bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cell Type **A**ware analysis of **R**NA-**seq** data (CARseq) reveals difference and similarities of the molecular mechanisms of Schizophrenia and Autism

Chong Jin
Department of Biostatistics
University of North Carolina at Chapel Hill

Mengjie Chen
Genetic Medicine
University of Chicago

Danyu Lin
Department of Biostatistics
University of North Carolina at Chapel Hill

Wei Sun*
Public Health Science Division
Fred Hutchinson Cancer Research Center
wsun@fredhutch.org

July 13, 2020

Abstract

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. We propose a new framework to address these limitations: **C**ell Type **A**ware analysis of **R**NA-**seq** (CARseq). After evaluating its performance in simulations, we apply CARseq to compare gene expression of schizophrenia/autism subjects versus controls. Our results show that

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

these two neurodevelopmental disorders differ from each other in terms of cell type composition changes and genes related with 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.

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 [1], identifying genes and cell types related to Zika virus infection [2] or melanoma [3]. Most methods for differential expression studies using bulk RNA-seq data [4–6] do not consider cell type compositions. A few exceptions include csSAM [7] and TOAST [8], 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 [9, 10].

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 [11, 12]. 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 to improve 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) [13], 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 [7, 8], 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.

We apply CARseq to assess gene expression difference of schizophrenia (SCZ) or autism spectrum disorder (ASD) subjects versus 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 [14]. Two diseases have shared vulnerability genes and overlapping symptoms [15]. For example, ASD is characterized by deficit social interaction and repetitive behaviors, which are similar to the negative symptom of SCZ including social withdrawal and impaired motivation. There are also many differences,

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 symptoms 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 [16]. 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 [17].

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 fresh 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.

Results

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 [18]. 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 1). The model also allows the inclusion of cell type-independent covariates; see Method section for details.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

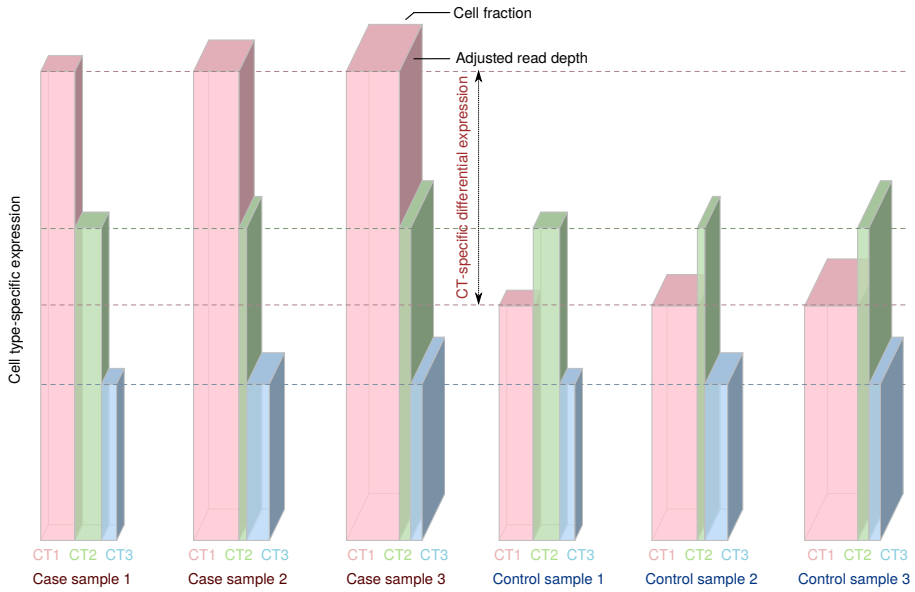

Figure 1: Each grouped bar illustrates the total expression of a gene in a bulk sample (the volume of the grouped bar) that is the summation of gene expression from individual cell types (each bar). The depth of each grouped bar is proportional to the covariate-adjusted read-depth. The width of each 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.

Benchmarking methods through simulations

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 [7] and TOAST [8]. 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 [8] and it cannot account for covariates. TOAST and CARseq are more similar since they both combine the estimation and 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 S4-S7).

We simulated CT-specific expression that roughly mirrors the gene expression derived from single nucleus RNA-seq (snRNA-seq) data derived from the middle temporal gyrus (MTG) of the human brain [19]. We simulated the cell fractions to resemble our estimates from the Common Mind Consortium (CMC) bulk RNA-sequencing data [20] (Figure S12). 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.

The benchmark consists of simulations in different sample sizes and different patterns of differential expression (Figure 2). 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., $n = 50$) for a relatively large effect size of 2. This is due to the uncertainty to estimate CT-specific expression.

CARseq: comparing schizophrenia subjects versus controls

Schizophrenia (SCZ) is a severe neuropsychiatric disorder that affects approximately 1% of word-wide population [21]. There is strong evidence from both human and animal studies 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%) [21]. 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 Common-Mind Consortium (CMC) [20], hereafter referred to as CMC-SCZ study. After filtering out the outlier samples reported by Fromer et al. [20], we had 250 SCZ subjects and 277 controls.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

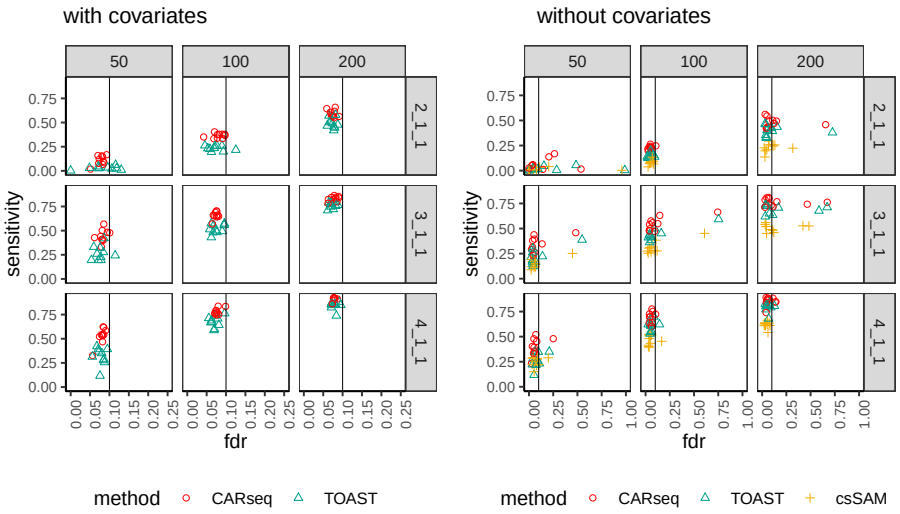

Figure 2: The FDR vs. sensitivity of several methods testing for CT-specific DE, when (a) the covariate is provided to the method, and (b) when the covariate is not provided to the method. There are 10 simulation replicates for each combination of sample size (columns) as the total number of case-control 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.

We estimated cell type proportions for six cell types: excitatory neurons (Exc), inhibitory neurons (Inh), astrocyte (Astro), microglia (Micro), oligodendrocyte (Oligo) using CIBERSORT [22] and ICeD-T [12]. The estimates from these two methods are highly correlated, though with some noticeable differences (Supplementary Materials Section B.2.1, Figures S12-S13). 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(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

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

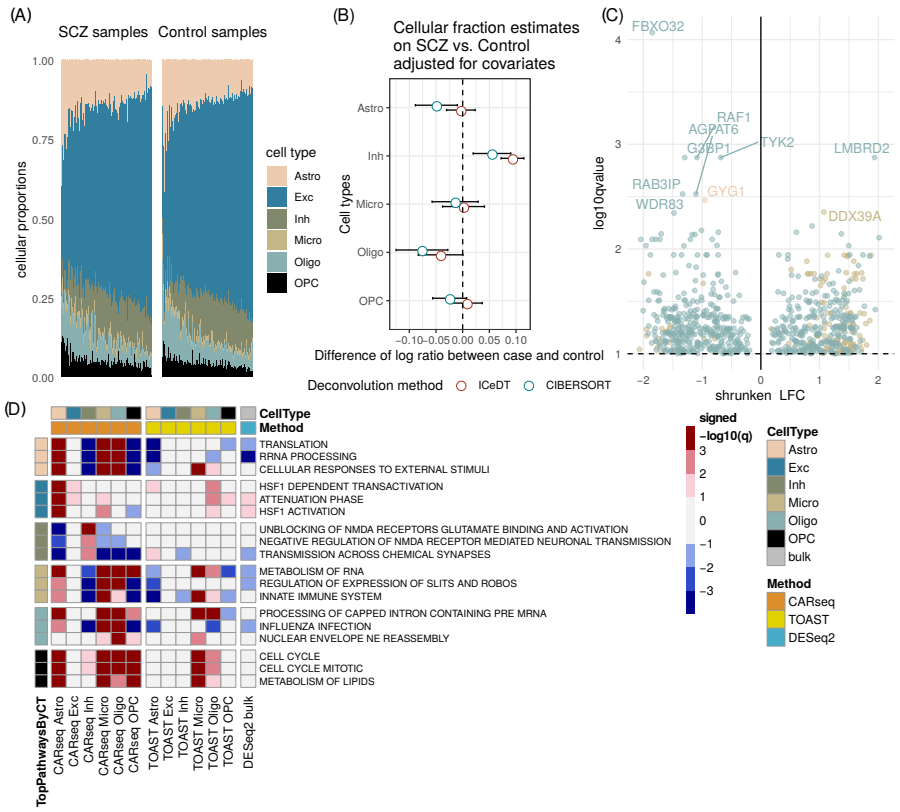

Figure 3: CARseq on gene expression data between schizophrenia (SCZ) and controls. (a) Estimated cell fractions by ICeD-T sorted by increasing fractions of excitatory neurons. (b) The effect size of case-control status on relative cell fractions against excitatory neurons (log ratio of the cell type of interest vs. excitatory neuron). The standard errors are denoted by bars. (c) Volcano plot of -log10(q-value) vs. shrunken log fold change (LFC) for differentially expressed genes (DEGs) 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. (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.

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).

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Since cell type fraction estimates from CIBERSORT and ICeD-T are highly correlated, we present the CARseq results using cell type fraction 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 by CARseq (see Figure 3(b) for a volcano plot and Figure S15 for p-value distributions). In contrast, TOAST identified 3 DEGs in inhibitory neurons, 30 DEGs in microglia, and 1 DEG in oligodendrocytes (See Figure S17 for p-value distribution). 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 (Figure S16 and S18). These results are consistent with our simulation results that CARseq can identify differentially expressed genes with a higher power than TOAST, while controlling FDR.

Although we did not find any DE genes in inhibitory neurons or excitatory neurons at q-value cutoff 0.1, gene set enrichment analysis (GSEA) using the rankings of all the genes based on CT-specific DE p-values recover some interesting pathways (Figure 3(C)). 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 [21]. The majority of the DE genes in NMDA pathways have lower expression levels in SCZ subjects than controls (Figure S23), 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 S23). This is consistent with previous findings that heat shock response plays a crucial role in the response of brain cells to prenatal environmental insults [23].

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(D), Figure S24), supporting the observations of activation of microglia in SCZ subjects [17]. It is interesting that these pathways are also enriched in oligodendrocyte, but they are down regulated in SCZ subjects than controls (Figure 3(D), Figure S24), suggesting inactivation of oligodendrocyte. We also found Slit-Robo signaling pathway is down/up -regulated in microglia and oligodendrocyte, respectively (Figure S24). 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 [24]. Our findings suggest microglia and oligodendrocyte may take different roles in this process in SCZ subjects.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CARseq: comparing ASD subjects versus controls

Autism spectrum disorder (ASD) affect more than 1% of population, with heritability estimated to be 68% to 96% [25]. Individuals with ASD are often impaired in social communication and social interaction, and limit themselves to repetitive behaviors and interests from an early age [25]. 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 [26]. We analyzed the bulk RNA-seq data from ASD subjects and controls, published by a UCLA group [27, 28], 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 [27]. 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 4(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 [29].

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 S28). TOAST recovers 2 DEGs in excitatory neurons and no DEGs in the other five cell types (Figure S30). We also sought to evaluate the FDR control by repeating our analysis after permuting case/control labels (Figure S29 and S31) 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

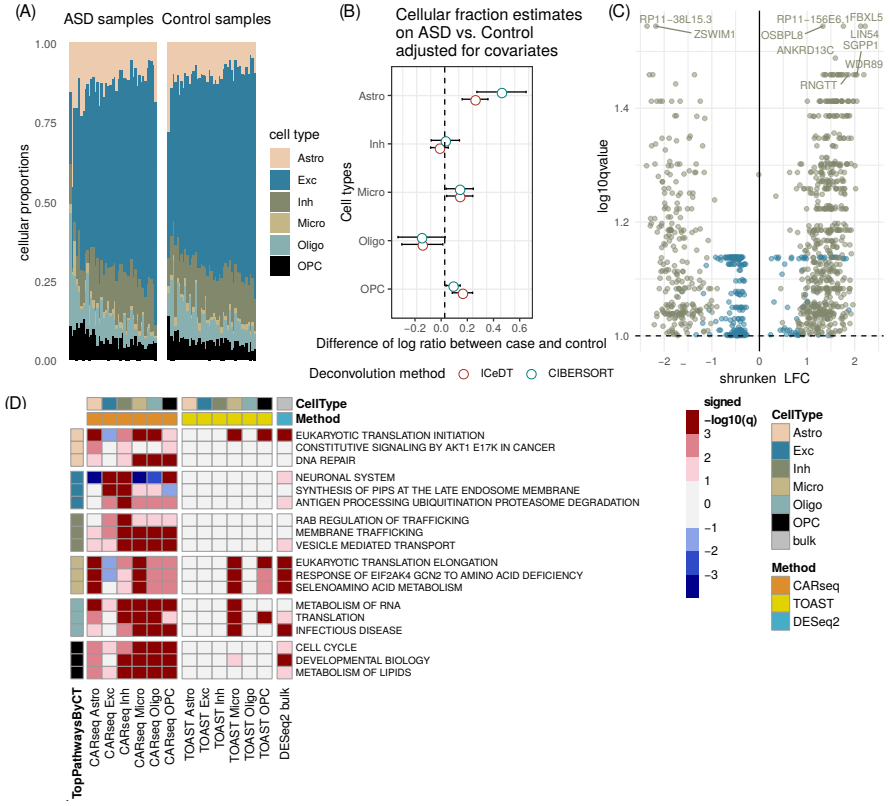

Figure 4: CARseq on the autism spectrum disorder (ASD) bulk expression data. (a) Estimated cell fractions by ICeD-T sorted by increasing fractions of excitatory neurons. (b) The effect size of case-control status on relative cell fractions against excitatory neurons (log ratio of the cell type of interest vs. excitatory neuron). The standard errors are denoted by bars. (c) Volcano plot of -log10(q-value) vs. shrunken log fold change (LFC) for differentially expressed genes (DEGs) 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. (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.

enriched among the DE genes in inhibitory neurons (p-value 5.8×10^{-7}) and excitatory neurons (p-value 3.3×10^{-4}), and they are significantly depleted among the DE genes in microglia (p-value 2.9×10^{-7}) (Table S5). Our findings are consistent with the results reported using snRNA-seq data [30]. Enrichment

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [6] (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 [31, 32]. The log fold changes of DE genes show that these two pathways are up-regulated in ASD inhibitory neurons than controls, but down-regulated in ASD excitatory neurons (Figure S36). 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 S36), 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 [33], which is corroborated by our findings that suggest their connections with ASD.

Comparing DE testing by CARseq versus DESeq2

DESeq2 [6] 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 S22 and S35). 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), 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), suggesting that many DEseq2 DE genes in ASD are indeed confounded by cell type compositions.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 5(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 S24(A) vs. S36(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 5(B), Table S6). 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 [34, 35].

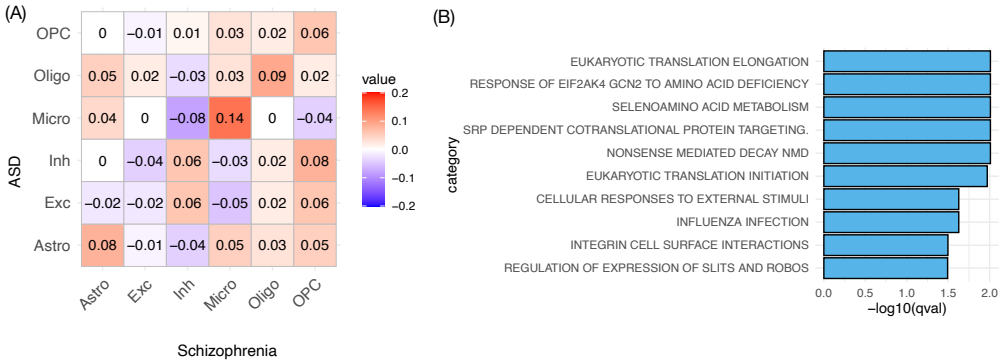

Figure 5: (A) The correlation matrix of -log10(Microglia-specific-DE p-values) (calculated by CARseq) between CMC-SCZ and UCLA-ASD studies. (B) The REACTOME pathways that are over-represented by the 65 genes with microglia-specific-DE p-values smaller than 0.05 in both CMC-SCZ and UCLA-ASD studies.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. 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 [14]. 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 [36] 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 [16], 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 [37].

Stress-associated signaling in SCZ and ASD has also 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 and ASD [17], 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 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 rela-

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tive 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, of which the estimation relies on reference of cell type-specific gene expression data. Part of the reasons that we focus on SCZ and ASD is that cell type-specific gene expression data are available in brain tissues. We expect with the development of human cell atlas [38], 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 uncertainly 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 use them to generate cell type-specific gene expression signatures and validate the results of CARseq.

Methods

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 $\widehat{\rho}_{hi}$. for and sample i.

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 [13], 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a log link function, which is commonly used for negative binomial regression: $\widetilde{\mu}_{jih} = d_i^{\beta_d} \exp\left(\sum_{m=1}^{M_0} \gamma_{jhm} x_{ihm}\right)$, 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:

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

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 [8], 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.

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 [8] 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. See more discussions of our method and TOAST in Section A.6.1 in Supplementary Materials. All the analysis

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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) [39]. 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 [11] 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 S4-S7).

Estimation of cell type compositions

We use two reference-based methods to estimate cell type compositions of bulk tissue samples: CIBERSORT [22] and ICeD-T [12]. 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 [19]. 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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` [39].

The covariates selected in our model are mostly similar to those included in the original analysis [20] 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 S1). Surrogate variables were computing using the R package "sva" [39]. 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 [6].

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. [27] 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. [27] 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 [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

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(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.

FDR control procedure

We use q-value to control FDR [40]. 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`:

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

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:

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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/

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).

Author Contributions

W.S. conceived of the approach. C.J. implemented the methods and performed analysis, with inputs from W.S., M.C., and D.L. W.S. and C.J. wrote the paper, with inputs from M.C. and D.L.

Competing Interests statement

None declared.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

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

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

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

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

[5] 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., Kendziorski, C.: EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics **29**(8), 1035–1043 (2013)

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

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

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

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

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

[11] 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., Steiner, D.,

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Diehn, M., Alizadeh, A.A.: Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature Biotechnology, 1 (2019)

[12] Wilson, D.R., Jin, C., Ibrahim, J.G., Sun, W.: 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 (2019)

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

[14] Cattane, N., Richetto, J., Cattaneo, A.: 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 (2018)

[15] Anttila, V., Bulik-Sullivan, B., Finucane, H.K., Walters, R.K., Bras, J., Duncan, L., Escott-Price, V., Falcone, G.J., Gormley, P., Malik, R., *et al.*: Analysis of shared heritability in common disorders of the brain. Science **360**(6395), 8757 (2018)

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

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

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

[19] 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., Penn, O., Yao, Z., Eggermont, J., Höllt, T., Levi, B.P., Shehata, S.I., Aevermann, B., Beller, A., Bertagnolli, D., Brouner, K., Casper, T., Cobbs, C., Dalley, R., Dee, N., Ding, S.-L., Ellenbogen, R.G., Fong, O., Garren, E., Goldy, J., Gwinn, R.P., Hirschstein, D., Keene, C.D., Keshk, M., Ko, A.L., Lathia, K., Mahfouz, A., Maltzer, Z., McGraw, M., Nguyen, T.N., Nyhus, J., Ojemann, J.G., Oldre, A., Parry, S., Reynolds, S., Rimorin, C., Shapovalova, N.V., Somasundaram, S., Szafer, A., Thomsen, E.R., Tieu, M., Quon, G., Scheuermann, R.H., Yuste, R., Sunkin, S.M., Lelieveldt, B., Feng, D., Ng, L., Bernard, A., Hawrylycz, M., Phillips, J.W., Tasic, B., Zeng, H., Jones, A.R., Koch, C., Lein, E.S.: Conserved cell types with divergent features in human versus mouse cortex. Nature **573**(7772), 61–68 (2019)

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[20] Fromer, M., Roussos, P., Sieberts, S.K., Johnson, J.S., Kavanagh, D.H., Perumal, T.M., Ruderfer, D.M., Oh, E.C., Topol, A., Shah, H.R., Klei, L.L., Kramer, R., Pinto, D., Gümüş, Z.H., Cicek, A.E., Dang, K.K., Browne, A., Lu, C., Xie, L., Readhead, B., Stahl, E.A., Xiao, J., Parvizi, M., Hamamsy, T., Fullard, J.F., Wang, Y.-C., Mahajan, M.C., Derry, J.M.J., Dudley, J.T., Hemby, S.E., Logsdon, B.A., Talbot, K., Raj, T., Bennett, D.A., De Jager, P.L., Zhu, J., Zhang, B., Sullivan, P.F., Chess, A., Purcell, S.M., Shinobu, L.A., Mangravite, L.M., Toyoshiba, H., Gur, R.E., Hahn, C.-G., Lewis, D.A., Haroutunian, V., Peters, M.A., Lipska, B.K., Buxbaum, J.D., Schadt, E.E., Hirai, K., Roeder, K., Brennand, K.J., Katsanis, N., Domenici, E., Devlin, B., Sklar, P.: Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nature Neuroscience **19**(11), 1442–1453 (2016)

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

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

[23] Lin, M., Zhao, D., Hrabovsky, A., Pedrosa, E., Zheng, D., Lachman, H.M.: 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), 94968 (2014)

[24] Kaneko, N., Herranz-Pérez, V., Otsuka, T., Sano, H., Ohno, N., Omata, T., Nguyen, H., Thai, T., Nambu, A., Kawaguchi, Y., *et al.*: New neurons use slit-robo signaling to migrate through the glial meshwork and approach a lesion for functional regeneration. Science advances **4**(12), 0618 (2018)

[25] Lord, C., Brugha, T.S., Charman, T., Cusack, J., Dumas, G., Frazier, T., Jones, E.J.H., Jones, R.M., Pickles, A., State, M.W., Taylor, J.L., Veenstra-VanderWeele, J.: Autism spectrum disorder. Nature Reviews Disease Primers **6**(1), 1–23 (2020)

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

[27] Parikshak, N.N., Swarup, V., Belgard, T.G., Irimia, M., Ramaswami, G., Gandal, M.J., Hartl, C., Leppa, V., Ubieta, L.d.l.T., Huang, J., Lowe, J.K., Blencowe, B.J., Horvath, S., Geschwind, D.H.: Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. Nature **540**(7633), 423–427 (2016)

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[28] Gandal, M.J., Zhang, P., Hadjimichael, E., Walker, R.L., Chen, C., Liu, S., Won, H., van Bakel, H., Varghese, M., Wang, Y., Shieh, A.W., Haney, J., Parhami, S., Belmont, J., Kim, M., Moran Losada, P., Khan, Z., Mleczko, J., Xia, Y., Dai, R., Wang, D., Yang, Y.T., Xu, M., Fish, K., Hof, P.R., Warrell, J., Fitzgerald, D., White, K., Jaffe, A.E., PsychENCODE Consortium, Peters, M.A., Gerstein, M., Liu, C., Iakoucheva, L.M., Pinto, D., Geschwind, D.H.: Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science **362**(6420), 8127 (2018)

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

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

[31] McCartney, A.J., Zolov, S.N., Kauffman, E.J., Zhang, Y., Strunk, B.S., Weisman, L.S., Sutton, M.A.: Activity-dependent pi (3, 5) p2 synthesis controls ampa receptor trafficking during synaptic depression. Proceedings of the National Academy of Sciences **111**(45), 4896–4905 (2014)

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

[33] Ishimura, R., Nagy, G., Dotu, I., Chuang, J.H., Ackerman, S.L.: Activation of gcn2 kinase by ribosome stalling links translation elongation with translation initiation. Elife **5**, 14295 (2016)

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

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

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

[37] Ajram, L., Horder, J., Mendez, M., Galanopoulos, A., Brennan, L., Wichers, R., Robertson, D., Murphy, C., Zinkstok, J., Ivin, G., *et al.*: Shifting brain inhibitory balance and connectivity of the prefrontal cortex of adults with autism spectrum disorder. Translational Psychiatry **7**(5), 1137–1137 (2017)

[38] Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., *et al.*: Science forum: the human cell atlas. Elife **6**, 27041 (2017)

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.13.201061; this version posted July 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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