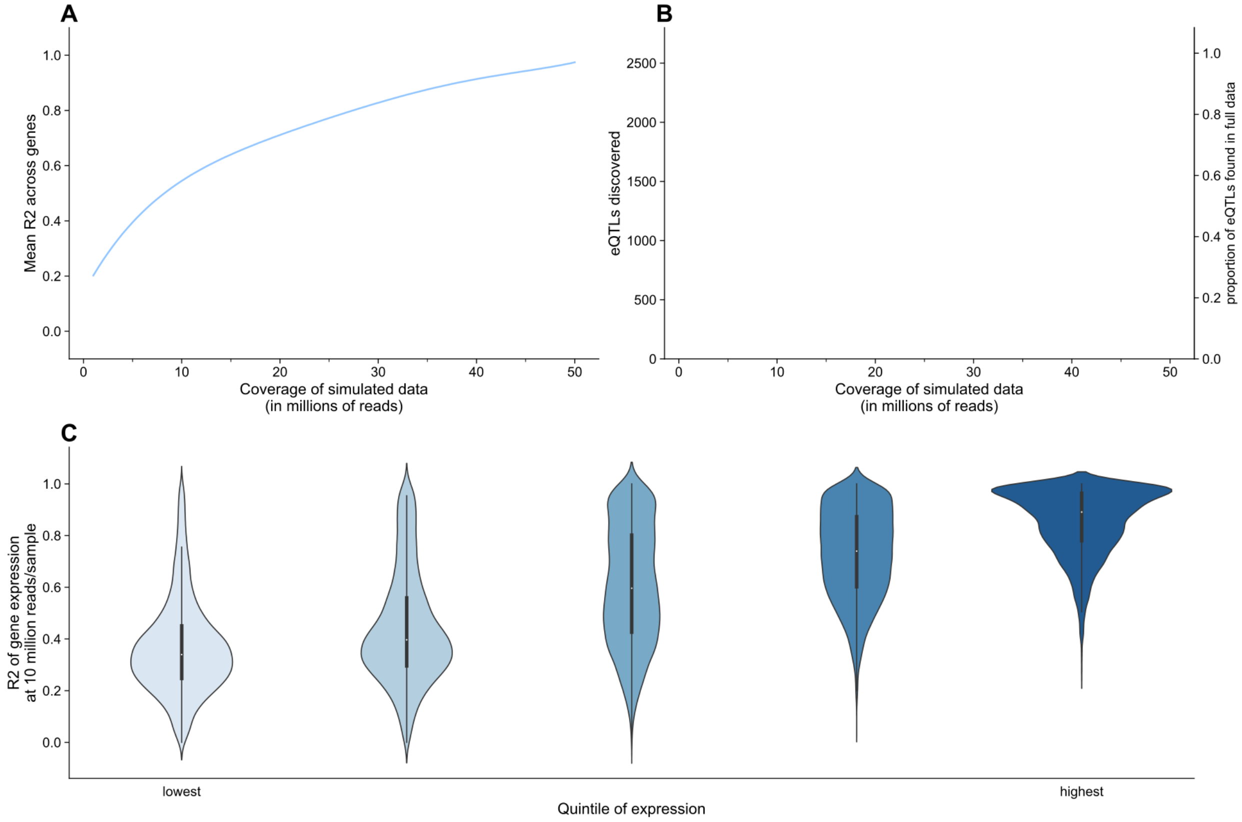


Figure 2.1: *Synthetic lower-coverage RNA-Seq captures expression signal*

(2.1A): On the x-axis, we show the level of simulated coverage, and on the y-axis we show the mean Pearson correlation of every gene. We calculate this value by finding the R^2 values for the TPM values of each of 45,910 genes across 155 samples between the high coverage data (average of 50 million reads per sample) and the simulated data, and reporting the mean R^2 value per gene. **(2.1B):** For a fixed number of individuals, absolute number and percentage of eGenes captured at 5% FDR, for synthetic RNA-Seq at varying levels of coverage. **(2.1C):** Gene expression accuracy as a function of relative gene expression observed in actual RNA-Seq data with 50 million reads/sample. 23,540 genes (with average expression > 0.1 TPM) are divided into five ascending quintiles of expression based on their average expression in 155 samples.

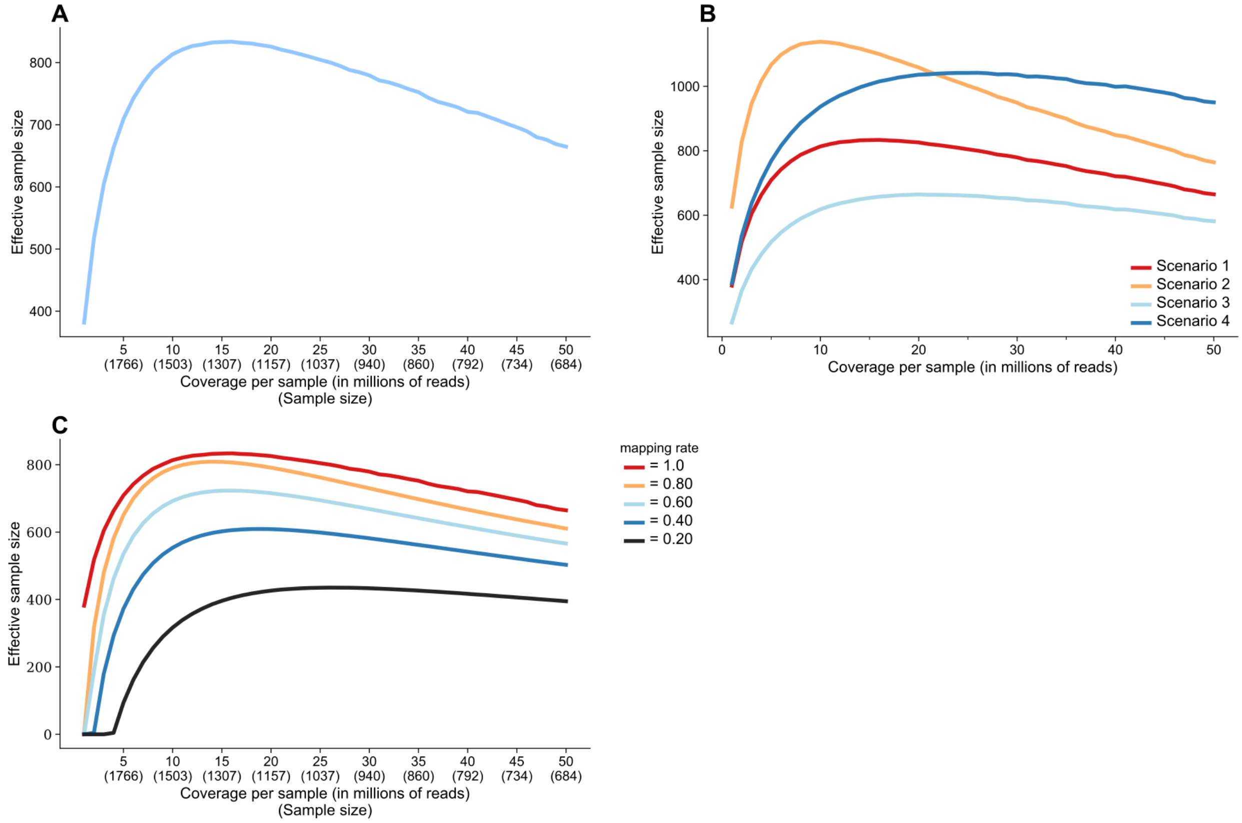


Figure 2.2: *Effective sample size under various budget parameters*

(2.2A): Effective sample size in RNA-Seq under a fixed budget ($300,000) as a function of the number of samples and the resulting coverage. Cost assumptions: $87 per library prep per sample, $1790 per lane of sequencing (300 million reads), $53 per genotyped sample. **(2.2B):** Effective sample size in RNA-Seq under a fixed budget ($300,000) as a function of the number of samples and the resulting coverage. Cost assumptions vary and are reflected in **Table 2.1**. **(2.2C):** Effective sample size under a fixed budget ($300,000) as a function of the number of samples and the results coverage. A global mapping rate parameter is used to simulate actual experimental conditions (**Methods**).

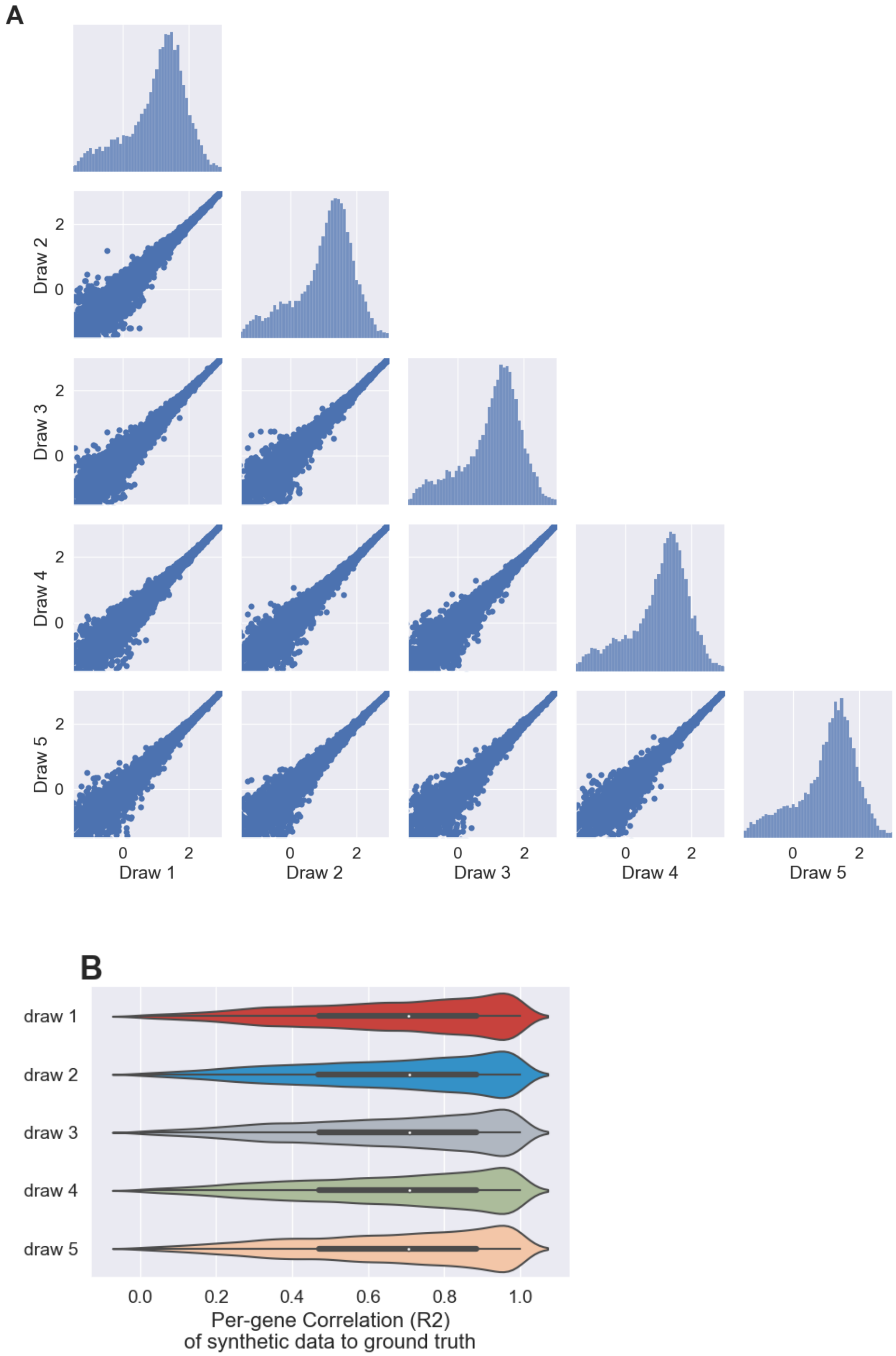

A
Draw 2
Draw 3
Draw 4
Draw 5
0
2
Draw 1
Draw 2
Draw 3
Draw 4
Draw 5
B
draw 1
draw 2
draw 3
draw 4
draw 5
0.0
0.2
0.4
0.6
0.8
1.0
Per-gene Correlation (R2)
of synthetic data to ground truth

Figure 2.3: V*ariability in correlations in synthetic data*. **(2.3A)** For synthetic data corresponding to one sample, a comparison of estimated log TPM values between five different uniform sampling draws at 10 million reads/sample, for 14,948 protein-coding genes. **(2.3B)** For 14,948 protein-coding genes estimated across five different uniform sampling draws at 10 million reads/sample, we compare the distribution of correlation (R^2) between the estimated expression of the samples and the ground truth gene expression.

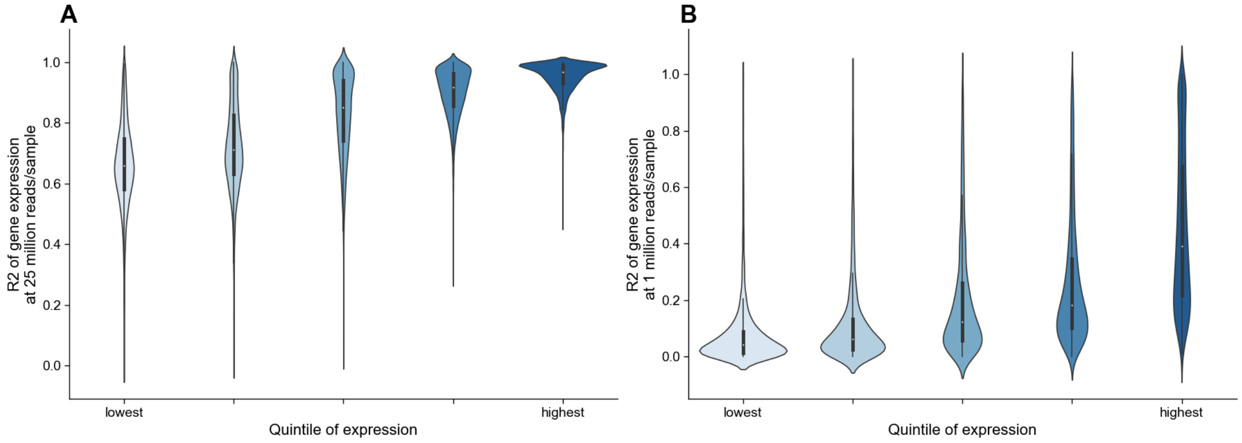


Figure 2.4: *Variability of correlations as a function of average expression in a given gene* **(2.4A)** Gene expression accuracy using data simulated with 1 million reads/sample, as a function of relative gene expression observed in actual RNA-Seq data with 50 million reads/sample. 23,043 genes (with average expression < 0.1 TPM) are divided into five ascending quintiles of expression based on their average expression in 155 samples. **(2.4B)** Gene expression accuracy using data simulated with 1 million reads/sample, as a function of relative gene expression observed in actual RNA-Seq data with 50 million reads/sample. 23,043 genes (with average expression < 0.1 TPM) are divided into five ascending quintiles of expression based on their average expression in 155 samples.

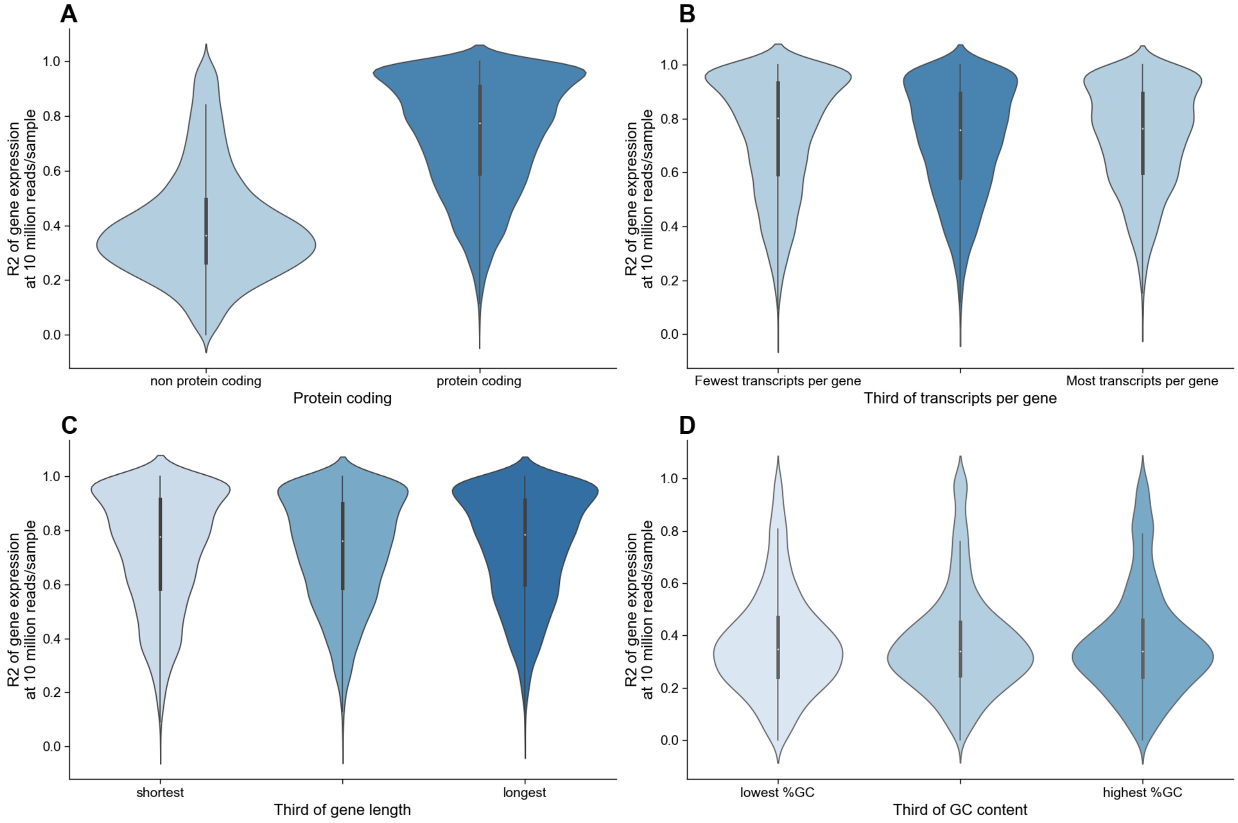


Figure 2.5: *Variability of correlation as a function of various gene characteristics*. **(2.5A)** Gene expression estimation accuracy simulated at 10 million reads/sample as a function of whether a gene codes for a protein. 24,093 genes (with average expression < 0.1 TPM) are divided into two groups. **(2.5B)** Gene expression estimation accuracy simulated at 10 million reads/sample as a function of how many transcripts each gene has. 23,540 genes (with average expression < 0.1 TPM) are divided into three ascending groups based on the number of transcripts contained in each gene. **(2.5C)** Gene expression estimation accuracy simulated at 10 million reads/sample as a function of relative gene length. 14,484 genes (with average expression < 0.1 TPM, protein coding) are divided into three groups based on the length of each gene. **(2.5D)** Gene expression accuracy as a function of relative GC content. 5,771 genes (with average expression < 0.1 TPM and GC content reported) are divided into three groups based on the length of each gene.

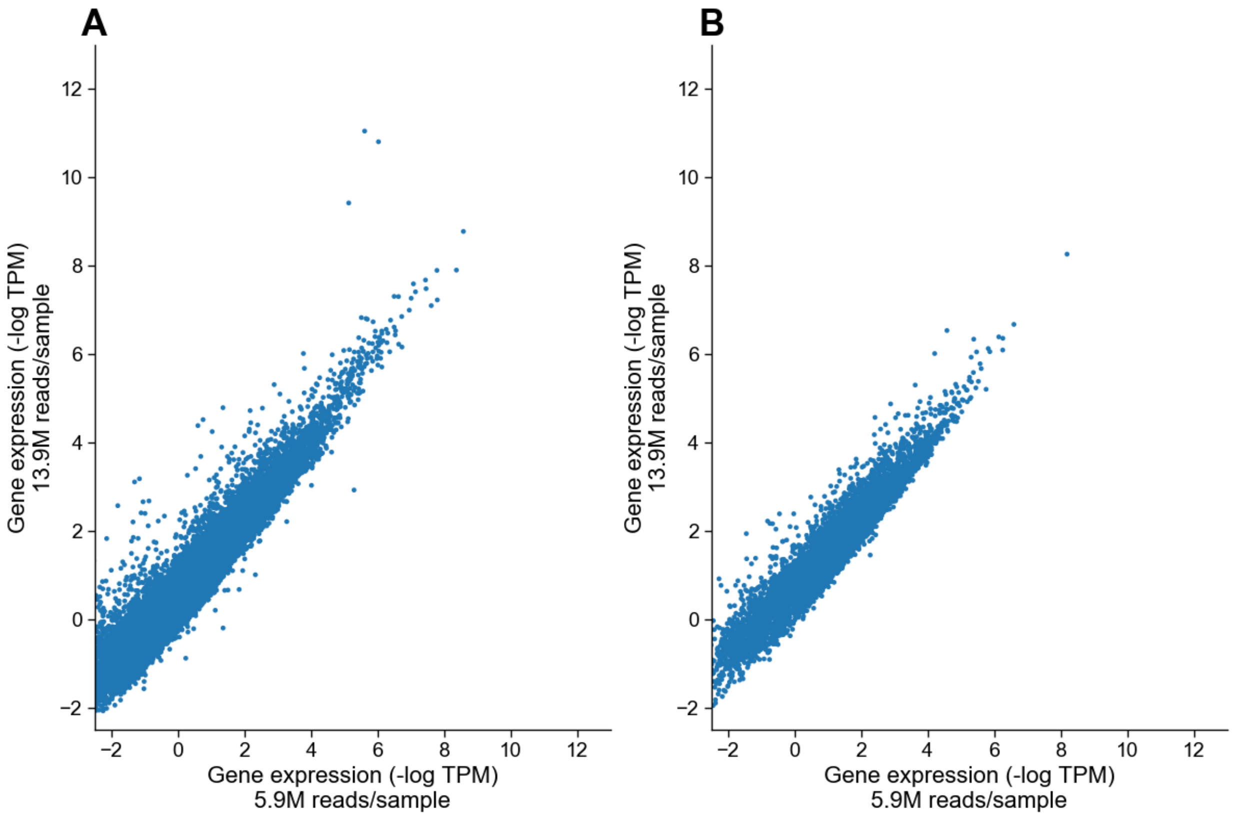


Figure 2.6: *Concordance of gene expression as a function of presence of a predominantly expressed transcript* **(2.6A)** Restricting to the 5894 genes with a transcript responsible for at least 50% of the gene's expression in the 13.9M read/sample dataset, comparison of the mean expression (log TPM) across samples, of every gene. R^2 = 0.91. **(2.6B)** Restricting to the 14711 genes without a transcript responsible for at least 50% of the gene's expression in the 13.9M read/sample dataset, comparison of the mean expression (log TPM) across samples, of every gene. R^2 = 0.90.

Chapter 3: eQTL mapping using low-coverage RNA-seq

3.1: Abstract

Mapping genetic variants that regulate gene expression (eQTL mapping) in large-scale RNA sequencing (RNA-seq) studies is often employed to understand functional consequences of

regulatory variants. However, the high cost of RNA-Seq limits sample size, sequencing depth, and therefore, discovery power in eQTL studies. In this work, we perform RNA-Seq of whole blood tissue across 1490 individuals at low-coverage (5.9 million reads/sample) and show that the effective power is higher than that of an RNA-Seq study of 570 individuals at moderate-coverage (13.9 million reads/sample). We perform rigorous analysis to show that the associations discovered in eQTL analysis using low-coverage RNA-seq are consistent with those from moderate-coverage and high-coverage (83 million reads/sample) RNA-seq. Our work suggests that lowering coverage to 5.9 million reads/sample, in practice, remains effective in accurately quantifying gene expression estimates.

3.2: Introduction

RNA-seq is usually done with higher expression

Low-coverage sequencing may introduce technical biases. Previous studies have shown that reduction in coverage can produce biases that mischaracterize RNA splicing in single cells.

In this work, we conduct low-coverage sequencing of 1490 samples derived from whole blood tissue (5.9M reads/sample). We leverage two existing datasets, one of moderate coverage (13.9M reads/sample) and publicly available data from GTEx (83M reads/sample), to validate the findings from low-coverage sequencing. We first show that expression estimates are consistent with those observed by using moderate-coverage sequencing. We conduct eQTL analyses with all three datasets and show that using low-coverage RNA-seq, we observe high concordance of effect sizes with eQTLs discovered using both moderate-coverage and high-coverage RNA-seq. Using coloc and TWAS, we show that using eQTLs derived from low-coverage RNA-seq, we implicate both new and existing risk variants from GWAS. Finally, we show that computational deconvolution tools to estimate cell type proportion perform very comparably between low-coverage and moderate-coverage RNA-seq.

3.3 Results

3.3.1: Whole blood RNA-seq datasets and background

To validate the utility of low-coverage RNA-sequencing, we sequenced whole blood tissue from N = 1490 unrelated individuals (Methods) (**Figure 3.3A** and **Figure 3.3B**). We target a sequencing coverage of 9.5 million reads per sample, yielding M = 5.9 million reads mapped to RefSeq genes on average (sd across samples of 1.96 million, **Figure 3.4**). We refer to this dataset as the lower-coverage RNA-Seq, or the M=5.9 million reads/sample dataset. We contrast this dataset with an RNA-Seq dataset obtained with a similar budget, but with 2.4-fold higher coverage (M = 13.9 reads) across N = 570 individuals (**Figure 3.3C** and **Figure 3.3D**) [22]. We refer to this as the moderate-coverage whole blood RNA-Seq, or the M = 13.9 million reads/sample dataset (**Table 3.1**).

Referred to as:	Coverage (million reads per sample)	Tissue	Number of samples	Library prep method
Lower-Coverage or M=5.9M reads/sample (Whole Blood)	5.9	Whole blood	1490	TruSeq Stranded plus rRNA and GlobinZero
Moderate-Coverage or M=13.9M reads/sample (Whole Blood) [19]	13.9	Whole blood	570	Meta-analysis of (1) TruSeq Stranded plus rRNA and GlobinZero and (2) TruSeq Stranded polyA selected
High-coverage (Fibroblast)	50.3	Fibroblast	150	TruSeq Stranded polyA selected
GTEX[12]	82	Whole blood	670	TruSeq Non-stranded polyA selected

eQTLGen[13]	N/A	Whole blood	31684	Meta-analysis consisting of RNA-Seq and microarray

Table 1: *RNA-Seq datasets discussed in this section*

The coverage refers to the average number of reads that successfully map to the transcriptome, except for GTEX, which refers to the median number of total reads per sample (average mapped not available). Further description of sample overlaps among cohorts in **Supplementary Note**.

3.3.2 Gene expression estimates using low-coverage RNA-seq are reliable

First, we assess the number of genes quantified in the two datasets. We observe 40459 genes with at least one mapped read on average across samples in the whole blood moderate-coverage dataset, and 27308 genes with at least one mapped read on average across samples in the whole blood lower-coverage dataset. Notably, when restricting to protein coding genes with at least one mapped read in both the moderate-coverage and lower-coverage datasets, we find more similar numbers between the data sets, with 18329 and 15605 genes quantified, respectively. This is likely due to the very sparse abundance of the non-protein coding genes, making them less likely to be detected in a lower coverage dataset. Indeed, we observe similar effects across the high vs low coverage datasets when assessing the genes with sufficient expression to be included in eQTL analysis (TPM > 0.1 in 20% of individuals, see **Methods**): 26566 genes (15496 protein coding genes) in moderate coverage data versus 19039 (13339 protein coding genes) in low coverage data. Most importantly, we observe a high correlation in the abundance levels across the two datasets. We calculate the median TPM across samples of 62487 gencode genes and restrict to the 20735 protein-coding genes that are detected in both datasets. Without recalculating TPM after these restrictions we observe a Pearson correlation (R^2) of 0.91, thus demonstrating that moderate and lower coverage RNA-Seq recover similar expression (**Figure 3.1A**).

3.3.3 Low-coverage RNA-seq can be used for powerful eQTL mapping

Next, we investigate the power of low-coverage RNA-Seq for eQTL mapping. We conducted cis-eQTL mapping with a 1 Mb window using QTLtools,[36] restricting to the 1490 unrelated individuals in the lower-coverage RNA-Seq data (Methods), to identify 7587 genes (eGenes) with a significant association at an FDR adjusted p-value < 0.05. As expected, eQTL distribution is concentrated at transcription start sites (TSS), with 73% of eGenes TSS within 250kb of the associated SNP (eSNP). Repeating this approach using the moderate-coverage whole blood data in 570 individuals, we only find 5971 genes with a significant association at FDR correction level of 5%. 4969 of the 7587 eGenes found using the lower-coverage data are also significant in the moderate-coverage data. Of these, 2163 of the eGenes are protein coding eGenes that share the same associated eSNP, and we see an extremely high level of concordance between effect sizes for these eGenes across the two datasets (R^2 = 0.93, **Figure 3.1B**). This further indicates that low-coverage RNA-Seq is robust in capturing eQTL effect sizes. Briefly, we tested to see whether the mean expression or number of transcripts differed between eGenes that shared the same eSNP between the two datasets (n = 2163) and those that did not (n = 4324) (**Figure 3.5**). We find slightly higher expression and a slight increase in the number of transcripts in the set of eGenes that do share the same eSNP. 1002 genes were found to be eGenes in the moderate-coverage eQTL analysis but not in the lower-coverage analysis, with 573 (of the 1002) not passing expression levels (TPM >0.1 in 20% individuals) to be included in the lower-coverage eQTL analysis; only 234 of the 573 were protein coding genes, suggesting that for most protein-coding genes, lower-coverage RNA-Seq can adequately capture their expression. Similar concordance is observed at the level of p values for the associations in both datasets (**Figure 3.1C**). Comparing the p values for eGenes detected in both eQTL analyses, the corresponding regression line has a slope of 0.39, consistent with the lower-coverage dataset having superior statistical power to detect associations over the moderate-coverage dataset, and consistent with overall number of significant eQTL discoveries. We report the results from using typed SNPs in these eQTL analyses (**Methods**), but observe similar patterns when using the full set of imputed SNPs.

3.3.4 Low-coverage RNA-seq successfully captures transcript-level expression

More recently, RNA-Seq data has been used to quantify gene expression at different resolutions, specifically at the transcript/isoform levels. To investigate whether lower-coverage RNA-Seq can be reliably used in this context, we use kallisto [33] to quantify transcript expression in both the 5.9M and 13.9M read/sample datasets (**Methods**). We quantify 227,046 transcripts between the two datasets and find strong concordance between transcript expression estimates across them (R^2 = 0.83), suggesting that lowering coverage to this degree does not strongly influence the ability to detect changes in transcript expression (**Figure 3.1D**). However, there does seem to be associations between transcript type and how well the transcript is quantified using lower-coverage RNA-Seq (**Figure 3.12** and **Table S3.3**).

3.3.5 Comparison of total reads in low-coverage and high-coverage RNA-seq design

To further validate the performance of eQTL analysis using low coverage RNA-Seq (coverage 5.9M, n = 1490), we compared the resulting eQTLs to the ones found by GTEx in whole blood [13] (**Figure 3.2**). Restricting to the 12247 protein coding genes with sufficient expression to be included in both studies (> 0.1 TPM in 20% of samples) we find that 3916 out of the 5538 protein coding genes (71%) with a significant association using the lower-coverage data also had a significant association in GTEx, correcting at an FDR level of 5%. We note that this is not an entirely equal comparison as the three datasets are generated from different budgets (**Table S3.2**). While GTEX (n = 668, 82M reads/sample) consists of 55.6B reads, the lower-coverage (n = 1490, 5.9M reads/sample) and moderate-coverage (n = 570, 13.9M reads/sample) datasets consist of just 8.8B and 7.9B reads, respectively. Considering the number of eGenes discovered using each of these datasets, we find that per 1 billion total reads, we discover 862 eGenes using the lower-coverage dataset, 756 eGenes using the moderate-coverage dataset, and just 190 eGenes in GTEx (**Figure 3.2A** and **Figure 3.2B**). Among eGenes shared by both datasets, we found that the leading eSNPs are in LD (average R^2= 0.41, sd = 0.39), showing that low-coverage RNA-Seq captures the same eQTL signal, either directly or by a nearby tagged SNP. Further restricting to eGenes with leading eSNPs with a LD R^2 value of at least 0.25 in both of these datasets (1927 genes) (**Figure 3.2C**), we observe a correlation (R^2) of 0.81 between their effect sizes. We find consistently high correlations regardless of the LD threshold used here (**Figure 3.6**). Looking into the 1622 protein coding genes with a significant association in eQTL analysis using the lower-coverage RNA-Seq but not in GTEx using an FDR adjusted p value cutoff of 0.05, we observe that 283 have a significant association in GTEx using an FDR adjusted p value cutoff of 0.10. To further ensure that these eGenes are not false positives, we compare the set of 1622 genes with eQTL analysis conducted by the eQTLGen Consortium [10] and find that 1498 of these genes (92.4%) have been found to have a significant association in eQTLGen. This suggests that the additional associations found using lower-coverage data that are not found in GTEx are not false positives, but fall just below the significance threshold in the GTEx analysis.

3.3.6 Low-coverage RNA-seq captures similar dynamic range of gene expression as moderate-coverage and high-coverage RNA-seq

Next, we investigate whether lower-coverage RNA-Seq “misses” genes with a low overall expression due to sequencing bias. To do this, we stratify the 19175 protein coding genes measured in GTEX into five groups by mean expression and report how many genes from each of these groups are discovered as eGenes using (1) GTEx, (2) lower-coverage sequencing, (3) both datasets, and (4) neither dataset (**Figure 3.2D**). At the lowest quintile of expression (3835 genes total), we observe that GTEx reports just 6 of these genes as eGenes, while using lower-coverage sequencing reports 78 to be eGenes. In the other four quintiles of higher expression, we observe fairly consistent numbers of eGenes identified only in GTEx (794, 997, 1000, 876, in increasing order), indicating that the lower-coverage sequencing performs consistently across coverage levels. We perform an analogous analysis comparing GTEx and the moderate-coverage dataset (**Figure 3.8A**), and find that the moderate-coverage RNA-Seq also does not detect many eGenes from the lowest expressed quintile of genes.

3.3.7 Effect sizes of eGenes found using low-coverage and moderate-coverage RNA-seq remain highly concordant after accounting for lowly expressed genes

Next, we look at whether the effect size comparison in real data between eGenes discovered using lower-coverage and moderate-coverage is inflated due to poor estimation of lowly expressed genes in both datasets. Similarly to the previous section, we stratify the 19175 protein coding genes measured in GTEX into five groups by mean expression and report how many genes from each of these groups are discovered as eGenes using (1) moderate-coverage, (2) lower-coverage RNA-Seq, (3) both datasets, and (4) neither dataset (**Figure 3.8B**). If the effect size concordance was in fact inflated, in real data, we would see either a lot of shared detected or shared missed eGenes among the lowly expressed gene quintiles in the lower- and moderate- coverage data that are detected in GTEx. However, **Figure 3.8B** shows that none of the three datasets reliably detect eQTLs in the quintile of lowest expression.

3.3.8 eQTLs found using low-coverage RNA-seq colocalize with GWAS loci

To demonstrate that these eQTLs are implicated in GWAS loci, we run colocalization analysis using GWAS statistics from several blood traits (mean corpuscular volume, mean cell hemoglobin, and systemic lupus) (**Table 3.2**). Using a PP4 threshold of 0.80 (Methods), we see that a total of 51 unique eGenes (0.67% of significant associations) colocalize with a total of 50 unique GWAS SNPs. This is especially encouraging, as we see that there does not exist a redundancy of GWAS loci explained by eQTL hits. When performing the same analysis using data from GTEx, we find that a total of 91 unique eGenes (0.86% of significant associations) colocalize with 82 unique GWAS SNPs. 14 eGenes are in common with 5 GWAS SNPs involved in a significant colocalization in both datasets.

trait	n coloc eGenes – lower-coverage (PP4 > 0.8)	n coloc GWAS SNPs – lower-coverage (PP4 > 0.8)	n coloc eGenes – GTEx (PP4 > 0.8)	n coloc GWAS SNPs – GTEx (PP4 > 0.8)
Mean Corpuscular Volume	36	27	54	45
Mean Cell Hemoglobin	33	29	52	42
Systemic Lupus	6	6	22	11
All of the above	51	50	91	82

Table 3.2: *Coloc results for selected blood traits*

The number of unique eGenes (columns 1 and 3) and GWAS SNPs (columns 2 and 4) with PP4 > 0.80 when running colocalization analysis on significant eQTLs from analyses using lower-coverage RNA-Seq (columns 1 and 2) and results from GTEx (columns 3 and 4).

3.3.9 eQTLs found using low-coverage RNA-seq can be used for TWAS

We perform a TWAS analysis for the same three traits (**Table 3.4**) and find that using the low-coverage data, there are 143 significant TWAS associations. Using GTEx, there are 311 significant TWAS associations. Between the two datasets, 59 eGenes are shared.

trait	Lower-coverage - n TWAS eGenes	GTEx - n TWAS eGenes
Mean Corpuscular Volume	104	219
Mean Cell Hemoglobin	96	191
Systemic Lupus	33	75
All of the above	143	311

Table 3.3: *TWAS results for selected blood traits*

The number of unique eGenes (columns 1 and 3) and GWAS SNPs (columns 2 and 4) significant (FDR < 0.05) in TWAS on eQTLs with significant heritability from analyses using lower-coverage RNA-Seq (columns 1 and 2) and results from GTEx (columns 3 and 4).

3.3.10 Low-coverage and moderate-coverage RNA-seq have comparable computationally estimated cell type proportion values

Finally, we explore the impact of RNA-Seq at lower coverages for cell type expression estimation. We use CIBERSORTx [44] to compare cell-type proportion estimates between the lower-coverage data and moderate-coverage data (Methods). We find that the median estimated cell type proportions are conserved across both datasets, suggesting that deconvolution of cell type specific signal from gene expression profiles of whole blood samples is not impacted when coverage is reduced by half (**Figure 3.9**).

3.4 Discussion

We conclude with some notes, caveats, and future directions. First, synthetic RNA-Seq via down-sampling reads is potentially limited in several ways. These synthetic datasets of lower coverage RNA-Seq are created by uniformly sampling from real RNA-Seq data with an average of 50 million reads mapped per sample. However, in practice, it is possible that sequencing biases are not captured by uniform sampling due to the different experimental setup compared to the dataset from which we sample [33,47]. Additionally, these synthetic datasets are based on data obtained from fibroblast tissue with different transcriptomic profiles from whole blood, potentially influencing the sequencing depth required to detect associations with gene expression. Finally, this approach is optimized for eQTL discovery. Other mechanisms that are detected using RNA-Seq, such as RNA splicing, have different mechanisms and will likely have different optimal coverages for detection. The fact that we identify different sets of eGenes depending on which gene expression measurements we consider (GTEx vs eQTLGen vs lower-coverage RNA-Seq), shows that we need to increase cohort sizes in order to fully understand the connection between genetics and gene expression in blood. Furthermore, the results in **Figure 3.2A** (figure showing effective sample size at various coverages) indicate that even including 1490 individuals under this fixed budget is not enough to achieve the optimal effective sample size. Current approaches are not sufficient to understand the full landscape of eQTLs in whole blood tissue, even while only considering a single genetic ancestry group. We compare the eGenes identified by GTEx, eQTLGen, and the lower-coverage RNA-Seq (**Figure 3.11**) and find that no single study is sufficient in capturing all of the associations in whole blood. We also see evidence of this in **Figure 3.2D**, **Figure 3.7** and **Figure 3.8**, where the lower-coverage, moderate-coverage, and GTEx datasets do not detect nearly as many eGenes from the lowest quintile of genes by mean expression. Furthermore, as observed by the relatively low levels of overlap in colocalization and TWAS hits between GTEx and the lower-coverage sequencing, larger sample sizes are necessary to understand the roles of eQTLs with respect to GWAS. As observed in GWAS, much larger sample sizes including far more ancestral diversity in these samples will enable discovery of novel associations in transcriptomics. Including non-European populations and considering the temporal aspect of gene expression will help us gain a more complete understanding of the blood transcriptome landscape in the entire population.

3.5 Methods

Genotyping pipeline

Genotypes for the lower-coverage whole blood samples were obtained from the following platforms: OmniExpressExome (N = 810), PSYCH (N = 523), and COEX (N = 163). Given that the SNP-genotype data for both the fibroblast and whole blood samples came from numerous studies using various genotyping platforms, the number of overlapping SNPs across all platforms was < 150k, prompting us to perform imputation separately for each genotyping platform (**Supplemental Note**). Genotypes were first filtered for Hardy-Weinberg equilibrium p value < 1.0e-6 for controls and p value < 1.0e-10 for cases, with minor allele frequency (MAF) > 0.01, and SNP-missingness < 0.05, leaving 148612 typed SNPs. **Table S3.1** provides the number of typed and imputed SNPs per platform after quality control.

Genotypes were imputed using the 1000 Genomes Project phase 3 reference panel [42] by chromosome using RICOPILI v.1 [43] separately per genotyping platform. These platform-specific genotypes were then subsequently merged after imputation, applying an individual-missingness threshold of 10% and SNP-missingness of 5% for post-merge quality control. We restricted to only autosomal SNPs due to sex chromosome dosage, as commonly done [13]. Imputation quality was assessed by filtering variants where genotype probability > 0.8 and INFO score > 0.1, resulting in 2289732 autosomal SNPs. The low final number of imputed SNPs stems from relatively disjoint starting sets of quality-controlled, typed genotypes per platform, leading to smaller sets of high-quality imputed variants that overlapped across platforms (with less than 5% SNP-missingness). Despite this, we were able to use over 15-fold more variants in the merged imputed set as compared to the typed merged set. Then subsets of genotypes for the fibroblast-specific individuals, lower-coverage-specific individuals, and higher-coverage specific individuals were extracted from the merged file set to be used in the eQTL analyses.

RNA-Seq processing pipeline

We used FASTQC to visually inspect the read quality from the lower-coverage whole blood RNA-Seq (5.9M reads/sample), the moderate-coverage whole blood RNA-Seq (13.9M reads/ sample), and the high-coverage fibroblast RNA-Seq (50M reads/sample). We then used kallisto to pseudoalign reads to the GRCh37 gencode transcriptome (v33) and quantify estimates for transcript expression. We aggregated transcript counts to obtain gene level read counts using scripts from the GTEx consortium (https://github.com/broadinstitute/gtex-pipeline) [13].

cis-eQTL mapping

Excluding related individuals (pi_hat > 0.2) from the analysis, we perform cis-eQTL analysis mapping using FastQTL [37], using a defined window of 1 Mb both up and downstream of every gene's TSS, for sufficiently expressed genes (TPM > 0.1 in 20% of individuals). We run the eQTL analysis in permutation pass mode (1000 permutations, and perform multiple testing corrections using the q value FDR procedure, correcting at 5% unless otherwise specified. We then restrict our associations to the top (or leading) SNP per eGene.

TWAS and Colocalization

We used the FUSION framework [4] to perform the transcriptome-wide association study and subsequent colocalization [24] analysis. We computed single-best eQTL models for all eGenes detected in the lower-coverage dataset with the FUSION.compute_weights.R script. As this framework is intended for cis-loci, for each gene we restricted to SNPs within a window of 250kb around the gene start and gene end position from the set of imputed genotypes. For the functional phenotypes (input through the –pheno flag), we used the gene-level TPMs generated by aggregating kallisto transcript expression estimates using scripts from GTEx [34,13]. Once the weights were generated, we input them in the FUSION.assoc_test.R script along with summary statistics from blood-related GWAS: Mean Corpuscular Volume (MCV [38]), Mean Cell Hemoglobin (MCH [38]), and Systemic Lupus Erythematosus (SLE [39]); the 1000 Genomes LD panel for European ancestries was used as the reference. Colocalization was performed on those gene-trait associations that had p value less than 0.05 (--coloc_P 0.05 flag). This pipeline was then repeated using the GTEx V8 whole-blood gene expression (using the GTEx pipeline) and corresponding SNP-genotypes from 668 unrelated donors**.**

Covariates

For eQTL analyses conducted using the moderate-coverage whole blood and synthetic data derived from fibroblasts, we include the top three genotype principal components and top 50 gene expression principal components, calculated separately for each synthetic dataset. For eQTL analyses conducted using the lower-coverage whole blood, we include the top 10 genotype PCs (to account for the differences across the multiple genotyping platforms used to genotype samples in this cohort), and the top 50 expression PCs. In eQTL analyses using synthetic data we also include sex and several cell line technical covariates (passage number and growth rate). In eQTL analyses using moderate-coverage whole blood, we include sex, disease status, and age. In eQTL analyses using lower-coverage whole blood, we include sex, disease status, genotyping platform, and several technical covariates regarding the tissue samples (RIN and concentration).

Cell type proportion estimation

We estimate the proportion of cell types of both the lower-coverage and moderate-coverage bulk whole blood RNA-seq datasets using CIBERSORTx [44] with batch correction applied and LM22 signature matrix as the reference gene expression profile. The LM22 signature matrix uses 547 genes to distinguish between 22 human hematopoietic cell phenotypes.

3.6 Acknowledgments

We thank the study subjects for their willingness to provide specimens and clinical data. We thank Yi Ding, Kathryn Burch, Ruthie Johnson, Arjun Bhattacharya, Malika Freund, and Claudia Giambartolomei for meaningful discussion in helping make this work possible. T.S. was supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under award no. T32NS048004. T.B. was supported by the NIH (grant no. 5T32HG002536-19). This research was supported by the National Institute of Mental Health of the National Institutes of Health under award no. 5R01MH115676-04. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

3.7 Figures

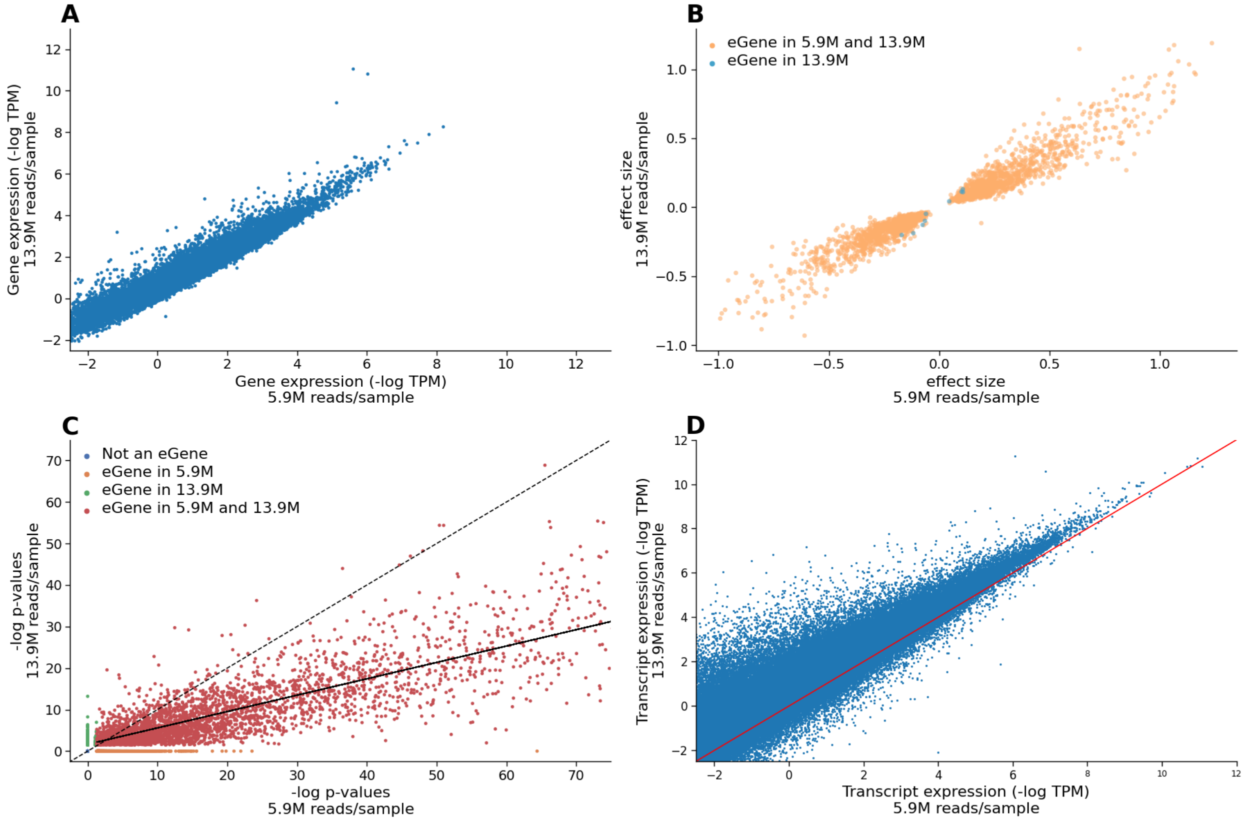


Figure 3.1: *Concordance of eQTL discovery when using lower-coverage RNA-Seq vs moderate-coverage RNA-Seq*

(3.1A): Restricting to the 20735 genes with sufficient expression levels to be included in eQTL analysis in both the 5.9M read/sample and 13.9M read/sample dataset, comparison of the median expression (log TPM) across samples, of every gene. $R^2 = 0.91$. **(3.1B):** In real data, scatterplot of effect sizes of most significant eQTL hits for the 2151 protein coding genes with the same eQTL hit in both eQTL analyses performed (lower-coverage and moderate-coverage). On the x-axis, we show the effect sizes for these genes using lower-coverage RNA-Seq, on the y-axis we show the effect sizes for these genes using moderate-coverage RNA-Seq. **(3.1C):** Real data p-value comparison scatterplot: In real data, scatterplot of -log p-values of most significant eQTL hit for 13950 genes included in both eQTL analyses performed (lower-coverage and moderate-coverage). On the x-axis, we show the -log p-values for these genes using lower-coverage RNA-Seq, on the y-axis we show the -log p-values for these genes using moderate-coverage RNA-Seq. The dotted line shows $y = x$, while the solid line shows the line of best fit for the 3985 protein-coding eGenes with a significant eQTL hit in both datasets. **(3.1D)** For the 227046 unique isoforms detected in the lower-coverage and moderate-coverage datasets, we show the mean expression across samples in each dataset ($R^2 = 0.83$).

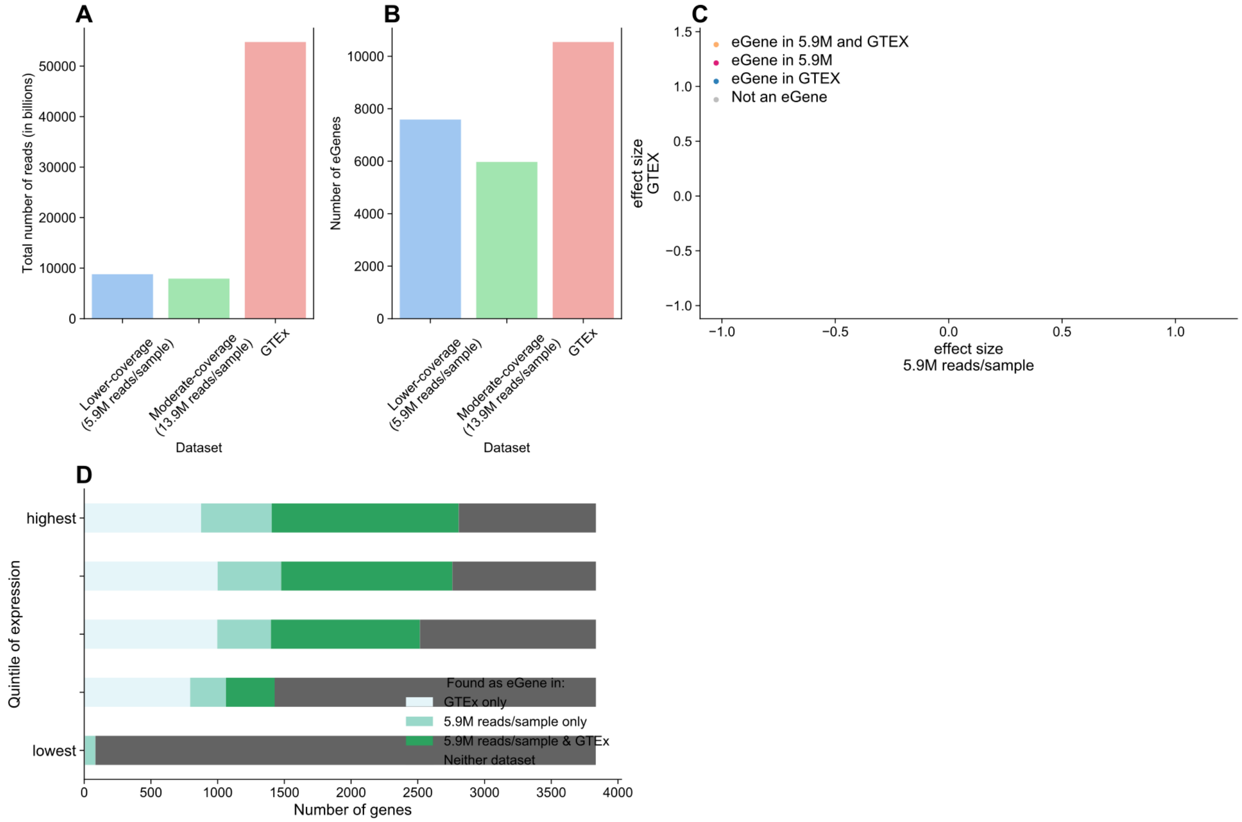


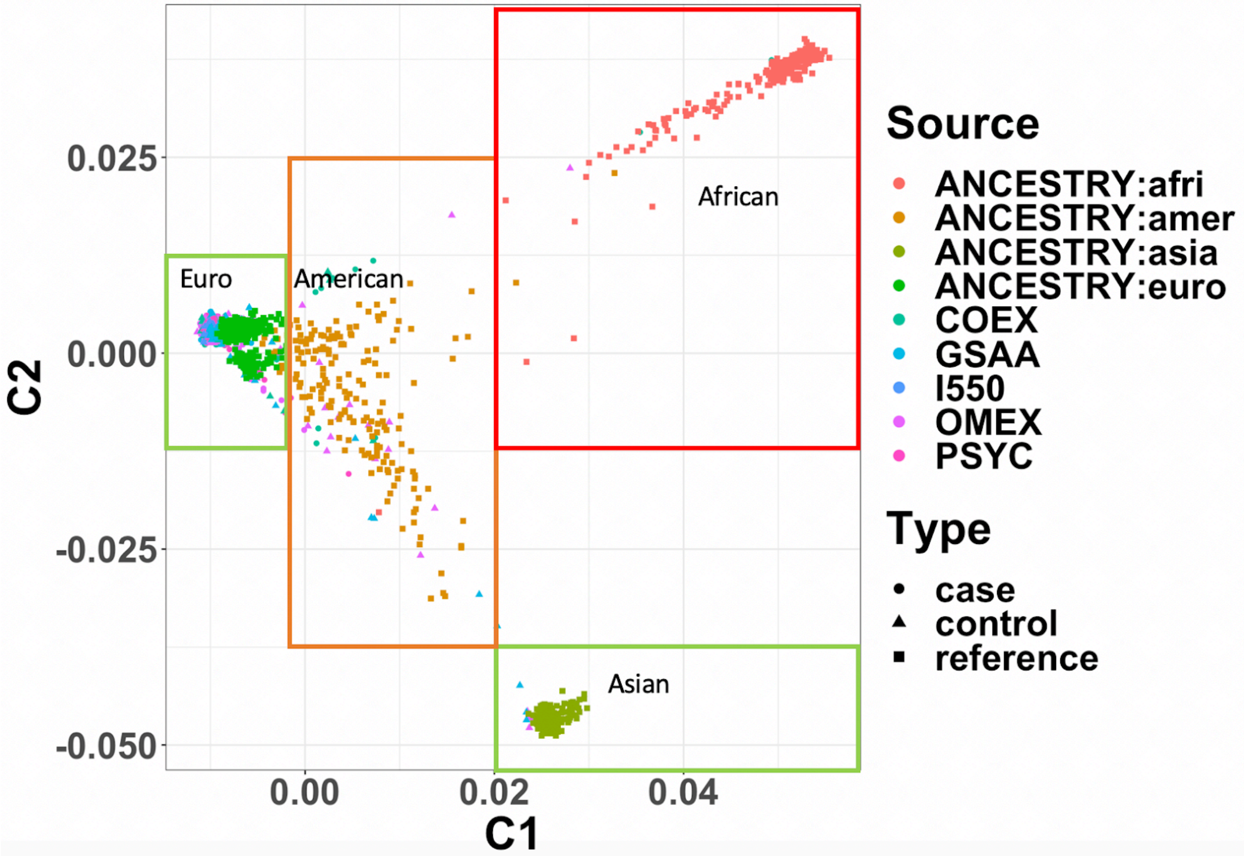
Figure 3.2: eQTL analysis using *lower-coverage RNA-Seq is comparable to eQTL analysis from the GTEx Consortium*

(3.2A): Estimates for the total number of reads (in billions) included in each of the three RNA-Seq experiments that we compare. **(3.2B):** Number of eGenes discovered at an FDR correction level of 0.05 in each of the three datasets that we compare. **(3.2C):** In real data, scatterplot of effect sizes of the most significant eQTL hit for the 1927 eGenes with leading eSNPs in LD with $R^2 > 0.25$ between the two datasets (lower-coverage RNA-Seq with 5.9M reads/sample and

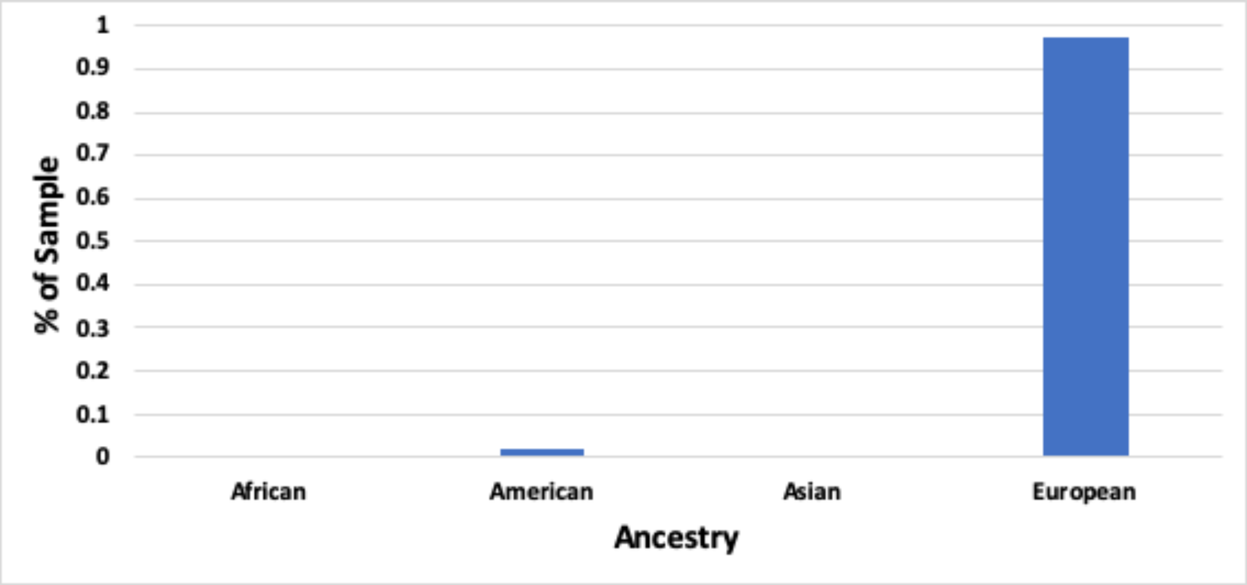
GTEX). On the x-axis, we show the effect size for these eGenes from eQTL analysis conducted using the 1490 individuals of EUR ancestry and typed genotypes, and on the y-axis we show the effect sizes for these eGenes from eQTL analysis published by the GTEX Consortium.

(3.2D) The overlap in eGenes identified in the lower-coverage RNA-Seq and GTEX, stratified into quintiles by the mean expression level observed in GTEX.

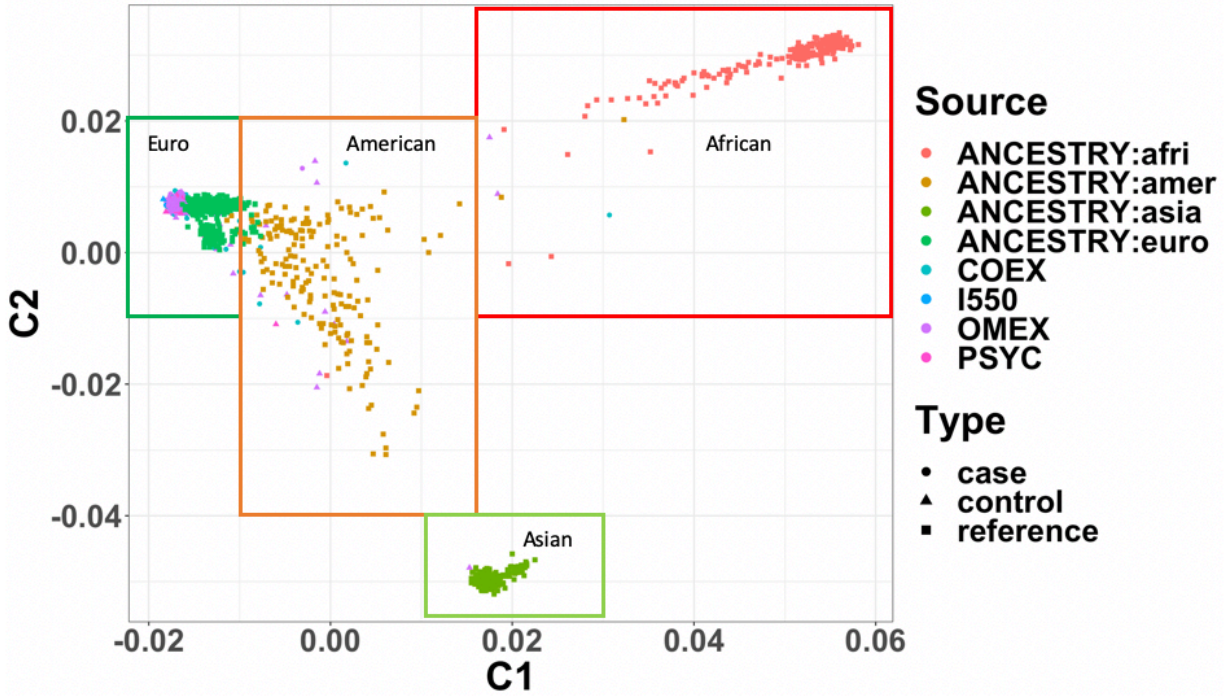
3.3A:



3 . 3 B :


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
% of Sample
African
American
Asian
European
Ancestry

3 . 3 C :


Euro
American
African
Asian
0.02
0.00
-0.02
-0.04
C2
-0.02
0.00
0.02
0.04
0.06
C1
Source
● ANCESTRY:afri
● ANCESTRY:amer
● ANCESTRY:asia
● ANCESTRY:euro
● COEX
● I550
● OMEX
● PSYC
Type
● case
▲ control
■ reference

3.3D:

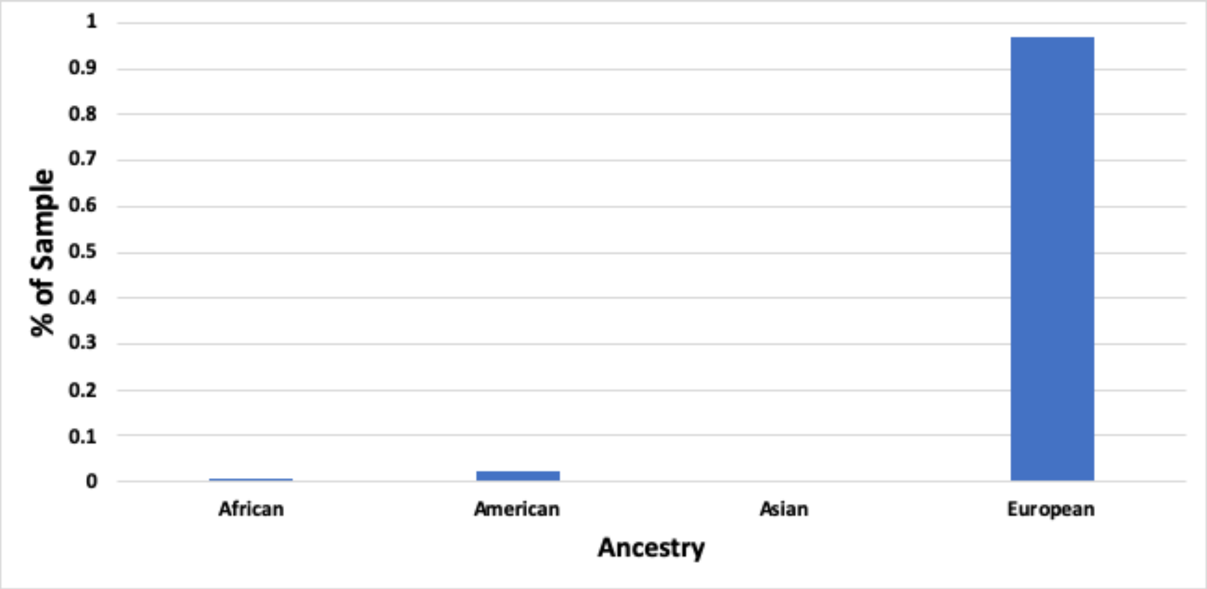


Figure 3.3: *Distribution of ancestry among samples.*

(3.3A) Genotype PC1 and PC2 are projected onto PCs from 1000 Genomes Project. Points labeled with “ANCESTRY” are from 1000 Genomes Project, remaining points designate the specific genotyping platform used in our cohort. Boxes are drawn around the centers to show where samples from the n = 2000 / 5.9M reads/sample cohort lie. **(3.3B)** A barplot showing the distribution of ancestry observed in the n = 2000 / 5.9M reads/sample cohort, according to the MDS plot. Note that only the 1963 samples that pass genotype QC thresholds are included here. Exact numbers of samples per ancestry group are: African - 4, American - 34, Asian - 9, European – 1916*.* **(3.3C)**: Genotype PC1 and PC2 are projected onto PCs from 1000 Genomes Project. Points labeled with “ANCESTRY” are from 1000 Genomes Project, remaining points designate the specific genotyping platform used in our cohort. Boxes are drawn around the centers to show where samples from the n = 759 / 13.9M reads/sample cohort lie. **(3.3D)** A barplot showing the distribution of ancestry observed in the n = 759 / 13.9M reads/sample cohort, according to the MDS plot. Exact numbers of samples per ancestry group are: African - 4, American - 19, Asian - 1, European – 735.

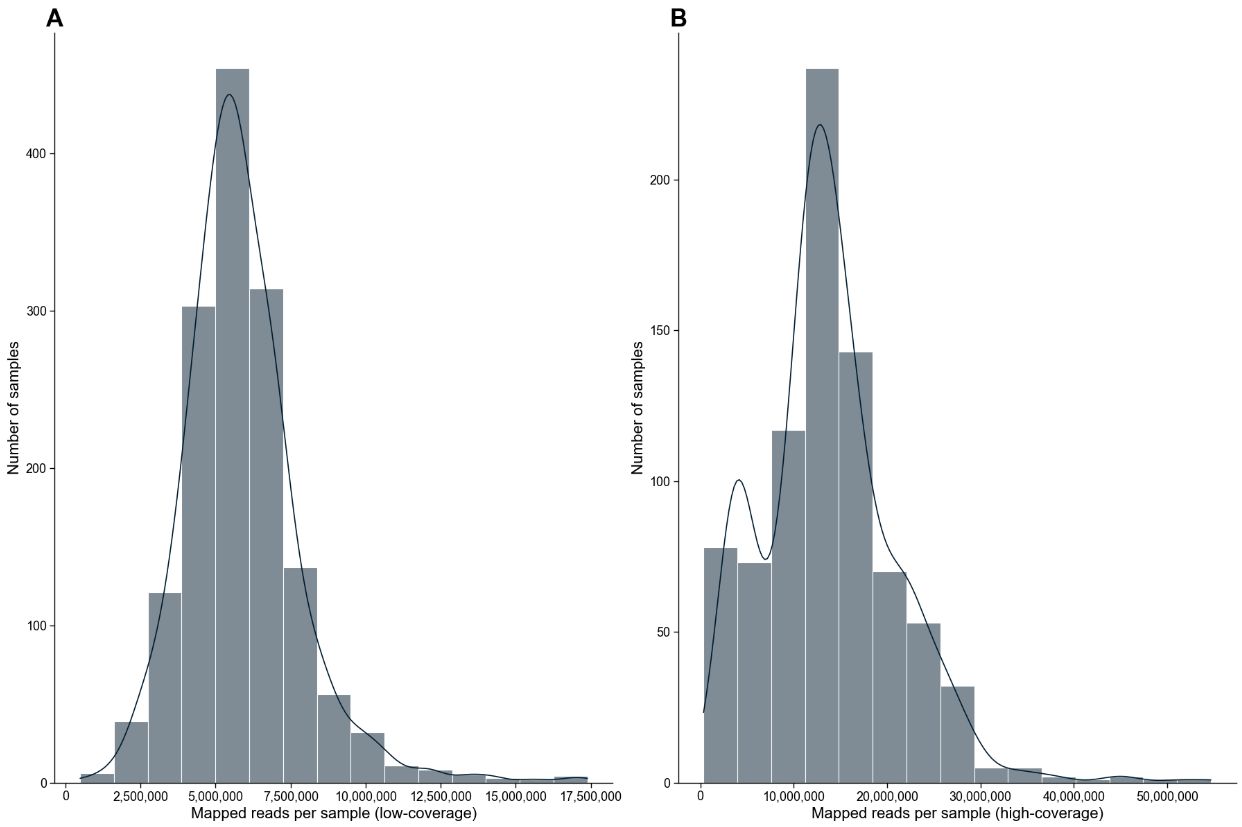


Figure 3.4: *Number of pseudoaligned reads per sample for low-coverage and high-coverage experiments*. **(3.4A)** In real data, a histogram showing the number of reads mapped to genes (or in kallisto terms: number of reads for which transcriptome successfully mapped), per sample. **(3.4B)** In real data, a histogram showing the number of reads mapped to genes (or in kallisto terms: number of reads for which transcriptome successfully mapped), per sample.

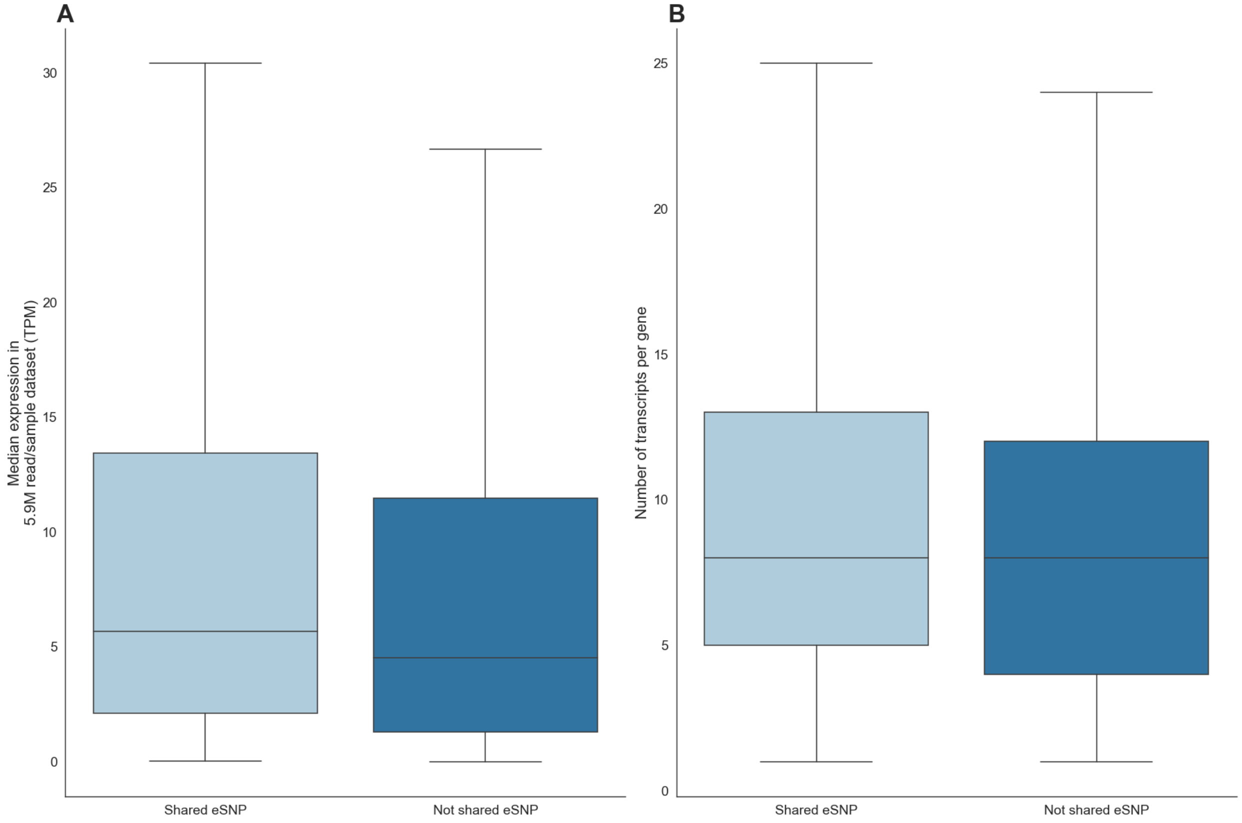


Figure 3.5: *Characteristics of eGenes that do/do not share the same eSNP between lower-coverage and moderate-coverage RNA-Seq.* **(3.5A)** eGenes with the same eSNP have an average expression of 15.6 TPM (sd = 51.1) while eGenes that do not share the same eSNP have an expression of 14.2 TPM (sd = 40.8), (p = 0.03). (**3.5B)** The average number of isoforms for eGenes that shared an eSNP was 10.0 (sd = 8.0), while it is only 9.6 (sd = 7.9) for those eGenes that do not share an eSNP (p = 0.03).

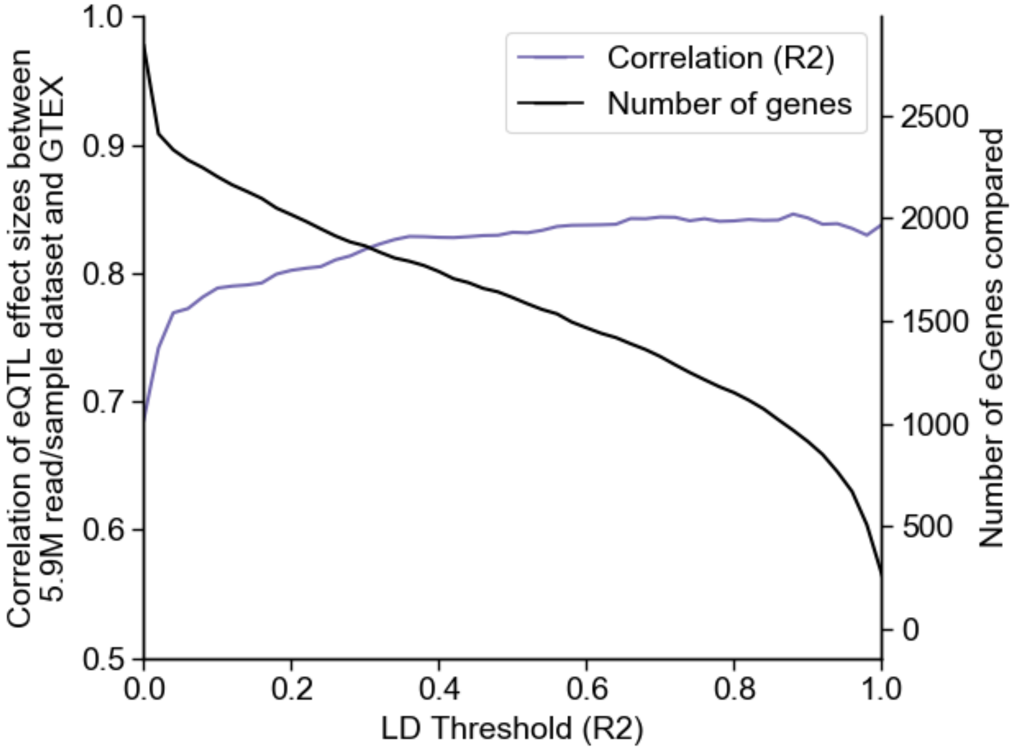


Figure 3.6: *Correlation of effect sizes between eQTL analyses low-coverage and GTEX with respect to LD-threshold between eSNPs.* **(3.6)** On the x-axis, we show the LD threshold used to restrict our comparisons of effect sizes between eQTLs found in the low-coverage dataset and GTEX. Only eGenes with the same eSNP, or eSNPs with an LD above the threshold are used. On the left-hand y-axis, we show the correlation of the effect sizes. On the right-hand y-axis, we show the number of eGenes compared under the given LD threshold.

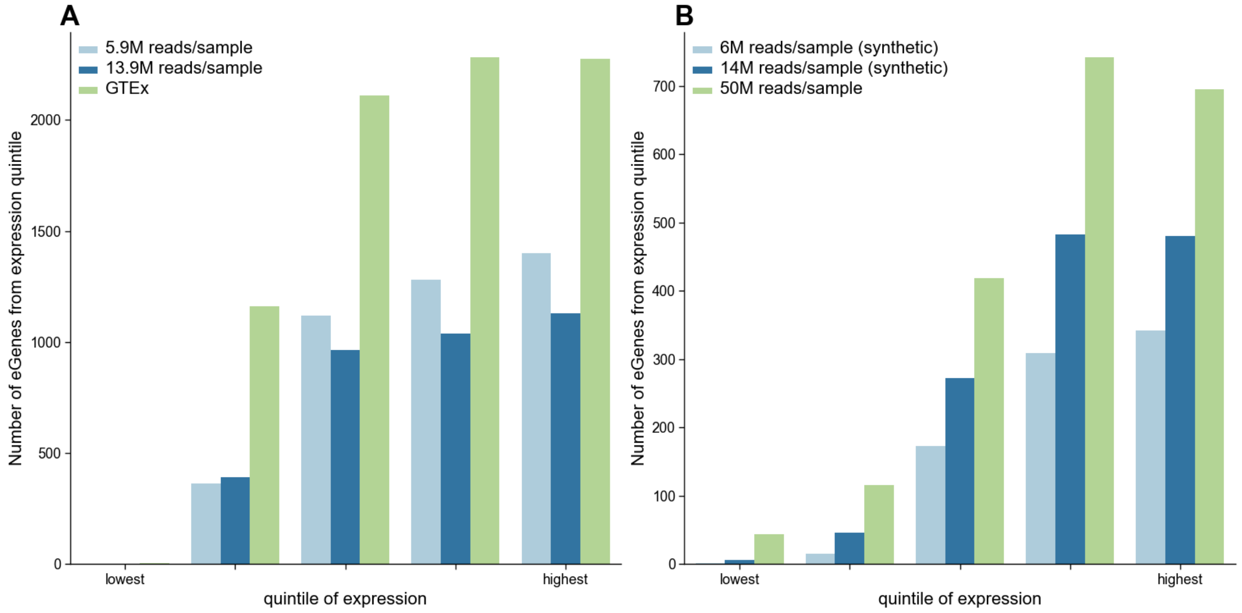


Figure 3.7: *Number of eGenes per mean expression quintile across datasets.* **(3.7A)** Stratifying the 19175 protein coding genes reported in GTEx into quintile groups by mean expression, on the x-axis, we show the quintile groups by increasing mean expression. On the y-axis, we show the number of eGenes found in the (1) low-coverage, (2) high-coverage, and (3) GTEx experiments, in each of these quintile groups. **(3.7B)** Stratifying the 24206 protein coding genes discovered in the high-coverage fibroblast dataset into quintile groups by mean expression, on the x-axis, we show the quntile groups by increasing mean expression. On the y-axis, we show the number of eGenes found in the (1) synthetic RNA-Seq at 6M reads/sample, (2) synthetic RNA-Seq at 14M reads/sample, and (3) RNA-Seq data at 50M reads/sample, in each of these quintile groups.

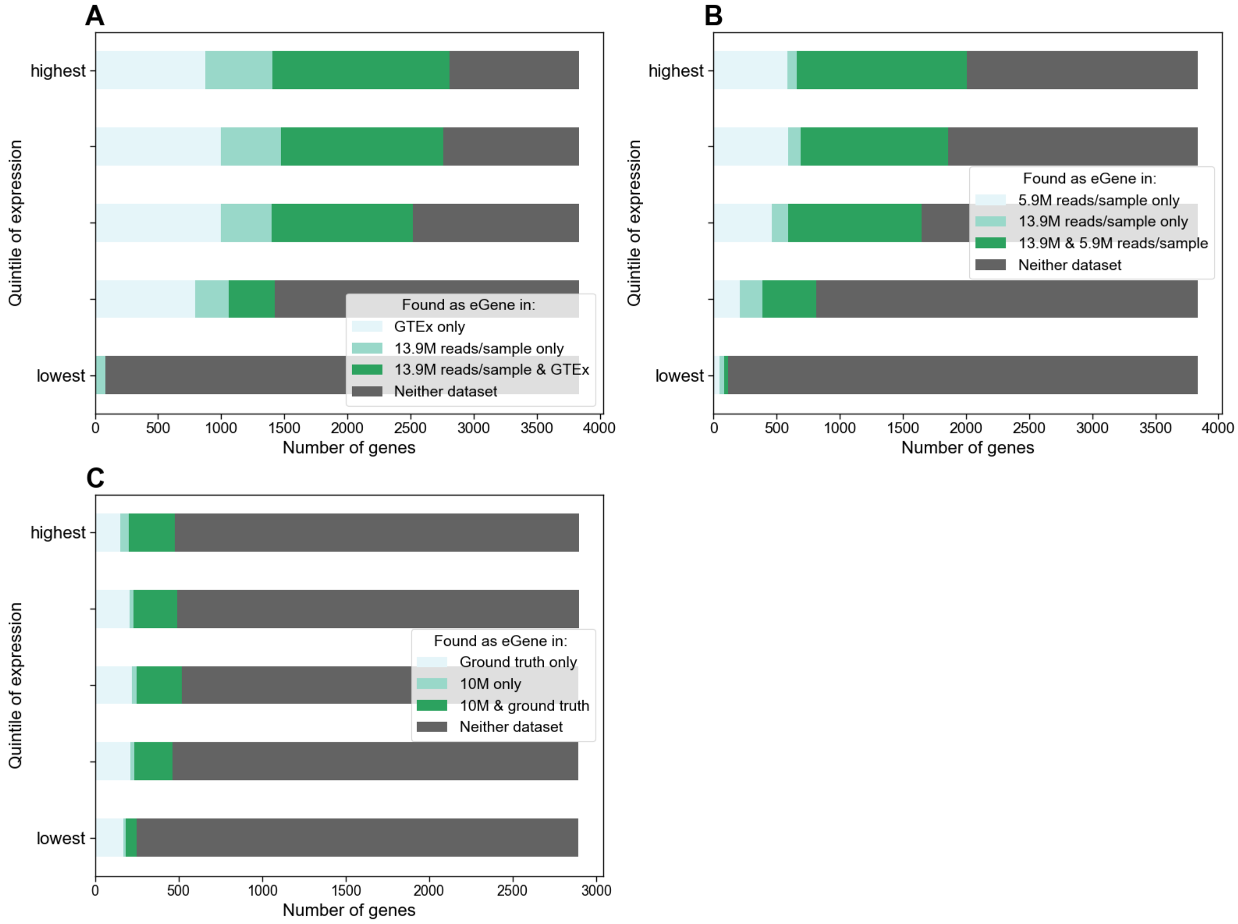


Figure 3.8: *eGene concordance by quintile in moderate-coverage RNA-Seq and GTEx:* **(3.8A):** The overlap in eGenes between moderate-coverage RNA-Seq and GTEX, stratified into quintiles by the mean expression level observed in GTEX. **(3.8B):** The overlap in eGenes between moderate-coverage RNA-Seq and lower-coverage RNA-Seq, stratified into quintiles by the mean expression level observed in GTEX. **(3.8C):** The overlap in eGenes between high-coverage "ground-truth" RNA-Seq and a 10M read/sample synthetic dataset, stratified into quintiles by the mean expression level observed in the high-coverage RNA-Seq

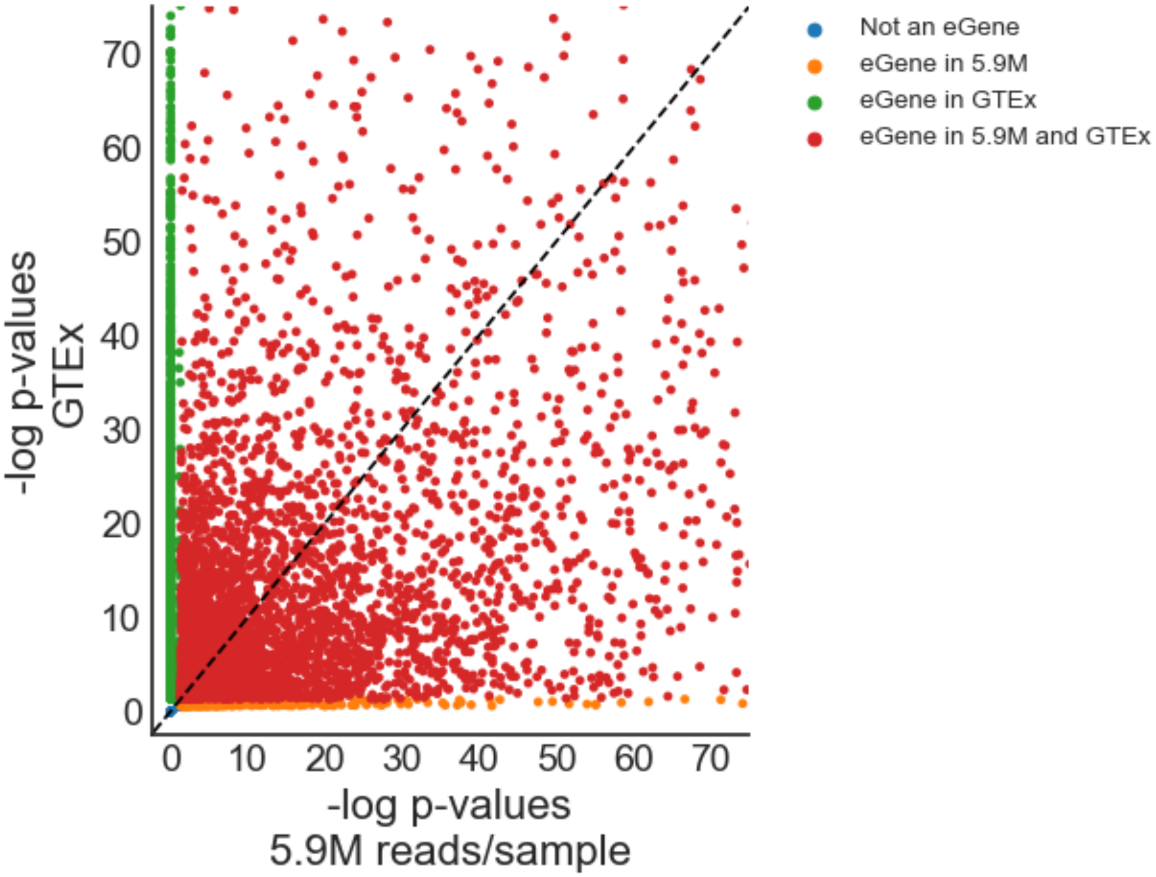


Figure 3.9: *Real data p-value comparison scatterplot with GTEX*. **(3.9A)** Using the 12,496 protein-coding genes included both in GTEX and the low-coverage datasets, on the x-axis, we show the -log p-values for leading SNP eQTL associations in the low-coverage dataset. On the y-axis, we show the -log p-values for leading SNP eQTL associations in the GTEX dataset.

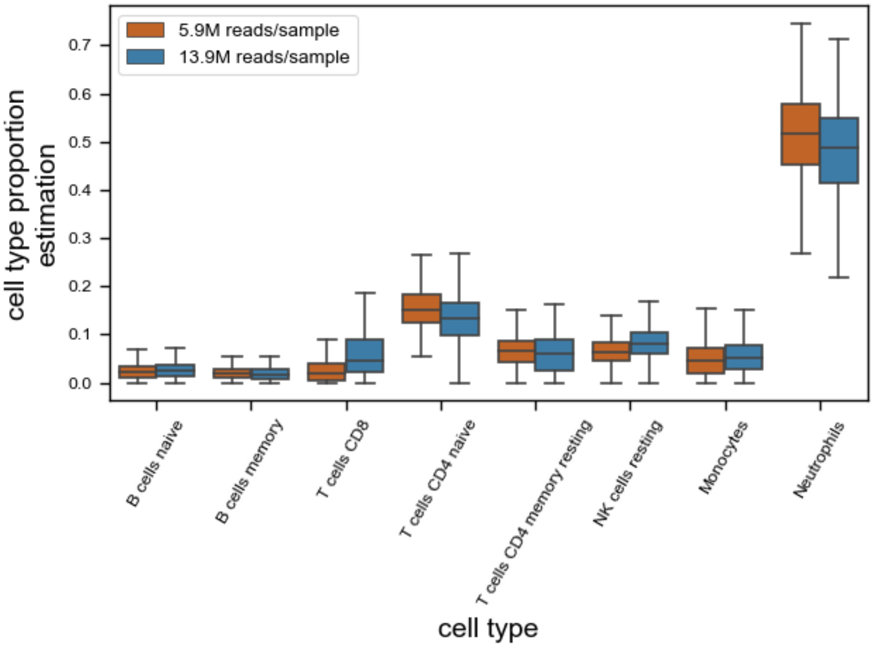


Figure 3.10: *Estimation of cell-type proportions.* **(3.10)** In real data, a comparison of estimated cell type proportions from CIBERSORTx between lower-coverage (5.9M reads/sample) and moderate-coverage (13.9M reads/sample) RNA-Seq data for the eight most common cell types in whole blood tissue.

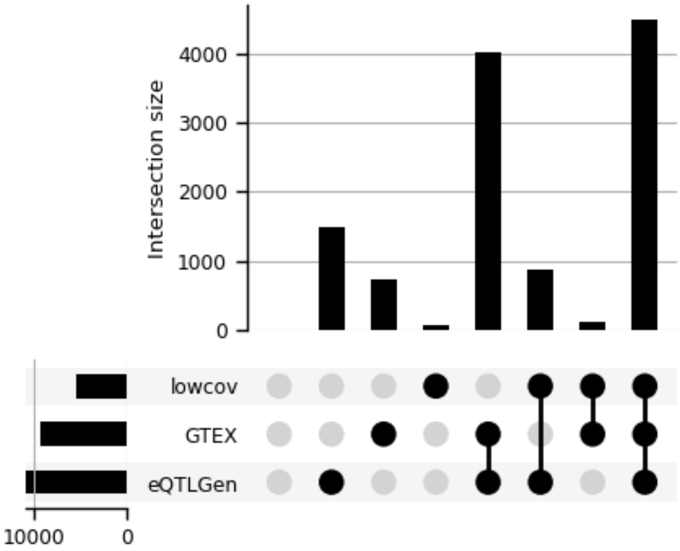


Figure 3.11: *Overlap of significant eGenes using RNA-Seq from three different datasets.* **(3.11)** Comparing number of genes with significant associations between three datasets: (1) Lower-coverage RNA-Seq (5.9M reads/sample on average, across 1,496 individuals), (2) GTEX (83M reads/sample on average, across 670 samples), (3) eQTLGen (31,684 individuals, mix of RNA-Seq and MicroArray assays used).

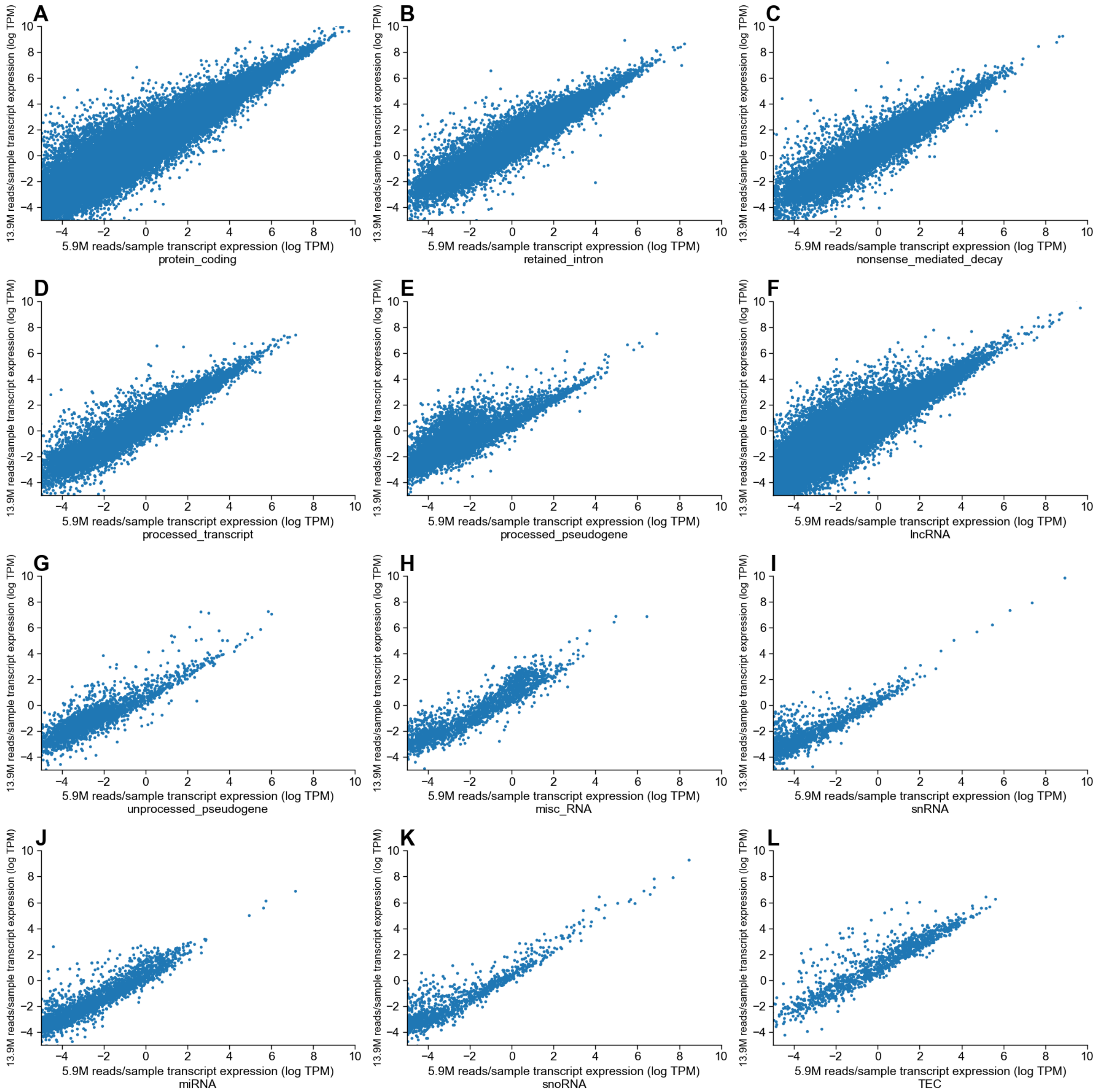


Figure 3.12: *Concordance of transcript expression estimation between lower-coverage RNA-Seq vs moderate-coverage RNA-Seq.* **(3.12)** For the 12 most highly represented transcript types (> 1000 transcripts quantified), we show mean expression estimates in lower-coverage RNA-Seq (x-axis) versus moderate-coverage RNA-Seq (y-axis).

3.8 Supplementary Tables

Platform	**N Typed SNPs**	**N Imputed SNPs**
OmniExpress Exome	619,690	5,576,428
Global Screening Array (GSA)	514,169	4,383,620
COEX	323,599	4,736,265
PsychArray	378,710	4,528,960
Merged	148,612	2,289,732

Table S3.1: *Number of typed and imputed SNPs after QC*

Dataset	Estimated number of total reads mapped in experiment	Estimated cost of experiment	Proportion of cost compared to lower-coverage RNA-Seq experiment	Proportion of eGenes identified compared to lower-coverage RNA-Seq
Lower-Coverage or M=5.9M reads/sample (Whole Blood)	~8.8B	~$292,000	1.0	1.0
GTEx	~55.6B	~$620,000	2.12	1.39

Table S3.2: *Whole blood RNA-Seq datasets and respective cost estimates*. We describe the lower-coverage and GTEx datasets in terms of estimated cost from our budget model and an

estimate for total number of reads used. We assume that the cost of genotyping is $53 per sample (per UCLA Neurogenetics Sequencing Core).

Transcript type	Number of transcripts	Correlation (R^2) of expression
Protein coding	83735	0.92
Retained intron	28411	0.89
Nonsense mediate decay	15856	0.88
Processed transcript	14128	0.89
Processed pseudogene	10055	0.72
lncRNA	59133	0.85
Unprocessed pseudogene	2644	0.81
miscRNA	1987	0.82
snRNA	1730	0.95
miRNA	2757	0.81
snoRNA	1367	0.94
TEC	1135	0.82

Table S3.3: *Transcript-level expression correlations by transcript type.* We quantify expression for 226,390 transcripts. For the 12 transcript types with at least 1000 transcripts represented, we calculate the correlations between mean transcript expression estimated using lower-coverage RNA-Seq and moderate-coverage RNA-Seq.

3.9 Supplementary Note

Notes about overlap in datasets:

- The samples in the low-coverage whole blood and high-coverage whole blood datasets are completely disjoint – no individuals overlap here.
- 97 individuals have RNA-seq data in both the high-coverage fibroblast dataset and low-coverage whole blood dataset
- 41 individuals have data in the high-coverage fibroblast dataset and the high-coverage whole blood dataset

- In total, 138 individuals overlap between the high-coverage fibroblast RNA-Seq samples and whole blood RNA-Seq samples (low-coverage and high-coverage)

Chapter 4: Computational deconvolution of bulk RNA-seq enables cell type biological insights

4.1 Abstract

eQTL mapping using expression estimates from bulk RNA-seq is a widely used tool for understanding GWAS. However, a limitation of bulk tissue expression is loss of cell type signal. Here, we leverage bMIND to compute cell type specific gene expression estimates using bulk RNA-Seq from 1,996 samples derived from whole blood tissue. Using the LM22 signature matrix from CIBERSORTx as a reference, we estimate gene expression for eight different immune cell types with average proportion across samples >= 0.05, including neutrophils, naïve B cells, memory B cells, CD8 T Cells, naïve CD4 T cells, memory CD4 T cells, NK resting cells, and monocytes. We show that these expression estimates can be used to conduct cell type eQTL analyses, identifying between 2,875 and 4,629 unique eGenes for each cell type, including 1,268 eGenes that are not found using bulk gene expression estimates. We find evidence of both shared and independent effects between cell type eQTLs and a standard eQTL analysis using estimates from bulk tissue. Finally, we investigate the effects of lithium use on cell type expression regulation and find 110 examples of genes whose cell type expression are differentially regulated dependent on lithium use, compared to just one whose bulk expression is differentially regulated dependent on lithium usage. Our study suggests that computational methods can be applied to large bulk RNA-Seq datasets to identify cell type gene expression signal and cell type specific biology.

4.2 Introduction

Bulk RNA-seq has enabled researchers to measure gene expression at scale at the tissue level. Among its many uses and applications, integrating bulk RNA-seq estimates with genetic information allows us to find where genetic variants may be associated with regulation of gene expression. This approach, known as eQTL analysis, has helped provide functional information to genetic variants that we would not see from GWAS. However, one limitation of standard eQTL studies is that they generally use expression estimates from bulk tissue. While this is informative, it is believed that there are many cell type specific mechanisms driving biology [55-58], which can be missed when looking at a collection of many cell types. In recent years, single cell RNA-Seq has enabled us to profile the gene expression of an individual cell, giving us a clearer picture of cell type gene expression [54]. However, single cell RNA-Seq experiments are considerably more expensive than bulk RNA-Seq, making it cost-prohibitive to perform these assays at the scales necessary to gain a complete picture of expression regulation at the cell type level. To leverage the advantages of each of these approaches, we can use computational methods to estimate cell type gene expression from bulk RNA-Seq expression.

There exist many methods to estimate cell type expression from bulk RNA-Seq. We elected to use CIBERSORTx [43] and bMIND [49] to estimate cell type proportions and cell type expression, respectively. Previous work comparing the effectiveness of various methods to estimate cell type proportion identify CIBERSORTx as one of the better performing methods across different contexts [47,53]. Computational methods for analyzing bulk gene expression data have the potential for being advantageous in some applications as it is possible to obtain much larger sample sizes using bulk RNA-Seq instead of single cell RNA-Seq. While most single cell RNA-Seq studies have sample sizes in the range of several hundreds of cells [35], leveraging low-coverage bulk RNA-Seq allows us to obtain samples from nearly 2,000 individuals. Recent studies have shown that there exists a strong shared “cis” component to expression regulation between cell types within a single tissue [48]. Larger sample sizes will better enable us to investigate both the shared and distinct cis-eQTL signal within tissues.

This cohort has been ascertained for individuals with BP. In addition to genetics and RNA-seq, the cohort also includes lithium use status at the time of blood draw. In Europe, where this cohort was recruited from, lithium is the most commonly prescribed treatment for BP. Previous studies have shown that bulk whole blood gene expression measurements in individuals with bipolar disorder are heavily confounded by lithium usage [59]. With increased sample size, we are curious to investigate the differential expression of genes in the context of lithium users and nonusers. Specifically, whether there exist eQTLs, both at the bulk and cell type level, whose effect size is significantly different dependent on lithium use status.

In this work we build a cell type decomposition pipeline, leveraging several publicly available tools, to derive cell type estimates for gene expression. We then use these results to conduct cell type cis-eQTL analyses, and compare the shared and unique cell type associations. We show that these cell type eQTL results derived from deconvoluted bulk RNA-Seq are consistent with eQTLs from scRNA-Seq [50, 52]. We go on to identify several examples of "opposite-effect" eQTLs, where a cell type eQTL signal demonstrates gene expression regulation in the opposite direction from that observed in a bulk eQTL study. Finally, we explore the effects of lithium use on cell type expression, and identify 110 examples of lithium-SNP interactions dictating the effect of an eQTL.

4.3 Results

4.3.1 Overview of cell type expression estimation pipeline

To estimate cell type gene expression in whole blood, we use results from analysis of bulk RNA-Seq [46] (N = 1,996) and computational deconvolution tools (**Figure 4.1**). First, we estimate cell type proportions using the LM22 signature matrix and CIBERSORTx [43] (**Figure 4.2A**). We find that these proportion estimates are consistent with those from other cohorts, namely with neutrophils in highest abundance, lymphocytes (including T cells, B cells, NK cells combined) in second highest abundance, and monocytes in lowest abundance, in general. However we note that blood cell type proportions vary widely per individual depending on numerous factors such as medication use, current illness, or age. We find that the proportions estimated via CIBERSORTx are consistent with the complete blood count measures taken in the clinic for a subset (N=143) of individuals in our dataset (**Figure 4.5**). For example, we observe a pearson correlation (R^2) of 0.76 for cell type proportions estimated in neutrophils using CIBERSORTx and proportions measured in clinic. These results suggest that the computationally estimated proportions are reliable.

4.3.2 Cell type expression estimation using computational tools

Next, we use these proportion estimates and an expression deconvolution software called bMIND (**Methods**) to estimate cell type expression. As expected, we find that R^2 of expression between different cell types is high, as all cell types are derived from the same tissue (**Figure 4.2B**). Next, we wanted to investigate whether despite there being an expectedly high correlation structure between different cell types, if computationally estimated cell type expression could successfully detect the differences in the expression between different cell types. We focused on the 548 genes included in the LM22 matrix (**Methods**) and found that these correlations ranged from 0.45 to 0.87. Finally, principal component analysis confirms that the major sources of variation in the dataset are attributable to differences in cell type expression (**Figure 4.8**). These results suggest that using large cohorts of bulk RNA-Seq, paired with computational deconvolution tools, finds differences in expression dependent on cell type composition.

4.3.3 Computationally estimated cell type proportions match sc-RNA-seq estimates

In order to validate whether the expression estimates we derive using computational methods sufficiently match expression estimates observed using single cell RNA-Seq (scRNA-Seq), we leverage two scRNA-Seq datasets. We compare median TPM estimates across six cell types and find moderate correlation between the reference single-cell expression and computationally derived expression, ranging from R^2 of 0.11 in naive B cells to R^2 of 0.27 in CD8 T cells (**Supplementary Table 4.1** and **Figure 4.7**). To further check how well computationally estimated expression compares to expression derived from scRNA-Seq, we look at how similar expression estimates are between the two reference scRNA-Seq datasets in monocytes, the one cell type with data available in both reference datasets. We find that the median TPM of the 2836 genes discovered to be eGenes in both datasets have an R^2 of 0.22, comparable to the correlations observed when comparing computationally estimated expression with scRNA-Seq.

4.3.4: Cell type expression accounts for varying proportions of variance of BP status

Among this cohort, there are 1,126 individuals diagnosed with BP, 104 individuals diagnosed with schizophrenia (SCZ), and 766 control individuals. We were interested to see whether there were BP-specific effects that could be observed using cell type deconvoluted expression and related information. We conduct a GREML analysis (**Methods**) to find the amount of variance in BP attributable to variance in gene expression (**Supplementary Table 4.3**). Using the bulk gene expression, we see that up to 87% of trait variance can be explained by variance in gene expression. At the cell type level, there exists varying degrees of trait heritability attributable to cell type level, the highest being Neutrophils, estimated at 95%. From this, we learn that cell type inferred gene expression can be a useful tool for trait prediction.

4.3.5: Cell type eQTL analysis reveals more refined biological signal

Next, we were interested in performing eQTL analyses on the resulting cell type expression estimates to find evidence of genetic regulation of cell type expression. Restricting to the eight cell types with average proportion > 2% (**Figure 4.2A**), naive B Cells, memory B Cells, CD4 naive T Cells, CD4 memory T cells, natural killer cells, monocytes, and neutrophils. Also restricting the analysis to 1,730 unrelated European individuals, we conduct cis-eQTL mapping with a 1 Mb window using QTLtools (**Methods**), to identify between 2,875 and 4,629 genes (eGenes) with a significant association at FDR correction level of 5%, across the eight different cell types (**Figure 4.3A**). In total, we identify 5,752 genes with a significant association in at least one of the eight main cell types. We go on to show that there exists a range of concordance of effect sizes for eGenes found in both the individual cell type analyses and the bulk eQTL analysis (**Figure 4.3B** and **4.3C**). This confirms findings from previous studies showing a strong shared genetic effect on gene expression across cell types. We observe that most ct-eGenes are detected as significant in either just one, or all eight cell types (**Figure 4.9**). Additionally, we find evidence of cell type "opposite-effect" eQTLs, where a SNP in a given cell type shows an association with the same eGene as detected using bulk RNA-Seq, but in the opposite direction. These examples are especially interesting as it supports the idea that looking at gene expression at the cell type level can uncover new biological mechanisms that go undetected when only using bulk tissue. Similar effects have been observed in other studies using both single cell RNA-Seq and deconvoluted bulk RNA-Seq.

4.3.6: Comparison of computationally estimated ct-eQTL effect sizes with ct-eQTL effect sizes derived from sc-RNA-seq

To further validate these cell type eQTLs, we compared the results of this analysis with results from eQTL analysis using single cell RNA-Seq from the BLUEPRINT consortium. We restrict to the protein coding genes identified as eGenes using the computational deconvolution approach. Generally, we find that the two approaches to cell type eQTL mapping show strong concordance. For example, in neutrophils, we find that 2,921 out of the 4,629 genes (63%) with a significant association using the computational deconvolution approach also had a significant association in using single-cell RNA-Seq, correcting at an FDR level of 5%. Among these eGenes, comparing the association with the same leading SNP in both of these datasets (**Figure 4.3D**), we observe a correlation (R^2) of 0.66 between their effect sizes. This suggests that the computational deconvolution approach to large scale bulk RNA-Seq projects can be used to obtain accurate cell type eQTL estimates.

4.3.7: Lithium-SNP interaction models at the bulk and cell type expression levels

To investigate lithium-dependent genetic regulation, we perform an interaction model eQTL scan between lithium users and nonusers, testing whether there exist SNPs whose cell type or cell type specific expression regulation is dependent on the presence of lithium. To do this, we include an interaction term for the genotypes and lithium status, in the regression model

(**Methods**). Using bulk expression, we only identify one gene with such an association (FDR p-value < 0.10). Looking at cell type expression derived from bMIND, we identify as many as 34 such eGenes (in monocytes), and a total of 110 examples of genes (Li-eGenes) that show differential regulation of cell type expression, compared to just one gene that shows differential regulation of bulk expression (**Supplementary Table 4.4**). We see that 97 of the eGenes that have significant differential lithium regulation exhibit opposite effect sizes between the lithium user and nonuser groups, at the cell type level. The remaining 13 Li-eGenes show same direction effect sizes between the lithium user and nonuser groups, with significantly different magnitudes. For example, in naïve B cells, KITLG (ENSG00000049130) shows opposite effect eQTLs based on rs11104703 (**Figure 4.4A**). While in monocytes we see that TNFRSF11A (ENSG00000105641) shows differential effect size, in the same direction, based on rs79143095 (**Figure 4.4B**). Due to the large number of samples used in this analysis, we are powered to detect small differences, like these.

4.4 Discussion

In this work, we leverage a large-scale bulk RNA-seq dataset and computational deconvolution methods to estimate cell type expression in eight major cell types in whole blood tissue. Using several published scRNA-seq datasets as a reference, we validate that expression is accurately estimated using the computational deconvolution approach. We then use these cell type expression estimates to conduct eQTL analyses at the cell type level and again use reference sc-RNA-seq datasets to validate effect size estimates observed in computationally deconvoluted inferred ct-eQTLs. Finally, we demonstrate that we can use these cell type expression estimates along with lithium-usage history from individuals in this cohort to study SNP-lithium interaction effects on expression, both at the bulk and cell type level, to uncover genes that have not previously been implicated in lithium regulation of whole blood gene expression.

Our study is, in part, motivated by trying to find approaches to boost statistical power in association studies, at the cell type level. Leveraging bulk RNA-seq to look at cell type biological questions enables use of larger sample sizes than would be possible when using scRNA-seq. We conclude with some notes, caveats, and future directions. We tested many computational tools for deconvolution of bulk RNA-seq data, and while gene expression estimates from CIBERSORTx + bMIND were the most consistent with scRNA-seq estimates, it is possible that this was true for this specific dataset. Furthermore, we note that correlations between multiple scRNA-seq datasets of the same cell type do not have extremely high per-gene correlations, such as one would observe when comparing bulk RNA-seq datasets from the same tissue type. To this point, current approaches for studying cell type transcriptomics are not sufficient, and would benefit from larger sample sizes. It is also important to note that the transcriptome is dynamic and responses to perturbations remain poorly understood. More resources will need to be invested in both bulk and single cell RNA-seq studies to further understand the landscape of the transcriptome.

4.5 Methods

Bulk RNA-Sequencing

We used FASTQC to visually inspect the read quality from the lower-coverage whole blood RNA- Seq (5.9M reads/sample), the moderate-coverage whole blood RNA-Seq (13.9M reads/ sample), and the high-coverage fibroblast RNA-Seq (50M reads/sample). We then used kallisto to pseudoalign reads to the GRCh37 gencode transcriptome (v33) and quantify estimates for transcript expression. We aggregated transcript counts to obtain gene level read counts using scripts from the GTEx consortium (https://github.com/broadinstitute/gtex-pipeline).

cis-eQTL mapping

Excluding related individuals (pi_hat > 0.2) from the analysis, we perform cis-eQTL analysis mapping using FastQTL [37], using a defined window of 1 Mb both up and downstream of every gene's TSS, for sufficiently expressed genes (TPM > 0.1 in 20% of individuals). We run the eQTL analysis in permutation pass mode (1000 permutations, and perform multiple testing corrections using the q value FDR procedure, correcting at 5% unless otherwise specified. We then restrict associations to the top (or leading) SNP per eGene.

Cohort Description

The samples included are from a study with individuals ascertained for bipolar disorder (BP). The cohort consists of 1,126 individuals with BP, 104 individuals with schizophrenia, and 766 controls (including first and second degree relatives of individuals with BP).

Genotyping pipeline

Genotypes for the low-coverage whole blood samples were obtained from the following platforms: OmniExpressExome (N = 810), PSYCH (N = 523), and COEX (N = 163). Given that the SNP-genotype data for both the fibroblast and whole blood samples came from numerous studies using various genotyping platforms (including GSA, Illumina550, OmniExpress Exome, COEX, and PsychChip) the number of overlapping SNPs across all platforms was < 80k, prompting us to perform imputation separately for each genotyping platform. Genotypes were first filtered for Hardy-Weinberg equilibrium p value < 1.0e-6 for controls and p value < 1.0e-10 for cases, with minor allele frequency (MAF) > 0.01, leaving 148613 typed SNPs.

Genotypes were imputed using the 1000 Genomes Project phase 3 reference panel[11] by chromosome using RICOPILI v.1 [12] separately per genotyping platform, then subsequently merged. Imputation quality was assessed by filtering variants where genotype probability > 0.8 and INFO score > 0.1, resulting in 2289732 autosomal SNPs. We restricted to only autosomal due to sex chromosome dosage, as commonly done[13].

Bulk RNA-Seq processing pipeline

We used FASTQC to visually inspect the read quality from the lower-coverage whole blood RNA-Seq (5.9M reads/sample) and the higher-coverage fibroblast RNA-Seq (13.9M reads/

sample). We then used kallisto [33] to pseudoalign reads to the GRCh37 transcriptome and quantify estimates for transcript expression. We aggregated transcript counts using scripts from the GTEX consortium (https://github.com/broadinstitute/gtex-pipeline).

Cell type proportion estimation

We estimate the proportion of cell types of both the lower coverage and higher coverage bulk whole blood RNA-seq datasets using CIBERSORTx, with batch correction applied and LM22 signature matrix as the reference gene expression profile. The LM22 signature matrix uses 547 genes to distinguish between 22 human hematopoietic cell phenotypes.

Complete blood counts were provided for a subset of the cohort, providing us ground truth measures for neutrophils, lymphocytes, monocytes, basophils, and eosinophils.

Cell type expression estimation

We log2-transform the matrix of bulk TPM measures and compute the first 50 principal components to be included as covariates. Using the cell type proportions derivedoutput from CIBERSORTx in conjunction with these log-transformed bulk expression measures, we use bMIND in order to derive cell type expression estimates, with flag np=TRUE. and 50 expression PCs included as covariates.

bMIND derived estimates

We use output from bMIND, we perform cis-eQTL analysis mapping using QTLtools, using a defined window of 1 Mb both up and downstream of every gene's TSS, for sufficiently expressed genes (TPM > 0.1 in 20% of individuals). We run the eQTL analysis in permutation pass mode (1000 permutations, and perform multiple testing corrections using the q value FDR procedure, correcting at 5% unless otherwise specified. We then restrict associations to the top (or leading) SNP per eGene.

Interaction model

To test whether there exists an interaction between SNP-lithium usage, we included an interaction component in the regression model, as such: $y = \beta *X + \beta*I +\beta*(X+I)$+covariates . We use MatrixEQTL [51] to implement this approach, using the “modelLINEAR_CROSS” setting. We used estimated counts from genes with at least 1.0 count in at least 40% of samples, including the first 25 expression PCs as covariates.

Estimating variance explained by gene expression

To estimate variance in case/control status attributed to variance in gene expression, we conduct a genomic REML analysis, found here (http://psoerensen.github.io/qgg/).

4.6 Acknowledgments

We thank the study subjects for their willingness to provide specimens and clinical data. We thank Arjun Bhattacharya, Yi Ding, Ella Petter, Mike Thompson, Bruna Balliu, Noah Zaitlen, Sasha Gusev, and Alkes Price for meaningful discussion in helping make this work possible. T.S. and T.B. are co-first-authors on this work and have both played a major role in study design and execution of analysis. T.B. created **Figure 4.1**, ran the bMIND software, and conducted extensive additional analyses that will be included in the preprint and article that is being written from this work. C.D. created **Figure 4.5** . T.S. was supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under award no. T32NS048004. T.B. was supported by the NIH (grant no. 5T32HG002536-19). This research was supported by the National Institute of Mental Health of the National Institutes of Health under award no. 5R01MH115676-04. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

4.7 Figures

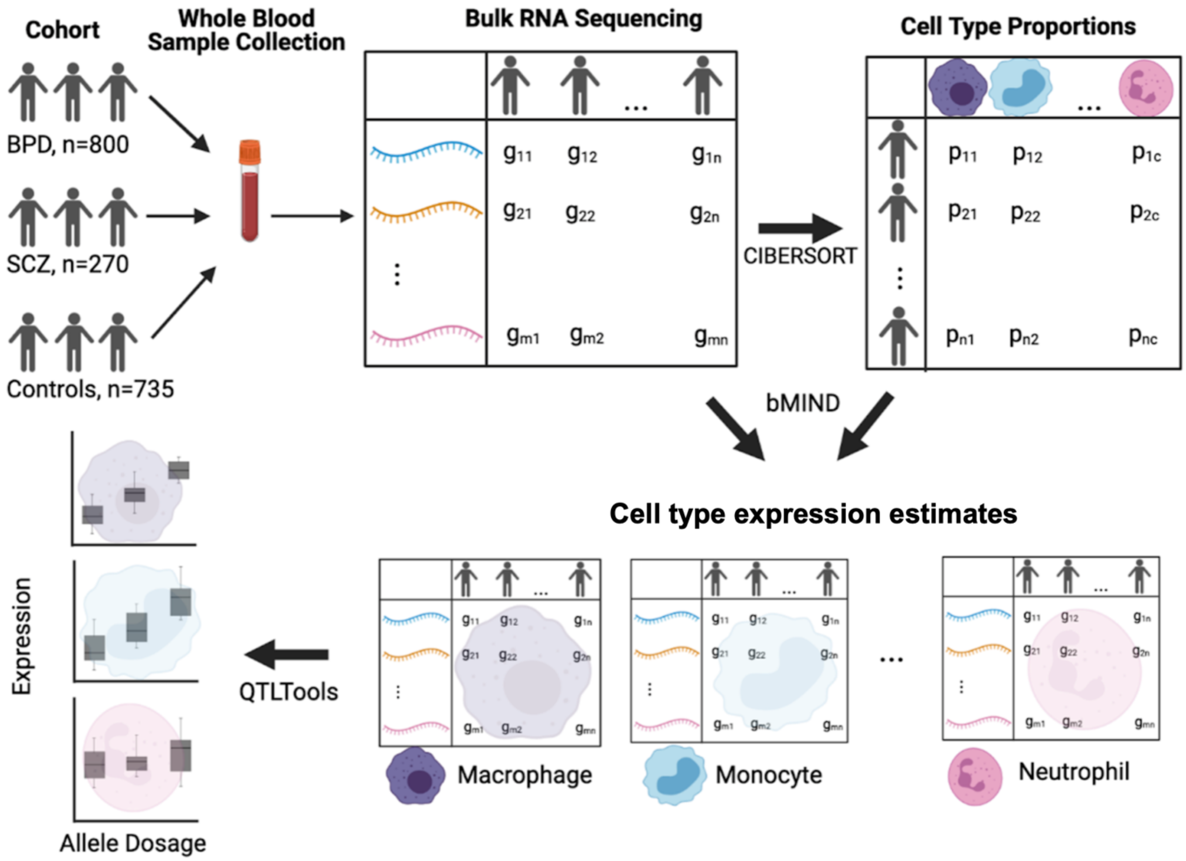


Figure 4.1: *Visual schematic of deconvolution pipeline*

This cohort is ascertained for individuals with bipolar disorder and has 800 individuals with BP, 270 individuals with SCZ, and 735 controls. RNA-seq was performed on whole blood tissue collected from each of the samples. CIBERSORTx was used to estimate cell type proportions for 22 cell types in whole blood tissues. Of the 22, the eight most abundant cell types were used for further analysis. Gene expression estimates and cell type proportion estimates were inputted into bMIND to obtain cell type expression estimates for the eight major cell types. cis-eQTL mapping was then performed separately for each of the eight cell types.

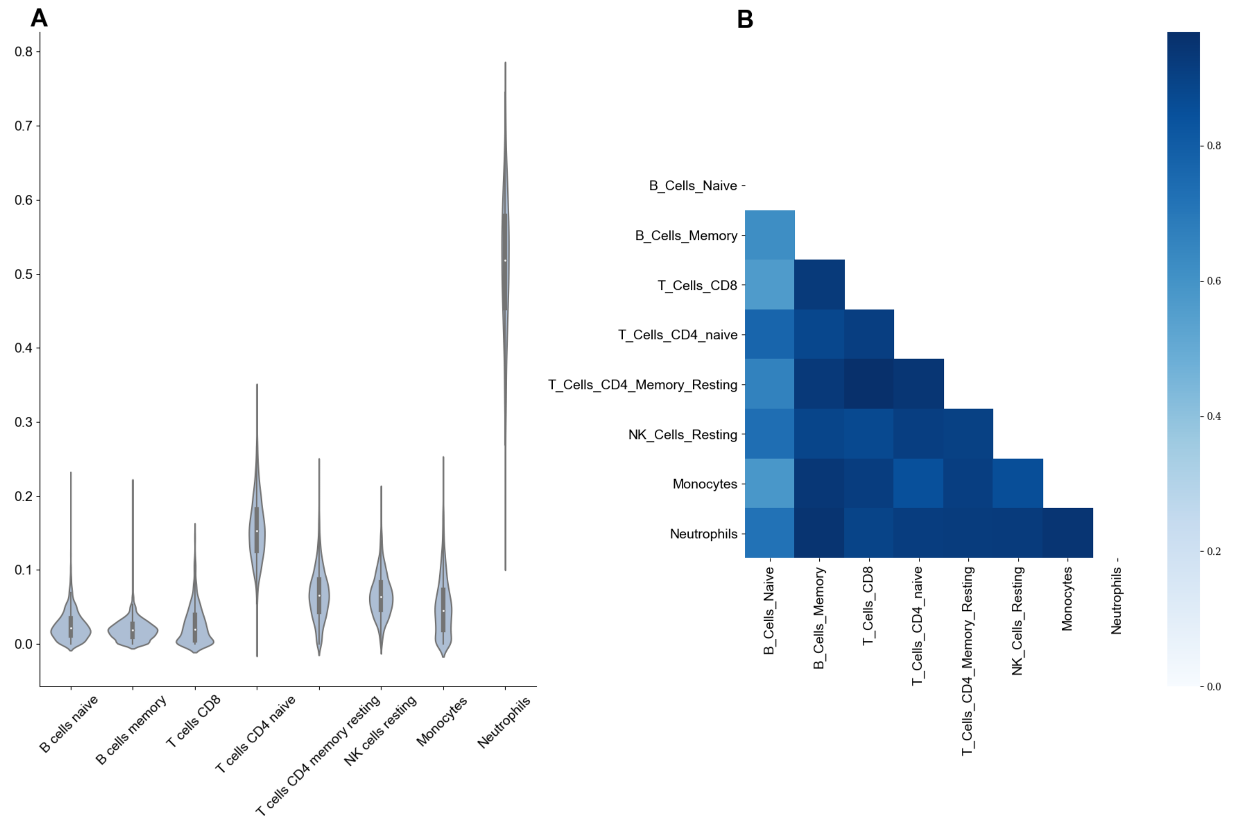


Figure 4.2: *Computationally estimated cell type expression*

(**4.2A**): Using CIBERSORTx, the predicted cell type proportion values of eight major cell types in whole blood tissue for all samples in the low-coverage RNA-seq cohort. (**4.2B**): R^2 values of computationally estimated expression, using bMIND, between all eight major cell types in whole blood tissue.

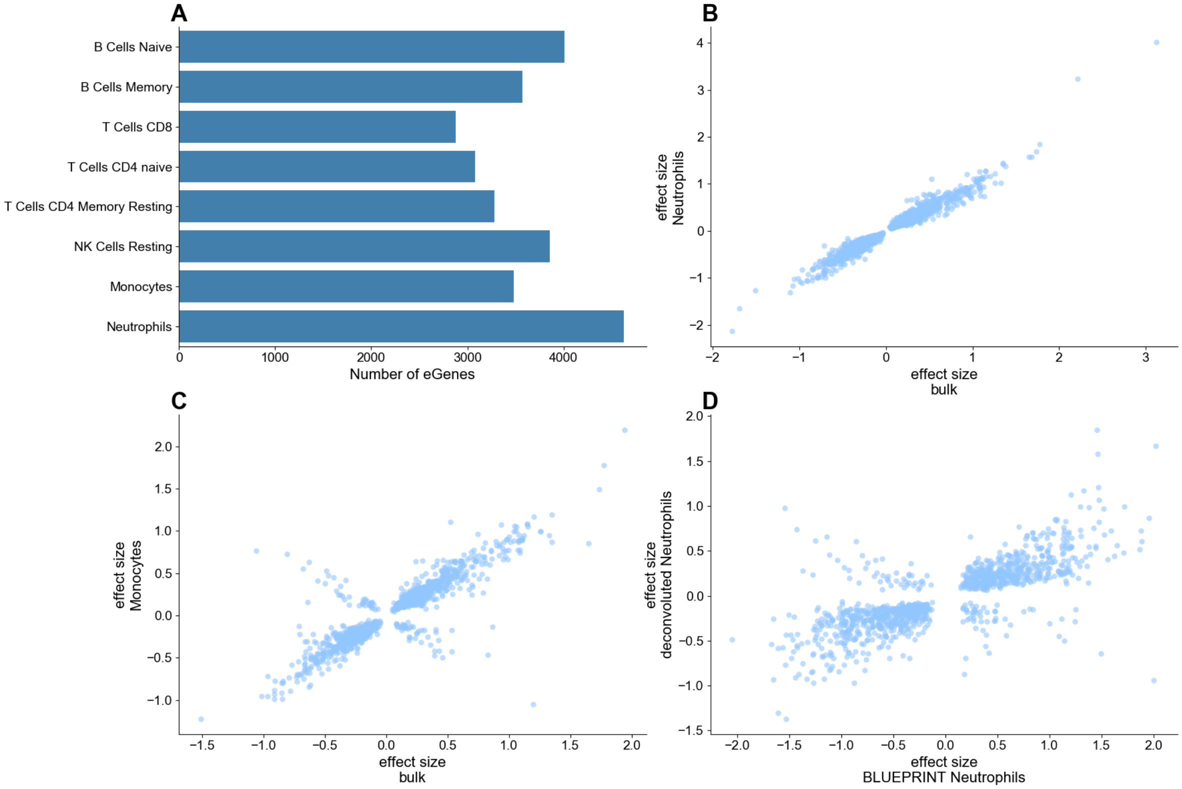


Figure 4.3: *Computationally estimated ct-expression and ct-eQTLs*

(**4.3A**): Using QTLtools, we conduct genome-wide ct-cis-eQTL scans and find between 2,875 and 4,629 ct-eGenes. (**4.3B**): Scatterplot of effect sizes of the leading eQTL hit for the 1,870 eGenes with a shared leading eSNP between the eQTL analysis using computationally deconvoluted neutrophil gene expression and bulk gene expression data. On the x axis, we show the effect size for the eGenes from eQTL analysis conducted using the bulk RNA-seq, on the y axis, we show the effect size for the eGenes from eQTL analysis conducted using computationally deconvoluted neutrophil gene expression. (**4.3C**): Scatterplot of effect sizes of the leading eQTL hit for the 1,291 eGenes with a shared leading eSNP between the eQTL analysis using computationally deconvoluted monocyte gene expression and bulk gene expression data. On the x axis, we show the effect size for the eGenes from eQTL analysis conducted using the bulk RNA-seq, on the y axis, we show the effect size for the eGenes from eQTL analysis conducted using computationally deconvoluted monocyte gene expression. (**4.4D**) Scatterplot of effect size comparisons between eQTLs identified using scRNA-seq reference data from the BLUEPRINT consortium compared with eQTLs identified using computationally deconvoluted data.

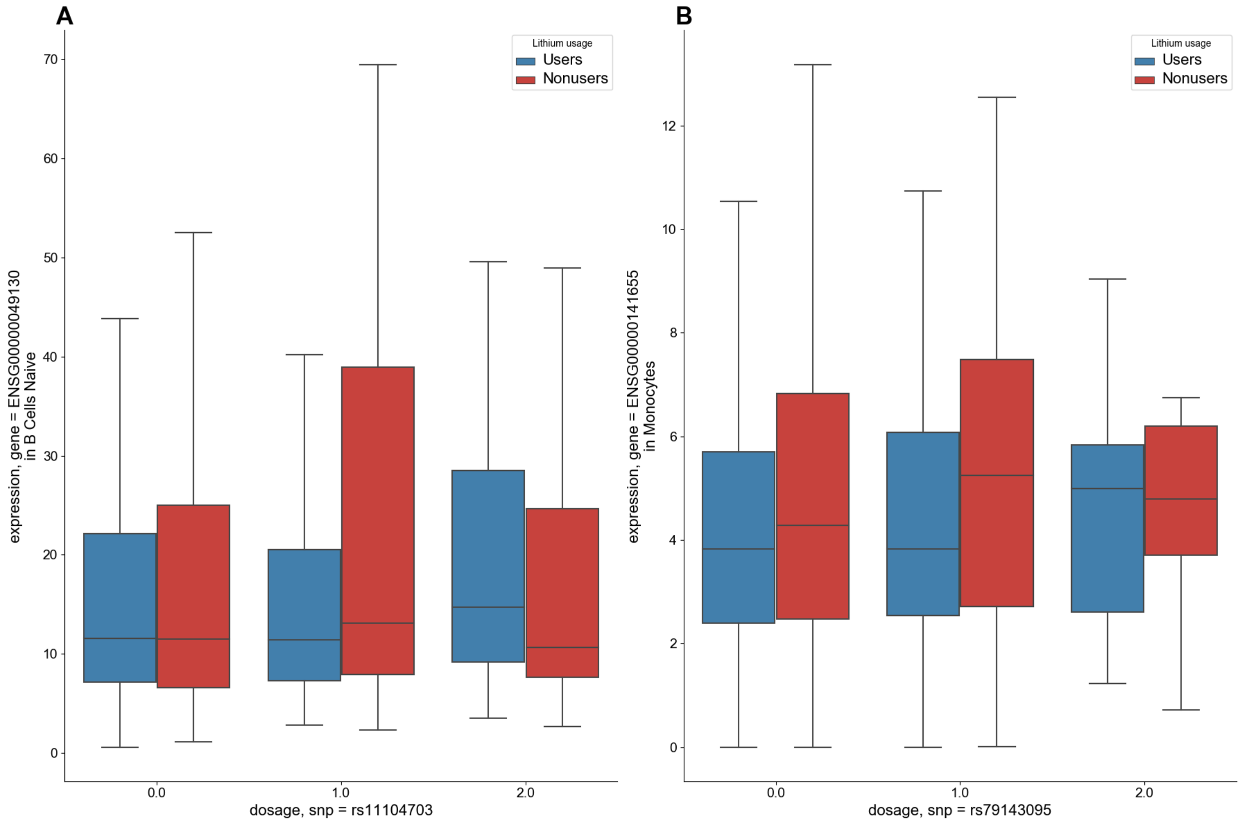


Figure 4.4: *SNP-lithium interaction eQTLs*

(**4.4A**): Boxplots showing the expression of KITLG (ENSG00000049130) in naïve B cells, stratified by dosage of SNP rs11104703 in lithium users versus nonusers. (**4.4B**): Boxplots showing the expression of TNFRSF11A (ENSG00000105641) in monocytes, stratified by dosage of SNP rs79143095 in lithium users versus nonusers.

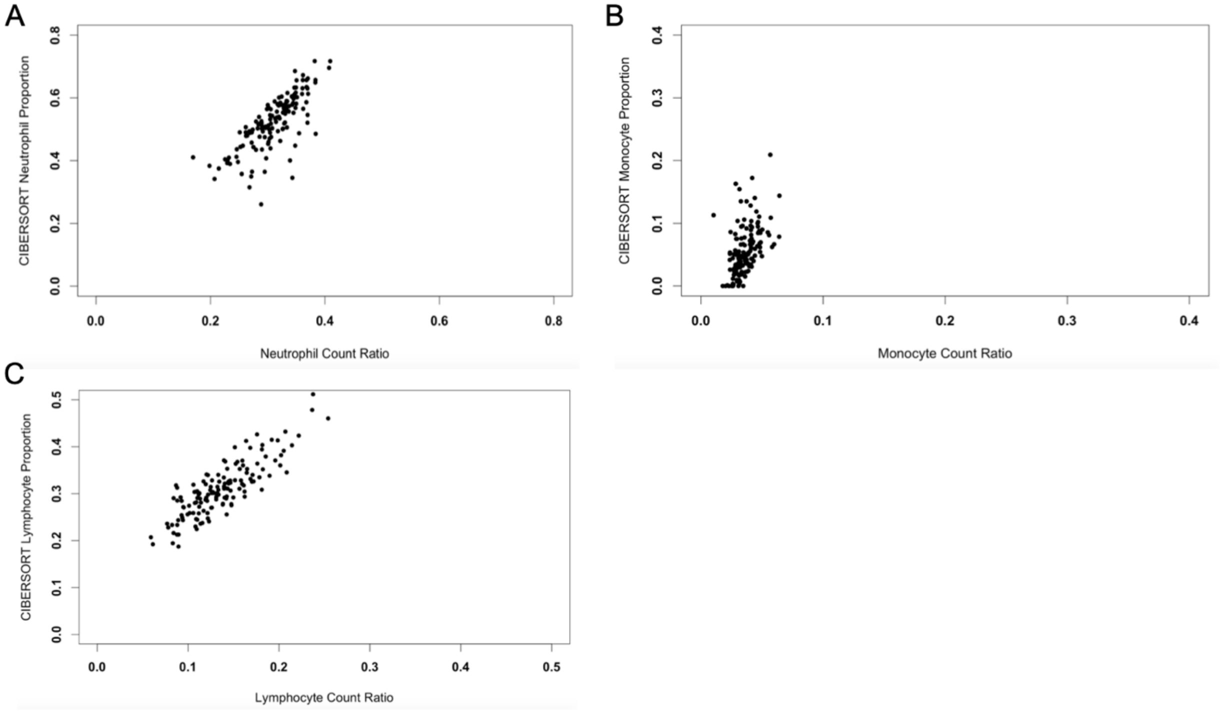


Figure 4.5: *Computationally estimated cell type proportions compared to laboratory measured cell type proportion values (N=143)*

(**4.5A**): On the x axis we show the laboratory measured cell type proportions for neutrophils, on the y axis we show the computationally estimated cell type proportions for neutrophils derived using CIBERSORTx. (**4.5B**): On the x axis we show the laboratory measured cell type proportions for monocytes, on the y axis we show the computationally estimated cell type proportions for monocytes derived using CIBERSORTx. (**S4.5C**): On the x axis we show the laboratory measured cell type proportions for lymphocytes, on the y axis we show the computationally estimated cell type proportions for lymphocytes derived using CIBERSORTx.

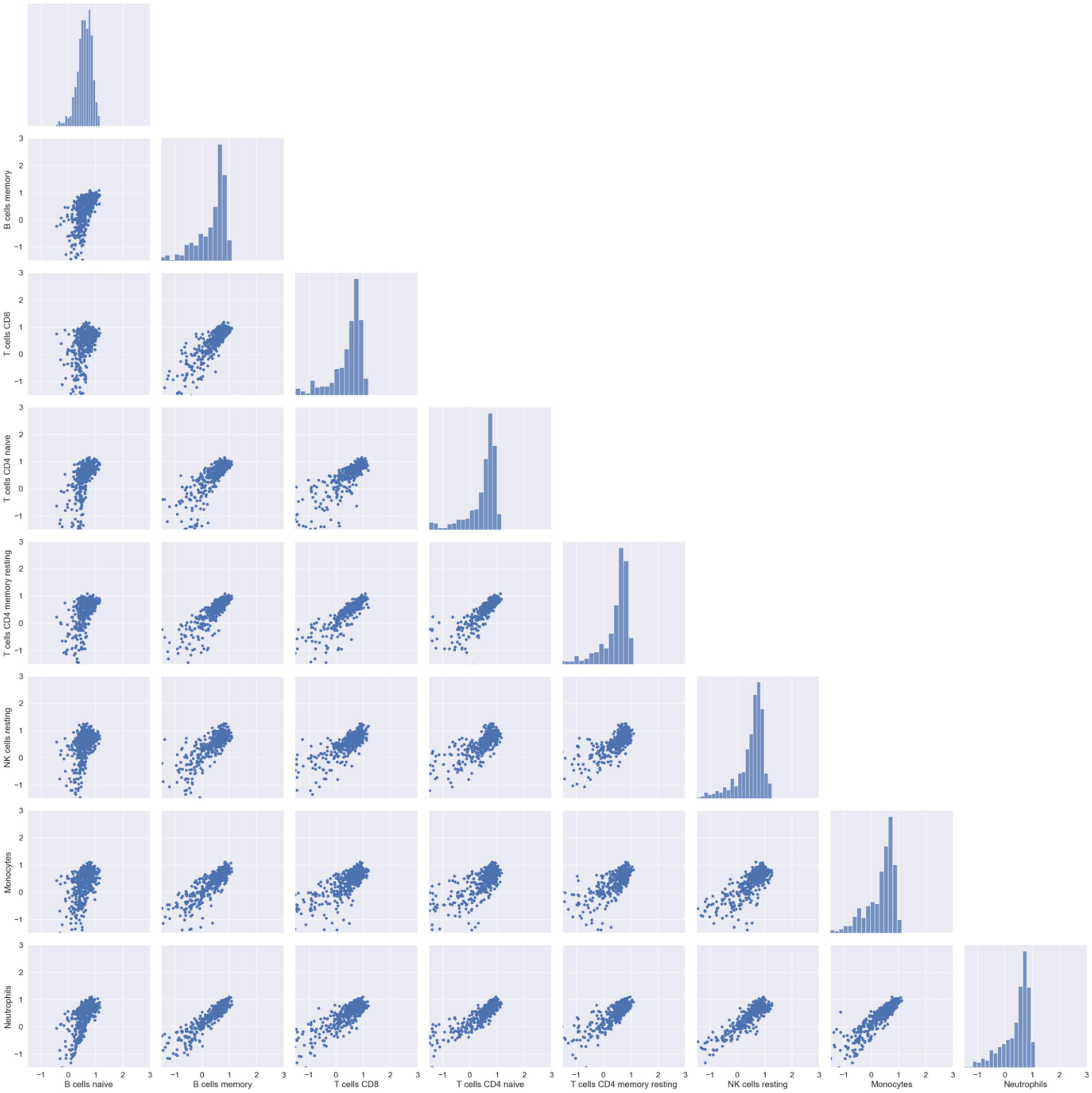


Figure 4.6: *Mean expression values of LM22 genes across all samples*

(**4.6**): Using the 548 genes included in the LM22 matrix, we compare mean expression values across samples in the eight major cell types.

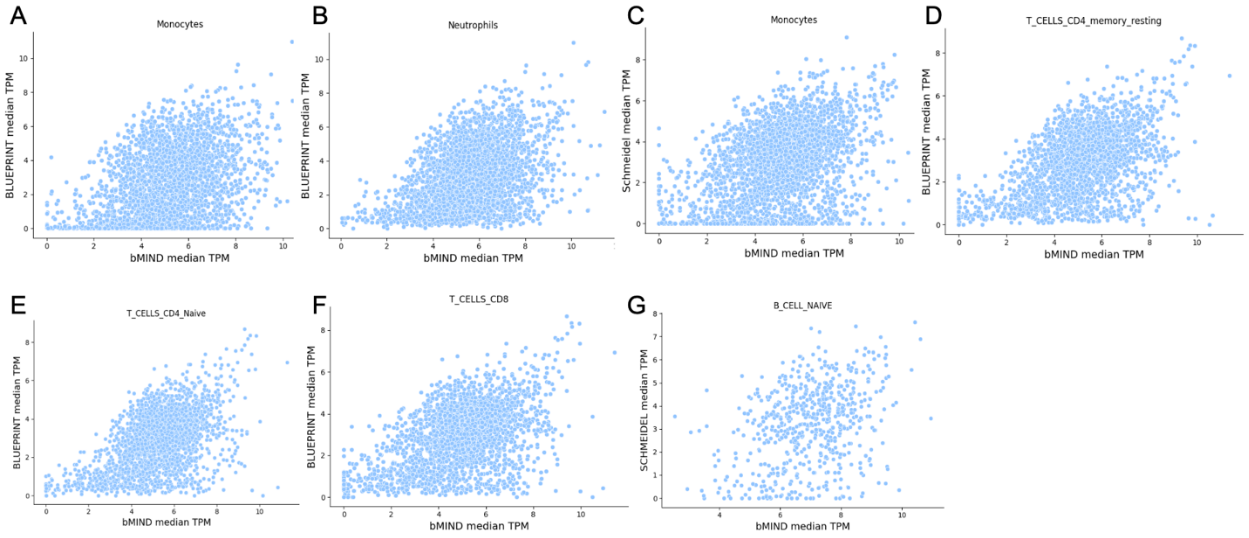


Figure 4.7: *Comparing computationally estimated cell type expression with scRNA-seq reference datasets*

(**4.7A**): Monocytes – x axis: bMIND estimated expression, y axis: BLUEPRINT reference.

(**4.7B**): Neutrophils – x axis: bMIND estimated expression, y axis: BLUEPRINT reference.

(**4.7C**): Monocytes – x axis: bMIND estimated expression, y axis: Schmeidel reference. (**4.7D**): Resting CD4 T cells – x axis: bMIND estimated expression, y axis: BLUEPRINT reference.

(**4.7E**): Naïve CD4 T cells – x axis: bMIND estimated expression, y axis: BLUEPRINT reference.

(**4.7F**): CD8 T cells – x axis: bMIND estimated expression, y axis: BLUEPRINT reference.

(**4.7G**): Naïve B cells – x axis: bMIND estimated expression, y axis: Schmeidel reference

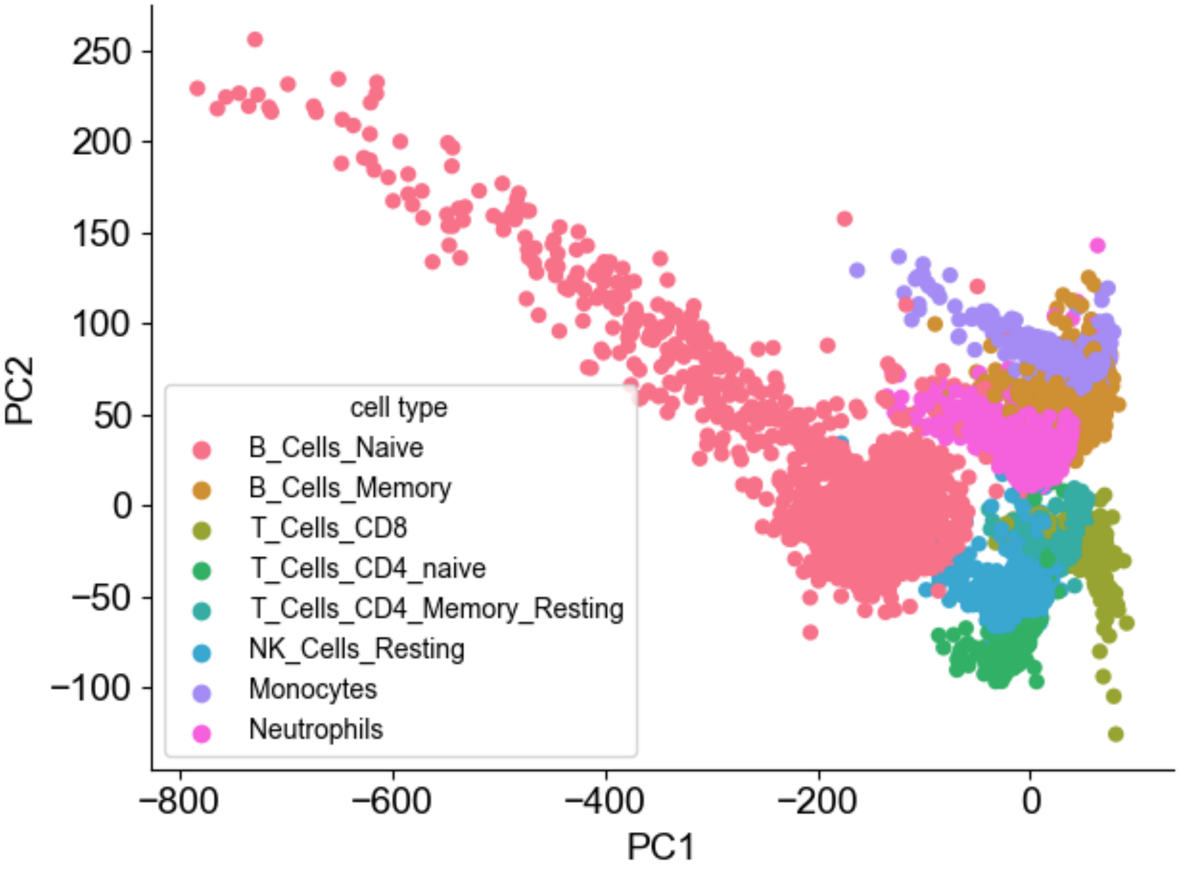


Figure 4.8: *PCA of computationally deconvoluted expression*

(**4.8**): Using computationally deconvoluted expression data of N = 1,996 individuals consisting of eight cell types from bMIND, PC1 and PC2 are plotted, with each point colored by cell type.

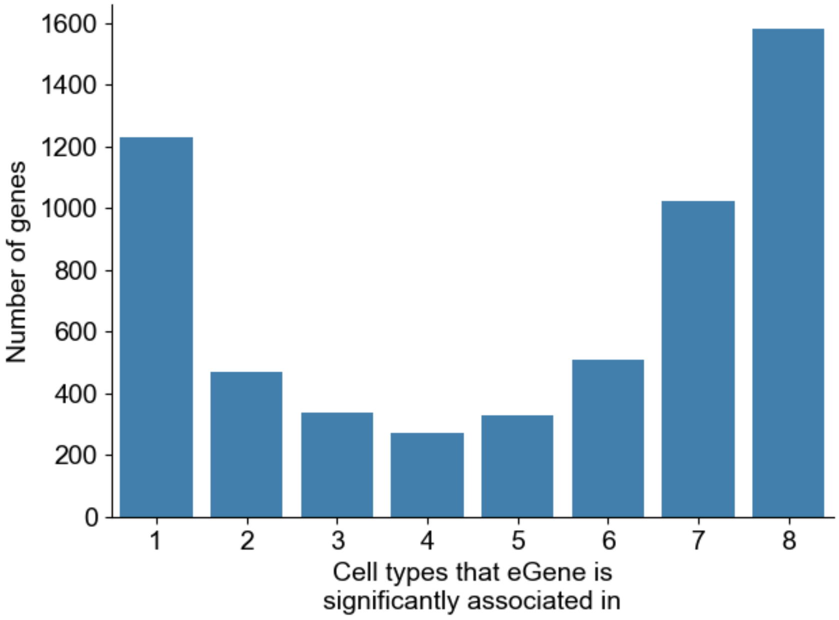


Figure 4.9: *Number of cell types eGenes are significant in*

(**4.9**): Of the 5,752 eGenes found in any of the eight primary whole blood cell types, we show how many cell types each is a significant ct-eGene in.

4.8 Supplementary Tables

Cell type	**Reference**	**R²**	**Number of genes**
Monocytes	BLUEPRINT	0.14	2896
Neutrophils	BLUEPRINT	0.48	3157
CD4 Memory T Cells	BLUEPRINT	0.26	2504
CD4 Naive T Cells	BLUEPRINT	0.27	2504
CD8 T Cells	BLUEPRINT	0.24	2504
B Cell Naive	Schmeidel	0.11	624
Monocytes	Schmeidel	0.15	2896

Monocytes *	Schmeidel/BLUEPRINT	0.22	2836

Supplementary Table 4.1: *Comparing computationally inferred ct-expression with scRNA-seq reference expression*

Restricting to the genes identified as eGenes using both the single cell RNA-Seq reference dataset and the computationally derived cell type expression, we report R2 values for the median TPM for genes across samples.

Cell type	**Number of eGenes**
Bulk	7302
Naive B Cells	4009
Memory B Cells	3571
Monocytes	3483
Neutrophils	4629
Resting NK Cells	3858
CD8 T Cells	3284
CD4 Memory T Cells	3284
CD4 Naive T Cells	3082

Supplementary Table 4.2: *Number of ct-eGenes per cell type*

Cell type	**GREML estimate**	**Variance**
Bulk	0.87	0.03

Naive B Cells	0.09	0.04
Memory B Cells	4.2e-9	0.0
Monocytes	0.40	0.06
Neutrophils	0.95	0.08
Resting NK Cells	0.24	0.06
CD8 T Cells	0.12	0.04
CD4 Memory T Cells	0.14	0.05
CD4 Naive T Cells	0.56	0.07

Supplementary Table 4.3: *GREML estimates*

Cell type	Number of Li-eGenes	Number of “same-direction” Li-eGenes	Number of “opposite-direction” Li-eGenes
Naive B cells	24	4	20
Memory B cells	15	3	12
CD8 T cells	2	1	1
Naive CD4 T cells	25	1	24
Memory CD4 T cells	2	0	2
Resting NK cells	5	0	5
Monocytes	34	3	31
Neutrophils	3	1	2

Supplementary Table 4.4: *SNP-Lithium interaction results*

Using an FDR cut-off of $p < 0.10$, we look at the number of eGenes with a significant SNP-lithium interaction. “Same-direction” Li-eGenes have the same direction of effect sizes between lithium users and nonusers, and “opposite-direction” Li-eGenes have the opposite direction of effect sizes between lithium users and nonusers.

Chapter 5: Concluding remarks

Linking genetic variation to risk for complex traits remains an important challenge in human genetics. GWAS has allowed us to find many risk regions for various complex traits, but results from these studies remain poorly understood, as noncoding regions account for much of the risk derived from GWAS. Noncoding regions of the genome are difficult to link to function, as the proteins they may be linked to are not as obvious as regions that lie in coding regions. One hypothesis is that risk for complex disease may be mediated through gene expression. We can link gene expression to genetic variation by conducting eQTL mapping. While the cost of genotyping has plummeted, enabling very large-scale genetics studies, measuring gene expression accurately remains cost prohibitive. The state-of-the-art method for doing so is RNA sequencing (RNA-seq). In this work, we provide a comprehensive investigation, using synthetic datasets, into the amount of gene expression signal we expect to obtain as we manipulate coverage in RNA-seq. We show that in these synthetic datasets, it is possible to significantly reduce coverage while still maintaining high accuracy in gene expression estimates. Though these results were promising in simulations, low-coverage RNA-seq data is not commonly used in practice, so possible shortcomings due to technical errors are largely unexplored. To address this, we design and implement a large-scale RNA-seq experiment at low coverage (5.9M reads/ sample, on average). We investigate whether eQTLs found using this novel approach are consistent with those that we find with traditional approaches to RNA-seq experiment design.

In Chapter 2, we conduct rigorous simulations via downsampling high-coverage RNA-seq data into various synthetic datasets, to show how manipulating coverage impacts our ability to quantify gene expression at various coverages. We create multiple synthetic datasets at each coverage level to account for variability in sampling and find that it is still possible to accurately quantify gene expression at lower levels of coverage than are typically used. For example, using

10% of the reads, we retain an R^2 of 0.40, or quantify gene expression with 40% accuracy, on average. We go on to show how different features of genes, including mean abundance, whether or not they are protein-coding, GC content, number of transcripts, among other features, impact quantification accuracy at lower coverage levels. Finally, we provide a formula to estimate an experiment's effective sample size, and provide a webtool to enable scientists to more easily explore the tradeoff between coverage and number of samples under a fixed budget.

In Chapter 3 we dive into exploring low-coverage RNA-seq and its implications in real data. We generate a large-scale, low-coverage sequencing dataset, including 1490 samples derived from whole blood tissue, with RNA-seq at an average sequencing depth of 5.9M reads/sample. We compare the results from low-coverage RNA-seq with two datasets of higher coverage, to ensure accuracy. We use an RNA-seq dataset including 570 samples of whole blood tissue sequencing at a mean sequencing depth of 13.9M reads/sample and use RNA-seq data from the GTEx consortium, including 670 samples derived from whole blood tissue. We compare estimated gene expression levels between low-coverage RNA-seq with moderate-coverage RNA-seq, and find a very high concordance of mean gene expression across datasets, suggesting that low-coverage RNA-seq is broadly effective in capturing gene expression. We conduct cis-eQTL mapping for each of the three datasets, and find that the resulting effect sizes show strong concordance between the eQTL analysis done using low-coverage RNA-seq compared to both moderate-coverage and high-coverage RNA-seq. Additionally, we show that despite having considerably fewer reads in total, the low-coverage RNA-seq approach captures a similar number of associations to GTEx. Finally, we demonstrate that the associations found using low-coverage RNA-seq are robust and colocalize with GWAS loci for several blood traits.

In Chapter 4, we leverage computational deconvolution methods to build a pipeline to estimate cell type expression in eight major cell types in whole blood tissue. We find that the expression estimates generated from this approach are consistent with those observed using sc-RNA-seq. Furthermore, we find that we have increased power to identify sc-eQTLs with this approach due to the large sample sizes obtainable in bulk RNA-seq experiments. We incorporate an interaction term in the eQTL model to investigate whether there exist SNP-drug interactions with respect to lithium usage. We find several examples of genes, both at the bulk and cell type level, that are differentially associated with nearby SNPs, dependent on lithium usage. We show that this approach can be coupled with

Both the low-coverage RNA-seq and GTEx consortium RNA-seq are well-powered to detect many eQTL associations. However, each dataset discovers additional associations that do not meet that statistical threshold cutoffs in the other, and neither is as well powered to detect associations as eQTLGen, a meta-analysis of many projects consisting of microarray data and RNA-seq, containing >30,000 samples in total. This leads us to believe that there is still much work to be done to characterize the full profile of transcriptional regulation through genetics by investing more resources in such experiments, especially in recruiting cohorts containing more genetic ancestral diversity. We also note that there is a temporal component to gene expression that is left largely unexplored. Our work will enable researchers to better design experiments without expending additional resources.

We provide the gene expression estimates from the low-coverage RNA-seq dataset generated in Chapter 3 as a publicly available resource (accession number phs002856.v1).

REFERENCES

1. Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E.T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature *464*, 773–777.

2. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768–772.

3. Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., and Pasaniuc, B. (2017). Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. Am. J. Hum. Genet. *100*, 473–487.

4. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48,* 245–252.

5. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., GTEx Consortium, Nicolae, D.L., Cox, N.J., and Im, H.K. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091– 1098.

6. Zhernakova, D.V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., van ’t Hof, P., Mei, H., van Dijk, F., Westra, H.J., et al. (2017). Identification of context-dependent expression quantitative trait loci in whole blood. Nat. Genet. *49*, 139–145.

7. Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

8. Rockman, M.V., and Kruglyak, L. (2006). Genetics of global gene expression. Nat. Rev. Genet. *7*, 862–872.

9. Gallagher, M.D., and Chen-Plotkin, A.S. (2018). The post- GWAS era: from association to function. Am. J. Hum. Genet. *102*, 717–730.

10. Vosa, U., Claringbould, A., Westra, H.J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., et al. (2021). Large-scale cis- and trans-eQTL analyses identify thou- sands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. *53*, 1300–1310. https://doi.org/ 10.1038/s41588-021-00913-z.

11. Hoffman, G.E., Bendl, J., Voloudakis, G., Montgomery, K.S., Sloofman, L., Wang, Y.C., Shah, H.R., Hauberg, M.E., Johnson, J.S., Girdhar, K., et al. (2019). CommonMind consortium pro- vides transcriptomic and epigenomic data for schizophrenia and bipolar disorder. Sci. Data *6*, 180.

12. Franzen, O., Ermel, R., Cohain, A., Akers, N.K., Di Narzo, A., Talukdar, H.A., Foroughi-Asl, H., Giambartolomei, C., Fullard, J.F., Sukhavasi, K., et al. (2016). Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. Science *353*, 827–830.

13. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.

14. Lepik, K., Annilo, T., Kukuskina, V., eQTLGen Consortium, Kisand, K., Kutalik, Z., Peterson, P., and Peterson, H. (2017). C-reactive protein upregulates the whole blood expression of CD59 - an integrative analysis. PLoS Comput. Biol. *13*, e1005766. https://doi.org/10.1371/journal.pcbi.1005766.

15. Pasaniuc,B.,Rohland,N.,McLaren,P.J.,Garimella,K.,Zaitlen, N., Li, H., Gupta, N., Neale, B.M., Daly, M.J., Sklar, P., et al. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat. Genet. *44*, 631–635.

16. Sims, D., Sudbery, I., Ilott, N.E., Heger, A., and Ponting, C.P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. *15*, 121–132.

17. CONVERGE consortium (2015). Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature *523*, 588–591.

18. Homburger, J.R., Neben, C.L., Mishne, G., Zhou, A.Y., Kathir- esan, S., and Khera, A.V. (2019). Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. Genome Med. *11*, 74.

19. Rubinacci, S., Ribeiro, D.M., Hofmeister, R.J., and Delaneau, O. (2021). Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat. Genet. *53*,412.

20. Baccarella, A., Williams, C.R., Parrish, J.Z., and Kim, C.C. (2018). Empirical assessment of the impact of sample number and read depth on RNA-Seq analysis workflow performance. BMC Bioinf. *19*, 423.

21. Pritchard, J.K., and Przeworski, M. (2001). Linkage disequilib- rium in humans: models and data. Am. J. Hum. Genet. *69*, 1–14.

22. Tarazona, S., Garc´ıa-Alcalde, F., Dopazo, J., Ferrer, A., and Con- esa, A. (2011). Differential expression in RNA-seq: a matter of depth. Genome Res. *21*, 2213–2223.

23. Robinson, D.G., and Storey, J.D. (2014). subSeq: determining appropriate sequencing depth through efficient read subsam- pling. Bioinformatics *30*, 3424–3426.

24. Williams, A.G., Thomas, S., Wyman, S.K., and Holloway, A.K. (2014). RNA-seq data: challenges in and recommendations for experimental design and analysis. Curr. Protoc. Hum. Genet. *83*, 11.13.1–20.

25. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. *10*, e1004383. https://doi.org/10.1371/journal.pgen.1004383.

26. Lloyd-Jones, L.R., Holloway, A., McRae, A., Yang, J., Small, K., Zhao, J., Zeng, B., Bakshi, A., Metspalu, A., Dermitzakis, M., et al. (2017). The genetic architecture of gene expression in pe- ripheral blood. Am. J. Hum. Genet. *100*, 228–237. https://doi. org/10.1016/j.ajhg.2016.12.008.

27. GTEx Consortium (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213. https://doi.org/ 10.1038/nature24277.

28. 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. https://doi. org/ 10.1038/s41588-020-0625-2.

29. Mu, Z., Wei, W., Fair, B., Miao, J., Zhu, P., and Li, Y.I. (2021). The impact of cell type and context-dependent regulatory var- iants on human immune traits. Genome Biol. *22*, 122. https:// doi.org/10.1186/s13059-021-02334-x.

30. Lachmann, A., Torre, D., Keenan, A.B., Jagodnik, K.M., Lee, H.J., Wang, L., Silverstein, M.C., and Ma'ayan, A. (2018). Massive mining of publicly available RNA-seq data from hu- man and mouse. Nat. Commun. *9*, 1366. https://doi.org/10. 1038/s41467-018-03751-6.

31. Krebs, C.E., Ori, A.P.S., Vreeker, A., Wu, T., Cantor, R.M., Boks, M.P.M., Kahn, R.S., Olde Loohuis, L.M., and Ophoff, R.A. (2020). Whole blood transcriptome analysis in bipolar disorder reveals strong lithium effect. Psychol. Med. *50*, 2575– 2586.

32. Frazee, A.C., Jaffe, A.E., Langmead, B., and Leek, J.T. (2015). Polyester: simulating RNA-seq datasets with differential transcript expression. Bioinformatics *31*, 2778–2784.

33. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

34. Liu, Y., Ferguson, J.F., Xue, C., Silverman, I.M., Gregory, B., Re- illy, M.P., and Li, M. (2013). Evaluating the impact of sequencing depth on transcriptome profiling in human adipose. PLoS One *8*, e66883.

35. Mandric, I., Schwarz, T., Majumdar, A., Hou, K., Briscoe, L., Perez, R., Subramaniam, M., Hafemeister, C., Satija, R., Ye, C.J., et al. (2020). Optimized design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. Nat. Commun. *11*, 5504.

36. Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T., and Dela- neau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics *32*, 1479–1485.

37. van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Ver- weij, N., Sehmi, J., Paul, D.S., Elling, U., Allayee, H., Li, X., et al. (2012). Seventy-five genetic loci influencing the human red blood cell. Nature *492*, 369–375.

38. Bentham, J., Morris, D.L., Graham, D.S.C., Pinder, C.L., Tombleson, P., Behrens, T.W., Mart ´ın, J., Fairfax, B.P., Knight, J.C., Chen, L., et al. (2015). Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat. Genet. *47*, 1457–1464.

39. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gi- lad, Y. (2008). RNA-seq: an assessment of technical reproduc- ibility and comparison with gene expression arrays. Genome Res. *18*, 1509–1517.

40. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

41. The 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

42. Lam, M., Awasthi, S., Watson, H.J., Goldstein, J., Panagiotaro- poulou, G., Trubetskoy, V., Karlsson, R., Frei, O., Fan, C.C., De Witte, W., et al. (2019). RICOPILI: rapid imputation for COnsortias PIpeLIne. Bioinformatics *36*, 930–933.

43. Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaud- huri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat. Biotechnol. *37*, 773–782.

44. Aguet, F., and Munoz Aguirre, M. (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

45. Daley, T., and Smith, A.D. (2013). Predicting the molecular complexity of sequencing libraries. Nat. Methods *10*, 325–327. https://doi.org/10.1038/nmeth.2375.

46. Schwarz T., Boltz T., Hou K., Bot M., Duan C., Olde Loohuis L.M., Boks M.P., Kahn R.S., Ophoff R.A. & Pasaniuc B. (2022) Powerful eQTL mapping through low coverage RNA Sequencing. *HGG Advances https://doi.org/10.1016/j.xhgg.2022.100103*

47. Cobos, F. A., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature Communications* vol. 11 (2020).

48. Perez RK, Gordon MG, Subramaniam M, Kim MC, Hartoularos GC, Targ S, Sun Y, Ogorodnikov A, Bueno R, Lu A, Thompson M, et al. Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science*. 2022

49. Wang, J., Roeder, K. & Devlin, B. Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data. *Genome Res.* (2021) doi:10.1101/gr.268722.120.

50. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).

51. Shabalin, A.A. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28, no. 10 (2012): 1353-1358.

52. Schmiedel, B. J. *et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* vol. 175 1701–1715.e16 (2018).

53. Jin, H., Liu, Z. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *Genome Biol* **22,** 102 (2021). https://doi.org/10.1186/s13059-021-02290-6

54. Haque, A., Engel, J., Teichmann, S.A. *et al.* A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med* **9,** 75 (2017). https://doi.org/10.1186/s13073-017-0467-4

55. Patel, D., Zhang, X., Farrell, J.J. *et al.* Cell-type-specific expression quantitative trait loci associated with Alzheimer disease in blood and brain tissue. *Transl Psychiatry* **11,** 250 (2021). https://doi.org/10.1038/s41398-021-01373-z

56. Neavin, D., Nguyen, Q., Daniszewski, M.S. *et al.* Single cell eQTL analysis identifies cell type-specific genetic control of gene expression in fibroblasts and reprogrammed induced pluripotent stem cells. *Genome Biol* **22,** 76 (2021). https://doi.org/10.1186/s13059-021-02293-3

57. Ding J, Gudjonsson JE, Liang L, Stuart PE, Li Y, Chen W, et al. Gene expression in skin and lymphoblastoid cells: refined statistical method reveals extensive overlap in cis-eQTL signals. Am J Hum Genet. 2010;87:779–89.

58. Nguyen QH, Lukowski SW, Chiu HS, Senabouth A, Bruxner TJC, Christ AN, et al. Single-cell RNA-seq of human induced pluripotent stem cells reveals cellular heterogeneity and cell state transitions between subpopulations. Genome Res. 2018;28:1053–66.

59. Krebs, C., Ori, A., Vreeker, A., Wu, T., Cantor, R., Boks, M., . . . Ophoff, R. (2020). Whole blood transcriptome analysis in bipolar disorder reveals strong lithium effect. *Psychological Medicine*, 50(15), 2575-2586. doi:10.1017/S0033291719002745

60. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A., Mikheenko, A., … Phillippy, A. (2022) The complete sequence of a human genome. *Science*, 376(6588), 44-53. doi:10.1126/science.abj6987

61. Collins FS, Fink L. The Human Genome Project. *Alcohol Health Res World.* 1995;19(3):190-195. PMID: 31798046; PMCID: PMC6875757.

62. Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. From DNA to RNA. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26887/

63. Fu W, O'Connor TD, Akey JM. Genetic architecture of quantitative traits and complex diseases. *Curr Opin Genet Dev*. 2013 Dec;23(6):678-83. doi: 10.1016/j.gde.2013.10.008. Epub 2013 Nov 26. PMID: 24287334; PMCID: PMC6764439.

64. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet.* 2003 Mar;33 Suppl:228-37. doi: 10.1038/ng1090. PMID: 12610532.

65. Mackay TF. The genetic architecture of quantitative traits. *Annu Rev Genet.* 2001;35:303-39. doi: 10.1146/annurev.genet.35.102401.090633. PMID: 11700286.

66. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH: Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 2010, 11:446–450.

67. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL, et al.: Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* 1989, 245:1066–1073.

68. Drumm ML, Ziady AG, Davis PB: Genetic variation and clinical heterogeneity in cystic fibrosis. *Annu Rev Pathol* 2012, 7:267–282.

69. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 2009, 106:9362–9367.

70. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al.: Finding the missing heritability of complex diseases. *Nature*2009, 461:747–753.

71. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, Suveges D, Vrousgou O, Whetzel PL, Amode R, Guillen JA, Riat HS, Trevanion SJ, Hall P, Junkins H, Flicek P, Burdett T, Hindorff LA, Cunningham F and Parkinson H. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 2019, Vol. 47 (Database issue): D1005-D1012.

72. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016 Jan 26;17:13. doi: 10.1186/s13059-016-0881-8.