STATISTICAL METHODS FOR THE ESTIMATION OF CELL-TYPE COMPOSITION AND CELL-TYPE SPECIFIC ASSOCIATION STUDIES

Douglas Roy Wilson, Jr.

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2018

Approved by:

Joseph G. Ibrahim

Wei Sun

Quefeng Li

Naim Rashid

Di Wu

ABSTRACT

Douglas Roy Wilson, Jr.: Statistical Methods for the Estimation of Cell-Type Composition and Cell-Type Specific Association Studies
(Under the direction of Wei Sun and Joseph Ibrahim)

Samples of human tissues used in biological research are often impure. Such samples contain cells of the type under study and multiple ancillary cell types, leading to inaccurate expression estimates and analysis for the cell type under study. While estimates of cell type abundance can be of interest on their own, their use is critical to the correction of differential expression testing in heterogeneous cell type samples to account for differential cell type abundance across conditions. This dissertation develops and examines three statistical models for the estimation or use of cell type abundance profiles in the analysis of RNA-seq data from heterogeneous cell type samples.

Regarding estimation of cell type abundance profiles, we propose two models: IsoDeconv and ICeD-T. The IsoDeconv model approaches abundance estimation using isoform-level expression. We extend the IsoDeconv model to allow for biological variability in isoform expression across samples. The IsoDeconv model is assessed via simulation and through use of *in silico* mixtures of genuine RNA-seq expression datasets from non-cancerous human cell lines. The ICeD-T model approaches abundance estimation deconvolution using gene-level expressions while allowing for aberrant gene behavior within mixed cell type samples. Estimation properties of ICeD-T are assessed via simulation and validated in both microarray and RNA-seq datasets.

Transitioning to the use of abundance profiles in the analysis of heterogeneous cell type samples, we propose pTReCASE. pTReCASE is an expression quantitative trait locus (eQTL) mapping technique for use in bulk tumor samples. pTReCASE extends current eQTL mapping methods for tumor tissues to estimate eQTLs within tumor and normal cells separately. The type I error and power of pTReCASE are assessed via simulation before application to the study of breast cancer data from 547 Caucasian women.

ACKNOWLEDGEMENTS

To my wife, Bailey Chu: your strength of purpose and calming spirit have been indispensable in making this dissertation possible. You are a boundless source of inspiration and a daily reminder to enjoy the life we are creating together. To my mother, Dr. Kelly Wilson: your constant counsel and encouragement have buoyed me during the most difficult moments of this process. Our many discussions regarding the nature of research and the scientific process have been invaluable to my growth as a researcher. To my father, Doug Wilson: You are a shining example of the man I aspire to be each and every day. No dog-walk is complete without your gentle ribbing and words of encouragement. I love each of you and thank you for your support over the past 6 years.

To Herbert Chu, Susan Marcinko, Greg Morse, and Melanie Rhoads: thank you for the many shared dinner conversations and trips to the movies. Words cannot express my gratitude for all of the assistance and care you have provided throughout my brief time as a member of your family.

I would like to thank my advisors, Dr. Joseph Ibrahim and Dr. Wei Sun. Your tireless support, optimism, and drive have propelled me beyond the limits of what I thought possible. You have helped me to grow as a statistician and as a researcher and for that I am endlessly grateful. I would also like to thank Dr. Quefeng Li, Dr. Naim Rashid, and Dr. Di Wu for serving on my committee. I would like to thank the National Cancer Institute for supporting 5 years of my study on the *Biostatistics for Research in Genomics and Cancer* Training Grant (NCI grant 5T32CA106209-07). Finally, thank you to the Lineberger Comprehensive Cancer Center for supporting my final year of research with the University Cancer Research Funds grant (UCRF).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ENCODE	Encyclopedia of DNA Elements
eQTL	Expression Quantiative Trait Loci
GWAS	Genome wide association studies
ICeD-T	Immune Cell Deconvolution in Tumor Tissues

CHAPTER 1: INTRODUCTION

Almost invariably, tissue samples collected from human sujects are not restricted to a single cell type. Instead, each sample is a mixture of cell types (e.g. immune cells, fibroblasts, etc.). In tumor tissues, mixtures arise naturally from incomplete separation of normal cells from the tumor section or from the infiltration of normal cells into the tumor itself (e.g. immune cell infiltration). It is often of interest to examine the composition of such mixtures with respect to the constituent cell types. In the case of immune cell infiltration within tumors, abundance profiles can provide biological insight into the body's response to cancer and can identify possible deficiencies in this response. When the abundance profile itself is not of interest, cell type composition is crucial to the study of cell-type specific differential expression. Failing to consider differential abundance profiles risks conflating expression differences due to shifting profiles with true differential expression.

Physical methods exist to separate the mixture tissues into their distinct, constituent cell types. However, these techniques can be prohibitively expensive and remove the cells from the natural environment in which they live. The use of computational methods to deconvolute gene expression from mixed cell type samples sidesteps these obstacles while still providing useful information in their analysis. It is the purpose of this dissertation to explicitly model cell type abundance profiles and incorporate such profiles into the study of heterogeneous cell types samples using gene expression profiles from RNA-seq experiments.

In chapter 3, we develop IsoDeconv, a model for cell type abundance estimation which utilizes isoform expression information. Since isoform expression is a more granular examination of the expression products utilized by a cell, it may be more sensitive to cell type differences existing between cells of similar lineages (e.g. CD8+ T-cells, $\gamma\delta$ T-cells) than gene-level expression. Thus, the proposed methodology would be more accurate than current gene-level methods for deconvolution of highly similar cell types. IsoDeconv is extended to allow for biological variation in isoform expression across samples. This model is assessed via simulation and through the use of *in silico* mixtures of genuine RNA-seq expression datasets from non-cancerous cell lines.

In chapter 4, the abundance estimation problem is revisited using gene-level expression data. We propose a model for the estimation of immune cell abundance within tumor tissues called ICeD-T. ICeD-T is designed to estimate immune cell abundance profiles in the presence of aberrant gene behavior using tumor purity estimates, if available. ICeD-T is validated on microarray expression data from human blood and an RNA-seq expression dataset from melanomas. ICeD-T is also applied to the examination of response-to-treatment for an immune checkpoint therapy.

Chapter 5 transitions to the use of cell type abundance profiles in the analysis of heterogeneous cell type samples rather than the estimation of such profiles. In this chapter, we propose a statistical model for the identification of expression quantitative trait loci (eQTL) within tumor tissues called pTReCASE. As outlined above, tumor tissues contain tumor cells and normal cells. pTReCASE extends current tumor-tissue eQTL mapping techniques to estimate eQTLs in tumor and normal cells separately. We demonstrate that pTReCASE provides proper Type I error control and improved power in the detection of eQTL. pTReCASE is applied to the study of gene expression data from the breast cancers of 547 Caucasian women.

CHAPTER 2: LITERATURE REVIEW

2.1 The Biology of Transcription and RNA-Seq

Next generation sequencing technologies (NGS) provide a wealth of knowledge for the analysis of the cell. For example, one can examine the genetic sequence, interrogate the use of DNA-protein interactions, and explore the transcriptional activity of the cell in order to uncover the biological determinants governing cell behavior and function. Ribonucleic acid sequencing (RNA-Seq) allows examination of the latter, providing a snapshot of the building blocks each cell uses to create proteins. These proteins, in turn, are the basic units used to perform work within a cell. Thus, by providing a snapshot of the transcriptional activity of a cell, RNA-Seq allows examination of a cells identity by examining the tools it will use to perform work.

Before discussing RNA-Seq, a discussion of the basic biology it examines is warranted. In order to create proteins, cells translate the genetic information contained in DNA into RNA through the process of transcription. Once the DNA has been transcribed into RNA, cellular mechanisms edit the RNA transcript. This editing removes regions of code, termed introns, which will not be used in creating proteins. The remaining regions of code, called exons, are retained in the final messenger RNA (mRNA) transcript and will be expressed through proteins. A single gene in the genome can produce multiple mRNA transcripts and thus multiple proteins through the use of alternative splicing [1]. Alternative splicing, occurring in at least 90% of human genes, is the process by which a single RNA transcript has various exons removed from the final mRNA transcript to create multiple, unique transcripts called isoforms [2]. RNA-Seq measures the presence and amount of the final, processed mRNA transcripts within the cell.

The process of RNA-Seq begins with the purification of the RNA sample. This typically involves processes to remove ribosomal RNA – a subset of non-coding RNAs which comprise 90%

of RNA product in the average cell – or the enrichment of the sample for messenger RNA (mRNA). Once a sample of purified RNA is obtained, these molecules are often sheared to reduce the length of fragments before sequencing and to remove structural impediments to downstream processing. These RNA sequence fragments are then reverse transcribed into complementary DNA (cDNA). This cDNA is then amplified for sequencing and sequenced on one of numerous massively parallel DNA sequencers including Illumina Genome Analyzer, ROCHE 454, or SOLiD sequencers. Often, these sequenced RNAs are then mapped back onto a reference human genome to determine their site of origin and summarized by genomic locus or isoform [3].

The output of the typical RNA-Seq experiment involves counts of the number of sequence fragments at various sites in the genome. However, comparison of counts both across and within samples requires normalization. Archetypical normalization attempts to remove imbalances in RNA-Seq read counts due to differences in the amount of RNA procured in a given experiment and/or the length of the genomic feature of interest. This is typically performed through use of Fragments per Kilobase per Million Mapped reads (FPKM) correction, which divides each observed count by the amount of mapped sequence present in each sample and multiplied by the length of the genomic feature [4]. In this way, counts across different samples and genomic features can be viewed on a similar scale.

Notable sources of bias in RNA-Seq counts at different genomic loci include mappability biases, GC-content biases, and nucleotide start site biases. Mappability bias implies that regions of DNA with fewer repetitive elements and longer stretches of unique arrangements of DNA will see larger observed RNA-Seq read counts [5]. GC content biases indicate that regions with a large proportion of GC nucleotides tend to experience larger RNA-Seq read counts [6, 7, 8]. Finally, though somewhat attributable to a preparatory technique known as random hexamer priming, it was observed that RNA-Seq reads tended to start at genomic loci with G or C nucleotides [8]. The correction for these biases is most critical in situations where attempts are made to compare transcription activity levels between different genomic locations, not for the examination of transcriptional activity at the same genomic loci across different conditions.

Compared to RNA microarrays, RNA-Seq provides numerous improvements. While microarrays can only consider pre-specified genomic locations or splice junctions, RNA-Seq provides a large dynamic range of reads covering the genome. This allows for the discovery of novel genes, transcripts, and splice sites [9, 10]. In addition, RNA microarray data processing often requires the use of opaque normalization techniques whose impacts on resulting measures of signal strength are not readily comprehensible. Current popular RNA-Seq normalization techniques are simple to apply and their effects on resulting signal strength measures are much clearer.

2.2 Count Models for RNA-Seq Expression

As discussed in the previous section, RNA-Seq output comes in the form of observed sequence counts at various genomic loci or transcripts. Typical models for RNA-Seq data have involved count distributions such as the multinomial or discrete model, the Poisson model, and the negative binomial model. Each of these will be discussed briefly.

Early models of RNA-Seq expression assumed that reads represented independent, random realizations from a selection process that was uniform across the length of a transcript and dependent on the activity level of the transcript or locus in question [11]. Thus, the multinomial model was a natural starting point. The multinomial models can occur on one of two scales which summarize the same data in different ways. The first scale is to model expression at each genomic locus. The second scale models the likelihood for each individual RNA read. RNA-Seq by Expectation Maximization (RSEM) is a good representative of the individual RNA read level model [12]. Let θ_i be the proportion of expression and l_i denote the length of a transcript i. The probability that one would observe a read fragment r that arises from transcript i is given by:

$$P\left\{r \in i\right\} = \frac{\theta_i l_i}{\sum \theta_j l_j}$$

This modeling technique accounts for the length of different isoforms in generating RNA-Seq reads. RSEM also incorporates corrections for positional biases and sequencing errors by adding weighting factors for each possible location that a given read can map within the genomic feature of interest [12].

As RNA-Seq expression libraries are quite large, and the expression of typical genes can be considered quite small in relation to the entire library, the multinomial model lends itself well to an approximation via Poisson counts. Marioni et al [13] assess the variability in read counts at certain loci across technical replicates of RNA-Seq expression. A technical replicate is data arising from sequencing of the same biological sample on differing lanes or runs of a DNA sequencer. They found that, at least in the technical replicate setting, a Poisson model accounting for library size and average transcript expression adequately captured the mean and variance for differential expression testing [13]. It has been proven that estimation of transcript abundance using the multinomial and Poisson counts is identical under identical bias models [14].

It has been noted, however, that the Poisson model does not adequately capture the variance in expression levels across biological replicates as compared to technical replicates. It was found that the observed variance in expression tended to exceed the mean across biological replicates, suggesting the presence of overdispersion [15, 16]. It is in this setting that the negative binomial models were introduced to correct differential expression analyses for this greater observed variation. Anders and Huber utilized a library-size and locus specific expression dependent model of mean expression. In a low replicate setting, Anders and Huber borrow information across genes to estimate overdispersion by hypothesizing a functional relationship between average expression and variance using local regression to obtain estimates [15]. Robinson and Smyth proposed an alternative estimator for the overdispersion using a weighted likelihood approach, which acts as a penalty on a genes estimate of overdispersion towards a common overdispersion using a likelihood function motivated by Empirical Bayesian estimates of penalty parameters [16].

In a model titled IsoDOT, Sun et al provided an extension to the negative binomial models often used in RNA-Seq that shifted focus from the modeling of gene or transcript level expression to exon set level expression. By breaking down expression within a gene into sets of reads which overlap particular groupings of exons, Sun et al were able to interrogate the alternative splicing mechanisms used in transcription regulation and thereby model isoform specific abundances across varying conditions (e.g. cell type, disease status) [17].

The following introduces the structure of the IsoDOT model for a single gene, which will set the stage for further extensions of this model to deconvolution of expression data from multiple cell type sources. Consider exon sets A and B. A and B are distinct groupings of non-overlapping regions of exonic code within a particular gene. Exon sets A and B may contain one or more exons in common, but their set difference is non-empty. Each read is partitioned to a single exon set if and only if it overlaps each exon in the set and no others [17].

One can expect the number of reads mapping to exon sets to depend on, at the very least, two factors. The first is the length of the exon set. As noted above, larger genomic features produce larger counts due to the fragmentation of RNA output for sequencing. Additionally, more reads will be expected for a particular exon set if the isoforms which use the exons in this set are more active [17].

Sun et al incorporate these features into the mean structure of a negative binomial model by defining an effective length of each exon set within each transcript. Thus, for exon sets with exons not composing an isoform I, the effective length of this exon set for this isoform is 0. Otherwise, the effective length of an exon set A in isoform I is the expected number of start sites for a read covering these exons assuming the ordered composition of an isoform I [17].

In summary, Sun et al model the count of reads across a group of exon sets in a single experiment the following way, where X is an $E \times I$ matrix of effective lengths, γ is the vector of normalized isoform expressions, and t_i is a measure of read depth within the sample:

$$y_i \sim \texttt{Neg.Bin.}\,(\mu = t_i X \gamma, \phi)$$
$$\mu = t_i \left(\sum_{i=1}^{I} x_u \gamma_u \right)$$

The true novelty of this approach is in its partitioning of reads at the gene level into varying exon sets which create the information necessary for probabilistically estimating the isoform expression values without probabilistically assigning reads to transcripts [17].

Sun et al also incorporate an iterative adaptive lasso penalty in the optimization of the γ parameters to limit the number of estimated isoform parameters above 0. In this way, it simultaneously performs variable selection while estimating isoform expression. Differential usage testing of isoforms across conditions uses a bootstrapped likelihood ratio statistic [17].

2.3 Expression Deconvolution Using RNA-Seq Data

Almost invariably, sequencing data derived from human tissue samples are contaminated. Such samples contain not only cells of the type desired for study, but also contaminating cell types such as blood or epithelium. The need for statistical procedures which can deconvolute these sequencing data into components from contaminating cell types and those cell types of interest has frequently been addressed in the literature. Accurate estimates of cell type abundance provide useful information in the correction of analyses for differential expression of genes or transcripts for differences in cell type composition, estimating the presence of immune cell contamination in tumor samples, and the guiding of patient care in the presence of intra-tumor heterogeneity or the presence of multiple competing cancers in a single tumor.

There are two main approaches to the deconvolution of mixed cell type samples into their constituent cell types expressions for analysis. The first is the mechanical separation of cell types into groups of purified cell types using techniques such as laser capture microdissection or fluorescence activated cell sorting (FACS). Concerns exist as to whether these physical procedures influence the cellular environment and thus the expression patterns of cells [18]. The second approach attempts to deconvolute expression in silico through the use of statistical models to mathematically predict cell type proportions. One can group the in silico approaches into a few different subcategories: ratio based models of expression, linear models of expression, infiltration score models, models which incorporate prior information about cell type abundance, and perturbed expression models.

Ratio based models of expression typically handle the two cell type case. These cell types can be two different normal cells or two subsets of cells (e.g. normal and tumor). Consider a mixture sample composed of cell types A and B. Let E_{iA} represent the expression of gene i in cell type A. Let E_{iAB} be the expression of gene i in the mixture tissue. It is assumed that $E_{iAB} = p_A E_{iA} + (1 - p_A)E_{iB}$. The idea behind ratio-based models is to compare the expression of a gene i in a pure sample of one cell type A, against the expression of this gene in a mixture sample of types A and B. Thus, define $R_{iAB} = E_{iAB}/E_{iA}$ where a reference value of E_{iA} is known a priori. In the absence of noise, one would expect the minimum of this ratio across genes to approach p_A since one considers genes more highly expressed in A and more lowly expressed in B. However, the presence of biological or experimental noise complicates the analysis and ruins estimation of the proportions.

Two different approaches to handle this noise were examined. In the approach used by Gosink et al, several realistic simulations are computed across a range of values for p_A. For each simulation, the R_{iAB} are ranked by magnitude. The fifth percentile R_{iAB} score is used in a regression equation to predict p_A. The model used was selected from multiple regression models considering the q-th percentile R_{iAB} and its square as covariates. The one which provided the best R^2 value was chosen [19]. Clarke and Seo, however, accounted for noise using a transformed version of the R_{iAB} values. The transformed version of the R_{iAB} $\left[tR_{iAB} = \frac{\log(1+\alpha E_{iAB})}{\log(1+\alpha E_{iA})}\right]$ values uses an optimized choice of correction factor (α) found by computing the "elbow" of a plot of the mean transformed R_{iAB} values against the correction factor. The minimum tR_{iAB} was found to be a good estimator of p_A [20].

Another ratio based method, UNDO uses the ratios of expression (R_i) between two mixed tumor/normal samples across all available genes i. A gene is considered a marker gene if it falls within some pre-specified small-ϵ neighborhood of the minimum and maximum R_i. As before, these maximum and minimum R_i are indicative of genes used exclusively in one cell type. Using a simple relationship between these marker genes expressions and cell type proportions in each of the mixtures, namely that the ratios are bounded by ratios of cell type abundances one can estimate cell type specific proportions for each cell type from sample mean expression values in the mixture tissues [21]. Without the use of pre-specified pure sample references, one could term this method unsupervised estimation.

Linear models of expression all begin with the following underlying model for the expression in a mixture tissue with K-constituent cell types. Let Y, a $g \times n$ matrix, be appropriately normalized gene expressions in n mixture tissues. Let C, a $g \times K$ matrix, be the appropriately normalized, Gold-standard reference expression at each gene for each of the K cell types. Let W, a $K \times n$ matrix, represent the mixing fractions of the different cell types across the k tissues. The model for expression in the mixture tissue is then given by $Y = CW$. The majority of these methods require a Gold-Standard reference expression profile, preferably composed of genes with differential expression across cell types.

The earliest methods in linear models were designed and fit using simulated annealing procedures to obtain estimates of the matrix W [18, 22]. In the Lu paper, simulated annealing to minimize prediction error was used to fit the linear model. Lu et al investigated yeast cell population dynamics under various environmental conditions [22]. Shen-Orr applied the technique to human blood samples [18]. Early linear modeling procedures required post-estimation correction of physically impossible results such as negative proportions or proportions which did not sum to one by renormalizing positively estimated proportions to fall between 0 and 1[23].

After the advent of RNA-Seq, Gong et al tested the applicability of the linear modeling procedures previously developed in the microarray setting. Advocating proper normalization, such as RPKM in the RNA-Seq setting, their method DeconRNASeq utilizes quadratic programming to find an estimate of W which minimizes the squared prediction error and is subject to the normal physical constraints of proportions. DeconRNASeq also refines the Gold-Standard expression matrix by utilizing a subset of rows of the reference matrix with the smallest condition number, thereby producing more stable estimates of the cell type proportions [24].

Much like DeconRNASeq, CIBERSORT uses a refined Gold-Standard reference profile A that is made robust through the use of condition number selection to remove uninformative genes from the selection matrix. However, CIBERSORT uses support vector regression to obtain estimates of cellular proportions [25].

The digital sorting algorithm (DSA) was designed by Zhong et al to incorporate the linear model in the absence of a reference profile matrix. Applied to gene-level, normalized microarray data, DSA requires a priori knowledge of marker genes genes expressed in only one of the K cell-types. Let C^* be a diagonal matrix with the average expression across all marker genes of cell type k at $C^*(k,k)$. Let Y^* be the corresponding observed average expression values in the mixture cell type samples. Assuming that the number of mixture samples exceeds the number of cell types in the mixtures, then $C^{*-1}Y^*$ creates an overdetermined system of equations for the $C^*(k,k)$ since each column of the matrix $C^{*-1}Y^*$ must sum to one. Once these values are estimated, the estimation of the W matrix is simply $W = C^{*-1}Y^*$ [26].

Finally, the DeMix model uses a similar linear structure for mean intensity in non-log transformed microarray data, but it does so only in the two cell type case of tumor versus normal cells. It proposes a log base-two normal distribution for microarray intensity in tumor and normal cells, leading to a mixture expression distribution that is the weighted convolution of these two distributions. Estimation proceeds through use of a mixture of Nelder-Mead optimization steps to cyclically update estimates of tumor purity, mean and variance parameters across samples for each genes tumoral and normal expression, and tumoral intensity expression at each gene until convergence of the likelihood. Multiple mixture samples are required for proper deconvolution [27].

In stark contrast to the majority of the previously described models, two methods of RNA expression deconvolution focus on the estimation of gene expression values from within a mixed cell type sample using prior information regarding cell type abundances. In the absence or pure sample expressions, prior information regarding cell type abundance allows for identifiable models of cell type specific gene expression. TEMT uses RNA-Seq data and a model similar to that discussed for RSEM to quantify cell type specific gene expressions. The authors of TEMT encourage strong information regarding cell type proportion for proper deconvolution of cell type specific gene expression signatures [28].

DSection also makes use of prior information regarding cell type proportions in the deconvolution of microarray intensity data. Under a normal distribution model, DSection utilizes a weighted sum of cell type specific expressions for the mean with heteroscedasticity across genes. It incorporates the use of normal priors on cell type specific expression, gamma priors for probe specific variance measures, and Dirichlet priors for cell type proportions. Parameters regarding sample specific contributions of cell type intensity, probe specific variance measures, and cell type proportions are updated through the use of MCMC sampling [29].

The last subset of deconvolution models can be termed perturbation models. In these models, the reference expression levels of various cell types in the mixture are perturbed versions of pure sample reference expression levels. These models attempt to correct for differences in mixture sample and reference sample profiles arising from environmental effects, cell culture effects, or mutational effects leading a cell to become a cancerous version of itself.

PERT, a model designed by Qiao et al, uses microarray RNA expression to estimate cell type abundances. PERT assumes that the latent reference gene expression profile (D^*) of the heterogeneous cell type sample is a multiplicative perturbation of a true expression reference (D) given by a renormalized version of $diag(P) * D$ where each column sums to 1 and P is a vector of perturbation factors for each gene. Maximum likelihood estimation of cell type abundances and perturbation factors is performed assuming a discrete model for each unit of intensity. Regularization of cell type abundances and perturbation factors is incorporated across multiple mixture samples using dirichlet and gamma priors, respectively [23].

IsoPure, the successor of ISOLATE [30], models mRNA microarray expression in a tumor in the following way [31]. As before, let C be a reference expression matrix for each normal cell type, normalized so that each column sums to one, d_n be a reference profile for sample ns particular cancer, and θ_n be the mixing proportions of sample n. Let ω define an indicator vector of the cell-type origin of the cancer being examined in the study. Then, we may model y_n, the gene-expression vector for the n-th sample as:

$$
\begin{aligned}
y_n &\sim \texttt{Multinomial}\left(\begin{bmatrix} C & d_n \end{bmatrix} \theta_n\right)\\
d_n &\sim \texttt{Dirichlet}\left(k_n m\right)\\
m &\sim \texttt{Dirichlet}\left(k' C\omega\right)\\
\theta_n &\sim \texttt{Dirichlet}\left(\nu\right)
\end{aligned}
$$

In this way, they model cancer as a random perturbation of a single cell type profile, allow for each individual's cancer profile to vary about the true cancer profile and regularize cell type proportions across individuals. The IsoPure model provides cancer purity estimates, an identification of the cell-type of origin of the cancer under study, while at the same time restricting cell-type abundances to follow a similar profile across subjects [31].

The majority of these proposed methods were designed specifically to handle microarray data. Only DeconRNASeq, Undo, Cibersort, TEMT, and IsoPure were designed specifically with RNA-Seq methodology in mind. Of these, only DeconRNASeq, TEMT, and EPIC specifically applied to and validated using RNA-Seq data. The remaining models suggest validity in the RNA-Seq case, but do not test for it. Thus, the proposed methodology for cell type expression deconvolution in RNA-Seq data is limited.

Additionally, some models require prior information or regularization procedures with respect to cell type abundances for deconvolution. Prior information with respect to cell-type abundance may difficult to obtain due to cost, infeasibility of good cell type separation, or a desire to maintain natural environmental conditions for the heterogeneous cell type samples. Models such as IsoPure or Isolate, which require regularization of cell type abundance parameters, are somewhat limited to the case where multiple mixed tissue samples exist or natural regularization parameters are available.

Finally, the majority of these models measure expression on the gene level. Thus, deconvolution among cell types necessitate cell types that express differently at the gene level. However, as noted earlier, at least 90% of genes utilize alternative splicing[2]. It may be the case that alternative splicing could be more sensitive to cell type differences than gene expression, especially in cells of highly similar lineage, for example different varieties of B-Cells. Extension of these models to the transcript level expression would require that accurate transcript level expression information is available. Often, however, reads map to multiple transcripts and the estimation of transcript level expression is inexact and this uncertainty in reference expression profiles would not be captured.

2.4 Immune Cell Expression Deconvolution in Tumor Tissues

The developing relationship between a cancer and its host's immune system is well summarized in a hypothesis known as immunoediting. In summary, the immune system's attack upon a burgeoning cancer cell population becomes a driving force behind the development of immunosubversive cancer cells [32, 33]. Such cells are capable of hiding from the immune system, killing its cytotoxic population or even recruiting its suppressors to reduce further response [34, 35]. Much therapeutic development has been focused on therapies, called immunotherapies, which enhance the bodys natural immune response by counteracting these self-defense mechanisms in cancer. Despite recent successes in prostate cancers and melanomas, immunotherapies have been demonstrated to be highly disease- and subject-specific. For example, immune checkpoint blockade therapies have produced response rates as low as 12% in head and neck squamous cell cancers and as high as 85% in Hodgkins lymphomas [35, 36]. As immune cell composition estimation within human cancers has demonstrated prognostic value [37, 38, 39], such an approach provides a potential avenue for the investigation of therapeutic efficacy of immunotherapies as well as the identification of mechanisms for novel interventions.

Bulk expression analyses such as RNA-seq and RNA microarrays have become standard tools in the interrogation of infiltrating immune cells within the tumor microenvironment (TME). However, as bulk expression experiments capture the totality of expression from all cells within the TME, there is a need for statistical models which can disentangle the individual contributions of each cell type in order to estimate the composition of the TME.

The previous section discussed the bulk-expression deconvolution problem and several methods used in its address for general tissue types. Here we focus on an additional subset of deconvolution models used specifically to estimate immune cell proportions in tumor tissues. Pioneering methods sought to quantify immune cell infiltration by proxy using infiltration scores. In this setting, infiltration scores are quantities which are correlated with immune cell proportion in the tumor body; the larger the score, the more abundant the immune cell population they measure. Two such methods utilized the normalized expressions of a select few immune-specific genes (e.g. CD8A, GZMA, PRF1) to examine infiltration of cytotoxic immune cells within the tumor body [40, 41]. A similar method, MCP-counter, extended these models to compute infiltration scores for an increased array of immune cell subsets including B-cells, CD4 T-cells, CD8 T-cells, Monocytes, macrophages and more [42]. Becht et al computed the infiltration score for any category as the log-2 average gene expressions computed across genes with cell-type specific expression in the given immune cell subset. Infiltration scores are limited in their interpretations. At best, they allow one to rank subjects with respect to the level of immune cell infiltration. However, they do not allow one to assess whether there are more cells of one type than another.

An alternative to infiltration score techniques, linear regression based models remain popular deconvolution techniques for use in tumor expression datasets. However, within tumor tissues, the previously described approaches must be modified to account for the presence of a tumor cell type. This modification is often performed by restricting the "gold-standard" reference matrices to immune-specific genes, or genes expressed only within immune cells and not within tumor cells. With an appropriate selection of immune-specific genes, the mixture expressions being modeled remain a linear combination of the expressions across non-tumor cells only. To this end, the CIBERSORT team developed the Leukocyte Matrix 22 (LM22) to characterize reference immune expression across 22 different immune cell subsets at genes expressed only within the immune population[25]. TIMER uses a similar approach by constructing a "gold-standard" reference expression matrix for several immune cell subsets and restricting genes to those anti-correlated with tumor sample purity. In addition, the TIMER methodology removes high-expressions genes as these such genes tend to have high variability and exert undue influence on model estimates[43].

The most recent linear model for tackling immune cell deconvolution in tumor tissues is called EPIC [44]. EPIC utilizes constrained, weighted least squares to simultaneously estimate the proportions of several immune cell types and the proportion of tumor cells within the sample. EPIC uses a weighting scheme which considers the ratio of average gene expression to gene expression variance. Higher weights are assigned to genes with lower expression variance in accordance with the magnitude of their average expression across reference profiles. For use within tumor samples, particularly melanomas, EPIC creates a reference expression matrix, TRef, constructed from single cell RNA-seq expression profiles conducted on melanoma samples. Racle et al argue that immune cell expression profiles developed from circulating immune cells may not adequately capture their behavior within the tumor microenvironment. By assessing expression profiles directly from tumor infiltrating immune cells, EPIC seeks to correct for this possible disagreement [44].

Previously proposed methods for immune cell deconvolution in the tumor microenvironment suffer from three main disadvantages. First, each method fails to utilize knowledge of the level of non-tumor cell infiltration into the TME should this information be available. Second, as expression mixing occurs in linear-space, each of these methods deconvolve expression in the linear space. However, this fails to incorporate the beneficial properties of the log-transformation for RNA-seq and microarray expression experiments. Finally, previously proposed methods fail to provide a mechanism which can both control for and identify loci within *individual* mixture tissues which are inconsistent with measured references.

2.5 Expression QTL Mapping Using Gene Expression Data

Genome wide association studies (GWASs) have long been used as a tool for establishing a link between genetic variation and phenotypes. Genetic variation in GWASs is examined through the use of single nucleotide polymorphisms (SNPs), or single base pairs which vary across human subjects. The phenotypes considered often involve indicators of disease (e.g. cancer) [45].

The complex regulatory landscape that governs phenotypic expression made interpretation of GWAS results difficult. While a link between a disease and a SNP could be established, the biological mechanisms behind this association remained unclear. The advent of expression microarrays and RNA-Sequencing expanded the phenotypes available for GWAS examination to include RNA expression of individual genes. Investigation of gene-level expression afforded by such techniques allows for direct examination of the functional role of genetic variation in gene expression and helps to interpret GWAS results [46].

SNPs linked to changes in gene-level expression are termed expression quantitative trait loci (eQTL). EQTL are categorized in two groups, *cis*-eQTL or *trans*-eQTL, distinguished by the patterns of expression change they induce in affected genes [47, 48]. In order to understand the distinction between these two types of eQTL, recall that humans are diploid organisms. Normal cells within humans contain two homologous copies of each chromosome: a maternal and a paternal copy.

A locus is considered a *cis*-eQTL if it regulates expression of the gene it effects in an allele-specific manner [48]. For example, consider a mutation on a single allele at a transcription factor binding site which inhibits the initiation of transcription. Alleles which lack this mutation are able to successfully bind the transcription factor and thus initiation of transcription proceeds uninhibited. Such a mutation acts in an allele-specific manner and is thus considered a *cis*-eQTL. *Cis*-eQTL are often found close to the gene whose expression they alter. As a result, *cis*-eQTL are often misleadingly labeled ”Local eQTL”[47].

On the other hand, *trans*-eQTL are distinguished by genetic variants which impact expression of both alleles of a gene [48]. A common mechanism employed by *trans*-eQTL is the mutation of loci involved in the production of transcription factors which bind to all copies of a gene. Altering the expression or functionality of a transcription factor would thereby impact both alleles of the gene. *Trans*-eQTL may be found near the genes whose expression they alter but can also be found at a distance, possibly from a different chromosome [47].

Traditional eQTL mapping methods implicitly assume that an eQTL has the same effect on all cells within a sample. This is a reasonable assumption for samples with a relatively homogeneous cell population. However, tumor samples invariably contain both tumor cells and infiltrating normal cells (e.g. immune cells) and eQTL effects could differ between these two cell types. Previous eQTL studies in tumors ignored tumor purity, defined as the proportion tumor cells among all cells within the tumor sample,thereby assuming that tumors are composed of homogeneous cell types. For example, such studies regressed normalized gene expression on eQTL genotype and other population stratifying factors such as genotype principal components or gender [49, 50, 51, 52]. Other regression-based techniques propose first regressing tumor gene expression on estimates of methylation and/or somatic copy number before regressing the resulting residual expression on eQTL genotype[53]. Identification of tumor-specific eQTL is generally ad-hoc, labeling an eQTL tumor-specific when analysis within tumor tissues suggests an eQTL where none is found in a separate analysis of normal tissues.

An extension of these linear models has been developed to incorporate measures of tumor purity or, in the normal tissue setting, cell-type specific abundance. Westra et al utilize a proxy for neutrophil abundance as a covariate in a linear regression model to identify neutrophil specific eQTL within whole blood samples[54].

As an alternative to tumor purity, several studies have incorporated allele-specific expression (ASE) in analyses of eQTL within tumor tissues to strengthen conclusions. For example, Li et al also propose use of RNA-Seq data to examine allelic imbalance, or the proportion of reads mapping to one allele or another. Deviation of this proportion from 0.5 is termed imbalance[53]. However, by failing to integrate the regression and ASE components of the model, such studies restrict ASE analysis to a supportive role in eQTL identification. In a framework developed by Sun et al [55], it has been demonstrated that concurrent incorporation of ASE and gene-level expression improves power in the detection of *cis*-eQTL within normal tissues.

We briefly describe the model proposed by Sun for eQTL identification in normal tissues[55]. Sun examines expression data from RNA-Seq by modeling two distinct, yet overlapping components:

total read count and allele specific read count. Employing a negative binomial model, the total number of reads mapping to a gene are examined to estimate *trans*- or *cis*-eQTL effects. The subset of total reads which are uniquely mappable to a single allele are modeled in a beta-binomial framework and can only identify *cis*-eQTL effects. For *cis*-eQTL, these two components of the model are linked through use of a common eQTL parameter which is defined for a composite, normal tissue type. Previous extensions of this model have attempted to remove its dependence on prior genotype imputation and haplotype phasing [56].

CHAPTER 3: ISODECONV: CELL TYPE ABUNDANCE ESTIMATION USING RNA ISOFORM EXPRESSION

3.1 Introduction

Sequencing data derived from human tissue samples are often mixtures of heterogeneous cell types. Such samples contain not only cells of the type desired for study (e.g. tumor cells, B-cells), but a milieu of additional cell types. It is often of interest to quantify the abundance of each constituent cell type found within a heterogeneous cell type sample. In some cases, the abundance profiles themselves contain relevant information regarding biological response, such as the case of immune infiltration within a tumor. In others, abundance profiles are crucial for proper cell type-specific differential expression analyses. Cell-sorting and other physical separation techniques exist to partition heterogeneous cell type samples into purified samples of their constituent cell populations, but such methods can be costly and may even induce changes to the cellular environment which impact expression profiles [18]. As an alternative to physical separation methods, the development of statistical models for the deconvolution of expression profiles from heterogeneous cell type samples has become an active area of research.

In silico expression deconvolution models can largely be separated into three main developments: ratio-based models, linear models, and infiltration scores. Ratio-based models rely upon computing expression ratios between a mixed expression profile and a “gold standard” reference for a single cell type. The minimum of these ratios across genes roughly approximates the proportion of the referent cell type [19, 20, 21]. These methods are often limited to the two cell group case (e.g. tumor vs normal). In response to the limited cell populations interrogated by ratio based methods, linear modeling of mixture expressions was introduced. The traditional linear model framework assumes that appropriately normalized mixture expressions can be modeled as a weighted sum of contributions of “gold-standard” expression profiles from two or more cell types [22, 24, 25, 26].

Recent deconvolution models have been applied to the study of immune infiltration in tumors and have focused on the computation of infiltration scores, or unitless quantities designed to reflect increasing abundance of a certain immune cell [42, 43].

Almost exclusively, the proposed methods have been designed to examine and validate on gene-level expression only. Thus, appropriate deconvolution requires that cell types express differently at the gene level. In the case of highly similar cell types (e.g. CD8+ T-cells vs. $\gamma\delta$ T-cells), however, it may be the case that gene-level expression differences are minimal. Through a process known as alternative splicing, a single unprocessed mRNA transcript produced by a gene can form multiple distinct processed transcripts, or isoforms. Isoform expression, therefore, represents a more granular examination of the expression products utilized by a single gene [2]. In the case of highly similar cell types, the differential usage of isoforms may be more sensitive to cell type identity than higher-level gene expression.

In this chapter, we outline the development of two statistical models for expression deconvolution in mixture tissues which are capable of utilizing isoform-level expression differences between cell types. The first, IsoDeconvNB, posits a negative binomial structure for mixture expressions across cell types. The second, IsoDeconvMM, is the successor to IsoDeconvNB which is designed to explicitly model biological variability in reference isoform expression profiles.

3.2 Statistical Methods

3.2.1 The Data

Consider a biological tissue sample composed of K different cell types. The abundance of each cell type k, or the proportion of cells of type k in the heterogeneous cell type sample, is unknown and must be estimated. In order to estimate these proportions, IsoDeconv requires a single RNA-Seq experiment performed on the mixed cell type sample. In addition, it is assumed that there exist RNA-Seq experiments for each cell type k performed on purified samples of cells of this type; for a single cell type k, these N_k sets of reads are considered cell-type specific RNA-seq profiles. For each experiment, read counts are summarized at the exon level by counting the number of reads overlapping various sets of exons.

IsoDeconv assumes that there exists a list of cell-type specific genes wherein there are gene- and/or isoform- expression differences across the K cell types. Such a list of genes can be found using one of many differential expression testing methods for RNA-seq data. For the following, the Cufflinks suite was used to determine differentially expressed and/or regulated genes for use in deconvolution [57].

Furthermore, it is assumed that a detailed gene and isoform construction model is available for each gene at which expression is assessed. The gene construction model assumes knowledge of all non-overlapping exons utilized by the gene and their locations within the gene body. The isoform-construction model for each gene assumes knowledge of all isoforms used by the gene and their construction with respect to the known exons. For further details regarding the formation of these gene and isoform construction models, please see the supplementary materials of Sun et al [17].

3.2.2 IsoDeconv - Negative Binomial Model (IsoDeconvNB)

Within the IsoDeconv model, estimation of cell-type abundances is first estimated for each gene and each sample separately. Then the final estimates for each sample are derived by aggregating the gene-specific estimates. In the following, we first describe the IsoDeconv model for a single gene.

Consider a hypothetical gene composed of m non-overlapping exons. These m exons are utilized by I isoforms, or distinct mRNA transcripts formed by unique combinations of these exons. As specified in the assumed gene- and isoform-construction models, the locations of these exons within the gene are known as are the identities and compositions of all isoforms used by this gene. In order to model isoform expression and cell-type abundance, IsoDeconv examines read counts at the exon-set level. We define the read count at any exon set e as the number of fragments which overlap each of the exons in e and only these exons.

To visualize the following setup, consider the hypothetical gene displayed in figure (3.1). This gene is composed of $m = 4$ exons utilized by $I = 3$ different isoforms. Suppose that isoforms 1, 2, and 3 compose the set of all isoforms used by this gene and that their structure with respect to the exons is as given in the figure. Consider the exon set $e := \{1, 2, 3\}$. The read count at e is defined as the number of RNA-Seq reads which, when mapped, overlap exons 1, 2, and 3 but do not overlap exon 4.

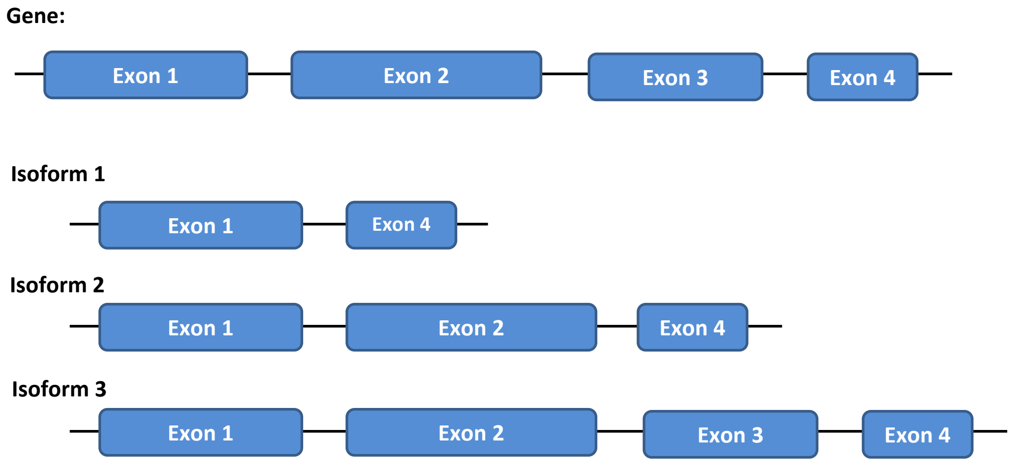


Figure 3.1: Hypothetical gene and isoform construction model.

The IsoDeconv model posits three main factors influencing the observed read count at an exon set e within purified reference sample j of cell type k: the read depth of the RNA-seq experiment (t_{kj}), the length of the exon-set feature e within each utilized isoform i (x_{ei}), and the expression levels of the isoforms used by cells of type k $(\gamma_k = (\gamma_{k1}, ..., \gamma_{kI})')$. As discussed in Chapter 2, RNA-seq expression is commonly corrected for read-depth and feature length. Previously, however, the notion of feature length pertained to the length of the genes or isoforms being measured and not to the lengths of exon sets. Sun et al. [17] extend the definition of feature length for exon-sets and name it as effective length of an exon set. It is calculated as the expected number of starting locations for an RNA-Seq fragment from that exon set. If an exon set is not included in an isoform, its effective length in the isoform is set as 0. For detailed information on the computation of the effective lengths of each exon set, please see the supplementary materials of Sun et al [17].

The negative binomial variant of IsoDeconv models the read count at an exon set e for pure sample j of cell type k, denoted by Y_{kje}, in the following way:

$$Y_{kje} \sim \texttt{NB}\left(\mu = t_{kj} u_{ke}, \phi_k\right) \text{ where } u_{ke} = \sum_{i=1}^{I} x_{ei}\gamma_{ki} = x_e^T \gamma_k$$

and ϕ_k is a cell-type specific overdispersion parameter. Given the isoform activity levels γ_k, read counts are assumed independent across exon sets within a single gene, across genes within a single sample, and across different samples.

This model is extended to the heterogeneous cell type mixture by introducing cell type specific abundance parameters. Let Y^*_{ke} be the unobserved read count at exon set e attributable to cells of type k within the mixture, Z_e be the observed read count of exon set e in the mixture, t_m be the read-depth of the mixture, and p_k be the proportion of expression attributable to cells of type k. Then:

$$Z_e = \sum_{k=1}^{K} Y^*_{ke}$$

$$Y^*_{ke} \sim \texttt{NB}\left(\mu = t_m p_k u_{ke}, \phi_k\right)$$

Assuming that the expression of cells of type k in the mixture is independent of the expressions across all other cell types, Z_e is a convolution of independent, negative-binomially distributed random variables with differing means and overdispersions.

3.2.2.1 Model Fit Algorithm

Model parameters are estimated through the maximum likelihood framework on a gene-by-gene basis. Thus, an estimate of p_k is obtained independently for each gene and aggregated across genes to provide a final estimate of cell type abundance. Within a single gene, optimization proceeds via block coordinate ascent. The steps are as follows:

(1) Assume the γ_k and ϕ_k are fixed, update p_k.

(2) Assume the p_k are fixed, update γ_k and ϕ_k.

Steps (1) and (2) are cycled until convergence of the cell type proportions p_k. Within step (1), optimization of the p_k proceeds using gradient-free LBFGS. A gradient is not supplied to avoid the intractability of the likelihood and its gradient in the case of convolved negative binomials. Optimization in step (2) proceeds using an EM algorithm described in the following section. This algorithm relies upon use of LBFGS methodology to optimize the γ_k and ϕ_k.

3.2.2.2 EM Algorithm

Optimization in the setting of a convolution of negative binomial random variables is a challenging problem. Within step (2), an EM algorithm was developed to transform the negative binomial likelihoods into hierarchical gamma-poisson mixtures to simplify the likelihood. Define λ_{kje} to be the unobserved mean read count at exon set A in pure sample j of cell type k and λ^*_{ke} to be the same quantity for the mixture sample. We can now recharacterize our model as:

$$\begin{aligned} Y_{kje}\big|\lambda_{kje} &\sim \texttt{Pois}(\lambda_{kje}) \\ \lambda_{kje} &\sim \texttt{Gamma}\left(\nu = \phi_k^{-1}, \mu = t_{kj}u_{ke}\right) \end{aligned}$$

and

$$\begin{aligned} Y^*_{ke}\big|\lambda^*_{ke} &\sim \texttt{Pois}\left(\lambda^*_{kje}\right) \\ \lambda^*_{ke} &\sim \texttt{Gamma}\left(\nu = \phi_k^{-1}, \mu = t_m p_k u_{ke}\right) \end{aligned}$$

The transformation of this problem into gamma-poisson mixtures allows separation of the complete-data log-likelihood into K components, one for each cell type, which can be optimized in parallel when cell type proportions are fixed.

To see this, consider the complete data log-likelihood (ℓ) given below, where: $\ell^{(j)}_{ke}$ is the complete data log-likelihood for pure sample j of cell type k at exon set e; $\ell^{(m)}_{e}$ is the complete data log-likelihood for the mixture sample at exon set e; $\ell_{C|D}$ represents the log-likelihood for a random variable R given another random variable D; $f(\cdot)$ is the density function of the specified random variable; and $\boldsymbol{\lambda}_{\boldsymbol{e}}^{*}$ is a $K \times 1$ random vector given by $[\lambda_{1e}^{*}, ..., \lambda_{Ke}^{*}]$.

$$\begin{aligned}
\ell &= \sum_{e=1}^{E} \left\{ \sum_{k=1}^{K} \left(\sum_{j=1}^{n_k} \ell_{ke}^{(j)} \right) + \ell_{e}^{(m)} \right\} \\
&= \sum_{e=1}^{E} \left\{ \sum_{k=1}^{K} \left[\left(\sum_{j=1}^{n_k} \ell_{Y_{kje}|\lambda_{kje}} + \ell_{\lambda_{kje}} \right) \right] + \log \left(f\left(Z_e | \boldsymbol{\lambda}_{\boldsymbol{e}}^{*} \right) \prod_{k=1}^{K} f(\lambda_{ke}^{*}) \right) \right\} \\
&= \sum_{e=1}^{E} \left\{ \sum_{k=1}^{K} \left[\left(\sum_{j=1}^{n_k} \ell_{Y_{kje}|\lambda_{kje}} + \ell_{\lambda_{kje}} \right) + \ell_{\lambda_{ke}^{*}} \right] + \ell_{Z_e|\boldsymbol{\lambda}_{\boldsymbol{e}}^{*}} \right\}
\end{aligned}$$

Under the gamma-poisson framework, Z_e is now a Poisson random variable conditional upon the missing data $\boldsymbol{\lambda}_{\boldsymbol{e}}^{*}$ since it is the sum of K Poisson variates. Thus, the complete data log-likelihood within the mixture sample becomes the sum of K independent gamma log-likelihoods $\left(\ell_{\lambda_{ke}^{*}}\right)$ for each λ_{ke}^{*} and a Poisson log-likelihood $\left(\ell_{Z_e|\boldsymbol{\lambda}_{\boldsymbol{e}}^{*}}\right)$ where the mean is given by $\sum_k \lambda_{ke}^{*}$. As a Poisson random variable, note that $\left(\ell_{Z_e|\boldsymbol{\lambda}_{\boldsymbol{e}}^{*}}\right)$ only depends upon the sum of the λ_{ke}^{*} but does not depend on the parameters γ_k or ϕ_k for any k. Thus, during optimization, this term may be discarded. For a similar reason, the Poisson likelihoods from the pure samples may be discarded as well.

According to the EM algorithm, parameter updates proceed by maximizing a Q-function given by $Q(\gamma_1, ..., \gamma_K, \phi_1, ..., \phi_K) = E\left[\ell \middle| Y_{11}, ..., Y_{KE}, Z_1, ..., Z_E\right]$ where this expectation is computed with respect to the unobserved λ_{kje} and λ_{ke}^{*}. Thus, four posterior means found in the remaining gamma log-likelihood terms are necessary for updates in the expectation step of the EM algorithm proposed above, namely $\lambda_{kje}, \lambda_{ke}^{*}, \log(\lambda_{kje})$ and $\log(\lambda_{ke}^{*})$. For the missing parameters λ_{kje}, it is possible to define closed form posterior expectations. For the λ^*_{ke}, no such closed forms are possible.

$$\begin{aligned}
E[\lambda_{kje} \big| Y_{kje}] &= \frac{\nu_k + Y_{kje}}{\nu_k/\mu_{kje} + 1} \\
E[\ln(\lambda_{kje}) \big| Y_{kje}] &= -\ln(\nu_k/\mu_{kje} + 1) + \Phi(\nu_k + Y_{kje}) \\
E[\lambda^*_{ke} \big| Z_e] &= E\left[\frac{\nu_k + Y^*_{ke}}{\nu_k/\mu^*_{ke} + 1}\middle| Z_e\right] \\
E[\ln(\lambda^*_{ke}) \big| Z_e] &= E\left[-\ln(\nu_k/\mu^*_{ke} + 1) + \Phi(\nu_k + Y^*_{ke})\middle| Z_e\right]
\end{aligned}$$

The latter posterior means for the heterogeneous cell type sample can be determined numerically since Y^*_{ke} only has mass on the set $\{0, 1, ..., Z_e\}$. A finite-summation, numerical approximation to the distribution of convolved negative binomials is utilized to compute the necessary conditional probabilities[58].

3.2.2.3 Simulation Study

The fit of the negative binomial variant of the IsoDeconv model was first assessed using 9 *in silico* mixtures of RNA-seq expression experiments from two cell lines, GM12878 and HSMM. GM12878 is a non-cancerous blood cell line and HSMM is a non-cancerous human skeletal muscle myoblasts dataset, both derived from human tissues. Expression experiments for these cell lines can be found at the Encyclopedia of DNA Elements online database (ENCODE)[59]. The fit was characterized in the low-sample size setting where only a single reference sample is available for estimation in each cell type.

To generate a set of simulated mixture and reference samples, two paired-end RNA-seq experiments each of GM12878 and HSMM were downloaded from the ENCODE database. From each cell type, read pairs overlapping chromosome 1 from a single replicate were randomly sampled to obtain files from 1 million to 9 million reads each. These downsampled files are merged *in silico* to create mixture files of 10 million reads each. Thus, to create a mixture sample of 30% GM12878 and 70% HSMM, 3 million reads are downsampled from the GM12878 experiment and 7 million from HSMM. Reference samples were generated in a similar manner using the remaining replicate for each cell type.

Composition	$\hat{P}_G$	**No. Clusters**	**Restricted**
		$(\phi_{max} \leq 1.5)$	$\hat{P}_G$
GM - 10 / HS - 90	0.395	9	0.137
GM - 20 / HS - 80	0.456	13	0.221
GM - 30 / HS - 70	0.551	11	0.287
GM - 40 / HS - 60	0.583	10	0.475
GM - 50 / HS - 50	0.652	9	0.419
GM - 60 / HS - 40	0.708	7	0.445
GM - 70 / HS - 30	0.750	7	0.687
GM - 80 / HS - 20	0.812	7	0.788
GM - 90 / HS - 10	0.880	6	0.813

Table 3.1: Examining estimation quality under biological replicate mixture generation. Composition defines the structure of the simulated mixture (i.e. GM-10/HS-90 refers to a sample that is 10% GM12878 and 90% HSMM) and $\hat{P}_G$ represents the estimated proportion of GM12878. The final two columns discuss model estimation characteristics when the estimating gene set is restricted to genes where maximum overdispersion between GM12878 and HSMM is limited, namely the number of such genes available and the proportion estimate across these genes.

Reducing the estimating set to approximately 300 genes identified by Cufflinks as differentially expressed or spliced, the model was fit to nine different simulated mixtures. Examining table (3.1), we note that the estimation of cell type abundance was poor when the entire estimating set was used. However, it was also noted that the estimated overdispersions for each cell type (ϕ_k) were quite large for the majority of genes. In fact, for most samples, fewer than 10 of the 300 genes had a maximum estimated overdispersion across GM12878 and HSMM cell types of less than 1.5. However, by restricting to genes with lower levels of estimated overdispersion, estimates were seen to improve (Table 3.1).

These discoveries led to the hypothesis that biological variability across samples was impacting the results of the IsoDeconvNB model fit. It was supposed that this extra variability was entering the model through use of different biological replicates to construct the mixtures and reference samples. In essence, it was supposed that the cell type specific isoform abundances and gene expressions varied across biological replicates of a cell type. Thus, the provided reference samples differed too much from the samples used to generate the mixture tissue. This discrepancy between the mixture cell types and the reference cell types resulted in poor model fit and inflated the observed variances in the exon-set level counts.

To test this hypothesis, we utilized the GM12878 cell line and a non-cancerous human mammary epithelial cell line (HMEC) used for early model validation. A single replicate each of GM12878 and HMEC were selected for examination. While true technical replicates were not available from ENCODE, pseudo-technical replicates were generated by randomly splitting the given files in half, generating two non-overlapping sets of reads from each cell type and sample. Mixture files were then produced by downsampling reads from one of these pseudo-technical replicates and combining them in the manner discussed previously. The remaining pseudo-technical replicate was used as the cell type reference. Due to the limited read count available in each file, the range of proportions used for generating mixture datasets is restricted.

In addition to the use of technical replicates for cell type abundance estimation, three different methods for restricting the estimating gene set were used. The first restriction limits consideration to genes noted by Cufflinks as having gene-level or isoform-level differential expression. The second restriction considers only genes with noted isoform level differential expression. Finally the last estimating set of genes is restricted to all genes under the first restriction which have estimated maximum overdispersions less than 1.5. These restrictions were designed to assess the impact of different estimating sets on the quality of abundance estimates and to examine the impact of technical replicate use on observed overdispersion.

Immediately, it is seen from table (3.2) that the quality of estimation improves dramatically over the biological replicate simulations examined in table (3.1). All estimates hover within 5% of the truth regardless of the estimating gene set used. Restriction sets (1) and (2) estimate similarly across all samples, suggesting that the use of genes with gene-level expression differences only is not damaging to IsoDeconvNB's abundance estimates. Additionally, examining restriction set (3), the number of available genes with estimated overdispersions less than 1.5 increased by 3- to 4-fold

	Gene or Isoform Diff.		**Isoform Diff. Only**		**Overdispersion Restriction**	
Composition	*# Genes*	$\hat{P}_G$	*# Genes*	$\hat{P}_G$	*# Genes*	$\hat{P}_G$
GM - 40 / HM - 60	354	0.398	158	0.396	42	0.403
GM - 50 / HM - 50	354	0.481	158	0.481	41	0.491
GM - 60 / HM - 40	354	0.580	158	0.581	38	0.598
GM - 70 / HM - 30	354	0.664	158	0.665	42	0.678
GM - 80 / HM - 20	354	0.760	158	0.761	32	0.801
GM - 90 / HM - 10	354	0.859	158	0.871	35	0.910

Table 3.2: Examining estimation quality under pseudo-technical replicate file generation. Results are presented for 3 different estimating gene sets: genes with gene or isoform expression differences, genes with isoform expression differences only, and genes with maximum overdispersions less than 1.5. # genes specifies the number of genes in the estimating set and $\hat{P}_G$ details the estimation of the proportion of GM12878 across these genes.

over the samples seen in table (3.1). These results suggest that the additional variability observed in previous simulations was likely due to the use of biological replicates instead of technical replicates in the creation of mixture and reference samples. However, due to to the limitations of the number of quality replicates for the HMEC cell line, a direct assessment of the impact of technical replicates compared to biological replicates is not possible in these cell lines.

Thus, a final set of RNA-seq samples were downloaded from ENCODE consisting of two RNA-seq read experiments from biological replicates of CD20+ monocytes and two biological replicates of GM12878 experiments. Using these experiments, two sets of reference and mixture files were generated. In the first set, the pseudo-technical replicate approach described previously was used to generate references and mixtures. In the second set, the biological replicate approach described previously was used to generate references and mixtures.

Table (3.3) reinforces the idea that the quality of estimation drops dramatically when independent biological replicates are used to produce the mixture and reference samples. While the the number of genes with maximum overdispersion less than 1.5 does not differ much between these technical and biological replicates, each cell type displays fewer genes with overdispersion values less than 1.5 when comparing biological to technical replicates across the majority of the simulations. We do note a slight discrepancy in this trend for the CD20 cell type at low proportions

	Technical Replicates				**Biological Replicates**			
Composition	ϕ_G	ϕ_C	$\max(\phi)$	$\hat{P}_G$	ϕ_G	ϕ_C	$\max(\phi)$	$\hat{P}_G$
GM - 30 / CDT - 70	41	85	13	0.274	33	58	19	0.232
GM - 40 / CDT - 60	42	69	11	0.371	40	54	18	0.210
GM - 50 / CDT - 50	50	55	19	0.452	42	51	19	0.357
GM - 60 / CDT - 40	50	46	17	0.582	50	45	21	0.380
GM - 70 / CDT - 30	69	41	15	0.658	51	40	15	0.598
GM - 80 / CDT - 20	83	22	14	0.783	63	37	13	0.710
GM - 90 / CDT - 10	108	21	14	0.895	84	30	13	0.760

Table 3.3: Examining change in estimation quality when using the various replicate strategies. ϕ_G and ϕ_C summarize the number of genes for each cell type where the overdispersion is less than 1.5. $\max(\phi)$ refers to the number of such genes where the maximum overdispersion is less than 1.5. $\hat{P}_G$ considers estimation quality within the set of genes with limited maximum overdispersion (≤ 1.5) to ensure comparability across replicate settings and previous experiments.

of CD20 in the sample. However, this is matched by a 33% and 25% increase in the number of such genes when comparing GM12878 technical to biological replicates.

3.2.2.4 Discussion

The preceding simulations suggest that the negative binomial variant of the IsoDeconv model is questionable for cell type deconvolution in the RNA-seq data setting. It has been demonstrated that introducing biological variability between the mixture and reference datasets results in poor estimation quality and a reduction in the number of minimally overdispersed genes. Rather than expanding the estimation set to consider more genes with sufficiently low overdispersion, the decision was made to address the extra biological variability observed by restructuring the cell type abundance model. In particular, an alternative variant of the IsoDeconv model would need to explicitly account for the biological variation in expression through a probabilistic mechanism for sample-specific gene and isoform expression parameters to improve estimation. Ultimately, this necessitated a switch to a multinomial structure for the distribution of read counts within genes with Dirichlet distributions over gene and isoform expression parameters. It also requires multiple purified RNA-seq reference experiments for each cell type k to capture both the mean and variance of these parameters across independent experiments. We detail this model in the following sections.

3.2.3 IsoDeconv - Multinomial Model with Dirichlet Penalty (IsoDeconvMM)

Pure Sample Expressions		
Value	**Dim.**	**Description**
$Y_{kj(E)}$	1×1	Total read count outside gene of interest in pure sample j of cell type k.
Y_{kjA}	1×1	Read count at exon set A in pure sample j of cell type k.
Y_{kj}	$E \times 1$	Collection of read counts across all exon sets in the given gene for pure sample j of cell type k.
γ_{kj}	$I \times 1$	Isoform expression parameters unique to pure sample j of cell type k.
τ_{kj}	1×1	Probability that a randomly selected read maps to the gene of interest in pure sample j of cell type k.
t_{kj}	1×1	The total read count in pure sample j of cell type k.
Mixture Sample Expressions		
Value	**Dim.**	**Description**
Z_A	1×1	Read count at exon set A in the mixture cell type sample.
Z	$E \times 1$	Collection of Z_A in a single vector.
$Z.$	1×1	Total number of reads mapping to gene of interest in the mixture $\left(Z. = \mathbf{1}^T Z = \sum_{e=1}^{E} Z_A\right)$.
Z_{kA^*}	1×1	Read count at exon set A in the mixture cell type sample attributable to cells of type k.
γ_k^*	$I \times 1$	Isoform expression parameters unique to cells of type k found within the mixture cell type sample.
τ_k^*	1×1	The probability that a randomly selected read from cells of type k in the mixture sample maps to the gene of interest which is unique to the cells in the mixture sample.
Cell-Type Specific and Cluster Level Parameters		
Value	**Dim.**	**Description**
X	$E \times I$	Matrix of effective lengths for each exon set within each of the isoforms.
X_{ij}	1×1	Effective length of gene i in isoform j.
$\tilde{l}$	$I \times 1$	Vector of complete effective lengths of each utilized isoform $\left(\tilde{l}_j = \sum_{i=1}^{E} X_{ij}\right)$.
p_k	1×1	Proportion of cell type k present in the mixture tissue.
p	$K \times 1$	Collection of abundances for each of the K cell types which compose the mixture.
α_k	$I \times 1$	Hyperparameters governing average isoform expression levels and variances within cells of type k.
β_k	2×1	Hyperparameters governing gene expression levels within cells of type k.
Value	**Dim.**	**Description**
$\circ$	NA	This operator indicates element-wise multiplication of two vectors.

Table 3.4: Notation for defining the IsoDeconv Model.

In order to specify the multinomial variant of the IsoDeconv model (IsoDeconvMM), revised definitions of several parameters must be specified and additional model parameters governing gene and isoform expression must be introduced. As before, the model specification will pertain to a single gene only. Each gene will be modeled independently and the per-gene cell type abundance estimates will be aggregated afterwards.

All parameters for IsoDeconvMM have been described in Table (3.4). Several notes are necessary to clarify the new meanings of these parameters. Firstly, the γ_{kj} values represent the isoform expression quantities for a single sample. These expressions are interpreted as per-unit-of-effective-length conditional probabilities that a read maps to isoform i given that it maps to the gene which utilizes isoform i. Secondly, the gene expression parameters τ are not normalized in a manner that allows for comparison across genes (e.g. FPKM). These parameters are raw probabilities that a randomly selected read pair, not a randomly selected transcript, maps to the gene of interest. The addition of the subscripts kj allows us to capture the biological variation across samples.

Using this notation, the cell type abundance model within purified reference samples can be detailed as follows:

$$\begin{aligned} \tau_{kj} &\sim \text{Beta}(\beta_k) \\ \tilde{l} \circ \gamma_{kj} &\sim \text{Dirichlet}(\alpha_k) \\ \begin{bmatrix} Y_{kj(E)} \\ Y_{kj} \end{bmatrix} \Bigg| \tau_{kj}, \gamma_{kj} &\sim \text{Multinomial} \left(t_{kj}, \begin{bmatrix} 1 - \tau_{kj} \\ \tau_{kj} X \gamma_{kj} \end{bmatrix} \right) \end{aligned} \tag{3.1}$$

where $\circ$ in $\tilde{l} \circ \gamma_{kj}$ denotes element-by-element product. In the mixture tissue, then, the cell type abundance model is given by:

$$\begin{aligned} \tau_k^* &\sim \text{Beta}(\beta_k) \\ \tilde{l} \circ \gamma_k^* &\sim \text{Dirichlet}(\alpha_k) \\ \begin{bmatrix} Z \end{bmatrix} \Big| \tau_k^*, \gamma_k^* &\sim \text{Multinomial} \left(Z_{\cdot}, \begin{bmatrix} \frac{\sum_{k=1}^K p_k \tau_k^* X \gamma_k^*}{\sum_{k=1}^K p_k \tau_k^*} \end{bmatrix} \right) \end{aligned} \tag{3.2}$$

Within the IsoDeconvMM model, independence is assumed across samples and across genes within samples.

3.2.3.1 Model Fit Algorithm

Within each gene, the model is fit using a staged estimation approach with three stages. In stage one, the gene and isoform expression parameters are estimated separately for each purified reference sample by maximum likelihood estimation. The likelihood used for stage 1 involves only the multinomial component of equation (3.1). Under such a framework, closed form estimates of τ_{kj} are obvious and a logarithmic adaptive barrier algorithm can be used to obtain estimates of the γ_{kj} subject to boundary constraints. Once obtained for each cell type and sample, these estimates are held fixed for all further stages.

Within stage 2, the estimated values of τ_{kj} and γ_{kj} are treated as observations from the Dirichlet component of equation (3.1). Estimates of α_k and β_k are obtained via maximum likelihood estimation within separate Dirichlet models. Once obtained, these estimates of α_k and β_k are fixed for stage 3.

Finally, in stage three, the α_k and β_k estimates are used in Dirichlet distributions as penalty functions in the estimation of the γ_k^*, τ_k^*, and p_k. In this way, we regularize estimates of γ_k^* and τ_k^* to be like those estimates obtained in the pure cell type samples. Use of an EM algorithm allows separation of the full likelihood into $K+1$ independent components. The first K components pertain to the isoform expression parameters from each of the K cell types. Each of these components is optimized using a Newton-Raphson algorithm on the $\log(\gamma_k^*)$ until convergence of isoform parameters. The last component contains information regarding the p_k and $\log(\tau_k^*)$ values, which are optimized using a quasi-Newton's method optimization procedure (BFGS). Estimation is seeded at various start points to identify global maxima. The EM algorithm is iterated until convergence in the proportion estimates. Proportion estimates across multiple genes are then aggregated using the spatial median to obtain final estimates of cell type proportions.

3.2.3.2 Explaining Modeling Decisions

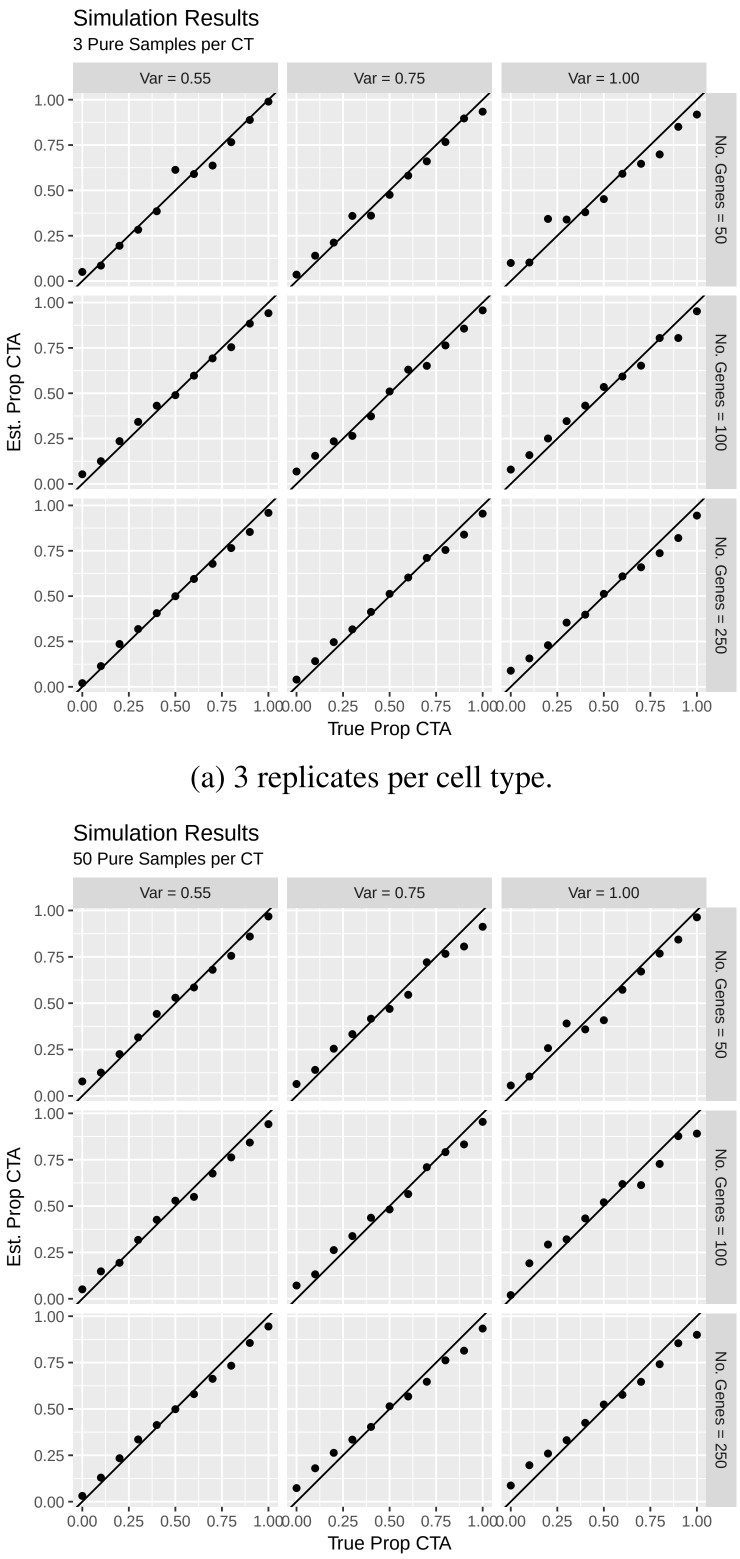
Several facets of the preceding discussion deserve illumination. First, consider the switch from the negative binomial model to the Multinomial-Dirichlet model. In order to incorporate the isoform expression parameters as conditional probabilities, the model within a gene must condition on the number of reads mapped to that gene in the purified reference samples. Supposing this conditioning is performed and that an independent negative binomial distribution is assumed at each exon set, the likelihood becomes inconsistent. This arises because the independent negative binomials could theoretically exceed the read count upon which the model is conditioned. The multinomial model maintains its consistency despite the conditioning argument.

Secondly, the use of the described staged estimation approach became necessary after an initial version of the model, which attempted to estimate α_k and γ_k values simultaneously, proved intractable. This approach led to unstable estimates of the γ_k and α_k parameters wherein the α_k parameters became unbounded. This would suggest little to no variability in the isoform expressions, an impossibility in the simulated data upon which the model was tested.

Finally, the incorporation of the $\log(\tau_k^*)$ and $\log(\gamma_k^*)$ transformations was performed after initial testing with untransformed parameters proved inaccurate. "Hill-climbing" estimation methods such as Newton Raphson and BFGS require that the likelihood is sufficiently stable across the parameter space so that the crest of the "hill" is not continually overstepped. The proposed optimization approach is more stable with respect to the log parameters since the log scale spreads out the small parameter values. Under these reparametrizations, model accuracy and the mobility of proportion estimates improved.

3.2.3.3 Simulation Study

In order to evaluate the fit properties of the IsoDeconvMM model, a simulation study was conducted for the two cell type case. To ensure that simulated gene- and isoform-construction models are sufficiently complex, the models estimated from the GM12878 and HMEC data are utilized as the "true" constructions from which to simulate. These simulations seek to capture the influence of three factors on model fit: variability in isoform expression across subjects, the number of purified reference samples per cell type, and the number of genes utilized for estimation.



(a) 3 replicates per cell type.

(b) 50 replicates per cell type

Figure 3.2: (a) Simulation Results for 3 replicates per cell type (b) Simulation Results for 50 replicates per cell type.

Simulations are conducted as follows. Suppose that there are 3 purified references per cell type and that 100 genes will be utilized for estimating cell type proportions. Across these 100 genes, gene expressions will be simulated from a normal distribution with a mean of 130 reads and a standard deviation of 33. Gene- and isoform expressions will be simulated as described in the supplementary materials. Of note, the variabilities in isoform expression are set to ensure that 90% of observations fall within $X\%$ of the cell-type average where X is allowed to vary from 55 to 100%. To simulate cell-type differences in average isoform expression profiles, average isoform expression levels are permuted to ensure that there is no overlap in the top 2 or 3 isoforms used by each cell type.

The results of these simulations are displayed in Figure (3.2). Here we see that IsoDeconvMM provides strong results across the range of parameters tested. In addition, the results appear consistent with expectations regarding the effects of larger gene set size and increased isoform expression variability. As one incorporates more genes into the estimation set, the accuracy and stability of the estimator improves. As one increases the variability in isoform expression, the accuracy and stability of the estimates decrease.

We do note that there is decrease in the accuracy of the model when one moves from 3 references per cell type to 50. This counterintuitive result is likely due to the following. Consider the 3 reference sample case. When one draws these 3 samples very little of the support of the underlying reference distribution is interrogated. Thus, while drawing three subjects may increase the variability across repeated sampling, the probability of drawing three highly similar subjects is greater than the probability of drawing 50 highly similar subjects. In this way, the model's perception of variance may decrease for genes in the 3 reference sample setting causing stronger penalties than those in the 50 reference sample case.

3.2.3.4 Discussion

We have presented two formulations for the IsoDeconv model for cell-type abundance estimation utilizing isoform expression information from RNA-seq data. IsoDeconvNB, the first variant, was a direct extension of the IsoDot framework proposed by Sun et al [17]. Extensive study of simulated mixtures generated from genuine RNA-seq read experiments found that estimation was sound only in the limited variability setting, as was the case when references and mixtures were generated from non-overlapping samples from a single read experiment. The introduction of additional variability due to the use of biological replicates saw diminished performance and an increased perception of variability in exon-set level counts. Due to the limited availability of RNA-seq experiments performed on purified samples of various tissues, it was decided to extend the IsoDeconv model to explicitly incorporate biological variation.

The second variant of the IsoDeconv model, IsoDeconvMM, involved a major restructuring to allow for subject-specific gene and isoform expressions. This restructuring necessitated a shift to a multinomial framework to maintain coherence in the proposed likelihood. The behavior of IsoDeconvMM was explored via simulation study. IsoDeconvMM was found to provide quality estimates of cell type proportions via simulation across a range of simulated parameters. The impact of increased isoform expression variability and diminished gene set size were consistent with expectations. Increasing the number of purified reference samples resulted in a counterintuitive result wherein increasing sample size saw decreasing stability in the proportion estimates.

Future research regarding IsoDeconvNB and IsoDeconvMM should focus on two primary avenues: decreasing the computational complexity and refining their application to real data. Both IsoDeconvNB and IsoDeconvMM are computationally complex algorithms. In the case of IsoDeconvNB, the convolution of negative binomial models with differing means and variances is a particular challenge. In this setting, we utilized numerical approximations techniques proposed by [58]. Such a technique introduces approximation error as well as high computational cost due to the need to compute large sums for K different cell types with each update. The multistage modeling approach used by IsoDeconvMM and the need for multiple estimation start points creates a slow down in model optimization. Future study will need to address these complexities to improve the behavior and time to solution for the model in each gene.

In addition, both IsoDeconvNB and IsoDeconvMM have shown limited utility in the real data setting. IsoDeconvNB's behavior has been thoroughly catalogued in the simulated mixture profiles developed from real data. IsoDeconvMM was also applied to such a dataset composed of CD8+ T-cell and CD4+ T-cell mixtures from real data with 3 pure sample references per cell type[60]. Results were highly unstable with the majority of genes optimizing at proportions of 0% CD8+ T-cells or 100%. Exploration to uncover the cause of this behavior is needed. It is suspected that this behavior could result from the limitations of the current pipeline to identify genes with differential expression of isoforms in the low replicate setting when two highly similar cell types (e.g. CD4+ and CD8+ cells) are considered. Without reliable information regarding which genes experience differential isoform expression, the model cannot be expected to perform well.

Finally, future research should examine simplifying the IsoDeconv framework to account for highly similar isoforms. Consider the design matrix X for any gene. Isoforms which differ by the removal of a single exon are likely to induce high correlation in X. This instability in the X matrix can cause unreliable isoform expression estimation which may influence model estimates. To this end, a consideration of the IsoDeconv framework which could model groups of highly similar isoforms or even utilize some form of dimension reduction (e.g. PCA) could improve model fit and time to solution.

Please see Appendix 1 for the mathematical supplement for this chapter. This supplementary material contains additional information regarding model optimization, the mathematical foundations behind IsoDeconvNB and IsoDeconvMM, and additional information regarding the simulation structure for IsoDeconvMM.

CHAPTER 4: ICED-T PROVIDES ACCURATE ESTIMATES OF IMMUNE CELL ABUNDANCE IN TUMOR SAMPLES BY ALLOWING FOR ABERRANT GENE EXPRESSION PATTERNS

4.1 Introduction

The evolving relationship between a cancer and its host's immune system is well summarized by a hypothesis known as immunoediting. Immunoediting stresses that the immune system not only suppresses tumor cells, but also shapes tumor immunogenicity in ways that may promote tumor growth [32, 61]. For example, consider the relationship between tumors and tumor-infiltrating T cells. Infiltrating T cells can be cytotoxic, contributing to death of cancer cell populations. However, these T cells express immune checkpoints which inhibit their function; such checkpoints prevent the immune system from indiscriminately attacking healthy host cells. Under selective pressure from the immune system, cancers can evolve defense mechanisms which activate immune checkpoints and thereby limit the anti-tumor activity of the infiltrating T cells.

Early strategies in immunotherapy were developed based on the insights of immunoediting [35]. Among the best known immunotherapy strategies, immune checkpoint inhibitors block immune inhibition pathways that restrict effective anti-tumor T cell responses [62]. Checkpoint inhibitors have achieved phenomenal successes in a fraction of cancer patients, exhibiting response rates around 40% and 20% for melanoma and lung cancer, respectively [63]. It is of great clinical interest to identify the subset of cancer patients who may respond to checkpoint inhibitors. Use of tumor-infiltrating immune cells to predict clinical response to therapy has shown promising results. Previous studies have shown that the patients with CD8+ T cells around tumor cells have higher response rate to checkpoint inhibitors [64]. In addition to benefiting development of precision immunotherapies, immune cell composition estimates of tumor samples have also demonstrated prognostic value [38, 39]. Therefore, studying immune cell composition in tumor samples is timely and potentially has high impact on cancer research.

Several groups have studied immune cell composition using gene expression data from bulk tumor samples [40, 41, 65, 25, 43, 44]. These pioneering works have demonstrated promising results, but also bear some limitations. For example, a subset of these works estimate immune cell presence using the expression of few genes [40, 41], or calculate average expression of the genes with cell type-specific expression[42] instead of estimating immune cell composition. As an alternative, several methods have been proposed to estimate immune cell composition using a regression-based approach, with gene expression from bulk tumor samples as the response variable and reference gene expression from purified cell types as covariates. CIBERSORT [25] employs support-vector regression. TIMER [43] uses a linear regression and removes the genes with very high expression due to their strong influence on model fitting. EPIC [44] is the most recent work. It uses weighted linear regression to give the genes with lower expression variation higher weights. These regression-based methods, when applied to tumor expression data, explicitly or implicitly assume that they start with a set of genes that have negligible expression in tumor cells, and that the expression of immune cells are conserved between purified reference samples and tumor samples. These assumptions are questionable as many environmental factors that affect gene expression may differ between tumor and reference samples.

In this paper, we propose a new statistical method for cell type deconvolution entitled ICeD-T, which stands for Immune Cell Deconvolution in Tumor tissues. ICeD-T is an extension of existing regression based methods [25, 43, 44] with two major novel features designed to overcome the limitations of these methods.

First, ICeD-T employs a likelihood based framework, which assumes that gene expression follows a log-normal distribution. Previous work has shown that deconvolution should be performed on linear-scale instead of log-scale of gene expression data since linear-scale mixing of gene expression better captures the biological realities of cell mixing in a bulk tissue sample [66]. However, since gene expression variation increases with expression level, genes with higher expression may become outliers with great influence on linear scale deconvolution models. Therefore one may need to remove genes with high expression for robust deconvolution analysis [43]. The log transformation, often used in expression studies, enjoys variance-stabilizing and skew-mitigation properties which counteract this relationship in expression data [67, 68]. ICeD-T is able to perform gene expression deconvolution on the linear-scale while simultaneously incorporating the beneficial properties of the log-transformation through our method design and the use of log-normal distribution.

Second, ICeD-T automatically identifies the genes whose expressions in tumor samples are inconsistent with reference profiles due to altered immune cell behavior in the mixture or unexpected tumor cell expression. Within its estimation algorithm, ICeD-T down-weights the contribution of such genes in cell type abundance estimation using a mixture model that separates all the genes into two groups: an "aberrant" group and a "consistent" group.

4.2 Statistical Methods

4.2.1 The Input Data

While ICeD-T can be applied on microarray data, we focus mainly on RNA-seq data as it is more popular now and in the foreseeable future. We assume that RNA-seq data from bulk tumor samples are available for n independent subjects. Gene expression from purified samples may be pre-computed or processed from raw RNA-seq data of multiple replicates for each cell type. Across reference expression profiles and bulk samples, the RNA-seq measurements of gene expression are appropriately normalized in a consistent manner using FPKM, FPKM-UQ, or TPM. More specifically, to calculate FPKM, we divide gene expression (# of RNA-seq fragments) by total number of mapped fragments (in millions) and the gene length (in kilo bases). FPKM-UQ is a variant of FPKM where sample-specific read-depth is measured by 75 percentile of gene level fragment counts across all genes, instead of the total number of mapped fragments. TPM reverses the order of the two normalization steps. It first divides the gene-level fragment counts by gene length, and then divides it by the summation of gene-length corrected fragment counts across all genes.

Additional information utilized by ICeD-T's deconvolution model includes a pre-selected gene set (ideally, genes with immune-specific expression) and tumor purity, if available. Several such gene sets have been prepared by previous work, such as the gene sets used by CIBERSORT of EPIC [25, 44]. Provision of tumor purity is optional, and it can be computed, for example, using somatic copy number aberration data [69].

4.2.2 Statistical Model

Specification of the ICeD-T model begins with a consideration of expression behavior in purified references samples of constituent cell types. Denote by Z_{jkh} the expression of gene j in the h-th purified sample of cell type k. ICeD-T assumes that the Z_{jkh} follows independent log-normal distributions, given by:

$$\log(Z_{jkh}) \sim \mathcal{N}(\mu_{jk}, \sigma_{jk}^2), \tag{4.1}$$

where

$$E[Z_{jkh}] = \gamma_{jk} = \exp(\mu_{jk} + \sigma_{jk}^2/2), \quad \text{and} \quad V[Z_{jkh}] = \gamma_{jk}^2 \left[\exp(\sigma_{jk}^2) - 1\right]. \tag{4.2}$$

Therefore, the distribution parameters for each cell type's gene expression (e.g., μ_{jk} and σ_{jk}^2) may be estimated by the mean and variance of the log-transformed Z_{jkh} values. Once estimated, these parameters represent expression profiles for each cell type in our deconvolution model. Optionally, ICeD-T accepts previously computed profiles which would replace the γ_{jk} above.

Shift focus to the n bulk tumor samples. Assuming that each sample is composed of K immune cell types and other extraneous cell types, the expression of gene j in bulk tumor sample i - denoted by Y_{ij} - is modeled by

$$Y_{ij} = \sum_{k=1}^{K} \rho_{ik} X_{ijk} + \epsilon_{ij},$$

where X_{ijk} represents the expression of gene j for cells of type k in the i-th sample, and ρ_{ik} is the proportion of expression attributable to cell type k. The residual error ϵ_{ij} represents signals from other cell types (e.g., tumor cells) or random noise. If tumor purity information is provided, $\sum_{k=1}^{K} \rho_{ik} = 1 - \rho_{iT}$, where ρ_{iT} is tumor purity. If tumor purity is not provided, $\sum_{k=1}^{K} \rho_{ik} \leq 1$.

One potential question for the above deconvolution model is: if we only consider genes expressed in various immune cells, and assume these genes are not expressed in other cell types (e.g., tumor cells), shouldn't $\sum_{k=1}^{K} \rho_{ik} = 1$? This is not true because gene expressions were normalized by FPKM or TPM using genome-wide gene expression data. Therefore, the expression of a immune-specific gene is affected by the expression of other genes. For example, if tumor purity is high, then the expression of some other genes that are expressed in tumor are high, and thus after FPKM or TPM normalization, the expression of those immune genes are relatively lower, which will lead to smaller values of ρ_{ik} estimates, hence reflecting the fact of higher tumor purity.

We begin to develop the probabilistic framework utilized by ICeD-T to model the relationship posited above by first assuming that there are no aberrant genes (i.e. gene expression of each cell type in reference samples is consistent with gene expression in tumor microenvironment). Under such an assumption, X_{ijk} has the same distribution as the Z_{jkh} for any i, h, and j (i.e. $X_{ijk} \sim Z_{jkh}$). The summation of independent log-normal random variables does not have a closed form distribution function. To address this issue, ICeD-T approximates the distribution of Y_{ij} using another log-normal:

$$\log(Y_{ij}) \sim \mathcal{N}\left(\tilde{\mu}_{ijC}, \Delta_j \sigma_{iC}^2\right), \text{ where } \tilde{\mu}_{ijC} = \log\left(\sum_{k=1}^{K} \rho_{ik} \gamma_{jk}\right) - \Delta_j \sigma_{iC}^2, \tag{4.3}$$

and Δ_j is the weight for the j-th gene.

The approximation used above is based upon the Fenton-Wilkinson approach which states that the summation of log-normals can be approximated by another log-normal whose parameters are obtained via moment-matching [70]. Under a strict Fenton-Wilkinson approach, the distribution of Y_{ij} would be given by:

$$\log\left(Y_{ij}\right) \sim \mathcal{N}\left(\tilde{\mu}_{ijC}, \sigma_{\tilde{ijC}}\right)$$

where

$$\begin{aligned}
\tilde{\mu}_{ijC} &= \log\left(\sum_{k=1}^{K} \rho_{ik}\gamma_{jk}\right) - \tilde{\sigma}^2_{ijC}/2, \\
\tilde{\sigma}^2_{ijC} &= \log\left(\sum_{k=1}^{K} \left(\rho_{ik}\gamma_{jk}\right)^2 \left[\exp\left(\sigma^2_{jk}\right) - 1\right] \Bigg/ \left[\sum_{k=1}^{K} \rho_{ik}\gamma_{jk}\right]^2 + 1\right).
\end{aligned}$$

We replace the variance structure posited by Fenton-Wilkinson with the weighted variance model of equation (4.3) as the weighted model demonstrated improved fit and stability in simulated data.

Regarding the variance weights used by ICeD-T, we implement two different options. One assumes a homogeneous weight for all genes, i.e., $\Delta_j = 1$ for all j. Later we refer to this option as "No Weights". The second option for the weight of each gene is termed maximal variance weights or "Max Var Weights". To define maximal variance weights, let σ_j^{*2} be the maximum expression variance across all cell types k at gene j:

$$\sigma_j^{*2} = \max_k \left(\hat{\sigma}^2_{jk}\right)$$

The weight of a gene j is then specified as follows:

$$\Delta_j = \frac{\sigma_j^{*2}}{\underset{j}{\text{median}}\left[\sigma_j^{*2}\right]}$$

Thus, a gene's weight compares its maximal expression variance to the median of all such maxima across genes. Under this construction, genes with larger variances will have larger variance weights. Larger variance weights ensure that residuals from such genes will have smaller impact on estimation of cell type composition.

The Δ_j specified above require slight modification to improve stability of the model fit. Unadjusted, this procedure can provide some genes with excessively small variance weights and some genes with excessively high variance weights. To control this extreme behavior, the bottom 15% of variance weights are replaced with the 15th percentile variance weight across all genes. Similarly, the top 15% of all variance weights are replaced by the 85th percentile variance weight. In this way, no genes are allowed to become too minimally or maximally important to model fit.

Return to the specification of Y_{ij} in equation (4.3). Now assume that some genes in the dataset are aberrant. For aberrant genes, ICeD-T borrows the expression structure proposed for consistent genes but inflates the variance. Thus, if gene j is aberrant, the expression of Y_{ij} is given by:

$$\log(Y_{ij}) \sim \mathcal{N}\left(\tilde{\mu}_{ijA}, \Delta_j \sigma^2_{iA}\right), \tag{4.4}$$

where

$$\tilde{\mu}_{ijA} = \log\left(\sum_{k=1}^{K} \rho_{ik}\gamma_{jk}\right) - \Delta_j \sigma^2_{iA} \quad \text{and} \quad \sigma^2_{iA} > \sigma^2_{iC}.$$

By allowing aberrant genes to have larger variance, the ICeD-T model flattens the likelihood for such genes, and thus down-weights their contributions to cell type proportion estimates.

Direct use of the likelihoods provided by equations (4.3) and (4.4) within bulk data is impossible since it is unknown whether a gene is consistent or aberrant *a priori*. Thus, ICeD-T must model expression at any gene as a mixture of the log-normal distributions pertaining to consistent and aberrant genes. The mixture likelihood utilized by ICeD-T is found below:

$$Y_{ij} \sim p_i \mathcal{LN}\left(\tilde{\mu}_{ijC}, \Delta_j \sigma^2_{iC}\right) + (1 - p_i)\mathcal{LN}\left(\tilde{\mu}_{ijA}, \Delta_j \sigma^2_{iA}\right),$$

where $\mathcal{LN}$ denotes the density function of a log-normal distribution, and p_i and $1 - p_i$ denotes the proportion of genes being consistent and inconsistent, respectively. This likelihood function can be maximized using an EM algorithm. Missing data necessary for the EM algorithm is introduced in the form of class membership indicators H_{ij}, where $H_{ij} = 0$ or 1 denotes an aberrant or consistent gene, respectively. Thus, the complete data log-likelihood for the i-th bulk tumor sample is given by:

$$\ell_i = \sum_{j=1}^{n_G} H_{ij} \left[\log(p_i) - (1/2)\log(\Delta_j \sigma_{iC}^2) - \left(1/2\Delta_j \sigma_{iC}^2\right)\left(\log(y_{ij}) - \tilde{\mu}_{ijC}\right)^2\right] +$$
$$(1 - H_{ij}) \left[\log(1 - p_i) - (1/2)\log(\Delta_j \sigma_{iA}^2) - \left(1/2\Delta_j \sigma_{iA}^2\right)\left(\log(y_{ij}) - \tilde{\mu}_{ijA}\right)^2\right],$$

where n_G is the number of genes used in our model.

Within each EM step, maximization of Q function with respect to $(\rho_{i1}, ..., \rho_{iK}, \sigma_{iC}^2, \sigma_{iA}^2)$ and p_i are separable. Given the other parameters, the estimate of p_i has a closed form. Given p_i, the remaining parameters are grouped into two blocks: the mixture proportions ρ_{ik}'s (block 1) and the two variance parameters $(\sigma_{iC}^2, \sigma_{iA}^2)$ (block 2), and the parameters of two blocks are iteratively updated. Given the estimates of $(\sigma_{iC}^2, \sigma_{iA}^2)$, the mixture proportions ρ_{ik} are estimated using numerical optimization (the BFGS algorithm) while the constraints are incorporated using the Augmented Lagrangian method (R function `auglag`). Given the estimates of the mixture proportions ρ_{ik}, the two variance terms $(\sigma_{iC}^2, \sigma_{iA}^2)$ are involved in separate pieces of the complete data log-likelihood, and thus can be estimated separately. Given variance weights, each of σ_{iC}^2 and σ_{iA}^2 is estimated by numerical optimization (R function `optimize`). Without variance weights, they can be estimated by closed form. See Appendix B Section 1.5.2 for details of the parameter estimation steps.

The ρ_{ik}'s estimated by any regression based deconvolution approach should be interpreted as the proportion of gene expression contributed by certain cell types. If one seeks to estimate the proportion of cells, these ρ_{ik}'s should be adjusted by cell size factors. We borrow the cell size factors, denoted by s_k, from Racle et al. [44] and construct revised relative abundance of immune cell types by $\rho^*_{ik} = (\rho_{ik}/s_k)/\sum_{i=1}^{K}(\rho_{ik}/s_k)$. Further details are provided in the Supplementary Materials (Section C.2).

4.3 Results

4.3.1 Simulation Study

We conducted a simulation study to evaluate the performance of ICeD-T, CIBERSORT, and EPIC. For each method, we seek to assess the estimation accuracy and the robustness of estimation in the presence of aberrant gene behavior. For ICeD-T only, we also assess its ability to identify aberrant genes.

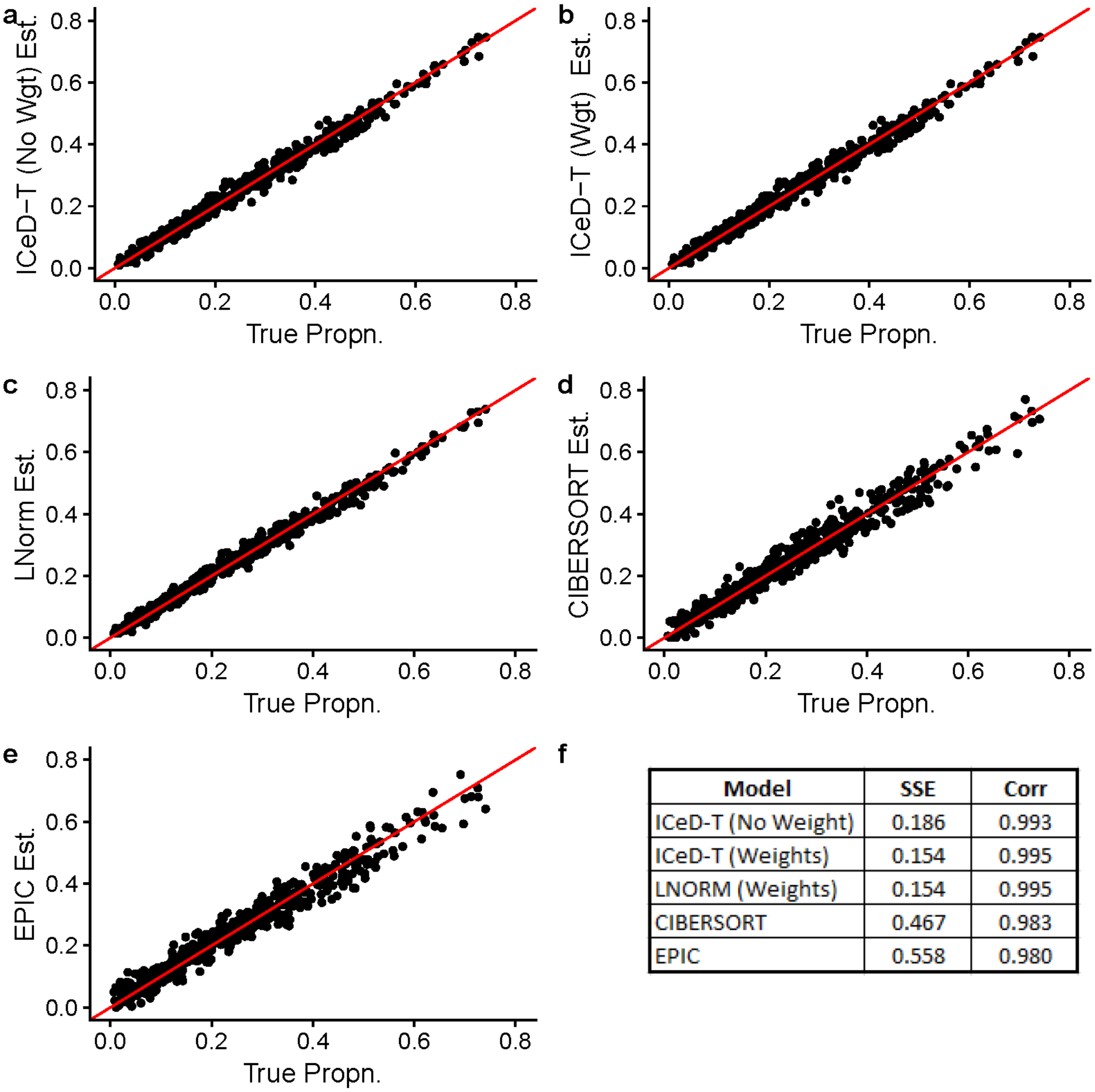
We simulated reference expression of 250 genes for 5 cell types: one tumor cell type and four immune cell types. Our simulations assume that these 250 genes were selected to be expressed in immune cells but not tumor cells. When there are no aberrant genes, the expression of these 250 genes in a bulk tumor sample was simulated by mixing the 4 immune cell types with known proportions. For each gene, we assume it is expressed in one of the four immune cell types and has low/background expression in the other three immune cell types. To better mimic the complexity of real data, we do not assume one homogeneous background expression. Instead, we assume the background expression has a three-tiered scale to reflect lowly, moderately or highly expressed genes (range: 2.0-8.0). Average log-transformed expression for the expressed cell type is simulated from by an up-shift of background expression level (range: 3.5-9.0). See Supplementary Materials Section B.1 for more details. Using RNA-seq expression data from immune cells taken from Linsley et al. [60], a mean-variance relationship was computed from FPKM-UQ normalized data across immune specific genes. The simulated average expression profiles are then mapped to a corresponding variance using this relationship with allowance for random error. Fifteen reference samples were simulated for each cell type from its unique expression profile using a log-normal distribution.

To generate the expression of a bulk tumor sample, a tumor purity value was simulated from a normal distribution (mean=0.60, sd=0.15) and truncated at endpoints of 0.17 and 0.95. The remaining immune cell proportions were then simulated from a Dirichlet distribution with average abundances ranging from 15% to 40%. For each gene in the bulk tumor sample, its expression in each immune cell type was simulated from a log-normal distribution and a weighted summation of these expression values was computed as the expression in the bulk tumor sample. These gene expression profiles are then perturbed to account for aberrant behavior. Zero or approximately twenty percent of genes were randomly selected as aberrant genes. Among them, 25% have down-regulated expression in the highly expressed cell type, 25% have up-regulated expression of the highly expressed cell type, and 50% have expression in tumor cells at a background level. See the Appendix B for further details regarding the construction of these simulations and additional simulation results.

The expression profile of each cell type was estimated from the 15 simulated samples of that cell type. This reference is used for deconvolution in each of the following models: ICeD-T without variance weights, ICeD-T with variance weights, LNORM with variance weights, CIBERSORT (version Jar 1.06), and EPIC. LNORM is a variant of the ICeD-T model which does not consider aberrant gene behavior.

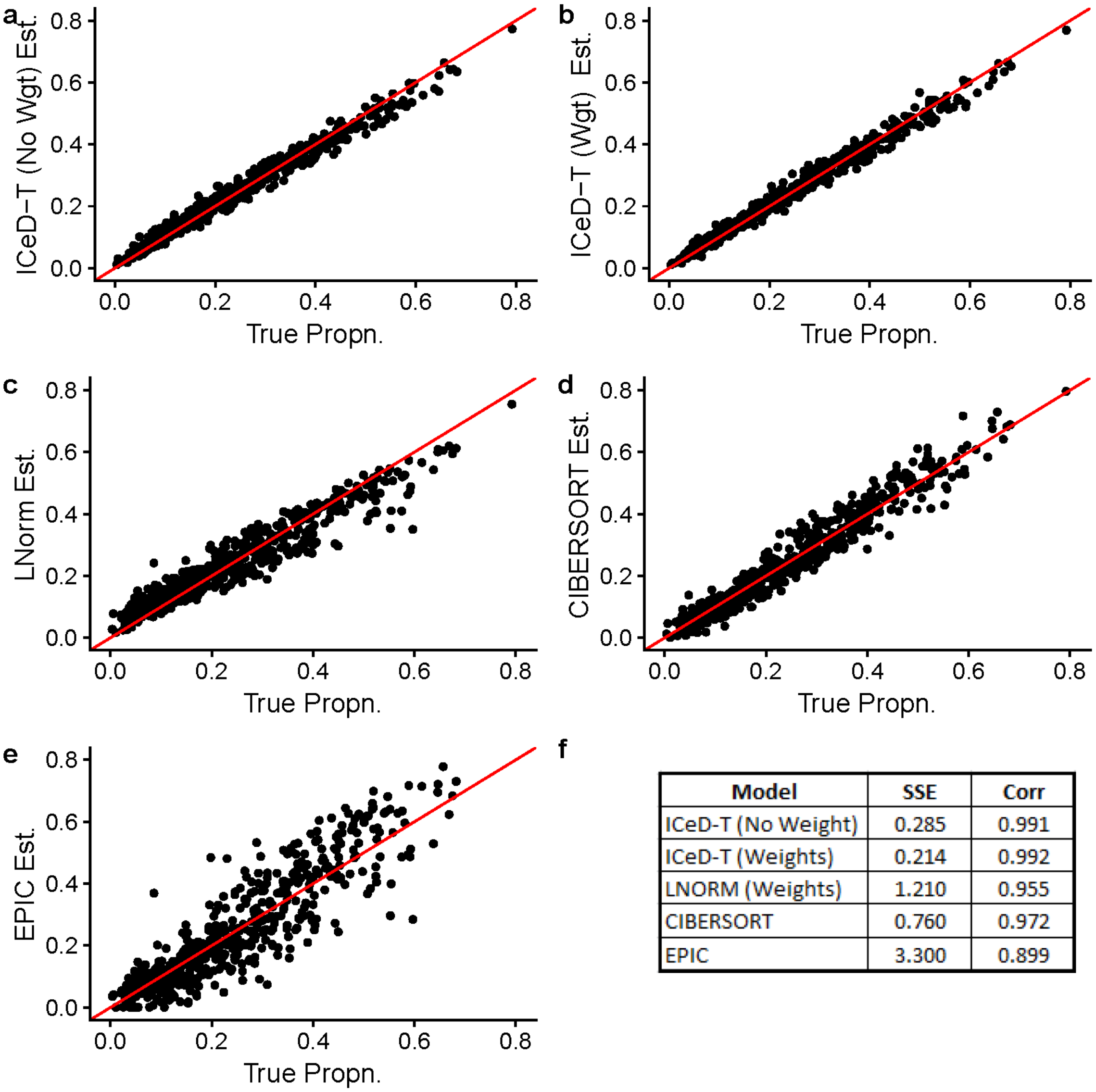
When there is no aberrance in gene expression, all methods perform well, while ICeD-T provides the most accurate estimates of cell type proportions (Figure 4.1). When 20% of the 250 genes are aberrant, the performance of LNorm, EPIC, and CIBERSORT all become worse, while the performance of ICeD-T method remain similar (Figure 4.2). Both EPIC and LNorm's cell type proportion estimates suffer from bias and larger variance in the presence of aberrant genes. CIBERSORT still performs relatively well, but has an apparent inflation of the estimation variance. While the weighted variant of ICeD-T provides the best results, both weighted and unweighted ICeD-T are able to maintain high accuracy with minimal estimation variance (Figure 4.2(a)-(b)).

To identify aberrant genes, ICeD-T computes the posterior probability of a gene being consistent. Examining the distribution of this quantity across consistent and aberrant genes, we see that both the



Model	SSE	Corr
ICeD-T (No Weight)	0.186	0.993
ICeD-T (Weights)	0.154	0.995
LNORM (Weights)	0.154	0.995
CIBERSORT	0.467	0.983
EPIC	0.558	0.980

Figure 4.1: Visualized results of model fits on simulated data without aberrance. Figure (f) summarizes the accuracy across all 135 subjects for each model.



Model	SSE	Corr
ICeD-T (No Weight)	0.285	0.991
ICeD-T (Weights)	0.214	0.992
LNORM (Weights)	1.210	0.955
CIBERSORT	0.760	0.972
EPIC	3.300	0.899

Figure 4.2: Visualized results of model fits on simulated data when $\sim 20\%$ of the genes are abberant. Figure (f) summarizes the accuracy across all 135 subjects for each model.

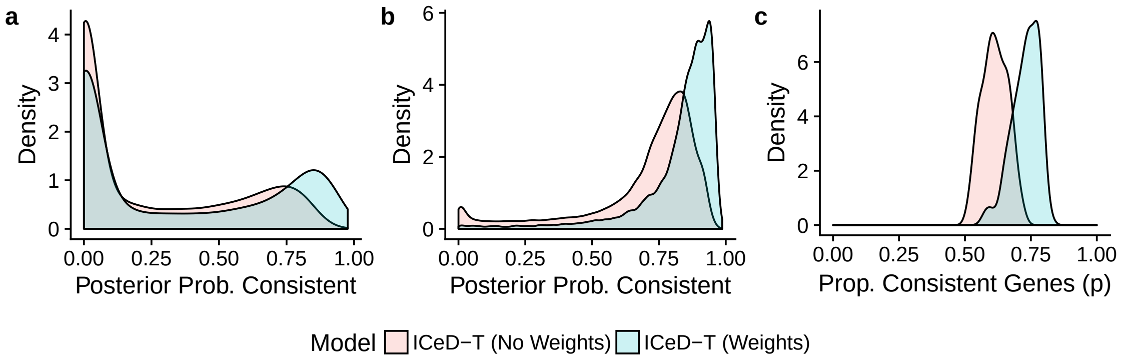


Figure 4.3: (a) The posterior probabilities of being consistent for those aberrant genes. (b) The posterior probabilities of being consistent for those consistent genes. (c) Estimates of the proportion of consistent genes.

weighted and unweighted versions of ICeD-T separate consistent and aberrant genes reasonably well (Figure 4.3). The weighted variant of ICeD-T provides more accurate estimate of the proportion of aberrant genes, and identify consistent genes with higher confidence. For aberrant genes, the posterior probability of being consistent show a bi-modal distribution, implying that a small proportion of aberrant genes are missed. This is partly due to our very challenging simulation setting, with three types of aberrant patterns and three tiers of expression levels for background genes. Such three tiers of background diminishes the difference between background cell types and expressed cell types, and further complicates the identification of aberrant genes.

4.3.2 Validation in Microarray Expression of PBMCs

In the CIBERSORT paper, Newman et al. [25] described the collection of peripheral blood mononuclear cell (PBMC) gene expression data from 20 healthy adults. After extraction of PBMC samples from each subject, these samples were subjected to microarray expression analysis and flow cytometric measurement to establish ground-truth cell type proportions. We use this dataset to evaluate our method and compare its performance with CIBERSORT and EPIC.

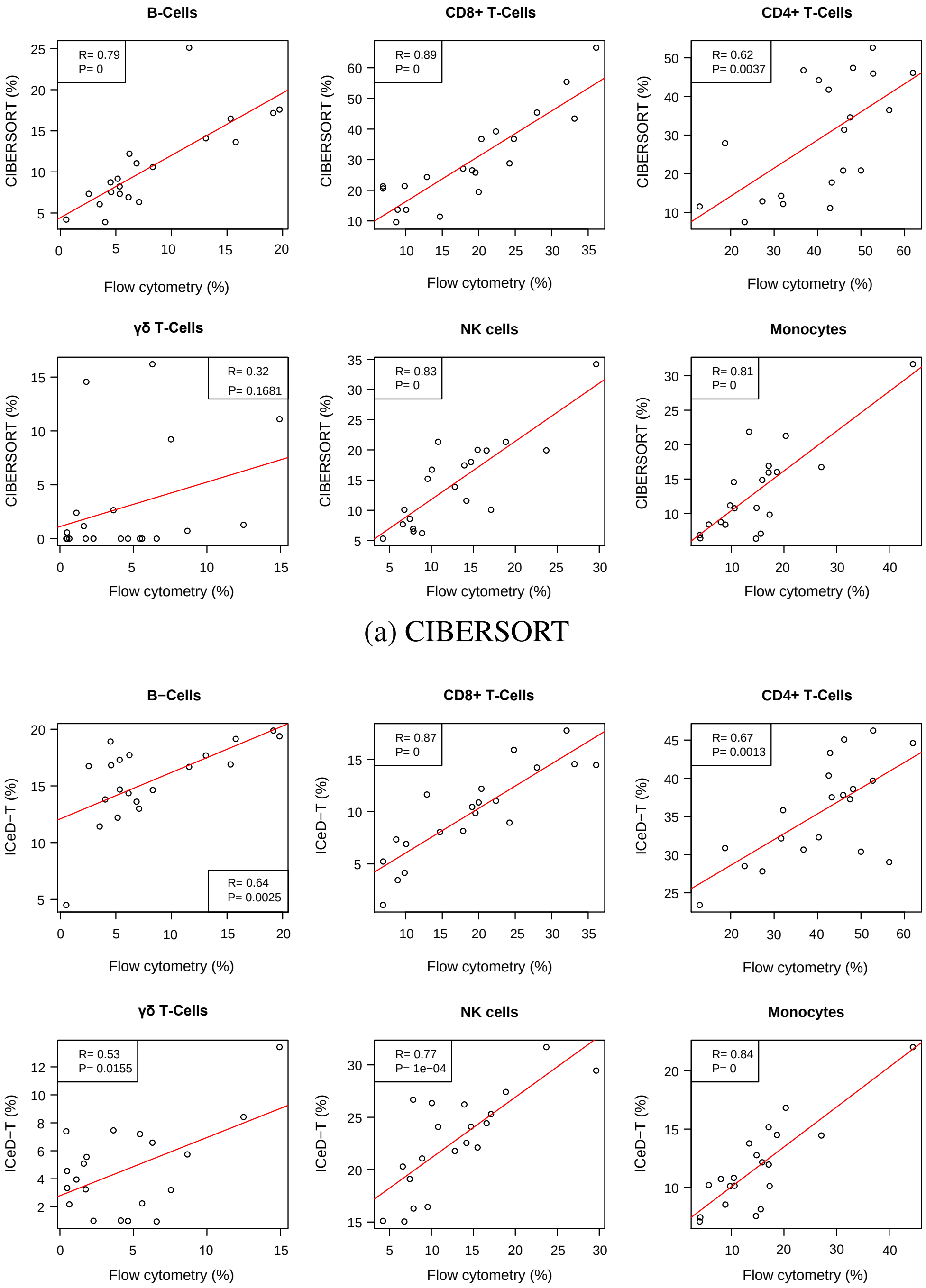
To be consistent with the approach used by Newman et al. [25], we use the their LM22 reference of cell type-specific gene expression for all methods. The LM22 reference matrix is derived from microarray gene expression data, and thus is consistent with the gene expression platform of the bulk tissue samples. EPIC had developed its own reference matrices from RNA-seq data (TRef for bulk tumor samples and BRef for bulk normal samples), but they are inappropriate in microarray settings. Because EPIC and ICeD-T both require that the gene expression from bulk samples and reference samples are measured on the same scale, gene expression data from bulk samples were quantile normalized to a target distribution established by the reference samples used to derive the LM22 matrix. The results of each method are then restricted to the nine cell-types examined in Newman et al. [25]: naive B-cells, memory B-cells, CD8+ T-cells, naive/memory resting/memory activated CD4+ T-cells, $\gamma\delta$ T-cells, Natural killer cells, and monocytes. Estimates for each mixture sample are renormalized so that their summation equals 100 after correction for cell size of different cell types. The accuracy of each method is assessed by comparing sums of squared errors and correlations between the expression-based cell type proportion estimates and flow-cytometry estimates. Correlations are computed by pooling all cell type proportions for all subjects and all cell types.

Model	**SSE**	**Cor**
ICeD-T (no weight)	13.10	0.53
ICeD-T (w/ weight)	12.05	0.59
CIBERSORT	14.15	0.65
EPIC	29.43	0.31

Model	**SSE**	**Cor**
ICeD-T (no weight)	10.48	0.75
ICeD-T (w/ weight)	9.44	0.78
CIBERSORT	11.02	0.77
EPIC	32.01	0.18

Table 4.1: Validation of immune cell proportion estimates by flow cytometry for 9 cell types [left] and 6 cell types after grouping naive B-cells and memory B-cells as B cells, and naive/memory resting/memory activated CD4+ T-cells as CD4+ T cells [right].

Examining the results of the 9 original cell types, ICeD-T provides the most accurate estimates of cell type proportions in terms of sum of squared errors. CIBERSORT, on the other hand, provides the most accurate estimates with respect to the correlations (Table 4.1, Figure 4.4). However, the superior correlation of CIBERSORT is due in part to several cell subsets with positive correlations but severe bias (e.g. memory activated CD4 T-cells, memory resting CD4 T-cells) (Supplementary Materials Section C.4). After grouping a few highly similar cell types (e.g., grouping naive B-cells and memory B-cells as B cells, and naive/memory resting/memory activated CD4+ T-cells as CD4+ T cells), ICeD-T achieves comparable or higher correlation between expression-based cell type



(a) CIBERSORT

(b) ICeD-T (with weight)

Figure 4.4: Comparison of cell type proportion estimates by CIBERSORT and ICeD-T versus the cell type proportions measured by flow cytometry. Red lines indicate the least squares model fit to the estimated immune proportions.

proportion estimates and flow-cytometry estimates while maintaining the smallest sum of squared errors (Table 4.1, Figure 4.4). In this dataset, EPIC has very poor performance, which may be due to the fact that it is designed for RNA-seq data.

4.3.3 Flow Cytometry Validation in Melanomas

In the EPIC paper, Racle et al. [44] obtained metastatic melanoma samples from the lymph nodes of four patients with stage III melanomas. A portion of each of these samples was used for a flow cytometric analysis while the remaining portion was used for bulk RNA-sequencing. Results from flow cytometry were used to establish a ground-truth cell type composition. TPM-normalized RNA-seq expressions and flow cytometry measured compositions were extracted directly from the EPIC R package.

We used EPIC's TRef matrix as reference gene expression for both EPIC and ICeD-T. ICeD-T was run in four different modes, with or without variance weights (denoted by wY and wN, respectively) and with or without sample purity as part of the inputs (denoted by pY and pN, respectively). For this analysis, purity is defined as the proportion of non-immune content plus the proportions of cells not assessed via flow cytometry (e.g. Macrophages, CAFs, and Endothelials, and others). CIBERSORT was fit using both the LM22 and TRef matrices directly to the TPM data. All cell type proportion estimates were corrected by cell size factors reported by Racle et al. [44]. To allow comparison of ICeD-T and EPIC with CIBERSORT that only computes relative immune cell abundance estimates, we obtain relative proportions for all methods by normalizing cell type proportions so that they add up to 1.

Overall EPIC provides more accurate estimates of the total proportion of all immune cells, while ICeD-T provides more accurate estimation of the relative proportions of immune cells among the modeled immune cell types (Table 4.2, Figure 4.5). Comparing non-relativized proportions of the remaining immune cells, ICeD-T (pY, wY) improves upon EPIC's fit in terms of the overall sum of squared error (0.043 vs 0.11) while preserving strong correlation (0.924 vs 0.918) across all subjects.

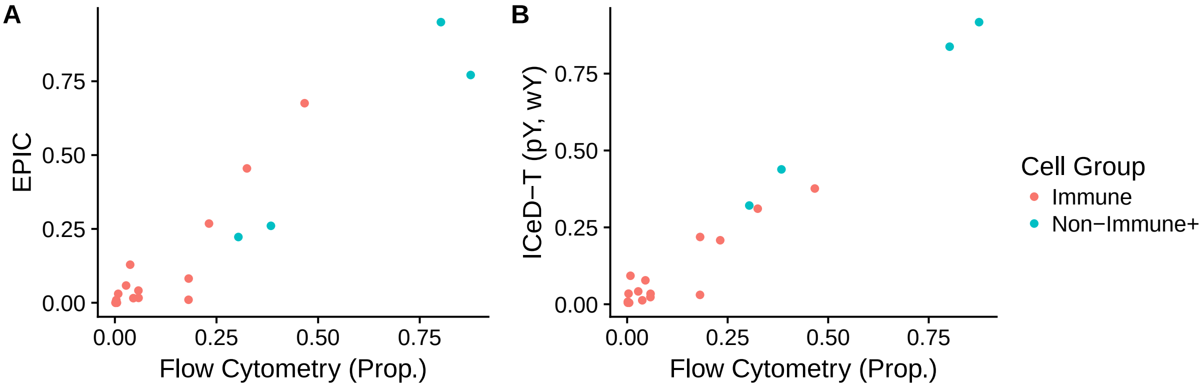


Figure 4.5: Plots of EPIC and ICeD-T model estimates against flow cytometry estimates. ICeD-T is fit using variance weights and sample purity.

Model	**LAU125**	**LAU1255**	**LAU1314**	**LAU335**
CIBERSORT (LM22)	0.12	0.16	0.003	0.010
CIBERSORT (TRef)	0.32	0.10	0.021	0.095
EPIC	0.86	0.15	0.066	0.013
ICeD-T (pN, wN)	1.03	0.10	0.042	0.003
ICeD-T (pN, wY)	1.07	0.14	0.005	0.004
ICeD-T (pY, wN)	0.85	0.08	0.039	0.008
ICeD-T (pY, wY)	0.85	0.14	0.020	0.002

Table 4.2: Sum of Squared Errors for relative immune proportions among all immune cell types. ICeD-T fits are labeled with (pX, wX) to indicate use of purity (pY=Yes and pN=No) and weight (wY=Yes and wN=No).

We also evaluated the performance of CIBERSORT versus the flow cytometry estimates. Compared with other methods, CIBERSORT has comparable or less accurate estimates of cell type proportions in three subjects, but much better performance than the other methods in subject LAU125 (Table 4.2). Based on flow cytometry estimates, this subject has somewhat unexpected immune cell proportion: almost entirely B-cells. All methods perform much worse in this subject than other subjects, with larger sum squared errors. CIBERSORT has relatively better performance for this challenging subject could be due to a combination of its objective function and use of LM22 reference matrix. CIBERSORT's performance becomes worse when using LM22 instead of TRef as reference matrix, though it still has much smaller sum squared error than EPIC and ICeD-T.

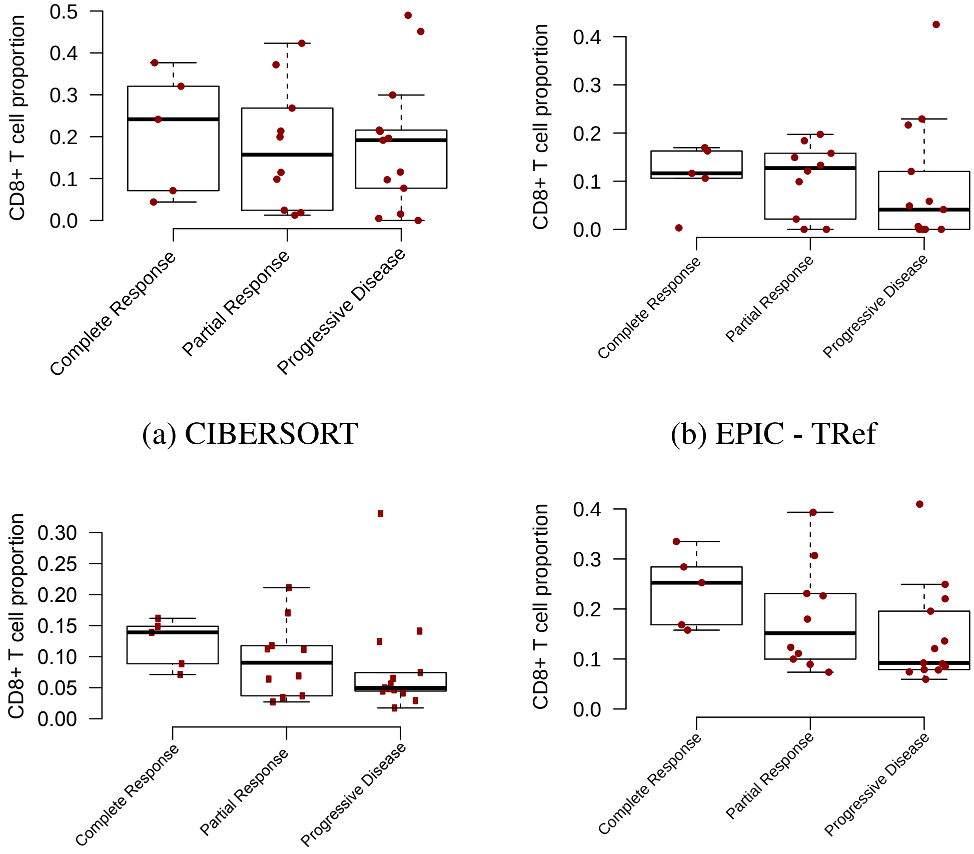
In addition, we also compare the cell type proportion estimate of one cell type across subjects. This is arguably more interesting when we want to use immune cell composition as predictor or treatment response. The limited sample size of this dataset does limit our ability to make comparison, though we do note that ICeD-T provides the best fit for the CD8+ T-cell subset across subjects (Supplementary Materials, Section D.3). CIBERSORT and EPIC particularly struggle to capture the CD8+ T cell proportion for subject LAU1255.

4.3.4 Application to anti-PD-1 Immunotherapy Data

Finally, we use ICeD-T, CIBERSORT, and EPIC to analyze an RNA-seq dataset from bulk tumor samples of melanoma patients [71]. The RNA-seq data are available in 28 patients before treatment with pembrolizumab. We seek to predict treatment response (Complete Response, Partial Response, or Non-response) using CD8+ cell type composition estimated by each of the three methods.

Fastq files of RNA-seq data were downloaded from NCBI Sequence Read Archive, mapped to human genome (hg38) and the number of RNA-seq fragment per gene were counted. Then such counts were normalized by TPM. We ran EPIC and ICeD-T using the TRef reference gene expression data. ICeD-T was fit without using tumor purity as this information was not available. CIBERSORT was fit using LM22 reference matrix. Abundance estimates across each method are corrected using EPIC's cell type size factors. In addition, to ensure comparability across all methods, immune cell proportions are renormalized so that their summation equals to 1.

Differences in relative CD8+ T-cell abundance across response categories was assessed using a Jonckheere-Terpstra test for trended differences. The Jonckheere-Terpstra test can be considered as an extension of non-parametric ANOVA tests (e.g. Kruskal-Wallis) to allow greater power to detect ordered population differences [72]. Previous studies have shown that those cancer patients with more CD8+ T cells within tumor microenviroment are more likely to respond to anti-PD-1 treatment [73]. Thus, as one moves across response categories from most to least responsive to therapy, one would expect to see a decrease in CD8+ T cell abundance.



(a) CIBERSORT

(b) EPIC - TRef

(c) ICeD-T (No Weight)

(d) ICeD-T (Max Var Weight)

Figure 4.6: Comparison of model fits to PD-1 Immunotherapy Data

CIBERSORT and EPIC capture the expected relationship between CD8+ T cell proportion and immunotherapy response to some extent, but have trouble in separating the members of at least two groups. For CIBERSORT, individuals in the partial response group behave similarly to those in the progressive disease group. For EPIC, individuals in the complete response group behave similarly to those who exhibited partial response. The Jonckheere-Terpstra tests provide numerical confirmation of these difficulties as the tests are not significant, with p-values for CIBERSORT and EPIC being 0.30 and 0.14, respectively.

ICeD-T, on the other hand, provides clear visual distinction between these three groups show less CD8+ T cells for those who do not response to anti-PD-1 treatment. This relationship is reinforced through consideration of the significant Jonckheere-Terpstra test (p=0.038). Introduction of variance weights further separates these categories (p=0.017), but does so at the expense of inflated contributions of CD8+ T-cells to the immune response in the TME. Cell type proportions estimates by either versions of ICeD-T have higher within group similarities than either CIBERSORT or EPIC.

4.4 Discussion

In this paper, we have outlined a novel statistical method for immune cell expression deconvolution within tumor tissues, ICeD-T. ICeD-T utilizes the variance stabilizing properties of the log-transformation while simultaneously controlling for aberrant gene behavior within the tumor tissue. In addition, ICeD-T incorporates a variance weighting structure which diminishes the impact of highly variable genes on abundance estimation. Optionally, ICeD-T can refine cell type abundance estimation through use of tumor purity information, if available.

We have demonstrated that ICeD-T is an accurate model in both simulated and real datasets. The robustness of ICeD-T to misbehaved genes and its ability to identify these genes was demonstrated in simulated data. ICeD-T's accuracy was reinforced in real datasets using both microarray and RNA-seq expression where it was consistently a top performer compared with other methods. In particular, it was noted that ICeD-T can provide more accurate estimates of the CD8+ T-cell proportions than other methods. We applied ICeD-T to study the relation between CD8+ T cell proportion and response to anti-PD-1 immunotherapy and found significant associations between CD8+ T cell proportions and patients' response to immunotherapy.

There is room to further improve the performance of ICeD-T. One direction is to refine the reference matrix of cell type-specific gene expression. In this paper, we have adopted the reference gene expression matrix (TRef) used by EPIC's. TRef was constructed using single cell RNA-seq (scRNA-seq) data from melanoma cancer samples. Cell type-specific expression was estimated by pooling cells of the same cell types, identified by clustering method. However, some technical limitations of scRNA-seq, such as dropout (expression of many genes were measured at 0 while they may be lowly expressed) [74]. Careful examination of such effects may improve the reference matrix of cell type-specific gene expression. On the other hand, techniques for scRNA-seq are a very active research area. New techniques and new data (e.g. Human Cell Atlas [75]) may help generate higher quality data for such a reference matrix.

Another future direction to improve ICeD-T is to refine the the variance weights. We have implemented the variance weight for each gene based on the maximum of cell type-specific variances. Other options that use the variances across all cell types may be more desirable. As is, some minimally variable genes may be overweighted since the maximal variance was utilized for weights. By refining the weighting structure, the perception of gene-expression variance in the mixture could improve and allow for genes to contribute to cell type composition estimation in a way which more closely mirrors their true behavior. However, with limited cell type-specific gene expression data, we have not yet identified a clear choice.

CHAPTER 5: MAPPING TUMOR-SPECIFIC EXPRESSION QTLS IN IMPURE TUMOR SAMPLES

5.1 Introduction

Genetic variants (e.g. Single Nucleotide Polymorphisms (SNPs)) that are associated with the expression of one or more genes are referred to as gene expression quantitative trait loci (eQTLs). Genome-wide eQTL study is a powerful tool for understanding the functional roles of genetic variants. For example, eQTL analyses can help interpret the results of genome-wide association studies (GWASs) [46].

There are two types of eQTL, *cis*-eQTL and *trans*-eQTL [47, 48], which are distinguished by the pattern of expression change they induce. To precisely define these eQTL types, we first define the term “allele”. Consider a diploid genome, which has two homologous copies of each chromosome: a maternal copy and a paternal copy. As such, each genetic locus (e.g., a SNP or a gene) has two copies within a cell, which are referred to as the two alleles of this locus. For a gene affected by a *cis*-eQTL, the expression of each allele is moderated by the genetic content of the corresponding homologous chromosome, which leads to allelic imbalance of gene expression. In contrast, for a gene affected by a *trans*-eQTL, the expression of both alleles are modified to the same extent.

The concepts of *cis*- and *trans*-eQTLs are crucial to our method development, and thus we further illustrate them by two examples. Consider a *cis*-eQTL, which is a SNP with A and T alleles. The A allele inhibits the binding of a transcription factor, which up-regulates the expression of a nearby gene. In contrast, the T allele does not affect transcription factor binding. If we refer to the two alleles of this gene by A or T allele (based on known phase between this *cis*-eQTL and and the nearby gene of interest), this *cis*-eQTL leads to lower expression in the A allele than the T allele. An example of a *trans*-eQTL could be a SNP that affects the activity of a transcription factor, which in turn regulates the expression of a gene and it has the same influence on the gene expression from both alleles.

Cis-eQTLs are often falsely conflated with local eQTLs since *cis*-eQTLs are often located nearby the genes they affect. *Trans*-eQTLs, on the other hand, can be located anywhere in the genome in relation to the genes which they regulate [47]. It is important to reinforce that the defining characteristics of *cis*-eQTLand *trans*-eQTLare not based on their proximity to their target genes, as local eQTLs can induce *cis*- or *trans*- patterns of expression change.

Traditional eQTL mapping methods implicitly assume an eQTL has the same effect on all cells within a sample. This is a reasonable assumption for samples with a relatively homogeneous cell population. However, tumor samples invariably contain both tumor cells and infiltrating normal cells (e.g., immune cells) and eQTL effects could differ between these two types of cells. To quantitatively capture this concept of inhomogeneity within a tumor cell population, we consider its tumor purity, defined as the proportion of tumor cells within the tumor sample. Previous eQTL studies in tumor samples often ignore tumor purity information and directly apply eQTL mapping methods that assume the tumor samples are composed of homogenous cells [49, 50, 51, 53, 76]. When tumor and normal eQTL are discordant, our results show that ignoring tumor purity may lead to severely inflated type I error in the identification of tumor-specific eQTL.

In this paper, we focus on eQTL mapping using germline genetic variants. The proposed methods may be extended to study eQTL mapping using somatic variants, but such extensions must address the challenge of intra-tumor heterogeneity with respect to somatic mutations. To the best of our knowledge, only one previous work has considered a similar problem of cell-type-specific eQTL mapping given cell type proportion estimates [54]. Specifically, Westra et al [54] identify neutrophil-specific eQTLs using a linear model: $y = \beta_0 + \beta_1 G + \beta_2 P + \beta_3 GP$ where y is gene expression, G is genotype, and P is an estimate or proxy of neutrophil proportion. Loci where eQTL effects are different between neutrophil and other cell types were identified by testing the hypothesis $\beta_3 = 0$. This approach does not directly estimate or assess cell-type-specific eQTL effects. We show in our analysis that a variant of this method that explicitly models a tumor-specific eQTL effect has lower power than our proposed method.

The proposed methods are applied to the genetic expression data of 547 women with breast cancer provided by The Cancer Genome Atlas. We examine the agreement and disagreement between each posited model with respect to eQTL identification as well as a discussion of some interesting eQTL identified by our method.

5.2 Model

Our model is an extension of the TReCASE method, which performs eQTL mapping using RNA-seq data [55]. The TReCASE method models RNA-seq data along two dimensions, Total Read Count (TReC) and Allele-Specific Expression (ASE), and simultaneously uses these two types of data for eQTL mapping [48, 55]. The TReC for a gene of interest is the total number of RNA-seq reads mapped to this gene. Under the TReCASE framework, TReCs across samples are modeled by a negative binomial distribution. The ASE of a gene is quantified by the number of allele-specific reads that match the genotype of one haplotype, but not the other haplotype of this gene. Thus, an RNA-seq read is allele-specific if it overlaps with at least one SNP that is heterozygous across the two haplotypes. The number of allele-specific reads from one allele given the total number of allele-specific reads follows a beta-binomial distribution in the TReCASE framework.

The TReCASE method jointly analyzes the TReC and ASE data for *cis*-eQTLs as these two types of data provide consistent information regarding the effect sizes of *cis*-eQTLs. In contrast, for a *trans*-eQTLthe eQTL effect is non-zero for TReC but zero for ASE, and thus only TReC data are used for mapping *trans*-eQTLs. The TReCASE model implicitly assumes eQTL effects are the same across all the cells within a sample, which may not be correct for tumor samples. In this paper, we extend the TReCASE model for tumor eQTL studies through the incorporation of tumor purity and separate tumor- and normal-specific eQTL effects into our likelihood model. We refer to this new model as pTReCASE.

5.2.1 The Data

We assume that phased germline genotype data and RNA-seq data from tumor samples are available for n independent subjects. Since germline genotype data have been phased, we have genotypes for each of a subjects' two haplotypes. We also assume that an estimate of tumor purity is available for each tumor sample. For example, one could estimate tumor purity using somatic copy number aberration data [69].

While pTReCASE is designed to be applied across multiple gene-snp pairs, in the following discussion, we consider the model for a specific gene of interest and a single potential eQTL of this gene. For clarity and simplicity in the following notation, we suppress subscripts related to gene and eQTL and note that the given structure applies across any Gene-SNP pairing. Let $G(i)$ be the genotype of subject i at the potential eQTL. $G(i)$ can take values in $\{AA, AB, BB\}$ where A and B denote two alleles of the potential eQTL. Let ρ_i, d_i, and $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$ be the tumor purity estimate, read depth measurement, and a vector of p covariates for the i-th sample respectively. We set d_i as the 75-th percentile of the TReCs across all genes in the i-th sample, which is a more robust way to measure read-depth than the summation of the TReCs across all genes.

5.2.2 Purity Corrected Total Read Count (pTReC) Model

The total read count Y_i is defined as the number of RNA-seq reads that are mapped to a given gene. We assume that Y_i follows a negative binomial distribution with over-dispersion ϕ and subject-specific mean μ_i, the likelihood for which is given by:

$$f(Y_i; \mu_i, \phi) = \frac{\Gamma(Y_i + 1/\phi)}{Y_i!\Gamma(1/\phi)} \left(\frac{1}{1 + \phi\mu_i}\right)^{1/\phi} \left(\frac{\phi\mu_i}{1 + \phi\mu_i}\right)^{Y_i}$$

with $E(Y_i) = \mu_i$ and $Var(Y_i) = \mu_i + \phi\mu_i^2$. Summarizing across all n subjects, the log-likelihood for the pTReC model is:

$$\ell_{TReC} = \sum_{i=1}^{n} \log\left[f(Y_i; \mu_i, \phi)\right]. \tag{5.1}$$

.

In impure tumor samples, pTReC captures the genetic effects of a potential eQTL on Y_i through its specification of μ_i (equation 5.3). In order to illuminate the structure of μ_i , we must first quantitatively define these genetic effects for both tumor and normal cells. Let μ_{iA} and μ_{iB} be the mean expression of alleles A and B for the i-th subject, and use superscripts $^{(T)}$ and $^{(N)}$ to denote measurements from tumor and normal cells, respectively. Values of μ_{iA} and μ_{iB} are allowed to vary across subjects, but we assume that the ratios of these quantities are fixed across subjects. Symbolically:

$$\text{For all } i, \qquad \mu_{iB}^{(N)}/\mu_{iA}^{(N)} = \eta, \quad \mu_{iB}^{(T)}/\mu_{iA}^{(T)} = \gamma, \quad \mu_{iA}^{(T)}/\mu_{iA}^{(N)} = \kappa.$$

Thus, γ represents an eQTL effect within tumor tissues that is common to all subjects and η is its counterpart for tumor tissues. The remaining parameter, κ, is a nuisance parameter that models the baseline gene expression difference between tumor and normal tissues.

To further elucidate these eQTL effects, focus on γ. When $\gamma = 1$, no eQTL effect exists within the tumor as the mean expression of alleles A and B are identical within a subject $\big($for all $i, \quad \mu_{iA} = \mu_{iB}\big)$. Suppose now that $\gamma < 1$; this implies that the B allele is under-expressed relative to A by a multiplicative factor of γ (e.g. $\mu_{iB} = \gamma\mu_{iA}$). On the other hand, $\gamma > 1$ implies over-expression of the B-allele relative to A. Identical rules govern the interpretation of η for normal tissues. Note that by specifying common values of η and γ across subjects, we imply that the ratio of over/under-expression of the B allele relative to A is consistent across subjects while allowing each subject a unique mean-expression level of alleles A and B.

Now let $\xi_i = \mu_{iB}/\mu_{iA}$. Assuming that the mean expression of an allele is the weighted summation of its expression in tumor and normal cells, we have:

$$\xi_i = \frac{\mu_{iB}}{\mu_{iA}} = \frac{(1-\rho_i)\mu_{iB}^{(N)} + \rho_i\mu_{iB}^{(T)}}{(1-\rho_i)\mu_{iA}^{(N)} + \rho_i\mu_{iA}^{(T)}} = \frac{(1-\rho_i)\eta + \rho_i\kappa\gamma}{(1-\rho_i) + \rho_i\kappa} = (1-c_i)\eta + c_i\gamma, \tag{5.2}$$

where $c_i = (\rho_i\kappa)/(1-\rho_i+\rho_i\kappa)$. The third equality is obtained by dividing both the numerator and denominator by $\mu_{iA}^{(N)}$. Therefore, the overall genetic effect in a tumor sample is a mixture of the genetic effects within tumor cells and normal cells.

Next we consider modeling the μ_i across different genotypes. First, if the i-th subject has genotype AA at the candidate eQTL,

$$\mu_i = \mu_{iA} + \mu_{iA} = 2\mu_{iA}^{(N)}\left[1-\rho_i+\rho_i\kappa\right].$$

We model $\log\left(2\mu_{iA}^{(N)}\right)$ using a linear function of log read-depth and p covariates: $\beta_0 + \beta_d\log(d_i) + \sum_{j=1}^{p}\beta_j x_{ij}$. Applying similar derivations for the subjects with genotypes AB and BB, we have:

$$\log(\mu_i) = \begin{cases} \beta_0 + \beta_d\log(d_i) + \sum_{j=1}^{p}\beta_j x_{ij} + \log(1-\rho_i+\rho_i\kappa), & \text{if } G(i) = AA \\ \beta_0 + \beta_d\log(d_i) + \sum_{j=1}^{p}\beta_j x_{ij} + \log(1-\rho_i+\rho_i\kappa) + \log\left(\frac{1+\xi_i}{2}\right), & \text{if } G(i) = AB \\ \beta_0 + \beta_d\log(d_i) + \sum_{j=1}^{p}\beta_j x_{ij} + \log(1-\rho_i+\rho_i\kappa) + \log(\xi_i), & \text{if } G(i) = BB \end{cases} \tag{5.3}$$

In the pTReC model, estimates of $\beta, \kappa, \eta, \gamma,$ and ϕ are obtained by maximizing equation 5.1 with respect to these parameters. We maximize this likelihood using a block coordinate ascent algorithm. Within block coordinate ascent, optimization proceeds by maximizing the likelihood with respect to a single set of parameters – called a block – at a time while holding all other blocks of parameters fixed. Each block of parameters is then optimized iteratively until covergence of parameter estimates. For pTReC, block 1 consists of parameters $\kappa, \eta,$ and γ; block 2 consists of parameters ϕ, β_d, β_j for $j = 0, 1, ..., p$. Thus, holding the values of all parameters in block 2 constant, a single update of block 1 is accomplished via a quasi-Newton method (LBFGS).

Then, holding the parameters of block 1 fixed, the parameters in block 2 are updated via negative binomial regression. As described, we then iteratively update the parameters in blocks 1 and 2 until convergence.

5.2.3 Purity Corrected Allele Specific Expression (pASE) Model

We refer the reader to Sun and Hu [48] for details on how allele specific reads are counted in RNA-Seq data. In the following, we briefly describe this process for a single candidate eQTL. For each subject, we have genotypes available for arbitrarily labeled haplotypes, haplotype 1 and haplotype 2. We extract all RNA-seq reads that overlap with at least one heterozygous SNP within the body of the gene and assign each of these reads to the haplotype that matches its nucleotide sequence. As haplotypes 1 and 2 are arbitrarily labeled for each subject, we ensure comparability across subjects by relabeling these haplotypes with respect to the genotype of the candidate eQTL. For subjects who are heterozygous at the candidate eQTL, haplotype A contains the A allele of the candidate eQTL and haplotype B contains the B allele. For subjects who are homozygous at the candidate eQTL, haplotypes A and B may be defined arbitrarily without affecting the likelihood function or statistical inference.

Let R_{iA} and R_{iB} be the number of allele specific RNA-seq reads assigned to haplotypes A and B, respectively. Let $R_i = R_{iA} + R_{iB}$ be the total number of allele-specific RNA-seq reads. We model R_{iB} given R_i using a beta-binomial distribution with probability of success π_i and over-dispersion ψ, the likelihood of which is given by:

$$g(R_{iB}; R_i, \pi_i, \psi) = \frac{R_i!}{R_{iA}!R_{iB}!} \frac{\Gamma\left(\psi^{-1}\right)\Gamma\left(\psi^{-1}\pi_i + R_{iB}\right)\Gamma\left(\psi^{-1}(1-\pi_i) + R_{iA}\right)}{\Gamma\left(\psi^{-1}\pi_i\right)\Gamma\left(\psi^{-1}(1-\pi_i)\right)\Gamma\left(\psi^{-1} + R_i\right)}.$$

Incorporating all individuals, we may express the log-likelihood for the ASE model as:

$$\ell_{ASE} = \sum_{i=1}^{n} \log\left[g\left(R_{iB}; R_i, \pi_i, \psi\right)\right],$$

where $\xi_{i,ASE} = \mu_{iB}/\mu_{iA}$ and

$$\pi_i = \begin{cases} \mu_{iB}/(\mu_{iA} + \mu_{iB}) = \xi_{i,ASE}/(\xi_{i,ASE} + 1), & \text{if } G(i) = AB \\ 0.5, & \text{otherwise.} \end{cases}$$

Recall that we had defined ξ_i similarly in equation (5.2). We introduce the slight change in notation for the pASE model in order to distinguish *cis*-acting and *trans*-acting eQTL. For *cis*-eQTL, $\xi_{i,ASE} = \xi_i$ as defined in equation (5.2). For *trans*-eQTL, however, $\xi_{i,ASE} = 0.5$ since expression of the A and B alleles are impacted to the same extent. A consequence of the above modeling strategy is that ASE is uninformative regarding κ, η, or γ when an eQTL is *trans*-acting. In addition, for *cis*-eQTL, subjects who are homozygous at the potential eQTL do not contribute to the estimation of the eQTL parameters κ, η, or γ. However, such subjects are informative regarding the over-dispersion parameter ψ.

As for pTReC, model fitting in pASE is also achieved via block coordinate ascent using two blocks of parameters: block 1 consists of parameters κ, η and γ; block 2 consists of the lone parameter ψ. We employ the cyclical algorithm described in the previous section to iteratively update the parameters of blocks 1 and 2 until convergence. Updates for each block are accomplished via LBFGS.

5.2.4 pTReCASE: Unifying pTReC and pASE Models

Restricting to *cis*-eQTLs, the pTReC and pASE models share the κ, η, and γ parameters allowing for unification into a single likelihood model:

$$\begin{aligned} P\left(Y_i, R_i, R_{iB}|\Theta\right) &= P\left(Y_i|\Theta\right) P\left(R_i|Y_i, \Theta\right) P\left(R_{iB}|Y_i, R_i, \Theta\right) \\ &= f\left(Y_i|\Theta\right) P\left(R_i|Y_i\right) g\left(R_{iB}|R_i, \Theta\right), \end{aligned}$$

where $\Theta = (\kappa; \eta; \gamma; \beta_j \text{ for } j = 0, 1, ..p; \phi; \eta_{ASE}; \gamma_{ASE}; \psi)$, the set of all parameters found in the pTReC and pASE models.

Note that the likelihood above explicitly relates Y_i and R_i. Since the set of allele-specific reads (R_i) is a subset of all reads mapping to the i-th gene (Y_i), it is clear that $R_i \leq Y_i$. Despite this relationship between these two variables, it is reasonable to assume that given Y_i, the distribution of R_i does not depend on our covariate or eQTL effects. Given the total read count at the i-th gene, the number of reads overlapping at least one heterozygous SNP in gene i is a function of the number of such SNPs present within the gene-body; thus, this value is not likely to be related to eQTL effects. Therefore, we may remove $P(R_i|Y_i)$ from the likelihood function. The log-likelihood of all n subjects is then given by:

$$\ell(\Theta) = \sum_{i=1}^{n} \log\left[f\left(Y_i|\Theta\right)\right] + \log\left[g\left(R_{iB}|R_i, \Theta\right)\right] = \ell_{TReC} + \ell_{ASE}.$$

Model fitting is achieved via block coordinate ascent using three blocks: block 1 consists of κ, η and γ; block 2 consists of ϕ, β_d and β_j for $j = 0, 1, ..., p$; and block 3 consists of ψ alone. A single update is defined by the following steps. First, given the parameters of blocks 2 and 3, the parameters of block 1 are updated using LBFGS. Then, given the parameters of blocks 1 and 3, the parameters of block 2 are updated via negative binomial regression. And finally, given the other parameters, the parameter of block 3 is updated using LBFGS. These cyclical updates are repeated until convergence.

5.2.5 Hypothesis Testing

Under the proposed models of sections 2.2 through 2.4, there are three critical questions of interest. Should we use the pTReC or pTReCASE model to assess an eQTL? Does an eQTL exist within normal tissue? Does an eQTL exist within tumor tissue?

Addressing the first question requires consideration of the biological mechanisms driving *cis*- and *trans*-eQTLs. For a *cis*-eQTL, the TReC and ASE components share the same parameters for eQTL effect sizes, and thus jointly modeling TReC and ASE (i.e., TReCASE) increases power. For a *trans*-eQTL, expression of both alleles of the affected gene are altered to the same extent, and thus ASE is not informative in the detection of eQTL or estimating eQTL effect size. Therefore, only TReC data should be used for eQTL mapping of *trans*-eQTL. We develop a “Cis-Trans” score test to aid in model selection by addressing a null hypothesis of consistent eQTL effects across the TReC and ASE components of the model.

To structure this test, let η_{ASE} and γ_{ASE} be the eQTL effects for a gene and a candidate eQTL within the ASE component of the model. We still use η and γ to model eQTL effects in TReC data. Define $\eta_{ASE} = \eta + \alpha_\eta$ and $\gamma_{ASE} = \gamma + \alpha_\gamma$ where α_η and α_γ reflect the discrepancy between ASE and TReC eQTL effects for normal and tumor tissues, respectively. The null hypothesis of equivalent eQTL effects in TReC and ASE components of the model is defined using the notation above by $\alpha_\eta = \alpha_\gamma = 0$. See the supplementary information for a detailed description and derivation of this “Cis-Trans” score test. The test statistic and its asymptotic distribution are provided below:

$$SC = \dot{\ell}\left(\hat{\Theta}\right)^T I(\hat{\Theta})^{-1}\dot{\ell}\left(\hat{\Theta}\right) \sim \chi^2_{(2)},$$

where $\hat{\Theta}$ are the MLEs of our parameters under the null hypothesis; $\dot{\ell}$ is the gradient of the TReCASE likelihood with respect to the parameters; and $I(\hat{\Theta})$ is the Fisher’s Information Matrix.

The presence of eQTL in normal tissue (i.e., $\eta \neq 1$) or tumor tissues (i.e., $\gamma \neq 1$) can be assessed using likelihood ratio tests (LRT). These test statistics and their asymptotic distributions take the form:

$$\Lambda = -2\left[\ell\left(\hat{\Theta}_0\right) - \ell\left(\hat{\Theta}\right)\right] \sim \chi^2_{(1)},$$

where $\hat{\Theta}_0$ represents parameter estimates under the null and $\hat{\Theta}$ represents estimates under the alternative. To test for the presence of an eQTL in normal or tumor tissue, $\hat{\Theta}_0$ is obtained by fitting the model under a null hypothesis of $\eta = 1$ or $\gamma = 1$, respectively.

To identify eQTL within a single gene-snp pair, we propose the following procedure.

(1) Conduct the “Cis-Trans” score test to determine use of pTReC or pTReCASE model.

(2) Under the prescription of the "Cis-Trans" test, conduct independent LRT of $\gamma = 1$ and $\eta = 1$ to determine the presence of eQTL effects.

The above algorithm is designed to ensure that inconsistent effects in the pTReC and pASE models do not limit the power to detect *trans*-eQTL. For *trans*-eQTL, the eQTL effect modeled by pASE should be 1 whereas that modeled by pTReC should be non-unity $(\neq 1)$. Thus, joint estimation using pTReCASE will dilute effect strength resulting in a loss of power. Since the goal of the "Cis-Trans" score test is to determine whether the eQTL effects modeled by pTReC and pASE are consistent and not to assess the presence of an eQTL effect, the sampling properties of the eQTL effect tests should remain unaffected.

5.3 Results

5.3.1 Simulation Study

We conducted a simulation study to compare the statistical power and type I error rate of pTReCASE and several other methods. Simulations were conducted across a range of eQTL effect sizes in normal and tumor cells. To assess Type I error, we set γ at 1 and allowed η to vary. To assess power to detect tumor-specific eQTL, we set η at 1 and allowed γ to vary. For each pair of η and γ, we simulated 400 replicates of gene expression and genotype data for 500 subjects. Genotypes were simulated assuming a minor allele frequency of 0.2. Read counts were simulated according to the pTReCASE model using the following algorithm:

(1) Randomly generate tumor purities from a uniform distribution on (0.5,1) for each of the 500 subjects.

(2) Simulate TReC via a negative binomial model with:

   (A) Mean of 100 reads for subjects with genotype AA and tumor purity of 0%.

   (B) $\kappa = 1.5$ and $\phi = 0.2$.

(3) Assume that 5% of the simulated TReC reads are allele specific reads, rounded to the nearest integer. Partition allele specific reads to haplotypes according to the established beta-binomial model using an overdispersion of $\psi = 0.2$.

Each considered eQTL model is then fit to the simulated data. For any given modeling procedure, the type I error is estimated by the proportion of simulations which incorrectly identify a tumor eQTL when none is present. Power is estimated by the proportion of simulations which correctly identify a tumor eQTL when one is present.

The competing eQTL models that we considered include the TReC/TReCASE model without correction for tumor purity, and the TReC model with tumor purity (pTReC). In addition, we also considered a naïve approach of linear regression ignoring tumor purity, labeled LR, and a modification of the approach adopted by Westra et al [54] denoted by pLR. To fit a linear model, we first applied a normal quantile transformation to (read-depth corrected) TReC values of each gene across n samples, and then used the transformed TReC as a response variable for linear regression. Specifically, we first replaced TReC values by their ranks across n samples, and then replaced the ranks by their corresponding normal quantiles. For example, rank r was replaced by the $r/(n + 1)$-th normal quantile. Letting $\bar{Y}$ be the transformed TReC data, the linear model is given by $E\left(\bar{Y}\right) = \beta_0 + \beta_1 G$, where G is the genotype of the candidate eQTL.

To test genotype and tumor purity interaction using pLR, we fit a linear model $E\left(\bar{Y}\right) = \beta_0 + \beta_1 G + \beta_2 \rho + \beta_3 G\rho$ where ρ is an estimate of tumor purity. The interaction test employed by Westra et al [54] (i.e. $\beta_3 = 0$) does not assess the strength of a tumor eQTL. Rather, it tests whether eQTL effects differ between tumor and normal tissues. Under pLR, we assessed tumor eQTL effects by testing $\beta_1 + \beta_3 = 0$ since $\beta_1 + \beta_3$ is the genetic effect of the candidate eQTL when tumor purity is 1.

All three methods that control for tumor purity (pTReCASE, pTReC, pLR) control Type I error at the desired level. As eQTL strength in the normal tissue increases, the methods that do not account for tumor purity see a rapid increase in Type I error (Figure 1A). In terms of power (Figure 1B), the methods that do not account for tumor purity exhibit the largest statistical power due to

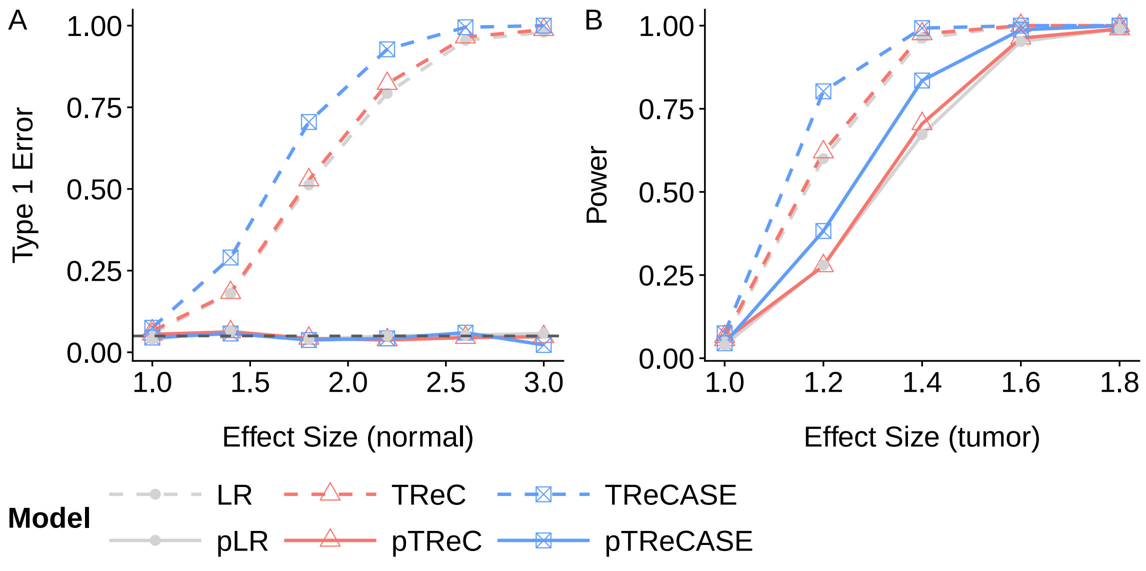


Figure 5.1: Examining Type I error [A] and Power [B] from simulation studies.

their anti-conservative control of Type I error. Among those methods that control Type I error (i.e. pLR, pTReC, pTReCASE), pTReCASE has the highest power. This is a result of its joint analysis of TReC and ASE.

5.3.2 The Cancer Genome Atlas (TCGA) Data

5.3.2.1 Data and Model Fitting

We applied the pTReCASE model to analyze gene expression and germline SNP genotype data from 550 breast cancer patients of The Cancer Genome Atlas (TCGA) project. All the data were downloaded from TCGA data portal (`https://tcga-data.nci.nih.gov/docs/publications/tcga/`), which has now been replaced by NCI Genomic Data Commons (`https://portal.gdc.cancer.gov/`). We started with 728 patients with RNA-seq data from tumor samples. In order to assess allele-specific gene expression, we downloaded raw RNA-seq data in bam file format. For genotype data, we downloaded the Affymetrix CEL files. We restricted our analysis 550 of 728 patients who had available genotype data, passed quality controls for both genotype and RNA-seq data, and were Caucasian females (See Supplementary Materials

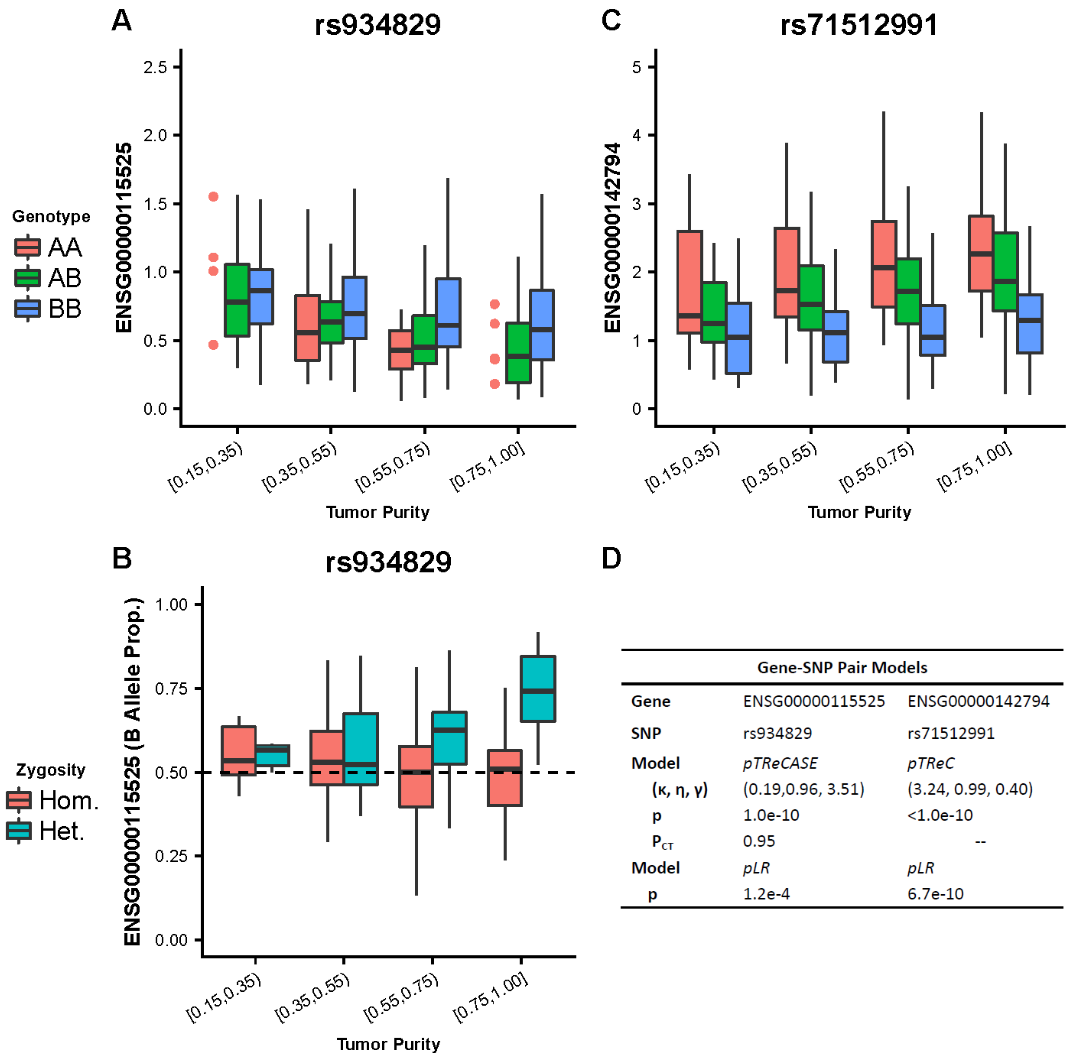
Section B for details). Males were excluded as breast cancer in men is rare and may have a different disease etiology. The restriction to Caucasian samples is not necessary, but it helps to eliminate possible confounders [77].

For the remaining 550 patients, genotype imputation and haplotype phasing was performed by MACH [78] using reference haplotypes from the 1000 Genomes Project. Starting with $\sim$800,000 SNPs genotyped by Affymetrix 6.0 array, we imputed the gneotypes for $\sim$36 million SNPs. For each sample, we used all the SNPs with heterozygous genotypes to estimate allele-specific expression (See Supplementary Materials Section B for details). For the purposes of eQTL mapping, we restricted our analysis to those SNPs with MAF $\geq$ 0.02 (6,825,065 SNPs after imputation) because there is limited power to detect eQTL at lower values of MAF. Tumor purities were estimated using ABSOLUTE [69], which led to the exclusion of three additional subjects lacking valid purity estimates. Estimated haplotypes and tumor purities were treated as truth in the subsequent pTReCASE and linear regression models.

Linear models for eQTL analysis and the revised Westra approach (i.e. pLR) were fit using matrixEQTL [79] and customized R code on normal quantile transformed RNA-Seq count data, respectively. TReC, TReCASE, pTReC and pTReCASE models were fit using our own package. The median analysis time for a single gene-SNP pair using pTReCASE was 2.71 seconds (IQR = 2.93 seconds). The covariates used for eQTL mapping include read-depth of RNA-seq experiments (Supplementary Figure 7), RNA sample plates, age, and the top two principal components derived from the genotype data of the 550 Caucasian samples. Since our method is designed to identify *cis*-eQTLs and most *cis*-eQTLs are local to the genes which they affect, we restricted our analysis to SNPs located within 100Kb of the gene of interest.

5.3.2.2 eQTL Identification

Figures 5.2A-B illustrate a tumor-specific eQTL identified by the pTReCASE model. The estimates of effect sizes (ratio of gene expression of the B allele versus the A allele) for normal and tumor-specific eQTLs are 0.96 (η) and 3.51 (γ), respectively. The fold change of gene expression in tumor versus normal cells (for genotype AA) is 0.19 (κ) (Figure 5.2D). In other words, gene



Gene-SNP Pair Models		
Gene	ENSG00000115525	ENSG00000142794
SNP	rs934829	rs71512991
Model	*pTReCASE*	*pTReC*
(κ, η, γ)	(0.19,0.96, 3.51)	(3.24, 0.99, 0.40)
p	1.0e-10	<1.0e-10
P_{CT}	0.95	--
Model	*pLR*	*pLR*
p	1.2e-4	6.7e-10

Figure 5.2: (A) Covariate-corrected total expression estimated via pTReCASE plotted against genotype and tumor purity. Outliers were suppressed for clarity. Dot plot instead of boxplot was used when sample size of a category is too small. (B) Examination of the allele specific expression corresponding to case shown (A). (C) Covariate-corrected total expression estimated via pTReC plotted against genotype and tumor purity. (D) Table providing Gene, SNP, and estimated parameters for the displayed assessments. p_{CT} references the value of the Cis-Trans score test.

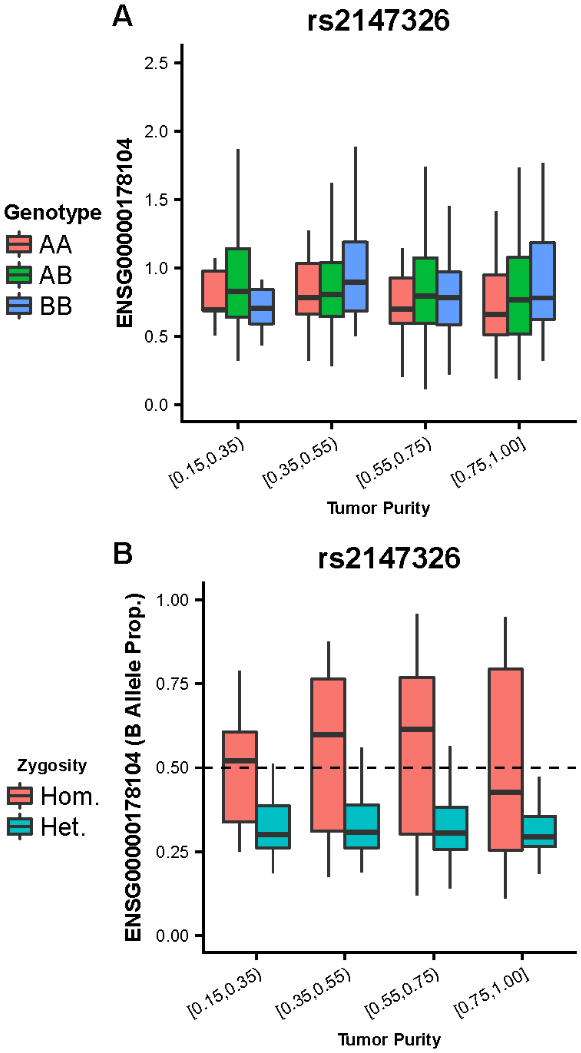
expression in tumor cells is lower than that in normal cells, but the eQTL effect is only present in tumor cells. These numerical estimates were well demonstrated by Figures 5.2A-B. As tumor purity increases, gene expression measured by TReC decreases (Figure 5.2A), and the strength of the eQTL increases. Both TReC and ASE show consistent signals that the B allele has higher expression, with a “Cis-Trans” test p-value of 0.95.

To highlight the functioning of the pTReC model, another example of tumor-specific eQTL was identified and shown in Figure 5.2C. In this example, gene expression from the ASE model was not used for eQTL mapping due to a significant “Cis-Trans” test using the full model. In this example, gene expression is higher in tumor compared to normal cells and the B allele has lower expression than the A allele in tumor cells, but not in normal cells. Note that we can still see some signals of an eQTL in the category with the lowest tumor purity. This results from TCGA samples being selected to have relatively higher tumor purity, thereby creating a categorization schema wherein even the lowest tumor purity category has a non-negligible amount of tumor cells.

We use another example to demonstrate the utility of the Cis-Trans score test (Figure 5.3). Considering only the TReC data, the B allele has slightly higher expression than the A allele when tumor purity is high (Figure 5.3A). In contrast, considering only the ASE data, the B allele has much lower expression than A across all tumor purity levels. This inconsistency between TReC and ASE data led to a highly significant Cis-Trans p-value (Figure 5.3C). In such cases, only the TReC data is trusted and used to estimate eQTL effects. ASE tends to be noisier in real data as mapping biases, incorrect genotype data, and/or other biological and technical factors can lead to the observed ASE imbalance as opposed to eQTL effects. Failure to consider the Cis-Trans test could lead to the acceptance of spurious eQTL results.

Next, we systematically compare the results for all eQTLs using the pTReCASE, TReCASE, and pLR approaches at various p-value cutoffs. One way to compare the results is to check the overlap of each significant eQTL association, i.e., each gene-SNP pair (Supplementary Table 2). However, due to LD, the expression of one gene may be associated with multiple SNPs that are in close proximity to one another and often represent redundant eQTL signals. Therefore, we focus on the eQTL results summarized at the gene level. In other words, for a given p-value cutoff, we count the number of genes with at least one eQTL with a p-value falling below the cutoff (Table 1).

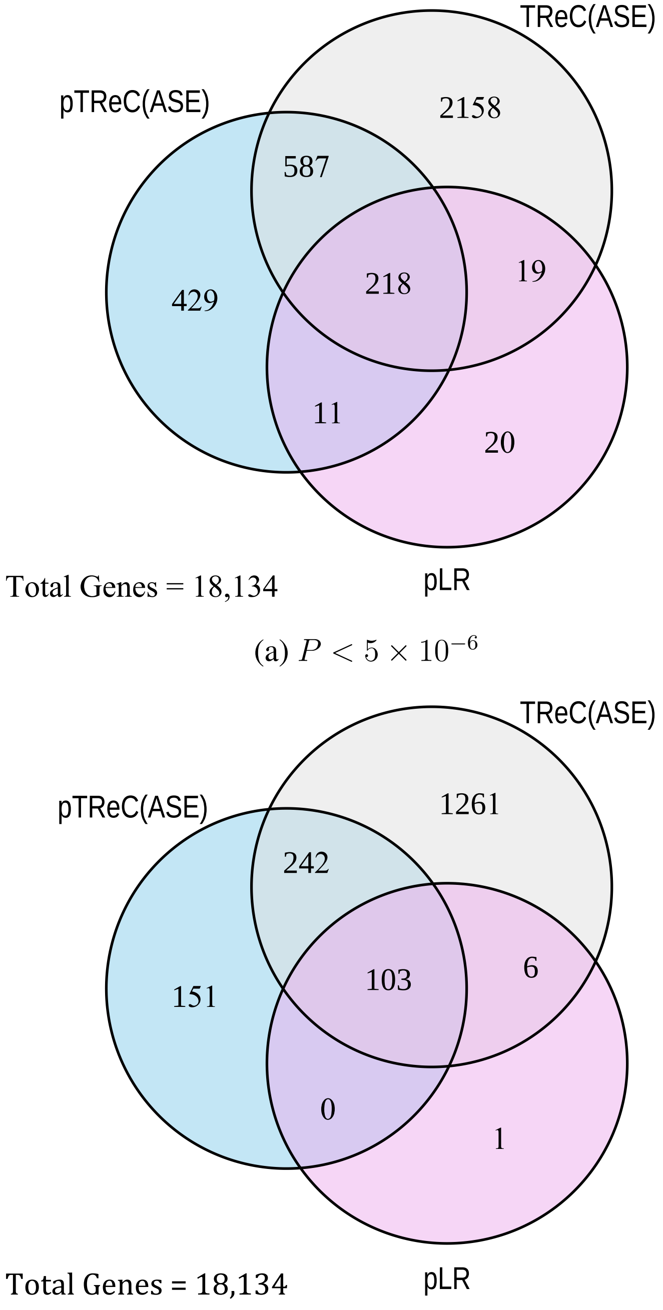
Compared to pTReCASE, the TReCASE model identifies eQTLs in a larger number of genes. For those genes where TReCASE identifies a significant eQTL and pTReCASE does not, the significant findings of the TReCASE model are most likely driven by an eQTL in normal tissue.



C

ENSG00000178104, rs2147326		
Model	*pTReCASE*	*pTReC*
(κ, η, γ)	(0.95,0.61,0.57)	(0.79, 0.86,1.08)
p (γ≠1)	<1.0e-10	0.67
P_{CT}	1.0e-10	--

Figure 5.3: Demonstrating the utility of the Cis-Trans score test. (A) Covariate-corrected total expression plotted as a function of genotype and tumor purity. (B) Allele Specific Expression with respect to genotype and tumor purity. (C) Table containing relevant modeling information for A and B. p_{CT} provides the p-value of the Cis-Trans score test.



(a) $P < 5 \times 10^{-6}$

(B) $P < 5 \times 10^{-8}$

Figure 5.4: Visual summary of the number of genes with at least significant eQTL at the prescribed p-value cutoff for each model fit and their overlaps.

		Genes		
P-value Cutoff	Category	pTReC(ASE)	TReC(ASE)	pLR
5×10^{-6}	# of Genes	1245	2982	268
	overlap/alternative	–	27.0	85.4
	overlap/pTReC(ASE)	–	64.7	18.4
5×10^{-8}	# of Genes	496	1612	110
	overlap/alternative	–	21.4	93.6
	overlap/pTReC(ASE)	–	69.6	20.8

Table 5.1: Summarizing the results of pTReC(ASE), TReC(ASE), the Westra-inspired models for TCGA data. Here the notation pTReC(ASE) indicates that we used the pTReCASE or pTReC model depending on the results of the Cis-Trans test. "Overlap" represents the genes with at least one significant eQTL identified by both pTReC(ASE) and an alternative method. "Overlap/alternative" is the number of overlaps divided by the number of findings using the alternative method. "Overlap/pTReC(ASE)" is the number of overlaps divided by the number of findings using pTReC(ASE). If we consider the results of pTReC(ASE) as true findings, then "overlap/alternative" is the true discovery rate and "overlap/pTReC(ASE)" is the sensitivity

TReCASE recaptures around two-thirds of eQTL findings identified by pTReCASE. The one-third missed by TReCASE are more likely to have weaker effect size and/or are only present in tumor cells.

Across p-value thresholds, the pLR model identifies relatively fewer significant gene-SNP pair relationships. Of those relationships identified by the pLR model, 85- to 93- percent are also identified by pTReCASE. The pLR model also misses at least 73% of significant results identified by pTReCASE. Possible reasons for the poor performance of the pLR model could arise from the fact that ASE is not incorporated and/or the relationship between transformed gene expression and tumor purity is not examined on the linear scale.

At the time of this writing, these authors were unable to find many studies seeking to identify eQTL in breast cancer tissues. Of those available, none utilized information regarding tumor purity when constructing analysis models which may limit the comparability of results. However, we identified one such study of breast cancer eQTL which controls for both somatic copy number and methylation [53]. We compare the results of pTReCASE against those in Li to determine the extent of overlap in their eQTL calls [53]. At the 5e-6 significance level, Li identifies 165 genes in common with pTReCASE whereas 91 genes would be expected by chance. In fact, the hypergeometric probability of observing 165 or more genes in common is 1.6.e-14. At the 5e-8 level, Li identifies 86 genes in common with pTReCASE whereas 37 genes would be expected by chance (hypergeometric $p = 3.5e - 14$). Note, Li et al examine eQTL using a reduced set of SNPs, a reduced sample size, and expression data derived from microarrays and not RNA-seq. The lack of stronger overlap between Li et al and pTReCASE could be due to Li's use of microarray expression instead of RNA-seq, smaller set of considered SNPs, and smaller sample size of breast cancers.

5.3.2.3 Assessing Copy Number Effects

Within tumor samples, copy number aberrations (CNA) are pervasive. Involving the addition or deletion of copies of genomic loci (e.g. SNPs or genes), CNA can increase or decrease the expression of various genes by virtue of adding or deleting copies of this gene. At present, the pTReC and pTReCASE methods do not assess the impact of copy number aberration on gene expression. Future extension of these models is needed to account for such impacts. However, we contend that CNA are unlikely to result in false positive results for pTReC and pTReCASE, instead resulting in a likely loss of power.

To examine the extent of copy number aberration within the TCGA dataset examined above, we define the following terms. C_{ij} is the total copy number of gene i in sample j. N_i is the ploidy of subject i. To determine the impact of CNA on a subject's gene expression, consider the metric D_{ij} where:

$$D_{ij} = C_{ij} - N_i.$$

Thus, D_{ij} represents the difference between a gene's copy number and the average copy number across all loci within the subject. For a single subject and gene, we define a copy number event to occur whenever $|D_{ij}| > 0.5$.

To motivate the use of D_{ij} to define copy number events, compare a sample a with no CNA across the entire genome to a sample b with CNA such that, on average, any gene is expected to have one additional copy (e.g. $N_i = 3$). Assuming identical expressions of each copy of a gene in the two samples and equivalent sampling depth during measurement, sample b would experience gene expressions 1.5 times larger than those in sample a. However, if one were to relativize expression within these samples (i.e. using FPKM, FPKM-UQ), the CNA across all genes within sample b would be eliminated. This is a direct result of the fact that the measure of read-depth would incorporate a consistent increase in expression due to the CNA seen in b. Thus, CNA should only impact relative gene expression when a gene's copy number is different from the average such copy number across the genome. Since pTReC and pTReCASE incorporate measures of read-depth within their mean structures, these models benefit from the properties of relativized expression comparisons.

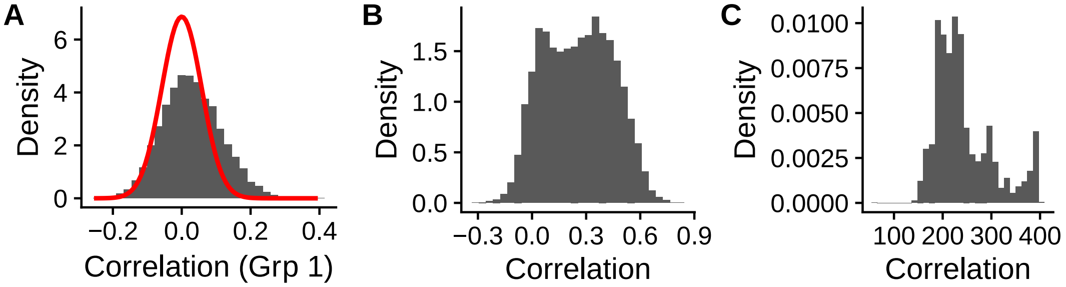


Figure 5.5: Evaluating the extent of copy number aberration within the TCGA dataset. [A] Distribution of the correlations between D_{ij} and C_{ij} for subjects where $|D_{ij}| \leq 0.5$ summarized across all 18,134 genes. Red line indicates density of N(0,1/$\sqrt{296}$). [B] Distribution of the correlations between D_{ij} and relative gene expression summarized across all 18,134 genes [C] The distribution of the number of subjects with $|D_{ij}| > 0.5$ across all 18,134 genes.

To justify use of $|D_{ij}| > 0.5$ to define copy number events, consider figure (5.5). Focusing on subjects such that $|D_{ij}| \leq 0.5$, or the group of subjects assumed not to have experienced a copy number event, we would expect to see no relationship between relative gene expression (Y_{ij}/d_i) and C_{ij}. Thus, assuming relative expression can help mitigate the effects of CNA, the distribution of the correlation between C_{ij} and (Y_{ij}/d_i) should be approximately 0. Indeed, panel (A) suggests this to be a reasonable assumption. While there is a slight positive skew, the correlation strengths remain low and are reasonably consistent with their expected distribution assuming the true mean is 0. As an additional note, the expected distribution is computed assuming that all correlations are computed with the same group size. However, the number of subjects not experiencing a CNA varies across gene and thus is not meant to provide an exact alignment.

Further, consider the relationship between a variable

$$G_{ij} = \begin{cases} 0 & \text{if } D_{ij} < -0.5, \\ 1 & \text{if } |D_{ij}| \leq 0.5 \\ 2 & \text{if } D_{ij} > 0.5 \end{cases}$$

and relativized expression (Y_{ij}/d_i). Under the argument posited above, one would expect to see a positive correlation between these two variables. This would indicate that as one increases the CN of a gene relative to the ploidy of the individual, the relative expression should also increase. Panel (B) provides support for this assertion, demonstrating positive correlations the bulk of which fall between 0.0 and 0.6.

Thus, considering the support displayed above, panel (C) suggests that approximately 75% of genes have 50% or fewer subjects impacted by copy number events. Indeed, copy number aberration is pervasive within the TCGA dataset.

However, the presence of copy number aberrations is unlikely to indicate the presence of false positive calls from pTReC and pTReCASE. If the correlation between copy number and eQTL genotype is weak or non-existent, CNA would add noise to the statistical models but would not induce false signal. To substantiate this claim within the TCGA dataset, we examine the correlation between D_{ij} and eQTL genotype. For each gene with at least one significant eQTL at the 5×10^{-6} level, we select its most significant SNP and compute the correlation between D_{ij} and the genotype of this SNP. Figure (5.6) plots the distribution of these correlations across the 1,271 significant genes with the red-line indicating the expected distribution assuming that these correlations have mean 0. Figure (5.6) demonstrates that these correlations are often weak, and thus pTReC and pTReCASE are unlikely to suffer from false positives due to CNA.

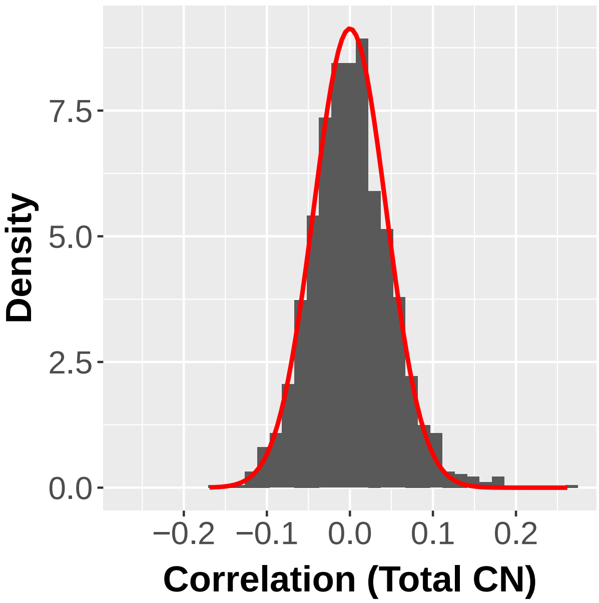


Figure 5.6: Computes the correlation between the eQTL genotype of a gene's most significant SNP and the gene's copy number difference D_{ij}

5.4 Discussion

Due to contamination of tumor samples with infiltrating normal cells, the identification of eQTL within tumor tissues poses several challenges. First and foremost, one needs to separately estimate the eQTL signals in tumor and normal cells. Second, while total gene expression has been widely used for transcriptome studies, it is important to leverage the additional information provided by allele-specific expression which can be effectively derived using RNA-seq data. We have developed a statistical model and software package, pTReCASE, to address these issues. The desirable performance of pTReCASE has been validated using simulations and real data analysis. In constrast, a naïve approach for eQTL mapping that ignores tumor purity may lead to a large fraction of false positives.

Readers may note that the mean structure utilized by pTReCASE involves two critical assumptions: (1) Expression in the tumor is composed of two cell groups, tumor and normal; and (2) the mean structure precludes the modeling of dominant and recessive eQTL effects. As mentioned in the introduction, intra-tumor heterogeneity is an ongoing challenge in the analysis of cancer expression data that provides a natural challenge to the assumption in (1). However, assumption

(1) allows pTReCASE to identify eQTL effects that are common to the majority of tumor cells across samples. Further refinement of eQTL effects into effects arising within different subclones is likely not possible due to the high degree of difference between subclones across cancers. Thus, assumption (1) does not seek to imply that it is impossible for certain subclones to experience different eQTL regulation. Instead it restricts the ability of pTReCASE to identify eQTL that are present in poorly represented subclones.

With regard to assumption (2), the additive structure used by pTReCASE is a natural consequence of cis-acting regulation. Should dominance and recessive relationships exist, it is unlikely to result from cis-acting regulation and thus one should not incorporate ASE information in the model. The pTReC model could be modified to capture dominance and recessive relationships in future studies.

In addition, the effects of γ and η on mean expression through formulas involving ξ_i rather than introducing a genotype covariate directly into the regression equation. The relationship between genotype and genetic expression is assumed to occur on the linear scale. The relationship between genetic expression and covariates is assumed to occur on the log-scale. Thus, replacing η and γ with extra elements in the covariate vector β is impossible due to the different scales.

Within the current established framework for pTReC(ASE), there are three additional avenues for further development and research. The first is to improve the computational efficiency of our software package. Using the current implementation, it takes about thousands of CPU hours for genome-wide local eQTL mapping. This can be easily done using a moderately sized computing cluster, but is not computationally feasible for a single computer. High computational costs also prevent us from using permutations to assess the significance of eQTL results. Thus, we recommend use of Bonferroni correction, Benjamini-Hochberg FDR control [80], or calculation of the number of independent tests by examining the correlation structure of the genotype data [81].

We have assumed that the haplotypes connecting candidate eQTLs and the SNPs within the gene body are known. In practice, such haplotypes are imputed/phased using statistical methods. Phasing is usually accurate within short genetic distances around the gene of interest. However, if we would like to consider potential eQTLs further from the gene, there is a possibility of phasing error. The second avenue for improving the posited model is to allow for uncertainty in the haplotype phasing by following the approach of Hu et al [56].

Lastly, both the TReC and ASE components of the pTReCASE model have assumed no copy number change across subjects. Suppose that an eQTL for a given gene modifies expression through copy number changes. More specifically, suppose the B allele at the candidate eQTL is associated with a larger copy number in tumor tissues. This would lead to increased expression from the B allele and would be interpreted by the pTReCASE model as higher expression of B allele. In such a scenario, the pTReCASE model is applicable without modification. However, if both eQTL and copy number affect gene expression independently, the pTReCASE model should be adjusted for copy number differences. We have demonstrated in section 3.2.3 that, despite not controlling for pervasive CNA in its analysis, pTReC(ASE) is unlikely to have experienced false positive calls as the correlation between SNP genotype and gene copy number are often weak.

Given estimates of allele-specific copy number, our model can be modified to address copy number variation across subjects. Estimation of allele-specific copy number in tumor samples is a very challenging task due to the confounding effects of tumor purity, ploidy, and the possibility of subclonal copy number changes [82, 83]. It is desirable to systematically study the effects of both germline SNPs, somatic copy number changes, and even somatic point mutations (single nucleotide variants or indels) while also accounting for intra-tumor heterogeneity, but such explorations are beyond the scope of this paper and warrant a series of future studies.

CHAPTER 6: CONCLUSIONS AND FUTURE WORK

Within this dissertation, we have explored the analysis of bulk expression experiments from heterogeneous cell type samples. Two main problems from such data have been examined: estimating the cell type abundance profile from which the mixture expressions were generated and conducting cell-type specific differential expression analyses controlling for cell type abundance. While the abundance estimation and differential expression analyses are of interest in general cell type samples, we focused the majority of our model development and application on tumor tissue, namely estimating immune cell abundance within tumor samples and identifying tumor-specific eQTL.

Within chapter 3, we posited an RNA-seq framework (IsoDeconv) for cell type abundance estimation using isoform expression information. The presented work displays the promise of isoform expression for use within the cell type deconvolution setting, but challenges remain. As noted in chapter 3, future work on the IsoDeconv model should focus on reducing its computational complexity and refining its application to real data settings. In particular, excessive variation has been shown to hamper the estimation process. Strategies for mitigating the impact of this variation have been proposed but require further development. Improving IsoDeconvs ability to handle higher levels of expression variance is critical for its use in capturing the additional information regarding cell type identity provided by isoform expression.

Chapter 4 provided an alternative cell type abundance estimation framework for use within tumor tissues. The presented model, ICeD-T, was found to be both accurate and robust to aberrant gene behavior within mixture expression experiments. Future work on the ICeD-T model should focus on the development of superior reference matrices for immune cell deconvolution and a refined set of variance weights for use in the estimation process. By refining the reference matrices and variance weights, the accuracy of ICeD-Ts estimates and its ability to detect aberrant gene behavior is likely to improve. Single cell RNA-seq is a promising technology for the development of superior reference matrices and variance weights. scRNA-seq experiments represent an opportunity to capture reference immune cell expressions within the tumor context. As the variability and accuracy of scRNA-seq improves, so too will our ability to accurately capture immune cell expression within references.

Finally, the problem of cell-type specific eQTL was addressed within chapter 5. Our model, pTReCASE, was demonstrated to properly control Type I error and provide superior power in the analysis of tumor-specific eQTL when compared to alternative modeling strategies. One avenue for future improvement of the pTReCASE model includes the incorporation of copy number aberration data. While it was found that copy number aberration data was unlikely to introduce false positives to the pTReCASE model, incorporating CNA data could improve its power in tumors types with high levels of CNA. In addition, extension of pTReCASE to the study of somatic mutations could reveal the biological processes behind the behavior of tumor subclones. While the study of somatic mutations is promising, such work must overcome the challenge of estimating intra-tumor heterogeneity, an ongoing topic of research.

APPENDIX A: SUPPLEMENT FOR CHAPTER 3

A.1 IsoDeconvNB: Supplementary Methods

A.1.1 Notation Table

<table>
<tr><th colspan="3">Pure Sample Expressions</th></tr>
<tr><th>Value</th><th>Dim.</th><th>Description</th></tr>
<tr><td>Y_{kje}</td><td>1×1</td><td>Total read count at exon set e in pure sample j of cell type k.</td></tr>
<tr><td>t_{kj}</td><td>1×1</td><td>Measure of read-depth (e.g. total read count) in pure sample j of cell type k.</td></tr>
<tr><td>γ_k</td><td>$I \times 1$</td><td>A vector of Isoform expression levels unique to cells of type k.</td></tr>
<tr><td>ν_k</td><td>1×1</td><td>An overdispersion parameter governing read count variance at exon sets for expression in cells of type k.</td></tr>
<tr><td>n_k</td><td>1×1</td><td>Number of pure samples of cells of type k.</td></tr>
<tr><th colspan="3">Mixture Sample Expressions</th></tr>
<tr><th>Value</th><th>Dim.</th><th>Description</th></tr>
<tr><td>Z_e</td><td>1×1</td><td>Read count at exon set e in the mixture cell type sample.</td></tr>
<tr><td>Y_{ke}^*</td><td>1×1</td><td>Unobserved read count attributable to cells of type k at exon set e in the mixture.</td></tr>
<tr><td>t_m</td><td>1×1</td><td>Measure of read-depth (e.g. total read count) in the mixture sample.</td></tr>
<tr><td>p_k</td><td>1×1</td><td>Abundance of cell type k in the mixture sample.</td></tr>
<tr><td>p</td><td>$K \times 1$</td><td>Vector of cell type abundances in the mixture across all K cell types.</td></tr>
<tr><th colspan="3">Cluster Level Parameters</th></tr>
<tr><th>Value</th><th>Dim.</th><th>Description</th></tr>
<tr><td>X</td><td>$E \times I$</td><td>Matrix of effective lengths for each exon set within each of the isoforms.</td></tr>
<tr><td>X_{ij}</td><td>1×1</td><td>Effective length of gene i in isoform j.</td></tr>
<tr><td>X_e</td><td>$I \times 1$</td><td>e-th row of effective length matrix X.</td></tr>
<tr><th colspan="3">Gamma-Poisson Mixture Parameters</th></tr>
<tr><th>Value</th><th>Dim.</th><th>Description</th></tr>
<tr><td>λ_{kje}</td><td>1×1</td><td>Unobserved, gamma-distributed, Poisson-mean read count at exon set e for pure sample j of cell type k.</td></tr>
<tr><td>λ_{ke}^*</td><td>1×1</td><td>Unobserved, gamma-distributed, Poisson-mean read count at exon set e for cells of type k in the mixture.</td></tr>
<tr><td>λ_e^*</td><td>$K \times 1$</td><td>Vector of the λ_{ke}^* values such that $\boldsymbol{\lambda}_e^* = [\lambda_{1e}^*, ..., \lambda_{Ke}^*]$.</td></tr>
<tr><th colspan="3">Function Definitions</th></tr>
<tr><th>Value</th><th>Dim.</th><th>Description</th></tr>
<tr><td>$\Psi_0()$</td><td>NA</td><td>The digamma function, defined as the first derivative of the lnΓ () function.</td></tr>
</table>

Table A.1: Notation for defining the IsoDeconv Model.

A.1.2 Optimization Algorithm

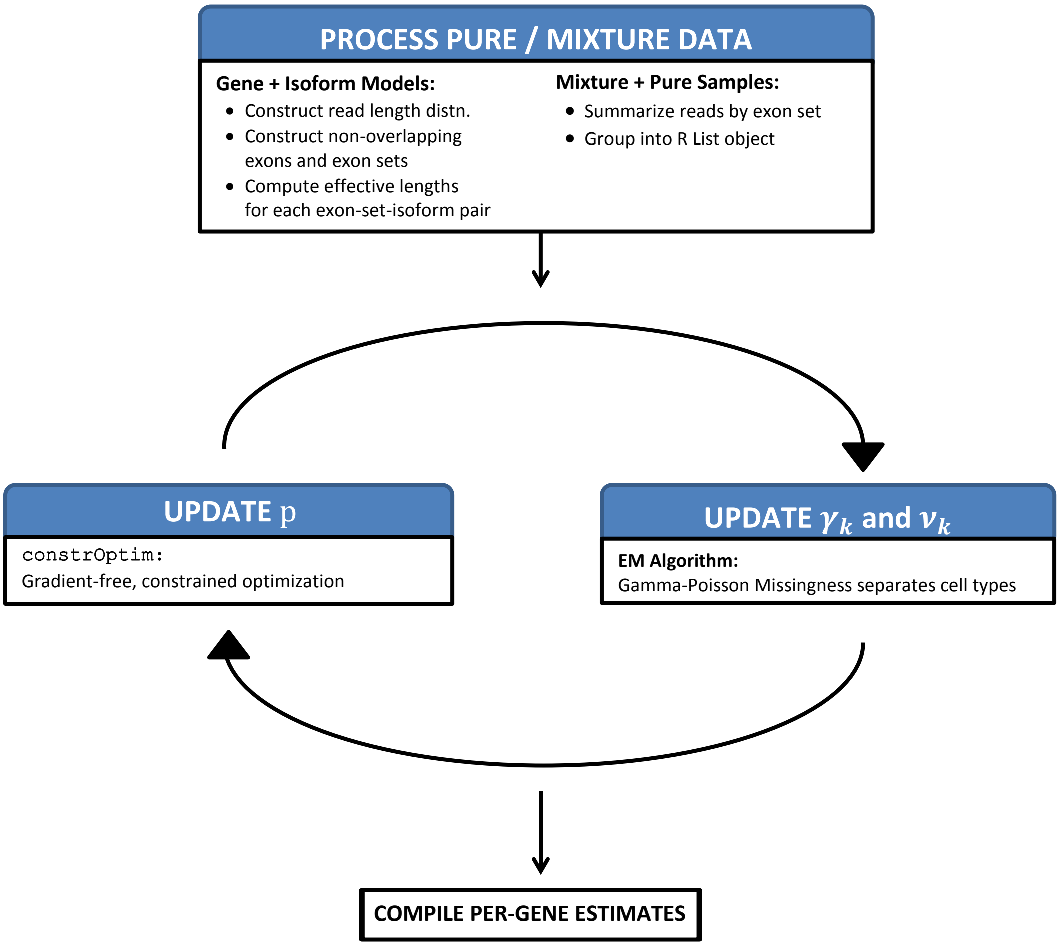

PROCESS PURE / MIXTURE DATA
Gene + Isoform Models:
• Construct read length distn.
• Construct non-overlapping exons and exon sets
• Compute effective lengths for each exon-set-isoform pair
Mixture + Pure Samples:
• Summarize reads by exon set
• Group into R List object
UPDATE p
constrOptim:
Gradient-free, constrained optimization
UPDATE γ_k and ν_k
EM Algorithm:
Gamma-Poisson Missingness separates cell types
COMPILE PER-GENE ESTIMATES

Figure A.1: Visual representation of the IsoDeconv Negative Binomial algorithm from early stage data processing to iterative update algorithm.

Model parameters in the IsoDeconv model are estimated through a maximum-likelihood framework on a gene-by-gene basis. An estimate of p_k is obtained for each gene and then aggregated across genes to provide a final abundance estimate. Within each gene, optimization proceeds via block coordinate descent. The steps are as follows:

(1) Assume the p_k are fixed, update γ_k and ν_k.

(2) Assume the γ_k and ν_k are fixed, update the p_k.

Steps (1) and (2) are cycled until convergence of the cell type proportions, p_k. Within step (2), optimization of the p_k terms proceeds using a gradient-free, numerical optimization routine, R's `constrOptim`. Use of the gradient free approach is incorporated to avoid the intractability of a gradient function in the convolution of negative binomials model. Within step (1), optimization proceeds under an EM algorithm, described in section 4 of Appendix 1. We derive some useful results first.

A.1.3 Derivations involving Gamma-Poisson Random Variables

The optimization procedure utilized by IsoDeconv relies upon the categorization of a negative binomial random variable as a gamma-poisson mixture. To see this, consider the following hierarchical framework for two random variables Y and λ:

$$Y\big|\lambda \sim \texttt{Poisson}\left(\lambda\right) \quad \text{and} \quad \lambda \sim \Gamma\left(\mu, \nu\right)$$

Integrating out the random variable λ, we obtain the marginal distribution of Y:

$$\begin{aligned}
f_Y(y) &= \int_0^\infty f_{Y|\lambda}(y,\lambda) f_\lambda(\lambda) \partial\lambda \\
&= \int_0^\infty \left(\frac{\lambda^Y e^{-\lambda}}{y!}\right)\left(\frac{\lambda^{\nu-1}(\nu/\mu)^\nu e^{-(\nu/\mu)\lambda}}{\Gamma(\nu)}\right)\partial\lambda \\
&= \left[\frac{(\nu/\mu)^\nu}{\Gamma(\nu)y!}\right]\int_0^\infty \lambda^{y+\nu-1}e^{-(\nu/\mu+1)\lambda}\partial\lambda \\
&= \left[\frac{(\nu/\mu)^\nu}{\Gamma(\nu)y!}\right]\left[\frac{\Gamma(y+\nu)}{(\nu/\mu+1)^{y+\nu}}\right]\int_0^\infty \frac{\lambda^{y+\nu-1}(\nu/\mu+1)^{y+\nu}e^{-(\nu/\mu+1)\lambda}}{\Gamma(y+\nu)}\partial\lambda \\
&= \left[\frac{\Gamma(y+\nu)}{\Gamma(\nu)y!}\right]\left(\frac{\nu}{\nu+\mu}\right)^\nu\left(\frac{\mu}{\nu+\mu}\right)^y
\end{aligned}$$

We recognize this as the density function of a negative binomially distributed random variable, as desired. It is clear from the properties of gamma random variables, poisson random variables, and conditional expectations that:

$$\begin{aligned}
E[Y] &= E\left\{E[Y|\lambda]\right\} \\
&= E\left[\lambda\right] \\
&= \mu \\
V[Y] &= E\left\{V[Y|\lambda]\right\} + V\left\{E[Y|\lambda]\right\} \\
&= E[\lambda] + V[\lambda] \\
&= \mu + (1/\nu)\mu^2
\end{aligned}$$

where ν can be characterized as our overdispersion parameter for the negative binomial distribution.

The optimization algorithm that follows relies upon certain conditional expectations developed under this framework. Thus, we specify the following additional necessities constructed from Gamma-Poisson mixtures.

(1) Conditional Distribution of $\lambda|Y$

(2) Conditional Expectation of $\lambda|Y$

(3) Conditional Expectation of $\log(\lambda)|Y$

(4) Sums of Gamma-Poisson Mixtures

Each of these necessities are developed below.

<u>D.1 - Conditional Distribution of $\lambda|Y$:</u>

By properties of conditional distributions, we know that

$$\begin{aligned} f_{\lambda|Y}(\lambda, y) &\propto f_{Y|\lambda}(y, \lambda) f_\lambda(\lambda) \\ &\propto \left(\lambda^Y e^{-\lambda}\right)\left(\lambda^{\nu-1} e^{-(\nu/\mu)\lambda}\right) \\ &= \lambda^{Y+\nu-1} e^{-(\nu/\mu+1)\lambda} \end{aligned}$$

We recognize this as the kernel of a gamma distribution. Thus, it is clear that:

$$\lambda|Y \sim \Gamma\left(\mu' = \frac{Y+\nu}{\nu/\mu+1}, \nu' = Y+\nu\right)$$

D.2 - Conditional Expectation of $\lambda|Y$:

By properties, of the gamma distribution we know that

$$E[\lambda|Y] = \mu' = \frac{Y+\nu}{\nu/\mu+1}$$

D.3 - Conditional Expectation of $\log(\lambda)|Y$:

Consider the moment generating function (MGF) of the random variable $R = \log(\lambda)$ given Y.

$$\begin{aligned}
M_{R|Y}(t) = E\left[e^{tR}\middle|Y\right] &= E\left[e^{t\log(\lambda)}\middle|Y\right] \\
&= E\left[\lambda^t|Y\right] \\
&= \int_0^\infty \lambda^t \left(\frac{\lambda^{\nu'-1}(\nu'/\mu')^{\nu'} e^{-(\nu'/\mu')\lambda}}{\Gamma(\nu')}\right)\partial\lambda \\
&= \left[\frac{\Gamma(\nu'+t)(\mu'/\nu')^t}{\Gamma(\nu')}\right]\int_0^\infty \frac{\lambda^{\nu'+t-1}\left(\nu'/\mu'\right)^{\nu'+t} e^{-(\nu'/\mu')\lambda}}{\Gamma(\nu'+t)} \quad \partial\lambda \\
&= \left[\frac{\Gamma(\nu'+t)(\mu'/\nu')^t}{\Gamma(\nu')}\right]
\end{aligned}$$

The existence of this MGF implies the existence of the moments of R. We compute the first such moment by taking the derivative of this MGF and evaluating at $t = 0$.

$$\begin{aligned}
E[R|Y] &= \left.\frac{\partial M_{R|Y}(t)}{\partial t}\right|_{t=0} \\
&= \left.\frac{\dot{\Gamma}(\nu'+t)}{\Gamma(\nu')}(\mu'/\nu')^t + \frac{\Gamma(\nu'+t)}{\Gamma(\nu')}(\mu'/\nu')^t\left[\log(\mu') - \log(\nu')\right]\right|_{t=0} \\
&= \Psi_0(\nu') + \log(\mu'/\nu') \\
&= \Psi_0(\nu+Y) - \log\left(\nu/\mu+1\right)
\end{aligned}$$

D.4 - Sums of Gamma-Poisson Mixtures:

Consider a framework wherein we have K independent Gamma-Poisson mixtures and are examining their sum Z. Symbolically, we are examining:

$$Z = \sum_{i=1}^{K} Y_k$$

where

$$\begin{aligned} Y_k|\lambda_k &\sim \texttt{Poisson}(\lambda_k) \\ \lambda_k &\sim \texttt{Gamma}(\mu_k, \nu_k) \end{aligned}$$

and

$$(Y_j, \lambda_j) \perp (Y_k, \lambda_k) \quad \forall j \neq k$$

Now, we'll consider variations of D.2 and D.3 defining instead the following conditional frameworks:

(D.3.1) Conditional Expectation of $\lambda_k|Z$

(D.3.2) Conditional Expectation of $\log(\lambda_k)|Z$

For the first conditional expectation, the properties of double expectation in combination with D.2 provide:

$$\begin{aligned} E[\lambda_k|Z] &= E\left\{E[\lambda_k|Y_1, ..., Y_K, Z]\Big|Z\right\} \\ &= E\left\{E[\lambda_k|Y_1, ..., Y_K]\Big|Z\right\} \\ &= E\left\{\frac{Y_k + \nu_k}{\nu_k/\mu_k + 1}\middle|Z\right\} \\ &= \sum_{y=0}^{Z} \left(\frac{y + \nu_k}{\nu_k/\mu_k + 1}\right) f_{Y_k|Z}(y) \end{aligned}$$

For the second conditional expectation, the properties of double expectation in combination with D.3 provide:

$$\begin{aligned} E\left[\log(\lambda_k)\middle|Z\right] &= E\left\{E[\log(\lambda_k)|Y_1, ..., Y_K, Z]\middle|Z\right\} \\ &= E\left\{E[\log(\lambda_k)|Y_1, ..., Y_K]\middle|Z\right\} \\ &= E\left[\Psi_0(Y_k + \nu_k)\middle|Z\right] - \log\left(\nu_k/\mu_k + 1\right) \\ &= \sum_{y=0}^{Z} \Psi_0(Y_k + \nu_k) f_{Y_k|Z}(y) \end{aligned}$$

We note that the remaining expectations are now finite sums over the support of Y_k from 0 to Z. Thus, in order to compute this value, we need only the following piece:

$$f_{Y_k|Z}(y, z) = \frac{f_{Y_k}(y) f_{Z-Y_k}(z - y)}{f_Z(z)}$$

Using an approximation to the distribution of a convolution of negative binomial pieces utilized by Efron, we can approximate the densities $f_{Z-Y_k}()$ and $f_Z()$. Thus, we render this expectation computable. If $K = 2$, no approximation is necessary as $f_Z()$, f_{Y_1}, and $f_{Z-Y_1}() = f_{Y_2}$ are easily computed.

A.1.4 EM Algorithm: Update γ_k and ν_k

Recall the likelihood framework established within the text of “IsoDeconv: Cell Type Abundance Estimation using RNA Isoform Expression”. We restate it briefly here for completeness utilizing the notation specified in the table in section A.1. For purified samples of cells of type k, we have:

$$Y_{kje}|\gamma_k, \nu_k \sim \texttt{Neg.Bin.}\left(\mu = t_{kj}X_e^T\gamma_k, \phi = 1/\nu_k\right)$$

Within the mixture tissue, reads attributable to cells of type k are assumed to arise from the following model:

$$\begin{aligned} Y_{ke}^*\big|\gamma_k, \nu_k &\sim \texttt{Neg.Bin.}\left(\mu = t_m p_k X_e^T\gamma_k, \phi = 1/\nu_k\right) \\ Z_e &= \sum_{k=1}^{K} Y_{ke}^* \end{aligned}$$

In order to construct an EM algorithm, we need to represent the given stochastic system using missing values. The missing values in this setting become evident when we re-characterize the negative binomials as Gamma-Poisson mixtures. Within these mixtures, the poisson mean for each cell type’s read count is unknown–specified by λ_{kje} for pure samples and λ_{ke}^* in the mixture. Thus, we can specify a complete data, log-likelihood as follows:

$$
\begin{aligned}
\ell &= \sum_{e=1}^{E} \left\{ \sum_{k=1}^{K} \left(\sum_{j=1}^{n_k} \ell_{ke}^{(j)} \right) + \ell_{e}^{(m)} \right\} \\
&= \sum_{e=1}^{E} \left\{ \sum_{k=1}^{K} \left[\left(\sum_{j=1}^{n_k} \ell_{Y_{kje}|\lambda_{kje}} + \ell_{\lambda_{kje}} \right) + \ell_{\lambda^{*}_{ke}} \right] + \ell_{Z_e|\boldsymbol{\lambda^{*}_{e}}} \right\} \\
&= \sum_{e=1}^{E} \Bigg\{ \sum_{k=1}^{K} \Bigg[\Bigg(\sum_{j=1}^{n_k} -\lambda_{kje} + Y_{kje} \log(\lambda_{kje}) - \log(Y_{kje}) + (\nu_k - 1)\log(\lambda_{kje}) + \\
&\qquad\qquad \nu_k \left[\log(\nu_k) - \log(\mu_{kje})\right] - (\nu_k/\mu_{kje})\lambda_{kje} - \ln\Gamma\left(\nu_k\right) \Bigg) + \\
&\qquad (\nu_k - 1)\log(\lambda^{*}_{ke}) + \nu_k \left[\log(\nu_k) - \log(\mu^{*}_{ke})\right] - (\nu_k/\mu^{*}_{ke})\lambda^{*}_{ke} - \ln\Gamma\left(\nu_k\right) \Bigg] - \\
&\quad \left(\sum_{k=1}^{K} \lambda^{*}_{ke} \right) + Z_e \log \left(\sum_{k=1}^{K} \lambda^{*}_{ke} \right) - \log(Z_e) \Bigg\}
\end{aligned}
$$

Grouping like terms, we have:

$$
\begin{aligned}
\ell = \sum_{e=1}^{E} \Bigg\{ \sum_{k=1}^{K} \Bigg[&(\nu_k - 1) \left(\sum_{j=1}^{n_k} \log(\lambda_{kje}) + \log(\lambda^{*}_{ke}) \right) + \\
&\nu_k \left((n_k + 1)\log(\nu_k) - \sum_{j=1}^{n_k} \log(t_{kj}) - \log(t_m) - \log(p_k) \right) - \\
&\nu_k (n_k + 1) \log(X_e^T \gamma_k) - \left(\frac{\nu_k}{X_e^T \gamma_k} \right) \left(\sum_{j=1}^{n_k} \lambda_{kje}/t_{kj} + \lambda^{*}_{ke}/t_m p_k \right) \Bigg] + \\
&f\left(\mathbf{Y}_e, \boldsymbol{\lambda}_e, Z_e, \boldsymbol{\lambda}^{*}_{e}\right) \Bigg\}
\end{aligned}
$$

Thus, without explicitly specifying the utilized functions, it is clear that we can regroup by cell type to construct:

$$
\ell = \sum_{k=1}^{K} \left\{ \sum_{e=1}^{E} \ell_{e}^{k}(\gamma_k, \nu_k) + (1/K) f\left(\mathbf{Y}_e, \boldsymbol{\lambda}_e, Z_e, \boldsymbol{\lambda}^{*}_{e}\right) \right\}
$$

This regrouping makes it explicitly clear that we can perform optimization separately within each cell type.

Optimization in k-th Cell Type: Isoform Gradient

To assist in the numerical optimization routines used within the R software, we compute the gradient of the likelihood with respect to the isoform parameters. To this end, we specify the restricted portion of our likelihood containing only that information necessary for the optimization of the isoform parameters:

$$\ell_{\gamma}^{(k)} = -\nu_k \left[\sum_{e=1}^{E} (n_k + 1) \log(X_e^T \gamma_k) + \frac{\left(\sum_{j=1}^{n_k} \lambda_{kje}/t_{kj} + \lambda_{ke}^{*}/t_m p_k \right)}{X_e^T \gamma_k} \right]$$

Computing the gradient of this function with respect to the isoform parameters, we have:

$$\begin{aligned} \dot{\ell}_{\gamma}^{(k)} &= -\nu_k \left[\sum_{e=1}^{E} \left(\frac{n_k + 1}{X_e^T \gamma_k} \right) X_e - \frac{\left(\sum_{j=1}^{n_k} \lambda_{kje}/t_{kj} + \lambda_{ke}^{*}/t_m p_k \right)}{\left(X_e^T \gamma_k \right)^2} X_e \right] \\ &= -\nu_k X^T \Delta \end{aligned}$$

where

$$\Delta = \begin{bmatrix} \left(\frac{n_k+1}{X_1^T \gamma_k} \right) - \frac{\left(\sum_{j=1}^{n_k} \lambda_{kj1}/t_{kj} + \lambda_{k1}^{*}/t_m p_k \right)}{\left(X_1^T \gamma_k \right)^2} \\ \vdots \\ \left(\frac{n_k+1}{X_E^T \gamma_k} \right) - \frac{\left(\sum_{j=1}^{n_k} \lambda_{kjE}/t_{kj} + \lambda_{kE}^{*}/t_m p_k \right)}{\left(X_E^T \gamma_k \right)^2} \end{bmatrix}$$

Optimization in K-th Cell Type: Overdispersion Gradient

To assist in the numerical optimization routine used with the R software, we compute the gradient of the likelihood with respect to the overdisperion parameter ν_k. To this end, we specify the restricted portion of our likelihood containing only that information necessary for the optimization of the overdispersion parameter:

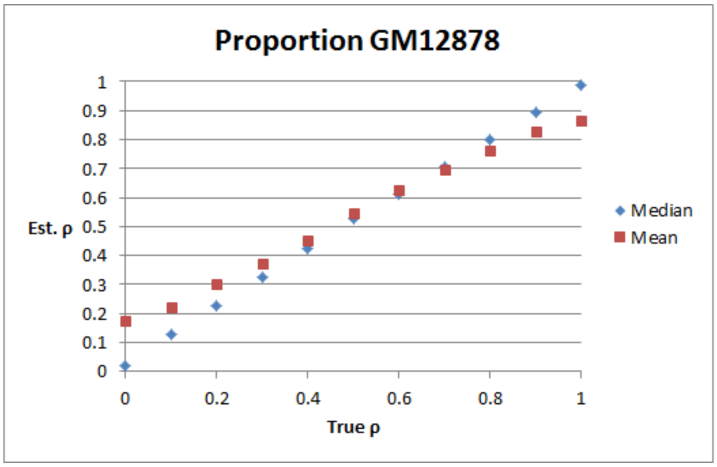
$$\ell_{\nu}^{(k)} = \sum_{e=1}^{E} \left[(\nu_k - 1) \left(\sum_{j=1}^{n_k} \log(\lambda_{kje}) + \log(\lambda_{ke}^{*}) \right) + \nu_k \left((n_k + 1) \log(\nu_k) - \sum_{j=1}^{n_k} \log(t_{kj}) - \log(t_m) - \log(p_k) \right) - \right.$$
$$\left. \nu_k (n_k + 1) \log(X_e^T \gamma_k) - \left(\frac{\nu_k}{X_e^T \gamma_k} \right) \left(\sum_{j=1}^{n_k} \lambda_{kje}/t_{kj} + \lambda_{ke}^{*}/t_m p_k \right) - (n_k + 1) \mathrm{ln}\Gamma\left(\nu_k\right) \right]$$

Computing the gradient of this function with respect to the overdispersion parameter ν_k, we have:

$$\dot{\ell}_{\nu}^{(k)} = \sum_{e=1}^{E} \left[\left(\sum_{j=1}^{n_k} \log(\lambda_{kje}) + \log(\lambda_{ke}^{*}) \right) + \right.$$
$$\left((n_k + 1) \log(\nu_k) - \sum_{j=1}^{n_k} \log(t_{kj}) - \log(t_m) - \log(p_k) \right) + (n_k + 1) -$$
$$\left. (n_k + 1) \log(X_e^T \gamma_k) - \left(\frac{\sum_{j=1}^{n_k} \lambda_{kje}/t_{kj} + \lambda_{ke}^{*}/t_m p_k}{X_e^T \gamma_k} \right) - (n_k + 1) \Psi_0(\nu_k) \right]$$

A.1.5 Choosing Aggregation Technique

Simulations similar to those in section 2.2.2.3 are performed to determine the best method for aggregating per-gene estimates of cell type abundance. To generate a set of simulated mixture and reference samples, one paired-end RNA-seq experiment each of GM12878 and HMEC were downloaded from the ENCODE database. Reference samples were generated similarly by again downsampling the GM12878 and HMEC files to 10 million reads apiece.



True P_g	Median	Mean
0.0	0.02	0.17
0.1	0.13	0.22
0.2	0.23	0.30
0.3	0.33	0.37
0.4	0.43	0.45
0.5	0.53	0.55
0.6	0.61	0.62
0.7	0.71	0.70
0.8	0.80	0.76
0.9	0.90	0.83
1.0	0.99	0.86

Figure A.2: Estimates of the proportion of GM12878 using Median and Mean aggregation of per-gene estimates.

The mixture sample generation methodology described here was found to be flawed. Oversampling of the GM12878 and HMEC files created references and mixtures which were exceedingly similar. Thus, the accuracy of the observed results was determined to be a function of this similarity and not the appropriateness of the IsoDeconvNB model. Despite this fact, these original examinations allowed determination of the most appropriate way to aggregate cell type abundance estimates across genes. As is seen in figure (A.2), the median per-gene estimate of cell type abundance provides superior estimation of the overall cell type abundance within the sample. Thus, in the following results, per-gene estimates of cell type abundance are aggregated using the median.

A.2 IsoDeconvMM: Supplementary Methods

A.2.1 Notation Table

Pure Sample Expressions		
Value	**Dim.**	**Description**
$Y_{kj(E)}$	1×1	Total read count outside gene of interest in pure sample j of cell type k.
Y_{kje}	1×1	Read count at exon set e in pure sample j of cell type k.
Y_{kj}	$E \times 1$	Collection of read counts across all exon sets in the given gene for pure sample j of cell type k.
γ_{kj}	$I \times 1$	Isoform expression parameters unique to pure sample j of cell type k.
τ_{kj}	1×1	Probability that a randomly selected read maps to the gene of interest in pure sample j of cell type k.
t_{kj}	1×1	The total read count in pure sample j of cell type k.
Mixture Sample Expressions		
Value	**Dim.**	**Description**
Z_e	1×1	Read count at exon set e in the mixture cell type sample.
Z	$E \times 1$	Collection of Z_e in a single vector.
$Z.$	1×1	Total number of reads mapping to gene of interest in the mixture $\left(Z. = \mathbf{1}^T Z = \sum_{e=1}^{E} Z_e\right)$.
Z_{ke^*}	1×1	Read count at exon set e in the mixture cell type sample attributable to cells of type k.
γ_k^*	$I \times 1$	Isoform expression parameters unique to cells of type k found within the mixture cell type sample.
τ_k^*	1×1	The probability that a randomly selected read from cells of type k in the mixture sample maps to the gene of interest which is unique to the cells in the mixture sample.
Cell-Type Specific and Cluster Level Parameters		
Value	**Dim.**	**Description**
X	$E \times I$	Matrix of effective lengths for each exon set within each of the isoforms.
X_{ij}	1×1	Effective length of exon set i in isoform j.
$\tilde{l}$	$I \times 1$	Vector of complete effective lengths of each utilized isoform across all exon sets $\left(\tilde{l}_j = \sum_{i=1}^{E} X_{ij}\right)$.
p_k	1×1	Proportion of cell type k present in the mixture tissue.
p	$K \times 1$	Collection of abundances for each of the K cell types which compose the mixture.
α_k	$I \times 1$	Hyperparameters governing average isoform expression levels and variances within cells of type k.
β_k	2×1	Hyperparameters governing gene expression levels within cells of type k.
Value	**Dim.**	**Description**
$\circ$	NA	This operator indicates element-wise multiplication of two vectors.

Table A.2: Notation for defining the IsoDeconv Model.

A.2.2 Lemmas Involving Multinomial Distribution

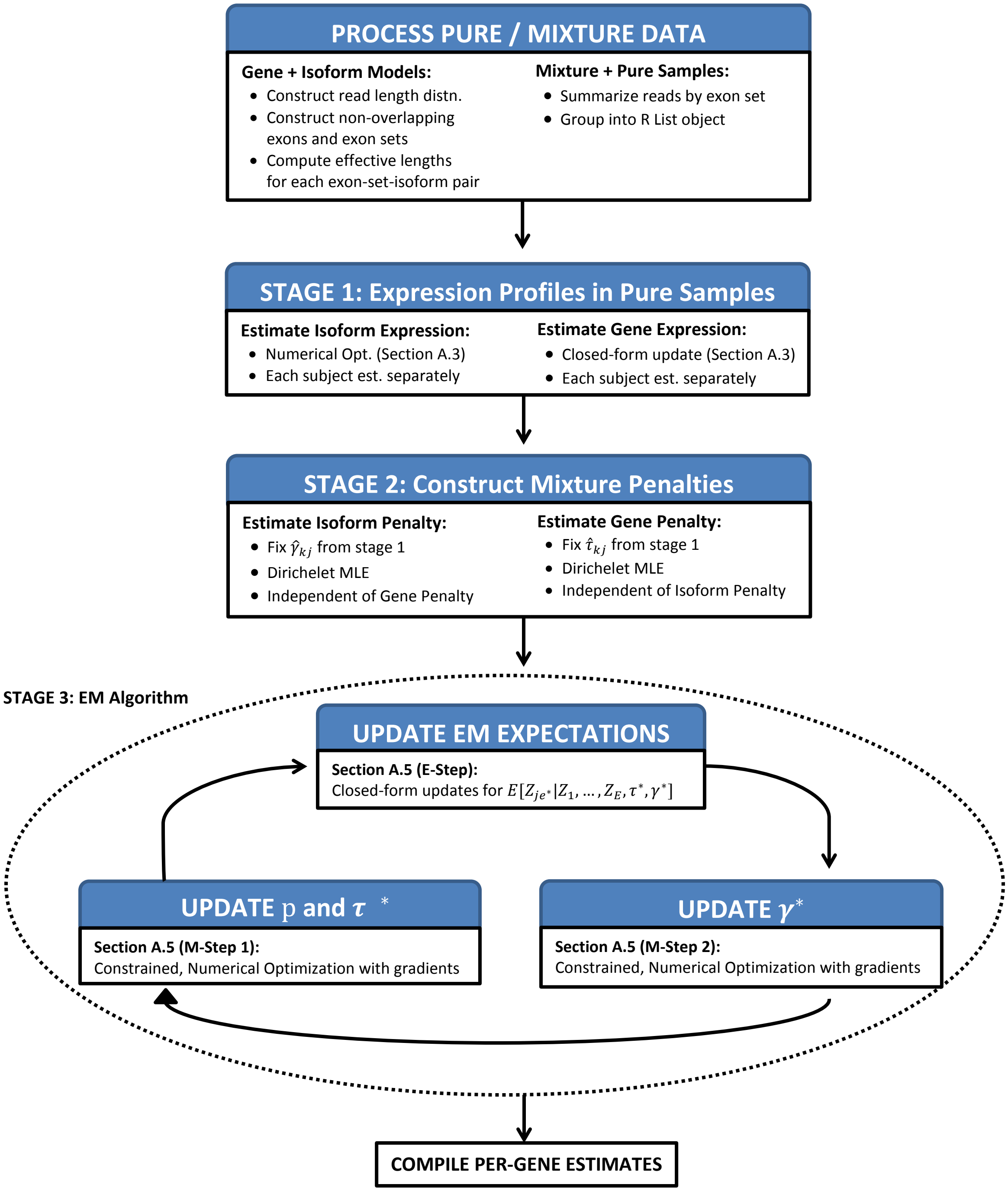

PROCESS PURE / MIXTURE DATA
Gene + Isoform Models:
Construct read length distn.
Construct non-overlapping exons and exon sets
Compute effective lengths for each exon-set-isoform pair
Mixture + Pure Samples:
Summarize reads by exon set
Group into R List object
STAGE 1: Expression Profiles in Pure Samples
Estimate Isoform Expression:
Numerical Opt. (Section A.3)
Each subject est. separately
Estimate Gene Expression:
Closed-form update (Section A.3)
Each subject est. separately
STAGE 2: Construct Mixture Penalties
Estimate Isoform Penalty:
Fix $\hat{\gamma}_{kj}$ from stage 1
Dirichelet MLE
Independent of Gene Penalty
Estimate Gene Penalty:
Fix $\hat{\tau}_{kj}$ from stage 1
Dirichelet MLE
Independent of Isoform Penalty
STAGE 3: EM Algorithm
UPDATE EM EXPECTATIONS
Section A.5 (E-Step):
Closed-form updates for $E[Z_{je^*}|Z_1,\ldots,Z_E,\tau^*,\gamma^*]$
UPDATE p and τ *
Section A.5 (M-Step 1):
Constrained, Numerical Optimization with gradients
UPDATE γ^*
Section A.5 (M-Step 2):
Constrained, Numerical Optimization with gradients
COMPILE PER-GENE ESTIMATES

Figure A.3: Visual representation of the IsoDeconv Multinomial algorithm from development of reference matrices to Stage 3 EM Updates.

A.2.3 Lemmas Involving Multinomial Distribution

Prior to specification of the IsoDeconv model, we develop a set of lemmas for the multinomial distribution which will allow easier specification in the following materials. For completeness, we define a multinomially distributed vector $X = (X_1, ..., X_R)$ with size n and proportions $\rho = (\rho_1, ..., \rho_R)$. The density function of $X \sim \texttt{Multinomial}\,(n, \rho)$ is given by:

$$P\left\{X_1, ..., X_R \middle| n, \rho\right\} = \begin{pmatrix} n \\ X_1, ..., X_R \end{pmatrix} \prod_{i=1}^{R} \rho_i^{X_i}$$

<u>*Lemma 1.1: Sum over Groups*</u>

W.L.O.G. construct the sum $X. = X_1 + ... + X_g$ and consider the grouped multinomial $X' = f(X) = (X., X_{g+1}, X_{g+2}, ..., X_R)$. Let S represent the set of vectors X such that $X' = f(X) = x$ where x is an arbitrary $(R - g + 1)$-dimensional non-negative vector summing to n. The density of this random variable is given by:

$$\begin{aligned} P\left\{X' = x \middle| n, \rho\right\} &= \sum_{X' \in S} \begin{pmatrix} n \\ X_1, ..., X_R \end{pmatrix} \prod_{i=1}^{R} \rho_i^{X_i} \\ &= \begin{pmatrix} n \\ x_1, x_2 ..., x_{R-g+1} \end{pmatrix} \prod_{i=g+1}^{R} \rho_i^{x_i} \left[\sum_{X \in S} \begin{pmatrix} x_1 \\ X_1, ..., X_g \end{pmatrix} \prod_{i=1}^{g} \rho_i^{X_i} \right] \\ &= \begin{pmatrix} n \\ x_1, x_2 ..., x_{R-g+1} \end{pmatrix} \left(\sum_{i=1}^{g} \rho_i \right)^{x_1} \prod_{i=g+1}^{R} \rho_i^{x_{i-g+1}} \end{aligned}$$

Thus, it is clear that $X' \sim \texttt{Multinomial}\,(n, \rho')$ where $\rho' = (\rho_1 + ... + \rho_g, \rho_{g+1}, ..., \rho_R)$.

<u>*Lemma 1.2: Marginal of a Single Element*</u>

We extend (1.1) to the case where $X. = X_1 + ... + X_{R-1}$ and consider the distribution of $X' = f(X) = (X., X_R)$. Using (1.1) it is clear that $X' \sim \texttt{Multinomial}\,(n, (1 - \rho_R, \rho_R))$. Thus, it is obvious that:

$$X_R \sim \text{Bin}\left(n, \rho_R\right)$$

Lemma 2.1: Conditional over Multiple Elements

W.L.O.G. consider conditioning on the first g elements. Thus, we seek to specify the conditional density of $X^* = (X_{g+1}, ..., X_R)$ given $(X_1, ..., X_g)$. By lemma (1.1), we know that:

$$P\left\{X_1, ..., X_g\right\} = \binom{n}{X_1, ..., X_g, n - X_1 - ... - X_g} \left[\prod_{i=1}^{g} \rho_i^{X_i}\right] (1 - \rho_1 - ... - \rho_g)^{n - X_1 - ... - X_g}$$

Thus, applying this to the definition of conditional densities, we have:

$$\begin{aligned}
P\left\{X^* \middle| X_1, ..., X_g\right\} &= \frac{P\left\{X^* \cap (X_1, ..., X_g)\right\}}{P\left\{X_1, ..., X_g\right\}} \\
&= \frac{\binom{n}{X_1, ..., X_R} \prod_{i=1}^{R} \rho_i^{X_i}}{\binom{n}{X_1, ..., X_g, , n - \sum_{r=1}^{g} X_r} \left[\prod_{i=1}^{g} \rho_i^{X_i}\right] (1 - \sum_{r=1}^{g} \rho_r)^{n - X_1 - ... - X_g}} \\
&= \binom{n - X_1 - ... - X_g}{X_{g+1}, ..., X_R} \left[\prod_{i=g+1}^{R} \left(\frac{\rho_i}{1 - \rho_1 - ... - \rho_g}\right)^{X_i}\right]
\end{aligned}$$

Thus, it is clear that:

$$X^* \Big| (X_1, ..., X_g) \sim \texttt{Multinomial}\left(n - X_1 - ... - X_g, \rho^*\right)$$

where $\rho^* = \left(\frac{\rho_{g+1}}{1 - \rho_1 - ... - \rho_g}, \cdots, \frac{\rho_R}{1 - \rho_1 - ... - \rho_g}\right)$.

Lemma 2.2: Conditional of a Single Element

We consider a specific case of lemma (2.1) where $X^* = (X_2, \cdots, X_R)$. Thus, it is clear that:

$$X^* \Big| X_1 \sim \texttt{Multinomial}\left(n - X_1, \rho_1^*\right) \quad \text{where} \quad \rho_1^* = \left(\frac{\rho_2}{1 - \rho_1}, \cdots, \frac{\rho_R}{1 - \rho_1}\right)$$

Lemma 3: Conditional Over Sums

Under the original framework, consider splitting the R elements of X into K distinct groups. W.L.O.G. we specify:

Group 1	**Group 2**	$\cdots$	**Group K**
$X_1, \cdots, X_{k_1}$	$X_{k_1+1}, \cdots, X_{k_2}$	$\cdots$	$X_{k_{K-1}}, \cdots, X_R$
$\rho_1, \cdots, \rho_{k_1}$	$\rho_{k_1+1}, \cdots, \rho_{k_2}$	$\cdots$	$\rho_{k_{K-1}}, \cdots, \rho_R$

For convenience, define $S_j = \sum_{i=k_{j-1}+1}^{k_j} X_i$ where $k_0 = 1$. Additionally, define $\rho_j^* = \sum_{i=k_{j-1}+1}^{k_j} \rho_i$. Thus, we examine the following conditional density:

$$\begin{aligned} P\left\{X_1, ..., X_R \middle| S_1, \cdots, S_K\right\} &= \frac{P\left\{X_1, \cdots, X_R\right\}}{P\left\{S_1, \cdots, S_K\right\}} \\ &= \frac{\begin{pmatrix} n \\ X_1, ..., X_R \end{pmatrix} \prod_{j=1}^{R} \rho_j^{X_j}}{\begin{pmatrix} n \\ S_1, ..., S_K \end{pmatrix} \prod_{j=1}^{K} \rho_j^{*S_j}} \\ &= \prod_{j=1}^{K} \left[\begin{pmatrix} S_j \\ X_{k_{j-1}}, ..., X_{k_j} \end{pmatrix} \prod_{l=k_{j-1}+1}^{k_j} \left(\frac{\rho_l}{\rho_j^*}\right)^{X_l} \right] \end{aligned}$$

The second equality holds through repeated application of Lemma 1.1. The final equality demonstrates that the desired conditional distribution is the product of independent multinomials. Symbolically, we have:

$$X_1, \cdots, X_R \middle| S_1, \cdots, S_K \sim \prod_{j=1}^{K} \texttt{Multinomial}\left(S_j, \rho_j'\right)$$

where $\rho_j' = \left(\rho_{k_{j-1}+1}, \cdots, \rho_{k_j}\right) / \rho_j^*$.

A.2.4 Stage 1 Estimation: Pure Sample Necessities

For the following, refer to Table A.2 regarding notation. Additionally, note that the following specification is performed for a single gene only; subscripts related to gene identity are omitted for clarity. The following structure holds for a single purified reference sample and gene:

$$\begin{bmatrix} Y_{kj(E)} \\ Y_{kj} \end{bmatrix} \sim \texttt{Multinomial} \left(t_{kj}, \begin{bmatrix} 1 - \tau_{kj} \\ \tau_{kj} X \gamma_{kj} \end{bmatrix} \right)$$

Implicit in this construction are restrictions upon the τ_{kj} and γ_{kj}. As a single probability value, it must be that $0 \leq \tau_{kj} \leq 1$. However, the γ_{kj} pose a more complicated set of restrictions. Consider the following:

$$\begin{aligned} 1 &= (1 - \tau_{kj}) + \tau_{kj} \mathbf{1}^T X \gamma_{kj} \\ &= \mathbf{1}^T X \gamma_{kj} \end{aligned}$$

It is clear from the above that the $X\gamma_{kj}$ are conditional probabilities and thus must be non-negative. To ensure this, we restrict the γ_{kj} to be non-negative since the elements of X_{kj} are non-negative by definition. Using our summation constraints, we have:

$$1 = \mathbf{1}^T X \gamma_{kj} = \sum_{i=1}^{I} \tilde{l}_i \gamma_{kji}$$

This shows that the $\tilde{l}_i \gamma_{kji}$ are probabilities that a randomly selected read is attributable to isoform i for reference j of cell type k. Thus, it is clear that the γ_{kj} are collections of per-unit of effective length conditional probabilities that a read belongs to isoform i given that it maps to the gene of interest.

Thus, the likelihood for sample j of cell type k is given by:

$$\begin{aligned}\ell_{kj} &= Y_{kj(E)} \log\left(1-\tau_{kj}\right) + \sum_{e=1}^{E} Y_{kje} \log\left(\tau_{kj} X_e^T \gamma_{kj}\right) \\ &= Y_{kj(E)} \log\left(1-\tau_{kj}\right) + \left(\mathbf{1}^T Y_{kj}\right) \log\left(\tau_{kj}\right) + \sum_{e=1}^{E} Y_{kje} \log\left(X_e^T \gamma_{kj}\right)\end{aligned}$$

Given the gene and isoform expressions, the reference samples within and across cell types are independent. Thus, we may estimate the τ_{kj} and γ_{kj} separately within each sample.

Estimate τ_{kj}:

The maximum likelihood estimate of τ_{kj} is given by:

$$\hat{\tau}_{kj} = \frac{\mathbf{1}^T Y_{kj}}{t_{kj}}$$

Estimate γ_{kj}:

In order to estimate the isoform expressions for a single subject, we make some simplifying alterations to the effective length matrix X and reparametrize the isoform expression parameters.

Consider the following, where X_{cj} refers to the j-th column of X.

$$
\begin{aligned}
\begin{bmatrix} X_{c1} & X_{c2} & \cdots & X_{cI} \\ \vert & \vert & & \vert \end{bmatrix}
\begin{bmatrix} \gamma_{kj1} \\ \vdots \\ \gamma_{kjI} \end{bmatrix}
&= \begin{bmatrix} X_{c1}/\tilde{l}_1 & X_{c2}/\tilde{l}_2 & \cdots & X_{cI}/\tilde{l}_I \\ \vert & \vert & & \vert \end{bmatrix}
\begin{bmatrix} \tilde{l}_1\gamma_{kj1} \\ \vdots \\ \tilde{l}_1\gamma kjI \end{bmatrix} \\
&= \left[\sum_{i=1}^{I-1}\left(\frac{X_{ci}}{\tilde{l}_i}\right)\left(\tilde{l}_i\gamma_{kji}\right)\right] + (1-\sum_{r=1}^{I-1}\tilde{l}_r\gamma_{kjr})\left(\frac{X_{cI}}{\tilde{l}_I}\right) \\
&= \left[\sum_{i=1}^{I-1}\left(\frac{X_{ci}}{\tilde{l}_i}-\frac{X_I}{\tilde{l}_I}\right)\left(\tilde{l}_i\gamma_{kji}\right)\right] + \left(\frac{X_{cI}}{\tilde{l}_I}\right) \\
&= \begin{bmatrix} X^*_{c1} & X^*_{c2} & \cdots & X^*_{cI} \\ \vert & \vert & & \vert \end{bmatrix}
\begin{bmatrix} e^{-\gamma_{kj1}} \\ \vdots \\ e^{-\gamma_{kj,I-1}} \\ 1 \end{bmatrix}
\end{aligned}
$$

where:

$$
\begin{aligned}
X^*_{cs} &= \left[X_s - \left(\tilde{l}_s/\tilde{l}_I\right)X_{cI}\right] && \text{for } j \in \{1,2,...,I-1\} \\
X^*_{cI} &= \frac{X_{cI}}{\tilde{l}_I} \\
X^* &= \begin{bmatrix} X^*_{c1} & \cdots & X^*_{cI} \end{bmatrix} \\
\gamma_{kji} &= e^{-\gamma^r_{kji}} && \text{for } i \in \{1,2,...,I-1\} \\
\gamma'_{kj} &= (\gamma_{kj1},...,\gamma_{kj,I-1},1)
\end{aligned}
$$

We optimize the likelihood with respect to these isoform expression parameters using R's `constrOptim` from the `alabama` package. To this end, we specify the derivative to improve efficiency of the routine. This derivative is found on the next page.

In the following, let X_e^* refer to the e-th row of the matrix X^* and $X_{e,(I)}$ be the truncated version of this row excluding the last column entry. :

$$\frac{\partial \ell_{kj}}{\partial \gamma'_{kj}} = -\sum_{e=1}^{E} \left(\frac{Y_{kje}}{X_e^{*T} \gamma'_{kj}} \right) \left[X^*_{e,(I)} \circ e^{-\gamma^r_{kj}} \right]$$

A.2.5 Stage 2 Estimation: Defining Penalties

We must now incorporate the estimates from purified reference samples to guide estimation within the mixture. We choose to accomplish this using a penalty function over the isoform expression parameters within the mixture. As we have allowed for biological variance in gene and expression parameters across subjects and because these parameters are probabilities, it is natural to propose a dirichelet distribution over these parameters.

Normally, by placing a dirichelet distribution over these parameters, one would construct a likelihood function containing both pieces simultaneously. This likelihood would then be optimized with respect to all parameters, including hyperparameters, at the same time. However, we found this approach to be unstable. Thus, we separate the estimation of individual expression parameters from the hyperparameters to improve results. Fixing the individual gene and isoform expression parameters, we construct a likelihood optimization using the dirichelet piece. Optimization of this likelihood proceeds numerically using quasi-Newton methods and non-negativity constraints. The following derivatives improve accuracy of the estimates obtained from R's `nlminb`.

Gene Expression Penalty:

The likelihood for this penalty is given below

$$\ell_k^{\tau} = \sum_{j=1}^{n_k} \left[\ln\Gamma\left(\alpha_{k1} + \alpha_{k2}\right) - \ln\Gamma\left(\alpha_{k1}\right) - \ln\Gamma\left(\alpha_{k2}\right) + \left(\alpha_{k1} - 1\right)\log(\tau_{kj}) + \left(\alpha_{k2} - 1\right)\log(1 - \tau_{kj})\right]$$

The necessary derivatives are provided here. Denote the digamma function by $\varphi()$ and trigamma by $\varphi_1()$ for this derivatives.

$$\nabla \ell_k^\tau = n_k \begin{bmatrix} \varphi(\alpha_{k1} + \alpha_{k2}) - \varphi(\alpha_{k1}) \\ \varphi(\alpha_{k1} + \alpha_{k2}) - \varphi(\alpha_{k2}) \end{bmatrix} + \begin{bmatrix} \sum_{j=1}^{n_k} \log(\tau_{kj}) \\ \sum_{j=1}^{n_k} \log(1 - \tau_{kj}) \end{bmatrix}$$

$$\text{Hess}\left(\ell_k^\tau\right) = n_k \begin{bmatrix} \varphi_1(\alpha_{k1} + \alpha_{k2}) - \varphi_1(\alpha_{k1}) & \varphi(\alpha_{k1} + \alpha_{k2}) \\ \varphi(\alpha_{k1} + \alpha_{k2}) & \varphi(\alpha_{k1} + \alpha_{k2}) - \varphi(\alpha_{k2}) \end{bmatrix}$$

Isoform Expression Penalty:

As for the gene expression penalty, we define the likelihood here. To clarify the following terms, define $\beta_{k\cdot} = \sum_{i=1}^{I} \beta_{ki}$ and utilize the same definitions for φ and φ_1.

$$\ell_k^\gamma = \sum_{j=1}^{n_k} \left[\ln\Gamma \left(\beta_{k\cdot} - \sum_{i=1}^{I} \ln\Gamma\left(\beta_{ki}\right) \right) + \sum_{i=1}^{I} \left(\beta_{ki} - 1\right) \log\left(\tilde{l}_i \gamma_{kji}\right) \right]$$

The necessary derivatives are specified below:

$$\nabla \ell_k^\gamma = n_k \begin{bmatrix} \varphi(\beta k\cdot) - \varphi(\beta_{k1}) \\ \vdots \\ \varphi(\beta k\cdot) - \varphi(\beta_{kI}) \end{bmatrix} + \begin{bmatrix} \sum_{j=1}^{n_k} \log\left(\tilde{l}_1 \gamma_{kj1}\right) \\ \vdots \\ \sum_{j=1}^{n_k} \log\left(\tilde{l}_I \gamma_{kjI}\right) \end{bmatrix}$$

$$\text{Hess}\left(\ell_k^\gamma\right) = n_k \left(\mathbf{1}\mathbf{1}^T \varphi_1\left(\beta_{k\cdot}\right) - \text{diag}_i\left(\varphi_1(\beta_{ki})\right)\right)$$

A.2.6 Stage 3 Estimation: Mixture Sample Estimation

To structure the likelihood model within the mixture sample, consider the following underlying likelihood model. In this model, we assume that the number of reads mapping to each cell type within each gene and outside of it can be observed and that t_m represents the total read count in the mixture.

$$\begin{bmatrix} Z_{1(E)^*} & \cdots & Z_{K(E)^*} \\ Z_{11^*} & & Z_{K1^*} \\ \vdots & & \vdots \\ Z_{1E^*} & & Z_{KE^*} \end{bmatrix} \Bigg| \tau_k^*, \gamma_k^* \sim \texttt{Multinomial}\left(t_m, \begin{bmatrix} p_1(1-\tau_1^*) & \cdots & p_K(1-\tau_K^*) \\ p_1\tau_1^* X \gamma_1^* & \cdots & p_K \tau_K X \gamma_k^* \end{bmatrix} \right)$$

When allowing IsoDeconv to consider genes mapping outside of the gene of interest, initial simulations demonstrated that these terms dominated estimation. This occurs since over 99% of all reads map outside the gene of interest and thus drown out the information within the gene due to sheer abundance. Restricting to reads within the gene of interest only, estimation behavior was seen to improve (not shown). Thus, using lemma 1.1 to combine all contributions of cell types outside the gene and then lemma 2.2 to condition on this quantity, we have:

$$\begin{bmatrix} Z_{11^*} & Z_{K1^*} \\ \vdots & \vdots \\ Z_{1E^*} & Z_{KE^*} \end{bmatrix} \Bigg| \tau_k^*, \gamma_k^* \sim \texttt{Multinomial}\left(Z_G, \begin{bmatrix} \frac{p_1\tau_1^* X \gamma_1^*}{\sum_{k=1}^K p_k \tau_k^*} & \cdots & \frac{p_K \tau_K X \gamma_k^*}{\sum_{k=1}^K p_k \tau_k^*} \end{bmatrix} \right) \tag{A.1}$$

However, due to the properties of bulk expression datasets, we do not observe the number of reads mapping to each cell type. Thus, we only observe the sums from all cell types at each exon set.

$$Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_E \end{bmatrix} \Bigg| \tau_k^*, \gamma_k^* \sim \texttt{Multinomial}\left(Z_G, \left[\frac{\sum_{k=1}^K p_k \tau_k^* X \gamma_1^*}{\sum_{k=1}^K p_k \tau_k^*} \right] \right)$$

The update of such a likelihood is a computationally difficult problem - we have I+2 parameters being measured for each cell type and all must be optimized simultaneously. To improve the tractability of such a numerical optimization technique, we utilize the EM algorithm.

For this problem, the missing data that we will assume is the expression from each individual cell type. Thus, we revert to the likelihood given above in equation (A.1). The complete data log-likelihood is given by:

$$\ell = \sum_{k=1}^{K} \Bigg\{ \sum_{e=1}^{E} \Big[Z_{ke^*} \Big(\log[p_1 \tau_1^*] - \log \Big[\sum p_r \tau_r^* \Big] + \log(X_e^T \gamma_k^*) \Big) \Big] + \\ \ln\Gamma\left(\alpha_{k\cdot}\right) - \ln\Gamma\left(\alpha_{k1}\right) - \ln\Gamma\left(\alpha_{k2}\right) + (\alpha_{k1} - 1)\log(\tau_k^*) + (\alpha_{k2} - 1)\log(1 - \tau_k^*) \\ \ln\Gamma\left(\beta_{k\cdot}\right) - \ln\Gamma\left(\beta_{k1}\right) - \cdots - \ln\Gamma\left(\beta_{kI}\right) + \sum_{i=1}^{I} (\beta_{ki} - 1) \log\left(\tilde{l}_i \gamma_{ki}^*\right) \Bigg\}$$

The EM algorithm utilized to solve this problem is composed of three separate steps.

1 E-Step: Update Posterior Means of Z_{ke^*}

2 M-Step (1): Update $(p_1, ..., p_k, \tau_k^*)$

2 M-Step (2): Update γ_k^*

These steps are outlined below.

E-Step: Update Posterior Means of Z_{ke^*}:

Recall that the observed expression values, the Z_e, represent the sum of all counts from each cell type. Thus, $Z_e = \sum_{k=1}^{K} Z_{ke^*}$. By grouping elements of the multinomial according to exon set, a simple application of lemma 3 provides:

$$(Z_{11^*}, ..., Z_{K1^*}, ..., Z_{1E^*}, ...Z_{KE^*}) \Big| Z_1, ..., Z_e, \tau^*, \gamma^* \sim \prod_{e=1}^{E} \texttt{Multinomial}\left(Z_e, \rho_e'\right)$$

where

$$\rho_e' = \left(\frac{p_1 \tau_1^* X \gamma_1^*}{\sum_{k=1}^{K} p_k \tau_k^* X \gamma_k^*}, \cdots, \frac{p_K \tau_K^* X \gamma_K^*}{\sum_{k=1}^{K} p_k \tau_k^* X \gamma_k^*} \right)$$

Thus, it becomes clear by property of the multinomial distribution that:

$$E\left[Z_{je^*}\Big|Z_1,\cdots,Z_E,\tau^*,\gamma^*\right]=Z_e\left(\frac{p_j\tau_j^*X\gamma_j^*}{\sum_{k=1}^K p_k\tau_k^*X\gamma_k^*}\right)$$

M-Step (1): Update $(p_1, ..., p_K, \tau_K^*)$:

It is clear from the complete data log-likelihood specified above that the cell type proportions and gene expression parameters must be updated simultaneously. These terms are inextricably linked within the log function. We do note that this set of parameters is separable from the isoform parameters as the likelihood can be partitioned into a sum of two independent pieces, one containing the gene expression parameters and cell type proportions and the other containing the isoform parameters. Thus, we consider recasting the likelihood to include only the cell type proportions and gene expression parameters.

$$\begin{aligned}
\ell\left(p,\tau^*\right) &= \sum_{k=1}^K\left\{\sum_{e=1}^E\left[Z_{ke^*}\left(\log[p_k\tau_k^*]-\log\left[\sum p_r\tau_r^*\right]\right)\right]+(\beta_{k1}-1)\log(\tau_k^*)+\right.\\
&\qquad\left.(\beta_{k2}-1)\log(1-\tau_k^*)\right\}\\
&=\sum_{k=1}^K\left\{Z_{k\cdot^*}\left(\log[p_k\tau_k^*]-\log\left[\sum p_r\tau_r^*\right]\right)+(\beta_{k1}-1)\log(\tau_k^*)+(\beta_{k2}-1)\log(1-\tau_k^*)\right\}\\
&=\left\{\sum_{k=1}^K Z_{k\cdot^*}\log[p_k\tau_k^*]+(\beta_{k1}-1)\log(\tau_k^*)+(\beta_{k2}-1)\log(1-\tau_k^*)\right\}-Z_.\log\left[\sum p_r\tau_r^*\right]\\
&=\left\{\sum_{k=1}^K Z_{k\cdot^*}\log[p_k\exp\{-\tau_k'\}]+(\beta_{k1}-1)\log(\exp\{-\tau_k'\})+\right.\\
&\qquad\left.(\beta_{k2}-1)\log(1-\exp\{-\tau_k'\})\right\}-Z_.\log\left[\sum p_r\exp\{-\tau_r'\}\right]
\end{aligned}$$

Taking the expectation of this likelihood will result in the use of quantities found in (1) to replace the Z_{ke^*} pieces. In the following, we leave the the $Z_{k\cdot}$ notation for simplicity of notation, but please note that these values have been replaced by their expectations.

Taking the derivative of $\ell(p, \tau^*)$ with respect to the reparametrized τ^*, we have:

$$\dot{\ell}_{\tau'_r}(p, \tau^*) = -Z_{r\cdot} - (\beta_{k1} - 1) + \frac{(\beta_{k2} - 1)\exp\{-\tau'_r\}}{1 - \exp\{-\tau'_r\}} + Z_{\cdot}\left(\frac{p_r \exp\{-\tau'_r\}}{\sum p_k \exp\{-\tau'_k\}}\right)$$

To consider the derivatives of the proportions, we consider the natural linearity constraints to rewrite the likelihood as follows and subsequently take the derivative:

$$\begin{aligned} \ell(p, \tau^*) \approxeq & \left[\sum_{k=1}^{K-1} Z_{k\cdot} \log\left(p_k \tau^*_k\right)\right] + Z_{K\cdot} \log\left((1 - p_1 - \ldots - p_{K-1})\tau^*_K\right) - \\ & Z_{\cdot} \log\left(\sum_{s=1}^{K-1} p_s(\tau^*_s - \tau^*_K) + \tau^*_K\right) \\ \dot{\ell}_{p_r}(p, \tau^*) = & \left(\frac{Z_{r\cdot}}{p_r}\right) - Z_{K\cdot}\left(\frac{1}{1 - p_1 - \ldots - p_{K-1}}\right) - Z_{\cdot}\left[\frac{\tau^*_r - \tau^*_K}{\sum_{s=1}^{K} p_s \tau^*_s}\right] \end{aligned}$$

The update of the procedures proceeds using a joint, constrained optimization approach using R's `constrOptim`.

M-Step (2): Update γ^*_k:

As noted above, we may update the γ^*_k separately from one another and from the proportion and gene expression parameters. The piece of the likelihood governing the update of isoform expression parameters for cells of type k is given by:

$$\begin{aligned}
\ell\left(\gamma_k^*\right) &= \left(\sum_{e=1}^{E} Z_{ke}\log\left(X_e^T\gamma_k^*\right)\right) + \ln\Gamma\left(\alpha_{k\cdot}\right) - \sum_{i=1}^{I}\ln\Gamma\left(\alpha_{ki}\right) + \sum_{i=1}^{I}\left(\alpha_{ki}-1\right)\log\left(\tilde{l}_i\gamma_{ki}\right) \\
&= \left(\sum_{e=1}^{E} Z_{ke}\log\left(X_e^T\gamma_k^*\right)\right) + \ln\Gamma\left(\alpha_{k\cdot}\right) - \sum_{i=1}^{I}\ln\Gamma\left(\alpha_{ki}\right) + \sum_{i=1}^{I-1}\left(\alpha_{ki}-1\right)\log\left(\tilde{l}_i\gamma_{ki}\right) + \\
&\qquad \left(\alpha_{kI}-1\right)\log\left(1-\tilde{l}_1\gamma_{k1}^* - \cdots - \tilde{l}_{I-1}\gamma_{k,I-1}^*\right)
\end{aligned}$$

For simplicity of notation in the following, we suppress the notation regarding expectations of the missing parameters. Note, however, that these values are replaced by their expectations derived in the E-step.

Recall the special definitions of X^*, X_e^* and $X_{e,(I)}^*$ from their use in the pure sample expression materials. In addition, we define reparameterized isoform expression parameters for the mixture given by $\gamma_{ki}^* = \exp\left\{-\gamma_{ki}^{*r}\right\}$ to simplify constraints. Finally, we define $\tilde{l}_{(I)}$ as the $\tilde{l}$ vector with the I-th entry removed. Taking the derivative, we have:

$$\begin{aligned}
\dot{\ell}_{\gamma_k^{r*}}\left(\gamma_k^*\right) = \sum_{e=1}^{E} -Z_{ke}\left(\frac{X_{e,(I)}^* \circ \exp\{-\gamma_k^{r*}\}}{X_e^{*T}\gamma_k^*}\right) - (\beta_k - 1) + \\
\left(\frac{\beta_{kI}-1}{1-\tilde{l}_1\exp\{-\gamma_{k1}^{*r}\} - \cdots - \tilde{l}_{I-1}\exp\{-\gamma_{k,I-1}^{*r}\}}\right)\left(\tilde{l}_{(I)} \circ \exp\{-\gamma_k^{r*}\}\right)
\end{aligned}$$

Thus, given the restrictions outlined for the pure sample case, we utilize this derivative in R's `constrOptim` to update the isoform expression parameters.

A.2.7 Simulation Supplement

To construct a single simulated mixture composed of two cell types, we must construct facsimiles to the following components of real RNA-seq experiments: Gene and Isoform construction models, Gene expression averages, Isoform expression level averages for an arbitrary cell type 1, and isoform expression level averages for an arbitrary cell type 2.

Gene and Isoform Construction Models:

Approximately 400 genes and corresponding gene/isoform construction models were extracted from the IsoDeconvNB *in silico* mixtures using GM12878 and HMEC cell lines. Therefore, these construction models contain genuine gene and isoform constructions as well as realistic distributions of RNA fragment lengths for construction of an effective length matrix.

Gene Expression Level:

To simulate expression at a single gene, the average read count for cell type 1, r_1, is randomly drawn from a normal random variable with mean 130 and standard deviation 33. From this, we may construct $\beta_1 = v\left(\frac{r}{1e7}, 1 - \frac{r}{1e7}\right)$ where v is a Chebyshev derived variance inflation factor. This provides a mean expression level for an arbitrary cell type 1. For 25% of simulated genes, the average expression of the gene in cell type 1 is upregulated by 20% for cell type 2. An additional 25% of genes see downregulated expression by 20% for cell type 2. The remaining genes exhibit no gene-level expression differences across cell types.

Isoform Expression Level (Cell Type 1):

In order to construct an isoform expression profile, simulation methods were developed according to the number of isoforms used by the gene being simulated. Genes with 1 or fewer isoforms are excluded from consideration.

Two Isoforms:

(A) Simulate isoform probability averages from a Dirichlet(7.5,2.5) distribution.

(B) Randomly permute these averages across isoform identities to obtain α_k^*.

(C) Utilize a Chebyshev Derived Variance Factor (v) to multiply isoform averages in order to control variance in simulated isoform expressions. Thus, $\alpha_1 = v\alpha_1^*$.

(D) Simulate 3 or 50 purified reference sample isoform expressions utilizing a Dirichlet(α_1)

Three Isoforms:

(A) Simulate isoform probability averages from a Dirichlet(6.0,3.0,1.0) distribution.

(B) Randomly permute these averages across isoform identities to obtain α_k^*.

(C) Utilize a Chebyshev Derived Variance Factor (v) to multiply isoform averages in order to control variance in simulated isoform expressions. Thus, $\alpha_1 = v\alpha_1^*$.

(D) Simulate 3 or 50 purified reference sample isoform expressions utilizing a Dirichlet(α_1)

Four + Isoforms:

(A) Simulate isoform probability averages from a Dirichlet(4.5,2.5,1.5,J) where J is a vector of length $I - 3$ with values $1.5/(I - 3)$ for each entry.

(B) Randomly permute these averages across isoform identities to obtain α_k^*.

(C) Utilize a Chebyshev Derived Variance Factor (v) to multiply isoform averages in order to control variance in simulated isoform expressions. Thus, $\alpha_1 = v\alpha_1^*$.

(D) Simulate 3 or 50 purified reference sample isoform expressions utilizing a Dirichlet(α_1)

Isoform Expression Level (Cell Type 2):

To simulate isoform expression level averages in cell type 2, the averages of cell type 1 are permuted in such a way that none of the top 2-3 isoforms of cell type 1 are the major isoforms of cell type 2. For genes with 2 or 3 isoforms, this is accomplished by permuting the largest element of α_1 to a new location in α_2 and randomly ordering the remainder. For genes with four of five isoforms, the top two largest elements of α_1 are permuted to new locations in α_2 and the remainder are randomly ordered. For genes with 6+ isoforms, the top 3 elements of α_1 are permuted to new locations in α_2 and the rest are randomly ordered.

Chebyshev Derived Variance Factor:

Chebyshev's rule states that 90% of observations fall with 3 standard deviations of the mean. To derive the variance control factor based on Chebyshev's rule and a desire to have 90% of observations fall within Z*100% of the truth, we have:

$$3\sqrt{p(1-p)(v+1)^{-1}} = Zp$$
$$v = \left[\frac{1-p}{p}\right](9/Z^2) - 1$$

For isoform expressions, variance is largest when $p = 0.5$. Thus, to specify the value v for isoforms, we utilize an assumption of $p = 0.5$ to control the variance in the worst-case setting.

Simulate Mixture Expression:

After the preceding steps have been accomplished, the Dirichlet structure specified in the paper is fully specified. Thus, we plug these simulated parameters into the multinomial structure to simulate a single mixture experiment.

APPENDIX B: SUPPLEMENT FOR CHAPTER 4

B.1 Supplementary Methods

B.1.1 Notations and overview

B.1.1.1 Notation Table

The following table contains the notation used to develop and mathematically interrogate the ICeD-T model and its variants. Subscripts for aberrant genes, denoted by $(\cdot)$ in the following table, may take values (A) or (C); (A) indexes quantities pertaining to aberrant genes and (C) indexes those in consistent genes.

Model Design Quantities		
Value	**Dimension**	**Description**
n	1×1	The number of mixed cell type samples for deconvolution.
n_k	1×1	The number of purified samples of cell type k.
K	1×1	The number of constituent cell types, excluding the tumor.
n_G	1×1	The number of signature genes used in cell type modeling.
Pure Sample Quantities		
Value	**Dimension**	**Description**
μ_{jk}	1×1	Mean log-transformed expression of gene j in cell type k.
σ^2_{jk}	1×1	The variance of log-transformed expression of gene j in cell type k.
γ_{jk}	1×1	The mean expression of gene j in cell type k on the untransformed scale.
$\boldsymbol{\gamma}$	$n_G \times K$	Matrix of mean expression across all genes and cell types.
$\boldsymbol{\gamma}_j$	$K \times 1$	Vector of mean expressions of gene j across the K cell types (j-th row of $\boldsymbol{\gamma}$).
Z_{jkh}	1×1	Normalized expression of gene j in purified sample h of cell type k.
$\mathbf{Z}_k$	$n_G \times n_q$	Collection of Z_{jkh} across all genes and purified samples.
Mixture Sample Quantities		
Value	**Dimension**	**Description**
$\tilde{\mu}_{ij(\cdot)}$	1×1	Mean expression of gene j in mixture sample i
ρ_{ik}	1×1	Proportion of RNA expression attributable to cells of type k in mixture i.
ρ_i	$K \times 1$	Collection of ρ_{ik} across cell types for subject i only.
$\sigma^2_{ij(\cdot)}$	1×1	Variance of expression for gene j in mixture sample i.
$\sigma^2_{i(\cdot)}$	1×1	Unweighted variance parameter governing expression in mixture sample i.
Δ_j	1×1	Optional variance weight for gene j.
Y_{ij}	1×1	Normalized expression of gene j in mixture sample i.
$\mathbf{Y}_i$	$n_G \times 1$	Collection of Y_{ij} across genes for subject i only.

Table B.1: Notation for defining the ICeD-T model.

B.1.1.2 Overview of Optimization Algorithm

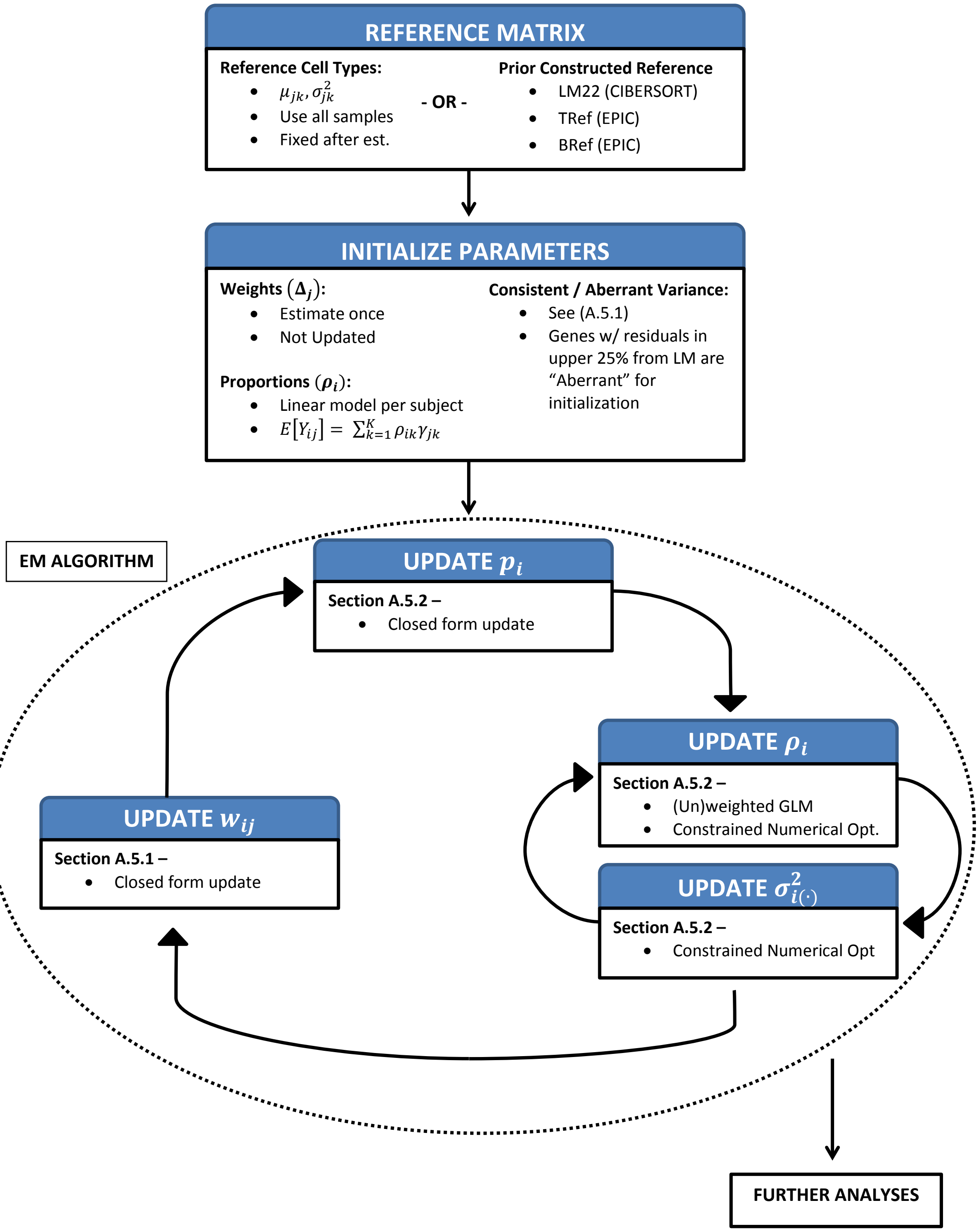


Figure B.1: Visual representation of the ICeD-T algorithm from development of reference matrices to EM algorithm.

B.1.2 Pure Sample Optimization

We focus first on estimation using purified reference samples. Recall that for reference sample h of cell type k, the expression at marker gene j is assumed to follow a log-normal distribution, given by:

$$Z_{jkh} \sim \mathcal{LN}\left(\mu_{jk}, \sigma^2_{jk}\right).$$

The first and second central moments of which are given by:

$$\begin{aligned} E[Z_{jkh}] &= \gamma_{jk} = \exp\left(\mu_{jk} + \sigma^2_{jk}/2\right), \\ V[Z_{jkh}] &= \gamma^2_{jk}\left(\exp\left(\sigma^2_{jk}\right) - 1\right). \end{aligned}$$

Assuming independence of expression across genes within a sample and across samples, the estimators of μ_{jk}, σ^2_{jk} and γ_{jk} are obvious:

$$\begin{aligned} \hat{\mu}_{jk} &= \frac{\sum_{r=1}^{n_q} \log(Z_{jkh})}{n_k}, \\ \hat{\sigma}^2_{jk} &= \frac{\sum_{r=1}^{n_q} \left[\log(Z_{jkh}) - \hat{\mu}_{jk}\right]^2}{n_k - 1}, \\ \hat{\gamma}_{jk} &= \exp\left(\hat{\mu}_{jk} + \hat{\sigma}^2_{jk}/2\right). \end{aligned}$$

In the low sample size setting, we may borrow information across cell types for estimating the variance. We do this in the following way:

$$\hat{\sigma}^2_j = \frac{\sum_{k=1}^{K}\sum_{h=1}^{n_k} \left[\log(Z_{jkh}) - \hat{\mu}_{jk}\right]^2}{n_P - 1},$$

giving

$$\hat{\sigma}^2_{jk} = \left(\frac{n_k}{n_P}\right)\hat{\sigma}^2_{jk} + \left(\frac{n_P - n_k}{n_P}\right)\hat{\sigma}^2_j.$$

B.1.3 Variance Weight Selection

The ICeD-T model allows for the provision of variance weights to be used in optimization of the log-normal model. In essence, variance weights increase or diminish the importance of residuals across various genes. The larger the weight, the more a residual at the given gene is discounted. We suggest the following weight schema be used in the ICeD-T model. In order to compute these weights, the user must provide the variance of the log-expression for each cell type and each gene.

Option 1: Homoscedastic Weights

"Homoscedastic Weights" is a misnomer as this corresponds to setting $\Delta_j = 1$ for all j. No discounting is performed as each gene is assumed to have the same variance.

Option 2: Maximal Variance Weights

The second option utilized by ICeD-T is termed "Maximal Variance Weights". For this weight structure, the weights are given by:

$$\Delta_j = \frac{\max\limits_k \left(\hat{\sigma}^2_{jk}\right)}{\operatorname*{median}\limits_j \left[\max\limits_k \left(\hat{\sigma}^2_{jk}\right)\right]}.$$

Robustness Considerations

To ensure that the variance weights do not make some genes too over- or under-influential, we let the top 15% of weights take 85th-percentile value and the bottom 15% take the 15th-percentile value. This will ensure that outliers causing inflated variances in certain genes are not overly influential upon results.

B.1.4 Fenton-Wilkinson Approximation

To simplify the maximum likelihood optimization used by the ICeD-T algorithm, we seek a closed-form approximation to the distribution of a sum of independent log-normals. The Fenton-Wilkinson approximation to the distribution of a sum of log-normals provides such an approach.

Consider a simplified framework of random variables Y_j for $j = 1, ..., K$ where $Y_j \sim \mathcal{LN}\left(\mu_k, \sigma_k^2\right)$ and the variable of interest $Y = \sum_{k=1}^{K} Y_k$. Fenton-Wilkinson approximates the distribution Y by another log-normal whose parameters are defined by moment-matching on the first and second central moments. Thus:

$$Y \sim \mathcal{LN}\left(\tilde{\mu}, \tilde{\sigma}^2\right)$$

Where we define:

$$\exp\left(\tilde{\mu} + \tilde{\sigma}^2/2\right) = \sum_{k=1}^{K} E[Y_k] = \sum_{k=1}^{K} \gamma_k,$$

and

$$\exp\left(2\tilde{\mu} + \tilde{\sigma}^2\right)\left[\exp\left(\tilde{\sigma}^2\right) - 1\right] = \sum_{k=1}^{K} V[Y_k] = \sum_{k=1}^{K} \gamma_k^2 \left(\exp\left(\sigma_k^2\right) - 1\right).$$

This provides the following closed forms for $\tilde{\mu}$ and $\tilde{\sigma}^2$:

$$\tilde{\mu} = \log\left(\sum_{k=1}^{K} \gamma_k\right) - \tilde{\sigma}^2/2,$$

and

$$\tilde{\sigma}^2 = \log\left(\sum_{k=1}^{K} \gamma_k^2 \left[\exp\left(\sigma_k^2\right) - 1\right] \Bigg/ \left[\sum_{k=1}^{K} \gamma_k\right]^2 + 1\right).$$

B.1.5 Mixture Sample Optimization

The ICeD-T model assumes that distribution of expression at a single gene in tumor sample i is a mixture over two log-normals, one component assuming the gene is a consistent gene and the other an aberrant one. This distribution is given by:

$$Y_{ij} \sim p_i \mathcal{LN}\left(\tilde{\mu}_{ijC}, \sigma^2_{ijC}\right) + (1-p_i)\mathcal{LN}\left(\tilde{\mu}_{ijA}, \sigma^2_{ijA}\right),$$

where:

$$\begin{aligned}\tilde{\mu}_{ij(\cdot)} &= \log\left(\sum_{k=1}^{K} \rho_{ik}\gamma_{jk}\right) - \sigma^2_{ij(\cdot)}/2,\\ \sigma^2_{ij(\cdot)} &= \Delta_j \sigma^2_{i(\cdot)}.\end{aligned}$$

ICeD-T can be run using two options. Option (1) represents a homoscedasticity assumption and assumes $\Delta_j = 1$ for all j. Option (2) allows for these variance weights to differ and must be specified before optimization. Utilizing these assumptions for variance provide superior performance in the estimation of cell type proportions compared to a direct application of Fenton-Wilkinson.

We also note that the separating feature between consistent and aberrant genes is the assumed variance. In particular, aberrant genes are assumed to have a larger variance. In essence, a larger variance for aberrant genes ”flattens” the observed likelihood, allowing for values inconsistent with the model proportions to become more likely.

In order to optimize this mixture distribution, we utilize an EM algorithm. We introduce missing data in the form of indicators of class membership, H_{ij}. When H_{ij} is 1, the gene is assumed consistent and when H_{ij} is 0, aberrant. Thus, a complete data log-likelihood for subject i is given by:

$$\ell_i = \sum_{j=1}^{n_G} H_{ij} \left[\log(p_i) - (1/2)\log(\tilde{\sigma}_{ij}^{2,(C)}) - \left(1/2\tilde{\sigma}_{ij}^{2,(C)}\right)\left(\log(y_{ij}) - \tilde{\mu}_{ij}^{(C)}\right)^2 \right] +$$
$$(1 - H_{ij}) \left[\log(1-p_i) - (1/2)\log(\tilde{\sigma}_{ij}^{2,(A)}) - \left(1/2\tilde{\sigma}_{ij}^{2,(A)}\right)\left(\log(y_{ij}) - \tilde{\mu}_{ij}^{(A)}\right)^2 \right].$$

The EM algorithm will replace H_{ij} with their posterior expectations $w_{ij} = E\left[H_{ij} \middle| Y_{ij}, \phi\right]$ prior to optimization at each iteration where ϕ is a collection of current estimates of abundances, individual variances, and aberrance proportions.

B.1.5.1 Update Posterior Means

For a set of parameter estimates arising from iteration (t), it is readily seen that when:

$$\begin{aligned} w_{ij} &= E\left[H_{ij}\middle|Y_{ij},\phi\right] \\ &= \frac{\left(\frac{p_i}{\tilde{\sigma}_{ij}^{(C)}}\right)\exp\left\{\left(\frac{-1}{2\tilde{\sigma}_{ij}^{2,(C)}}\right)\left(\log(Y_{ij})-\tilde{\mu}_{ij}^{(C)}\right)^2\right\}}{\left(\frac{p_i}{\tilde{\sigma}_{ij}^{(C)}}\right)\exp\left\{\left(\frac{-1}{2\tilde{\sigma}_{ij}^{2,(C)}}\right)\left(\log(Y_{ij})-\tilde{\mu}_{ij}^{(C)}\right)^2\right\}+\left(\frac{1-p_i}{\tilde{\sigma}_{ij}^{(A)}}\right)\exp\left\{\left(\frac{-1}{2\tilde{\sigma}_{ij}^{2,(A)}}\right)\left(\log(Y_{ij})-\tilde{\mu}_{ij}^{(A)}\right)^2\right\}}. \end{aligned}$$

B.1.5.2 Update p_i, $\tilde{\sigma}_i^{2,((\cdot))}$, and $\boldsymbol{\rho}_i$

It is simple to show that the likelihood is separable with respect to p_i and $\left(\boldsymbol{\rho}_i^T,\sigma_{iC}^2,\sigma_{iA}^2\right)$. Thus, we may estimate these parameters separately.

Update p_i:

We update p_i with its MLE estimate, given by:

$$\hat{p}_i = \frac{\sum_{j=1}^{n_G} E[H_{ij}\big|Y_{ij},\phi]}{n_G}.$$

Update $\left(\boldsymbol{\rho}_i^T,\sigma_{iC}^2,\sigma_{iA}^2\right)$

The cell type proportions and variance parameters are not separable within the likelihood and must be updated simultaneously. We opt for a block coordinate ascent algorithm consisting of two blocks; cell type proportions compose block 1 and variance terms compose block 2. Block 1 is updated while block 2 is held fixed, then block 2 is updated while block 1 is fixed. This process is repeated until convergence.

Consider first the variance terms without variance weights. Holding the cell type proportions fixed, the terms pertaining to aberrant and consistent genes are separable. We focus on the portion of the complete data log-likelihood pertaining to the consistent variance term, though similar results hold for σ^2_{iA}:

$$\begin{aligned}
\ell_i\left(\tilde{\sigma}_i^{2,(C)}\right) &= \sum_{j=1}^{n_G} w_{ij}\left[-(1/2)\log\left(\Delta_j\tilde{\sigma}_i^{2,(C)}\right) - \left(1/2\Delta_j\tilde{\sigma}_i^{2,(C)}\right)\left(\log(Y_{ij}) - \tilde{\mu}_{ij}^{(C)}\right)^2\right] \\
&= \sum_{j=1}^{n_G} w_{ij}\left[-(1/2)\log\left(\Delta_j\tilde{\sigma}_i^{2,(C)}\right) - \left(1/2\Delta_j\tilde{\sigma}_i^{2,(C)}\right)\left(\nu_{ij} + \Delta_j\tilde{\sigma}_i^{2,(C)}/2\right)^2\right]
\end{aligned}$$

where $\nu_{ij} = \log\left(Y_{ij}\right) - \log\left(\sum_{k=1}^{K}\rho_{ik}\gamma_{jk}\right)$.

Taking the first derivative with respect to the consistent variance term, we have:

$$\begin{aligned}
\dot{\ell}\left(\tilde{\sigma}_i^{2,(C)}\right) = \sum_{j=1}^{n_G} w_{ij}\Bigg[&\left(\frac{-1}{2\tilde{\sigma}_i^{2,(C)}}\right) + \left(\frac{1}{2\Delta_j\tilde{\sigma}_i^{4,(C)}}\right)\left(\nu_{ij} + \Delta_j\tilde{\sigma}_i^{2,(C)}/2\right)^2 - \\
&\left(\frac{1}{\Delta_j\tilde{\sigma}_i^{2,(C)}}\right)\left(\nu_{ij} + \Delta_j\tilde{\sigma}_i^{2,(C)}/2\right)\Bigg].
\end{aligned}$$

Under option (2), we did not find a closed form update for these variance terms opting to use numerical optimization. Under option (1), we can further reduce this equation by plugging in $\Delta_j = 1$ for all j:

$$\begin{aligned}
\dot{\ell}\left(\tilde{\sigma}_i^{2,(C)}\right) &= \sum_{j=1}^{n_G}\left(\frac{1}{2\tilde{\sigma}_i^{4,(C)}}\right)\left[-\tilde{\sigma}_i^{2,(C)} + \nu_{ij}^2 + \nu_{ij}\tilde{\sigma}_i^{2,(C)} + \tilde{\sigma}_i^{4,(C)}/4 - \tilde{\sigma}_i^{2,(C)}\nu_{ij} - \tilde{\sigma}_i^{4,(C)}/2\right] \\
&= \sum_{j=1}^{n_G}\left(\frac{1}{2\tilde{\sigma}_i^{4,(C)}}\right)\left[-\left(\tilde{\sigma}_i^{4,(C)}/4 + \tilde{\sigma}_i^{2,(C)}\right) + \nu_{ij}^2\right] \\
&= \sum_{j=1}^{n_G}\left(\frac{1}{2\tilde{\sigma}_i^{4,(C)}}\right)\left[\left(\tilde{\sigma}_i^{2,(C)}/2 + 1\right)^2 + \nu_{ij}^2\right]
\end{aligned}$$

Setting equal to 0 and solving, we have a closed form update for $\tilde{\sigma}_i^{2,(C)}$:

$$\tilde{\sigma}_i^{2,(C)} = 2\left[\sqrt{\frac{\sum_{j=1}^{n_G} w_{ij}\nu_{ij}^2}{\sum_{j=1}^{n_G} w_{ij}} + 1} - 1\right].$$

We now turn to the cell type proportions piece, assuming the variance terms are held fixed. The complete data log-likelihood pertaining to these parameters is given by:

$$\ell_i = \sum_{j=1}^{n_G} w_{ij}\left[\log(p_i) - (1/2)\log(\tilde{\sigma}_i^{2,(C)}) - \left(1/2\tilde{\sigma}_i^{2,(C)}\right)\left(\log(Y_{ij}) - \tilde{\mu}_{ij}^{(C)}\right)^2\right] +$$
$$(1 - w_{ij})\left[\log(1 - p_i) - (1/2)\log(\tilde{\sigma}_i^{2,(A)}) - \left(1/2\tilde{\sigma}_i^{2,(A)}\right)\left(\log(Y_{ij}) - \tilde{\mu}_{ij}^{(A)}\right)^2\right].$$

Before constructing the derivative of this likelihood with respect to our cell type proportions, we examine derivatives an interior of the likelihood to improve clarity of the full derivation. In the following, let $\eta_{ij} = \sum_{k=1}^{K} \rho_{ik}\gamma_{jk}$.

$$\frac{\partial \mu_{ij(\cdot)}}{\partial \rho_i} = \frac{\partial}{\partial \rho_i}\left[\log\left(\sum_{k=1}^{K} \rho_{ik}\gamma_{jk}\right) - \sigma_{ij(\cdot)}^2/2\right] = \frac{\boldsymbol{\gamma_j}}{\sum_{k=1}^{K} \rho_{ik}\gamma_{jk}} = \frac{\boldsymbol{\gamma}_j}{\eta_{ij}}$$

Plugging this value into the gradient for the complete data log-likelihood, we have:

$$
\begin{aligned}
\dot{\ell}_i = \sum_{j=1}^{n_G} & w_{ij} \left[(1/\tilde{\sigma}_{ij}^{2,(C)}) \left(\log(Y_{ij}) - \tilde{\mu}_{ij}^{(C)} \right) \right] \left(\frac{\partial \mu_{ij}^{(C)}}{\partial \rho_i} \right) + \\
& \qquad (1 - w_{ij}) \left[(1/\tilde{\sigma}_{ij}^{2,(A)}) \left(\log(Y_{ij}) - \tilde{\mu}_{ij}^{(A)} \right) \right] \left(\frac{\partial \mu_{ij}^{(A)}}{\partial \rho_i} \right) \\
= \sum_{j=1}^{n_G} & \gamma_j \left[\left(\frac{w_{ij} \left(\log(Y_{ij}) - \tilde{\mu}_{ij}^{(C)} \right)}{\tilde{\sigma}_{ij}^{2,(C)} \eta_{ij}} \right) + \left(\frac{(1 - w_{ij}) \left(\log(Y_{ij}) - \tilde{\mu}_{ij}^{(A)} \right)}{\tilde{\sigma}_{ij}^{2,(A)} \eta_{ij}} \right) \right] \\
= \boldsymbol{\gamma}^T & \left[\left(\frac{w_{ij} \left(\log(Y_{ij}) - \tilde{\mu}_{ij}^{(C)} \right)}{\tilde{\sigma}_{ij}^{2,(C)} \eta_{ij}} \right) + \left(\frac{(1 - w_{ij}) \left(\log(Y_{ij}) - \tilde{\mu}_{ij}^{(A)} \right)}{\tilde{\sigma}_{ij}^{2,(A)} \eta_{ij}} \right) \right]_j .
\end{aligned}
$$

To ensure proper constraints during fit, numerical optimization routines from R's `constrOptim` function in the `alabama` package are used to optimize the log-likelihood with respect to ρ_i.

When no information is assumed for the proportion of a $(K + 1)$-st cell type (e.g. a tumor cell type), these proportions are non-negative and allowed to sum to a value less than 1. If the proportion of a $(K + 1)$-st cell type is assumed (e.g. tumor purity), the proportions are constrained to sum to $1 - \rho_{K+1}$. As noted in the main paper, the $(K + 1)$-st cell type is assumed not to express or to express at a minimal level across the n_G genes used for optimization.

B.2 Simulations Supplement

The first assessment of the estimation properties of the ICeD-T model was performed on *in silico* simulated datasets. For each simulation, we constructed two sets of expression pseudo-experiments: reference expression datasets from 5 simulated reference cell types and reference expression datasets from 135 mixture datasets composed of expression from 4 of these 5 cell types. Each expression experiment consists of expression values across 250 common loci. Within the mixtures, one cell type represents a ”missing” cell type for each sample; this cell type is known to be present in the mixture but it does not express at the 250 modeled loci.

These simulations were built in three main steps: Step (1) generates purified reference sample expressions and variance measures; Step (2) generates mixture expression files for deconvolution; and step (3) edits the output expressions from step 2 to allow for aberrant gene behavior.

B.2.1 Step 1 - Generating Pure Sample Expressions

The first element in generating pure sample expressions is to define profiles from which each ”purified reference” sample is simulated. For each locus separately, it is randomly determined whether the locus is lowly, moderately, or highly expressed. In addition, one of the four expressed cell types is labeled the indicated cell type for this locus while the remaining cell types are considered background. We then simulate a mean log-expression for each gene and cell type according to the following table:

Level	**Pct. Loci**	**Background**	**Indicated**
Low	33%	$Uniform(2.0, 4.0)$	$Uniform(3.5, 5)$
Moderate	33%	$Uniform(4.0, 6.0)$	$Uniform(5.5, 7)$
High	33%	$Uniform(6.0, 8.0)$	$Uniform(7.5, 9)$

Once the mean log-expressions are simulated, we must construct a reasonable variance schema for these average log-expression profiles. We construct a mean-variance relationship in the log-expression setting by mirroring an example found in FPKM-normalized RNA-seq data.

Read counts from purified samples of B-cells (20), CD4 T-cells (20), CD8 T-cells (20), Monocytes (20), Neutrophils (20) and Natural Killers (14) were downloaded from the Array Express website from the Linsley et al study [60]. The read counts for each sample are FPKM normalized, utilizing the (75th-percentile read count/1000) instead of total read depth for each subject. The mean-variance relationship is modeled across 441 immune-related genes for each of these six cell types using a Loess curve, similar to the procedure utilized by VOOM [67]. This Loess curve was used to map the simulated log-expression means for each gene and cell type to a data-supported variance measure. Random error was also introduced.

Following the generation of the mean and variance profiles, the 5 or 15 purified, reference-sample pseudo-experiments are generated for each cell type from its profile via a log-normal distribution.

B.2.2 Step 2 - Generating Mixture Expressions

We must now generate the mixture expression pseudo-experiments. We first generate the proportion of the missing cell type from a standard normal distribution with mean 0.60 and standard deviation 0.15. In addition, any of these proportions falling below 17% or above 95% are set at 17% and 95% respectively. The remaining proportions are simulated from a Dirichlet distribution with average abundances ranging from 15% to 40%.

With the proportions generated for each of the 5 cell types and each subject, we turn to simulating the expression experiments. For each subject individually, we construct mixture experiments according to the following algorithm.

(1) *Simulate Pure Sample Expression for non-missing cell types*

$$X_{ijk} \sim \exp\left(\mathcal{N}\left(\mu_{jk}, \sigma^2_{jk}\right)\right)$$

(2) *Mix Pure Sample Expressions*

$$Y_{ij} \sim \sum_{k=1}^{4} \rho_{ik} X_{ijk}$$

Thus, these mixture expression experiments are simulated as true convolutions of independent log-normals. In this way, we can examine the adequacy of our approximated distribution.

B.2.3 Step 3 - Edit Mixtures to Create Aberrance

The final step in the mixture experiments is to allow loci to misbehave. We allow 3 mechanisms for misbehavior. Mechanism 1 takes the expression of the indicated cell type and downregulates it to 25% - 75% of its true level; mechanism 2 takes the expression of the indicated cell type and upregulates it to 133% - 400% of its true level; and mechanism 3 allows the missing cell type to express at the background levels established above. The table below summarizes these mechanisms.

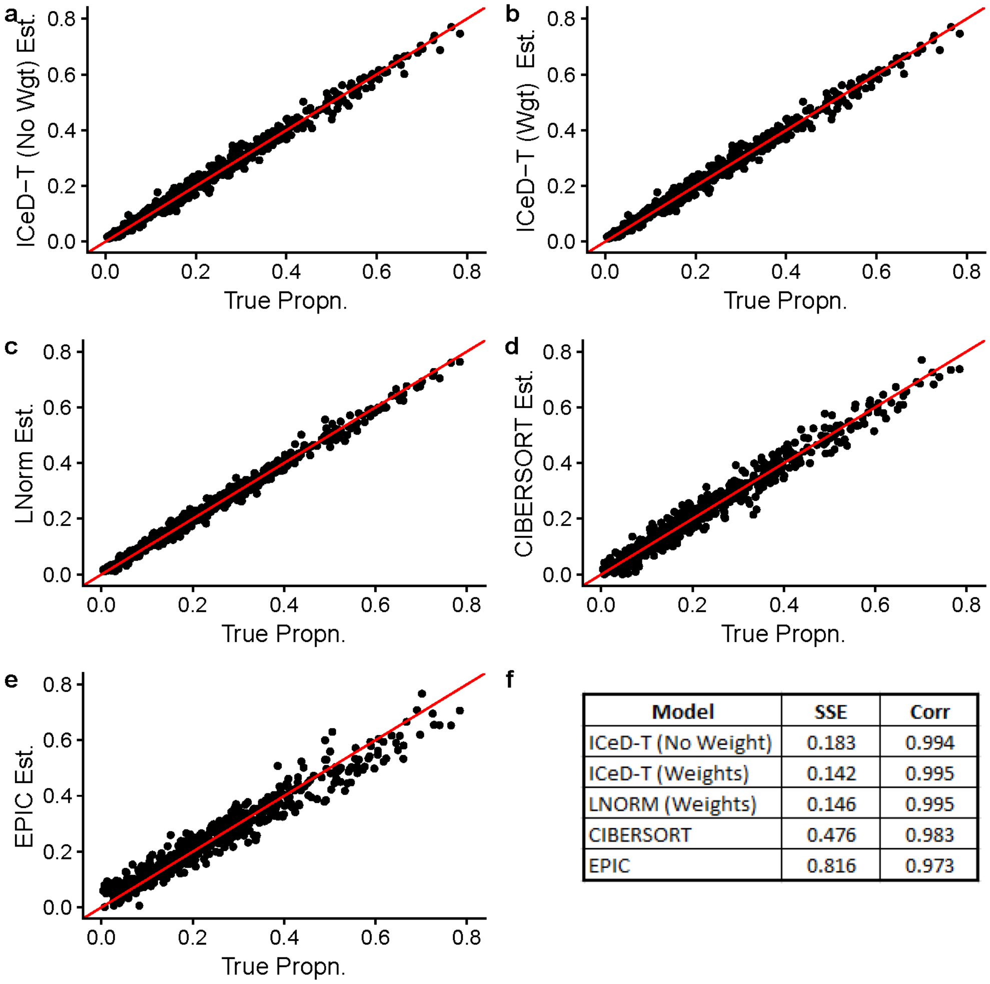
Mechanism	Pct. Ab. Loci	Indicated CT Effect	Missing CT Exp.
1 - Downregulate	25%	$Uniform(25\%, 75\%)$	0
2 - Upregulate	25%	$Uniform(133\%, 400\%)$	0
3 - Missing Exp.	50%	0	$Uniform(.,.)$

Table B.2: Pct Ab. Loci = Percentage of Aberrant Loci Effected, Indicated CT Effect = Effect on the expression of Indicated Cell Type, Missing CT Exp = Expression Level of Missing Cell Type

For impacted loci, expression is resimulated as in B.2 with the revised expression profiles. The number of loci impacted is allowed to vary from 0% to 30% of the total expression and the resulting estimates are examined.

B.2.4 Results

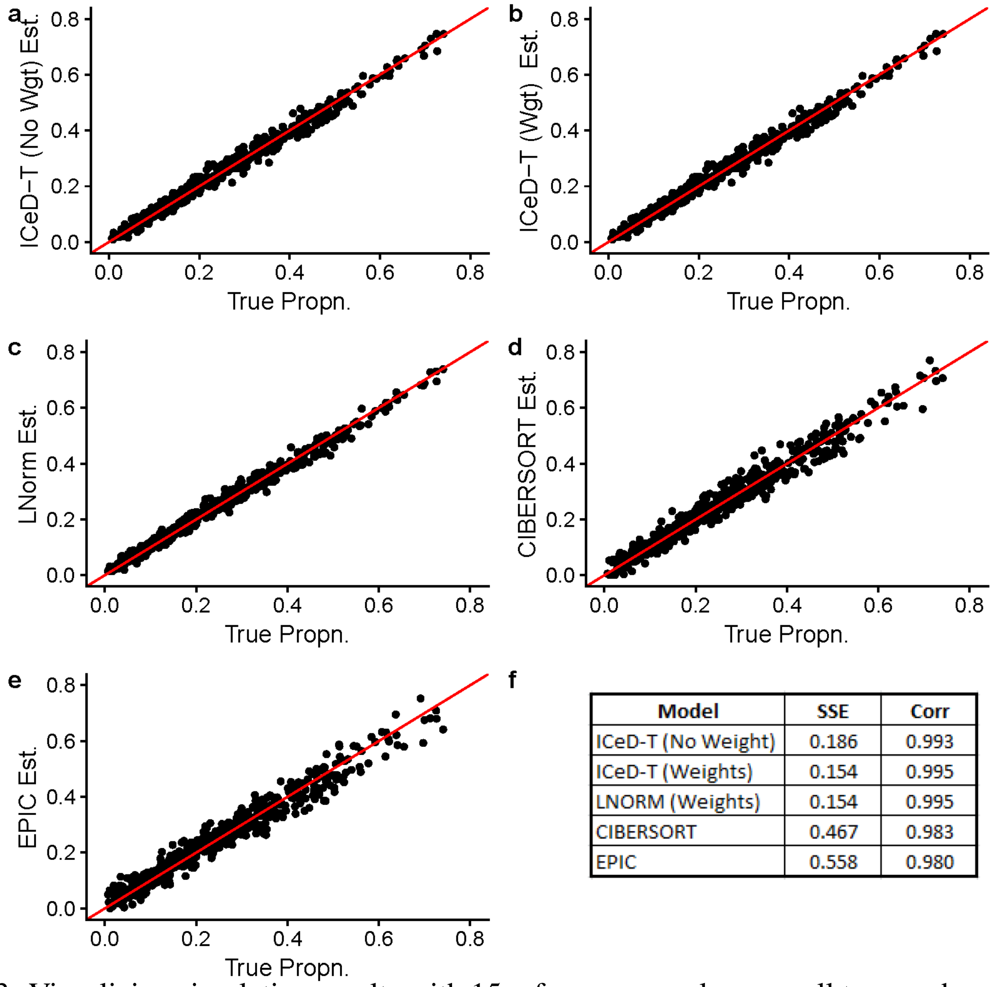
Pct Ab. = 0%, No. Rep. = 5



Model	SSE	Corr
ICeD-T (No Weight)	0.183	0.994
ICeD-T (Weights)	0.142	0.995
LNORM (Weights)	0.146	0.995
CIBERSORT	0.476	0.983
EPIC	0.816	0.973

Figure B.2: Visualizing simulation results with 5 reference samples per cell type and no aberrance.

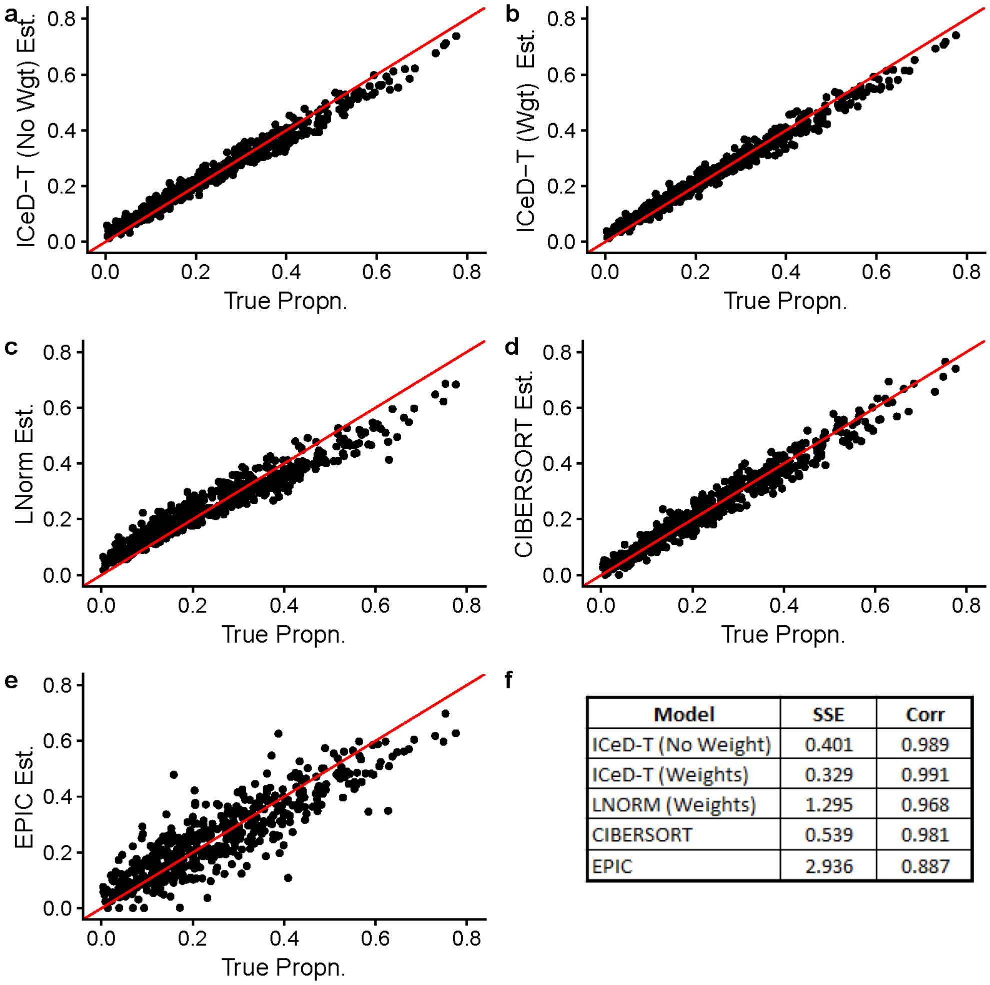
Pct Ab. = 0%, No. Rep. = 15



Model	SSE	Corr
ICeD-T (No Weight)	0.186	0.993
ICeD-T (Weights)	0.154	0.995
LNORM (Weights)	0.154	0.995
CIBERSORT	0.467	0.983
EPIC	0.558	0.980

Figure B.3: Visualizing simulation results with 15 reference samples per cell type and no aberrance.

Pct Ab. = 15%, No. Rep. = 5



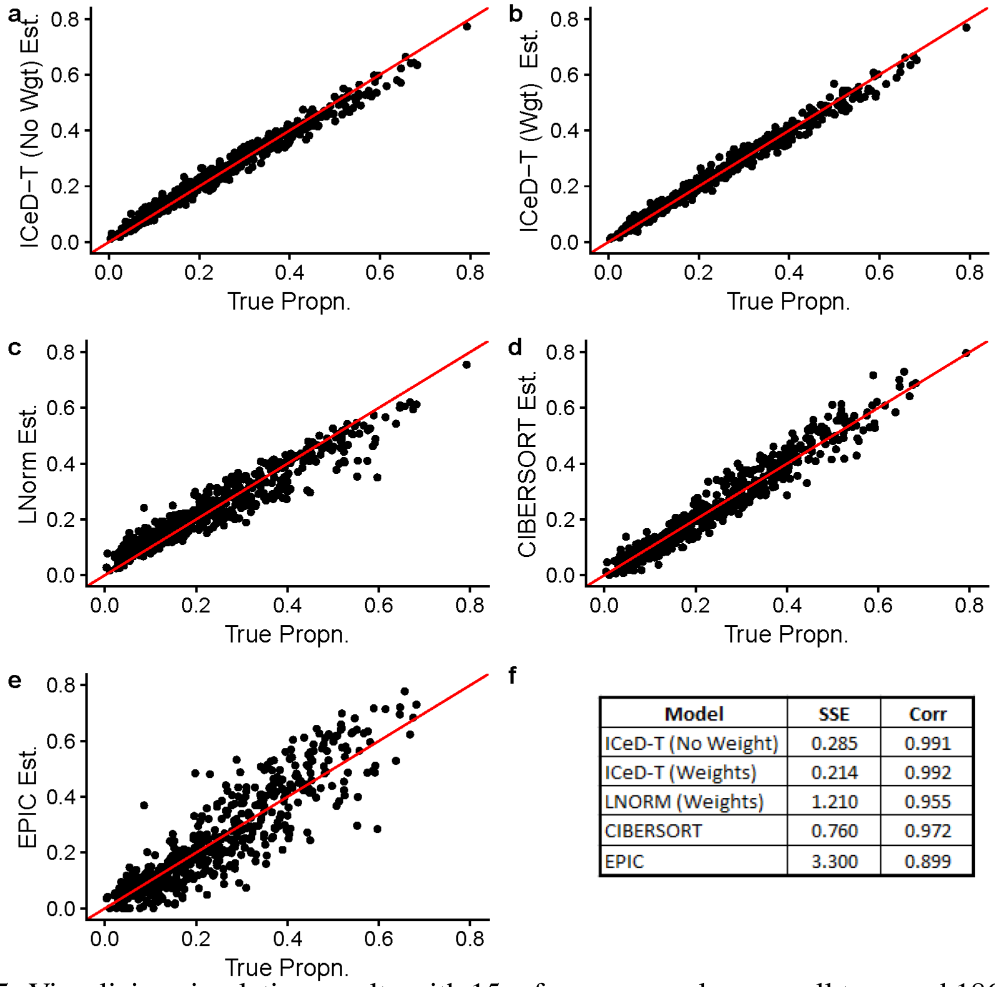
Model	SSE	Corr
ICeD-T (No Weight)	0.401	0.989
ICeD-T (Weights)	0.329	0.991
LNORM (Weights)	1.295	0.968
CIBERSORT	0.539	0.981
EPIC	2.936	0.887

Figure B.4: Visualizing simulation results with 5 reference samples per cell type and 15% of genes behaving aberrantly.

Model	**Aberrant**	**1Q**	**Med**	**3Q**
ICeD-T (No Weight)	Yes	0.000	0.114	0.607
	No	0.625	0.748	0.824
	p_i	0.572	0.612	0.655
ICeD-T (Weights)	Yes	0.004	0.468	0.823
	No	0.804	0.884	0.931
	p_i	0.718	0.768	0.803

Table B.3: Summarizing ICeD-T's ability to detect aberrant gene behavior (Pct. Ab. = 15%, No. Rep. = 5).

Pct Ab. = 18%, No. Rep. = 15



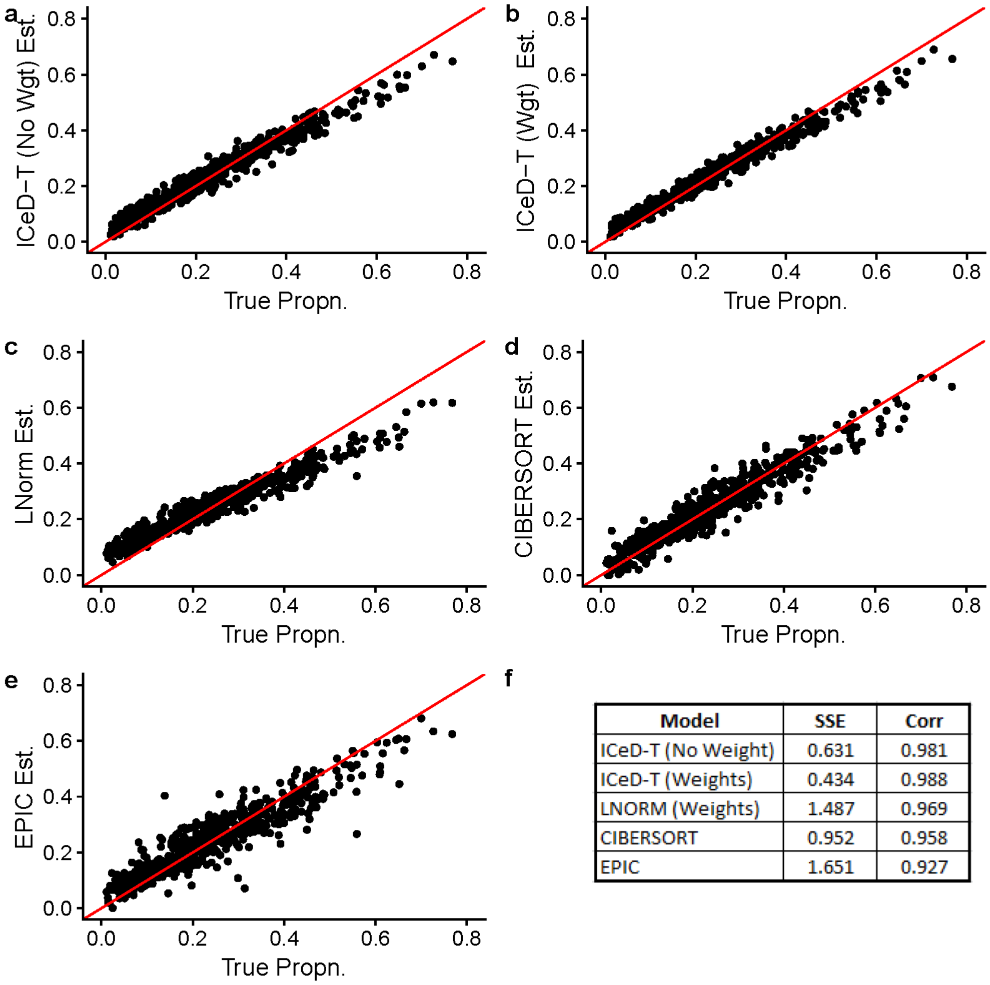
Model	SSE	Corr
ICeD-T (No Weight)	0.285	0.991
ICeD-T (Weights)	0.214	0.992
LNORM (Weights)	1.210	0.955
CIBERSORT	0.760	0.972
EPIC	3.300	0.899

Figure B.5: Visualizing simulation results with 15 reference samples per cell type and 18% of genes behaving aberrantly.

Model	**Aberrant**	**1Q**	**Med**	**3Q**
ICeD-T (No Weight)	Yes	0.000	0.043	0.538
	No	0.647	0.769	0.838
	p_i	0.579	0.613	0.657
ICeD-T (Weights)	Yes	0.000	0.114	0.744
	No	0.797	0.872	0.920
	p_i	0.697	0.738	0.772

Table B.4: Summarizing ICeD-T's ability to detect aberrant gene behavior (Pct. Ab. = 18%, No. Rep. = 15).

Pct Ab. = 30%, No. Rep. = 5



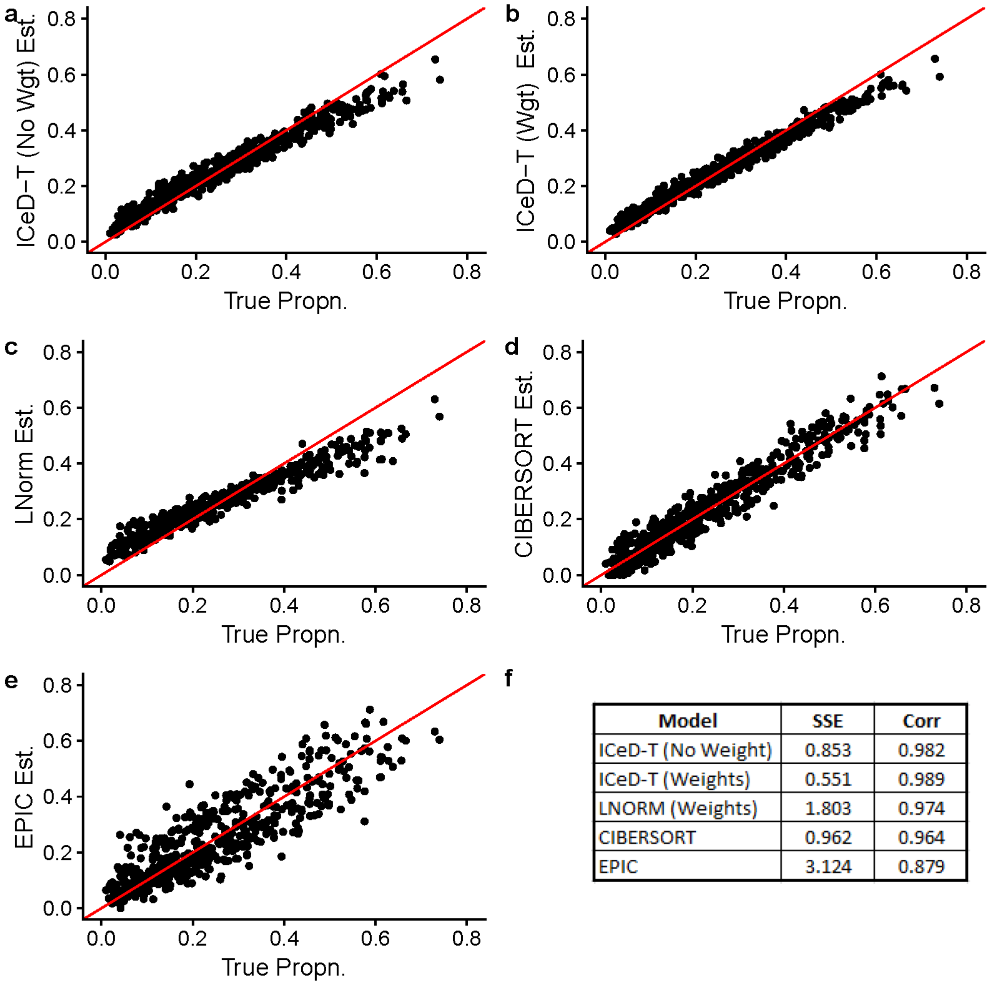
Model	SSE	Corr
ICeD-T (No Weight)	0.631	0.981
ICeD-T (Weights)	0.434	0.988
LNORM (Weights)	1.487	0.969
CIBERSORT	0.952	0.958
EPIC	1.651	0.927

Figure B.6: Visualizing simulation results with 5 reference samples per cell type and 30% of genes behaving aberrantly.

Model	**Aberrant**	**1Q**	**Med**	**3Q**
ICeD-T (No Weight)	Yes	0.001	0.194	0.645
	No	0.618	0.659	0.791
	p_i	0.530	0.555	0.587
ICeD-T (Weights)	Yes	0.011	0.552	0.824
	No	0.780	0.862	0.897
	p_i	0.673	0.707	0.725

Table B.5: Summarizing ICeD-T's ability to detect aberrant gene behavior (Pct. Ab. = 30%, No. Rep. = 5).

Pct Ab. = 35%, No. Rep. = 15



Model	SSE	Corr
ICeD-T (No Weight)	0.853	0.982
ICeD-T (Weights)	0.551	0.989
LNORM (Weights)	1.803	0.974
CIBERSORT	0.962	0.964
EPIC	3.124	0.879

Figure B.7: Visualizing simulation results with 15 reference samples per cell type and 35% of genes behaving aberrantly.

Model	Aberrant	1Q	Med	3Q
ICeD-T (No Weight)	Yes	0.002	0.202	0.582
	No	0.574	0.678	0.734
	p_i	0.480	0.503	0.529
ICeD-T (Weights)	Yes	0.013	0.406	0.730
	No	0.714	0.790	0.837
	p_i	0.597	0.621	0.643

Table B.6: Summarizing ICeD-T's ability to detect aberrant gene behavior (Pct. Ab. = 30%, No. Rep. = 15).

We see from the above that the ICeD-T model with and without weights provides the best fit for these simulated data in terms of both sum of squared error and correlation. The aberrance model adequately handles the misbehavior across loci even up to 30% aberrance, with the weighted model providing the strongest estimation. It most closely estimates the proportion of aberrant genes and provides stronger distinctions in the probabilities of aberrance given the data.

As we reach 30% aberrance, we do note a slight bias in ICeD-T's results beginning to become evident near the tails. However, even when compared against CIBERSORT–a method which provides a very strong runner-up in these simulated data– we see that ICeD-T is superior. This is a classic case of the bias-variance trade-off. ICeD-T allows some bias to impact results as aberrance increases, but maintains a strong linear relationship. CIBERSORT, on the other hand, experiences increasing variability and a slightly diminshed linear relationship as the amount of aberrance increases.

We also fit the ICeD-T model without using estimates of tumor purity (data not shown). The model performs well up to 30% aberrance, however, at around 30% aberrance it begins to struggle to capture aberrant genes appropriately. Regardless of this fact, the ICeD-T model with weights continues to be one of the strongest performers even up to 30% aberrance.

B.3 CIBERSORT Flow Cytometry Validation

The second assessment of the performance properties of the ICeD-T model is performed in real data. In their paper ”Robust Enumeration of Cell Subsets from Tissue Expression Profiles,” the creators of CIBERSORT validate their modeling procedure on peripheral blood mononuclear cells (PBMCs) extracted from 20 adult subjects. We reanalyze this dataset using CIBERSORT's web application, the ICeD-T model, and EPIC.

B.3.1 Data

PBMCs were extracted from each of 20 adult subjects. For each sample, expression profiles were created using microarray expression analysis. Additionally, each sample was examined using flow cytometry to measure the ground-truth abundance of each of the immune cell types composing the PBMCs. The resulting datasets were provided to us directly by Newman et al. In addition, the microarray expression data from purified samples of 22 immune cell types used to construct LM22 were also provided.

B.3.2 Cell Type Size Correction

The authors of EPIC advocate the use of cell size factors to correct regression results for differences in the productivity of various cell types composing mixture experiments. In their work, “Simultaneous Enumeration Of Cancer And Immune Cell Types From Bulk Tumor Gene Expression Data,” they note that cells of various types produce differing levels of mRNA. We borrow these cell size factors here and use them to correct the results of CIBERSORT and ICeD-T as was performed below. The cell size factors utilized here are provided below. Cell size factors are incorporated into model estimates after running the ICeD-T or CIBERSORT models as was done in EPIC. Define s_k to be the cell size factor for a cell type k. Then the revised estimate of abundance for cell type k is

Cell	**Size Factor**	**Extensions**
B-Cells	0.40	Naive and memory B-cells
T-Cells	0.40	Naive, memory-resting and memory-activated CD4 T-cells; CD8 T-cells; Delta-Gamma T-cells
NK cells	0.42	None
Monocytes	1.40	Macrophages, Dendritic Cells
Neutrophils	0.15	Eosinophils, Mast Cells

Table B.7: EPIC-derived cell type size factors with extensions to cell types not explicitly measured.

given by:

$$\rho_k^* = \frac{\rho_k/s_k}{\sum_{i=1}^{K}(\rho_k/s_k)}.$$

B.3.3 Model Fit Description

CIBERSORT:

The CIBERSORT web application (Version: CIBERSORT Jar 1.06) was used to fit these microarray data. The model was fit using the LM22 signature matrix run with quantile normalization and 500 permutations.

EPIC:

The EPIC library was downloaded from https://github.com/GfellerLab/EPIC in February 2018. The mixture expression data is quantile normalized and fit to the LM22 reference matrix using EPIC with default options, except scaleExprs set to FALSE.

ICeD-T:

The ICeD-T model was fit to the LM22 reference without specifying the proportions of extraneous cell types in the model and no weights, maximal variance weights, and maximal expression variance weights. Variance weights were computed using the variance of log-transformed expression across all purified references of a given cell type.

Quantile Normalization:

EPIC and ICeD-T require that the modeled mixture data be measured on the same scale as the design matrix utilized for modeling. To this end, the purified references used to compose the LM22 matrix are quantile normalized. The mixture data are then quantile normalized to the target distribution specified by the purified references using the `preprocessCore` library and its functions `normalize.quantiles.determine.target` and `normalize.quantiles.use.target`. This normalization is performed prior to specification of gene and cell type variance measures.

Result Renormalization:

Results are handled in the manner suggested by Newman et al in personal correspondence as was performed for their manuscript. All estimated cell type proportions are restricted to the ten examined cell types: B-cells naive, B-cells memory, CD8+ T-cells, naive CD4+ T-cells, resting memory CD4+ T-cells, activated memory CD4+ T-cells, Delta-gamma T-cells, Activated and resting natural killer cells, and Monocytes (including the modeled macrophage populations). These proportions are then renormalized to sum to 100.

B.3.4 Fit Comparison

The following table details the correlations and sum of squared errors for each of the fit models. As noted above, each of these measures use cell size corrected proportions for examination.

Model	SSE	Cor
ICeD-T (No Wgt)	13099.93	0.53
ICeD-T (Max Var Wgt)	12050.67	0.59
CIBERSORT	14146.59	0.65
EPIC	29427.74	0.31

Table B.8: Fit summary statistics for each model compared against flow cytometry measured ground-truth.

We note from the above that the CIBERSORT model provides the best results in terms of correlations. However, each of the fit ICeD-T models provide superior fit in terms of sums of squared errors. When using variance weights, the correlations between ICeD-T estimates and CIBERSORT become comparable as well ($\sim$0.60 vs. 0.65). Thus, it appears that the ICeD-T method is comparable to CIBERSORT in terms of correlation and provides superior results in terms of error.

In the following considerations, we will focus on the ICeD-T model with maximal variance weights. Despite the fact that the maximal expression weights produced the best fit for both overall correlation and sum of squared errors, it has notably weaker fit for many important cell types (e.g. CD4, CD8). Compared to CIBERSORT, in addition to having lower overall error, ICeD-T appears to provide superior performance for memory B-cells, naive CD4 T-cells, and gamma-delta T-cells. Both CIBERSORT and ICeD-T provide comparable performance with respect to monocyte expression. Both models struggle with CD8 T expression despite being well correlated for this cell type as CIBERSORT tends to overestimate expression by in the upper tail where ICeD-T seems to underestimate.

The results provided by the EPIC model are very poor for this dataset. However, this is not a condemnation of EPIC's use in real data. EPIC was designed for RNA-seq data, not for microarrays.

Thus, the weighting structure and gene selection for the fit shown here may not be suitable for EPIC's off-the-shelf options.

ICeD-T, No Weights

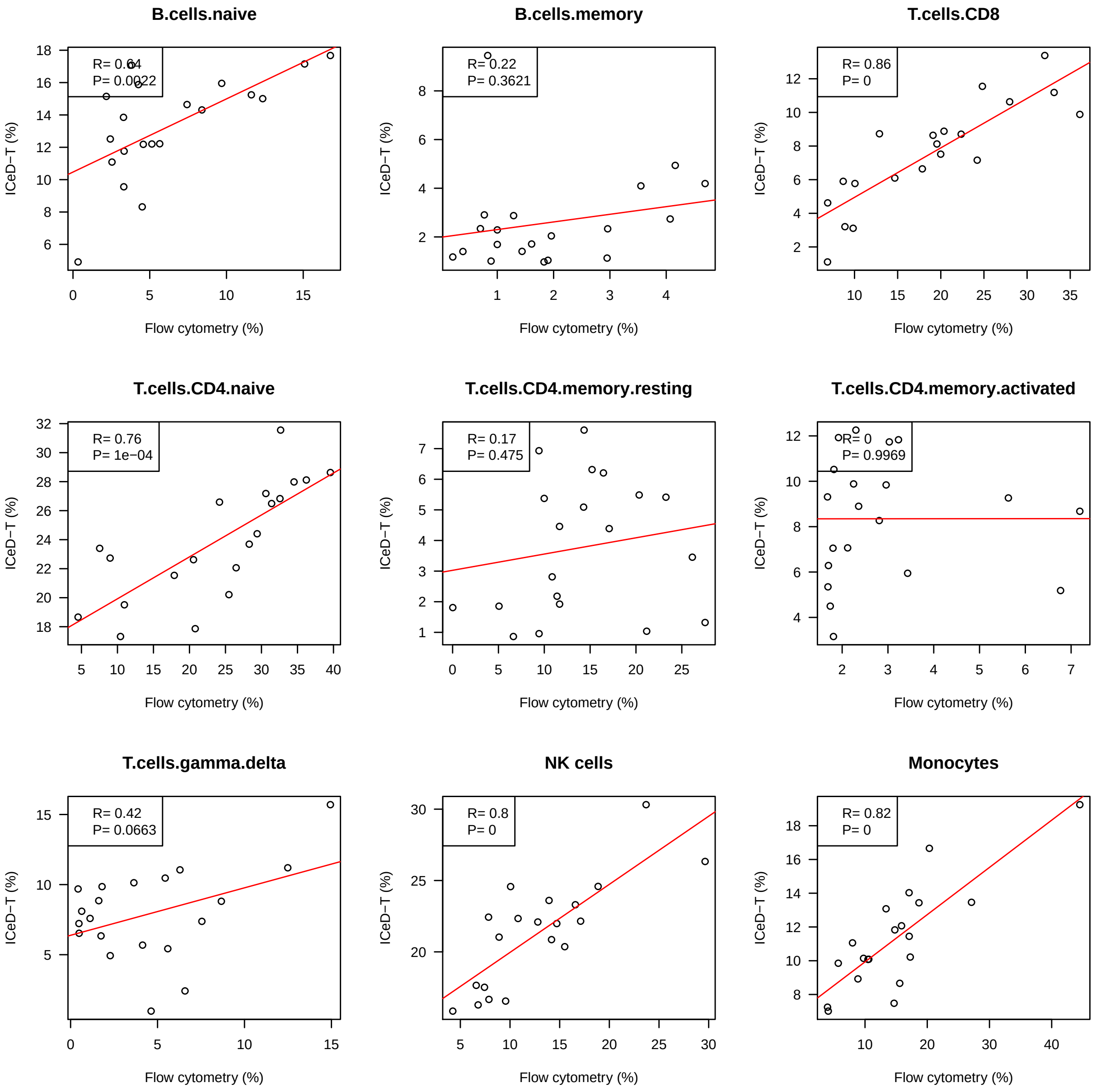


Figure B.8: Plotting true, relative abundances of 9 immune cell subpopulations against ICeD-T (no weights) estimates.

ICeD-T, Max Variance Weights

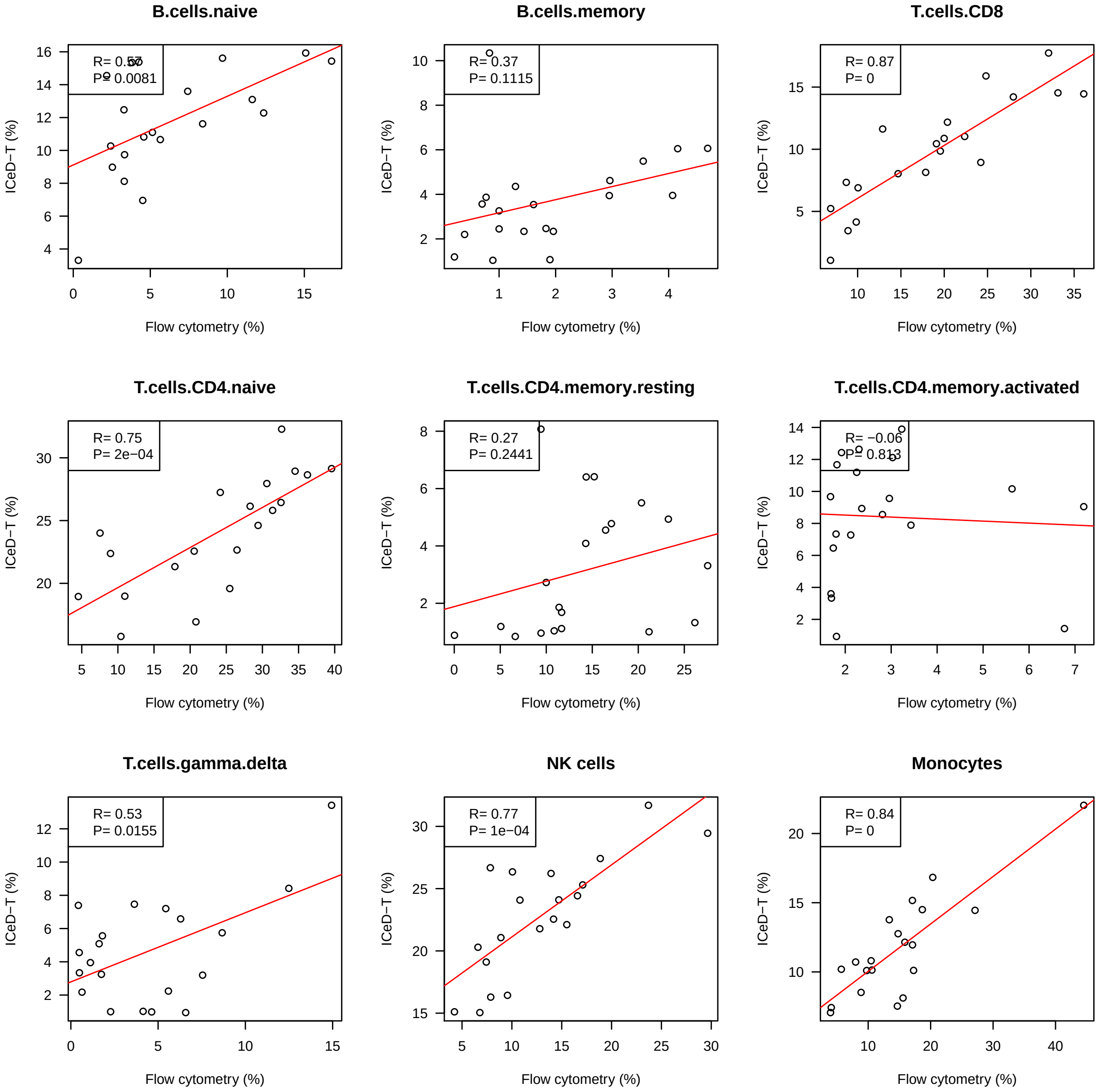


Figure B.9: Plotting true, relative abundances of 9 immune cell subpopulations against ICeD-T (weights) estimates.

CIBERSORT

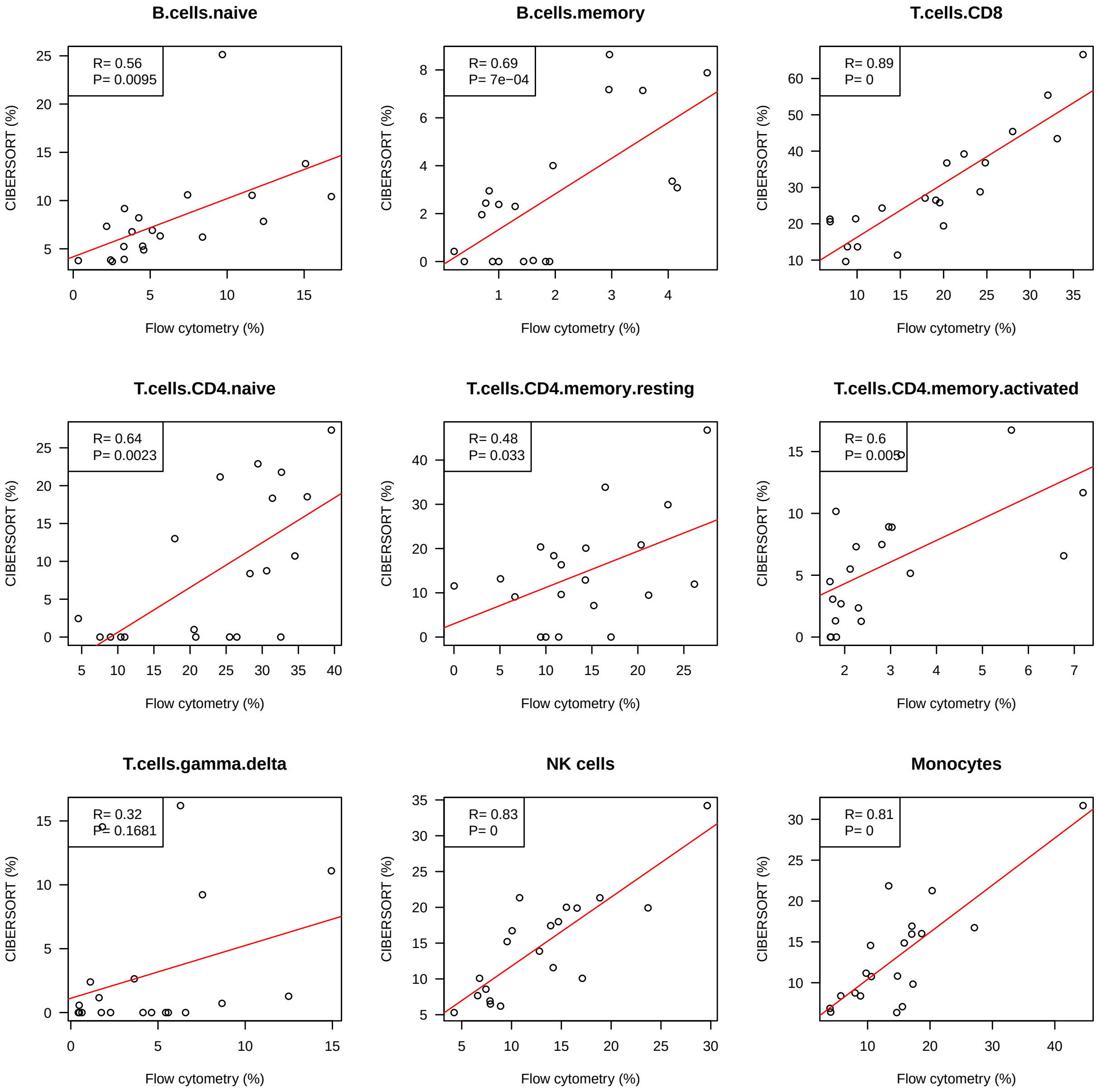


Figure B.10: Plotting true, relative abundances of 9 immune cell subpopulations against CIBERSORT estimates.

B.4 EPIC Melanoma Data Validation

The third examination of the estimation properties of ICeD-T is performed on validation data provided by Racle et al. It offers an opportunity to evaluate the performance of ICeD-T on RNA-seq experiments from tumor samples.

B.4.1 Data

For more information regarding this dataset, see ’Simultaneous Enumeration of Cancer and Immune Cell Types from Bulk Tumor Gene Expression Data’ from Racle et al. In brief, cells were extracted from the lymph nodes of four patients with stage III melanomas. A portion of each of the single cell suspensions obtained from these subjects was used for a flow cytometric analysis while the remaining portion was used for bulk RNA-sequencing.

The data was extracted directly from the EPIC library, file accession pathway given here: `EPIC-master/data/melanoma_data.rda`. This RData files contains a single list object `melanoma_data`, which houses fields containing the TPM-normalized RNA-seq expression for each subject, the flow-cytometry measured cell type proportions, and the predicted EPIC cell type proportions obtained using the TRef reference matrix.

B.4.2 Model Fit Description

CIBERSORT:

The CIBERSORT web application (Version: CIBERSORT Jar 1.06) was used to fit these TPM normalized RNA-seq data. The model was fit using the LM22 signature matrix and run with quantile normalization disabled.

EPIC:

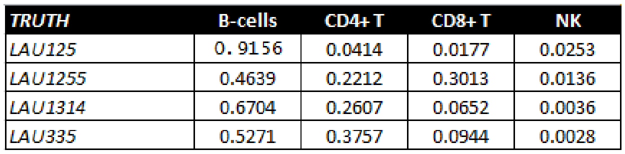
The EPIC model was fit to these TPM normalized RNA-seq data using its TRef reference matrix and all default options.

ICeD-T:

The ICeD-T model is fit using all 4 combinations of the following options: (1) Use Tumor Purity: Yes or no? (2) Use maximal variance weights: Yes or No?. For the purposes of this analysis, tumor purity is obtained from the flow cytometry results by combining the proportions of cancer cells and other cells.

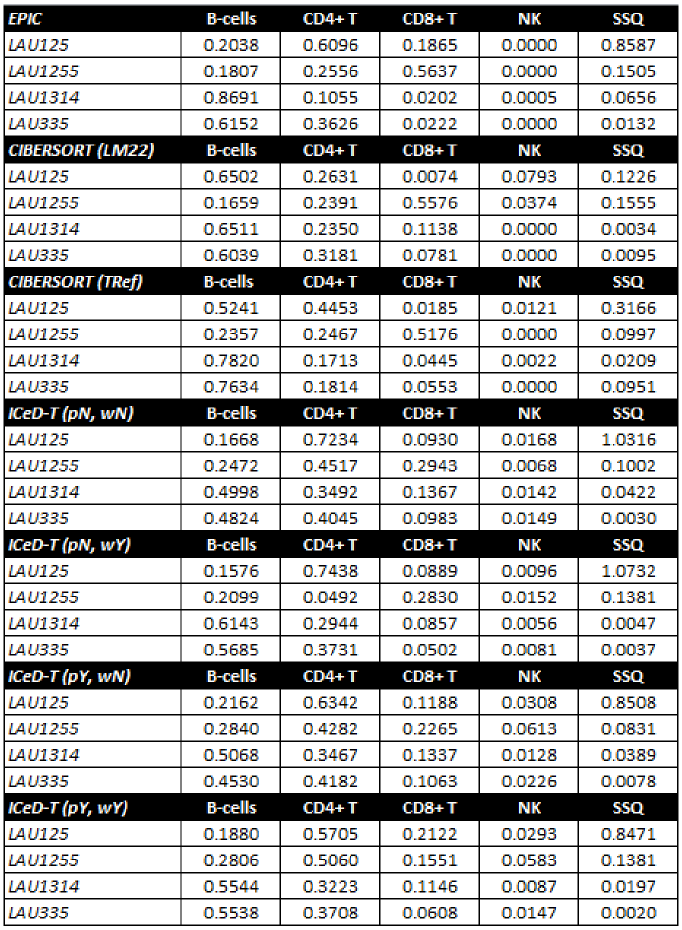
B.4.3 Fit Results

For the results shown below, all immune content is corrected for cell type size and renormalized so that proportions are computed with respect to the immune cells in the mixture (B-cells, CD4+ T-cells, CD8+ T-cells, and Natural Killers).



TRUTH	B-cells	CD4+ T	CD8+ T	NK
LAU125	0. 9156	0.0414	0.0177	0.0253
LAU1255	0.4639	0.2212	0.3013	0.0136
LAU1314	0.6704	0.2607	0.0652	0.0036
LAU335	0.5271	0.3757	0.0944	0.0028

Table B.9: Melanoma Data - True relative proportions of Immune cells



EPIC	B-cells	CD4+ T	CD8+ T	NK	SSQ
LAU125	0.2038	0.6096	0.1865	0.0000	0.8587
LAU1255	0.1807	0.2556	0.5637	0.0000	0.1505
LAU1314	0.8691	0.1055	0.0202	0.0005	0.0656
LAU335	0.6152	0.3626	0.0222	0.0000	0.0132
CIBERSORT (LM22)	B-cells	CD4+ T	CD8+ T	NK	SSQ
LAU125	0.6502	0.2631	0.0074	0.0793	0.1226
LAU1255	0.1659	0.2391	0.5576	0.0374	0.1555
LAU1314	0.6511	0.2350	0.1138	0.0000	0.0034
LAU335	0.6039	0.3181	0.0781	0.0000	0.0095
CIBERSORT (TRef)	B-cells	CD4+ T	CD8+ T	NK	SSQ
LAU125	0.5241	0.4453	0.0185	0.0121	0.3166
LAU1255	0.2357	0.2467	0.5176	0.0000	0.0997
LAU1314	0.7820	0.1713	0.0445	0.0022	0.0209
LAU335	0.7634	0.1814	0.0553	0.0000	0.0951
ICeD-T (pN, wN)	B-cells	CD4+ T	CD8+ T	NK	SSQ
LAU125	0.1668	0.7234	0.0930	0.0168	1.0316
LAU1255	0.2472	0.4517	0.2943	0.0068	0.1002
LAU1314	0.4998	0.3492	0.1367	0.0142	0.0422
LAU335	0.4824	0.4045	0.0983	0.0149	0.0030
ICeD-T (pN, wY)	B-cells	CD4+ T	CD8+ T	NK	SSQ
LAU125	0.1576	0.7438	0.0889	0.0096	1.0732
LAU1255	0.2099	0.0492	0.2830	0.0152	0.1381
LAU1314	0.6143	0.2944	0.0857	0.0056	0.0047
LAU335	0.5685	0.3731	0.0502	0.0081	0.0037
ICeD-T (pY, wN)	B-cells	CD4+ T	CD8+ T	NK	SSQ
LAU125	0.2162	0.6342	0.1188	0.0308	0.8508
LAU1255	0.2840	0.4282	0.2265	0.0613	0.0831
LAU1314	0.5068	0.3467	0.1337	0.0128	0.0389
LAU335	0.4530	0.4182	0.1063	0.0226	0.0078
ICeD-T (pY, wY)	B-cells	CD4+ T	CD8+ T	NK	SSQ
LAU125	0.1880	0.5705	0.2122	0.0293	0.8471
LAU1255	0.2806	0.5060	0.1551	0.0583	0.1381
LAU1314	0.5544	0.3223	0.1146	0.0087	0.0197
LAU335	0.5538	0.3708	0.0608	0.0147	0.0020

Table B.10: Melanoma Data - True relative proportions of Immune cells

It is clear from the above that CIBERSORT would produce the minimum sum of squared error among all model fits due in chief to the manner in which it handles subject LAU125. ICeD-T with use of Tumor information (both with weights and maximal variance weights), produced the second best fit by sum of squared error. EPIC would produce the third best fit by sum of squares. Finally, ICeD-T without tumor purity would produce the worst results.

Examining subject LAU125, this subject is highly anomalous. This subjects immune response in this sample is composed almost entirely of B-cells. Both EPIC and ICeD-T struggle to estimate the B-cell proportions for this subject - a likely consequence of their use of the same reference matrix. CIBERSORT does not struggle as greatly with this single subject and thus experiences smaller sums of squared error.

Across the remaining individuals, ICeD-T using any option produces the best results for LAU1255 and LAU335. ICeD-T without tumor purity and using maximal variance weights produces the best results for LAU1255, LAU1314 and LAU 335. Thus, outside of the strange subject LAU125, ICeD-T is able to provide the most competitive results across remaining subjects.

Focus now on the estimation of CD8 T-cell abundance across all methods. The use of ICeD-T without Tumor purity provides the best fit across the singular cell type among all individuals.

B.5 PD-1 Checkpoint Therapy Use in Melanomas

The final validation datasets for the ICeD-T method examine its application to a set of RNA-seq experiments derived from patients on PD-1 Checkpoint inhibitor therapies [71].

B.5.1 Data

The raw fastq files of RNA-seq data were downloaded from Sequence Read Archive (`https://www.ncbi.nlm.nih.gov/sra`), under the accession numbers SRP067938 and SRP090294. We mapped the RNA-seq reads to hg38 reference genome using STAR with gene annotation from GENCODE version 27. Then the number of RNA-seq fragments per gene were counted using R function `GenomicAlignments/summarizeOverlaps`.

B.5.2 Fit Method

CIBERSORT:

The CIBERSORT web application (Version: CIBERSORT Jar 1.06) was used to fit these TPM normalized RNA-seq data. The model was fit using the LM22 signature matrix and run with quantile normalization disabled.

EPIC:

The EPIC model is fit to the TRef reference matrix using TPM-normalized RNA-seq data.

ICeD-T:

The ICeD-T model is fit to the TRef reference matrix using TPM-normalized RNA-seq data. It is fit both without weights and with maximal variance weights derived from the TRef reference data. This is made possible through a function, `EPIC.Extract`, which extracts the fitted data and reference matrix from the EPIC library's function and outputs them in a form usable by ICeD-T.

TPM-normalization:

As noted above, data were provided in gene count form. As such, computation of TPM values using software such as RSEM is not possible. Thus, we transform the raw counts into TPM values using the following formula:

$$TPM_j = 10^6 \left(\frac{r_j/l_j}{\sum_{j=1}^{n_G} r_j/l_j} \right).$$

APPENDIX C: SUPPLEMENT FOR CHAPTER 5

C.1 Supplementary Methods

C.1.1 Notation Table

The following table contains the notation used to develop the TReC and TReCASE models for an arbitrary gene a candidate eQTL of this gene. Subscripts specifying the gene and eQTL are suppressed. The A allele and B allele are defined based on the genotype of the candidate eQTL.

<table>
<tr><th colspan="3">TReC + ASE Quantities</th></tr>
<tr><th>Value</th><th>Dimension</th><th>Description</th></tr>
<tr><td>$G(i)$</td><td>NA</td><td>The genotype of subject i at the specified eQTL. Can take values:
AA – homozygous for A allele
AB – heterozygous
BB – homozygous for B allele</td></tr>
<tr><td>ρ_i</td><td>1×1</td><td>Estimate of the tumor purity for the tumor sample of subject i, defined as the proportion of cells that are tumor cells.</td></tr>
<tr><th colspan="3">TReC Only Quantities</th></tr>
<tr><th>Value</th><th>Dimension</th><th>Description</th></tr>
<tr><td>Y_i</td><td>1×1</td><td>Total read count at the given gene in the tumor sample of subject i.</td></tr>
<tr><td>μ_{iA}</td><td>1×1</td><td>The mean TReC for subject i at A allele.</td></tr>
<tr><td>μ_{iB}</td><td>1×1</td><td>The mean TReC for subject i at B allele.</td></tr>
<tr><td>μ_i</td><td>1×1</td><td>The mean TReC for subject i.</td></tr>
<tr><td>ϕ</td><td>1×1</td><td>The overdispersion parameter for the distribution of TReC.</td></tr>
<tr><td>$\mathbf{x}_i$</td><td>$P \times 1$</td><td>Vector of covariate values for subject i</td></tr>
<tr><td>$\boldsymbol{\beta}$</td><td>$P \times 1$</td><td>Vector of covariate impacts on log total read count.</td></tr>
<tr><td>d_i</td><td>1×1</td><td>Read depth of RNA-Seq experiment for subject i.</td></tr>
<tr><th colspan="3">ASE Only Quantities</th></tr>
<tr><th>Value</th><th>Dimension</th><th>Description</th></tr>
<tr><td>R_i</td><td>1×1</td><td>The total number of allele specific reads for subject i.</td></tr>
<tr><td>R_{iB}</td><td>1×1</td><td>The number of allele specific reads mapped to the B allele for subject i.</td></tr>
<tr><td>ψ</td><td>1×1</td><td>The overdispersion parameter for the distribution of the ASE.</td></tr>
<tr><th colspan="3">eQTL Parameters</th></tr>
<tr><th>Value</th><th>Dimension</th><th>Description</th></tr>
<tr><td>η</td><td>1×1</td><td>The eQTL effect in normal tissue: $\mu_{iB}^{(N)}/\mu_{iA}^{(N)}$.</td></tr>
<tr><td>γ</td><td>1×1</td><td>The eQTL effect in tumor tissue: $\mu_{iB}^{(T)}/\mu_{iA}^{(T)}$.</td></tr>
<tr><td>κ</td><td>1×1</td><td>An over-expression effect in the tumor for A allele: $\mu_{iA}^{(T)}/\mu_{iA}^{(N)}$.</td></tr>
<tr><td>ξ_i</td><td>1×1</td><td>The ratio of gene expression of B allele versus A allele for subject i, defined as μ_{iB}/μ_{iA}.</td></tr>
</table>

Table C.1: Notation for defining the TReC and TReCASE models.

C.1.2 Optimization Algorithm

As mentioned in main text, the optimization routine for solving the TReC and TReCASE models uses a coordinate block ascent routine with the following steps.

(0) Select initial estimates for κ, η, and γ.

(1) Holding κ, η, γ, and ψ constant, use negative binomial regression to update $\boldsymbol{\beta}$ and ϕ.

(2) Holding $\boldsymbol{\beta}, \phi$ and ψ constant, use a Quasi-Newton method (LBFGS) to update κ, η, and γ.

(3) Holding $\boldsymbol{\beta}, \phi, \kappa, \eta$, and γ constant, update ψ using a Quasi-Newton method (LBFGS).

(4) Iterate steps (1)-(3) until convergence

The algorithm above is specified for the TReCASE model. A similar algorithm is used for TReC model except that we need to remove step (3) and iterate steps (1) and (2) repeatedly (while removing ψ from estimation procedures) until convergence.

To fully define the algorithm above, a discussion of Step (0) is warranted. Under the null hypothesis $\eta = 1$, model fit proceeds following the above algorithm starting at position $\kappa = 1$ and $\gamma = 1$ and holding η fixed at 1. Under the null hypothesis $\gamma = 1$, model fit proceeds as above, starting at position $\kappa = 1$ and $\eta = 1$ and holding γ at 1 throughout. To fit the full model, we choose initial values for κ, η, and γ in accordance with the fit of the null hypothesis, either $\eta = 1$ or $\gamma = 1$, which gives larger likelihood value at its MLE. This initialization method ensures that the suggested likelihood ratio tests are well-defined by avoiding situations where the likelihood of full model is less than the likelihood of a restricted model.

C.1.3 Mathematical Details for Optimization

Mathematical details for section (A.2) are presented in the following. Note that, as defined, κ, η and γ are strictly positive parameters. Thus, we estimate log(η), log(γ), and log(κ) in the optimization process to guarantee that κ, η and γ are all positive, and avoid constrained optimization when working directly with κ, η and γ.

C.1.3.1 Total Read Count (TReC) Model Component

To motivate the structure of the TReC model, consider the ratio of the mean expressions for alleles B versus allele A for subject i. Assume that the expression of each allele is a weighted sum of its expression in normal and tumor tissues, weighted by the proportional composition of the sample with respect to each type. One can then specify this ratio for subject i as:

$$\begin{aligned}
\xi_i = \frac{\mu_{iB}}{\mu_{iA}} &= \frac{(1-\rho_i)\mu_{iB}^{(N)} + \rho_i\mu_{iB}^{(T)}}{(1-\rho_i)\mu_{iA}^{(N)} + \rho_i\mu_{iA}^{(T)}} \\
&= \frac{(1-\rho_i)\left(\mu_{iB}^{(N)}/\mu_{iA}^{(N)}\right) + \rho_i\left(\mu_{iB}^{(T)}/\mu_{iA}^{(T)}\right)\left(\mu_{iA}^{(T)}/\mu_{iA}^{(N)}\right)}{1-\rho_i+\rho_i\left(\mu_{iA}^{(T)}/\mu_{iA}^{(N)}\right)} \\
&= \frac{(1-\rho_i)\eta + \rho_i\kappa\gamma}{1-\rho_i+\rho_i\kappa} = (1-c_i)\eta + c_i\gamma,
\end{aligned}$$

where $c_i = (\rho_i\kappa)/(1-\rho_i+\rho_i\kappa)$. Assuming now that the total expression for subject i is the sum of the expressions from each constituent allele and modelling $\mu_{i,AA}^{(N)} = \exp(x_i^T\boldsymbol{\beta})$, the above implies that our mean takes the following form:

$$\mu_i = \begin{cases}
e^{x_i^T\boldsymbol{\beta}}(1-\rho_i+\rho_i\kappa), & \text{if } G(i) = AA \\
e^{x_i^T\boldsymbol{\beta}}(1-\rho_i+\rho_i\kappa)(1+\xi_i)/2, & \text{if } G(i) = AB \\
e^{x_i^T\boldsymbol{\beta}}(1-\rho_i+\rho_i\kappa)\xi_i, & \text{if } G(i) = BB
\end{cases}$$

Under a negative binomial distribution, the likelihood component for the TReC model for a single subject is given by:

$$f(Y_i;\mu_i,\phi) = \frac{\Gamma(Y_i + 1/\phi)}{Y_i!\Gamma(1/\phi)}\left(\frac{1}{1+\phi\mu_i}\right)^{1/\phi}\left(\frac{\phi\mu_i}{1+\phi\mu_i}\right)^{Y_i}.$$

Thus, the log-likelihood for this component takes the form:

$$\begin{aligned}\ell_{TReC} &= \sum_{i=1}^{N} \ell_{TReC}^{(i)} \\ &= \sum_{i=1}^{N}\left[\ln\left\{\frac{\Gamma(y_i+1/\phi)}{y_i!\Gamma(1/\phi)}\right\} - (1/\phi + y_i)\log(1+\phi\mu_i) + y_i\log(\phi) + y_i\log(\mu_i)\right].\end{aligned}$$

Letting λ denote one of κ, η, or γ, we have:

$$\frac{\partial \ell_{TReC}}{\partial \log(\lambda)} = \sum_{i=1}^{N}\left(\frac{\partial \ell_{TReC}^{(i)}}{\partial \mu_i}\right)\left(\frac{\partial \mu_i}{\partial \lambda}\right)\left(\frac{\partial \lambda}{\partial \log(\lambda)}\right).$$

We derive each of these components in turn. First, consider $\partial\ell_{TReC}^{(i)}/\partial\mu_i$:

$$\frac{\partial \ell_{TReC}^{(i)}}{\partial \mu_i} = \frac{y_i}{\mu_i} - \frac{1+\phi y_i}{1+\phi\mu_i}.$$

For utility in later steps, lets consider derivatives of the form $\frac{\partial \xi_i}{\partial \lambda}$ and $\frac{\partial c_i}{\partial \kappa}$:

$$\frac{\partial \xi_i}{\partial \kappa} = (\gamma - \eta)\left(\frac{\partial c_i}{\partial \kappa}\right), \quad \frac{\partial c_i}{\partial \kappa} = \kappa^{-1}c_i(1-c_i), \quad \frac{\partial \xi_i}{\partial \gamma} = c_i, \text{ and } \frac{\partial \xi_i}{\partial \eta} = 1 - c_i.$$

Next, consider $\partial\mu_i/\partial\lambda$. It is easiest to consider this component separately for each genotype. For $G(i) = AA$, μ_i is dependent on κ, but free of η and γ. Thus:

$$\frac{\partial \mu_i}{\partial \kappa} = e^{x_i^T\boldsymbol{\beta}}\rho_i, \text{ and } \frac{\partial \mu_i}{\partial \eta} = \frac{\partial \mu_i}{\partial \gamma} = 0.$$

For $G(i) = AB$, we have:

$$\begin{aligned}
\frac{\partial \mu_i}{\partial \kappa} &= e^{x_i^T \beta}\left[\rho_i \left(\frac{1+\xi_i}{2}\right) + (1-\rho_i+\rho_i\kappa)(1/2)\left(\frac{\partial \xi_i}{\partial \kappa}\right)\right] = e^{x_i^T\beta}(\rho_i/2)(1+\gamma),\\
\frac{\partial \mu_i}{\partial \eta} &= e^{x_i^T \beta}(1-\rho_i+\rho_i\kappa)(1/2)\left(\frac{\partial \xi_i}{\partial \eta}\right) = e^{x_i^T\beta}(1/2)(1-\rho_i),\\
\frac{\partial \mu_i}{\partial \gamma} &= e^{x_i^T \beta}(1-\rho_i+\rho_i\kappa)(1/2)\left(\frac{\partial \xi_i}{\partial \gamma}\right) = e^{x_i^T\beta}(\rho_i/2)\kappa.
\end{aligned}$$

Finally, for $G(i) = BB$, we have:

$$\begin{aligned}
\frac{\partial \mu_i}{\partial \kappa} &= e^{x_i^T \beta}\left[\rho_i \xi_i + (1-\rho_i+\rho_i\kappa)\left(\frac{\partial \xi_i}{\partial \kappa}\right)\right] = e^{x_i^T\beta}\rho_i\gamma,\\
\frac{\partial \mu_i}{\partial \eta} &= e^{x_i^T \beta}(1-\rho_i+\rho_i\kappa)\left(\frac{\partial \xi_i}{\partial \eta}\right) = e^{x_i^T\beta}(1-\rho_i),\\
\frac{\partial \mu_i}{\partial \gamma} &= e^{x_i^T \beta}(1-\rho_i+\rho_i\kappa)\left(\frac{\partial \xi_i}{\partial \eta}\right) = e^{x_i^T\beta}\rho_i\kappa.
\end{aligned}$$

While not used for the C++ implementation of the model, the R-version uses the Hessian matrix with respect to the κ, η, and γ variables. We derive it here for completeness. Let $\dot{\ell}_{TReC} = \frac{\partial \ell_{TReC}}{\partial log(\lambda)}$ where λ is one of κ, η, and γ. As specified above:

$$\dot{\ell}_{TReC} = \lambda \sum_{i=1}^{N} \left\{ \left(\frac{\partial \ell_{TReC}^{(i)}}{\partial \mu_i}\right)\left(\frac{\partial \mu_i}{\partial \lambda}\right)\right\}.$$

Then:

$$
\begin{aligned}
\frac{\partial^2 \dot{\ell}_{TReC}}{\partial \log(\lambda)^2} &= \left(\frac{\partial \dot{\ell}_{TReC,\kappa}}{\partial \kappa}\right)\left(\frac{\partial \kappa}{\partial \log(\kappa)}\right) = \kappa \left(\frac{\partial \dot{\ell}_{TReC,\kappa}}{\partial \kappa}\right) \\
&= \dot{\ell}_{TReC,\kappa} + \kappa^2 \sum_{i=1}^{N} \left(\frac{\partial^2 \ell^{(i)}_{TReC}}{\partial \mu_i \partial \kappa}\right)\left(\frac{\partial \mu_i}{\partial \kappa}\right) + \kappa^2 \sum_{i=1}^{N} \left(\frac{\partial \ell^{(i)}_{TReC}}{\partial \mu_i}\right)\left(\frac{\partial^2 \mu_i}{\partial \kappa^2}\right) \\
&= \dot{\ell}_{TReC,\kappa} + \kappa^2 \sum_{i=1}^{N} \left(\frac{\partial^2 \ell^{(i)}_{TReC}}{\partial \mu_i^2}\right)\left(\frac{\partial \mu_i}{\partial \kappa}\right)^2 + \kappa^2 \sum_{i=1}^{N} \left(\frac{\partial \ell^{(i)}_{TReC}}{\partial \mu_i}\right)\left(\frac{\partial^2 \mu_i}{\partial \kappa^2}\right) \\
&= \dot{\ell}_{TReC,\kappa} + \kappa^2 \sum_{i=1}^{N} \left(\frac{\partial^2 \ell^{(i)}_{TReC}}{\partial \mu_i^2}\right)\left(\frac{\partial \mu_i}{\partial \kappa}\right)^2 .
\end{aligned}
$$

The last equality holds since $\frac{\partial^2 \mu_i}{\partial \kappa^2} = 0$ and we may plug in:

$$
\frac{\partial^2 \ell^{(i)}_{TReC}}{\partial \mu_i^2} = -\left(\frac{y_i}{\mu_i^2}\right) + \frac{\phi + \phi^2 y_i}{(1+\phi\mu_i)^2}.
$$

Similar results hold for η and γ and are given below:

$$
\begin{aligned}
\frac{\partial^2 \ell_{TReC}}{\partial \log(\eta)^2} &= \dot{\ell}_{TReC,\eta} + \eta^2 \sum_{i=1}^{N} \left(\frac{\partial^2 \ell^{(i)}_{TReC}}{\partial \mu_i^2}\right)\left(\frac{\partial \mu_i}{\partial \eta}\right)^2, \\
\frac{\partial^2 \ell_{TReC}}{\partial \log(\gamma)^2} &= \dot{\ell}_{TReC,\gamma} + \gamma^2 \sum_{i=1}^{N} \left(\frac{\partial^2 \ell^{(i)}_{TReC}}{\partial \mu_i^2}\right)\left(\frac{\partial \mu_i}{\partial \gamma}\right)^2 .
\end{aligned}
$$

To complete the Hessian, we compute the remaining results:

$$
\begin{aligned}
\frac{\partial^2 \ell_{TReC}}{\partial \log(\kappa) \partial \log(\eta)} &= \eta\kappa \sum_{i=1}^{N} \left(\frac{\partial^2 \ell^{(i)}_{TReC}}{\partial \mu_i^2}\right)\left(\frac{\partial \mu_i}{\partial \kappa}\right)\left(\frac{\partial \mu_i}{\partial \eta}\right), \\
\frac{\partial^2 \ell_{TReC}}{\partial \log(\kappa) \partial \log(\gamma)} &= \gamma\kappa \sum_{i=1}^{N} \left[\left(\frac{\partial^2 \ell^{(i)}_{TReC}}{\partial \mu_i^2}\right)\left(\frac{\partial \mu_i}{\partial \kappa}\right)\left(\frac{\partial \mu_i}{\partial \gamma}\right) + \left(\frac{\partial \ell^{(i)}_{TReC}}{\partial \mu_i}\right)\left(\frac{\partial^2 \mu_i}{\partial \kappa \partial \gamma}\right)\right], \\
\frac{\partial^2 \ell_{TReC}}{\partial \log(\eta) \partial \log(\gamma)} &= \eta\gamma \sum_{i=1}^{N} \left(\frac{\partial^2 \ell^{(i)}_{TReC}}{\partial \mu_i^2}\right)\left(\frac{\partial \mu_i}{\partial \eta}\right)\left(\frac{\partial \mu_i}{\partial \gamma}\right),
\end{aligned}
$$

where $\frac{\partial^2 \mu_i}{\partial\kappa\partial\eta} = \frac{\partial^2 \mu_i}{\partial\eta\partial\gamma} = 0$ and

$$\frac{\partial^2 \mu_i}{\partial\kappa\partial\gamma} = \begin{cases} 0, & \text{if } G(i) = AA \\ (1/2)e^{x_i^T\beta}\rho_i, & \text{if } G(i) = AB \\ e^{x_i^T\beta}\rho_i, & \text{if } G(i) = BB \end{cases}$$

C.1.3.2 Allele Specific Expression (ASE) Model Component

In the following, let μ_{i1} represent the number of reads that are expressed by allele 1 on average for subject i and μ_{i2} be its counterpoint for allele 2. Within a sample prepped for RNA-seq, the pool of reads for the given gene contains $\mu_{i1} + \mu_{i2}$ reads. The proportion of reads belonging to allele 1 on average is then given by:

$$\pi_i = \frac{\mu_{i1}}{\mu_{i1} + \mu_{i2}} = \frac{(\mu_{i1}/\mu_{i2})}{1 + \mu_{i1}/\mu_{i2}}.$$

Thus, viewing the RNA-Seq sampling procedure as drawing a group of reads at random and allowing for extra-binomial variation, we can model the data-generation mechanism via a beta-binomial distribution. Extra-binomial variation is often observed in genetic studies and in the case of ASE reads can in part be attributed to incorrectly genotyped alleles resulting from genotyping or imputation error.

In order to model a consistent eQTL effect within the TReC and ASE components of the model, define allele 1 as that containing the minor allele B for heterozygous subjects. In homozygous subjects, an arbitrary allele is selected as the expression between the two alleles is assumed to be equal on average. Thus, by the statement above and previous definitions, we may model the average reads for allele 1 as:

$$\pi_i = \begin{cases} \xi_i/(1+\xi_i), & \text{if } G(i) = BB \\ (1/2), & \text{otherwise} \end{cases}$$

Thus, the likelihood for the ASE component of the model is given by:

$$f(r_{iB}; r_i, \pi_i, \psi) = \binom{r_i}{r_{iB}} \left[\frac{\Gamma\left(\psi^{-1}\right)}{\Gamma\left(\psi^{-1}\pi_i\right)\Gamma\left(\psi^{-1}(1-\pi_i)\right)}\right] \times \\ \left[\frac{\Gamma\left(\psi^{-1}\pi_i + r_{iB}\right)\Gamma\left(\psi^{-1}(1-\pi_i) + r_i - r_{iB}\right)}{\Gamma\left(\psi^{-1} + r_i\right)}\right].$$

Define $\ell_{ASE}^{(i)}$ be the ASE likelihood from the $i-th$ sample. Then:

$$\ell_{ASE} = \sum_{i=1}^{n} \ell_{ASE}^{(i)} = \sum_{i=1}^{n} \log\left[f(r_{iB}; r_i, \pi_i, \psi)\right].$$

It can be seen that the gradient functions for π_i and ψ are given by:

$$\begin{aligned} \frac{\partial \ell_{ASE}^{(i)}}{\partial \pi_i} &= \psi^{-1}\Bigg[\Psi_0\left(\psi^{-1}\pi_i + r_{iB}\right) - \Psi_0\left(\psi^{-1}(1-\pi_i) + r_i - r_{iB}\right) - \Psi_0\left(\psi^{-1}\pi_i\right) + \\ &\qquad \Psi_0\left(\psi^{-1}(1-\pi_i)\right)\Bigg], \\ \frac{\partial \ell_{ASE}}{\partial \psi} &= \sum_{i=1}^{n} -\psi^{-2}\pi_i\left[\Psi_0\left(\psi^{-1}\pi_i + r_{iB}\right) - \Psi_0\left(\psi^{-1}\pi_i\right)\right] - \\ &\qquad \sum_{i=1}^{n} \psi^{-2}(1-\pi_i)\left[\Psi_0\left(\psi^{-1}(1-\pi_i) + r_i - r_{iB}\right) - \Psi_0\left(\psi^{-1}(1-\pi_i)\right)\right] - \\ &\qquad \sum_{i=1}^{n} \psi^{-2}\left[\Psi_0\left(\psi^{-1}\right) - \Psi_0\left(\psi^{-1} + r_i\right)\right]. \end{aligned}$$

Before deriving the remaining components necessary for the gradient, we note that only individuals of heterozygous genotype contribute to the gradient of κ, η and γ, whereas all individuals contributed to the gradient of ψ. Thus, we have:

$$\begin{aligned}\frac{\partial \ell_{ASE}}{\partial \log(\lambda)} \equiv \dot{\ell}_{ASE,\lambda} &= \sum_{i;G(i)=AB} \left(\frac{\partial \ell_{ASE}^{(i)}}{\partial \pi_i}\right) \left(\frac{\partial \pi_i}{\partial \xi_i}\right) \left(\frac{\partial \xi_i}{\partial \lambda}\right) \left(\frac{\partial \lambda}{\partial \log(\lambda)}\right) \\ &= \lambda \sum_{i;G(i)=AB} \left(\frac{\partial \ell_{ASE}^{(i)}}{\partial \pi_i}\right) \left(\frac{\partial \pi_i}{\partial \xi_i}\right) \left(\frac{\partial \xi_i}{\partial \lambda}\right).\end{aligned}$$

To calculate the above quantity, we need:

$$\frac{\partial \pi_i}{\partial \xi_i} = (1+\xi_i)^{-2}, \quad \frac{\partial \xi_i}{\partial \eta} = 1 - c_i, \quad \frac{\partial \xi_i}{\partial \gamma} = c_i, \quad \frac{\partial \xi_i}{\partial \kappa} = (\gamma - \eta) c_i (1 - c_i) \kappa^{-1}.$$

As noted in the previous section, the C++ fit routine does not utilize the Hessian but we provide its derivation here for completeness. We will make repeated use of the following terms, so they are presented here for later reference.

$$\begin{aligned}\frac{\partial^2 \ell_{ASE}^{(i)}}{\partial \pi_i^2} &= \psi^{-2} \Bigg[\Psi_1\left(\psi^{-1}\pi_i + r_{iB}\right) + \Psi_1\left(\psi^{-1}(1-\pi_i) + r_i - r_{iB}\right) - \Psi_1\left(\psi^{-1}\pi_i\right) - \\ &\qquad \Psi_1\left(\psi^{-1}(1-\pi_i)\right) \Bigg] \\ \frac{\partial^2 \pi_i}{\partial \xi_i^2} &= -2(1+\xi_i)^{-3} \\ \frac{\partial^2 \xi_i}{\partial \kappa^2} &= (\gamma - \eta) \left[\kappa^{-1}(1-2c_i)\left(\frac{\partial c_i}{\partial \kappa}\right) - \kappa^{-2} c_i (1-c_i) \right],\end{aligned}$$

where

$$\Psi_0(x) = \frac{\partial \ln\Gamma(x)}{\partial x} \text{ and } \Psi_1(x) = \frac{\partial^2 \ln\Gamma(x)}{\partial x^2}.$$

We complete the derivation in the following.

$$\begin{aligned}
\frac{\partial^2 \ell_{ASE}}{\partial \log(\kappa)^2} &= \left(\frac{\partial \dot{\ell}_{ASE,\kappa}}{\partial \kappa}\right)\left(\frac{\partial \kappa}{\partial \log(\kappa)}\right) = \kappa\left(\frac{\partial \dot{\ell}_{ASE,\kappa}}{\partial \kappa}\right) \\
&= \dot{\ell}_{ASE,\kappa} + \kappa^2 \times \\
&\quad \sum_{i;G(i)=AB}\left\{\left(\frac{\partial^2 \ell^{(i)}_{ASE}}{\partial \pi_i^2}\right)\left(\frac{\partial \pi_i}{\partial \xi_i}\right)^2\left(\frac{\partial \xi_i}{\partial \kappa}\right)^2 + \left(\frac{\partial \ell^{(i)}_{ASE}}{\partial \pi_i}\right)\left(\frac{\partial^2 \pi_i}{\partial \xi_i^2}\right)\left(\frac{\partial \xi_i}{\partial \kappa}\right)^2 + \right. \\
&\qquad \left.\left(\frac{\partial^2 \ell^{(i)}_{ASE}}{\partial \pi_i^2}\right)\left(\frac{\partial \pi_i}{\partial \xi_i}\right)\left(\frac{\partial^2 \xi_i}{\partial \kappa^2}\right)\right\}.
\end{aligned}$$

Similarly for η and γ, we have:

$$\begin{aligned}
\frac{\partial^2 \ell_{ASE}}{\partial \log(\eta)^2} &= \dot{\ell}_{ASE,\eta} + \eta^2 \sum_{i;G(i)=AB}(1-c_i)^2\left[\left(\frac{\partial^2 \ell^{(i)}_{ASE}}{\partial \pi_i^2}\right)\left(\frac{\partial \pi_i}{\partial \xi_i}\right)^2 + \left(\frac{\partial \ell^{(i)}_{ASE}}{\partial \pi_i}\right)\left(\frac{\partial^2 \pi_i}{\partial \xi_i^2}\right)\right], \\
\frac{\partial^2 \ell_{ASE}}{\partial \log(\gamma)^2} &= \dot{\ell}_{ASE,\gamma} + \gamma^2 \sum_{i;G(i)=AB} c_i^2\left[\left(\frac{\partial^2 \ell^{(i)}_{ASE}}{\partial \pi_i^2}\right)\left(\frac{\partial \pi_i}{\partial \xi_i}\right)^2 + \left(\frac{\partial \ell^{(i)}_{ASE}}{\partial \pi_i}\right)\left(\frac{\partial^2 \pi_i}{\partial \xi_i^2}\right)\right].
\end{aligned}$$

Finally, for the "mixed" second derivatives, we have:

$$\begin{aligned}
\frac{\partial^2 \ell_{ASE}}{\partial \log(\eta)\partial \log(\kappa)} &= \kappa\eta \sum_{i;G(i)=AB}\left[\left\{\left(\frac{\partial^2 \ell^{(i)}_{ASE}}{\partial \pi_i^2}\right)\left(\frac{\partial \pi_i}{\partial \xi_i}\right)^2 - \left(\frac{\partial \ell^{(i)}_{ASE}}{\partial \pi_i}\right)\left(\frac{\partial^2 \pi_i}{\partial \xi_i^2}\right)\right\}\left(\frac{\partial \xi_i}{\partial \kappa}\right)(1-c_i) - \right. \\
&\qquad \left.\left(\frac{\partial \ell^{(i)}_{ASE}}{\partial \pi_i}\right)\left(\frac{\partial \pi_i}{\partial \xi_i}\right)\left(\frac{\partial c_i}{\partial \kappa}\right)\right], \\
\frac{\partial^2 \ell_{ASE}}{\partial \log(\gamma)\partial \log(\kappa)} &= \kappa\gamma \sum_{i;G(i)=AB}\left[\left\{\left(\frac{\partial^2 \ell^{(i)}_{ASE}}{\partial \pi_i^2}\right)\left(\frac{\partial \pi_i}{\partial \xi_i}\right)^2 + \left(\frac{\partial \ell^{(i)}_{ASE}}{\partial \pi_i}\right)\left(\frac{\partial^2 \pi_i}{\partial \xi_i^2}\right)\right\}\left(\frac{\partial \xi_i}{\partial \kappa}\right)c_i + \right. \\
&\qquad \left.\left(\frac{\partial \ell^{(i)}_{ASE}}{\partial \pi_i}\right)\left(\frac{\partial \pi_i}{\partial \xi_i}\right)\left(\frac{\partial c_i}{\partial \kappa}\right)\right], \\
\frac{\partial^2 \ell_{ASE}}{\partial \log(\gamma)\partial \log(\eta)} &= \gamma\eta \sum_{i;G(i)=AB} c_i(1-c_i)\left[\left(\frac{\partial^2 \ell^{(i)}_{ASE}}{\partial \pi_i^2}\right)\left(\frac{\partial \pi_i}{\partial \xi_i}\right)^2 + \left(\frac{\partial \ell^{(i)}_{ASE}}{\partial \pi_i}\right)\left(\frac{\partial^2 \pi_i}{\partial \xi_i^2}\right)\right].
\end{aligned}$$

C.1.4 Cis-Trans Score Test

Recall that eQTL come in two varieties: *cis-* and *trans-*eQTL. *cis-*eQTLs induce allelic imbalance of gene expression whereas *trans-*eQTLs affect the expression of two alleles to the same degree. [55] and [56] have developed and refined a "Cis-Trans test" to identify whether eQTL act in a *cis-* or *trans-* fashion. Under the null hypothesis (*cis-*), the eQTL effect sizes are the same between TReC and ASE models. A small p-value using this test leads to rejection of the null hypothesis, and thus the conclusion that the given Gene-SNP pair behave in a *trans*-eQTL manner. In that case, only the TReC data should be used for eQTL mapping.

To develop this test for eQTL mapping in tumor tissues, we follow [56] by extending the likelihood framework through the introduction of new parameters which allow eQTL effects to differ between TReC and ASE components. Specifically, we define:

$$\eta_{ASE} = \eta + \alpha_\eta, \text{ and } \gamma_{ASE} = \gamma + \alpha\gamma,$$

where η and γ are the TReC-specific eQTL effects in normal and tumor tissues, respectively; η_{ASE} and γ_{ASE} are the ASE-specific counterparts; α_η and α_γ are the discrepancies of eQTL effects between ASE and TReC components of the model in normal and tumor tissues, respectively. Then to test *cis-* versus *trans*-eQTL, we employ a score test for the two-dimensional hypothesis: $\alpha_\eta = \alpha_\gamma = 0$.

C.1.4.1 Structure of the Score Test

Define the following groups of parameters: $\epsilon = (\kappa, \eta, \gamma)^T$; $\alpha = (\alpha_\eta, \alpha_\gamma)^T$; and $\Theta = (\boldsymbol{\beta}^T, \epsilon^T, \alpha^T, \phi, \psi)$. Let $\ell = \ell_{TReC} + \ell_{ASE}$ be the full data log-likelihood, $\dot{\ell}$ be the first derivative of the log-likelihood with respect to the parameters, and $I(\Theta)$ be the Fisher's Information

Matrix. We may specify the Fisher's Information Matrix in the following way:

$$I(\Theta) = \begin{pmatrix} I_{\boldsymbol{\beta},\boldsymbol{\beta}} & I_{\boldsymbol{\beta},\epsilon} & I_{\boldsymbol{\beta},\phi} & I_{\boldsymbol{\beta},\psi} & I_{\boldsymbol{\beta},\alpha} \\ I_{\epsilon,\boldsymbol{\beta}} & I_{\epsilon,\epsilon} & I_{\epsilon,\phi} & I_{\epsilon,\psi} & I_{\epsilon,\alpha} \\ I_{\phi,\boldsymbol{\beta}} & I_{\phi,\epsilon} & I_{\phi,\phi} & I_{\phi,\psi} & I_{\phi,\alpha} \\ I_{\psi,\boldsymbol{\beta}} & I_{\psi,\epsilon} & I_{\psi,\phi} & I_{\psi,\psi} & I_{\psi,\alpha} \\ I_{\alpha,\boldsymbol{\beta}} & I_{\alpha,\epsilon} & I_{\alpha,\phi} & I_{\alpha,\psi} & I_{\alpha,\alpha} \end{pmatrix} = \begin{pmatrix} M_1 & M_2 \\ M_2^T & I_{\alpha,\alpha} \end{pmatrix},$$

where M_1 is the upper-left block of the Fisher's Information matrix through $I_{\psi,\psi}$ and M_2 is the remaining block excluding $I_{\alpha,\alpha}$.

Following the developments of [84], we may compute the score test of $\alpha_\eta = \alpha_\gamma = 0$ in the following way:

$$\begin{aligned} SC &= \dot{\ell}\left(\hat{\Theta}\right)^T I\left(\hat{\Theta}\right)^{-1} \dot{\ell}\left(\hat{\Theta}\right) \\ &= \begin{pmatrix} \frac{\partial \ell}{\partial \alpha_\eta} & \frac{\partial \ell}{\partial \alpha_\gamma} \end{pmatrix} \left(I_{\alpha,\alpha} - M_2^T M_1^{-1} M_2\right)^{-1} \begin{pmatrix} \frac{\partial \ell}{\partial \alpha_\eta} \\ \frac{\partial \ell}{\partial \alpha_\gamma} \end{pmatrix} \Bigg|_{\Theta=\hat{\Theta}}, \end{aligned}$$

where $\hat{\theta}$ is the estimate of our parameters under the null. SC is asymptotically chi-squared with two degrees of freedom under the null.

C.1.4.2 TReC Derivatives

Preceding development of the gradients and Hessians of the TReC components in the following section, it will be helpful to compose a list of definitions and useful derivatives for later use. Recall that μ_i is the mean read count in the TReC component of the model, given by:

$$\mu_i = \begin{cases} e^{x_i^T \boldsymbol{\beta}} \left[1 - \rho_i + \rho_i \kappa\right], & \text{if } G(i) = AA, \\ e^{x_i^T \boldsymbol{\beta}} \left[1 - \rho_i + \rho_i \kappa\right] \left[\frac{1+\xi_i}{2}\right], & \text{if } G(i) = AB, \\ e^{x_i^T \boldsymbol{\beta}} \left[1 - \rho_i + \rho_i \kappa\right] \xi_i, & \text{if } G(i) = BB, \end{cases}$$

where $\xi_i = (1 - c_i)\eta + c_i\gamma$ and $c_i = (\rho_i\kappa)/(1 - \rho_i + \rho_i\kappa)$. It is clear that:

$$\begin{aligned} \frac{\partial c_i}{\partial \kappa} &= \kappa^{-1} c_i (1 - c_i) \\ \frac{\partial^2 c_i}{\partial \kappa^2} &= \kappa^{-1}(1 - 2c_i)\left(\frac{\partial c_i}{\partial \kappa}\right) - \kappa^{-2} c_i (1 - c_i) \end{aligned}$$

This allows us to compose the following derivatives for ξ_i:

$$\begin{aligned} \frac{\partial \xi_i}{\partial \kappa} &= (\gamma - \eta)\left(\frac{\partial c_i}{\partial \kappa}\right) \\ \frac{\partial \xi_i}{\partial \eta} &= (1 - c_i) \\ \frac{\partial \xi_i}{\partial \gamma} &= c_i \end{aligned}$$

The Hessian for ξ_i is provided by the following

$$\frac{\partial \xi_i}{\partial \epsilon \partial \epsilon^T} = \begin{pmatrix} (\gamma - \eta)\left(\frac{\partial^2 c_i}{\partial \kappa^2}\right) & -\frac{\partial c_i}{\partial \kappa} & \frac{\partial c_i}{\partial \kappa} \\ & 0 & 0 \\ & & 0 \end{pmatrix}$$

The gradient of μ_i with respect to ϵ is provided below:

$$\left.\frac{\partial \mu_i}{\partial \epsilon}\right|_{G(i)=AA} = e^{x_i^T \boldsymbol{\beta}} \begin{pmatrix} \rho_i \\ 0 \\ 0 \end{pmatrix}$$

$$\left.\frac{\partial \mu_i}{\partial \epsilon}\right|_{G(i)=AB} = e^{x_i^T \boldsymbol{\beta}} \begin{pmatrix} \left[\rho_i \left(\frac{1+\xi_i}{2}\right) + (1-\rho_i+\rho_i\kappa)(1/2)\left(\frac{\partial \xi_i}{\partial \kappa}\right)\right] \\ \left[(1-\rho_i+\rho_i\kappa)(1/2)\left(\frac{\partial \xi_i}{\partial \eta}\right)\right] \\ \left[(1-\rho_i+\rho_i\kappa)(1/2)\left(\frac{\partial \xi_i}{\partial \gamma}\right)\right] \end{pmatrix}$$

$$\left.\frac{\partial \mu_i}{\partial \epsilon}\right|_{G(i)=BB} = e^{x_i^T \boldsymbol{\beta}} \begin{pmatrix} \left[\rho_i \xi_i + (1-\rho_i+\rho_i\kappa)\left(\frac{\partial \xi_i}{\partial \kappa}\right)\right] \\ \left[(1-\rho_i+\rho_i\kappa)\left(\frac{\partial \xi_i}{\partial \eta}\right)\right] \\ \left[(1-\rho_i+\rho_i\kappa)\left(\frac{\partial \xi_i}{\partial \gamma}\right)\right] \end{pmatrix}$$

The Hessian for μ_i is identically 0 for genotype AA. However, for genotypes AB and BB, we have the following where we define $\delta_i = 1 - \rho_i + \rho_i\kappa$.

$$\frac{\partial^2 \mu_i}{\partial \epsilon \partial \epsilon^T} = e^{x_i^T \boldsymbol{\beta}} \begin{pmatrix} \left[\rho_i \left(\frac{\partial \xi_i}{\partial \kappa}\right) + (1/2)\delta_i \left(\frac{\partial^2 \xi_i}{\partial \kappa^2}\right)\right] & (1/2)\left[\rho_i \left(\frac{\partial \xi_i}{\partial \eta}\right) + \delta_i \left(\frac{\partial^2 \xi_i}{\partial \kappa \partial \eta}\right)\right] & (1/2)\left[\rho_i \left(\frac{\partial \xi_i}{\partial \gamma}\right) + \delta_i \left(\frac{\partial^2 \xi_i}{\partial \kappa \partial \gamma}\right)\right] \\ (1/2)\left[\rho_i \left(\frac{\partial \xi_i}{\partial \eta}\right) + \delta_i \left(\frac{\partial^2 \xi_i}{\partial \kappa \partial \eta}\right)\right] & 0 & 0 \\ (1/2)\left[\rho_i \left(\frac{\partial \xi_i}{\partial \gamma}\right) + \delta_i \left(\frac{\partial^2 \xi_i}{\partial \kappa \partial \gamma}\right)\right] & 0 & 0 \end{pmatrix}$$

$$\frac{\partial^2 \mu_i}{\partial \epsilon \partial \epsilon^T} = e^{x_i^T \boldsymbol{\beta}} \begin{pmatrix} \left[2\rho_i \left(\frac{\partial \xi_i}{\partial \kappa}\right) + \delta_i \left(\frac{\partial^2 \xi_i}{\partial \kappa^2}\right)\right] & \left[\rho_i \left(\frac{\partial \xi_i}{\partial \eta}\right) + \delta_i \left(\frac{\partial^2 \xi_i}{\partial \kappa \partial \eta}\right)\right] & \left[\rho_i \left(\frac{\partial \xi_i}{\partial \gamma}\right) + \delta_i \left(\frac{\partial^2 \xi_i}{\partial \kappa \partial \gamma}\right)\right] \\ \left[\rho_i \left(\frac{\partial \xi_i}{\partial \eta}\right) + \delta_i \left(\frac{\partial^2 \xi_i}{\partial \kappa \partial \eta}\right)\right] & 0 & 0 \\ \left[\rho_i \left(\frac{\partial \xi_i}{\partial \gamma}\right) + \delta_i \left(\frac{\partial^2 \xi_i}{\partial \kappa \partial \gamma}\right)\right] & 0 & 0 \end{pmatrix}$$

To simplify the notation in our derivation, we define the following $n \times n$ diagonal matrices, Δ_1 through Δ_6. Elements on the diagonal are contained within the diag() notation below and are specified for a single subject.

$$\begin{aligned}
\Delta_1 &= \text{diag}\left(\frac{\mu_i}{Var[Y_i]}\right)\\
\Delta_2 &= \text{diag}\left(\frac{\mu_i^2}{Var[Y_i]}\right)\\
\Delta_3 &= \text{diag}\left(\frac{\mu_i^3(y_i-\mu_i)}{Var[Y_i]^2}\right)\\
\Delta_4 &= \text{diag}\left(\frac{\mu_i^2(y_i-\mu_i)}{Var[Y_i]^2}\right)\\
\Delta_5 &= \text{diag}\left(\frac{1}{Var[Y_i]}\right)\\
\Delta_6 &= \text{diag}\left(\frac{(y_i-\mu_i)(1+2*\phi\mu_i)}{Var[Y_i]^2}\right)
\end{aligned}$$

The log-likelihood for the TReC component is given by:

$$\ell_{TReC} = \sum_{i=1}^{N} \text{ln}\Gamma\left(y_i + 1/\phi\right) - \text{ln}\Gamma\left(1/\phi\right) - \text{ln}\Gamma\left(y_i+1\right) - \left[1/\phi + y_i\right]\ln\left(1+\phi\mu_i\right) + \\ y_i\left(\ln(\phi)+\ln(\mu_i)\right)$$

It can be shown that the following hold for derivatives involving $\boldsymbol{\beta}$:

$$\begin{aligned}
\frac{\partial \ell}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{N}\left(\frac{y_i-\mu_i}{1+\phi\mu_i}\right)x_i = X^T\Delta_1(Y-\mu)\\
\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} &= -\sum_{i=1}^{N}\left[\frac{\mu_i}{1+\phi\mu_i} + \frac{\phi\mu_i\left(y_i-\mu_i\right)}{(1+\phi\mu_i)^2}\right]x_i x_i^T = -\left[X^T\Delta_2 X + \phi X^T\Delta_3 X\right]\\
\frac{\partial \ell}{\partial \boldsymbol{\beta}\partial \epsilon^T} &= -\sum_{i=1}^{N}\left[\frac{1}{1+\phi\mu_i} + \frac{\phi(y_i-\mu_i)}{(1+\phi\mu_i)^2}\right]x_i\frac{\partial \mu_i}{\partial \epsilon}^T = -\left[X^T\Delta_1 D_\mu(\epsilon) + \phi X^T\Delta_4 D_\mu(\epsilon)\right]\\
\frac{\partial \ell}{\partial \boldsymbol{\beta}\partial \phi} &= -\sum_{i=1}^{N}\left[\frac{(y_i-\mu_i)\mu_i}{(1+\phi\mu_i)^2}\right]x_i = -X^T\Delta_3 J_N
\end{aligned}$$

Regarding derivatives involving ϵ, we have:

$$
\begin{aligned}
\frac{\partial \ell_{TReC}}{\partial \epsilon} &= \sum_{i=1}^{N} \left[\frac{y_i - \mu_i}{\mu_i + \phi\mu_i^2}\right] \frac{\partial \mu_i}{\partial \epsilon} = D_\mu(\epsilon)^T \Delta_5 \left(Y - \mu\right), \\
\frac{\partial \ell_{TReC}}{\partial \epsilon \partial \epsilon^T} &= \sum_{i=1}^{N} - \left[\frac{1}{\mu_i + \phi\mu_i^2} + \frac{(y_i - \mu_i)(1 + 2\phi\mu_i)}{\left(\mu_i + \phi\mu_i^2\right)^2}\right] \left(\frac{\partial \mu_i}{\partial \epsilon}\right) \left(\frac{\partial \mu_i}{\partial \epsilon}\right)^T + \left(\frac{y_i - \mu_i}{\mu_i + \phi\mu_i^2}\right) \left(\frac{\partial^2 \mu_i}{\partial \epsilon \partial \epsilon^T}\right) \\
&= - \left[D_\mu(\epsilon)^T \Delta_5 D_\mu(\epsilon) + D_\mu(\epsilon)^T \Delta_6 D_\mu(\epsilon)\right] + \sum_{i=1}^{N} \left(\frac{y_i - \mu_i}{\mu_i + \phi\mu_i^2}\right) \left(\frac{\partial^2 \mu_i}{\partial \epsilon \partial \epsilon^T}\right), \\
\frac{\partial \ell}{\partial \epsilon \partial \phi} &= - \sum_{i=1}^{N} \frac{(y_i - \mu_i)\mu_i^2}{(\mu_i + \phi\mu_i^2)^2} = -D_\mu(\epsilon)^T \Delta_4 J_N.
\end{aligned}
$$

Finally, derivatives involving ϕ are provided below:

$$
\begin{aligned}
\frac{\partial \ell}{\partial \phi} &= \sum_{i=1}^{N} -\phi^{-2} \left[\Psi_0(y_i + \phi^{-1}) - \Psi_0(\phi^{-1}) - \ln(1 + \phi\mu_i)\right] - (\phi^{-1} + y_i) \left[\frac{\mu_i}{1 + \phi\mu_i}\right] + \frac{y_i}{\phi} \\
\frac{\partial \ell}{\partial \phi^2} &= \sum_{i=1}^{N} 2\phi^{-3} \left[\Psi_0(y_i + \phi^{-1}) - \Psi_0(\phi^{-1}) - \ln(1 + \phi\mu_i)\right] + \phi^{-4} \left[\Psi_1(y_i + \phi^{-1}) - \Psi_1(\phi^{-1})\right] + 2\phi^{-2} \left[\frac{\mu_i}{1 + \phi\mu_i}\right] \\
&\qquad - \frac{y_i}{\phi^2} + (\phi^{-1} + y_i) \left[\frac{\mu_i^4}{V[Y_i]^2}\right]
\end{aligned}
$$

C.1.4.3 ASE Derivatives

Preceding development of the gradients and Hessians of the ASE component in the following section, it will be helpful to compose a list of definitions and useful derivatives for later use. Recall the definitions of ξ_i^A and π_i:

$$
\begin{aligned}
\xi_i^A &= (1 - c_i)(\eta + \alpha_\eta) + c_i(\gamma + \alpha_\gamma) \\
\pi_i &= \begin{cases} \xi_i^A/(1 + \xi_i^A) & \text{, if } G(i) = AB \\ 0.5 & \text{, otherwise} \end{cases}
\end{aligned}
$$

For genotypes AA and AB, π_i is independent of our parameters. Only genotype AB will be considered. Thus, consider the gradient of ξ_i^A with respect to our parameters.

$$\frac{\partial \xi_i^A}{\partial(\epsilon, \alpha)} = \begin{pmatrix} [(\gamma + \alpha_\gamma) - (\eta + \alpha_\eta)] \left(\frac{\partial c_i}{\partial \kappa}\right) \\ 1 - c_i \\ c_i \\ 1 - c_i \\ c_i \end{pmatrix}$$

The Hessian of ξ_i is presented below:

$$\frac{\partial \xi_i^A}{\partial(\epsilon, \alpha)\partial(\epsilon, \alpha)^T} = \begin{pmatrix} [(\gamma + \alpha_\gamma) - (\eta + \alpha_\eta)] \left(\frac{\partial^2 c_i}{\partial \kappa^2}\right) & -\frac{\partial c_i}{\partial \kappa} & \frac{\partial c_i}{\partial \kappa} & -\frac{\partial c_i}{\partial \kappa} & \frac{\partial c_i}{\partial \kappa} \\ -\frac{\partial c_i}{\partial \kappa} & 0 & 0 & 0 & 0 \\ \frac{\partial c_i}{\partial \kappa} & 0 & 0 & 0 & 0 \\ -\frac{\partial c_i}{\partial \kappa} & 0 & 0 & 0 & 0 \\ \frac{\partial c_i}{\partial \kappa} & 0 & 0 & 0 & 0 \end{pmatrix}$$

Then for an arbitrary λ, we have:

$$\frac{\partial \pi_i}{\partial \lambda} = (1 + \xi_i^A)^{-2} \left(\frac{\partial \xi_i^A}{\partial \lambda}\right)$$

$$\frac{\partial^2 \pi_i}{\partial \lambda_1 \partial \lambda_2} = -2(1 + \xi_i^A)^{-3} \left(\frac{\partial \xi_i}{\partial \lambda_1}\right) \left(\frac{\partial \xi_i}{\partial \lambda_2}\right) + (1 + \xi_i^A)^{-2} \left(\frac{\partial^2 \xi_i^A}{\partial \lambda_1 \partial \lambda_2}\right)$$

The log-likelihood for the ASE component of the data is given below:

$$\begin{aligned} \ell_{ASE} = \sum_{i=1}^{n} & \ln\Gamma\left(r_i + 1\right) - \ln\Gamma\left(r_{iB} + 1\right) - \ln\Gamma\left(r_i - r_{iB} + 1\right) + \ln\Gamma\left(\psi^{-1}\right) - \ln\Gamma\left(\psi^{-1}\pi_i\right) - \\ & \ln\Gamma\left(\psi^{-1}(1 - \pi_i)\right) + \ln\Gamma\left(\psi^{-1}\pi_i + r_{iB}\right) + \ln\Gamma\left(\psi^{-1}(1 - \pi_i) + r_i - r_{iB}\right) - \\ & \ln\Gamma\left(\psi^{-1} + r_i\right) \end{aligned}$$

Let λ represent a single parameter from either ϵ or α. For such terms, it can be shown that:

$$\begin{aligned}
\frac{\partial \ell_{ASE}}{\partial \lambda} &= \sum_{i=1}^{n} \psi^{-1} B_i \left(\frac{\partial \pi_i}{\partial \lambda} \right) \\
\frac{\partial^2 \ell_{ASE}}{\partial \lambda_1 \partial \lambda_2} &= \sum_{i=1}^{n} \psi^{-1} B_i \left(\frac{\partial^2 \pi_i}{\partial \lambda_1 \partial \lambda_2} \right) + \psi^{-1} \left(\frac{\partial B_i}{\partial \pi_i} \right) \left(\frac{\partial \pi_i}{\partial \lambda_1} \right) \left(\frac{\partial \pi_i}{\partial \lambda_2} \right) \\
\frac{\partial \ell_{ASE}}{\partial \lambda \partial \psi} &= \sum_{i=1}^{n} -\psi^{-2} B_i \left(\frac{\partial \pi_i}{\partial \lambda} \right) + \psi^{-1} \left(\frac{\partial B_i}{\partial \psi} \right) \left(\frac{\partial \pi_i}{\partial \lambda} \right)
\end{aligned}$$

Where we define B_i and it's derivatives in the following way:

$$\begin{aligned}
B_i &= -\Psi_0\left(\psi^{-1}\pi_i\right) + \Psi_0\left(\psi^{-1}(1-\pi_i)\right) + \Psi_0\left(\psi^{-1}\pi_i + r_{iB}\right) - \Psi_0\left(\psi^{-1}(1-\pi_i) + r_i - r_{iB}\right) \\
\frac{\partial B_i}{\partial \pi_i} &= \psi^{-1}\Bigg[- \Psi_1\left(\psi^{-1}\pi_i\right) - \Psi_1\left(\psi^{-1}(1-\pi_i)\right) + \Psi_1\left(\psi^{-1}\pi_i + r_{iB}\right) + \\
&\qquad \Psi_1\left(\psi^{-1}(1-\pi_i) + r_i - r_{iB}\right)\Bigg] \\
\frac{\partial B_i}{\partial \psi} &= -\psi^{-2}\pi_i\left[-\Psi_1\left(\psi^{-1}\pi_i\right) + \Psi_1\left(\psi^{-1}\pi_i + r_{iB}\right)\right] + \\
&\qquad -\psi^{-2}(1-\pi_i)\left[\Psi_1\left(\psi^{-1}(1-\pi_i)\right) - \Psi_1\left(\psi^{-1}(1-\pi_i) + r_i - r_{iB}\right)\right]
\end{aligned}$$

Derivatives involving ψ are specified below:

$$\begin{aligned}
\frac{\partial \ell_{ASE}}{\partial \psi} &= \sum_{i=1}^{N_{AS}} -\psi^{-2} A_i, \\
\frac{\partial^2 \ell_{ASE}}{\partial \psi^2} &= \sum_{i=1}^{N_{AS}} 2\psi^{-3} A_i - \psi^{-2}\left(\frac{\partial A_i}{\partial \psi}\right),
\end{aligned}$$

where A_i and its derivatives are specified by:

$$\begin{aligned}
A_i &= \pi_i \left[-\Psi_0\left(\psi^{-1}\pi_i\right) + \Psi_0\left(\psi^{-1}\pi_i + r_{iB}\right)\right] + \\
&\quad (1-\pi_i)\left[-\Psi_0\left(\psi^{-1}(1-\pi_i)\right) + \Psi_0\left(\psi^{-1}(1-\pi_i) + r_i - r_{iB}\right)\right] + \\
&\quad \left[\Psi_0\left(\psi^{-1}\right) - \Psi_0\left(\psi^{-1} + r_i\right)\right], \\
\frac{\partial A_i}{\partial \psi} &= -\psi^{-2}\pi_i^2\left[-\Psi_1\left(\psi^{-1}\pi_i\right) + \Psi_1\left(\psi^{-1}\pi_i + r_{iB}\right)\right] - \\
&\quad \psi^{-2}(1-\pi_i)^2\left[-\Psi_1\left(\psi^{-1}(1-\pi_i)\right) + \Psi_1\left(\psi^{-1}(1-\pi_i) + r_i - r_{iB}\right)\right] - \\
&\quad \psi^{-2}\left[\Psi_1\left(\psi^{-1}\right) - \Psi_1\left(\psi^{-1} + r_i\right)\right].
\end{aligned}$$

C.1.4.4 Fisher's Information: Observed or Expected

The traditional form of the score test involves use of the expected Fisher's Information Matrix. In the case where the expected value of the Fisher's Information Matrix is difficult to compute, the observed Fisher's Information Matrix is often used [85]. In some situations, while use of the observed Fisher's Information Matrix still provides a statistically valid test under the null, it can be unstable and produce inconsistent estimates of the variance matrix for MLEs [85]. In the likelihood framework proposed by this paper, there is an inherent, stochastic dependence of R_i on Y_i. Namely, the value of R_i depends on the number of heterozygous SNPs present within the gene body and cannot exceed Y_i. This makes computing the expected Fisher's Information Matrix challenging as it becomes an infinite sum of finite sums containing the digamma and trigamma functions.

As such, we may compute an approximation to the expected Fisher's Information Matrix which assumes that Y_i and R_i are stochastically independent or we may use the observed Fisher's Information Matrix. The observed Fisher's Information Matrix can be computed as in the previous section using untransformed κ, η, γ, and ψ or the log-transformations of these quantities. The log transformation variant of the observed score test, termed Observed Score test (log), is slightly more stable than its untransformed competitor. A comparison of these three methods [observed, observed (log), expected] on simulated data is provided below. To evaluate Type I error of the Cis-Trans score test, simulations follow the structure provided for the power simulations in the body of Chapter 5. To evaluate power, $\xi_{i,ASE}$ is set to 0.5 for all subjects regardless of eQTL genotype and eQTL effect strength. This behavior is designed to mimic trans-eQTL behavior. In the case of numerical instability for the observed information Cis-Trans score tests, the expected information variant is substituted.

	Observed Score Test		**Observed Score Test (log)**		**Expected Score Test**	
γ Value	**Power**	**Type I Error**	**Power**	**Type I Error**	**Power**	**Type I Error**
1.0	–	8.4 (7)	–	8.9 (4)	–	6.6 (5)
1.2	25.8 (6)	8.0 (2)	25.3 (4)	8.0 (4)	15.5 (0)	5.3 (1)
1.4	66.8 (38)	9.5 (10)	63.5 (35)	8.5 (3)	48.8 (0)	3.5 (0)
1.6	89.0 (88)	9.3 (2)	86.3 (66)	8.5 (4)	82.3 (0)	4.3 (0)
1.8	99.0 (185)	8.5 (2)	98.5 (164)	10.3 (3)	97.3 (0)	3.8 (0)

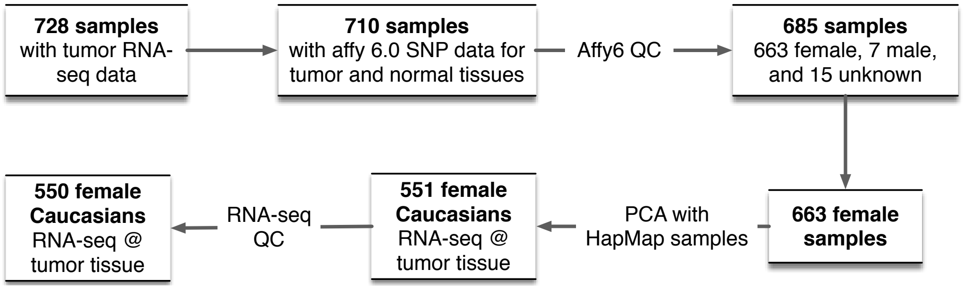
Table C.2: Summarizing the power and Type I error of the derived score tests. Number in parentheses represents the number of failures due to numerical instability.

As we can see from Supplementary Table C.2, the observed information matrix variants of the Cis-Trans Score test display superior power to the expected information variant at the cost of an inflated type I error ($\sim$8%). In addition, we note that the numerical instability of the observed information variants leads to a high rate of computation failure for the Cis-Trans score test. Due to its superior stability and Type I error, we opt to use the approximated expected Fisher's Information matrix within the real data analysis.

C.2 Supplementary Results for Real Data Analysis

C.2.1 Sample Size

Among these 728 patients, 178 were excluded from our analysis: 18 did not have genotype data (Affymetrix 6.0 array) from both tumor and paired normal samples, 35 failed Affymetrix genotype quality control (QC), 22 were male or of unknown gender, and 112 were non-Caucasian individuals, and 1 failed RNA-seq QC (Supplementary Figure C.1).



728 samples with tumor RNA-seq data
710 samples with affy 6.0 SNP data for tumor and normal tissues
Affy6 QC
685 samples 663 female, 7 male, and 15 unknown
663 female samples
PCA with HapMap samples
551 female Caucasians RNA-seq @ tumor tissue
RNA-seq QC
550 female Caucasians RNA-seq @ tumor tissue

Figure C.1: Sample size after each step of filtering.

C.2.2 Genotype Data Preparation

C.2.2.1 Genotype calling and quality control (QC)

We started our genotype data analysis with raw data in CEL files. After downloading all the CEL files of Affymetrix 6.0 arrays, we saved the file locations of these CEL files into file `cel_files_normal.txt` and ran the following APT (Affymetrix Power Tools) command to check genotype quality.

```
apt-geno-qc \
    --cdf-file  /path_to_lib_files/GenomeWideSNP_6.cdf \
    --qcc-file  /path_to_lib_files/GenomeWideSNP_6.r2.qcc \
    --qca-file  /path_to_lib_files/GenomeWideSNP_6.r2.qca \
    --cel-files /path_to_working_folder/cel_files_normal.txt \
    --out-file  /path_to_working_folder/apt-geno-qc.txt
```

Low quality samples were determined via low contrast QC (contrast.qc ≤ 0.4) or low QC call rate (qc.call.rate.all ≤ 0.8) (Supplementary Figure C.2).

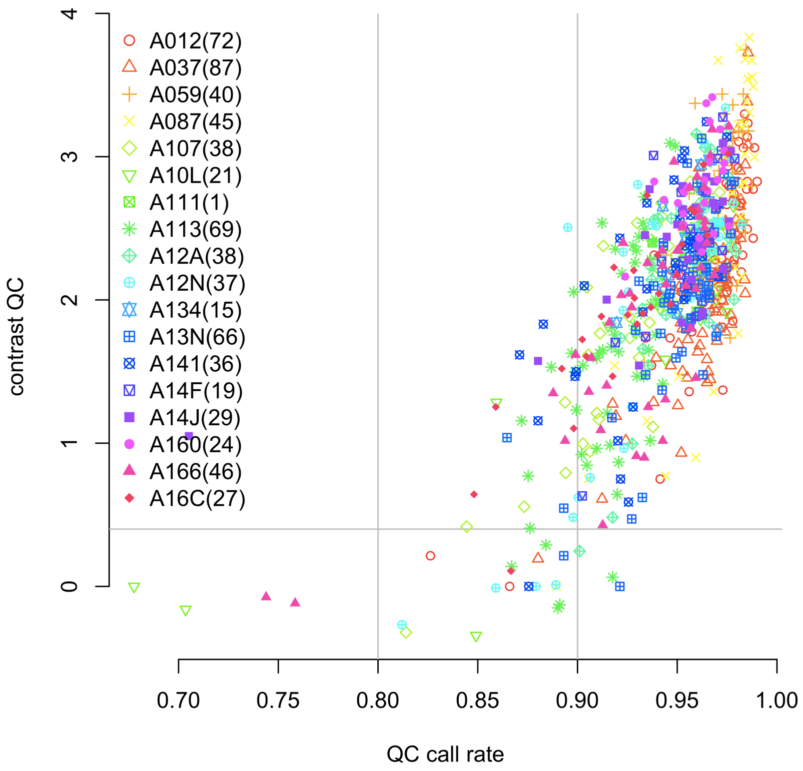


Figure C.2: Results of genotype QC by APT. Each sample is labelled by the plate it belongs to. The cutoff we use to select samples are QC call rate > 0.8 and contrast QC > 0.4.

After removing low quality samples, the new list of 685 remaining CEL files were recorded in file `cel_files_normal_after_qc.txt`. We called genotypes and genders for these 685 samples using birdseed-v2 implemented as part of APT.

```
apt-probeset-genotype \
  -o ../genotype_normal \
  -c  /path_to_lib_files/GenomeWideSNP_6.cdf \
  --set-gender-method cn-probe-chrXY-ratio \
  --chrX-probes /path_to_lib_files/GenomeWideSNP_6.chrXprobes \
  --chrY-probes /path_to_lib_files/GenomeWideSNP_6.chrYprobes \
  --special-snps /path_to_lib_files/GenomeWideSNP_6.specialSNPs \
  --read-models-birdseed /path_to_lib_files/GenomeWideSNP_6.birdseed-v2.models \
  -a birdseed-v2 \
  --cel-files /path_to_working_folder/cel_files_normal_after_qc.txt
```

To determine sample ethnicity, we performed PCA using genotype from TCGA samples together with genotypes from HAPMAP CEU (Caucasian), YRI (African), and CHB (Asian) samples. The PC1 versus PC2 plot clearly separated CEU, YRI, and CHB samples, and the TCGA samples that were clustered with CEU samples in the PC1 versus PC2 plot were classified as Caucasian samples (Supplementary Figure C.3).

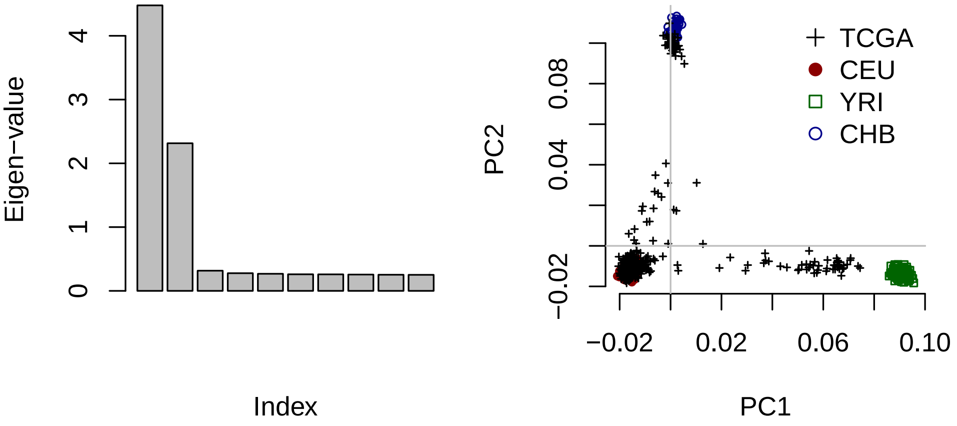


Figure C.3: The left panel shows eigen-values of the PCA, and the right panel shows PC1 versus PC2 plot. Based on this plot, we choose the Caucasian samples as those with PC1 $<$ 0 and PC2 $<$ 0.

C.2.2.2 Genotype Imputation

We imputed genotype data for the 551 samples that passed all the genotype-related filters. The output of birdseed includes genotype calls for 909,622 SNPs. We removed those SNPs without chromosome location information or with more than 5% of missing values leaving 832,334 SNPs which passed these filters. We used MACH [78] (mach.1.0.18.Linux) to phase and impute the genotypes using the 1000 Genome Reference ($\sim$36 million SNPs), which were downloaded from

MACH website (`http://csg.sph.umich.edu/abecasis/MaCH/download/1000G.2012-02-14.html`).

C.2.3 RNA-seq Data Preparation

We downloaded RNA-seq bam files from the TCGA data portal. First, we pre-processed these bam files using the R function `prepareBAM` of R package `asSeq` (`http://research.fhcrc.org/sun/en/software/asSeq.html`), to remove duplicated reads, or reads with average sequencing quality or mapping quality lower than 10. Next the expression of each gene in a sample is calculated as the number of RNA-seq reads that overlap with the exonic regions of this gene, obtained using R function `asSeq/countReads`. Annotations of exonic regions of each gene were obtained from Ensembl (version Homo_sapiens.GRCh37.66). Based on this version of gene annotation, we obtained read counts for 53,561 genes. Many of these genes have zero expression across most of the samples. We selected the 18,827 genes for which the 75 percentile of gene expression is equal or larger than 20. In other words, we remove those genes whose expression is less than 20 in more than 75% of samples.

To obtain allele-specific read counts for each sample, we first extracted all the heterozygous SNPs per sample, and then extracted those RNA-seq reads that overlap with at least one heterozygous SNP by R function `asSeq/extractAsReads`. Such RNA-seq reads were saved into three bam files, one for reads that match haplotype 1, one for those that match haplotype 2, and one for those with conflicts. For example, a conflicting read may overlap with more than one heterozygous SNPs, and its haplotype assignment is not consistent across these heterozygous SNPs. Usually the number of reads assigned to the conflict bam file is much smaller than the number of reads assigned to the two other bam files, otherwise it indicates errors in the data files or the data processing pipeline. Approximately 3.4% of the RNA-seq reads are classified as allele-specific reads (Supplementary Figure C.4) across all 551 samples, with one apparent outlier (sample ID: A15R), which is labeled

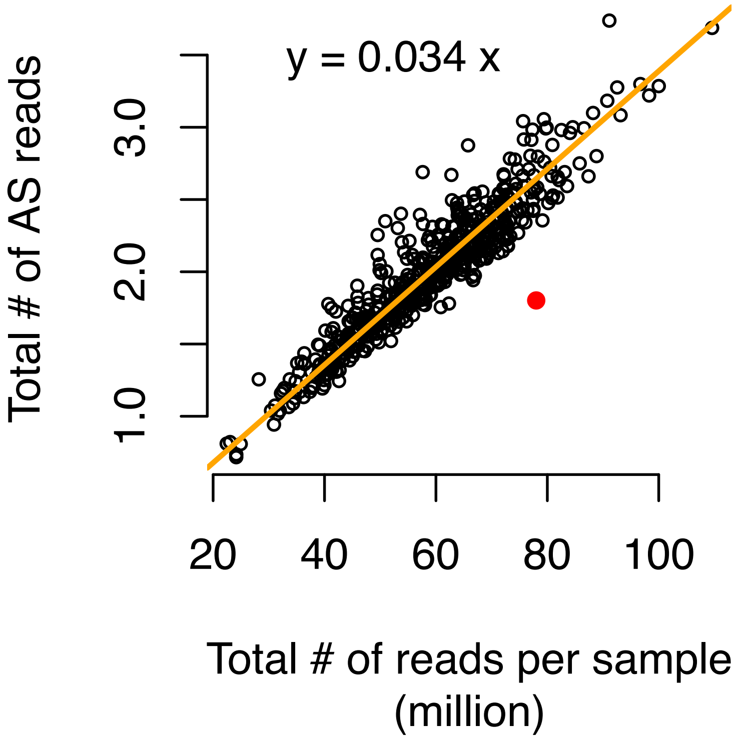


Figure C.4: The total number of reads (across all genes) per sample versus the total number of allele-specific reads per sample. The red point indicates a sample (A15R) that has unexpected low proportion of allele-specific reads and it is excluded from further analysis.

as red in Supplementary Figure C.4. We removed this sample in the following analysis.

For any association analysis using TReC per gene, one has to account for read-depth difference across samples. One way to quantify read-depth of a sample is to simply add up the total number of reads of this sample. Here we adopted a more robust approach, to quantify read-depth using 75 percentile of TReC across all the genes of a sample. In fact, in this data set, the two measurements of read depth are highly correlated (Supplementary Figure C.5).

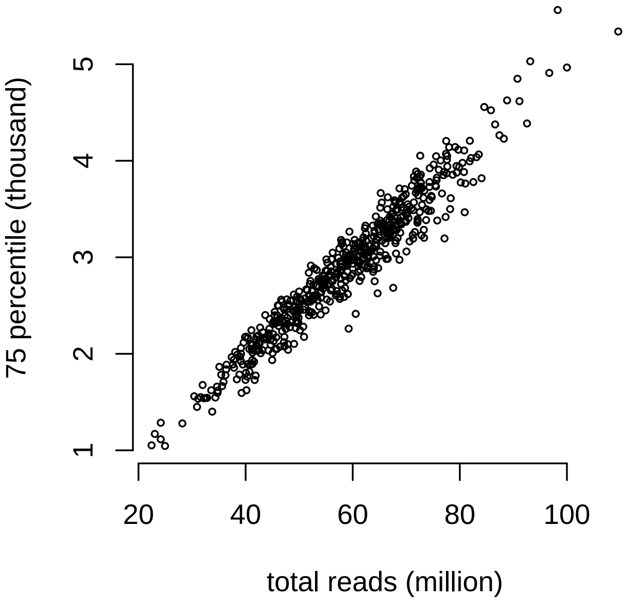


Figure C.5: The total number of reads (across all genes) per sample versus the 75 percentile of the TReC of all the genes within a sample.

C.2.4 eQTL mapping results

We summarize the agreement and disagreement of each tested model in the table below. This table summarizes the agreement of each model with respect to individual gene-SNP pairs.

Gene-SNP Pairs				
P-value Cutoff	**Category**	**pTReC(ASE)**	**TReC(ASE)**	**pLR**
5×10^{-4}	# of gene-SNP pairs	133,599	436,021	48,717
	overlap/alternative	–	19.8%	69.4%
	overlap/pTReC(ASE)	–	64.6%	25.3%
5×10^{-6}	# of gene-SNP pairs	43,605	208,546	14,285
	overlap/alternative	–	16.8%	80.5%
	overlap/pTReC(ASE)	–	80.2%	26.4%
5×10^{-8}	# of gene-SNP pairs	19,867	131,795	6,593
	overlap/alternative	–	13.5%	78.5%
	overlap/pTReC(ASE)	–	89.8%	26.0%

Table C.3: Summarizing the results of pTReC(ASE), TReC(ASE) and Westra models for TCGA data analysis. Here the notation pTReC(ASE) indicate that we use pTReCASE or pTReC model, depending on the results of Cis-Trans test. "overlap" represents the gene-SNP pairs identified by both pTReC(ASE) and an alternative method. "overlap/alternative" is the number of overlaps divided by the number of findings by the alternative method. "overlap/pTReC(ASE)" is the number of overlaps divided by the number of findings by pTReC(ASE). If we consider the results of pTReC(ASE) as true findings, then "overlap/alternative" is true discovery rate and "overlap/pTReC(ASE)" is sensitivity.

In addition, we perform the same summary with respect gene-level estimates. In this table, we summarize the number of genes with at least 1 significant eQTL and the prescribed p-value cutoff.

		Genes		
P-value Cutoff	**Category**	**pTReC(ASE)**	**TReC(ASE)**	**pLR**
5×10^{-4}	# of Genes	4788	7793	2055
	overlap/alternative	–	42.7	70.2
	overlap/pTReC(ASE)	–	69.5	30.3
5×10^{-6}	# of Genes	1245	2982	268
	overlap/alternative	–	27.0	85.4
	overlap/pTReC(ASE)	–	64.7	18.4
5×10^{-8}	# of Genes	496	1612	110
	overlap/alternative	–	21.4	93.6
	overlap/pTReC(ASE)	–	69.6	20.8

Table C.4: Summarizing the results of pTReC(ASE), TReC(ASE), the Westra models for TCGA data at gene level. The results are presented in the same format as Table 5.1, though the results are summarized at gene level instead of the level of SNP-gene pairs.

C.2.5 Additional results

Using the omic data prepared by [86], we examined the correlation between gene expression before and after removing copy number effects. Such correlations are very high for most of the genes. For example, it is larger than 0.8 for 86% of 15,284 genes with both gene expression and copy number data.

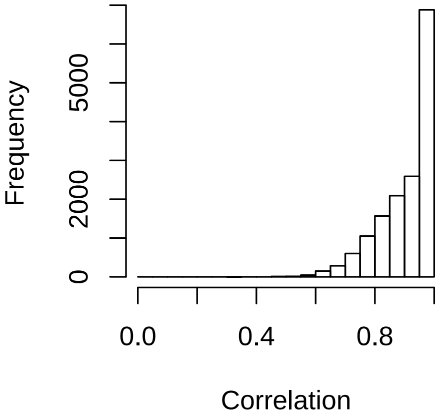


Figure C.6: The distribution of correlations between gene expression before and after removing copy number effects using a linear regression.

We also checked whether copy number of DNA methylation may confound the eQTLs reported in Figure 2 of the main paper. The first example is about gene ENSG00000115525 (ST3GAL5). Its expression is not associated with its copy number (p-value 0.22, $R^2 = 0.039$), but is associated with the methylation level of two CpG's: cg10017626 (p-value 6.2e-05, $R^2 = 0.039$) and cg07214715 (p-value 2.8e-05, $R^2 = 0.043$) after correcting for tumor purity and cell type compositions [86]. The second example is about gene ENSG00000142794 (NBPF3). Its expression is not associated with DNA methylation but is associated with its copy number (p-value 1.3e-07, $R^2 = 0.067$). These associations are illustrated in Figure C.7.

Next we check whether the associations between eQTL SNP genotype and gene expression are affected after controlling DNA methylation of gene expression measurement. We conducted this

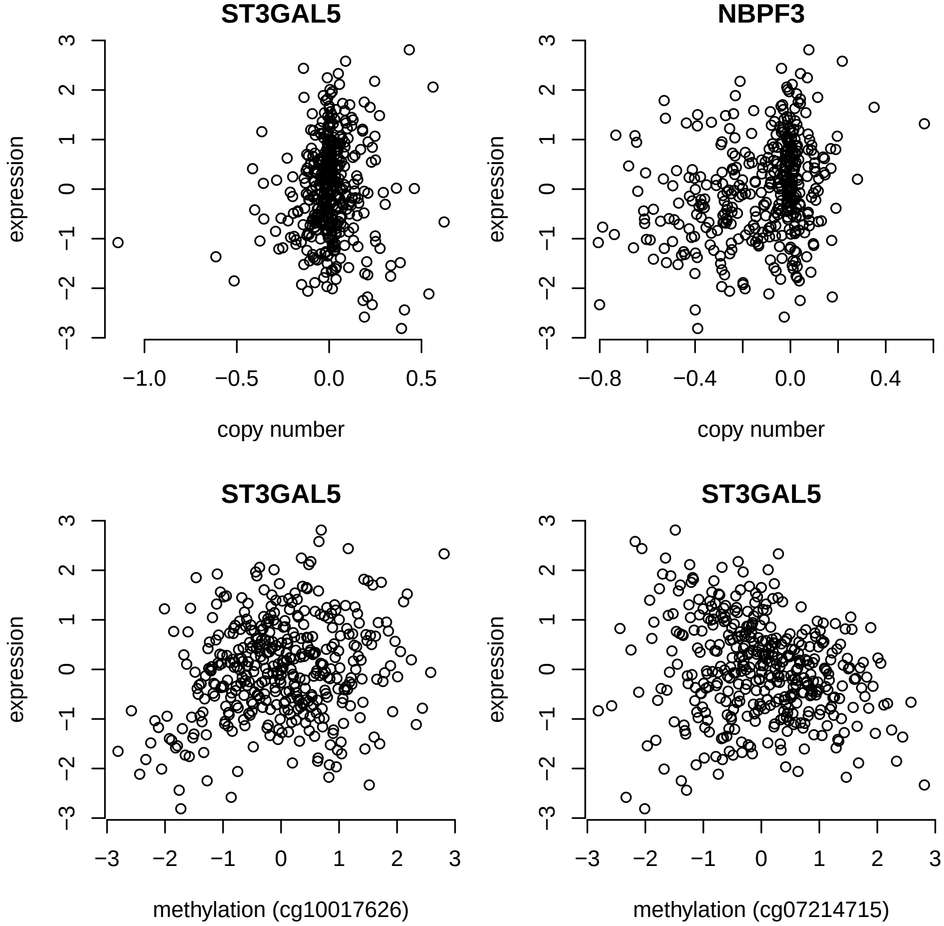


Figure C.7: Scatter plots demonstrate the associations between gene expression and copy number of two genes ST3GAL5 and NBPF3 (upper panel), and the associations between gene expression of ST3GAL5 and DNA methylation of two CpG's.

analysis in 328 samples (a subset of the 550 samples in main analysis) with all the data needed: SNP genotype, copy number, gene expression, and DNA methylation. Using a simple linear regression of gene expression versus SNP genotype (without using allele-specific expression), the eQTL p-value for ST3GAL5 is 2.3e-4, and after controlling for methylation, the p-values remain similar (1.8e-4 for cg10017626 and 5.4e-4 for cg07214715). The eQTL p-value for NBPF3 is also similar before and after controlling for copy number (t-statistics being 9.309 and 9.295 before and after controlling for copy number and p-value < 2e-16 in both cases).

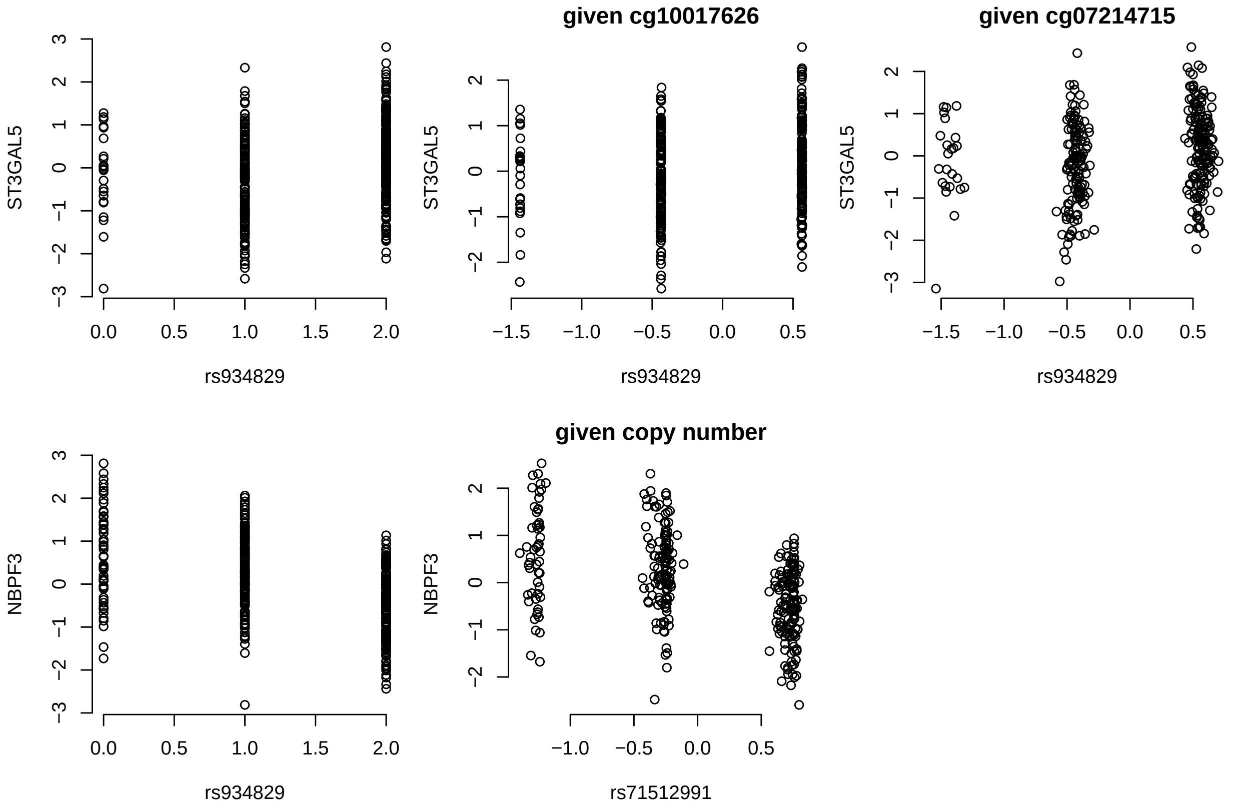


Figure C.8: Scatter plots demonstrate the associations between eQTL and gene expression before or after conditioning on two CpG's for gene ST3GAL5 (upper panel), and associations between eQTL and gene expression before or after conditioning on copy number alteration for gene NBPF3 (lower panel).

In the following, we provide visual justification for considering a copy number event as the difference between a gene's copy number and a samples ploidy. Each figure utilizes a different cutoff for this different in determining a copy number event. These figures also examine the extent of copy number events in the breast cancer data.

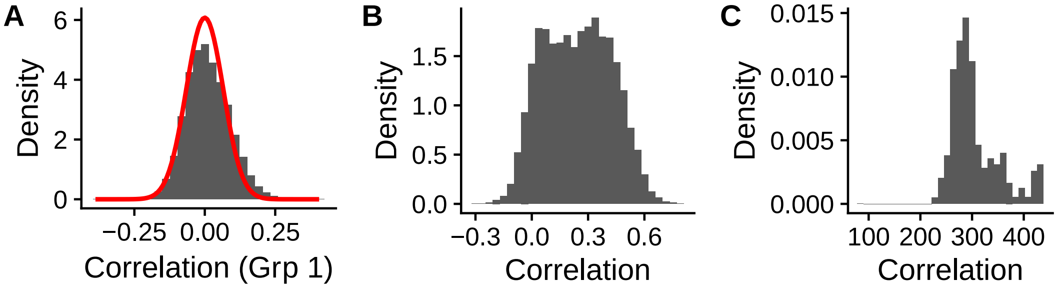


Figure C.9: Evaluating the extent of copy number aberration within the TCGA dataset. [A] Distribution of the correlations between D_{ij} and C_{ij} for subjects where $|D_{ij}| \leq 0.25$ summarized across all 18,134 genes. Red line indicates density of N(0,1/$\sqrt{232}$). [B] Distribution of the correlations between D_{ij} and relative gene expression summarized across all 18,134 genes [C] The distribution of the number of subjects with $|D_{ij}| > 0.25$ across all 18,134 genes.

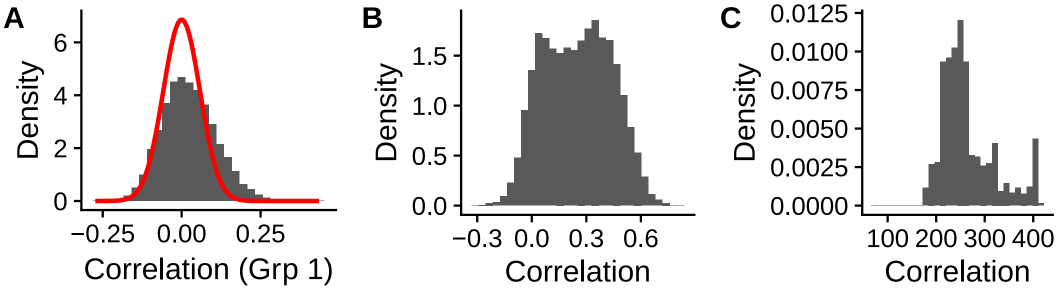


Figure C.10: Evaluating the extent of copy number aberration within the TCGA dataset. [A] Distribution of the correlations between D_{ij} and C_{ij} for subjects where $|D_{ij}| \leq 0.40$ summarized across all 18,134 genes. Red line indicates density of N(0,1/$\sqrt{296}$). [B] Distribution of the correlations between D_{ij} and relative gene expression summarized across all 18,134 genes [C] The distribution of the number of subjects with $|D_{ij}| > 0.40$ across all 18,134 genes.

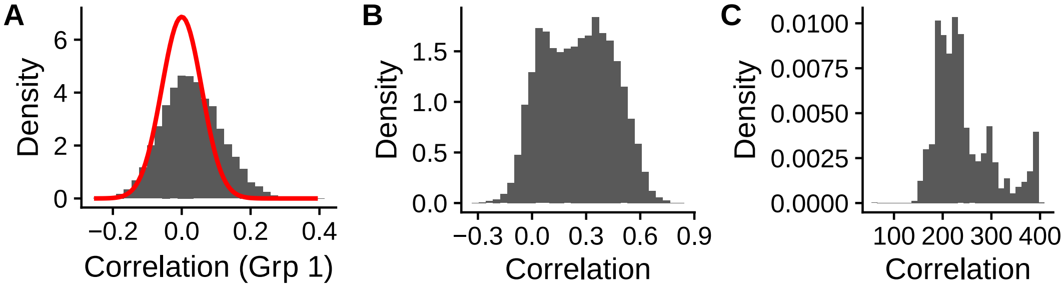


Figure C.11: Evaluating the extent of copy number aberration within the TCGA dataset. [A] Distribution of the correlations between D_{ij} and C_{ij} for subjects where $|D_{ij}| \leq 0.5$ summarized across all 18,134 genes. Red line indicates density of N(0,1/$\sqrt{296}$). [B] Distribution of the correlations between D_{ij} and relative gene expression summarized across all 18,134 genes [C] The distribution of the number of subjects with $|D_{ij}| > 0.5$ across all 18,134 genes.

BIBLIOGRAPHY

[1] B. Alberts, *Molecular biology of the cell.* Garland Science Taylor and Francis, 2008.

[2] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge, "Alternative isoform regulation in human tissue transcriptomes," *Nature*, vol. 46, no. 7221, p. 470476, 2008.

[3] B. T. Wilhelm and J.-R. Landry, "Rna-seqquantitative measurement of expression through massively parallel rna-sequencing," *Methods*, vol. 48, no. 3, p. 249257, 2009.

[4] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell, "Computational methods for transcriptome annotation and quantification using rna-seq," *Nature Methods*, vol. 8, no. 6, p. 469477, 2011.

[5] S. Schwartz, R. Oren, and G. Ast, "Detection and removal of biases in the analysis of next-generation sequencing reads," *PLoS ONE*, vol. 6, no. 1, 2011.

[6] W. Zheng, L. M. Chung, and H. Zhao, "Bias detection and correction in rna-sequencing data," *BMC Bioinformatics*, vol. 12, no. 1, p. 290, 2011.

[7] Y. Benjamini and T. P. Speed, "Summarizing and correcting the gc content bias in high-throughput sequencing," *Nucleic Acids Research*, vol. 40, Sep 2012.

[8] K. D. Hansen, S. E. Brenner, and S. Dudoit, "Biases in illumina transcriptome sequencing caused by random hexamer priming," *Nucleic Acids Research*, vol. 38, no. 12, 2010.

[9] M. L. Metzker, "Sequencing technologies - the next generation," *Nature Reviews Genetics*, vol. 11, no. 1, p. 3146, 2009.

[10] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: A revolutional tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, p. 5763, 2009.

[11] L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang, "Degseq: an r package for identifying differentially expressed genes from rna-seq data," *Bioinformatics*, vol. 26, no. 1, p. 136138, 2009.

[12] B. Li and C. N. Dewey, "Rsem: Accurate transcript quantification from rna-seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, no. 1, 2011.

[13] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "Rna-seq: An assessment of technical reproducibility and comparison with gene expression arrays," *Genome Research*, vol. 18, no. 9, p. 15091517, 2008.

[14] L. Pachter, "Models for transcript quantification from rna-seq," *Arxiv Database: q-bio*, vol. 1104, no. 2889, 2011.

[15] S. Anders and W. Huber, “Differential expression analysis for sequence count data,” *Genome Biology*, vol. 11, no. 10, 2010.

[16] M. D. Robinson and G. K. Smyth, “Moderated statistical tests for assessing differences in tag abundance,” *Bioinformatics*, vol. 23, no. 21, p. 28812887, 2007.

[17] W. Sun, Y. Liu, J. J. Crowley, T.-H. Chen, H. Zhou, H. Chu, S. Huang, P.-F. Kuan, Y. Li, D. Miller, and et al., “Isodot detects differential rna-isoform expression/usage with respect to a categorical or continuous covariate with high sensitivity and specificity,” *Journal of the American Statistical Association*, vol. 110, p. 975986, Mar 2015.

[18] S. S. Shen-Orr, R. Tibshirani, P. Khatri, D. L. Bodian, F. Staedtler, N. M. Perry, T. Hastie, M. M. Sarwal, M. M. Davis, A. J. Butte, and et al., “Cell type-specific gene expression differences in complex tissues,” *Nature Methods*, vol. 7, no. 4, p. 287289, 2010.

[19] M. M. Gosink, H. T. Petrie, and N. F. Tsinoremas, “Electronically subtracting expression patterns from a mixed cell population,” *Bioinformatics*, vol. 23, no. 24, p. 33283334, 2007.

[20] J. Clarke, P. Seo, and B. Clarke, “Statistical expression deconvolution from mixed tissue samples,” *Bioinformatics*, vol. 26, p. 10431049, Apr 2010.

[21] N. Wang, T. Gong, R. Clarke, L. Chen, I.-M. Shih, Z. Zhang, D. A. Levine, J. Xuan, and Y. Wang, “Undo: a bioconductor r package for unsupervised deconvolution of mixed gene expressions in tumor samples,” *Bioinformatics*, vol. 31, no. 1, p. 137139, 2014.

[22] P. Lu, A. Nakorchevskiy, and E. M. Marcotte, “Expression deconvolution: A reinterpretation of dna microarray data reveals dynamic changes in cell populations,” *Proceedings of the National Academy of Sciences*, vol. 100, no. 18, p. 1037010375, 2003.

[23] W. Qiao, G. Quon, E. Csaszar, M. Yu, Q. Morris, and P. W. Zandstra, “Pert: A method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions,” *PLoS Computational Biology*, vol. 8, no. 12, 2012.

[24] T. Gong and J. D. Szustakowski, “Deconrnaseq: a statistical framework for deconvolution of heterogeneous tissue samples based on mrna-seq data,” *Bioinformatics*, vol. 29, no. 8, p. 10831085, 2013.

[25] A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, and A. A. Alizadeh, “Robust enumeration of cell subsets from tissue expression profiles,” *Nature Methods*, vol. 12, no. 5, p. 453457, 2015.

[26] Y. Zhong, Y.-W. Wan, K. Pang, L. M. Chow, and Z. Liu, “Digital sorting of complex tissues for cell type-specific gene expression profiles,” *BMC Bioinformatics*, vol. 14, no. 1, p. 89, 2013.

[27] J. Ahn, Y. Yuan, G. Parmigiani, M. B. Suraokar, L. Diao, I. I. Wistuba, and W. Wang, “Demix: deconvolution for mixed cancer transcriptomes using raw measured data,” *Bioinformatics*, vol. 29, no. 15, p. 18651871, 2013.

[28] Y. Li and X. Xie, “A mixture model for expression deconvolution from rna-seq in heterogeneous tissues,” *BMC Bioinformatics*, Apr 2014.

[29] T. Erkkil, S. Lehmusvaara, P. Ruusuvuori, T. Visakorpi, I. Shmulevich, and H. Lhdesmki, “Probabilistic analysis of gene expression measurements from heterogeneous tissues,” *Bioinformatics*, vol. 26, no. 20, p. 25712577, 2010.

[30] G. Quon and Q. Morris, “Isolate: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing,” *Bioinformatics*, vol. 25, no. 21, p. 28822889, 2009.

[31] C. V. Anghel, G. Quon, S. Haider, F. Nguyen, A. G. Deshwar, Q. D. Morris, and P. C. Boutros, “Isopurer: an r implementation of a computational purification algorithm of mixed tumour profiles,” *BMC Bioinformatics*, vol. 16, no. 1, 2015.

[32] G. P. Dunn, C. M. Koebel, and R. D. Schreiber, “Interferons, immunity and cancer immunoediting,” *Nature Reviews Immunology*, vol. 6, no. 11, p. 836, 2006.

[33] V. Shankaran and et al, “Ifn gamma and lymphocytes prevent primary tumor development and shape tumor immunogenecity,” *Nature*, vol. 410, pp. 1107–1111, Apr 2001.

[34] e. a. Dunn, G. P., “Interferons, immunity and cancer immunoediting leading to impaired immune function in cancer patients,” *Nature Reviews Immunology*, vol. 7, no. 1, pp. 1–2, 2007.

[35] S. Farkona, E. P. Diamandis, and I. M. Blasutig, “Cancer immunotherapy: the beginning of the end of cancer?,” *BMC medicine*, vol. 14, no. 1, p. 73, 2016.

[36] C. L. Ventola, “Cancer immunotherapy, part 2: Efficacy, safety, and other clinical considerations,” *Pharmacy and Therapeutics*, vol. 42, Jul 2017.

[37] T. F. Gajewski, H. Schreiber, and Y. X. Fu, “Innate and adaptive immune cells in the tumor microenvironment,” *Nature Immunology*, vol. 14, no. 10, pp. 1014–1022, 2013.

[38] E. Sato, S. H. Olson, J. Ahn, B. Bundy, H. Nishikawa, F. Qian, A. A. Jungbluth, D. Frosina, S. Gnjatic, C. Ambrosone, *et al.*, “Intraepithelial CD8+ tumor-infiltrating lymphocytes and a high CD8+/regulatory T cell ratio are associated with favorable prognosis in ovarian cancer,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 51, pp. 18538–18543, 2005.

[39] G. Bindea, B. Mlecnik, M. Tosolini, A. Kirilovsky, M. Waldner, A. C. Obenauf, H. Angell, T. Fredriksen, L. Lafontaine, A. Berger, *et al.*, “Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer,” *Immunity*, vol. 39, no. 4, pp. 782–795, 2013.

[40] S. D. Brown, R. L. Warren, E. A. Gibb, S. D. Martin, J. J. Spinelli, B. H. Nelson, and R. A. Holt, "Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival," *Genome research*, vol. 24, no. 5, pp. 743–750, 2014. PMCID: PMC4009604 .

[41] M. S. Rooney, S. A. Shukla, C. J. Wu, G. Getz, and N. Hacohen, "Molecular and genetic properties of tumors associated with local immune cytolytic activity," *Cell*, vol. 160, no. 1, pp. 48–61, 2015.

[42] E. Becht, N. A. Giraldo, L. Lacroix, B. Buttard, N. Elarouci, F. Petitprez, J. Selves, P. Laurent-Puig, C. Sautes-Fridman, W. H. Fridman, and et al., "Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression," *Genome Biology*, vol. 17, no. 218, 2016.

[43] B. Li, E. Severson, J.-C. Pignon, H. Zhao, T. Li, J. Novak, P. Jiang, H. Shen, J. C. Aster, S. Rodig, *et al.*, "Comprehensive analyses of tumor immunity: implications for cancer immunotherapy," *Genome Biology*, vol. 17, no. 1, p. 174, 2016.

[44] J. Racle, K. D. Jonge, P. Baumgaertner, D. E. Speiser, and D. Gfeller, "Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data," *eLife*, vol. 6, 2017.

[45] W. S. Bush and J. H. Moore, "Chapter 11: Genome-wide association studies," *PLoS Computational Biology*, vol. 8, no. 12, 2012.

[46] W. Cookson, L. Liang, G. Abecasis, M. Moffatt, and M. Lathrop, "Mapping complex disease traits with global gene expression," *Nature Reviews Genetics*, vol. 10, no. 3, pp. 184–194, 2009.

[47] M. Rockman and L. Kruglyak, "Genetics of global gene expression," *Nature Reviews Genetics*, vol. 7, no. 11, pp. 862–872, 2006.

[48] W. Sun and Y. Hu, "eqtl mapping using rna-seq data," *Statistics in biosciences*, vol. 5, no. 1, pp. 198–219, 2013.

[49] L. W. M. Loo, I. Cheng, M. Tiirikainen, A. Lum-Jones, A. Seifried, L. M. Dunklee, J. M. Church, R. Gryfe, D. J. Weisenberger, R. W. Haile, and et al., "cis-expression qtl analysis of established colorectal cancer risk variants in colon tumors and adjacent normal tissue," *PLoS ONE*, vol. 7, no. 2, 2012.

[50] C. Grisanzio, L. Werner, D. Takeda, B. C. Awoyemi, M. M. Pomerantz, H. Yamada, P. Sooriakumaran, B. D. Robinson, R. Leung, A. C. Schinzel, and et al., "Genetic and functional analyses implicate the nudt11, hnf1b, and slc22a3 genes in prostate cancer pathogenesis," *Proceedings of the National Academy of Sciences*, vol. 109, no. 28, pp. 11252–11257, 2012.

[51] Q.-R. Chen, Y. Hu, C. Yan, K. Buetow, and D. Meerzaman, "Systematic genetic analysis identifies cis-eqtl target genes associated with glioblastoma patient survival," *PLoS ONE*, vol. 9, no. 8, 2014.

[52] Q. Li, A. Stram, C. Chen, S. Kar, S. Gayther, P. Pharoah, C. Haiman, B. Stranger, P. Kraft, and M. L. Freedman, “Expression qtl-based analyses reveal candidate causal genes and loci across five tumor types,” *Human molecular genetics*, p. ddu228, 2014.

[53] Q. Li, J.-H. Seo, B. Stranger, A. McKenna, I. Pe’Er, T. LaFramboise, M. Brown, S. Tyekucheva, and M. L. Freedman, “Integrative eqtl-based analyses reveal candidate causal genes and loci across five tumor types,” *Cell*, vol. 152, no. 3, pp. 633–641, 2013.

[54] H.-J. Westra, D. Arends, T. Esko, M. J. Peters, C. Schurmann, K. Schramm, J. Kettunen, H. Yaghootkar, B. P. Fairfax, A. K. Andiappan, and et al., “Cell specific eqtl analysis without sorting cells,” *PLoS Genetics*, vol. 11, 5 2015.

[55] W. Sun, “A statistical framework for eqtl mapping using rna-seq data,” *Biometrics*, vol. 68, pp. 1–11, 12 2011.

[56] Y.-J. Hu, W. Sun, J.-Y. Tzeng, and C. M. Perou, “Proper use of allele-specific expression improves statistical power for cis-eqtl mapping with rna-seq data,” *Journal of the American Statistical Association*, vol. 110, pp. 962–974, 3 2015.

[57] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, L. Pachter, and et al., “Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks,” *Nature Protocols*, vol. 7, p. 562578, Jan 2012.

[58] E. Furman, “On the convolution of the negative binomial random variables,” *Statistics and Probability Letters*, vol. 77, no. 2, p. 169172, 2007.

[59] E. P. Consortium, “An integrated encyclopedia of dna elements in the human genome,” *Nature*, vol. 489, no. 7414, p. 5774, 2012.

[60] P. S. Linsley, C. Speake, E. Whalen, and D. Chaussabel, “Copy number loss of the interferon gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis,” *PloS one*, vol. 9, no. 10, p. e109760, 2014.

[61] R. D. Schreiber, L. J. Old, and M. J. Smyth, “Cancer immunoediting: integrating immunity?s roles in cancer suppression and promotion,” *Science*, vol. 331, no. 6024, pp. 1565–1570, 2011.

[62] P. Sharma and J. P. Allison, “The future of immune checkpoint therapy,” *Science*, vol. 348, no. 6230, pp. 56–61, 2015.

[63] M. Yarchoan, A. Hopkins, and E. M. Jaffee, “Tumor mutational burden and response rate to PD-1 inhibition,” *New England Journal of Medicine*, vol. 377, no. 25, pp. 2500–2501, 2017.

[64] S. L. Topalian, J. M. Taube, R. A. Anders, and D. M. Pardoll, “Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy,” *Nature Reviews Cancer*, vol. 16, no. 5, p. 275, 2016.

[65] M. Angelova, P. Charoentong, H. Hackl, M. L. Fischer, R. Snajder, A. M. Krogsdam, M. J. Waldner, G. Bindea, B. Mlecnik, J. Galon, *et al.*, “Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy,” *Genome biology*, vol. 16, no. 1, p. 64, 2015. PMCID: PMC4377852.

[66] Y. Zhong and Z. Liu, “Gene expression deconvolution in linear space,” *Nature Methods*, vol. 9, no. 1, p. 89, 2012.

[67] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth, “voom: Precision weights unlock linear model analysis tools for rna-seq read counts,” *Genome biology*, vol. 15, no. 2, p. R29, 2014.

[68] M. Gierliński, C. Cole, P. Schofield, N. J. Schurch, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. Simpson, T. Owen-Hughes, *et al.*, “Statistical models for rna-seq data derived from a two-condition 48-replicate experiment,” *Bioinformatics*, vol. 31, no. 22, pp. 3625–3630, 2015.

[69] S. L. Carter, K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P. W. Laird, R. C. Onofrio, W. Winckler, B. A. Weir, *et al.*, “Absolute quantification of somatic dna alterations in human cancer,” *Nature biotechnology*, vol. 30, no. 5, pp. 413–421, 2012.

[70] L. Fenton, “The sum of log-normal probability distributions in scatter transmission systems,” *IRE Transactions on Communications Systems*, vol. 8, no. 1, pp. 57–67, 1960.

[71] W. Hugo, J. M. Zaretsky, L. Sun, C. Song, B. H. Moreno, S. Hu-Lieskovan, B. Berent-Maoz, J. Pang, B. Chmielowski, G. Cherry, *et al.*, “Genomic and transcriptomic features of response to anti-pd-1 therapy in metastatic melanoma,” *Cell*, vol. 165, no. 1, pp. 35–44, 2016.

[72] V. Bewick, L. Cheek, and J. Ball, “Statistics review 10: Further nonparametric methods,” *Critical Care*, vol. 8, p. 196199, Apr 2004.

[73] D. S. Chen and I. Mellman, “Elements of cancer immunity and the cancer–immune set point,” *Nature*, vol. 541, no. 7637, p. 321, 2017.

[74] P. Angerer, L. Simon, S. Tritschler, F. A. Wolf, D. Fischer, and F. J. Theis, “Single cells make big data: New challenges and opportunities in transcriptomics,” *Current Opinion in Systems Biology*, vol. 4, p. 8591, 2017.

[75] A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, *et al.*, “Science forum: the human cell atlas,” *Elife*, vol. 6, p. e27041, 2017.

[76] Q. Li, A. Stram, C. Chen, S. Kar, S. Gayther, P. Pharoah, C. Haiman, B. Stranger, P. Kraft, M. L. Freedman, and et al., “Expression qtl-based analyses reveal candidate causal genes and loci across five tumor types,” *Human Molecular Genetics*, vol. 23, pp. 5294–5302, 6 2014.

[77] C. R. Baquet, S. I. Mishra, P. Commiskey, G. L. Ellison, and M. DeShields, “Breast cancer epidemiology in blacks and whites: disparities in incidence, mortality, survival rates and histology,” *Journal of the National Medical Association*, vol. 100, no. 5, pp. 480–489, 2008.

[78] Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis, “Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes,” *Genetic epidemiology*, vol. 34, no. 8, pp. 816–834, 2010.

[79] A. A. Shabalin, “Matrix eqtl: ultra fast eqtl analysis via large matrix operations,” *Bioinformatics*, vol. 28, no. 10, pp. 1353–1358, 2012.

[80] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: A practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society. Series B Methodological*, vol. 57, no. 1, pp. 289–300, 1995.

[81] X. Gao, L. C. Becker, D. M. Becker, J. D. Starmer, and M. A. Province, “Avoiding the high bonferroni penalty in genome-wide association studies,” *Genetic Epidemiology*, 1 2010.

[82] P. V. Loo, S. H. Nordgard, O. C. Lingjaerde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, and et al., “Allele-specific copy number analysis of tumors,” *Proceedings of the National Academy of Sciences*, vol. 107, no. 39, pp. 16910–16915, 2010.

[83] R. Shen and V. E. Seshan, “Facets: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput dna sequencing,” *Nucleic acids research*, vol. 44, no. 16, pp. e131–e131, 2016.

[84] C. Radhakrishna Rao, “Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation,” *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 44, no. 1, p. 5057, 1948.

[85] D. A. Freedman, “How can the score test be inconsistent?,” *The American Statistician*, vol. 61, no. 4, pp. 291–295, 2007.

[86] W. Sun, P. Bunn, C. Jin, P. Little, V. Zhabotynsky, C. M. Perou, D. N. Hayes, M. Chen, and D.-Y. Lin, “The association between copy number aberration, dna methylation and gene expression in tumor samples,” *Nucleic acids research*, vol. 46, no. 6, pp. 3009–3018, 2018.