STATISTICAL METHODS IN INTRA-TUMOR HETEROGENEITY

Chong Jin

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 in the Gillings School of Global Public Health.

Chapel Hill
2020

Approved by:

Wei Sun

Danyu Lin

Mengjie Chen

Yuchao Jiang

Michael Love

Leonard McMillan

ABSTRACT

Chong Jin: Statistical Methods in Intra-Tumor Heterogeneity
(Under the direction of Wei Sun and Danyu Lin)

A tumor sample of a single patient often includes a conglomerate of heterogeneous cells. While its genetic and transcriptomic sequencing data represent a mixture of signals from different cell types, they can be further deconvolved so that we can get down to the level of each homogeneous component and test the association between the composition and some interesting response variables.

Understanding intra-tumor heterogeneity through deconvolution of genetic data may help us identify useful biomarkers to guide the practice of precision medicine. While popular methods exist, they usually do not jointly consider copy number aberrations and somatic point mutations and their timings under a valid statistical framework.

Differential expression using RNA sequencing data of bulk tissue samples (bulk RNA-seq) is a very popular and effective approach to study many biomedical problems. However, most tissue samples are composed of different cell types. Differential expression analysis without accounting for cell type composition cannot separate the changes due to cell type composition or cell type-specific expression. In addition, cell type-specific signals may be masked or even misrepresented, especially for relatively rare cell types.

In Chapter 2 of the proposed dissertation, we develop a new statistical method, SHARE (Statistical method for Heterogeneity using Allele-specific REads and somatic point mutations), that reconstructs clonal evolution history using whole-exome sequencing data of matched tumor and normal samples. Our method jointly models copy number aberrations and somatic point mutations using both total and allele-specific read counts. Cellular prevalence, allele-specific copy number and multiplicity of point mutations within each subclone can be estimated by maximizing the model likelihood. We apply our method to infer the subclonal composition in tumor samples from TCGA colon cancer patients.

In Chapter 3, we propose a new framework to address these limitations: **C**ell Type **A**ware analysis of **R**NA-**seq** (CARseq). CARseq employs a negative binomial regression approach to fully utilize the count features of RNA-seq data to improve statistical power. After evaluating its performance in simulations, we apply CARseq to compare gene expression of schizophrenia/autism subjects versus controls. Our results show that these two neurodevelopmental disorders differ from each other in terms of cell type composition changes and genes related to different types of neurotransmitter receptors were differentially expressed in neuron cells. We also discover some overlapping signals of differential expression in microglia, supporting the two diseases' similarity through immune regulation.

To Song and Meng

ACKNOWLEDGEMENTS

First, I would like to thank my advisors Drs. Wei Sun and Danyu Lin for their support—intellectually, emotionally, and financially. I would not have accomplished this dissertation, especially since there are many times when I am stuck in the middle of nowhere and cannot see any light at the end of the tunnel. I would also like to thank Dr. Jen Jen Yeh for providing me a research assistant position for a year so that I can have some crucial experience in collaborative work in cancer genomics.

Second, I would like to thank my committee members for their time and effort in the long process and their valuable suggestions, and I would like to thank my fellow graduate students from whom I learned most knowledge and got career advice.

Finally, I owe a lot to my family. I would like to thank my parents for their help regardless of the circumstances. I would like to thank Song and Meng who have brought me so much happiness in the gloomy days. Most importantly, I would like to give Dr. Mengqi Zhang, my wife, a special thank for her loving support that I am forever grateful to.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: LITERATURE REVIEW

1.1 Intra-tumor heterogeneity reconstruction using DNA sequencing data

Tumor initiation and progression involves gradual evolution from normal cells to cancer cells (Hanahan and Weinberg, 2000; Greaves and Maley, 2012). Tumor sequencing can provide important insights for tumor characteristics and treatment options. For example, recently it was shown that somatic mutation load estimated from sequencing data is associated with the efficacy of immunotherapy (Schumacher and Schreiber, 2015). Sequencing of multiple tumor biopsies from each patient, both from primary and metastasis regions, portrays a complex cancer genomic landscape highlighted by a branched evolutionary path of tumor growth (Gerlinger et al., 2012).

Collecting multiple tumor samples per patient is often impractical and thus many researchers and clinicians often face the daunting challenge to infer intra-tumor heterogeneity (ITH) from one tumor sample. A promising solution is to sequence a larger number of single cells. However, single-cell DNA sequencing data have poor quality due to very limited input materials (Wang et al., 2014). In addition, it is cost prohibitive to sequence a large number of cells, and even if one can afford to sequence tens of thousands of cells, it is still challenging to unbiasedly select these cells from billions of cells of the bulk sample. In this paper, we focus on *in-silico* approaches that estimate the subclonal structure of tumor cells based on DNA sequencing data from a single tumor sample.

Consider a evolutionary tree of tumor progression. The root of this tree is a normal cell that acquires somatic mutations gradually and evolves into a cancerous cell, which is the founding subclone of all tumor cells. Following previous works, we use the infinite site model for somatic point mutations (SPMs), which assumes one locus can only be mutated once. For somatic copy number aberrations (SCNAs), we assume each of the two alleles of a locus can only be mutated once, which is slightly more general than the assumption that each locus can only be mutated once (Li and Li, 2014).

1.1.1 Overview of computational methods for ITH

The study of intra-tumor heterogeneity (ITH) has attracted considerate amount of attention in the recent years. Many methods have been developed to infer ITH using somatic copy number aberration (SCNA) only (Van Loo et al., 2010; Carter et al., 2012; Oesper et al., 2013, 2014; Shen and Seshan, 2016) or somatic point mutation (SPM) only while either assuming SCNA is known (Roth et al., 2014; Deshwar et al., 2015; Chedom-Fotso et al., 2016) or only applicable to SCNA free regions (Miller et al., 2014; Hajirasouliha et al., 2014; El-Kebir et al., 2015; Yuan et al., 2015; Malikic et al., 2015). A few methods have jointly analyzed both SCNA and SPM. CHAT (Li and Li, 2014) estimates the cellular frequencies of SCNA and SPM, while accounting for the uncertainty of their order of occurrences, yet subclonal configurations are not estimated. BayClone2 (Lee et al., 2014) considers both SCNAs and SPMs and develops a new class of Bayesian feature allocation models to characterize subclonal copy number, but stops short of exploiting the phylogenetic relation between SCNAs and SPMs. CloneHD (Fischer et al., 2014) employs hidden Markov models (HMM) to couple SCNAs and SPMs without defining a phylogenetic tree. Instead of formulating a statistical model. EXPANDS (Andor et al., 2014) minimizes a heuristic error-term to combine SCNAs and SPMs to estimate subclonal cellular proportions.

Collecting tumor sequencing data becomes increasingly popular in clinical settings and typically one sample is collected per patient (Frampton et al., 2013). To assess the association between somatic mutations and clinical outcomes, it is important to take into account of ITH. Towards this goal, a statistical method to discern ITH of SCNA and SPM in a likelihood framework will be very useful, which can be subsequently combined with hypothesis testing step in association studies. Canopy (Jiang et al., 2016) can almost serve this purpose but it was mainly designed for in-depth analysis of multiple samples of one patient and its computational efficiency limits its application for large-scale studies of thousands of patients. This motivates us to develop this new method, Statistical method for Heterogeneity using Allele-specific REads and somatic point mutations (SHARE), which provides a likelihood inference framework to study ITH of SCNA and SPM, while emphasizing the situation that there is only one sample per patient.

1.1.2 GenoCNV and GenoCNA

Being an earlier attempt in literature to reconstruct copy number in the presence of normal cells, GenoCNV and genoCNA are developed to extract genotype and copy number calls from Illumina SNP arrays (Sun et al., 2009). As their names suggest, genoCNV is used to detect copy number change caused by germline mutation while genoCNA is suited to detect copy number change arising from somatic mutations, which is prevalent in tumor tissues. The novelty of the method is that genoCNA takes normal tissue contamination into account, and can also utilize paired tumor-normal tissue information when available. Both genoCNV and genoCNA can determine genotype of specific regions. The key input data are LRR (Log R Ratio) and BAF (B Allele Frequency) for all SNPs specified. LRR quantifies the overall copy number while BAF measures the allelic contrast. Other important parameters include the SNP names, positions of all the SNPs, population frequency of B allele, etc. GenoCNA also allows for adding paired tumor-normal tissue information to improve the result.

In the modeling strategy of genoCNV and genoCNA, two continuous time HMMs with discrete states representing genotypes are set up to describe LRR and BAF output. The two HMMs share the same states and transition probability but different emission probability. For SCNA, the nine states include the consideration of normal tissue contamination. LRR is modeled using the mixture of a uniform distribution to represent noise and a normal distribution of signal, while BAF can be decomposed into a uniform component and several truncated normal components. To solve the MLEs of the parameters in HMM, the Baum-Welch algorithm is employed. This computation intensive part is implemented using the C language in the genoCN software package. The Baum-Welch algorithm is iterated until the parameters converge.

In real data, genoCNV has a performance comparable to PennCNV for HapMap individuals, and genoCNA performs better than PennCNV in SCNA analysis using TCGA data. This is mainly because of the presence of normal tissue contamination invalidates some of the assumptions specifically designed for CNV studies in PennCNV, while genoCNA prevails at this point. The tumor purity estimation tested on TCGA data shows that the data-driven method gives a lower estimate than clinically-estimated tumor purity, which suggests that the latter may not accurate. However, it is important to mention that genoCNA may fail when the chromosomal background is not diploid.

1.1.3 ASCAT

Similar to genoCNA, although ASCAT is technically not a method for ITH (Van Loo et al., 2010), it is still worth mention since the copy numbers derived from ASCAT is still widely used as the input of some popular ITH methods that accept a precomputed copy number, e.g., pyclone (Roth et al., 2014). The authors develop a method called ASCAT (Allele-Specific Copy number Analysis of Tumors) to compute the allel-specific copy number of tumors from Illumina SNP array data, with normal tissue infiltration considered. The program can report aberrant tumor cell fraction, ploidy, gains, losses, loss of heterozygosity (LOH), copy number-neutral events, and allelic skewness.

As a preprocessing step before running ASCAT, ASPCF (allele-specific piecewise constant fitting) is used to preprocess SNP data to reduce noise. The main component of ASPCF is a penalized optimization criterion, which seeks to approximate LRR and BAF data using two piecewise constant functions with as few change points as possible. The change points that agree between piecewise constant functions derived from LRR and BAF are kept. The input parameters are designated beforehand.

After LRR and BAF are preprocessed with ASPCF, ASCAT is used to give estimates of ρ, percentage of aberrant cells, and ϕ_t, tumor ploidy. The LRR and BAF at each locus can be written as a function of ρ, ϕ_t and the underlying copy numbers of the two alleles. Thus the underlying copy numbers can be calculated given given LRR, BAF, ρ and ϕ_t. Then ρ and ϕ_t are estimated as the ones that bring the corresponding copy number profile the minimum distance to a integer number solution. To evaluate the solution, aberration reliability score measures how well the ASCAT-predicted integer copy numbers match the data.

To validate ASCAT, it is first applied to a dilution series of tumor sample mixed with different portions of normal sample. The derived allele-specific copy number is consistent. Second, the copy number measured by Feulgen photocytometry gives credence to the computational method. Finally, ASCAT is proven by copy numbers obtained using FISH experiments, which tends to slightly underestimate copy number when compared with ASCAT. Notably, there are cases that ASCAT would fall, especially when LRR is overly noisy even after ASPCF segmentation.

1.1.4 CloneHD

The cloneHD software (Fischer et al., 2014), written in C++, is a likelihood-based HMM model that integrates three data sources: copy number aberrations, B allele frequency and somatic SNVs from whole-genome sequencing data with 1 kb resolution. It can work on multiple samples from one patient. Unlike other methods that performs segmentation first, it relies on filterHD, a fuzzy data segmentation scheme that gives state transition probabilities for each locus.

The cloneHD software uses geometric distributions as prior to induce penalties. They derive posterior of SCNA, use it to obtain BAF prior, and then use the SCNA and BAF posterior to get SNV prior. They use a heuristic BIC to perform model selection.

Limitations of this method include not considering tree structure, which would impose more restrictions on the subclonal copy numbers and multiplicities of point mutations.

1.1.5 EXPANDS

EXPANDS (Andor et al., 2014), provided as an R package, is a method to estimate the cellular frequency of subpopulations (SP) in tumor samples. Conceptually, it is somewhat similar to the pyclone method, since both of them attempt to obtain the distribution of variant allele frequencies (VAFs) across somatic point mutations, and then cluster those VAFs. However, the cluster means provide an estimate of the cellular prevalence of mutations instead of subclones. For example, if a set of mutations occur in two subclones, assuming there is no SCNA and the mutations occur in one of the two alleles, then the cellular prevalence of these mutations is the one half of the summation of the cellular prevalence of the two subclones. The "one half" factor is due the assumption that the mutations occur on one of the two alleles.

Their nomenclature is somewhat unorthodox, in that they group somatic point mutations together with LOH events, and they use B allele ploidy to refer to multiplicity of point mutations. For each locus in each SP, they define total ploidy in tumor cells, total ploidy in normal cells, B allele ploidy in tumor cells, and B allele ploidy in normal cells. Then they define an error-term that comprises of SCNA and SNV, convert that to $P_l(f)$, the probability that the mutation of locus l occurs in a fraction f of cells, and then use K-L divergence as a measure of distance to cluster on the distributions associated with each locus.

Not explained in their journal article, the R package provides more functionalities. The vignette shows how to get phylogenetic relations between the SPs, i.e., infer copy number of each segment of each SP by assigning copy number of clustered SNVs, and then do a neighbor-joining clustering. Also provided in their vignette is a guide on how to use the method on multiple samples, but they do not actually provide the data to run their example script.

1.1.6 Canopy

Canopy (Jiang et al., 2016) is a Bayesian method to recover tree structure from tumor samples. Unlike our proposed method, it is mainly designed to jointly analyze multiple samples from the same subject, where Canopy can unleash its full potential when the clustering steps of SPMs provides enough simplification in the otherwise computationally demanding steps. The input of Canopy is very similar to what we designate as input of our program, except that Canopy takes segmental major and minor copy numbers instead of our program taking segmental LRR and BAF.

The following lists the input of Canopy:

R A matrix of mutant allele read depth. Each column is a sample and each row is a SPM. This is equivalent to X_{Mki} in our program.

X A matrix of total read depth. Each column is a sample and each row is a SPM. This is equivalent to $X_{Mki} + X_{Nki}$ in our program.

WM A matrix of observed major copy number. Each column is a sample and each row is a segment. This is similar to n_{Al} in our program, although in Canopy we know the major copy number for each segment using other methods. In simulation, we can provide the true n_{Al}.

Wm A matrix of observed minor copy number. Each column is a sample and each row is a segment. This is similar to n_{Bl} in our program.

epsilonM Standard deviation of **WM**. Related to σ_k in our method, it can be provided as a scalar or a matrix the same dimension as **WM**.

epsilonm Standard deviation of **Wm**.

C A matrix that specifies which segments have multiple SCNA events. It is based on a manual inspection of the LRR-BAF plots of multiple samples , and does not need to be specified in our simulation settings if we analyze one sample at a time.

Y A matrix that specifies which SPMs are on known SCNA segments. Each column is a segment and each row is a SPM. This is related to m_k in our program, but in a different coding scheme.

To improve the convergence of MCMC, especially when the VAF of mutations show clear cluster patterns, `canopy.cluster` can be used to perform a binomial clustering of SPMs using EM algorithm. It can then create psuedo-SPMs corresponding to each cluster that can be used in subsequent MCMC sampling.

The most time-consuming step is `canopy.sample`, where multiple chains with random initiations are sampled according to multiple update rules every iteration: sample SPM positions, sample clonal proportions, sample major and minor copy number, and sample whether SPM falls in the major or minor copy number. The results are saved every `writeskip` iterations.

To collect the sampling results, `canopy.post` merges the chains by only keeping the positions with larger likelihood values, so that the number of positions left is no more than five times the length of a chain after burn-in and thinning. Subsequently, the function removes the extremes of the likelihood values and remove the less visited configurations with the cutoff specified by `post.config.cutoff`.

1.1.7 PhyloWGS

PhyloWGS (Deshwar et al., 2015) is an earlier method that is similar to Canopy. In their modeling decision, normal cells count as a subclone. The differences are:

- PhyloWGS conceptualizes SCNA in a similar way as SPM. In fact, they need to transform SCNA into “pseudo” SPM in their model. Fortunately, they have provided a parser to transform a more widely used SCNA input file as their input.

- They use tree-structured stick-breaking process. This enables them to travel through all possible number of subclones within a single MCMC chain. (The default number of burn-in samples is 1000 and the default number of MCMC samples is 2500. They also don’t run multiple chains by default.)

In PhyloWGS, while not a method as foolproof as Canopy, it is possible to compute entropy by extracting cellular frequency and tree structure from the JSON output.

1.2 Cell type-specific differential expression from RNA sequencing data

Next-generation RNA sequencing has been generating massive amounts of expression data that deepen our knowledge across a spectrum of diseases, and it is still going strong after more than a decade of success. Towards the long term goal of effective prevention and treatment, researchers often compare transcriptomes of samples grouped by different clinical outcomes, hoping to elucidate the molecular characteristics and mechanisms of the underlying disease.

Differential expression analysis aims to identify the association between mRNA expression and covariates of interest. As an indispensable tool to hunt down possible biological mechanisms, it has long been part of the repertoire of a bioinformatician since its inception. The common practice in bulk RNA-seq analysis involves performing a transcriptome-wide screening for possible signals, and then a pathway-level enrichment analysis would ensue, generally based on some combination of gene-level (adjusted) p-value and log fold change (LFC) estimates. They can be obtained from some widely used methods, including edgeR (Robinson et al., 2010), voom (Law et al., 2014), and DESeq2 (Love et al., 2014), where borrowing information across genes through empirical Bayes methods to improve small sample performance is a recurring theme. While being able to regress on a couple of covariates, these testing methods do not offer the option to incorporate cell type composition, and any interpretation will rest upon the evidence gathered at the mixture level. Thus their results could be driven by differential cell type composition, and cell type-specific expression could be obscured especially when a cell type is relatively rare.

Cell type-specific differential expression test seeks to answer whether a gene is differentially expressed within a given cell type as if we can isolate purified cells under a microscopic scale. Single cell RNA-seq data, with cell types annotations from either marker genes or external references such as SingleR (Aran et al., 2019), virtually meet the canonical definition of transcriptome profiling after cell type purification. There are already a wealth of methods analyzing cell type-specific differential expression through single cell RNA-seq data (Wang et al., 2019). Methods specially tailored to single cell RNA-seq data require attention to the excess of zeros in read counts, e.g. by using a zero-inflated Poisson model in SCDE (Kharchenko et al., 2014) or by using a hurdle model in MAST (Finak et al., 2015). While the advancement of single cell RNA-seq techniques can provide unparalleled direct knowledge about individual cell types, large-scale single cell RNA-seq studies are less feasible than those using bulk RNA-seq due to budget considerations. It has also been reported that single-cell isolation and sequencing may introduce artifacts that can skew cellular proportions when compared to traditionally recognized approaches such as immunohistochemistry (IHC) estimates (Newman et al., 2019).

In the following subsections, we review a few methods about deconvolution of bulk RNA sequencing data and then highlight a couple of methods about cell type-specific differential expression analysis.

1.2.1 Cell type composition deconvolution

1.2.1.1 CIBERSORT

CIBERSORT (Newman et al., 2015) is a machine learning method developed to estimate cell type proportions from gene expression profiles in bulk tissue. It relies on a linear assumption that the observed expression is the sum of cell type-specific expression weighted by cellular proportions. Following this idea, they carefully constructed a signature matrix that contains only a subset of genes which are differentially expressed between different leukocyte cell types while removing genes with a high expression in solid tumor cell lines. To improve the stability of the deconvolution, number of signature genes is chosen to minimize the condition number. A support vector regression method is used to estimate cell type proportions. While CIBERSORT is validated in the deconvolution of microarray expression data, it can also be tailored to work on RNA-seq data, though the normalization process is different.

CIBERSORTx(Newman et al., 2019), the successor to CIBERSORT, is more conscious of the common practice of constructing the reference from single cell RNA-seq data, and supports B-mode and S-mode to remove unwanted platform-specific effects.

We will focus on the explanation of a special feature of CIBERSORTx, high-resolution purification, which allows for expression reconstruction at the sample level. The high-resolution expression purification is similar to NMF, where a matrix of expression is decomposed as the sum of multiple matrices of expression.

According to Page 31 of Supplementary Materials of CIBERSORTx, they have:

$$\text{diag}(\mathbf{G}_{i,\cdot,\cdot}\mathbf{F}) = \mathbf{M}_{i,\cdot}, 1 \leq i \leq n,$$

where for each gene i, $\mathbf{G}_{i,\cdot,\cdot}$ is a $k \times c$ expression matrix, with c indexing the cell types, $\mathbf{F}$ is a $c \times k$ cellular frequency matrix, $\mathbf{M}_{i,\cdot}$ is the mixture expression across k samples.

NMF will not give a unique solution unless we have more regularity conditions (Gillis, 2014). For identifiability purposes, CIBERSORTx high-resolution expression purification requires several assumptions:

1. The algorithm is applied to each gene separately.

2. Underlying phenotypic class structure can be revealed by ordering bulk gene expression data. In Supplementary Fig. 8 of CIBERSORTx, using an example by reconstructing bulk tissue by mixing scRNAseq data of five major pancreatic islet cell types of Type 2 Diabetes and normal subjects, they showed that fold change of cell type-specific gene expression was highly correlated with fold change

of reconstructed bulk gene expression, and it was also highly correlated with enrichment of case vs. control samples and high vs. low expression groups.

Assumption 2 is not expected to be easily justifiable in very general mixture datasets. Also detrimental to the logic of ordering bulk gene expression data, we can see in Supplementary Fig. 8(a) of CIBERSORTx that the scale of fold change in bulk is much smaller than the scale of fold change in purified cell types, as suggested by the regression line in that figure.

The steps in their high-resolution purification—performed gene by gene—can be summarized as follows:

1. Estimate cell type-specific expression. This is similar to how they obtain group-mode expression. They first apply NNLS on bootstrapped samples with a sample size around 4-5 fold greater than the number of cell types. Then they get two kinds of p-values testing whether the cell type-specific expression is zero or not. One is based on the empirical proportion of zeros observed using bootstrapping, and the other is based on a one group t-test based on the mean and variance estimated using bootstrapping. They combine two p-values using Fisher's method. If the p-value of a gene/cell type pair is less than 0.05 (by default), the cell type is included in high-resolution purification.

2. Sort the bulk gene expression. The samples remain sorted in subsequent steps.

3. When assuming a change point t can partition the samples into two groups, t is chosen so that the L2 reconstruction loss is minimized when the groupwise cell type-specific expression is estimated using subjects partitioned by the change point.

4. Cell type-specific test between two groups are partitioned by the optimal t. They get two kinds of p-values when testing whether the cell type-specific expression are different between two groups. One is based on the distribution of $|\mathbf{g}_2^* - \mathbf{g}_1|$, where $\mathbf{g}_2^*$ is the bootstrapped cell type-specific expression in group 2 (note that they didn't opt to use the bootstrapped $\mathbf{g}_1^*$ instead of $\mathbf{g}_1$ in the statistic), and the other is based on a two group t-test based on the mean and variance estimated using bootstrapping (This is actually similar to what CARseq would do if Wald test is used and there is no batch effects to adjust for). They combine two p-values using Fisher's method.

5. Refine the groupwise cell type-specific expression. If a cell type is not significantly expressed, it is removed from the model. Then the groupwise cell type-specific expression is shrinked closer to each other, the shrinkage being determined by the cell type-specific differential expression test in Step 4.

6. Smooth cell type-specific expression. Before step 6, the expression only has two values according to which side of t a sample resides after their expression is sorted. To obtain expression that is different in each sample, they use a sliding window of length w to obtain cell type-specific expression in each window. They pad the results so that we now have cell type-specific expression for every sample. Afterwards, they said they use least squares to individually fit the sample-specific expression coefficients to the groupwise cell type-specific expression for each cell type. It is not quite clear how it is done, but I suspect that rescaling the sample-specific expression coefficients to the groupwise cell type-specific expression would work.

7. Restore the original ordering of samples.

8. Repeat steps 1 to 7 for all genes.

While the high-resolution purification leads to a fairly straightforward application of cell type-specific differential expression testing, this application is not recommended due to possible inflation of type I error.

1.2.1.2 ICeD-T

ICeD-T is a more recent method of reference-based cell type deconvolution (Wilson et al., 2019).

Compared to CIBERSORT, it offers a few features:

1. ICeD-T is a likelihood-based framework where the gene expression deconvolution is done on a non-log scale while enjoying the benefits of variance stabilization through log-normal distribution.

2. ICeD-T explicitly models the aberrant genes through the incorporation of a component with inflated variance, therefore automatically down-weighting their contribution to the cell fraction estimates. In contrast, since CIBERSORT relies on a support vector regression model, the genes that fit the model very well (within some margin) do not actually affect the cell fraction estimates.

3. In ICeD-T, while the cell fraction estimates are non-negative, we actually rarely get really small cell fraction estimates even for the less abundant cell types. In CIBERSORT, however, the raw cell fraction estimates from support vector regression can even be negative, only later forced to be zero. Cell fraction estimates being zero will greatly reduce the applicability of cell type-specific differential expression analysis methods.

4. Optionally, ICeD-T allows for the user-specified tumor purity and variance weights.

CHAPTER 2: INFERRING INTRA-TUMOR HETEROGENEITY

SCNA is abundant in many solid tumors, such as breast cancer or colon cancer, and a SCNA may cover a long segment of the genome, such as one chromosome arm. Whenever a SPM occurs in a SCNA region, one needs to disentangle the influence of subclone structure and SCNA on this SPM. Multiplicity of point mutations are complicated by the order of somatic point mutation (SPM) and somatic copy number alteration (SCNA). For example, consider a SPM on A allele and a SCNA that doubles A allele. If the SPM happens first, the multiplicity of the SPM is 2 after SCNA, and if the SCNA happens first, the multiplicity of the SPM is 1. This motivates us to develop this new method, Statistical method for Heterogeneity using Allele-specific REads and somatic point mutations (SHARE), which provides a likelihood inference framework to study ITH of SCNA and SPM, while emphasizing the situation that there is only one sample per patient.

2.1 Likelihood model

2.1.1 Likelihood Model for SCNA

We have developed a complete pipeline to work with raw exome-seq data (Fig. 2.3). First consider the SCNA data from bulk tissue. We assume the exome-seq data are available from both tumor and matched normal samples. After running pipelines for quality control (QC), read mapping and counting, we obtain the total number of reads for the ith gene, denoted by Y_i, and the number of allele-specific reads, denoted by Y_{iA} and Y_{iB} for the two alleles A and B respectively. Denote the corresponding counts in matched normal sample by Y_{Ni}, Y_{NiA}, and Y_{NiB}, respectively. We filter out those genes with small number of reads or with significant allelic imbalance in normal sample, which implies mapping bias. Let d and d_N be the total number of reads in the normal and tumor sample, respectively. We define Log R Ratio (LRR) as $r_i = \log_2[(Y_i/d)/(Y_{Ni}/d_N)]$, and assume r_i follows a normal distribution.

Because B allele is arbitrarily decided, we model the allele-specific read count Y_{iB} given $Z_i = Y_{iA} + Y_{iB}$ by a mixture of binomial distribution:

$$0.5 f_{\mathscr{B}}(Y_{iB}|Z_i, b_i) + 0.5 f_{\mathscr{B}}(Y_{iB}|Z_i, 1 - b_i)$$

with success probability $b_i < 0.5$. b_i is often referred to as B allele frequency (BAF). We jointly segment LRR (r_i) and BAF (b_i) using a modified circular binary segmentation (CBS) method (Olshen et al., 2004). Given the likelihood function for LRR and BAF, we segment LRR and BAF simultaneously and evaluate change-points by likelihood ratio statistic. We have implemented this algorithm in an R package named `blueHouse`.

We define a copy number state by the allele-specific copy numbers of two alleles, denoted by n_{Al} and n_{Bl} for the lth state. By setting a maximum copy number *a priori*, the total number of copy number states, denoted by L, is known. For example, if the maximum copy number is 3 and there is only one subclone, there are 6 copy number states: $Null$ (homozygous deletion), A/B, AA/BB, AB, AAA/BBB, and AAB/ABB. This concept can be generalized to the situation with more than one subclone, where the copy number states are dictated by an array of n_{Asl} and n_{Bsl} with the subscript $s \in \{1, \dots, S\}$ for the sth subclone.

If the kth segment belongs to the lth copy number state, we assume the mean value of LRR and BAF of the segment, referred to as segmental LRR and BAF, follows normal distribution:

$$\bar{r}_k \sim f_{\mathcal{N}}(\mu_l, \sigma_k^2) \text{ and } \bar{b}_k \sim f_{\mathcal{N}}(\pi_l, \tau_k^2),$$

where $\mu_l = \log_2 \left[\{2(1 - \rho) + \rho(n_{Al} + n_{Bl})\} / \phi \right]$, ρ is tumor purity (the proportion of tumor cells in this tumor sample), ϕ is ploidy (the genome-wide average copy number of all cells in the tumor sample), the variance terms σ_k^2 and τ_k^2 are within-segment variances of LRR and BAF, and

$$\pi_l = [1 - \rho + \rho n_{Bl}]/[2 - 2\rho + \rho(n_{Al} + n_{Bl})].$$

Although $\bar{b}_l$ is bounded by $[0, 1]$, the normal distribution assumption is reasonable because the observed BAF does not reach these boundaries unless a tumor sample is 100% of pure tumor cells, which is unlikely in practice. After we flip $\bar{b}_l$ around 0.5, $\bar{b}_l$ is bounded by $[0, 0.5]$, and we need to use some heuristic rules to set $\bar{b}_l = 0.5$ in segments where evidence favors allelic balance. The log likelihood function $\bar{\mathbf{r}} = (\bar{r}_1, ..., \bar{r}_K)^T$ and $\bar{\mathbf{b}} = (\bar{b}_1, ..., \bar{b}_K)^T$ in the K segments can be written as

$$\mathscr{L}(\rho, \phi, \boldsymbol{\alpha} | \{\sigma_k^2\}, \{\tau_k^2\}, \bar{\mathbf{r}}, \bar{\mathbf{b}}) = \prod_{k=1}^{K} \sum_{l=1}^{L} \left\{ \alpha_l f_{\mathcal{N}} \left(\bar{r}_k \middle| \mu_l, \sigma_k^2 \right) f_{\mathcal{N}} \left(\bar{b}_k \middle| \pi_l, \tau_k^2 \right) \right\},$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_L)^T$ are the mixture proportion of the L mixture components.

Due to subclones, the copy numbers n_{Al} and n_{Bl} are likely fractions instead of integers. Assuming there are S subclones in total. For any genes that belong to the lth component, let n_{Asl} and n_{Bsl} be their *integer* copy numbers in the sth subclone. Let q_s be the proportion of cells from the sth subclone, then $\sum_{s=1}^{S} q_s = 1$, $n_{Al} = \sum_{s=1}^{S} q_s n_{Asl}$, and $n_{Bl} = \sum_{s=1}^{S} q_s n_{Bsl}$.

To encourage the selection of a model with moderate and infrequent copy number change, we add two geometric distributions to the model to control the magnitude and multitude of possible change. Because we assume for each allele, there is at most one SCNA event, it follows that $n_{Asl} = 1$ or one other value, denoted by n^*_{Al}, and similarly we can define n^*_{Bl}. That is to say, for copy state l, we define n^*_{Al} (or n^*_{Bl}) as the copy number of A (or B) allele after the most recent change. We introduce $c_l := |n^*_{Al} - 1| + |n^*_{Bl} - 1|$ to quantify the magnitude of copy number change and let d_l denote if subclonal allele-specific copy number changes is present. This inspires us to specify $\alpha_l := P(l|\xi, \theta) = \frac{e^{-(\xi c_l + \theta d_l)}}{\sum_{l=1}^{L} e^{-(\xi c_l + \theta d_l)}}$. The parameters ξ and θ can be estimated in our method.

2.1.2 Likelihood model for SPM

Next, we consider the model for somatic point mutation (SPM) data from bulk tissue. Let X_{Nki} and X_{Mki} be the number of sequence reads harboring normal (germline) and mutant allele of the ith $(i = 1, ..., m_k)$ somatic point mutation of the kth segment. Assuming the kth segment belongs to the

sth clone of the lth copy number state, we model X_{Mki} by a mixture of binomial distribution

$$h(X_{Mki}|X_{ki}, \{n_{Asl}\}, \{n_{Bsl}\}, \rho, \epsilon, \{\omega_{lj}\}) = \prod_{i=1}^{m_k} \left[\frac{1}{X_{ki}} \epsilon + \sum_{j=1}^{t_l} f_{\mathscr{B}}(X_{Mki}; X_{ki}, p_{lj}) (1-\epsilon)\omega_{lj} \right], \qquad (2.1)$$

where

$$p_{lj} = \frac{\rho \sum_{s=1}^{S} q_s h_{ljs}(n_{Asl}, n_{Bsl})}{2(1-\rho) + \rho \sum_{s=1}^{S} q_s (n_{Asl} + n_{Bsl})}.$$

Here j indexes four situations: the somatic mutation does not occur, it occurs after SCNA event and thus with copy number 1, or it occurs before SCNA event, and thus with copy number n_{Asl} or n_{Bsl}. We employ indicators $U_{kl} \in \{1, \ldots, t_l\}$ to specify possible patterns of multiplicities of point mutations of all subclones under copy state l. When $U_{kl} = j$, the multiplicity of point mutations can be computed by functions $h_{ljs}(n_{Asl}, n_{Bsl}) \in \{0, 1, n_{Asl}, n_{Bsl}\}$.

The mixture of binomial distributions is intended to be flexible to adapt possible patterns of multiplicities of point mutations. The mixture proportion $\omega_{lj} = \Pr(U_{kl} = j)$ is the probability that a somatic point mutation on segment k belongs to the jth pattern of multiplicities of point mutations assuming copy state l, satisfying $\sum_{j=1}^{t_l} \omega_{lj} = 1$. We also introduce a known error rate ϵ to accommodate circumstances where a somatic point mutation cannot be explained by multiplicities of point mutations predicted according to copy state. In other words, somatic point mutations are uninformative so we need to apply a discrete uniform distribution. For example, a somatic point mutation located on a focal copy number change not captured by joint segmentation cannot provide hints on subclone proportions. We assume $\epsilon = 0.01$. To simplify the notation, we denote $\Theta_{\texttt{SPM}} = (\{\omega_{lj}\})$ and input data $\Psi = (\{\sigma_k^2\}, \{\tau_k^2\}, \bar{\mathbf{r}}, \bar{\mathbf{b}}, \{X_{Mki}\}, \{X_{ki}\})$.

2.1.3 Model specification

We explicitly specify the computational problem to obtain maximum likelihood estimate. We will maximize:

$$\mathscr{L} = \prod_{k=1}^{K} \sum_{l=1}^{L} \left\{ \alpha_l f_{\mathcal{N}} \left(\bar{r}_k \middle| \mu_l, \sigma_k^2 \right) f_{\mathcal{N}} \left(\bar{b}_k \middle| \pi_l, \tau_k^2 \right) \prod_{i=1}^{m_k} \left[\frac{1}{X_{ki}} \epsilon + \sum_{j=1}^{t_l} f_{\mathscr{B}}(X_{Mki}; X_{ki}, p_{lj}) (1-\epsilon)\omega_{lj} \right] \right\}, \tag{2.2}$$

with respect to

$$\boldsymbol{\alpha},\ \rho,\ \phi,\ \boldsymbol{q},\ \{\omega_{lj}\},$$

where

$$f_{\mathcal{N}} \left(\bar{r}_k \middle| \mu_l, \sigma_k^2 \right) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp[-(\bar{r}_k - \mu_l)^2)/(2\sigma_k^2)],$$

$$f_{\mathcal{N}} \left(\bar{b}_k \middle| \pi_l, \tau_k^2 \right) = \frac{1}{\sqrt{2\pi\tau_k^2}} \exp[-(\bar{b}_k - \pi_l)^2)/(2\tau_k^2)],$$

$$p_{lj} = \frac{\rho \sum_{s=1}^{S} q_s h_{ljs}(n_{Asl}, n_{Bsl})}{2(1-\rho) + \rho \sum_{s=1}^{S} q_s (n_{Asl} + n_{Bsl})},$$

$$f_{\mathscr{B}}(X_{Mki}; X_{ki}, p_{lj}) = \binom{X_{ki}}{X_{Mki}} p_{lj}^{X_{Mki}} (1-p_{lj})^{X_{ki}-X_{Mki}},$$

and

$$\mu_l = \log_2 \left[\{2(1-\rho) + \rho(n_{Al} + n_{Bl})\} / \phi \right],$$

$$\pi_l = [1 - \rho + \rho n_{Bl}]/[2 - 2\rho + \rho(n_{Al} + n_{Bl})],$$

$$q_s \geq 0,\ \sum_{s=1}^{S} q_s = 1,$$

$$\alpha_l = \frac{e^{-(\xi c_l + \theta d_l)}}{\sum_{l=1}^{L} e^{-(\xi c_l + \theta d_l)}},\ \alpha_l \geq 0,\ \sum_{l=1}^{L} \alpha_l = 1,$$

$$n_{Al} = \sum_{s=1}^{S} q_s n_{Asl},\ n_{Bl} = \sum_{s=1}^{S} q_s n_{Bsl},$$

$$n_{Al} \geq n_{Bl},$$

$$\omega_{lj} \geq 0,\ \sum_{j=1}^{t_l} \omega_{lj} = 1.$$

As an extension, we derive the model when there are multiple tumor samples with indices $n \in \{1, \ldots, N\}$ sequenced from the same subject:

$$\begin{aligned}\mathscr{L} = \prod_{k=1}^{K} \sum_{l=1}^{L} \Bigg\{ \alpha_l \prod_{n=1}^{N} \Bigg[f_{\mathcal{N}} \left(\bar{r}_{nk} \middle| \mu_{nl}, \sigma_{nk}^2 \right) f_{\mathcal{N}} \left(\bar{b}_{nk} \middle| \pi_{nl}, \tau_{nk}^2 \right) \Bigg] \\ \prod_{i=1}^{m_k} \sum_{j=1}^{t_l} \omega_{lj} \prod_{n=1}^{N} \left[\frac{1}{X_{nki}} \epsilon + (1 - \epsilon) f_{\mathscr{B}}(X_{Mnki}; X_{nki}, p_{nlj}) \right] \Bigg\}.\end{aligned} \tag{2.3}$$

2.1.4 Choice of n_{Asl} and n_{Bsl}

One intuitive approach to maximize the likelihood function is to choose the number of L and initial values of n_{Asl} and n_{Bsl} through a joint clustering of segmental LRR and BAF, and then devise a way to update the variables in each iteration. However, this leads to a very challenging integer programing problem since n_{Asl} and n_{Bsl} take integer values. Our solution to this problem is to choose fixed n_{Asl} and n_{Bsl}. Note that n_{Asl} and n_{Bsl} can be expressed in a single matrix if we consider each of them as a matrix and concatenate the two matrices by rows:

$$\begin{bmatrix} n_{A11} & n_{A12} & n_{A13} & \cdots & n_{A1L} \\ n_{A21} & n_{A22} & n_{A23} & \cdots & n_{A2L} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_{AS1} & n_{AS2} & n_{AS3} & \cdots & n_{ASL} \\ \hline n_{B11} & n_{B12} & n_{B13} & \cdots & n_{B1L} \\ n_{B21} & n_{B22} & n_{B23} & \cdots & n_{B2L} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_{BS1} & n_{BS2} & n_{BS3} & \cdots & n_{BSL} \end{bmatrix}$$

Each column of the matrix above corresponds to a copy number state by which segments can be grouped. Each row of the matrix corresponds to each subclone restricted to either the A allele (upper half matrix) or B allele (lower half matrix). $2S$ integer subclonal allele-specific copy numbers jointly determine a copy state.

Suppose $n_{\max}$ is a known upper bound of subclonal total copy number $n_{Asl} + n_{Bsl}$ so that their possible values are $\{0, 1, \ldots, n_{\max}\}$. When we consider all possible combinations, except when $S = 1$, i.e., when the tumor cell population is homogeneous, the exhaustive approach would be computationally intractable.

Therefore, to reduce the number of copy states to a manageable level, we assume for each segment, the allele-specific copy number of one allele can be altered at most once. To formalize, consider a binary phylogenetic tree of tumor cell evolution. Normal cells are defined to be the root in the binary tree, and for notation simplicity they are defined as subclone 0, with $n_{A0l} = n_{B0l} = 1$. At most one edge connecting any two subclones r and s in the phylogenetic tree is allowed to have $n_{Arl} \neq n_{Asl}$, and at most one edge connecting any two subclones r and s is allowed to have $n_{Brl} \neq n_{Bsl}$.

Furthermore, since the bounds $[0, 0.5]$ of BAF imply $n_{Al} \geq n_{Bl}$, we only include copy states which satisfy $n_{Asl} > n_{Bsl}$ for some s, or $n_{Asl} = n_{Bsl}$ for all $s \in \{1, \ldots, S\}$. During the optimization, we enforce $n_{Al} \geq n_{Bl}$ by swapping $\{n_{Asl} : s \in \{1, \ldots, S\}\}$ with $\{n_{Bsl} : s \in \{1, \ldots, S\}\}$ when for some l, $n_{Al} < n_{Bl}$.

Under the assumptions, L exhibits quadratic growth with respect to $n_{\max}$ and exhibits exponential growth with respect to S. In practice, $S = 5$ would be more than adequate for sequencing data without an ultra high coverage.

2.2 Optimization

2.2.1 EM algorithm to calculate the profile likelihood

We need to optimize:

$$\sum_{k=1}^{K} \log \left\{ \sum_{l=1}^{L} \left[a_{kl}(\xi, \theta) \prod_{i=1}^{m_k} \left(\sum_{j=1}^{t_l+1} f_{kilj} \omega_{lj} \right) \right] \right\}$$

where

$$a_{kl}(\xi, \theta) = \frac{e^{-(\xi c_l + \theta d_l)}}{\sum_{l=1}^{L} e^{-(\xi c_l + \theta d_l)}} f_{\mathcal{N}} \left(\bar{r}_k \middle| \mu_l, \sigma_k^2 \right) f_{\mathcal{N}} \left(\bar{b}_k \middle| \pi_l, \tau_k^2 \right),$$

and

$$f_{kilj} = f_{\mathcal{B}}(X_{Mki}; X_{ki}, p_{lj}) (1 - \epsilon) \text{ for } j \in \{1 \dots t_l\}, \quad f_{kil(t_l+1)} = \frac{1}{X_{ki}} \epsilon,$$

and for notation simplicity, we let $\omega_{l(t_l+1)} = 1$ for all l.

Under the multi-sample scenario, we need to optimize:

$$\sum_{k=1}^{K} \log \left\{ \sum_{l=1}^{L} \left[a_{kl}(\xi, \theta) \prod_{n=1}^{N} \prod_{i=1}^{m_k} \left(\sum_{j=1}^{t_l+1} f_{nkilj} \omega_{lj} \right) \right] \right\}$$

where

$$a_{kl}(\xi, \theta) = \frac{e^{-(\xi c_l + \theta d_l)}}{\sum_{l=1}^{L} e^{-(\xi c_l + \theta d_l)}} \prod_{n=1}^{N} f_{\mathcal{N}} \left(\bar{r}_{nk} \middle| \mu_{nl}, \sigma_{nk}^2 \right) f_{\mathcal{N}} \left(\bar{b}_{nk} \middle| \pi_{nl}, \tau_{nk}^2 \right),$$

and

$$f_{nkilj} = f_{\mathcal{B}}(X_{Mnki}; X_{nki}, p_{nlj}) (1 - \epsilon) \text{ for } j \in \{1 \dots t_l\}, \quad f_{nkil(t_l+1)} = \frac{1}{X_{nki}} \epsilon.$$

We iterate between the step to update ω_{lj}, and the step to update ξ and θ.

The step to update ω_{lj}:

We maximize the following Q-function with respect to $\{\omega_{lj}\}$:

$$Q(\boldsymbol{\omega}; \boldsymbol{\omega}^{(m)}) = \sum_{l=1}^{L} \sum_{k=1}^{K} \gamma_{kl}^{(m)} \left[\sum_{i=1}^{m_k} \log \left(\sum_{j=1}^{t_l+1} f_{kilj} \omega_{lj} \right) \right], \quad \text{s.t.} \sum_{j=1}^{t_l} \omega_{lj} = 1, \ \forall l \in \{1 \dots t_l\},$$

where in the $(m+1)$th iteration, the posterior probability of the kth segment under copy state $l \in \{1, \dots, L\}$ is given by:

$$\begin{aligned}\gamma_{kl} &\coloneqq \Pr\left(z_k = l \middle| \bar{r}_k, \bar{b}_k, \alpha_l^{(m)}, \beta_{kl}^{(m)}\right)\\ &= \frac{p_{\mathcal{N}}\left(\bar{r}_k \middle| \mu_l, \sigma_k^2\right) p_{\mathcal{N}}\left(\bar{b}_k \middle| \pi_l, \tau_k^2\right) \alpha_l^{(m)} \beta_{kl}^{(m)}}{\sum_{l=1}^{L} p_{\mathcal{N}}\left(\bar{r}_k \middle| \mu_l, \sigma_k^2\right) p_{\mathcal{N}}\left(\bar{b}_k \middle| \pi_l, \tau_k^2\right) \alpha_l^{(m)} \beta_{kl}^{(m)}}\\ &= \frac{a_{kl}(\xi,\theta) \prod_{i=1}^{m_k} \left(\sum_{j=1}^{t_l+1} f_{kilj}\omega_{lj}\right)}{\sum_{l=1}^{L} a_{kl}(\xi,\theta) \prod_{i=1}^{m_k} \left(\sum_{j=1}^{t_l+1} f_{kilj}\omega_{lj}\right)},\end{aligned} \tag{2.4}$$

where unobserved variable $z_k = l$ if and only if the kth segment is under copy state l.

Fortunately, this optimization problem is separable. For every l, we maximize:

$$Q(\boldsymbol{\omega}_{l\cdot}; \boldsymbol{\omega}_{l\cdot}^{(m)}) = \sum_{k=1}^{K} \gamma_{kl}^{(m)} \left[\sum_{i=1}^{m_k} \log\left(\sum_{j=1}^{t_l+1} f_{kilj}\omega_{lj}\right)\right], \quad \text{s.t.} \sum_{j=1}^{t_l} \omega_{lj} = 1.$$

In the step to update ξ and θ, we get $\alpha_l^{(m+1)}$ using Newton or quasi-Newton methods.

2.2.2 Optimization of the profile likelihood

To maximize the likelihood, we simplify by maximizing the profile likelihood with respect to ρ, ϕ and $\boldsymbol{q}$:

$$\max_{\boldsymbol{\alpha}, \{\omega_{lj}\}} \mathscr{L}(\boldsymbol{\alpha}, \rho, \phi, \boldsymbol{q}, \{\omega_{lj}\}),$$

since the profile likelihood is easy to calculate using the EM algorithm explained in the previous section.

Since the profile likelihood is continuous, we recommend the global and gradient-free optimization algorithm `NLOPT_GN_DIRECT_L` (Gablonsky and Kelley, 2001) wrapped in R package `nloptr` (Johnson, 2014). We use 500 as the maximum number of the profile likelihood evaluation and bound $\rho \in [10^{-5}, 1 - 10^{-5}]$, $\phi \in [1, 5]$, and $r_s \in [10^{-5}, 1]$. The bounds on the ploidy can be justified by the ploidy range reported in the literature.

After the optimization, we can calculate an alternate version of ploidy $\tilde{\phi}$ by

$$\tilde{\phi} = 2(1-\rho) + \rho \sum_{k=1}^{K} w_k \cdot \{\text{the most probable } (n_{Al} + n_{Bl}) \text{ for } k\text{th segment}\}.$$

In whole exome sequencing, the w_k's are the proportion of genes, making ϕ the ploidy of the tumor sample restricted to the exome. This makes sense because we don't have any information for the non-coding regions. $\tilde{\phi}$ is arguably a more robust version of ploidy than the model-based ϕ since it is based on estimated copy numbers instead of how much LRR has shifted away from 0, which can be influenced by read-depth corrections.

2.3 Model selection

The number of subclones, the evolution tree structure, and the maximum total copy number need to be predetermined before we optimize the profile likelihood. We use BIC to perform model selection:

$$-2\log\mathscr{L} + [(S-1)N+4]\log(2NK) + [(S-1)N+S+4]\log\left(1 + N\sum_{k=1}^{K} m_k\right),$$

where $2NK$ is the number of observations providing copy number information, $N\sum_{k=1}^{K} m_k$ is the number of observations providing information from somatic point mutations, $(S-1)N+4$ is the number of free parameters attributed to $\boldsymbol{q}$, $\boldsymbol{\alpha}$, ρ and ϕ, and somatic point mutations belonging to the lth copy state has additional degrees of freedom of $t_l - 1$, approximately S, from the mixing proportions ω_{lj}.

In practice, as the number of subclones is unknown, we iterate through a range of number of subclones, e.g., from one to five. Usually BIC will first decrease and then increase with the number of subclones. We record the number of subclones that will give us the lowest BIC. Then the problem boils down to determining BIC given number of subclones, where we go through all possible evolution tree structures, and record the tree structure that will give us the lowest BIC. Now we arrive at the innermost layer of the model selection problem. To determine BIC given number of subclones and evolution tree structure, we pick a moderately large number, say 10, as the maximum total copy number $n_{\max}$ so that the model is complex enough to explain a majority segments and simple enough so that optimization speed is not compromised. We maximize the profile likelihood to obtain the optimal ρ, ϕ and $\boldsymbol{q}$. After that, we expand the copy number states by setting $n_{\max}$ to a large number, say 30, and run the EM algorithm to get updated $\boldsymbol{\alpha}$, ω_{lj}, σ^2 and τ^2.

2.4 Simulation

We use PhyloWGS to compare with our method, SHARE, under three scenarios:

1. Samples have a large number of subclonal copy number aberrations;
2. Samples have a large number of copy number aberrations but no subclonal copy number aberrations;
3. Samples have no copy number aberration at all.

In PhyloWGS, different SCNAs can have different cellular prevalences. This is similar to our approach, but they only allow one SCNA on each segment. Their program accepts (integer) allele-specific copy numbers and their respective cellular prevalences obtained from other software.

We compare our method with PhyloWGS by simulating multiple batches of 500 samples with 250 somatic point mutations, with varying levels of subclonal copy number aberration in each batch. Since PhyloWGS requires integer-valued copy number of tumor cells and cellular proportions of SCNAs as part of its input, these values are computed using ASCAT, a copy number calling method that accounts for purity, modified to get input in the form of simulated segmented LRR and BAF. Using ASCAT to provide copy numbers essentially restricts PhyloWGS from considering subclonal SCNA. In fact, methods such as Battenberg (Nik-Zainal et al., 2012), FACETS (Shen and Seshan, 2016) and TITAN (Ha et al., 2014) are all equipped with some functionalities of calling subclonal SCNA without the help from SPM, but there is no consensus that they are better than ASCAT on exome-seq data (Shen and Seshan, 2016; Van Loo, 2018).

The result of SHARE in Fig. 2.4 and result of PhyloWGS in Fig. 2.5 are very similar when there is only clonal CNA. When there is subclonal CNA (subfigures b and e), since ASCAT does not model subclonal CNA, this will affect the estimation of purity; ASCAT favors a solution with a higher ploidy even though consideration of subclonal CNA could be a better fit (subfigures a and d). Moreover, when there is no copy number change, ASCAT always gives 1 as the purity estimate. Since PhyloWGS does not have a native way to estimate purity, we resort to use the cellular proportion of all subclones inferred by PhyloWGS as its purity estimate. The estimation of purity in PhyloWGS is less accurate compared to SHARE in the simulations (subfigures c and f).

The simulation results show that modeling SCNA and SPM together can help to resolve some of the identifiability issues of subclonal SCNA and SPM at the expense of a more complicated model and a longer running time.

2.5 Results

2.5.1 TCGA COAD samples

We apply our model to 400 TCGA COAD whole-exome sequencing samples. The details of processing the data are described in the Online Supplementary Materials.

First, we provide a case study of sample TCGA-A6-5656, one of the 400 TCGA COAD samples, in Fig. 2.6. Using our method, the estimated purity is 0.816 and the estimated ploidy is 2.023. Three subclones are identified, forming a linear clonal tree with cellular proportions of 0.444, 0.446 and 0.110, respectively.

Fig. 2.6 (a) illustrates gene-level Log R Ratio (LRR), segmental LRR and model-predicted segmental LRR. In obtaining LRR, the normal sample that can generate most signals in LRR (“best match”) is chosen instead of the matched normal sample. Fig. 2.6 (b) illustrates gene-level B Allele Frequency (BAF), segmental BAF, and model-predicted segmental BAF. The agreement between segmental and model-predicted values demonstrates that except for focal copy number changes, our model successfully captures most of copy number alterations.

Fig. 2.6 (c) illustrates mutant allele frequency along the genome. A strength of our model is the incorporation of somatic point mutations, which can crack open the door of intra-tumor heterogeneity not characterized by somatic copy number alterations. From the figure, most of the observed mutant allele frequency can be explained by possible levels of mutant allele frequency predicted by our fitted model.

Fig. 2.6 (d), (e) and (f) show the total and allele-specific copy number in the subclones. SHARE picks the best model as the model with 3 subclones and a linear clonal tree structure using BIC. The majority of copy number alterations takes place before formation of the founding subclone, which favors the hypothesis that copy number alteration is a driving force in tumor progression. Additional copy number alterations take place on chromosome 7 and 13 in subclone 3. Our model can also recover the timing of somatic point mutations.

After displaying details about the result of one colon cancer sample, we now describe the result of the joint analysis of 400 TCGA colon cancer samples.

Fig. 2.7 provides a summary of results of the COAD samples. In Fig. 2.7 (a), the majority of samples have a purity from 0.4 to 0.9 and a ploidy from 2 to 3. The estimated distribution of ploidy ϕ justifies a lower bound of 1 and an upper bound of 5 imposed on ploidy during optimization. We also provide an alternative measure of ploidy $\tilde{\phi}$ in Fig. 2.7 (b), which is, simply put, a weighted average of the segmental copy numbers. This measurement is more robust than the model-based ploidy estimate since it does not rely on read-depth corrections in obtaining LRRs. The alternative measurement of ploidy suggests that the majority of samples have a ploidy from 2 to 4, which is consistent with a tumor history experiencing either no whole-genome doubling, or one whole-genome doubling with subsequent deletions.

Fig. 2.7 (c) summarizes the distribution of estimated number of subclones and clonal tree structure index of the TCGA COAD samples being investigated. The maximum number of subclones considered in the experiment is 5. We see the prevalence of intra-tumor heterogeneity from the overwhelming number of samples with >1 number of subclones. Inferred from the data, the most frequent number of subclones is 3. For the estimated clonal tree structure, we see that nearly a half of the samples fall into the “tree structure index = 1” category, suggesting that these tumor samples may have undergone a linear evolution.

Fig. 2.7 (d) illustrates tumor entropy, defined as $-\sum_{s=1}^{S} q_s \log q_s$, as a measure of heterogeneity of tumor to predict patient survival (Jiang et al., 2014). The violin plot implies that tumor entropy is a better measurement than the number of subclones to characterize heterogeneity, as there is some overlap of entropy distribution between tumor samples with different number of subclones.

Figs. 2.7 (e) and (f) constitute a comparison of SHARE results with ABSOLUTE results reported in a pan-cancer study (Zack et al., 2013). Despite the fact that ABSOLUTE takes processed data from SNP 6.0 array while our SHARE takes whole exome sequencing data, the comparison reveals a very similar ploidy estimate for a majority of samples. A comparison of SHARE purity and ABSOLUTE purity also reveals similar estimates for a majority of samples. ABSOLUTE has a greater tendency to provide an estimate with purity = 1, the unlikely case that the bulk tumor consists of pure tumor cells, as ABSOLUTE is not blessed with the information from point mutations. We conclude that while SHARE can provide ploidy and purity estimates generally consistent with ABSOLUTE results, it has the advantage of using whole exome sequencing data to explicitly model subclones and provide subclonal copy number estimates.

As a comparison to SHARE entropy estimates, we also derive entropy estimates from PhyloWGS in a similar way as we use it to obtain entropy estimates from simulation data. While the purity estimates from SHARE and ASCAT already start to differ in Fig. 2.8 (a), the entropy estimates from SHARE and PhyloWGS are more different, probably since in a two-step approach such as ASCAT plus PhyloWGS, we tend to overestimate purity and underestimate entropy when subclonal SCNA is not accounted for.

We investigate whether entropy is associated with cellular proportion of immune cells derived from bulk RNA-seq data using CIBERSORT (Newman et al., 2015). Full results are reported in Supplementary Materials. The top hits among immune cell types to correlate with SHARE entropy are CD8+ T cells and M1 Macrophages (Figs. 2.8 (c) and (d)), which can participate in immune surveillance. When there is a higher proportion of CD8+ T cells and M1 Macrophages, more neoantigens are suppressed. Therefore, intra-tumor heterogeneity measured by entropy is decreased. Entropy estimates from PhyloWGS (Figs. 2.8 (e) and (f)), however, does not have a significant correlation with cellular proportion estimates to support this hypothesis.

We also fit a Cox proportional hazards model to investigate whether entropy, either from SHARE or PhyloWGS, is associated with patient survival. Since PhyloWGS fails to provide an entropy estimate in a few samples, there are slightly fewer observations. Age, tumor stage, mutation burden, entropy and entropy by mutation burden are included as covariates in the model. We use a likelihood ratio test to see if entropy is significantly associated with survival. After eliminating the entropy and entropy by mutation burden terms and re-fitting the model, we take the two fold of the likelihood difference between the full model and the reduced model. Using PhyloWGS entropy, the likelihood ratio statistic is 5.72 and we have a p-value of 0.057 under a χ^2 distribution with 2 degrees of freedom. As a comparison, using SHARE entropy, the likelihood ratio statistic is 7.45 and we have a p-value of 0.024 under a χ^2 distribution with 2 degrees of freedom. This analysis suggests that SHARE entropy is more predictive of patient survival than PhyloWGS entropy.

2.5.2 Hugo et al. Melanoma Anti-PD-1 Response Data

Hugo et al. (Hugo et al., 2016) reported tumor response after anti-PD-1 treatment in melanoma patients. It has been suggested that clonal mutation burden (McGranahan et al., 2016), defined as log10 number of somatic point mutations that appear in more than 99% of the tumor cells, could be a better predictor of immunotherapy response. To compare SHARE and Canopy, we use each of them on whole-exome sequencing data of pre-treatment melanoma samples to reconstruct tumor phylogeny using a resample of somatic point mutations. Then we infer which somatic point mutations are clonal and calculate clonal mutation burden using all somatic point mutations, the number of which can be as high as tens of thousands.

We have decided to remove SRR4289744 from our analysis. There is almost no copy number aberration in this sample and the VAF of most somatic point mutations are all very low, which suggests that it is very difficult to infer intra-tumor heterogeneity of this sample.

We now have three measures of neoantigen quantity: mutation burden (Fig. 2.9 (a)), SHARE mutation burden (Fig. 2.9 (b)), and PhyloWGS mutation burden (Fig. 2.9 (c)). Uisng a two-sided Jonckeheere-Terpstra test (Jonckheere, 1954; TERPSTRA, 1952), we test for whether there is a significant trend among three three response groups for each definition of mutation burden. The results show that using SHARE, we can define a clonal mutation burden that associates with tumor response better than PhyloWGS mutation burden, which is actually slightly worse than the vanilla version of mutation burden in terms of the association.

2.6 *Tables and figures*

Property	EXPANDS	CHAT	BayClone2	CloneHD	PhyloWGS	PyClone	Canopy	SHARE
Models subclonal SCNA	+	+	+	+	+	−	+	+
Constructs phylogeny	−	−	−	−	+	−	+	+
Considers the order of SCNA and SPM	−	+	−	−	+	+	+	+
Does not require known CN estimates	+	+	+	+	−	−	−	+
Provides cellular frequency of subclones	−	−	+	+	+	−	+	+

Table 2.1: Overview of some subclone reconstruction methods that jointly infer from SCNA (somatic copy number alterations) and SPM (somatic point mutations): EXPANDS (Andor et al., 2014), CHAT (Li and Li, 2014), BayClone2 (Lee et al., 2016), CloneHD (Fischer et al., 2014), PhyloWGS (Deshwar et al., 2015), PyClone (Roth et al., 2014), Canopy (Jiang et al., 2016), and SHARE (our proposed method).

Covariate	Est.	p-value
Age	0.03	1.3e-3
Stage II vs. I	0.45	0.36
Stage III vs. I	1.37	5.0e-3
Stage IV vs. I	2.42	1.2e-6
MB	−1.56	4.5e-2
SHARE entropy	5.58	7.2e-3
SHARE entropy by MB	2.16	7.1e-3

Table 2.2: Survival analysis of TCGA colon cancer patients using Cox proportional hazards model. This is the model with SHARE entropy as a covariate. The number of observations is 386 and the number of events is 86, with 11 observations deleted due to missingness. The log-likelihood of the model with SHARE entropy is −414.30. MB: Mutation Burden defined as total number of somatic point mutations in log10 scale.

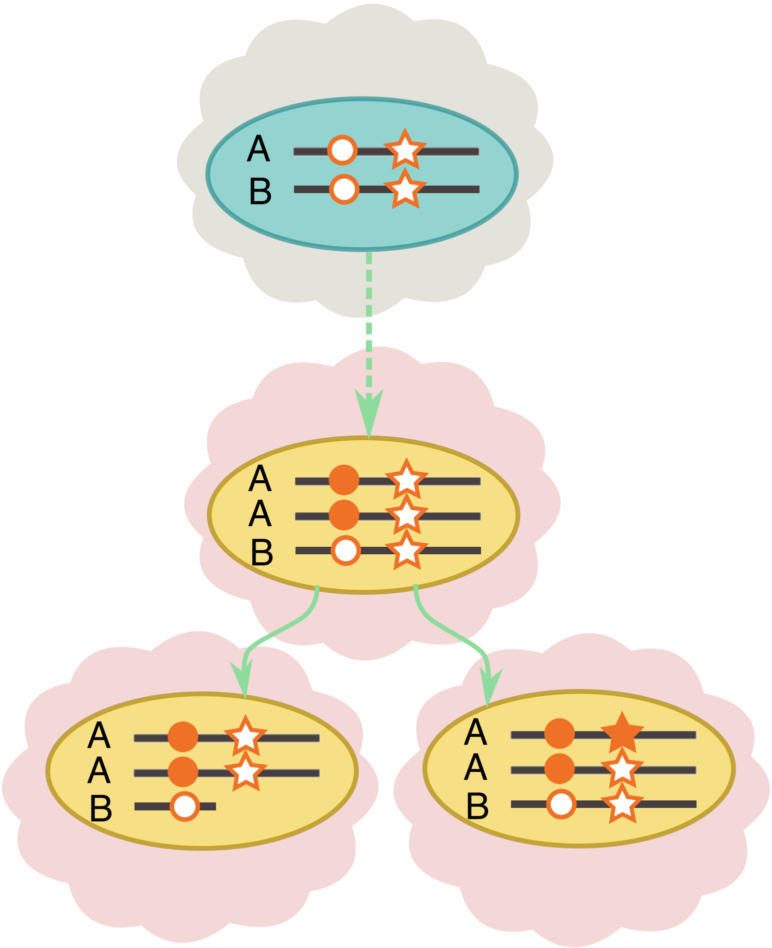


Figure 2.1: A fabricated example of clonal evolution tree. A cell in a homogeneous aggregate is represented by an eclipse in a cloud. The gray cloud indicates the normal cells in the bulk tumor while the pink clouds indicate tumor subclones. Circles and stars indicate genomic loci; hollow shapes are for reference alleles and filled shapes are for mutant alleles. In the founding subclone, the locus represented by circle is mutated and the *A* allele is amplified. Two subclones are derived from the founding subclone, one by a deletion on a segment of the *B* allele and the other one by a mutation on the *A* allele.

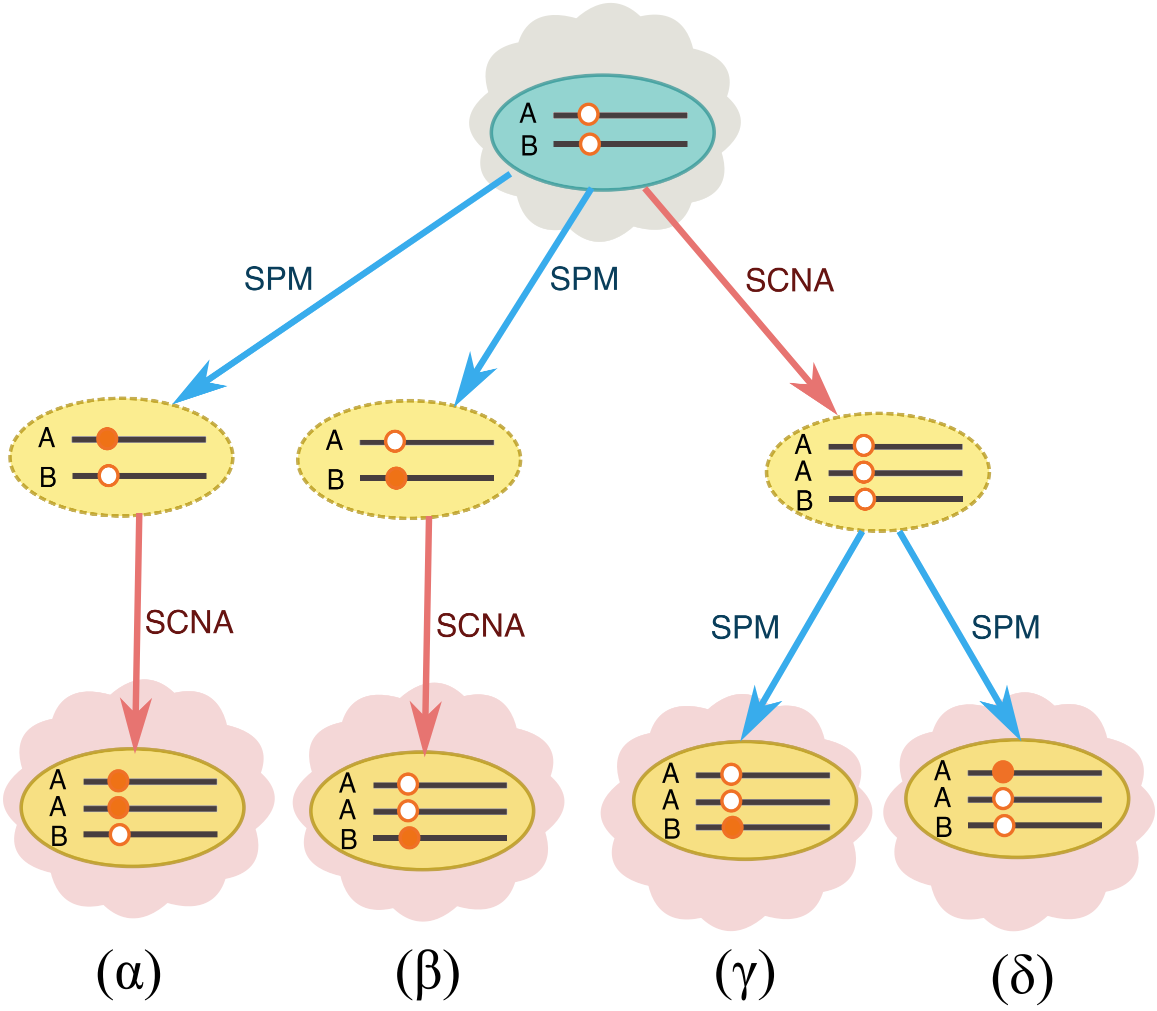


Figure 2.2: Multiplicity of point mutations are complicated by the order of somatic point mutation (SPM) and somatic copy number alteration (SCNA). If we already know that at a genomic location, SPM happens and SCNA affects the genome by doubling A allele, timing and the allele where the point mutation is located jointly determine the multiplicity of point mutation. In our model, (α) can be distinguished from the other three scenarios by multiplicity of point mutation. (δ) can be distinguished from (β) and (γ) through informative reads that harbor both somatic variants and phased germline variants (Chedom-Fotso et al., 2016) (not provided in our model). The scenarios (β) and (γ) are indistinguishable. Cells marked by eclipse with dashed borders are transient and cannot be detected in bulk cells. At any given genomic location, only one of the four scenarios can be the underlying mechanism.

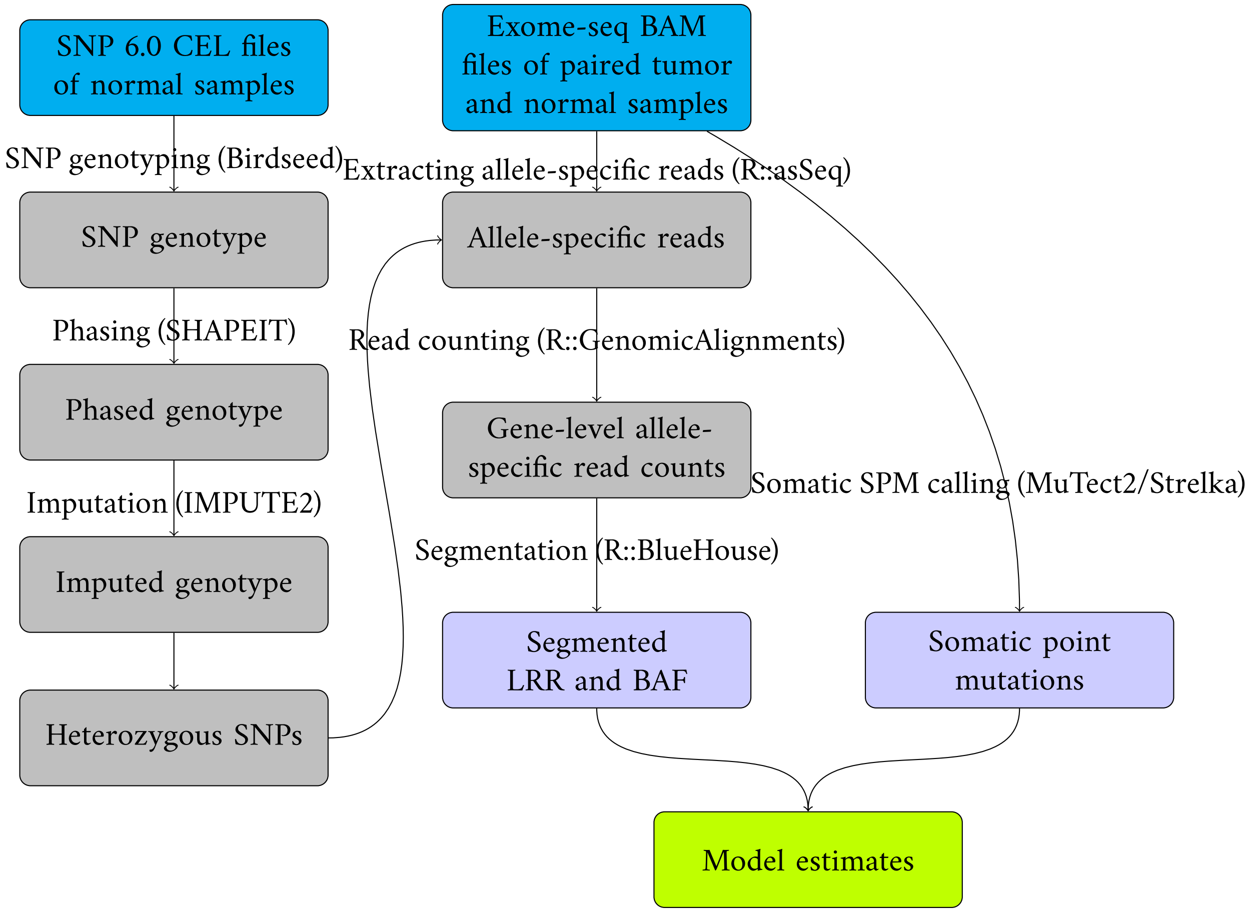
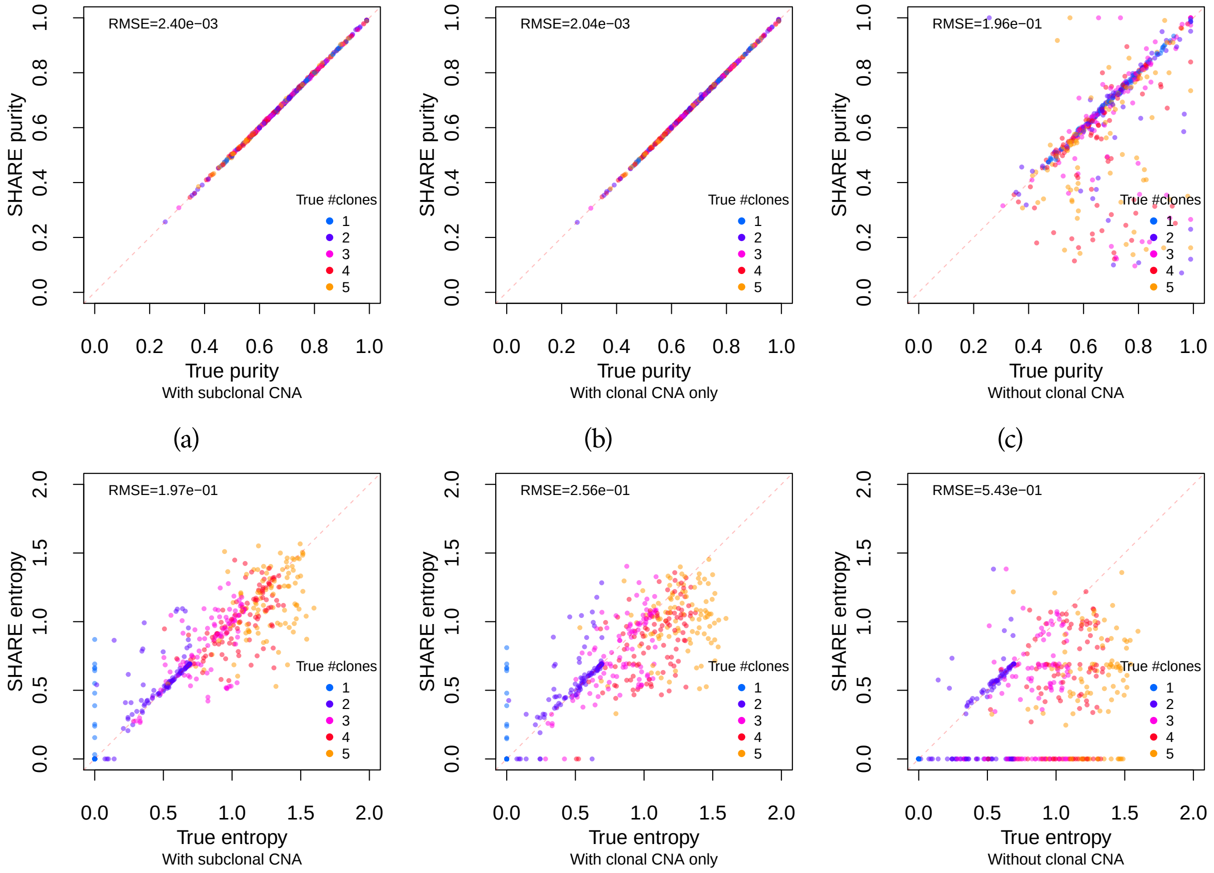

SNP 6.0 CEL files of normal samples
Exome-seq BAM files of paired tumor and normal samples
SNP genotyping (Birdseed)
Extracting allele-specific reads (R::asSeq)
SNP genotype
Allele-specific reads
Phasing (SHAPEIT)
Read counting (R::GenomicAlignments)
Phased genotype
Gene-level allele-specific read counts
Somatic SPM calling (MuTect2/Strelka)
Imputation (IMPUTE2)
Segmentation (R::BlueHouse)
Imputed genotype
Segmented LRR and BAF
Somatic point mutations
Heterozygous SNPs
Model estimates

Figure 2.3: A schematic diagram of pipeline to prepare the input data for our model. Rounded rectangles represent data and arrows represent data processing steps. Here we obtained heterozygous SNPs from SNP array data. If SNP array data are not available, GATK has an established pipeline to perform germline variant calling using whole exome sequencing data.

Covariate	Est.	p-value
Age	0.03	1.3e-3
Stage II vs. I	0.48	0.34
Stage III vs. I	1.35	5.6e-3
Stage IV vs. I	2.40	1.4e-6
MB	2.25	1.1e-2
PhyloWGS entropy	4.76	5.2e-2
PhyloWGS entropy by MB	−2.00	3.2e-2

Table 2.3: Survival analysis of TCGA colon cancer patients using Cox proportional hazards model. This is the model with PhyloWGS entropy as a covariate. The number of observations is 379 and the number of events is 84, with 11 observations deleted due to missingness. The log-likelihood of the model with SHARE entropy is -403.19. MB: Mutation Burden defined as total number of somatic point mutations in log10 scale.



(a)

(b)

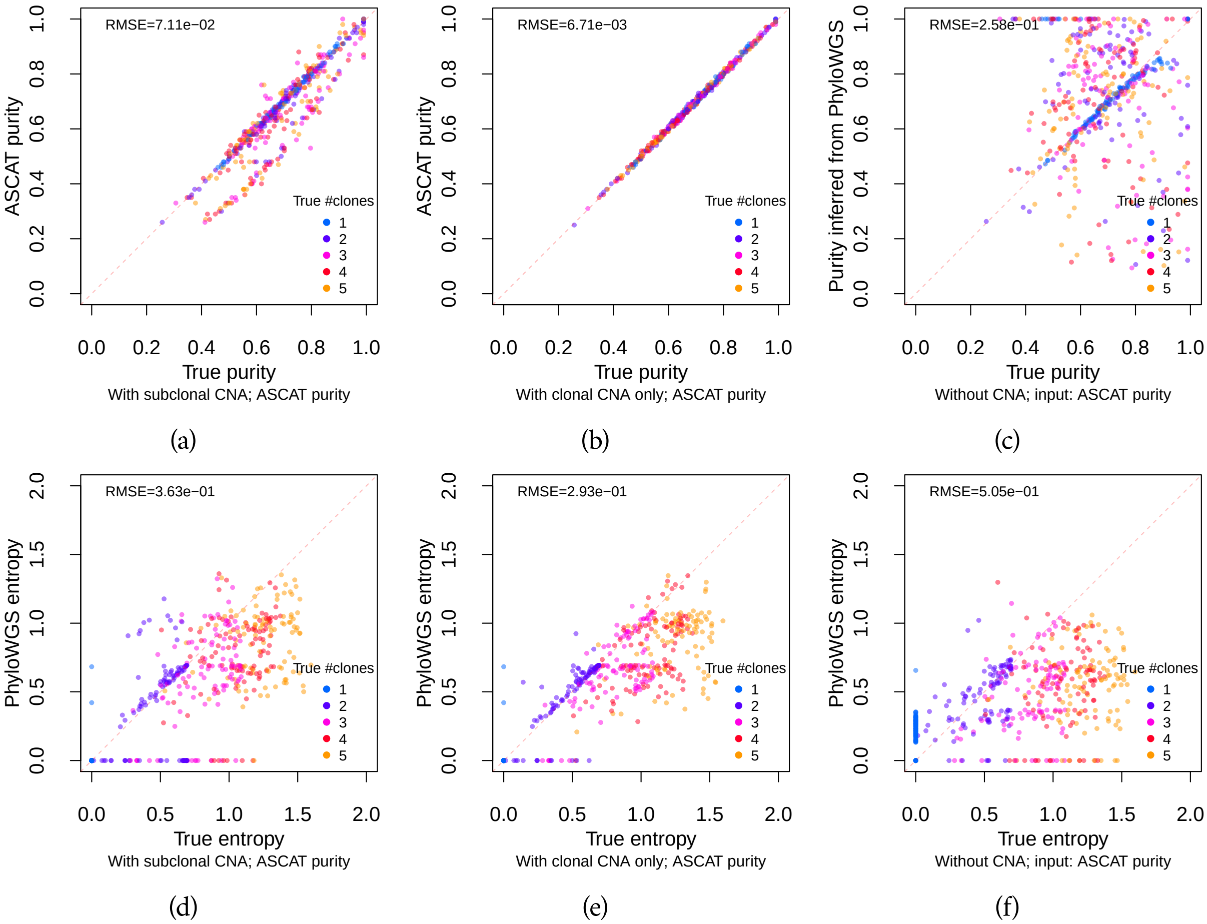
(c)

(d)

(e)

(f)

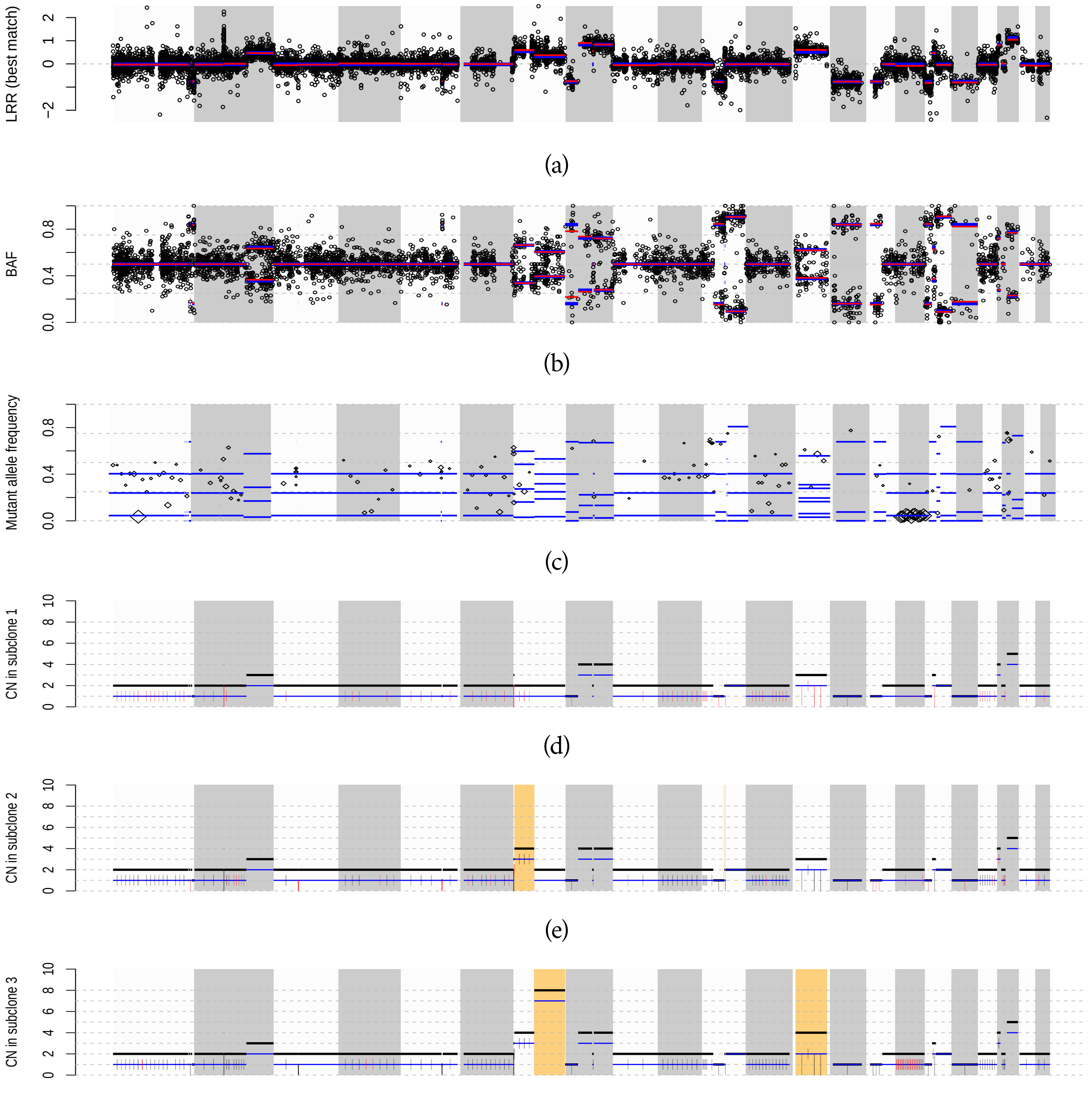
Figure 2.4: Results of simulations on SHARE. (a) (d) SHARE purity and entropy compared to true values on 500 simulations with subclonal CNA. (b) (e) SHARE purity and entropy compared to true values on 500 simulations with clonal CNA only. (c) (f) SHARE purity and entropy compared to true values on 500 simulations without clonal CNA.



(a) (b) (c)

(d) (e) (f)

Figure 2.5: Results of simulations on PhyloWGS. Purity and copy number estimates are taken from ASCAT given segmented LRR and BAF. (a) (d) ASCAT purity and PhyloWGS entropy compared to true values on 500 simulations with subclonal CNA. (b) (e) ASCAT purity and PhyloWGS entropy compared to true values on 500 simulations with clonal CNA only. (c) (f) PhyloWGS purity and entropy compared to true values on 500 simulations without clonal CNA. Since ASCAT will not give valid purity estimates when there is no copy number change, the purity is the proportion of the founding subclone in PhyloWGS result and entropy is calculated from the proportion of other subclones.



(a)

(b)

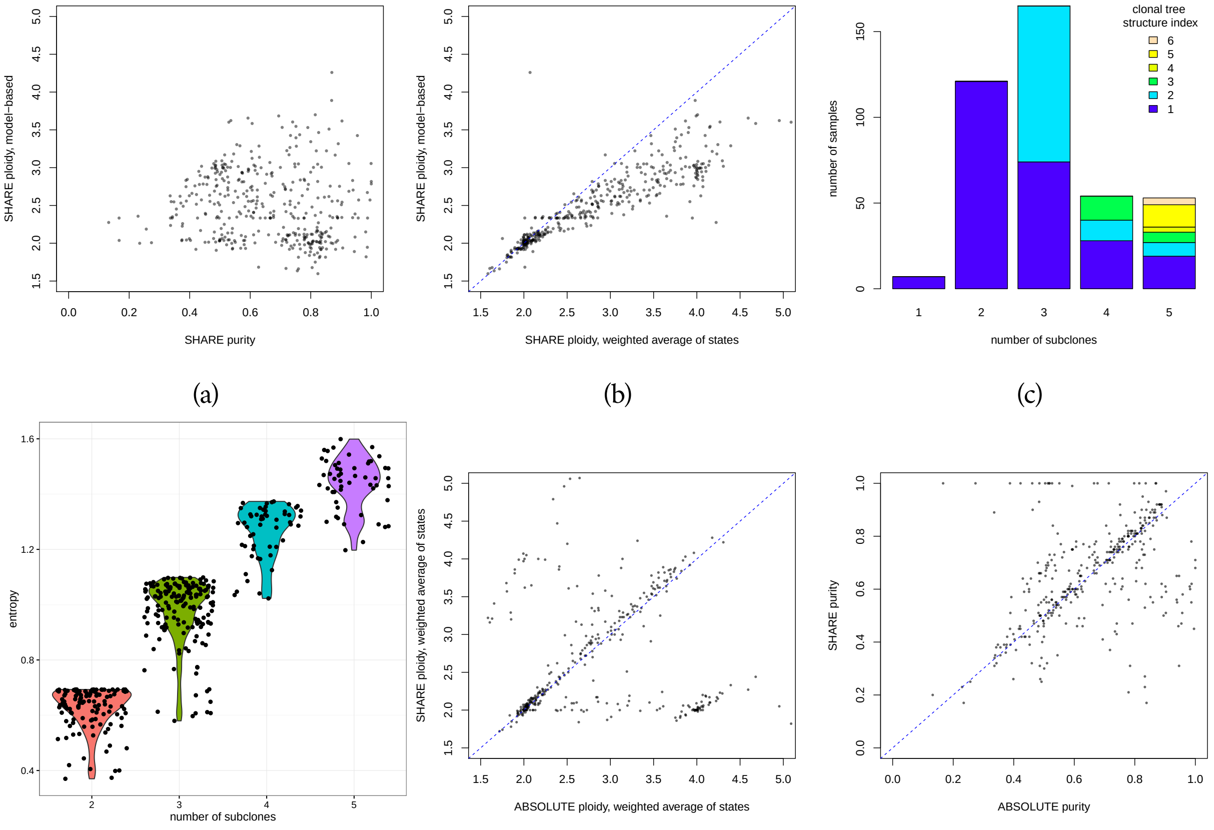
(c)

(d)

(e)

(f)

Figure 2.6: Results of SHARE on sample TCGA-A6-5656. (a) Log R Ratio (LRR). Each point represents a gene. Red segments represent segmental LRRs taken as the input in our model. Blue segments represent predicted segmental LRRs based on the output of our model. (b) B Allele Frequency (BAF). Each point represents a gene. Red segments represent segmental BAFs taken as the input in our model. Blue segments represent predicted segmental BAFs based on the output of our model. (c) Mutant allele frequency. Bigger diamonds indicate a somatic point mutation with higher read depth. Blue segments represent predicted segmental levels of mutant allele frequency based on the output of our model. (d), (e), and (f) show estimated copy number (CN) of all the subclones in our model. Black, thick segments represent total copy number while blue, thin segments represent allele-specific copy number of *A* allele. The estimated purity is 0.816 and the estimated ploidy is 2.023. Three subclones are identified, forming a linear clonal tree with cellular proportions of 0.444, 0.446 and 0.110, respectively.



(a)

(b)

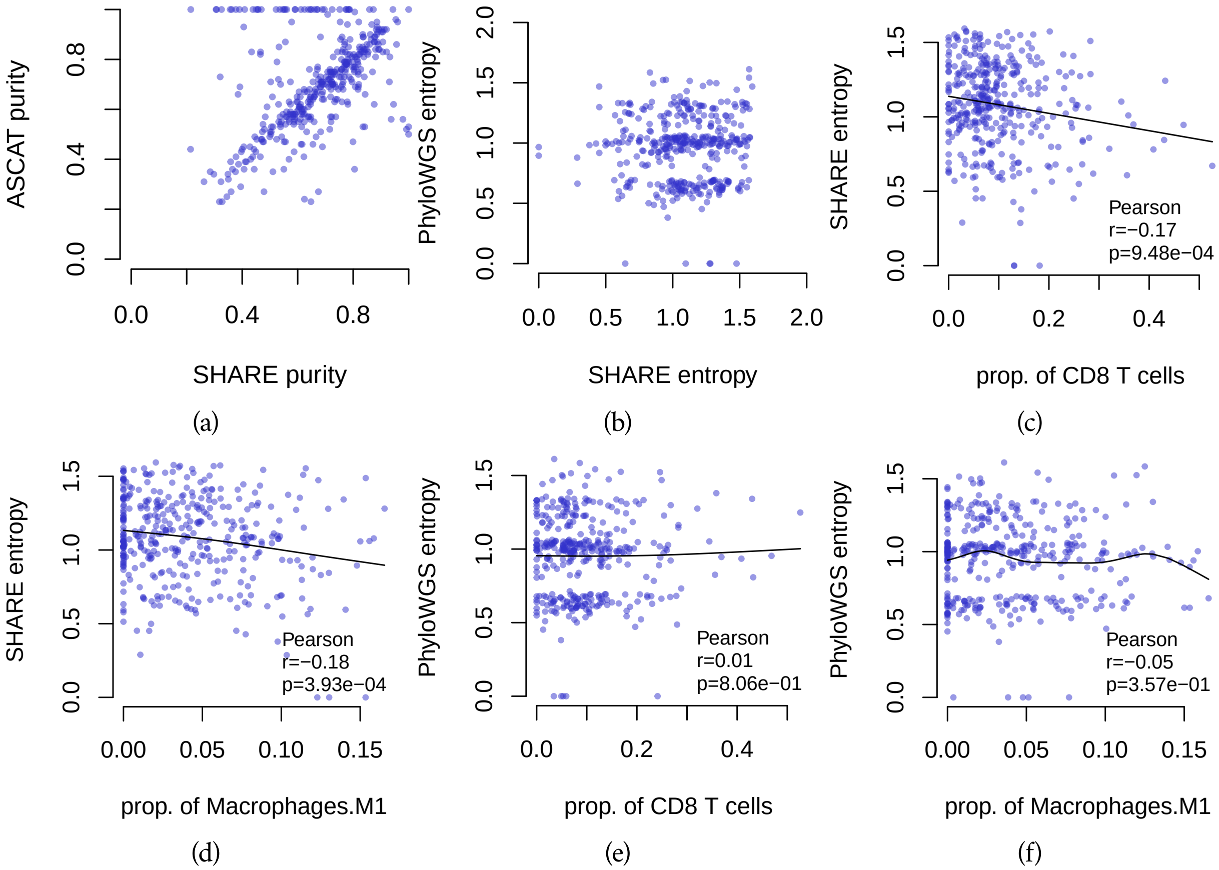
(c)

(d)

(e)

(f)

Figure 2.7: Results of SHARE on TCGA COAD samples. (a) Scatterplot of model-based ploidy of all cells (y axis) and estimated tumor purity (x axis) using SHARE. Each point indicates a tumor sample. (b) Model-based SHARE ploidy of all cells (y axis) compared with SHARE ploidy calculated by taking a weighted average of copy number states (x axis), which is an alternative version. (c) Barplots of estimated number of subclones. Samples with the same estimated number of subclones but different clonal tree structure index are stacked together and colored differently. Grey color stands for linear evolution tree; for other clonal tree structures, see Supplementary Materials. (d) Violin plots of tumor entropy. Each point indicates a tumor sample. Clonal tumor samples, whose entropy equal to zero, are not displayed. (e) A comparison of estimated tumor purity using SHARE (y axis) and estimated tumor purity using ABSOLUTE (x axis). (f) A comparison of estimated tumor ploidy using SHARE (y axis) and estimated tumor ploidy using ABSOLUTE (x axis).



(a)

(b)

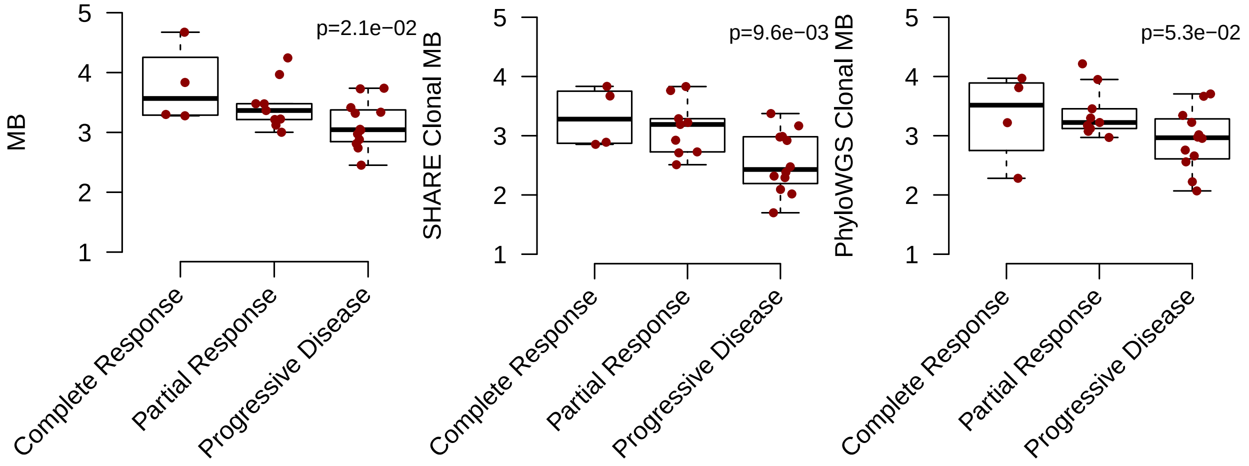
(c)

(d)

(e)

(f)

Figure 2.8: Comparison of SHARE and PhyloWGS on TCGA COAD samples. Copy number and purity are called using a modified version of ASCAT. (a) SHARE purity vs. ASCAT purity of the 400 TCGA colon cancer samples. (b) PhyloWGS entropy vs. PhyloWGS entropy. (c) Proportion of CD8 T cells inferred by CIBERSORT vs. SHARE entropy. (d) Proportion of M1 macrophages inferred by CIBERSORT vs. SHARE entropy. (e) Proportion of CD8 T cells inferred by CIBERSORT vs. PhyloWGS entropy. (f) Proportion of M1 macrophages inferred by CIBERSORT vs. PhyloWGS entropy.



(a)

(b)

(c)

Figure 2.9: Mutation burden (MB) (a), SHARE mutation burden (b) and PhyloWGS mutation burden (c) of Hugo et al. dataset, broken down by response. The p-values are calculated using a two-sided Jonckheere-Terpstra test for ordered differences.

CHAPTER 3: CELL TYPE-AWARE DIFFERENTIAL EXPRESSION ANALYSIS

Differential expression analysis using RNA-seq data is a widely used approach to identify the association between gene expression and covariates of interest. RNA-seq data are often collected from bulk tissue samples, most of which comprise a heterogeneous population of different cell types. Several recent studies have demonstrated that studying cell type-specific gene expression and cell type composition is crucial for many scientific and clinical questions; for example, classifying neuron subtypes (Pavlov and Tracey, 2017), identifying genes and cell types related to Zika virus infection (Nowakowski et al., 2016) or melanoma (Zhang et al., 2018). Most methods for differential expression studies using bulk RNA-seq data (Robinson et al., 2010; Leng et al., 2013; Love et al., 2014) do not consider cell type compositions. A few exceptions include csSAM (Shen-Orr et al., 2010) and TOAST (Li et al., 2019), which are designed for continuous gene expression data and do not fully utilize the count features of RNA-seq data. There are also a few methods with similar goals that were developed for DNA methylation data (Zheng et al., 2018; Luo et al., 2019).

We develop a framework of cell type aware analysis of RNA-seq data (CARseq). We assume cell type compositions have been estimated by an existing method based on reference gene expression of purified cells (Newman et al., 2019; Wilson et al., 2019). CARseq takes the input of bulk RNA-seq data and cell type fraction estimates and performs two tasks: comparison of cell type compositions and cell type-specific differential expression (CT-specific-DE). For CT-specific-DE, CARseq employs a negative binomial regression approach to fully utilize the count features of RNA-seq data, which can substantially improve the statistical power. CARseq is a tribute to both the tradition that the gene expression of a mixture is the summation of non-negative expression of each cell type (i.e., deconvolution on a linear scale) (Zhong and Liu, 2012), and that cell type-independent covariates are adjusted on a log scale. Our shrunken estimates of log fold change (LFC), currently unaddressed in other methods (Shen-Orr et al., 2010; Li et al., 2019), produces a robust and interpretable quantification of CT-specific DE. We benchmark CARseq together with other methods under various simulation setups, illustrating CARseq has the highest power while maintaining type I error control. For example, in a comparison versus TOAST by simulated data with 25 cases vs. 25 controls, CARseq can improve the power by 2 to 4 folds.

We apply CARseq to assess gene expression difference of schizophrenia (SCZ) or autism spectrum disorder (ASD) subjects versus healthy controls. SCZ and ASD are two severe neuropsychiatric disorders that are likely caused by disruption of brain development in early life (particularly in the prenatal and early postnatal period) due to environmental exposure combined with genetic predispositions (Cattane et al., 2018). Two diseases have shared vulnerability genes and overlapping symptoms (Anttila et al., 2018). For example, ASD is characterized by deficit social interaction and repetitive behaviors, which are similar to the negative symptoms ("negative" means taking away from normal state) of SCZ including social withdrawal and impaired motivation. There are also many differences, however, between the two diseases. For example, ASD is an early childhood disease (onset at 6 months to 3 years old) and most SCZ are diagnosed at young adulthood. Compared with ASD, SCZ has additional positive symptoms ("positive" means addition to the normal state) of delusions and hallucinations. The underlying biological mechanisms of the two diseases are not very well understood yet. Our results bring some new insight into the difference and connection between the two diseases. For example, we have observed an imbalance of excitation/inhibition neurons in SCZ but not ASD, which may explain the hallucination symptom in SCZ but not ASD (Jardri et al., 2016). We have also found these two diseases have overlapping signals of CT-specific DE in microglia, supporting the connections of the two diseases through inflammation and oxidative stress (Prata et al., 2017).

Analyzing single cell RNA-seq (scRNA-seq) data is a promising solution for cell type-aware analysis. However, due to high cost and logistic difficulties (e.g., collection of high quality tissue samples, unbiased sampling of single cells), currently, it is very challenging, if not infeasible, to collect scRNA-seq data from large cohorts. If the massive amount of existing bulk RNA-seq data could be re-analyzed to study CT-specific expression and cell type composition, it could bring paradigm-shifting changes to many fields. Our work is one step towards this goal and our results on SCZ and ASD illustrate the power of this CARseq framework, which can be applied to other diseases or conditions.

3.1 Results

3.1.1 Introduction to cell type-aware analysis

To assess the associations between cell type fractions and the covariate of interest, one needs to pay attention to the compositional nature of the data, e.g., we cannot modify the proportion of one cell type without altering the proportion of at least one other cell type (Aitchison and Egozcue, 2005). Therefore, following a commonly used practice for compositional data analysis, we transform the k cell type fractions to $k - 1$ log ratios: log of the fraction of each cell type (other than the reference) vs. a reference cell type. We choose excitatory neuron as our reference cell type because it is the most abundant cell type in our studies and the results are easier to explain (e.g., when studying excitation/inhibition imbalance).

The more challenging part is to assess CT-specific-DE, while we only observe expression in bulk samples where the variability can come from both CT-specific expression and cell type fractions. Our model is built around the assumption that the expression in bulk samples is the summation of CT-specific expression weighted by cell fractions in linear scale (Figure 3.10). The model also allows the inclusion of cell type-independent covariates, such as age, gender, batch etc.

3.1.2 Benchmarking methods through simulations

3.1.2.1 Simulation setup

We first use a simulation study to evaluate the power and type I error of CT-specific-DE by our method and two existing methods: csSAM (Shen-Orr et al., 2010) and TOAST (Li et al., 2019). csSAM assesses CT-specific-DE by a two-step approach: estimation of CT-specific expression followed by testing by permutations. It has lower power than TOAST (Li et al., 2019) and it cannot account for covariates. TOAST and CARseq are more similar since they both combine the estimation of CT-

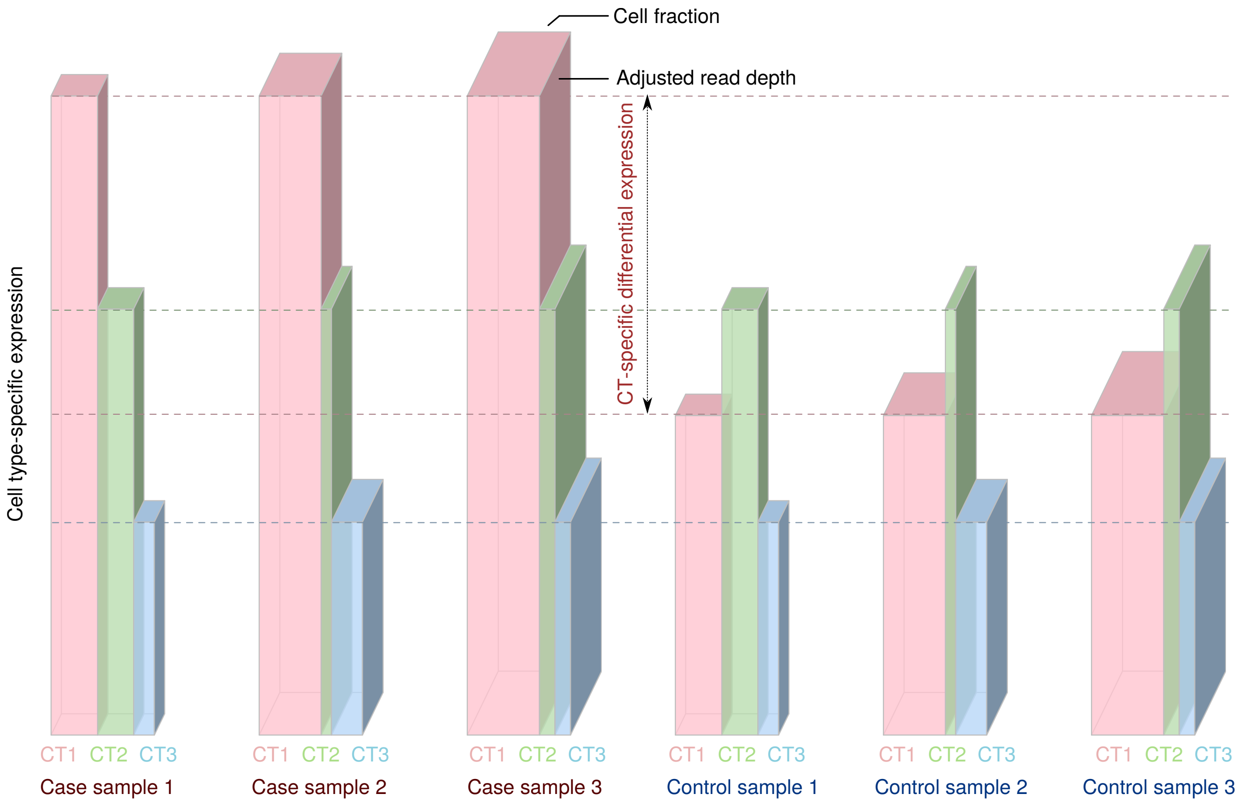


Figure 3.10: Each grouped bar illustrates the total expression of a gene in a bulk sample (the volume of the grouped bar) that is the summation of gene expression from individual cell types (each bar). The depth of each grouped bar is proportional to the covariate-adjusted read-depth. The width of each bar is proportional to its cell fraction, and the height of each bar is proportional to the CT-specific expression. The left/right three columns show three case/control samples, respectively. Our method estimates the mean value of CT-specific expression for case and control groups separately. In this toy example, cell type 1 (pink) has twice expression in cases than controls, while cell type 2 (green) and cell type 3 (blue) are not differentially expressed.

specific expression and CT-specific-DE testing in one likelihood framework that allows adjustment for covariates. The difference is that CARseq adopts the negative binomial model that allows modeling of gene expression decomposition on a linear scale. In contrast, TOAST uses a linear model that is less desirable to model count data. An alternative is to use TPM (transcripts per million) to replace count, which is a linear transformation of counts after adjusting for gene length and read-depth. We evaluated the performance of TOAST using both counts and TPM and the latter delivers better results, so we reported the results of TOAST using TPM and left the results using counts in Supplementary Materials (Figures 3.19-3.22).

We simulated CT-specific expression data that mirror the gene expression data from single nucleus RNA-seq (snRNA-seq) of human brains (Hodge et al., 2019). We simulated the cell fractions to resemble our estimates from the Common Mind Consortium (CMC) bulk RNA-seq data (Fromer et al., 2016) (Figure 3.31). Three cell types were simulated. Cell type 1, intended to imitate the excitatory neuron, taking the lion's share of around 60% of the cells in each sample. The other two cell types, with much smaller fractions, were intended to represent inhibitory neurons and non-neuron cells. We also simulated a covariate in the mold of RNA integrity number (RIN) and specified the distribution of its effect size based on estimates of RIN effect from the CMC data. More details of the simulation procedure can be found in Section B.1 of the Supplementary Materials.

3.1.2.2 CARseq has substantially higher power than other methods

The benchmark consists of simulations in different sample sizes and different patterns of differential expression (Figure 3.11). With covariates provided, both CARseq and TOAST can control false discovery rate (FDR) very well, with CARseq having an edge in power. When covariates are missing, the model misspecification might result in inflated type I error under some replicates, regardless of the method being used. Nevertheless, the simulation results demonstrated that CARseq is more powerful than TOAST, which is more powerful than csSAM, and correct specification of covariates can improve power and ensure the control of FDR. It worth noting the power of CT-specific-DE can be low when the sample size is small (e.g., $n = 50$, 25 cases vs. 25 controls), due to the uncertainty to estimate CT-specific expression. This is also the situation where CARseq shows much higher power than TOAST, with two to four folds of improvement (Figure 3.11(A)).

3.1.2.3 CARseq is robust to noise in cell fraction estimates or cell size factors

We use the true cell type fractions in the above simulations. Next we demonstrate that plugging in the estimates of cell type fractions that have reasonable deviations from true values will not lead

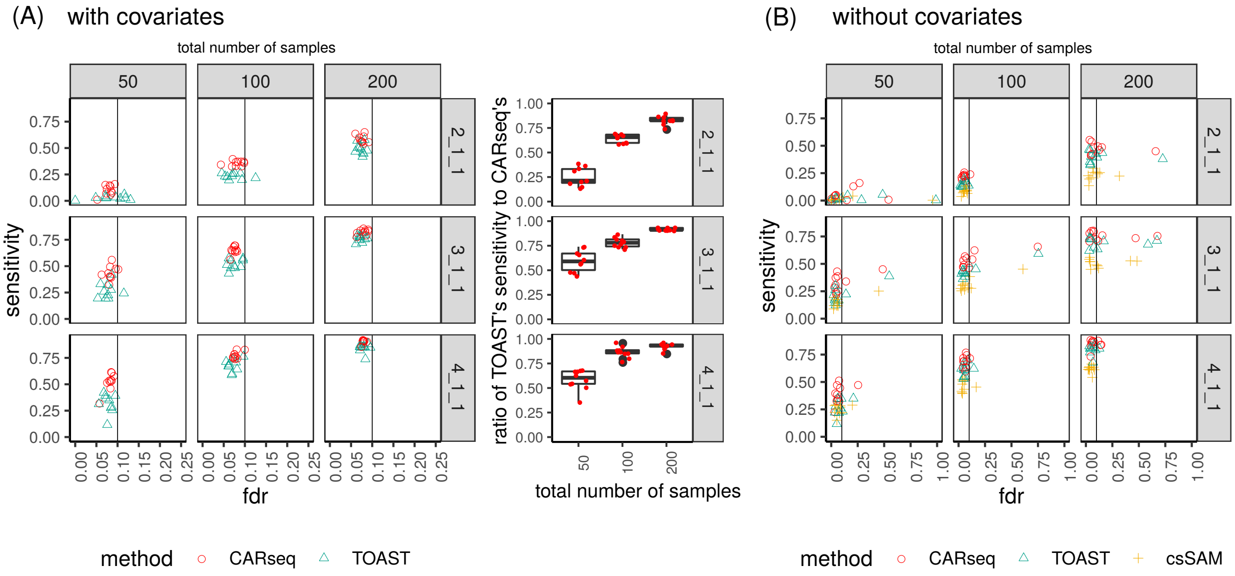


Figure 3.11: The FDR vs. sensitivity of several methods testing for CT-specific DE, when (a) the covariate is provided to the method, and (b) when the covariate is not provided to the method. The ratio of TOAST's sensitivity to CARseq's when the covariate is provided (hence type I error has been controlled) is illustrated in the boxplot. There are 10 simulation replicates for each combination of sample size (columns) as the total number of case-control samples and patterns of differential expression (rows). For each replicate, there are 2,000 genes following the pre-specified pattern of differential expression and 8,000 genes with no differential expression in any of the three cell types. In the notation for the pattern of differential expression, the three numbers separated by underscores each represent the fold change in each cell type. In this simulation setup, only cell type 1 is differentially expressed. The vertical line indicates the intended FDR level of 0.1. Note that csSAM does not support the inclusion of covariates; the scales of the x-axis in the two subfigures are different.

to a discernible decrease in power or increase in type I error. Specifically, we added a zero-centered Gaussian noise with a standard deviation of 0.1 to the cell fractions on a logit scale and rescaled the cell fractions so that their summation is 1 for each sample (Figure 3.23). Using this noisy cell type fraction estimates, the sensitivity and FDR for CT-specific-DE are very similar to the noise-free scenario in Figure 3.11. A major limiting factor for accurate cell type fraction estimation is the availability of reference data of CT-specific expression. With the quick development of single cell techniques and large scale project such as human cell atlas (Regev et al., 2017), we expect that relatively accurate estimation of cell type fractions will be available in more tissue types.

Cell size factor is another source of uncertainty. Most computational methods estimate the fraction of gene expression instead of the fraction of cells for each cell type. If one cell type has on average more expression per cell, a cell size factor correction is needed to estimate cell fractions. The cell fractions are needed for CARseq since it directly models count data. In contrast, when using TPM to quantify expression level in TOAST, there is no need to adjust for cell size factor. To interrogate the performance of CARseq when the cell size factor is misspecified, we intentionally applied wrong size factors (1.2, 1, 1) instead of the true ones (1, 1, 1) when evaluating CT-specific-DE. This misspecification of cell size factor slightly reduces the power of CARseq, though it still has higher power than TOAST. Only under extreme and unrealistic misspecification of cell size factor (e.g., (2, 1, 1) vs. true values of (1,1,1)) does the power of CARseq drops to become similar to that of TOAST (Figure 3.26).

3.1.2.4 CARseq delivers more accurate and reproducible estimates of effect sizes

CARseq quantify the effect size of CT-specific-DE by log fold change or shrunken log fold change (see Method Section for more details). TOAST defines the effect size as $\beta/(\mu + \beta/2)$, where μ is base-line expression in one group, and β is the gene expression difference between two groups. To make the results more comparable between CARseq and TOAST, we amend the effect size definition in TOAST and propose to define LFC as $\log(|\mu + \beta|) - \log(|\mu|)$. To examine the reproducibility of effect size estimation, we divided the samples in a simulation replicate into two subsets of equal sizes and then compare the effect size estimates in the two subsets. It is clear that CARseq's shrunken log fold change is best reproduced between the two subsets (Figure 3.12). For example, when sample size is 25 cases vs. 25 controls for each subset (middle panel of Figure 3.12), the Spearman correlation of effect size estimates between the two replicates are 0.71, 0.53, 0.13, and 0.08 for effect size qualified by CARseq shrunken LFC, CARseq LFC, TOAST LFC, and TOAST effect size, respectively.

3.1.3 CARseq: comparing schizophrenia subjects versus controls

Schizophrenia (SCZ) is a severe neuropsychiatric disorder that affects approximately 1% of word-wide population (Owen et al., 2016). There is strong evidence from both human and animal studies

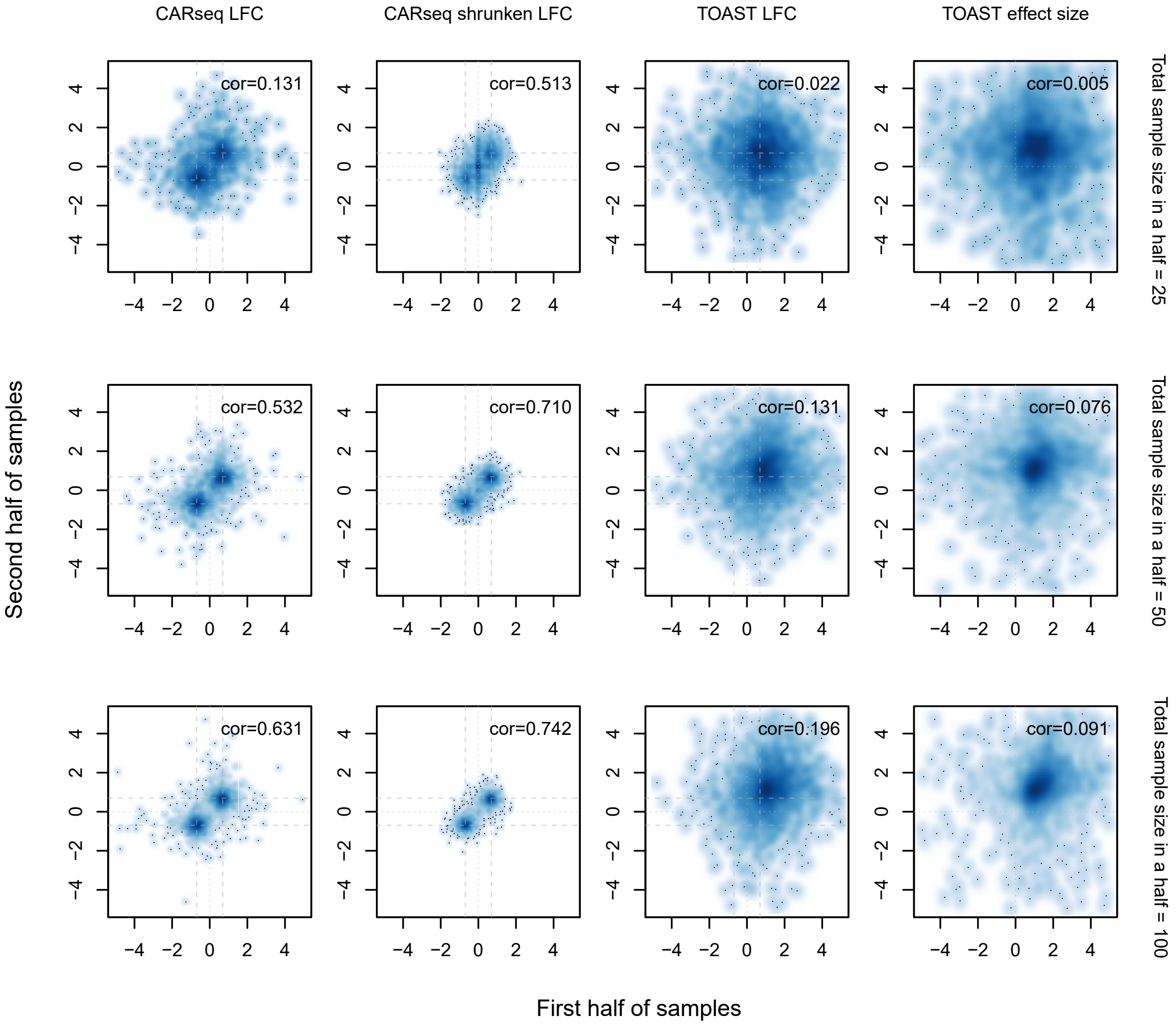


Figure 3.12: The reproducibility of effect size estimation among 2,000 differentially expressed genes in cell type 1 (fold change of 2 or log fold change of 0.7) when CARseq and TOAST are applied to a simulation dataset of a mixture of three cell types where only cell type 1 is differentially expressed. A Spearman correlation coefficient for the reproducibility of effect size estimates is added at the top right corner of each plot.

that support a neurodevelopmental model of SCZ: perturbation of early neurodevelopment during pregnancy (e.g., by environmental factors such as maternal stress or infections), combined with a genetic predisposition (the heritability of SCZ is estimated to be roughly 80%) (Owen et al., 2016). We applied CARseq to study the gene expression of SCZ patients vs. controls using the bulk RNA-seq data of prefrontal cortex samples, generated by the CommonMind Consortium (CMC) (Fromer et al., 2016), hereafter referred to as CMC-SCZ study. After filtering out the outlier samples reported by Fromer et al. (Fromer et al., 2016), we had 250 SCZ subjects and 277 controls.

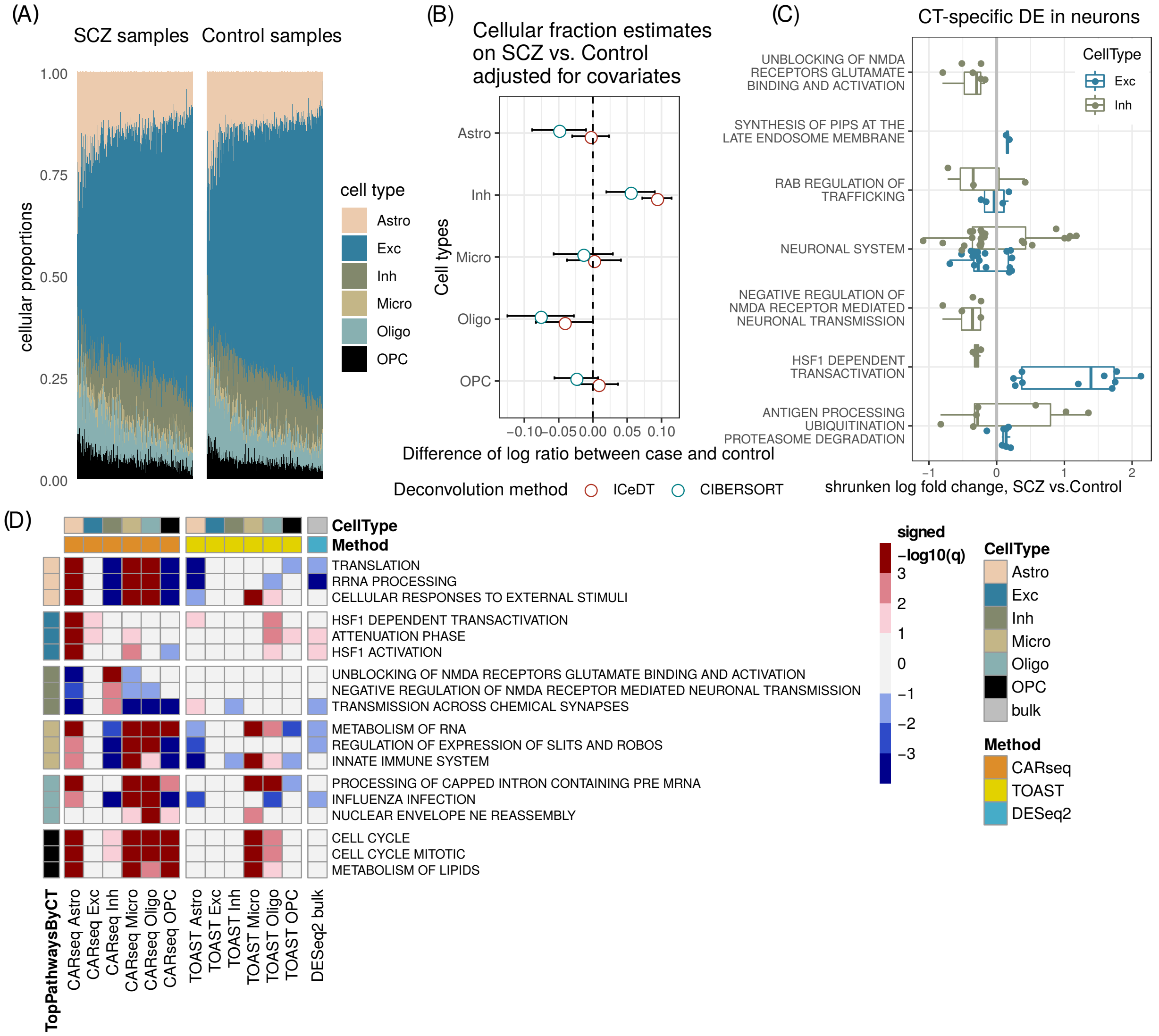


Figure 3.13: CARseq on gene expression data between schizophrenia (SCZ) and controls. (A) Estimated cell fractions by ICeD-T sorted by increasing fractions of excitatory neurons. (B) The effect size of case-control status on relative cell fractions against excitatory neurons (log ratio of the cell type of interest vs. excitatory neuron). The standard errors are denoted by bars. (C) CT-specific DE in excitatory neurons and inhibitory neurons in significantly enriched pathways. Only genes with a *p*-value less than 0.05 are shown. (D) Gene set enrichment analysis results on REACTOME pathways. Three top pathways were shown for each cell type, ranked by -log10 q value with the sign of normalized enrichment score (NES). Positive NES indicates enrichment of genes with small p-values.

We estimated cell type proportions for six cell types: excitatory neurons (Exc), inhibitory neurons (Inh), astrocyte (Astro), microglia (Micro), oligodendrocyte (Oligo) using CIBERSORT (Newman et al., 2015) and ICeD-T (Wilson et al., 2019). The estimates from these two methods are highly correlated, though with some noticeable differences (Figures 3.31-3.32). We examined whether relative cell fractions with respect to excitatory neuron are associated with the case-control status while accounting for a set of covariates including log transformed read depth, age, gender, RNAseq QC metrics, batch effects, genotype PCs, as well as two surrogate variables that were estimated conditioning on cell type fractions (Figure 3.13(A), see Method section for details of the covariates used in our analysis). We found that the relative cellular abundance of the inhibitory neuron quantified by ICeD-T is significantly higher in SCZ samples than control samples (p-value 1.5×10^{-5}), and there is a similar trend for cell fraction estimates by CIBERSORT, though the difference is not significant (p-value 0.12). There is also a trend of relative depletion of oligodendrocyte, though it is not significant (p-values 0.32 for ICeD-T and 0.12 for CIBERSORT).

Since cell type fraction estimates from CIBERSORT and ICeD-T are highly correlated, we present the CARseq results using cell type fractions estimated by ICeD-T for simplicity. CARseq found 1 differentially expressed gene (DEG) (q-value < 0.1) in astrocytes, 138 DEGs in microglia, and 656 DEGs in oligodendrocytes (see Figure 3.34 for p-value distributions). In contrast, TOAST identified 3 DEGs in inhibitory neurons, 30 DEGs in microglia, and 1 DEG in oligodendrocytes (See Figure 3.36 for p-value distributions). Both methods could control type I error/FDR, indicated by the fact that if the case-control label was permuted, the only false discovery (q-value < 0.1) is 1 gene in microglia reported by CARseq, and the p-value distribution is uniform (Figures 3.35 and 3.37). These results are consistent with our simulation results that CARseq can identify DEGs with a higher power than TOAST, while controlling FDR.

Although we did not find any DEGs in inhibitory neurons or excitatory neurons at q-value cutoff 0.1, gene set enrichment analysis (GSEA) using the rankings of all the genes by their CT-specific DE p-values recover some interesting pathways (Figure 3.13(D)). For inhibitory neurons, we found that genes involved in unblocking or negative regulation of NMDA receptors are enriched. This is very relevant since NMDA hypofunction is a key contributor to the SCZ disease process and they are involved in excitation/inhibition imbalance (Owen et al., 2016). The majority of the inhibitory-neuron-DE genes in NMDA pathways have lower expression levels in SCZ subjects than controls (Figure 3.13(C), Figure 3.42), consistent with the hypofunction of NMDA. We found that the heat shock related genes are enriched in the DE genes in excitatory neurons, and they tend to have higher expression in SCZ subjects than controls (Figure 3.13(C), Figure 3.42). This is consistent with previous findings that heat shock response plays a crucial role in the response of brain cells to prenatal environmental insults (Lin et al., 2014).

Next, we shift our attention to glial cells. For microglia, we found the pathways of innate immune system and cell cycle are enriched in the CT-specific-DE genes and they are over-expressed in SCZ subjects than controls (Figure 3.13(D), Figure 3.43), supporting the observations of activation of microglia in SCZ subjects (Prata et al., 2017). It is interesting that these pathways are also enriched in oligodendrocyte, but they are down regulated in SCZ subjects than controls (Figure 3.13(D), Figure 3.43), suggesting inactivation of oligodendrocyte. We also found Slit-Robo signaling pathway is down/up -regulated in microglia and oligodendrocyte, respectively (Figure 3.43). Slit-Robo signaling pathway is involved in the neurogenesis and migration of neuronal precursors toward the lesions, and glial cells are also involved in these processes (Kaneko et al., 2018). Our findings suggest microglia and oligodendrocyte may take different roles in this process in SCZ subjects.

3.1.4 CARseq: comparing ASD subjects versus controls

Autism spectrum disorder (ASD) affect more than 1% of population, with heritability estimated to be 68% to 96% (Lord et al., 2020). Individuals with ASD are often impaired in social communication and social interaction, and limit themselves to repetitive behaviors and interests from an early age (Lord et al., 2020). There is abundant evidence supporting the neurodevelopmental model for ASD. Large-scale ASD genetic studies have identified hundreds of ASD risk genes that are mutated more frequently in ASD subjects than controls (Satterstrom et al., 2020). We analyzed the bulk RNA-seq data from ASD subjects and controls, published by a UCLA group (Parikshak et al., 2016; Gandal et al., 2018), hereafter referred to as UCLA-ASD study. They reported findings on 251 post-mortem samples of frontal and temporal cortex and cerebellum for 48 ASD subjects versus 49 controls and found significantly differentially expressed genes in cortex but not in cerebellum (Parikshak et al., 2016). In this study, we focus on frontal cortex region based on positive findings of DE genes in this earlier study and that it matches the brain region of SCZ data analyzed in this paper. After filtering by brain regions, we ended up with 42 ASD subjects and 43 control subjects (See Method section for details).

Comparing relative cell type fractions (with respect to excitatory neurons) between ASD subjects and controls, we found the relative abundance of astrocyte is significantly higher in ASD subjects than controls (p = 0.021 and 0.024 for cell type fractions estimated by ICeD-T and CIBERSORT, respectively, Figure 3.14(A)). Microglia also show a trend of higher relative abundance in ASD subjects than controls. These observations support the hypothesis that pro-inflammatory maternal cytokines in the developing brain can lead to neuroinflammation and proliferation of astrocyte and microglia (Petrelli et al., 2016).

CARseq reports 232 DEGs (q-value < 0.1) in excitatory neurons and 855 DEGs in inhibitory neurons, and no DEGs in the other four cell types (Figure 3.47). TOAST recovers 2 DEGs in excitatory neurons and no DEGs in the other five cell types (Figure 3.49). We also sought to evaluate the FDR control by repeating our analysis after permuting case/control labels (Figure 3.48 and 3.50) and noticed inflation of type I error in some permutations. This is likely due to the fact that the model is mis-specified after permuting case/control labels, and small sample size and/or unaccounted covariates could further exaggerate such effects. Thus the results of this analysis should be interpreted with caution. Nevertheless, as discussed next, we observed expected functional category enrichment and some consistent signals between SCZ and ASD, suggesting our analysis in this dataset with relatively small sample size still captures meaningful signals.

First, we considered a list of 328 autism risk genes curated by Simons Foundation Autism Research Initiative (SFARI). Most of these risk genes were identified because they harbor more disruptive mutations in the ASD cases than the general population. We found that these ASD risk genes are significantly enriched among the DE genes in inhibitory neurons (p-value 5.8×10^{-7}) and excitatory

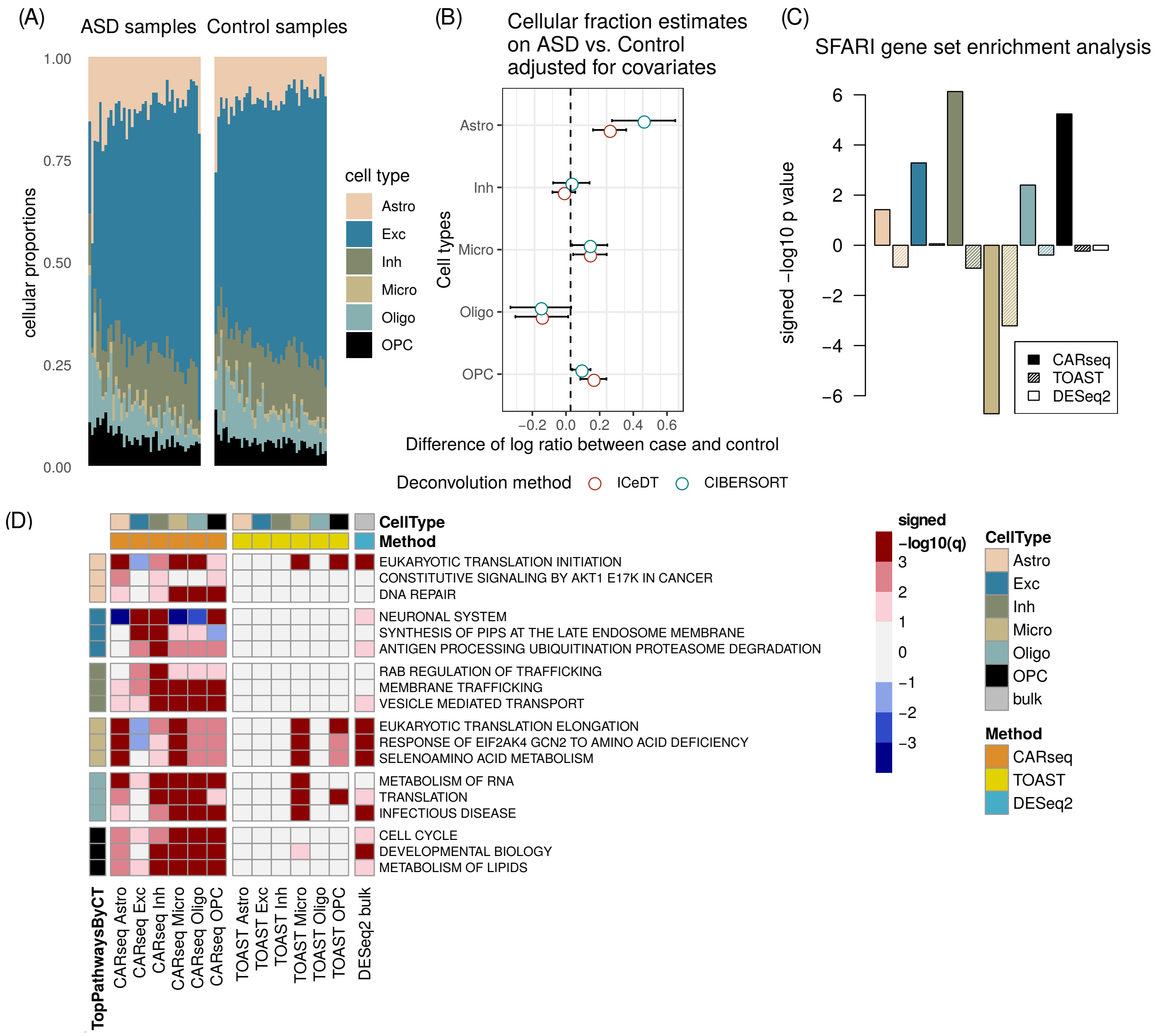


Figure 3.14: CARseq on the autism spectrum disorder (ASD) bulk expression data. (A) Estimated cell fractions by ICeD-T sorted by increasing fractions of excitatory neurons. (B) The effect size of case-control status on relative cell fractions against excitatory neurons (log ratio of the cell type of interest vs. excitatory neuron). The standard errors are denoted by bars. (C) Gene set enrichment analysis results in -log10 p value with the sign of normalized enrichment score (NES) of the SFARI gene set, a curated list of 328 autism risk genes. (D) Gene set enrichment analysis results on REACTOME pathways. Three top pathways were shown for each cell type, ranked by -log10 q value with the sign of normalized enrichment score (NES). Positive NES indicates enrichment of genes with small p-values.

neurons (p-value 3.3×10^{-4}), and they are significantly depleted among the DE genes in microglia (p-value 2.9×10^{-7}) (Figure 3.14(C), Table 3.8). Our findings are consistent with the results reported using snRNA-seq data (Velmeshev et al., 2019). Enrichment by TOAST results is consistent with CARseq for microglia (4.1×10^{-4}), but not significant for inhibitory neurons (p-value 0.14) or excitatory neurons (p-value 0.89). In contrast, no enrichment is found from DE analysis on bulk tissue by DESeq2 (Love et al., 2014) (p-value 0.66).

The pathways enriched in DE genes of excitatory or inhibitory neurons include more generic and broad pathways such as "neuronal system" and "antigen processing", and more specific ones such as "synthesis of PIPs " and "RAB regulation of trafficking". Both "synthesis of PIPs" and "RAB regulation of trafficking" are related to one type of glutamate receptors named AMPA receptor (McCartney et al., 2014; Hausser and Schlett, 2019). The log fold changes of DE genes show that these two pathways are up-regulated in inhibitory neurons of ASD subjects, but down-regulated in excitatory neurons of ASD subjects (Figure 3.55). Such up/down-regulation pattern is much cleaner in "synthesis of PIPs" than "RAB regulation of trafficking". This suggests the relevance of AMPA activity in the pathophysiology of ASD. Genes in the antigen processing pathways tend to be up-regulated in inhibitory neurons but down-regulated in excitatory neurons (Figure 3.55), suggesting increased/decreased interactions with the immune system in inhibitory neurons and excitatory neurons, respectively. The enriched pathways in glial cells include those related to translation initiation, elongation, and "Response of EIF2AK4 (GCN2) to amino acid deficiency". It has been shown that dysregulation of translation can cause neurodegeneration (Ishimura et al., 2016), which is corroborated by our findings that suggest their connections with ASD.

3.1.5 Comparing DE testing by CARseq versus DESeq2

DESeq2 (Love et al., 2014) is a good representative of existing methods for DE analysis of bulk tissue samples. In both SCZ and ASD analyses, most findings from DESeq2 were not identified as CT-specific-DE genes (Figure 3.41 and 3.54). An immediate question is whether those DESeq2 DE genes are differentially expressed in one or more cell types, or they may reflect confounding effect due to cell type compositions. In our default analysis, DESeq2 does not take any cell type composition as covariates. After accounting for cell type compositions (by including log ratios of cell type compositions), DEseq2 identified more DE genes using the CMC-SCZ data (from 1,009 to 1,888, with an intersection of 810 at q-value 0.1, Table 3.5), suggesting that these DE genes are differentially expressed in one or more cell types, but with a relatively small effect sizes and thus were missed by CARseq. In contrast, for the UCLA-ASD data, DESeq2 identified much less DE genes after accounting for cell type compositions (from 1063 to 481, with an intersection of 185 at q-value 0.1, Table 3.5), suggesting that many DEseq2 DE genes in ASD are indeed confounded by cell type compositions.

3.1.6 Concordant microglia-specific DE genes between SCZ and ASD

We found an interesting pattern that genome-wide microglia-specific-DE p-values show significant correlations between SCZ and ASD (Figure 3.15(A), correlation 0.14 and p-value $< 2 \times 10^{-16}$). In addition, the fold changes of microglia DE genes in different pathways also show consistent patterns between SCZ and ASD (comparing Figure 3.43(A) vs. 3.49(A)): up-regulation in innate immune system and cell cycle, and down-regulation in translation, slit-robo signaling pathway, and influenza infection. We further study the overlapping DE genes. Using a liberal p-value cutoff of 0.05, we identified 1,674 and 355 microglia-specific-DE genes in SCZ and ASD studies, respectively, with an overlap of 65 genes. This overlap is significantly larger than 33 overlaps expected by chance (p-value 9.6×10^{-9} by Chi-squared test). Several REACTOME pathways are over-represented by these 65 genes (by R package `goseq`, Figure 3.15(B), Table 3.9). Note that this over-representation analysis is different from GSEA, which uses genome-wide rankings, and here we only consider these 65 genes. In addition to the pathways that have been identified by GSEA, we found some new ones. One interesting finding is "Selenoamino acid metabolism". Since selenium-dependent enzymes prevent and reverse oxidative damage in brain, our findings support that selenium-dependent enzymes could mediate the relation between antioxidants and SCZ/ASD (Raymond et al., 2014; Greenhalgh et al., 2020).

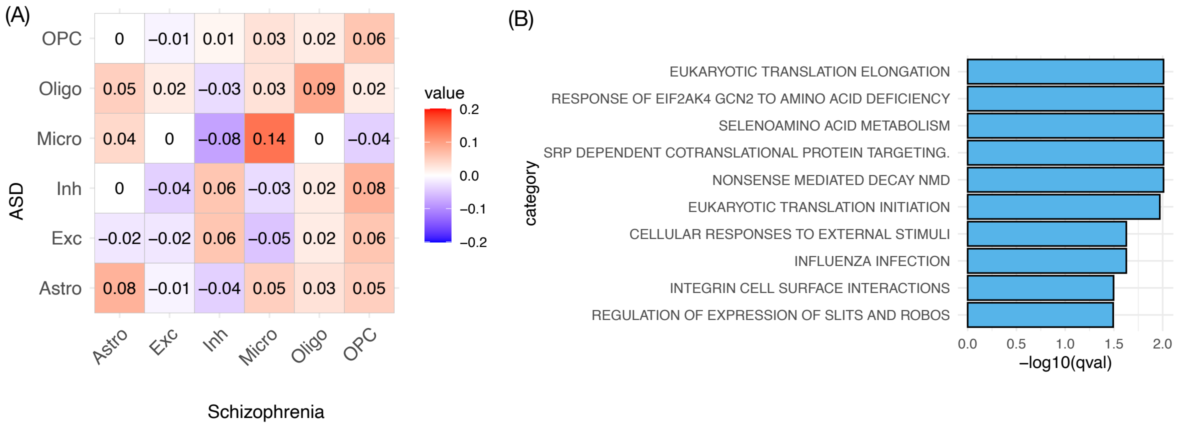


Figure 3.15: (A) The correlation matrix of -log10(Microglia-specific-DE p-values) (calculated by CARseq) between CMC-SCZ and UCLA-ASD studies. (B) The REACTOME pathways that are over-represented by the 65 genes with microglia-specific-DE p-values smaller than 0.05 in both CMC-SCZ and UCLA-ASD studies.

3.2 Discussion

SCZ and ASD are two prevalent neuropsychiatric disorders with profound burdens on the affected individuals, their families, and society. Previous studies have identified genetic and environmental factors that contribute to disease risks. However, our understanding of the molecular mechanisms that connect these risk factors with disease onset is still incomplete, and insight on such molecular mechanisms is crucial for more effective treatment and prevention. For example, several SCZ drugs work at molecular level to interact with neurotransmitters. However, it is not clear which cell types are more relevant for the functioning of the drugs. Analysis of gene expression in postmortem brain samples is an effective approach to study such molecular mechanisms, though traditional methods for DE analysis cannot separate the effects of cell type compositions and cell type-specific gene expression changes. To the best of our knowledge, we present the first cell type aware analysis of postmortem gene expression data from SCZ and ASD.

The molecular mechanisms underlying SCZ and ASD can be divided into two categories: alterations in neurotransmitter systems and stress-associated signaling including immune/inflammatory-related processes and oxidative stress (Cattane et al., 2018). NMDA and AMPA are two types of receptors for neurotransmitter glutamate. We found evidence for hypofunction of NMDA in SCZ (particularly in inhibitory neurons) and dysregulation of AMPA in ASD. While excitation-inhibition (E-I) imbalance has been suggested as a common feature of SCZ and ASD, we found the relative fractions of inhibitory neurons versus excitatory neurons is higher in SCZ than controls, but are similar between ASD and controls. This is consistent with previous finding that the hypofunction of NMDA could cause E-I imbalance (Kehrer et al., 2008) and our finding that NMDA pathway is perturbed in SCZ but not in ASD. In addition, it has been shown that E-I imbalance is the underlying mechanism for hallucination (Jardri et al., 2016), which is a symptom of SCZ but not ASD. Thus our finding of E-I imbalance in SCZ but not in ASD may explain part of the symptom difference between the two diseases. A recent study also found no E-I difference between ASD and controls (Ajram et al., 2017).

Stress-associated signaling in SCZ and ASD has been widely studied. For example, animal studies show that prenatal stress referred as maternal immune activation (regardless the cause such as infection by different pathogens or immune stimulation) can lead to SCZ or ASD (Prata et al., 2017), implying the role of immune system in disease pathology. Microglia is the tissue resident macrophages in brain and plays a central role in immune response in brain. We found microglia-specific DE genes have significant overlap between SCZ and ASD and they have higher expression in SCZ/ASD subjects than in controls, suggesting microglia are in more active states in SCZ/ASD than controls, and pointing to the relevance of several biological processes including translation regulation and oxidative damage. The relative proportion of astrocyte is significantly higher in ASD than controls, and there is a trend of higher abundance of microglia in ASD than controls. The relative fractions of astrocyte and microglia are similar between SCZ and controls, though.

These analyses are enabled by our computational framework CARseq, a framework for cell type-aware analysis of RNA-seq data from bulk tissue samples. CARseq require the estimates of cell type fractions, which relies on reference of cell type-specific gene expression data. We expect with the development of human cell atlas (Regev et al., 2017), such resource in other tissues will be generated in the near future, and thus enable CARseq analysis in broader tissues and relevant diseases.

A practical consideration of using CARseq is that it may have limited power when the sample size is small. This is the price that we have to pay for the uncertainty of estimating CT-specific expression from bulk RNA-seq data. As a rule of thumb, we do not recommend using CARseq when the sample size minus the number of covariates is smaller than 20. For large studies, e.g., with hundreds of samples, it may worth considering a new study design to generate scRNA-seq data in a subset of samples, and generate bulk RNA-seq data from all the samples. The scRNA-seq data can be used to generate cell type-specific gene expression reference for cell type fraction estimation, which can be used for the CARseq analysis on bulk RNA-seq data. In addition, the scRNA-seq data can also be used to validate the results of CARseq.

3.3 Methods

3.3.1 Likelihood function of CARseq model

Let T_{ji} be the RNA-seq read count (or fragment count for paired-end reads) for gene $j \in \{1, \ldots, J\}$ and sample $i \in \{1, \ldots, n\}$, where J is the total number of genes and n is the number of bulk samples. We denote the cell fraction for cell type $h \in \{1, \ldots, H\}$ in the i-th sample by $\widehat{\rho}_{hi}$.

We assume T_{ji} follows a negative binomial distribution: $T_{ji} \sim f_{NB}(\mu_{ji}, \phi_j)$, with mean value μ_{ji} and dispersion parameter ϕ_j. Since deconvolution on a linear (non-log) scale yields better accuracy (Zhong and Liu, 2012), we let:

$$\mu_{ji} = \sum_{h=1}^{H} \rho_{hi} \widetilde{\mu}_{jih},$$

where $\widetilde{\mu}_{jih}$ is the mean expression of the j-th gene in the h-th cell type of the i-th sample. The above deconvolution states that the expected total read count is the summation of expected CT-specific read count weighted by cell fractions across all cell types $h \in \{1, \ldots, H\}$. In practice, cell fraction estimates $\hat{\rho}_{hi}$ are used in place of ρ_{hi}.

We model the relation between $\widetilde{\mu}_{jih}$ and M_0 CT-specific covariates through a log link function, which is commonly used for negative binomial regression: $\widetilde{\mu}_{jih} = d_i^{\beta_{j0}} \exp\left(\sum_{m=1}^{M_0} \gamma_{jhm} x_{ihm} \right)$, where d_i is the sequencing read depth of sample i, γ_{jhm} and x_{ihm} are the regression coefficient observed data for the m-th covariate. In all the analyses of this paper, we use 75 percentile of the expression across all the genes within a sample.

The effect sizes of many covariates may not vary across cell types. For example, since RNA integrity number (RIN) quantifies sample RNA quality, it would associate with observed gene expression in the same way regardless of the original cell type. By separating cell type-independent covariates from CT-specific covariates, we can construct a model with less degrees of freedom. Suppose M out of M_0 parameters are CT-specific and the rest $K = M_0 - M$ parameters are cell type-independent, we have:

$$\widetilde{\mu}_{jih} = d_i^{\beta_{j0}} \exp\left(\sum_{k=1}^{K} \beta_{jk} w_{ik}\right) \exp\left(\sum_{m=1}^{M} \gamma_{jhm} x_{ihm}\right),$$

where w_{ik} is the value of the k-th cell type-independent covariate in sample i.

The log-likelihood can be maximized using iteratively weighted least squares (IWLS) with some tweaks (Supplementary Materials Section A). After that, we can construct likelihood ratio statistics to conduct the CT-specific-DE tests. Although the likelihood-based testing framework can be generalized to accommodate a variety of tasks (Li et al., 2019), e.g., test for a continuous variable or test for a linear combination of regression coefficients, our main focus in the article is CT-specific-DE tests among two or more groups.

We noted that CARseq reports large LFC estimates in some cell types, which probably reflect estimation uncertainty, particularly for the cell types with low proportion or when the sample size is small. To mitigate this problem, we developed a shrunken LFC estimation procedure, see Section A.5 in Supplementary Materials for details.

3.3.2 Comparison with other methods.

There are a few alternatives to our methods, though they were indeed developed for different purposes and not best suited for RNA-seq data analysis. TOAST (Li et al., 2019) is a method for CT-specific DE or differential methylation analysis. It uses a linear model that is more flexible than our negative binomial model to handle different types of data, though for RNA-seq count data with a very strong mean-variance relationship, a linear model that assumes homogeneous variance has to choose between variance stabilization (e.g., by log-transformation of gene expression) or deconvolution in linear scale. All the analysis performed in real data has been done using both CARseq and TOAST and additional results for TOAST are available in Supplementary Figures.

Accounting for observed/unobserved confounding covariates is crucial for DE analysis, and the unobserved covariates can be estimated by surrogate variable analysis (SVA) (Leek et al., 2012). In simulation data, we found that not accounting for relevant covariates can lead to inflated type I error. This limits the application of csSAM that cannot adjust for covariates in a lot of practical settings. For this reason, we did not apply csSAM in real data analysis.

It is worth mentioning that CIBERSORTx (Newman et al., 2019) provides a high-resolution mode that can provide CT-specific expression estimates for each sample. It is very helpful for data exploration, but not appropriate for differential expression testing since it relies on many assumptions that create some dependency in the final output. The authors of CIBERSORTx did not recommend using such high-resolution mode for differential expression analysis. However, it could be a convenient and attractive approach for practitioners. To warn against it, we demonstrate this approach could bring inflated type I error by simulation studies (Figures 3.19-3.22).

3.3.3 Estimation of cell type compositions

We use two reference-based methods to estimate cell type compositions of bulk tissue samples: CIBERSORT (Newman et al., 2015) and ICeD-T (Wilson et al., 2019). CIBERSORT is a popular method that use a support vector regression to estimate cell type proportions. ICeD-T is a likelihood-based method that model gene expression using log-normal distribution. It allows a subset of genes to be "aberrant" in the sense that the CT-specific gene expression of such genes are inconsistent between bulk tissue samples and external reference. Such aberrant genes are down-weighted in estimating cell type proportions.

We generated CT-specific gene expression reference using snRNA-seq data from the middle temporal gyrus (MTG) of the human brain (Hodge et al., 2019). This is not a perfect match for the bulk RNA-seq data that are from pre-frontal cortex (PFC). We have compared MTG with another snRNA-seq data generated from human PFC as well as other brain regions using DroNC technique. The CT-specific gene expression is similar between the two datasets, except for endothelial cells. We chose to use MTG data to generate reference since it has much higher depth and better coverage, making it more similar to bulk RNA-seq data. We exclude endothelial in our analysis since there are only 8 endothelial cells in MTG data and its expression has very weak similarity to the endothelial cells from DroNC data. See Supplementary Materials Section B.2.1 for more details.

A related question is that when estimating cell type fractions, an implicit assumption is that the signature genes' cell type-specific expression level does not change in different conditions. Then it seems to be a contradiction when we assess its DE. A more rigorous approach is to assess DE only for non-signature genes. However, since cell fractions are estimated using hundreds of genes with robust models, removing any one signature gene will not lead to a noticeable change of cell type fraction estimates. Therefore, it is as if we assess DE of a signature gene without using it as part of the signature matrix. On the other hand, if too many genes within the signature gene set are detected to be differentially expressed, the accuracy of the cell type fraction estimates is questionable, and an alternative signature gene set should be selected.

3.3.4 CARseq analysis for SCZ

The gene expression data and sample characteristics data were downloaded from CommonMind Consortium (CMC) Knowledge Portal (see section URLs). We include the following covariates in our CARseq CT-specific DE analysis:

- log transformed read-depth (75 percentile of gene expression across all the genes within a sample),
- institution (a factor of three levels for the three institutes where the samples were collected),
- age, gender, and PMI (Post-mortem interval),
- RIN (RNA integrity number) and its square transformation RIN^2,

- a batch variable "Libclust", which is clusters of library batches into 8 groups,
- two genotype PCs, and two surrogate variables.

The surrogate variables were calculated after accounting for cell type compositions. Specifically, we add the log ratios of cell type compositions (with excitatory neuron as baseline) as the covariates and then calculate surrogate variables using R function `sva` from R package `sva` (Leek et al., 2012).

The covariates selected in our model are mostly similar to those included in the original analysis (Fromer et al., 2016) except two differences. One is that we included two instead of five genotype PCs in our analysis since other PCs are not associated with gene expression data (Figures 3.16). Surrogate variables were computing using the R package "sva" (Leek et al., 2012). Two surrogate variables are included because adding these two surrogate variables increased the variance explained (R^2) in a linear model to fit log-scale mixture expression from 0.55 to 0.68, while more surrogate variables offered a comparably limited increase in R^2. Prior to inclusion in the model, all the continuous covariates were scaled to ensure numerical stability (Love et al., 2014).

3.3.5 CARseq analysis for ASD

The gene expression data (expected read counts derived from RSEM) were downloaded from Freeze1 of PsychENCODE Consortium (PEC) Capstone Collection, and the accompanying meta data and clinical data were downloaded from PsychENCODE Knowledge Portal, see Section URLs for the exact links. There are 341 samples from 100 individuals. We kept the samples from BrodmannArea 9 (BA9), including 89 samples from 85 individuals. Four individuals have duplicated samples and we chose the one with higher RIN. These 85 individuals include 42 ASD subjects and 43 controls, and they were from two brain banks: 53 from Autism Tissue Program (ATP) and 32 from NICHD, see Parikshak et al. (Parikshak et al., 2016) for more details of this dataset.

We examined the association between each potential covariates and genome-wide gene expression and found PMI and Sex are not associated with gene expression, as evidenced by a uniform distribution of p-values, therefore we removed these two covariates and used the following covariates in our analysis.

- log transformed read-depth (75 percentile of gene expression across all the genes within a sample),
- BrainBank (a factor of two levels),
- SequencingBatch (a factor of 3 levels),
- age, RIN (RNA integrity number),
- four sequencing surrogate variables (SeqSVs).

The SeqSVs, which are the notations used by Parikshak et al. (Parikshak et al., 2016) are PCs derived from sequencing QC metrics. We used 4 principal components because they explained 99% of the variance of the sequencing metrics. Prior to inclusion in the model, all the continuous covariates were scaled to ensure numerical stability (Love et al., 2014).

3.3.6 Gene set enrichment analysis

The gene set enrichment was done on REACTOME pathways downloaded from `https://www.gsea-msigdb.org/gsea/msigdb/download_file.jsp?filePath=/msigdb/release/7.1/c2.cp.reactome.v7.1.symbols.gmt`. There are originally 1,532 pathways, of which 1,090 pathways have a size between 10 and 1,000 genes.

For each cell type, we used "`fgseaMultilevel`" function in "`fgsea`" R package to simultaneously calculate p-values and normalized enrichment scores (NES) across the 1,090 pathways without any weights (fgseaMultilevel argument gseaParam = 0) in a gene list ranked by potentially CT-specific p-values from the differential expression analysis. The p-values across the 1,090 pathways were then converted to q-values using "`get_qvalues_one_inflated`" in our CARseq package. Next, we collected in a table all the candidates of the pathway-cell type pairs satisfying NES > 0 (genes in the pathway tend to have smaller p-values), and sorted them by the rank of increasing q-values and decreasing NES within each cell type. We then deduplicated the pathways by only retaining the first appearance of each pathway in the table. The top N pathway-cell type pairs were subsequently chosen. For illustrative purposes, N was picked to be 3 in our paper.

In the Main Figures, the primary differential expression method was CARseq, and the top pathways were defined by GSEA results from genes ranked by CARseq CT-specific-DE p-values. In the Supplementary Figures, we also reported heatmaps featuring top pathways defined by GSEA results based on rankings by TOAST CT-specific-DE p-values.

3.3.7 Optimization of regression coefficients through IWLS

For each gene, we have $K + MH$ regression coefficients and an overdispersion parameter ϕ_j to estimate. We will have an initial value of ϕ_j, and estimate regression coefficients and the overdispersion parameter until convergence.

From now on, assume we already know the overdispersion parameter. The problem is to maximize the log-likelihood with respect to regression coefficients β_{jk} and γ_{jhm}.

Let the cell type-specific mean expression weighted by cell fractions be

$$\mu_{jih} := d_i \exp\left(\sum_{k=1}^{K} \beta_{jk} w_{ik}\right) \hat{\rho}_{hi} \exp\left(\sum_{m=1}^{M} \gamma_{jhm} x_{ihm}\right),$$

satisfying

$$\mu_{ji} = \sum_{h=1}^{H} \mu_{jih},$$

then

$$\frac{\partial \mu_{jih}}{\partial \beta_{jk}} = w_{ik} \mu_{jih},$$

$$\frac{\partial \mu_{jih}}{\partial \gamma_{jhm}} = x_{ihm} \mu_{jih} = z_{jihm} \mu_{jih},$$

when we introduce $z_{jihm} := x_{ihm}$ for generality.

The bulk RNA-seq log-likelihood function is:

$$\ell_j = \sum_{i=1}^{n} \Big[\log\Gamma(T_{ji} + \phi_j) - \log\Gamma(\phi_j) - \log\Gamma(T_{ji} + 1) + \phi_j \log\phi_j + T_{ji} \log\mu_{ji} - (T_{ji} + \phi_j)\log(\phi_j + \mu_{ji}) \Big].$$

We further define

$$d_{ji} := \frac{\phi_j}{\mu_{ji}(\phi_j + \mu_{ji})},$$

$$r_{ji} := \frac{(T_{ji} - \mu_{ji})\phi_j}{\mu_{ji}(\phi_j + \mu_{ji})}.$$

Once we have defined μ_{jih}, z_{jihk} and r_{ji}, we can rewrite the score function:

$$\frac{\partial \ell_j}{\partial \beta_{jk}} = \sum_{i=1}^{n} w_{ik} r_{ji} \mu_{ji},$$

$$\frac{\partial \ell_j}{\partial \gamma_{jhm}} = \sum_{i=1}^{n} z_{jihm} r_{ji} \mu_{jih},$$

Now we derive the information matrix. First, let

$$B_{jik} := \frac{\partial r_{ji}}{\partial \beta_{jk}} = \frac{w_{ik}\phi_j \mu_{ji}(\mu_{ji}^2 - 2T_{ji}\mu_{ji} - T_{ji}\phi_j)}{\mu_{ji}^2(\mu_{ji} + \phi_j)^2},$$

$$G_{jigp} := \frac{\partial r_{ji}}{\partial \gamma_{jgp}} = \frac{z_{jigp}\phi_j \mu_{jig}(\mu_{ji}^2 - 2T_{ji}\mu_{ji} - T_{ji}\phi_j)}{\mu_{ji}^2(\mu_{ji} + \phi_j)^2}.$$

Second,

$$\frac{\partial^2 \ell_j}{\partial \beta_{jp} \beta_{jk}} = \sum_{i=1}^{n} \left[B_{jip} w_{ik} \mu_{ji} + \frac{\partial (w_{ik}\mu_{ji})}{\partial \beta_{jp}} r_{ji} \right] = \sum_{i=1}^{n} \left[B_{jip} w_{ik} \mu_{ji} + w_{ik} w_{ip} \mu_{ji} r_{ji} \right],$$

$$\frac{\partial^2 \ell_j}{\partial \gamma_{jgp} \beta_{jk}} = \sum_{i=1}^{n} \left[G_{jigp} w_{ik} \mu_{ji} + \frac{\partial (w_{ik}\mu_{ji})}{\partial \gamma_{jgp}} r_{ji} \right] = \sum_{i=1}^{n} \left[G_{jigp} w_{ik} \mu_{ji} + w_{ik} z_{jigp} \mu_{jig} r_{ji} \right],$$

$$\frac{\partial^2 \ell_j}{\partial \gamma_{jgp} \gamma_{jhm}} = \sum_{i=1}^{n} \left[G_{jigp} z_{jihm} \mu_{jih} + \frac{\partial (z_{jihm}\mu_{jih})}{\partial \gamma_{jgp}} r_{ji} \right].$$

Then we calculate the expectations with respect to the observed read counts T_{ji}. Observe that

$$\mathrm{E}[r_{ji}] = 0,$$

$$\mathrm{E}[B_{jik}] = -\frac{w_{ik}\phi_j\mu_{ji}}{\mu_{ji}(\mu_{ji}+\phi_j)},$$

$$\mathrm{E}[G_{jigp}] = -\frac{z_{jigp}\phi_j\mu_{jig}}{\mu_{ji}(\mu_{ji}+\phi_j)}.$$

It follows that

$$\mathrm{E}\left[\frac{\partial^2\ell_j}{\partial\beta_{jp}\beta_{jk}}\right] = -\sum_{i=1}^{n}\frac{w_{ik}w_{ip}\phi_j\mu_{ji}^2}{\mu_{ji}(\mu_{ji}+\phi_j)} = -\sum_{i=1}^{n} w_{ik}w_{ip}d_{ji}\mu_{ji}^2,$$

$$\mathrm{E}\left[\frac{\partial^2\ell_j}{\partial\gamma_{jgp}\beta_{jk}}\right] = -\sum_{i=1}^{n}\frac{w_{ik}z_{jigp}\phi_j\mu_{ji}\mu_{jip}}{\mu_{ji}(\mu_{ji}+\phi_j)} = -\sum_{i=1}^{n} w_{ik}z_{jigp}d_{ji}\mu_{ji}\mu_{jip},$$

$$\mathrm{E}\left[\frac{\partial^2\ell_j}{\partial\gamma_{jgp}\gamma_{jhm}}\right] = -\sum_{i=1}^{n}\frac{z_{jihm}z_{jigp}\phi_j\mu_{jim}\mu_{jip}}{\mu_{ji}(\mu_{ji}+\phi_j)} = -\sum_{i=1}^{n} z_{jihm}z_{jigp}d_{ji}\mu_{jih}\mu_{jip}.$$

We now have the information matrix. Let us define D_j as the n-dimension diagonal matrix with diagonal elements as $d_{ji},\ i \in \{1,2,\ldots,n\}$. Define X_j as matrix with n rows and $K+HM$ columns:

$$\begin{pmatrix}
\mu_{j1}w_{11}\cdots & \mu_{j1}w_{1k} & \cdots & \mu_{j1}w_{1K} & \mu_{j11}z_{j111} & \mu_{j12}z_{j121} & \cdots & \mu_{j1h}z_{j1hm} & \cdots & \mu_{j1H}z_{j1HM} \\
\vdots & & & \vdots & \vdots & & & & & \vdots \\
\mu_{ji}w_{i1}\cdots & \mu_{ji}w_{ik} & \cdots & \mu_{ji}w_{iK} & \mu_{ji1}z_{ji11} & \mu_{ji2}z_{ji21} & \cdots & \mu_{jih}z_{jihm} & \cdots & \mu_{jiH}z_{jiHM} \\
\vdots & & & \vdots & \vdots & & & & & \vdots \\
\mu_{jn}w_{n1}\cdots & \mu_{jn}w_{nk} & \cdots & \mu_{jn}w_{nK} & \mu_{jn1}z_{jn11} & \mu_{jn2}z_{jn21} & \cdots & \mu_{jnh}z_{jnhm} & \cdots & \mu_{jnH}z_{jnHM}
\end{pmatrix}$$

Then the information matrix is

$$X_j^T D_j X_j.$$

Define the column vector with $(K+HM)$ rows when we concatenate $\beta_{jk},\ k \in \{1,2,\ldots,K\}$ and $\gamma_{jhm},\ h \in \{1,2,\ldots,H\},\ m \in \{1,2,\ldots,M\}$ as $\tilde{\boldsymbol{\beta}}_j$, and the column vector with n rows as $\boldsymbol{Y}_j^*$ whose each entry is $T_{ji} - \mu_{ji}$. Then the score function is

$$X_j^T D_j \boldsymbol{Y}_j^*.$$

With the information matrix and score function formulated, using Fisher scoring method,

$$\tilde{\boldsymbol{\beta}}_j^{(k+1)} = \tilde{\boldsymbol{\beta}}_j^{(k)} + \left(X_j^{(k)^T} D_j^{(k)} X_j^{(k)}\right)^{-1} X_j^{(k)^T} D_j^{(k)} \boldsymbol{Y}_j^{*(k)},$$

or equivalently

$$\tilde{\boldsymbol{\beta}}_j^{(k+1)} = \left(X_j^{(k)^T} D_j^{(k)} X_j^{(k)}\right)^{-1} X_j^{(k)^T} D_j^{(k)} \left(X_j^{(k)} \tilde{\boldsymbol{\beta}}_j^{(k)} + \boldsymbol{Y}_j^{*(k)}\right).$$

Define

$$\boldsymbol{Z}_j^{(k)} = X_j^{(k)} \tilde{\boldsymbol{\beta}}_j^{(k)} + \boldsymbol{Y}_j^{*(k)},$$

Now we see that $\tilde{\boldsymbol{\beta}}_j^{(k+1)}$ is the solution of a weighted least squares with response vector $\boldsymbol{Z}_j^{(k)}$, design matrix $X_j^{(k)}$, and weights $D_j^{(k)}$. Applying this iteratively, we should now have an optimization algorithm with quadratic convergence speed.

3.3.8 Practical considerations in IWLS

In practice, however, the vanilla version of IWLS will not work as intended, to put it mildly. From now on, we will focus on a typical use case where x_{ihm}'s are group labels taking binary $\{0, 1\}$ values. Unlike negative binomial regression without cell types, where IWLS will generally converge nicely to a finite estimate as in a typical GLM framework, now we would start to see cases that the estimates of some γ_{jhm}'s tend to negative infinity. This phenomenon is more prominent when the sample size is smaller or the expected cell fraction is lower. Intuitively, since deconvolution is done on a non-log scale (Zhong and Liu, 2012), when the standard error of cell type-specific expression on a non-log scale is large, the estimate of cell type-specific expression may become zero or even negative without a non-negativity constraint implied through the log-scale modeling of expression. Meanwhile, some μ_{jih}'s may decrease to 0, which means some columns of $X_j^{(k)}$ may decrease to 0's, thus the condition number of $X_j^{(k)^T} D_j^{(k)} X_j^{(k)}$ becomes either extremely large or infinity. As a result, the IWLS iterations will get stuck somewhere when some regression coefficient estimates reach the vicinity of negative infinity, and where it may get stuck depends on the initial values.

One issue standing out is that the negative log-likelihood is not always decreasing during the iterations. To solve the issue, recall that the search direction is:

$$\boldsymbol{p}^{(k)} = \left(X_j^{(k)^T} D_j^{(k)} X_j^{(k)}\right)^{-1} X_j^{(k)^T} D_j^{(k)} \boldsymbol{Y}_j^{*(k)}.$$

We find the step size ρ^u that guarantees sufficient decrease in negative log-likelihood, also called the Armijo condition (Nocedal and Wright, 2006):

$$-\ell_j(\tilde{\boldsymbol{\beta}}_j^{(k)} + \rho^u \boldsymbol{p}^{(k)}) \leq -\ell_j(\tilde{\boldsymbol{\beta}}_j^{(k)}) + c\rho^u[-\dot{\ell}_j(\tilde{\boldsymbol{\beta}}_j^{(k)})]^T \boldsymbol{p}^{(k)},$$

where we take c as 1e−4, ρ as 0.5, and u as the smallest non-negative integer that makes the inequality hold. The process of finding the right step size is called backtracking line search.

While backtracking line search will facilitate the convergence of iterations, it does not directly address the singularity problem. To solve this, our approach ("optimization on a non-log scale") is to re-parametrize the cell type-specific regression coefficients by letting $\widetilde{\gamma}_{jhm} = \exp(\gamma_{jhm})$, $\widetilde{\gamma}_{jhm} \geq 0$:

$$\widetilde{\mu}_{jih} = d_i \exp\left(\sum_{k=1}^{K} \beta_{jk} w_{ik}\right) \prod_{m=1}^{M} \widetilde{\gamma}_{jhm}^{x_{ihm}}.$$

Then we minimize the negative log-likelihood using IWLS exactly as what is stated above, except that we need to set $z_{jihm} = x_{ihm}/\gamma_{jhm}$ and that the weighted least squares performed at each iteration need to be replaced with non-negative weighted least squares. Bounded variable least squares (Stark and L. Parker, 1995) specifying $\widetilde{\gamma}_{jhm} \geq$ 1e−30 is the actual implementation being adopted here. When boundary restraints are taken good care of, the IWLS algorithm with backtracking line search and non-negative least squares will converge to the MLE fairly easily. We further note that although not the default option used in the analyses, setting an uninformative prior of $N(0, \text{1e6})$ and using iteratively reweighted ridge regression will also lead to the convergence to the same MLE as we obtained using optimization on a non-log scale. This is the optimization strategy used in DESeq2, and more on that will be covered in the "Shrunken log fold change" section.

3.3.9 Estimation of the overdispersion parameter

To initialize the overdispersion parameter, we take the overdispersion parameter estimated by sampling x^*_{im} from x_{ihm}, $h \in \{1, \ldots, H\}$ and solving a negative binomial regression: $T_{ji} \sim f_{NB}(\mu_{ji}, \phi_j)$ and $\mu_{ji} = d_i \exp\left(\sum_{k=1}^{K} \beta_{jk} w_{ik} + \sum_{m=1}^{M} \gamma^*_{jm} x^*_{im}\right)$. We take the MLE of ϕ_j as the initial value of the overdispersion parameter. The MLE of β_{jk} is also used to initialize the regression coefficients β_{jk}, and the MLE of γ^*_{jm} is replicated across H cell types to populate the initial value of γ_{jkm}.

The bulk RNA-seq log-likelihood is:

$$\ell_j = \sum_{i=1}^{n} \Big[\log\Gamma(T_{ji}+\phi_j) - \log\Gamma(\phi_j) - \log\Gamma(T_{ji}+1) + \phi_j \log\phi_j + T_{ji}\log\mu_{ji} - (T_{ji}+\phi_j)\log(\phi_j+\mu_{ji}) \Big].$$

And define the Cox-Reid adjusted profile log-likelihood as:

$$\ell_j^{\mathrm{CR}}(\phi_j) = \ell_j(\phi_j) - \frac{1}{2}\log(\det(X_j^T D_j(\phi_j) X_j))$$

where $D_j(\phi_j)$ and X_j have been defined in the previous section "Practical considerations in IWLS" and their calculation involves combining cell type-specific covariates, cell type-independent covariates, and estimated cell fractions.

The Cox-Reid approximate conditional inference (Cox and Reid, 1987) was proposed to estimate negative binomial overdispersion in the analysis of SAGE data to control type I error in small-sample tests (Robinson and Smyth, 2008). The adjustment term $-\frac{1}{2}\log(\det(X_j^T D_j(\phi_j) X_j))$ is derived from the observed information of ϕ_j. Since then, the conditional maximum likelihood estimate has been used by edgeR (Robinson et al., 2010) and DESeq2 (Love et al., 2014) to reduce the bias of overdispersion estimation, and thus the hypothesis testing of differential expression becomes more conservative than using the MLE of overdispersion parameter.

The overdispersion parameter is updated after the regression coefficients are estimated using IWLS by doing a one-dimensional optimization on the adjusted profile likelihood of ϕ_j introduced above. The process of updating regression coefficients and then the overdispersion parameter will continue until the overdispersion parameter will not be changed for more than 0.1 on a log scale.

Since cell type-specific differential expression usually requires a larger sample size to detect cell type-specific expression variability, overdispersion estimation is done gene by gene. This is different from the default options of edgeR (Robinson et al., 2010) and DESeq2 (Love et al., 2014), where mean-overdispersion relationship across genes is leveraged to generate moderated estimation of overdispersion parameters. This approach could improve the sensitivity and specificity of differential expression tests when dealing with very small sample sizes. However, with sample of modest size, there is not a clear-cut case supporting its necessity (Zhou et al., 2011).

3.3.10 Cell type-specific tests using likelihood ratio statistics

We introduce likelihood ratio statistics to test cell type-specific expression. Although the general testing framework is very flexible to be inclusive of all kinds of tests discussed in TOAST (Li et al., 2019), we only emphasize on the application that the test is about whether there is cell type-specific differential expression analysis across groups.

Consider the cell type-specific expression:

$$\widetilde{\mu}_{jih} := d_i \exp\left(\sum_{k=1}^{K} \beta_{jk} w_{ik}\right) \exp\left(\sum_{m=1}^{M} \gamma_{jhm} x_{ihm}\right).$$

Under the full model, γ_{jhm} is the cell type-specific mean expression in cell type h and group m, controlled for other covariates. Then $x_{ihm} = 1$ when sample i belongs to group m, and $x_{ihm} = 0$ otherwise.

Under the reduced model, within a certain cell type h, there is no differential expression across groups. Instead, the expression is always γ_{jhm} regardless of which group sample i belongs to. Without loss of generality, let us suppose that the cell type-specific expression stays at γ_{jh1}. In this way, $x_{ih1} = 1$ and $x_{ihm} = 0, m \in \{2, \ldots, M\}$, and the corresponding γ_{jhm}'s are no longer possible to estimate since they are not included in the reduced model as a matter of fact. Thus, during estimation of regression coefficients, these pairs of h and m are left out of the IWLS algorithm.

Once we have the log-likelihood under the full model and the reduced model, we can calculate the likelihood ratio statistics and compare it to a chi-squared distribution with $M - 1$ degrees of freedom. This approximation to the distribution of the statistics is asymptotic (a t distribution has been proposed as an alternative in the literature), and is only appropriate when the true parameter is not on the boundary. Biologically speaking, the cell type-specific expression should indeed be always positive. However, in a finite sample case, the closeness to a boundary case might change the null distribution into a mixture distribution without a closed-form expression and more complex than a mixture chi-squared distribution (Self and Liang, 1987; Molenberghs and Verbeke, 2007). With that being said, when the sample size is moderately large, any departure from the asymptotic distribution is commonly too small to warrant much scrutiny.

3.3.11 Shrunken log fold change

While a sufficiently small p-value from a differential expression test can be interpreted as statistical significance in the association between expression and the parameter of interest, it does not indicate the biological strength of association. Log fold change (LFC) is up to this task. It is customary to plot LFC and p-value of genes on a scatterplot—aptly named as volcano plot—and dictate which genes are candidates for further investigation using both LFC and p-value thresholds. LFC can also be used in gene ontology (GO) or gene set enrichment analysis (GSEA).

Nevertheless, the interpretation of LFC can be difficult especially when the sample size is small and the variability is large. Smaller studies tend to report large LFCs even when the differential expression is absent (The Brainstorm Consortium et al., 2018). Shrunken log fold change is a stabilized estimation of the strength of differential expression. Notably, since deconvolution is done on a non-log scale, without any shrinkage, the estimated cell type-specific expression can go to zero due to the (implicit) constraint that cell type-specific expression is non-negative, and the raw LFC goes to positive or negative infinity. A similarly hard-to-interpret scenario can be found in TOAST results, where the estimated cell type-specific expression can even become negative, and thus LFC has no definition. These ill-posed problems severely limit the use cases of raw LFC to quantify the strength of cell type-specific expression. With a prior that is sufficiently informative, such as an empirical Bayes prior, the shrunken log fold change is generally finite and easier to interpret. Furthermore, if the whole experiment is replicated or if it is replicated with a larger sample size, the consistency between replicated studies would be better if the shrunken LFC, instead of the raw LFC, is used as the measure of the strength of differential expression (Love et al., 2014).

The shrunken log fold change implemented in CARseq largely follows that in DESeq2 (Love et al., 2014), though there are some subtleties of CARseq implementation. The additive structure of mixture expression depending on cell type-specific expression in CARseq requires some weak shrinkage on the cell type-specific group mean. Though an unconventional choice, it compresses the estimated group mean towards zero on a log scale and mitigates the optimization difficulties when any estimate of cell type-specific expression approaches the boundary. Another pertinent distinction lies in CARseq forgoing the moderation of overdispersion parameters. Although this does not affect the raw LFC, the adaptation of moderated overdispersion parameter can further stabilize the shrunken LFC, resulting the impression that DESeq2 produces more aggressively shrunken log fold change.

Recall that when setting the design matrix, to code a factor with multiple levels, the default choice is to set the first level as intercept and code each other level as a contrast with the first. While the designation of which level as intercept will not affect estimation and testing without a prior, the asymmetric application of prior to contrasts will be problematic. Therefore, we borrow the concept of expanded design matrix from DESeq2. Essentially, we set the group mean γ_{jh0} of cell type-specific expression as intercept, and code each level as a constrast with the group mean by γ_{jhm}. Without setting a prior on the contrasts, the design matrix is singular and the linear model is not estimable. With a prior on the contrasts, however, the estimation can still proceed. The contrasts have been applied with a comparatively strong prior and the group mean has a weaker prior.

In a normal design matrix, there are HM cell type-specific variables. In an expanded design matrix, for each cell type h, there are M contrasts γ_{jhm} and one group mean γ_{jh0}, so there are $H(M+1)$ cell type-specific variables altogether. When applying the empirical Bayes prior to log fold change, the K cell type-independent variables are not being penalized on. This is because we assume the parameters of interest are the cell type-specific ones. The posterior of LFC will be shrunken towards zero with a prior of zero-centered normal distribution:

$$\widehat{\boldsymbol{\beta}_j}^{\text{MAP}} = \arg\max \left[\ell_j(\tilde{\boldsymbol{\beta}}_j^*) + P(\tilde{\boldsymbol{\beta}}_j^*)\right],$$

where $\tilde{\boldsymbol{\beta}}_j^*$ is the column vector with $K + H(M+1)$ rows when we concatenate $\beta_{jk},\ k \in \{1, 2, \ldots, K\}$ and $\gamma_{jhm},\ h \in \{0, 1, 2, \ldots, H\},\ m \in \{1, 2, \ldots, M\}$, and

$$P(\tilde{\boldsymbol{\beta}}_j^*) = \sum_{h=1}^{H} \frac{-\gamma_{jh0}^2}{2\sigma_0^2} + \sum_{h=1}^{H} \sum_{m=1}^{M} \frac{-\gamma_{jhm}^2}{2\tau_h^2}.$$

To estimate the empirical Bayes prior, we first estimate the MLEs on a log scale of cell type-specific expression across genes. Any infinite values are excluded in the calculation of empirical distribution afterwards. We set a comparatively weak prior on the cell type-specific group means as $N(0, 10^4)$. The inclusion of such a prior is to require the posterior of the cell type-specific group means to be finite. Then we compute all combinations of contrasts between levels of each cell type h, collect the constrasts belonging to each cell type together, and align the 0.95 quantile of the empirical distribution of the absolute value of all the cell type-specific contrasts with the 0.975 quantile of a zero-centered normal distribution $N(0, \sigma^2)$. Note that the prior is the same for different cell types, because we do not have a preponderance of evidence favoring the other choice. We now have a comparatively strong prior on the cell type-specific contrasts.

The update rule in iteratively reweighted ridge regression (Park, 2006; Love et al., 2014) is:

$$\tilde{\boldsymbol{\beta}}_j^{*(k+1)} = \left(X_j^{*(k)^T} D_j^{(k)} X_j^{*(k)} + \boldsymbol{\lambda} I\right)^{-1} X_j^{*(k)^T} D_j^{(k)} \left(X_j^{*(k)} \tilde{\boldsymbol{\beta}}_j^{(k)} + \boldsymbol{Y}_j^{*(k)}\right),$$

where I is the identity matrix, $\boldsymbol{\lambda}$ is a vector of length $K + H(M+1)$ obtained by taking $\tilde{\boldsymbol{\beta}}_j^*$ and replacing β_{jk} with 0 and γ_{jhm} with $1/\sigma_m^2$, D_j is the n-dimension diagonal matrix with diagonal elements as

$d_{ji} = \frac{\phi_j}{\mu_{ji}(\phi_j + \mu_{ji})}$, $i \in \{1, 2, \ldots, n\}$, $\boldsymbol{Y}_j^*$ is a column vector with n rows whose each entry is $T_{ji} - \mu_{ji}.$, and X_j^* is a matrix with n rows and $K + H(M + 1)$ columns:

$$\begin{pmatrix}
\mu_{j1}w_{11} & \cdots & \mu_{j1}w_{1k} & \cdots & \mu_{j1}w_{1K} & \mu_{j11}x_{110} & \mu_{j12}x_{120} & \cdots & \mu_{j1h}x_{1hm} & \cdots & \mu_{j1H}x_{1HM} \\
\vdots & & & & \vdots & \vdots & & & & & \vdots \\
\mu_{ji}w_{i1} & \cdots & \mu_{ji}w_{ik} & \cdots & \mu_{ji}w_{iK} & \mu_{ji1}x_{i10} & \mu_{ji2}x_{i20} & \cdots & \mu_{jih}x_{ihm} & \cdots & \mu_{jiH}x_{iHM} \\
\vdots & & & & \vdots & \vdots & & & & & \vdots \\
\mu_{jn}w_{n1} & \cdots & \mu_{jn}w_{nk} & \cdots & \mu_{jn}w_{nK} & \mu_{jn1}x_{n10} & \mu_{jn2}x_{n20} & \cdots & \mu_{jnh}x_{nhm} & \cdots & \mu_{jnH}x_{nHM}
\end{pmatrix}.$$

3.3.12 Miscellaneous notes on the CARseq model

3.3.12.1 Relation between CARseq model and TOAST model

The major difference between CARseq and TOAST is that the former employs a negative binomial distribution while the latter uses a linear model that is consistent with normal distribution assumption. In addition, the way to connect the mean expression in bulk tissue to CT-specific expression are different. Here we demonstrate the relation of these two approaches.

Note that μ_{ji} is the expected total read count of gene j and sample i. As TOAST does not allow for cell type-independent variables β_{jk}'s, which are actually possible to be recast into cell type-specific variables at the expense of added degrees of freedom, and it does not model sample-level read depth d_i's, we will adjust for them when comparing the CARseq model and the TOAST model.

$$\frac{\mu_{ji}}{d_i \exp\left(\sum_{k=1}^{K} \beta_{jk} w_{ik}\right)} = \sum_{h=1}^{H} \hat{\rho}_{hi} \exp\left(\sum_{m=1}^{M} \gamma_{jhm} x_{ihm}\right),$$

Let $\xi_{jhm} = \exp(\gamma_{jhm}) - 1$. It follows that

$$\frac{\mu_{ji}}{d_i \exp\left(\sum_{k=1}^{K} \beta_{jk} w_{ik}\right)} = \sum_{h=1}^{H} \left[\hat{\rho}_{hi} \prod_{m=1}^{M} (1 + \xi_{jhm})^{x_{ihm}}\right].$$

When $|\xi_{jhm}| \ll 1$ and $|x_{ihm}\xi_{jhm}| \ll 1$, the above equation can be approximated by:

$$\sum_{h=1}^{H}\left[\hat{\rho}_{hi}\prod_{m=1}^{M}(1 + x_{ihm}\xi_{jhm})\right] \approx \sum_{h=1}^{H}\left[\hat{\rho}_{hi}\Big(1 + \sum_{m=1}^{M} x_{ihm}\xi_{jhm}\Big)\right].$$

In the typical circumstance where $x_{ihm} \in \{0, 1\}$ are group indicators, the "approximately equal to" relations are all "equal to" relations.

We now arrive at a form very close to how the measurement in $TOAST$ is modeled with cellular proportion as main effects, which are parametrized below using η_{hi}'s, and proportion by covariate as interactions:

$$\sum_{h=1}^{H}\left[\sum_{m=1}^{M}(\hat{\rho}_{hi}\eta_{jh} + x_{ihm}\hat{\rho}_{hi}\xi_{jhm})\right].$$

3.3.12.2 Connections between cell fractions, cell type-specific transcript fractions and cell size

We explain how to convert between cell fractions in the model involving read counts, ρ_{hi}, and cell fractions in the model involving TPM, ρ_{hi}^{TPM}. Suppose j is a subscript for gene, i is a subscript for sample, and h is a subscript for cell types. The reference of purified cell types are denoted using γ_{jh}. Gene lengths are ℓ_j.

We can obtain cell fractions when we have counts from both mixture and purified cell types. Specifically, the counts of purified cell types are the average of total counts of all cells belonging to a cell type. The reference of purified cell types can be prepared using single cell RNA-seq data. In this way, cell types with a higher expression will have a higher total read counts, and there is no need to adjust for different cell sizes of cell types.

When we obtain cell fractions by deconvolving mixture expression in TPM, such as by CIBERSORT or ICeD-T, we need to adjust for cell size to obtain cell fractions in the literal sense.

Suppose the total read counts follow a negative binomial distribution:

$$T_{ji} \sim f_{NB}(d_i \sum_{h=1}^{H} \rho_{hi}\gamma_{jh}, \phi_j).$$

Taking expectation gives:

$$\mathrm{E}[T_{ji}] = d_i \sum_{h=1}^{H} \rho_{hi}\gamma_{jh},$$

$$\mathrm{E}[T_{ji}/\ell_j] = d_i \sum_{h=1}^{H} (\rho_{hi}\gamma_{jh}/\ell_j).$$

Since the total number of genes G is very large, using Approximations for Mean and Variance of a Ratio, we get:

$$\mathrm{E}\Big[\frac{T_{ji}/\ell_j}{\sum_{j=1}^{G}(T_{ji}/\ell_j)}\Big] \approx \frac{\mathrm{E}[T_{ji}/\ell_j]}{\mathrm{E}[\sum_{j=1}^{G}(T_{ji}/\ell_j)]} = d_i \sum_{h=1}^{H} \rho_{hi} \frac{\sum_{j=1}^{G}(\gamma_{jh}/\ell_j)}{\mathrm{E}[\sum_{j=1}^{G}(T_{ji}/\ell_j)]} \frac{(\gamma_{jh}/\ell_j)}{\sum_{j=1}^{G}(\gamma_{jh}/\ell_j)}.$$

Define

$$r_{hi} = \frac{\sum_{j=1}^{G}(\gamma_{jh}/\ell_j)}{\mathrm{E}[\sum_{j=1}^{G}(T_{ji}/\ell_j)]},$$

and we have mixture expression and cell type-specific expression in TPM:

$$\mathrm{E}[T_{ji}^{\mathrm{TPM}}] = d_i \sum_{h=1}^{H} \rho_{hi} r_{hi} \gamma_{jh}^{\mathrm{TPM}}.$$

If we add up all the genes

$$\sum_{j=1}^{G} \mathrm{E}[T_{ji}^{\mathrm{TPM}}] = \sum_{j=1}^{G} d_i \sum_{h=1}^{H} \rho_{hi} r_{hi} \gamma_{jh}^{\mathrm{TPM}},$$

we get

$$\sum_{h=1}^{H} d_i \rho_{hi} r_{hi} = 1.$$

Define

$$\rho_{hi}^{\mathrm{TPM}} = d_i \rho_{hi} r_{hi},$$

which follows that

$$\rho_{hi}^{\mathrm{TPM}} \propto \rho_{hi} \sum_{j=1}^{G} (\gamma_{jh}/\ell_j).$$

This justifies us to define the total number of transcripts in a cell:

$$s_h = \sum_{j=1}^{G} (\gamma_{jh}/\ell_j)$$

as the cell size to convert between ρ_{hi}^{TPM} and ρ_{hi}.

3.3.13 FDR control procedure

We use q-value to control FDR (Storey and Tibshirani, 2003). The calculation of q-value requires an estimate of the overall proportion of null p-values $\hat{\pi}_0$. We use the following formula that specifically accommodates the situation where a proportion of p-values equal to 1, implemented in function `get_qvalues_one_inflated` of R package `CARseq`:

$$\hat{\pi}_0 = (\text{proportion of p value} = 1) + 2 \times (\text{proportion of p value} > 0.5 \text{ and } < 1).$$

3.3.14 URLs

snRNA-seq data for CT-specific expression reference,

file `human_MTG_gene_expression_matrices_2018-06-14.zip`, downloaded from

`http://celltypes.brain-map.org/api/v2/well_known_file_download/694416044`.

CommonMind Consortium (CMC) Knowledge Portal:

`https://www.synapse.org/#!Synapse:syn2759792/wiki/69613`.

CMC gene expression data:

`https://www.synapse.org/#!Synapse:syn3346749`

CMC gene expression meta data:

`https://www.synapse.org/#!Synapse:syn18103174`

CMC clinical data:

`https://www.synapse.org/#!Synapse:syn3275213`.

PEC Capstone Collection:

`https://www.synapse.org/#!Synapse:syn12080241`

UCLA-ASD gene expression data:

`https://www.synapse.org/#!Synapse:syn8365527`

UCLA-ASD gene expression meta data:

`https://www.synapse.org/#!Synapse:syn5602933`

UCLA-ASD clinical data:

`https://www.synapse.org/#!Synapse:syn5602932`

SFARI ASD risk genes:

`https://gene.sfari.org/database/human-gene/`

3.3.15 Code availability

The codes for generating CT-specific gene expression reference panel are included in GitHub repository `scRNAseq_pipelines` (`https://github.com/Sun-lab/scRNAseq_pipelines`). we have analyzed three scRNA-seq datasets: MTG, dronc, and psychENCODE, and the codes were saved in corresponding folders. The codes to compare different references and generate final references were saved in folder `_brain_cell_type`.

The codes for CARseq analyses (including simulation, and analyses of SCZ and ASD datasets) were included in GitHub repository `CARseq_pipelines` (`https://github.com/Sun-lab/CARseq_pipelines`). The file `reproducible_figures.html` has the code to generate most Figures in this paper. The R package CARseq were deposited at GitHub repository `CARseq` (`https://github.com/Sun-lab/CARseq`).

3.4 Supplementary Materials for Data Analyses

3.4.1 Details of simulations

3.4.1.1 Steps of generating simulations

1. For every gene, fit real mixture read count using negative binomial (NB) regression with an offset term of log read depth, with an effect of RNA integrity number (RIN), but without the effects representing the cell type composition.

2. Collect the intercept terms across all the genes, and fit the mean and standard deviation of gene expression.

3. Cell type-specific gene expressions in the reference matrix for each gene, along with the RIN effect size, overdispersion parameter and the gene length, are simulated using a multivariate log-normal distribution. The correlation between reference expression of different cell types is set as the median of all pairwise correlations between reference gene expression of different cell types. This is to mirror the correlation structure of the reference matrix, whose entries are computed from MTG single cell data with 6 labeled cell types. The correlations account for gene lengths and similarity between cell types.

4. Compute expected mixture expression using the reference matrix from (3), cell fractions we have simulated before, and RIN per sample and its gene-specific effect size we have simulated before.

5. Generate mixture read counts using NB distribution with overdispersions simulated from a log-normal distribution fitted using overdispersion estimates from (1).

3.4.1.2 Details of simulations

We benchmarked CARseq against csSAM 1.2.4, TOAST 0.99.8, CIBERSORTx high-resolution mode, and DESeq2 1.24.0. When expression is quantified from next-generation sequencing data, these methods of testing differential expression typically either require read counts (modeled by CARseq and DESeq2) or TPM (modeled by CIBERSORTx, csSAM, and TOAST). If cell fractions are known, we can conduct cell type-specific differential expression tests. While reference-free methods have been documented (Zaitsev et al., 2019), in most practices, reference matrix containing the expression of purified cells are preferred to estimate the cell fractions of cell types.

We generated read count data of bulk tissue as a mixture of three cell types. The pre-specified number of samples was selected to be 50, 100, and 200, and then was divided equally into case and control groups. The total number of genes was 10,000, among which 2,000 genes had spiked-in cell type-specific differential expression between groups.

The read counts in mixture samples were generated from a negative binomial distribution:

$$t_{ji} \sim \mathrm{NB}(\mu_{ji}, \theta_j),$$

with the mean structure being

$$\mu_{ji} = d_i \exp\left(\sum_{k=1}^{K} \beta_{jk} w_{ik}\right) \sum_{h=1}^{H} \left(\hat{\rho}_{hi} \prod_{m=1}^{M} \gamma_{jhm}^{x_{ihm}}\right).$$

We set the simulation to include $H = 3$ cell types, $M = 2$ effects that are cell type-specific (case/control groups), and $K = 1$ batch effect that is cell type-independent (RIN).

We generated cell fractions ρ_{hi} using Dirichlet distribution with parameters estimated from cell fractions estimated using ICeD-T without weights and adjusted by cell sizes. Here the cell fractions ρ_{hi} reflected the actual proportion of cells of each cell type, rather than the proportion of transcripts that can be assigned to a cell type ρ_{hi}^{TPM} that resulted from the deconvolution on the scale of TPM. If the cell fraction estimates were obtained from the deconvolution of TPM, which would be the case when ICeD-T or CIBERSORT was used, an adjustment of cell size was applied through $\rho_{hi} = \frac{\rho_{hi}^{\mathrm{TPM}}/s_h}{\sum_h \rho_{hi}^{\mathrm{TPM}}/s_h}$. See “Notes about cell fractions and cell size” for a more detailed description.

The number of cell types H was 3, a simplification of the reference matrix with 6 cell types constructed from MTG single cell data. This was achieved by retaining excitatory (“Exc”) and inhibitory

("Inh") neurons while bundling the other cell types ("Other"), mainly glial cells, together after using ICeD-T or CIBERSORT to estimate cell fractions in real data analysis. The collapsed cell fractions are then used to fit a Dirichlet distribution. The estimate of the concentration parameter of the Dirichlet distribution was (30.10, 4.94, 12.91), which was subsequently used to generate cell fractions of each sample.

To set realistic parameters in data simulation, we used CARseq to fit a model of read counts with CMC data as the mixture data and MTG single cell data as the reference. For every gene, we fitted the model without cell types and differential expression between sample groups:

$$t_{ji} \sim \mathrm{NB}(\mu_{ji}, \theta_j),$$

and the mean structure is

$$\mu_{ji} = d_i \exp(\beta_{j1} w_{i1}) \gamma_j.$$

d_i is sample-level read depth taken as the third quartile of highly expressed genes, which are 20,614 genes with the third quartile of read counts across samples being larger than 20. w_{i1} is the RIN per sample, a batch effect to adjust for. Both d_i and w_{i1} are known when fitting the model using CMC data.

For observed read counts t_{ji}, we fitted a negative-binomial model using CARseq to obtain estimates of the triple $(\beta_{j1}, \gamma_j, \theta_j)$.

In the scatterplot of the triple $(\beta_{j1}, \gamma_j, \theta_j)$ across 20,614 highly expressed genes in the log scale (one exception is β_{j1}, which is already parametrized in the log scale), we noticed that multivariate Gaussian distribution is a good approximation after we remove 31 genes with a much higher overdispersion than the others. We compute and base our simulation on the estimated covariance matrix of expression, batch effect, and overdispersion.

After getting an idea of how multiple parameters jointly distributed in real data, we went on to generate data that could mirror what we had found. The β_{j1} estimates above do not involve cell types. This was a deliberate choice: if we directly fitted a model with cell type-specific expression using

CARseq, we would be facing with a large proportion of zeros in estimated cell type-specific expression. These zeros arise because it would be hard to accurately estimate cell type-specific expression when the variance of cell fractions was low.

To mitigate the problem, when generating the reference matrix, we assumed that the expected expression of every cell type γ_{jh} followed a log-normal distribution with the same mean and variance across cell types, and the mean and variance of mixture expression γ_j can be used as a substitute. However, we needed to add correlation between cell types to make the expression pattern to be more realistic, as in a large number of genes, the gene expression would be similar across cell types, the most prominent examples being the housekeeping genes. We also noticed that gene expression and gene lengths are positively correlated. Gene lengths were needed to calculate TPM to generate the input data of csSAM and CIBERSORTx.

To estimate the correlation between cell type-specific gene expression γ_{jh}, and the correlation between gene expression γ_{jh} and gene lengths ℓ_j, we used MTG single cell data to create an average cell expression for 6 cell types, and estimated the correlation structure between cell type-specific expression $\gamma_{jh},\ h \in \{1, \dots, 6\}$, and gene lengths ℓ_j, all in log scale. We then took the median of all pairwise correlations of gene expression between cell types and fill it into the correlation matrix to simulate data, and likewise take and use in simulation the median of all correlations between cell type-specific expression and gene lengths. We assumed the correlation was zero between gene lengths and batch effect, and between gene lengths and overdispersion. This completed the specification of covariance matrix of three cell types, batch effect, overdispersion and, gene lengths. Thus we generated the tuple $(\gamma_{j1}, \gamma_{j2}, \gamma_{j3}, \beta_{j1}, \theta_j, \ell_j)$ under log scale except for β_{j1} using a multivariate Gaussian distribution:

$$\begin{bmatrix} \log \gamma_{j1} \\ \log \gamma_{j2} \\ \log \gamma_{j3} \\ \beta_{j1} \\ \log \theta_j \\ \log \ell_j \end{bmatrix} \sim N\left(\begin{bmatrix} -0.60 \\ -0.60 \\ -0.60 \\ -0.09 \\ 2.40 \\ 7.98 \end{bmatrix}, \begin{bmatrix} 3.39 & 2.58 & 2.58 & -0.24 & 0.19 & 0.35 \\ 2.58 & 3.39 & 2.58 & -0.24 & 0.19 & 0.35 \\ 2.58 & 2.58 & 3.39 & -0.24 & 0.19 & 0.35 \\ -0.24 & -0.24 & -0.24 & 0.05 & 0.07 & 0.00 \\ 0.19 & 0.19 & 0.19 & 0.07 & 0.64 & 0.00 \\ 0.35 & 0.35 & 0.35 & 0.00 & 0.00 & 0.83 \end{bmatrix} \right).$$

The batch effect variable, RIN w_{i1}, was generated using a normal distribution with mean and variance estimated from variables collected in CMC data:

$$\log w_{i1} \sim N(7.61, 0.89).$$

We generated read depths d_i using log-normal distribution with mean and variance estimated from mixture expression of CMC data:

$$\log d_i \sim N(6.70, 0.27).$$

To create input data for CIBERSORTx, we needed the signature matrix of a few hundred of signature genes where only one of the several cell types dominated the mixture expression. To select the signature genes, for each cell type, we collected the top 100 genes with the largest fold change of its gene expression compared to gene expression from all other cell types in the non-log scale. As an example, the signature genes associated with cell type 1 are the top 100 genes sorted in descending order by the fold change $\gamma_{j1}/(\gamma_{j2} + \gamma_{j3})$.

The signature genes were then merged into a list and deduplicated, a step that might not be needed since we did not find any duplicated genes in the simulation. The TPM of both mixture and reference gene matrices were calculated with all genes. We also incorporated the batch effect from RIN to construct adjusted reference matrix for CIBERSORTx, which brought relatively accurate

CIBERSORTx cell fraction estimates:

$$\widetilde{\gamma}_{jh} = \frac{1}{n}\sum_{i=1}^{n} \exp(\beta_{j1} w_{i1})\gamma_{jh},$$

$$\widetilde{\gamma}_{jh}^{\text{TPM}} = \frac{\widetilde{\gamma}_{jh}/\ell_j}{\sum_{j=1}^{G}(\widetilde{\gamma}_{jh}/\ell_j)}.$$

We specified the pattern of differential expression we would like to spike in as $\mathbf{c} = (c_1, c_2, c_3)$ so that we were able to obtain group-specific mean γ_{jhm} from reference mean γ_{jh}. Among 10,000 genes, suppose 8,000 genes were not differentially expressed between case and control samples, which means $\gamma_{jhm} = \gamma_{jh}$; 1,000 genes expressed differently (measured by fold change) in one cell type among case samples, which means $\gamma_{jh2} = c_h\gamma_{jh1} = c_h\gamma_{jh}$; and 1,000 genes expressed differently in the same cell type among control samples, which means $\gamma_{jh1} = c_h\gamma_{jh2} = c_h\gamma_{jh}$. Then we calculated μ_{ji} and generated the observed mixture read count matrix with each entry t_{ji} following a negative binomial distribution. Using the same approach, we implemented more complicated patterns of differential expression where more cell types were differentially expressed. The test we implemented in the simulation studies has the null hypothesis $\gamma_{jh1} = \gamma_{jh2}$ for any gene j and cell type h, which says the cell type-specific mean are the same across two groups when controlled for other covariates.

3.4.1.3 Methods to benchmark

1. CARseq (w/ and w/o clinical variables)

   `R/simulation_test_CARseq.R`

   Input 1: read counts; true cell fractions.

   Input 2: read counts; cell fractions from CIBERSORTx without using B mode and adjusted for cell size.

   Output: p-value for every gene/cell type pair.

   To use CARseq, the cell fraction estimates are required. We recommend to use ICeD-T or CIBERSORT to obtain the cell fraction estimates $\hat{\rho}_{hi}$ that sum up to 1 for every sample i.

Here the cell fractions are true cell fractions. Since CIBERSORTx will first estimate cell fractions before high-resolution expression analysis, and to reflect the fact that we actually would not know true cell fractions when using CARseq, we also include a set of analysis of CARseq accounting for RIN while using cell fractions from CIBERSORTx without B mode correction.

CARseq has the a higher power than csSAM while controlling for type I error. One possible reason is CARseq can leverage discrete distributions, which would provide a higher power. Another reason is csSAM does not allow for the incorporation of known batch effects. The third reason is csSAM uses permutation test, and the power is generally lower than model-based tests when the model holds for the data.

2. csSAM (w/o clinical variables) csSamWrapper with nonNeg = TRUE as other methods does not allow negative expression values

   `R/simulation_test_csSAM.R`

   Input: TPM; the true cell fractions in the scale of TPM (can be alternatively called the cell type-specific transcript fractions).

   When we assume the cell sizes are all the same across different cell types, the cell fractions do not need to be adjusted.

   Output: FDR for every gene/cell type pair.

3. CIBERSORTx high resolution "Tutorial 5 - Impute Gene Expression, High-Resolution Mode" (w/ and w/o clinical variables; w/ and w/o batch correction)

   `R/simulation_test_CIBERSORTx.R`

   Input: TPM; signature matrix adjusted for batch effect RIN in TPM.

   Output: p-value for every gene/cell type pair after we conduct linear regression for purified expression on sample label while adjusting for the batch effect RIN.

CIBERSORTx high-resolution mode can be considered as a more heuristic approach instead of to a rigorous statistical model. Therefore, it is no surprise that type I error cannot be properly controlled in other cell types not being differentially expressed themselves.

4. TOAST (w/ and w/o clinical variables; using TPM)

   `R/simulation_test_TOAST_TPM.R`

   Input 1: read counts; true cell fractions.

   Input 2: TPM; true cell fractions in the scale of TPM.

   Output: p-value for every gene/cell type pair.

   TOAST uses a linear regression model. Based on our experience, it is better to supply TPM instead of read counts, as TPM is already adjusted for read depth, while the read depth needs to be included as a covariate when modeling read counts. Read counts, in contrast, is better to be modeled using a generalized linear model to fully utilize the mean-variance structure. Unless specifically mentioned, we use TPM instead of read counts as the input of TOAST. The code to run TOAST using read counts is included as `R/simulation_test_TOAST_count.R`.

   As TOAST used a linear regression model, it actually allows for negative estimates of cell type-specific expression. While this is not biologically possible, it does not affect statistical modeling. This is not the case in CARseq, where the cell type-specific expression has to be non-negative to satisfy the model requirements of a non-negative mean in a negative binomial regression.

   TOAST claimed that there is no consensus on whether to deconvolve under log scale or non-log scale. Based on our knowledge, for RNA-seq data, it is a general practice to deconvolve under a non-log scale while modeling the covariates under a log scale.

   The authors of TOAST wanted to make the effect of additional covariates proportional to baseline expression, but due to the limitation of linear regression, they had to model them using interaction terms. While this makes their model more flexible, we worry about the splurge on degrees of freedom especially when these parameters are not of interest in the sense of testing.

3.4.1.4 Explanation of rare occurrences of inflated type I error when the covariate is not provided

To only highlight major problems in the figures, we need to display the FDR in a more stable way. When its definition is ill-defined (0/0 = NaN) or almost ill-defined (the denominator, or the total discoveries, is no greater than 5), the FDR is set to 0 and the sensitivity is set to 0.

In the plots, FDR and type I error refer to false discoveries in the 8000 non-DE genes in a certain cell type. Be careful that “power” and “sensitivity” refer to discoveries in the 2000 DE genes, which can be either true or false depending on which cell type is differentially expressed; the information is coded in the “DE pattern” of fold changes between two groups.

In general, the simulation performs as expected. CARseq strikes a good balance between controlling type I error and being powerful.

We find that CARseq would produce inflated type I error in very few cases when the covariate (RIN) is not provided. The inflated type I error we see in simulation setup `n_200_DE_pattern_2_1_1_replicate_1` (without covariates) is caused by the collinearity between RIN and cell fractions of cell type 2 in one group. In general, in a regression framework, if there is considerable correlation between an effect to test and an effect not included in the model, then the problem of inflated type I error could arise. When RIN is not included in the model, the variation in expression attributed to RIN effect is instead falsely attributed to differential expression in cell type 2:

```
> cor.test(clinical_variables[101:200], rho[101:200,2])


    Pearson's product-moment correlation


data:  clinical_variables[101:200] and rho[101:200, 2]
t = -3.0656, df = 98, p-value = 0.002807
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4652033 -0.1055285
```

```
sample estimates:
      cor
-0.295815


> cor.test(clinical_variables[1:100], rho[1:100,2])


    Pearson's product-moment correlation


data:  clinical_variables[1:100] and rho[1:100, 2]
t = 1.5444, df = 98, p-value = 0.1257
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.04359692  0.34025758
sample estimates:
      cor
0.1541411
```

As this symptom is not a methodological pitfall of CARseq, other algorithms can also manifest inflation of type I error when the variable to test is correlated to a batch effect not incorporated in the model. Since the cell type-specific test in csSAM is conceptually the same as CARseq without batch effects, when `CARseq_without_RIN` is plagued with inflated type I error, csSAM is also bound to fail. An example can be found in simulation setup `n_200_DE_pattern_2_1_1_replicate_1`. When randomly generated RIN is highly correlated with group labels, DESeq2 could have inflated type I error among non-DE genes. An example can be found at `DESeq2_without_RIN` in simulation setup `n_100_DE_pattern_4_1_1_replicate_1`.

3.4.1.5 Code description for simulations

First, R/simulation_step1_get_distribution_of_parameters_from_real_data.R fits models from CMC data using MTG single cell data as reference. Second, the joint distribution of fitted parameters is used to generate simulation data in R/simulation_step2_simulate_data.R. There are separate code snippets starting with R/simulation_step3 to run each method in the simulation folder. Then we use R/simulation_step4_compare_methods_multiple_replicates.R to calculate metrics to compare the methods across ten replicates. The methods of using CIBERSORTx high resolution mode and TOAST with read counts are summarized in R/simulation_step4_compare_methods_multiple_replicates.R where only one replicate has been investigated.

3.4.2 Details of real data analyses

3.4.2.1 Deconvolution of bulk tissue using reference from single cell data

Estimates of cell fractions were obtained using a reference of cell type-specific expression constructed from MTG snRNA-seq data, which was generated using SMART-Seq v4 Ultra Low Input RNA Kit, which is an improved version of SMART-seq2 protocol (Hodge et al., 2019). We used both ICeD-T (without weights) and CIBERSORT to estimate cell fractions.

We analyzed the MTG snRNA-seq data (https://github.com/Sun-lab/scRNAseq_pipelines/blob/master/MTG/human_MTG.html followed by an R code step1_expression_signature.R in the same folder) using a pipeline that is slightly different from the original paper, mostly based on Bioconductor workflows for scRNAseq (Lun et al., 2016). Then we clustered all the cells/nuclei using K-means, ranging from 10 to 20 clusters. Based on manual inspection, we choose to compare the annotated cell type labels with K-means with 15 clusters, as shown below.

```
      Astro Endo  Exc  Inh Micro Oligo  OPC unknown
  1       0    0    0 1279     0     0    0      15
  2       0    0 1867    0     0     0    0      24
  3       0    1    8   11     0   310    1      12
  4     0    0  260    0      0     0    0        22
  5   287    0   12    1      0     2    4        21
  6     0    0 1494    0      0     0    0        73
  7     0    0 1483    0      0     0    0        17
  8     0    0 1552    1      0     0    0        15
  9     0    0    1 1210      0     0    0         4
  10    0    0    2    4     62     1    0         9
  11    0    0    0  835      0     0    0         3
  12    1    8   16  807      1     0  233        35
  13    0    0  326    1      0     0    0         7
  14    0    0 1798    1      0     0    0        38
  15    0    0 1654    1      0     0    0        28
```

There is a high consistency between our clustering results and annotated cell types. To generate the cell type-specific gene expression profile for each cell type, we select those clusters that are either is the largest cluster for this cell type or includes more than 200 cells of this cell type. This helps us filter out some cells for each cell type. In total, we kept 15,465 nuclei for the following analysis and they are separated into 7 cell types:

```
  Cell_Type nCells_All
1       Inh       4131
2       Exc      10434
3     Oligo        310
4       OPC        233
5     Astro        287
6     Micro         62
7      Endo          8
```

We compared this MTG snRNAseq data (SMART-Seq v4) with another snRNA-seq dataset generated using drop-seq technique named DroNc-seq (Habib et al., 2017) (hereafter referred to as DroNC data). We have re-analyzed DroNC data using a pipeline similar to the one for MTG data (`https://github.com/Sun-lab/scRNAseq_pipelines/blob/master/dronc/dronc_seq.html` followed by an R code `step1_expression_signature.R` in the same folder) and selected the cells belonging to each cell type by taking intersections of clustering results and cell type labels provided by Habib et al. (Habib et al., 2017). DroNC data includes human hippocampus and PFC from five adults. There are 11,585 nuclei from 11 cell types, and we also consider a subset of them (4,536 nuclei) from PFC since it matches the tissues of bulk RNA-seq data.

```
   Cell_Type nCells_All nCells_PFC
1      exCA3        630          8
2      exCA1        179          0
3      exPFC       3107       3072
4        ASC       1584        339
5       GABA       1154        758
6        ODC       2582        167
7       exDG       1380          0
8        END         68          8
9         MG        223         19
10       OPC        533        165
11       NSC        145          0
```

In the above table, exCA3, exCA1, exPFC, and exDG are four types of excitatory neurons or glutamatergic neurons. ASC is astrocyte. GABA is GABAergic interneuron or inhibitory neuron. ODC is oligodendrocyte. END is endothelial cell. MG is microglia. OPC is oligodendrocyte precursor cell. NSC is neuronal stem cell.

Then we compare the cell type-specific gene expression data from DroNC (all cells or only the cells in PFC) vs. MTG data. We added up all the counts per gene across all the cells and make comparison, and collapse four types of excitatory neurons of DroNC data into one category. Detailed comparison of astrocyte and endothelial cell are shown in Supplementary Figures 3.27 and 3.28, respectively, and similar figures for other cell types can be found at `https://github.com/Sun-lab/scRNAseq_pipelines/tree/master/_brain_cell_type/figure`. It is clear that MTG data does not have much more cells than DroNC data, but the total read count is often 100 times more than DroNC data. Overall, gene expression of the same cell type are similar between MTG and DroNC, except endothelial, possibly due to small number endothelial cells (Supplementary Figure 3.29). We chose to use MTG data to generate reference since it has much higher depth and better coverage, making it more similar to bulk RNA-seq data. We exclude endothelial in our analysis since there are only 8 endothelial cells in MTG data and its expression has very weak similarity to the endothelial cells from DroNC data. The small number of cells also make the next step of selecting genes with cell type-specific expression much harder.

Next we selected around 120-130 genes per cell type for 6 cell types: excitatory neurons (Exc), inhibitory neurons (Inh), astrocyte (Astro), microglia (Micro), oligodendrocyte (Oligo), and oligodendrocyte precursor cell (OPC) by differential expression testing using MAST (Finak et al., 2015). All the genes we chose have FDR < 0.001 and fold change large than 2. Among these genes, we chose those with smallest FDR and largest fold changes. Specifically, we did a percentile grid search of FC > quantile(FC,pp) and FDR < quantile(FDR,1-pp) until there are anywhere between 120 and 130 genes for some percentile pp.

Since the deconvolution was done in the scale of TPM, which have adjusted for cell level read-depth, the cell type fraction estimates from CIBERSORT of ICeD-T are the fraction of expression from each cell types, not necessary the fraction of cells, if the total amount of RNA molecules are different across cell types. For each cell type, we estimate cell size factor and such it to adjust cell fraction estimates. Similar procedure has been used in earlier studies for immune cell types (Racle et al., 2017). See section 3.3.12.2 for details on transformation between cell fraction by expression and cell fraction by the number of cells.

3.4.2.2 Additional details of real data analyses

In the differential expression analysis, only about 20,000 genes whose third quartile of read counts are larger than 20 are included. The sample depth is defined by the third quartile of read counts among the aforementioned genes.

The surrogate variables were calculated using the "sva" R package using a linear model with transformed data matrix as response and covariates included in the differential expression as predictors. To control for cell fractions, the log transformed cell fraction estimates from ICeDT are also included.

To annotate the genes, particularly to match gene names and calculate gene lengths to calculate TPMs, we used "`Homo_sapiens.GRCh37.70.processed.gtf.gz`" for the SCZ dataset and "`gencode.v19.annotation.gtf.gz`" for the ASD dataset as the GENCODE GTF files. Gene lengths are defined by the length of the union of all exons. The annotated gene expressions were wrapped in the "SummarizedExperiment" container.

Cell fractions were estimated using both CIBERSORT and ICeD-T. The input of the cell type deconvolution methods were bulk gene expression in TPMs and reference matrix from MTG. For ICeD-T, we used the unweighted version. For CIBERSORT, we used their website interface of `https://cibersort.stanford.edu/index.php`.

Default options were adoped in running CARseq, TOAST, and DESeq2. The gene expression was supplied in the form of TPM for TOAST, and the cell fraction estimates were not adjusted for cell sizes. CARseq and DESeq2 required read counts, and CARseq also demanded cell fraction estimates adjusted for cell sizes. The p-values obtained from each method and cell type were transformed to q-values using a code snippet in the "CARseq" package.

	SCZ			ASD		
	CARseq	TOAST	DESeq2	CARseq	TOAST	DESeq2
Astro	1	0		0	0	
Exc	0	0		232	2	
Inh	0	3		835	0	
Micro	138	30		0	0	
Oligo	656	1		0	0	
OPC	0	0		0	0	
bulk			1009/1888/810			1063/481/185

Table 3.4: Number of differentially expressed genes by cell type and method using a q-value cutoff of 0.1, including DESeq2 for bulk samples. The results of DESeq2 include 3 numbers: the number of findings without accounting for cell type compositions, after accounting for cell type compositions, and their intersection.

	SCZ (#case = 250, #ctrl = 277)			ASD (#case = 42, #ctrl = 43)		
	Estimate	Std. error	p-value	Estimate	Std. error	p-value
Astro						
ICeDT	−0.003	0.027	0.899	0.217	0.092	0.021
CIBERSORT	−0.049	0.039	0.210	0.406	0.176	0.024
Inh						
ICeDT	0.093	0.021	<0.001	−0.037	0.063	0.557
CIBERSORT	0.055	0.035	0.121	−0.005	0.101	0.965
Micro						
ICeDT	0.002	0.039	0.968	0.107	0.093	0.253
CIBERSORT	−0.014	0.043	0.744	0.107	0.097	0.278
Oligo						
ICeDT	−0.041	0.042	0.323	−0.160	0.147	0.279
CIBERSORT	−0.076	0.048	0.116	−0.165	0.169	0.331
OPC						
ICeDT	0.008	0.029	0.777	0.127	0.073	0.085
CIBERSORT	−0.024	0.032	0.444	0.060	0.051	0.244

Table 3.5: Associations between relative cell type fractions and case-control status, assessed by a linear model with log ratio of cell fractions (with Exc neurons as baseline) as response, adjusted for other covariates.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Astro.ICeDT	527	0.146	0.046	0.023	0.121	0.158	0.416
Exc.ICeDT	527	0.592	0.078	0.303	0.551	0.642	0.751
Inh.ICeDT	527	0.109	0.032	0.021	0.087	0.134	0.191
Micro.ICeDT	527	0.021	0.012	0.005	0.014	0.026	0.087
Oligo.ICeDT	527	0.086	0.045	0.018	0.053	0.112	0.308
OPC.ICeDT	527	0.046	0.020	0.006	0.033	0.054	0.129
Astro.CIBERSORT	527	0.103	0.062	0.000	0.071	0.108	0.493
Exc.CIBERSORT	527	0.784	0.096	0.350	0.752	0.846	0.957
Inh.CIBERSORT	527	0.017	0.012	0.000	0.008	0.024	0.066
Micro.CIBERSORT	527	0.007	0.008	0.000	0.000	0.010	0.059
Oligo.CIBERSORT	527	0.088	0.056	0.012	0.049	0.113	0.405
OPC.CIBERSORT	527	0.001	0.005	0	0	0	0.50

Table 3.6: Summary of deconvolution results from ICeDT (Wilson et al., 2019) and CIBERSORT (Newman et al., 2019) for CMC data (Fromer et al., 2016), combining schizophrenia patients and controls.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Astro.ICeDT	85	0.121	0.056	0.042	0.089	0.132	0.456
Exc.ICeDT	85	0.585	0.086	0.226	0.552	0.634	0.710
Inh.ICeDT	85	0.130	0.036	0.013	0.109	0.151	0.227
Micro.ICeDT	85	0.018	0.014	0.005	0.011	0.020	0.110
Oligo.ICeDT	85	0.084	0.058	0.010	0.045	0.102	0.359
OPC.ICeDT	85	0.062	0.022	0.025	0.047	0.071	0.137
Astro.CIBERSORT	85	0.066	0.077	0.000	0.022	0.069	0.551
Exc.CIBERSORT	85	0.794	0.121	0.177	0.755	0.864	0.939
Inh.CIBERSORT	85	0.040	0.020	0.001	0.028	0.051	0.107
Micro.CIBERSORT	85	0.003	0.008	0	0	0.001	0
Oligo.CIBERSORT	85	0.097	0.078	0.007	0.048	0.117	0.516
OPC.CIBERSORT	85	0.0001	0.001	0	0	0	0.006

Table 3.7: Summary of deconvolution results from ICeDT (Wilson et al., 2019) and CIBERSORT (Newman et al., 2019) for the UCLA-ASD data (Parikshak et al., 2016), combining ASD patients and controls.

Cell type	Astro	Exc	Inh	Micro	Oligo	OPC
CARseq	0.051 (1.6)	3.3e−4 (2.4)	5.8e−7 (3.2)	2.9e−7 (−3.2)	3.6e−3 (2.0)	1.9e−5 (2.8)
TOAST	0.12 (−1.4)	0.89 (0.66)	0.14 (−1.3)	4.1e−4 (−2.3)	0.36 (1.1)	0.55 (−0.91)

Table 3.8: Gene Set Enrichment Analysis (GSEA) results for 328 ASD risk genes curated by SFARI. GSEA were run using the rankings of all the genes by CT-specific-DE p-values. The number is the parenthesis is the normalized enrichment score (NES). Positive NES means ASD risk genes tend to have smaller p-values, while negative NES means ASD risk genes tend to have larger p-values.

category	pval	nDE	nCat	qval
EUKARYOTIC TRANSLATION ELONGATION	1.70E-05	6	90	0.0098
RESPONSE OF EIF2AK4 GCN2 TO AMINO ACID DEFICIENCY	2.50E-05	6	98	0.0098
SELENOAMINO ACID METABOLISM	3.70E-05	6	105	0.0098
SRP DEPENDENT COTRANSLATIONAL PROTEIN TARGETING TO MEMBRANE	4.50E-05	6	109	0.0098
NONSENSE MEDIATED DECAY NMD	4.60E-05	6	113	0.0098
EUKARYOTIC TRANSLATION INITIATION	6.00E-05	6	117	0.0107
CELLULAR RESPONSES TO EXTERNAL STIMULI	1.70E-04	10	487	0.0236
INFLUENZA INFECTION	1.80E-04	6	149	0.0236
INTEGRIN CELL SURFACE INTERACTIONS	2.70E-04	4	73	0.0319
REGULATION OF EXPRESSION OF SLITS AND ROBOS	3.00E-04	6	160	0.0322
EXTRACELLULAR MATRIX ORGANIZATION	5.70E-04	6	237	0.0548
RRNA PROCESSING	6.20E-04	6	188	0.0548
SIGNALING BY ROBO RECEPTORS	8.60E-04	6	204	0.0708
FCGR ACTIVATION	1.30E-03	2	11	0.0999

Table 3.9: Over-representation of functional categories among the 65 genes with marginal DE signals (p-value < 0.05) in microglia in both SCZ and ASD studies. The column of nDE is the number of DE genes within each category and the column nCat is the number of genes within that category.

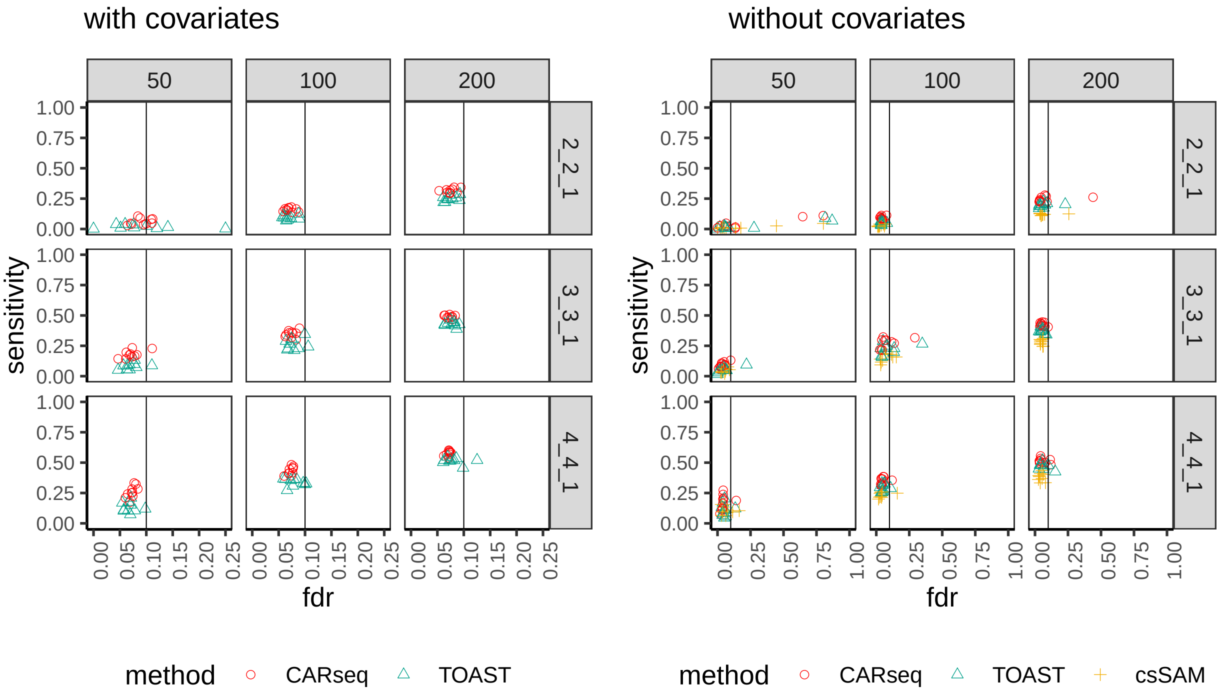


Figure 3.16: The FDR vs. sensitivity of several methods testing for CT-specific DE, when a confounding covariate is provided (a) or it is missing (b). There are 10 simulation replicates for each combination of total sample size with equal number of cases and controls (columns, e.g., $n = 50$ means 25 cases + 25 controls) and pattern of differential expression (rows). The notation for each pattern represents the fold changes in cell type 1, 2, and 3, respectively. For example, 2_2_1 indicates that both cell type 1 and cell type 2 are differentially expressed between the case and control groups by a fold change of 2 and cell type 3 are equivalently expressed between cases and controls. For each replicate, there are 2,000 genes following the pre-specified pattern of differential expression and 8,000 genes with no differential expression in any of the three cell types. The vertical line indicates the intended FDR level of 0.1. Note that csSAM does not support the inclusion of covariates, and that the scales of the x axis in the two subfigures are different.

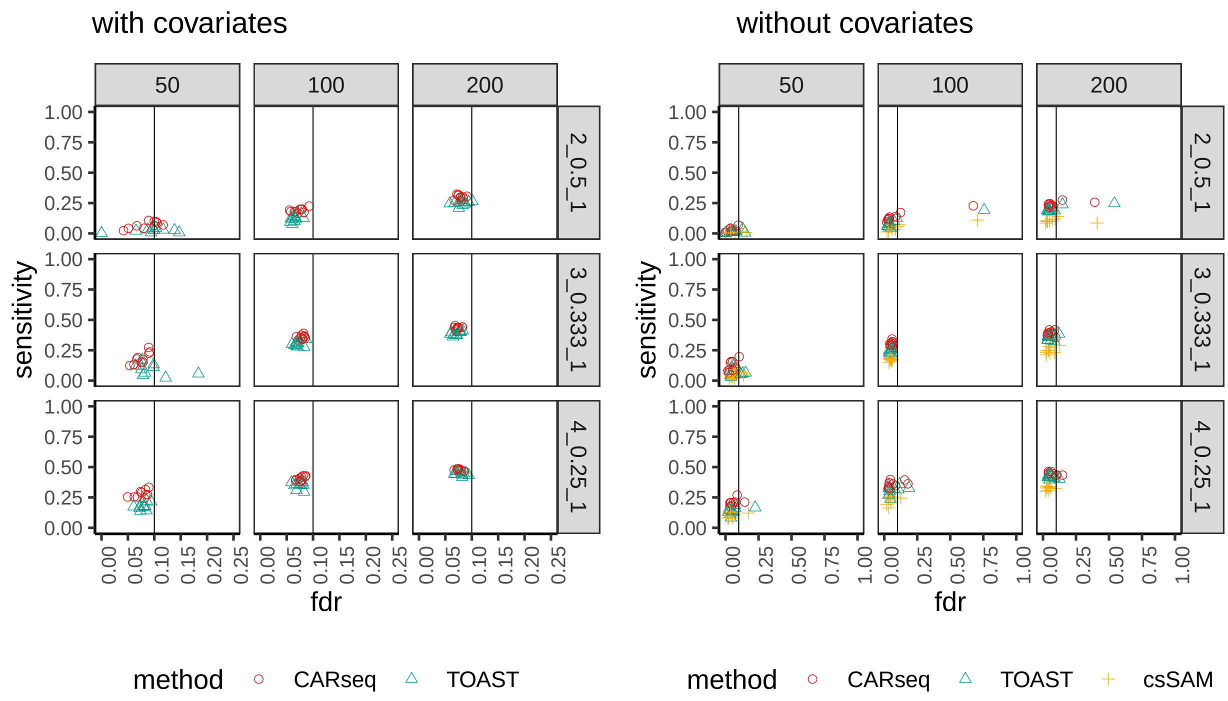


Figure 3.17: Similar to Figure 3.16, but with different patterns of differential expression where cell type 1 is over-expressed and cell type 2 is under-expressed.

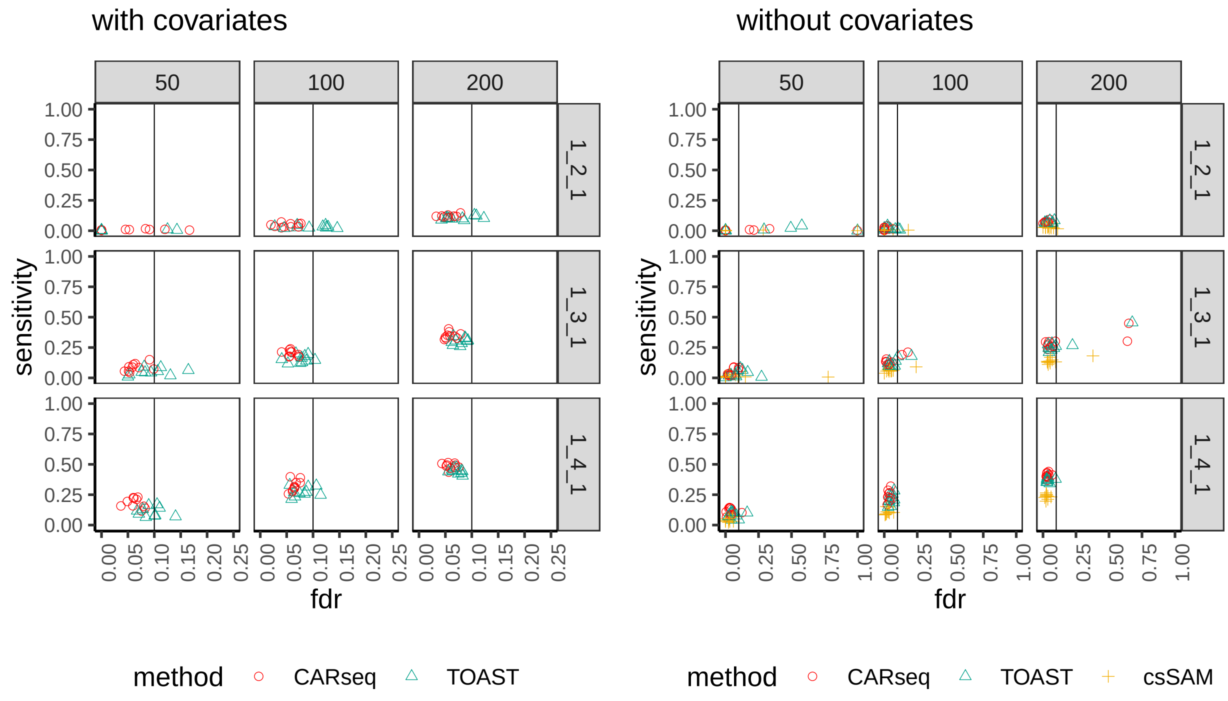


Figure 3.18: Similar to Figure 3.16, but with different patterns of differential expression where only cell type 2 is differentially expressed.

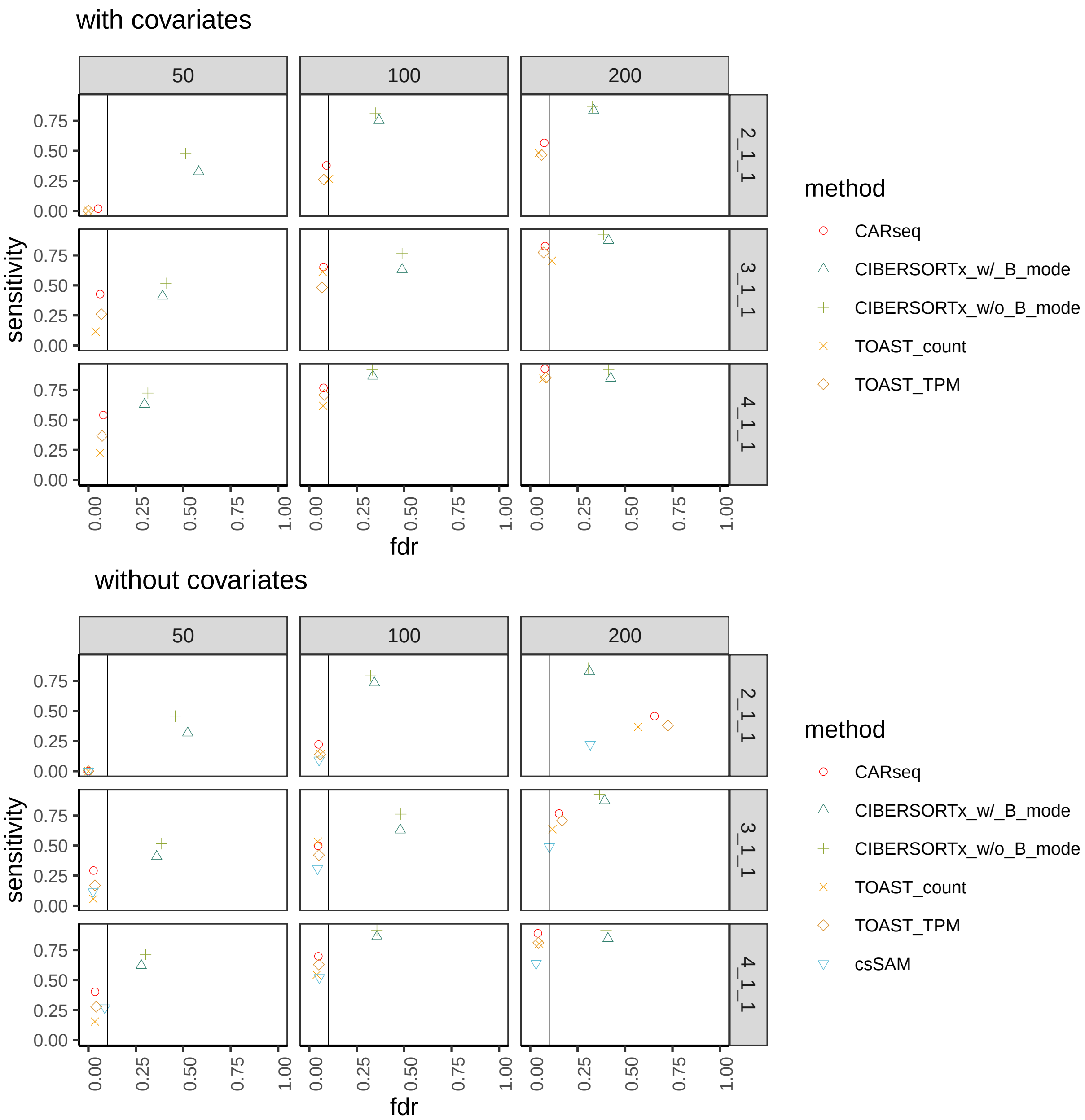


Figure 3.19: Results of this figure use the same simulation setup up as Figure 3.11 in main text, in terms sample sizes and effect sizes. The difference is that here we have one replicate per method (still with 2,000 DE genes and 8,000 non-DE genes per replicate), and more methods to compare. B_mode for CIBERSORT refers to their Bulk-mode batch correction (B-mode) that seek to correct the batch effect between reference and bulk RNA-seq data.

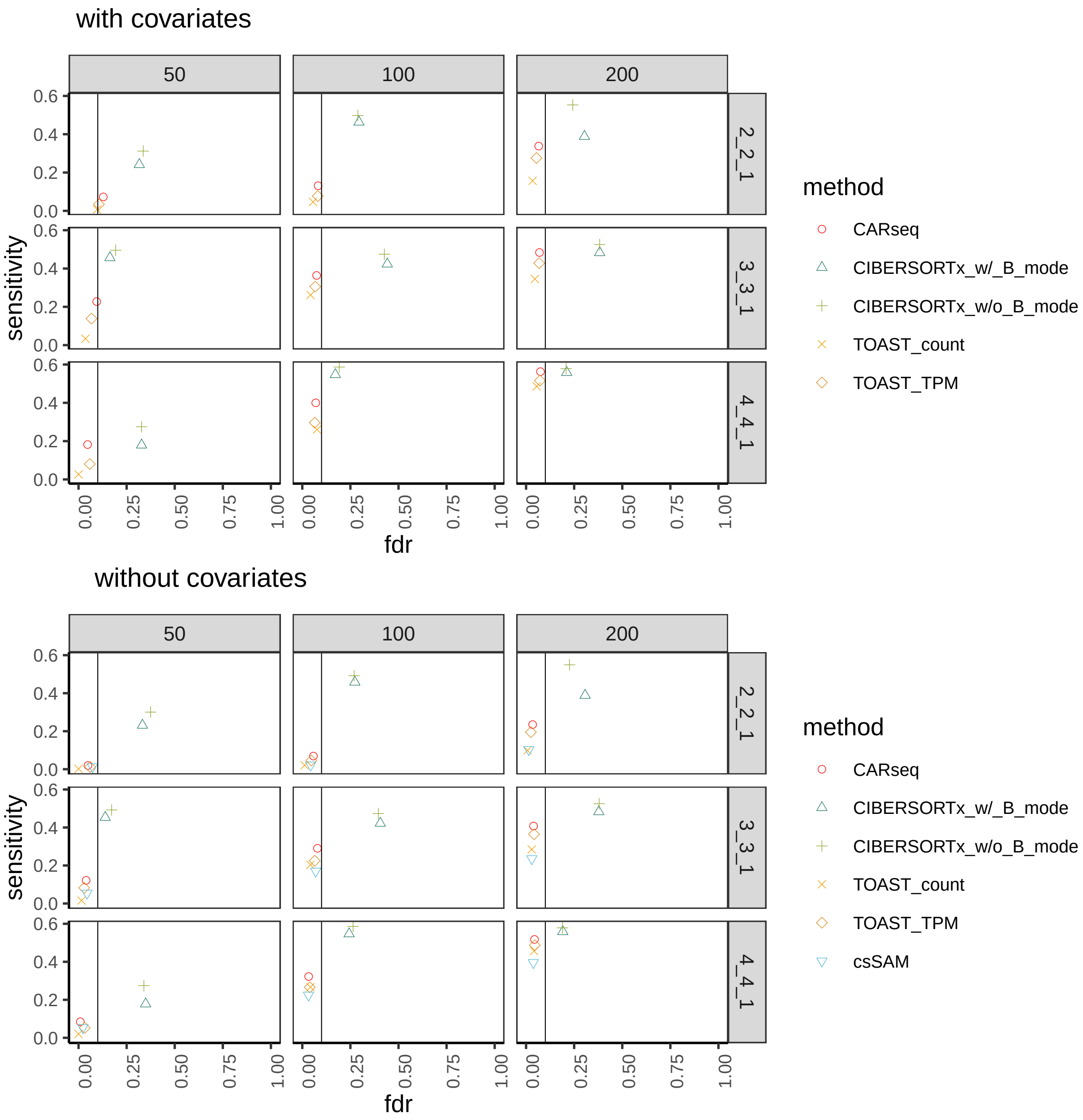


Figure 3.20: Similar to Figure 3.16 except comparing more methods with one replicate per method.

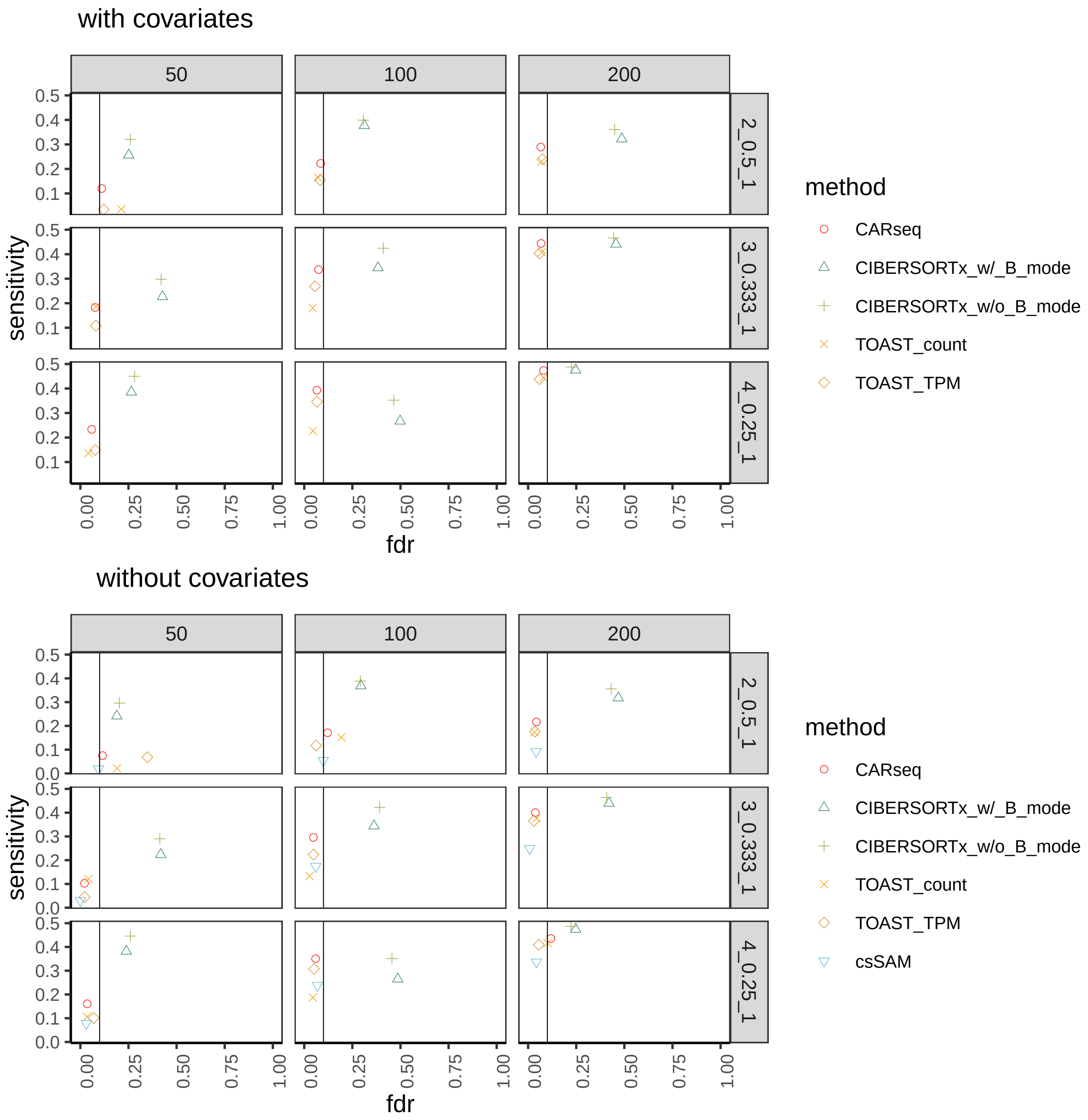


Figure 3.21: Similar to Figure 3.17 except comparing more methods with one replicate per method .

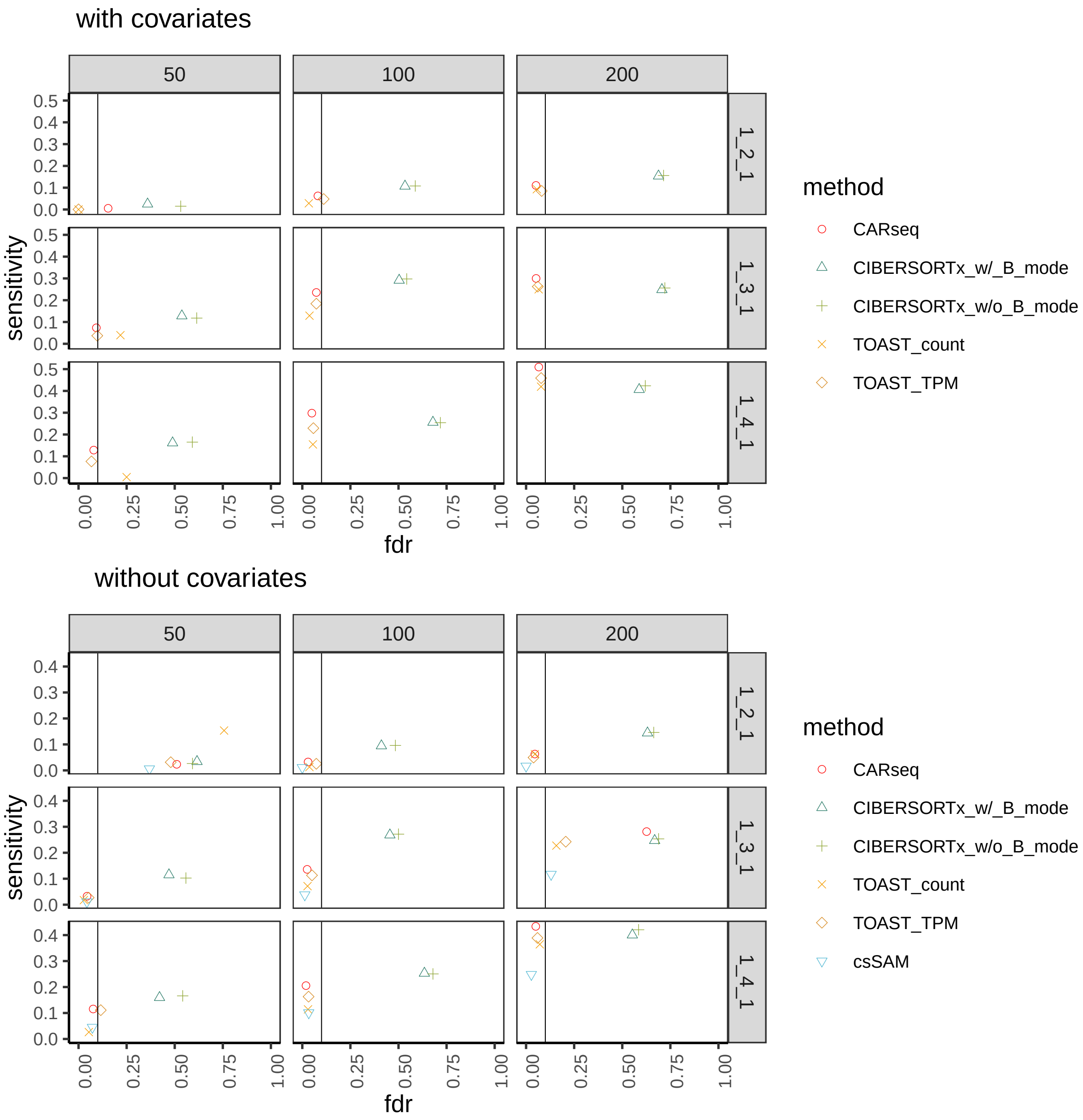


Figure 3.22: Similar to Figure 3.18 except comparing more methods with one replicate per method .

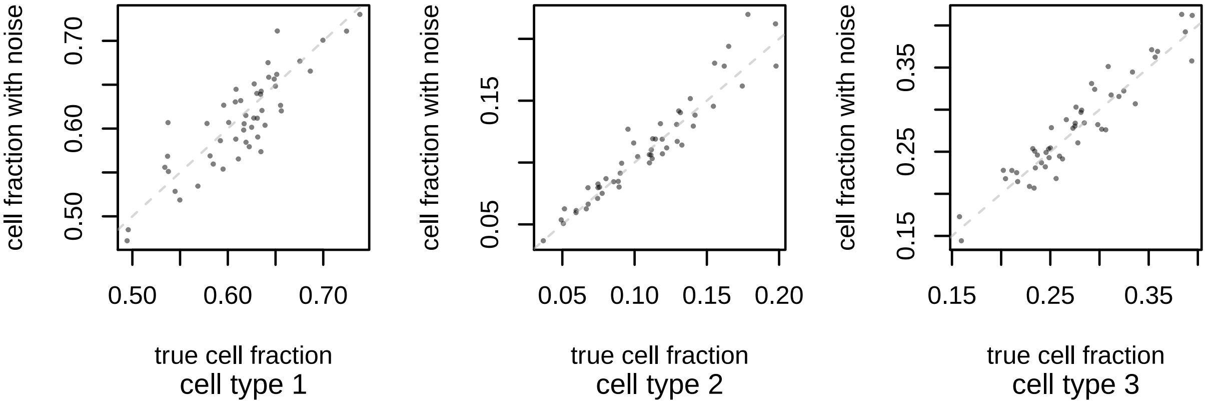


Figure 3.23: Noise in cell fraction estimates compared to the truth in a simulation with sample size 50 (25 cases vs. 25 controls).

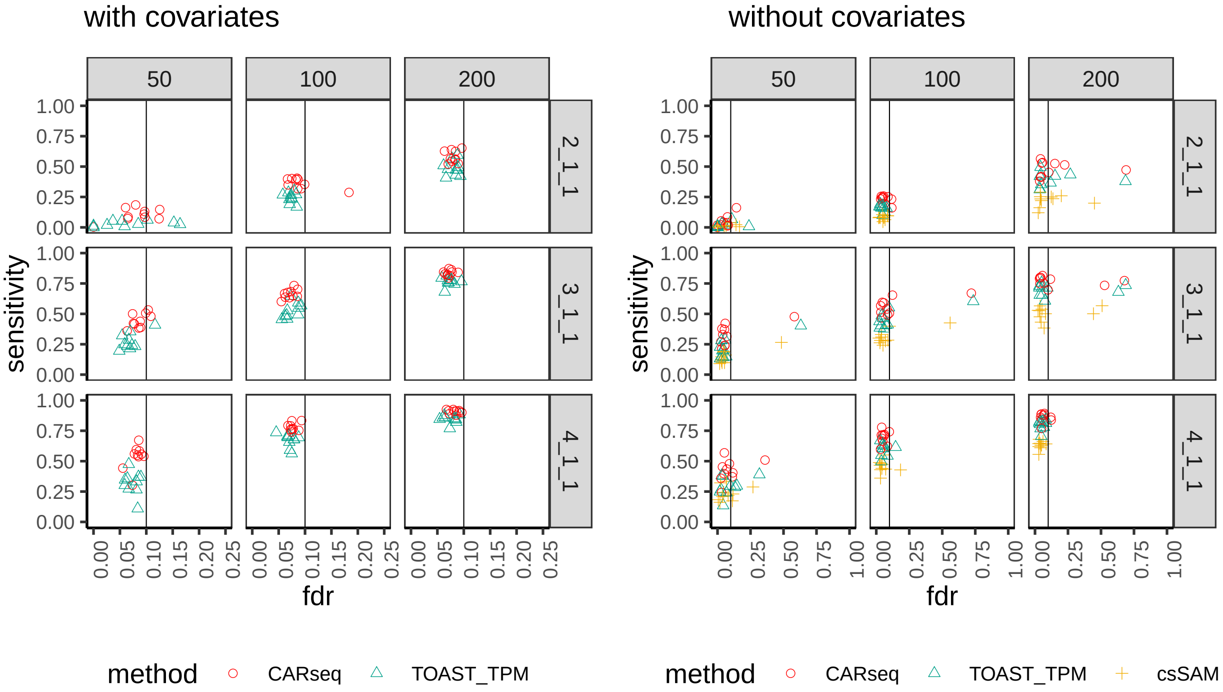


Figure 3.24: The FDR vs. sensitivity of several methods testing for CT-specific DE, when a confounding covariate is provided (a) or it is missing (b). A zero-centered Gaussian noise with a standard deviation of 0.1 is added to the cell fractions estimates on a logit scale. There are 10 simulation replicates for each combination of total sample size with equal number of cases and controls (columns, e.g., $n = 50$ means 25 cases + 25 controls) and pattern of differential expression (rows). The notation for each pattern represents the fold changes in cell type 1, 2, and 3, respectively. For example, 2_1_1 indicates that both cell type 1 is differentially expressed between the case and control groups by a fold change of 2 and cell type 2 and cell type 3 are equivalently expressed between cases and controls. For each replicate, there are 2,000 genes following the pre-specified pattern of differential expression and 8,000 genes with no differential expression in any of the three cell types. The vertical line indicates the intended FDR level of 0.1. Note that csSAM does not support the inclusion of covariates, and that the scales of the x axis in the two subfigures are different.

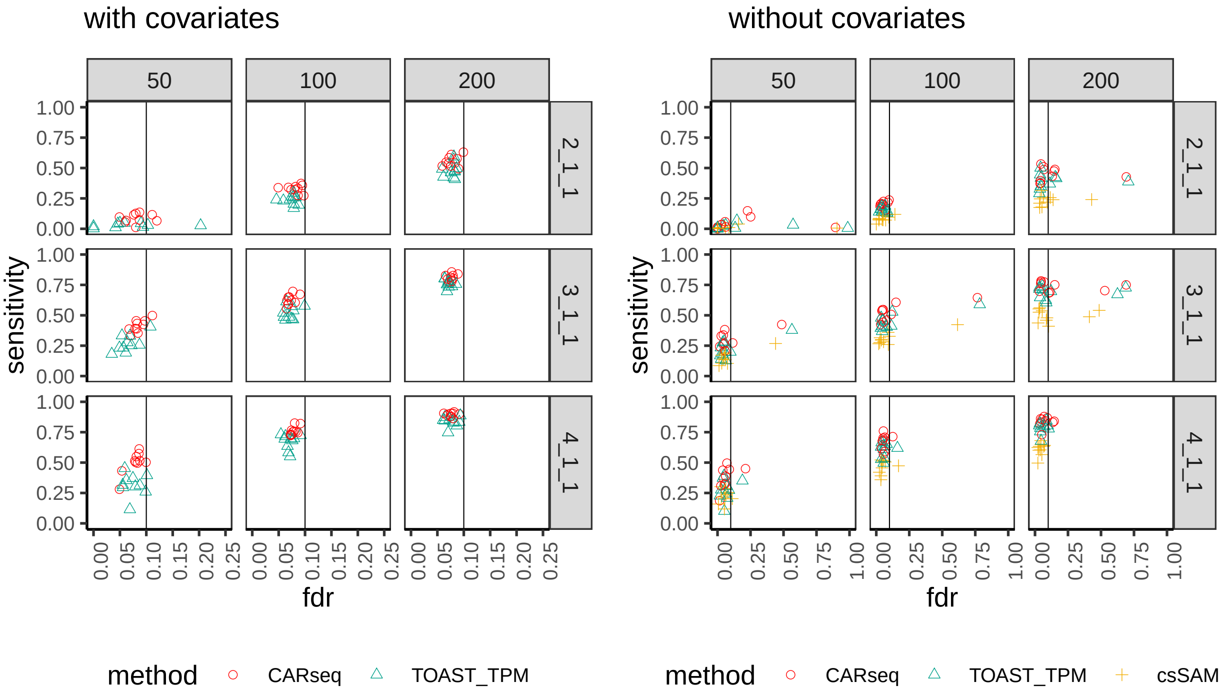


Figure 3.25: The FDR vs. sensitivity of several methods testing for CT-specific DE, when a confounding covariate is provided (a) or it is missing (b). The cell size factor is misspecified; we intentionally apply a list of wrong size factors (1.2, 1, 1) in the inference and testing instead of the true size factors (1, 1, 1) used in data simulation. There are 10 simulation replicates for each combination of total sample size with equal number of cases and controls (columns, e.g., $n = 50$ means 25 cases + 25 controls) and pattern of differential expression (rows). The notation for each pattern represents the fold changes in cell type 1, 2, and 3, respectively. For example, 2_1_1 indicates that both cell type 1 is differentially expressed between the case and control groups by a fold change of 2 and cell type 2 and cell type 3 are equivalently expressed between cases and controls. For each replicate, there are 2,000 genes following the pre-specified pattern of differential expression and 8,000 genes with no differential expression in any of the three cell types. The vertical line indicates the intended FDR level of 0.1. Note that csSAM does not support the inclusion of covariates, and that the scales of the x axis in the two subfigures are different.

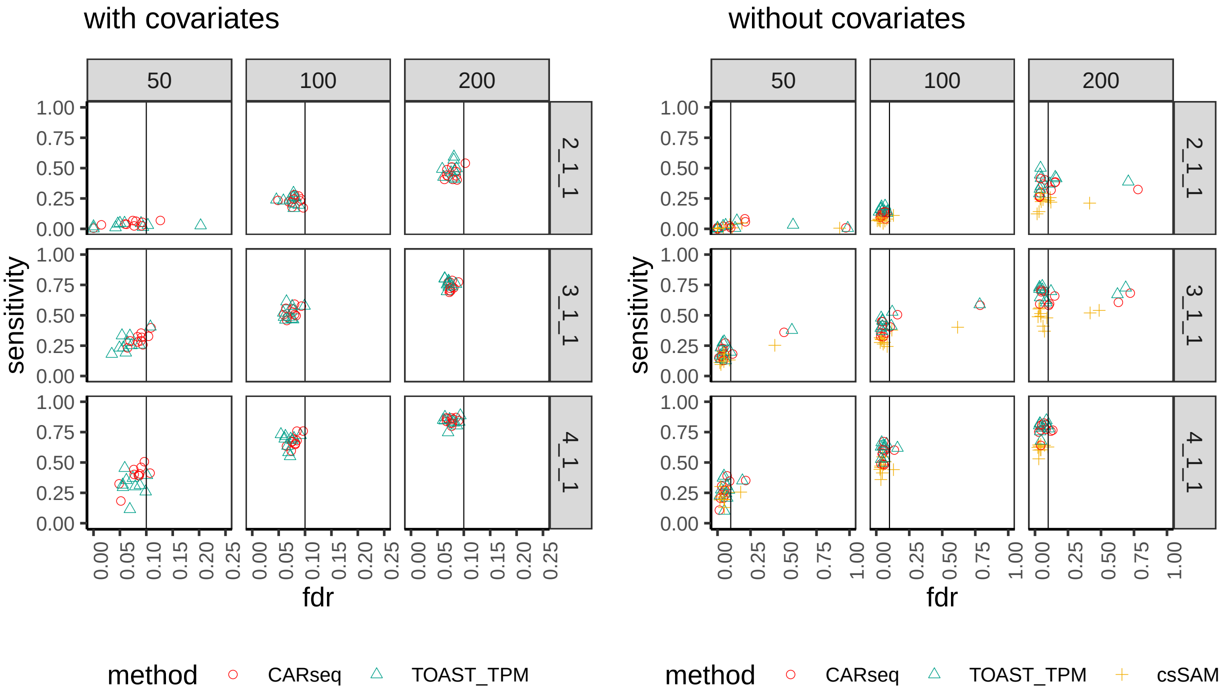


Figure 3.26: Similar to Figure 3.25 except the cell size factors are extremely misspecified to be 2, 1, 1.

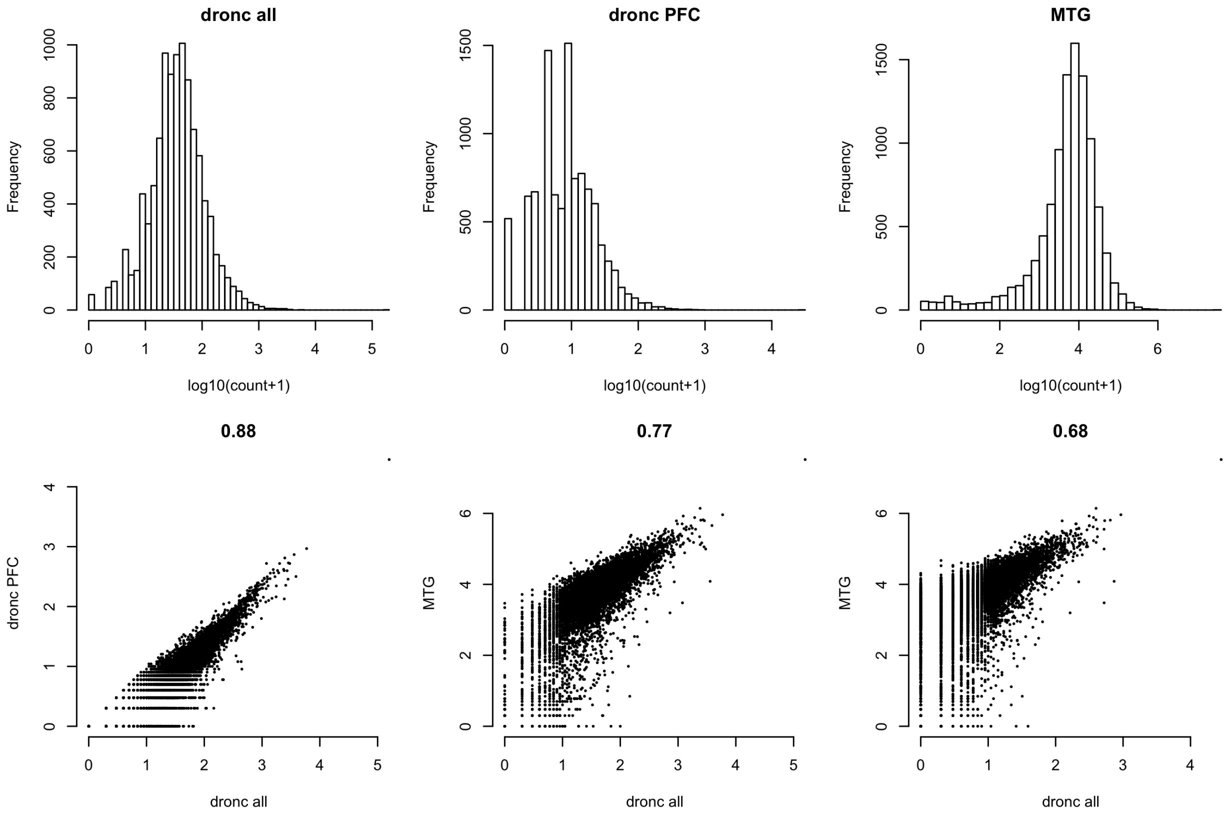


Figure 3.27: Compare astrocyte-specific gene expression of 10,151 genes derived from DroNC and MTG data. The top panel shows the distribution of gene expression in log transformed raw counts. Note that the counts are summation across 1584, 339, and 287 cells for all DroNC data, DroNC PFC only, and MTG, respectively. The lower panel shows pair-wise scatter plots.

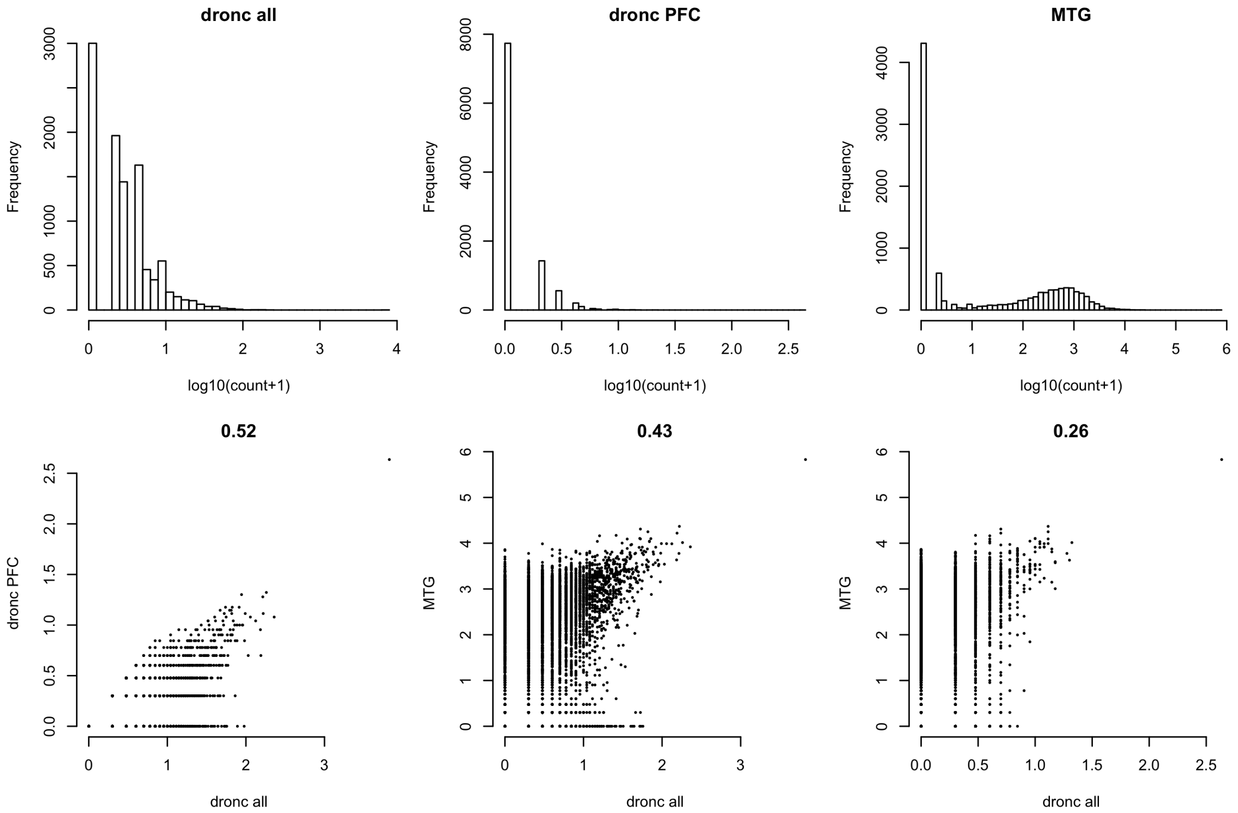


Figure 3.28: Compare endothelial-specific gene expression of 10,151 genes derived from DroNC and MTG data. The top panel shows the distribution of gene expression in log transformed raw counts. Note that the counts are summation across 68, 8, and 8 cells for all DroNC data, DroNC PFC only, and MTG, respectively. The lower panel shows pair-wise scatter plots.

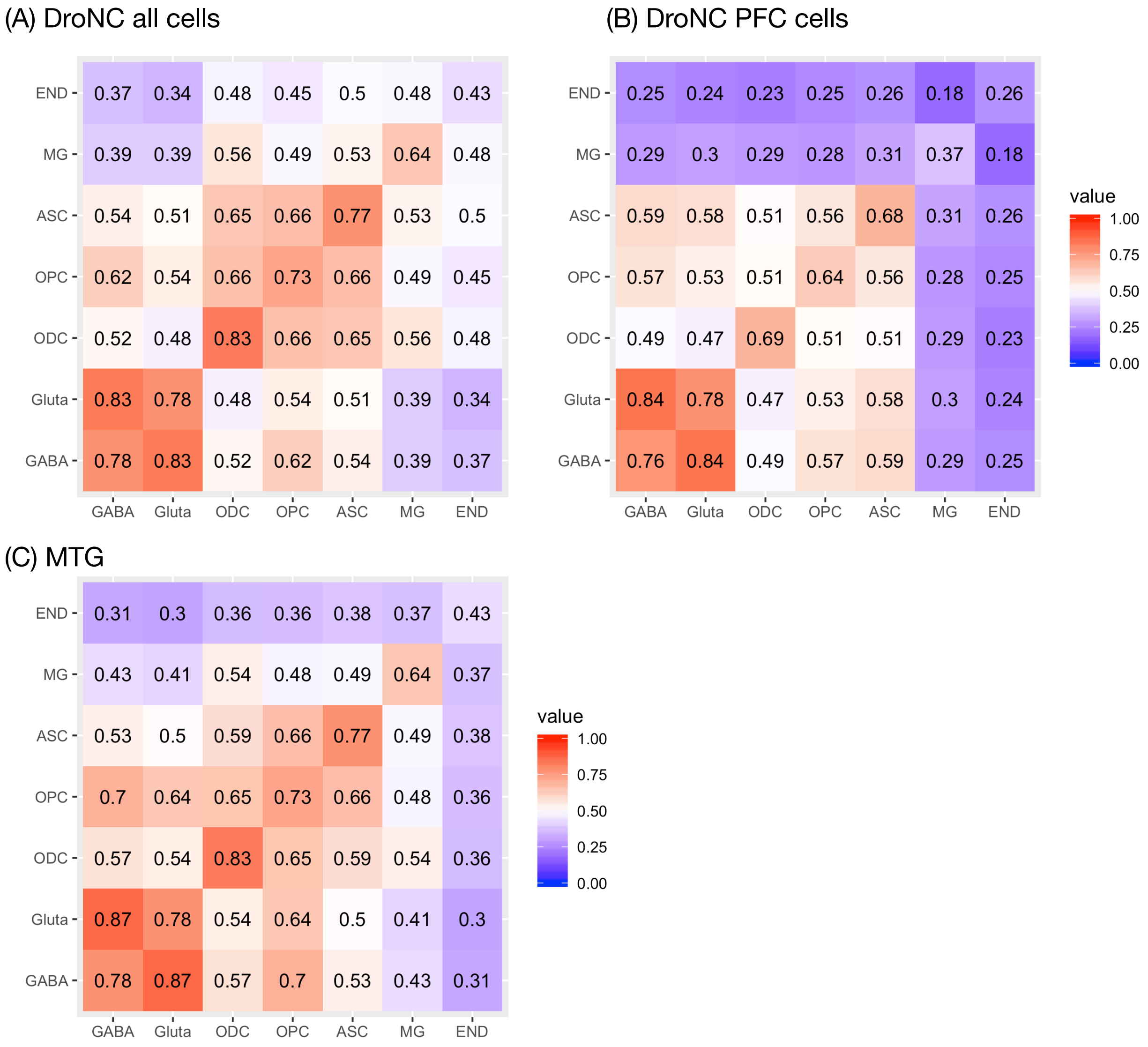


Figure 3.29: Correlations of genome-wide gene expression (10,151 genes) within each dataset, while the diagonals of the three panels were replaced by the correlation of (A) DroNC (all cells) vs. MTG (B) DroNC (PFC cells) vs. MTG, and (C) DroNC (all cells) vs. MTG. GABA: GABAergic interneurons or inhibitory neurons; Gluta: glutamatergic neurons or excitatory neurons; ODC: oligodendrocytes; OPC: oligodendrocyte precursor cells; ASC: astrocytes; MG: microglia; END: endothelial cells.

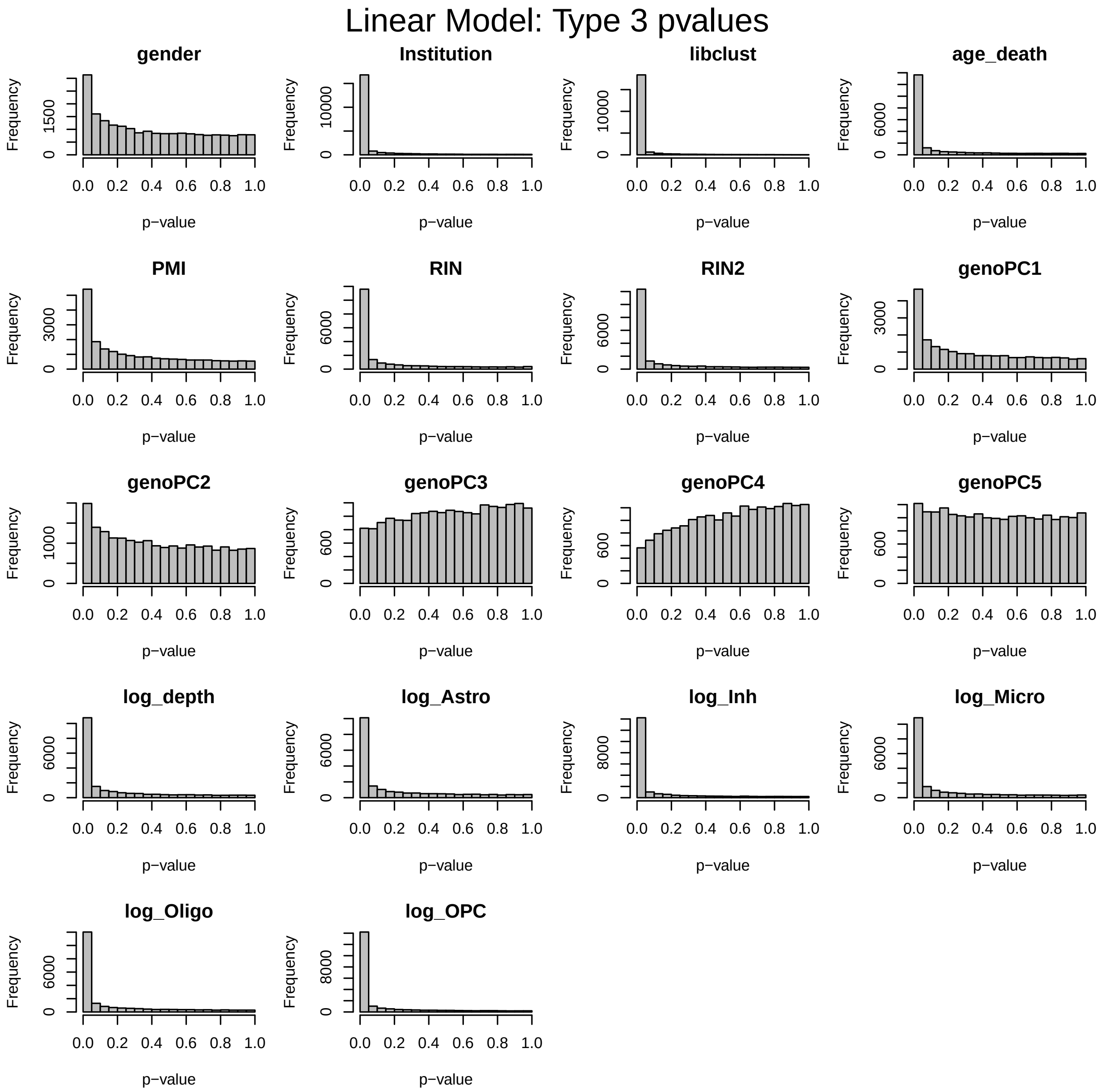


Figure 3.30: The p-values for each covariate vs. genome-wide gene expression for the schizophrenia study, assessed by a linear model for log-transformed gene expression. Cell type proportions were included as log ratios, e.g., `log_Astro` is log(Astro proportion/Excitatory neuron proportion).

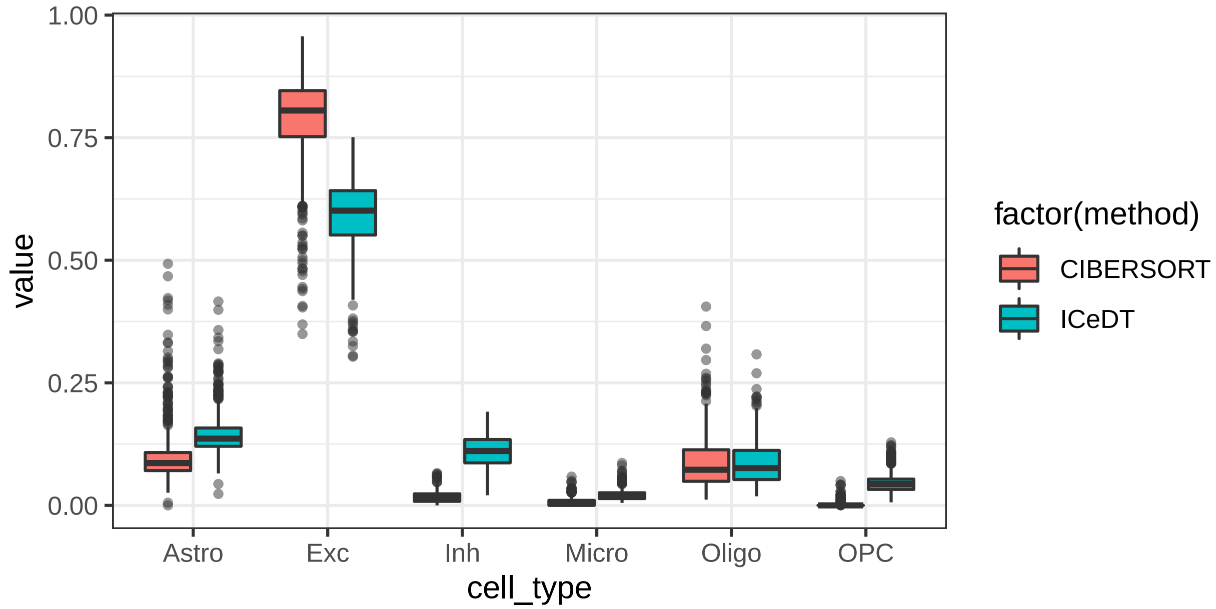


Figure 3.31: Box plots of cell type fraction estimates for CMC data.

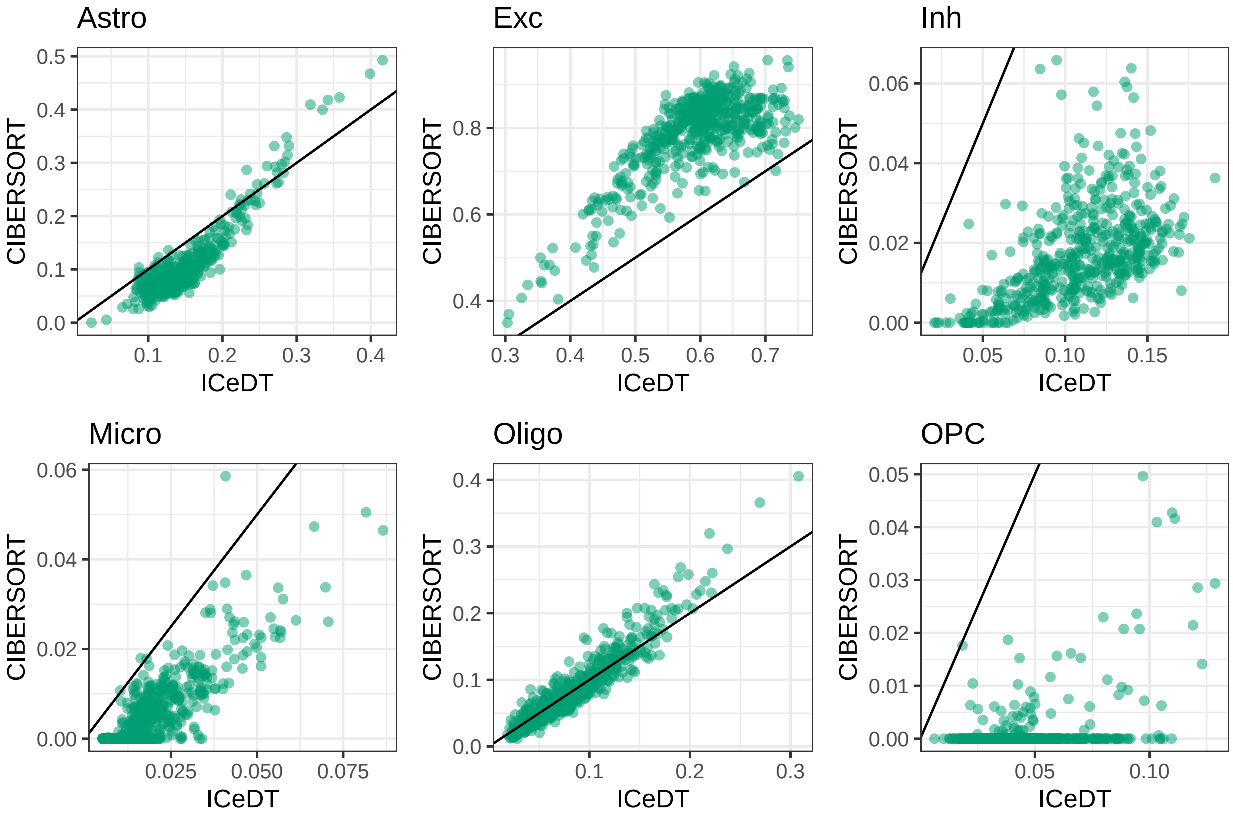


Figure 3.32: Scatter plots of cell type fraction estimates for CMC data.

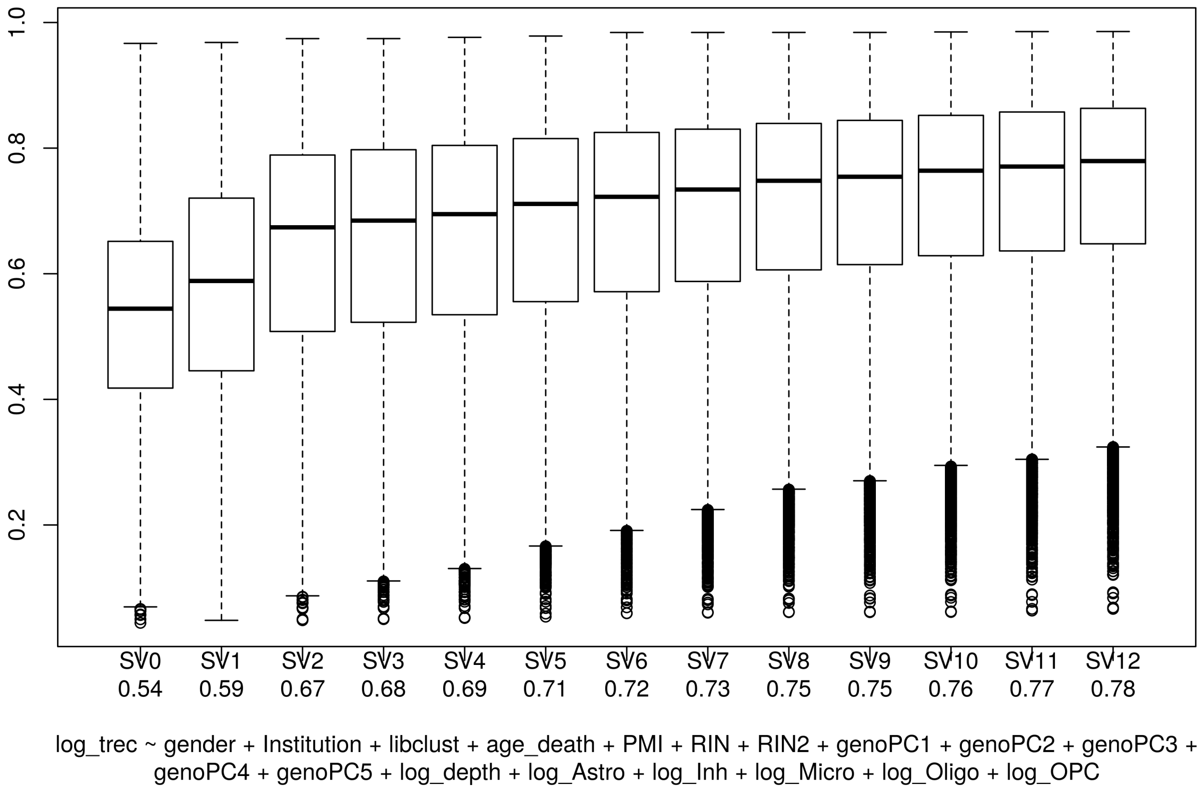


Figure 3.33: R^2 explained by increasing number of surrogate variables for CMC data (Fromer et al., 2016).

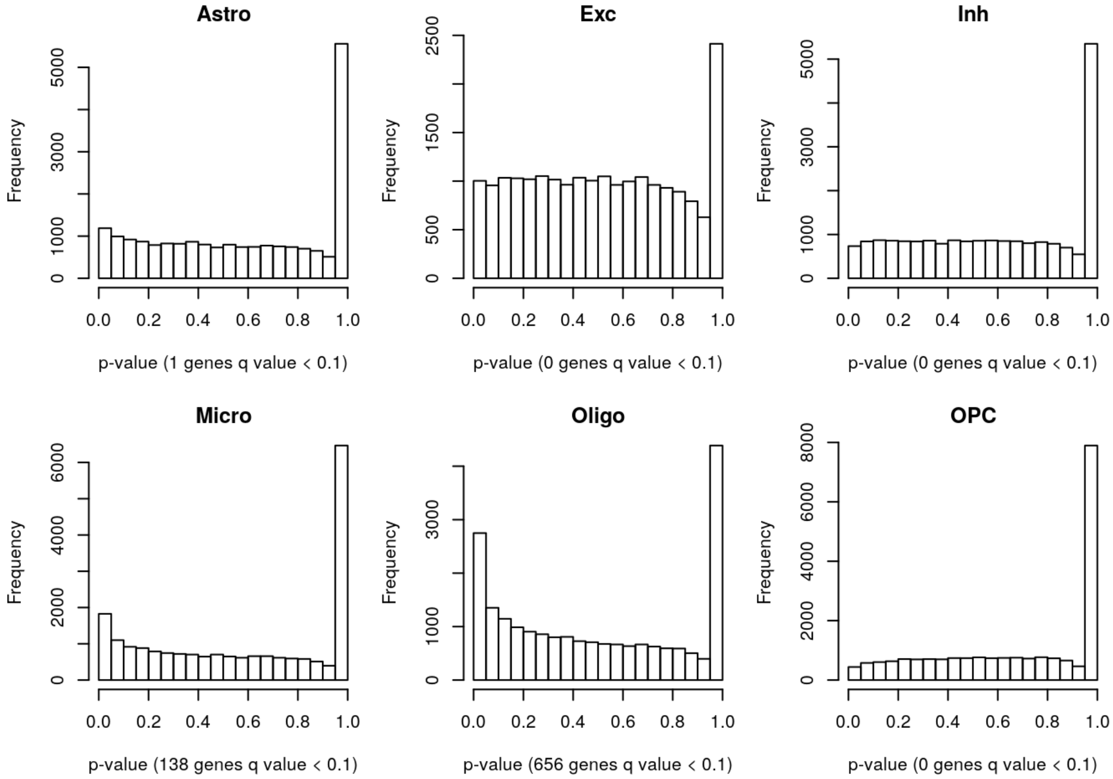


Figure 3.34: CARseq p-value distribution in the SCZ study.

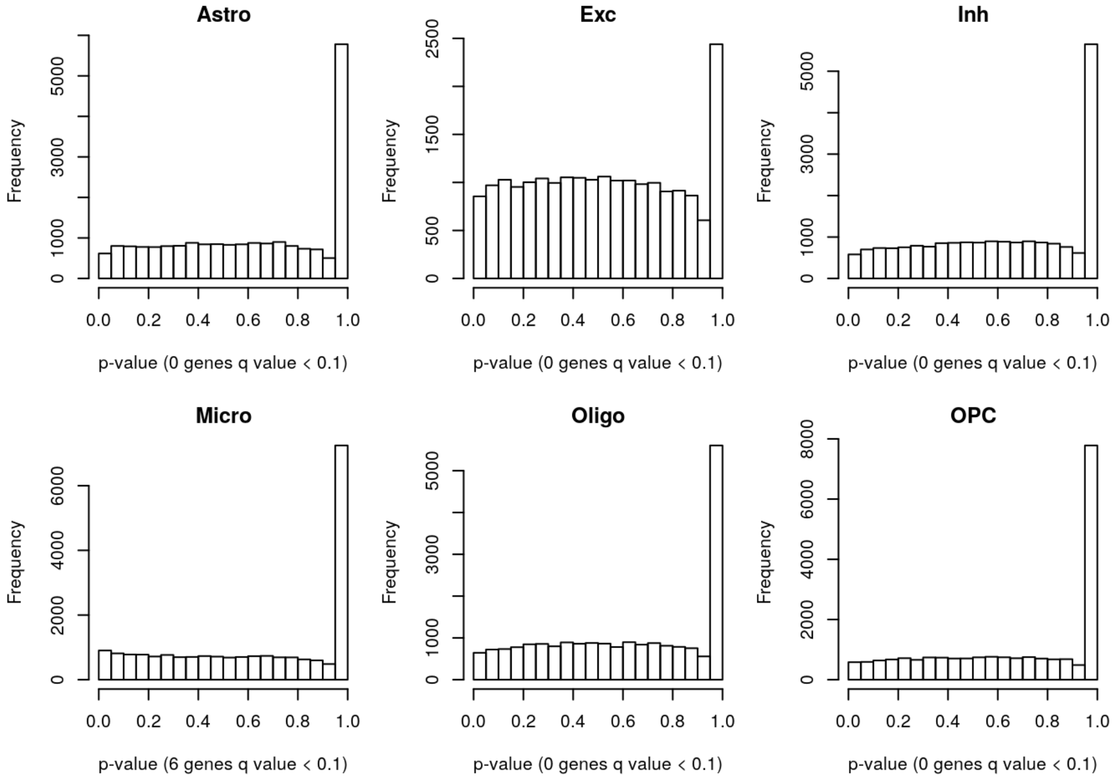


Figure 3.35: CARseq p-value distribution in the SCZ study where the case-control label has been permuted to reflect the null distribution.

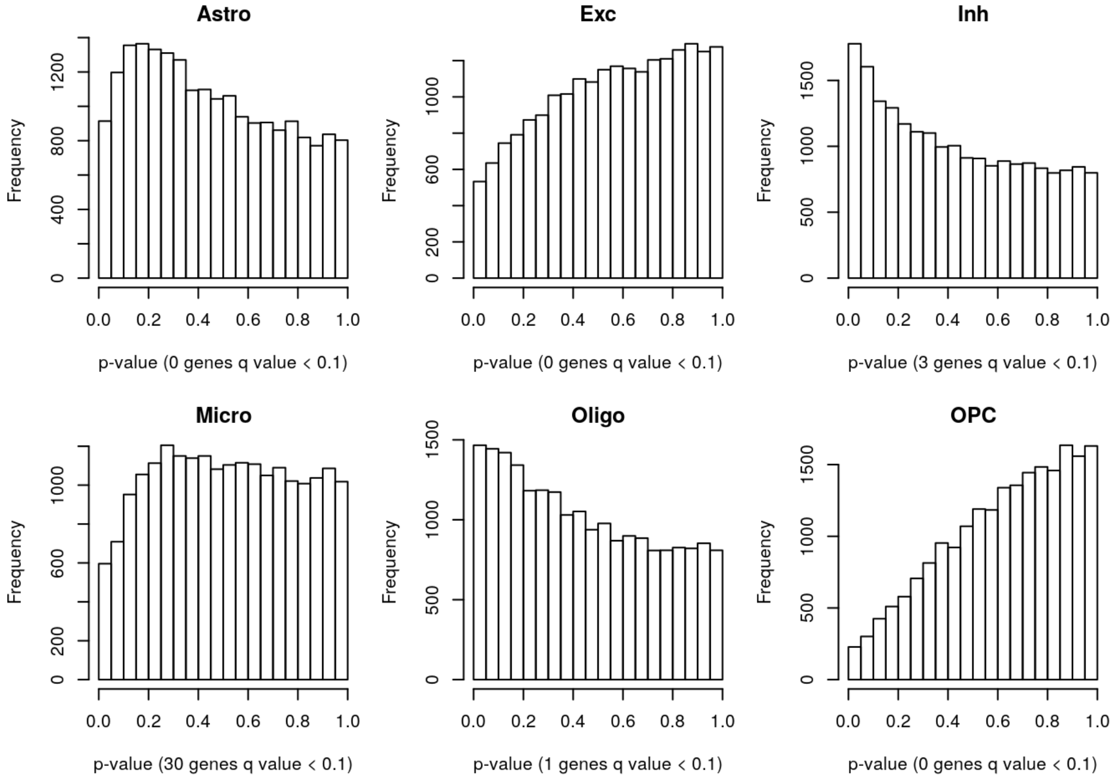


Figure 3.36: TOAST p-value distribution in the SCZ study.

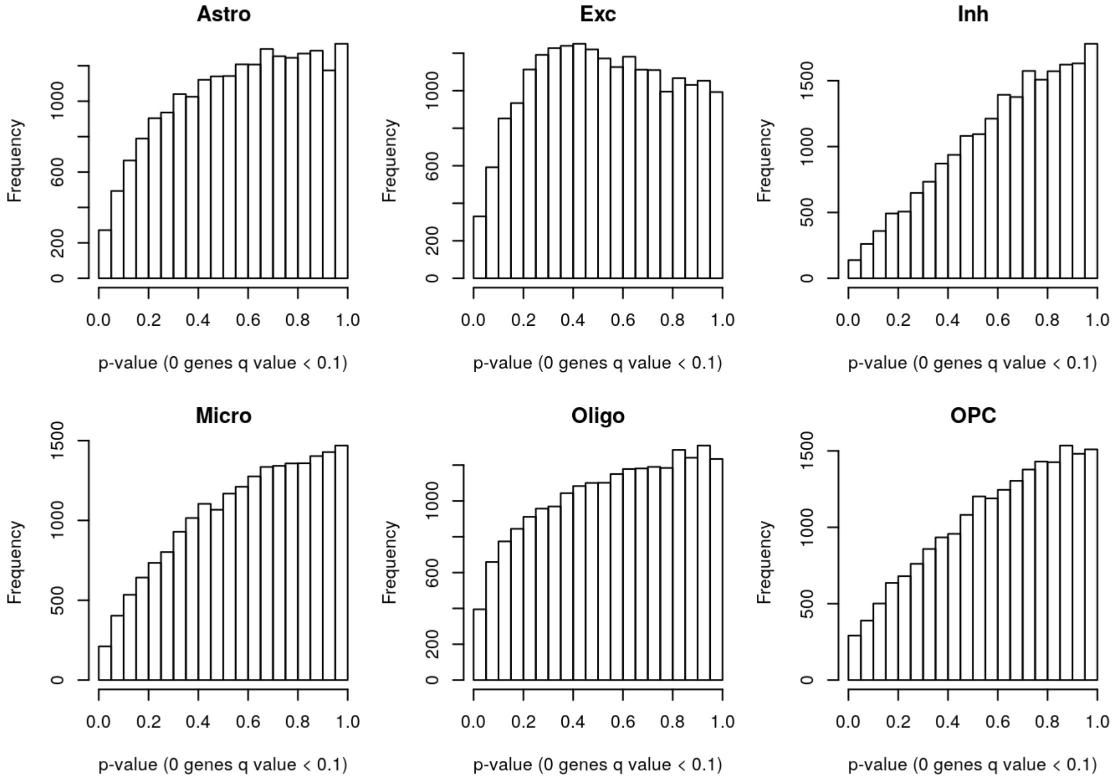


Figure 3.37: TOAST p-value distribution in the SCZ study where the case-control label has been permuted to reflect the null distribution.

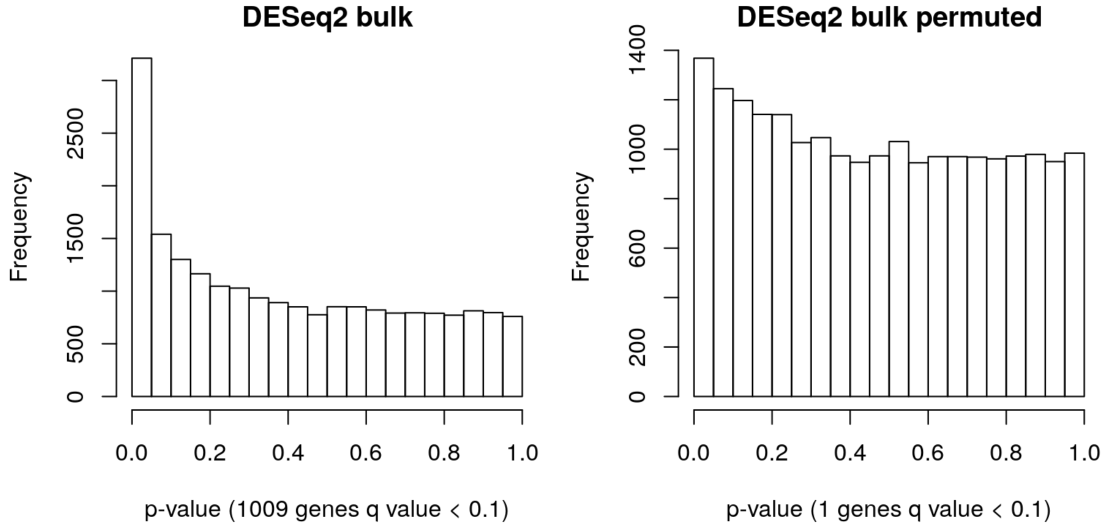


Figure 3.38: DESeq2 p-value distribution in the SCZ study where the case-control label is either unpermuted or permuted.

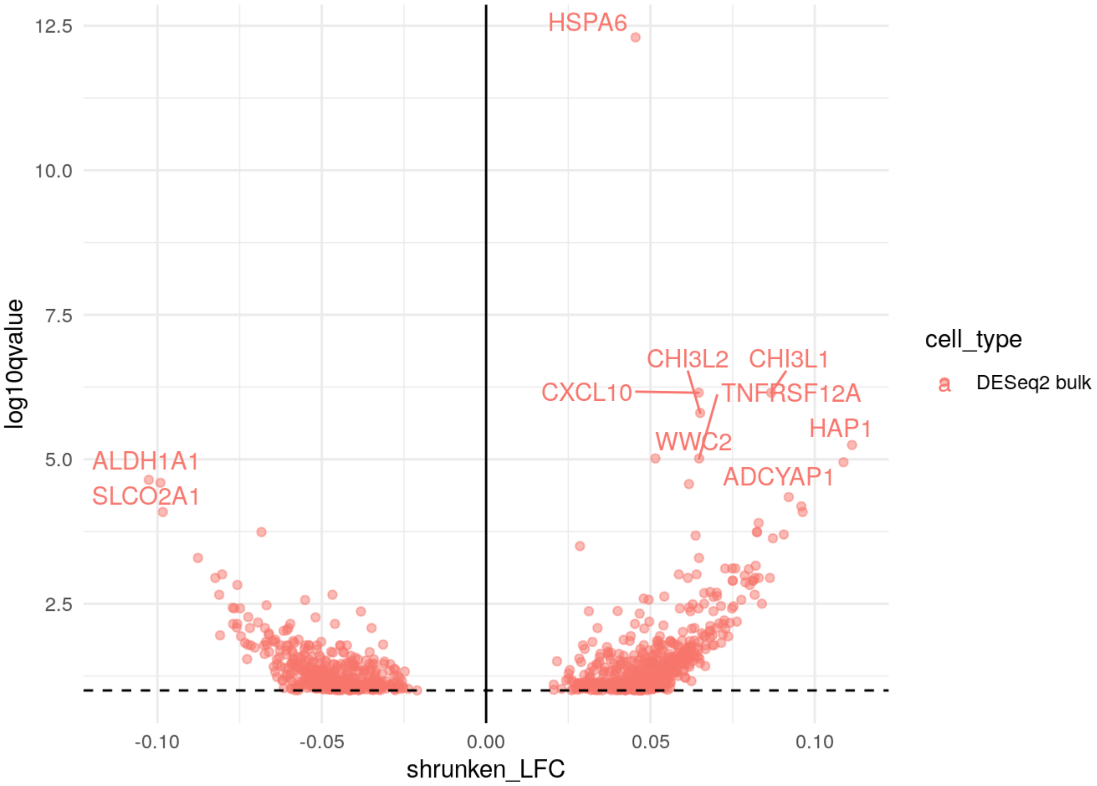


Figure 3.39: DESeq2 volcano plot in the SCZ study.

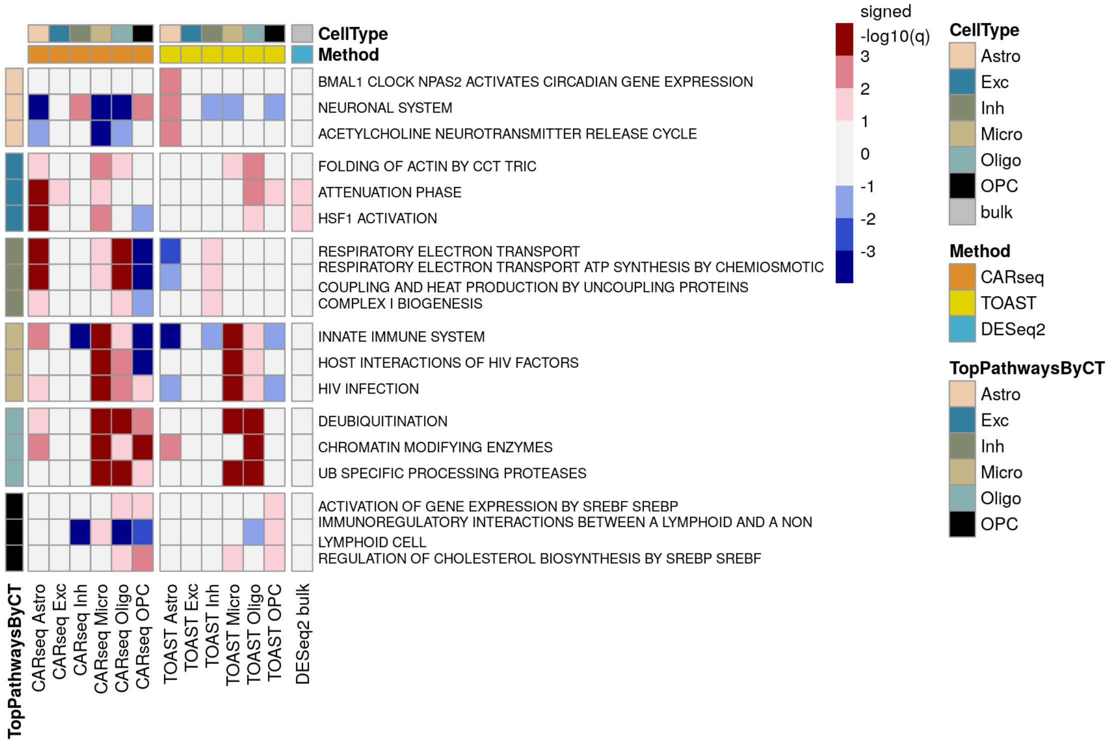


Figure 3.40: REACTOME GSEA ranked by TOAST in the SCZ study.

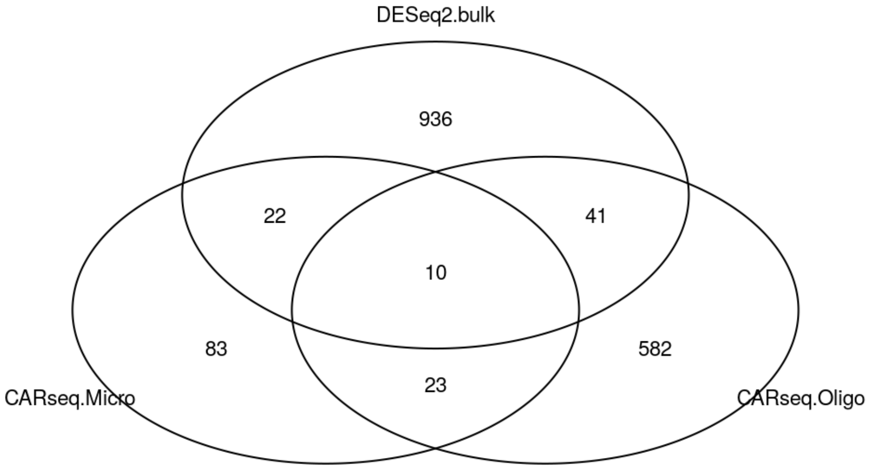


Figure 3.41: Venn plot of DEGs (q-value < 0.1) in the SCZ study.

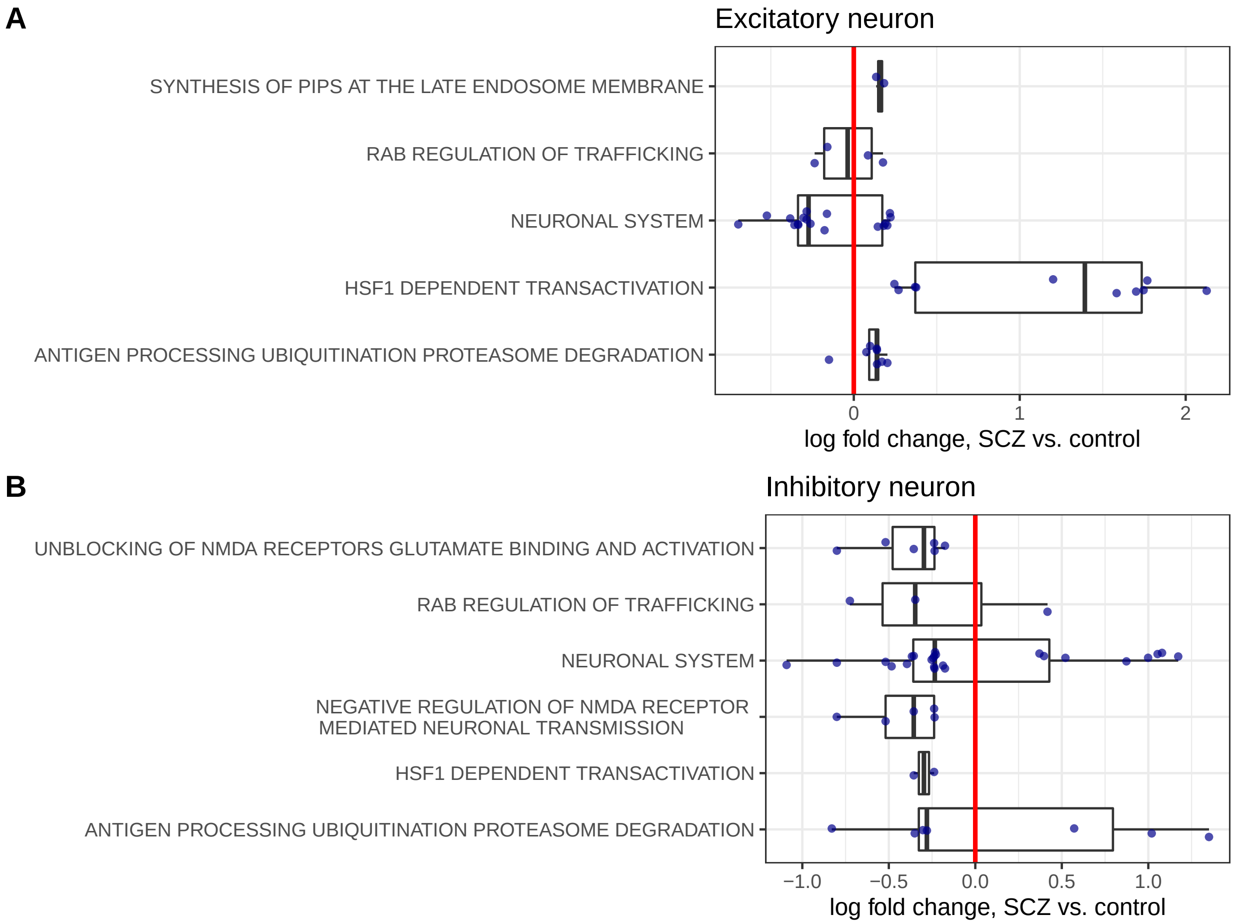


Figure 3.42: Shrunken log fold change estimates (SCZ vs. controls) for genes belonging to some REACTOME pathways. Panel A includes all the pathways related with NDMA. Panels B and C include the pathway identified by GSEA in excitatory/inhibitory neurons, from either SCZ or ASD studies. Only the genes with CT-specific-DE p-value (comparing SCZ vs. controls) smaller than 0.05 are shown.

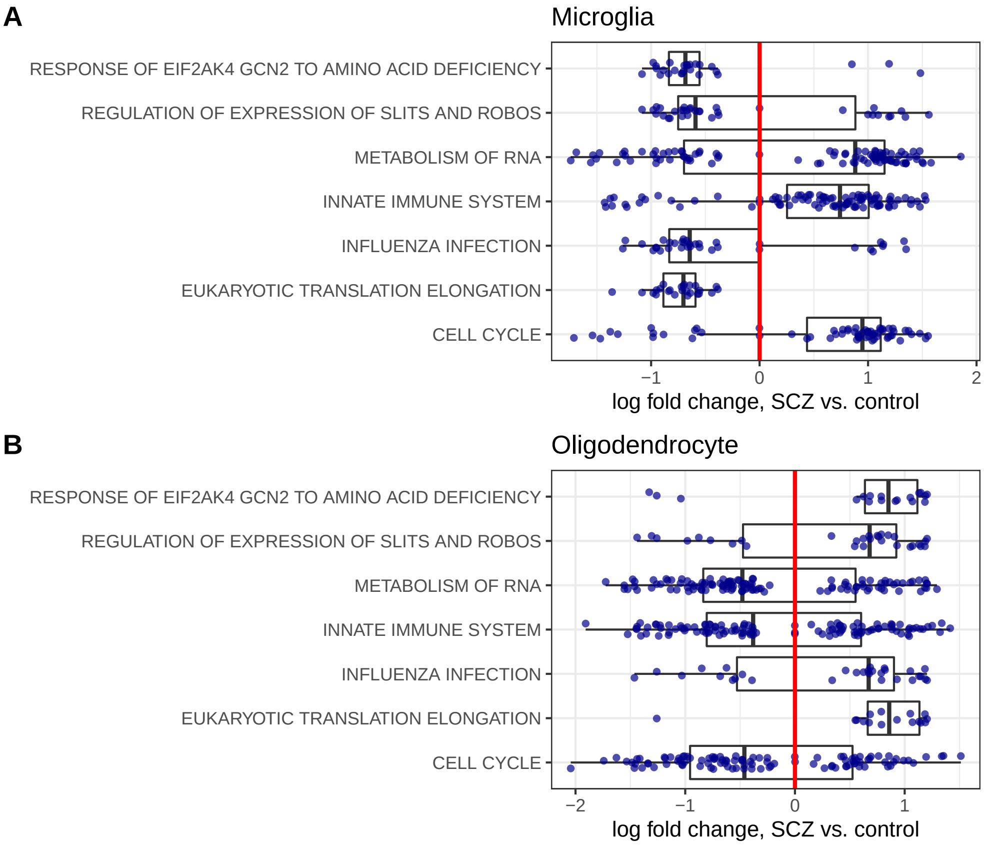


Figure 3.43: Shrunken log fold change estimates (SCZ vs. controls) for genes belonging to the REACTOME pathways identified by GSEA in microglias or oligodendrocytes, from either SCZ or ASD studies. Only the genes with CT-specific-DE p-value (comparing SCZ vs. controls) smaller than 0.05 are shown.

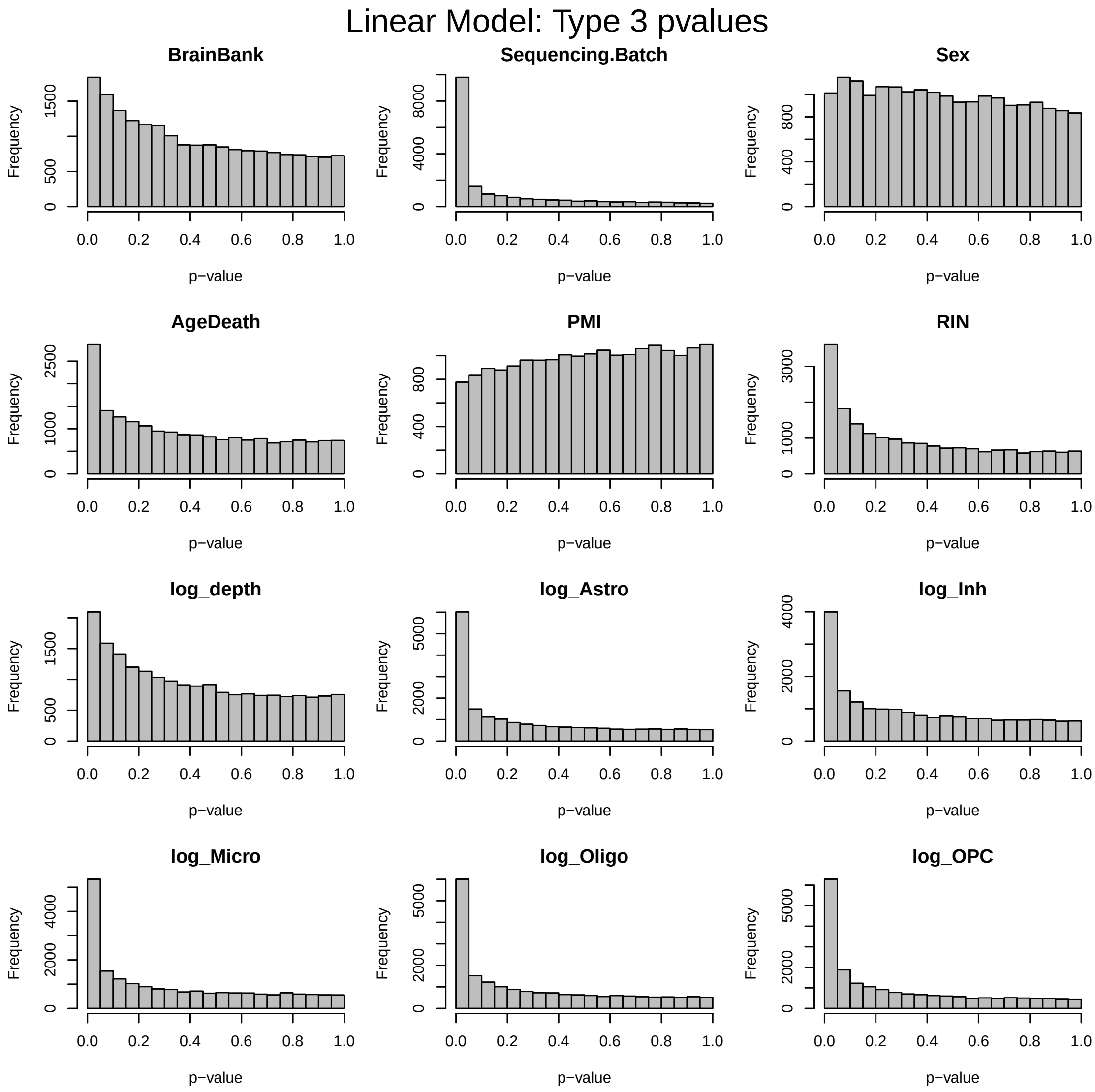


Figure 3.44: The p-values for each covariate vs. genome-wide gene expression for the ASD study, assessed by a linear model for log-transformed gene expression. Cell type proportions were included as log ratios, e.g., `log_Astro` is log(Astro proportion/Excitatory neuron proportion). The analyses for this figure were done using 82 out of 85 samples with non-missing PMI values.

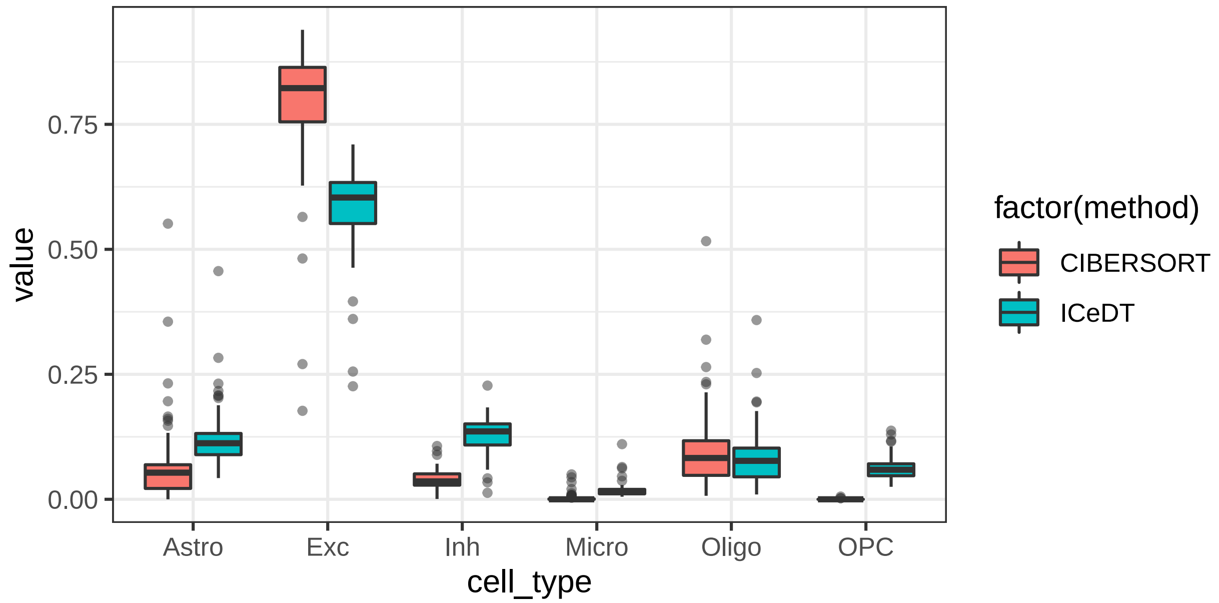


Figure 3.45: Box plots of cell type fraction estimates for UCLA-ASD data.

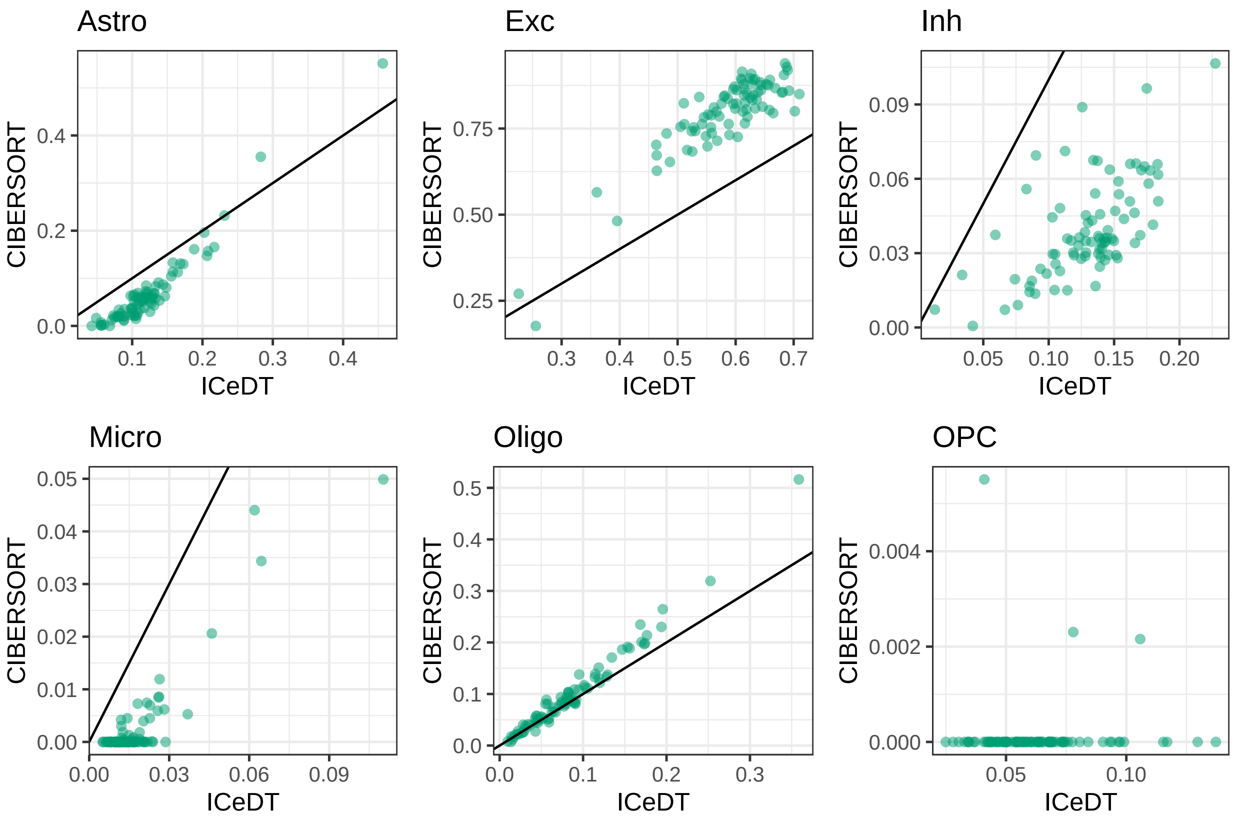


Figure 3.46: Scatter plots of cell type fraction estimates for UCLA-ASD data.

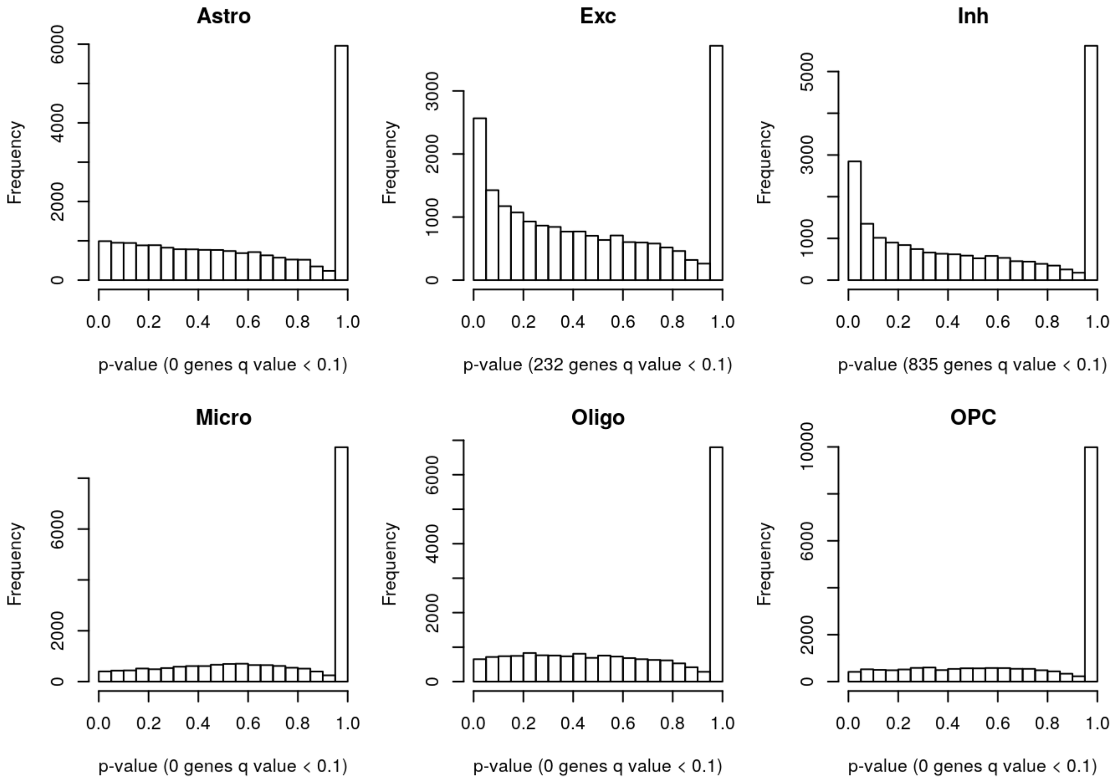


Figure 3.47: CARseq p-value distribution in the ASD study.

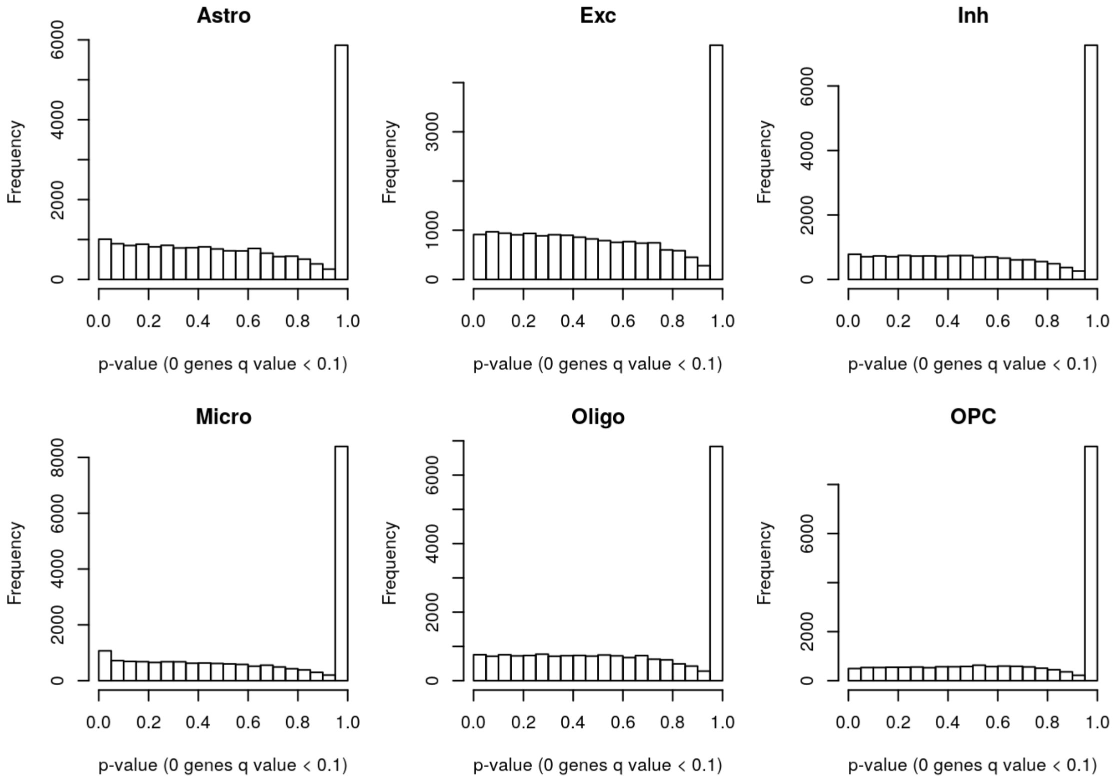


Figure 3.48: CARseq p-value distribution in the ASD study where the case-control label has been permuted to reflect the null distribution.

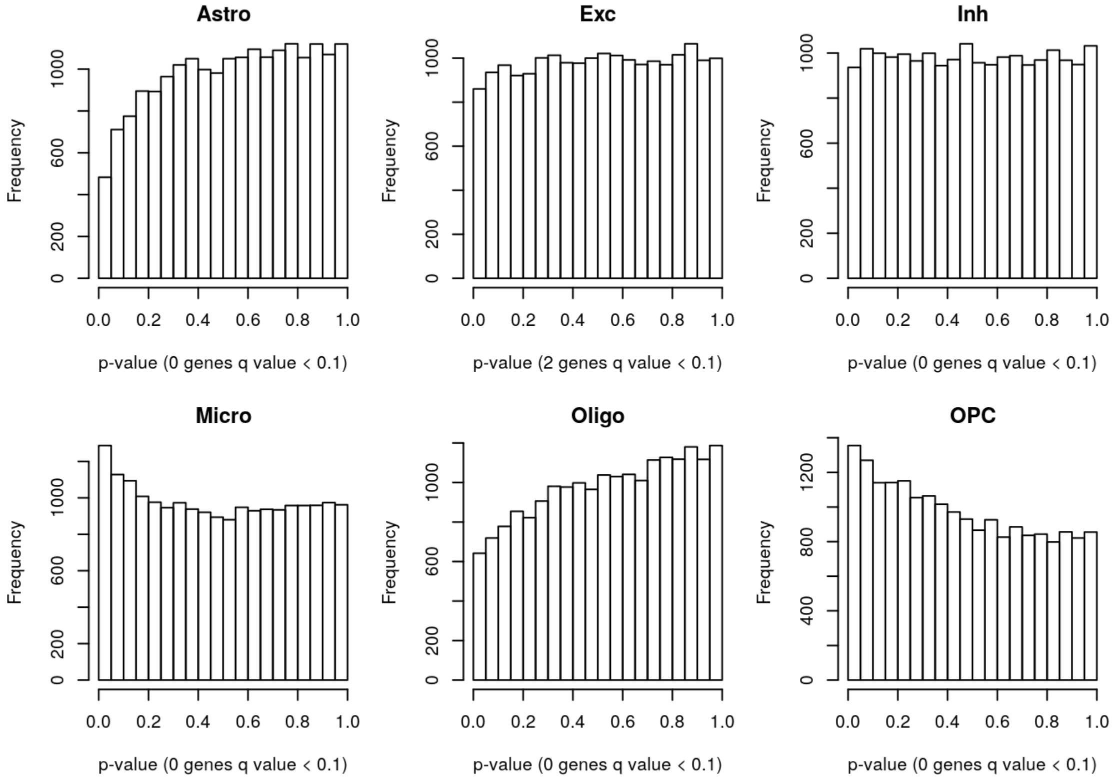


Figure 3.49: TOAST p-value distribution in the ASD study.

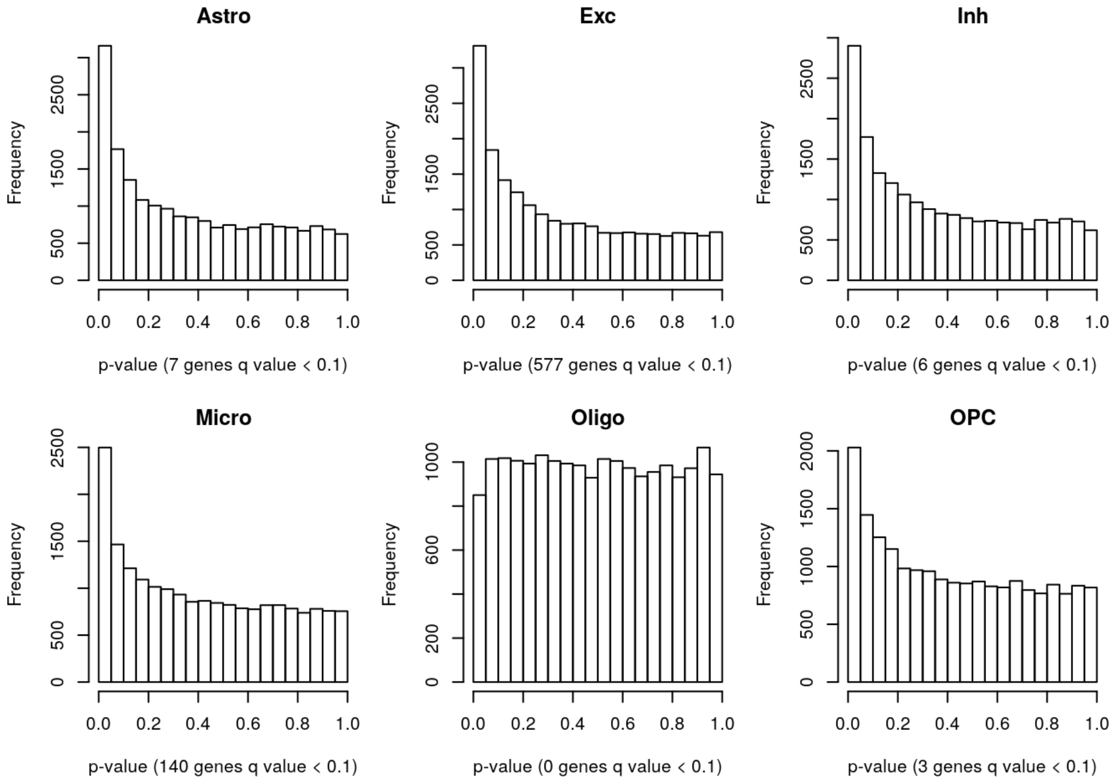


Figure 3.50: TOAST p-value distribution in the ASD study where the case-control label has been permuted to reflect the null distribution.

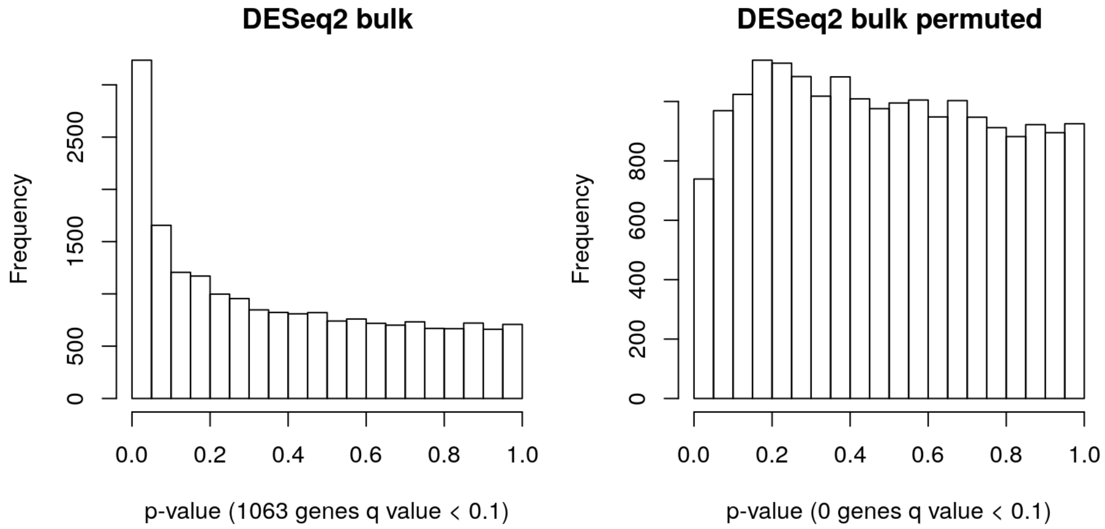


Figure 3.51: DESeq2 p-value distribution in the ASD study where the case-control label is either unpermuted or permuted.

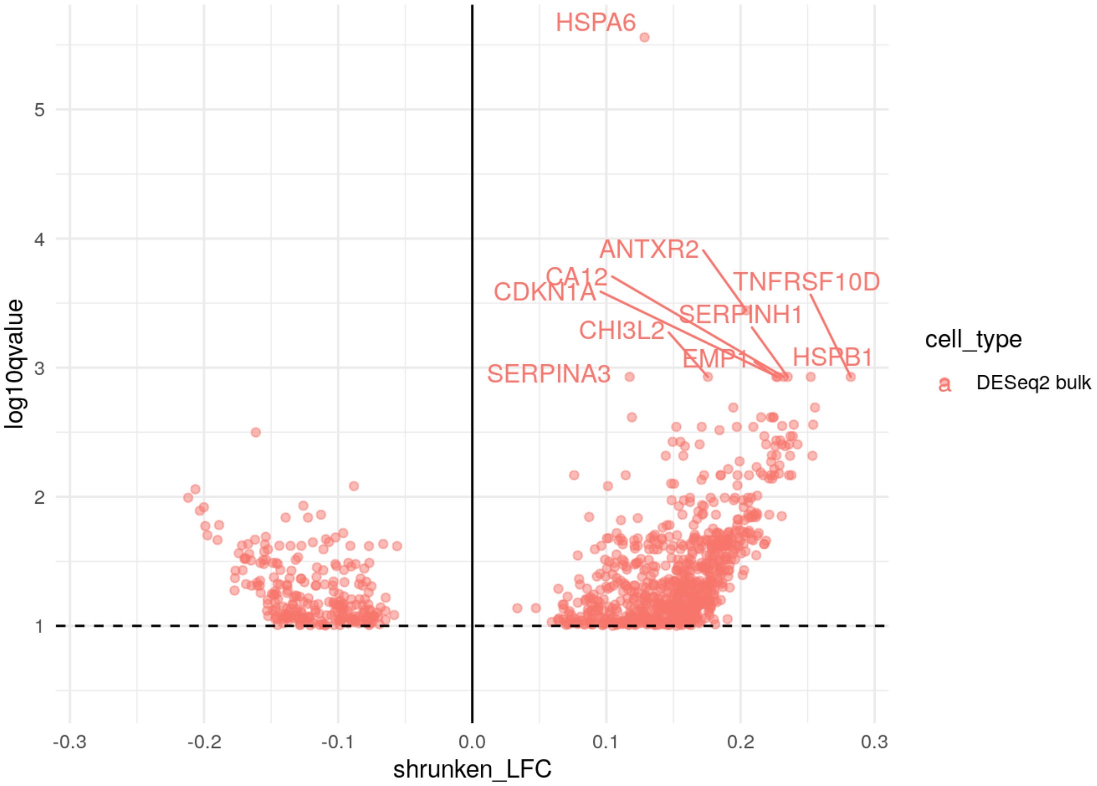


Figure 3.52: DESeq2 volcano plot in the ASD study.

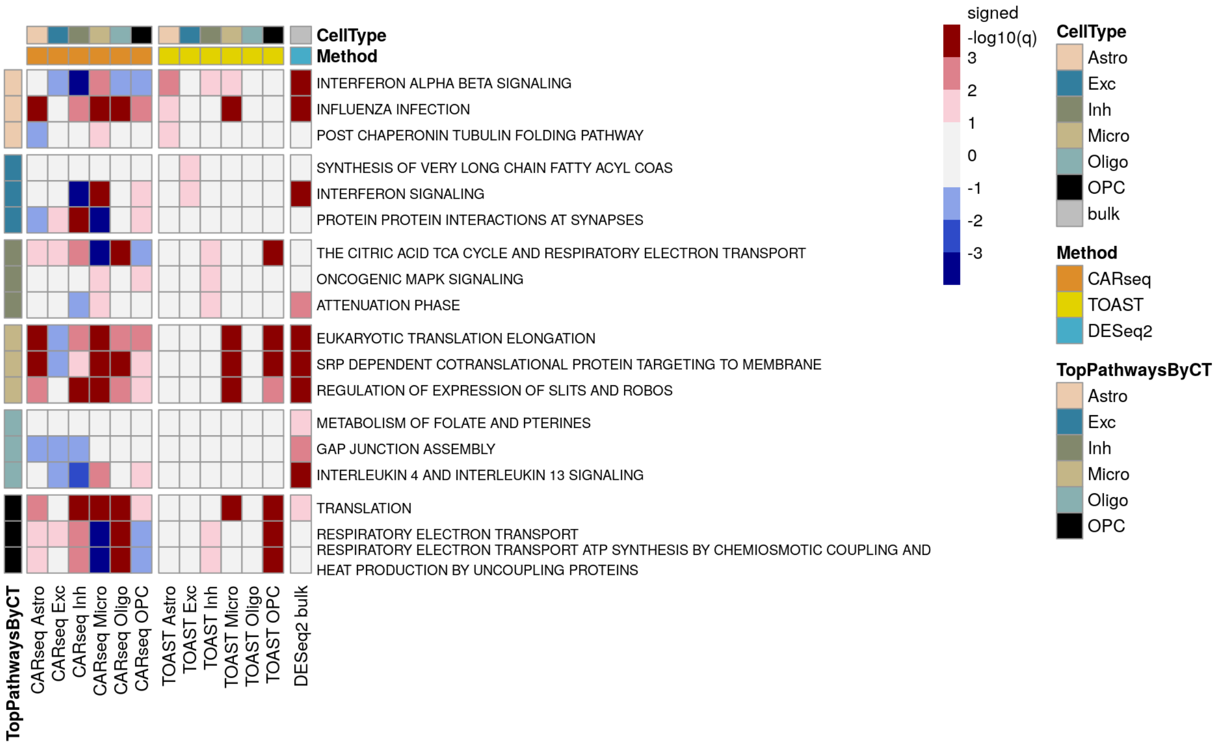


Figure 3.53: REACTOME GSEA ranked by TOAST in the ASD study.

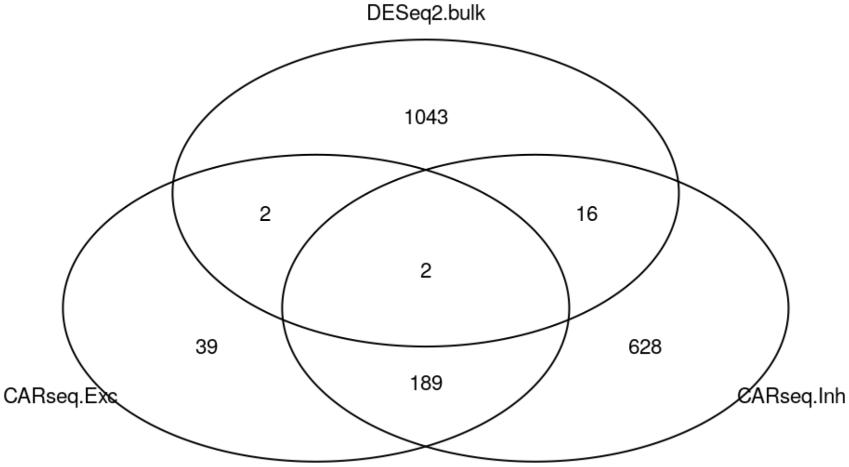


Figure 3.54: Venn plot of DEGs (q-value < 0.1) in the ASD study.

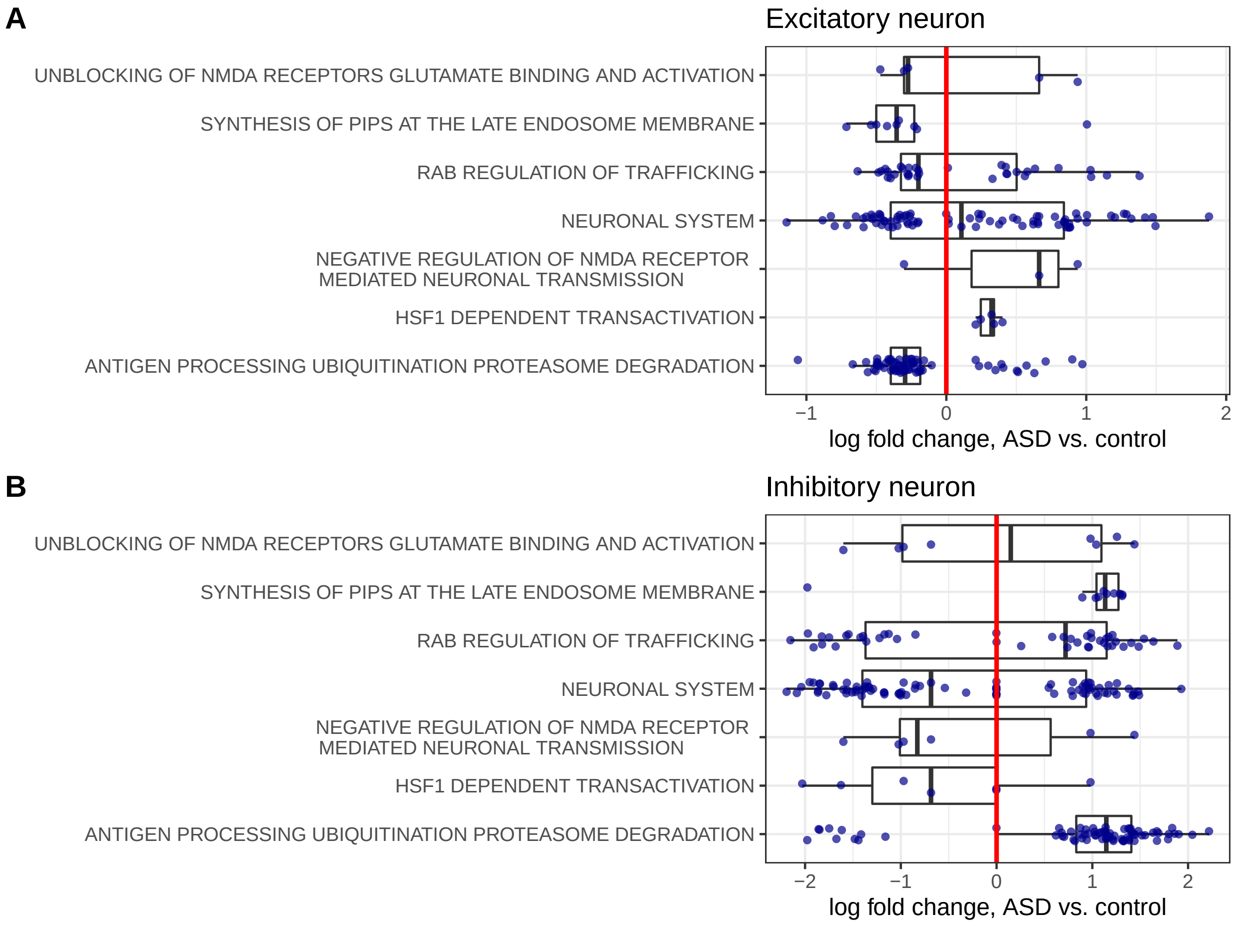


Figure 3.55: Shrunken log fold change estimates (ASD vs. controls) for genes belonging to some REACTOME pathways. Panel A includes all the pathways related with NDMA. Panels B and C include the pathways identified by GSEA in excitatory/inhibitory neurons, from either SCZ or ASD studies. Only the genes with CT-specific-DE p-value (comparing ASD vs. controls) smaller than 0.05 are shown.

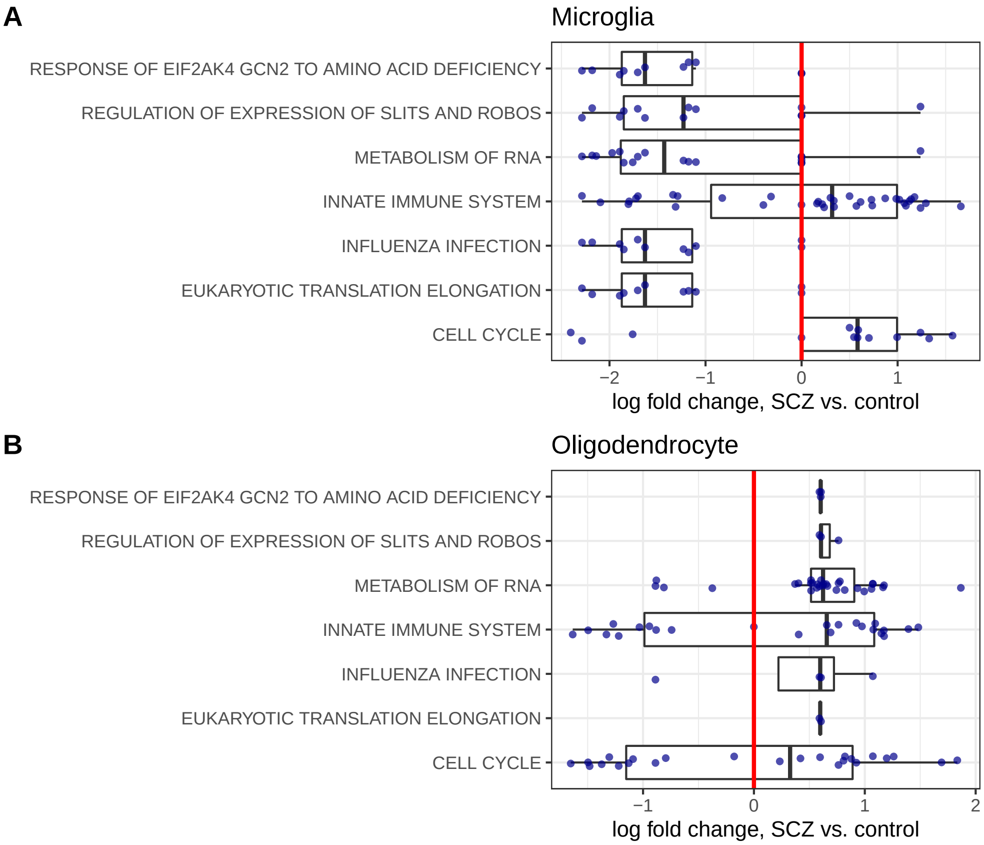


Figure 3.56: Shrunken log fold change estimates (ASD vs. controls) for genes belonging to the REACTOME pathways identified by GSEA in microglias or oligodendrocytes, from either SCZ or ASD studies. Only the genes with CT-specific-DE p-value (comparing ASD vs. controls) smaller than 0.05 are shown.

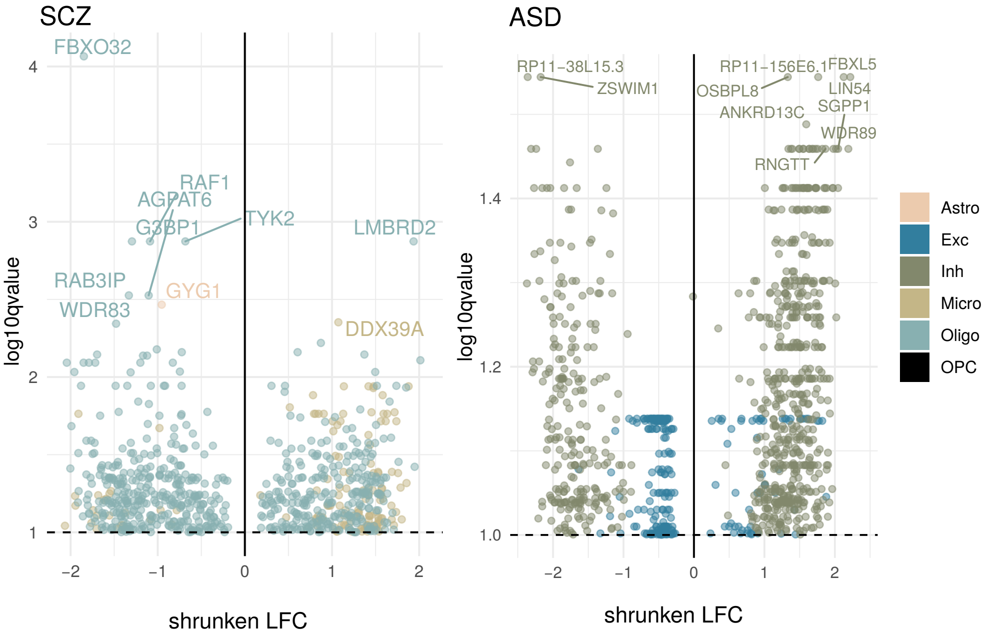


Figure 3.57: Volcano plot of -log10(q-value) vs. shrunken log fold change (LFC) for differentially expressed genes (DEGs) in SCZ (left panel) and ASD (right panel) inferred by CARseq. Only genes passing a q-value cutoff of 0.1 and an absolute value of LFC > 0.01 are shown. Top 10 CT-specific DEGs are labeled for each study.

BIBLIOGRAPHY

Aitchison, J. and Egozcue, J. J. (2005). Compositional data analysis: where are we and where should we be heading? *Mathematical Geology*, 37(7):829–850.

Ajram, L., Horder, J., Mendez, M., Galanopoulos, A., Brennan, L., Wichers, R., Robertson, D., Murphy, C., Zinkstok, J., et al. (2017). Shifting brain inhibitory balance and connectivity of the prefrontal cortex of adults with autism spectrum disorder. *Translational Psychiatry*, 7(5):e1137–e1137.

Andor, N., Harness, J. V., MÃijller, S., Mewes, H. W., and Petritsch, C. (2014). EXPANDS: expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics*, 30(1):50–60.

Anttila, V., Bulik-Sullivan, B., Finucane, H. K., Walters, R. K., Bras, J., Duncan, L., Escott-Price, V., Falcone, G. J., Gormley, P., et al. (2018). Analysis of shared heritability in common disorders of the brain. *Science*, 360(6395):eaap8757.

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, 20(2):163–172.

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, 30(5):413–421.

Cattane, N., Richetto, J., and Cattaneo, A. (2018). Prenatal exposure to environmental insults and enhanced risk of developing schizophrenia and autism spectrum disorder: focus on biological pathways and epigenetic mechanisms. *Neuroscience & Biobehavioral Reviews*.

Chedom-Fotso, D., Ahmed, A. A., and Yau, C. (2016). OncoPhase: Quantification of somatic mutation cellular prevalence using phase information. *bioRxiv*, page 046631.

Cox, D. R. and Reid, N. (1987). Parameter Orthogonality and Approximate Conditional Inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(1):1–39.

Deshwar, A. G., Vembu, S., Yung, C. K., Jang, G., Stein, L., and Morris, Q. (2015). PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):35.

El-Kebir, M., Oesper, L., Acheson-Field, H., and Raphael, B. J. (2015). Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16(1):278.

Fischer, A., VÃązquez-GarcÃŋa, I., Illingworth, C., and Mustonen, V. (2014). High-Definition Reconstruction of Clonal Composition in Cancer. *Cell Reports*, 7(5):1740–1752.

Frampton, G. M., Fichtenholtz, A., Otto, G. A., Wang, K., Downing, S. R., He, J., Schnall-Levin, M., White, J., Sanford, E. M., et al. (2013). Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nature Biotechnology*, 31(11):1023–1031.

Fromer, M., Roussos, P., Sieberts, S. K., Johnson, J. S., Kavanagh, D. H., Perumal, T. M., Ruderfer, D. M., Oh, E. C., Topol, A., et al. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature Neuroscience*, 19(11):1442–1453.

Gablonsky, J. M. and Kelley, C. T. (2001). A locally-biased form of the DIRECT algorithm. *Journal of Global Optimization*, 21(1):27–37.

Gandal, M. J., Zhang, P., Hadjimichael, E., Walker, R. L., Chen, C., Liu, S., Won, H., van Bakel, H., Varghese, M., et al. (2018). Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science*, 362(6420):eaat8127.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., et al. (2012). Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine*, 366(10):883–892.

Gillis, N. (2014). The why and how of nonnegative matrix factorization. *Regularization, optimization, kernels, and support vector machines*, 12(257):257–291.

Greaves, M. and Maley, C. C. (2012). Clonal evolution in cancer. *Nature*, 481(7381):306–313.

Greenhalgh, A. D., David, S., and Bennett, F. C. (2020). Immune cell regulation of glia during cns injury and disease. *Nature Reviews Neuroscience*, pages 1–14.

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., et al. (2014). TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.*, 24(11):1881–1893.

Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S. R., Aguet, F., Gelfand, E., et al. (2017). Massively parallel single-nucleus rna-seq with dronc-seq. *Nature methods*, 14(10):955–958.

Hajirasouliha, I., Mahmoody, A., and Raphael, B. J. (2014). A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, 30(12):i78–i86.

Hanahan, D. and Weinberg, R. A. (2000). The Hallmarks of Cancer. *Cell*, 100(1):57–70.

Hausser, A. and Schlett, K. (2019). Coordination of ampa receptor trafficking by rab gtpases. *Small GTPases*, 10(6):419–432.

Hodge, R. D., Bakken, T. E., Miller, J. A., Smith, K. A., Barkan, E. R., Graybuck, L. T., Close, J. L., Long, B., Johansen, N., et al. (2019). Conserved cell types with divergent features in human versus mouse cortex. *Nature*, 573(7772):61–68.

Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., et al. (2016). Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell*, 165(1):35–44.

Ishimura, R., Nagy, G., Dotu, I., Chuang, J. H., and Ackerman, S. L. (2016). Activation of gcn2 kinase by ribosome stalling links translation elongation with translation initiation. *Elife*, 5:e14295.

Jardri, R., Hugdahl, K., Hughes, M., Brunelin, J., Waters, F., Alderson-Day, B., Smailes, D., Sterzer, P., Corlett, P. R., et al. (2016). Are hallucinations due to an imbalance between excitatory and inhibitory influences on the brain? *Schizophrenia bulletin*, 42(5):1124–1134.

Jiang, Y., Qiu, Y., Minn, A. J., and Zhang, N. R. (2016). Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proceedings of the National Academy of Sciences*, 113(37):E5528–E5537.

Jiang, Y., Redmond, D., Nie, K., Eng, K. W., Clozel, T., Martin, P., Tan, L. H., Melnick, A. M., Tam, W., and Elemento, O. (2014). Deep sequencing reveals clonal evolution patterns and mutation events associated with relapse in B-cell lymphomas. *Genome Biology*, 15:432.

Johnson, S. G. (2014). *The NLopt nonlinear-optimization package*.

Jonckheere, A. R. (1954). A Distribution-Free k-Sample Test Against Ordered Alternatives. *Biometrika*, 41(1/2):133–145.

Kaneko, N., Herranz-Pérez, V., Otsuka, T., Sano, H., Ohno, N., Omata, T., Nguyen, H., Thai, T., Nambu, A., et al. (2018). New neurons use slit-robo signaling to migrate through the glial meshwork and approach a lesion for functional regeneration. *Science advances*, 4(12):eaav0618.

Kehrer, C., Maziashvili, N., Dugladze, T., and Gloveli, T. (2008). Altered excitatory-inhibitory balance in the nmda-hypofunction model of schizophrenia. *Frontiers in molecular neuroscience*, 1:6.

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29.

Lee, J., Mueller, P., Sengupta, S., Gulukota, K., and Ji, Y. (2014). Bayesian inference for tumor subclones accounting for sequencing and structural variants. *arXiv preprint arXiv:1409.7158*.

Lee, J., Müller, P., Sengupta, S., Gulukota, K., and Ji, Y. (2016). Bayesian inference for intratumour heterogeneity in mutations and copy number variation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(4):547–563.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883.

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., Haag, J. D., Gould, M. N., Stewart, R. M., and Kendziorski, C. (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, 29(8):1035–1043.

Li, B. and Li, J. Z. (2014). A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biology*, 15(9):1.

Li, Z., Wu, Z., Jin, P., and Wu, H. (2019). Dissecting differential signals in high-throughput data from complex tissues. *Bioinformatics*, 35(20):3898–3905.

Lin, M., Zhao, D., Hrabovsky, A., Pedrosa, E., Zheng, D., and Lachman, H. M. (2014). Heat shock alters the expression of schizophrenia and autism candidate genes in an induced pluripotent stem cell model of the human telencephalon. *PloS one*, 9(4):e94968.

Lord, C., Brugha, T. S., Charman, T., Cusack, J., Dumas, G., Frazier, T., Jones, E. J. H., Jones, R. M., Pickles, A., et al. (2020). Autism spectrum disorder. *Nature Reviews Disease Primers*, 6(1):1–23.

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

Lun, A. T., McCarthy, D. J., and Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. *F1000Research*, 5.

Luo, X., Yang, C., and Wei, Y. (2019). Detection of cell-type-specific risk-CpG sites in epigenome-wide association studies. *Nature Communications*, 10(1):1–12.

Malikic, S., McPherson, A. W., Donmez, N., and Sahinalp, C. S. (2015). Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31(9):1349–1356.

McCartney, A. J., Zolov, S. N., Kauffman, E. J., Zhang, Y., Strunk, B. S., Weisman, L. S., and Sutton, M. A. (2014). Activity-dependent pi (3, 5) p2 synthesis controls ampa receptor trafficking during synaptic depression. *Proceedings of the National Academy of Sciences*, 111(45):E4896–E4905.

McGranahan, N., Furness, A. J., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S. K., Jamal-Hanjani, M., Wilson, G. A., Birkbak, N. J., et al. (2016). Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*, 351(6280):1463–1469.

Miller, C. A., White, B. S., Dees, N. D., Griffith, M., Welch, J. S., Griffith, O. L., Vij, R., Tomasson, M. H., Graubert, T. A., et al. (2014). SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *PLoS Computational Biology*, 10(8):e1003665.

Molenberghs, G. and Verbeke, G. (2007). Likelihood Ratio, Score, and Wald Tests in a Constrained Parameter Space. *The American Statistician*, 61(1):22–27.

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., and Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12(5):453–457.

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., Khodadoust, M. S., Esfahani, M. S., Luca, B. A., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*, page 1.

Nik-Zainal, S., VanÂăLoo, P., Wedge, D., Alexandrov, L., Greenman, C., Lau, K., Raine, K., Jones, D., Marshall, J., et al. (2012). The Life History of 21 Breast Cancers. *Cell*, 149(5):994–1007.

Nocedal, J. and Wright, S. J. (2006). *Numerical optimization*. Springer series in operations research. Springer, New York, 2nd ed edition.

Nowakowski, T. J., Pollen, A. A., Di Lullo, E., Sandoval-Espinosa, C., Bershteyn, M., and Kriegstein, A. R. (2016). Expression analysis highlights AXL as a candidate Zika virus entry receptor in neural stem cells. *Cell Stem Cell*, 18(5):591–596.

Oesper, L., Mahmoody, A., and Raphael, B. J. (2013). THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biology*, 14(7):R80.

Oesper, L., Satas, G., and Raphael, B. J. (2014). Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics*, 30(24):3532–3540.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572.

Owen, M. J., Sawa, A., and Mortensen, P. B. (2016). Schizophrenia. *The Lancet*, 388(10039):86–97.

Parikshak, N. N., Swarup, V., Belgard, T. G., Irimia, M., Ramaswami, G., Gandal, M. J., Hartl, C., Leppa, V., Ubieta, L. d. l. T., et al. (2016). Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature*, 540(7633):423–427.

Park, M. Y. (2006). Generalized linear models with regularization. *PhD Thesis*, Stanford University, Department of Statistics.

Pavlov, V. A. and Tracey, K. J. (2017). Neural regulation of immunity: molecular mechanisms and clinical translation. *Nature Neuroscience*, 20(2):156–166.

Petrelli, F., Pucci, L., and Bezzi, P. (2016). Astrocytes and microglia and their potential link with autism spectrum disorders. *Frontiers in cellular neuroscience*, 10:21.

Prata, J., Santos, S. G., Almeida, M. I., Coelho, R., and Barbosa, M. A. (2017). Bridging autism spectrum disorders and schizophrenia through inflammation and biomarkers-pre-clinical and clinical investigations. *Journal of neuroinflammation*, 14(1):179.

Racle, J., Jonge, K. d., Baumgaertner, P., Speiser, D. E., and Gfeller, D. (2017). Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife*, 6:e26476.

Raymond, L. J., Deth, R. C., and Ralston, N. V. (2014). Potential role of selenoenzymes and antioxidant metabolism in relation to autism etiology and pathology. *Autism Research and Treatment*, 2014:1–15.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., et al. (2017). Science forum: the human cell atlas. *Elife*, 6:e27041.

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

Robinson, M. D. and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–332.

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-CÃťtÃľ, A., and Shah, S. P. (2014). PyClone: statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4):396–398.

Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., et al. (2020). Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell*, 180(3):568–584.

Schumacher, T. N. and Schreiber, R. D. (2015). Neoantigens in cancer immunotherapy. *Science*, 348(6230):69–74.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions. *Journal of the American Statistical Association*, 82(398):605–610.

Shen, R. and Seshan, V. E. (2016). FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Research*, 44(16):e131–e131.

Shen-Orr, S. S., Tibshirani, R., Khatri, P., Bodian, D. L., Staedtler, F., Perry, N. M., Hastie, T., Sarwal, M. M., Davis, M. M., and Butte, A. J. (2010). Cell typeâĂŞspecific gene expression differences in complex tissues. *Nature Methods*, 7(4):287–289.

Stark, P. and L. Parker, R. (1995). Bounded-Variable Least-Squares: an Algorithm and Applications. *Computational Statistics*, 10.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445.

Sun, W., Wright, F. A., Tang, Z., Nordgard, S. H., Loo, P. V., Yu, T., Kristensen, V. N., and Perou, C. M. (2009). Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Research*, 37(16):5365–5377.

TERPSTRA, T. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indag Math*, 14:327–333.

The Brainstorm Consortium, Anttila, V., Bulik-Sullivan, B., Finucane, H. K., Walters, R. K., Bras, J., Duncan, L., Escott-Price, V., Falcone, G. J., et al. (2018). Analysis of shared heritability in common disorders of the brain. *Science*, 360(6395).

Van Loo, P. (2018). When should I use ASCAT? When should I use Battenberg?

Van Loo, P., Nordgard, S. H., LingjÃęrde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., et al. (2010). Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915.

Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., Bhaduri, A., Goyal, N., Rowitch, D. H., and Kriegstein, A. R. (2019). Single-cell genomics identifies cell typeâĂŞspecific molecular changes in autism. *Science*, 364(6441):685–689.

Wang, T., Li, B., Nelson, C. E., and Nabavi, S. (2019). Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics*, 20(1):40.

Wang, Y., Waters, J., Leung, M. L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., et al. (2014). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155–160.

Wilson, D. R., Jin, C., Ibrahim, J. G., and Sun, W. (2019). ICeD-T Provides Accurate Estimates of Immune Cell Abundance in Tumor Samples by Allowing for Aberrant Gene Expression Patterns. *Journal of the American Statistical Association*, 0(0):1–11.

Yuan, K., Sakoparnig, T., Markowetz, F., and Beerenwinkel, N. (2015). BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biology*, 16(1):36.

Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., Lawrence, M. S., Zhang, C.-Z., Wala, J., et al. (2013). Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, 45(10):1134–1140.

Zaitsev, K., Bambouskova, M., Swain, A., and Artyomov, M. N. (2019). Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nature Communications*, 10(1):2209.

Zhang, T., Choi, J., Kovacs, M. A., Shi, J., Xu, M., Goldstein, A. M., Iles, M. M., Duffy, D., MacGregor, S., et al. (2018). Cell-type specific eQTL of primary melanocytes facilitates identification of melanoma susceptibility genes. *Genome Research*, 28:1621–1635.

Zheng, S. C., Breeze, C. E., Beck, S., and Teschendorff, A. E. (2018). Identification of differentially methylated cell types in epigenome-wide association studies. *Nature Methods*, 15(12):1059–1066.

Zhong, Y. and Liu, Z. (2012). Gene expression deconvolution in linear space. *Nature Methods*, 9(1):8–9.

Zhou, Y.-H., Xia, K., and Wright, F. A. (2011). A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics*, 27(19):2672–2678.