THE UNIVERSITY OF CHICAGO

STATISTICAL METHODS FOR TRANSCRIPTOME DATA

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY

TAE HYUN KIM

CHICAGO, ILLINOIS
AUGUST 2020

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTARY FIGURES

LIST OF SUPPLEMENTARY TABLES

ACKNOWLEDGMENTS

I thank my advisors for their guidance in research without which I could not have produced this work. Dan Nicolae has been endlessly thoughtful and patient in guiding me, not just in statistical research but in academic and personal growth; his keen intuition in statistics has contributed greatly to my research, and I am more than grateful. I also feel very fortunate to have joined Mengjie Chen's group. She has, by example, given me a role model as a scholar through her creative approach in research as well as her resilience and positive attitude. I also thank Mary Sara McPeek for her valuable insights and generosity with her time and guidance. I thank Haky Im and other faculty members in the Statistics and Genetic Medicine department for providing wonderful resources for my research.

I have made many great friends along the way who helped me get through graduate school. I am thankful to all of them, especially Minju Kim who has become my confidant and best friend. Her self-discipline continues to inspire me, and I am grateful for her friendship. I also thank Rockefeller Chapel Choir through which I learned what a supportive community looks like at its best. The time I spent in the chapel will have a place in my heart wherever I go. I genuinely thank all FitChicago instructors, especially Cruz, for keeping me healthy both physically and mentally.

I thank Junyoung Park for his unconditional love and support. I am grateful to have a partner with such dedication and compassion. I thank my parents Hyosun and Kyungshik Kim for prioritizing their children's education over so many other things, for consistently encouraging them to dream big, and for being both life mentors and best friends. I also thank my brother Kangpyo Kim for teaching me math when all other teachers failed; none of my achievements would have been possible without him. Lastly, I send my love to my nephew Chanyoon and niece Jaewon who have been giving me a renewed motivation for good work.

ABSTRACT

Transcriptome data provides key information about molecular mechanism for phenotypic diversity. Advances in technology have made transcriptome data available in improved quality and quantity, calling for new statistical methods that can account for large data size, complex dependence structure, and technical artifacts. This dissertation proposes statistical methods that tackle those challenges in transcriptome data analysis while addressing important biological questions. Chapters 2, 3, and 4 focus on the analysis of African American gene expression levels. African American samples' genetic ancestry is investigated regarding its relationship with their gene expression levels. Chapter 5 introduces a statistical method for data sets that became more recently available — single cell transcriptome. Chapter 5 provides a comprehensive analysis tool for UMI scRNA-seq data by modeling the noise structure. Although the proposed methods are developed for the specific purpose of analyzing gene expression level data, some of them can be potentially applied to diverse fields. For example, Chapter 2 develops a multivariate Bayesian variable selection tool that can account for data sets with random missing values. Chapter 4 develops a covariance analysis tool that expands traditional heteroskedasticity analysis to dynamically varying covariance analysis. These methods focus on accounting for inter-tissue or inter-gene correlations, so they can be applied to correlated data from other fields. Most of the methods have been implemented as open-source softwares to promote their applicability. We believe these methods contribute not only to future research in molecular biology but also to the field of large and complex, modern data analysis.

CHAPTER 1

INTRODUCTION

If all cells of an organism have the same genetic material, why do they vary in their types and roles? This phenotypic diversity exists because, for each cell, different sets of genes are activated and different sets of proteins are synthesized, leading to different cellular fates (Morozova et al., 2009). Gene expression levels, a measure for gene activation and transcription, explain the heterogeneity in traits that the genome cannot explain. Hence, analyzing gene expression levels can help answer many important biological questions about when and why organisms show different patterns of traits. Gene expression levels can be measured through transcriptome data by quantifying the abundance of transcripts from each gene.

Novel sequencing technology, such as RNA sequencing (RNA-seq) and single-cell RNA sequencing (scRNA-seq), made transcriptome data available in improved quantity and quality, opening up possibilities for understanding molecular biology in a more granular level. As a response to the rise of these data sets, this work develops statistical methods that can analyze gene expression levels in depth. Chapters 2, 3, and 4 study the relationship between gene expression levels and genetic ancestry in African American samples. Chapter 5 develops a statistical tool for processing 10X scRNA-seq that suffers from noise due to low sequencing depth. The following two sections respectively introduce the motivation and the data sets for the two topics.

1.1 African American Transcriptome Analysis

1.1.1 Motivation

Studies in the past have shown that many phenotypic traits vary with the proportion of African ancestry in the genome of African Americans (Fyr et al., 2007; Nalls et al., 2008; Reiner et al., 2007). Such results motivate us to investigate the molecular mechanism of how

ancestry is correlated with those difference in traits. The first three chapters address the problem by studying the relationship between ancestry and gene expression levels.

Population geneticists have strived to understand the genetic and phenotypic variations across different populations, and many of them have depended on the self-reported race to identify the population structure. However, race is a complicated concept of social construct and does not necessarily reflect the genetic ancestry of individuals. Moreover, most phenotypes are not only affected by genetic difference but also by non-genetic, environmental factors, and these confounders cannot be handled well if the population structure is treated as categorical as in the self-reported race. Admixed populations, therefore, hold valuable information; they have diverse genetic profiles, but they are from a relatively homogeneous environment (Winkler et al., 2010a; Seldin et al., 2011).

Figure 1.1 illustrates how population admixture occurs. When separate and isolated populations (light green and dark green) begin to interbreed, recombinations would occur, and their offspring after several generations would have mosaic patterns of chromosomes where each segments are from different ancestries (Winkler et al., 2010b). These patterns of admixture can be inferred by existing statistical tools. Inferred admixture patterns allow us to map genetic ancestry into a continuous space rather than into a category. For example, someone could be 80% African and 20% European as opposed to 70% African and 30% European, a difference that would not have been easily noticed without the data and methodology for accurate inference of ancestry. In other words, this measure of ancestry allows the investigation of genetic and phenotypic variability within people who self-identify as the same race. Moreover, ancestry can be defined locally. Different locations in a chromosome hold different ancestral information as illustrated in Figure 1.2. Local ancestry is counted as the number of chromosomes — 0, 1, or 2 — that are from a certain ancestry. Local ancestry allows more granular investigation of the population differences in genomics and transcriptomics. For

example, Europeans and Africans could have developed different variants of a transcription factor which can lead to population differences in the local gene regulation process.

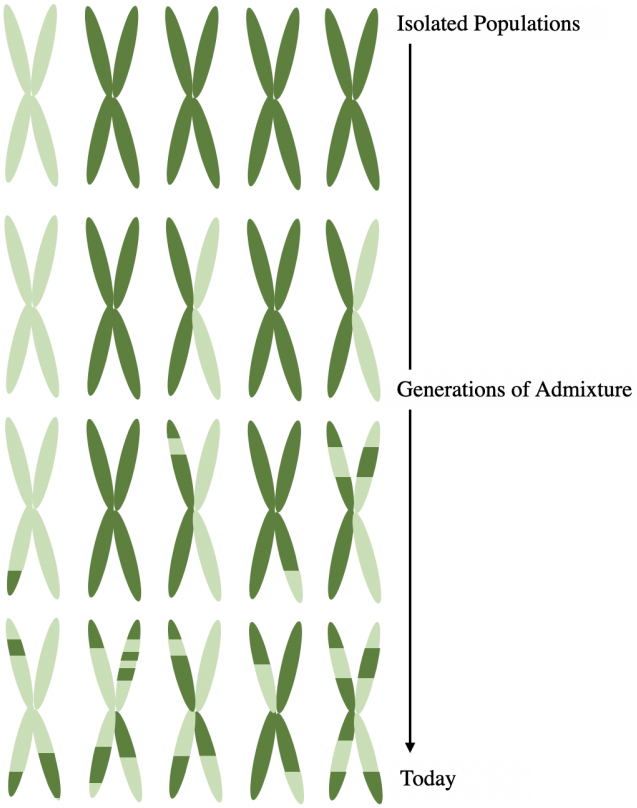


Figure 1.1: Illustration of recombination and admixture. When Africans (light green chromosomes) and Europeans (dark green chromosomes) were isolated, their chromosomes have distinct colors. When they start to interbreed, the next generation will have one chromosome that is enitrely light green and the other chromosome completely dark green, assuming no recombination events. In generations further down, more and more recombinations would happen, and the chromosomes develop a mosaic pattern of alternating segments of different ancestries. (Winkler et al., 2010a)

Chapter 2, 3, and 4 are made possible by the resources provided by Genotype-Tissue Expression (GTEx) project which shares a comprehensive public data that allows the investigation of gene regulation process. We study the African American samples included in GTEx data V6P, in particular their gene expression levels. We also use their genotype data to infer their genetic ancestry. Below, we introduce the data set in detail.

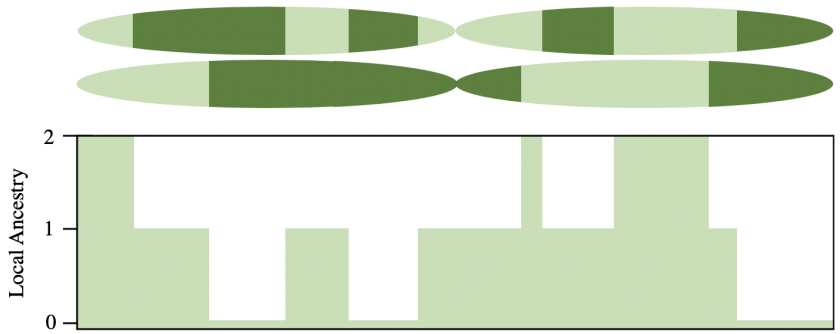


Figure 1.2: Visualization of local ancestry. We count African chromosomes (light green) at a given loci. We treat local ancestry as a quantitative trait, so it is a value ranging from 0 to 2, rather than a categorical variable of (0, 1, 2). (Winkler et al., 2010a).

1.1.2 GTEx Data

Figure 1.3 summarizes the GTEx data. The expression levels are measured for N subjects for $K = 38498$ genes in the autosomal chromosomes across $T = 44$ tissues through RNA-seq. They are pre-processed by GTEx: truncated for having at least 0.1 RPKM, normalized, log-transformed, and corrected for technical artifacts The numbers of genes are summarized in Table 1.2. For this dissertation, N is equal to 356 which is the number of samples who self-declared as either European Americans or African Americans, or close to 70 when only dealing with African Americans. Each sample has different tissue availability. For an hypothetical example, one sample could have expression levels measured at lung and heart tissues, while another sample at five different brain tissues only. Inevitably, the gene expression level data has missing values. Therefore, when we look at the data tissue by tissue, the dimension N could be different. For instance, whole blood tissue has 326 available samples while liver tissue has only 96. Therefore, the sample size N in each chapter can refer to different numbers.

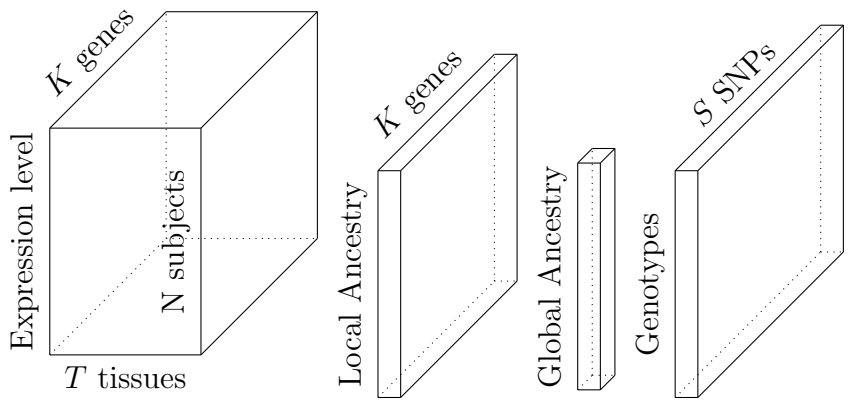


Figure 1.3: Summary of GTEx data. The expression level is a 3-dimensional tensor for N subjects, K genes, and T tissues. There are many missing values within the tensor due to tissue availability of each subject. Local Ancestry is a matrix measured for N subjects across K genes. Global Ancestry on the other hand is a length N vector as it is the same for all tissues and all genes. The genotypes are observed across S SNPs for N subjects, and genotypes are invariant to genes.

GTEx provides the genotype data of all the subjects. The 356 subjects were genotyped through Illumina OMNI 5M SNP array. Only the SNPs with minor allele frequency $> 5\%$ were included in the analysis, and only the first two most frequent alleles were used. The numbers of SNPs are summarized in Table 1.1. Local ancestry and global ancestry for N samples are inferred through external software using the genotype data in GTEx. Local ancestry is defined for each of K genes, while global ancestry is a sample-specific value that is constant across genes.

1.1.3 Ancestry Inference

There are various softwares today that can infer local ancestry given the genotype data. One of the first softwares to address local ancestry inference was Pritchard et al. (2000), but it was very limited because it takes loci that are far apart and independent from each other. This is irrelevant given today's technology where we have genotype information about very dense SNP sets. New methods such as (Halder et al., 2008; Tian et al., 2006) were developed

Number of SNPs			
chr1	502,003	chr12	335,827
chr2	576,281	chr13	264,149
chr3	488,450	chr14	225,404
chr4	514,658	chr15	201,010
chr5	447,505	chr16	210,634
chr6	471,029	chr17	188,733
chr7	408,784	chr18	202,046
chr8	382,608	chr19	164,372
chr9	298,030	chr20	153,560
chr10	350,319	chr21	99,372
chr11	343,657	chr22	95,734
		Total	6,954,165

Table 1.1: Number of SNPs per chromosomes MAF$>$ 0.5%

Number of genes			
chr1	2,387	chr12	1,190
chr2	1,544	chr13	393
chr3	1,315	chr14	753
chr4	841	chr15	810
chr5	1,088	chr16	1,048
chr6	1,174	chr17	1,386
chr7	1,141	chr18	328
chr8	818	chr19	1,512
chr9	946	chr20	550
chr10	910	chr21	267
chr11	1,268	chr22	579
		Total	22,248

Table 1.2: Number of genes per chromosome RPKM$>$ 0.1

that use Ancestry Information Markers (AIMs), which are the SNPs that are highly differentiated by populations. Hidden Markov Models were mainly applied to learn the latent variables of ancestry of each locus, assuming that each SNP is independent, not accounting for the linkage disequilibrium. Then, SNP data became even more dense so Markov Hidden Markov Model was next applied in works such as Tang et al. (2009) so that observed data are correlated.

This work applies the software LAMP that implements a moving window approach. LAMP arbitrarily assume that for a given window, there is no ancestry switch. Ancestry is determined as the one that gets the largest number of votes. Recent versions of LAMP such as WINPOP allow overlapping windows and use adaptive window sizes to make the software more flexible and more accurate. LAMP reaches as high as 98% accuracy level for distinguishing YRI and CEU ancestry. We choose LAMP for its advantage in speed compared to other softwares that use Bayesian methods while compromise in accuracy is negligible.

LAMP receives as input the minor allele frequency of pure population CEU and YRI, the genotype data of the subjects from GTEx, and the SNP positions. For reference allele

frequency of pure population, we use 186 Yoruban (YRI) subjects and 183 Northern Europeans from Utah (CEU) subjects from 1000 Genome Project Clarke et al. (2017). Therefore, only the SNPs that were genotyped both in 1000GP and GTEx were included for the local ancestry inference. SNPs with more than half of the haplotype values missing were excluded from the ancestry inference procedure. For other parameters, we use 7 for the number of generations of admixture, 0.2 and 0.8 for initial proportion of CEU and YRI population, and 10^{-8} for recombination rate. The inference results however were robust to these initial parameters. LAMP returns 0, 1, or 2 African chromosomes for each locus. An example output is in Figure 1.4 (c).

In this work, local ancestry is counted as the number of African chromosome at a given loci of a sample's chromosome, as illustrated in Figure 1.1 and 1.2. In other words, for a given locus, local ancestry is 2 if both chromosomes are African, 1 if heterozygous, and 0 if both European. Also, we define local ancestry of a gene as the local ancestry of the SNP that is closest to the center of the gene. This is designed to estimate the average of local ancestry of the gene from the start site to the end site (Price et al., 2008). In most cases, the genes had the same ancestry information overall, without any recombination event. When there are also genes in which we have no SNP information between the start site and end site, we use our best approximation of the local ancestry of the gene, which is the local ancestry of the SNP that is the most closely located to the gene, or equivalently, to the center of the gene. When there are genes where 1 or more subjects have recombination event within the gene, we still use the ancestry information of the center of the gene. Table 1.3 catalogs the number of samples that show recombination events. For instance, 20,358 genes had no recombination event for all the subjects, and 1,329 genes had one individual who had a recombination event within the gene. According to the table, around 92% of the genes show no recombination events in all the subjects, and only less than 3% of the genes have more than one individuals with recombination events. Based on these data, we conclude that our definition of the local

ancestry of a gene is a justifiable approximation of the average local ancestry of the gene (Table 1.3).

# of subjects	0	1	2	3	4	5	6	7	8	9	13
# of genes	20358	1329	379	120	34	12	9	3	2	1	1

Table 1.3: Number of genes that show recombination events. 20,358 genes show no recombination event within the gene (from transcription start site to the end) for all samples, 1329 genes show recombination event for one of the samples, etc. This table shows that our approximation of local ancstry is reasonably accurate.

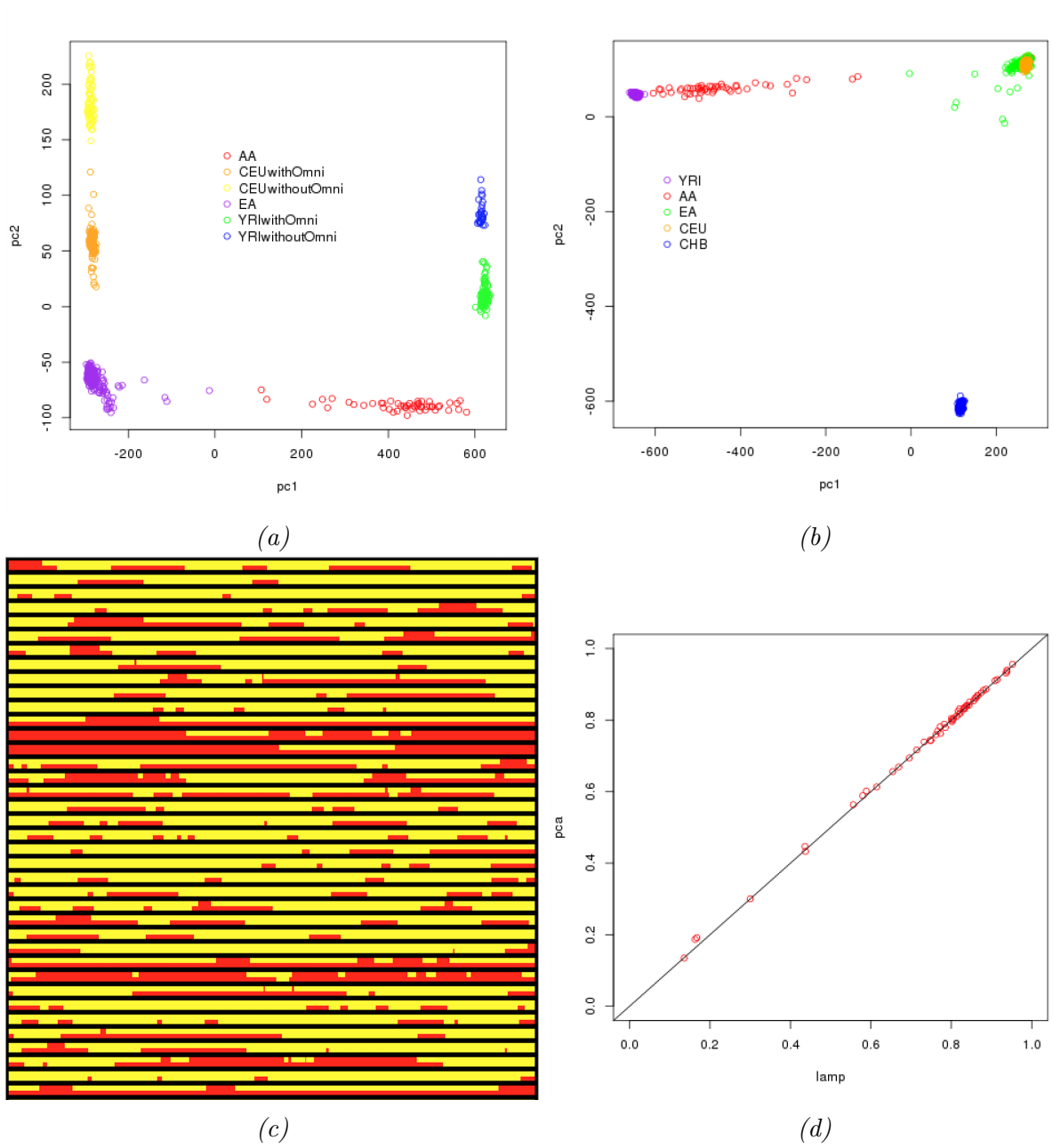


Figure 1.4: (a) PCA result of pure populations CEU and YRI from 1000 Genome Project and African Americans (AA) and European Americans (EA) from GTEx. First component identifies the ancestry while the second component identifies technical element, "Omni", which is a sequencing platform indicator. (b) PCA result, similar to (a), but with CHB population. The second PC identifies the Asian ancestry instead of technical element. (c) Sample output from LAMP. Each row is chromosome 1 of one subject, where yellow means African and red means European ancestry. (d) Comparing global ancestry inferred from LAMP and from PCA. Both local ancestry inference and PCA tell the same story regarding the global ancestry of GTEx AA samples.

Global ancestry, a value between 0 and 1 that quantifies the proportion of African chromosome in an individual, is computed by averaging the local ancestry, and it is cross-checked with the result of principal component analysis. The covariance matrix of the genotypes of AAs and EAs from GTEx and YRIs and CEUs from 1000GP was eigen-decomposed to create the plot in Figure 1.4 (a) and (b), once with CHB (Han Chinese in Beijing, China) and once without. In Figure (a), the first principal component distinguishes YRI from CEU population. As predicted, African Americans are placed between CEU and YRI. There are four outliers among self-reported GTEx European Americans. For the rest of the paper, we treat these 4 outliers as African Americans, making our pool of subjects a mix of 55 AAs and 301 EAs. The second principal component divides the data based on the genotyping procedures. The GTEx individuals were placed at the bottom. The individuals with 1000GP were also divided in that the group that was genotyped with the OMNI chip was placed below. This means that principal component analysis is capable of detecting technical artifacts (Leek et al., 2010), and verifies that there is no serious issue with the quality of the genotype data from GTEx. Figure 1.4 (c) shows that when Asian population is added, the second principal component shows the Asian ancestry rather than the technical artifacts.

Figure 1.4 (d) compares the global ancestry inferred by PCA and by LAMP. The global ancestry of individual i inferred from PCA was computed by

$$\text{global ancestry}_i = \frac{\text{pc1}_i - \overline{\text{pc1}_{CEU}}}{\overline{\text{pc1}_{YRI}} - \overline{\text{pc1}_{CEU}}}$$

where pc1 indicates the value of the first principal component in Figure 1.4 (a), and $\overline{\text{pc1}_{CEU}}$ means the average of the first principal component values of CEU population. The same goes with $\overline{\text{pc1}_{YRI}}$. This plot shows that the returned local ancestry in LAMP is verified through PCA by matching global ancestry.

1.1.4 Outline

Each chapter progressively paints a bigger picture of the transcriptome of African Americans.

Chapter 2 aims to find genes whose expression levels are correlated with their local ancestry. The two main challenges are estimating the complex tissue-tissue covariance structure and circumvent the issues introduced by missing values that arise due to different tissue availability. In Chapter 3, we expand the scope of our analysis to model the interaction between ancestry and genotype. We apply methods specialized in large-scale linear systems, and apply permutation test to preserve the correlation among the SNPs. Due to the computational burden, we look at Muscle Skeletal tissue for application example. Lastly, Chapter 4 studies how the gene coexpression pattern changes with respect to global ancestry. We propose a score test that is computationally simple and robust to model misspecification of the covariance term ρ. Subsequently, we expand the method to test relationships between one highly connected gene, such as transcription factors, and several other genes to obtain a more global view of the dynamic of the coexpression network.

Throughout the three chapters, we tackle the complicated dependence structure among the SNPs, genes, and tissues, as well as the computational difficulty that comes with the large size of the data sets.

1.2 Single Cell Transcriptome Analysis

1.2.1 Motivation

Although the RNA sequencing technology (RNA-seq) has led to countless valuable biological insights, it implicitly assumes that the sequenced cells are from a homogeneous population. In GTEx data, for example, gene expression levels are measured from a bulk of cells from one tissue. However, cells are heterogeneous even in one tissue; for example, blood tissue comprises of T cells, B cells, monocytes, and other cell types that have distinct roles. Failing to account for this cell type heterogeneity can obscure the biological signals present in the data. This issue has been overcome by the recent advances in single-cell RNA sequencing technology (scRNA-seq) that provides expression level measurements at the single-cell level. Through scRNA-seq data, cell-to-cell variation can be investigated even within one tissue.

Recently, droplet-based scRNA-seq methods produce single cell resolution data at affordable costs and essentially changed the landscape of genomics research in complex biological system. The expression levels from these methods tend to have a higher proportion of zeros compared to RNA-seq data, so these data sets have been analyzed predominantly through zero-inflated models. However, in the state-of-the-arts protocols, a step called barcoding unique molecular identifiers (UMI) removes amplification bias and further improves data quality (Islam et al., 2014). Some literature (Chen et al., 2018; Townes et al., 2019) suggests that the UMI barcoding leads to a different data structure from read count data structure, but many tools remain to not acknowledge the difference between the count data produced with and without barcoding.

In chapter 5, we propose a method that addresses the unique noise model of UMI data. This novel pipeline streamlines the pre-processing step and uses a computationally simple, easily interpretable methods that can select biologically important features and cluster the

cells into different groups while preserving the hierarchical structure of cell types.

1.2.2 scRNA-seq Data

ID	Data Set	Species	Protocol
10x	5KNeuron	Mouse Neuron	10x (v3.1) CR* 3.0.2
10x	10KHeart	Mouse Heart	10x (v3) CR 3.0.0
GSE111108	Tian2018	Human Cell Lines	10x Chromium
GSE115189	Freytag2018	Human PBMC	10x (v2)
10x	1KNeuron	Mouse Neuron	10x (v2) CR 2.1.0
SRP073767	Zhengmix4eq	Human PBMC	10x (v1) CR 1.1.0
SRP073767	Zhengmix4uneq	Human PBMC	10x (v1) CR 1.1.0
SRP073767	Zhengmix8efq	Human PBMC	10x (v1) CR 1.1.0
SRP073767	PBMC3k	Human PBMC	10x (v1) CR 1.1.0
SRP073767	PBMC4k	Human PBMC	10x (v1) CR 1.1.0
SRP073767	PBMC68k	Human PBMC	10x (v1) CR 1.1.0
GSE84133	Baron2016	Human Pancreas	inDrop
GSE114724	AziziPatient09Rep1	Human Breast Tumor	10x CR 2.1.1
GSE114724	AziziPatient09Rep2	Human Breast Tumor	10x CR 2.1.1
GSE114724	AziziPatient10Rep1	Human Breast Tumor	10x CR 2.1.1
GSE114724	AziziPatient11Rep1	Human Breast Tumor	10x CR 2.1.1
GSE114724	AziziPatient11Rep2	Human Breast Tumor	10x CR 2.1.1
SDY998	Zhang2019	Human Joint Synovial	CEL-seq2
GSE63473	Macosko2015	Mouse	Drop-seq
GSE77288	Tung2017	Human iPSC	HiSeq 2500
Tabula Muris	Tabula Muris	Mouse	10x (v2)

*Table 1.4: List of data sets used in the main text and supplementary materials. * CR: Cell Ranger*

Throughout the analysis, we use publicly available single cell UMI sequencing data most of which use 10X protocol. Most analysis in the main text is focused on SRP073767 which is also available in 10x Genomics, and it sequences 68,000 PBMC cells using Cell Ranger 1.1.0 (Zheng et al., 2017). We use different subsets of this data sets, namely Zhengmix4eq, Zhengmix4uneq, and Zhengmix8eq as defined in Duò et al. (2018). Other data sets used in the main text are GSE111108 from Tian et al. (2018), GSE115189 from Freytag et al. (2018), and GSE114724 from Azizi et al. (2018). Supplementary Materials include analyses of more data sets from 10X including 5k Cells from a combined cortex, hippocampus and

subventricular zone of an E18 mouse (v3 chemistry), 1k Brain Cells from an E18 Mouse (v2 chemistry), and 10k Heart Cells from an E18 mouse (v3 chemistry). We also use GSE84133 (Baron et al., 2016) as an example of in-Drop and GSE63473 (Macosko et al., 2015) as an example of Drop-seq. To include more data sets from different protocols and tissues, we use Tabula Muris data from Tabula Muris Consortium et al. (2018) that sequences multiple mouse tissues, Tung2017 data from Tung et al. (2017) that uses Hi-Seq 2500, and Zhang2019 data from Zhang et al. (2019) that uses CEL-seq2. All the data sets were analyzed after their own filtering process. The data sets are summarized in Table 1.4.

1.2.3 Outline

Chapter 5 shows extensive analysis results of scRNA-seq data that uses unique molecular identifiers (UMI), especially those using 10X protocol, and propose an analysis pipeline that reflect the findings. First, we show that after accounting for cell type heterogeneity, the zero inflation disappears, so that the noise model does not need the zero inflation parameter. The section demonstrates that over-correcting the zero inflation might introduce wanted noise in downstream analyses. Next section introduces the novel analysis pipeline HIPPO (Heterogeneity Inspired Pre-Processing tOol) that incorporates iterative feature selection and clustering procedure.

CHAPTER 2

EFFECTS OF LOCAL ANCESTRY ON GENE EXPRESSION LEVEL OF AFRICAN AMERICANS

2.1 Introduction

In this chapter, we study the relationship between the genetic ancestry and gene expression levels of multiple tissues. This analysis presents two important statistical challenges. First, it is difficult to fully account for the complex dependence structure across the tissues. Second, the tissue accessibility is different for each sample, introducing high proportion of missing values. Given the low sample size of admixed population, missing values make statistical inference particularly demanding.

This chapter proposes a Bayesian hierarchical modeling and inference method to analyze the effect of genetic ancestry of African American genome on incomplete, correlated gene expression level data. Our method not only accounts for the covariance structure of the response variable but also provides a flexible interpretation of the effect size β while circumventing the issues introduced by missing values. Simulations show that including the variance structure in the model improves statistical power by borrowing information across the observations from multiple tissues. The model is fitted through Markov Chain Monte Carlo (MCMC) algorithm.

Bayesian variable selection methods have been studied in various contexts, and many works use spike and slab prior as a widely accepted method. This prior models the regression coefficients as a mixture of a point mass at 0 and a normal distribution (George & McCulloch, 1997; Hernández-Lobato et al., 2013; Narisetty et al., 2014). There have been extensions to a multivariate linear regression with the spike-and-slab prior as well (Brown et al., 1998; Lee et al., 2017), but they tend to put a restrictive prior on the effect size.

For example, some assume that the effect sizes follow the same variance structure as the response variable or that they come from an independent normal distribution with an arbitrarily fixed amount of variance (Lee et al., 2017; Brown et al., 1998; Liang et al., 2008). Meanwhile, (Guan & Stephens, 2011), under the context of univariate Bayesian variable selection, parametrizes the effect size using the concept of proportion of variance explained (PVE) and therefore offers direct and intuitive interpretation. Here, this prior is extended to the multivariate context to allow more flexibility and better interpretation.

Also, most past works that attempt to analyze incomplete multivariate data have focused on imputations. There are methods to effectively impute the gene expression level matrix, one specifically for the missing values driven by tissue accessibility in GTEx, but not many methods attempt to analyze only the available data (Wang et al., 2016; Oba et al., 2003; Li et al., 2017). In this work, we instead make the "Missing at Random" (MAR) assumption (Little & Rubin, 2014; Rubin, 1976) that allows analysis without imputation. Sometimes, certain data-generating process drives the correlation between the missing pattern and the unobserved data. For example, gene expression levels, especially in single-cell RNA-seq data, are recorded as missing in RNA-seq because not enough reads have been mapped, in which case the missingness suggests that the underlying true value is low. In such case, the missing pattern holds relevant information about the data. However, the missing pattern in GTEx only depends on the tissue availability of the patients. This is mostly technical rather than biological. Possible factors include whether the sample is post-mortem or surgical, which surgery the subject went through, and which tissue is difficult to maintain fresh samples. Therefore, we assume that tissue availability holds no information regarding the underlying true gene expression level. Simulations show that the proposed method performs well even when the majority of the data is missing.

The remainder of this chapter is organized as follows. Section 2.2 describes the Bayesian

linear regression framework, including the including Markov chain Monte Carlo algorithm that works around the missing values. In Section 2.3, we present the simulation results that examine the effectiveness of our proposed method under various settings. We prove the algorithm's superiority to traditional linear analysis and show the algorithm's robustness to hyperparameter mis-specifications. Section 2.4 shows the results of analyzing the GTEx data. We conclude the chapter with a discussion in Section 2.5.

2.2 Methods

This section introduces the proposed Bayesian normal regression model, explains the prior distributions and hyperparameter specifications, and presents the MCMC algorithm.

2.2.1 Model

For each gene, we model the effect of local ancestry L on gene expression $\boldsymbol{y}$, assuming all other covariates have been accounted for. The gene expression levels $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iT})$ are measured for each subject $i = 1, \cdots, N$. Each sample i has missing values in at least one of T tissues. We consider a linear regression model with multivariate response $\boldsymbol{y}_i$ with a predictor vector $L = (L_1, \cdots L_N)^T$, the coefficient parameter $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_T)^T$, and the mean parameter $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_T)^T$. We assume $\boldsymbol{y}_i$, including the unobserved values, follows a multivariate normal distribution, $\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathcal{N}_T(\boldsymbol{\mu} + L_i\boldsymbol{\beta}, \boldsymbol{\Sigma})$, with a dense, unstructured covariance matrix $\Sigma \in \mathbb{R}^{T\times T}$. When $\boldsymbol{y}_i$ and L are centered, the calculation is simplified while posterior for $\boldsymbol{\beta}$ is unaffected, so we assume the response variable $\boldsymbol{y}_i$ and the covariate L are centered henceforth. The mean term disappears, and our final model is

$$\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathcal{N}_T(L_i\boldsymbol{\beta}, \boldsymbol{\Sigma}). \tag{2.1}$$

2.2.2 Prior Distributions

We use "spike-and-slab" prior for the coefficients $\boldsymbol{\beta}$ as suggested by past literature for Bayesian variable selection (Mitchell & Beauchamp, 1988; George & McCulloch, 1993, 1997). The latent binary variable γ_t for $t = 1, \cdots, T$ indicates whether the variable t is included in the model. If $\gamma_t = 1$, β_t is non-zero and comes from the distribution $\mathcal{N}(0, \sigma_\beta^2)$, and if $\gamma_t = 0$, β_t comes from a point mass at 0.

$$\beta_t \sim \gamma_t \mathcal{N}(0, \sigma_\beta^2) + (1 - \gamma_t)\delta_0. \tag{2.2}$$

The latent variables γ_t independently follow Bernoulli distribution $\gamma_t \sim Ber(\pi)$, and $\pi \sim \text{Be}(a, b)$ where a and b are hyperparameters to be specified. For the algorithm's simplicity, we obtain the marginal prior of $\boldsymbol{\gamma}$ by integrating $p(\boldsymbol{\gamma}, \pi)$ over $p(\pi)$. The marginal prior only depends on the size of the vector: $|\boldsymbol{\gamma}| = \sum_{t=1}^{T} \gamma_t$.

$$p(\gamma_t) = \frac{\Gamma(a+b)\Gamma(|\boldsymbol{\gamma}| + a)\Gamma(T + b - |\boldsymbol{\gamma}|)}{\Gamma(T + a + b)\Gamma(a)\Gamma(b)}. \tag{2.3}$$

We explain our choice of the hyperparameters a and b in the next section.

It is possible to model $\boldsymbol{\gamma}$ to have a non-trivial correlation structure. One natural way is to define a latent variable following a multivariate normal distribution with mean 0 and variance same as that of Y, and make an arbitrary threshold for each dimension to obtain correlated binary variables. However, this requires the computation of cumulative distribution of multivariate normal which does not have a closed form. Moreover, even when γ_t is independent a priori, the correlation structure inferred from the data will decide the selection of γ_t. Using an independent prior for γ_t is equivalent to only using the correlation information that comes from the data, and therefore we believe it is a valid choice.

Next, we model the covariance matrix Σ to follow the inverse Wishart prior distribution

$$\Sigma \sim W^{-1}(\nu, \nu\Phi). \tag{2.4}$$

which is conjugate to the multivariate normal variance. The posterior mean is a weighted average of the hyperparameter Φ and the empirical covariance matrix, and the weights are decided by the degrees of freedom ν and sample size N.

The set-up so far is quite standard and is backed up by past literature (George & McCulloch, 1993; Mitchell & Beauchamp, 1988). There have been various suggestions for the specification of σ_β^2 without a general consensus on a natural choice. Some use an arbitrary fixed number or an estimate of the coefficients from the data (Brown et al., 1998; Lee et al., 2017). Some attempt to put a prior on $\boldsymbol{\gamma}$ to make the model more flexible (Liang et al., 2008). Here, we choose an option that best aids the interpretation. Guan & Stephens (2011) focuses on what the prior implies about the proportion of variance in $\boldsymbol{y}$ explained by L (PVE) under a univariate setting for GWAS. Other priors have assumed the independence of $\boldsymbol{\gamma}$ and σ_β^2, which implies that more complex models are expected to have higher PVE. However, both in GWAS and in gene expression level analysis, it seems plausible a priori that a simple model has a higher PVE and a complex model has a low PVE. For example, local ancestry can mainly drive the variation of a gene's expression level in one tissue but has no effect in other tissues. In this case, a large proportion of variance of $\boldsymbol{y}$ is explained by L although $|\boldsymbol{\gamma}|$ is only 1.

So we expand this concept to a multivariate version with correlated response variables. In univariate linear regression, PVE is defined as R^2, but it does not have a natural extension to the multivariate setting because there is no scalar representation of the covariance matrix. As an approximation, we use the trace of the covariance matrix. It has an intuitive

interpretation of the amount of variance explained because the trace of a matrix is equal to the sum of its eigenvalues. For example, in the principal component analysis, a normalized eigenvalue is the proportion of variance explained by each principal component.

To formalize the multivariate PVE, let $V(\boldsymbol{\beta}) = \frac{1}{N} tr[(L\boldsymbol{\beta}^T)^T(L\boldsymbol{\beta}^T)]$ denote the trace of the empirical variance of $L\boldsymbol{\beta}^T$. In the beginning of the paper, we centered both response and the covariate, so there is no need to consider the mean term. Then $\text{PVE}(\boldsymbol{\beta}) = \frac{V(\beta)}{V(\beta)+\text{tr}(\Sigma)}$. We define h as the approximation of the expectation of PVE,

$$h := \frac{E(V(\boldsymbol{\beta}))}{E(V(\boldsymbol{\beta})) + \text{tr}(\Sigma)} \approx E(PVE(\boldsymbol{\beta}) \mid \Sigma, \gamma, \sigma_\beta^2)$$

where

$$E(V(\boldsymbol{\beta}) \mid \Sigma, \sigma_\beta^2, \gamma) = \sum_{t:\gamma_t=1} \sigma_\beta^2 \sum_{i=1}^{N} \frac{L_i^2}{N}$$

with expectation being taken over $\boldsymbol{\beta}$. Note that we approximate the expectation of a ratio as a ratio of expectations. This approximation is equivalent to approximating $1-\text{tr}(\Sigma)E\left(\frac{1}{V(\beta)+\text{tr}(\Sigma)}\right)$ as $1-\text{tr}(\Sigma)\left(\frac{1}{E(V(\beta))+\text{tr}(\Sigma)}\right)$, and Jensen's inequality tells us that this formulation of h systematically overestimates the PVE. The error is 0 when $\sigma_\beta^2 = 0$ and the error grows as σ_β^2 becomes larger. However, we don't expect σ_β^2 to be very large in our application, and so we believe that this approximation works well as a proxy to PVE. Therefore, h can be represented in terms of σ_β^2.

$$h(\sigma_\beta^2|\boldsymbol{\gamma}, \Sigma) = \frac{\sum_{t:\gamma_t=1} \sigma_\beta^2 \sum_{i=1}^{N} \frac{L_i^2}{N}}{\sum_{t:\gamma_t=1} \sigma_\beta^2 \sum_{i=1}^{N} \frac{L_i^2}{N} + \text{tr}(\Sigma)}.$$

Then we can parametrize σ_β^2 in terms of h, $\sigma_\beta^2(h|\boldsymbol{\gamma}, \Sigma) = \frac{h \cdot \text{tr}(\Sigma)}{(\sum_t \gamma_t)(1-h)\sum_{i=1}^{N} \frac{L_i^2}{N}}$. When no variables are selected, $\sum_t \gamma_t$ is 0, and $\sigma_\beta^2(h|\mathbf{0}, \Sigma)$ becomes non-finite, so we add a "pseudo-count" in the denominator. Our final specification of σ_β^2 given h, $\boldsymbol{\gamma}$, and Σ is

$$\sigma_\beta^2(h|\boldsymbol{\gamma}, \Sigma) = \frac{h \cdot \mathrm{tr}(\Sigma)}{(\sum_t \gamma_t + 1)(1-h)\sum_{i=1}^N \frac{L_i^2}{N}}. \tag{2.5}$$

With these adjustments, we can parametrize the model with h instead of with σ_β^2 according to an almost non-informative prior that helps interpretation. We use a uniform prior on h, and the specification of its hyperparameters is discussed in the next section.

2.2.3 Hyperparameter Specifications

The hyperparameters ν and Φ are easy to interpret because the posterior mean is a weighted average of the empirical covariance matrix $\hat{\Sigma}$ and Φ, and N and ν respectively decides each weight. Φ can be estimated empirically for the application of GTEx data. We average the empirical covariance matrices computed from the available gene expression measurements from $\sim 10,000$ genes to use as Φ, effectively leveraging information across many genes. This is from an assumption that tissue-tissue correlation is similar across many genes. For example, for any gene, the expression level from sun-exposed skin tissue will be more correlated to not-sun-exposed skin tissue than to a brain tissue. For other applications, standard choices such as $T \times T$ identity matrix are acceptable as well, as long as it is well conditioned.

Since we have a reasonable choice of Φ, we set the degrees of freedom ν to N, giving the same weight to the prior and to the data. Although it is possible to use smaller ν to give minimal weight to the prior and allow flexibility in the choice of Φ, in the particular case of gene expression level of African Americans where N is small and data is highly incomplete, it is difficult to get a good estimate of the covariance structure, so we decide to give more weight to the prior. We believe this is a valid choice mainly because we have a biological explanation for this prior. Moreover, we have enough genes (more than 10,000) to dilute the effect of using a data-driven prior.

Next, we fix a and b for the prior distribution of π that reflects the sparsity of the model. For a well-justified prior, we look at past eQTL analyses with GTEx data that studies multi-tissue gene expression level. (The GTEx Consortium, 2015) tested the SNPs for their effects on gene expression level in various tissues, and the result showed that much more SNPs were related to only 1 or all of the tissues than to a few tissues, showing a U-shaped pattern with respect to the number of tissues. Although the profiles involving only a few tissues have many more possible combinatorial patterns, eQTLs show high tissue specificity and high tissue ubiquity. We expect similar behavior from the effects of local ancestry. We first expect that most of the genes will show signal in no tissues. For the rest of the genes, we expect many of them to show signals on either 1 or all T tissues.

Figure 2.1 shows the marginal prior of $\boldsymbol{\gamma}$ for different values of (a, b). When $a = b = 0.1$, $p(\boldsymbol{\gamma})$ is symmetrically U-shaped with the highest density at $|\boldsymbol{\gamma}| = 0$ and $|\boldsymbol{\gamma}| = T$, but it doesn't give particularly large weight to the null case. When $a = 0.1$ and $b = 5$, more weight is given to $|\boldsymbol{\gamma}| = 0$, but the graph is not U-shaped. When $a = 0.01$ and $b = 0.5$, the expected $|\boldsymbol{\gamma}|$ is same as before, around 90% of the weight is given to $|\boldsymbol{\gamma}| = 0$, and it also keeps the U-shape among the rest of the cases. We believe this reflects our prior belief about the effect of local ancestry on multi-tissue gene expression level, so we use this setting for the algorithm. Although this may seem very restrictive compared to a non-informative prior of $\pi \sim \text{Be}(1, 1)$, given that we are testing the 30,000 genes separately, we believe a more conservative prior is appropriate for a reasonable error control.

We next choose the prior for h. Guan & Stephens (2011) uses non-informative uniform prior on h, but h near the boundary of this support can be problematic in the multivariate context. For example, when $h = 0$, σ_β^2 becomes 0, and the algorithm would no longer add any variables. Also, when h is small and σ_β^2 is too close to zero, the normal distribution $N(0, \sigma_\beta^2)$ does not have much discriminating power from the point mass at 0, disabling the

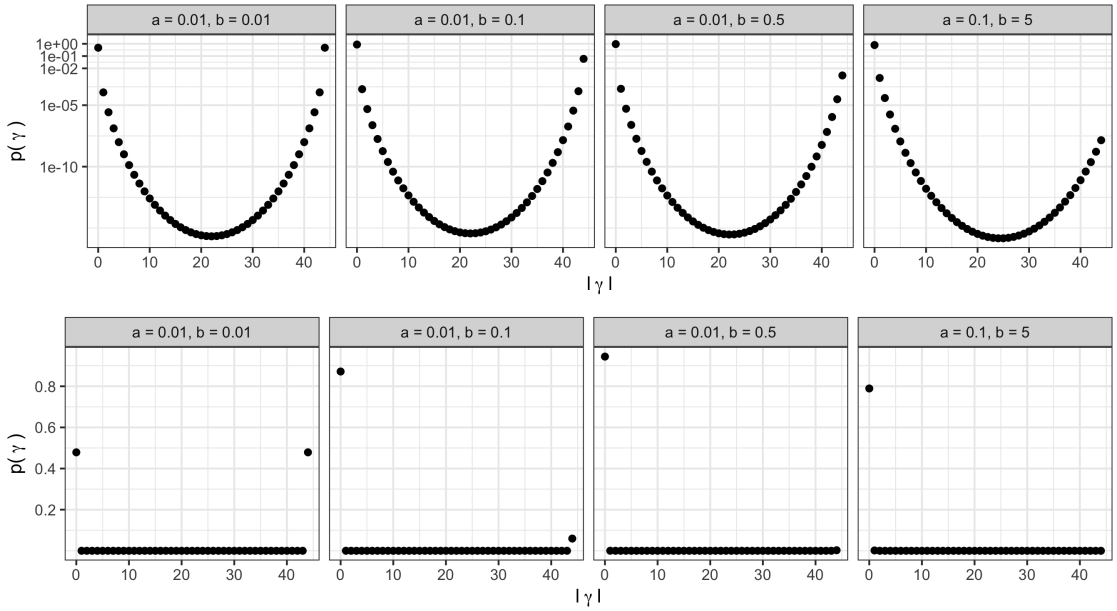


Figure 2.1: Marginal distribution of $\boldsymbol{\gamma}$ for different hyperparameter settings. The y-axis shows the probability mass of $\boldsymbol{\gamma}$ with given $|\boldsymbol{\gamma}|$, where the first row is in log-scale and the second row is in the plain scale. When $a = 0.01$ and $b = 0.5$, $p(\boldsymbol{\gamma})$ puts around 90% of the weight on the null ($\boldsymbol{\gamma} = \mathbf{0}$) and distributes the rest of the weight on the rest with a rough U-shape.

spike-and-slab prior of $\boldsymbol{\beta}$ (George & McCulloch, 1997). On the other hand, when h becomes close to 1, the denominator becomes close to 0. Moreover, PVE value close to 1 is unrealistic in most biological applications. To account for these boundary cases, we use Unif$(0.1, 0.9)$ as the prior for h. Having a lower bound on h effectively puts an appropriate lower bound on σ_β^2, and restricting the range of h by putting an upper bound can decrease the search space and can expedite the algorithm while reflecting our belief that local ancestry explains less than 90% of the variance of gene expression level. In other applications, the upper bound can be extended to higher values, as long as it is strictly less than 1.

2.2.4 MCMC algorithm

The Markov-chain Monte Carlo algorithm is based on the following factorization of the joint model, $p(\boldsymbol{y}|\boldsymbol{\beta}, \Sigma)p(\boldsymbol{\beta}|\sigma_\beta, \boldsymbol{\gamma})p(\boldsymbol{\gamma}|\pi)p(\pi)p(\sigma_\beta^2)p(\Sigma)$. Replacing $p(\sigma_\beta)$ with $p(h)$ and integrating

out γ leads to the following form

$$\prod_{i=1}^{N} p(\boldsymbol{y}_i|\boldsymbol{\beta}, \Sigma)p(\boldsymbol{\beta}|h, \boldsymbol{\gamma})p(\boldsymbol{\gamma})p(h)p(\Sigma). \tag{2.6}$$

This is equivalent to the product of the likelihood, prior for $\boldsymbol{\beta}$, prior for $\boldsymbol{\gamma}$, prior for h, and prior for Σ. This serves as the target distribution in our MCMC algorithm. We initialize $\boldsymbol{\gamma}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$ as $\mathbf{0}$, $\Sigma^{(0)}$ as Φ, and the $\sigma_\beta^{2(0)}$ as 1. At each iteration j, we repeat the following steps of updating $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, updating Σ, and updating h. The details below are for any jth iteration.

We first update $\boldsymbol{\beta}^{(j)}$ and $\boldsymbol{\gamma}^{(j)}$ simultaneously using the Metropolis-Hastings algorithm (MH) given fixed $\Sigma^{(j-1)}$ and $\sigma_\beta^{2(j-1)}$ (Robert & Casella, 2013; Hastings, 1970). We assign with $\boldsymbol{\beta}^{(j)} = \boldsymbol{\beta}^{(j-1)}$ and $\boldsymbol{\gamma}^{(j)} = \boldsymbol{\gamma}^{(j-1)}$, propose $\boldsymbol{\beta}^*$ and $\boldsymbol{\gamma}^*$ 500 times, and accept the proposals whenever the acceptance probability, defined later in the section, is high enough. We propose $\boldsymbol{\gamma}^*$ by changing the status of one variable from the values of $\boldsymbol{\gamma}^{(j)}$. Each variable $t = 1, \cdots, T$ has an equal chance $(1/T)$ of being selected. If the picked variable t is already in the model $\left(\gamma_t^{(j-1)} = 1\right)$, then proposed value γ_t^* is 0. If the variable t is not in the model $\left(\gamma_t^{(j-1)} = 0\right)$, then proposed value γ_t^* is 1.

Based on $\boldsymbol{\gamma}^*$, we draw β_t^* from for all t such that $\gamma_t^* = 1$. We need to pre-fix u_β^2 and w_β^2 as prespecified variances of the proposal density (Lee et al., 2017). The β_t of the newly added variable t is drawn from $N(0, u_\beta^2)$. The variables that are already in the model and are not removed are drawn from $N(\boldsymbol{\beta}^{(j)}_{t:\gamma_t^{(j)}=1,\gamma_t^*=1}, w_\beta^2)$. The reason we use two separate variances for the update is to expedite the mixing and convergence of the algorithm. u_β^2 reflects an approximate size of σ_β^2 as it determines the initial guess of the effect size of a variable. Empirically, we use the variance of linear regression coefficients for the variables with p-values less than 0.05. One might suggest using $\sigma_\beta^{2(j-1)}$ for u_β^2, but this unnecessarily slows down the convergence because the posterior of σ_β^2 has high variance. w_β^2 reflects the size of per-

turbation that fine-tunes the coefficients that are already in the model, and around 1/100 of u_β^2 works well.

With the proposed $\boldsymbol{\gamma}^*$ and $\boldsymbol{\beta}^*$, we compute the acceptance probability as the product of the likelihood ratio, prior ratio for $\boldsymbol{\beta}$, prior ratio for $\boldsymbol{\gamma}$, and proposal ratio. Note that the prior for Σ is canceled out. First, the likelihood ratio is,

$$\begin{aligned} L &= \left(\frac{\prod_{i=1}^N f(\boldsymbol{y}_i \mid \boldsymbol{\beta}^*, \Sigma^{(j-1)})}{\prod_{i=1}^N f(\boldsymbol{y}_i \mid \boldsymbol{\beta}^{(j)}, \Sigma^{(j-1)})} \right) \\ &= \frac{\prod_{i=1}^N \exp\left(-\frac{1}{2}(\boldsymbol{y}_i - L_i\boldsymbol{\beta}^*)^T \Sigma^{-1(j-1)} (\boldsymbol{y}_i - L_i\boldsymbol{\beta}^*)\right)}{\prod_{i=1}^N \exp\left(-\frac{1}{2}(\boldsymbol{y}_i - L_i\boldsymbol{\beta}^{(j)})^T \Sigma^{-1(j-1)} (\boldsymbol{y}_i - L_i\boldsymbol{\beta}^{(j)})\right)} \end{aligned} \tag{2.7}$$

Note that since $\boldsymbol{y}_i$ has missing values, we need some adjustments in computing L. We discuss this in detail in the next section. The prior ratio for $\boldsymbol{\beta}$ is

$$B = \frac{p(\boldsymbol{\beta}^* | \sigma_\beta^{2(j-1)})}{p(\boldsymbol{\beta}^{(j)} | \sigma_\beta^{2(j-1)})} = \frac{\prod_{t:\gamma_t^*=1} \exp\left(-\frac{\beta_t^{2*}}{2\sigma_\beta^{2(j-1)}}\right)}{\prod_{t:\gamma_t^{(j-1)}=1} \exp\left(-\frac{\beta_t^{2(j)}}{2\sigma_\beta^{2(j-1)}}\right)}$$

The prior ratio for $\boldsymbol{\gamma}$ is

$$G = \frac{p(\boldsymbol{\gamma}^*)}{p(\boldsymbol{\gamma}^{(j)})} = \frac{\Gamma(a + |\boldsymbol{\gamma}^*|)\Gamma(T + b - |\boldsymbol{\gamma}^*|)}{\Gamma(a + |\boldsymbol{\gamma}^{(j)}|)\Gamma(T + b - |\boldsymbol{\gamma}^{(j)}|)}.$$

The proposal ratio when adding a variable is like below. First, we define g as the number of non-zero values in $\gamma^{(j-1)}$, and t as the index of the variable that we are adding ($\gamma_t^{(j)} = 0, \gamma_t^* = 1$). q_1 is the proposal distribution of $\boldsymbol{\beta}$ (Hastings, 1970).

$$P_{\text{add}} = \frac{1/(g+1)}{1/(T-g)} \times \frac{q_1(\boldsymbol{\beta}^* \to \boldsymbol{\beta}^{(j)} \mid \boldsymbol{\gamma}^*)}{q_1(\boldsymbol{\beta}^{(j)} \to \boldsymbol{\beta}^* \mid \boldsymbol{\gamma}^{(j-1)})}$$

$$= \frac{1/(g+1)}{1/(T-g)} \times \frac{1}{\frac{1}{\sqrt{2\pi u_\beta^2}}\exp(-\frac{\beta_t^{*2}}{2u_\beta^2})}$$

Similarly, the proposal ratio when deleting a variable t is ($\gamma_t^{(j)} = 1$, $\gamma_t^* = 0$)

$$P_{\text{delete}} = \frac{1/(T-g+1)}{1/g} \times \frac{\frac{1}{\sqrt{2\pi u_\beta^2}}\exp\left(-\frac{\beta_t^{(j)2}}{2u_\beta^2}\right)}{1}$$

To summarize, the acceptance probability is $\min(A, 1)$ where A is either $L \times B \times P_{\text{add}}$ or $L \times B \times P_{\text{delete}}$. With probability of $\min(A, 1)$, we make an update $\beta^{(j)} = \beta^*$ and $\gamma^{(j)} = \gamma^*$. Otherwise, we propose another set of $\boldsymbol{\beta}^*$ and $\boldsymbol{\gamma}^*$ values without changing $\boldsymbol{\beta}^{(j)}$ or $\boldsymbol{\gamma}^{(j)}$. We repeat the process until we exhaust all the 500 different proposals.

Then we update $\Sigma^{(j)}$ given $\boldsymbol{\beta}^{(j)}$ and $\boldsymbol{\gamma}^{(j)}$, and $h^{(j-1)}$ by drawing from the posterior distribution. Since inverse Wishart distribution is the conjugate prior for multivariate normal variance, we can easily get the closed form posterior for $\Sigma^{(j)}$:

$$\begin{aligned} &p(\Sigma^{(j)}|\boldsymbol{\beta}^{(j)}, h^{(j-1)}, \boldsymbol{\gamma}^{(j)}, \boldsymbol{y}) \\ &= W^{-1}\left(N+\nu, \sum_{i=1}^{N}(\boldsymbol{y}_i - L_i\boldsymbol{\beta}^{(j)})(\boldsymbol{y}_i - L_i\boldsymbol{\beta}^{(j)})^T + \nu\Phi\right) \end{aligned} \tag{2.8}$$

where $\frac{1}{N}\sum_i(\boldsymbol{y}_i - L_i\boldsymbol{\beta}^{(j)})(\boldsymbol{y}_i - L_i\boldsymbol{\beta}^{(j)})^T$ is the MLE of the covariance matrix of the data. This is problematic because $\boldsymbol{y}_i$ is not a complete vector. Next section about missing data explains the adjustment to this posterior in detail.

Next, we update $h^{(j)}$ through Metropolis-Hastings algorithm. First specify $h^{(j)} = h^{(j-1)}$ and test for 100 proposals $h^* = h^{(j)} + \delta$ where δ is randomly drawn from Unif$(-0.1, 0.1)$. h^* is reflected around the boundary of the support $(0.1, 0.9)$. We compute σ_β^* from the proposed h^* and calculate thacceptance probability. The updates are symmetric, so proposal probability

is ignored. Then the acceptance probability is $min(C, 1)$ where

$$C = \frac{p(\boldsymbol{\beta}^{(j)} \mid \sigma_\beta^{*2})}{p(\boldsymbol{\beta}^{(j)} \mid \sigma_\beta^{2(j)})} = \frac{\prod_{t:\gamma_t^{(j)}=1} \exp\left(-\frac{\beta_t^{2(j)}}{2\sigma_\beta^{2*}}\right)}{\prod_{t:\gamma_t^{(j)}=1} \exp\left(-\frac{\beta_t^{2(j)}}{2\sigma_\beta^{2(j)}}\right)}$$

With probability of $min(C, 1)$ update $h^{(j)} = h^*$ and $\sigma_\beta^{(j)} = \sigma_\beta^*$. Otherwise, we propose new h^* without changing $h^{(j)}$. We repeat the process until we exhaust all the 100 proposals.

2.2.5 Missing Data

The multi-tissue expression level data from GTEx has many missing values, and the proposed computational algorithm is not feasible if the data is incomplete. The two main challenges are computing the acceptance probability in the Metropolis-Hastings algorithm and computing the empirical covariance matrix for updating Σ. In this section, we propose a way to work around the missing values using MAR assumption.

Past works that attempt to analyze GTEx's multi-tissue gene expression level matrix have focused on imputations. Many of them also focused only on some of the tissues that have plenty of observations (Li et al., 2017; The GTEx Consortium, 2015). Here, we adopt a classical approach by modeling M (Rubin, 1976), a binary random variable indicating data availability, and use all the available tissues even with less than 10 observations.

We define $M = (M_1, \cdots, M_n)$ sa a matrix random variable of missing data indicator. Each M_i is a length T vector with values of 0 or 1, indicating tissue availability for individual i. The probability that M takes the value $\boldsymbol{m} = (\boldsymbol{m_1}, \cdots, \boldsymbol{m_N})$ given $\boldsymbol{y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N)$ is $g(\boldsymbol{m}|\boldsymbol{y})$. As mentioned in the introduction, we assume that the tissue availability holds no information regarding the gene expression level, either observed or unobserved. This condition where the missing pattern is independent of the underlying true values is called missing

at random (MAR). Under the MAR assumption, $g(\boldsymbol{m_i} \mid \boldsymbol{y}_i) = g(\boldsymbol{m_i})$ takes the same value for all $\boldsymbol{y}_i$, and this allows simpler analysis of incomplete data (Rubin, 1976). For notational convenience, consider a separation of $\boldsymbol{y}_i$ into the observed part $\boldsymbol{y}_{io}$ and the missing part $\boldsymbol{y}_{im}$.

One challenge of the current version of the algorithm is the computation of the acceptance probability, especially the likelihood ratio (2.7), when we update $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. The likelihood $\prod_{i=1}^{N} L(\boldsymbol{\beta}|\boldsymbol{y}_i, \Sigma) = \prod_{i=1}^{N} f(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \Sigma)$ cannot be computed when $\boldsymbol{y}_i$ is not a complete vector. The full posterior distribution of the parameter $\boldsymbol{\beta}$ accounting for $\boldsymbol{m}$ is proportional to

$$p(\boldsymbol{\beta}) \prod_{i=1}^{N} \int f(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \Sigma) g(\boldsymbol{m_i}) d\boldsymbol{y}_{im}$$

and, under MAR, this is equivalent to

$$c \cdot p(\boldsymbol{\beta}) \prod_{i=1}^{N} \int f(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \Sigma) d\boldsymbol{y}_{im}$$

where c is some constant that is canceled out in likelihood ratio. In our context, $f(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \Sigma)$ is multivariate normal $N_T(\boldsymbol{y}_i; \boldsymbol{\beta}, \Sigma)$, and so $\int f(\boldsymbol{y}_i \mid \boldsymbol{\beta}) d\boldsymbol{y}_{im}$ is equivalent to the marginal density of multivariate normal $N_T(\boldsymbol{y}_{io}; \boldsymbol{\beta_{i,o}}, \Sigma_{i,o})$, where $\boldsymbol{\beta}_{i,o}$ is the subvector of coefficient $\boldsymbol{\beta}$ only at the observed index of individual i, and $\Sigma_{i,o}$ is similarly the submatrix of covariance matrix Σ with rows and columns indexed at the observed part of individual i. This shows that replacing the full joint likelihood with the marginal likelihood does not influence the posterior of our parameters of interest (Rubin, 1976), and therefore the likelihood ratio in (2.7) becomes

$$\frac{\prod_i \exp\left(-\frac{1}{2}(\boldsymbol{y}_{io} - L_i \boldsymbol{\beta}^*_{\boldsymbol{io}})^T \Sigma_{io}^{-1(j-1)} (\boldsymbol{y}_{io} - L_i \boldsymbol{\beta}^*_{\boldsymbol{io}})\right)}{\prod_i \exp\left(-\frac{1}{2}(\boldsymbol{y}_{io} - L_i \boldsymbol{\beta}^{(j)}_{\boldsymbol{io}})^T \Sigma_{io}^{-1(j-1)} (\boldsymbol{y}_{io} - L_i \boldsymbol{\beta}^{(j)}_{\boldsymbol{io}})\right)}.$$

Another challenge from the missing values is the empirical covariance matrix required to update Σ with conditioning on fixed $\boldsymbol{\beta}$, and we can apply the same process. The full posterior

for Σ is proportional to

$$p(\Sigma)\prod_{i=1}^{N}\int f(\boldsymbol{y}_i \mid \Sigma)g(m_i \mid \boldsymbol{y}_{io})d\boldsymbol{y}_{im}$$
$$= c \cdot p(\Sigma)\prod_{i=1}^{N}\int f(\boldsymbol{y}_i \mid \Sigma)d\boldsymbol{y}_{im}$$

The likelihood $\int f(\boldsymbol{y}_i \mid \Sigma)d\boldsymbol{y_{im}}$ is no longer a function of the full matrix Σ but rather a submatrices $\Sigma_{i,o}$ for $i = 1, \cdots, n$, and it is impossible to obtain a closed form posterior. We propose to use the EM algorithm to estimate the MLE $\hat{\Sigma}$ and maintain the posterior formula (2.8). We find MLE $\hat{\Sigma}$ by solving the following optimization function through the EM algorithm, whose details can be found in (Little & Rubin, 2014).

$$\arg\max_{\Sigma} -\frac{1}{2}\sum_{i=1}^{N} log|\Sigma_{i,o}| - \frac{\sum_{i=1}^{N}(\boldsymbol{y}_i - L_i\boldsymbol{\beta}_{i,o})^T\Sigma_{i,o}^{-1}(\boldsymbol{y}_i - L_i\boldsymbol{\beta}_{i,o})}{2}. \tag{2.9}$$

The first obvious benefit of this approximation is the algorithm's simplicity. There is no intuitive proposal distribution for the covariance matrix to implement the Metropolis-Hastings algorithm, and especially when T is large, calculating the acceptance probability for a large number of covariance matrices can be computationally intractable. Another benefit is that this EM algorithm can return a valid result even when two variables share no common subjects. For example, there is no subject that has observations both in Testis and Uterus tissues, but the MLE $\hat{\Sigma}$ can return a valid correlation value between the two using other variables, while the posterior mean in (2.9) only receives information from the prior at such indices. Also, simulations show this update procedure does not interfere with the correct inference on $\boldsymbol{\gamma}$.

2.3 Simulation Studies

We evaluate the proposed method's performance on data sets simulated under multiple settings. The results show that the algorithm performs well even in difficult settings with low sample size, high rate of missing values, and weak signals, all of which are expected in the GTEx data. We also demonstrate that the method is robust to hyperparameter misspecification, and that posterior inclusion probability (PIP) is well calibrated in that variables with higher PIP has higher proportions of true positives. The simulation results show that the proposed method improves the statistical power compared to the univariate linear regression that assumes independence among the outcomes.

We construct data sets to resemble the real GTEx data. For sample size, we use $N = 71$ which is the number of African American samples and $T = 24$ which matches the number of tissues in GTEx data with available expression levels from more than 20 subjects. The missing pattern is inherited from the tissue availability of the real data. Some outcomes have more missing values than others, and on average, around 53% of the entries are missing. The covariate L comes from one of the gene's local ancestry data where age, sex, and global ancestry have been regressed out. We fix Σ with 1 at the diagonals and 0.5 elsewhere. Fixed error level creates a consistent environment so that we can observe how the power varies with the effect sizes.

We generate the data sets as follows. For each simulation, we first draw π from the specified Beta distribution, and then draw $\gamma_t|\pi \sim \text{Ber}(\pi)$ for $t = 1, \cdots, T$. Then according to $\boldsymbol{\gamma}$ we draw $\beta_t|(\gamma_t = 1) \sim N(0, \sigma_\beta^2)$, or $\beta_t|(\gamma_t = 0) = 0$. Lastly, we draw error $\epsilon \sim N_T(\mathbf{0}, \Sigma)$ and construct $Y = L\boldsymbol{\beta}^T + \epsilon$ with fixed L. We use two effect sizes, $\sigma_\beta^2 = 5$ and $\sigma_\beta^2 = 1$. We suspect that the scenario with small effect sizes resemble the real data application. We also use two distributions $\pi \sim \text{Be}(0.05, 0.5)$ and $\pi \sim \text{Be}(1, 1)$ to see how parameter misspecification affects the performance of the algorithm. The simulation scenarios are summarized in

Table 2.1, and we generate 500 different Y for each scenario to run the algorithm with constant hyperparameter specifications. Note that for the real data application, we use a more conservative prior $a = 0.01, b = 0.5$. However, drawing $\pi \sim \text{Be}(0.01, 0.5)$ creates too few true signals, making it difficult to examine the algorithm's performance. So we deliberately create more true signals for our simulations by using a more liberal prior for π.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5 (Null)
N	71				
T	44				
missing proportion	67%				
Σ	$\Sigma_{ii} = 1$, $\Sigma_{ij} = 0.5$				
π	Be(0.05, 0.5)	Be(0.05, 0.5)	Be(1, 1)	Be(1,1)	0
σ^2_β	5	1	5	1	NA
(a, b)	(0.05, 0.5)				
Φ	$\Phi_{ii} = 1$, $\Phi_{ij} = 0$				
ν	N				

Table 2.1: Simulation settings. Scenarios 1 to 4 compare the algorithm's behavior for different effect sizes (σ^2_β) and hyperparameters (a, b). Scenario 5 runs a null simulation with $\boldsymbol{\beta} = 0$, and we observe the algorithm's resistance toward Type I error.

We first define some of the terms that are used to analyze the result of the algorithm. We observe at each iteration $\hat{\gamma}_{jst}$ and $\hat{\beta}_{jst}$ where $j = 1, \cdots, J$ is the iteration index after removing burn-in, $s = 1, \cdots, S$ is simulation index, and $t = 1, \cdots, T$ is the outcome index. We define posterior inclusion probability (PIP) as

$$p(\hat{\gamma}_{st} = 1) = \frac{\sum_{j=1}^{J} \hat{\gamma}_{sjt}}{J}.$$

We use the following definitions to analyze Type I and Type II errors given the PIP threshold

c.

$$\text{True Positive} : p(\hat{\gamma}_{st} = 1) \geq c \text{ and } \gamma_{st} = 1$$

$$\text{False Positive} : p(\hat{\gamma}_{st} = 1) \geq c \text{ and } \gamma_{st} = 0$$

$$\text{True Negative} : p(\hat{\gamma}_{st} = 0) \geq c \text{ and } \gamma_{st} = 0$$

$$\text{True Negative} : p(\hat{\gamma}_{st} = 0) \geq c \text{ and } \gamma_{st} = 1$$

We also define FDR and power given c. Although these concepts are fundamentally frequentist, they are useful when we compare the result with the univariate linear regression. We reject the null $\gamma_{st} = 0$ if PIP $\geq c$, and not reject the null if PIP $< c$.

$$\text{FDR}_c = \frac{\sum_{s,t} \mathbf{1}\{p(\hat{\gamma}_{st} = 1) \geq c \text{ and } \gamma_{st} = 0\}}{\sum_{s,t} \mathbf{1}\{p(\hat{\gamma}_{st} = 1) \geq c\}}$$

$$\text{Power}_c = \frac{\sum_{s,t} \mathbf{1}\{p(\hat{\gamma}_{st} = 1) \geq c \text{ and } \gamma_{st} = 1\}}{\sum_{s,t} \mathbf{1}\{\gamma_{st} = 1\}}$$

We also define posterior mean of β,

$$\bar{\beta}_{st} = \sum_{j:\hat{\gamma}_{sjt}=1} \frac{\hat{\beta}_{sjt}}{\sum_j \hat{\gamma}_{sjt}}$$

to check the algorithm's performance on the estimation of the effect size.

We first examine the number of false and true positives for varying PIP threshold c. Figure 2.2 (a) shows that all scenarios show consistent behaviors. Scenarios 1 and 2 has more false discovery rate at low 0 because π is drawn from $\text{Be}(0.05, 0.5)$ and there are not that many true signals in the data. When the effect sizes are small ($\sigma_\beta^2 = 1$, scenarios 2 and 4), power decreases more quickly as c increases. Even when hyperparameters are mis-specified ($\pi \sim \text{Be}(1, 1)$, secnarios 3 and 4), the algorithm performs well.

To evaluate the inference on β, for each simulation s and variable t, we compute the posterior mean and plot it against the true value in Figure 2.2 (b). For each scenario, we compute the FDR and power as defined in the previous section, and use PIP threshold c where FDR_c reaches 0.05. The red points are the ones not selected by the algorithm($p(\hat{\gamma}_{st} = 1) < c$), and the blue points selected ($p(\hat{\gamma}_{st} = 1) \geq c$). We also divide the true β values into bins and investigate the change in power in Table 2.2. PIP increases as the effect size increases, proving the calibration of PIP for variable selection.

Figure 2.2 (c) shows the power improvement compared to the univariate analysis. For the marginal linear regression, we record the $-log_{10}(p)$ values for the $24 \times 500 = 12000$ variables to test the null hypothesis $\beta = 0$. We discretize the log-transformed p-value threshold and compute FDR and power just as we do with posterior inclusion probability thresholds. Then we match the FDR level with the multivariate result to create Figure 2.2 (d) plot. The power of the proposed method is consistently higher than that of the marginal result when FDR level is fixed.

Figure 2.2 (d) shows the calibration of PIP as a selection criterion. We divide PIP into 5 bins and compute the mean of PIP and the proportion of true positives for each bin for each scenario. The higher the PIP, the higher the proportion of true positives. This means that we can decide on a PIP threshold to effectively control for type I error. Scenarios 1 and 2 show more inconsistent pattern compared to scenarios 3 and 4, and this is simply due to the size of true positives in the simulations. When π is drawn from $\text{Be}(0.05, 0.5)$, only around 20% of the variables are non-zero and they're divided into 10 bins. Especially in scenario 2, since the effect sizes are small ($\sigma_\beta^2 = 1$), a very small number of variables are placed into bins with PIP greater than 0.5.

We also run a null simulation where $\pi = 0$ and the rest of the data generating process

is the same. This is designed to check the algorithm's susceptibility to false positives. The result returned no variables with PIP higher than 0.95, and only one variable returned PIP higher than 0.9 out of $500 \times 24 = 12,000$ variables. This shows that the the algorithm is quite robust to false positives, and we believe 0.95 is a conservative enough threshold that can effectively control the error.

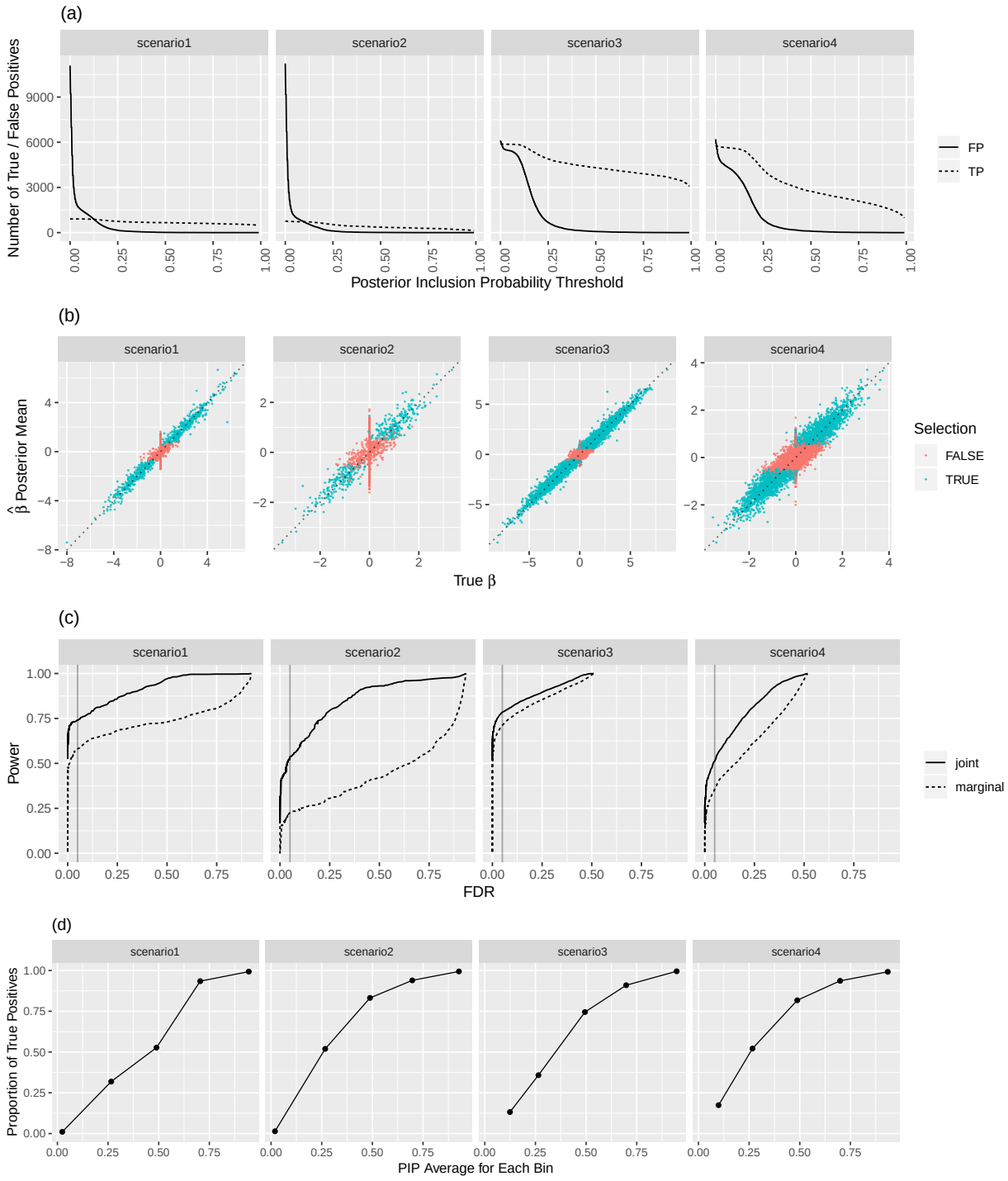


Figure 2.2: Results from simulation studies. (a) Number of false positives and true positives on varying PIP thresholds. (b) Average of $\hat{\beta}_j$ for iterations j with $\gamma_j = 1$. Red if $\hat{\gamma} = 0$, and blue if $\hat{\gamma} = 1$. (c) Power comparison with univariate analyses at a given FDR. (d) Calibration of PIP. We place each variable into one of 5 bins. Each point on the graph represents a single bin. x coordinate is the mean of the PIPs and y coordinate being the proportion of true positives within the bin.

	Scenario1	Scenario2	Scenario3	Scenario4
	$E(\hat{\gamma})$ (SD)	$E(\hat{\gamma})$ (SD)	$E(\hat{\gamma})$ (SD)	$E(\hat{\gamma})$ (SD)
$\|\beta\| = 0$	0.061 (0.001)	0.058 (0.001)	0.084 (0.001)	0.084 (0.001)
$\|\beta\| \in (0, 0.2]$	0.227 (0.009)	0.226 (0.01)	0.22 (0.009)	0.229 (0.008)
$\|\beta\| \in (0.2, 0.4]$	0.282 (0.011)	0.257 (0.011)	0.27 (0.01)	0.282 (0.009)
$\|\beta\| \in (0.4, 0.6]$	0.348 (0.014)	0.373 (0.014)	0.348 (0.012)	0.357 (0.011)
$\|\beta\| \in (0.6, 0.8]$	0.455 (0.015)	0.505 (0.014)	0.491 (0.014)	0.497 (0.013)
$\|\beta\| \in (0.8, 1]$	0.6 (0.015)	0.643 (0.014)	0.645 (0.013)	0.633 (0.013)
$\|\beta\| \in (1, 1.2]$	0.79 (0.011)	0.809 (0.01)	0.776 (0.011)	0.792 (0.009)
$\|\beta\| \in (1.2, 1.4]$	0.881 (0.008)	0.875 (0.009)	0.873 (0.007)	0.861 (0.008)
$\|\beta\| \in (1.4, 1.6]$	0.921 (0.006)	0.941 (0.005)	0.95 (0.003)	0.934 (0.005)
$\|\beta\| \in (1.6, 1.8]$	0.976 (0.002)	0.962 (0.003)	0.966 (0.002)	0.966 (0.002)
$\|\beta\| \in (1.8, 2]$	0.99 (0.001)	0.989 (0.001)	0.987 (0.001)	0.988(0.001)
$\|\beta\| > 2$	1 (0)	0.999 (0)	0.999 (0)	0.999 (0)

Table 2.2: Average of posterior inclusion probability (PIP) of each scenario given the effect size in simulated data along with the coverage probability of posterior distribution of β.

2.4 Application to GTEx Data

In this section, we aim to discover the genes whose expression levels are affected by the local or global ancestry. We study the expression levels of 32.006 genes from 71 African American individuals with reliable measurements of local and global ancestry across 24 tissues with more than 20 observations.

We use the proposed method to test the effects of both local and global ancestry for 32,006 genes. These genes were expressed in at least 2 tissues. We use the hyperparameters (Φ, ν, a, b) as specified in Section 2.3. Based on the simulations, we use the PIP threshold 0.95. The results are summarized in Table 2.3 and Table 2.4.

We also compare the result with simple linear regression that assumes inter-tissue independence of the expression levels and analyze the data separately for each tissue t. We used

the same demographical covariates including the two principal components of the expression level. We do not find any signal that stood out when we use FDR threshold 0.05 with Benjamini Hochberg procedure (Benjamini & Hochberg, 1995).

Tissue	Local	Global
Adipose Subcutaneous	Z98048.1, **FO393419.3**	
Adipose Visceral Omentum	TRAV21, AL354989.1	
Adrenal Gland	CYP3A5, AC139495.1, AC026369.2	**AP000255.1**, AL356966.1
Artery Aorta	MFGE8, RPS15AP36	AL096803.2 , AC011444.1 , AL589765.6
Artery Coronary	**CHP2**, AC090044.1	VN1R81P, BX255923.1 , **IGBP1P1**
Artery Tibial	IGLV1-51	SGK494, AL445435.1 , **IGBP1P1** , AL450263.1, LINC00930
Breast Mammary Tissue	MIR635	AC135507.1
Colon Transverse	**APCDD1L**	
Esophagus Mucosa	ASCL2	AL121655.1 , **AP000255.1**
Esophagus Muscularis		B4GALT6
Heart Atrial Appendace		MYOT
Lung	**CHP2**	
Nerve Tibial	KLB, ADPGK-AS1	
Skin Not Sun Exposed Suprapubic		PLN
Skin Sun Exposed Lower Leg	ZNF788, ADGRG5, **APCDD1L**, CBR3-AS1	CTSV, SEC14L6, AC015914.1
Stomach	**FO393419.3**	
Testis	AC011444.3, AC010327.3	
Thyroid	MYPN, LINC01301	SORCS1, HEMGN
Whole Blood	AC131056.3	

Table 2.3: Genes with PIP greater than 0.95 using the proposed Bayesian Variable Selection method. Left column is measuring the effect of local ancestry on gene expression level, and the right of global ancestry. Genes that occur more than once across many tissues are bold-faced.

Pathway	Genes
Immune Response	TRAV21, IGLV1-51
Metabolism	CYP3A5, MFGE8

Table 2.4: Among the genes selected in Table 2.3, genes that are part of two pathways of interest.

2.5 Discussion

We have developed a Bayesian variable selection method that can explain the relationship between a covariate and correlated multiple phenotypes. The non-informative prior for the proportion of variance explained (PVE) aids the interpretation, and the algorithm can also analyze highly incomplete data. This method allows us to analyze multi-tissue expression level data against a covariate, for example local ancestry, and it has a wide range of other possible applications.

The simulation section shows that the proposed method works better than the linear regression that assumes independence across the tissues, especially in scenario 2 where the signals are scarce ($\pi \sim \text{Be}(0.05, 0.5)$) and small ($\sigma_\beta^2 = 1$). In recent challenges in biology, single variable rarely explains a significant portion of trait variability, and it is common to search for weak and sparse signals, in which our proposed method shows an advantage.

The algorithm could take a few possible other directions. First, as briefly mentioned, we could use a prior $\boldsymbol{\gamma}$ that allows correlation. However, this can induce false confidence in the selection of $\boldsymbol{\gamma}$ compared to relying on the data to infer correlation. It can also impose more computational burden to the algorithm. Second, it is possible to expand this algorithm to consider multiple covariates simultaneously. However, it is common to focus on one explanatory variable, and it is straightforward to regress out other covariates beforehand.

This algorithm lacks an effective error control, which is difficult when the number of tests is as large as the number of genes. The simulation shows that the empirical FDR reaches 0.05 at approximately PIP=0.5 threshold. However, FDR is a frequentist concept, and it is difficult to apply it to the Bayesian variable selection problem. The null simulation of 500 data sets with 24 variables each returns 1 case with PIP higher than 0.95, and this multiplies fast as the number of tests increases to more than 30,000 genes. It is possible to use a much

more conservative prior, but it would violate our assumption that some genes have many cross-tissue signals. It is also possible to pre-select some of the genes to reduce the number of tests. Still, our method allows us to observe the top genes with the strongest signals, and it gives valuable biological insights to better understand African American genome and the effects of genetic ancestry.

CHAPTER 3

EFFECTS OF LOCAL ANCESTRY ON GENE EXPRESSION LEVEL AND THE EQTLS OF AFRICAN AMERICANS

3.1 Introduction

In this chapter, we expand our study to jointly analyze the transcriptome and genome of African Americans. We study the interaction effects of local ancestry and single nucleotide polymorphisms (SNPs) on the gene expression level to find "ancestry-eQTLs" that we define as the SNPs that influence the gene expression level differently based on the gene's local ancestry. We are ultimately interested in "ancestry-eGenes," defined as the genes that have at least one ancestry-eQTL in their cis-region. One biological illustration of this phenomenon would be when Africans and Europeans have developed different structures for a certain transcription factor (TF) as a result of evolutionary divergence, and African TF binds particularly well to the minor allele while European TF does not. In this case, the minor allele would have positive relationship with the expression level for African Americans, but not so for European Americans.

The main challenge in this analysis is computational. The genetic variation space is so large that even for simple linear models, we need methods that scale. Moreover, the SNPs are intricately correlated with one another by a mechanism called "linkage disequilibrium." To find the ancestry-eGenes and ancestry-eQTLs, we tackle two statistical challenges. First, we adopt computational methods called 'matrix-eQTLs' suggested in (Shabalin, 2012) to our studies of the interaction effects between SNPs and ancestry. Matrix-eQTL provides the method for a large-scale linear systems, and it can be applied to the our variation of the eQTL model. Second, we apply permutation test that preserves the linkage disequilibrium within the genomic data. We also apply approximate the null distribution of the test statistics to further improve the precision of inference without the computational burden that

comes with permutation tests.

We include European American samples for the analysis of admixed transcriptome by assigning 0 to both their global ancestry and local ancestry, in this chapter only. Expanding the sample is aimed at increasing the statistical power. We are interested in the interaction effects between genotypes and local ancestries, and since most of our African American subjects have global ancestry around 0.8, most of their genes have 1 or 2 for local ancestry and there are many genetic variants where no individual has local ancestry 0. Therefore, increasing sample size, if possible, can improve power, and it would especially help to include Europeans who have local ancestry 0 everywhere. An illustrative example of increase in power is presented in the Section 3.4.

The remainder of this chapter is organized as follows. In Section 2, we describe the model and present the methods that circumvent the two statistical issues of complex dependence structure and computational burden. In Section 3, we present the real data analysis result for discovering ancestry-eGenes using the Muscle Skeletal tissue of GTEx data. We present the discovered genes' functional clustering results for meaningful biological interpretations. We end with a discussion about the strength and limitation of our analysis.

3.2 Methods

In this section, we introduce our model and present exploratory analyses to show that the model is statistically sound.

3.2.1 Model

We use the following full model for our primary goal of finding gene k and its cis-SNP s where $\theta_{k,s}$ is non-zero.

$$y_k \sim \mu_{k,s} + \alpha_{k,s}\mathbb{1}_{EA} + \beta_{k,s}A + \gamma_{k,s}L_k + \lambda_{k,s}G_s + \theta_{k,s}LG_{k,s} + \nu_{k,s}X + \epsilon_{k,s}$$

$y_k \in \mathbb{R}^{N\times 1}$ is a vector of the gene expression level of $N = 356$ individuals for gene k. $\mu_{k,s}$ is a mean term, and $\mathbb{1}_{EA}$ is an indicator function assigning 1 to European Americans and 0 to African Americans. This allows different means for African Americans and European Americans. A is a $N \times 1$ vector for global ancestry of those individuals where each value is between 0 and 1, and $L_k \in \{0, 1, 2\}^{N\times 1}$ is the vector for the subjects' local ancestry of the gene k, defined as the local ancestry of the SNP that is closest to the middle of the gene. G_s is a $N \times 1$ vector for the genotype of SNP s. Here we limit s to be in a cis-region of the gene k, where 'cis' area is between (*gene start site - 1 million bases*) and (*gene end site + 1 million bases*). $LG_{k,s}$ is also a $N \times 1$ vector for the interaction term, which is an element-wise multiplication of G_s and L_k. Lastly, $\epsilon_{k,s} \sim N(0, \sigma_k^2)$ is the $N \times 1$ vector for the error term. $\alpha_{k,s}$, $\beta_{k,s}$, $\gamma_{k,s}, \lambda_{k,s}, \theta_{k,s}$ are all scalar coefficients. X is the covariate matrix $\in \mathcal{R}^{N\times 4}$ of age, sex and the first two principal components of the expression level matrix, and ν is a coefficient vector of length 4.

Here, we formally define the ancestry-eGenes using the notation and the model above. Ancestry e-Genes are genes k with non-zero interaction effect ($\theta_{k,s} \neq 0$) for at least one SNP s. The interpretation of non-zero $\theta_{k,s}$ is that minor allele has different effects on the expression of a gene with different local ancestry. An example is illustrated in Figure 3.1. One individual can have purely European ancestry ($L = 0$) on a gene HLA-C, and another individual can have purely African ancestry ($L = 2$) for the same gene. Then, a minor allele on the SNP rs2523578 can decrease the gene expression level of the first individual

and increase that of the second individual. We define ancestry-eQTLs as these SNPs that differentially affect genes by their local ancestries, and ancestry-eGenes as the genes that have at least one ancestry-eQTL in its cis-region.

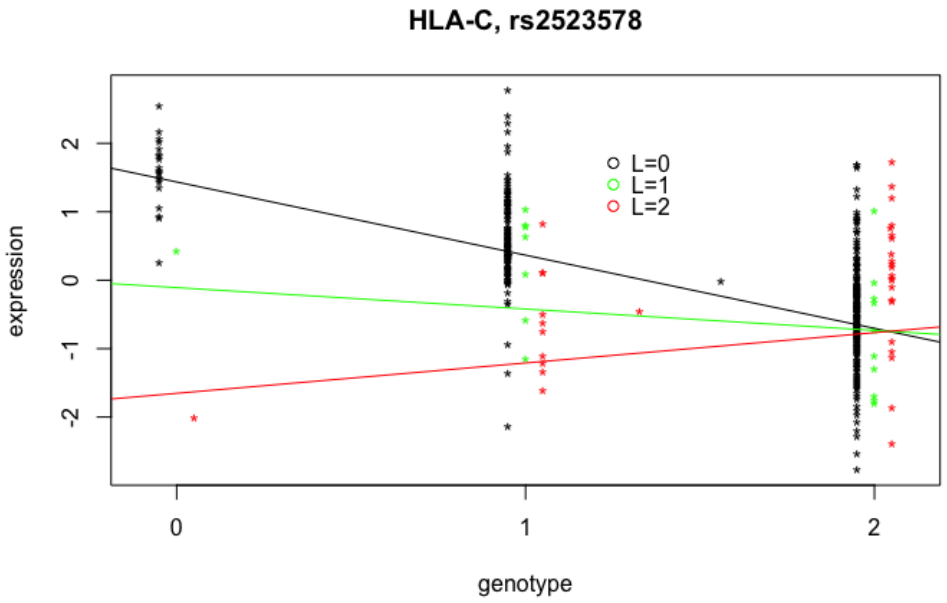


Figure 3.1: The interaction effect on the expression of HLA-C between genotype of SNP rs2523578 and the gene's local ancestry. Some genotypes have been imputed by GTEx (The GTEx Consortium, 2015)

We are assuming additive effects for both local ancestry and genotypes, so we will treat them as quantitative variables. The interaction variable $LG_{k,s}$ is defined as the element-wise multiplication of the genotype vector G_s and local ancestry L_k vector.

Below, we present a series of tests designed to verify that under the above model, AA and EA are homogeneous.

Constant Variance

In order to check that EAs and AAs can be included in the same linear model without breaking the assumption of homogeneity, we tested if the two groups' residuals have the same variance under the model $y_k \sim \mu_k + \beta_k A + \epsilon_k$. We used the fact that the expression level follows normal distribution very closely: 22241 genes out of 22248 have Shapiro statistic > 0.99 and all genes have Shapiro statistic over 0.96. Denoting r_{AA} as the residual vector in African Americans and r_{EA} as the residual vector in European Americans, we tested the ratio of the variance of the residuals, $\frac{Var(r_{AA})}{Var(r_{EA})}$ which should follow $F_{N_{AA}-1,n_{EA}-1}$ given normality of r_{AA} and r_{EA} where N_{AA} is the sample size of African Americans and European Americans, i.e. 55 and 301 respectively. The 22,248 p-values from the F test for each gene is summarized in S.Figure 6.1. The distribution of p-values is very close to uniform, and there was no particularly strong signal under FDR=0.05 indicating non-constant variance from any genes.

Same Mean Test

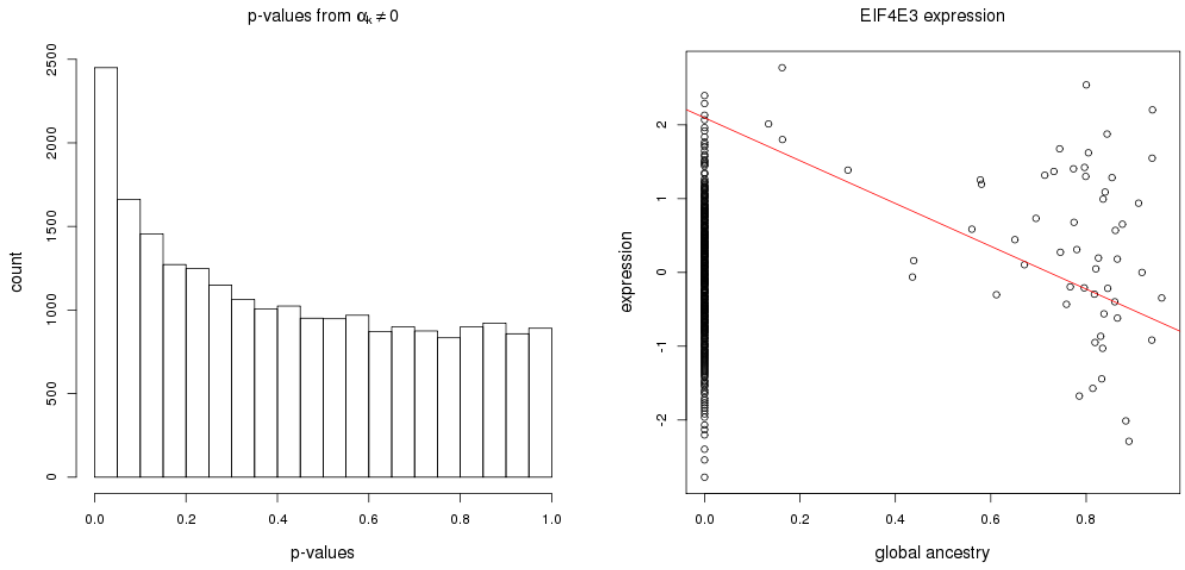
Next, we checked if an extra term of indicator $\mathbb{1}_{EA}$ is necessary. More specifically, we compared the following two models.

$$y_k \sim \mu_k + \beta_k A$$

$$y_k \sim \mu_k + \alpha_k \mathbb{1}_{EA} + \beta_k A$$

This aims to see if the intercept of the linear model built on only African Americans is consistent with the mean expression level of European Americans whose global ancestry is set to 0. The resulting histogram of p-values testing $H_0 : \alpha_k = 0$ is shown in Figure 3.2. The histogram shows clear signals of two different linear models and suggests that excluding the indicator term can lead to a completely different model. The worst case scenario (lowest p-value) is observed in gene ST20, as illustrated in Figure 3.2 (b), where excluding the indicator term incorrectly identifies the sign of the ancestry effect, which can be problematic. The red

line is the model only for African Americans, and the black lines is when AA and EA are jointly modeled.



(a) *(b)*

Figure 3.2: (a) Histogram of p-values from the t-test for the significance of the extra mean term. (b) The gene with the strongest signal from the indicator variable. The red line is the fitted regression line only using African American samples.

Note that, in the model with only global ancestry as a main variable, having an extra indicator term essentially builds a separate model for African Americans. Consider the following two models.

$$\text{M1: } y_k \sim \mu_k + \alpha_k \mathbb{1}_{EA} + \beta_k A$$

$$\text{M2: } y_{k,1:55} \sim \mu_{k,1:55} + \beta_{k,1:55} A_{1:55}$$

where M2 is the linear model fit only to the 55 African American subjects. By construction, the estimated mean term μ and the slope for global ancestry β are identical for M1 and M2. Therefore, the power to detect the significance of $\hat{\beta}$ only depends on the standard error of the estimated β which is defined as $\hat{\sigma}\sqrt{(X^T X)^{-1}_{ii}}$ where X is the design matrix for covariates

and i is the column index of the variable of interest, i.e. $i = 2$ for M1 and $i = 1$ for M2. Also note that, again by construction, A and $\mathbb{1}_{EA}$ are orthogonal, meaning that $(X^T X)^{-1}$ in model M1 is a diagonal matrix. Therefore, the difference between standard error of estimated β in two models M1 and M2 only depends on the estimated error variance $\hat{\sigma}$ defined as

$$\hat{\sigma}^2 = \frac{(y - \hat{y})^T (y - \hat{y})}{N - p}$$

which is the sum of residuals squared divided by degrees of freedom. Therefore, under the constant variance assumption that we verified earlier, adding an indicator variable is equivalent to modeling with the 55 African American individuals when we only have global ancestry as the covariate.

Two Sample t-test

Next, we checked if treating European Americans with local ancestry zero is appropriate. This means, if our model is correct, when global ancestry is controlled, African Americans with local ancestry 0 should have the same gene expression level as European Americans for that gene. To check this, we first obtained residuals from the linear regression of expression level against global ancestry, the indicator variable for European Americans, and other covariates of gender, age, and principal components of the expression level. Then we conducted the two sample t test to see if the two groups have the same mean under the assumption of equal variance. Note that out of 22,248 genes, 510 had only 0 or 1 African American subject who had 0 as local ancestry, so that the two sample t test was not even feasible. The result is summarized in S.Figure 6.2. No gene was rejected with FDR level at 0.05. The curved shape of p-value histogram is due to gene-gene correlations.

This shows that jointly modeling African Americans and European Americans maintains the intended interpretation. When the added indicator variable and the global ancestry is controlled, as we are planning to do so, we can say that African Americans with local ancestry

0 is homogeneous with European Americans who always have local ancestry 0. Therefore, this allows us to have a much bigger sample size to detect the interaction between local ancestry and genotype.

3.2.2 Computational Challenge

Individually testing the model

$$y_k \sim \mu_{k,s} + \alpha_{k,s}\mathbb{1}_{EA} + \beta_{k,s}A + \gamma_{k,s}L_k + \lambda_{k,s}G_s + \theta_{k,s}LG_{k,s} + \nu_{k,s}X + \epsilon_{k,s}$$

for gene k and SNP s for each gene-SNP pair requires around 10^{10} linear regressions. Studies have suggested various methods to speed up the process. Matrix eQTL software by (Shabalin, 2012), for example, redesigns the linear regression into matrix operations. This can be applied to the above model with a small modification.

We denote N as the sample size 356 and S_k as the number of cis-SNPs in gene k. Below, I consider the method for only one gene. Since we can parallelize the procedure in a gene-level process, it is computationally tractable.

Our test statistic for each gene-SNP pair is the t-value computed from the Pearson correlation coefficient between the interaction term $LG_{k,s}$ and the expression y_k. Note that if two vectors have mean 0 and sums of squares 1, the inner product is the same as the Pearson correlation coefficient. Adjusting this marginal correlation with degrees of freedom can easily lead to the t-value, which we will call $\rho_{k,s}$. Our goal is therefore to get the marginal correlation between the two terms while controlling for other covariates.

First, we center all the variables to mean 0 so that we have no need to consider the intercept μ term. Then marginal correlation between two variables of interest with other covariates

controlled can be acquired by comparing the residuals

$$r_{k,s}^{(1)} = y_k - \hat{y_k}$$

$$r_{k,s}^{(2)} = LG_{k,s} - \hat{LG}_{k,s}$$

where $\hat{y}$ is the fitted values from linear model

$$y_k \sim \mu_{k,s} + \alpha_{k,s}\mathbb{1}_{EA} + \beta_{k,s}A + \gamma_{k,s}L_k + \lambda_{k,s}G_s + \nu_{k,s}X$$

and $\hat{LG}_{k,s}$ is the fitted values from linear model

$$LG_{k,s} \sim \mu_{k,s} + \alpha_{k,s}\mathbb{1}_{EA} + \beta_{k,s}A + \gamma_{k,s}L_k + \beta_{k,s}G_s + \nu_{k,s}X$$

Then, $r_s^{(1)}$ and $r_s^{(2)}$ will again be standardized, so that the inner product $< r_{k,s}^{(1)}, r_{k,s}^{(2)} >$ is the Pearson correlation coefficient. This computation can be done in a large matrix with all s; matrix $r_k^{(1)}$ and $r_k^{(2)}$ with dimension $N \times S_k$ can be multiplied element-wise, and then the column sums will be the inner products for each SNP.

However, getting the matrix $r_k^{(1)}$ requires running S_k number of linear regressions because the covariates include the genotype for each SNP. Alternatively, we can manually compute the residuals by solving normal equations, but that includes inverting the covariance matrix which can be burdensome. So, we use Gram-Schmidt orthogonalization of all the covariates, including the genotypes. First, create a covariate matrix with only the variables that do not change for each SNP:

$$C_k = \begin{bmatrix} L_k & A & X \end{bmatrix} \in \mathbb{R}^{N \times p}$$

and make $\tilde{C_k}$ by orthogonalizing each column. In C_k, L and A are vectors but X is a matrix with relevant covariates including gender, age, and the principal components of the

expression level matrix. Then we will orthogonalize each column of the cis-SNP matrix of G. More specifically, first orthogonalize C_k by QR decomposition. C has the dimension of only 356×6 (A, L, gender, age, two principal components of expression level matrix), and this needs to be done only once, so the efficiency of orthogonalization doesn't matter much in our process. Call the six orthogonalized columns $\tilde{c}_1$ .. $\tilde{c}_6$. Then for each column of $G_{k,s}$ of the genotype matrix G_k, use Gram Schmidt algorithm to make

$$\tilde{G}_{k,s} = \frac{1}{r_{ss}} \left(G_{k,s} - \sum_{j=1}^{6} \tilde{c}_j^T G_{k,s} \tilde{c}_j \right)$$

where r_{ss} is the normalizing constant. We can easily vectorize this to apply for all SNPs s:

$$\tilde{G}_k = \left(G_k - \tilde{C}_k (\tilde{C}_k^T G_k) \right) \cdot \text{diag}(1/r_{ss})_{s=1,..,S_k}$$

where $\tilde{G}_k, G_k \in \mathbb{R}^{N \times S_k}$ and $\tilde{C}_k \in \mathbb{R}^{N \times 6}$.

Now, we have orthogonalized all covariates for the linear regression for each SNP s, so we can compute the residuals $r_{k,s}^{(1)}$:

$$r_{k,s}^{(1)} = y_k - \tilde{C}_k (\tilde{C}_k^T y) - \tilde{G}_{k,s} (\tilde{G}_{k,s}^T y)$$

This can also be easily vectorized by replacing $\tilde{G}_{k,s}$ by $\tilde{G}_k$ by using element-wise operations.

Computing $r_{k,s}^{(2)}$ is also a very similar process, and we can re-use the orthogonalized matrices of genotypes and covariates. Therefore, through matrix operations, we can efficiently compute $\rho_{k,s}$ for all gene-SNP pairs.

3.2.3 Inference

Although we are computing the t-value $\rho_{k,s}$ for marginal correlation for each gene-SNP pair, our initial goal is to find the ancestry-eGenes. Therefore, we define the test statistic for each

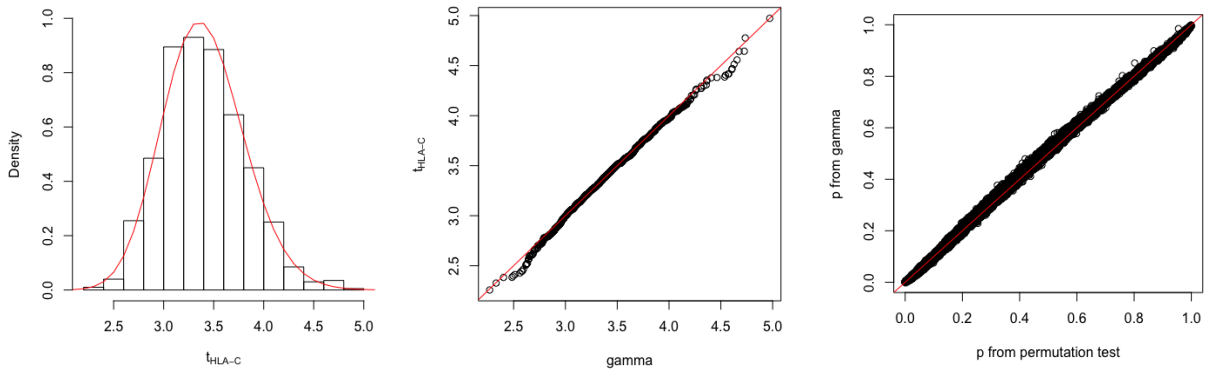
gene as the maximum of the absolute value of $\rho_{k,s}$ and define this test statistic for gene k as $t_k = max_s(|\rho_{k,s}|)$.

To get the significance of this test statistic t_k, we re-arranged all variables that are not affected by LD to perform the permutation test. In this case, they are everything except the interaction term and the genotype term. Then we conducted the same test for each permutation, and recorded the test statistic t_{k,π_1}, .. $t_{k,\pi_{1000}}$ where π_1, .., π_{1000} shows the 1000 different permutations for each gene k.

Normally we would simply calculate the quantile of t_k against $t_{k,\pi}$, but note that in order to correct for the multiple comparisons, we need p-value to be finely tuned. That means we will have to conduct at least a million permutations for each gene, which is practically infeasible. Therefore, we only conducted 1,000 permutations and fitted an gamma distribution on the test statistics under the null, i.e. under random permutation. Figure 3.3 shows that, for gene HLA-C, the $t_{k,\pi}$ matches gamma distribution very closely, where parameters for gamma distribution were fitted through the method of moments. Figure 3.3 also shows that the p-values derived from fitting gamma is very close to the p-values the traditional permutation test.

3.3 Application to GTEx data

The histogram of p-values are shown in Figure 3.4. There were 30 genes who had no genotyped cis-SNPs or the interaction term included a perfect collinearity, so they were removed from the analysis, leaving us 22,218 genes. The tail part near $p = 1$ shows an unusual behavior, and shows that the null distribution of p-value might not be uniformly distributed. This usually suggests a possible unexplained confounder during our modeling procedure, but given that we controlled for population structure through global and local ancestry and for the technical



(a) *(b)*

Figure 3.3: (a) Comparing the gamma distribution fitted by method of moments and the empirical distribution of test statistics from permutation test. This is the result of the same gene with Figure 3.1, HLA-C. (b) Plot of the p-values from the permutation test versus the p-values from the fitted gamma distribution. As we expected, they are very close to the $y = x$ line.

artifacts through the principal components of the expression level matrix, and also given the Figure 3.3 (b) where our fitted p-values through gamma distribution closely matches the p-values from the permutation test, there is also a high possibility that the signals are real. Taking these p-values as true result, we used Benjamini Hochberg procedure with $\alpha = 0.05$ to identify 201 genes as ancestry-eGenes, and with $\alpha = 0.01$ to identify 110 genes.

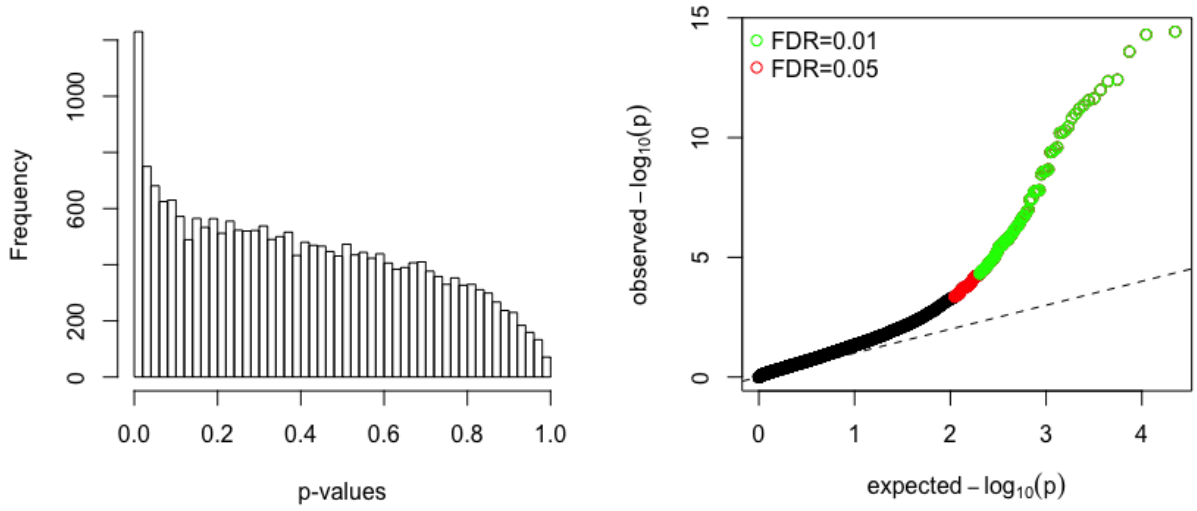


Figure 3.4: The left plot shows the histogram of the p-values. The plot on the right shows the qq-plot of p-values compared to the uniform distribution. The green points are rejected with FDR level 0.01 and green and red points are rejected with FDR level 0.05.

Software DAVID takes a gene list as an input and returns a result of functional analysis by extracting biological roles of the genes. It also lets the users to specify the background alleles for a context-specific interpretation of a gene list (Huang et al., 2009). DAVID recommends the users to have more than a hundred genes, so we used the 201 signals that were found with FDR level 0.05. For background reference genes, due to the size restriction of the software, we used 3,000 randomly selected genes from the genes that were expressed (RPKM > 0.1) in muscle skeletal tissue. DAVID's functional clustering result showed that the two most enriched clusters' keywords were 'MHC' and 'glutathione'. To reaffirm the result, since it is easy to identify the MHC genes by their locations, (chromosome 6 28,477,797-33,448,354), we counted all signals in MHC regions and conducted the Fisher Exact Test to verify the enrichment. Out of 201 rejected genes, 14 were in the MHC region, and with the odds ratio 12.23 and the p-value of 8.56e-13, the test showed enough evidence for the high enrichment

	MHC	non-MHC
Signals (FDR<0.05)	14	187
Non-signals	165	21852

Table 3.1: Fisher Exact Test for the enrichment of genes in the MHC region in the signals of ancestry-eGenes.

of MHC genes within the signals (Table 3.1). The 14 genes in the MHC region were HLA-C, HLA-DRB5, HLA-DRB1, HLA-DQA2, HLA-DQB2, HLA-K, HLA-U, HLA-DQB1, STK19B, MICA, XXbac-BPG181B23.7, HLA-DPA1, C4A, HLA-DPB1, in the order of the signal strength.

Past works have shown that autoimmune disorders display widely heterogeneous behavior based on different ethnicities (Mori et al., 2005)(Davidson & Diamond, 2001). Given that the MHC region and human leukocyte antigens (HLA) genes are responsible for making receptors for pathogens and play a large role in immunity, this result can be an interesting addition that corroborates such phenomenon.

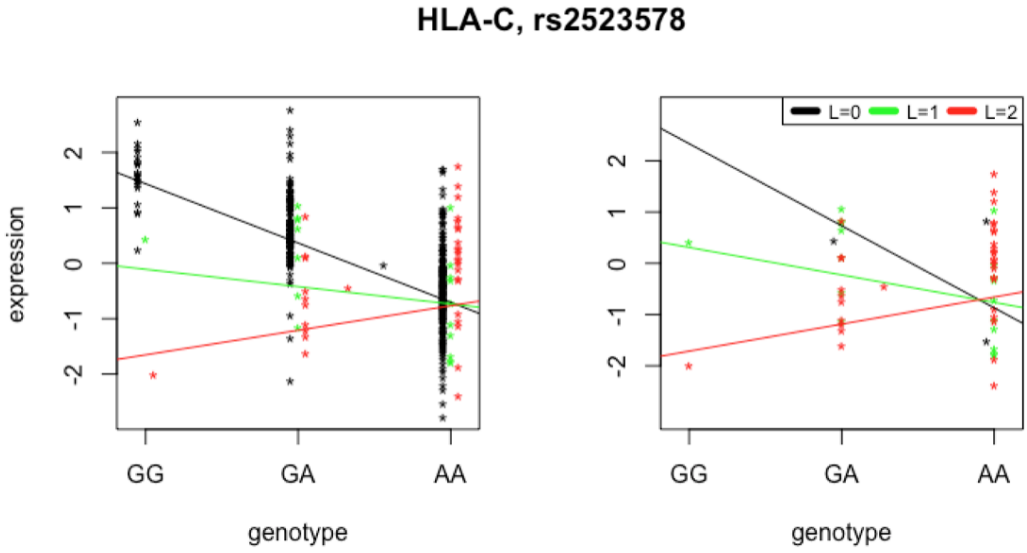
The next enriched group was the keyword glutathione, an important endogenous antioxidant. Especially in the muscle skeletal tissue, it has been suggested that it reduces the risk of cellular injury, improves performance, and delays muscle fatigue (Powers et al., 1999), but the mechanism through which ancestry shows an interaction effect with minor allele should be investigated further.

3.4 Discussion

We have modeled the expression data and ancestry not only with admixed individuals but also with pure European Americans by assigning 0 to their ancestry, and Figure 3.5 illustrates the impact in power. Under joint modeling, the interaction effect is significant with p-value $1.97 \cdot 10^{-12}$, but when only the admixed population was used, the p-value was 0.003 which is not low enough after correcting for the multiple comparisons. Therefore, we successfully

increased power by including European Americans in our model.

However, we still believe that more samples of admixed population would be helpful for our tests, and although we laid some grounds on the validity of joint modeling with European Americans, only investigating the genome of admixed populations could lead to different results due to unexplained non-genetic variables like lifestyle difference.



	Estimate(SE)	p-value
Intercept	1.44 (0.41)	6e-04
A	0.71 (0.60)	0.24
L	-1.54 (0.26)	5e-09
G	-1.07 (0.06)	9e-50
LG	0.76 (0.10)	2e-12

	Estimate(SE)	p-value
Intercept	2.33 (1.07)	0.03
A	0.30 (0.74)	0.68
L	-2.02 (0.56)	7e-04
G	-1.60 (0.55)	0.01
LG	1.06 (0.34)	3e-03

Figure 3.5: Increase in power after adding European Americans. The left plot and the table show the result of fitting linear model on both African Americans and European Americans, while the right plot and table are the result of fitting on only 55 African Americans. The interaction effect and the genotype effect both lose significance after multiple testing adjustments.

There could be several biological interpretations for genotype-ancestry interaction. For example, any difference between populations in regulatory elements such as transcription factors

or binding sites could have caused the interaction. Another possible explanation is a SNP-SNP interaction. If one variant is a strong ancestry-informative-marker, an interaction with this variant would mimic the effect of the interaction with ancestry. Yet another scenario is different linkage disequilibrium pattern among populations. Consider an eQTL (effective regardless of ancestry) and a SNP that is differentially correlated with eQTL by population — for instance, Africans have strong LD but Europeans have weak LD. Then, the SNP would have strong effects on the expression level for Africans butnot in Europeans, leading to the signal for interaction. In order to uncover the exact molecular mechanism of the interaction effects, further research is required.

We believe our gene list of significant ancestry-genotype interaction provides new insights about the population structure difference in gene regulatory network. We hope further empirical research uncovers the in-depth relationship between ancestry and gene regulatory procedures for these candidate SNP-gene pairs to better understand population differentiation.

CHAPTER 4

EFFECTS OF GLOBAL ANCESTRY ON THE GENE CO-EXPRESSION OF AFRICAN AMERICANS

4.1 Introduction

As a next step of analyzing the transcriptome data of admixed population, we examine the gene coexpression, the covariance structure of gene expression data. Gene coexpression shows how genes are functionally connected and provides insights into the design of the transcriptional regulatory system. Ideally, such a complicated biological system can be fully understood through longitudinal observations in multiple and diverse cell types that capture the dynamics of the system. In reality, however, such comprehensive measurements are often unavailable or too expensive, and the expression dynamics must be captured instead through cross-sectional or tissue-specific data sets. In such cases, investigating the dependence structure can be useful. The dependence structure can be especially valuable for characterizing how few key genes are connected to the rest of the transcriptome. For example, we can focus on one transcription factor — genes that help turn transcription of genes on and off — and study how it is connected to its target genes. To further investigate this problem, we define "local connectivity" of a transcription factor as its overall connectivity to its target genes.

This chapter investigates how local connectivity varies across genetic ancestry. More specifically, it studies the gene coexpression of African American subjects to identify candidate transcription factors whose effects on their target genes vary with the proportion of African ancestry in their genome. This analysis will lead to a better comprehension of how genes are differentially regulated in distinct populations.

The above biological problem can be investigated using multivariate statistical models of gene expression with a covariance structure (characterizing connectivity) that depends on

one or more features (such as ancestry). This chapter focuses on testing the contribution of ancestry on the covariance matrix, and we start from its simplest form by studying the expression levels of two genes. We construct a statistical model that can explain how their correlation varies against genetic ancestry and use that to test if the correlation is constant across conditions. We generalize it to the local connectivity of a transcription factor by meta-analyzing the pairwise statistics. Note that covariance modeling for multivariate data is important in many applications outside the field of genetics. Variance modeling has been widely studied in the context of heteroskedasticity (Breusch & Pagan, 1979; Glejser, 1969; White et al., 1980), and correlation modeling under discrete conditions has been studied in the context of the differential network (Ideker & Krogan, 2012), but dynamic correlation modeling has been less explored.

Li (2002) and Li et al. (2004) addresses the most similar scientific problem to ours, using the term "liquid association" (LA) to conceptualize the internal evolution of the coexpression pattern for a pair of genes. They analyze the coexpression that changes across different unobserved cellular states that are represented by the expression level of another gene as a proxy. Other studies have built on the liquid association to better identify cell states that affect coexpression (Yan et al., 2017; Yu, 2018), most focusing on expanding the test to genome-scale. However, methods based on liquid association have some limitations. First, it restricts the covariate to be a 1-dimensional vector, and cannot be generalized to more realistic scenarios. Second, it treats the covariate as a random variable that follows a normal distribution, which genetic ancestry does not, so it cannot be used for our application. Third, it only tests the linear relationship between the covariate and the coexpression. Lastly, the corresponding test statistic does not have a closed-form null distribution and requires a permutation test, leading to computational inefficiency.

This chapter propose a methodology for the continuously-varying covariance problem.

We apply traditional Rao's score test for heteroskedasticity (Breusch & Pagan, 1979) where the null hypothesis is that the coexpression does not vary with the covariate. This method is generalizable to non-normal, multivariate covariates, and it is also applicable to a non-linear relationship between the variance and the covariate. Moreover, the score test statistic asymptotically follows a chi-squared distribution, and hence it is easily expandable to a large number of tests without excessive computational burden. Subsequently, we tackle the local connectivity problem by expanding the scope of the problem from the relationship of two genes to the relationships between one gene and multiple genes by combining pairwise test statistics. When the number of genes in the local cluster is smaller than the sample size, the desired statistical properties apply to the new combined test statistic as well.

The rest of this chapter is organized as follows. First, we lay out the framework for the score test that investigates whether the covariance between bivariate normal variables varies against a continuous covariate X. Then we propose a way to combine the pair-wise test statistics for one gene and test the global null that the local connectivity of one variable does not change with genetic ancestry. In the simulation section, we show that the proposed method has distinct advantages compared to alternatives such as the likelihood ratio test or liquid association. Finally, we share the analysis results of African American's transcriptome using GTEx data for African Americans' transcriptome and genome. We end with a discussion about limitations of the method, possible future directions, and potential applications to fields outside genetics.

4.2 Methods

4.2.1 Test for Two Genes

Consider 2-dimensional data for N subjects $\boldsymbol{y}_i \in \mathbb{R}^2$, $i = 1, 2, \cdots, N$ independently following a bivariate normal distribution. There are two covariate matrices $Z \in \mathbb{R}^{N \times R}$ and $X \in$

$\mathbb{R}^{N\times P}$, for the mean term and the variance term, respectively. X is the vector or matrix that quantifies genetic ancestry, while Z is an appropriate covariate matrix that takes into account the mean effect, with the unity in its first column as the intercept. They are both assumed to be full rank. We notate each element of X and Z as $\{x_{ip}\}_{i=1,p=1}^{N,P}$ and $\{z_{ir}\}_{i=1,r=1}^{N,R}$, and consider the following model

$$\begin{bmatrix} y_{i1} \\ y_{i2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{z}_i^T\boldsymbol{\beta}_1 \\ \boldsymbol{z}_i^T\boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix} \\ \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho(\boldsymbol{x}_i) \\ \rho(\boldsymbol{x}_i) & \sigma_2^2 \end{bmatrix} \right) \tag{4.1}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are length R vectors for the mean term parameters. The variances σ_1^2 and σ_2^2 are scalars while ρ is a function. We define $\boldsymbol{\alpha} \in \mathbb{R}^P$ and a scalar α_0 as the parameters that characterize the impact of the covariates on the correlation through the function ρ:

$$\rho(\boldsymbol{x}_i) = \rho(\boldsymbol{x}_i^T\boldsymbol{\alpha} + \alpha_0) \tag{4.2}$$

The parameter of interest is $\boldsymbol{\alpha}$ while all others — α_0, $\boldsymbol{\beta}$, σ_1^2, σ_2^2 — are nuisance parameters. We develop here methodology for testing the following null hypothesis:

$$H_0 : \boldsymbol{\alpha} = \mathbf{0}. \tag{4.3}$$

Under the null hypothesis, $\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha} + \alpha_0) = \rho(\alpha_0)$ is a constant. ρ can take any linear or non-linear form, and the only assumptions required are linearity and additivity in (4.2) which are standard, so (4.1) is a flexible framework that can take be used for many forms of heteroskedasticity.

In the context of gene coexpression in admixed populations, $\boldsymbol{y}_i$ is gene expression level of an admixed individual i at two selected genes, and $\boldsymbol{x}_i$ is a P-dimensional covariate matrix for individual i that holds information about genetic ancestry. It can be a scalar that represents

the proportion of ancestry in the genome, a vector of the first few principal components of the genotypes, or a vector of local ancestry at multiple loci. In the application example in section 4.4, we focus on the global ancestry scalar $\boldsymbol{x}_i$ for straightforward interpretability.

There are two well-established tools for testing the null hypothesis (4.3): the likelihood ratio test and Rao's score test (Breusch & Pagan (1979)). The likelihood ratio test has a few disadvantages compared to the score test. It requires the full specification of the function ρ to estimate the maximum likelihood estimate (MLE) of $\boldsymbol{\alpha}$ both under the null hypothesis and under the alternative hypothesis. Possible choices for ρ include any kind of sigmoid function bound to $(-\sqrt{\sigma_1^2\sigma_2^2}, \sqrt{\sigma_1^2\sigma_2^2})$ such as logistic function, hyperbolic tangent function, or any cumulative distribution supported on the whole real line. This modeling strategy leads to one disadvantage. If ρ is highly misspecified, we sacrifice statistical power. Another disadvantage is that most of the reasonable assumptions for ρ, such as the sigmoid functions mentioned above, do not lead to a closed form MLE of $\boldsymbol{\alpha}$ under the alternative hypothesis. It would require us to numerically optimize the likelihood, leading to computational inefficiency, especially when the test space is large as in our application of gene coexpression. These limitations are also demonstrated in Section 4.3.

On the other hand, Rao's score test, unlike the likelihood ratio test, only requires the MLE of $\boldsymbol{\alpha}$ under the null hypothesis (Rao & Statistiker, 1973). Moreover, under our linear and additive model ($\rho(\boldsymbol{x}_i) = \rho(\alpha_0 + \boldsymbol{x}_i^T\boldsymbol{\alpha})$), the test statistic does not depend on the form of ρ while maintaining its asymptotic properties as long as ρ is twice differentiable. In order to test (4.3), we expand the result from Breusch & Pagan (1979) to derive the test statistic.

Proposition 1 *Consider the model in (4.1). If the non-diagonal term of Σ follows the function ρ as defined in (4.2), then Rao's score statistic does not depend on the unknown function ρ and asymptotically follows χ^2_P under the null hypothesis (4.3).*

Rao's score statistic is denoted by q and satisfies:

$$q = \phi^T \Psi^{-1} \phi \tag{4.4}$$

where

$$\phi = \frac{\hat{\rho}(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)\sum \tilde{\boldsymbol{x}}_i + (\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}^2)\sum_i \hat{u}_{i1}\hat{u}_{i2}\tilde{\boldsymbol{x}}_i - \hat{\sigma}_1^2\hat{\rho}\sum \hat{u}_{i2}^2\tilde{\boldsymbol{x}}_i - \hat{\sigma}_2^2\hat{\rho}\sum_i \hat{u}_{i1}^2\tilde{\boldsymbol{x}}_i}{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)},$$

$$\Psi = (\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}^2)(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)\sum_i \tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^T - \frac{4\hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}^2}{N}(\sum \tilde{\boldsymbol{x}}_i)(\sum \tilde{\boldsymbol{x}}_i^T),$$

and $\hat{u}_{i1}$ and $\hat{u}_{i2}$ are OLS residuals from (4.1) under the null (4.3), and $\hat{\sigma}_1$, $\hat{\sigma}_2$, and $\hat{\rho}$ are maximum likelihood estimates also under the null (4.3).

$$\hat{\sigma}_1^2 = \frac{1}{N}\sum_{i=1}^{N}\hat{u}_{i1}^2, \quad \hat{\sigma}_2^2 = \frac{1}{N}\sum_{i=1}^{N}\hat{u}_{i2}^2, \quad \hat{\rho} = \rho(\hat{\alpha}_0) = \frac{1}{N}\sum_{i=1}^{N}\hat{u}_{i1}\hat{u}_{i2}$$

$$\hat{\boldsymbol{u}}_1 = \boldsymbol{y}_1 - Z(Z^TZ)^{-1}Z^T\boldsymbol{y}_1, \qquad \hat{\boldsymbol{u}}_2 = \boldsymbol{y}_2 - Z(Z^TZ)^{-1}Z^T\boldsymbol{y}_2, \quad \tilde{\boldsymbol{x}}_i = \begin{bmatrix} 1 & \boldsymbol{x}_i \end{bmatrix}^T$$

Above statistic has two advantages. It does not depend on the unknown function ρ, so it is flexible under many shapes of heteroskedasticity. It also has low computational cost; every component of q is easily acquired from the data, and it asymptotically follows χ_P^2 (Breusch & Pagan (1979)).The detailed derivation and proof for Proposition 1 are in Appendix A (Section 6.2.1).

Note that the introduced test statistic has convenient asymptotic properties, but the inference might not be precise under a finite sample size. The error is in the order of N^{-1} (Harris, 1985), and various Monte Carlo experiments showed that the test rejects the null hypothesis less frequently than indicated by its nominal size (Godfrey, 1978; Griffiths & Surekha, 1986). In response, corrections have been suggested (Cribari-Neto & Ferrari, 2001;

Harris, 1985), and we apply the method from Honda (1988) to ensure the validity of the asymptotic properties under smaller sample sizes. Details about the small sample correction are in the Appendix B (Section 6.2.2).

4.2.2 Test for Local Connectivity

Section (4.2) introduced a statistic used to assess the evidence that pair-wise correlation changes with the covariate X. As a natural extension to the pair-wise test statistic, this section expands the method to more variables to study local connectivity. Consider the extension of (4.1) to K-dimensional multivariate normal.

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ \cdots \\ y_{iK} \end{bmatrix} = \begin{bmatrix} \boldsymbol{z}_i^T \boldsymbol{\beta}_1 \\ \boldsymbol{z}_i^T \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} u_{i1} \\ u_{i2} \\ \cdots \\ u_{iK} \end{bmatrix}$$
$$\begin{bmatrix} u_{i1} \\ u_{i2} \\ \cdots \\ u_{iK} \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}(\boldsymbol{x}_i) & \cdots & \rho_{1K}(\boldsymbol{x}_i) \\ \rho_{12}(\boldsymbol{x}_i) & \sigma_2^2 & \cdots & \rho_{2K}(\boldsymbol{x}_i) \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{K1}(\boldsymbol{x}_i) & \rho_{K2}(\boldsymbol{x}_i) & \cdots & \sigma_K^2 \end{bmatrix} \tag{4.5}$$

$$\rho_{k\ell}(\boldsymbol{x}_i) = \rho_{k\ell}(\boldsymbol{x}_i^T \boldsymbol{\alpha}_{k\ell} + \alpha_{0,k\ell}), \quad k, \ell = 1, \cdots, K, \quad k \neq \ell$$

The global null hypothesis for variable 1 is constructed from the first row of the variance matrix Σ of (4.5).

$$\boldsymbol{H}_0^{(1)} : \boldsymbol{\alpha}_{12} = \boldsymbol{\alpha}_{13} = \cdots = \boldsymbol{\alpha}_{1K} = \mathbf{0}, \tag{4.6}$$

Under $\boldsymbol{H}_0^{(1)}$, no other variables' correlation with variable 1 changes with the different values of X. We believe testing the global null (4.6) improves the statistical power when the "hot spot" variables or "hub" variables are connected to a lot of other nodes forming cliques or modules. In the context of gene coexpression network, transcription factors (TFs) are considered "hubs" that regulate the gene expression of multiple genes. TFs are good candidate genes for the local connectivity analysis because a given TF is likely to affect its coexpression with many of its targets simultaneously.

Here, we propose simple linear combination of pair-wise statistics to test local connectivity of variable 1,

$$d_1 = q_{12} + q_{13} + \cdots + q_{1K} = \sum_{k=2}^{K} q_{1k}. \tag{4.7}$$

The statistic d_1 in (4.7) is designed to improve statistical power when pair-wise effect sizes are too small but become substantial when combined. The choice of adding them without additional weighting reflects the lack of prior knowledge about the relationships between the variables.

Appendix C (Section 6.2.3) in shows that q_{1k} for gene pair 1 and k can be written as a $\sum_{p=1}^{P} r_{1k,p}^2$ where $r_{1k,p}$ individually follows the standard normal distribution. If expanded to K genes, we can show the following.

$$\boldsymbol{r}_{1,p} = \begin{bmatrix} r_{12,p} \\ r_{13,p} \\ \cdots \\ r_{1K,p} \end{bmatrix} \rightarrow N_{K-1}(\mathbf{0}, H_1) \quad \forall p = 1, \cdots, P \tag{4.8}$$

where H_1 is a $(K-1) \times (K-1)$ matrix. Appendix C (Section 6.2.3) shows element-wise mapping from Σ to H_1 and shows what is the asymptotic form of H_1. Let $H_1 = U_1 \Lambda_1 U_1^T$ be the eigen-decomposition of the covariance matrix H_1 in (4.8), where the diagonal matrix Λ has eigenvalues $\lambda_{12}, \cdots, \lambda_{1K}$ in a decreasing order. Then, consider the transformation $\boldsymbol{r}_{1,p}^* = U\boldsymbol{r}_{1,p}$ that follows normal distribution with diagonal covariance matrix Λ_1. Note that $\|\boldsymbol{r}_{1,p}\|_2^2 = \|U\boldsymbol{r}_{1,p}\|_2^2 = \|r_{1k,p}^*\|_2^2$ due to the orthogonal invariance of L2 norm. Then, d_1 asymptotically follows a sum of independent Gamma distributions

$$d_1 = \sum_{k=2}^{K} q_{1k} \xrightarrow{d} \sum_{k=2}^{K} \Gamma\left(\frac{P}{2}, \frac{\lambda_{1k}}{2}\right) \tag{4.9}$$

Assuming that we know the true, symmetric, positive definite H_1, we can acquire positive

λ_{1k} for $k = 2, \cdots, K$, and we have expressed the null distribution of d_1 as the sum of distributions of gamma variables. We can computationally simulate this null distribution easily.

Proposition 2 *Under the setting of (4.5), if none of the variables $y_1, \cdots, y_K$ are perfectly correlated, d_1 asymptotically follows the finite sum of Gamma distributions as defined in (4.9) under the global null hypothesis (4.6).*

Detailed proof of Proposition 2 is in Appendix C (Section 6.2.3).

H_1 is acquired from Σ if Σ is known, but Σ is often unknown in practice. When N is sufficiently larger than K, the maximum likelihood estimate $\hat{\Sigma}$ is a consistent estimator for Σ. Using $\hat{\Sigma}$ guarantees that the test statistic in (4.7) converges in distribution to (4.9).

However, when K is larger than N, there is no consistent estimator Σ that does not require regularization conditions such as sparsity or low-rank. Then, permutation test can be an alternative; testing Y against randomly shuffled X can simulate the null distribution of d_1 while preserving the dependence structure.

Permutation tests lead to a limited resolution of p-values. Imprecise p-values prevent accurate inference especially when we need to correct for a large number of hypotheses. Performing a large number of permutations can lead to better resolution of p-values but could be computationally wasteful. To strike a balance, we use the sequential precision-improvement permutation test, similar to one suggested by Chen et al. (2012).

Precision-improvement permutation test terminates the procedure early if the signal is not strong enough. The detailed procedure is as follows. For every permutation $b = 1, \cdots, B$, we permute the rows (samples) of the covariate matrix X to create X_b, so that any existing

link between X and Y is broken. Then we compute the degree statistic for gene k d_{kb} using X_b and the data matrix Y; d_{kb} should follow the null distribution. Then, p-value for d_k is computed as a quantile of d_k compared to the empirical distribution of d_{kb} for each b. After the minimum number of permutations pre-defined by the user (100 in the Section 4.4), we count the number of permutations b where d_{kb} is larger than d_k. If there are two or more such cases, we terminate the permutation procedure early. Most genes fall into this category leading to p-values greater than 0.02. If there are less than 2 such cases observed, we iteratively perform 100 more permutations and re-check the number of d_{kb} with values larger than d_k. We repeat until the number of permutations B reaches the predefined maximum number of permutation (10^6 in Section 4.4), which is designed to give a good enough resolution of p-value given the number of tests that we are performing.

4.3 Simulation Studies

Here, we evaluate the proposed method through simulations. We focus on the pairwise analysis and compare the performance of the proposed score test with two other alternatives - liquid association and the likelihood ratio test.

First, we check the calibration of test statistics under the null hypothesis. We sample X from the univariate standard normal distribution to match the required setting of liquid association. We simulate the data matrix Y from below.

$$\boldsymbol{y}_i \sim \mathcal{N}_2\left(\boldsymbol{b}_0 + \boldsymbol{z}_i^T\boldsymbol{\beta}, \begin{bmatrix} 1 & \bar{\rho} \\ \bar{\rho} & 1 \end{bmatrix}\right)$$

where $\bar{\rho}$ was randomly selected from uniform distribution ranging from -1 and 1, and $\boldsymbol{\beta} = \mathbf{0}$. We test different sample sizes of $N = 500, 100, 30$ to check the behavior of each method under the null hypothesis. For each N, we sample X once, and generate Y 1,000

times. The likelihood ratio test was designed to assume hyperbolic tangent model for ρ,

$$\rho(\tilde{\boldsymbol{x}}_i^T \tilde{\boldsymbol{\alpha}}) = \frac{e^{\tilde{\boldsymbol{x}}_i^T \tilde{\boldsymbol{\alpha}}} - 1}{e^{\tilde{\boldsymbol{x}}_i^T \tilde{\boldsymbol{\alpha}}} + 1}, \tag{4.10}$$

with $\tilde{\boldsymbol{x}}_i = \begin{bmatrix} 1 & \boldsymbol{x}_i \end{bmatrix}$, $\tilde{\boldsymbol{\alpha}} = \begin{bmatrix} \alpha_0 & \boldsymbol{\alpha} \end{bmatrix}$.

(4.10) is the inverse of Fisher transformation, $\frac{1}{2}\tilde{\boldsymbol{x}}_i^T \tilde{\boldsymbol{\alpha}} = \frac{1}{2} log\left(\frac{1+\rho}{1-\rho}\right)$. Fisher-transformed ρ asymptotically follows normal distribution, so it works well when X is drawn from normal distribution. We use *optim* function in R to find $\hat{\boldsymbol{\alpha}}_{\text{MLE}}$ under the alternative hypothesis.

The results show that all three methods control the type I error at the nominal size well, where score and likelihood ratio test statistics both follow χ_1^2 closely.

Next, we generate the data under the alternative hypothesis to compare the statistical power. We use $N = 70$ to reflect the sample size of GTEx data. We again draw X from standard normal distribution. Then, for $i = 1, \cdots, N$, we generate $\rho(\tilde{\boldsymbol{x}}_i^T \tilde{\boldsymbol{\alpha}})$ from hyperbolic tangent function in (4.10). We draw Y from (4.1) with varying levels of $\boldsymbol{\alpha}$, 1000 times each. The hyperbolic tangent model places the likelihood ratio test at an advantage because the model is correctly specified, so as a contrasting case, we use a quadratic model to generate ρ as follows,

$$\rho(\alpha_0 + \tilde{\boldsymbol{x}}_i^T \tilde{\boldsymbol{\alpha}}) = (-0.1 + \tilde{\boldsymbol{x}}_i^T \tilde{\boldsymbol{\alpha}})^2 - 0.99, \tag{4.11}$$

where subtracting 0.99 is to ensure numerical stability. Since the likelihood ratio test assumes a wrong model, it is expected to lose power. Also, since quadratic function is highly nonlinear, liquid association is expected to have poor performance. Figure 4.1 (a) and (b) show the shape of ρ with respect to X with varying levels of $\boldsymbol{\alpha}$.

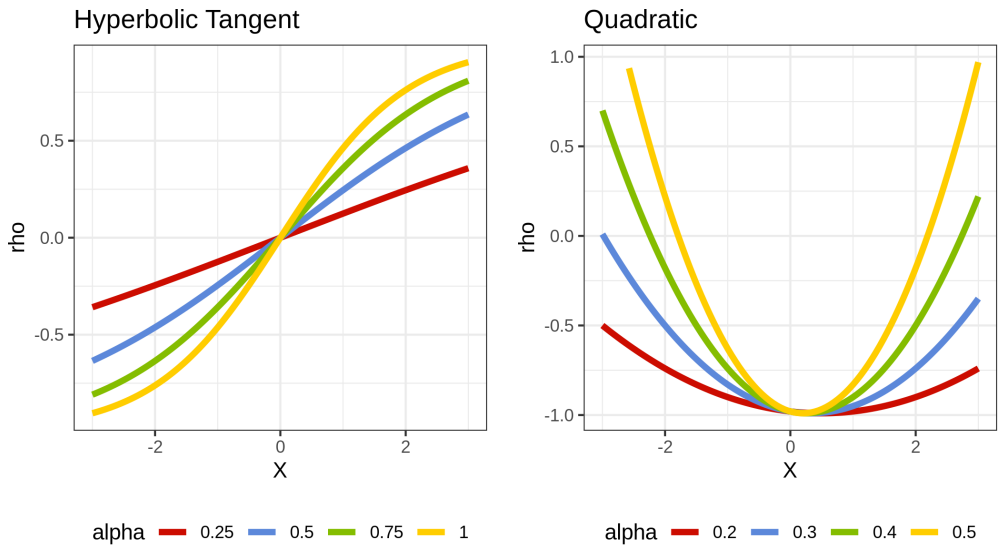


Figure 4.1: Generation of ρ for the simulations. The two example functions ρ above show the relationship between X and ρ for the two example functions and varying levels of α.

Table 4.1 summarizes the result. It counts the proportion of simulations that show p-values less than 0.05 out of 1,000 total simulations. When ρ is generated from hyperbolic tangent function, likelihood ratio test generally outperforms the other two methods, as expected, since the model is correctly specified in LR test. Score test and liquid association perform similarly. Meanwhile, when ρ is generated from quadratic function, score function clearly outperforms the other two methods. Hence, the simulations show that the proposed score test is robust to the shape of heteroskedasticity. Figure 4.1 (c) shows the distribution of computation times of each method to compute the test statistic once in the scale of log_{10} for 1000 simulations under quadratic model with $\alpha = 0.5$. The score test is the most efficient, because the likelihood ratio test requires numerical estimation of MLEs both under the null and under the alternative hypothesis while liquid association requires permutation test for inference. The simulations were done sequentially (in a non-parallel manner) with LAMBDA QUAD workstation with Intel Xeon W-2175 processor.

ρ	tanh					quadratic			
$\boldsymbol{\alpha}$	0	0.25	0.5	0.75	1	0.2	0.3	0.4	0.5
score	0.047	0.148	0.442	0.755	0.911	0.912	0.831	0.735	0.699
LA	0.041	0.145	0.468	0.767	0.919	0.036	0.017	0.012	0.02
LR	0.056	0.174	0.527	0.838	0.964	0.164	0.27	0.356	0.501

Table 4.1: Proportion of simulations for each method that shows p-value < 0.05 at each data generating model and α level. We use two functions for ρ, hyperbolic tangent and quadratic function. The likelihood ratio test was conducted under the assumption that ρ is generated by hyperbolic tangent function. The intercept α_0 is 0 in all cases.

Test	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Score	2.62e-05	2.96e-05	3.03e-05	3.13e-05	3.12e-05	2.21e-03
LA	3.94e-03	4.06e-03	4.22e-03	4.43e-03	4.28e-03	6.96e-02
LR	3.04e-01	5.32e-01	6.51e-01	6.94e-01	8.051e-01	2.29e-01

Table 4.2: Distribution of time (in seconds) for getting one score statistic for the each method from 1,000 simulations. The proposed method is faster than the likelihood ratio test in the order of 10^4, and than the liquid association in the order of 10^2.

4.4 Applications to GTEx Data

We next apply this method to the expression levels in muscle skeletal tissue in GTEx data where 71, highest among all tissues, African American samples are available. We aim to find transcription factors that change their coexpression pattern with their target genes as the global ancestry changes. We acquire a list of transcription factors from TF checkpoint database from Chawla et al. (2013). We also acquire a list of target genes for each transcription factors from TF2DNA database in Pujato et al. (2014). We only take into consideration target genes with the highest binding scores.

For each of the $k = 1, \cdots, K = 848$ transcription factor encoding gene, we compute the pair-wise statistic q_{kj} for all its target genes $j = 1, \cdots, J_k$, where J_k is the number of target genes for each transcription factor k. Then, we compute $d_k = \sum_{j=1}^{J_k} q_{kj}$ to test the

hypothesis that the correlation between the transcription factor k and its targets remain the same across different genetic industry. We first divide d_k with the number of targets J_k to compute the average score of all the target genes for the given TF k, and we make a heuristic comparison against χ^2_1 distribution. Under the null hypothesis, the expectation of d_k/J_k is 1, although the variance is not trivial due to high dependence. Then, we choose 10 genes with the top d_k/J_k values to perform the permutation test.

Table 4.3 summarizes the top 5 transcription factors with the highest average d_k values and their p-values computed from sequential permutation tests. The adjusted p-values were computed as below using Benjamini Hochberg procedure.

$$\text{adjusted } p = p \times (\text{number of transcription factors})/(\text{rank of } p).$$

Gene Name	p-value	Adjusted p-value
ATOH8	1.47×10^{-5}	0.015
ZNF678	3.26×10^{-5}	0.014
ZNF638	2.15×10^{-4}	0.060
ZBTB32	5.55×10^{-4}	0.118
FARSA	1.53×10^{-3}	0.261

Table 4.3: Top 5 transcription factors out of 848 with the lowest p-value in muscle skeletal tissue. Each transcription factor's local connectivity with its target genes was investigated through the proposed statistic d. The p-values were obtained through the sequential precision-improvement permutation test , and adjusted p-values were obtained via Benjamini-Hochberg procedure.

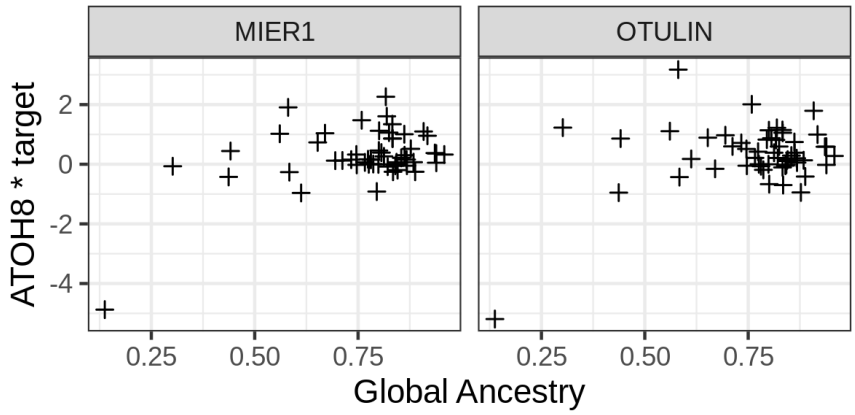


Figure 4.2: Top two target genes of ATOH8. The product of the expression levels of ATOH8 and two targets with the highest scores (MIER1 and OTULIN) are plotted on the y axis against global ancestry. The y axis shows the product which is an unbiased estimator of the correlation between two genes' expression levels. For both genes, the correlation between two genes become stronger as samples have higher proportion of global ancestry.

Two genes, ATOH8 and ZNF678, maintain their significance level of 0.05 under the Bonferroni criterion for 848 transcription factors. For the top gene ATOH8, the two highest contributing target genes were MIER1 and OTULIN, each having scores of 58.27 and 29.02 (Figure 4.2).

4.5 Discussion

We propose a method to test whether the covariance between bivariate normal variables changes with quantitative covariates. We further expanded our scope of analysis by looking at local connectivity — how one variable's connectivity with multiple other variables change with continuous covariates. We provided a real data example by identifying major transcription factor genes that are differentially connected with their targets by genetic ancestry.

Our method is more flexible than other alternatives for covariance testing, but it still has some limitations. First, when there are more variables than the available sample size, as often is the case for many modern data sets, we ultimately turn to a permutation test,

not being able to take advantage of the theoretical results. It is challenging, to say the least, to find consistent estimators for the eigenvalues of Σ in high-dimensional data, but there could be other ways to estimate Σ depending on application. For example, one could impose a sparsity assumption on the covariance matrix, and that can lead to a reliable estimate of Σ and subsequently H in (4.8). Such assumption is too restrictive in our context, but in other applications, one can take the liberty to make structural assumptions on the covariance matrix. Second, the score statistic q was derived under the normality assumption of the data set, although simulations show that the result is quite robust to distribution mis-specification. These limitations of the methods propose possible future research topics: (1) better way to estimate H in (4.8) when $K > N$ to preserve asymptotic results, and (2) a non-parametric version of correlation analysis that can generalize to any underlying distributions.

In practice, users of this method should note the method's sensitivity to correct mean modeling. For example, if correct data generating procedure is $\boldsymbol{z}_i = \boldsymbol{x}_i + \boldsymbol{x}_i^2$, missing the quadratic term (mis-specifying $\boldsymbol{z}_i$ as $\boldsymbol{x}_i$) would incorrectly assign the effects of $\boldsymbol{x}_i^2$ to the residuals. Then, the score statistic computed from the residuals would show remaining effects from the mean term, and hence the type I error will be inflated. Therefore, the users are expected to be careful in regressing out all the relevant terms so that the mean term effects do not spill over to the variance term.

Alternative ways have been proposed for some components of our methodology. An important aspect is the formulation of the combined statistic d. Chen et al. (2012) discusses two ways to construct the alternative hypothesis for testing the global null in (4.7). One way, called a sparse alternative, is to test whether only a small number among all tests have non-zero effects while all other tests are null. Another way is to test if at least one test has a non-zero effect size, as we did in (4.7). Based on prior knowledge in biology and coexpression network, we assume that there are many small signals instead of few large, so we choose the

latter, which is equal to (4.7). Another aspect is about deriving the null distribution of the sum of correlated χ^2_P. Moschopoulos (1985) provides another way to derive the distribution of (4.9) by expressing its cumulative distribution in a form of infinite sum, but we believe it is impractical.

We believe the proposed method can be applied to data problems in diverse domains, especially where certain central variables are connected to many other variables as in the transcriptional regulatory networks. Many network problems, such as protein interactions, metabolic networks, co-authorship networks, and semantic networks, are known to have, or have something close to, a scale-free topology, indicating that the important variables in those networks can be tested against other variables. The proposed correlation analysis will provide insights into the building blocks of diverse network problems by looking at the pairwise and more than pairwise relationships among the variables.

CHAPTER 5

ANALYSIS OF SINGLE CELL TRANSCRIPTOME

5.1 Introduction

In this chapter, we explore a different type of data by analyzing cell-level expression level data set, and present a new scRNA-seq analysis pipeline that accounts for the error structure of UMI data.

Many pipelines have been built for scRNA-seq UMI data analysis. Despite subtle differences in these pipelines, the general order of a scRNA-seq analysis is as follows: quality control (filtering), cleaning (normalization, imputation, de-noising, batch-correction, etc.), feature selection which often involves dimension reduction, and downstream analysis such as clustering and lineage analysis. In this chapter, we do not discuss filtering and focus on the later three steps. First, the challenge of scRNA-seq data cleaning has led to the development of a wide array of tools. Some methods adjust for sequencing depths using size factors (Butler et al., 2018; Vallejos et al., 2017). Some impute the reads directly using a zero inflated model, to reduce the noise from drop-outs (Gong et al., 2018). Some try to de-noise the entire data set by fitting parametric models, where one example is sctransform that uses the residuals from negative binomial regression (Hafemeister & Satija, 2019) and another example is SAVER that uses Poisson LASSO regression (Huang et al., 2018). Despite the diversity of proposed methods, the general consensus has been reached to use one of the following distributions to model the counts: Poisson, Negative Binomial, or Zero inflated Negative Binomial distribution. Secondly, methods for feature selection have been less controversial. Most tools use some form of gene variance to mean ratio to identify genes that are highly dispersed, where the dispersion level is interpreted as a signal of biological heterogeneity (Butler et al., 2018; Hafemeister & Satija, 2019; Chen et al., 2018). Another less recognized approach is to use the zeros in the read or UMI counts; genes with inflated zeros are interpreted as

biologically important signals (Andrews & Hemberg, 2018). Lastly, after data cleaning and feature selection, the pre-processed data will then be piped into downstream analysis tools for clustering analysis (Wang et al., 2017; Eraslan et al., 2019; Freytag et al., 2018), trajectory inference (Saelens et al., 2019; Trapnell et al., 2014), or differential expression analysis (Robinson et al., 2010; Love et al., 2014). Currently, pre-processing and downstream analysis have been mostly considered as separate and consecutive steps (Eraslan et al., 2019; Butler et al., 2018; Huang et al., 2018; Wang et al., 2019).

Here, we present extensive analyses of publicly available UMI data sets that challenge most existing pre-processing tools' assumption, mainly that pre-processing is a necessary step before feature selection and downstream analysis. Our results suggest that clustering, or resolving the cell heterogeneity, should be the foremost step of the scRNA-seq analysis pipeline, not as part of the downstream analysis. Normalizing or imputing the data set before resolving the heterogeneity can lead to adverse consequences in downstream analysis. Adding to the arguments that the UMI data is much cleaner than the read count data Chen et al. (2018); Townes et al. (2019), our analyses demonstrate that the simple Poisson distribution is sufficient to fully leverage the biological information contained in the UMI data if the cell-type heterogeneity has been appropriately accounted for. As a result, we provide a new perspective on scRNA-seq data analysis by integrating the pre-processing step and clustering, which was classified as part of the down-stream analysis. The proposed procedures have been implemented in software HIPPO (*https://github.com/tk382/HIPPO*).

5.2 Results

5.2.1 Demystifying Drop-outs

We started by exploring zero detection rates in three UMI datasets generated by 10X protocols for both homogeneous and heterogeneous cell populations. Taking a subset of data in

Zheng 2017 Zheng et al. (2017) as created in Freytag 2018 Freytag et al. (2018) as an example, we computed zero proportions, defined as the proportion of cells with zero counts per gene, across 15,568 genes, in CD19+ B cells, CD4+/CD25 Regulatory T cells, and combined. The obtained statistics were plotted against gene-level average count and were compared with expected zero proportions under the Poisson, Negative Binomial and Zero-Inflated Negative Binomial distribution, respectively (Figure 5.1 A). For a homogeneous cell population, we observe most genes align well with the expected curve under the Poisson assumption. Few genes can benefit from using the Negative Binomial model to account for extra dispersion from the Poisson, but our results strongly suggest that to model the drop-outs by introducing an extra zero-inflation component by the Zero-Inflated Negative Binomial distribution is unnecessary. For example, in Zheng dataset, 257 genes out of 5,568 genes would benefit from Negative Binomial modeling (p-values pass Bonferroni criterion in likelihood ratio test at 0.05 type I error level), but no gene would benefit from extra zero inflation parameter. The p-values are not calibrated to the uniform distribution because there are many genes that have UMI count of 1 in one cell and 0 in everywhere else, in which case p-value is close to 1 (Figure 5.1 B). This result shows that drop-outs are within the range of natural Poisson sampling noise in UMI data for a homogeneous cell population, and they do not introduce excessive zero inflation, which is contradictory to prevalent opinions Eraslan et al. (2019); Risso et al. (2018); Tung et al. (2017); Vallejos et al. (2017). Extra zero inflation can be measured by comparing observed zero to expected zero counts under Poisson distribution within a homogeneous cell population(Methods). Through the following analysis, we show zero proportions are as effective measures for cell-type heterogeneity as other widely used alternatives, gene variance, coefficient of variation (CV), or dispersion parameter in negative binomial distribution Andrews & Hemberg (2018) (Figure 5.1 A). It provides simplicity and interpretability in particular for data sets with low UMI counts and reasonable number of zeros, as zero-inflation is meaningless when no zero is observed.

Analysis in multiple UMI data sets shows that zero proportions in most genes can be effectively modeled by the Poisson distribution, as more than 95% of absolute z-values (Methods) are below 2. For mixed cell types, zero proportions considerably deviate from expected values under the Poisson model, as only less than 30% of the genes have z-values below 2. This shows that the zero inflation test is an effective way to find genes that contribute to cellular heterogeneity. On the contrary, gene variance of mixed cell types does not always surpass those of a single cell type. In Zheng data, 62% of the genes had higher variance in pure Naive Cytotoxic cells than in mixed PBMC cells. On average, gene variance is similarly distributed for homogeneous and heterogeneous cell populations (Supplementary Table 6.2, Supplementary Figure 6.13). Therefore, the gene variance is rather more of a gene-specific characteristic while being less informative about the characteristics of the entire cell population. CV, on the other hand, suffers from an inherent numerical instability issue when gene mean is close to 0, because when mean is close to 0, CV estimates have high variability. Another popular option is to conduct model selection to assign genes to one of three candidate distributions of Poisson, NB, and ZINB, but measuring over-dispersion also suffers from a similar problem in selecting biologically meaningful genes Choi et al. (2020); Chen et al. (2018). When we used statistics from likelihood ratio test and select top genes from the resulting statistics, the selected genes were very different from those selected when we used zero proportion (Figure 5.2 F). For three data sets of Azizi2016, Zheng2017, and Freytag2018 (median sequencing depth of 4371, 1298, and 2393.5 respectively), the likelihood ratio test selects genes that are overly focused on those with mean close to 0. Intuitively, the dispersion parameter scales with the ratio of gene mean to the gene variance^2, and in nature very similar to CV. These genes with very low mean are likely to have little information about the cells. Still, dispersion parameter can be more useful than zero-inflation statistic when the data set has high UMI counts, so deviance has been implemented in HIPPO as an alternative feature selection method (Supplementary Figure 6.14, 6.15).

We expand the data to study all 68,579 cells from Zheng dataset Zheng et al. (2017). When we aligned the zero proportions with the expected Poisson curve according to the provided cell type labels: CD14+ Monocyte, CD19+ B, CD34+, CD4+ T Helper2, CD4+/CD25 T Reg, CD45RA+/CD25- Naive T, CD4+/CD45RO+ Memory, CD56+ NK, CD8+ Cytotoxic T, CD8+/CD45RA+ Naive Cytotoxic, and Dendritic cells. Most of these cell types look relatively homogeneous. However, one cell type, CD34+, was particularly noisy with very high zero proportions, indicating cellular heterogeneity (Figure 5.1 C). Based on the diagnosis from t-SNE plots, we identified three subtypes within the CD34+ cells. The alignment of zero proportions against the Poisson curve was immediately improved according to the inferred subtype labels. This indicates the effectiveness of zero proportions as metrics to evaluate cellular heterogeneity and their potentials to discern cell types.

We further checked how zero proportions could be dispersed from the Poisson distribution for genes with various functional annotations across all PBMC datasets (Figure 5.1 D). Specifically, we calculated the difference between observed zero proportions and expected proportions (under Poisson) for each functional group using reference data from GENCODE of GRCh38 Harrow et al. (2012). The vast majority of genes are categorized as "protein-coding genes". Their zero proportions cover a wide range from 0 to 0.7, but centered at 0, indicating variability in zero proportions but no systematic inflation of zero proportions. In contrast, immune-related genes are consistently zero-inflated, with the interquartile range as high as 10% to 20%. The enrichment analysis (Supplementary Table 6.4) shows that immune-related genes have significantly higher proportion of zero-inflated genes compared to genes that are not related to immune function. The top-ranked annotations for zero inflation include IG C genes, TR C genes, and HLA genes. IG C genes are immunoglobulin genes of the constant (C) region, while TR C genes are T cell receptor genes of the constant region, and hence both gene types are deeply connected to immune system. HLA genes are genes in human leukocyte antigen system that is responsible for the regulation of the immune

system. Genes involved in immune functions are expected to be inherently heterogeneous Spurgin & Richardson (2010). For example, HLA genes are highly polymorphic than others, and TR-C genes go through VDJ recombinations that lead to more diverse sequences across cells. This result corroborates with the notion that cellular heterogeneity is the main driver of zero-inflation. Higher level of heterogeneity in immune genes explain the past studies' results that even within one cell type, there are zero-inflated genes Clivio et al. (2019). The high heterogeneity of cells in certain genes suggests that it is difficult, to say the least, to fully account for the cell type for every gene; as the cells are finely clustered, a point will be reached where the number of cells left in each cluster is not enough to do statistically meaningful analysis. Therefore, we use the following stopping criteria for iterative clustering where the procedure stops when the number of genes with zero-inflated is less than a certain threshold. This criterion is designed to take into account the remaining granular biological heterogeneity in certain genes that cannot be fully resolved through cell-type.

5.2.2 Zero inflation test for cellular heterogeneity

Based on the above observations, we propose a new feature selection strategy that uses detected zero proportion of a given gene as the statistic to test for cellular heterogeneity. Under the null hypothesis, where complete cellular homogeneity is assumed, the proportion of zeros is equal to the expected zero proportion under Poisson distribution. Under the alternative hypothesis, zero proportion is inflated, as if the count data follows mixture of Poisson (Methods). Formally, our framework can be presented as follows:

$$H_0 : \quad p_g = e^{-\lambda_g}, \tag{5.1}$$

$$H_A : \quad p_g > e^{-\lambda_g} \tag{5.2}$$

where g is gene index and λ_g is the mean UMI count for gene g. Above testing framework is based on an assumption that whether UMI count being 0 follows the Bernoulli distribution.

Test statistic z_g follows a standard normal under the null hypothesis (Methods). Genes with rejected null hypotheses will be selected for downstream analysis. For example, the CD34+ cell population within Zheng2017 dataset Zheng et al. (2017) has 2.7% of the genes with significant zero inflation at 5% Type I error level after Bonferroni correction. But after clustering into subtypes, each subtype had 1.3%, 0.3%, and 1.2% of genes with zero inflation respectively. We demonstrate the intuition of this test procedure in Table 5.1 using gene PPBP as an example. PPBP was identified with a high zero proportion of 26% and an average mean UMI count of 25.89 within CD34+ cells, indicating very high zero inflation with z-score greater than 10^6 when the proposed test is applied. After we separate CD34+ cells into three subtypes, the test within each subtype is no longer statistically significant. We observe PPBP is highly expressed in subtype 2 and 3 and is almost unexpressed in subtype 1. This shows how cellular heterogeneity can drive excessive zeros and how zero proportions can be used to discern cell types.

The proposed framework significantly differs from existing ones in several ways. First of all, only the proportion of zeros (p_g), but not that of other non-zero count values, is used in the test. We empirically show that this statistic is sufficient for cellular heterogeneity analysis in many data sets with low UMI counts. This allows us not to search for a particular parametric distribution to fit all non-zero values, which can be computationally more burdensome. In terms of clustering, analysis shows that deviance and zero-inflation both lead to feature selection with similar performance (Supplementary Figure 6.17, 6.18); our software can use deviance test for feature selection when a data set has high UMI counts and zeros alone do not hold enough information. Secondly, this framework allows each gene to have different grouping structure across cells. Most existing methods select one set of genes to cluster all the cells Butler et al. (2018); Duò et al. (2018); Kiselev et al. (2017); Wang et al. (2017), which implicitly assume cell types can be well-defined biologically by a common set of genes. This is not realistic given the fact that each gene's heterogeneity level varies with its

function. For example, house-keeping genes are expected to behave similarly in all cells but immune-related genes, known to have more diverse genetic profiles with highly polymorphic nucleotides Hughes (1997); Hurst & Smith (1999), might be more finely differentiated among the sampled cells. Our approach acknowledges this type of variability. Finally, our approach provides a much more optimistic view of the UMI data analysis. No complicated modeling is needed for resolving the cellular heterogeneity.

We observe that the droplet-based data-generating process in the 10X protocols affect UMI counts in different cell types and even across datasets in a similar fashion (Supplementary Figure 6.5, 6.6). Regardless of samples or cell types, all cells show the same distribution of zero proportions with respect to mean UMI count. This means the technical noise affects each and every cell fairly, and hence, biological heterogeneity alone can largely explain the zero inflation phenomenon. Once the heterogeneity is accounted for, without any other pre-processing steps, zero proportions of UMI data closely follow the expected curve under a Poisson distribution. These observations urge us to re-evaluate some widely used pre-processing methods under this scope.

5.2.3 Inappropriate pre-processing introduces unwanted noise in the downstream analysis

One of the most most popular method for normalization is to divide UMI counts by a cell-specific scaling factor so that total UMI counts are equal across cells Vallejos et al. (2017). This strategy implicitly assumes sequencing depth effects are purely technical. Total UMI count needs to be carefully corrected in case of data integration. Data sets collected from different protocols and batches have different distribution of UMI counts. However, when there is no batch effect, the total UMI count has valuable biological information. Here, we show sequencing depths are confounded with cell types and size factor-based adjustment can obscure biological information (Fig 5.2 A). For a given gene, dividing UMI counts by

cell-specific factors does not change its zero proportion across cells but changes its mean. As a consequence, zero proportions across genes no longer follow the expected curve under a Poisson distribution (Fig 5.2) A), and the two curves from each cell type are separated from one another. For example, in 6 PBMC data sets (Azizi and Zheng) Azizi et al. (2018); Zheng et al. (2017), monocytes have lower UMI counts than B cells. The median UMI counts for monocytes and B cells, respectively, are 787 and 1180 for Zheng data for 68,000 PBMC cells, 4831 and 5575 for Azizi 2018 data for breast cancer tumors patient 9 (replication 1), 4891 and 5372 for patient 10, 5093 and 5722 for patient 11 (replication 1). When they are forced to match the median UMI count of all cells, the counts for the monocytes are inflated while those for the B cells deflated. In addition, cell types are stratified on the zero proportion plot after adjustment, indicating that total UMI counts of each cell contain valuable information about its cell type (Figure 5.2 A, Supplementary Figure 6.16). The UMI counts for different cell types do not need to be consistent across different tissues or organisms. For example, in Zhang2019 data, B cells and Monocytes have similar total UMI counts (Supplementary Figure 6.16), but fibroblast has more UMI counts than all other types. Forcing fibroblasts to have the same UMI counts as other types would reduce the signal strengths of the markers for fibroblasts.

Sctransform is one recent influential UMI analysis method Hafemeister & Satija (2019). The key idea of sctransform pre-processing is to remove sequencing depth effects by introducing log-scale sequencing depth as a covariate and regressing it out from each cell under a Negative Binomial model. Similarly, this approach destroys the natural Poisson structure for zero proportions. We show in Figure 5.2 B an example of how normalization can further interfere with detection of biological signals. Across all 6 PBMC datasets mentioned above, we observe B cells always have more UMI counts than Monocytes before pre-processing. Applying sctransform barely modifies the sequencing depths of Monocytes but shrinks the UMI counts of B cells to match those of Monocytes. Due to the artificial shrinkage, biological

markers for B cells, such as MS4A1, CD79A, and CD79B lose their power to discern B cells and Monocytes Schelker et al. (2017). This suggests cell type differences could be potentially compromised due to excessive cleaning from sctransfrom.

Another popular pre-processing step is to apply deep learning based de-noising tools such as Deep Count Autoencoder (DCA) and SAVER, which de-convolute the technical effects from biological effects and impute zero accounts due to drop-outs at the same time. DCA implements deep neural network with flexible parametric options for noise distributions. Similarly, we observe DCA blurs the distinction among cell types, because de-noising methods essentially regularize each cell to resemble one another. We illustrate its negative impacts on downstream analysis by comparing differential expression analysis results using two imputation strategies. We selected Naive T cells and regulatory T cells from Zhengmix8eq for the experiments, which clustering algorithms often struggle to differentiate because of their similarity. In the first experiment, we imputed Naive T cells and regulatory T cells together. In the second, we imputed Naive T cells and regulatory T cells separately. Then we performed DE analysis on imputed data sets using edgeR's likelihood ratio test Robinson et al. (2010). We observe much greater log fold change values between Naive T and regulatory T cells from data imputed separately than data imputed together. Overall, the signal strength of DE analysis is greatly compromised across all the genes if imputing two cell types together (Fig 5.2 E). Using Type I error level of 0.05, 320 genes pass the Bonferroni criterion if clustering is performed first, while only 156 does if imputation is performed first. Known markers including CD4, CTLA4, FOXP3, and IL2RA Schelker et al. (2017) lost significant amount of biological signals by showing weaker log-fold change (Figure 5.2 D, Supplementary Figure 6.19). When the cells were first clustered and then imputed, the p-values were 1e-04, 8e-04, 2e-07, and 4e-11 respectively for those genes. When the cells were imputed first through DCA, the p-values were 3e-01, 4e-02, 4e-02, and 6e-07. Hence, three of the 4 genes lost statistical significance at a very liberal p-value threshold of 0.05. This analysis suggests

imputing the UMI data without resolving cell heterogeneity can lead to loss of important biological information.

5.2.4 HIPPO: Heterogeneity-Induced Pre-Processing Tool

The above analyses suggest the first and foremost step in pre-processing is to account for the cellular heterogeneity. Imputation or normalization before resolving the cellular heterogeneity may lead to inevitable loss of biological signals. We implement this new perspective into a computational tool called HIPPO, where we integrate the proposed zero inflation test into a hierarchical clustering framework. Specifically, we first selected genes with strong indication for cellular heterogeneity. We use a cut-off of 2 on z-score for selection of genes. The selected features were then used to cluster the cells into 2 groups using PCA + K-means. Then each cluster was evaluated with their intra-variability using the mean Euclidean distance from the centers of K-mean algorithm. The group with the highest intra-variability was selected and assigned for next round of clustering. The feature selection and clustering steps are iteratively repeated until one of the two ending criteria are met: K round of clustering for pre-determined number of clusters K, or the number of zero-inflated genes is less than a certain percentage of the genes. The former one can be difficult to set in real practice without any prior knowledge and the later one offers a more natural stopping criterion. HIPPO is computationally cheap because fewer and fewer features will be left for the next round of clustering, and the Poisson based test statistic has closed-form expression (Figure 5.3 E, F). In Figure 5.3 B, we show the results from each iteration of HIPPO on Zhengmix8eq data. HIPPO successfully identifies Monocytes, Natural Killer cells, B cells, and T cells in the respective order. Then it further separates naive cytotoxic cells, memory T cells, and Naive T cells from a group of Regulatory T cells and Helper T cells. However, when forced to separate into one more group, instead of clustering the remaining T cells, it created another subgroup of natural killer cells. Meanwhile, Seurat and Sctransform fails to separate the Memory T cells, Regulatory T cells, and Helper T cells, grouping them as one cluster. The

adjusted rand index for the three methods show that HIPPO performs the best throughout the different K specification (Figure 5.3 A, D). When the selected features' characteristics were studied through CV, gene variance, and zero proportion, Seurat and Sctransform selected more features (2000 and 3000 respectively while HIPPO selected 950), but they are highly concentrated where gene means are near 0. This is because their feature selection focuses on coefficient of variation which becomes numerically unstable as gene mean becomes near zero. HIPPO selects fewer but more relevant genes by using the zero proportion as the selection metric (Figure 5.3 C). This result is repeated in a different data set from muscular heart tissue in Figure 5.4 A. Genes selected by both methods are those with non-zero mean UMI counts, but Seurat selects extra number of genes that have mean count very close to 0. These genes are likely to add noise instead of contributing to real biological signals detection.

HIPPO's iterative procedures naturally offer strong interpretability through sequential visualization of the analysis at each round of clustering. We use HIPPO results on an unlabeled 10X UMI data set of 10K E18 mouse heart cells for illustration. Sequential feature selection can be monitored through the visualization of the changing relationships between zero proportions and gene means. As cells are clustered into finer distinct groups, or as more cellular heterogeneity is resolved, regression lines between zero proportions and gene means get more closely aligned with the expected Poisson curve (Figure 5.4 B). Simultaneously, we can use a heatmap to visualize top features that contribute most at each round of clustering (Figure 5.4). In addition to biomarkers identified based on zero inflation, HIPPO also implements a differential expression test based on all count values to extract more features (Methods, Figure 5.4 D). The differential analysis can be viewed together with a t-SNE plot constructed with the same color code(Figure 5.4 E).

5.2.5 Discussion

We have provided a new perspective on the analysis of single cell UMI data sets of multiple tissues and protocols (Supplementary Figures 6.5, 6.6, 6.7, 6.8, 6.11). Extensive analyses confirm the claims of recent literature (Chen et al., 2018) that different tool must be applied to the UMI data set from the tools for read count data set; UMI data set is free from amplification bias, so the level of technical noise is much lower. The results also show that cell-type heterogeneity must be tackled as the first step of analysis for more reliable downstream analyses. Moreover, through a streamlined feature selection method that reflects the dynamic nature of cellular process, the proposed method provides a computationally and mathematically simple analysis tool with great interpretability.

There are remaining challenges that are important in the future development of single cell UMI data analysis. First, lack of labeled data restricts the analyses in certain protocols such as Drop-seq. There is strong evidence for our method in 10X data sets. Supplementary Figures also show that the claims hold in Tung2018 data that uses Hi-Seq 2500 (**?**) and Baron2016 data that uses in-Drop (Baron et al., 2016). In Drop-seq, the noise level was too high to assume the zero proportions follow the exponential curve relative to the gene mean (Supplementary Figure 6.11). It is either that Drop-seq data sets have different noise structure from the 10X data sets, or in particular Macosko data (Macosko et al., 2015) of muscular retina cells have excessively high cellular heterogeneity (Klein et al., 2015). Future new Drop-seq data could help resolve the discrepancy between 10X and Drop-seq. Secondly, although HIPPO is computationally simple compared to existing tools, the computational bottleneck is the principal component analysis, which could be slow for large cell numbers. In that case, advanced computing techniques such as sub-sampling or more rigorous filtering should be applied. Thirdly, the zero inflation statistics are appropriate for data sets with relatively low sequencing depth. For example, when all gene counts are high and there are no cells with recorded 0, zero-inflation test is not valid.

We focus on the pre-processing with resolving cellular heterogeneity in our analysis tool, but this novel perspective on the noise structure of UMI data can be extended to other steps of analysis pipeline. Batch correction, lineage analysis or trajectory inference can all benefit from the simpler noise structure not only computationally but also by avoiding unnecessary normalizing steps that can introduce unwanted bias and noise.

Cell Population	gene mean	expected p	observed p	z-score
CD34+	25.89	5.69e-12	0.26	1838203
Subtype 1	0.5625	0.57	0.91	6.19
Subtype 2	22.36	1.93e-10	0	0
Subtype 3	38.96	1.25e-17	0	0

Table 5.1: Zero inflation test statistics of PPBP gene in CD34+ cells in Zheng data before and after clustering into subtypes.

5.3 Methods and materials

5.3.1 Datasets

Throughout the analysis, we used publicly available single cell UMI sequencing data most of which used 10X protocol. Most analysis in the main text is focused on SRP073767 which is also available in 10x Genomics, and it sequences 68,000 PBMC cells using Cell Ranger 1.1.0 (Zheng et al., 2017). We use different subsets of this data sets, namely Zhengmix4eq, Zhengmix4uneq, and Zhengmix8eq as defined in Duò et al. (2018). Other data sets used in the main text are GSE111108 (Tian et al., 2018) and GSE115189 (Freytag et al., 2018), and GSE114724 (Azizi et al., 2018). Supplementary data includes more data sets from 10X including 5k Cells from a combined cortex, hippocampus and subventricular zone of an E18 mouse (v3 chemistry), 1k Brain Cells from an E18 Mouse (v2 chemistry), and 10k Heart Cells from an E18 mouse (v3 chemistry). We also use GSE84133 (Baron et al., 2016) as an example of in-Drop and GSE63473 (Macosko et al., 2015) as an example of Drop-seq. All the data sets were analyzed after their own filtering process. (Table 5.2)

ID	Data Set	Species	Protocol
10x	5KNeuron	Mouse Neuron	10x (v3.1) CR* 3.0.2
10x	10KHeart	Mouse Heart	10x (v3) CR 3.0.0
GSE111108	Tian2018	Human Cell Lines	10x Chromium
GSE115189	Freytag2018	Human PBMC	10x (v2)
10x	1KNeuron	Mouse Neuron	10x (v2) CR 2.1.0
SRP073767	Zhengmix4eq	Human PBMC	10x (v1) CR 1.1.0
SRP073767	Zhengmix4uneq	Human PBMC	10x (v1) CR 1.1.0
SRP073767	Zhengmix8efq	Human PBMC	10x (v1) CR 1.1.0
SRP073767	PBMC3k	Human PBMC	10x (v1) CR 1.1.0
SRP073767	PBMC4k	Human PBMC	10x (v1) CR 1.1.0
SRP073767	PBMC68k	Human PBMC	10x (v1) CR 1.1.0
GSE84133	Baron2016	Human Pancreas	inDrop
GSE114724	AziziPatient09Rep1	Human Breast Tumor	10x CR 2.1.1
GSE114724	AziziPatient09Rep2	Human Breast Tumor	10x CR 2.1.1
GSE114724	AziziPatient10Rep1	Human Breast Tumor	10x CR 2.1.1
GSE114724	AziziPatient11Rep1	Human Breast Tumor	10x CR 2.1.1
GSE114724	AziziPatient11Rep2	Human Breast Tumor	10x CR 2.1.1
SDY998	Zhang2019	Human Joint Synovial	CEL-seq2
GSE63473	Macosko2015	Mouse	Drop-seq
GSE77288	Tung2017	Human iPSC	HiSeq 2500
Tabula Muris	Tabula Muris	Mouse	10x (v2)

*Table 5.2: List of data sets used in the main text and supplementary materials. * CR: Cell Ranger*

5.3.2 Benchmarked Methods

In Figure 5.3, we benchmark Seurat 3.0.0 (Stuart et al., 2019) and SCTransform version 0.2.0 that is integrated with Seurat platform. Seurat was implemented following its guided tutorial (*https://satijalab.org/seurat/v3.1/pbmc3k_tutorial.html*), and SCTransform through a vignette (*https://rawgit.com/ChristophH/sctransform/master/inst/doc/seurat.html*). All parameters were selected through software's default except resolution parameter for clustering to generate results for various number of clusters. Seurat used in Figure 5.4 A was also the same version with the default parameters for feature selection. In Figure 5.3 A, the t-SNE plots were created using the features selected by the first round of HIPPO because they reflected the division of true cell labels the most accurately.

DCA was installed through Conda and imputation was performed following the tutorial on (*https://github.com/theislab/dca*). In one experiment, we first divide the data set into correct labels, and then impute them separately using DCA (imputing homogeneous cell population). In the other experiment, we impute both cell types together (imputing heterogeneous cell population). One property of DCA is that it automatically removes genes that are 0 in all the cells. Naturally, there are more such genes in homogeneous cell populations. Especially, some of the biomarkers are not expressed at all when cell population is divided into subtype. In that case, we imputed zero to those genes, assuming DCA didn't perform any imputation. (Figure 5.2 D, E). SAVER was downloaded from CRAN with version 1.1.1. In Supplementary Figure 6.20 transcriptome-level statistics were compared only using genes that had at least one positive count in each cell type. In both DCA and SAVER, all the parameters the default values as suggested by the software.

For likelihood ratio test in Figure 5.1 B and Figure 5.2 was conducted by fitting the distributions using the fitdistr function from *MASS* package (Venables & Ripley, 2002), and zero-inflated negative binomial distribution was fitted using *pscl* package (Jackman et al.,

2007).

5.3.3 Poisson Mixture Model

Consider a gene by cell matrix if UMI counts X for gene $g = 1, \cdots, G$ and cell $c = 1, \cdots, C$. To understand the behavior of the zeros for each gene, the first step is to reduce the information from each gene to the proportion of zeros across the cells

$$\hat{p}_g = \sum_{c=1}^{C} \frac{\mathbb{1}_{X_{gc}=0}}{C} \tag{5.3}$$

which is an estimator for the true zero proportion of gene g: p_g. We study its relationship against the mean expression for the set of cells, because p_g would decrease as the expression level increases. With the test statistic above, we test a one-sided hypothesis for each gene g, whether the zero proportion is higher than the expected rate under the Poisson model. For the alternative hypothesis, we believe that UMI counts follow finite Poisson Mixture. The hypotheses for each gene g are formally specified below.

$$H_0 : p = e^{-\lambda_g}, \qquad H_A : p = \sum_{k=1}^{K_g} \pi_k e^{-\lambda_{kg}}$$

In practice, we re-frame the hypotheses as $H_0 : K_g = 1, H_A : K_g > 1$ when $p = \sum_{k=1}^{K_g} \pi_k e^{-\lambda_{kg}}$. In other words, zero inflation indicates there is cell heterogeneity across the samples. If the cell population is truly homogeneous, the count data follows Poisson data with expected zero proportion $e^{-\lambda_g}$.

Chen (2018) and Sarkar (2020) demonstrates that most genes in UMI data follow Poisson distribution (Chen et al., 2018; Sarkar & Stephens, 2020) while other noisy genes follow Negative Binomial or Zero-Inflated Binomial distribution. Such model, although fundamentally different, is closely tied to the Poisson mixture model because Negative Binomial is the

limiting distribution of Gamma-Poisson. If λ_{cg} for each cell is drawn independently from the gamma distribution $\Gamma(r_g, \frac{1-p_g}{p_g})$, then $\sum_{c=1}^{C} X_{cg} \sim \frac{1}{C} Pois(\lambda_{cg}) \Leftrightarrow X_{cg} \sim NB\left(r, \frac{1-p_g}{p_g}\right)$. While Negative Binomial assumes a continuous mixture of Poisson, the proposed model assumes a finite mixture of Poisson, which is simpler and more directly addresses the source of zero inflation.

In practice, we do not explicitly estimate π_k, but instead simply test if observed $\hat{p}_g$ is larger than expected p with estimated gene mean λ. ($H_A : p_g > e^{-\lambda_g}$). It might seem counterintuitive that this test statistic does not fully leverage the specification of the alternative hypothesis; we never estimate the mixture parameters π_k. Alternatively, for example, one might suggest that we can conduct a likelihood ratio test of Poisson versus Poisson mixture. The main strength of the proposed reduced test statistic is its robustness to the modeling assumptions. Table 5.3 shows that the proportion of zeros are always larger than expected under different alternative hypotheses. Under the proposed alternative, mixture of Poisson, the proportion of zeros under the null hypothesis would be $e^{-\lambda}$ where λ is the weighted mean of the gene mean for each cell-type. Due to Jensen's inequality, p under alternative hypothesis is always greater than that under the H_0.

Underlying Distribution	p under H_0	p under H_A
Mixture of Poisson	$e^{-\lambda} = e^{-\sum_k \pi_k \lambda_k}$	$\sum_k \pi_k e^{-\lambda_k}$
Negative Binomial	$e^{-\lambda}$	$\left(\frac{r}{r+\lambda}\right)^r$
Zero-inflated Negative Binomial	$e^{-\lambda}$	$\left(\frac{r}{r+\lambda}\right)^r + \pi_0$

Table 5.3: The alternative hypothesis $H_A : p_g > e^{-\lambda}$ is robust to different model hypotheses. In the first row, the right column is larger than the left column due to Jensen's inequality. For negative binomial, the dispersion parameter r is constructed so that the variance is $\frac{\lambda^2}{r} + \lambda$, so that Poisson is a special case of Negative Binomial with $r = \infty$. The zero-inflated negative binomial distribution is parameterized as $\pi_0 \delta_0 + (1 - \pi_0)$ NB(λ, r)

Alternative Hypothesis	variance under H_0	variance under H_A
Mixture of Poisson	$\lambda = \sum_k \pi_k \lambda_k$	$\sum_k \pi_k(\lambda_k + \lambda_k^2) - (\sum_k \pi_k \lambda_k)^2$
Negative Binomial	λ	$\frac{\lambda^2}{r} + \lambda$
Zero-inflated Negative Binomial	λ	$(1-\pi_0)^2 \left(\frac{\lambda^2}{r} + \lambda\right)$

Table 5.4: High gene variance is not a good indicator of cell type heterogeneity under the alternative hypothesis of zero-inflated negative binomial, because the variance can be lower under the alternative hypothesis

5.3.4 Feature selection and Inference

For gene g with count data X_{gc} for cells $c = 1, \cdots, C$, consider an estimate for the proportion of zeros $\hat{p}_g$ as

$$\hat{p}_g = \frac{\sum_{c=1}^{C} \mathbb{1}_{X_{gc}=0}}{C}$$

The gene mean is estimated as the average UMI counts $\bar{X}_g = \frac{1}{C}\sum_{c=1}^{C} X_{gc}$ and is treated as a fixed number. Then,

$$\hat{p}_g = \mathcal{N}\left(e^{-\bar{X}}, \frac{\hat{p}_g(1-\hat{p}_g)}{C}\right)$$

The test statistic z-score for gene g is as below.

$$z_g = \frac{\hat{p}_g - e^{-\bar{X}}}{\frac{\hat{p}_g(1-\hat{p}_g)}{C}}$$

In reality, the gene mean $e^{-\bar{X}}$ is also a random variable that follows log normal variable. The inference is not trivial, and further discussion is in Supplementary Text 1.

5.3.5 Hierarchical Clustering

Algorithm 1 outlines the iterative procedure of HIPPO's hierarchical clustering. Several stopping criteria can be determined by the user: the maximum number of clusters K, the feature

selection statistic threshold z, and outlier gene proportion o. The algorithm first computes the number of outlier genes to allow, $G \times o$. For example, if there are 30,000 genes in total and o is specified as $1\% = 0.01$, then the algorithm allows 300 features to have zero inflation. During the clustering procedure, HIPPO terminates in either scenarios: there are K identified clusters or if there are less than $G \times o$ genes that exceed the specified z-value threshold.

HIPPO takes all the cells and select the features whose zero inflation statistic z exceeds the threshold. Then it is natural-log transformed (log(X)+1), and it goes through principal component decomposition, scaled and centered, only using the zero-inflated features. Then, using the cell embeddings (user can choose how many dimensions to use; the default is 10), HIPPO clusters the cells into two groups using the dimension-reduced cell embeddings through K-means, performed multiple times (user-defined, default 10) for stability.

Meanwhile, during the hierarchical clustering, HIPPO keeps track of intra-cluster variation by performing un-scaled, un-centered PCA. HIPPO computes the first 10 dimensions (user-defined) of cell embeddings and sum the sample variance of each component. The PCA for recording the intra-cluster variability is because the scaling introduces bias according to the cell population size. Since a subset of cells are considered for clustering at each round, fewer and fewer cells are applied to the dimension reduction. When scaled, their dimension-reduced cell embeddings would be artificially more far apart compared to when more cells were considered for clustering. The intra-cluster variance is the criterion for selecting which cell group to cluster in the next round.

It repeats the procedure so that each cluster in the end have the least intra-cluster variation.

Algorithm 1 Cell-Type Hierarchical Clustering

K: upper limit of cluster number

z-threshold: threshold for feature selection

$\ell = 1$

for $k = 2, \cdots, K$ **do**

  if Less than the designated number of genes exceed z threshold **then**

    stopping criterion; terminate algorithm

  else

    update the matrix by selecting new features

    log transformation + centered/scaled PCA + Kmeans

    divide cells into two groups, one with label ℓ and another with label k

    log transformation + un-centered/un-scaled PCA

    update intra-cluster variation by taking sample variance of un-scaled PCs

    update ℓ = cluster with the highest intra-cluster distance

  end if

end for

return cluster labels for each k

5.3.6 Differential Expression Testing

After the group labeling, we can do a similar but simpler hypothesis test to see if a certain gene is differentially expressed in two groups.

$$X_{cg}|c \in \mathcal{C}_1 \sim \text{Poisson}(\lambda_1), \qquad X_{cg}|c \in \mathcal{C}_2 \sim \text{Poisson}(\lambda_2)$$

$$H_0 : \lambda_1 = \lambda_2$$

We can use a 2-sample t-test and order the genes in the order of significance in the mean difference of two groups.

$$t = \frac{\bar{X}_{\mathcal{C}_1 g} - \bar{X}_{\mathcal{C}_2 g}}{\sqrt{\frac{\bar{X}_{\mathcal{C}_1 g}}{|\mathcal{C}_1|} + \frac{\bar{X}_{\mathcal{C}_2 g}}{|\mathcal{C}_2|}}}$$

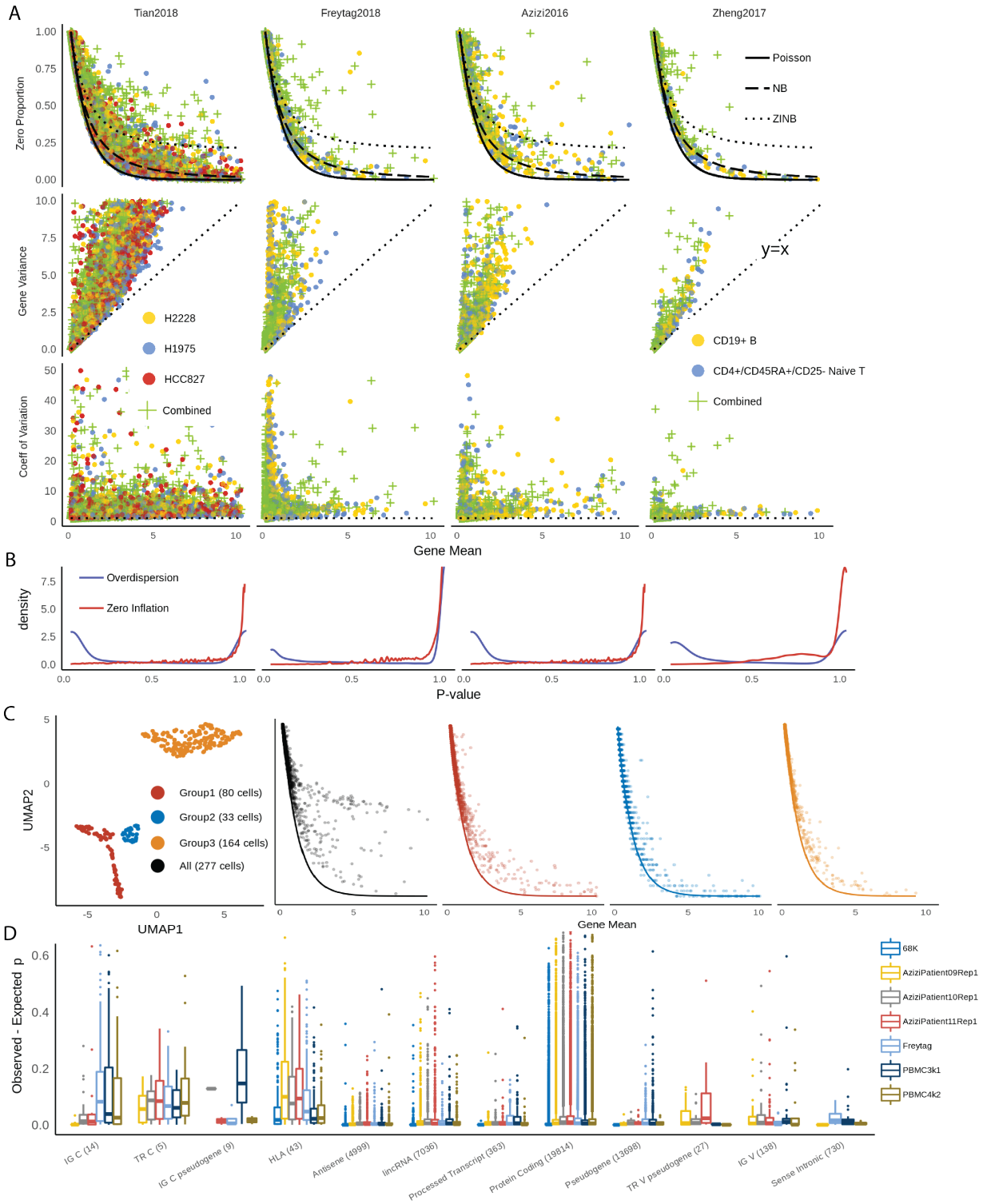


Figure 5.1: A. Comparisons of zero proportion, gene variance and CV as indicators for cellular heterogeneity in different UMI data sets. B. Distributions of p-values from likelihood ratio test for over-dispersion and zero-inflation. C. t-SNE plots of CD34+ cells in Zheng data, and relationship between zero proportions and gene means before (black) and after (colors) clustering of CD34+ cells. D. Distributions of zero inflation in different PBMC data sets. The x-axis labels represent gene types from GENCODE annotations and the number of genes within each type.

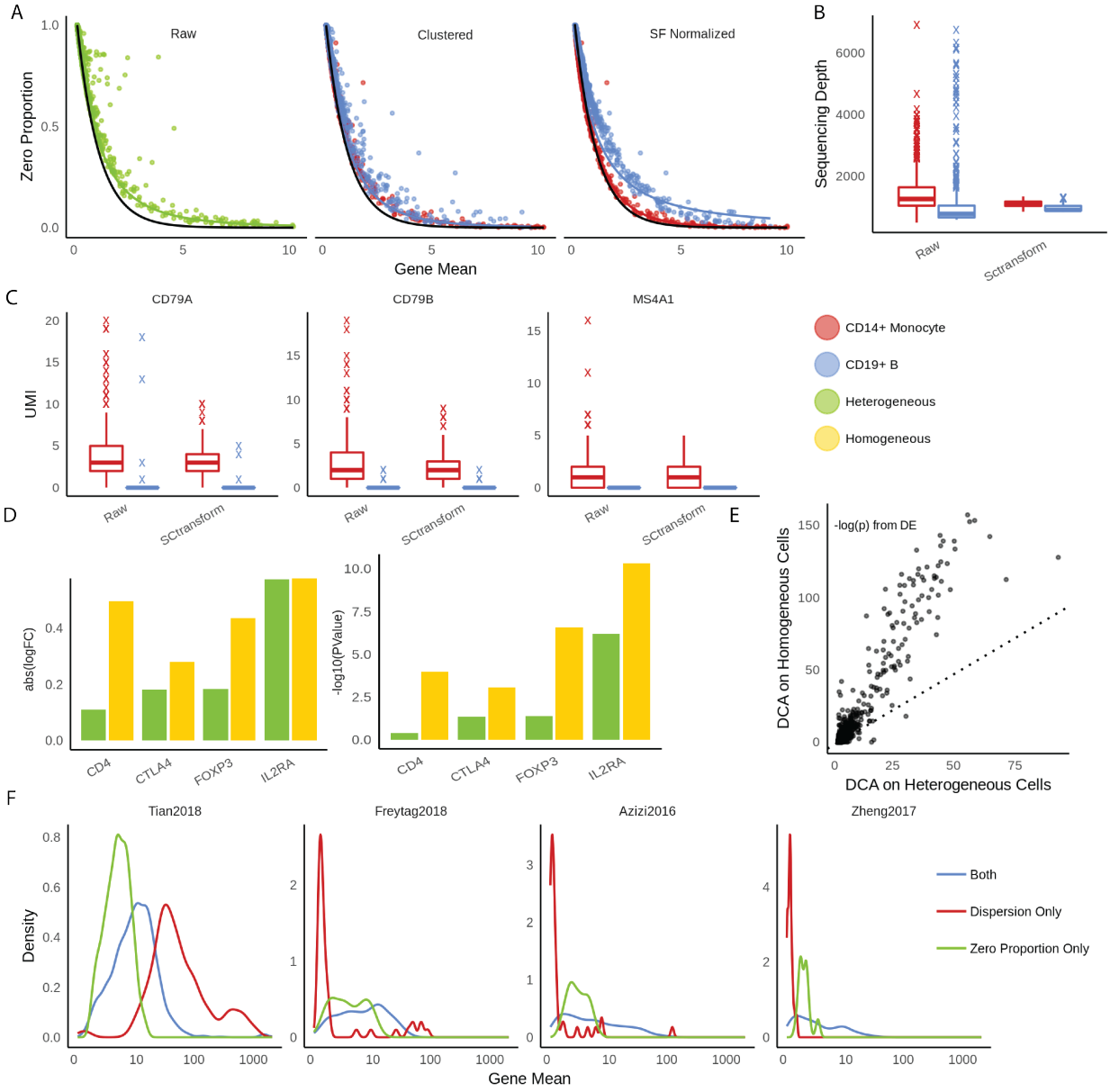


Figure 5.2: A. Scatterplots between gene means and zero proportions across genes calculated from raw UMI data, clustered data, and data after sequencing-depth normalization, respectively. Fitted line is negative binomial curve. B-C. Evaluations of pre-processing in Sctransform. B. Distributions of sequencing depths across cells in raw UMI data vs. data cleaned by SCtransform. C. Comparisons of three Monocyte markers in raw UMI data vs. data cleaned by SCtransform. D-E. Evaluations of pre-processing in DCA. D. Log fold changes and log p-values from differential expression analysis using the same data set but imputed by DCA as heterogeneous and homogeneous cell populations, respectively. E. The p-values comparisons between two different imputation strategies show the general deflation of biological signals of DCA when applied to heterogeneous cells. F. Comparisons of selected features using likelihood ratio test and zero proportions. Dispersion tends to select the genes that have mean UMI count close to 0.

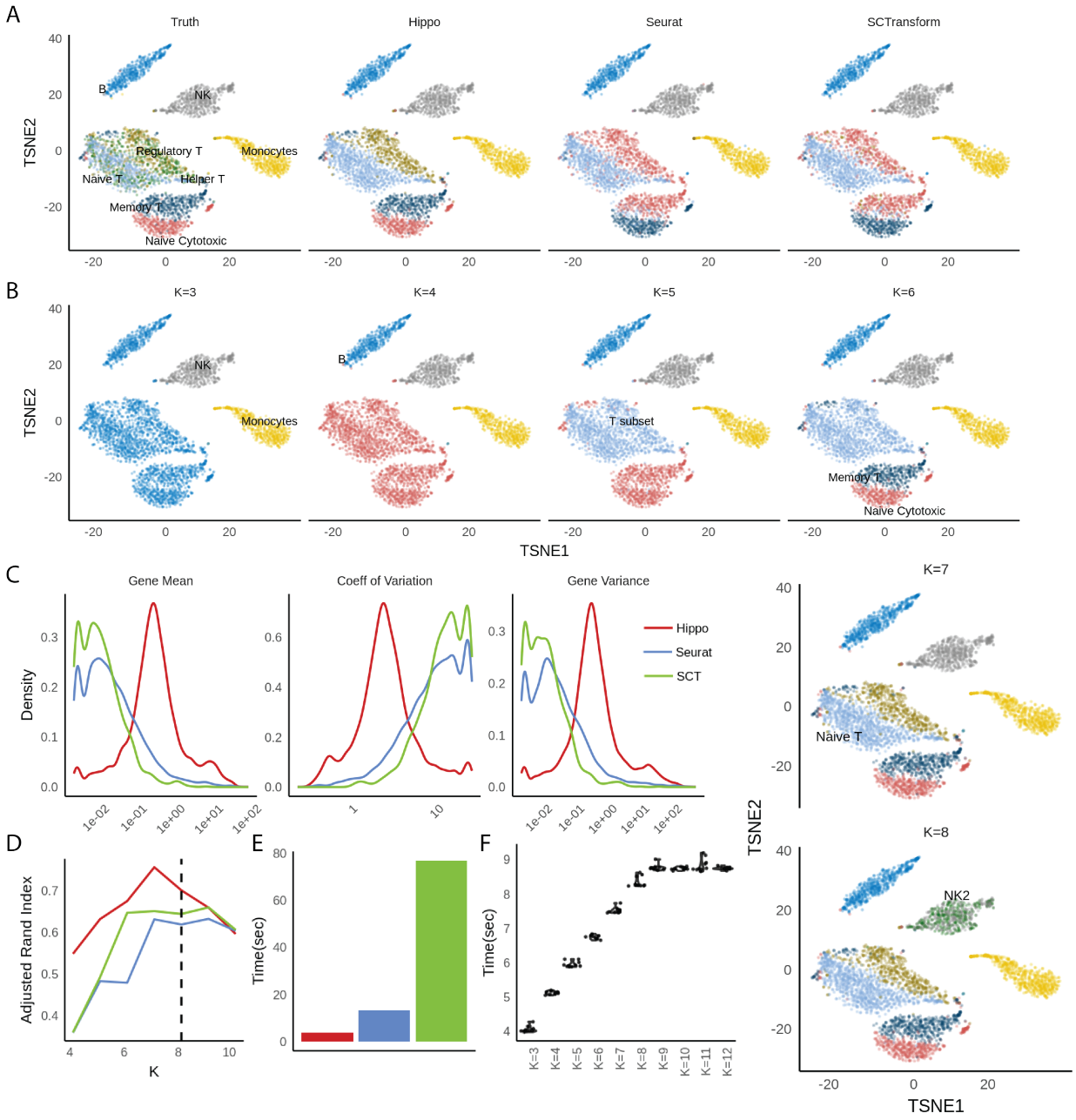


Figure 5.3: HIPPO framework applied to Zhengmix8eq data. A. t-SNE plots for clustering results from three methods: HIPPO, Seurat, and SCTransform, compared to true labels. Seurat and SCTransform cannot differentiate Helper T/Regulatory T and Memory T cells. B. HIPPO's sequential clustering results for $K = 3, \cdots, 8$. C. Comparisons of features selected by different methods for their gene mean, CV and variance. Seurat and SCTransform use CV as the selection criteria, and hence their features weigh heavily on genes with small mean expression and variance. D. Clustering results comparisons using Adjusted Rand Index. E. Computing time for each method using LAMBDA QUAD workstation with Intel Xeon W-2175 processor sequentially (non-parallel). F. Computing time for HIPPO using different k.

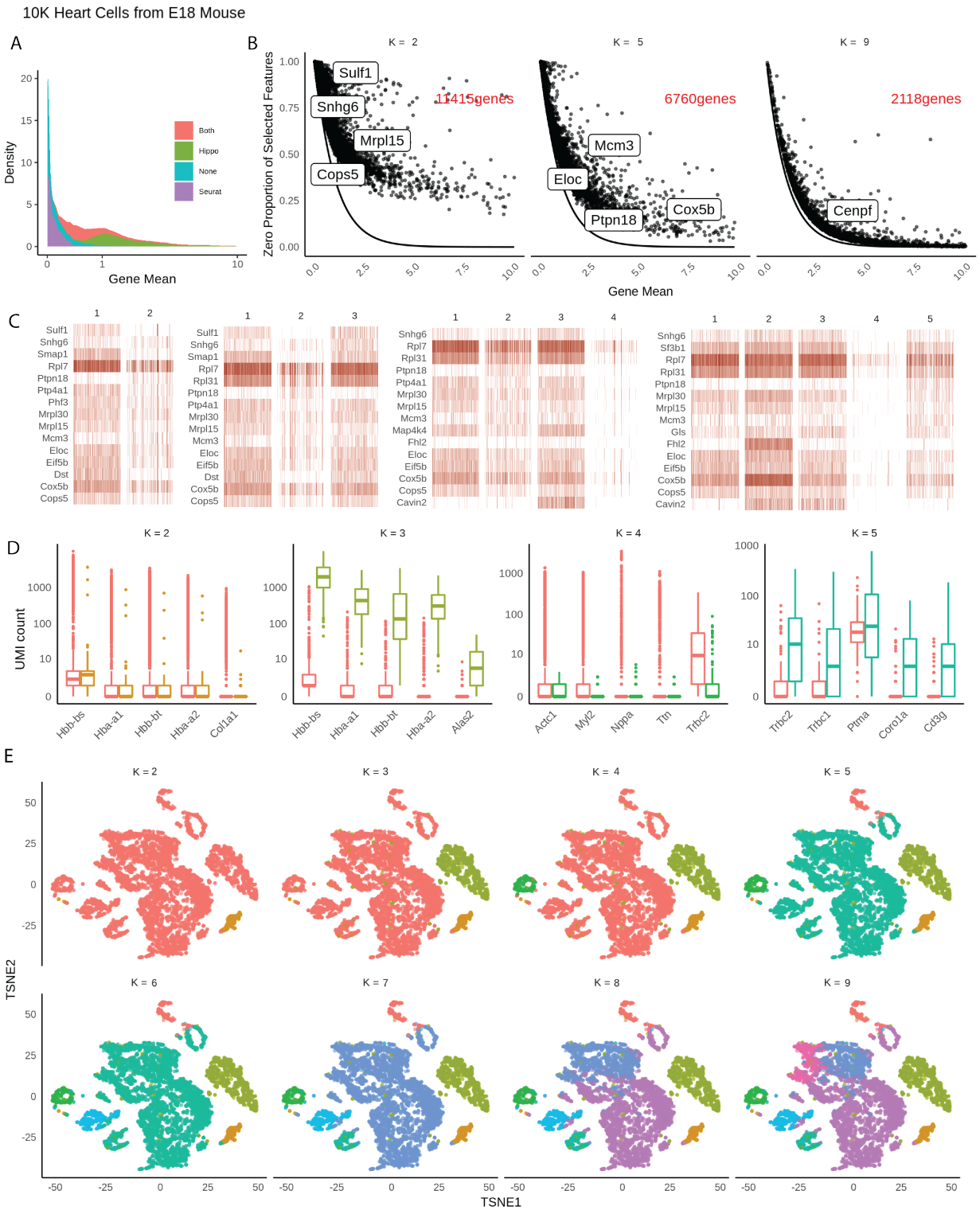


Figure 5.4: HIPPO framework applied to 10K E18 mouse heart cells. A. Distributions of means across gene features selected by Seurat, Hippo, both, or none. B. Sequential feature selection visualizes how genes gradually align closer to the expected Poisson line as more heterogeneity is accounted for. C. Heatmaps of top features selected at the first 5 rounds of clustering. D. Top differential expression genes obtained at each round of clustering. E. Visualization of sequential clustering using t-SNE plots.

BIBLIOGRAPHY

Amemiya, T. (1977): A note on a heteroscedastic model. *Journal of Econometrics* **6**(3):365–370.

Andrews, T. S. & Hemberg, M. (2018): M3Drop: Dropout-based feature selection for scRNASeq. *Bioinformatics* .

Azizi, E.; Carr, A. J.; Plitas, G.; Cornish, A. E.; Konopacki, C.; Prabhakaran, S.; Nainys, J.; Wu, K.; Kiseliovas, V.; Setty, M. et al. (2018): Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* **174**(5):1293–1308.

Baron, M.; Veres, A.; Wolock, S. L.; Faust, A. L.; Gaujoux, R.; Vetere, A.; Ryu, J. H.; Wagner, B. K.; Shen-Orr, S. S.; Klein, A. M. et al. (2016): A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems* **3**(4):346–360.

Benjamini, Y. & Hochberg, Y. (1995): Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* pages 289–300.

Breusch, T. S. & Pagan, A. R. (1979): A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society* pages 1287–1294.

Brown, P. J.; Vannucci, M. & Fearn, T. (1998): Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**(3):627–641.

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

Chawla, K.; Tripathi, S.; Thommesen, L.; Lægreid, A. & Kuiper, M. (2013): TFcheckpoint: a curated compendium of specific DNA-binding RNA polymerase II transcription factors. *Bioinformatics* **29**(19):2519–2520.

Chen, L. S.; Hsu, L.; Gamazon, E. R.; Cox, N. J. & Nicolae, D. L. (2012): An exponential combination procedure for set-based association tests in sequencing studies. *The American Journal of Human Genetics* **91**(6):977–986.

Chen, W.; Li, Y.; Easton, J.; Finkelstein, D.; Wu, G. & Chen, X. (2018): UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome biology* **19**(1):70.

Choi, K.; Chen, Y.; Skelly, D. A. & Churchill, G. A. (2020): Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. *bioRxiv* .

Clarke, L.; Fairley, S.; Zheng-Bradley, X.; Streeter, I.; Perry, E.; Lowy, E.; Tassé, A.-M. & Flicek, P. (2017): The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic acids research* **45**(D1):D854–D859.

Clivio, O.; Lopez, R.; Regier, J.; Gayoso, A.; Jordan, M. I. & Yosef, N. (2019): Detecting Zero-Inflated Genes in Single-Cell Transcriptomics Data. *bioRxiv* page 794875.

Cribari-Neto, F. & Ferrari, S. L. (1995): An improved Lagrange multiplier test for heteroskedasticity. *Communications in Statistics-Simulation and Computation* **24**(1):31–44.

Cribari-Neto, F. & Ferrari, S. L. (2001): Monotonic improved critical values for two $\chi 2$ asymptotic criteria. *Economics Letters* **71**(3):307–316.

Davidson, A. & Diamond, B. (2001): Autoimmune diseases. *New England Journal of Medicine* **345**(5):340–350.

Duò, A.; Robinson, M. D. & Soneson, C. (2018): A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research* **7**.

Eraslan, G.; Simon, L. M.; Mircea, M.; Mueller, N. S. & Theis, F. J. (2019): Single-cell RNA-seq denoising using a deep count autoencoder. *Nature communications* **10**(1):390.

Freytag, S.; Tian, L.; Lönnstedt, I.; Ng, M. & Bahlo, M. (2018): Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Research* **7**.

Fyr, C. L. W.; Kanaya, A. M.; Cummings, S. R.; Reich, D.; Hsueh, W.-C.; Reiner, A. P.; Harris, T. B.; Moffett, S.; Li, R.; Ding, J. et al. (2007): Genetic admixture, adipocytokines, and adiposity in Black Americans: the Health, Aging, and Body Composition study. *Human genetics* **121**(5):615–624.

George, E. I. & McCulloch, R. E. (1993): Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**(423):881–889.

George, E. I. & McCulloch, R. E. (1997): Approaches for Bayesian variable selection. *Statistica sinica* pages 339–373.

Glejser, H. (1969): A new test for heteroskedasticity. *Journal of the American Statistical Association* **64**(325):316–323.

Godfrey, L. G. (1978): Testing for multiplicative heteroskedasticity. *Journal of Econometrics* **8**(2):227–236.

Gong, W.; Kwak, I.-Y.; Pota, P.; Koyano-Nakagawa, N. & Garry, D. J. (2018): DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC bioinformatics* **19**(1):220.

Griffiths, W. & Surekha, K. (1986): A Monte Carlo evaluation of the power of some tests for heteroscedasticity. *Journal of Econometrics* **31**(2):219–231.

Guan, Y. & Stephens, M. (2011): Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics* pages 1780–1815.

Hafemeister, C. & Satija, R. (2019): Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *bioRxiv* page 576827.

Halder, I.; Shriver, M.; Thomas, M.; Fernandez, J. R. & Frudakis, T. (2008): A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. *Human mutation* **29**(5):648–658.

Harris, P. (1985): An asymptotic expansion for the null distribution of the efficient score statistic. *Biometrika* **72**(3):653–659.

Harrow, J.; Frankish, A.; Gonzalez, J. M.; Tapanari, E.; Diekhans, M.; Kokocinski, F.; Aken, B. L.; Barrell, D.; Zadissa, A.; Searle, S. et al. (2012): GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* **22**(9):1760–1774.

Hastings, W. K. (1970): Monte Carlo sampling methods using Markov chains and their applications .

Hernández-Lobato, D.; Hernández-Lobato, J. M. & Dupont, P. (2013): Generalized spike-and-slab priors for Bayesian group feature selection using expectation propagation. *The Journal of Machine Learning Research* **14**(1):1891–1945.

Honda, Y. (1988): A size correction to the Lagrange multiplier test for heteroskedasticity. *Journal of Econometrics* **38**(3):375–386.

Huang, D. W.; Sherman, B. T. & Lempicki, R. A. (2009): Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**(1):44–57.

Huang, M.; Wang, J.; Torre, E.; Dueck, H.; Shaffer, S.; Bonasio, R.; Murray, J. I.; Raj,

A.; Li, M. & Zhang, N. R. (2018): SAVER: gene expression recovery for single-cell RNA sequencing. *Nature methods* **15**(7):539.

Hughes, A. L. (1997): Rapid evolution of immunoglobulin superfamily C2 domains expressed in immune system cells. *Molecular biology and evolution* **14**(1):1–5.

Hurst, L. D. & Smith, N. G. (1999): Do essential genes evolve slowly? *Current biology* **9**(14):747–750.

Ideker, T. & Krogan, N. J. (2012): Differential network biology. *Molecular systems biology* **8**(1).

Islam, S.; Zeisel, A.; Joost, S.; La Manno, G.; Zajac, P.; Kasper, M.; Lönnerberg, P. & Linnarsson, S. (2014): Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature methods* **11**(2):163.

Jackman, S.; Kleiber, C.; Zeileis, A. et al. (2007): Regression Models for Count Data in R. Technical report.

Kiselev, V. Y.; Kirschner, K.; Schaub, M. T.; Andrews, T.; Yiu, A.; Chandra, T.; Natarajan, K. N.; Reik, W.; Barahona, M.; Green, A. R. et al. (2017): SC3: consensus clustering of single-cell RNA-seq data. *Nature methods* **14**(5):483.

Klein, A. M.; Mazutis, L.; Akartuna, I.; Tallapragada, N.; Veres, A.; Li, V.; Peshkin, L.; Weitz, D. A. & Kirschner, M. W. (2015): Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**(5):1187–1201.

Lee, K. H.; Tadesse, M. G.; Baccarelli, A. A.; Schwartz, J. & Coull, B. A. (2017): Multivariate Bayesian variable selection exploiting dependence structure among outcomes: Application to air pollution effects on DNA methylation. *Biometrics* **73**(1):232–241.

Lee, Y. (2015): Generalized principal component analysis. *Journal of Educational Psychology* **24**(6):417–441.

Leek, J. T.; Scharpf, R. B.; Bravo, H. C.; Simcha, D.; Langmead, B.; Johnson, W. E.; Geman, D.; Baggerly, K. & Irizarry, R. A. (2010): Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* **11**(10):733–739.

Li, G.; Shabalin, A. A.; Rusyn, I.; Wright, F. A. & Nobel, A. B. (2017): An empirical Bayes approach for multiple tissue eQTL analysis. *Biostatistics* **19**(3):391–406.

Li, K.-C. (2002): Genome-wide coexpression dynamics: theory and application. *Proceedings of the National Academy of Sciences* **99**(26):16875–16880.

Li, K.-C.; Liu, C.-T.; Sun, W.; Yuan, S. & Yu, T. (2004): A system for enhancing genome-wide coexpression dynamics study. *Proceedings of the National Academy of Sciences* **101**(44):15561–15566.

Liang, F.; Paulo, R.; Molina, G.; Clyde, M. A. & Berger, J. O. (2008): Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association* **103**(481):410–423.

Little, R. J. & Rubin, D. B. (2014): *Statistical analysis with missing data*, volume 333. John Wiley & Sons.

Love, M. I.; Huber, W. & Anders, S. (2014): Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**(12):550.

Macosko, E. Z.; Basu, A.; Satija, R.; Nemesh, J.; Shekhar, K.; Goldman, M.; Tirosh, I.; Bialas, A. R.; Kamitaki, N.; Martersteck, E. M. et al. (2015): Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**(5):1202–1214.

Mitchell, T. J. & Beauchamp, J. J. (1988): Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83**(404):1023–1032.

Mori, M.; Yamada, R.; Kobayashi, K.; Kawaida, R. & Yamamoto, K. (2005): Ethnic differences in allele frequency of autoimmune-disease-associated SNPs. *Journal of human genetics* **50**(5):264–266.

Morozova, O.; Hirst, M. & Marra, M. A. (2009): Applications of new sequencing technologies for transcriptome analysis. *Annual review of genomics and human genetics* **10**:135–151.

Moschopoulos, P. G. (1985): The distribution of the sum of independent gamma random variables. *Annals of the Institute of Statistical Mathssematics* **37**(1):541–544.

Nalls, M. A.; Wilson, J. G.; Patterson, N. J.; Tandon, A.; Zmuda, J. M.; Huntsman, S.; Garcia, M.; Hu, D.; Li, R.; Beamer, B. A. et al. (2008): Admixture mapping of white cell count: genetic locus responsible for lower white blood cell count in the Health ABC and Jackson Heart studies. *The American Journal of Human Genetics* **82**(1):81–87.

Narisetty, N. N.; He, X. et al. (2014): Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics* **42**(2):789–817.

Oba, S.; Sato, M.-a.; Takemasa, I.; Monden, M.; Matsubara, K.-i. & Ishii, S. (2003): A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* **19**(16):2088–2096.

Powers, S. K.; Ji, L. L. & Leeuwenburgh, C. (1999): Exercise training-induced alterations in skeletal muscle antioxidant capacity: a brief review. *Medicine and science in sports and exercise* **31**(7):987–997.

Price, A. L.; Patterson, N.; Hancks, D. C.; Myers, S.; Reich, D.; Cheung, V. G. & Spielman, R. S. (2008): Effects of cis and trans genetic ancestry on gene expression in African Americans. *PLoS Genet* **4**(12):e1000294.

Pritchard, J. K.; Stephens, M. & Donnelly, P. (2000): Inference of population structure using multilocus genotype data. *Genetics* **155**(2):945–959.

Pujato, M.; Kieken, F.; Skiles, A. A.; Tapinos, N. & Fiser, A. (2014): Prediction of DNA binding motifs from 3D models of transcription factors; identifying TLX3 regulated genes. *Nucleic acids research* **42**(22):13500–13512.

Rao, C. R. & Statistiker, M. (1973): *Linear statistical inference and its applications*, volume 2. Wiley New York.

Reiner, A. P.; Carlson, C. S.; Ziv, E.; Iribarren, C.; Jaquish, C. E. & Nickerson, D. A. (2007): Genetic ancestry, population sub-structure, and cardiovascular disease-related traits among African-American participants in the CARDIA Study. *Human genetics* **121**(5):565–575.

Risso, D.; Perraudeau, F.; Gribkova, S.; Dudoit, S. & Vert, J.-P. (2018): A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature communications* **9**(1):1–17.

Robert, C. & Casella, G. (2013): *Monte Carlo statistical methods*. Springer Science & Business Media.

Robinson, M. D.; McCarthy, D. J. & Smyth, G. K. (2010): edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1):139–140.

Rubin, D. B. (1976): Inference and missing data. *Biometrika* **63**(3):581–592.

Saelens, W.; Cannoodt, R.; Todorov, H. & Saeys, Y. (2019): A comparison of single-cell trajectory inference methods. *Nature biotechnology* **37**(5):547.

Sarkar, A. K. & Stephens, M. (2020): Separating measurement and expression models clarifies confusion in single cell RNA-seq analysis. *BioRxiv* .

Schelker, M.; Feau, S.; Du, J.; Ranu, N.; Klipp, E.; MacBeath, G.; Schoeberl, B. & Raue,

A. (2017): Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nature communications* **8**(1):2032.

Seldin, M. F.; Pasaniuc, B. & Price, A. L. (2011): New approaches to disease mapping in admixed populations. *Nature Reviews Genetics* **12**(8):523.

Shabalin, A. A. (2012): Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**(10):1353–1358.

Spurgin, L. G. & Richardson, D. S. (2010): How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings of the Royal Society B: Biological Sciences* **277**(1684):979–988.

Stuart, T.; Butler, A.; Hoffman, P.; Hafemeister, C.; Papalexi, E.; Mauck III, W. M.; Hao, Y.; Stoeckius, M.; Smibert, P. & Satija, R. (2019): Comprehensive integration of single-cell data. *Cell* **177**(7):1888–1902.

Tabula Muris Consortium et al. (2018): Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**(7727):367.

Tang, F.; Barbacioru, C.; Wang, Y.; Nordman, E.; Lee, C.; Xu, N.; Wang, X.; Bodeau, J.; Tuch, B. B.; Siddiqui, A. et al. (2009): mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods* **6**(5):377.

The GTEx Consortium (2015): The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**(6235):648–660.

Tian, C.; Hinds, D. A.; Shigeta, R.; Kittles, R.; Ballinger, D. G. & Seldin, M. F. (2006): A genomewide single-nucleotide–polymorphism panel with high ancestry information for African American admixture mapping. *The American Journal of Human Genetics* **79**(4):640–649.

Tian, L.; Dong, X.; Freytag, S.; Le Cao, K.-A.; Su, S.; Amann-Zalcenstein, D.; Weber, T. S.; Seidi, A.; Naik, S. & Ritchie, M. E. (2018): scRNA-seq mixology: towards better benchmarking of single cell RNA-seq protocols and analysis methods. *BioRxiv* page 433102.

Townes, F. W.; Hicks, S. C.; Aryee, M. J. & Irizarry, R. A. (2019): Feature Selection and Dimension Reduction for Single Cell RNA-Seq based on a Multinomial Model. *bioRxiv* page 574574.

Townes, F. W. & Street, K. (2020): *scry: Small-Count Analysis Methods for High-Dimensional Data*. R package version 1.1.0.

Trapnell, C.; Cacchiarelli, D.; Grimsby, J.; Pokharel, P.; Li, S.; Morse, M.; Lennon, N. J.; Livak, K. J.; Mikkelsen, T. S. & Rinn, J. L. (2014): The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* **32**(4):381.

Tung, P.-Y.; Blischak, J. D.; Hsiao, C. J.; Knowles, D. A.; Burnett, J. E.; Pritchard, J. K. & Gilad, Y. (2017): Batch effects and the effective design of single-cell gene expression studies. *Scientific reports* **7**:39921.

Vallejos, C. A.; Risso, D.; Scialdone, A.; Dudoit, S. & Marioni, J. C. (2017): Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nature methods* **14**(6):565.

Venables, W. N. & Ripley, B. D. (2002): *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wang, B.; Zhu, J.; Pierson, E.; Ramazzotti, D. & Batzoglou, S. (2017): Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nature methods* **14**(4):414.

Wang, J.; Agarwal, D.; Huang, M.; Hu, G.; Zhou, Z.; Ye, C. & Zhang, N. R. (2019): Data

denoising with transfer learning in single-cell transcriptomics. *Nature methods* **16**(9):875–878.

Wang, J.; Gamazon, E. R.; Pierce, B. L.; Stranger, B. E.; Im, H. K.; Gibbons, R. D.; Cox, N. J.; Nicolae, D. L. & Chen, L. S. (2016): Imputing gene expression in uncollected tissues within and beyond GTEx. *The American Journal of Human Genetics* **98**(4):697–708.

White, H. et al. (1980): A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *econometrica* **48**(4):817–838.

Winkler, C. A.; Nelson, G. W. & Smith, M. W. (2010a): Admixture mapping comes of age. *Annual review of genomics and human genetics* **11**:65–89.

Winkler, C. A.; Nelson, G. W. & Smith, M. W. (2010b): Admixture mapping comes of age. *Annual review of genomics and human genetics* **11**:65–89.

Yan, Y.; Qiu, S.; Jin, Z.; Gong, S.; Bai, Y.; Lu, J. & Yu, T. (2017): Detecting subnetwork-level dynamic correlations. *Bioinformatics* **33**(2):256–265.

Yu, T. (2018): A new dynamic correlation algorithm reveals novel functional aspects in single cell and bulk RNA-seq data. *PLoS computational biology* **14**(8):e1006391.

Zappia, L. & Oshlack, A. (2018): Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *GigaScience* **7**(7).

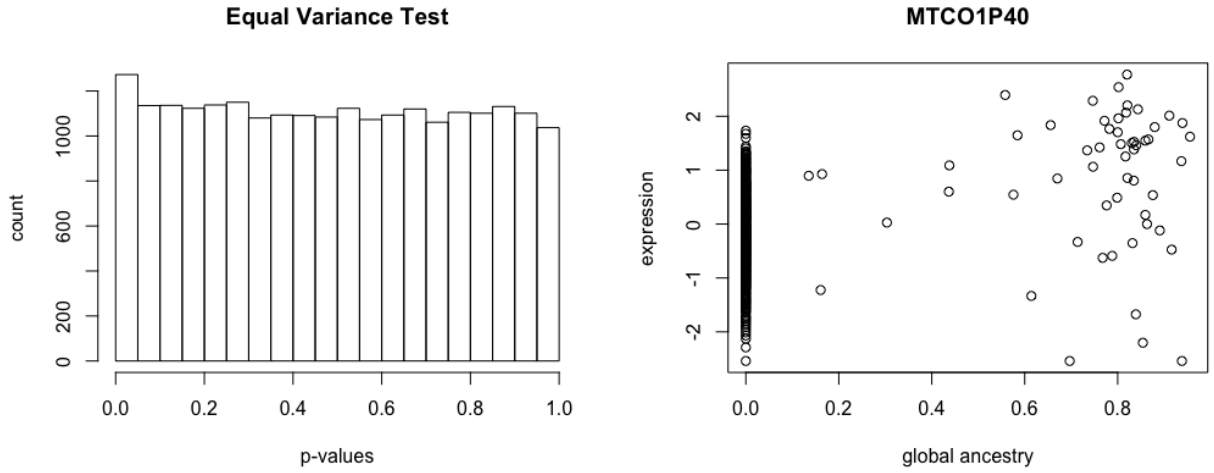
Zhang, F.; Wei, K.; Slowikowski, K.; Fonseka, C. Y.; Rao, D. A.; Kelly, S.; Goodman, S. M.; Tabechian, D.; Hughes, L. B.; Salomon-Escoto, K. et al. (2019): Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. *Nature immunology* **20**(7):928–942.

Zheng, G. X.; Terry, J. M.; Belgrader, P.; Ryvkin, P.; Bent, Z. W.; Wilson, R.; Ziraldo, S. B.; Wheeler, T. D.; McDermott, G. P.; Zhu, J. et al. (2017): Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**:14049.

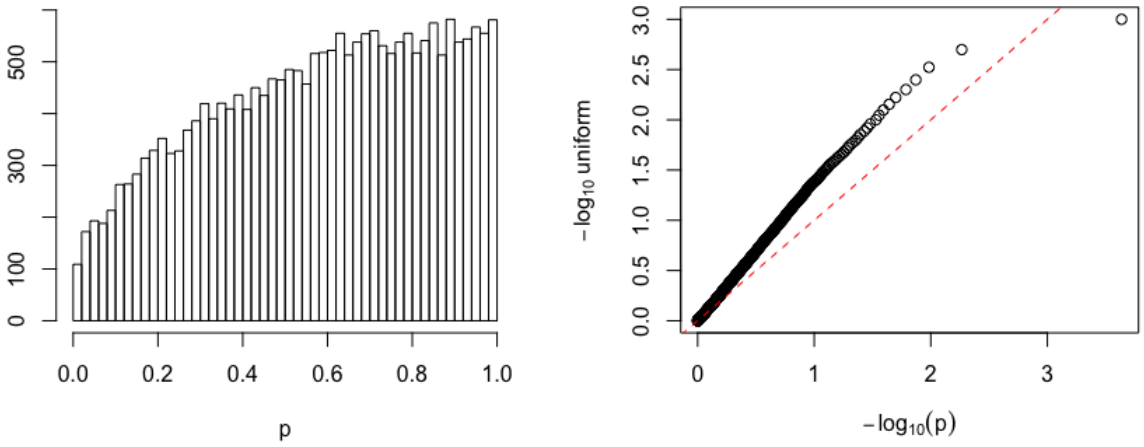
CHAPTER 6

SUPPLEMENTARY PLOTS AND FIGURES

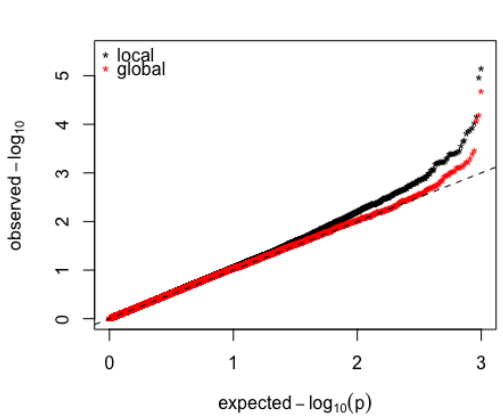
6.1 Supporting Evidence for Chapter 3



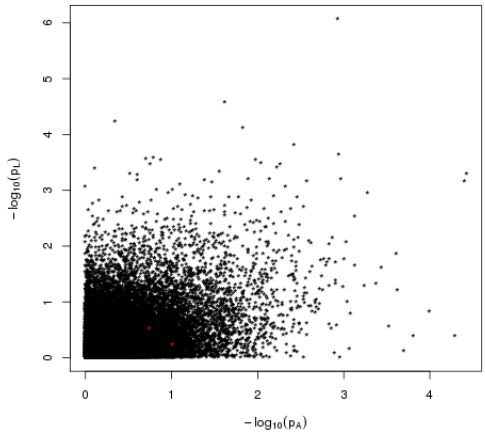
Supplementary Figure 6.1: (a) Histogram of p-values from variance ratio F-test of 22,248 genes. (b) The worst case gene with p-value of $7.04e-06$ *from the F-test.*



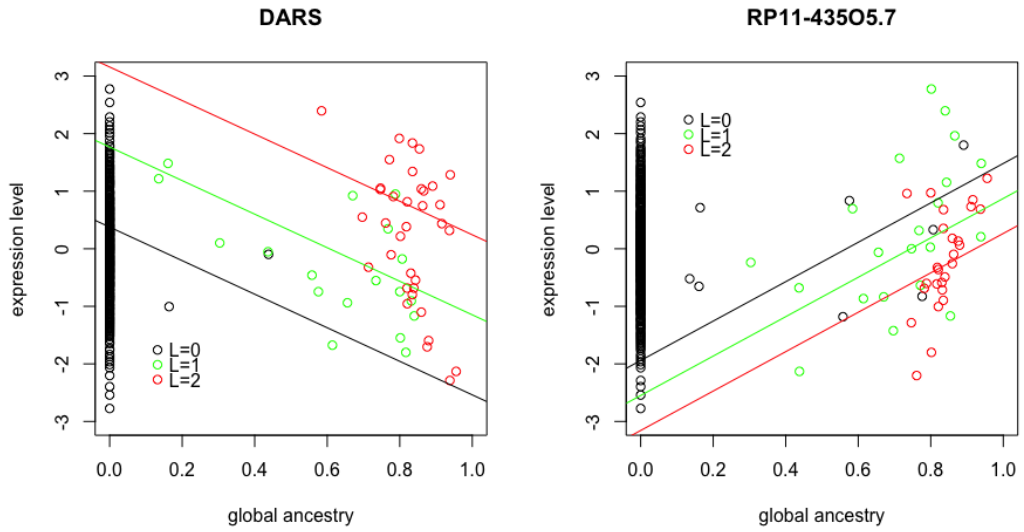
Supplementary Figure 6.2: The p-values from the same mean t-test between European Americans and African Americans with local ancestry 0.



(a) QQ-plot for p-values.



(b) p-values for local and global ancestry.

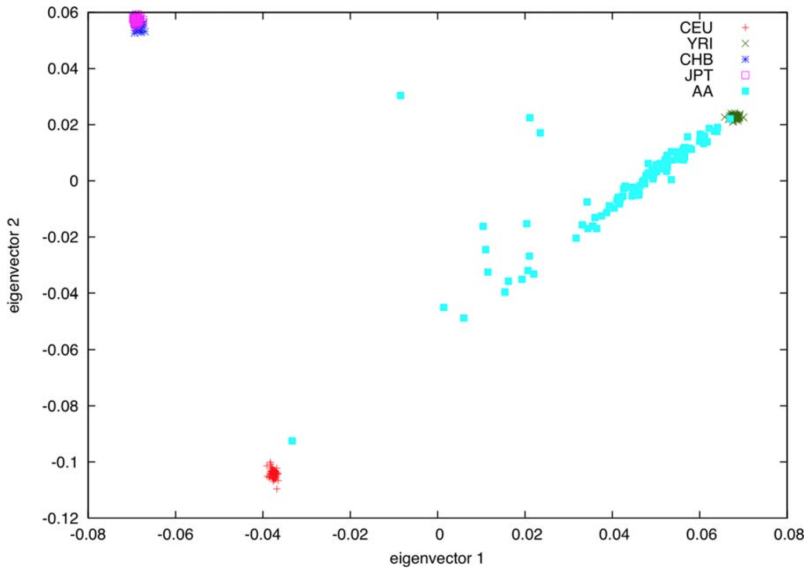


(c)

Supplementary Figure 6.3: (a),(b): Strengths of the marginal effects of local and global ancestry. (c) Genes that showed the strongest local ancestry effect and global ancestry effect respectively. Their fitted models are summarized in Supplementary Table 6.1

	DARS			RP11-435O5.7		
	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	0.38	0.59	0.51	-1.94	0.56	0.54e-04
$\mathbb{1}_{EA}$	0.013	0.52	0.98	1.75	0.51	0.67e-04
Global	-2.92	0.89	1.15e-03	3.42	0.79	2.12e-05
Local	1.39	0.31	7.37e-06	-0.61	0.21	4.2e-03

Supplementary Table 6.1: The fitted models for the plotted genes. The covariates like gender and age were fitted but not shown, and not taken into account in the plots (c).



Supplementary Figure 6.4: PCA of pure and admixed populations' genotypes The reason we focus particularly on CEU and YRI populations is shown in this principal component analysis plot. Here, the red dots are CEU and the green crosses are YRI, and you can see that most of the African Americans, the blue squares, are lying strictly between those two populations. The paper mentions that four genetic outliers were removed from the sample, and I'm guessing those four are these ones that deviate a lot from the line.

6.2 Supporting Evidence for Chapter 4

6.2.1 Appendix A. Derivation of the test statistic

Consider the model in (4.1).

$$\begin{bmatrix} y_{i1} \\ y_{i2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{z}_i^T \boldsymbol{\beta}_1 \\ \boldsymbol{z}_i^T \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix}$$

$$\begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_i = \begin{bmatrix} \sigma_1^2 & \rho(\tilde{\boldsymbol{x}}_i^T \tilde{\boldsymbol{\alpha}}) \\ \rho(\tilde{\boldsymbol{x}}_i^T \tilde{\boldsymbol{\alpha}}) & \sigma_2^2 \end{bmatrix} \right)$$

where $\tilde{\boldsymbol{\alpha}} = \begin{bmatrix} \alpha_0 & \boldsymbol{\alpha}^T \end{bmatrix}$, and $\tilde{\boldsymbol{x}}_i$ is $\begin{bmatrix} 1 & \boldsymbol{x}_i \end{bmatrix}^T$. Let $\ell(\theta)$ be the log likelihood depending on a vector of parameters $\theta^T = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \sigma_1^2 & \sigma_2^2 & \alpha_0 & \boldsymbol{\alpha} \end{bmatrix}$, with $d = \frac{\partial \ell}{\partial \theta}$ as the first derivative (score) vector and $\mathcal{I} = -E(\partial^2 \ell / \partial\theta \partial\theta^T)$ as the information matrix. Then following Rao & Statistiker (1973), the score statistic for testing the null hypothesis represented by parametric constraints $\phi(\theta) = 0$, equivalent to $\boldsymbol{\alpha} = 0$ in our context, is given by

$$q = \hat{d}^T \hat{\mathcal{I}}^{-1} \hat{d}$$

where the hats indicate that the quantities are evaluated with $\hat{\theta}$, the restricted maximum likelihood estimate satisfying $\phi(\hat{\theta}) = 0$:

$$\hat{\theta} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 & \hat{\boldsymbol{\beta}}_2 & \hat{\sigma}_1^2 & \hat{\sigma}_2^2 & \hat{\alpha}_0 & \mathbf{0} \end{bmatrix},$$

where $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ are linear regression coefficients. The error term u can be replaced with the OLS residuals $\hat{u}$. Then, $\hat{\sigma}_1^2 = \sum_{i=1}^N u_{i1}^2/N$ and $\hat{\sigma}_2^2 = \sum_{i=1}^N u_{i2}^2/N$. $\hat{\alpha}_0$ is a value that makes $\rho(\hat{\alpha}_0) = \hat{\rho} = \sum_{i=1}^N u_{i1} u_{i2}/N$.

Consider partitioning θ into $\begin{bmatrix}\theta_1; & \theta_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{\beta}_1 & \boldsymbol{\beta}_2; & \sigma_1^2 & \sigma_2^2 & \alpha_0 & \boldsymbol{\alpha}\end{bmatrix}$. Then, the constraints refer to only one of the subsets $\phi(\theta_2) = 0$. The score vector d can be partitioned similarly as $d = \begin{bmatrix}d_1 & d_2\end{bmatrix}$. Furthermore, the information matrix is block diagonal between θ_1 and θ_2, so that $\mathcal{I}_{21} = -E(\partial^2\ell/\partial\theta_2\partial\theta_1^T) = 0$. $\hat{d}_1$ is $\mathbf{0}$. Hence, the score statistic becomes

$$q = \hat{d}_2^T \hat{\mathcal{I}}_{22}^{-1} \hat{d}_2.$$

Score vector

This section derives the score vector with respect to σ_1^2, σ_2^2, $\tilde{\boldsymbol{\alpha}}$. For the variance parameter, for example $\tilde{\boldsymbol{\alpha}}$, the derivative of the log likelihood is as follows:

$$\frac{\partial \ell}{\partial \tilde{\boldsymbol{\alpha}}} = -\frac{1}{2}\sum_{i=1}^{N}\left(\frac{\partial log|\Sigma_i|}{\partial \tilde{\boldsymbol{\alpha}}} + \frac{\partial \boldsymbol{u}_i^T \Sigma^{-1} \boldsymbol{u}_i}{\partial \tilde{\boldsymbol{\alpha}}}\right)$$

First term:

$$\begin{aligned}
\frac{\partial log(|\Sigma_i|)}{\partial \sigma_1^2} &= \frac{\sigma_2^2}{\sigma_1^2\sigma_2^2 - \rho_i^2} \\
\frac{\partial log(|\Sigma_i|)}{\partial \sigma_2^2} &= \frac{\sigma_1^2}{\sigma_1^2\sigma_2^2 - \rho_i^2} \\
\frac{\partial log(|\Sigma_i|)}{\partial \tilde{\boldsymbol{\alpha}}} &= \frac{-2\rho_i\rho_i'}{\sigma_1^2\sigma_2^2 - \rho_i^2} x_i
\end{aligned}$$

where $\rho_i' = \frac{\partial \rho(\tilde{\boldsymbol{\alpha}}^T \tilde{\boldsymbol{x}}_i)}{\partial \tilde{\boldsymbol{\alpha}}}$.

$$
\begin{aligned}
\boldsymbol{u}_i^T \Sigma_i^{-1} \boldsymbol{u}_i &= \frac{1}{\sigma_1^2\sigma_2^2 - \rho_i^2} \begin{bmatrix} u_{i1} & u_{i2} \end{bmatrix} \begin{bmatrix} \sigma_2^2 & -\rho_i \\ -\rho_i & \sigma_1^2 \end{bmatrix} \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix} \\
&= \frac{1}{\sigma_1^2\sigma_2^2 - \rho_i^2} \begin{bmatrix} u_{i1}\sigma_2^2 - u_{i2}\rho_i & -u_{i1}\rho_i + u_{i2}\sigma_1^2 \end{bmatrix} \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix} \\
&= \frac{1}{\sigma_1^2\sigma_2^2 - \rho_i^2} \left(u_{i1}^2\sigma_2^2 - u_{i1}u_{i2}\rho_i - u_{i1}u_{i2}\rho_i + u_{i2}^2\sigma_1^2 \right) \\
&= \frac{1}{\sigma_1^2\sigma_2^2 - \rho_i^2} (u_{i1}^2\sigma_2^2 + u_{i2}^2\sigma_1^2 - 2u_{i1}u_{i2}\rho_i)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \boldsymbol{u}_i^T \Sigma^{-1} \boldsymbol{u}_i}{\partial \sigma_1^2} &= \frac{u_{i2}^2(\sigma_1^2\sigma_2^2 - \rho_i^2) - (u_{i2}^2\sigma_1^2 + u_{i1}^2\sigma_2^2 - 2u_{i1}u_{i2}\rho_i)\sigma_2^2}{(\sigma_1^2\sigma_2^2 - \rho_i^2)^2} \\
\frac{\partial \boldsymbol{u}_i^T \Sigma^{-1} \boldsymbol{u}_i}{\partial \sigma_2^2} &= \frac{u_{i1}^2(\sigma_1^2\sigma_2^2 - \rho_i^2) - (u_{i2}^2\sigma_1^2 + u_{i1}^2\sigma_2^2 - 2u_{i1}u_{i2}\rho_i)\sigma_1^2}{(\sigma_1^2\sigma_2^2 - \rho_i^2)^2} \\
\frac{\partial \boldsymbol{u}_i^T \Sigma^{-1} \boldsymbol{u}_i}{\partial \tilde{\boldsymbol{\alpha}}} &= \frac{-2u_{i1}u_{i2}\rho_i'(\sigma_1^2\sigma_2^2 - \rho_i^2) - (u_{i2}^2\sigma_1^2 + u_{i1}^2\sigma_2^2 - 2u_{i1}u_{i2}\rho_i)(-2\rho_i)(\rho_i')}{(\sigma_1^2\sigma_2^2 - \rho_i^2)^2} \tilde{\boldsymbol{x}}_i \\
&= \frac{u_{i1}u_{i2}(-2\rho_i'(\sigma_1^2\sigma_2^2 - \rho_i^2) - 4\rho_i^2\rho_i') + (2u_{i2}^2\sigma_1^2\rho_i + 2u_{i1}^2\sigma_2^2\rho_i)\rho_i'}{(\sigma_1^2\sigma_2^2 - \rho_i^2)^2} \tilde{\boldsymbol{x}}_i \\
&= 2\rho_i' \frac{u_{i2}^2\sigma_1^2\rho_i + u_{i1}^2\sigma_2^2\rho_i - u_{i1}u_{i2}(\sigma_1^2\sigma_2^2 - \rho_i^2 + 2\rho_i^2)}{(\sigma_1^2\sigma_2^2 - \rho_i^2)^2} \tilde{\boldsymbol{x}}_i
\end{aligned}
$$

Plugging in the MLEs for each nuisance parameter, we can replace

$$\sigma_1^2 \text{ with } \hat{\rho}_{11} = \sum_{i=1}^{N} \hat{u}_{i1}^2/N$$

$$\sigma_2^2 \text{ with } \hat{\sigma}_2^2 = \sum_{i=1}^{N} \hat{u}_{i2}^2/N$$

$$\tilde{\boldsymbol{\alpha}} \text{ with } \hat{\tilde{\boldsymbol{\alpha}}} = (\hat{\alpha}_0, \mathbf{0})$$

$$\rho \text{ with } \hat{\rho} = \sum_{i=1}^{N} \hat{u}_{i1}\hat{u}_{i2}/N$$

$$\rho' \text{ with } \hat{\rho}' = \frac{\partial \rho(\tilde{\boldsymbol{\alpha}}^T \tilde{\boldsymbol{x}}_i)}{\partial \tilde{\boldsymbol{\alpha}}} \mid_{\theta=\hat{\theta}}$$

$$\begin{aligned}
\hat{d}_{\sigma_1^2} &= \frac{\partial \ell}{\partial \sigma_1^2} \mid_{\theta=\hat{\theta}} \\
&= -\frac{1}{2}\sum_{i=1}^{N}\left(\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2} + \frac{\hat{u}_{i2}^2(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2) - (\hat{u}_{i2}^2\hat{\sigma}_1^2 + \hat{u}_{i1}^2\hat{\sigma}_2^2 - 2\hat{u}_{i1}\hat{u}_{i2}\hat{\rho})\hat{\sigma}_2^2}{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2}\right) \\
&= -\frac{1}{2}\left(\frac{\hat{\sigma}_2^2 N}{\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2} + \frac{\sum \hat{u}_{i2}^2(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2) - (\hat{\sigma}_1^2\sum_i \hat{u}_{i2}^2 + \hat{\sigma}_2^2\sum \hat{u}_{i1}^2 - 2\hat{\rho}\sum \hat{u}_{i1}\hat{u}_{i2})\hat{\sigma}_2^2}{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2}\right) \\
&= -\frac{N}{2}\left(\frac{\hat{\sigma}_2^2(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)}{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2} + \frac{\hat{\sigma}_2^2(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2) - (\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\sigma}_1^2\hat{\sigma}_2^2 - 2\hat{\rho}^2))}{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2}\right) \\
&= -\frac{\hat{\sigma}_2^2 N}{2}\left(\frac{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2) + (\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2) - (2\hat{\sigma}_1^2\hat{\sigma}_2^2 - 2\hat{\rho}^2)}{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2}\right) \\
&= 0
\end{aligned}$$

Similarly, $\hat{d}_{\sigma_2^2} = 0$.

$$\begin{aligned}
\hat{d}_{\tilde{\boldsymbol{\alpha}}} &= -\frac{1}{2}\sum_{i=1}^{N}\left(\frac{-2\hat{\rho}\hat{\rho}'}{\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2} + \frac{2\hat{u}_{i2}^2\hat{\sigma}_1^2\hat{\rho} + 2\hat{u}_{i1}^2\hat{\sigma}_2^2\hat{\rho} - 2\hat{u}_{i1}\hat{u}_{i2}(\hat{\sigma}_1^2\hat{\sigma}_2^2 - 3\hat{\rho}^2)}{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2}\hat{\rho}'\right)\tilde{\boldsymbol{x}}_i \\
&= \frac{\hat{\rho}\hat{\rho}'\sum \tilde{\boldsymbol{x}}_i}{\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2} - \frac{\hat{\sigma}_1^2\hat{\rho}\hat{\rho}'\sum \hat{u}_{i2}^2\tilde{\boldsymbol{x}}_i + \hat{\sigma}_2^2\hat{\rho}\hat{\rho}'\sum_i \hat{u}_{i1}^2\tilde{\boldsymbol{x}}_i - (\hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}' + \hat{\rho}^2\hat{\rho}')\sum \hat{u}_{i1}\hat{u}_{i2}\tilde{\boldsymbol{x}}_i}{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2} \\
&= \hat{\rho}' \cdot \frac{(\hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho} - \hat{\rho}^3)\sum \tilde{\boldsymbol{x}}_i - \hat{\sigma}_1^2\hat{\rho}\sum \hat{u}_{i2}^2\tilde{\boldsymbol{x}}_i - \hat{\sigma}_2^2\hat{\rho}\sum_i \hat{u}_{i1}^2\tilde{\boldsymbol{x}}_i + (\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}^2)\sum_i \hat{u}_{i1}\hat{u}_{i2}\tilde{\boldsymbol{x}}_i}{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2} \\
&= \hat{\rho}' \cdot \left(\frac{\hat{\rho}(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)\sum \tilde{\boldsymbol{x}}_i + (\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}^2)\sum_i \hat{u}_{i1}\hat{u}_{i2}\tilde{\boldsymbol{x}}_i - \hat{\sigma}_1^2\hat{\rho}\sum \hat{u}_{i2}^2\tilde{\boldsymbol{x}}_i - \hat{\sigma}_2^2\hat{\rho}\sum_i \hat{u}_{i1}^2\tilde{\boldsymbol{x}}_i}{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2}\right)
\end{aligned}$$

Consider another partitioning into $\hat{d}_{\tilde{\boldsymbol{\alpha}}} = \left[\hat{d}_{\alpha_0};\ \ \hat{d}_{\boldsymbol{\alpha}}\right]$. Below, we show that $\hat{d}_{\alpha_0} = 0$.

$$\begin{aligned}
&\hat{\rho}(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)\sum_{i=1}^{N} 1 + (\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}^2)\sum_i \hat{u}_{i1}\hat{u}_{i2} - \hat{\sigma}_1^2\hat{\rho}\sum \hat{u}_{i2}^2 - \hat{\sigma}_2^2\hat{\rho}\sum_i \hat{u}_{i1}^2 \\
&= \hat{\rho}(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)N + (\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}^2)(N\hat{\rho}) - \hat{\sigma}_1^2\hat{\rho}(N\hat{\sigma}_2^2) - \hat{\sigma}_2^2\hat{\rho}(N\hat{\sigma}_1^2) \\
&= 2N\hat{\rho}(\hat{\sigma}_1^2\hat{\sigma}_2^2) - 2N\hat{\rho}\hat{\sigma}_1^2\hat{\sigma}_2^2 \\
&= 0
\end{aligned}$$

Fisher Information

The Fisher information for two variance parameters, for example σ_1^2 and σ_2^2, is

$$\mathcal{I}_{\sigma_1^2,\sigma_2^2} = \frac{1}{2}tr\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\sigma_1^2}\Sigma^{-1}\frac{\partial\Sigma}{\partial\sigma_2^2}\right)$$

$$\begin{aligned}
\mathcal{I}(\theta)_{\sigma_1^2\sigma_1^2} &= \frac{1}{2}\sum_i tr\left(\frac{1}{(\sigma_1^2\sigma_2^2 - \rho_i^2)^2}\begin{bmatrix}\sigma_2^2 & -\rho_i \\ -\rho_i & \sigma_1^2\end{bmatrix}\begin{bmatrix}1 & 0 \\ 0 & 0\end{bmatrix}\begin{bmatrix}\sigma_2^2 & -\rho_i \\ -\rho_i & \sigma_1^2\end{bmatrix}\begin{bmatrix}1 & 0 \\ 0 & 0\end{bmatrix}\right) \\
&= \frac{1}{2}tr\left(\frac{1}{(\sigma_1^2\sigma_2^2 - \rho_i^2)^2}\begin{bmatrix}\sigma_2^2 & 0 \\ -\rho_i & 0\end{bmatrix}^2\right) \\
&= \frac{1}{2}\frac{\sigma_2^4}{(\sigma_1^2\sigma_2^2 - \rho_i^2)^2}
\end{aligned}$$

$$
\begin{aligned}
\mathcal{I}(\theta)_{\sigma_2^2\sigma_2^2} &= \frac{1}{2}\sum_i tr\left(\frac{1}{(\sigma_1^2\sigma_2^2-\rho_i^2)^2}\begin{bmatrix}\sigma_2^2 & -\rho_i\\ -\rho_i & \sigma_1^2\end{bmatrix}\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}\sigma_2^2 & -\rho_i\\ -\rho_i & \sigma_1^2\end{bmatrix}\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}\right)\\
&= \frac{1}{2}tr\left(\frac{1}{(\sigma_1^2\sigma_2^2-\rho_i^2)^2}\begin{bmatrix}0 & -\rho_i\\ 0 & \sigma_1^2\end{bmatrix}^2\right)\\
&= \frac{1}{2}\frac{\sigma_1^4}{(\sigma_1^2\sigma_2^2-\rho_i^2)^2}
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{I}(\theta)_{\sigma_1^2\sigma_2^2} &= \frac{1}{2}\sum_i tr\left(\frac{1}{(\sigma_1^2\sigma_2^2-\rho_i^2)^2}\begin{bmatrix}\sigma_2^2 & -\rho_i\\ -\rho_i & \sigma_1^2\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\sigma_2^2 & -\rho_i\\ -\rho_i & \sigma_1^2\end{bmatrix}\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}\right)\\
&= \frac{1}{2}tr\left(\frac{1}{(\sigma_1^2\sigma_2^2-\rho_i^2)^2}\begin{bmatrix}\sigma_2^2 & 0\\ -\rho_i & 0\end{bmatrix}\begin{bmatrix}0 & \rho_i\\ 0 & -\sigma_1^2\end{bmatrix}\right)\\
&= \frac{1}{2}\sum_i\left(\frac{1}{(\sigma_1^2\sigma_2^2-\rho_i^2)^2}\begin{bmatrix}0 & \sigma_2^2\rho_i\\ 0 & -\rho_i^2\end{bmatrix}\right)\\
&= \frac{1}{2}\sum_i\frac{-\rho_i^2}{(\sigma_1^2\sigma_2^2-\rho_i^2)^2}
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{I}(\theta)_{\tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}} &= \frac{1}{2}\sum_i tr\left(\frac{1}{(\sigma_1^2\sigma_2^2-\rho_i^2)^2}\begin{bmatrix}\sigma_2^2 & -\rho_i\\ -\rho_i & \sigma_1^2\end{bmatrix}\begin{bmatrix}0 & \rho_i'\\ \rho_i' & 0\end{bmatrix}\begin{bmatrix}\sigma_2^2 & -\rho_i\\ -\rho_i & \sigma_1^2\end{bmatrix}\begin{bmatrix}0 & \rho_i'\\ \rho_i' & 0\end{bmatrix}\right)\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^T\\
&= \frac{1}{2}\sum_i tr\left(\frac{1}{(\sigma_1^2\sigma_2^2-\rho_i^2)^2}\begin{bmatrix}-\rho_i\rho_i' & \rho_i'\sigma_2^2\\ \sigma_1^2\rho_i' & -\rho_i\rho_i'\end{bmatrix}^2\right)\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^T\\
&= \frac{1}{2}\sum_i tr\left(\frac{1}{(\sigma_1^2\sigma_2^2-\rho_i^2)^2}\begin{bmatrix}\rho_i^2\rho_i'^2 & -2\sigma_2^2\rho_i^2\rho_i'\\ -2\sigma_1^2\rho_i^2\rho_i' & \rho_i^2\rho_i'^2\end{bmatrix}\right)\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^T\\
&= \sum_{i=1}^N \frac{\rho_i^2\rho_i'^2+\sigma_1^2\sigma_2^2\rho_i'^2}{(\sigma_1^2\sigma_2^2-\rho_i^2)^2}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^T
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{I}(\theta)_{\sigma_1^2\tilde{\boldsymbol{\alpha}}} &= \frac{1}{2}\sum_i tr\left(\frac{1}{(\sigma_1^2\sigma_2^2-\rho_i^2)^2}\begin{bmatrix}\sigma_2^2 & -\rho_i\\ -\rho_i & \sigma_1^2\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\sigma_2^2 & -\rho_i\\ -\rho_i & \sigma_1^2\end{bmatrix}\begin{bmatrix}0 & \rho_i'\\ \rho_i' & 0\end{bmatrix}\right)\tilde{\boldsymbol{x}}_i\\
&= -\sum_i \frac{\rho_i\rho_i'\sigma_2^2}{(\sigma_1^2\sigma_2^2-\rho_i^2)^2}\tilde{\boldsymbol{x}}_i
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{I}(\theta)_{\sigma_2^2\tilde{\boldsymbol{\alpha}}} &= \frac{1}{2}\sum_i tr\left(\frac{1}{(\sigma_1^2\sigma_2^2-\rho_i^2)^2}\begin{bmatrix}\sigma_2^2 & -\rho_i\\ -\rho_i & \sigma_1^2\end{bmatrix}\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}\sigma_2^2 & -\rho_i\\ -\rho_i & \sigma_1^2\end{bmatrix}\begin{bmatrix}0 & \rho_i'\\ \rho_i' & 0\end{bmatrix}\right)\tilde{\boldsymbol{x}}_i\\
&= -\sum_i \frac{\rho_i\rho_i'\sigma_1^2}{(\sigma_1^2\sigma_2^2-\rho_i^2)^2}\tilde{\boldsymbol{x}}_i
\end{aligned}
$$

Putting them altogether, $\mathcal{I}(\sigma_1^2,\sigma_2^2,\tilde{\boldsymbol{\alpha}}) =$

$$
\sum_{i=1}^N \frac{1}{(\sigma_1^2\sigma_2^2-\rho_i^2)^2}\begin{bmatrix}\frac{1}{2}\sigma_2^4 & -\frac{1}{2}\rho_i^2 & -\rho_i\rho_i'\sigma_2^2\tilde{\boldsymbol{x}}_i^T\\ -\frac{1}{2}\rho_i^2 & \frac{1}{2}\sigma_1^4 & -\rho_i\rho_i'\sigma_1^2\tilde{\boldsymbol{x}}_i^T\\ -\rho_i\rho_i'\sigma_2^2\tilde{\boldsymbol{x}}_i & -\rho_i\rho_i'\sigma_1^2\tilde{\boldsymbol{x}}_i & (\rho_i^2\rho_i'^2+\sigma_1^2\sigma_2^2\rho_i'^2)\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^T\end{bmatrix}
$$

Now replace all the nuisance parameters as their respective MLES: σ_1^2 to $\hat{\sigma}_1^2$, and so on. $\tilde{\boldsymbol{\alpha}} = (\hat{\alpha}_0, \alpha)$: then the Fisher information matrix is

$$\tilde{\mathcal{I}} = \frac{1}{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2} \begin{bmatrix} \frac{\hat{\sigma}_2^4 N}{2} & -\frac{\hat{\rho}^2 N}{2} & -\hat{\sigma}_2^2\hat{\rho}\hat{\rho}' \sum_i \tilde{\boldsymbol{x}}_i \\ -\frac{\hat{\rho}^2 N}{2} & \frac{\hat{\sigma}_1^4 N}{2} & -\hat{\sigma}_1^2\hat{\rho}\hat{\rho}' \sum_i \tilde{\boldsymbol{x}}_i \\ -\hat{\sigma}_2^2\hat{\rho}\hat{\rho}' \sum_i \tilde{\boldsymbol{x}}_i & -\hat{\sigma}_1^2\hat{\rho}\hat{\rho}' \sum_i \tilde{\boldsymbol{x}}_i & (\hat{\rho}^2\hat{\rho}'^2 + \hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}'^2) \sum_i \tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^T \end{bmatrix}$$

Consider the above matrix as a 2 by 2 block matrix of $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$, where

$A = \frac{1}{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2} \begin{bmatrix} \hat{\sigma}_2^4 N/2 & -\hat{\rho}^2 N/2 \\ -\hat{\rho}^2 N/2 & \hat{\sigma}_1^4 N/2 \end{bmatrix}$, $D = \frac{(\hat{\rho}^2\hat{\rho}'^2 + \hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}'^2) \sum_i \tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^T}{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2}$.

Using Schur's formula, $\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} E & F \\ G & H \end{bmatrix}$ where

$$E = (A - BD^{-1}C)^{-1}$$

$$F = -(A - BD^{-1}C)^{-1}BD^{-1}$$

$$G = -D^{-1}C(A - BD^{-1}C)^{-1}$$

$$H = (D - CA^{-1}B)^{-1}$$

The score statistic is $\hat{d}^T\mathcal{I}\hat{d}$ where $\hat{d} = (\hat{d}_{\sigma_1^2}, \hat{d}_{\sigma_2^2}, \hat{d}_{\tilde{\boldsymbol{\alpha}}}) = (0, 0, \hat{d}_{\tilde{\boldsymbol{\alpha}}})$. Therefore, E, F, G are irrelevant to our score statistic; we only care about the block matrix H. Below is step-by-step derivation.

$$A^{-1} = \begin{bmatrix} \frac{\hat{\sigma}_2^4 N}{2(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2} & -\frac{\hat{\rho}^2 N}{2(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2} \\ -\frac{\hat{\rho}^2 N}{2(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2} & \frac{\hat{\sigma}_1^4 N}{2(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2} \end{bmatrix}^{-1}$$

$$= \frac{4(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^4}{N^2(\hat{\sigma}_1^4\hat{\sigma}_2^4 - \hat{\rho}^4)} \begin{bmatrix} \frac{\hat{\sigma}_1^4 N}{2(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2} & \frac{\hat{\rho}^2 N}{2(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2} \\ \frac{\hat{\rho}^2 N}{2(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2} & \frac{\hat{\sigma}_2^4 N}{2(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2} \end{bmatrix} = \frac{2(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)}{N(\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}^2)} \begin{bmatrix} \hat{\sigma}_1^4 & \hat{\rho}^2 \\ \hat{\rho}^2 & \hat{\sigma}_2^4 \end{bmatrix}$$

$$CA^{-1}B = \frac{1}{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^4} \begin{bmatrix} -\hat{\sigma}_2^2\hat{\rho}\hat{\rho}' & -\hat{\sigma}_1^2\hat{\rho}\hat{\rho}' \end{bmatrix} A^{-1} \begin{bmatrix} -\hat{\sigma}_2^2\hat{\rho}\hat{\rho}' \\ -\hat{\sigma}_1^2\hat{\rho}\hat{\rho}' \end{bmatrix} (\sum_i \tilde{\boldsymbol{x}}_i)(\sum_i \tilde{\boldsymbol{x}}_i^T)$$

$$= \frac{1}{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^4} \frac{2(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)}{N(\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}^2)} \begin{bmatrix} -\hat{\sigma}_2^2\hat{\rho}\hat{\rho}' & -\hat{\sigma}_1^2\hat{\rho}\hat{\rho}' \end{bmatrix} \begin{bmatrix} \hat{\sigma}_1^4 & \hat{\rho}^2 \\ \hat{\rho}^2 & \hat{\sigma}_2^4 \end{bmatrix} \begin{bmatrix} -\hat{\sigma}_2^2\hat{\rho}\hat{\rho}' \\ -\hat{\sigma}_1^2\hat{\rho}\hat{\rho}' \end{bmatrix} (\sum_i \tilde{\boldsymbol{x}}_i)(\sum_i \tilde{\boldsymbol{x}}_i^T)$$

$$= \frac{2(\sum_i \tilde{\boldsymbol{x}}_i)(\sum_i \tilde{\boldsymbol{x}}_i^T)}{N(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^3(\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}^2)} (\hat{\sigma}_1^4\hat{\sigma}_2^4\hat{\rho}^2\hat{\rho}'^2 + \hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}^4\hat{\rho}'^2 + \hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}^4\hat{\rho}'^2 + \hat{\sigma}_1^4\hat{\sigma}_2^4\hat{\rho}^2\hat{\rho}'^2)$$

$$= \frac{2(\sum_i \tilde{\boldsymbol{x}}_i)(\sum_i \tilde{\boldsymbol{x}}_i^T)}{N(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^3(\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}^2)} 2\hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}^2\hat{\rho}'^2(\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}^2)$$

$$= \frac{4\hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}^2\hat{\rho}'^2(\sum_i \tilde{\boldsymbol{x}}_i)(\sum_i \tilde{\boldsymbol{x}}_i^T)}{N(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^3}$$

Then, $H = (D - CA^{-1}B)^{-1} =$

$$\left(\frac{\hat{\rho}^2\hat{\rho}'^2 + \hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}'^2}{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2} \sum_i \tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^T - \frac{4\hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}^2\hat{\rho}'^2(\sum x_i)(\sum x_i^T)}{N(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^3} \right)^{-1}$$

$$= (\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^3 \left((\hat{\rho}^2\hat{\rho}'^2 + \hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}'^2)(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2) \sum_i \tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^T - \frac{4\hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}^2\hat{\rho}'^2}{N}(\sum \tilde{\boldsymbol{x}}_i)(\sum \tilde{\boldsymbol{x}}_i^T) \right)^{-1}$$

$$= \frac{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^3}{\hat{\rho}'^2} \left((\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}^2)(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2) \sum_i \tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^T - \frac{4\hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}^2}{N}(\sum \tilde{\boldsymbol{x}}_i)(\sum \tilde{\boldsymbol{x}}_i^T) \right)^{-1}$$

Consider centering of the covariates so that $\sum_{i=1}^N \boldsymbol{x}_i = \mathbf{0}$, and replacing the intercept con-

stant from 1 to w to create $\check{X}$. Then, the above Fisher information matrix becomes block diagonal.

$$\sum_i \check{\boldsymbol{x}}_i \check{\boldsymbol{x}}_i^T = \begin{bmatrix} Nw & \mathbf{0} \\ \mathbf{0} & \check{X}^T\check{X} \end{bmatrix}$$

$$\left(\sum \check{\boldsymbol{x}}_i\right)\left(\sum \check{\boldsymbol{x}}_i^T\right) = \begin{bmatrix} Nw & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where w is

$$w = \frac{(\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}^2)(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2) - \frac{4\hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}^2}{N}}{(\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}^2)(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)} \tag{6.1}$$

The Fisher Information matrix is equivalent to

$$\begin{aligned} \hat{I}_{\tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}} &= \frac{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^3}{\hat{\rho}'^2}\left((\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}^2)(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)\sum_i \check{\boldsymbol{x}}_i\check{\boldsymbol{x}}_i\right)^{-1} \\ &= \frac{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}^2)^2}{(\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}^2)\hat{\rho}'^2}(\check{X}^T\check{X})^{-1} \end{aligned}$$

With the modified covariates $\check{X}$ the score statistic can be re-written as

$$q_{12} = (\sum_i \check{\boldsymbol{x}}_i f_{12,i})(\sum_i \check{\boldsymbol{x}}_i\check{\boldsymbol{x}}_i^T)^{-1}(\sum_i \check{\boldsymbol{x}}_i f_{12,i}) \tag{6.2}$$

where

$$f_{12,i} = \frac{(\hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}_{12} - \hat{\rho}_{12}^3) + (\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}_{12}^2)\hat{u}_{i1}\hat{u}_{i2} - \hat{\sigma}_1^2\hat{\rho}_{12}\hat{u}_{i2}^2 - \hat{\sigma}_2^2\hat{\rho}_{12}\hat{u}_{i1}^2}{\sqrt{(\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}_{12}^2)(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}_{12}^2)^2}}$$

Alternatively, collect all N observations by defining $\tilde{X} = \begin{bmatrix} \check{\boldsymbol{x}}_1 & \cdots & \check{\boldsymbol{x}}_N \end{bmatrix}^T \in \mathbb{R}^{N\times P}$, $f_{12} = \begin{bmatrix} f_{12,1} & \cdots & f_{12,N} \end{bmatrix}^T$, and write

$$q_{12} = f_{12}^T\check{X}(\check{X}^T\check{X})^{-1}\check{X}^T f_{12} \tag{6.3}$$

Following proof from Bruesch and Pagan, construct length N vector g_{12} where $g_{12,i} = f_{12,i} + 1$. Then, q_{12} is equal to the explained sum of squares from OLS regression that regresses X from g. Note

$$\sum_{i=1}^{N} f_{12,i} = \frac{N(\hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}_{12} - \hat{\rho}_{12}^3) + (\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}_{12}^2)N\hat{\rho}_{12} - \hat{\sigma}_1^2\hat{\rho}_{12}N\hat{\sigma}_2^2 - \hat{\sigma}_2^2\hat{\rho}_{12}N\hat{\sigma}_1^2}{\sqrt{(\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}_{12}^2)(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}_{12}^2)^2}} = 0$$

And therefore, $\sum g_i = \sum f_i + 1 = N$ and $\bar{g} = 1$.

$$\begin{aligned}
ESS &= \sum_{i=1}^{N} (\hat{g}_i - \bar{g})^2 \\
&= \sum_i \left(\check{X}(\check{X}^T\check{X})^{-1}\check{X}^T g_i - \bar{g} \right)^2 \\
&= \sum \left(\check{X}(\check{X}^T\check{X})^{-1}\check{X}^T g_i - \mathbf{1} \right)^2 \\
&= \sum_i g_i^T \check{X}(\check{X}^T\check{X})^{-1}\check{X}^T g_i - 2\mathbf{1}^T\check{X}(\check{X}^T\check{X})^{-1}\check{X}^T g_i + \mathbf{1}^T\check{X}(\check{X}^T\check{X})^{-1}\check{X}^T\mathbf{1} \\
&= (f_i + \mathbf{1})^T\check{X}(\check{X}^T\check{X})^{-1}\check{X}^T(f_i + \mathbf{1}) - 2\mathbf{1}^T\check{X}(\check{X}^T\check{X})^{-1}\check{X}^T(f_i + \mathbf{1}) + \mathbf{1}^T\check{X}(\check{X}^T\check{X})^{-1}\check{X}^T\mathbf{1} \\
&= \sum_i f_i^T\check{X}(\check{X}^T\check{X})^{-1}\check{X}^T f_i \\
&= q_{12}
\end{aligned}$$

Consider the regression g onto X and OLS coefficient estimator $\hat{\delta}$. Then, according to Amemiya (1977),

$$\frac{1}{\sqrt{N}}(\hat{\delta} - \delta) \xrightarrow{d} N\left(0, \lim_{N\to\infty} \frac{1}{N}(\check{X}^T\check{X})^{-1}\tau\right)$$

where

$$\tau = Var\left(\frac{(\hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}_{12} - \hat{\rho}_{12}^3) + (\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}_{12}^2)u_{i1}u_{i2} - \hat{\sigma}_1^2\hat{\rho}_{12}u_{i2}^2 - \hat{\sigma}_2^2\hat{\rho}_{12}u_{i1}^2}{\sqrt{(\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}_{12}^2)(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}_{12}^2)^2}}\right)$$
$$= E\left(\left(\frac{(\hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}_{12} - \hat{\rho}_{12}^3) + (\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}_{12}^2)u_{i1}u_{i2} - \hat{\sigma}_1^2\hat{\rho}_{12}u_{i2}^2 - \hat{\sigma}_2^2\hat{\rho}_{12}u_{i1}^2}{\sqrt{(\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}_{12}^2)(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}_{12}^2)^2}}\right)^2\right)$$

Using higher moments of multivariate normal: $E((u_{i1}^2u_{i2})^2) = \sigma_1^2\sigma_2^2 + \rho_{12}^2$, $E(u_{i1}^4) = 3\sigma_1^4$, $E(u_{i2}^4) = 3\sigma_2^4$, $E(u_{i1}^3u_{i2}) = \sigma_1^2\rho_{12}$, $E(u_{i1}u_{i2}^3) = \sigma_2^2\rho_{12}$,

$$\lim_{N\to\infty}\tau = E\left(\left(\frac{(\sigma_1^2\sigma_2^2\rho_{12} - \rho_{12}^3) + (\sigma_1^2\sigma_2^2 + \rho_{12}^2)u_{i1}u_{i2} - \hat{\sigma}_1^2\hat{\rho}_{12}u_{i2}^2 - \sigma_1^2\rho_{12}u_{i2}^2}{(\sigma_1^2\sigma_2^2 + \rho_{12}^2)(\sigma_1^2\sigma_2^2 - \rho_{12}^2)^2}\right)^2\right)$$
$$= 1$$

Here, we show that if $\hat{\delta}$ follows the normal distribution, then the ESS of the regression g_i onto $\check{X}$ follows χ_P^2. The ESS of the regression can be written as Y^TQLQY where $L = I - \mathbf{1}\mathbf{1}^T/N$ and $Q = \check{X}(\check{X}^T\check{X})^{-1}\check{X}^T$.

$$\begin{aligned}(\hat{g} - \bar{g})^T(\hat{g} - \bar{g}) &= \left(\hat{g} - \frac{1}{N}\mathbf{1}\mathbf{1}^Tg\right)^T\left(\hat{g} - \frac{1}{N}\mathbf{1}\mathbf{1}^Tg\right)\\ &= \left(\hat{g} - \frac{1}{N}\mathbf{1}\mathbf{1}^T\hat{g}\right)^T\left(\hat{g} - \frac{1}{N}\mathbf{1}\mathbf{1}^T\hat{g}\right)\\ &= \left(Qg - \frac{1}{N}\mathbf{1}\mathbf{1}^TQg\right)^T\left(Qg - \frac{1}{N}\mathbf{1}\mathbf{1}^TQg\right)\\ &= (Qg)^TL^TL(Qg)\\ &= g^TQLQg\end{aligned}$$

We use the following three intermediary results.

- $(Q - L)$ is idempotent because $(I - Q) + (Q - L) + L = I$, and $(I - Q)$ and L are both idempotent.

- $(Q-L)^2 = Q^2 - 2QL + L^2 = Q - 2QL + L$ which is equal to $Q - L$ due to above. Therefore, QL = L, which leads to $QLQLQ = QLLQ = QLQ$.
- rank$(LQ) = tr(LQ) = tr(Q - \mathbf{1}\mathbf{1}^T Q/N) = (P+1) - 1 = P$

Therefore, we can show that q_{12} follows χ^2_P distribution asymptotically.

$$q_{12} = g^T QLQg = (X\hat{\delta})^T L(X\hat{\delta}) \text{ where } X\hat{\delta} \to N(X\delta, Q)$$

and if $\delta = 0$, q_{12} follows the chi-squared distribution with degrees of freedom of rank(LQ) = P.

6.2.2 Appendix B. Small Sample Correction

Although the introduced test statistic q asymptotically follows χ^2_1, the statistic has its error in the order of N^{-1} (Harris, 1985) with finite sample size N, and many Monte Carlo experiments show that the test rejects the null hypothesis less frequently than indicated by its nominal size (Godfrey, 1978; Griffiths & Surekha, 1986; Honda, 1988). In response, Harris (1985) used Edgeworth expansion to obtain the distribution and moment generating function to order N^{-1} of the test statistic (Harris, 1985). Building on this expansion, Honda (1988) and Cribari-Neto & Ferrari (1995) proposed corrections to the critical value or to the test statistic that allows better inference even when the sample size is small while preserving the asymptotic properties.

Honda (1988) provided a closed-form formula to adjust the critical value in the order of $O(N^{-1})$ to correct the type I error of the test. This adjustment, only depending on the covariate, sample size, and the degrees of freedom, but not on the data, is a cubic function with respect to C_γ. C_γ is the critical value at the level of type I error γ, i.e. $P(\chi^2_P \geq C_\gamma) = \gamma$. The size-corrected critical value is a cubic function f of the original critical value C_γ as

follows.

$$f(C_\gamma) = C_\gamma + C_\gamma\left(\frac{A_3 - A_2 + A_1}{12NP}\right) + C_\gamma^2\left(\frac{A_2 - 2A_3}{(12NP(P+2)}\right) + C_\gamma^3\left(\frac{A_3}{12NP(P+2)(P+4)}\right) \tag{6.4}$$

where the scalars A_1, A_2, and A_3 follow the notation of Honda (1988) directly. Breusch & Pagan (1979).

One of the desirable properties of f would be monotonicity, because regardless of sample size, same ordering of the strength of evidence against the null is expected and is interpretable. Below, we show that monotonicity asymptotically holds and is almost always true in practice. The derivative of $f(C_\gamma)$ is

$$f'(C_\gamma) = \frac{A_3}{12NP}\left(\frac{A_3 - A_2 + A_1 + 12NP}{A_3} + \frac{2(A_2 - 2A_3)}{(P+2)A_3}C_\gamma + \frac{3}{(P+2)(P+4)}C_\gamma^2\right)$$

A_3 is strictly positive by definition, and we can solve the above quadratic equation to see in which case the derivative is positive (Cribari-Neto & Ferrari, 1995). In other words, we study whether the following discrimant is complex.

$$\sqrt{\left(\frac{2(A_2 - 2A_3)^2}{(P+2)A_3}\right)^2 - 4 \cdot \frac{3A_3(A_3 - A_2 + A_1)}{(P+2)(P+4)A_3} - 4 \cdot \frac{3 \cdot 12NP}{(P+2)(P+4)}}$$

The first two terms inside the square root are $O(1)$ and the last term is $O(N)$, so we can see that the discriminant becomes complex quickly as N increases.

We aim to adjust the test statistic so that the overall shape of the null distribution is closer to χ^2_P. We assume a large enough sample size for monotonicity of g and define the inverse function of g to propose the new adjusted test statistic $\tilde{q} = f^{-1}(q)$

$$\gamma = P(\chi^2_P \geq C_\gamma) = P(q \geq f(C_\gamma)) = P(f^{-1}(q) \geq C_\gamma).$$

The test statistic q is replaced by solving the following equation

$$q : q - f(C_\gamma) = 0$$

which is guaranteed to be unique by the monotonicity of f. The cubic equation can be solved given the covariate X.

6.2.3 Appendix C. Derivation of the distribution of d_1

Here, we derive the distribution of $d_1 = q_{12} + q_{13} + \cdots q_{1K}$ under the global null hypothesis (4.6).

We first scale and orthogonalize $\check{X}$ so that $\sum \check{x}_{ip}^2 = 1$ and $\sum_i \check{x}_{ip_1}\check{x}_{ip_2} = 0$ for all $p_1 \neq p_2$ in $1, \cdots, P$. This transformation does not affect the error u or the residuals $\hat{u}$ and allows to replace $(\check{X}^T\check{X})^{-1}$ with $\frac{1}{N}I$. Then, the score statistic can be re-written as below.

$$\begin{aligned} q_{12} &= \frac{1}{N}\left(\sum_{i=1}^{N} \check{\boldsymbol{x}}_i^T f_{12,i}\right)^T \left(\sum_{i=1}^{N} \check{\boldsymbol{x}}_i^T f_{12,i}\right) \\ &= \sum_{p=1}^{P}\left(\sum_i \check{x}_{ip} f_{12,i}\right)^2 \\ &= \sum_{p=1}^{P}\left(\sum_i \frac{1}{\sqrt{N}}\frac{1}{W_{12}} V_{12,ip}\right)^2 \end{aligned}$$

where

$$\begin{aligned} W_{12} &= \sqrt{(\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}_{12}^2)(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}_{12}^2)^2} \\ V_{12,ip} &= \check{x}_{ip}\left((\hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}_{12} - \hat{\rho}_{12}^3) + (\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}_{12}^2)\hat{u}_{i1}\hat{u}_{i2} - \hat{\sigma}_1^2\hat{\rho}_{12}\hat{u}_{i2}^2 - \hat{\sigma}_2^2\hat{\rho}_{12}\hat{u}_{i1}^2\right). \end{aligned}$$

Then,

$$\begin{aligned} d_1 &= \sum_{k=2}^{K} q_{1k} \\ &= \sum_{k=2}^{K} \sum_{p=1}^{P} \left(\sum_{i=1}^{N} \frac{1}{\sqrt{N}} \frac{1}{W_{1k}} V_{1k,ip} \right)^2 \\ &= \sum_{p=1}^{P} \sum_{k=2}^{K} r_{1k,p}^2 = \sum_{p=1}^{P} \|\boldsymbol{r}_{1,p}\|^2 \end{aligned}$$

where $\boldsymbol{r}_{1,p} = \begin{bmatrix} r_{12,p} & r_{13,p} & \cdots & r_{1K,p} \end{bmatrix}^T$. Below, we show that $\boldsymbol{r}_{1,p}$ asymptotically follows multivariate normal distribution $\mathcal{N}(\mathbf{0}, H_1)$, where the covariance matrix H_1 has 1 in the diagonals and $\eta_{k\ell}$ for $k \neq \ell$ in non-diagonals. We derive $\eta_{k\ell}$ below in a closed form as well.

First, $r_{1k,p} = \sum_{i=1}^{N} \frac{1}{\sqrt{N}} \frac{1}{W_{1k}} V_{1k,ip}$ asymptotically follows standard normal distribution. W_{1k} converges in probability to a constant. Using the definitions, $\hat{\sigma}_1^2 = \frac{1}{N} \sum_{i=1}^{N} \hat{u}_{i1}^2 \xrightarrow{P} \sigma_1^2$, $\hat{\sigma}_2^2 = \frac{1}{N} \sum_{i=1}^{N} \hat{u}_{i2}^2 \xrightarrow{P} \sigma_2^2$, and $\hat{\rho}_{12} = \frac{1}{N} \sum_{i=1}^{N} \hat{u}_{i1} \hat{u}_{i2} \xrightarrow{P} \rho_{12}$. Continuous Mapping Theorem shows that

$$\frac{1}{W_{12}} \xrightarrow{P} \frac{1}{\sqrt{(\sigma_1^2 \sigma_2^2 + \rho_{12}^2)(\sigma_1^2 \sigma_2^2 - \rho_{12}^2)}}$$

$V_{1k,ip}$ is asymptotically independent across $i = 1, \cdots, N$ so that we can use Central Limit Theorem in $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} V_{1k,ip}$. $V_{1k,ip}$ is a function of the regression residuals $\hat{u}_{i1}$ and $\hat{u}_{i2}$ that are not independent across i in finite sample N. The residual can be written as

$$\hat{u}_1 = (I - \check{X}(\check{X}^T \check{X})^{-1} \check{X}^T) u_1 = (I - \frac{1}{N} \check{X} \check{X}^T) u_1$$

where u_1 is i.i.d normal error vector. The non-diagonal term of the matrix $\left(I - \frac{1}{N} \check{X} \check{X}^T \right)$ is $-\frac{1}{N} \check{\boldsymbol{x}}_i^T \check{\boldsymbol{x}}_j$. Meanwhile, asymptotically, the non-diagonal elements converge in probability to 0 under mild conditions about the covariates and their dimension P, i.e. when P is finite and

fixed, and every element of matrix X is finite and fixed. Therefore, asymptotically, $\hat{u}_1 = u_1$ and $\hat{u}_2 = u_2$, making the residuals asymptotically independent across i.

$$
\begin{aligned}
&E(\lim_{N\to\infty}\sum_{i=1}^{N} V_{1k,ip}) \\
&= E\lim_{N\to\infty}\sum_{i=1}^{N}\check{x}_{ip}\left((\hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}_{1k}-\hat{\rho}_{1k}^3)+(\hat{\sigma}_1^2\hat{\sigma}_2^2+\hat{\rho}_{1k}^2)\hat{u}_{i1}\hat{u}_{i2}-\hat{\sigma}_1^2\hat{\rho}_{1k}\hat{u}_{i2}^2-\hat{\sigma}_2^2\hat{\rho}_{1k}\hat{u}_{i1}^2\right) \\
&= \lim_{N\to\infty}\sum_{i=1}^{N}\check{x}_{ip}E(\rho_{1k}(\sigma_1^2\sigma_k^2-\rho_{1k}^2)+(\sigma_1^2\sigma_k^2+\rho_{1k}^2)\hat{u}_{i1}\hat{u}_{i2}\sigma_1^2\rho_{1k}\hat{u}_{i2}^2-\sigma_k^2\rho_{1k}\hat{u}_{i1}^2) \\
&= \lim_{N\to\infty}\sum_{i=1}^{N}\check{x}_{ip}(\rho_{1k}(\sigma_1^2\sigma_k^2-\rho_{1k}^2)+(\sigma_1^2\sigma_k^2+\rho_{1k}^2)\rho_{1k}-\sigma_1^2\rho_{1k}\sigma_k^2-\sigma_k^2\rho_{1k}\sigma_1^2) \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
&Var(\lim_{N\to\infty}\sum_{i=1}^{N} V_{1k,ip}) \\
&= \sum_{i=1}^{N} E(\lim_{N\to\infty}\check{x}_{ip}^2\left((\hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}_{1k}-\hat{\rho}_{1k}^3)+(\hat{\sigma}_1^2\hat{\sigma}_2^2+\hat{\rho}_{1k}^2)\hat{u}_{i1}\hat{u}_{i2}-\hat{\sigma}_1^2\hat{\rho}_{1k}\hat{u}_{i2}^2-\hat{\sigma}_2^2\hat{\rho}_{1k}\hat{u}_{i1}^2\right)^2) \\
&= N(\rho_{1k}^2(\sigma_1^2\sigma_k^2-\rho_{1k}^2)^2+(\sigma_1^2\sigma_k^2+\rho_{1k}^2)^2(\sigma_1^2\sigma_k^2+2\rho_{1k}^2)+\sigma_1^4\rho_{1k}^2\cdot 3\sigma_k^4+\sigma_k^4\rho_{1k}^2\cdot 3\sigma_1^4 \\
&+ 2\rho_{1k}(\sigma_1^2\sigma_k^2-\rho_{1k}^2)(\sigma_1^2\sigma_k^2+\rho_{1k}^2)\cdot\rho_{1k}-2\rho_{1k}(\sigma_1^2\sigma_k^2-\rho_{1k}^2) \\
&\quad -\sigma_1^2\rho_{1k}\sigma_k^2-2\rho_{1k}(\sigma_1^2\sigma_k^2-\rho_{1k}^2)\sigma_k^2\rho_{1k}\cdot\sigma_1^2-2\sigma_1^2\rho_{1k}(\sigma_1^2\sigma_k^2+\rho_{1k}^2)\cdot 3\sigma_k^2\rho_{1k} \\
&\quad -2\sigma_k^2\rho_{1k}(\sigma_1^2\sigma_k^2+\rho_{1k}^2)\cdot 3\sigma_1^2\rho_{1k}+2\sigma_1^2\sigma_k^2\rho_{1k}^2\cdot(\sigma_1^2\sigma_k^2+2\rho_{1k}^2)) \\
&= N(\sigma_1^2\sigma_k^2+\rho_{1k}^2)(\sigma_1^2\sigma_2^2-\rho_{1k}^2)^2
\end{aligned}
$$

and therefore,

$$
\frac{1}{W_{1k}}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} V_{1k,ip}\right)\to\mathcal{N}(0,1) \tag{6.5}
$$

Next, we derive the asymptotic correlation between $r_{1k,p}$ and $r_{1\ell,p}$ by deriving

$$cov(\lim_{n\to\infty} V_{1k,ip}, \lim_{n\to\infty} V_{1\ell,i})$$

and hence derive the non-diagonal elements of H_1, $\eta_{k,\ell}$. We first derive the intermediary result of $h_{k,\ell}$:

$$\begin{aligned}
h_{k,\ell} &= \lim_{N\to\infty} cov\left(\sum_{i=1}^{N} V_{1k,ip}, \sum_{i=1}^{N} V_{1\ell,ip}\right) \\
&= \lim_{N\to\infty} E(\sum_{i=1}^{N} V_{1k,ip} \sum_{i=1}^{N} V_{1\ell,ip}) \\
&= \lim_{N\to\infty} E(\sum_{i=1}^{N} V_{1k,ip} V_{1\ell,ip}) \quad \textit{(asymptotic independence across } i\textit{)} \\
&= \lim_{N\to\infty} \sum_{i=1}^{N} \check{x}_{ip}^2 E(\Big((\hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}_{12} - \hat{\rho}_{12}^3) + (\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}_{12}^2)\hat{u}_{i1}\hat{u}_{i2} - \hat{\sigma}_1^2\hat{\rho}_{12}\hat{u}_{i2}^2 - \hat{\sigma}_2^2\hat{\rho}_{12}\hat{u}_{i1}^2\Big))\cdot \\
&\qquad \Big((\hat{\sigma}_1^2\hat{\sigma}_2^2\hat{\rho}_{12} - \hat{\rho}_{12}^3) + (\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}_{12}^2)\hat{u}_{i1}\hat{u}_{i2} - \hat{\sigma}_1^2\hat{\rho}_{12}\hat{u}_{i2}^2 - \hat{\sigma}_2^2\hat{\rho}_{12}\hat{u}_{i1}^2\Big) \\
&= N(2\sigma_1^4\sigma_k^2\sigma_\ell^2\rho_{1k}\rho_{1\ell} + 3\sigma_1^2\sigma_\ell^2\rho_{1k}^3\rho_{1\ell} + 3\sigma_1^2\sigma_k^2\rho_{1k}\rho_{1\ell}^3 - 2\sigma_1^4\sigma_k^4\sigma_\ell^2\rho_{1\ell} \\
&\quad - \sigma_1^4\sigma_k^2\rho_{1\ell}^2\rho_{k\ell} - \sigma_1^2\sigma_k^4\rho_{1\ell}^3 + \sigma_1^6\sigma_k^2\sigma_\ell^2\rho_{k\ell} - 2\sigma_1^2\sigma_k^2\sigma_\ell^2\rho_{1k}^2\rho_{1\ell} - \sigma_k^2\rho_{1k}^2\rho_{1\ell}^3 \\
&\quad - \sigma_1^4\sigma_\ell^2\rho_{1k}^2\rho_{k\ell} + 2\rho_{1k}^3\rho_{1\ell}^3 - \sigma_1^4\sigma_k^2\sigma_\ell^2\rho_{1k}^2 - \sigma_1^2\sigma_k^2\rho_{1k}^2\rho_{1\ell}^2 + 2\sigma_1^4\rho_{1k}\rho_{1\ell}\rho_{k\ell}^2)
\end{aligned}$$

This leads to the non-diagonal elements of H_1,

$$\eta_{k,\ell} = \frac{h_{k,\ell}}{\sqrt{(\sigma_1^2\sigma_k^2 + \rho_{1k}^2)(\sigma_1^2\sigma_k^2 - \rho_{1k}^2)^2(\sigma_1^2\sigma_\ell^2 + \rho_{1\ell}^2)(\sigma_1^2\sigma_\ell^2 - \rho_{1\ell}^2)^2}} \tag{6.6}$$

Therefore, we have shown that $\boldsymbol{r}_{1,p} \xrightarrow{D} \mathcal{N}(\mathbf{0}, H_1)$ with closed-form definition of H_1. Finally, we derive the null distribution of $d_1 = \sum_{p=1}^{P} \|\boldsymbol{r}_{1,p}\|_2^2$. Let $H_1 = U_1\Lambda_1 U_1^T$ be the eigen-decomposition of H_1, where Λ is a diagonal matrix with elements $\lambda_{12}, \cdots, \lambda_{1K}$. Due to the

orthogonal invariance of L2 norm, $d_1 = \|\boldsymbol{r}_{1,p}\|^2 = \|U_1\boldsymbol{r}_{1,p}\|^2$. Then,

$$U_1\boldsymbol{r}_{1,p} \sim \mathcal{N}(0, \Lambda_1)$$

Each component of the left hand side is independent with known variance λ_{1k} for $k = 2, \cdots, K$. The sum of the square of normal variables can be written as a sum of independent gamma variables:

$$\sum_{p=1}^{P} r_{1k,p}^2 \sim \Gamma\left(P/2, \lambda_{1k}/2\right)$$

$$d_1 = \sum_{k=2}^{K}\sum_{p=1}^{P} r_{1k,p}^2 \sim \sum_{k=2}^{K} \Gamma\left(\frac{P}{2}, \frac{\lambda_{1k}}{2}\right)$$

This concludes the proof of Proposition 2.

6.3 Supporting Evidence for Chapter 5

This section includes figures and tables that supplement analyses in the main manuscript.

We first study the relationship between zero proportions and gene means in the publicly available, labeled UMI data sets of Zheng2017, Azizi2016, Tabula Muris, Tung2017, Baron2016 (Supplementary Figures 6.5, 6.6, 6.7, 6.8, 6.9). We explore common cell types across different datasets to emphasize that the sampling noise affects different data sets and different cell types in the same way. We study the zero proportion and gene mean relationships in Supplementary Figure 6.10 and 6.11 for data generated from CEL-seq and Drop-seq. In Drop-seq, the noise level was too high to assume the zero proportions follow the exponential curve relative to the gene mean. It is either that Drop-seq data sets have different noise structure from the 10X data sets, or in particular Macosko data of muscular retina cells have excessively high cellular heterogeneity. In Zhang2017 data, cells from different disease types show a layered pattern of zero inflation. This means that the Poisson noise model cannot separate the technical effect from the biological confounder of disease type.

Next, we show why zero proportion is a better test statistic compared to gene variance and dispersion for cell-type heterogeneity using 4 data sets of Freytag2018, Zhengmix8eq, Tian2018, and Azizi2018.

We provide the details of the analysis that show immune-related genes are more zero-inflated than others. Supplementary Table 6.3 shows the number of genes present within each functional annotation for each data set. Supplementary Table 6.4 shows the result of enrichment analysis for AziziPatient09Rep01 data set to demonstrate that zero-inflated genes are particularly enriched in immune-related genes.

Then the feature selection test statistics are compraed between HIPPO and scry package

that uses deviance statistic. We in particular consider two scenarios with data sets with high UMI counts and low UMI counts, and show that in which case each method is more appropriate (Supplementary Figures 6.14, 6.15).

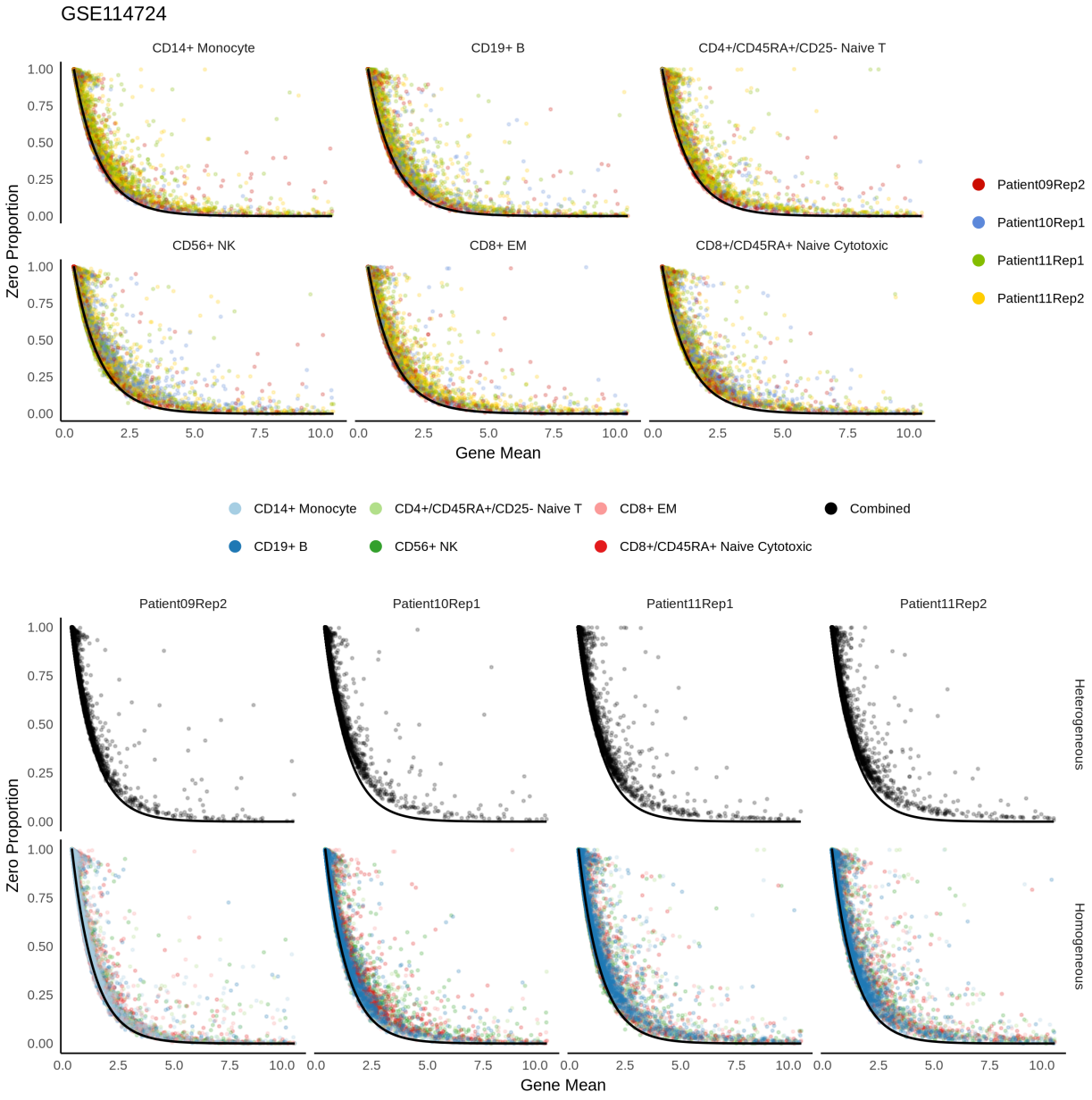
We show additional analyses that pre-processing steps before imputation and normalization lead to adversarial consequences in downstream analyses. Supplementary Figure 6.19 expands the result of Figure 5.2 E by showing the differential expression analysis for known markers after DCA in two cases: on homogeneous cell population and on heterogeneous population. Figure 6.20 shows the similar result for both DCA and SAVER but transcriptome-wide statistics for log fold change, likelihood ratio, and p-values.

The clustering performances are evaluated for more data sets: Tian2018, Zhengmix4eq, Zhengmix4uneq, Zhengmix8eq, PBMC3k1, and PBMC4k1. Supplementary Figure 6.21 shows the adjusted rand index for the available labeled data sets. Supplementary Figure 6.22 shows the sequential visualization of HIPPO's clustering method for all of those data sets. Supplementary Figure 6.23 evaluates the performance of generalized PCA (gPCA) that can account for the count structure directly Lee (2015).

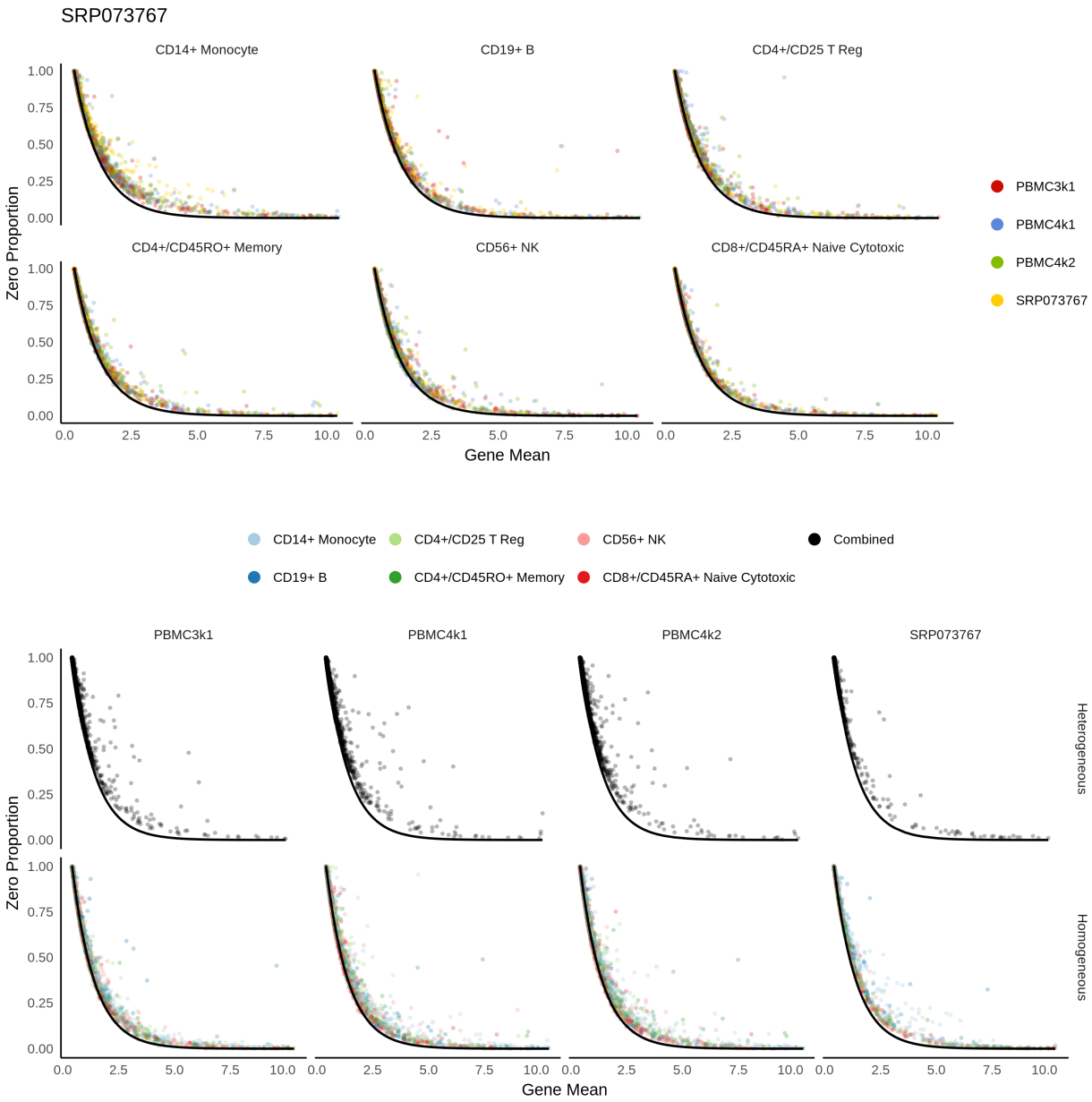
Applications of HIPPO to two different data sets of cells from muscular brain tissue (1k Brain Cells from an E18 Mouse and 5k Cells from a combined cortex, hippocampus and subventricular zone of an E18 mouse) are shown. Supplementary Figure 6.24 shows the clustering performance of HIPPO in two different data of un-labeled cells from brain tissue which are known to have a high level of heterogeneity. Supplementary Figure 6.25 shows an example analysis pipeline implemented in HIPPO. Supplementary Figure 6.26 visualizes the hierarchical structure of the clustering result of HIPPO through an external tool "clustree". Zappia & Oshlack (2018).

Lastly, we discusses in detail about the choice of zero-inflation statistic. It discusses the consequence of the simplification of the null distribution of the test statistic.

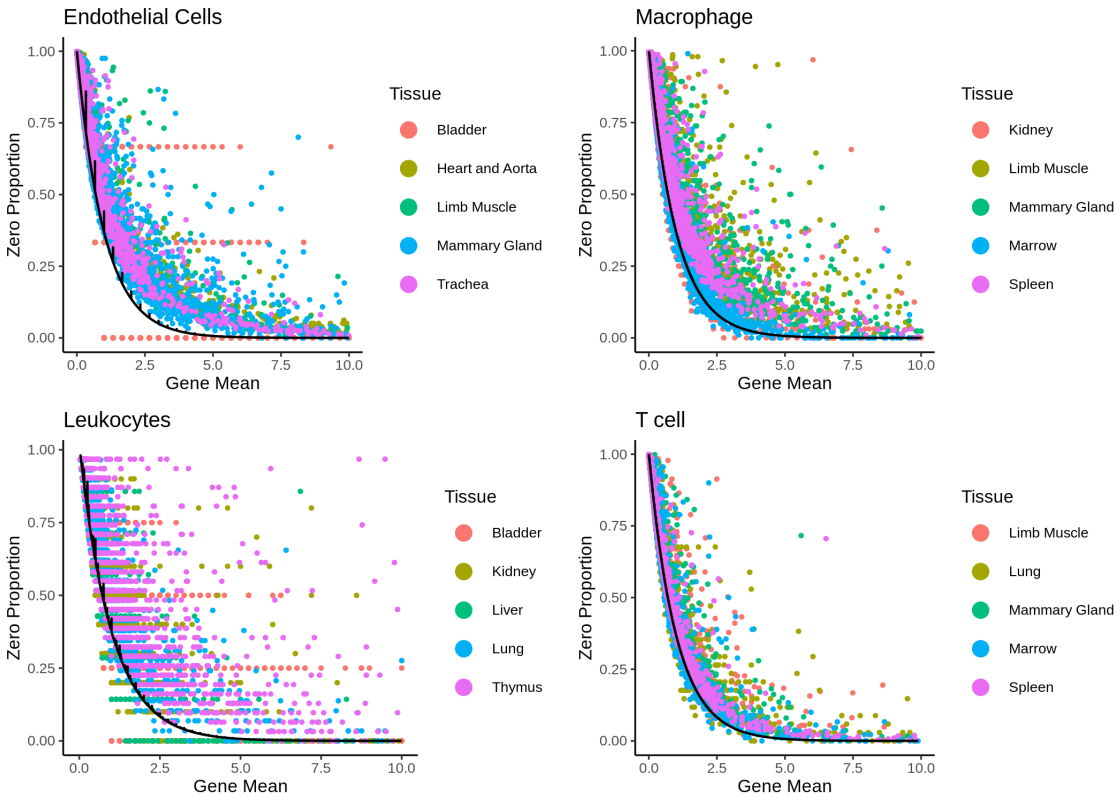
6.3.1 Supplementary Figures



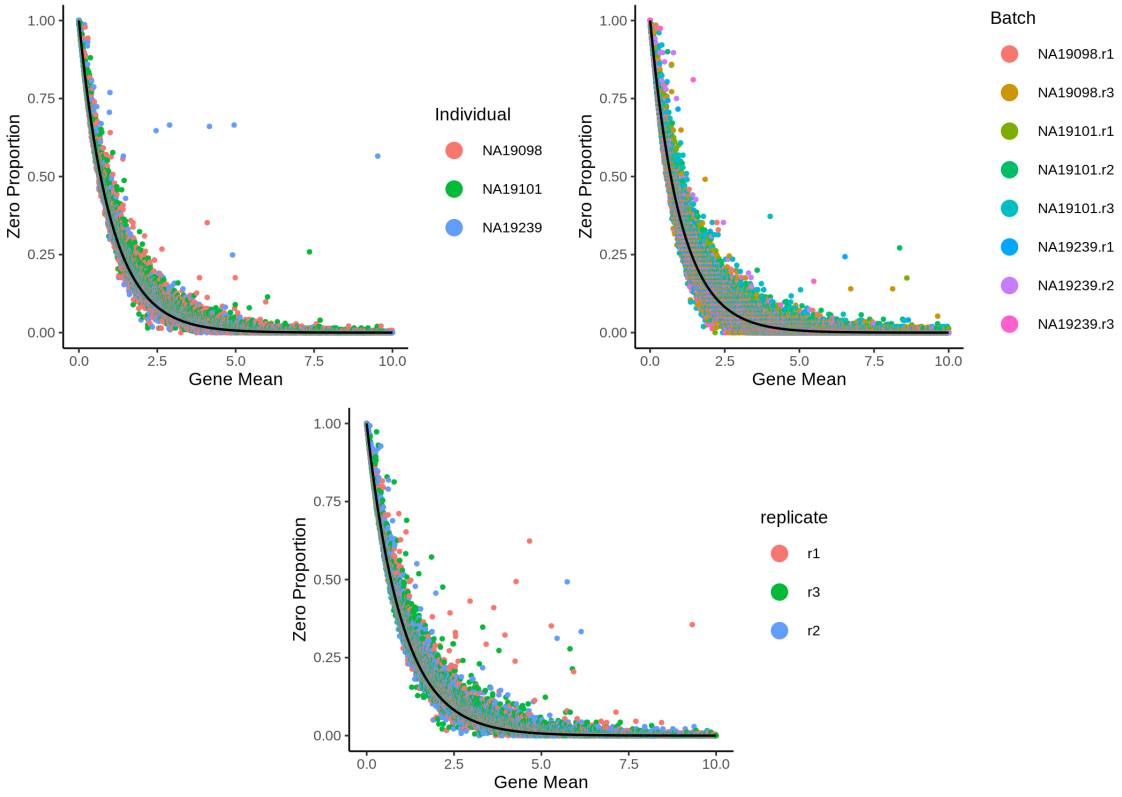
Supplementary Figure 6.5: Zero proportions against gene means for Azizi data Azizi et al. (2018) for multiple samples and replicates. The top plot shows that the zero proportion matches the curve across the data sets for each cell type, while bottom plot across the cell types for each data set. The bottom plot also shows that the zero proportions are off the curve in heterogeneous cell populations. The consistent plots show that the sampling noise is the same across cell types and across data sets.



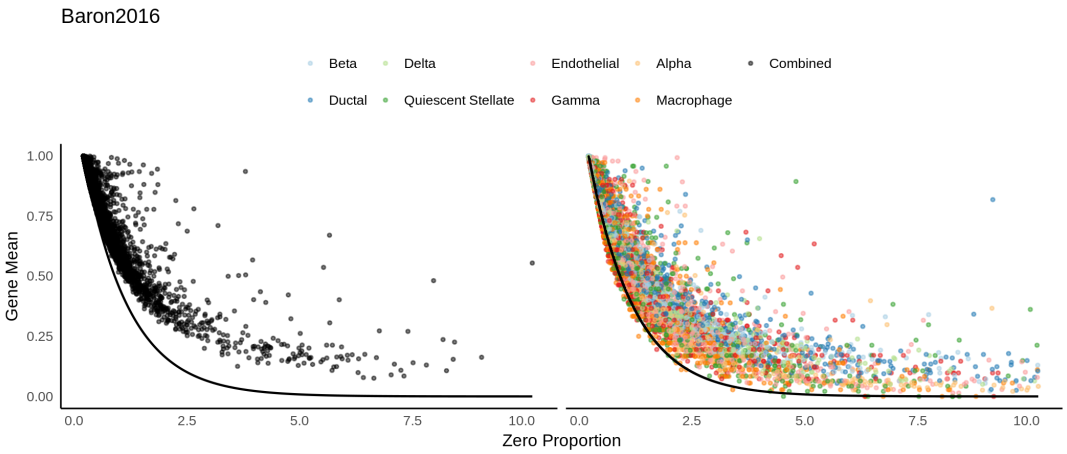
Supplementary Figure 6.6: Same analysis as Supplementary Figure 6.5 with Zheng2017 data Zheng et al. (2017)



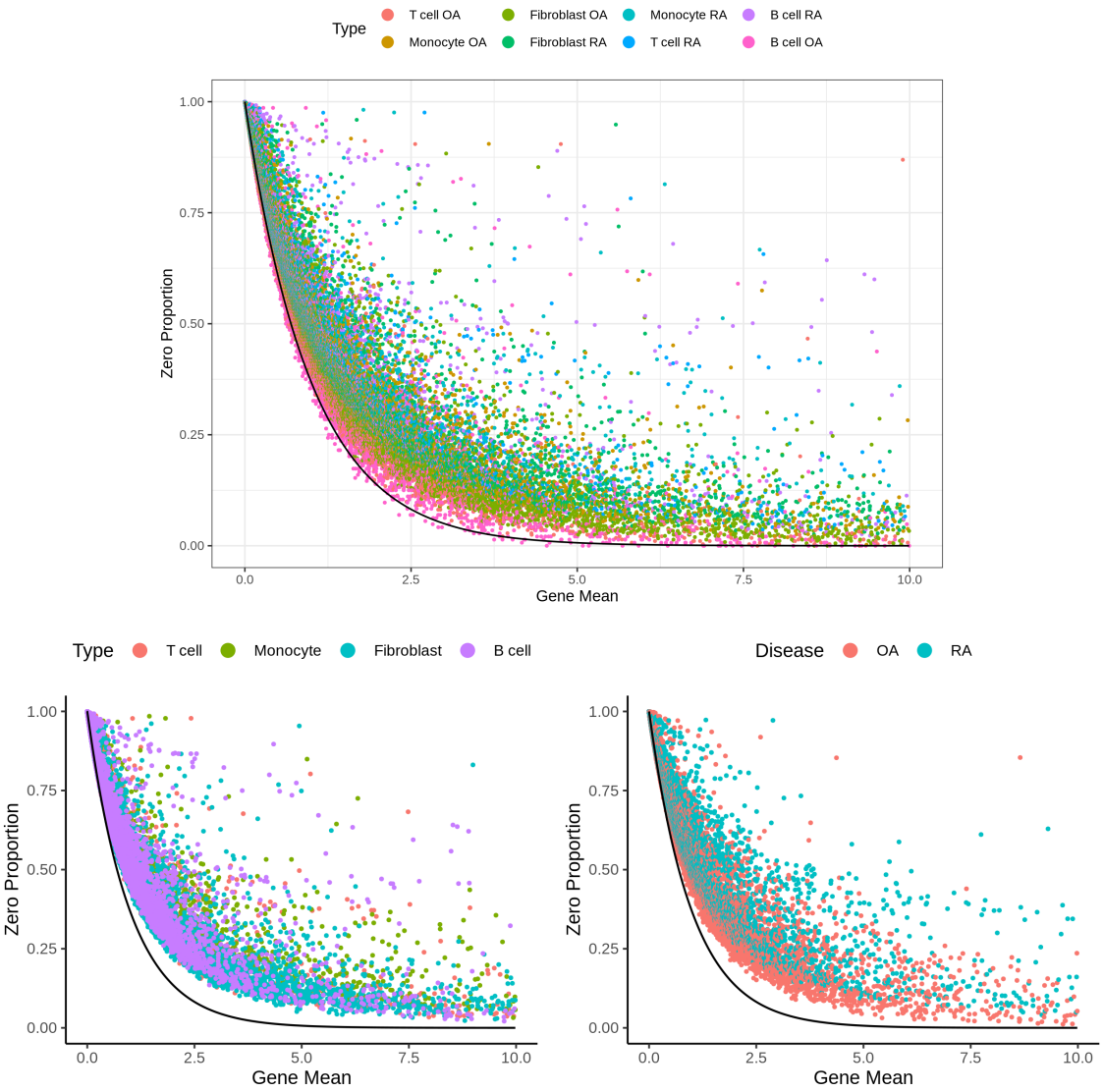
Supplementary Figure 6.7: Same analysis as Supplementary Figure 6.5 with Tabula Muris data Tabula Muris Consortium et al. (2018) . The color codes are not tissue-specific to maximize visibility.



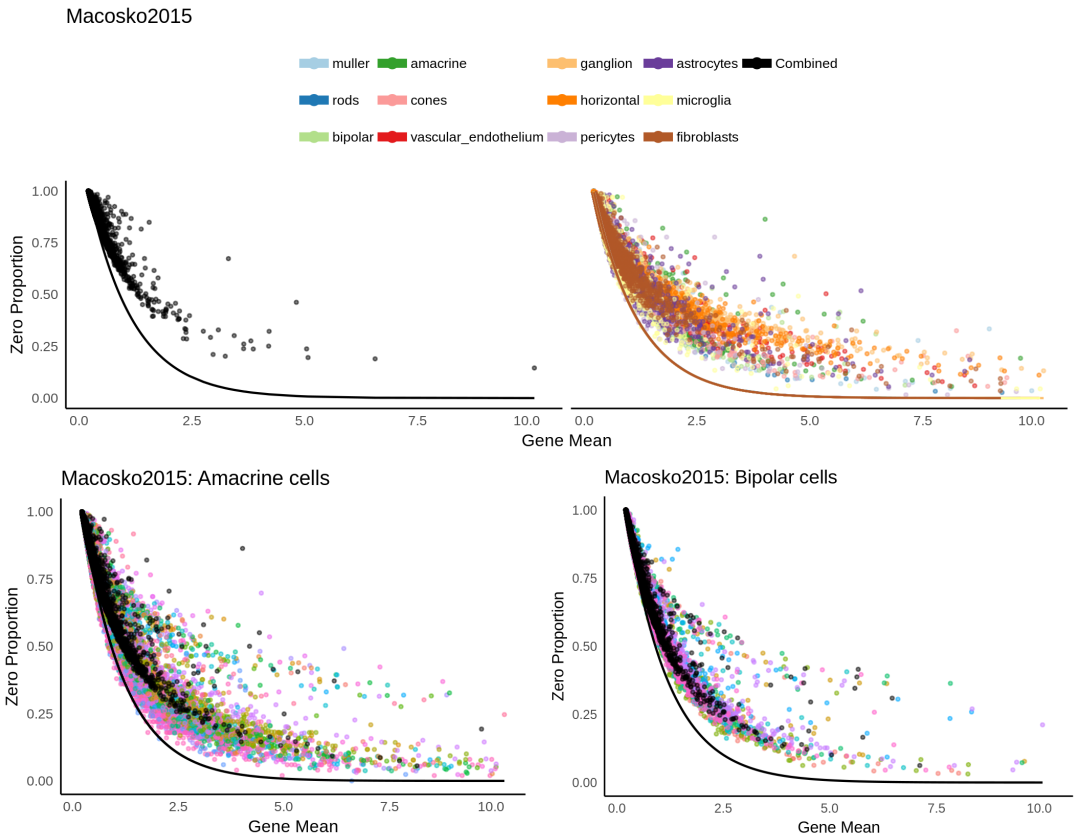
Supplementary Figure 6.8: Results from Tung2017 Tung et al. (2017) that uses Hi-Seq 2500. This data consists of homogeneous cell population of iPSC cell lines from three different individuals.



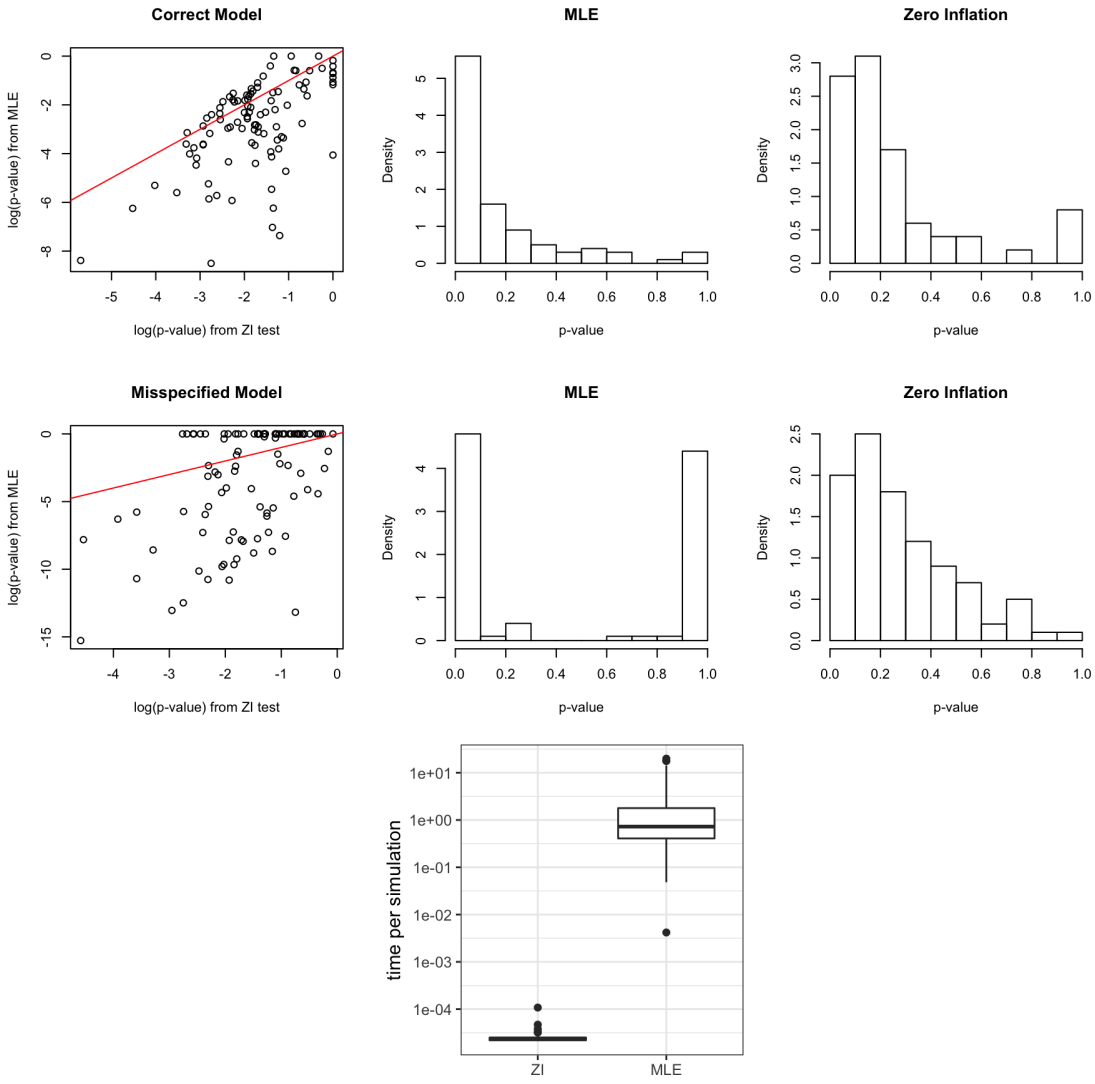
Supplementary Figure 6.9: In-drop is promising that it can be modeled using Poisson.



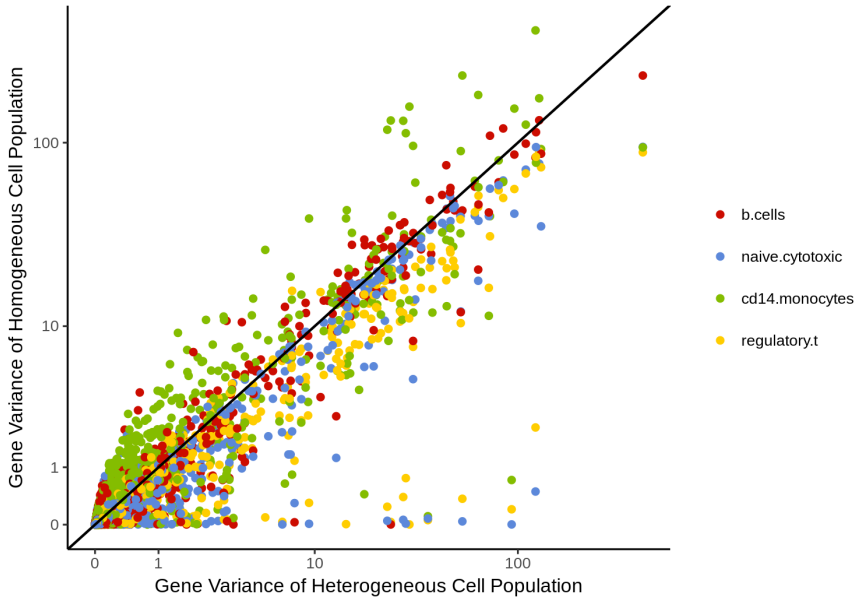
Supplementary Figure 6.10: Results from Zhang2017 Zhang et al. (2019) that uses CEL-SEQ2 2500.



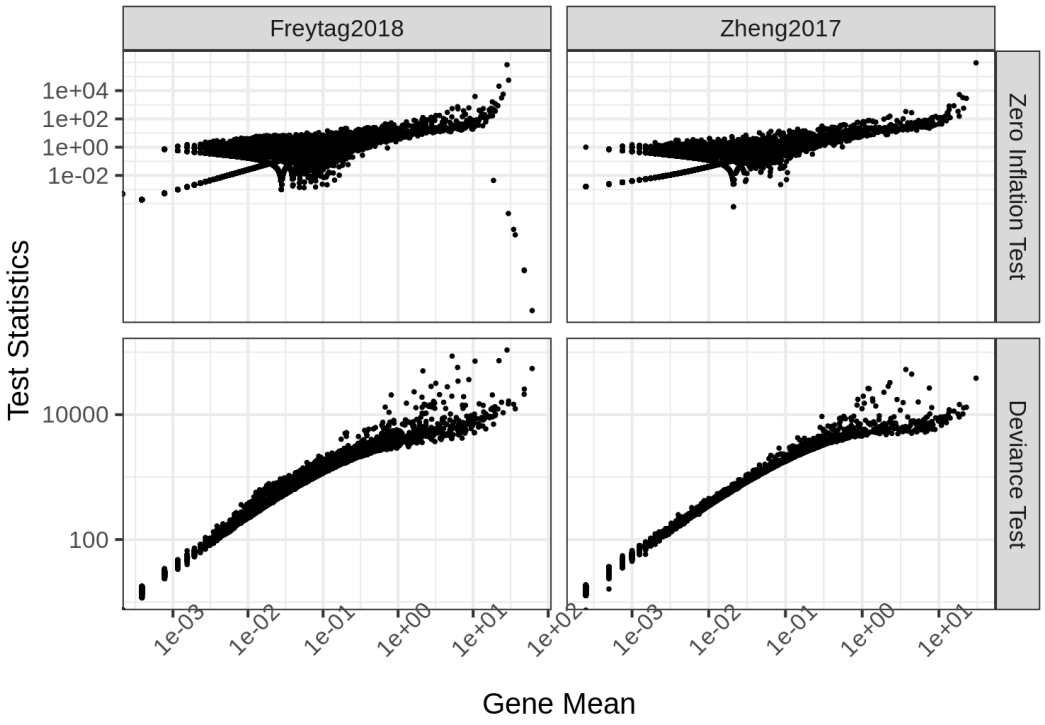
Supplementary Figure 6.11: Results from Drop-seqMacosko et al. (2015). However, the sampling noise of drop-seq data is too high, and zero-inflation element seems necessary. When amacrine cells were taken out and further clustered into subtypes, the noise level is closer to Poisson, so the culprit could be the particularly higher level of cell type diversity. The black points are plotted using heterogeneous cell population.



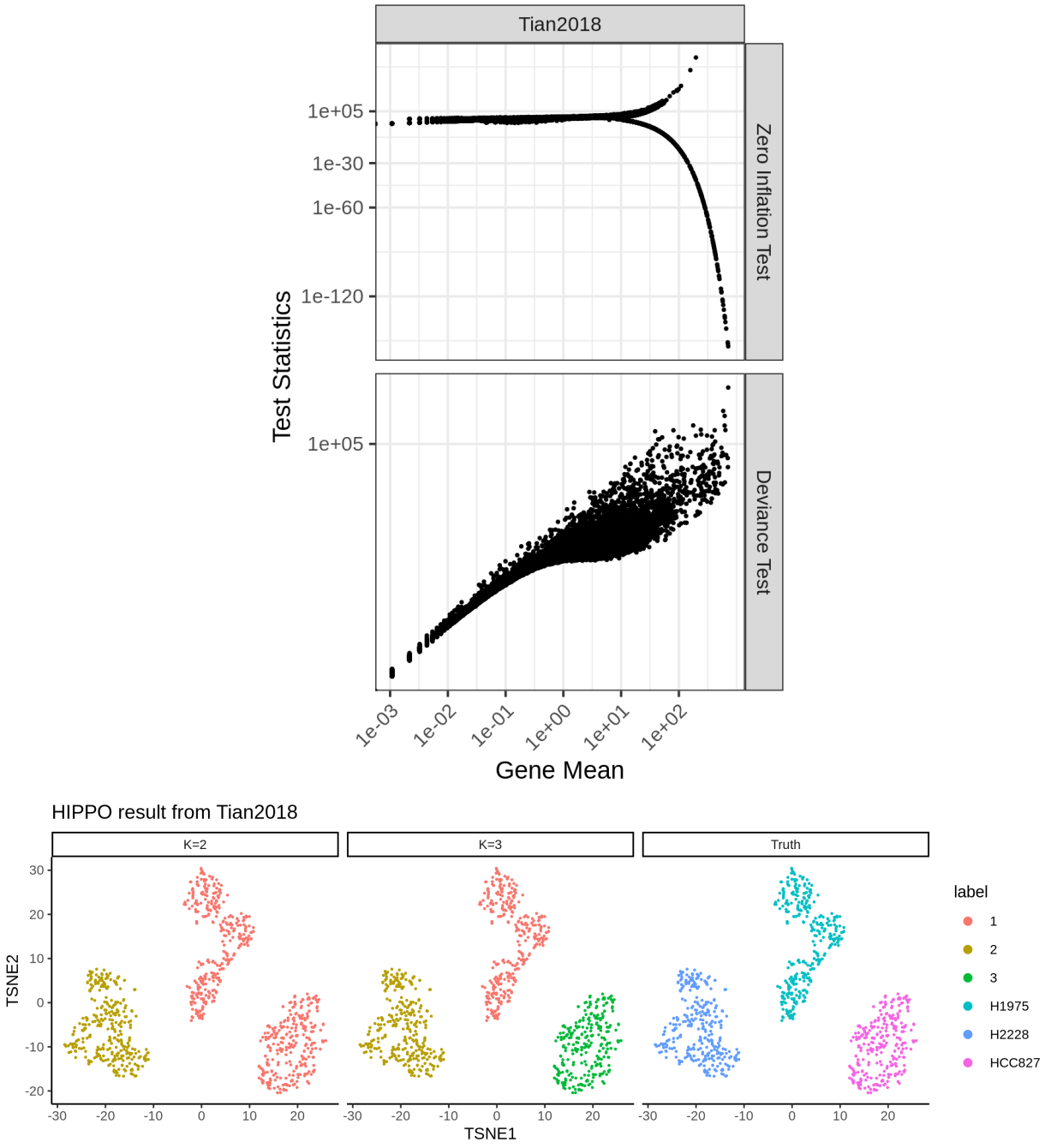
Supplementary Figure 6.12: Comparison of using maximum likelihood estimates of Poisson mixture and using proposed zero inflation test. When data is generated from Negative Binomial, EM algorithm for mixture estimate often breaks down, leading to very unstable result. Moreover, EM algorithm is much more computationally intensive in the order of 10^4 *to* 10^5.



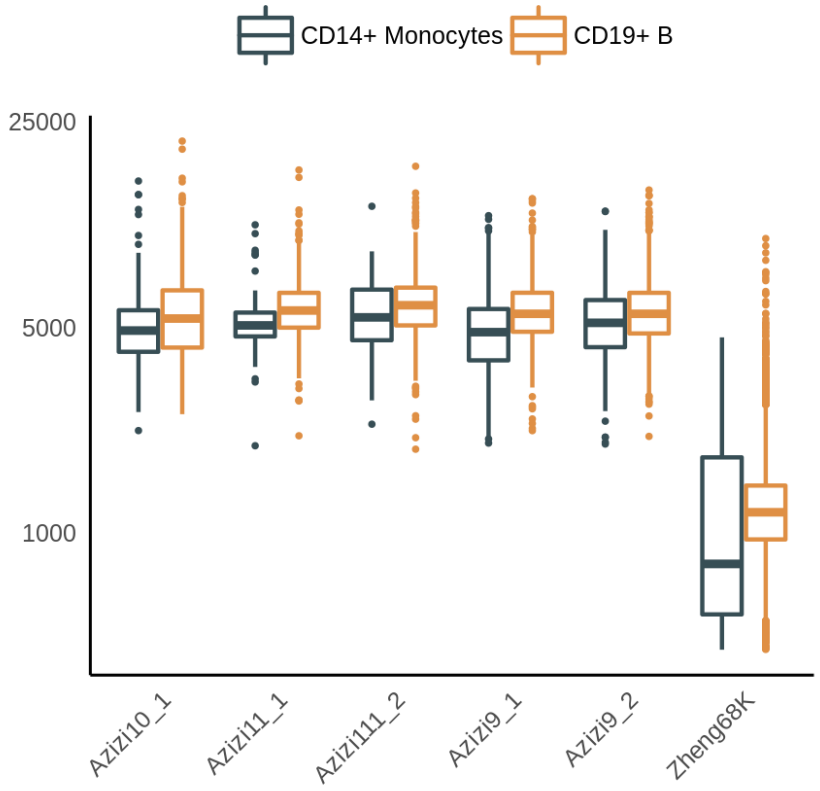
Supplementary Figure 6.13: Gene variance for homogeneous cell population (y axis) and heterogeneous cell population (x-axis). For most genes, gene variance is similar for both heterogeneous and homogeneous cells. Further quantifications are provided in Supplementary Table 6.13.



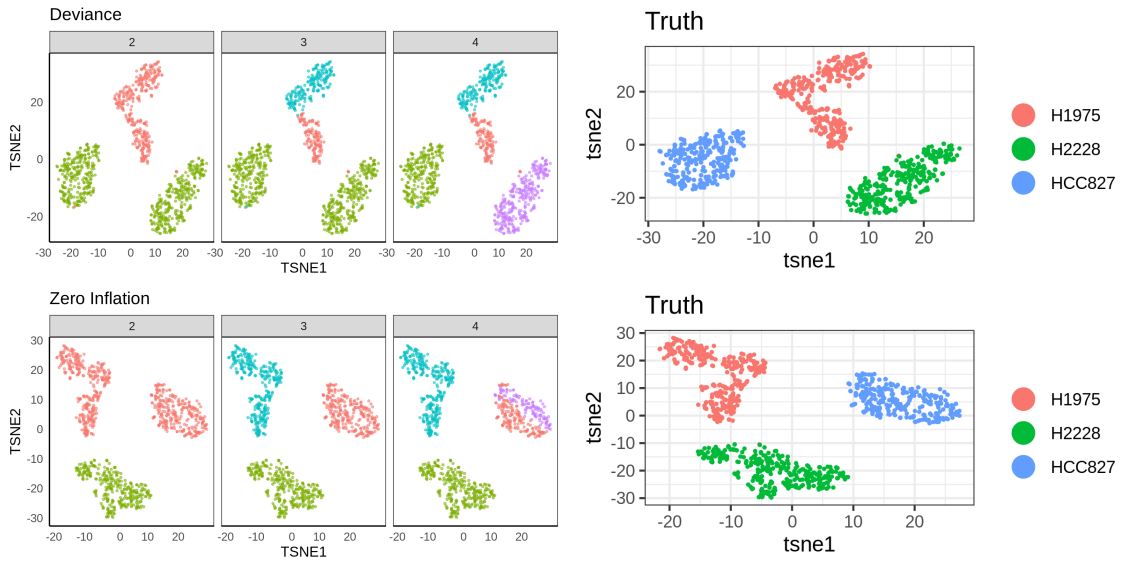
Supplementary Figure 6.14: Comparison of the test statistics between the proposed package HIPPO and package scry Townes & Street (2020). The ordering of the statistic is similar between zero inflation test statistic and deviance statistic, although the zero inflation test does not take account into the entire distribution of the gene counts. There are a few genes that have high deviance but low zero inflation in Freytag data. Those cases occur when there are no zeros recorded across all the cells. Zero inflation test statistic becomes lower as gene mean increases.



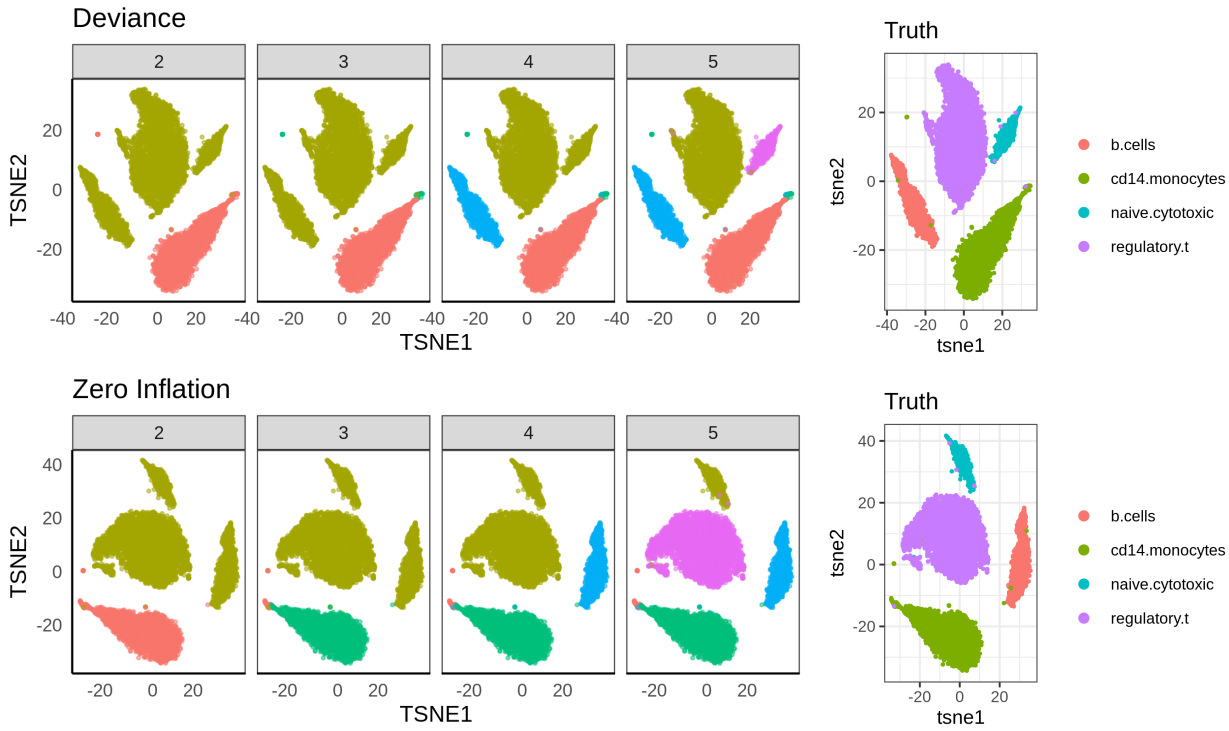
Supplementary Figure 6.15: Same analysis as Supplementary Figure 8 with Tian2018 data which has higher UMI counts. This analysis shows different relationship between zero inflation test and deviance test. When the mean counts become large, the zero inflation test statistic is either extremely low (there are no zeros recorded) or extremely high (there is at least 1 zero recorded). The problem is more severe when there are fewer cells as arguments for the test statistic are asymptotic. However, the HIPPO result shows that the zero-inflated genes still hold rich information for reliable clustering.



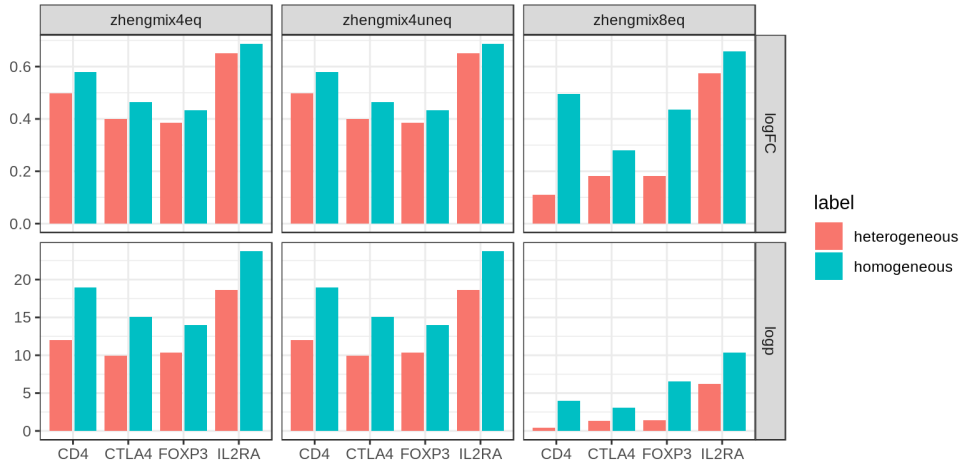
Supplementary Figure 6.16: Sequencing depth of monocytes and B cells. Monocytes have consistently higher total UMI counts than B cells in these particular data sets, and forcing all the cells to have the same sequencing depth (size factor normalization) would either shrink the counts of B cells or inflate the counts of monocytes.



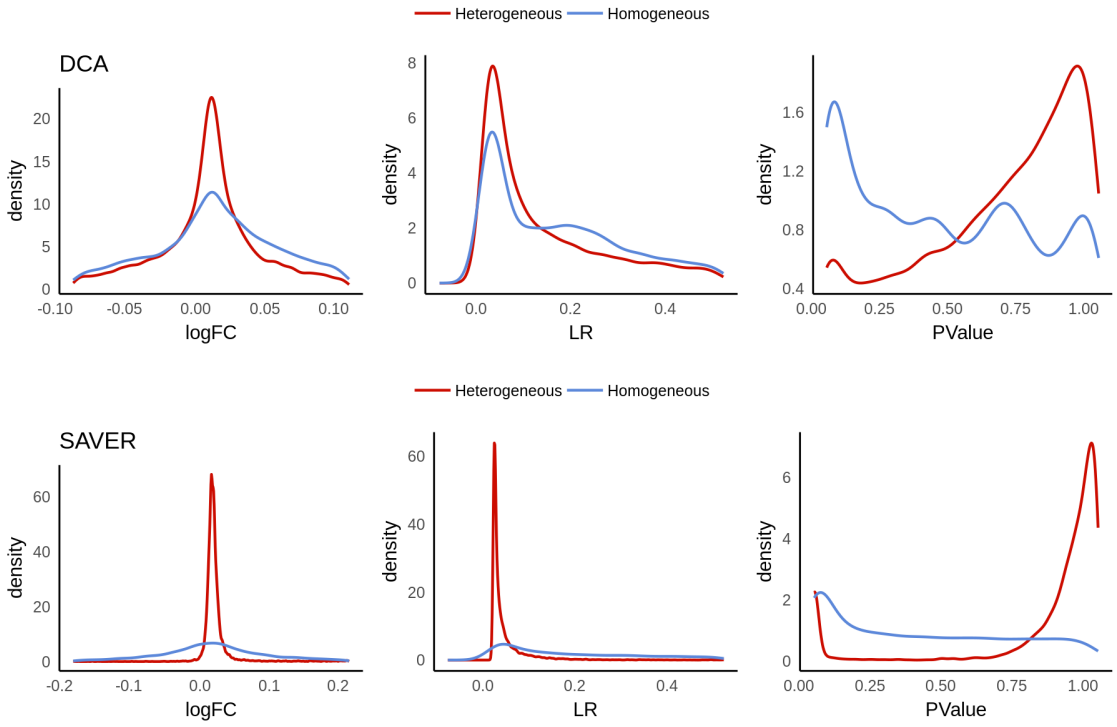
Supplementary Figure 6.17: Clustering results for two feature selection methods - zero inflation and deviance with Tian2018 data that has high UMI counts. The truth labels are shown for both dimension reductions using different sets of features.



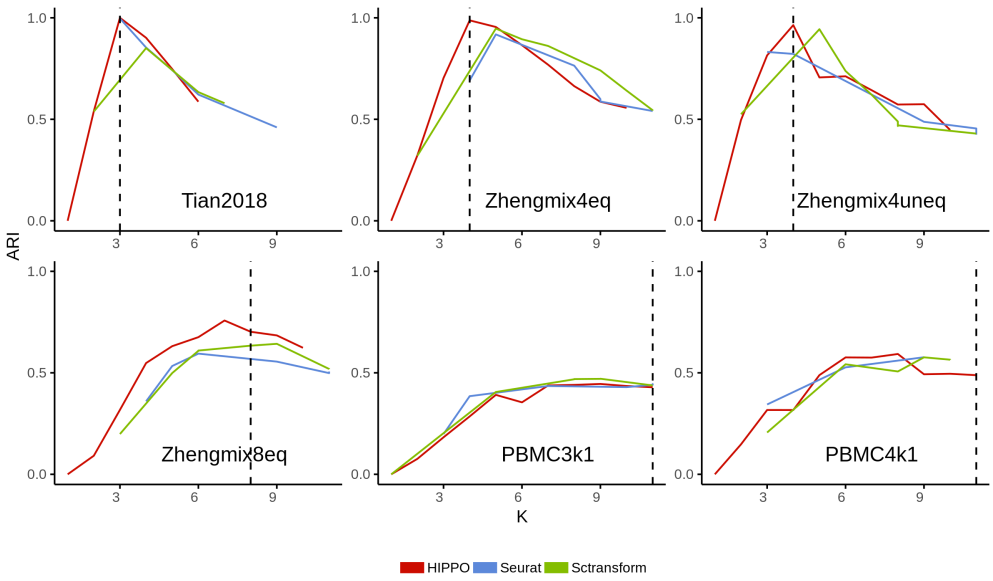
Supplementary Figure 6.18: Clustering results for two feature selection methods in Zhengmix4uneq data that has low UMI counts. The truth labels are shown for both dimension reductions using different sets of features. The performance is very similar using two different methods.



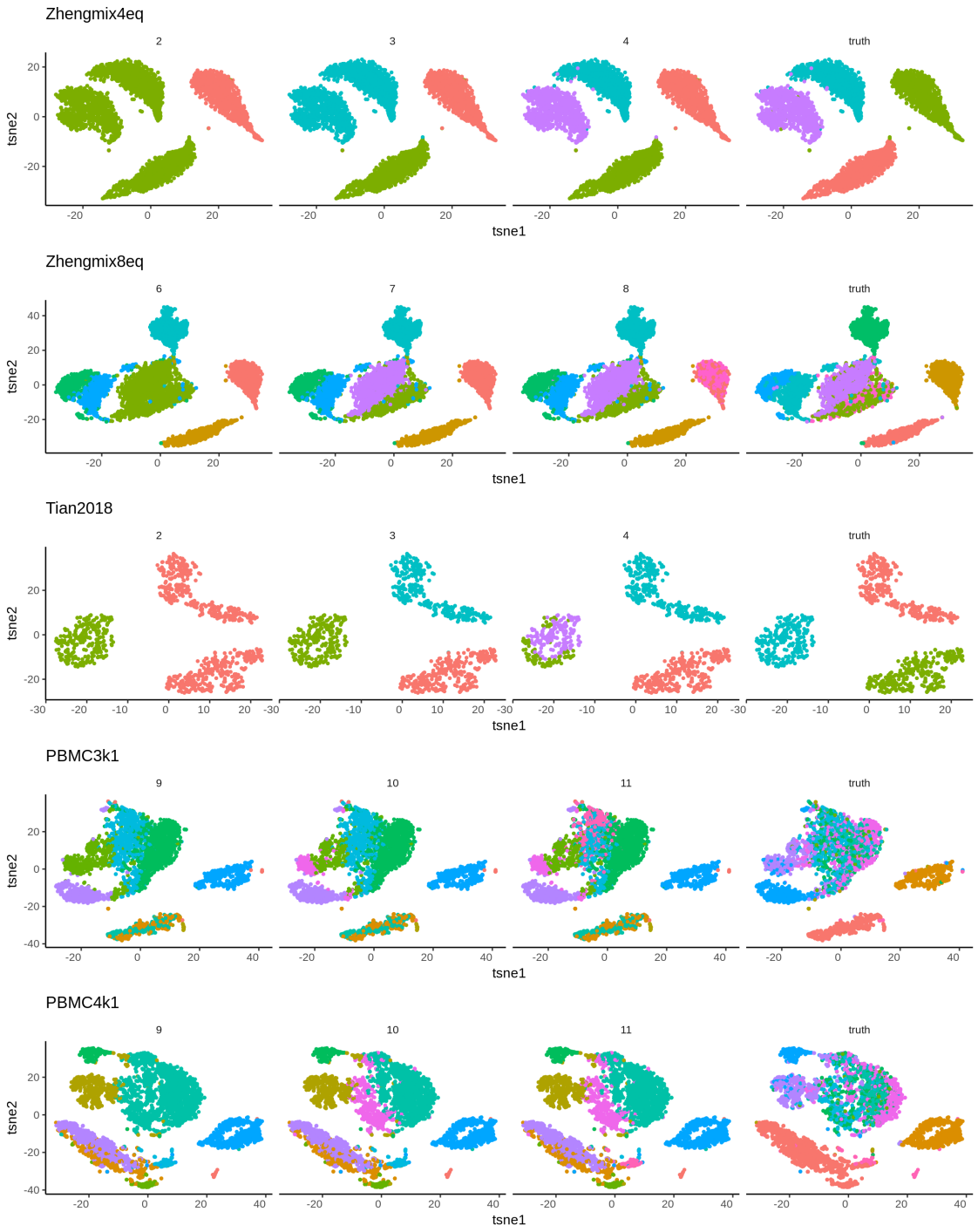
Supplementary Figure 6.19: Extension of Figure 5.2 E in the main text. The log-fold change is consistently lower across data sets if DCA Eraslan et al. (2019) is performed before the cell heterogeneity is accounted for.



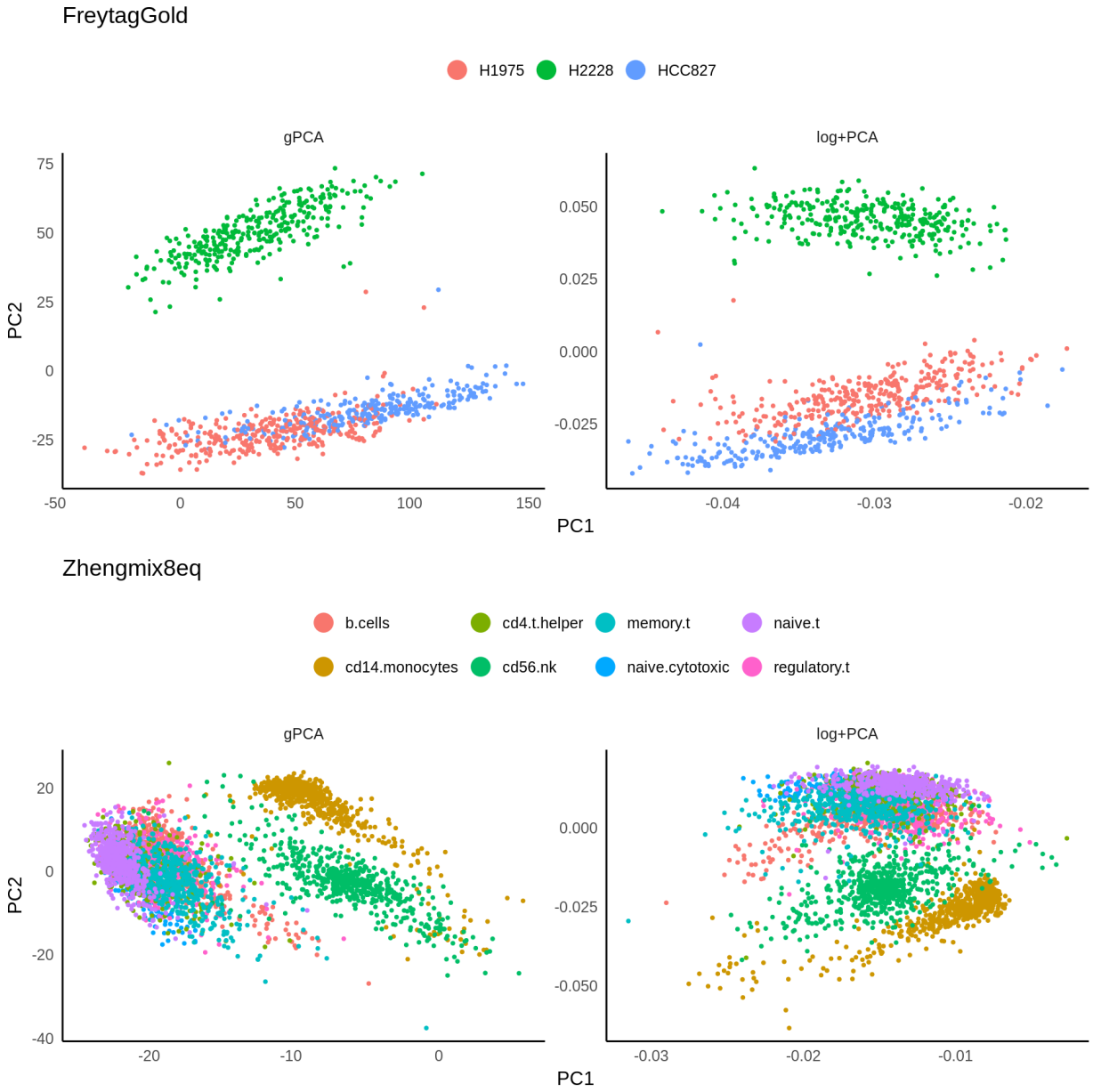
Supplementary Figure 6.20: Extension of Supplementary Figure 6.19. Overall distribution of various statistics (log fold change, likelihood ratio, and p-value) from differential expression test using edgeR's likelihood ratio test Robinson et al. (2010) after DCA Eraslan et al. (2019) and SAVER Huang et al. (2018). Overall signal size is deflated if we perform imputation first.



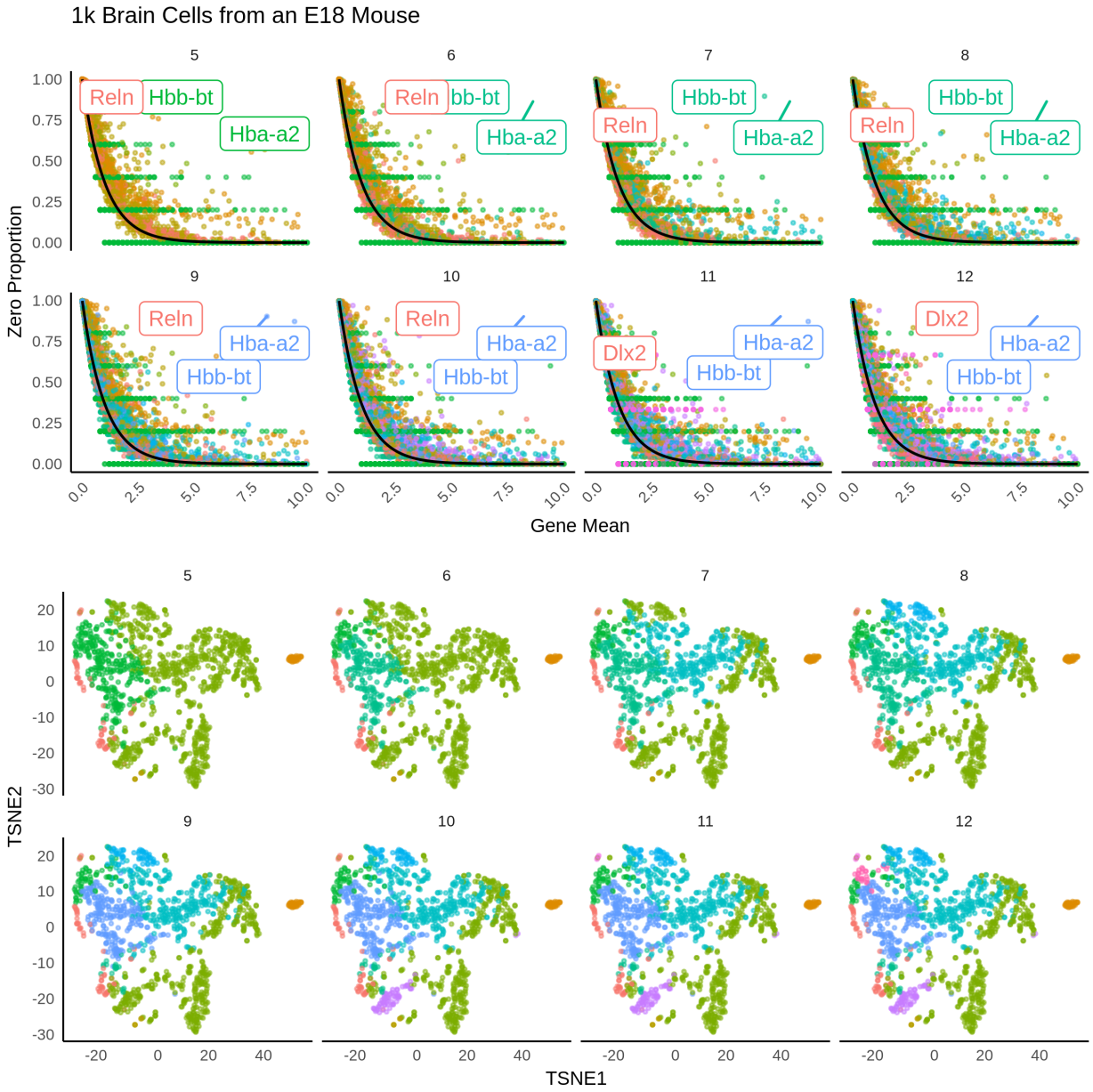
Supplementary Figure 6.21: Adjusted Rand Index for various data sets comparing three methods. HIPPO tends to work at least as well as Sctransform and Seurat.

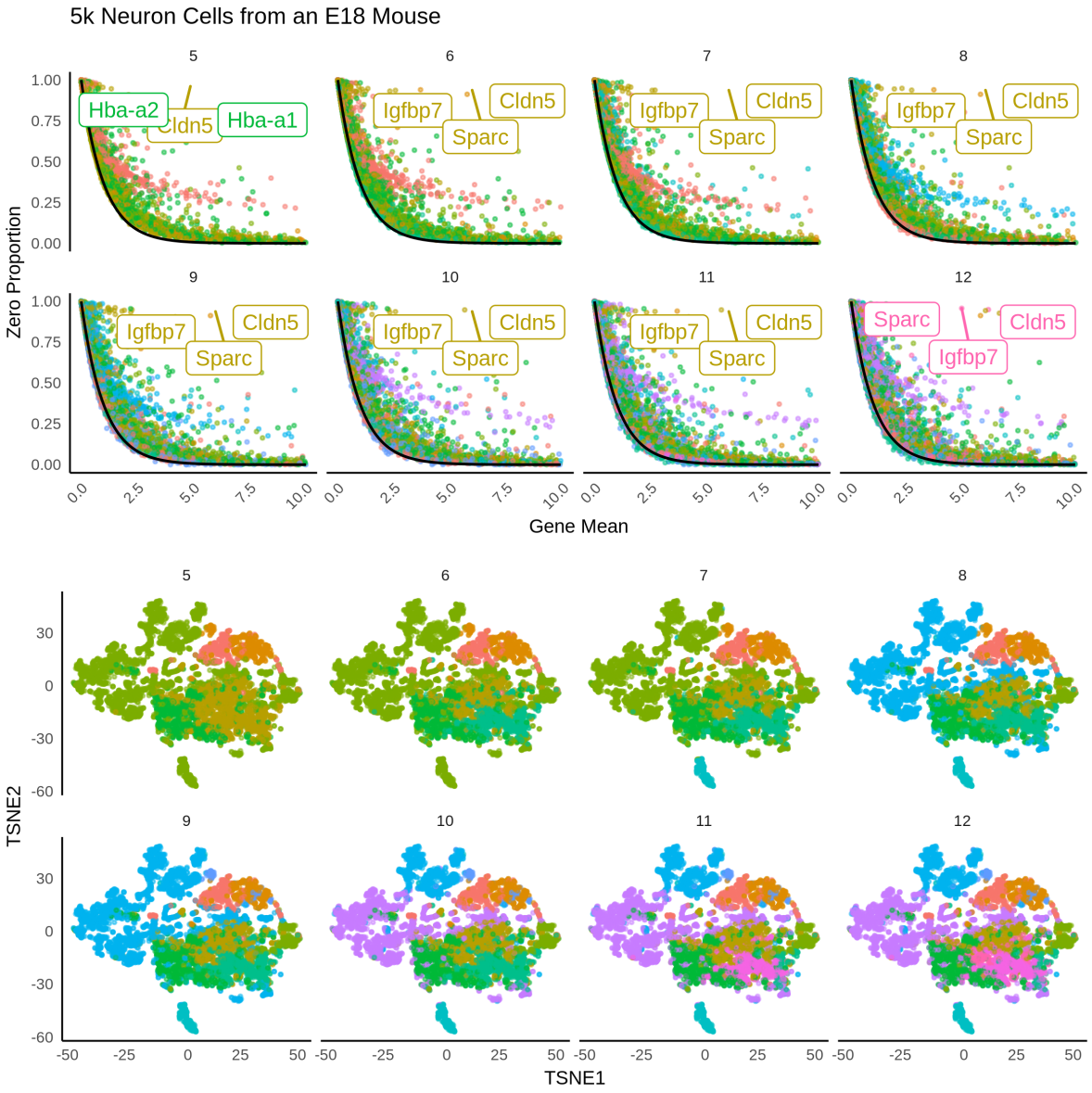


Supplementary Figure 6.22: Visualization of the step-by-step clustering of HIPPO in various data sets. One drawback is that when it can no longer identify distinct clusters and forced to cluster into more groups, it can divide existing groups into subsets and drive down the adjusted rand index.

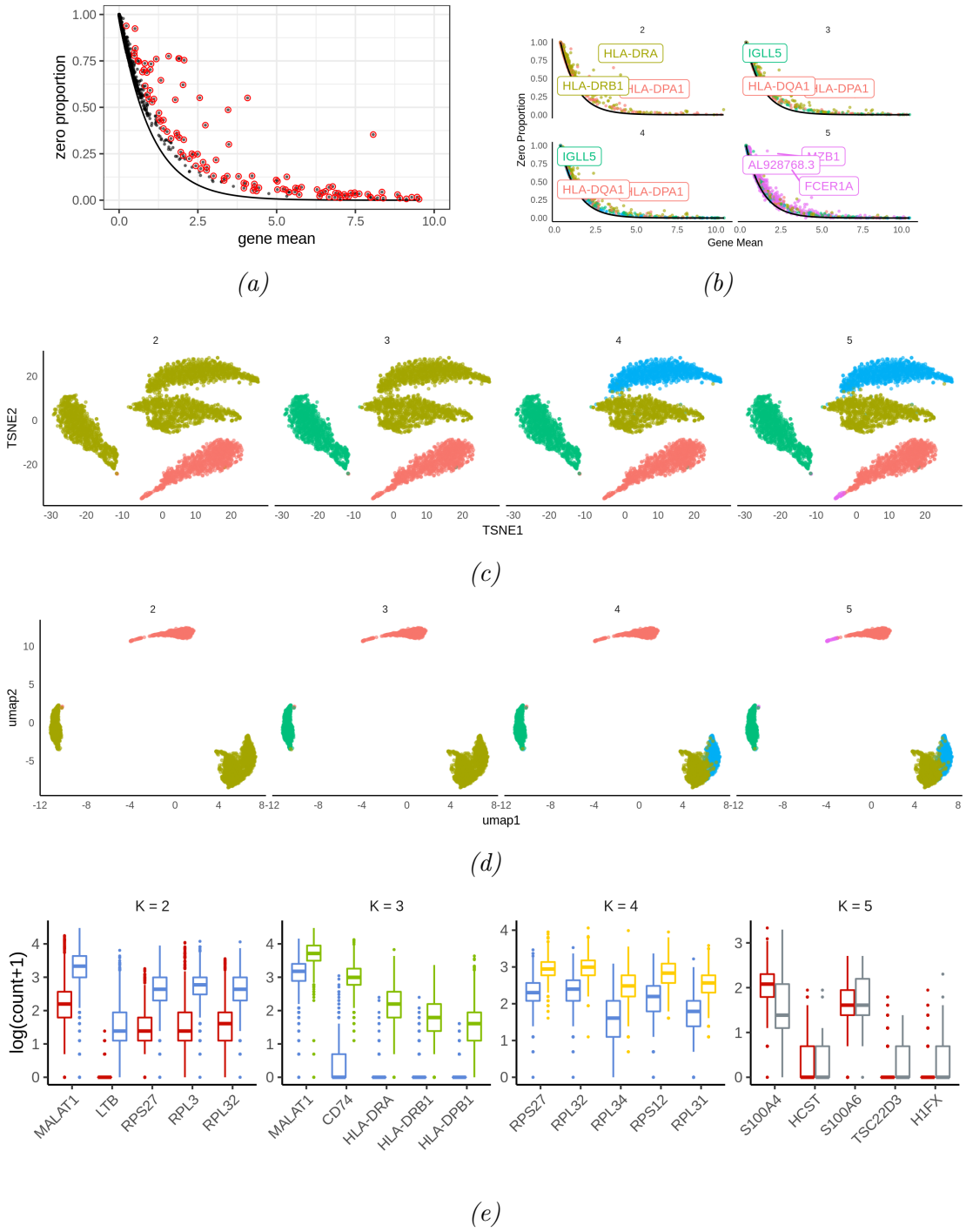


Supplementary Figure 6.23: Generalized PCA (gPCA) Lee (2015) takes into acount the count structure of the data to reduce the dimensions, and could be integrated into HIPPO procedure. However, empirically, its results are similar to the result of log transformation + PCA, and the result does not make up for the computational burden of gPCA.

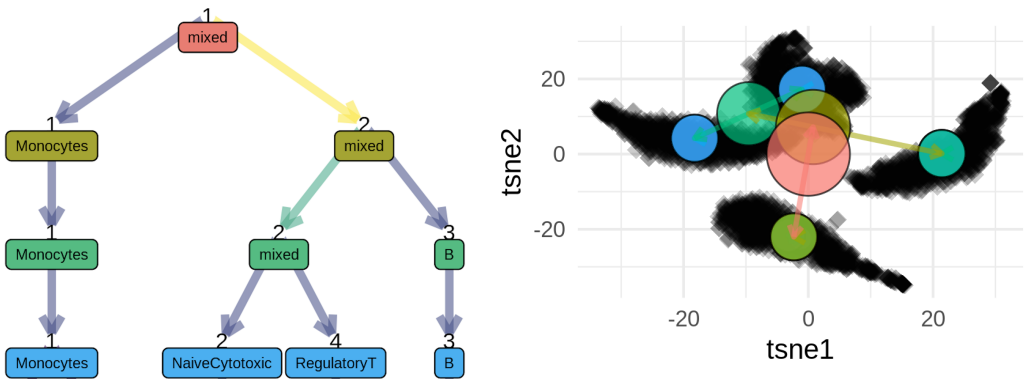

1k Brain Cells from an E18 Mouse
5
6
7
8
9
10
11
12
Reln
Hbb-bt
Hba-a2
Dlx2
Zero Proportion
1.00
0.75
0.50
0.25
0.00
0.0
2.5
5.0
7.5
10.0
Gene Mean
TSNE2
TSNE1
20
10
0
-10
-20
-30
-20
0
20
40



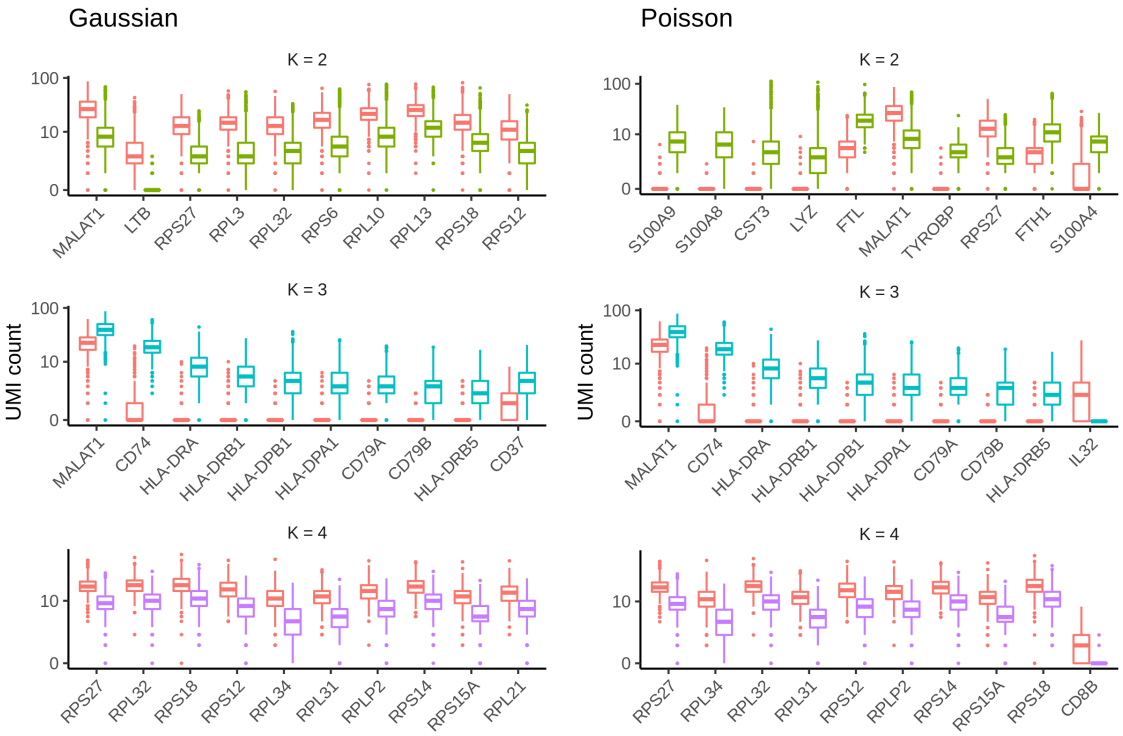
Supplementary Figure 6.24: Example analysis of HIPPO for cells from two examples of brain tissues with higher number of clusters. For each round of clustering, zero proportions are more aligned to the Poisson line. The t-SNE plot is more finely separated as the number of clusters increase. HIPPO can show the differentiation of cell types in sequencing manner.



Supplementary Figure 6.25: Sample analysis of Zhengmix4eq using HIPPO. The software first shows the diagnostic plot where zero-inflated genes are marked in red. Then it performs the clustering which leads to three sequential plots: zero proportions, t-SNE, and UMAP. Lastly, it shows the sequential differential expression analysis where color-coding matches the t-SNE and UMAP plots.



Supplementary Figure 6.26: Clustree Zappia & Oshlack (2018) package allows the tree-like visualization of the clustering result. The hierarchical structure gives insight to the overall structure of cell types and subtypes.



Supplementary Figure 6.27: Example of two differential expression methods in Zhengmix4eq data. Results are very similar, and their test statistics' spearman correlations are 0.98, 0.99, and 0.99 respectively for $K = 2$, $K = 3$, and $K = 4$.

6.3.2 Supplementary Tables

Cell Type	Proportion of Genes with Higher Variance
B cells	0.4939857
Naive Cytotoxic	0.3700687
Monocytes	0.4969385
Regulatory T	0.4766347
Helper T	0.4620756
NK	0.7163943
Memory T	0.5348796
Naive T	0.3535812

Supplementary Table 6.2: Proportion of genes that have higher variance in heterogeneous population than in homogeneous population. Using gene variance as feature selection would not be effective for detecting cellular heterogeneity.

	68K	Azizi09	Azizi10	Azizi11	Freytag	PBMC3k1	PBMC4k2
Antisense	2356	1701	1571	1537	269	305	329
HLA	23	21	21	21	29	31	21
IG-C	0	9	11	10	13	13	13
IG-C pseudo	0	1	2	2	5	4	6
IG-J gene	0	0	1	3	1	2	0
IG-V gene	0	60	49	77	61	77	67
IG-V pseudo	0	6	6	5	1	7	6
lincRNA	2202	1258	1140	1170	183	210	216
miRNA	0	1	2	1	1	2	0
misc RNA	1	0	0	0	96	190	0
Mt-rRNA	0	0	0	0	2	2	0
MT-tRNA	0	0	0	0	12	11	0
Polymorphic pseudo	0	6	6	6	11	11	7
Processed transcripts	3	58	52	55	86	88	46
Protein coding	15338	13684	13493	14080	12326	13025	13347
rRNA	0	0	0	0	14	18	0
Sense intronic	0	4	5	4	31	51	1
Sense overlapping	0	1	1	1	3	3	0
snoRNA	2	0	0	0	17	38	0
snRNA	0	0	0	0	71	134	0
TR-C gene	0	5	5	5	5	5	5
TR-J gene	0	1	2	46	0	0	0
TR-V gene	0	92	90	90	69	76	80
TR-V pseudogene	0	12	10	9	4	3	5

Supplementary Table 6.3: Gene counts for each data set and each gene type for PBMC data Azizi et al. (2018); Freytag et al. (2018); Zheng et al. (2017). Most of the genes are categorized as protein coding genes.

Azizi Patient 9 Rep1		$z \leq 3$	$z > 3$	Proportion	χ^2_1 statistic
Before Clustering	Immune genes	98	109	52.66%	553.66
	Others	15476	1262	7.54%	$p < 2.2e-16$
After Clustering	Immune Genes	826	678	45.08%	10960
	Others	145941	5152	3.41%	$p < 2.2e-16$

Supplementary Table 6.4: Azizi Patient9 Replication 1. Immune-related genes include HLA-gene, IG C gene, IG C pseudogene, IG V gene, IG V pseudogene, TR C gene, TR J gene, TR V gene, and TR V pseudogenes. The χ^2_1 statistic is computed through Pearson's chi squared test for independence of the two by two table. Clustering was performed using the true labels provided by the original paper Azizi et al. (2018). Each gene is recorded once for each cell type, explaining the increase of the number of genes. By repeating the Pearson's chi squared test for the combined data for each cell type, we are implicitly assuming that each cell types are independent.

Supplementary Text

There are two ways to estimate the zero proportions. The observed zero proportion is $\hat{p}_g = \frac{\sum_{c=1}^{C} \mathbb{1}_{X_{gc}=0}}{C}$ which measures the proportion of cells with zero counts across all cells. The expected zero proportion is, under Poisson assumption, $e^{-\bar{X}_g}$ which is the exponential of the negative average count $\bar{X}_g$ which is used as a proxy for the true gene mean.

The distributions of the two estimates are as below.

- $\hat{p}_g \sim \mathcal{N}(0, \frac{p_g(1-p_g)}{C})$, meaning

$$E(\hat{p}_g) = p_g$$

$$Var(\hat{p}_g) = \frac{p_g(1-p_g)}{C}$$

- $e^{-\bar{X}_g} \sim log\mathcal{N}(\lambda_g, \lambda_g/C)$, which means

$$E(e^{-\bar{X}_g}) = p_g^{\frac{2C-1}{2C}}$$

$$Var(e^{-\bar{X}}) = (p^{-\frac{1}{C}} - 1)p^{\frac{2C-1}{C}}$$

The distribution of the difference of normal and log-normal distribution is not trivial, especially because $\hat{p}_g$ and $\bar{X}_g$ are not independent. For practical convenience, $e^{-\bar{X}_g}$ is assumed to be equal to $e^{-\lambda_g}$, the expected zero proportion using the unobserved, true gene mean.

One consequence of this method is that there is a small bias. $E(e^{-\bar{X}_g}) = p_g^{\frac{2C-1}{2C}}$ is greater than the true expected zero proportion $e^{-\lambda_g} = p_g$. However, as we deal with more than 1,000 cells, this difference is negligible and disappears asymptotically.

The second issue is the underestimation of variance — the variance of $e^{-\bar{X}}$ is ignored when the distribution of $\hat{p}_g - e^{-\bar{X}_g}$. However, the ultimate goal of this method is to select

the features rather than making correct inferences. The z-score threshold is defined by the users, and they can alternatively choose to select top 2000 genes, in which case the variance does not have an impact on feature selection as the ordering of the important genes are not affected much by the variance. The method is still a valid approach for feature selection.