bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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.

The critical role of composition-independent temporal and autism pathologic changes in the human brain transcriptome

Qianhui Yu[1,2], Zhisong He[1*]

[1] CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology (PICB), Shanghai Institutes for Biological Sciences (SIBS), Chinese Academy of Sciences (CAS), Shanghai, China

[2] University of Chinese Academy of Sciences, Beijing, China

Correspondence:

Zhisong He

zhisong.he@gmail.com

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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.

Abstract

The functions of human brains highly depend on the precise temporal regulation of gene expression, and substantial transcriptome changes across lifespan have been observed. The substantial transcriptome alteration in neural disorders like autism has also been observed and is thought to be important for the pathology. While the cell type composition is known to be variable in brains, it remains unclear how it contributes to the temporal and pathological transcriptome changes in brains. Here, we applied the transcriptome deconvolution procedure to the age series RNA-seq data of healthy and autism samples, to quantify the contribution of cell type composition in shaping the temporal and autism pathological transcriptome in human brains. We estimated that composition change was the primary factor of both transcriptome changes. On the other hand, genes with substantial composition-independent expression changes were also observed in both cases. Those temporal and autism pathological composition-independent changes, many of which are related to synaptic functions, indicate the important intracellular regulatory changes in human brains in both processes.

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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.

Introduction

The development and aging of human brains are complex processes, which are shaped by anatomical and molecular changes [1, 2, 3, 4]. Autism spectrum disorder (ASD), as a common neurodevelopmental disorder, disrupts the critical developmental processes and results in the disruption of cognitive functions [5, 6, 7]. With the emergence of high-throughput measurement of different molecules, dozens of studies have been conducted to characterize age-related molecular changes in human brains, especially at the transcriptome level [8, 9, 10, 11]. Meanwhile, several studies have also been conducted in order to investigate the transcriptome alteration in ASD [12, 13, 14].

The human brain, however, is a highly complex and heterogeneous organ comprised of numerous different cell types, including neurons and multiple classes of non-neuronal glial cells – such as astrocytes, oligodendrocytes, oligodendrocyte precursor cells and microglia – as well as vascular, such as endothelial, cells. Each of those cell types expresses a distinct set of genes [15] and plays a unique and essential role in the development and functions of the brain [16]. Different cell types are also known to show different spatial-temporal distributions. Such complexity thus raises the unanswered questions: how much of the age-related molecular change in human brains, or specifically the age-related transcriptome change, is the direct consequence of the cell type composition change? If composition changes contribute to age-related expression changes, what's the biological meaning of the rest? Similar questions about the molecular changes in neural disorders, including ASD, also remain opened.

In order to comprehensively answer these questions, an accurate estimation of cell type composition in healthy and autistic human brains is required. Although experiments including stained cell counting [17, 18, 19, 20] and large scale single-cell RNA-seq [21, 22] have the potential to provide these data, these methods are either too labor-intensive or costly at present. Thus, the computational method of inversing sample heterogeneity, *i.e.* deconvolution, is one of the best alternative solutions to estimate the mixing percentage of different cell types [23, 24, 25]. The application of transcriptome deconvolution to human brain was long limited by the lack of comprehensive transcriptome data for the varied cell types in human brains. The recent emergence of the human brain single cell RNA-seq data, covering all the main cell types in the human brain [21], now renders this approach more feasible.

In this study, we used a two-step deconvolution procedure to estimate the cell type composition changes in the human brain after birth, as well as to quantitatively decompose the human brain transcriptome profiles across lifespan into composition-dependent and composition-independent components. The estimated composition changes matched well with the previous observations. The delineation of

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

the composition-dependent component from gene expression suggested that cell-type composition explained about 25% of the total expression temporal variance, and greatly contributes to the age-related expression pattern in brain. Meanwhile, although to a lesser extent, the composition-independent component also significantly contributes to age-related expression pattern. In addition, we used the similar framework to analyze the contribution of cell-type composition alteration in the brain transcriptome alteration of ASD patients. While composition-dependent component, especially the decreased proportion of RNA contributed by neurons, appeared to be the main power driving the difference, one group of synaptic genes with altered composition-independent component presented strong enrichment of ASD-associated genetic variants, implying the active regulatory disruption of synaptic functions in ASD.

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

Results

Cell type composition changes in brains across the human lifespan

To investigate the temporal cell type composition changes across lifespan in human brains, we applied the deconvolution procedure to the age series RNA-seq data set of the human prefrontal cortex (PFC) consisting of 40 postnatal human brain samples aged from two-day old to 61.5 year-old [11]. The published human brain single cell RNA-seq data [21] was used to obtained the gene expression information of the eight main cell types in human brains, quantifying the expression level of 14,054 protein-coding genes. CIBERSORT [24] was used to select genes distinguishing cell types, resulted in 904 genes (referred as CIBERSORT-selected markers, Supplementary Table S1). Two well-developed transcriptome deconvolution algorithms, CIBERSORT [24] and quadratic programming (QP), were applied based on the CIBERSORT-selected markers.

In addition, diffusion ratio (DR) based deconvolution was also used based on the 1491 cell type signature genes. This deconvolution algorithm was based on the simple assumption that the expression of a cell type signature gene in the bulk tissue can be seen as its expression in the cell type scaled by the cell type's mixing proportion. It is similar to the recent deconvolution algorithm MCP-counter [27] but simplified (see Methods). The cell type signature genes were identified by requiring at least ten-fold higher expression level in one cell type comparing to any of the remaining, including 319 signature genes for astrocytes, 288 for endothelial cells, 224 for microglia, 99 for oligodendrocytes, 71 for oligodendrocyte progenitor cells (OPC), 166 for adult neurons, 92 for fetal quiescent neurons, and 232 for fetal replicating neurons (Supplementary Table S1).

The three deconvolution algorithms resulted in similar composition patterns across the postnatal lifespan (Figure 1 and Supplementary Figure S1). The estimated composition changes were consistent with previous studies, e.g. the elimination of fetal neurons soon after birth accompanying with the increase of adult neurons [28], and the increase of oligodendrocytes with the decrease of OPC which may due to the myelination process [29]. The remaining cell types including astrocyte, endothelial cells and microglia did not show significant changes across the lifespan.

Nevertheless, the three deconvolution algorithms provided slightly different results. Therefore, two complementary measurements, namely the similarity measurement and discrimination measurement (see Methods), were used to compare the three estimated composition patterns. According to Pearson's correlation coefficient (PCC)

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

and root-mean-square deviation (RMSD), both of which were similarity measurements, CIBERSORT out-performed among the three algorithms, while DR performed slightly better than QP (Supplementary Figure S2). On the other hand, the discrimination measurement (see Methods), suggested that the cell type gene expression identities were reconstructed in the most discriminating way by DR, while CIBERSORT performed better than QP (Supplementary Figure S2).

In addition to the postnatal brain transcriptome, we further applied the three algorithms to the human embryonic developmental brain RNA-seq data obtained from BrainSpan database (Figure 1 and Supplementary Figure S1). DR-estimated composition pattern successfully recapitulated the domination of fetal replicating neurons dominate before 12 post conception weeks (pcw) and its dramatic elimination after that. This observation, coupling with the increase of fetal quiescent neurons, well matched with the neuronal proliferation that occurred during four pcw to 12 pcw [30]. Moreover, the estimated compositions by DR algorithm, especially those of fetal and adult neurons, presented a successive pattern with the postnatal composition changes estimated above, which further indicated the reliability and robustness of the composition estimation. On the other hand, CIBERSORT, although out-performed both DR and QP deconvolution according to the similarity measurement, failed to reproduce those scenarios.

Composition change is the primary power shaping the human brain postnatal temporal transcriptome

To comprehensively study the contribution of cell type composition changes in the human brain temporal transcriptome, we dissected the bulk gene expression into the composition-dependent and composition-independent components, based on the composition pattern estimated by DR algorithm (see Methods). Their contributions to the overall gene expression variance were estimated. On average, the composition dependent component, which represents the upper limit of transcriptome variance caused by variation of cell type composition in human brains, explained 26.4% of the total variance in the human postnatal age series data (Figure 2a). Focusing on the 5,119 genes with age-related expression (referred as age-related expressed genes, age test BH-corrected $FDR<0.05$), the composition-dependent component contributed 34.7% of the total variance which was significantly higher than the other genes (Figure 2b, permutation test, $P<0.001$).

To further investigate the roles of the composition-dependent and independent component in shaping the temporal gene expression pattern, we calculated the variances explained by age (age-explained variance) for each of the two components separately. Interestingly, the relative contribution of age-explained variance from the

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

composition component (55.8%) was much larger than the proportion of composition variance among total variance (34.7%) for age-related expressed genes, while the difference was much smaller for the non-age-related expressed genes (27.8% vs. 22.1%) (Figure 2b). Altogether, these observations implied that the change of the composition-dependent component, *i.e.* the cell-type composition changes, was the main power shaping the observed temporal expression in human brains.

Additionally, the proportion of age-explained variance in the composition-dependent component was dramatically higher for the age-related expressed genes than for other genes (median=66.4%, permutation test, $P<0.001$, Figure 2c); meanwhile for the same genes, a moderate but significant increase of age-related variance proportion was observed for the composition-independent component (median=28.9%, permutation test, $P<0.001$, Figure 2c). These observations not only point to the significance of composition-dependent changes, but also to that of the composition-independent changes, some of which may represent the changes of molecular features in one or several cell types, also participate in shaping the temporal transcriptome in human brains. It is worth to mention, that all the above observations were reproducible based on CIBERSORT-based composition pattern (Supplementary Figure S3), suggesting that it is not artifact due to the DR-based deconvolution procedure.

The age-related changes in the composition-independent components are important and well regulated

To better understand the biological significance of the age-related changes in the composition-dependent and composition-independent component explicitly, we applied age tests to each of the two components (see Methods), to identify genes with significant age-related changes in either component. 8,156 and 1,455 genes were found with age-related changes in the composition-dependent and independent component, respectively (Figure 3a, Supplementary Table S2). Both of the two gene sets were largely overlapped with the age-related expressed genes (Fisher's exact test, $P<0.0001$). However, no significant overlap was observed between them (Fisher's exact test, $P=0.216$, odds ratio=0.932), implying the independent contribution of the two components to the temporal transcriptome in human brains. CIBERSORT-based composition pattern provided similar estimation, resulting in 6,671 and 1,129 genes with age-related changes in the composition-dependent and independent component, respectively. Majority of those genes were overlapping with genes identified based on DR-based estimation (Supplementary Figure S4).

We further grouped the 8,719 genes with age-related changes in at least one of the two components into three categories: G1 – genes with age-related changes in both components; G2 – genes with age-related changes only in the composition-dependent

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

components; and G3 – genes with age-related changes only in the composition-independent components. Interestingly, these three groups of genes showed distinct temporal expression patterns (Figure 3b). G1 and G2 genes, and especially the latter gene, showed higher expression levels in early postnatal development stages, while G3 genes were highly expressed in adult stages. Different groups of genes were also enriched in different cell types: G1 – adult neuron, G2 – fetal quiescent neuron, fetal replicating neuron and adult neuron, and G3 – astrocyte and endothelial cells (Figure 3b).

More importantly, the three groups of genes showed distinctive functional enrichments (DAVID [31], Supplementary Table S3). In brief, G1 genes were enriched for synapse and translation-related functions. G2 genes were enriched in transcription regulation, protein degradation, and cell cycle. Lastly, G3 genes were significantly involved in extracellular regions and metabolism. These results indicated the distinct biological significance of the age-related composition dependent and independent changes.

The expression of these three groups of genes should have been modulated by certain regulatory mechanisms, such as transcription factors (TFs). To test this, we estimated the enrichment of TF binding motifs in the promoter regions of genes in each category. We observed significant excess of enriched TF binding motifs (hypergeometric test, $P<0.1$) in the G2 and G3 genes (permutation test, $P<0.001$) (Figure 3c). In addition, the expression of representative TFs of the enriched TF binding motifs in groups showed significantly better correlation (Wilcoxon test, $P<0.05$) with their targets in the respective groups than expected by chance (permutation test, $P<0.001$) (Figure 3c). TFs with TF binding motifs enriched in G2 genes, e.g. CUX1 and E2F1, were mostly negatively correlated with their targets and had been shown to be relevant to cell migration, cell cycles and neuronal development and maturation [32, 33, 34]. On the other hand, most of the TFs with binding motifs enriched in G3 genes, e.g. SMAD3, SREBF1 and NR2F2, were positively correlated with their G3 target genes, many of which had been reported participating in signal transduction and metabolism of astrocytes [35, 36, 37].

Significance of composition-independent component in ASD pathogenesis

Autism spectrum disorder (ASD) is a common neurodevelopmental disorder. Previous transcriptomic studies have suggested dramatic changes of gene expression [12, 13, 14]. For instance, using the same human PFC postnatal age series RNA-seq data set as described above, coupled with RNA-seq data of 34 PFC samples of autistic patients, Liu et al. identified 1,775 genes with differential expression (DE) between ASD cases and unaffected controls [14]. It is worthwhile to estimate the contribution of cell type

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

composition as well as to identify the composition-independent alteration which may represent the active regulatory alternation in ASD patients.

We applied CIBERSORT, QP and DR algorithms to estimate composition patterns in the autistic samples. The three algorithms provided similar results, and interestingly, all showed significant smaller proportion of neurons as well as larger proportions of non-neuron cells (Figure 4a and Supplementary Figure S5). The amplitudes of composition changes were significantly correlated among the estimations based on the three algorithms (Supplementary Figure S5), eliminating the possibility of algorithm artifact. The increase amplitudes of different non-neuron cell types were proportional to their estimated proportion in healthy samples (Figure 4b), suggesting that they were likely to be the passive consequence of neuron proportion decrease.

To further investigate the contribution of composition differences in the autistic brain transcriptome alteration, we dissected the bulk gene expressions of all the PFC samples, including the 34 ASD cases and 30 healthy controls within the same sample age range, into the composition-dependent and composition-independent components, based on the DR-based composition pattern. Natural spline based ANCOVA (Methods) was applied to each of the two components to identify genes with DE between ASD cases and healthy controls. Among the 14,032 detected genes, 74% of them (10,358) showed DE in their composition-dependent components (Benjamini-Hochberg corrected FDR<20%), while only no more than 5% of the genes (644) showed DE in composition-independent components under the same criteria (Supplementary Table S4). The two numbers based on CIBERSORT-based composition pattern were 8,030 and 275, with majority overlapping with the DR-based ones (Supplementary Figure S4). Both the two sets of genes were significantly overlapping with the 1,775 DE genes identified based on bulk expression (referred as bulk DE genes, Figure 5a, Fisher's exact test, $P<0.0001$), suggesting that both of them were important contributors in transcriptome alteration in ASD brains.

Although majority of the transcriptome alterations in ASD brains were related to the cellular composition changes, the alterations in the composition-independent components of brain transcriptome were of particular interest. They might represent the active regulatory disruptions in particular brain cells. The functional enrichments (DAVID [31]) suggested that they were enriched for membrane and synaptic proteins (Figure 4b, Supplementary Table S5). Consistently, those genes showed significant expression enrichment in adult and fetal quiescent neurons (Figure 4b). More interestingly, the genes with significant alterations in their composition-independent components, but not those with significant alterations only in their composition-dependent components, contained a significant excess of genes with

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

genetic variants associated with ASD, based on analysis of genes extracted from SFARI AutDB [38] (Fisher's exact test, odds ratio=1.52, P=0.009, Figure 4b), as well as genes from AutismKB database [39] (Fisher's exact test, odds ratio=1.37, P=0.008, Figure 4b). Furthermore, those genes were also consistently identified as differentially expressed genes in autistic brains in previous studies [12, 13] (Fisher's exact test, P<0.005, Figure 4b). It is also worth to note, that 40 microRNA (miRNA) families showed enrichment of predicted binding at the 3'UTRs of those genes (TargetScan v7.1 [40]; hypergeometric test, P<0.05), which was significantly more than by chance (permutation test, P<0.01, FDR=0.65). Some of the enriched miRNAs, including has-miR-363-3p which was also reported to be dysregulated in Alzheimer's disease [41], have been reported as differentially expressed in autistic brains [42]. All the above observations suggested that the disruptions of neuronal transcriptional regulations in autistic brains, which were likely to be partially caused by the dysregulations of certain neuronal miRNAs, may result in pathological changes in synapses and largely contribute to ASD pathology.

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

Discussions

In this study, we attempted to investigate the role of changes related or independent to cell type composition in human brain development and pathogenesis of ASD using computational approaches. With the gene expression of eight main cell types in human brain estimated using the human brain single cell RNA-seq data [21], we applied a two-step transcriptome deconvolution procedure to the human brain postnatal age series RNA-seq data [11] and the ASD brain RNA-seq data [14]. Based on the estimated composition patterns, gene expression was dissected into the composition-dependent and independent component, allowing the further investigation on each of them.

The transcriptome deconvolution in healthy samples resulted in temporal composition patterns consistent with the prior studies. On the other hand, we also notice that our estimated proportion of total neurons in the adult samples reached around 60-70%, which is much higher than the estimation based on cell counting [43, 44] or DNA methylation deconvolution [45]. This may be explained by the fact, that the estimated composition of transcriptome deconvolution actually reflects the proportion of RNAs contributed by each cell type instead of the proportion of cell numbers. In fact, the RNA content in neurons is reported to be close-to-two-fold as much as in glia cells [46].

Based on the variance analysis, we regard the composition change to be the main source of the age-related expression change in human brain. This is a consistent conclusion drawn by previous studies based on DNA methylation [45]. The proportion of variance explained by cell types, on the other hand, is smaller: only about 30% for the age-related expressed genes, while it is around 50% for the DNA methylation. Although the influence of sampling and measurement noise could not be ruled out, such observation may imply the additional regulatory signal that show age-related manner and is independent from the DNA methylation, which might be an interesting focus for further study.

Interestingly, genes with age-related changes in the composition-dependent components show high expression level in the infant stages, *i.e.* < 2 years old, with significantly enriched expression in the three types of neurons. This suggests that although some other cell types such as oligodendrocytes and OPCs presented mixing proportion changes in the age-related manner, most of the composition-dependent changes are due to the transition from fetal neurons to adult neurons, *i.e.* the neuron maturation process. It therefore implies that the neuron maturation is the primary factor creating the age-related expression, especially during the infant stage.

On the other hand, the age-related changes to the composition-independent component is appealing. Unlike the genes with age-related composition-dependent

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

changes which mostly express highly in the early postnatal development, genes with age-related composition-independent changes tend to have higher expression in adults. What's more, these composition-independent changes are significantly related to either synapse in neurons (G1), or extra-cellular regions and signal peptides in astrocytes and endothelial cells (G3), both of which are relevant to cell-to-cell communications. This is unlikely to be a coincidence. As an information processing system, such communications are critically important for human brains [47, 48], and our results suggest that the complexity growth of the communication system not only depends on the increased number of computational units, *i.e.* neurons, but also greatly relies on the enhanced inter-cellular communications which are independent from the cell type composition changes across lifespan. Such communications may not limit to the synaptic connections between neurons, but also include the neuron-glia and glia-glia communications which are also critical to the neuronal network functions [49].

In autistic brains, the contradictory statements of unchanged density of both neuronal and glial cell pools [50] and even more neuron numbers [51] has been reported. Therefore, it is a surprising observation that the estimated neuron proportions in autistic brains are significantly smaller than that in healthy brains. While the influence of unknown technical issue cannot be ruled out, this result implies the possibility that the general transcriptional activity of neurons is inhibited in ASD. In such case, the total mRNA content in neurons reduces, therefore resulting in smaller contribution of neuronal mRNAs to the bulk mRNA pool. This assumption is supported by the depletion of neuronal marker MAP2 in autistic brains [50]. However, further experimental measurement is also awaited for validation.

Considering that neuroinflammation is thought to play important roles in ASD pathology [52, 53], it is also unexpected to find no evidence suggesting general activation of microglia, the macrophage cells participating neuroinflammation in the central nervous system. Although the increase of microglia proportion is observed, no significant excess of increase is observed when comparing to other glial cell types after normalizing to their relative proportion in healthy brains. Nevertheless, our results did not rule out the importance of neuroinflammation in ASD, considering that the strongest activation of microglia in ASD patients happens in cerebellum and white matters [53, 54]. Although higher microglia density was also reported in ASD dorsolateral PFC [55], they may not be fully activated, making our analysis lack of power to capture such signal. More investigation is required to comprehensively study the role of PFC microglia in ASD brains.

While our results suggest that majority of the transcriptome alteration detected in PFC of ASD patients are passive alteration due to the changes of relative mRNA

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

contribution from cell types, a group of genes attracts most of the focus. Those 644 genes, enriching for membrane protein genes and synaptic functions in neurons, represent the transcriptome alterations which are at least partially independent to the composition changes. This is an interesting observation, suggesting the pathological changes happened to the regulatory network in neurons regulating synaptic functions in ASD. This linkage is further supported by their enriched genetic association with ASD. It is also concordant with previous studies, which report the abnormalities of synaptic functions as well as neuronal dendrites and spines in autistic humans and autistic mouse models [56, 57]. More comprehensive investigations are awaited for the roles of synaptic disruptions in ASD, which may lead to effective pharmacological treatment in the future.

Although our framework paves a way for more exhaustive analyses regarding the contribution of cell type composition to the transcriptome changes, we are well aware of the limitations of our method. For instance, our analysis failed to squarely pinpoint the one or several cell types with the greatest influence on either of the two components, though cell-type enrichment analysis was capable to give a glance. Our analysis greatly relied on the accurate transcriptome measurement of all or at least close to all of the cell types in the bulk tissue, which limits its applications. The experimental validation is relatively difficult. Nevertheless, the fast-developing single-cell technology, especially the single-cell RNA-seq, should be able to comprehensively, simultaneously and separately quantify changes happened in cell type composition, cell type RNA content and cell type molecular profiles. Indeed, the single cell RNA-seq technology has been widely used to describe cell type compositions in different organs including brains [21, 22, 58], as well as to characterize the molecular processes driving neurogenesis and somatic reprogramming to neurons [59, 60]. The further application of this technology to characterize human brain development, ageing, and pathology of neural disorders including ASD would be sufficient to verify our observation and of great value to comprehensively understand those important processes.

If nothing else, we hope that our attempt to dissect expression into composition-dependent and composition-independent components will inspire further studies to elucidate transcriptome changes in a more comprehensive manner.

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

Methods

Data

The human brain single cell RNA-seq data was retrieved from SRA (SRP057196). The human postnatal age-series brain RNA-seq data with 40 samples was retrieved from GEO (GSE51264). All the RNA-seq reads were mapped to the human genome hg38 with STAR 2.3.0e using the default parameters. The number of reads covering the exonic regions of each protein-coding gene annotated in GENCODE v21 was counted and normalized using the R package DESeq2 for each data set separately. The autistic brain RNA-seq data retrieved from GEO (GSE59288), were processed in the same way. The pre-calculated RPKM of the fetal human brain samples were downloaded from BrainSpan (http://www.brainspan.org/static/download.html).

Deconvolution for cell-type composition

Three different strategies were used for deconvolution for cell-type composition. The first method, quadratic programming (QP) based deconvolution, was to model the gene expression of each cell type signature genes in the bulk tissue sample as a linear combination of its expression in each cell type according to the cell type mixing proportion. Thus, the deconvolution problem for each bulk tissue sample was represented as a constrained linear least-square problem, which was:

$$\min(\|\boldsymbol{Cf} - \boldsymbol{x}\|^2), s.t. \begin{cases} \sum_i f_i = 1 \\ f_i \geq 0, \forall i \end{cases}.$$

Here, $\boldsymbol{f}$ was the vector of cell type mixing proportion, and $\boldsymbol{C}$ was the matrix of gene expression of the cell type signature genes in each cell type, while $\boldsymbol{x}$ was the known expression level of the cell type signature genes. This model was widely used in the deconvolution problem [23, 25], and can be solved using quadratic programming [26].

The second method, CIBERSORT [24], which was developed by Newman et al. in 2015, was also used. Instead of the solving the least square problem, CIBERSORT estimates the relative cell type composition based on nu-support vector regression, minimizing a linear ε-insensitive loss function with L2 regularization.

The third method, namely diffusion ratio (DR) based deconvolution, was based on the simple assumption that the expression of a cell type signature gene in the bulk tissue can be seen as its expression in the cell type scaled by the cell type's mixing proportion. It is similar as the recent algorithm MCP-counter [27] but was implemented in a simpler way. Under this assumption, the proportion of cell type i (represented as f_i)

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

was simply estimated as:

$$f_i = median_g(\frac{x_g}{c_{ig}}).$$

Here, g represented each cell type signature gene of cell type i. After calculating f_i for all the cell types, $\boldsymbol{f}$ was normalized so that $\sum_i f_i = 1$.

Deconvolution for cell-type expression profile re-estimation

The similar model as the QP-based deconvolution for cell-type composition described above was used for the second deconvolution to re-estimate cell-type expression profile, which was used for estimation of deconvolution performance and variance analysis. Similarly, the deconvolution problem for each gene can be seen as a constrained linear least-square problem, that is:

$$\min(\|\boldsymbol{Fc} - \boldsymbol{x}\|^2), s.t.\ \ c_i \geq 0, \forall i.$$

Here, $\boldsymbol{c}$ was the vector of re-estimated gene expression level in different cell types, and $\boldsymbol{x}$ was the vector of gene expression across the bulk tissue samples. $\boldsymbol{F}$ was the composition matrix with each row representing the estimated mixing proportion in the corresponding bulk tissue sample. This problem was also solved by using quadratic programming as described above.

Performance measurement of transcriptome deconvolution

Two different types of deconvolution performance measurements were used to compare different deconvolution methods.

The first type of performance measurement, namely similarity measurement, aimed to measure the general similarity between the real bulk tissue gene expression and the predicted gene expression based on the cell type gene expression level and the cell type mixing proportion. It is the commonly used measurement of deconvolution performance. In this case, two measurements were applied: Pearson's correlation coefficient (PCC), and root-mean-square deviation (RMSD) which was calculated as the root mean squares of log10-transformed fold changes between the observed FPKMs and the predicted FPKMs (products of estimated cell-type proportions and re-estimated cell type expressions) across all genes.

The second type of performance estimation, namely discrimination measurement, was established based on the following assumption: the accurate deconvolution can result in re-estimated cell type expression estimations which are not only similar to the measured cell type expressions, but also able to recover the cell type identities discriminating one cell type from the others. Using PCC as the proxy of similarity

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

between two cell type expression profiles, the discriminating recovery score (DRS) of cell type *i* was defined as:

$$DRS_i = r(\hat{\boldsymbol{e}}_i, \boldsymbol{e}_i) - \max_j r(\hat{\boldsymbol{e}}_i, \boldsymbol{e}_j)$$

Here, $\boldsymbol{e}_i$ was the vector of observed log10-transformed FPKM of cell type *i*, and $\hat{\boldsymbol{e}}_i$ was the vector of re-estimated log10-transformed FPKM of cell type *i*. The mean DRS across different cell types was then used as the proxy of the overall DRS.

Dissection of composition-dependent and composition-independent expression components

With **X** denoted as the $N \times M$ gene expression matrix with N genes and M samples, and the $d \times M$ estimated composition matrix **F** by any of the mentioned transcriptome deconvolution method with d cell types, the re-estimated cell type expression matrix **C** with N rows and d columns was obtained as described above. The composition-dependent component was then represented as the matrix $\mathbf{X_d} = \mathbf{CM}$, while the composition-independent component was represented as $\mathbf{X_i} = \mathbf{X} - \mathbf{X_d}$.

ANCOVA based on natural spline with variable degree of freedom

To identify age-related changes, for each gene, with its observed expression level, composition-dependent component, or composition-independent component as the response variable, the age test, an ANCOVA employing the F-test was used to compare the null model: a linear model only with intercept, to a series of alternative models: the natural spline with degree of freedom from two to eight in response to square root transformed sample ages (sqrt-age). The best alternative model was chosen by applying the adjusted r^2 criterion [8]. Genes with BH corrected $P<0.05$ were considered as genes with its expression, composition-dependent or composition-independent component changed in the age-related manner. For calculating the proportion of variance explained by age, the natural spline model with degree of freedom equaling to eight in response to sqrt-age was used.

The similar ANCOVA framework was also applied to identify changes between autistic samples and healthy samples, by considering the sample ages. The null model was defined as the unified natural spline model with no discrimination of autistic and healthy samples, while the alternative model with varied coefficient in the two groups was compared by employing the F test. Genes with BH corrected $P < 0.2$ were considered as genes with composition-dependent or composition-independent component changed between the two groups of samples. Similarly, sqrt-age was used as the independent variable.

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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 expression enrichment analysis

The cell type enrichment analysis was performed based on the average RPKM of each gene in the eight major cell types in human brains including fetal replicating neurons, fetal quiescent neurons, adult neurons, astrocytes, oligodendrocytes, endothelial cells, microglia and oligodendrocyte precursor cells (OPCs), based on the human brain single cell RNA-seq data [21]. The expression specificity of one gene in one cell type was represented by the log10-transformed fold change between its expression level in the cell type and its average expression level in the other cell types. For each cell type, a one-sided Wilcoxon's rank test was applied to compare the expression specificities of genes in the gene list, to the other genes that were detected in both data.

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

Acknowledgement

We thank Prof. Dr. P. Khaitovich, Dr. Y. Wei, Dr. S. Feng, Dr. H. Hu, Ms. H. Bammann, Ms. Q. Li, and Mr. C. Xu for the valuable discussions. We thank Dr. G. L. Banes for his helpful comments on the manuscript. This study was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant XDB13010200); and the National Natural Science Foundation of China (grants 91331203, 31171232, 31501047 and 31420103920).

Author Contributions Statement

Q.Y. and Z.H. conceived and designed the study, analyzed the data, and wrote the manuscript. Both the authors have read and approved the final manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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. Huffman K. The developing, aging neocortex: how genetics and epigenetics influence early developmental patterning and age-related change. *Front Genet* **3**, 212 (2012).

2. Alexander-Bloch A, Raznahan A, Bullmore E, Giedd J. The convergence of maturational change and structural covariance in human cortical networks. *J Neurosci* **33**, 2889-2899 (2013).

3. Nie J, Li G, Shen D. Development of cortical anatomical properties from early childhood to early adulthood. *Neuroimage* **76**, 216-224 (2013).

4. Giedd JN, Raznahan A, Alexander-Bloch A, Schmitt E, Gogtay N, Rapoport JL. Child psychiatry branch of the National Institute of Mental Health longitudinal structural magnetic resonance imaging study of human brain development. *Neuropsychopharmacology* **40**, 43-49 (2015).

5. Lai MC, Lombardo MV, Baron-Cohen S. Autism. *Lancet* **383**, 896-910 (2014).

6. Rogers SJ, DiLalla DL. Age of symptom onset in young children with pervasive developmental disorders. *J Am Acad Child Adolesc Psychiatry* **29**, 863-872 (1990).

7. Kanner L. Autistic disturbances of affective contact. *Acta Paedopsychiatr* **35**, 100-136 (1968).

8. Somel M, *et al.* Transcriptional neoteny in the human brain. *Proc Natl Acad Sci U S A* **106**, 5743-5748 (2009).

9. Kang HJ, *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483-489 (2011).

10. Liu X, *et al.* Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques. *Genome Res* **22**, 611-622 (2012).

11. He Z, Bammann H, Han D, Xie G, Khaitovich P. Conserved expression of lincRNA during human and macaque prefrontal cortex development and maturation. *RNA* **20**, 1103-1111 (2014).

12. Gupta S, *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat Commun* **5**, 5748 (2014).

13. Voineagu I, *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380-384 (2011).

14. Liu X, *et al.* Disruption of an Evolutionarily Novel Synaptic Expression Pattern in Autism. *PLoS Biol* **14**, e1002558 (2016).

15. Zhang Y, *et al.* An RNA-sequencing transcriptome and splicing database of glia, neurons, and

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

vascular cells of the cerebral cortex. *J Neurosci* **34**, 11929-11947 (2014).

16. Allen NJ, Barres BA. Neuroscience: Glia - more than just brain glue. *Nature* **457**, 675-677 (2009).

17. Mittelbronn M, Dietz K, Schluesener HJ, Meyermann R. Local distribution of microglia in the normal adult human central nervous system differs by up to one order of magnitude. *Acta Neuropathol* **101**, 249-255 (2001).

18. Herculano-Houzel S, Lent R. Isotropic fractionator: a simple, rapid method for the quantification of total cell and neuron numbers in the brain. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **25**, 2518-2521 (2005).

19. Sherwood CC, *et al.* Evolution of increased glia-neuron ratios in the human frontal cortex. *Proc Natl Acad Sci U S A* **103**, 13606-13611 (2006).

20. Bandeira F, Lent R, Herculano-Houzel S. Changing numbers of neuronal and non-neuronal cells underlie postnatal brain growth in the rat. *Proc Natl Acad Sci U S A* **106**, 14108-14113 (2009).

21. Darmanis S, *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci U S A* **112**, 7285-7290 (2015).

22. Zeisel A, *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138-1142 (2015).

23. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* **29**, 1083-1085 (2013).

24. Newman AM, *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453-457 (2015).

25. Gong T, *et al.* Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS One* **6**, e27156 (2011).

26. Lawson CL, Hanson RJ. *Solving Least Squares Problems*. Society for Industrial and Applied Mathematics (1987).

27. Becht E, *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol* **17**, 218 (2016).

28. Sanai N, *et al.* Corridors of migrating neurons in the human brain and their decline during infancy. *Nature* **478**, 382-386 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

29. Paus T, *et al.* Structural maturation of neural pathways in children and adolescents: in vivo study. *Science* **283**, 1908-1911 (1999).

30. Tau GZ, Peterson BS. Normal development of brain circuits. *Neuropsychopharmacology* **35**, 147-168 (2010).

31. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).

32. Cubelos B, *et al.* Cux1 and Cux2 regulate dendritic branching, spine morphology, and synapses of the upper layer neurons of the cortex. *Neuron* **66**, 523-535 (2010).

33. Cooper-Kuhn CM, *et al.* Impaired adult neurogenesis in mice lacking the transcription factor E2F1. *Mol Cell Neurosci* **21**, 312-323 (2002).

34. Wang L, Wang R, Herrup K. E2F1 works as a cell cycle suppressor in mature neurons. *J Neurosci* **27**, 12555-12564 (2007).

35. Tichauer JE, Flores B, Soler B, Eugenin-von Bernhardi L, Ramirez G, von Bernhardi R. Age-dependent changes on TGFbeta1 Smad3 pathway modify the pattern of microglial cell activation. *Brain Behav Immun* **37**, 187-196 (2014).

36. Medina JM, Tabernero A. Astrocyte-synthesized oleic acid behaves as a neurotrophic factor for neurons. *J Physiol Paris* **96**, 265-271 (2002).

37. Li Y, *et al.* Sonic hedgehog (Shh) regulates the expression of angiogenic growth factors in oxygen-glucose-deprived astrocytes by mediating the nuclear receptor NR2F2. *Mol Neurobiol* **47**, 967-975 (2013).

38. Basu SN, Kollu R, Banerjee-Basu S. AutDB: a gene reference resource for autism research. *Nucleic Acids Res* **37**, D832-836 (2009).

39. Xu LM, Li JR, Huang Y, Zhao M, Tang X, Wei L. AutismKB: an evidence-based knowledgebase of autism genetics. *Nucleic Acids Res* **40**, D1016-1022 (2012).

40. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**, (2015).

41. Lau P, *et al.* Alteration of the microRNA network during the progression of Alzheimer's disease. *EMBO Mol Med* **5**, 1613-1634 (2013).

42. Wu YE, Parikshak NN, Belgard TG, Geschwind DH. Genome-wide, integrative analysis implicates microRNA dysregulation in autism spectrum disorder. *Nat Neurosci* **19**, 1463-1476

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

(2016).

43. Pelvig DP, Pakkenberg H, Stark AK, Pakkenberg B. Neocortical glial cell numbers in human brains. *Neurobiol Aging* **29**, 1754-1762 (2008).

44. Herculano-Houzel S. The human brain in numbers: a linearly scaled-up primate brain. *Front Hum Neurosci* **3**, 31 (2009).

45. Jaffe AE, *et al.* Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat Neurosci* **19**, 40-47 (2016).

46. Filipchenko RE, Pevzner LZ, Slonim AD. RNA content in the neurons and glia of the hypothalamic nuclei after intermittent cooling. *Neurosci Behav Physiol* **7**, 69-71 (1976).

47. Volterra A, Meldolesi J. Astrocytes, from brain glue to communication elements: the revolution continues. *Nat Rev Neurosci* **6**, 626-640 (2005).

48. Bullmore E, Sporns O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci* **10**, 186-198 (2009).

49. Araque A, Navarrete M. Glial cells in neuronal network function. *Philos Trans R Soc Lond B Biol Sci* **365**, 2375-2381 (2010).

50. Mukaetova-Ladinska EB, Arnold H, Jaros E, Perry R, Perry E. Depletion of MAP2 expression and laminar cytoarchitectonic changes in dorsolateral prefrontal cortex in adult autistic individuals. *Neuropathol Appl Neurobiol* **30**, 615-623 (2004).

51. Courchesne E, *et al.* Neuron number and size in prefrontal cortex of children with autism. *JAMA* **306**, 2001-2010 (2011).

52. Young AM, Chakrabarti B, Roberts D, Lai MC, Suckling J, Baron-Cohen S. From molecules to neural morphology: understanding neuroinflammation in autism spectrum condition. *Mol Autism* **7**, 9 (2016).

53. Pardo CA, Vargas DL, Zimmerman AW. Immunity, neuroglia and neuroinflammation in autism. *Int Rev Psychiatry* **17**, 485-495 (2005).

54. Vargas DL, Nascimbene C, Krishnan C, Zimmerman AW, Pardo CA. Neuroglial activation and neuroinflammation in the brain of patients with autism. *Ann Neurol* **57**, 67-81 (2005).

55. Morgan JT, *et al.* Microglial activation and increased microglial density observed in the dorsolateral prefrontal cortex in autism. *Biol Psychiatry* **68**, 368-376 (2010).

56. Lepeta K, *et al.* Synaptopathies: synaptic dysfunction in neurological disorders - A review from

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

students to students. *J Neurochem* **138**, 785-805 (2016).

57. Martinez-Cerdeno V. Dendrite and spine modifications in autism and related neurodevelopmental disorders in patients and animal models. *Dev Neurobiol,* (2016).

58. Lake BB*, et al.* Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586-1590 (2016).

59. Treutlein B*, et al.* Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature* **534**, 391-395 (2016).

60. Shin J*, et al.* Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. *Cell Stem Cell* **17**, 360-372 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

Figures

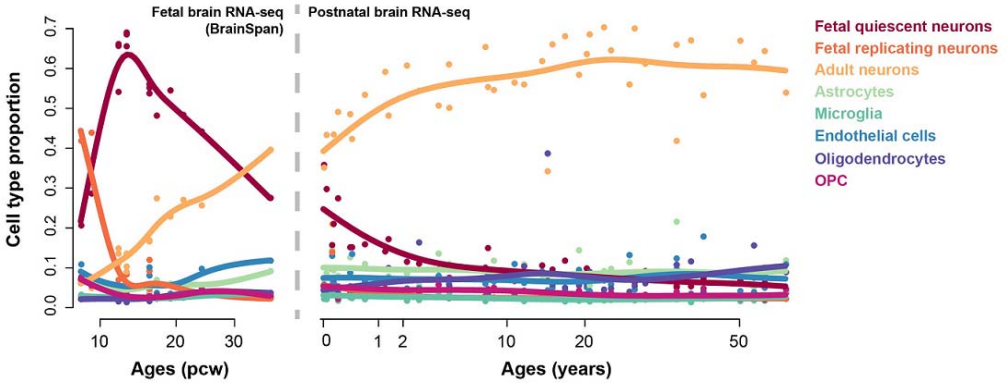


Figure 1. The cell type compositions across lifespan in human brains, estimated using the DR-based deconvolution. The dots show the estimated proportions of samples, and the curves show the spline interpolation results. The left panel shows the cell type composition in human fetal brains, based on the human embryonic developmental brain RNA-seq data from BrainSpan; the right panel shows the cell type composition in human postnatal brains, based on the human postnatal brain data set.

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

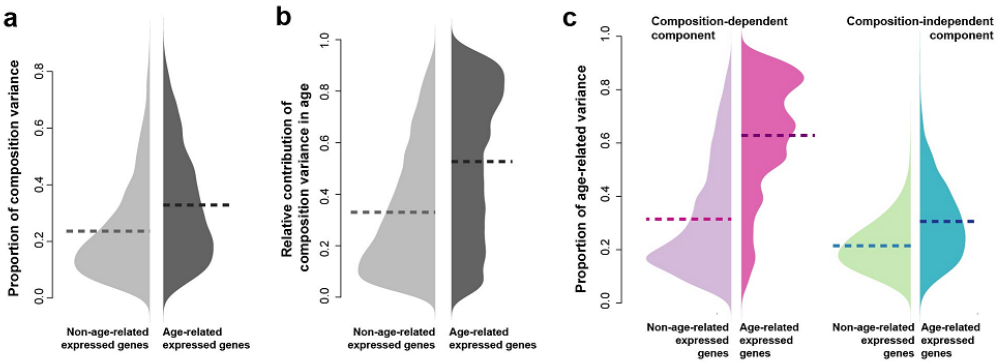


Figure 2. The cell type composition change was the primary driving force of the age-related expression. (a) The proportion of gene expression variance explained by cell type composition (composition-related variance) in the human postnatal brain data set. Light grey – genes without age-related expression; dark grey – genes with age-related expression. The horizontal dash lines show the mean proportion of composition-related variance. (b) The relative contribution of the composition-dependent component to the variance explained by age (age-explained variance), measured as ratio of age-explained variance in the composition-dependent component to the sum of age-explained variance in both components. Light grey – genes without age-related expression; dark grey – genes with age-related expression. The horizontal dash lines show the mean proportion of composition-related variance. (c) The proportion of variance explained by ages (age-related variance) in each of the two components of expression: pink – the composition-dependent component; green – the composition-independent component. The light colors represent the proportions of age-related variance in genes without age-related expression changes; the dark colors represent the proportions of age-related variance in genes with age-related expression changes. The horizontal dash lines show the means of age-related variance proportions.

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

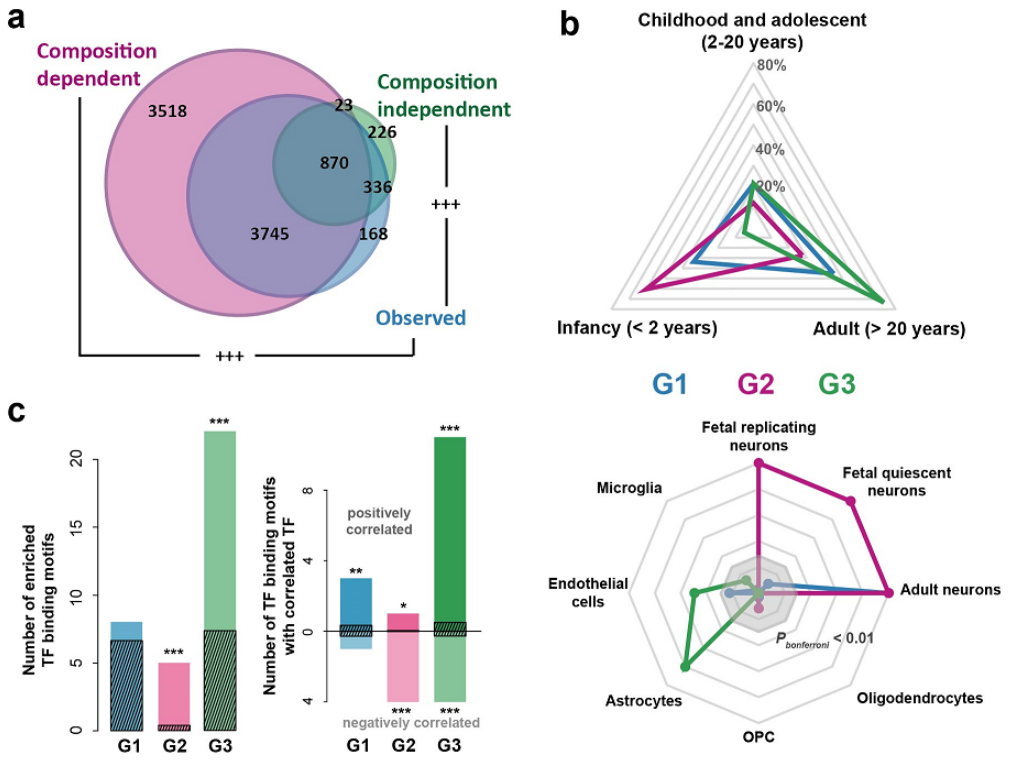


Figure 3. Age-related changes happened in the composition-dependent and composition-independent components. (a) The number of genes with age-related changes in each component or the observed expression level, based on the human postnatal brain data set (blue – observed expression level; pink – composition-dependent component; green – composition-independent component). (b) The expression properties of genes with age-related changes in the composition-dependent or composition-independent components. G1 – genes with age-related changes in both components; G2 – genes with age-related changes only in the composition-dependent component; G3 – genes with age-related changes only in the composition-independent component. Top: proportion of genes with highest expression level at each of the three lifespan stages. Bottom: expression enrichment in each of the eight cell types for the three groups of genes, represented as –log10(*P*), where *P* being the p-value of one-sided Wilcoxon's rank test of log10-transformed fold change from the particular cell type to the remaining cell types, between each of the three groups of genes and all the expressed protein-coding genes. The grey octagonal boxes represent –log10(*P*) equaling to values from ten (the outermost box) to two (the inner most box). Strong expression enrichment with –log10(*P*)>10 was presented as ten. (c) Regulation of genes with age-related changes in either component by transcription factors (TFs). (Left) the number of TF binding motifs enriched among genes within each group. The dark streaked bars represent the mean

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

number of enriched TF binding sites expected by chance, calculated by 1000 random assignment of the expressed genes into the three groups. (Right) the number of TF binding motifs with its representative TF correlated with the targets (correlated TF binding motifs) in the same group. The dark streaked bars represent the mean number of correlated TF binding motifs expected by chance, calculated by 1000 random assignment of the expressed genes into the three groups. The asterisks show significance of the numbers (* $P<0.05$, ** $P<0.01$, *** $P<0.001$, Bonferroni corrected).

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

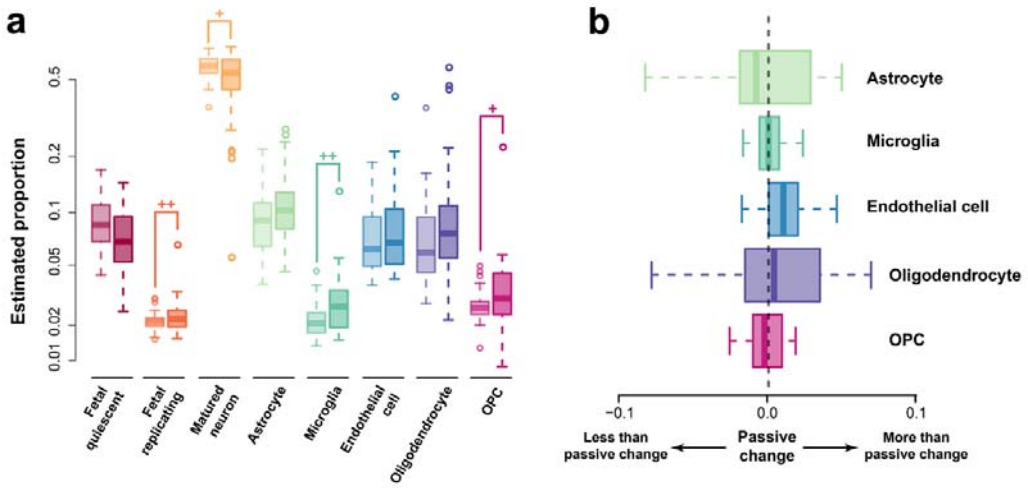


Figure 4. Alteration of cell type composition in ASD. (a) The estimated proportions of each cell types in ASD case samples and healthy control samples with ages at the same age range. Each pair of boxes with the same color show the proportions of one cell type, with the left box represents the proportions in healthy samples, and the right box represents the proportions in ASD samples. The plus signs mark significant difference according to Wilcoxon's test (+ – $P<0.1$; ++ – $P<0.05$). (b) The disparity of observed altered non-neuron cell type compositions and the predicted altered compositions with the assumption of passive change of non-neuron cell type composition. Each box with a different color showed the observed altered proportion of one non-neuron cell type subtracted by the predicted altered proportion of the cell type based on the altered proportion of neurons scaled by the relative proportions of the non-neuron cell type in healthy samples.

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

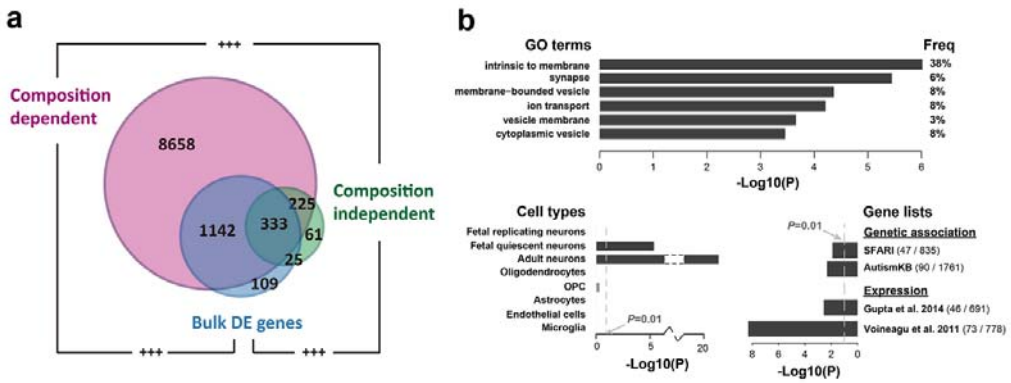
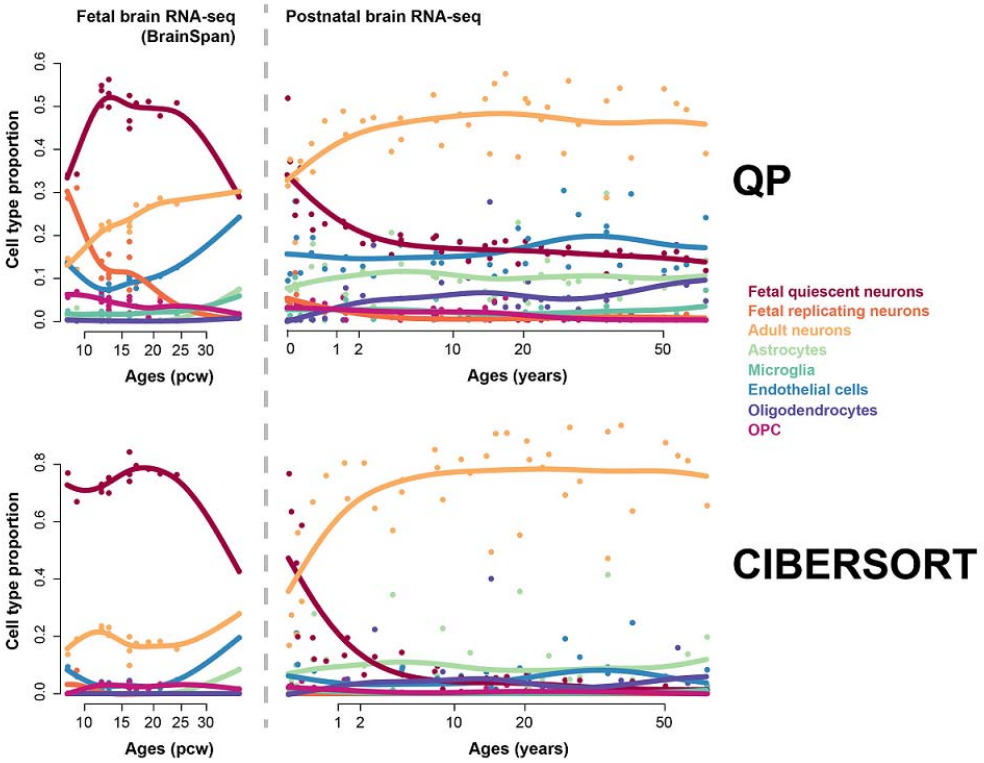


Figure 5. Composition-dependent and independent transcriptome alteration in ASD. (a) The number of genes with ASD alterations in each component or the bulk expression levels (blue – bulk expression levels; pink – composition-dependent component; green – composition-independent component). The plus signs mark the significance of overlap (+++ - Fisher's exact test $P<0.0001$). (b) Functional investigations of genes with ASD alterations in the composition-independent components, including GO enrichments (upper), cell type expression enrichments (button left) and enrichments of ASD-associated genes (button right). The bars show the log10-transformed *P*-values. The values within brackets in the button right panel show the numbers of all associated genes (right) and the sizes of the overlap with genes with ASD alterations in their composition-independent components (left).

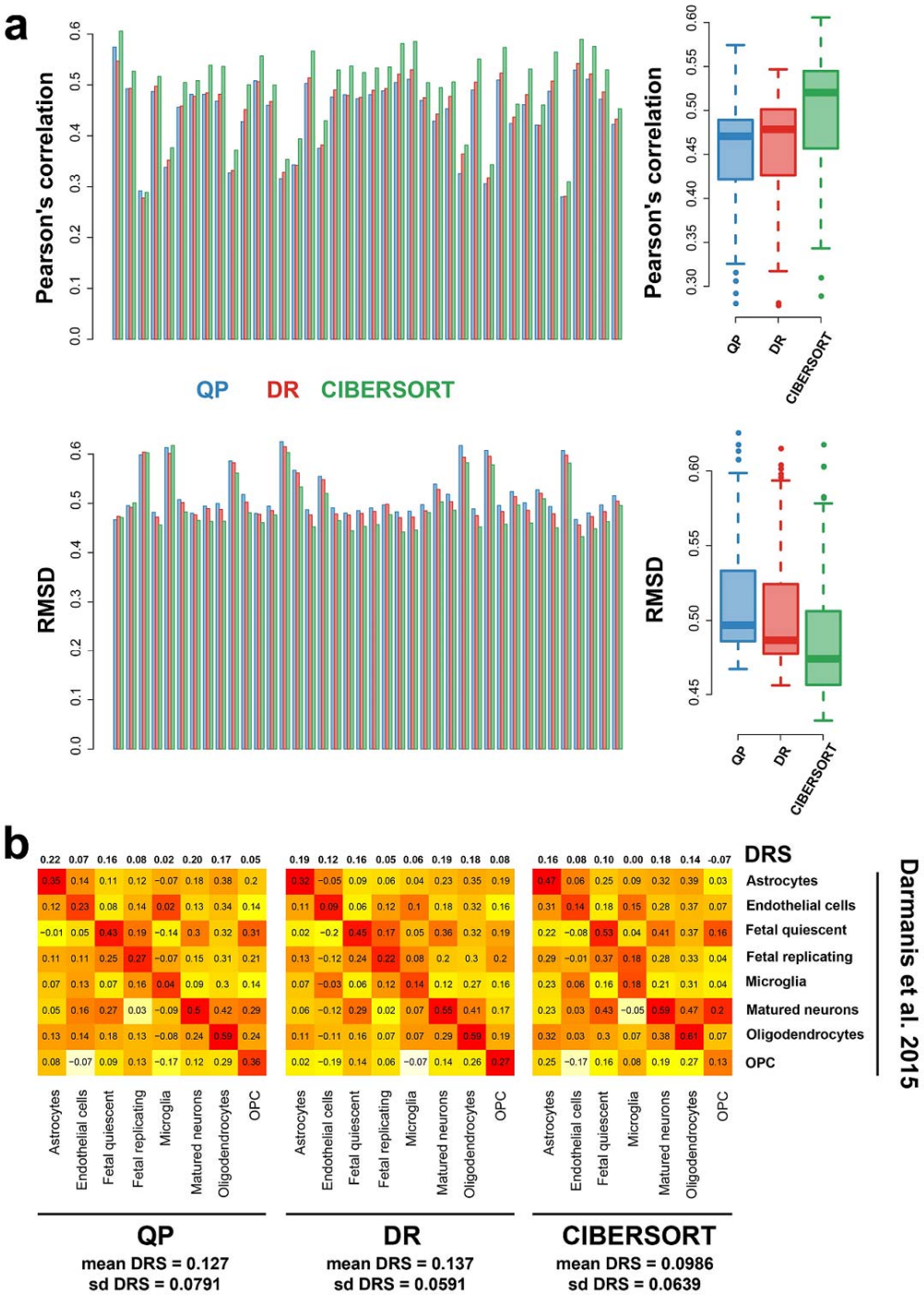
bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary materials



Supplementary Figure S1. The cell type compositions across lifespan in human brains, using QP (upper) and CIBERSORT (lower) algorithms. The dots show the estimated proportions of samples, and the curves show the spline interpolation results. The left panel shows the cell type compositions in human fetal brains, based on the human embryonic developmental brain RNA-seq data from BrainSpan; the right panel shows the cell type compositions in human postnatal brains, based on the human postnatal brain data set.

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

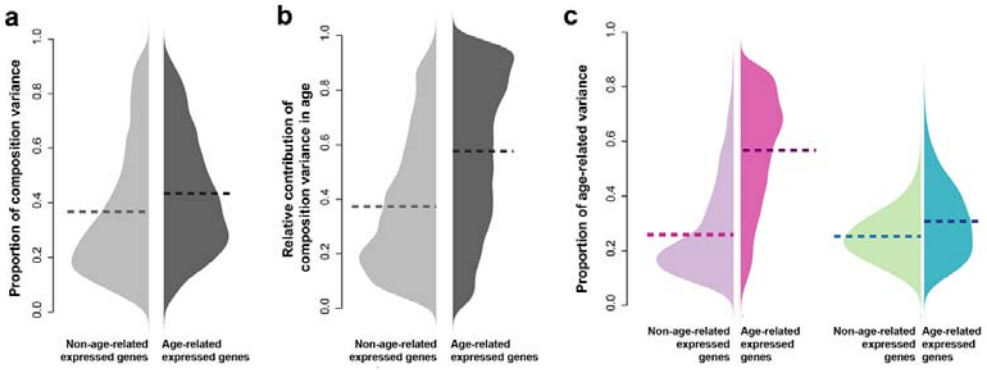


Supplementary Figure S2. Comparison of the three deconvolution algorithms: QP, DR and CIBERSORT. (a) Performance estimation using the similarity measurements, including the Pearson's correlation coefficient (PCC, upper) and root-mean-square deviation (RMSD, lower) between the observed bulk expression and the predicted bulk expression based on estimated cell type compositions. Colors of bars and boxes represent different algorithms. Each triple of bars in the left panel shows PCC and

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

RMSD of each sample. Boxes in the right panel show PCC and RMSD across all samples. (b) Performance estimated based on the discrimination measurement, i.e. discriminating recovery score (DRS). The heatmaps show PCC between the re-estimated cell type transcriptome (columns) and the measured cell type transcriptome by Darmanis et al. 2015 [21] (rows). Each of the three heatmaps represents the results based on QP-based (left), DR-based (middle) and CIBERSORT-based (right) composition patterns, respectively. Numbers in the heatmaps show the PCC, with colors representing the relative PCCs in each column. The DRS of each cell type is shown above each column of the heatmaps, with the mean DRSs with the standard deviations shown at the bottom.

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.



Supplementary Figure S3. The impact of cell type composition changes in age-related expressions, based on CIBERSORT-based composition patterns. (a) The proportions of gene expression variance explained by cell type composition (composition-related variance) in the human postnatal brain data set. Light grey – genes without age-related expression; dark grey – genes with age-related expression. The horizontal dash lines show the mean proportion of composition-related variance. (b) The relative contribution of the composition-dependent component to the variance explained by age (age-explained variance), measured as ratio of age-explained variance in the composition-dependent component to the sum of age-explained variance in both components. Light grey – genes without age-related expression; dark grey – genes with age-related expression. The horizontal dash lines show the mean proportion of composition-related variance. (c) The proportion of variance explained by ages (age-explained variance) in each of the two components of expression: pink – the composition-dependent component; green – the composition-independent component. The light colors represent the proportions of age-related variance in genes without age-related expression changes; the dark colors represent the proportions of age-related variance in genes with age-related expression changes. The horizontal dash lines show the means of age-related variance proportions.

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

Postnatal development

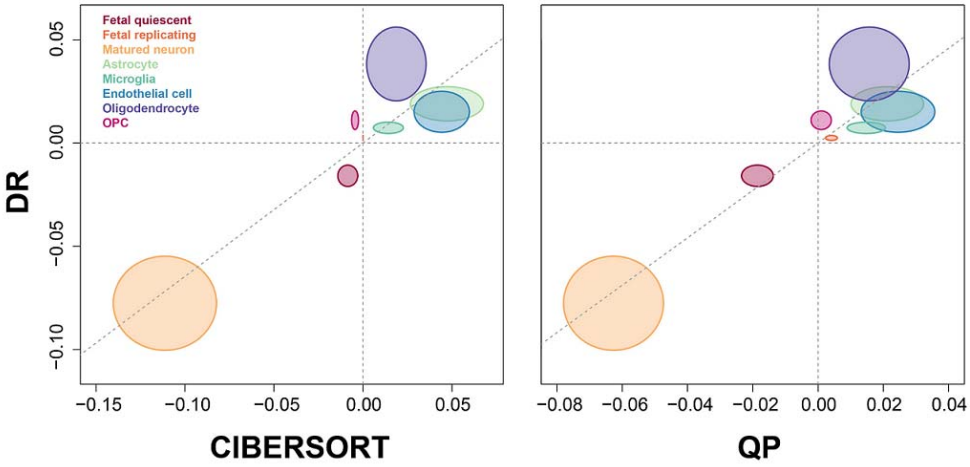
		DR				
		G1	G2	G3	Others	Total
CIBERSORT	G1	**400**	177	33	4	614
	G2	408	**5686**	37	246	6377
	G3	56	49	**256**	154	515
	Others	29	1351	236	**3855**	5471
	Total	893	7263	562	4259	

ASD alteration

		DR			
		Composition-independent	Composition-dependent only	Others	Total
CIBERSORT	Composition-independent	**229**	39	7	275
	Composition-dependent only	351	**6824**	647	7822
	Others	64	2937	**2934**	5935
	Total	644	9800	3588	

Supplementary Figure S4. DR and CIBERSORT based cell type composition patterns resulted in consistent classification of genes with composition-dependent and/or composition-independent component showing (a) age-related changes in postnatal human brains and (b) ASD alteration. The red shadows indicate significant overlap (hypergeometric test, $P<0.0001$).

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.



Supplementary Figure S5. Three deconvolution algorithms provide similar estimations of cell type composition alterations in ASD samples. Different ellipses represent the eight cell types. The center of each ellipse shows the average altered proportion of each cell type in ASD samples comparing to the healthy samples, based on the DR-based composition patterns (y-axis) or CIBERSORT (left panel) / QP (right panel) (x-axis) based composition patterns. The major and minor radius of each ellipse is proportional to the standard deviation of the alteration based on the corresponding algorithm.

bioRxiv preprint doi: https://doi.org/10.1101/065292; this version posted February 12, 2017. 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.

Supplementary Table S1. The CIBERSORT-selected markers used in CIBERSORT and QP algorithms, and the cell type signature genes of the eight main cell types in human brains used in DR algorithms.

Supplementary Table S2. ANCOVA results indicating age-related changes in postnatal human brains in composition-dependent or composition-independent components.

Supplementary Table S3. Functional enrichment (DAVID) of each of the three groups of age-related changed genes.

Supplementary Table S4. ANCOVA results indicating ASD alteration in composition-dependent or composition-independent components.

Supplementary Table S5. Functional enrichment (DAVID) of genes with altered expressions in ASD in their composition-independent components.