Cell type composition is the primary – but far from the only – power shaping temporal transcriptome of human brains

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

Corresponding authors

Correspondence to Zhisong He: zhisong.he@gmail.com

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. While cell type composition is known to be temporally variable in brains, it remains unclear whether it is the primary cause of age-related transcriptome changes. Here, taking advantage of published human brain single-cell RNA-seq data, we applied a two-step transcriptome deconvolution procedure to the public age series RNA-seq data to quantify the contribution of cell type composition in shaping the temporal transcriptome in human brains. We estimated that composition change contributed to around 25% of the total variance and was the primary factor of age-related transcriptome changes. On the other hand, genes with substantial composition-independent temporal expression changes were also observed, which had diverged expression properties, functions and regulators as genes with temporal expression changes related to composition. This indicates a second independent mechanism shaping the human brain's temporal transcriptome properties, which is important for human brain functions.

Introduction

The development and aging of human brains are complex processes, which are shaped by anatomical and molecular changes [1-4]. 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 [5-8].

The human brain, however, is a highly complex and heterogeneous organ comprised of numerous different cell types, including neurons, 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 [9] and plays a unique and essential role in the development and functions of the brain [10]. Different cell types are also known to show different spatial-temporal distributions. Neurons, for instance, are well-known to emerge in early embryonic development, while the remaining glia cells appear much later [11]. The cell-type composition of a brain also keeps on changing after birth. For example, myelination – a process largely linked to oligodendrocytes – is known to continue for at least 10 years after birth [12]. 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? And, besides the age-related changes caused by composition changes, what's the biological meaning of the rest?

In order to comprehensively answer these questions, an accurate estimation of cell type composition in human brains is required. Although experiments including stained cell counting [13-16] and large scale single-cell RNA-seq [17,18] 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 [19,20]. The recent emergence of the human brain single cell RNA-seq data, covering all the main cell types in the human brain [17], now renders this approach more feasible.

Meanwhile, quantifying the contribution of cell type composition changes in human brain differences, especially age-related molecular changes, also requires the proper decomposition of molecular profiles into components related or not to the cell type composition change. The typically used linear regression model [21,22] does not account for the nonnegative nature of the molecular signatures, e.g. gene expression levels measured in RNA-seq. Furthermore, most of those studies only scratched the surface, and provided no further investigation into the underlying biological characteristics

and meaning, especially for variance independent of composition.

In this study, we used two-step nonnegative deconvolutions 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 the composition-dependent component from gene expression suggested that cell-type composition explained about 25% of the total expression 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. More interestingly, distinct expression properties, functions and potential regulators of genes with age-related changes in the composition-dependent and independent components were observed, implying the diverged functional contributions of these two components in the human brain. Additionally, although not as the primary force in shaping the age-related expression pattern in human brains, genes with age-related changes independent of composition changed their expression in autism, a neurodevelopmental disorder characterized by patterns of behaviors and impaired social communication and interaction, which further suggested their important functions in brains.

Results

Cell type composition estimation in simulated data and RNA-seq data of human cortical layers

To obtain the gene expression information of cell types in human brains required for transcriptome deconvolution, we used the published human brain single cell RNA-seq data [17] and estimated the expression level of 14054 protein-coding genes in eight main cell types in human brains. Meanwhile, 1491 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).

To estimate the brain cell type composition given the bulk brain tissue RNA-seq data based on the human brain cell type expression profiles, we tried two different methods: quadratic programming (QP) based deconvolution and diffusion ratio (DR) based deconvolution. In brief, the QP-based deconvolution modeled the expression of a cell type signature gene in the bulk tissue as a linear combination of its expression in different cell types according to the unknown cell type mixing proportion. This method has been widely used for deconvolution in transcriptome data [19,20]. The DR-based deconvolution, on the other hand, 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. Simulations suggested that both methods provided good estimations regarding cell-type composition, while the QP-based deconvolution surpassed the DR-based deconvolution (Supplementary Fig. S1).

To test the application to real-world data, we applied both deconvolution methods to an RNA-seq data set representing the transcriptome of human cortical layers (SRP065273) which were known to have varied cell type constitutions [23]. Both methods provided similar cell type composition patterns across the cortical sections, each of which represents parts of one cortical layer or the mixture of two layers (Fig. 1a). The estimated composition patterns were concordant with the known composition difference, e.g. the high abundance of oligodendrocytes and low abundance of neurons in the deep layers. On the other hand, the DR-based deconvolution resulted in better exactness than the QP-based deconvolution (Methods, Fig. 1b), indicating more robustness of DR-based deconvolution in the noisy data set which was processed and normalized separately.

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 (age-DS1) [8]. Both the QP-based and DR-based deconvolution provided similar composition patterns (Fig. 2a and Supplementary Fig. S2). Another unpublished RNA-seq data including 72 samples aged from 0 days old to 98 years old (age-DS2) resulted in similar estimations as well (Supplementary Fig. S2). 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 [24], and the increase of oligodendrocytes with the decrease of OPC which may due to the myelination process [12]. The remaining cell types including astrocyte, endothelial cells and microglia did not show significant changes across the lifespan. The comparison between the two estimated compositions using two different deconvolution methods suggested the better reproduction of the bulk tissue gene expression for the DR-based deconvolution, which was similar to the observation in the cortical layer data (Supplementary Fig. S2).

We next applied the DR-based deconvolution to the human embryonic developmental brain RNA-seq data obtained from Allen Brain Atlas (Fig. 2a). A large proportion of fetal replicating neurons was observed in samples before 12 post conception weeks (pcw) but decreased dramatically since 12 pcw. This observation, coupling with the increase of fetal quiescent neurons, well matched with the neuronal proliferation that occurred during four pcw to 12 pcw [11]. Intriguingly, the estimated compositions, 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.

Cell type expression calibration and estimation of composition component variance

Despite of the reliable composition estimation, we noticed a huge discrepancy between the observed and predicted bulk tissue gene expression based on cell type expression and estimated composition. We hypothesized several reasons, including batch effect between the bulk tissue RNA-seq and the single-cell RNA-seq data, and the intrinsic molecular profile change across the lifespans. Such discrepancy had to be corrected for proper decomposition of gene expression variance into variance due to cell type composition. Therefore, we adopted the second deconvolution based on quadratic programming to calibrate cell type gene expression. Applying the method to age-DS1 with the DR-based composition resulted in calibrated cell type expression

which was concordant with the cell type expression by the single cell RNA-seq (Fig. 2b), with fidelity factor 0.125 (permutation test, $P<0.001$). The discrepancy between the observed and predicted bulk tissue gene expression was eliminated significantly when the calibrated cell type expression was used (Fig. 2c).

Based on the estimated cell type composition and the calibrated cell type expression, we decomposed the bulk gene expression into the composition-dependent and composition-independent components. Their contributions to the overall gene expression variance were estimated. On average, the composition component explained 22.6%-26.4% of the total variance in the human postnatal age series data (Fig. 3a). This proportion was much larger in the human cortical layer data (47.2%-54.3%, Supplementary Fig. S3), where the cell type composition had been known to be varied [23]. 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 29.7%-34.7% of the total variance which was significantly higher than the other genes (Fig. 3b, 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 composition component (55.8%) was much larger than the proportion of composition variance among total variance (29.7%-34.7%) for age-related expressed genes, while the difference was much smaller for the non-age-related expressed genes (27.8% vs. 19.3%-22.1%) (Fig. 3b).

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, $P<0.001$, Fig. 3b); 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, $P<0.001$, Fig. 3b). 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. Notably, however, 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.

The age-related changes in the composition-dependent and independent components

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, 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 (Fig. 4a). 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.

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 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 (Fig. 4b). G1 and G2 genes, and especially the latter gene, showed higher expression levels in early postnatal development, while G3 genes were highly expressed in the 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 (Fig. 4b).

More importantly, the three groups of genes showed distinctive functional enrichments (DAVID [25], Supplementary Table S2). 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, $P<0.001$) (Fig. 4c). 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, $P<0.001$) (Fig. 4c). 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 [26-28]. 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 [29-31].

Significance of composition-independent component in autism pathogenesis

Autism spectrum disorder (ASD) is a common neurodevelopmental disorder. In the previous study based on microarray technology [32], 444 genes have been reported with differential expression in autistic brains (ASD-DE genes), 343 of which were detected in both the single-cell RNA-seq and age-DS1 data. Among them, 158 genes showed increased expression level in autistic brains, while 185 showed decreased expression level. Intriguingly, we observed strong enrichment of genes with decreased expression in autistic brains in G1 (Fisher's exact test, odds ratio=4.00, $P<10^{-10}$), and in contrast, enrichment of genes with increased expression in autistic brains in G3 (Fisher's exact test, odds ratio=2.776, P=0.0002). Although with distinct functions, these ASD-DE genes were enriched for genes with age-related changes in the composition-independent component rather than in the composition-dependent component. This result indicated the important functions of genes with age-related changes in the composition-independent component. It also implied that the molecular pathology of autism alters the regulatory machinery regulating the molecular profiles of certain cell types instead of main cell type composition, and harms the cell-cell communications in brains.

Discussions

In this study, we refined and implemented a two-step transcriptome deconvolution procedure to estimate cell type composition and its contribution to the sample variance. With the gene expression of eight main cell types in human brain estimated using the human brain single cell RNA-seq data [17], we applied the deconvolution procedure to the human postnatal age series RNA-seq data [8], which resulted in composition patterns consistent with the prior studies. On the other hand, however, we also noticed that our estimated proportion of total neurons in the adult samples reached around 70%, which was much higher than the estimation based on cell counting [33,34] or DNA methylation deconvolution [22]. We suspected the primary cause to be the close-to-two-fold as much as RNA content in neurons than glia cells, as reported previously [35].

Based on the variance analysis, we regarded the composition change to be the main source of the age-related expression change in human brain. This was a consistent conclusion drawn by previous studies based on DNA methylation [22]. The proportion of variance explained by cell types, on the other hand, was smaller: only about 30% for the age-related expressed genes, while it was 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 showed age-related manner and independent from the DNA methylation, which might be an interesting focus for further study.

By decomposing the expression level into the composition-dependent component, *i.e.* expression explained by the cell type composition changes, and the composition-independent component which relies on other factors, we reported the first attempt to our knowledge to study the age-related transcriptome changes in brain in a more comprehensive way. In such a way, we distinguished changes due to cell type composition and changes due to other factors, manifesting changes happened in transcriptome of one or multiple cell types which were independent from the composition. Interestingly, along with the distinct biological implications of the two components, genes with age-related changes in the two components showed entirely diverged expression properties and functions, implying two independent machineries shaping the human brain transcriptome in the age-related manner.

Interestingly, genes with age-related changes in the composition-dependent components showed high expression level in the infant stages, *i.e.* < 2 years old, with significantly enriched expression in the three types of neurons. This suggested 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 were due to the transition from fetal neurons to adult neurons, *i.e.* the neuron maturation process. This thus implied that the neuron maturation was the primary factor creating the age-related expression, especially during the infant stage.

It is worth noting that, while the cell type composition was apparently the primary driving force of the age-related transcriptome change in human brain, especially during the early postnatal development, the age-related changes to the composition-independent component is appealing. With the major contribution of compositional change, these changes were easily overwhelmed when the two components remain mixed. Interestingly, unlike the genes with age-related composition-dependent changes which mostly expressed highly in the early postnatal development, genes with age-related composition-independent changes tended to have higher expression in adults. What's more, these composition-independent changes were significantly related to either synapse in neurons (G1), or extra-cellular regions and signal peptides in astrocytes and endothelial cells (G3), both of which were relevant to the cell-to-cell communications. This is unlikely to be a coincidence. As a creative information-processing system, such communications are critically important for human brains [36,37], 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 [38]. Additionally, our intuitive enrichment analysis of genes with differential expression in autistic brains suggested that the composition-independent component was more likely to be the primary contributor to the transcriptome alteration in autistic brains, which further supported their importance in human brains. However, further analysis of decomposition and comparison of gene expression in autistic and healthy brains would be necessary to answer this question directly.

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. If nothing else, we hope that our attempt to decompose expression into composition-dependent and composition-independent components will inspire further studies to elucidate

transcriptome changes in a more comprehensive manner.

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 two unpublished data sets, including the human cortical layers data set (SRP065273) and the other human postnatal age-series brain RNA-seq data set with 72 samples, were processed in the same way. The pre-calculated RPKM of the fetal human brain samples were downloaded from Allen Brain Atlas (http://www.brainspan.org/static/download.html).

Deconvolution for cell-type composition

Two 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:

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

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 [19,20], and can be solved using quadratic programming [39].

The second 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. Under this assumption, the proportion of cell type i (represented as f_i) was simply estimated as:

$$f_i = median_g\left(\frac{x_g}{C_{ig}}\right)$$

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 $\sum_i f_i = 1$.

Simulations for composition estimation

For each cell type, the cell type gene expression level was firstly estimated based on each of the 100 times of cell bootstrapping. Assuming the lognormal distribution of gene expression, we estimated the mean and standard deviation for the log-transformed gene expression (in FPKM) for each gene in each cell type.

For every simulation, we randomly simulated the mixing proportion of the cell types. The simulated bulk gene expression level for each gene was then generated as its mean expression across the cell types weighted by the simulated mixing proportion. Note that for each gene, its expression level in each cell type was randomly generated based on the expression level mean and standard deviation mentioned above. Finally, a white noise with standard variance proportional to the average gene expression level in all cell types was also added to the simulated expression level.

To evaluate the accuracy of the deconvolution results, the Pearson's correlation coefficient was calculated between the estimated cell type mixing proportions and the simulated cell type mixing proportions. This was done for each simulation, and altogether 1000 simulations were run to estimate the performance of the deconvolution for composition estimation.

Deconvolution for cell-type expression profile calibration

The similar model as the QP-based deconvolution for cell-type composition described above was used for the second deconvolution to calibrate cell-type expression profile. Similarly, the deconvolution problem for each gene can be seen as a constrained linear least-square problem, that is:

$$\min\left(\left\|Fc - x\right\|^2\right), st.c_i \geq 0, \forall i .$$

Here, $\boldsymbol{c}$ was the vector of calibrated 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.

Measurement of fidelity factor for the deconvolutions

Two different measurements of fidelity factor were used for the two deconvolution tasks. For the first deconvolution, that is, the deconvolution to estimate cell type composition, the difference 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 was used to describe the fidelity factor. In more detail, for each gene that was detected in both the bulk tissue data and the human brain single cell data, the absolute value of difference between the observed log10-transformed FPKM and the predicted log10-transformed FPKM was calculated, and then summed up across all the genes. This was used as the fidelity factor of one sample, and the overall fidelity factor was estimated as the average fidelity factor of all the bulk tissue samples.

For the second deconvolution, that is, the deconvolution to calibrate cell type expression level, and fidelity factor for cell type *i* was represented as:

$$ff_i = cor(e_i, \hat{e}_i) - \max_j cor(e_j, \hat{e}_i)$$

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 calibrated log10-transformed FPKM of cell type *i*. Using the correlation with other cell types' expression profiles as a control, this value thus represented the similarity between the calibrated cell type transcriptome profiles and the real one. The mean across different cell types was then used as the proxy of the overall fidelity factor.

Age test: ANCOVA based on natural spline with variable degree of freedom

For each gene, with its observed expression level, composition-dependent component, or composition-independent component as the response variable, 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 ages (sqrt-age). The best alternative model was chosen by applying the adjusted r^2 criterion [5]. Genes with BH corrected FDR<5% 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.

Acknowledgement

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). We thank Prof. Dr. P. Khaitovich for providing the unpublished RNA-seq data. We thank 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.

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.

Competing financial interests

The authors declare no competing financial interests.

References

1 Huffman, K. The developing, aging neocortex: how genetics and epigenetics influence early developmental patterning and age-related change. *Front Genet* **3**, 212, doi:10.3389/fgene.2012.00212 (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, doi:10.1523/JNEUROSCI.3554-12.2013 (2013).

3 Nie, J., Li, G. & Shen, D. Development of cortical anatomical properties from early childhood to early adulthood. *Neuroimage* **76**, 216-224, doi:10.1016/j.neuroimage.2013.03.021 (2013).

4 Giedd, J. N. *et al.* Child psychiatry branch of the National Institute of Mental Health longitudinal structural magnetic resonance imaging study of human brain development. *Neuropsychopharmacology* **40**, 43-49, doi:10.1038/npp.2014.236 (2015).

5 Somel, M. *et al.* Transcriptional neoteny in the human brain. *Proc Natl Acad Sci U S A* **106**, 5743-5748, doi:10.1073/pnas.0900544106 (2009).

6 Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483-489, doi:10.1038/nature10523 (2011).

7 Liu, X. *et al.* Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques. *Genome Res* **22**, 611-622, doi:10.1101/gr.127324.111 (2012).

8 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, doi:10.1261/rna.043075.113 (2014).

9 Zhang, Y. *et al.* An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J Neurosci* **34**, 11929-11947, doi:10.1523/JNEUROSCI.1860-14.2014 (2014).

10 Allen, N. J. & Barres, B. A. Neuroscience: Glia - more than just brain glue. *Nature* **457**, 675-677, doi:10.1038/457675a (2009).

11 Tau, G. Z. & Peterson, B. S. Normal development of brain circuits. *Neuropsychopharmacology* **35**, 147-168, doi:10.1038/npp.2009.115 (2010).

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

13 Mittelbronn, M., Dietz, K., Schluesener, H. J. & 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).

14 Herculano-Houzel, S. & Lent, R. Isotropic fractionator: a simple, rapid method for the quantification of total cell and neuron numbers in the brain. *J Neurosci* **25**, 2518-2521, doi:10.1523/JNEUROSCI.4526-04.2005 (2005).

15 Sherwood, C. C. *et al.* Evolution of increased glia-neuron ratios in the human frontal cortex. *Proc Natl Acad Sci U S A* **103**, 13606-13611, doi:10.1073/pnas.0605843103 (2006).

16 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, doi:10.1073/pnas.0804650106 (2009).

17 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, doi:10.1073/pnas.1507125112 (2015).

18 Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell

RNA-seq. *Science* **347**, 1138-1142, doi:10.1126/science.aaa1934 (2015).

19 Gong, T. *et al.* Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS One* **6**, e27156, doi:10.1371/journal.pone.0027156 (2011).

20 Gong, T. & Szustakowski, J. D. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* **29**, 1083-1085, doi:10.1093/bioinformatics/btt090 (2013).

21 Shen-Orr, S. S. *et al.* Cell type-specific gene expression differences in complex tissues. *Nat Methods* **7**, 287-289, doi:10.1038/nmeth.1439 (2010).

22 Jaffe, A. E. *et al.* Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat Neurosci* **19**, 40-47, doi:10.1038/nn.4181 (2016).

23 Defelipe, J. The evolution of the brain, the human nature of cortical circuits, and intellectual creativity. *Front Neuroanat* **5**, 29, doi:10.3389/fnana.2011.00029 (2011).

24 Sanai, N. *et al.* Corridors of migrating neurons in the human brain and their decline during infancy. *Nature* **478**, 382-386, doi:10.1038/nature10487 (2011).

25 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).

26 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, doi:10.1016/j.neuron.2010.04.038 (2010).

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

28 Wang, L., Wang, R. & Herrup, K. E2F1 works as a cell cycle suppressor in mature neurons. *J Neurosci* **27**, 12555-12564, doi:10.1523/JNEUROSCI.3681-07.2007 (2007).

29 Tichauer, J. E. *et al.* Age-dependent changes on TGFbeta1 Smad3 pathway modify the pattern of microglial cell activation. *Brain Behav Immun* **37**, 187-196, doi:10.1016/j.bbi.2013.12.018 (2014).

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

31 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, doi:10.1007/s12035-013-8395-9 (2013).

32 Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380-384, doi:10.1038/nature10110 (2011).

33 Pelvig, D. P., Pakkenberg, H., Stark, A. K. & Pakkenberg, B. Neocortical glial cell numbers in human brains. *Neurobiol Aging* **29**, 1754-1762, doi:10.1016/j.neurobiolaging.2007.04.013 (2008).

34 Herculano-Houzel, S. The human brain in numbers: a linearly scaled-up primate brain. *Front Hum Neurosci* **3**, 31, doi:10.3389/neuro.09.031.2009 (2009).

35 Filipchenko, R. E., Pevzner, L. Z. & Slonim, A. D. RNA content in the neurons and glia of the hypothalamic nuclei after intermittent cooling. *Neurosci Behav Physiol* **7**, 69-71 (1976).

36 Volterra, A. & Meldolesi, J. Astrocytes, from brain glue to communication elements: the revolution continues. *Nat Rev Neurosci* **6**, 626-640, doi:10.1038/nrn1722 (2005).

37 Bullmore, E. & Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci* **10**, 186-198, doi:10.1038/nrn2575 (2009).

38 Araque, A. & Navarrete, M. Glial cells in neuronal network function. *Philos Trans R Soc Lond B Biol Sci* **365**, 2375-2381, doi:10.1098/rstb.2009.0313 (2010).

39 Lawson, C. L. & Hanson, R. J. *Solving Least Squares Problems*. (Society for Industrial and Applied Mathematics, 1987).

Figures

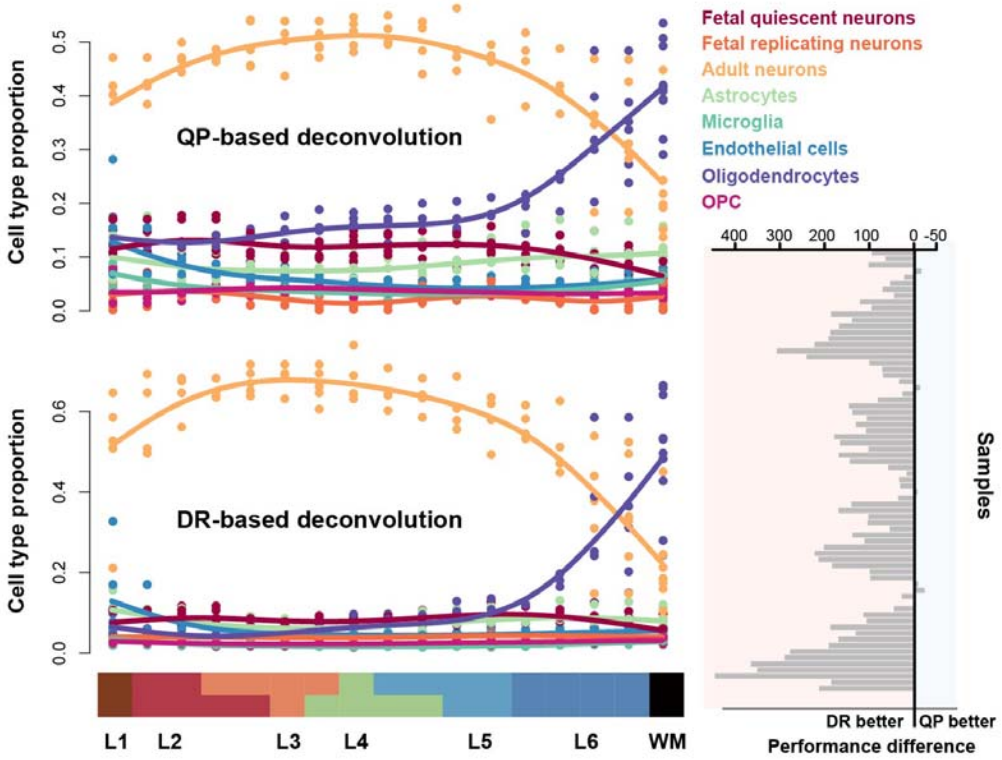

Figure 1. The cell type composition across human cortical layers estimated using quadratic programming (QP) based deconvolution and diffusion ratio (DR) based deconvolution method. (a) The estimated cell type proportion of each of the eight brain cell types in cortical layers. The dots show the samples' estimated proportions, and the curves show the spline interpolation results. The bars on the bottom show the cortical layers assigned to each of the 17 aligned sections, from the left most representing the most superficial layer 1 (L1), to the right most representing the most deep layer 6 (L6) and the adjacent white matter (WM). (b) The difference of deconvolution performance, represented as the discrepancy of predicted bulk tissue gene expression from the observed bulk tissue gene expression, between the QP-based deconvolution and the DR-based deconvolution.

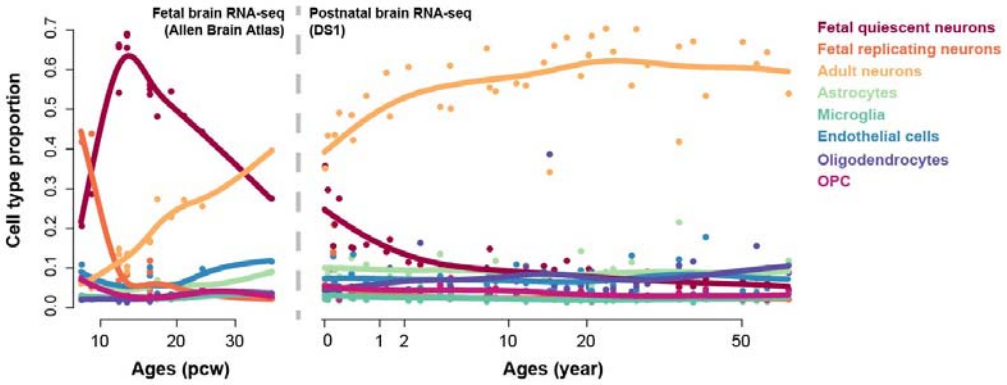

Figure 2. The cell type composition across lifespan in human brain, estimated using the DR-based deconvolution. The dots show the estimated proportions of samples, and the curves show the spline interpolation results. The left panel shows the cell type composition in human fetal brains, based on the human embryonic developmental brain RNA-seq data from the Allen Brain Atlas; the right panel shows the cell type composition in human postnatal brains, based on the human brain age series data set 1 (age-DS1).

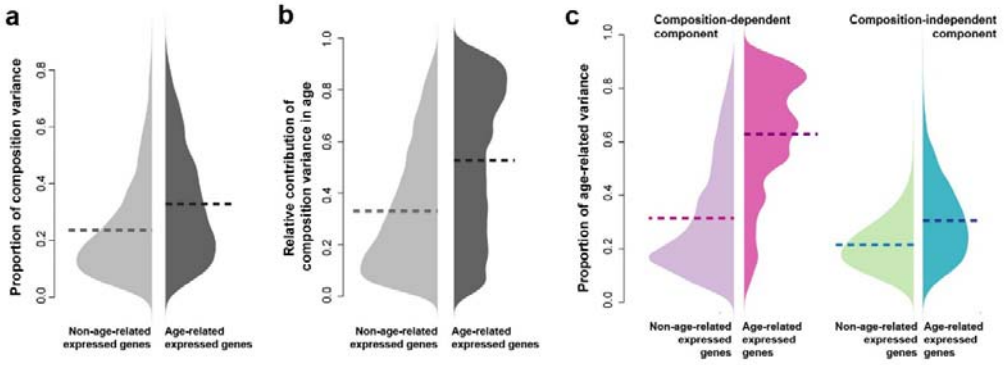

Figure 3. The cell type composition change was the primary driving force of the age-related expression. (a) The proportion of gene expression variance explained by cell type composition (composition-related variance) in age-DS1. Light grey – genes without age-related expression; dark grey – genes with 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.

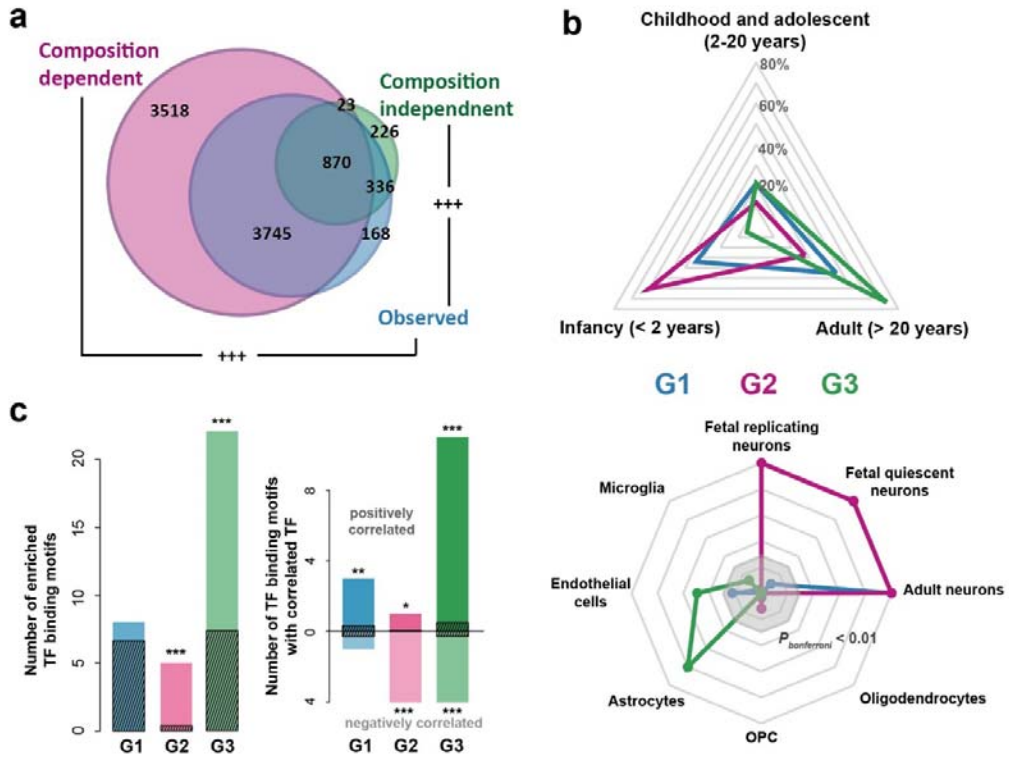

Figure 4. Age-related changes happened in the composition-dependent and composition-independent components. (a) The number of genes with age-related changes in each component or the observed expression level, based on age-DS1 (blue – observed expression level; pink – composition-dependent component; 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 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 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).