Title

Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes

Journal

Nature Neuroscience, 21(9)

ISSN

1097-6256

Authors

Kelley, Kevin W

Nakao-Inoue, Hiromi

Molofsky, Anna V

et al.

Publication Date

2018-09-01

DOI

10.1038/s41593-018-0216-z

Peer reviewed

Published in final edited form as:

Nat Neurosci. 2018 September ; 21(9): 1171–1184. doi:10.1038/s41593-018-0216-z.

Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes

Kevin W. Kelley[1,2,3,4,5], **Hiromi Nakao-Inoue**[2,3,4], **Anna V. Molofsky**[2,3,4], and **Michael C. Oldham**[1,2,3,*]

[1]Department of Neurological Surgery, University of California at San Francisco, San Francisco, CA, USA

[2]The Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California at San Francisco, San Francisco, CA, USA

[3]Weill Institute for Neurosciences, University of California at San Francisco, San Francisco, CA, USA

[4]Department of Psychiatry, University of California at San Francisco, San Francisco, CA, USA

[5]Medical Scientist Training Program and Neuroscience Graduate Program, University of California at San Francisco, San Francisco, CA, USA

Abstract

It is widely assumed that cells must be physically isolated to study their molecular profiles. However, intact tissue samples naturally exhibit variation in cellular composition, which drives covariation of cell-class-specific molecular features. By analyzing transcriptional covariation in 7221 intact CNS samples from 840 neurotypical individuals representing billions of cells, we reveal the core transcriptional identities of major CNS cell classes in humans. By modeling intact CNS transcriptomes as a function of variation in cellular composition, we identify cell-class-specific transcriptional differences in Alzheimer's disease, among brain regions, and between species. Among these, we show that *PMP2* is expressed by human but not mouse astrocytes and significantly increases mouse astrocyte size upon ectopic expression *in vivo*, causing them to more closely resemble their human counterparts. Our work is available as an online resource (http://oldhamlab.ctec.ucsf.edu/) and provides a generalizable strategy for determining the core molecular features of cellular identity in intact biological systems.

*Correspondence: Michael.Oldham@ucsf.edu.

AUTHOR CONTRIBUTIONS

K.W.K. and M.C.O. conceived and designed the analytical strategies and wrote the manuscript. K.W.K. performed most data analyses and histological experiments. K.W.K. and H.N. performed *PMP2* expression experiments under supervision from A.V.M.

ACCESSION CODES

Accession codes and URLs for all gene expression datasets analyzed in this study are provided in Table S1.

COMPETING INTERESTS

The authors declare no competing interests.

INTRODUCTION

Identifying the molecular features that define cellular identities is a fundamental goal of biological research. Consequently, several 'bottom-up' methods have been developed to isolate cells for molecular profiling, including fluorescence-activated cell sorting (FACS), immunopanning (IP), and sorting of single cells (SC) or nuclei (SN). Although these methods are readily applied to many biological systems, their applicability to the adult human CNS is limited by technical factors and practical considerations. For example, FACS, IP, and SC typically require fresh tissue and have therefore been mostly limited to surgical samples from a handful of CNS regions and individuals[1–3]. SN[4, 5] is compatible with frozen tissue but, like SC, suffers from technical noise caused by tissue dissociation, nucleus/cell capture, cDNA preamplification, and stochastic transcript coverage[6]. Furthermore, there is a trade-off between sequencing depth and the number of nuclei/cells that can be analyzed.

The adult human CNS is large, heterogeneous, and difficult to dissociate due to extensive myelin. It consists of ~170 billion cells, about half of which are neurons[7]. The remaining cells consist mostly of oligodendrocytes, astrocytes, and microglia, which are collectively referred to as glia. Identifying transcriptional differences among neuronal and glial subtypes is an important goal, since heterogeneity within CNS cell classes is incompletely understood. However, it is equally important to understand what CNS cell subtypes have in common. For example, is there a core set of genes whose expression is shared among all neurons? All astrocytes? Etc. Answering these questions will fill critical gaps in our understanding of CNS cell biology, produce novel experimental and analytical strategies, and provide important insights into the cellular origins of CNS pathologies.

Most studies of human CNS transcriptomes have analyzed intact postmortem samples. Because these samples are heterogeneous and cells must be destroyed to extract RNA, it is often assumed that these datasets contain no information about the cellular origins of gene expression. However, it is axiomatic that intact tissue samples from any biological system will exhibit variable cellular composition. Therefore, when many intact tissue samples are analyzed, genes expressed with the greatest sensitivity and specificity in the same cell class should appear highly correlated, since their expression levels depend primarily on the proportion of that cell class in each sample[8]. In support of this reasoning, which has motivated numerous *in silico* deconvolution strategies[9–15], we previously discovered highly reproducible gene coexpression modules in microarray data from intact human brain samples that were significantly enriched with markers of major CNS cell classes[16]. These findings were replicated in studies of intact CNS transcriptomes from mice[17], rats[18], zebra finches[19], macaques[20], and humans[21].

Gene coexpression modules corresponding to major cell classes are therefore robust and predictable features of CNS transcriptomes derived from intact tissue samples. Furthermore, the same genes consistently show the strongest affinities for these modules, offering substantial information about the molecular correlates of cellular identity[16]. Over the past decade, thousands of intact, neurotypical human samples from every major CNS region have been transcriptionally profiled. These data provide an unprecedented opportunity to determine the core transcriptional features of cellular identity in the human CNS from the

'top down' by integrating cell-class-specific gene coexpression modules from many independent datasets.

RESULTS

Gene coexpression analysis of synthetic brain samples accurately predicts differential expression among CNS cell classes

To illustrate the premise of our approach, we aggregated SC RNA-seq data from adult human brain[1] to create synthetic samples that mimic the heterogeneity of intact tissue (Fig. 1A). We performed unsupervised gene coexpression analysis to identify gene coexpression modules in each synthetic dataset that were maximally enriched with published markers[22, 23] of astrocytes, oligodendrocytes, microglia, or neurons ('cell-class modules'; Fig. 1A). Intuitively, expression variation in a cell-class module primarily depends on the representation of that cell class in each sample. Mathematically, the vector that explains the most variation in a coexpression module is its first principal component, or module 'eigengene' (Fig. 1A)[24]. This reasoning suggests that a cell-class module eigengene should approximate the relative abundance of that cell class in each sample. Because the precise cellular composition of each synthetic sample was known, we tested this hypothesis and found that actual cellular abundance was nearly indistinguishable from that predicted by cell-class module eigengenes (Fig. S1A).

To determine the affinity of each gene for each cell-class module, we calculated the Weighted Gene Coexpression Network Analysis measure of intramodular connectivity, k_{ME}[25]. k_{ME} is defined as the Pearson correlation between the expression pattern of a gene and a module eigengene. In the special situation of a cell-class module, k_{ME} therefore quantifies the similarity between the expression pattern of a gene and the relative abundance of that cell class in each sample. Because each sample is a heterogeneous mixture of cells, high k_{ME} for a cell-class module suggests that expression of the gene in that cell class is sensitive and specific. We tested this hypothesis by performing differential expression analysis of SC RNA-seq data for each cell class, restricting our analysis to exactly the same cells used to construct the synthetic samples. As shown in Fig. 1B, the genes that are most significantly up-regulated in a cell class also have the highest k_{ME} values for the corresponding cell-class module. We obtained nearly identical results by aggregating SC RNA-seq data from adult mouse brain[26] (Fig. S1B,C). These findings demonstrate that gene coexpression analysis of intact CNS samples can determine which genes are most differentially expressed among CNS cell classes. More generally, our results suggest that it is not always necessary to physically isolate cells in order to ascertain their defining transcriptional features.

Integrative gene coexpression analysis of intact tissue samples reveals consensus transcriptional profiles of major CNS cell classes in humans

To determine consensus transcriptional profiles of human CNS cell classes, we analyzed 7221 CNS transcriptomes from 840 neurotypical adult humans by combining data from eight studies[21, 27–33] and one resource (http://www.brainspan.org/). These data were generated from intact postmortem tissue samples using diverse technology platforms (Table

S1) and collectively represent billions of cells. Each sample was assigned to one of 19 broad neuroanatomical regions, resulting in 62 regional datasets (Fig. 1C). After data preprocessing and quality control, each dataset consisted of ≥25 samples (median: 76) (Table S1). For each dataset, we performed unsupervised gene coexpression analysis and identified the module that was maximally enriched with published markers[22, 23] of astrocytes, oligodendrocytes, microglia, or neurons (Fig. 1D, Table S2). PC1 of these modules was used to estimate the relative abundance of each cell class over all samples and calculate genome-wide k_{ME} values (Fig. 1E,F). Finally, we combined k_{ME} values for significant cell-class modules from all 62 datasets, producing a single value (z-score) for each gene that quantifies its global expression *fidelity* for each cell class (Fig. 1G). Importantly, estimates of fidelity were highly robust to the choice of gene set used for enrichment analysis (especially for glia; Fig. S2). Canonical markers consistently had high fidelity for the expected cell class and low fidelity for other cell classes (Fig. 2A-D). High-fidelity genes were also significantly and specifically enriched with expected cell-class markers from multiple independent studies (Fig. 2A-D). Compared to glia, the distribution of expression fidelity for neurons was compressed (Fig. 2A-D), likely reflecting neuronal heterogeneity among CNS regions. Genome-wide estimates of expression fidelity for major cell classes are provided in Table S3 and on our web site (http://oldhamlab.ctec.ucsf.edu/).

To further explore how estimates of gene expression fidelity derived from intact tissue relate to gene expression in individual cells, we analyzed droplet-based SN RNA-seq data from neurotypical adult human brains. Habib et al.[4] analyzed 14963 nuclei from cortical and hippocampal samples from five individuals, detecting a median of 529 unique genes/nucleus. Lake et al.[34] analyzed 35289 nuclei from cortical and cerebellar samples from six individuals, detecting a median of 719 unique genes/nucleus. In general, expression patterns of high-fidelity genes were conserved in SN RNA-seq data (Fig. 2E). We extended these comparisons by examining concordance among top high-fidelity genes and top differentially expressed genes for each cell class from each SN study (Fig. 2F-I). For all comparisons, overlap was highly significant ($p < 10^{-15}$), but less so for neurons, which likely reflects differences among CNS regions analyzed in each study. Given the shallow coverage that characterizes droplet-based SC/SN methods, we hypothesized that discordant results might also represent type II error in SN RNA-seq data (i.e. dropouts). To test this hypothesis, we compared expression levels of discordant and concordant genes. For all cell classes, discordant genes were expressed significantly lower than concordant genes in our collection of 7221 intact human CNS samples (Fig. 2F-M). Furthermore, discordant genes were detected far less frequently than concordant genes in single nuclei from both studies (Fig. S3). For an orthogonal perspective, we analyzed expression of discordant genes in cell classes purified by immunopanning from adult human temporal lobe surgical resections[3]. In all cases, gene expression fidelity correctly predicted the dominant cellular source of mRNA (Fig. 2N-Q). These results underscore the sparse nature of current droplet-based SN RNA-seq data. Comparisons of fidelity and differential expression results for all genes are reported in Table S4.

High-fidelity genes reveal the core transcriptional identities of major CNS cell classes in humans

The genes with the highest expression fidelity for major CNS cell classes are consistently coexpressed across regions and technology platforms (Fig. S4). We visualized the top 50 genes ranked by expression fidelity for each cell class to compare their expression levels, mutation intolerance, literature citations, cellular localization, and protein-protein interactions (PPI) (Fig. 3A-D). Overall, expression levels of high-fidelity genes were highest for neurons and lowest for microglia (Fig. 3A-D, red tracks). However, each cell class had a wide range of expression levels for high-fidelity genes, suggesting parallel regulatory mechanisms and/or differential transcript stability.

To assess the tolerance of high-fidelity genes to loss-of-function (LoF) mutations, we analyzed data from the Exome Aggregation Consortium (ExAC), which summarizes the prevalence of coding mutations in ~61K human exomes[35]. Unexpectedly, high-fidelity neuronal genes were significantly less tolerant to LoF mutations than high-fidelity glial genes (Fig. 3A-D, black tracks). We then searched PubMed to determine whether high-fidelity genes have been studied in their respective cell classes (Fig. 3A-D, green tracks). Interestingly, many searches returned no citations, highlighting critical gaps in our understanding of CNS cell biology. For example, the top microglial gene (amyloid beta precursor protein binding family B member 1 interacting protein, or *APBB1IP*) is unstudied in microglia.

We also examined the cellular localization of proteins[36] encoded by high-fidelity genes. Among those shown in Fig. 3A-D, membrane localization was reported for 33 in astrocytes, 22 in oligodendrocytes, and 30 in microglia, but only 13 in neurons (inside track). This result may reflect the homeostatic functions of glia as regulators of extracellular CNS environments. More generally, the non-random distributions of cellular localizations suggest that high-fidelity genes are expressed as proteins in their corresponding cell classes. Indeed, PPI[37] among high-fidelity gene products for each cell class revealed significantly more interactions than expected by chance (Fig. 3A-D, interior lines).

Because high-fidelity genes should encode optimal biomarkers, we searched for high-fidelity genes in the Human Protein Atlas (http://www.proteinatlas.org) to identify novel reagents for labeling human CNS cell classes. We identified validated antibodies for PON2 (astrocytes), DBNDD2 (oligodendrocytes), APBB1IP (microglia), and CELF2 (neurons) (Fig. 3A-D). Dual immunostaining with canonical markers revealed almost perfect concordance in human frontal cortex (Fig. 3E-H; Fig. S5).

Gene coexpression analysis of intact tissue samples reveals the core transcriptional features of diverse CNS cell classes

Variation among intact tissue samples can also reveal transcriptional features of less abundant human CNS cell classes. Following the general strategy outlined in Fig. 1, we calculated genome-wide expression fidelity for human cholinergic neurons, midbrain dopaminergic neurons, endothelial cells, ependymal cells, choroid plexus cells, mural cells, oligodendrocyte progenitor cells, and Purkinje neurons (Figs. 4, S6; Table S3). This analysis

correctly assigned high fidelity scores for canonical markers of these cells. For example, choline acetyltransferase (*CHAT*), the high-affinity choline transporter (*SLC5A7*), and the vesicular acetylcholine transporter (*VACHT*) all ranked within the top ~0.2% for genome-wide cholinergic neuron expression fidelity, while claudin 5 (*CLDN5*), tyrosine kinase with immunoglobulin like and EGF like domains 1 (*TIE1*), and platelet and endothelial cell adhesion molecule 1 (*PECAM1*) all ranked within the top ~0.3% for genome-wide endothelial cell expression fidelity (Table S3). High-fidelity genes were significantly and specifically enriched with published markers of each cell class from multiple independent studies; furthermore, novel markers predicted by our analysis were validated by *in situ* hybridization in adult mouse brain[38] (Figs. 4, S6).

High-fidelity genes enable predictive modeling of gene expression in transcriptomes from intact tissue samples

The reproducibility of gene coexpression modules corresponding to major cell classes (Table S2, Fig. S4) suggests that transcriptomes of intact CNS samples can be modeled as a function of cellular abundance. We explored this topic systematically by performing multiple linear regression in 47 CNS datasets with ≥40 samples to determine how much expression variation in a shared set of ~9600 genes could be explained by variation in the abundance of neurons, astrocytes, oligodendrocytes, and microglia. To estimate the relative abundance of each cell class in each dataset, we summarized the expression patterns of high-fidelity genes (Fig. 5A). To avoid circularity, we used a leave-one-out cross-validation strategy to redefine high-fidelity genes for each dataset by recalculating expression fidelity for each cell class using the remaining 46 datasets (as in Fig. 1C-G). More aggressive exclusion criteria (e.g. excluding 90% of datasets before redefining high-fidelity genes) produced nearly identical results (Fig. S7). Before modeling, each dataset was downsampled (n=40) to facilitate comparisons of results; this process was performed iteratively to ensure robustness (Fig. 5A).

Implementing this strategy, we obtained several important results (Fig. 5B). First, using only one gene (with the highest fidelity) as a surrogate for each cell class, our models explained 32.2% of total transcriptional variation averaged over all datasets and up to ~50% in some datasets (vs. ~0.1% for permuted data). Second, increasing the number of gene surrogates/cell class (e.g. using the top 10 or top 50 high-fidelity genes) provided only modest performance improvements (unless otherwise stated, subsequent models used the top 10 high-fidelity genes). Third, prediction accuracy depended strongly on technology platform ($p<10^{-7}$, ANOVA) but not CNS region (p=0.92, ANOVA). Among microarrays, older platforms fared substantially worse than newer platforms, while RNA-seq generally outperformed all microarrays.

Despite their simplicity, our models explained >50% of expression variation, averaged over all datasets, for ~2000 genes (Fig. 5C). Over all genes, the average amount of expression variation explained by our models followed a sigmoid function (Fig. 5C). We benchmarked model performance against the maximal explanatory power of any four predictors by using PC1–4 from each dataset as covariates for multiple regression. On average, PC1–4 explained 49.6% of total gene expression variation over all datasets (Fig. 5B). Thus, modeling gene

expression in the human CNS as a function of neuron, astrocyte, oligodendrocyte, and microglia abundance achieved, on average, 72.0% of the maximal explanatory power for all datasets and 80.1% for RNA-seq datasets (Fig. 5B).

We reasoned that model performance for RNA-seq might exceed that for microarrays since the latter have many probes for transcripts unlikely to be expressed in the CNS. We therefore stratified genes by expression levels and examined model performance. As expected, predictive power decreased at lower expression levels, with the sharpest decline between the 3rd and 4th quartiles (Fig. 5D).

We next explored how transcriptional variation related to variation in the abundance of individual cell classes, sex, and age. Neuronal abundance explained more transcriptional variation than glial abundance (Fig. 5E). After controlling for variation in the abundance of major cell classes, model performance did not substantially improve by including sex or age as covariates (Fig. 5E). We further explored this topic by correlating estimated cellular abundance with age in 32 CNS datasets. Neuronal and oligodendroglial abundance were negatively correlated with age, while astrocytic and microglial abundance were positively correlated (Fig. 5F). These results suggest that age-related changes in gene expression in bulk CNS transcriptomes are primarily driven by age-related changes in cellular composition.

Gene expression modeling applications

The ability to predict gene expression in transcriptomes from intact CNS samples has important implications. We illustrate the versatility of this approach through comparative analysis of gene expression models in disease, among CNS regions, and between species.

Application #1: Contextualizing disease genes and modeling gene expression in pathological samples

We asked whether genes associated with CNS diseases[39] are enriched among genes primarily expressed by astrocytes, oligodendrocytes, microglia, or neurons (Fig. 6A-B). Clustering select CNS diseases by enrichment p-values revealed several interesting findings. First, except for ALS, genes associated with neurodegenerative disorders were most enriched among genes expressed by microglia and astrocytes. Second, genes associated with neurodevelopmental disorders, epilepsy, and psychiatric disorders were most enriched among genes expressed by astrocytes and neurons. Third, genes expressed by astrocytes consistently showed the greatest enrichment with candidate CNS disease genes.

Beyond broad associations between diseases and cell classes, gene expression modeling can also reveal which cell classes are most likely to express candidate disease genes. For example, we modeled gene expression for Alzheimer's diseases (AD) risk genes as a function of neuronal, oligodendroglial, astrocytic, and microglial abundance in transcriptomes from intact neurotypical adult human temporal cortex (Fig. 6C). Expression levels of early-onset AD risk genes *APP* and *PSEN1* were mostly explained by variation in neuronal and oligodendroglial abundance, respectively. In contrast, expression levels of late-onset AD risk genes *APOE* and *TREM2* were mostly explained by variation in astrocytic

and microglial abundance, respectively. These results were consistent across 47 CNS datasets (Fig. 6D).

Compared to control (CTRL) human brain samples, AD samples should contain fewer neurons and proportionately more glia. We tested this hypothesis by using expression patterns of high-fidelity genes to infer the relative abundance of neurons, astrocytes, microglia, and oligodendrocytes in three gene expression datasets from intact postmortem brain samples of CTRL and AD subjects[40–42]. We observed a highly significant decrease in neuronal abundance in AD in all datasets (Figs. 6E, S8A-B). In two out of three datasets, there were significant increases in the relative abundance of astrocytes and microglia in AD, with similar trends in the third (Figs. 6E, S8A-B). Interestingly, there were no significant differences in oligodendrocyte abundance between CTRL and AD in any dataset (Figs. 6E, S8A-B). Importantly, simulations revealed that estimates of cellular abundance were robust to a wide range of transcriptional dysregulation among genes used to derive these estimates (Fig. S9). This strategy can help determine whether variable cellular composition is associated with diverse CNS disorders.

Because AD brain samples have fewer neurons and proportionately more astrocytes/microglia than CTRL, differential expression analysis of intact tissue samples will reveal down-regulation of neuronal transcripts and up-regulation of astrocytic/microglial transcripts. However, predictive modeling can identify cell-intrinsic transcriptional differences between CTRL and AD that are independent of changes in cellular composition. This strategy is analogous to that of Kuhn et al.[11], except we use expression patterns of high-fidelity genes to estimate cellular abundance. Surprisingly, after controlling for differences in cellular composition between CTRL and AD, we identified many genes that were consistently up-regulated in AD neurons (Fig. 6F, Table S5). These genes did not include canonical AD risk genes (Fig. S8C), but rather genes involved in protein ubiquitination, catabolism, proteasome degradation, and mitochondrial function (Fig. S8D). Examples are shown in Figs. 6G, S8.

Application #2: Identifying transcriptional differences in major cell classes among CNS regions

We recalculated expression fidelity for each CNS region with ≥3 datasets and clustered each cell class (Fig. 7A-D). Regional differences in expression fidelity were greatest for neurons, with bifurcation between cortical/subcortical structures (Fig. 7D-E). In contrast, oligodendrocyte expression fidelity was very similar among brain regions (Fig. 7B,E). Comparatively, microglia and astrocytes exhibited more regional variation in expression fidelity than oligodendrocytes, but less than neurons (Fig. 7A,C,E).

We developed a conservative strategy to identify binary expression differences in major cell classes among human brain regions (Fig. 7F-G, Table S6). Many genes were predicted to distinguish regional subpopulations of neurons, but we found no evidence for binary expression differences among regional subpopulations of microglia or oligodendrocytes (Figs. 7H, S10). However, we did predict binary expression differences among regional subpopulations of human astrocytes (Fig. 7H-I). For example, *CHRDL1* was predicted to be expressed by astrocytes in frontal cortex and striatum, but not diencephalon or midbrain

(Fig. 7I-K). To validate this prediction, we performed single-molecule fluorescent *in situ* hybridization (FISH) for *Chrdl1* and *Aldh1l1* in mouse cortex and thalamus. *Aldh1l1* is expressed ubiquitously by astrocytes[22] and was detected in both regions (Fig. 7J-L). Expression of *Chrdl1* colocalized with *Aldh1l1* in mouse cortex but not thalamus (Fig. 7L), as predicted.

Application #3: Identifying transcriptional differences in major CNS cell classes between species

We analyzed 1346 mouse brain transcriptomes to determine genome-wide expression fidelity for astrocytes, oligodendrocytes, microglia, and neurons (Tables S1, S7; Fig. S11). Expression fidelity was significantly correlated between mice and humans for each cell class, with the greatest similarity for neurons (Fig. 8A). We note that evolutionary conservation of neuronal expression fidelity relative to glia is mirrored at the protein level (Fig. 3A-D, black tracks). These findings may indicate that neurons are under greater evolutionary constraint than glia.

We applied stringent criteria and identified 50 genes predicted to be ‘on’ in human CNS cell classes and ‘off’ in the corresponding mouse CNS cell classes, as well as six genes with the opposite pattern (Fig. 8B, Table S8). ~85% of these differences were predicted to occur in glia (Fig. 8B). We also analyzed 476 outgroup samples from chimpanzee and macaque brains (Table S1). Of the 50 genes predicted to be expressed in human but not mouse cell classes, 29 were significantly associated with the same cell class in at least one primate dataset; conversely, of the six genes with the opposite pattern, none was significantly associated with the same cell class in any primate dataset (Table S8). For example, expression variation of *MRVI1* was largely explained by astrocyte abundance in primates, but not mice (Figs. 8B, S12A-B). Conversely, expression variation of *PLA2G7* was largely explained by astrocyte abundance in mice, but not primates (Figs. 8B, S12A-B). Single-molecule FISH confirmed that expression of *MRVI1* and *PLA2G7* is specific to human and mouse astrocytes, respectively (Fig. S12C-D).

To demonstrate the ability of our analyses to deliver functional insights into the unique biology of human brains, we focused on the unexplained fact that human astrocytes are much larger than mouse astrocytes (and non-human primate astrocytes)[43]. This phenotype has important implications for neuronal function, since one human astrocyte can encompass ~2MM synapses vs. ~100K synapses for one mouse astrocyte[43]. We reasoned that genes expressed by human but not mouse astrocytes might contribute to this phenotype. We were particularly intrigued by peripheral myelin protein 2 (*PMP2*; Fig. 8B), which encodes a fatty-acid binding protein that maintains membrane lipid composition in Schwann cells[44]. In the human CNS, *PMP2* expression was extremely high (mean percentile: 96.2) and largely explained by astrocyte abundance, while in the mouse CNS *Pmp2* expression was effectively absent (mean percentile: 11.2) and unrelated to astrocyte abundance (Fig. 8B-D). Furthermore, independent RNA-seq data from human, chimpanzee, macaque, and mouse neocortex[45] revealed a monotonic increase in *PMP2* expression from mouse to human (Fig. 8E).

Immunostaining revealed widespread PMP2 in human neocortical astrocytes but no PMP2 in mouse neocortex (Fig. 8F), despite robust expression by Schwann cells (Fig. S12E). To test whether PMP2 could increase mouse astrocyte size *in vivo*, we delivered a viral construct expressing *PMP2* under an astrocyte-specific promoter to neonatal mouse brains and analyzed the morphology of transduced astrocytes after 42d (Fig. 8G). Forced expression of *PMP2* in mouse astrocytes significantly increased their maximum diameter and number of primary processes (Fig. 8H-I). The increase in maximum diameter corresponded to an increase in mouse astrocyte volume of ~50% (assuming sphericity). We repeated this experiment with a different viral construct and obtained nearly identical results (Fig. S12F). To our knowledge, these data provide the first molecular explanation for morphological differences between human and mouse astrocytes. More generally, our findings illustrate how variation among intact tissue samples can predict cell-class-specific transcriptional features with important functional implications for human neurobiology.

DISCUSSION

We have described an approach to reveal the core transcriptional features of cellular identity via integrative gene coexpression analysis of intact tissue samples. Compared to ‘bottom-up’ methods such as FACS, IP, and SC/SN, the main advantages of our ‘top-down’ approach include: i) elimination of the need for fresh tissue; ii) applicability to huge amounts of existing data; iii) elimination of technical variability caused by tissue dissociation/cDNA preamplification; iv) elimination of sampling bias associated with cell/nucleus capture; v) ability to estimate the relative abundance of cell classes among intact tissue samples; and vi) ability to derive highly robust inferences about the core transcriptional features of cellular identity based on aggregate analysis of billions of cells.

Our approach also has important limitations. False-positive associations can result from technical factors (e.g. batch effects) or biological factors such as cellular collinearity. For example, we consistently observed that genes with high expression fidelity for oligodendrocytes had higher expression fidelity for microglia (and vice versa) than they did for astrocytes or neurons. Because oligodendrocytes and microglia are more abundant in white matter than gray matter[46], variation in the ratio of white matter to gray matter in CNS samples drives covariation in the abundance of these cell classes and the genes they express. False-negative associations can result from technical factors such as limitations in dynamic range/transcriptome coverage or probe failures, as well as biological factors like alternative splicing. Notwithstanding these limitations, the genes with the highest expression fidelity for major CNS cell classes are already remarkably stable.

It is interesting to consider the ability of our approach to detect transcriptional signatures of less abundant cell classes (e.g. Figs. 4, S6). The ability to discern the transcriptional signature of a cell class in intact tissue samples depends on many factors, including its representation, the uniqueness and abundance of its transcripts, its stoichiometry with other cell classes, the technology platform, the algorithmic approach, and the sampling strategy[8]. Some of these factors can be optimized to improve sensitivity. Ultimately, we envision combining top-down and bottom-up strategies to fully deconstruct the transcriptional architecture of biological systems.

Gene expression fidelity estimates were highly robust to the choice of gene set used for enrichment analysis, but more so for glia than neurons. This result indicates that neuronal diversity may require additional strategies to optimize estimates of neuronal expression fidelity, particularly on a regional basis. For example, the neuronal gene sets we used do not capture the transcriptional profile of cerebellar granule neurons, which is highly distinct[21, 29, 33]. To better account for neuronal diversity, future studies may combine neuron subtype-specific gene sets for enrichment analyses.

Our results suggest that the functional identity of a cell class can be conceived as a vector of genes ranked by the fidelity with which they are expressed in that cell class relative to all other cells in the biological system of interest. An advantage of this framing is that it is inherently context-dependent. Beyond revealing novel biomarkers and cellular phenotypes, such definitions can provide 'molecular rulers' for measuring the validity of human cells derived *in vitro* for disease modeling/cell-replacement therapies. Furthermore, these definitions can be tested in *de novo* CNS transcriptomes for their ability to predict gene expression through mathematical modeling.

Multivariate analyses of CNS transcriptomes often use module detection/clustering methods or projection methods such as principal component analysis. Although these methods have produced many important insights, they are inherently descriptive and do not facilitate comparisons among independent datasets. Because the building block of any biological system is the cell, and cells are distinguished by the genes they express, an alternative approach is to model gene expression as a function of cellular composition. We have shown how expression patterns of high-fidelity genes can be used for this purpose. The resulting models are highly robust, grounded in biology, easily compared among independent datasets, and capable of extracting cell-class-specific insights from intact tissue samples. Using this approach, we explored how predictive models of gene expression in transcriptomes from intact CNS samples can inform studies of aging, disease genes, pathological samples, regional heterogeneity, and species differences (Supplementary Note).

Our study is based on a simple idea: variation in cellular composition among intact tissue samples will drive covariation of transcripts that are uniquely or predominantly expressed in specific kinds of cells. Although we have focused here on gene expression, our approach can also be applied to other types of molecular data, thereby offering a generalizable strategy for determining the core molecular features of cellular identity in intact biological systems.

ONLINE METHODS

1. Integrative analysis of human CNS transcriptomes

We obtained publicly available gene expression data from eight studies[21, 27–33] and one resource (http://www.brainspan.org/) that profiled large numbers of postmortem CNS bulk tissue samples from neurotypical humans. Expression profiling was performed on six technology platforms, including RNA-seq and various commercial microarrays. Samples from each of the nine sources were separated into 62 datasets representing 19 major neuroanatomical regions. Each regional dataset consisted of at least 25 samples, all of which

came from adults (≥ 18 years). After removing outliers (see below), we analyzed a total of 7221 transcriptomes (Table S1).

1.1 Data preprocessing and quality control—Preprocessing was performed from raw data when possible. Affymetrix microarray raw data (.CEL files) were downloaded from Gene Expression Omnibus (GEO: http://www.ncbi.nlm.nih.gov/geo/) using the following accession IDs: GSE11882[28], GSE25219[29], GSE3790[27], GSE45642[31], and GSE46706[32]. Probe-level data from Affymetrix Exon 1.0 microarrays (GSE25219 and GSE46706) were summarized using the Robust Multi-Array (RMA) algorithm[50] at the gene level with Affymetrix Power Tools software (APT 1.15.2) and reverse log-transformed for further processing.

Affymetrix U133A and U133Plus2 microarray probes from GSE3790, GSE45642, and GSE11882 were pruned to eliminate non-specific and mis-targeted probes using the ProbeFilter R package (Dai et al. 2005) with mask files obtained from http://masker.nci.nih.gov/ev/[51]. After applying the mask files, only probe sets with at least seven remaining probes were retained for further analysis. Expression values were generated in R using the expresso function of the affy R package[52] with “mas” settings and no normalization, followed by scaling of arrays to the same average intensity (200). For GSE45642, gene expression was not scaled and technical replicates were removed (AMY samples restricted to site I; DLPFC samples restricted to site D; HIP and NUAC samples restricted to site M). Non-normalized Illumina microarray data were obtained from GEO for GSE36192[30]. Normalized expression data from GTEx, BrainSpan, and the Allen Brain Institute (ABI) were downloaded from their respective websites (http://www.gtexportal.org/, V6 data release; http://www.brainspan.org/, Oct2013 data release; and http://human.brain-map.org/, March2013 data release). For the RNA-seq datasets (GTEx and BrainSpan), RPKM gene expression values were used.

Sample information for datasets with GEO accession IDs was obtained using the GEOquery R package[53] with the exception of hybridization batch information, which was extracted from the header information of Affymetrix .CEL files when available. Each of the 62 regional datasets was individually processed using the SampleNetwork R function[54], which is designed to identify and remove sample outliers, perform data normalization, and adjust for batch effects[55]. We defined sample outliers as those that were more than four standard deviations below the mean connectivity of all samples measured over all features ($Z.K < -4$). Iterative pruning was applied for each regional dataset to remove all samples with $Z.K < -4$ (Table S1). For non-normalized data (GSE11882, GSE3790, GSE4542, and GSE36192), quantile normalization[56] was then performed. If a batch effect was present (defined as a significant association between the 1st principal component of the expression data and a technical batch covariate), batch correction was performed using the ComBat R function[55], which is implemented in SampleNetwork. For GTEx data, we detected a large batch effect due to center site. We therefore restricted our analysis to samples acquired by centers ‘B1, A1’ or ‘C1, A1’. Lastly, prior to coexpression analysis, probes / genes that had zero variance across all samples were removed. For GTEx, we further restricted our analysis to 27,540 transcripts that were detected (≥ 0.1 RPKM) in at least 200 CNS samples. Table S1 provides additional details on data preprocessing and quality control.

1.2 Unsupervised gene coexpression module detection—Gene expression datasets can exhibit different correlation structures due to biological and technical factors such as cellular heterogeneity and sample size. This variability makes it difficult to apply a single set of parameters for coexpression analysis across many independent datasets. To address this challenge, we analyzed gene coexpression relationships in each regional dataset in the R statistical computing environment (http://cran.us.r-project.org) using a four-step approach[57, 58]. First, pairwise biweight midcorrelations (bicor) were calculated for all possible pairs of probes / genes over all samples in each dataset using the bicor function in the WGCNA R package[25]. Bicor is a robust correlation metric that is less sensitive to outliers than Pearson correlation but often more powerful than Spearman correlation[59, 60]. Second, probes / genes were clustered using the flashClust[25] implementation of a hierarchical clustering procedure with complete linkage and 1 - bicor as a distance measure. Third, the resulting dendrogram was cut at a series of heights corresponding to the top 0.01%, 0.1%, 1%, 2%, 3%, 4%, or 5% of pairwise correlations for the entire dataset. Moreover, for each of these thresholds, we modified the minimum module size to require 8, 10, 12, 15, or 20 members. This approach yielded 7 × 5 = 35 different gene coexpression networks for each regional dataset. Third, initial modules in each network were summarized by their module eigengenes, which is defined as the first principal component obtained by singular value decomposition of the coexpression module[24]. Fourth, highly similar modules were merged if the correlations of their module eigengenes exceeded an arbitrary threshold (0.85). This procedure was performed iteratively for each network such that the pair of modules with the highest correlation (> 0.85) was merged first, followed by recalculation of module eigengenes, followed by recalculation of all correlations, until no pairs of modules exceeded the threshold. The WGCNA measure of intramodular connectivity (k_{ME}) was then calculated for each probe / gene with respect to all modules by correlating its expression pattern across all samples with each module eigengene[16, 24].

1.3 Module enrichment analysis with cell-class-specific gene sets—To identify cell-class-specific gene coexpression modules in each regional dataset, we cross-referenced module composition with gene sets consisting of published markers of major cell classes. For astrocytes, oligodendrocytes, and neurons, we used sets of genes that were expressed >10x higher in each cell class (purified from mouse forebrain) vs. the other two (Tables S4–6 from Cahoy et al.[22]; n=184 astrocyte genes, 130 oligodendrocyte genes, 319 neuron genes). For microglia, we used a set of genes expressed significantly higher in purified mouse microglia vs. whole brain (Table S2 from Hickman et al.[23]; n=99 genes). For mural and ependymal cells, we used gene sets from adult mouse forebrain single-cell RNA-seq data (Table S1 from Zeisel et al.[61]; n=155 mural genes, 484 ependymal genes). For endothelial cells, we used a set of genes that was significantly associated with known endothelial markers across 32 human organs (Table S3, tab 1, from Butler et al.[62]; n=237 genes). For OPCs, we used a set of genes identified in developing human midbrain single-cell RNA-seq data (Table S1 from La Manno et al.[63]; n=48 genes). For choroid plexus cells, we used a set of genes from the Allen Mouse Brain Atlas (Table S1 from Lein et al.[38]; n=101 genes). For Purkinje neurons, we used a set of genes identified from human cerebellar samples (Table S6 from Kuhn et al.[64]; n=80 genes). For dopaminergic and cholinergic neurons, we created sets of genes based on the top 50 genes ranked by correlation to *TH* and

CHAT in our human midbrain and striatum expression datasets, respectively. Modules were defined as all unique genes with positive k_{ME} values that were significant after applying a Bonferroni correction for multiple comparisons ($p < 0.05$ / (# probes or genes in the regional dataset × # of modules in the network)). If a probe / gene was significantly correlated with more than one module it was assigned to the module for which it had the highest k_{ME} value. For each regional dataset, enrichment analysis was performed for all modules in all (n = 35) networks using a one-sided Fisher's exact test as implemented by the fisher.test R function. The module with the most significant enrichment for each cell-class gene set was identified in each regional dataset.

1.4 Data integration and calculation of gene expression fidelity—Operationally, we define the transcriptional 'profile' of a cell class in a given dataset as a list of probes / genes ranked by descending k_{ME} values for the most significant cell-class module. To create 'consensus' transcriptional profiles for each cell class, transcriptional profiles from individual datasets were combined using the following approach. First, datasets that did not contain a module that was significantly enriched with markers of a given cell class after applying a Bonferroni correction for multiple comparisons ($p < 0.05$ / # modules) were excluded (Table S2). To reduce collinearity among endothelial, mural, choroid plexus, and ependymal cell-class signatures, we further pruned regional datasets in which a given cell-class module also showed significant enrichment for any other cell-class module. For dopaminergic neurons, fidelity calculations were restricted to midbrain datasets. Second, probe / gene identifiers from all datasets were mapped to a common identifier (HomoloGene ID data build 68). Third, k_{ME} vectors for each cell-class module in microarray datasets (in which individual transcripts are often targeted by multiple probes) were collapsed to unique identifiers by retaining for each HomoloGene ID the probe with the highest k_{ME} value. Because k_{ME} values are correlation coefficients, they cannot be averaged directly over independent datasets of different sample sizes. Therefore, to aggregate cell-class-specific k_{ME} values for a given HomoloGene ID across regional datasets, we used Fisher's method for combining correlation coefficients from independent datasets[65]. We implemented this method by initially transforming k_{ME} values using the Fisher transformation:

$$Z_{gdc} = \frac{1}{2} ln\left(\frac{1 + k_{ME.gdc}}{1 - k_{ME.gdc}}\right) \quad (1)$$

where g indexes the gene, d indexes the regional dataset, and c indexes the cell class. An average of the resulting z-scores (weighted by sample size) was then determined with the following equation:

$$\bar{Z}_{gc} = \frac{\sum_{d=1}^{D} z_{gdc}(n_d - 3)}{\sum_{d=1}^{D}(n_d - 3)} \quad (2)$$

where n denotes the number of samples in dataset d. The sampling standard deviation of $\bar{z}_{gc}$ is:

$$SD\left(\bar{Z}_{gc}\right) = \sqrt{\frac{1}{\sum_{d=1}^{D}\left(n_d - 3\right)}}. \quad (3)$$

Dividing the 'average' z-scores by the sampling standard deviation yields the genome-wide statistics displayed in Fig. 2, Fig. 4, Fig. S6 and Fig. S11A-D, or *gene expression fidelity*:

$$fidelity = Z_{gc} = \frac{\bar{z}_{gc}}{SD\left(\bar{z}_{gc}\right)}. \quad (4)$$

For interpretability, we also convert $\bar{z}_{gc}$ into an 'average' correlation coefficient by performing the reverse Fisher transformation:

$$\bar{r}_{gc} = \frac{\exp\left(2\bar{z}_{gc}\right) - 1}{\exp\left(2\bar{z}_{gc}\right) - 1}, \quad (5)$$

which is reported as 'Mean.r' along with expression fidelity for all genes with respect to all cell classes for humans (Table S3) and mice (Table S7). It is important to note that gene expression fidelity, as defined here, is robust to the choice of gene set used for enrichment analysis, as illustrated in Fig. S2.

1.5 Calculation of gene expression fidelity in distinct CNS regions—We calculated gene expression fidelity in individual CNS regions represented by at least three independent datasets (FCX, STR, HIP, DI, and MID), as described above. We only included regional datasets that contained a significant cell-class module, as defined above.

1.6 Enrichment analysis of high-fidelity genes—The following gene sets were used to demonstrate that high-fidelity genes were significantly and specifically enriched (one-sided Fisher's exact test) with expected cell-class markers from multiple independent studies: Fig. 4 (A1, O1, M1, N1 [n=100 genes]: Zhang et al.[47]; A2, O2, N2 [n=184, 130, and 319 genes]: Cahoy et al.[22]; M2 [n=99 genes]: Hickman et al.[23]; C1 [n=126 genes]: Doyle et al.[66]; C2 [n=92 genes]: Mancarci et al.[67]; D1, D2 [n=69 and 71 genes]: La Manno et al.[63]; E1 [n=100 genes]: Zhang et al.[47]; E2 [n=237 genes]: Butler et al.[62]; Ep1 [n=484 genes]: Zeisel et al.[61]; Ep2 [n=50 genes]: Gokce et al.[68]); Fig. S6 (Cp1 [n=101 genes]: Lein et al.[38]; Cp2 [n=50 genes]: Gokce et al.[68]; Mu1 [n=155 genes]: Zeisel et al.[61]; Mu2 [n=260 genes]: He et al.[69]; Op1 [n=48 genes]: La Manno et al.[63]; Op2 [n=100 genes]: Zhang et al.[47]; P1 [n=80 genes]: Kuhn et al.[64]; P2 [n=43 genes]: Mancarci et al.[67]); Fig. 2 (A1, O1, M1, N1 [n=100 genes]: Zhang et al.[47]; A2, O2, N2 [n=184, 130, 319 genes]: Cahoy et al.[22]; M2

[n=99 genes]: Hickman et al.[23]; A3, O3, N3 [n=38, 73, 67 genes]: Lein et al.[38]; M3 [n=152 genes]: Butovsky et al.[48]).

1.7 Visualization of core cell-class genes and their characteristics—The top 50 genes ranked by expression fidelity for each cell class were visualized with custom software using the ggplot2 package[70] in R (Fig. 3A-D, Fig. S11E-H). Mean gene expression levels were calculated for all samples in each regional dataset, converted to percentile ranks, and averaged across datasets ('CNS expression'). Regional percentile ranks (used in the analysis presented in Fig. 7) were similarly calculated for all samples from a given CNS region using ABI, GSE46706, and GTEx datasets. PubMed citations were extracted for each gene symbol using the RISmed package in R on 10/27/2016. To obtain total citations with potential relevance to specific cell classes we queried '[gene symbol]' and 'cell class' (e.g. 'neuron'). Cellular localization data were extracted from the COMPARTMENTS resource[36]. Human protein-protein interaction (PPI) data were downloaded from the STRING database (file: 9606.protein.links.detailed.v10.txt)[37]. PPI links in Fig. 3 represent STRING combined score values > 350, which correspond to medium-confidence PPI predictions. Empirical p-values for over-representation of putative protein interactions (PPI score > 350) among the top 50 cell-class genes were calculated from n = 100,000 permutations of 50 randomly sampled genes from a background of 18,451 HomoloGene identifiers (IDs present in at least one regional dataset). Visualizations were similarly made for the top 50 mouse cell-class genes ranked by expression fidelity (Fig. S11) except PubMed citations were extracted on 12/28/2016 and the following STRING database file was used: 10090.protein.links.detailed.v10.txt.

1.8 Figure schematic licensing information—Human brain (Figs. 1A,C and 7G) and cell-class schematics (Figs. 1A,G, and 8B) were obtained from Servier Medical Art under a Creative Commons Attribution 3.0 Unported License (https://creativecommons.org/licenses/by/3.0/legalcode). Color and dimensions were modified.

2. Single-cell (SC) / single-nucleus (SN) RNA-sequencing analysis

We obtained publicly available SC and SN RNA-seq data from four studies of adult human and mouse brains[1, 4, 26, 34]. Count data and sample information were downloaded from GEO using the accession IDs provided in Table S1. The SC dataset from Darmanis et al.[1] consisted of 332 cells isolated from temporal cortices of eight individuals undergoing surgery for epilepsy, while the SC dataset from Tasic et al.[26] consisted of 1375 'core' cells isolated from primary visual cortices of various transgenic mice (see Supplementary Data Table 3 from Tasic et al.[26]). The SN dataset from Lake et al. consisted of 35289 nuclei from postmortem samples of cortex and cerebellum from six individuals, while the SN dataset from Habib et al.[4] (which was obtained from https://www.gtexportal.org/home/datasets) consisted of 14963 nuclei from postmortem samples of cortex and hippocampus from five individuals. In all cases, we used the authors' original classifications to categorize cells.

2.1 Cell-class module eigengenes predict relative cellular abundance—To test the ability of cell-class module eigengenes to predict variation in cellular abundance, we analyzed gene coexpression relationships in synthetic mixtures of single-cell RNA-seq data

from human[1] and mouse[26] brain. For the human data, counts were normalized by dividing the total number of reads for each gene by the total number of reads for the sample. Normalized data were then multiplied by one million to yield counts per million (CPM), added by 1, and $\log_{10}$-transformed. For the mouse data, RPKM was used. We created synthetic mixtures by summing single-cell expression data over all genes from randomly sampled combinations of astrocytes, oligodendrocytes, microglia, and neurons. Each synthetic sample consisted of 16 or 22 total cells for the human and mouse data, respectively, which was based on the minimum number of unique cells that were available for any cell class (in both cases, microglia). We generated 100 synthetic samples and performed gene coexpression analysis and cell-class module detection as described above. We chose a sample size of n=100 for the synthetic tissue sample coexpression analysis to best match the sample sizes of our acquired human regional datasets. We repeated this process 10 times. For each iteration, estimates of relative cellular abundance were determined from cell-class module eigengenes (i.e. PC1 of the most significant cell-class modules).

In principle, the ability of cell-class module eigengenes to estimate relative cellular abundance across conditions (e.g. control samples versus Alzheimer's disease samples) could be confounded by concordant changes in the expression levels of high-fidelity genes that are used to estimate cellular abundance. For example, given two conditions, if the expression levels of high-fidelity genes uniformly increase in one condition but not the other it could appear that the relative cellular abundance had changed when it had not. To estimate the fraction of high-fidelity genes that would need to concordantly change to confound estimates of cellular abundance, we performed a simulation study of synthetic samples. Specifically, we created a 'control' dataset and a 'condition' dataset, each consisting of 100 synthetic samples, by aggregating single-cell RNA-seq data from human brain[1] as described above. To construct module eigengenes, we identified the top 10 astrocyte, oligodendrocyte, microglia, and neuron genes via differential expression analysis (see below) that were also expressed in at least 50% of cells in their respective cell class. We systematically perturbed gene expression in the 'condition' dataset by increasing the expression of randomly selected sets of genes (i.e. by selecting 1 to 10 out of the top 10 genes to perturb) by 100-, 50-, 25-, 10-, 5-, or 2-fold for each cell class of interest. We then combined the synthetic condition data with the synthetic control data and calculated module eigengenes. We repeated this analysis a total of 10 times. We then compared these perturbed module eigengenes to the actual cellular abundance. As shown in Fig. S9, these perturbations had little effect on cell-class composition estimates across a wide range of expression level changes as long as at least half of the genes used to calculate the eigengenes remain unaltered.

2.2 Gene coexpression analysis predicts differential expression among CNS cell classes—To test the relationship between cell-class module membership (k_{ME}) and differential expression, we calculated differential expression statistics from SC RNA-seq data using Monocle v2.2.0[71]. Restricting our analysis to exactly the same cells that were used to construct each synthetic dataset (described above), we calculated genome-wide differential expression statistics for each cell class compared to all other cell classes. Because the Monocle output statistics do not indicate directionality (i.e. up- or down-

regulation), we ordered genes based on p-value and fold-change. Therefore, genes with the highest differential expression percentile are the most significantly up-regulated in a given cell class, and vice versa. Figs. 1, S1 compare k_{ME} values from synthetic cell-class modules with differential expression percentiles for the corresponding cell class (averages of 10 iterations).

2.3 Comparing gene expression fidelity with adult human SN RNA-seq data

—To examine the relationship between estimates of gene expression fidelity from intact tissue samples and estimates of differential expression among single cells, we calculated differential expression statistics from SN RNA-seq data for astrocytes, oligodendrocyte, microglia, and neurons, using the cell assignments provided by the authors of the original studies[4, 34]. We used count data as input and calculated the significance of differential expression using Monocle v2.2.0[71], ordering genes based on p-value and fold-change, as described above. Datasets were merged using HGNC gene symbol identifiers. To compare discordant results between fidelity and SN RNA-seq with orthogonal data (Fig. 2N-Q), we analyzed RNA-seq data for cell classes that were purified by immunopanning from human temporal lobe surgical resections[3] (FPKM data from GEO; see Table S1). For each cell class (astrocytes, microglia, neurons, and mature [i.e. not O4+] oligodendrocytes), we averaged the expression values across replicate samples. To calculate the 'proportion of transcripts' for a given gene and cell class (Fig. 2N-Q), we divided the average expression value for that gene in a particular cell class by the sum of its average expression values for all cell classes.

3. Modeling gene expression in the human CNS

3.1 Modeling gene expression in the neurotypical CNS—

To determine how much of the variation in human CNS transcriptomes can be explained by variation in the abundance of major cell classes, we performed multiple linear regression of gene expression (γ_s) in each regional dataset composed of s samples using the lm function in R:

$$\gamma_s = \beta_o + \sum_{c=1}^{C} \beta_c a_{cs} + \varepsilon_s \quad (6)$$

where a_{cs} represents the inferred abundance in sample s of cell class c (where c = astrocyte, oligodendrocyte, microglia, or neuron), β_c is the regression coefficient for cell class c, β_0 is a constant, and a_{cs} is the error term. a_{cs} was estimated for each cell class using high-fidelity genes. Specifically, we calculated the 1st principal component (PC1) of the expression matrix for the top x high-fidelity genes (where x = 10 unless otherwise stated) over all samples in each regional dataset. PC1 was obtained through singular value decomposition of the scaled expression data using the svd (nu=1, nv=1) function in R, which is identical to the module eigengene metric[24]. When obtained in this fashion, PC1 of high-fidelity genes is a good predictor that is robust even when x is small (Fig. 5B).

To standardize modeling across datasets, we restricted our analysis to regional human CNS datasets with at least 40 samples (n = 47, Table S1). We restricted modeling to datasets with 40 samples, to provide ~85% power to identify large effect sizes ($f^2 > 0.4$, $P < .05$) when testing the hypothesis that the proportion of gene expression variance explained by variation

in the abundance of four cell types is zero. We performed gene expression modeling for ~9600 genes that mapped to HomoloGene IDs that were present in all 47 datasets. If multiple probes mapped to a given HomoloGene ID, we retained the probe with the highest mean expression in the subset expression data. For each dataset, we randomly selected 10 sets of 40 samples for gene expression modeling and averaged the adjusted r^2 values from the ~9600 gene expression models for the 10 random subsets. We report adjusted r^2 values because they facilitate comparisons of models with different numbers of parameters[72]. To avoid circularity in gene expression modeling, we redefined consensus cell-class signatures using a leave-one-out approach. Specifically, we calculated gene expression fidelity for each cell class as described below using 46 datasets, then used the resulting high-fidelity genes to infer cellular abundance and model gene expression in the 47th dataset (as illustrated in Fig. 5A–C). This procedure was performed iteratively for all 47 datasets. In all cases, genes that were used to infer cellular abundance were excluded from gene expression modeling. For example, if the top 10 high-fidelity genes were used to infer cellular abundance, those 10 genes were excluded from modeling to avoid circularity. Therefore, the total number of modeled genes depends on the total number of genes used to infer cellular abundance (i.e. 1, 10, or 50 genes, as illustrated in Fig. 5B).

To further demonstrate the robustness of gene expression modeling as a function of inferred cellular abundance, we replicated the results shown in Fig. 5B after redefining consensus cell-class signatures using a leave-X-out approach, where X corresponded to 10, 20, 40, 60, 80, or 90 percent of datasets, which were sampled at random (n = 10 iterations per value of X). Subsequent analysis and modeling was performed as described above. As shown in Fig. S7, gene expression modeling results were nearly identical for all values of X. These findings highlight the robust nature of gene expression fidelity, the resulting estimates of variation in cellular abundance, and the ensuing predictions of gene expression levels as a function of cellular composition in intact tissue samples. Furthermore, we did not observe consistent collinearity of estimated cellular abundance across datasets with the exception of oligodendrocytes and microglia, which tended to exhibit moderate positive correlations (~0.4) due to the increased concentration of these cells in white matter versus gray matter[46]. However, this relationship did not preclude the identification of high-fidelity genes for each cell class.

3.2 Modeling gene expression in Alzheimer's disease (AD) brain samples—

Publicly available gene expression data from AD and control (CTRL) samples from three additional datasets (GSE48350, GSE44770, and GSE36980)[40–42] were downloaded from GEO and preprocessed using the SampleNetwork R function[54] to identify sample outliers. We defined sample outliers as those that were more than three standard deviations below the mean connectivity of all samples measured over all features ($Z.K < -3$). Iterative pruning was applied until all samples with $Z.K < -3$ were removed.

For GSE44770 gene expression data, which was acquired with a Rosetta Human 44k microarray platform, we focused our analysis on the dorsolateral prefrontal cortex (PFC) samples, which were found in the original publication to have more transcriptional changes associated with AD compared to visual cortex and cerebellum[41]. Starting with the raw CY5 probe intensity values, we performed imputation of missing values[73], quantile

normalization[56], and batch correction[55]. We removed 20 sample outliers out of 230 PFC samples. Batch correction was performed using batch information provided in the original study.

For GSE48350 gene expression data, which was acquired with an Affymetrix U133 plus 2.0 microarray platform, normalization was performed using the RMA algorithm[50] with the justRMA function of the affy R package and reverse log-transformed for further processing. We removed 4 sample outliers out of 253 samples. Batch correction was performed[55] using the hybridization batch date extracted from the .CEL file header information.

For GSE36980 gene expression data, which was acquired with an Affymetrix Gene 1.0 ST microarray platform, normalization was performed using the RMA algorithm[50] with the justRMA function of the affy R package[52] and reverse log-transformed for further processing. Due to small sample size (n = 79 samples), we did not remove any sample outliers.

To estimate relative cellular abundance in CTRL and AD samples (Fig. 6E, Fig. S8A,B) we calculated module eigengenes (PC1 obtained by singular value decomposition) using microarray probes that matched our consensus top 10 high-fidelity genes for each cell class. When multiple probes matched a high-fidelity gene, we retained the probe with the highest correlation to other high-fidelity genes. Module eigengenes were calculated from the combined expression data of CTRL and AD samples, which were jointly normalized as described above.

We performed gene expression modeling in these datasets for every probe using the top 10 high-fidelity genes for each cell class as described above. For transcripts of high-fidelity genes that were targeted by multiple microarray probes, we retained the probe with the highest adjusted r^2 value when modeled as a function of the inferred relative abundance of astrocytes, oligodendrocytes, microglia, and neurons.

Gene expression modeling was performed separately for CTRL and AD samples. We ensured that the number of samples and regions sampled were matched between AD and CTRL samples. For GSE44770, which contained more AD samples than CTRL samples, we took a random subset of AD samples to match the number of CTRL samples (95 out of 112). For GSE48350, we first subset the CTRL samples to donors ≥ 70 years old in order to match the ages of the AD and CTRL samples, which gave 71 CTRL and 79 AD samples. We subsequently took a random subset of the AD samples, attempting to match the region of the samples (16 out of 24 postcentral gyrus samples) which gave 14 CTRL and 15 AD entorhinal cortex samples, 20 CTRL and 19 AD hippocampus samples, 17 CTRL and 16 AD postcentral gyrus samples, and 20 CTRL and 21 AD superior frontal gyrus samples. For GSE36980, we first subset the CTRL samples to donors ≥ 75 years old in order to match the ages of the AD and CTRL samples, which gave 35 CTRL and 32 AD samples. We subsequently took a random subset of the CTRL samples, attempting to match the regions of the samples (10 out of 12 temporal cortex samples) which gave 15 CTRL and 15 AD frontal cortex samples, 7 CTRL and 7 AD hippocampus samples, and 10 CTRL and 10 AD temporal cortex samples.

To identify genes that were differentially expressed between CTRL and AD after controlling for variation in cellular abundance, we calculated the differences in *t*-statistics for cell-class model coefficients in CTRL and AD datasets. We then randomly split each dataset (using the sample subsets above) into two equal-sized groups of samples (where each half included CTRL and AD samples) and performed gene expression modeling for each half (n = 1000 permutations). For each permutation, we calculated the differences in *t*-statistics for cell-class model coefficients of the two randomly separated groups, which produced a null distribution of *t*-statistic differences for each cell-class parameter and each gene. Using the null distributions for each gene, we calculated empirical p-values for the measured *t*-statistic differences between CTRL and AD gene models. Genes were considered differentially expressed in a given cell class between CTRL and AD if the resulting p-values were < 0.05 in each of the three independent datasets.

We note that in principle, the interpretation of modeling results can be confounded by concordant changes in cellular abundance and the expression levels of high-fidelity genes that are used to estimate cellular abundance. Simulation studies suggest that such a scenario is unlikely to occur unless more than half of the genes used to calculate an eigengene are transcriptionally dysregulated (Fig. S9). In practice, this possibility can be mitigated by comparing estimates of cellular abundance derived from different groups of high-fidelity genes, or by including a sufficiently large number of high-fidelity genes (e.g. 50) such that up- or down-regulation of any individual gene (or several genes) is unlikely to have an effect on estimates of cellular abundance.

Below we present example R code for resolving changes in cellular composition from changes in cell-class-specific transcriptional identity:

```
#Example illustrating how to identify cell-class-specific expression changes
in intact tissue

#samples while controlling for differences in cellular composition:

#Load expression data. In this example rows are genes and columns are
samples. Control samples are

#the first 100 columns (1:100) and condition samples are the second 100
columns (101:200):

datExpr=read.table(“data.csv”)

#Create cell-class module eigengenes (i.e. relative cellular abundance
estimates).

#Subset to cell-class genes of interest with a subset vector:

cell.expr=datExpr[subset,]

#Standardize expression values for each gene:
```

```
cell.expr=t(scale(t(cell.expr)))

#Calculate module eigengene (ME, aka relative cellular abundance) using
singular value

#decomposition:

ME=svd(cell.expr,nu=1,nv=1)$v[,1]

#To ensure positive module eigengene values correspond to higher expression
levels,

#we assess the correlation with ME and gene expression of a cell-class gene:

if (cor(NME,as.numeric(datExpr[is.element(rownames(datExpr),subset[1]),]))<0)
{ME=−1*ME}

#To determine if relative cellular abundance is significantly altered
between the conditions:

Condition=c(rep(“ctrl”,100),rep(“cond”,100))

wilcox.test(ME~Condition,data=tmp)$p.value

#To determine if a gene is differentially expressed between sample cohorts
after controlling for

#differences in relative cellular abundance, we perform linear modeling in
each condition with ME

#as the predictor:

#Recalculate MEs separately for each condition:

ME.ctrl=svd(t(scale(t(datExpr[is.element(rownames(datExpr),subset),1:100]))))
$v[,1]

ME.cond=svd(t(scale(t(datExpr[is.element(rownames(datExpr),subset),
101:200]))))$v[,1]

#Subset to gene of interest for each condition:

ctrl=datExpr[gene.subset,1:100]

cond=datExpr[gene.subset,101:200]

#Calculate modeling t-value for each condition:
```

```
ctrl.tvalue=summary(lm(as.numeric(ctrl)~ME.ctrl))$coef[2,3]

cond.tvalue=summary(lm(as.numeric(cond)~ME.cond))$coef[2,3]

#To determine the significance of the t-value difference between condition
and control,

#one can perform a permutation analysis of randomly sampled ‘control’ and
‘condition’ datasets,

#using example code above, to calculate an empirical p-value.
```

3.3 Modeling gene expression to identify regional differences in transcriptional identities of major cell classes—

To explore the regional heterogeneity of gene expression in major CNS cell classes, we adopted a conservative strategy to identify genes with binary expression patterns (i.e. ON in a given cell class in one region but OFF in the same cell class in another region). To reduce technical confounds, we restricted our analysis to pairwise comparisons of five brain regions (DI, FCX, HIP, MID, and STR) that were transcriptionally profiled in three independent studies (GTEx, GSE46707, and ABI). For each of these regional datasets (n = 15), we performed gene expression modeling as described above for all genes and samples using the top 10 high-fidelity genes to infer the relative abundance of each cell class. A gene was considered to be ‘regionally expressed’ in a specific cell class if it met the following criteria: i) it was significantly associated with the cell class in region 1 ($p < 2.67\times10^{-8}$, corresponding to a Bonferroni corrected p-value for the cell-class model coefficient based on the total number of gene models: 0.05/([4 cell classes]×[5 regions] × [48,170 ABI probes + 17,868 GSE46706 genes + 27,526 GTEx genes])); ii) it was not significantly associated with the same cell class in region 2; iii) it was differentially expressed between region 1 and region 2 (i.e. the mean expression percentile rank was >20 percentile ranks higher in region 1 vs. region 2; and iv) the preceding criteria were replicated in all three studies.

4. Analysis of CNS transcriptomes from non-human species

We obtained publicly available gene expression data from intact tissue samples of mouse, rhesus macaque, and chimpanzee brains (Table S1). Data preprocessing is described below.

4.1 Mouse brain expression data preprocessing, integration, and modeling—

Mouse gene expression data were obtained from 22 studies[17, 22, 47, 66, 74–91] and one resource (http://www.genenetwork.org). Preprocessing was performed from raw data when possible. Affymetrix microarray raw data (.CEL files) were downloaded from GEO using the accession IDs provided in Table S1. We only processed wild-type samples from studies with multiple conditions. Expression values for Affymetrix 430A, 430 2.0, and 430A 2.0 microarray probes were generated in R using the expresso function of the affy R package[52] with “mas” settings and no normalization. Probe-level data from Affymetrix Exon 1.0 and Gene 1.1 microarrays were summarized using the Robust Multi-Array (RMA) algorithm[50] at the gene level with Affymetrix Power Tools software (APT 1.15.2) and reverse log-transformed for further processing. Non-normalized Illumina microarray data were obtained

from GEO. For RNA-seq datasets, FPKM gene expression values (Ms.Barres, Ms.Fertuzinhos, Ms.GSE63078) or count data (Ms.GSE60312, Ms.GSE62669) were downloaded from GEO or the study's website (Table S1). We normalized raw count data from Ms.GSE60312 by dividing the total number of counts for each gene by the total number of counts for the sample.

Sample information for datasets with GEO accession IDs was obtained using the R package GEOquery[53] with the exception of hybridization batch information, which was extracted from the header information of Affymetrix .CEL files when available. Each of the mouse datasets was individually processed using the SampleNetwork R function[54], which is designed to identify and remove sample outliers, perform data normalization, and adjust for batch effects[55]. We defined sample outliers as those that were more than four standard deviations below the mean connectivity of all samples measured over all features ($Z.K < -4$). Iterative pruning was applied to remove all samples with $Z.K < -4$. For non-normalized Affymetrix 430A, 430 2.0, 430A 2.0, and Illumina microarray data, quantile normalization[56] was then performed. If a batch effect was present (defined as a significant association between the 1st principal component of the expression data and a technical batch covariate), batch correction was performed using the ComBat R function[55], which is implemented in SampleNetwork. See Table S1 for the number of sample outliers removed and datasets that were batch corrected.

Mouse coexpression analysis, enrichment analysis to identify cell-class modules, and data integration to determine consensus gene expression fidelity for major cell classes was performed as described for human data, with one exception. Due to heterogeneity introduced by variable sample preparation methods for mouse datasets (Table S1), we iteratively pruned outlier datasets before consensus fidelity calculations. Specifically, for each cell class we performed hierarchical clustering of genome-wide k_{ME} values for all datasets using 1 – Pearson correlation as the distance measure with average linkage. Datasets with distance > 0.9 were iteratively pruned. Datasets included in the consensus fidelity calculation are indicated in Table S1. Gene expression modeling was performed as described above using the top 10 genes from each mouse consensus expression fidelity cell-class list.

4.2 Human vs. mouse comparison—We implemented a conservative strategy to identify cell-class-specific gene expression differences between humans and mice (i.e. ON in a given cell class in humans but OFF in the same cell class in mice, or vice versa). To identify genes expressed in human but not mouse cell classes, we started with the sets of genes represented by the Venn diagram in Fig. 6A (i.e. genes that were significantly associated with the same cell class in a majority of human regional datasets using a genome-wide, Bonferroni-corrected p-value). We imposed three criteria to predict species differences: i) the gene was consistently well modeled with respect to the same cell class in humans (human median adjusted r^2 values ≥ 0.4); ii) the gene was not well modeled with respect to the same cell class in mice (mouse median adjusted r^2 values ≤ 0.05); iii) the gene was expressed substantially higher in humans vs. mice (mean expression percentile difference > 30).

To identify genes expressed in mouse but not human cell classes, we started with mouse genes that were significantly associated with the same cell class in at least three independent datasets (using the same definition of 'significant association' that was used for humans). We imposed three criteria to predict species differences: i) the gene was well modeled with respect to the same cell class in mice (mouse median adjusted r^2 values ≥ 0.4); ii) the gene was not well modeled with respect to the same cell class in humans (human median adjusted r^2 values ≤ 0.05); iii) the gene was expressed substantially higher in mice vs. humans (mean expression percentile difference > 30).

4.3 Primate brain expression data preprocessing and modeling—Affymetrix gene expression data from chimpanzee cerebral cortex were obtained from eight studies. Four studies analyzed samples using Affymetrix U95A/v2 microarrays[92–95] and four studies analyzed samples using Affymetrix U133Plus2 microarrays[96–99]. Because probes on these arrays were designed from human sequences, we created a custom mask file to exclude probes that did not have perfect alignment to the chimpanzee genome (panTro3: http://hgdownload.cse.ucsc.edu/goldenPath/panTro3/bigZips/panTro3.fa.gz) and chimpanzee RefSeq mRNAs (http://hgdownload.cse.ucsc.edu/goldenPath/panTro3/bigZips/refMrna.fa.gz). After excluding all unknown, random, haplotype, and mitochondrial sequences, the chimpanzee genome was concatenated into a single file. This file, along with the RefSeq mRNA file, was formatted for BLAST from the command line using the formatdb function from the ncbi_tools package (installed via MacPorts). Using a local BLAST installation, we retained a probe if it aligned perfectly to both the chimpanzee genome (e-value = 2.0e-05) and chimpanzee RefSeq mRNAs (e-value = 2.0e-08). We also excluded probes that were identified as mis-targeted or non-specific with respect to human sequences (http://masker.nci.nih.gov/ev) based on a previous re-annotation study[51]. The resulting mask files (one for each microarray platform) were used with the ProbeFilter R package[100] to exclude probes without perfect and specific alignment to chimpanzee sequences.

Expression data were generated from masked .CEL files using the expresso function from the affy package with 'mas' settings and no normalization, followed by scaling each sample to the same mean intensity (200). Only probe sets with at least half of their probes remaining after mask application were retained for further analysis (n = 9,178 [U95A/v2]; n = 35,754 [U133Plus2]). Further quality control and preprocessing was performed with the SampleNetwork R function[54]. Samples from each study were examined separately and no outliers were evident. After combining all samples for a given platform, data were quantile normalized[56] and batch-corrected for hybridization date using the ComBat R function[55]. The final, processed datasets consisted of 30 samples (U95A/v2) and 26 samples (U133Plus2) from chimpanzee cerebral cortex. RNA-seq RPKM gene expression data[45] from chimpanzee brain were downloaded from GEO (GSE49379). Further quality control and preprocessing was performed with the SampleNetwork R function[54]. No outliers were evident.

Affymetrix Macaque genome array normalized expression data [20, 101] from ABI were downloaded from http://www.blueprintnhpatlas.org (2014-03-06 data release). Further quality control and preprocessing was performed with the SampleNetwork R function[54]. We defined sample outliers as those that were more than four standard deviations below the

mean connectivity of all samples ($Z.K < -4$) measured over all features. Iterative pruning was applied to remove all samples with $Z.K < -4$. See Table S1 for the number of sample outliers removed.

Due to the limited availability of gene expression data from non-human primate (NHP) brains, we did not attempt to calculate consensus gene expression fidelity statistics. Instead, we used a semi-supervised approach to identify top cell-class biomarkers for gene expression modeling in NHP brain samples. Parsimony and similarities between human and mouse gene expression fidelity suggest that there is likely to be strong conservation of high-fidelity genes between human and NHP CNS cell classes. To determine the microarray probes that best matched our high-fidelity human genes in NHP expression data, we took a three-step approach: i) we identified NHP probes that targeted the top 50 high-fidelity genes for each human cell class; ii) we calculated the Pearson correlations among these probes over all NHP samples for each dataset; iii) we identified the top 10 probes with the highest average correlations to the others that mapped to unique gene symbols. These top 10 probes (genes) were summarized by PC1 through singular value decomposition of the scaled expression data using the svd (nu=1, nv=1) function in R and used for estimating the relative abundance of each NHP cell class in gene expression modeling.

5. Gene Ontology and pathway enrichment analysis

The ToppGene (https://toppgene.cchmc.org/) suite[102] contains an extensive list of databases and was used to calculate enrichment p-values from hypergeometric tests corrected for multiple comparisons. Specifically, we used the ToppFun application with default parameters and report Benjamini-Hochberg adjusted p-values. We present data from Gene Ontology (GO) annotations (biological process, cellular component and molecular function) and pathway annotations (Biosystems, BIOCYC, KEGG, and REACTOME).

For disease annotations (Fig. 6B), we used the Phenopedia database[39], which is a curated collection of records retrieved weekly from an automatic literature screening program[103] of PubMed abstracts for gene-disease associations. Staff at NCBI manually select abstracts that meet inclusion criteria. We analyzed enrichment of our human cell-class genes (Fig. 6A) with disease annotations consisting of ≥ 10 genes using a one-sided Fisher's exact test as implemented in the fisher.test R function. These p-values were corrected for multiple comparisons by controlling for the false-discovery rate, as implemented by the R package qvalue[49]. Hierarchical clustering of the log-transformed p-values was performed using Pearson's correlation as the similarity metric and average linkage. Phenopedia data were accessed on 12.14.15.

6. Immunohistochemistry

Human brain tissue was collected during autopsy with postmortem interval < 48 hours. Tissue was collected with previous patient consent in strict observance of legal and institutional ethical regulations in accordance with the University of California San Francisco Committee on Human Research. Brains were cut into ~1.5 cm coronal or sagittal blocks, fixed in 4% paraformaldehyde for 2 days, cryoprotected in a 30% sucrose solution, and embedded in optimal cutting temperature (OCT) compound (Tissue-Tek). Samples

contained no evidence of brain disease as assessed by a neuropathologist (Eric J. Huang). 14μm cryosections were collected on superfrost slides (VWR) using a Leica CM3050S cryostat. Cryosections were subjected to heat-induced antigen retrieval in 10mM sodium citrate (pH = 6) for 10 min and permeabilized and blocked for 1 hour at room temperature in PBS supplemented with 0.2% Triton X-100 and 10% goat serum. Primary incubations were overnight at 4 °C. Washes (3 × 10min) and secondary incubations (1 hour) were performed at room temperature.

We searched for high-fidelity genes in the Human Protein Atlas[104] and identified validated antibodies for PON2 (astrocytes), DBNDD2 (oligodendrocytes), APBB1IP (microglia), and CELF2 (neurons) (Fig. 3A-D). Antibodies used included goat PON2 (R&D systems: AF4344, 1:200), mouse ALDH1L1 (Neuromab: 73-140, 1:200), rabbit DBNDD2 (Sigma: HPA043991, 1:200), rabbit APBB1IP (Sigma: HPA017009, 1:100), rabbit CELF2 (Sigma: HPA035813, 1:200), rat GFAP (Fisher: 13-0300, 1:500), mouse NogoA (11c7, gift from M. Schwab, Zurich, Switzerland, 1:5000), goat AIF1 (Abcam: ab5076, 1:500), chicken NeuN (Millipore: ABN91, 1:500), rabbit PMP2 (Proteintech: 12717-1-AP, 1:100), chicken V5 (Abcam: AB9113, 1:500), chicken RFP (Rockland: 600-901-379S, 1:1000), rabbit DsRed (Clontech: 632496, 1:500), and chicken GFP (Aves: GFP-1020, 1:500).

Images were acquired on a Leica TCS SPE laser confocal microscope with detection settings normalized to a secondary-only control. For IHC data (Fig. 3E-H; Fig. S5), a 20× objective was used (1024×1024 pixels).

7. Fluorescent *in situ* hybridization (FISH)

Human brain tissue was acquired as described above. All mouse strains were maintained in the University of California San Francisco specific pathogen–free animal facility, and all animal protocols were approved by and in accordance with the guidelines established by the Institutional Animal Care and Use Committee and Laboratory Animal Resource Center. Mouse brain tissue (postnatal day 30) was acquired from animals perfused with 4% PFA and post-fixed for 24 hours followed by cryoprotection in 30% sucrose and embedded in OCT. Cryosections were prepared in the same manner described above.

Due to variability in RNA quality from human brain cases, we screened a number of samples for robust positive RNA signal using the RNAscope 2.0 HD brown assay (Advanced Cell Diagnostics, Hayward, CA; catalog #: 310033) for PPIB, a positive control. A sample from occipital cortex of a 19 month-old case demonstrated positive signal and was subsequently used for RNAscope multiplex fluorescent imaging (catalog #: 320851) of candidate species differences in astrocyte expression (Fig. S12 C,D), according to manufacturer's instructions. RNAscope probe information is provided below.

Images were acquired on a Leica TCS SPE laser confocal microscope with detection settings normalized to a negative control probe (DapB, catalog: 310043). A 40x objective was used (1024×1024 pixels). Images shown in Figs. 7, S12 were processed using the spots function (with default parameters) in the imaging software program Imaris (Bitplane).

RNAscope single-molecule FISH probe information

Gene	Species	Product	Catalog	Entrez ID	Accession No.
ALDH1L1	Human	Probe- Hs-ALDH1L1-C2	438881-C2	10840	NM_001270364.1
Aldh1l1	Mouse	Probe- Mm-Aldh1l1-C2	405891-C2	107747	NM_027406.1
Chrdl1	Mouse	Probe- Mm-Chrdl1	442811	83453	NM_001114385.1
MRVI1	Human	Probe- Hs-MRVI1	453841	10335	NM_001098579.2
Mrvi1	Mouse	Probe- Mm-Mrvi1	453821	17540	NM_010826.5
PLA2G7	Human	Probe- Hs-PLA2G7	453831	7941	NM_005084.3
Pla2g7	Mouse	Probe- Mm-Pla2g7	453811	27226	NM_013737.5
PPIB	Human	Probe- Hs-PPIB	313901	5479	NM_000942.2
Ppib	Mouse	Probe- Ms-Ppib	313911	19035	NM_011149.2

8. *PMP2* ectopic expression

All mice were maintained in the University of California San Francisco specific pathogen–free animal facility, and all animal protocols were approved by and in accordance with the guidelines established by the Institutional Animal Care and Use Committee and Laboratory Animal Resource Center. Swiss Webster mice were used in all viral experiments (Simonsen Laboratories, Gilroy, CA). For AAV-PMP2 ectopic expression, two females and two males were used in each group. For lentiviral-PMP2 ectopic expression, three females and one male were used in each group. Otherwise, male mice were used. Animals were randomly allocated into experimental groups. No statistical methods were used to pre-determine sample sizes but our sample sizes are similar to those reported in previous publications. Data collection and analysis were not performed blind to the conditions of the experiments as co-staining with PMP2 allowed the experimenter to deduce the experimental groups. To generate adeno-associated virus (AAV), we subcloned the human *PMP2* transcript (NM_002677.4) into an AAV plasmid backbone containing a *GFAP* minimal promoter (gfaABC1D) and tdTomato reporter. The AAV backbone was obtained from Addgene (pZac2.1 gfaABC1D-tdTomato, catalog # 44332) and Cyagen performed the subcloning. *PMP2* was subcloned upstream of tdTomato and separated by a t2A self-cleaving peptide allowing for bicistronic expression (Fig. 8G). Control (gfaABC1d-tdTomato) and *PMP2* (gfaABC1d-PMP2-t2A-tdTomato) AAV5 viruses were generated by the Penn Vector Core (1.03e14 and 2.096e13 genome copies per milliliter, respectively). To generate lentivirus, the human *PMP2* transcript (NM_002677.4) was synthesized with a 3’ t2A self-cleaving peptide by GenScript. This transcript was subcloned into a 3rd generation lentivirus backbone that contained a *GFAP* minimal promoter (gfaABC1D), spaghetti monster (sm) V5-tag reporter[105], and a Lck membrane-targeting domain. The plasmid was a generous gift from Amy Gleichman (Carmichael lab). *PMP2* was subcloned upstream of smV5 and separated by a t2A self-cleaving peptide (Fig. S12F). High-concentration lentivirus was created by the UCSF viral core (~10^8 colony-forming units per milliliter) for control (gfaABC1D-Lck-SmV5) and *PMP2* (gfaABC1D-PMP2-t2A-Lck-SmV5) viruses.

To infect mouse astrocytes *in vivo*, one μL of virus (for AAV experiments, the control vector was diluted 1:5 in PBS) was injected into the lateral ventricles of postnatal day 1 mice (coordinates from lambda: x = 0.8 mm lateral, y = 1.5 mm caudal, z = −1.6 mm from surface). Forty-two days after injection, mice were euthanized and perfused (as above) for histological analysis. Thick sections were created using a sliding microtome (Microm HM 450; 140 μm and 50 μm for the AAV and lentivirus experiments, respectively). After staining for V5 or tdTomato (see antibodies above), astrocytes were imaged with a confocal microscope (CSU-W1 spinning disk or Leica TCS SPE for AAV and lentivirus experiments, respectively) at 40x and ~0.3 μm step size. The maximum astrocyte diameter was defined as the greatest linear distance between two points in a 2-dimensional slice that crossed through the nucleus. Primary branches were counted manually through the z-stacked images.

9. Website information

Genome-wide estimates of expression fidelity for CNS cell classes are provided in Table S3 for humans and Table S7 for mice. To facilitate interactions with our findings, we have also created a website where users can search by CNS region, cell class, and gene to retrieve information about gene expression fidelity and associated measures (http://oldhamlab.ctec.ucsf.edu/). Major CNS cell classes (astrocytes, oligodendrocytes, microglia, and neurons) are currently supported.

10. Statistics and data presentation

All statistical analyses were performed in the R statistical computing environment (http://cran.us.r-project.org). Fisher's exact test was used to assess the significance of gene set enrichment, and Wilcoxon signed-rank and rank-sum tests were used to assess the significance of median differences (for paired and unpaired data, respectively). For PMP2 experiments, the normality of the data was verified using the Shapiro-Wilk test and unpaired one-sided Welch's t-tests were used to assess the significance of mean differences between the two groups. Linear regression was used to assess the predictive significance of cellular abundance with respect to gene expression patterns. The following assumptions were made: linear relationship between dependent and independent variables, normality of the variables, little multicollinearity between the independent variables, statistical independence of the residuals, and homoscedasticity (i.e. constant variance of the residuals). These assumptions were not formally tested for all models (i.e. all genes). However, scatter plots of select models (e.g. Figs. 6C, 6G, 7J, 8C, S8, S10, S12, and others not shown) suggest that these assumptions are valid. Violin plots outline the Gaussian kernel probability density and are trimmed to the range of the data. Edges of boxplots denote interquartile range (25th-75th percentile) with whiskers denoting 1.5 times the interquartile range and black line denoting the median value; notches denote 1.58 times the interquartile range divided by the square root of the number of samples.

11. Reporting summary

Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

12. Data availability

All gene expression datasets analyzed in this study are publicly available (accession codes and URLs are provided in Table S1). Genome-wide estimates of expression fidelity for major human CNS cell classes are provided on our web site (http://oldhamlab.ctec.ucsf.edu/). All other data that support the findings of this study are available from the corresponding author upon reasonable request.

13. Code availability

Example R code for resolving changes in cellular composition from changes in cell-class-specific transcriptional identity is provided in Online Methods (Section 3.2). All other code that supports the findings of this study is available from the corresponding author upon reasonable request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We are grateful to B. Dispensa (UCSF), J. Hesse (UCSF), D. Kleinhesselink (UCSF), and J. Jed (UCSF) for technical support. We thank A. Molinaro (UCSF) for statistical consultations, D. Rowitch (UCSF) for astrocyte discussions, and E. Huang (UCSF) and M. Paredes (UCSF) for human brain samples. Due to space limitations, we apologize that many relevant and important publications are not cited. This work was supported by the UCSF Program for Breakthrough Biomedical Research (to M.C.O.), which is funded in part by the Sandler Foundation, a Scholar Award from the UCSF Weill Institute for Neurosciences (to M.C.O.), NIMH R01MH113896 (to M.C.O.), a Pew Scholars Award (to A.V.M.), NIMH K08MH104417 (to A.V.M.), and the Burroughs Wellcome Fund (to A.V.M.).

REFERENCES

1. Darmanis S, et al. A survey of human brain transcriptome diversity at the single cell level. PNAS 112, 7285–7290 (2015). [PubMed: 26060301]
2. Paul G, et al. The adult human brain harbors multipotent perivascular mesenchymal stem cells. PLoS One 7, e35577 (2012). [PubMed: 22523602]
3. Zhang Y, et al. Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. Neuron 89, 37–53 (2016). [PubMed: 26687838]
4. Habib N, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods 14, 955–958 (2017). [PubMed: 28846088]
5. Lake BB, et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science 352, 1586–1590 (2016). [PubMed: 27339989]
6. Liu S & Trapnell C Single-cell transcriptome sequencing: recent advances and remaining challenges. F1000Res 5 (2016).
7. Azevedo FA, et al. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. J. Comp. Neurol. 513, 532–541 (2009). [PubMed: 19226510]
8. Oldham MC Transcriptomics: from differential expression to coexpression. in The OMICs: applications in neurosciences (ed. G. Coppola) 85–113 (Oxford, 2014).
9. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z & Clark HF Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS One 4, e6098 (2009). [PubMed: 19568420]

10. Gong T & Szustakowski JD DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics 29, 1083–1085 (2013). [PubMed: 23428642]
11. Kuhn A, Thu D, Waldvogel HJ, Faull RL & Luthi-Carter R Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. Nat Methods 8, 945–947 (2011). [PubMed: 21983921]
12. Newman AM, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 12, 453–457 (2015). [PubMed: 25822800]
13. Shen-Orr SS, et al. Cell type-specific gene expression differences in complex tissues. Nat Methods 7, 287–289 (2010). [PubMed: 20208531]
14. Zhong Y, Wan YW, Pang K, Chow LM & Liu Z Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC Bioinformatics 14, 89 (2013). [PubMed: 23497278]
15. Zuckerman NS, Noam Y, Goldsmith AJ & Lee PP A self-directed method for cell-type identification and separation of gene expression microarrays. PLoS Comput. Biol. 9, e1003189 (2013). [PubMed: 23990767]
16. Oldham MC, et al. Functional organization of the transcriptome in human brain. Nat. Neurosci. 11, 1271–1282 (2008). [PubMed: 18849986]
17. Fertuzinhos S, et al. Laminar and temporal expression dynamics of coding and noncoding RNAs in the mouse neocortex. Cell Rep 6, 938–950 (2014). [PubMed: 24561256]
18. Ponomarev I, Rau V, Eger EI, Harris RA & Fanselow MS Amygdala transcriptome and cellular mechanisms underlying stress-enhanced fear learning in a rat model of posttraumatic stress disorder. Neuropsychopharmacology 35, 1402–1411 (2010). [PubMed: 20147889]
19. Hilliard AT, Miller JE, Fraley ER, Horvath S & White SA Molecular microcircuitry underlies functional specification in a basal ganglia circuit dedicated to vocal learning. Neuron 73, 537–552 (2012). [PubMed: 22325205]
20. Bakken TE, et al. A comprehensive transcriptional map of primate brain development. Nature 535, 367–375 (2016). [PubMed: 27409810]
21. Hawrylycz M, et al. Canonical genetic signatures of the adult human brain. Nat. Neurosci. 18, 1832–1844 (2015). [PubMed: 26571460]
22. Cahoy JD, et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. J. Neurosci. 28, 264–278 (2008). [PubMed: 18171944]
23. Hickman SE, et al. The microglial sensome revealed by direct RNA sequencing. Nat. Neurosci. 16, 1896–1905 (2013). [PubMed: 24162652]
24. Horvath S & Dong J Geometric interpretation of gene coexpression network analysis. PLoS Comput. Biol. 4, e1000117 (2008). [PubMed: 18704157]
25. Langfelder P & Horvath S WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008). [PubMed: 19114008]
26. Tasic B, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat. Neurosci. 19, 335–346 (2016). [PubMed: 26727548]
27. Hodges A, et al. Regional and cellular gene expression changes in human Huntington's disease brain. Hum. Mol. Genet. 15, 965–977 (2006). [PubMed: 16467349]
28. Berchtold NC, et al. Gene expression changes in the course of normal brain aging are sexually dimorphic. PNAS 105, 15605–15610 (2008). [PubMed: 18832152]
29. Kang HJ, et al. Spatio-temporal transcriptome of the human brain. Nature 478, 483–489 (2011). [PubMed: 22031440]
30. Hernandez DG, et al. Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. Neurobiol. Dis. 47, 20–28 (2012). [PubMed: 22433082]
31. Li JZ, et al. Circadian patterns of gene expression in the human brain and disruption in major depressive disorder. PNAS 110, 9950–9955 (2013). [PubMed: 23671070]
32. Ramasamy A, et al. Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat. Neurosci. 17, 1418–1428 (2014). [PubMed: 25174004]

33. GTExConsortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348, 648–660 (2015). [PubMed: 25954001]
34. Lake BB, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat. Biotechnol. 36, 70–80 (2018). [PubMed: 29227469]
35. Lek M, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016). [PubMed: 27535533]
36. Binder JX, et al. COMPARTMENTS: unification and visualization of protein subcellular localization evidence. Database 2 25, bau012 (2014).
37. Szklarczyk D, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43, D447–452 (2015). [PubMed: 25352553]
38. Lein ES, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature 445, 168–176 (2007). [PubMed: 17151600]
39. Yu W, Clyne M, Khoury MJ & Gwinn M Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. Bioinformatics 26, 145–146 (2010). [PubMed: 19864262]
40. Cribbs DH, et al. Extensive innate immune gene activation accompanies brain aging, increasing vulnerability to cognitive decline and neurodegeneration: a microarray study. J. Neuroinflammation 9, 179 (2012). [PubMed: 22824372]
41. Zhang B, et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell 153, 707–720 (2013). [PubMed: 23622250]
42. Hokama M, et al. Altered expression of diabetes-related genes in Alzheimer's disease brains: the Hisayama study. Cereb. Cortex 24, 2476–2488 (2014). [PubMed: 23595620]
43. Oberheim NA, Goldman SA & Nedergaard M Heterogeneity of astrocytic form and function. Methods Mol. Biol. 814, 23–45 (2012). [PubMed: 22144298]
44. Zenker J, et al. A role of peripheral myelin protein 2 in lipid homeostasis of myelinating Schwann cells. Glia 62, 1502–1512 (2014). [PubMed: 24849898]
45. Bozek K, et al. Exceptional evolutionary divergence of human muscle and brain metabolomes parallels human cognitive and physical uniqueness. PLoS Biol. 12, e1001871 (2014). [PubMed: 24866127]
46. 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). [PubMed: 11307625]
47. 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 (2014). [PubMed: 25186741]
48. Butovsky O, et al. Identification of a unique TGF-beta-dependent molecular and functional signature in microglia. Nat. Neurosci. 17, 131–143 (2014). [PubMed: 24316888]
49. Storey JD & Tibshirani R Statistical significance for genomewide studies. PNAS 100, 9440–9445 (2003). [PubMed: 12883005]

ONLINE METHODS REFERENCES

50. Irizarry RA, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4, 249–264 (2003). [PubMed: 12925520]
51. Zhang J, Finney RP, Clifford RJ, Derr LK & Buetow KH Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach. Genomics 85, 297–308 (2005). [PubMed: 15718097]
52. Gautier L, Cope L, Bolstad BM & Irizarry RA affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20, 307–315 (2004). [PubMed: 14960456]
53. Davis S & Meltzer PS GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 23, 1846–1847 (2007). [PubMed: 17496320]
54. Oldham MC, Langfelder P & Horvath S Network methods for describing sample relationships in genomic datasets: application to Huntington's disease. BMC Syst. Biol. 6, 63 (2012). [PubMed: 22691535]

55. Johnson WE, Li C & Rabinovic A Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127 (2007). [PubMed: 16632515]
56. Bolstad BM, Irizarry RA, Astrand M & Speed TP A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19, 185–193 (2003). [PubMed: 12538238]
57. Molofsky AV, et al. Expression profiling of Aldh1l1-precursors in the developing spinal cord reveals glial lineage-specific genes and direct Sox9-Nfe2l1 interactions. Glia 61, 1518–1532 (2013). [PubMed: 23840004]
58. Lui JH, et al. Radial glia require PDGFD-PDGFRbeta signalling in human but not mouse neocortex. Nature 515, 264–268 (2014). [PubMed: 25391964]
59. Hardin J, Mitani A, Hicks L & VanKoten B A robust measure of correlation between two genes on a microarray. BMC Bioinformatics 8, 220 (2007). [PubMed: 17592643]
60. Song L, Langfelder P & Horvath S Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinformatics 13, 328 (2012). [PubMed: 23217028]
61. Zeisel A, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347, 1138–1142 (2015). [PubMed: 25700174]
62. Butler LM, et al. Analysis of Body-wide Unfractionated Tissue Data to Identify a Core Human Endothelial Transcriptome. Cell Syst 3, 287–301 e283 (2016). [PubMed: 27641958]
63. La Manno G, et al. Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. Cell 167, 566–580 e519 (2016). [PubMed: 27716510]
64. Kuhn A, et al. Cell population-specific expression analysis of human cerebellum. BMC Genomics 13, 610 (2012). [PubMed: 23145530]
65. Fisher RA Statistical Methods for Research Workers (Hafner Publishing Company, 1970).
66. Doyle JP, et al. Application of a translational profiling approach for the comparative analysis of CNS cell types. Cell 135, 749–762 (2008). [PubMed: 19013282]
67. Mancarci BO, et al. NeuroExpresso: A cross-laboratory database of brain cell-type expression profiles with applications to marker gene identification and bulk brain tissue transcriptome interpretation. bioRxiv (2016).
68. Gokce O, et al. Cellular Taxonomy of the Mouse Striatum as Revealed by Single-Cell RNA-Seq. Cell Rep 16, 1126–1137 (2016). [PubMed: 27425622]
69. He L, et al. Analysis of the brain mural cell transcriptome. Sci. Rep. 6, 35108 (2016). [PubMed: 27725773]
70. Wickham H ggplot2: Elegant graphics for data analysis (Springer-Verlag, New York, 2009).
71. Trapnell C, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. 32, 381–386 (2014). [PubMed: 24658644]
72. Yin, P.a.F X Estimating R2 shrinkage in multiple regression: A comparison of different analytical methods. The Journal of Experimental Education 69, 203–224 (2001).
73. Troyanskaya O, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 17, 520–525 (2001). [PubMed: 11395428]
74. Benoit J, Ayoub A & Rakic P Epigenetic stability in the adult mouse cortex under conditions of pharmacologically induced histone acetylation. Brain Struct Funct (2015).
75. Benton CS, et al. Evaluating genetic markers and neurobiochemical analytes for fluoxetine response using a panel of mouse inbred strains. Psychopharmacology (Berl.) 221, 297–315 (2012). [PubMed: 22113448]
76. Chu PL, Keum S & Marchuk DA A novel genetic locus modulates infarct volume independently of the extent of collateral circulation. Physiol. Genomics 45, 751–763 (2013). [PubMed: 23800850]
77. Iancu OD, et al. Cosplicing network analysis of mammalian brain RNA-Seq data utilizing WGCNA and Mantel correlations. Front Genet 6, 174 (2015). [PubMed: 26029240]
78. Jiang P, et al. A systems approach identifies networks and genes linking sleep and stress: implications for neuropsychiatric disorders. Cell Rep 11, 835–848 (2015). [PubMed: 25921536]
79. Kasukawa T, et al. Quantitative expression profile of distinct functional regions in the adult mouse brain. PLoS One 6, e23228 (2011). [PubMed: 21858037]

80. Kleiman RJ, et al. Dendritic spine density deficits in the hippocampal CA1 region of young Tg2576 mice are ameliorated with the PDE9A inhibitor PF-04447943. Alzheimer's & Dementia 6, S563–S564 (2010).
81. Ling KH, et al. Functional transcriptome analysis of the postnatal brain of the Ts1Cje mouse model for Down syndrome reveals global disruption of interferon-related molecular networks. BMC Genomics 15, 624 (2014). [PubMed: 25052193]
82. Mackiewicz M, et al. Macromolecule biosynthesis: a key function of sleep. Physiol. Genomics 31, 441–457 (2007). [PubMed: 17698924]
83. Matarin M, et al. A genome-wide gene-expression analysis and database in transgenic mice during development of amyloid or tau pathology. Cell Rep 10, 633–644 (2015). [PubMed: 25620700]
84. Parente MK, Rozen R, Cearley CN & Wolfe JH Dysregulation of gene expression in a lysosomal storage disease varies between brain regions implicating unexpected mechanisms of neuropathology. PLoS One 7, e32419 (2012). [PubMed: 22403656]
85. Peixoto LL, et al. Memory acquisition and retrieval impact different epigenetic processes that regulate gene expression. BMC Genomics 16 Suppl 5, S5 (2015).
86. Segall SK, et al. Comt1 genotype and expression predicts anxiety and nociceptive sensitivity in inbred strains of mice. Genes Brain Behav 9, 933–946 (2010). [PubMed: 20659173]
87. Stark KL, et al. Altered brain microRNA biogenesis contributes to phenotypic deficits in a 22q11-deletion mouse model. Nat. Genet. 40, 751–760 (2008). [PubMed: 18469815]
88. Stevens SL, et al. Multiple preconditioning paradigms converge on interferon regulatory factor-dependent signaling to promote tolerance to ischemic brain injury. J. Neurosci. 31, 8456–8463 (2011). [PubMed: 21653850]
89. Vanderlinden LA, et al. Whole brain and brain regional coexpression network interactions associated with predisposition to alcohol consumption. PLoS One 8, e68878 (2013). [PubMed: 23894363]
90. Wes PD, et al. Tau overexpression impacts a neuroinflammation gene expression network perturbed in Alzheimer's disease. PLoS One 9, e106050 (2014). [PubMed: 25153994]
91. Wolen AR, et al. Genetic dissection of acute ethanol responsive gene networks in prefrontal cortex: functional and mechanistic implications. PLoS One 7, e33575 (2012). [PubMed: 22511924]
92. Enard W, et al. Intra- and interspecific variation in primate gene expression patterns. Science 296, 340–343 (2002). [PubMed: 11951044]
93. Khaitovich P, et al. Regional patterns of gene expression in human and chimpanzee brains. Genome Res. 14, 1462–1473 (2004). [PubMed: 15289471]
94. Caceres M, et al. Elevated gene expression levels distinguish human from non-human primate brains. PNAS 100, 13030–13035 (2003). [PubMed: 14557539]
95. Fraser HB, Khaitovich P, Plotkin JB, Paabo S & Eisen MB Aging and gene expression in the primate brain. PLoS Biol. 3, e274 (2005). [PubMed: 16048372]
96. Khaitovich P, et al. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science 309, 1850–1854 (2005). [PubMed: 16141373]
97. Franz H, et al. Systematic analysis of gene expression in human brains before and after death. Genome Biol. 6, R112 (2005). [PubMed: 16420671]
98. Khaitovich P, et al. Positive selection on gene expression in the human brain. Curr. Biol. 16, R356–358 (2006). [PubMed: 16618540]
99. Somel M, et al. Transcriptional neoteny in the human brain. PNAS 106, 5743–5748 (2009). [PubMed: 19307592]
100. Dai M, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res. 33, e175 (2005). [PubMed: 16284200]
101. Bernard A, et al. Transcriptional architecture of the primate neocortex. Neuron 73, 1083–1099 (2012). [PubMed: 22445337]
102. Chen J, Bardes EE, Aronow BJ & Jegga AG ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 37, W305–311 (2009). [PubMed: 19465376]

103. Yu W, et al. GAPscreener: an automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. BMC Bioinformatics 9, 205 (2008). [PubMed: 18430222]

104. Uhlen M, et al. Proteomics. Tissue-based map of the human proteome. Science 347, 1260419 (2015). [PubMed: 25613900]

105. Viswanathan S, et al. High-performance probes for light and electron microscopy. Nat Methods 12, 568–576 (2015). [PubMed: 25915120]

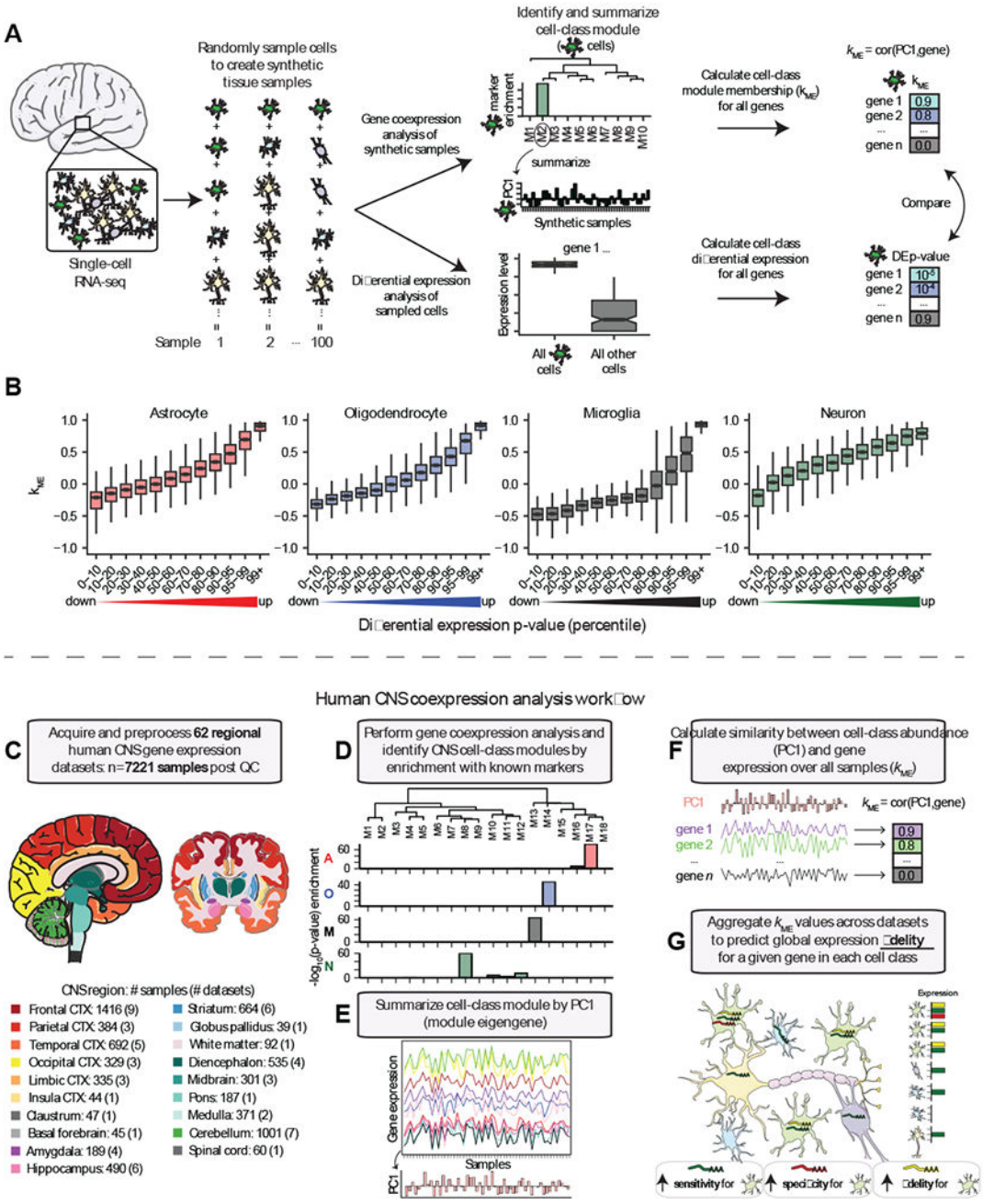

Fig. 1 |. Rationale and workflow.

A) Left: Single-cell RNA-seq data from adult human brain samples[1] were randomly aggregated to create 100 synthetic tissue samples. Right (top): Unsupervised gene coexpression analysis of synthetic samples revealed CNS cell-class modules that were highly enriched with markers of major cell classes. Cell-class module membership strength (k_{ME}) was calculated for all genes. Right (bottom): Using the same cells that were selected to create synthetic samples, single-cell differential expression analysis was performed for all genes with respect to each cell class. **B)** k_{ME} values for synthetic cell-class modules

accurately predicted the results of differential expression analysis for each cell class (n=10 synthetic datasets; ‘up’ / ‘down’ denote up- and down-regulated genes for each cell class). **C)** 62 gene expression datasets consisting of 7221 non-pathological samples from 19 adult human CNS regions were acquired and preprocessed (Table S1). **D)** Unsupervised gene coexpression analysis was performed for each dataset to identify modules of genes with similar expression patterns. Published markers of astrocytes (A), oligodendrocytes (O), microglia (M), and neurons (N) were cross-referenced with all modules (one-sided Fisher’s exact test; Table S2). **E)** For each dataset, the module that was most significantly enriched with markers of a given cell class was summarized by its 1st principal component (PC1, or module eigengene). **F)** Cell-class module eigengenes were used to calculate the similarity between cellular abundance and genome-wide expression patterns (k_{ME}) over all samples in each dataset. **G)** Genome-wide k_{ME} values for significant cell-class modules from all datasets were combined to yield a global measure of expression ‘fidelity’ for each gene with respect to each cell class. Schematic: A gene has high fidelity for a cell class if its expression is sensitive (it is consistently expressed by members of that cell class) and specific (it is not expressed by members of other cell classes).

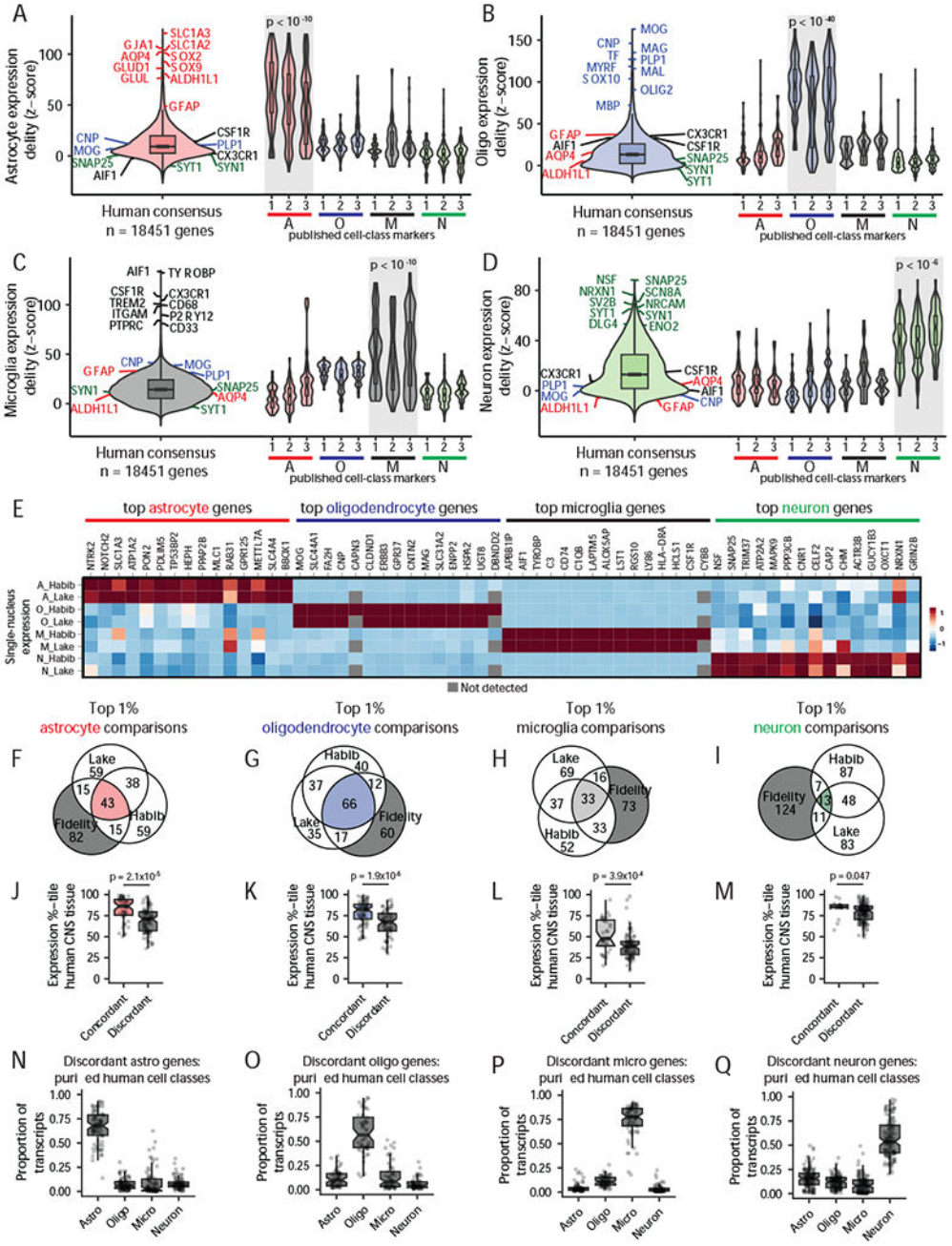

Fig. 2 |. Integrative gene coexpression analysis of intact CNS transcriptomes reveals consensus transcriptional profiles of human astrocytes, oligodendrocytes, microglia, and neurons.

A-D) Left: consensus gene expression fidelity distributions for human astrocytes (A), oligodendrocytes (O), microglia (M), and neurons (N). Canonical markers are labeled in red (A), blue (O), black (M), and green (N). Right: gene expression fidelity distributions for published cell-class markers (A1, O1, M1, N1: [47]; A2, O2, N2: [22]; M2: [23]; A3, O3, N3: [38]; M3: [48]) were cross-referenced with high-fidelity genes (z-score >50). Gray shading: significant enrichment (one-sided Fisher's exact test). Note that A2, O2, M2, and N2 were

the gene sets used for module enrichment analysis (Table S2). The number of independent samples used to calculate fidelity for each gene is provided in Table S3. **E)** Single-nucleus (SN) RNA-seq data from adult human brain[4, 34] for the top 15 high-fidelity genes from each cell class. Standardized expression levels are shown, with red/blue = high/low expression. **F-I)** Overlap among the top 1% of high-fidelity genes and the top 1% of differentially expressed genes for each cell class in SN RNA-seq data from adult human brain[4, 34]. The central colored sectors denote concordant genes, while the peripheral dark grey sectors denote discordant genes. **J-M)** Mean expression percentiles in intact human CNS samples (n=7221 biologically independent tissue samples, Fig. 1C and Table S3) for concordant and discordant genes (**F-I**). Statistical significance was assessed with a one-sided Wilcoxon rank-sum test. **N-Q)** Expression patterns of discordant genes (**F-I**) in cell classes that were purified by immunopanning from adult human temporal lobe surgical resections[3] (purifications from n=12 [astrocyte], three [oligodendrocyte], three [microglia], and one [neuron] biologically independent samples).

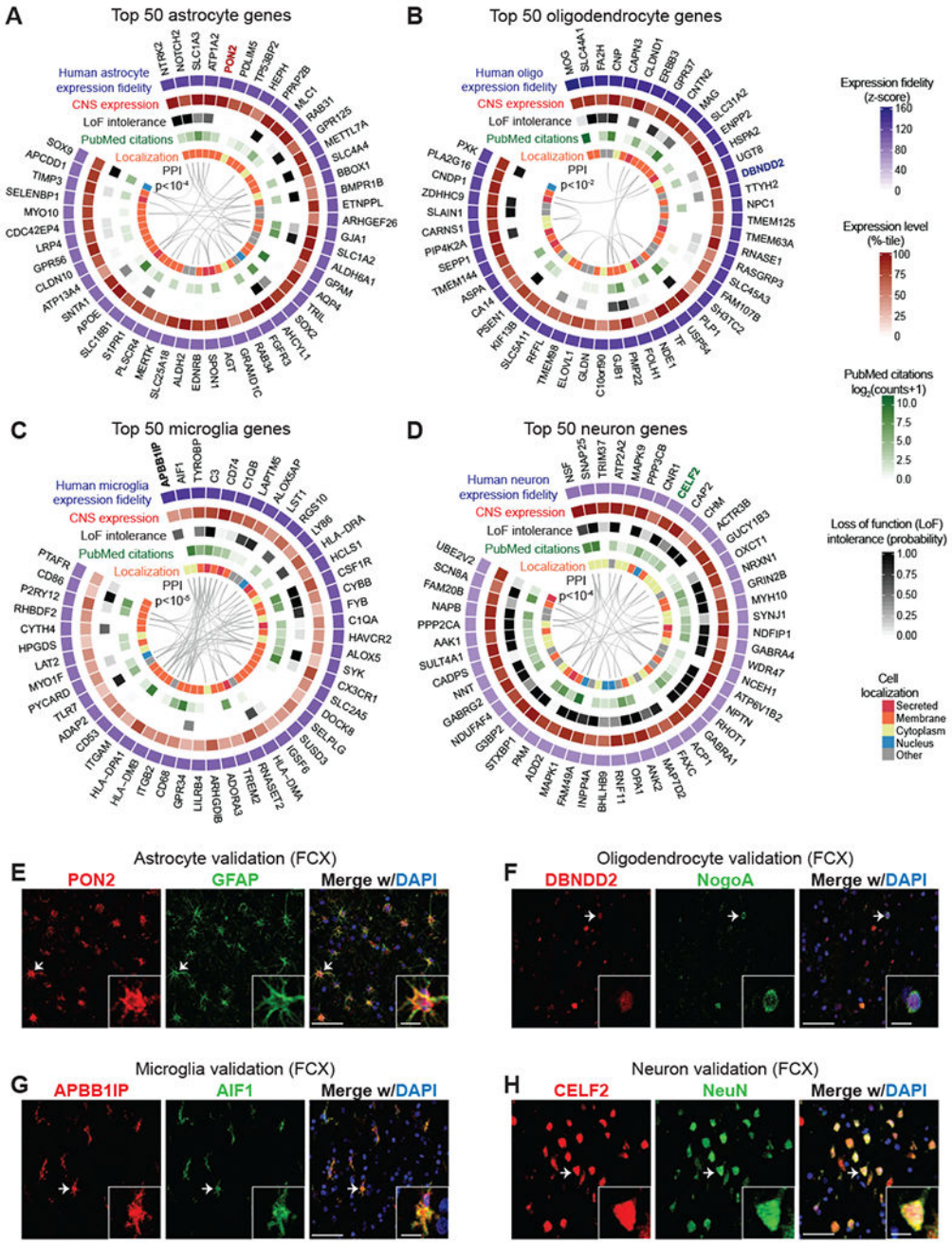

Fig. 3 |. The core transcriptional identities of human astrocytes, oligodendrocytes, microglia, and neurons include known and novel biomarkers.

A-D) The top 50 genes ranked by consensus expression fidelity for astrocytes, oligodendrocytes, microglia, or neurons. Expression levels represent averages of mean percentile ranks for all regional datasets where gene data were present. Mutation intolerance data were obtained from ExAC[35]. PubMed citations were obtained by queries with gene symbol and cell type (e.g. gene symbol and 'astrocyte'). Cellular localization data were extracted from COMPARTMENTS[36]. Predicted protein-protein interactions (PPI) were

obtained from STRING[37]. A link is shown if the combined score between two proteins was > 350. The probability of observing the number of depicted links by chance was determined by resampling (n = 100,0000 random samples of 50 genes). **E-H)** Novel markers of human astrocytes (PON2), oligodendrocytes (DBNDD2), microglia (APBB1IP), and neurons (CELF2) in adult human dorsolateral prefrontal cortex (DLPFC; **E**: L5/6; **F**,**G**: white matter; **H**: L2/3). Immunostaining was repeated at least twice on independent samples with similar results. Arrowhead: cell in inset. Scale bar: 50μm; inset scale bar: 10μm.

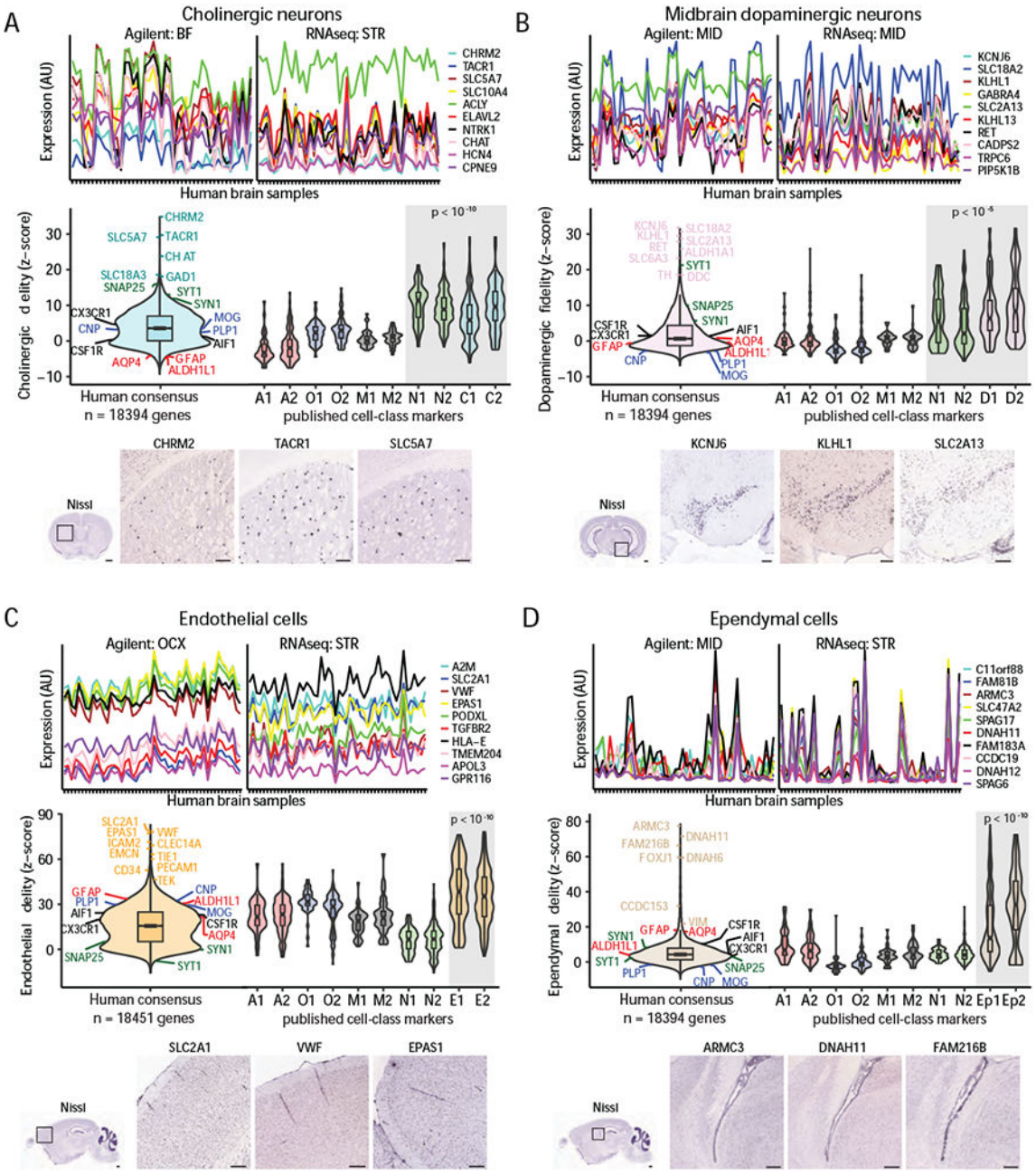

Fig. 4 |. Variation among intact tissue samples reveals transcriptional signatures of human cholinergic neurons, midbrain dopaminergic neurons, endothelial cells, and ependymal cells.

A-D) Top: high-fidelity genes for each cell class (top 10 are shown) are consistently coexpressed in independent datasets (Table S1). Middle: consensus gene expression fidelity distributions for each cell class with canonical markers of major cell classes labeled in green (neurons), red (astrocytes), blue (oligodendrocytes), and black (microglia). Gene expression fidelity distributions for published sets of markers (Al, A2, O1, O2, M1, M2, N1, N2, C1, C2, D1, D2, E1, E2, Ep1, Ep2; Methods) were cross-referenced with high-fidelity genes (top

three percentiles). Gray shading: significant enrichment (one-sided Fisher's exact test). Note that E2 and Ep1 were gene sets used for module enrichment analysis (Table S2). The number of independent samples used to calculate fidelity for each gene is provided in Table S3. Bottom: mouse *in situ* hybridization data from the Allen Mouse Brain Atlas[38] for high-fidelity genes in dorsal striatum (**A**), ventral midbrain (**B**), cortex (**C**), and lateral ventricle (**D**). Scale bar: 200μm; inset scale bar: 500μm.

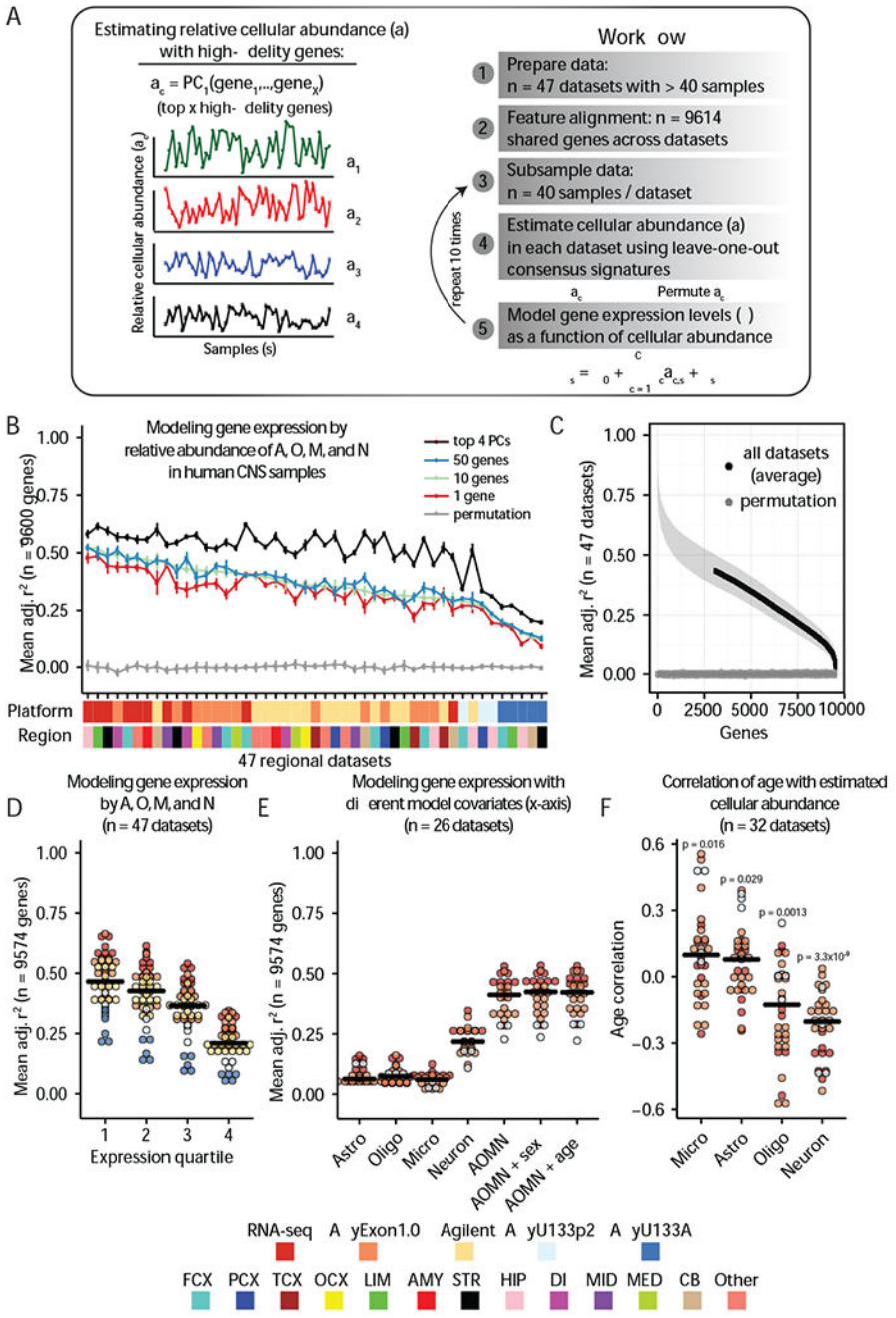

Fig. 5 |. Variation in cellular abundance predicts gene expression in transcriptomes from intact CNS samples

A) Strategy for modeling gene expression in intact human CNS samples as a function of inferred cellular abundance. **B)** Total % variance explained (mean adj. r^2) for ~9600 genes whose expression levels were modeled as a function of inferred astrocyte, oligodendrocyte, microglia, and neuron abundance in each of 47 regional datasets (subset to 40 samples; values are mean +/− 2 s.e.m., 10 iterations). **C)** Mean adj. r^2 values for individual genes from (**B**) over the 47 datasets. Grey envelope: loess smoothed C.I. (+/− 2 s.e.m., 10

iterations). **D)** Mean adj. r^2 values for genes from (**B**) grouped by mean expression quartiles (each point is one dataset). **E)** Mean adj. r^2 values for 7 different models (restricted to datasets w/ sex and age information: GSE46706, GTEx, GSE11882, GSE25219). **F)** Pearson correlation of inferred cellular abundance with donor sample age (one-sample Wilcoxon signed-rank test). Horizontal bars (**D-F**): median; points colored by technology platform.

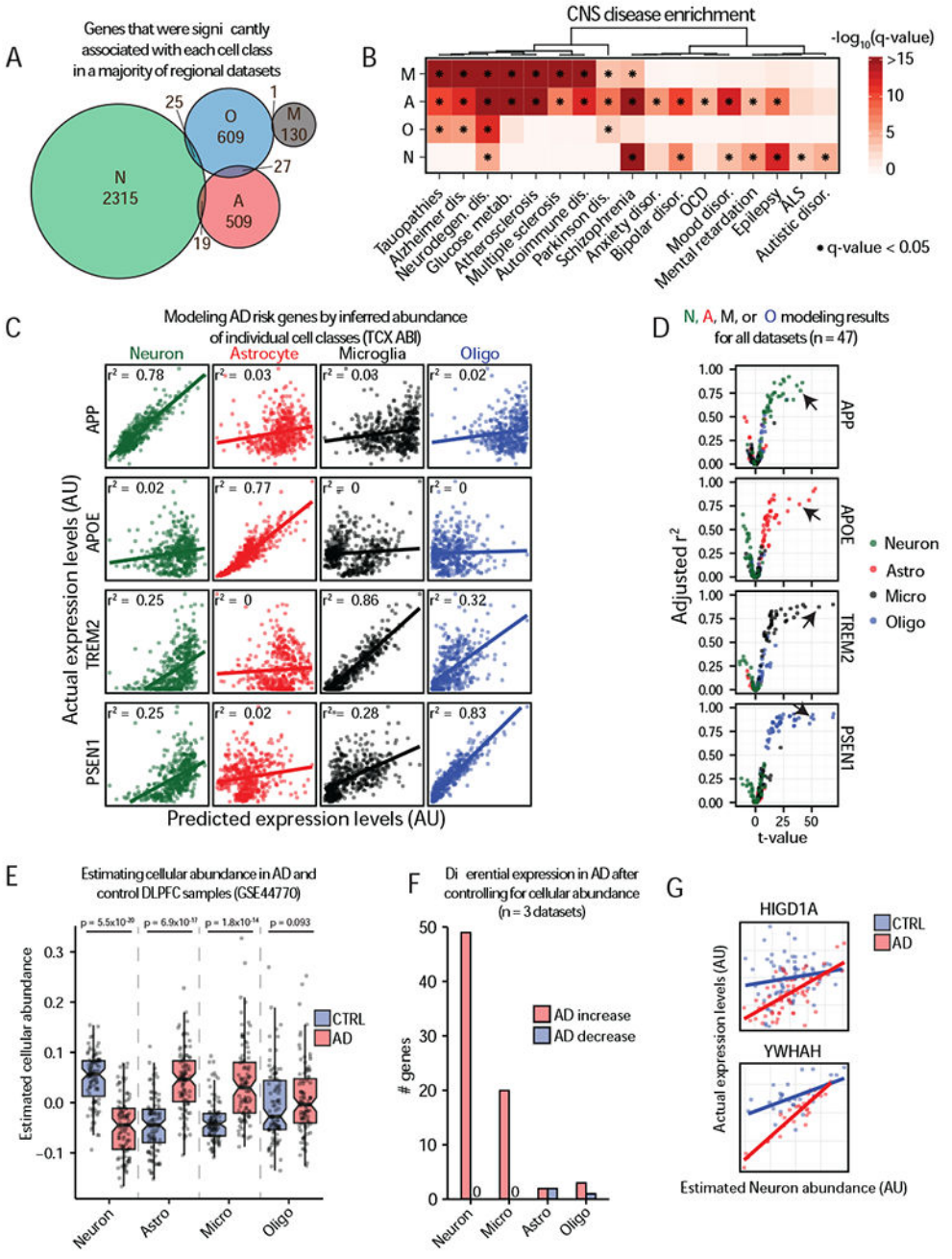

Fig. 6 |. Gene expression modeling offers new avenues for studying human CNS diseases.

A) Genes that were significantly associated with a cell class by linear regression modeling in a majority of human CNS datasets ($p < 8.37\times10^{-9}$, corresponding to a Bonferroni correction based on the total number of gene models). **B)** Enrichment analysis (one-sided Fisher's exact test) of genes from (**A**) with human CNS disease genes from Phenopedia[39]. FDR-adjusted p-values (q-values) are shown[49]. **C)** Linear regression modeling results in human temporal cortex (TCX ABI; i.e. one dataset consisting of 465 samples) for four Alzheimer's disease (AD) risk genes. **D)** Modeling results for genes from (**C**) in 47 datasets (≥40

samples). **E)** Top 10 high-fidelity genes were used to estimate the relative abundance of neurons, astrocytes, microglia, and oligodendrocytes in DLPFC from control (CTRL) and AD[41] (Fig. 5A; n=95 CTRL and 95 AD biologically independent tissue samples). P-values: two-sided Wilcoxon rank-sum test. **F)** Gene expression modeling in three datasets[40–42] reveals consistent cell-class-specific expression changes in AD after controlling for differences in cellular abundance ($p<0.05$ based on 1000 permutations of sample labels). Shown are the total number of genes that were significantly altered for each cell class in all three datasets. **G**) Examples of two genes that are up-regulated in AD neurons (top[41]; bottom[42]).

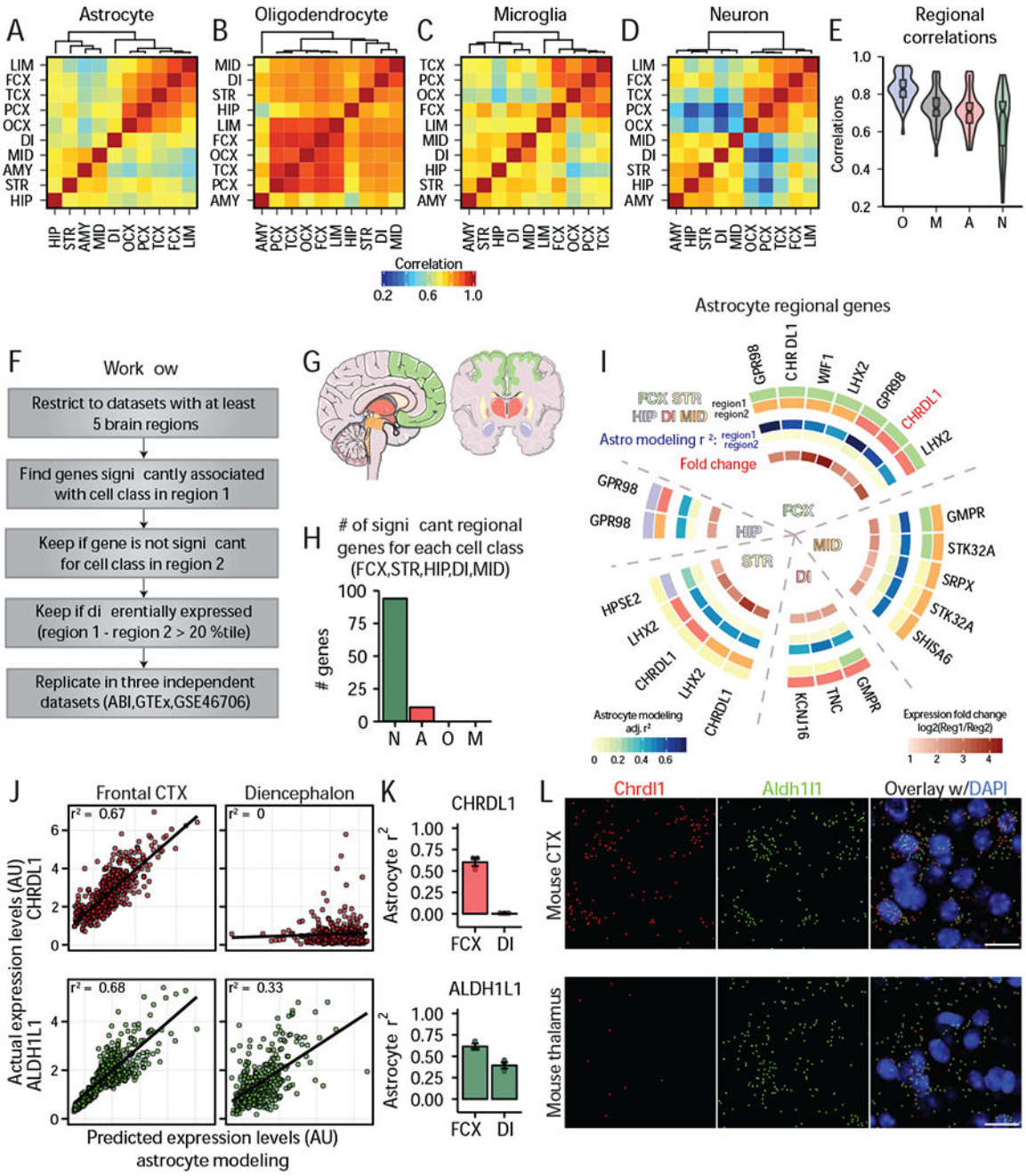

Fig. 7 |. Regional expression fidelity and predictive modeling reveal astrocyte heterogeneity in the human brain

(A-D) Hierarchical clustering of human brain regions (excluding cerebellum) based on Pearson correlations among regional expression fidelity for each cell class (n=18451 genes, ≥3 datasets/region; see Table S1 for the number of datasets and samples per region). **E)** Distributions of correlations in (**A-D**), n=10 regions. **F)** Workflow to predict regional expression differences in specific cell classes. Significance threshold: $p<2.67\times10^{-8}$ (Bonferroni correction for total # of gene models). **G)** Schematic of analyzed brain regions:

frontal cortex (FCX), striatum (STR), hippocampus (HIP), diencephalon (DI), and midbrain (MID). **H)** Total # of region-specific genes conservatively predicted for each cell class. **I)** Genes predicted to be expressed by human astrocytes in restricted brain regions. **J)** Modeling of *CHRDL1* (candidate from (**I**)) and *ALDH1L1* (positive control) as a function of inferred astrocyte abundance in example datasets (FCX/DI from ABI; see Table S1 for sample sizes). **K)** Linear regression modeling results for same genes in three datasets (ABI, GTEx, and GSE46706; Table S1). Barplots: mean values; error bars: s.e.m. **L)** Single-molecule FISH of *Chrdl1* and *Aldh1l1* in mature mouse brain (P30). Scale bar: 20μm. FISH was repeated at least twice on independent samples with similar results.

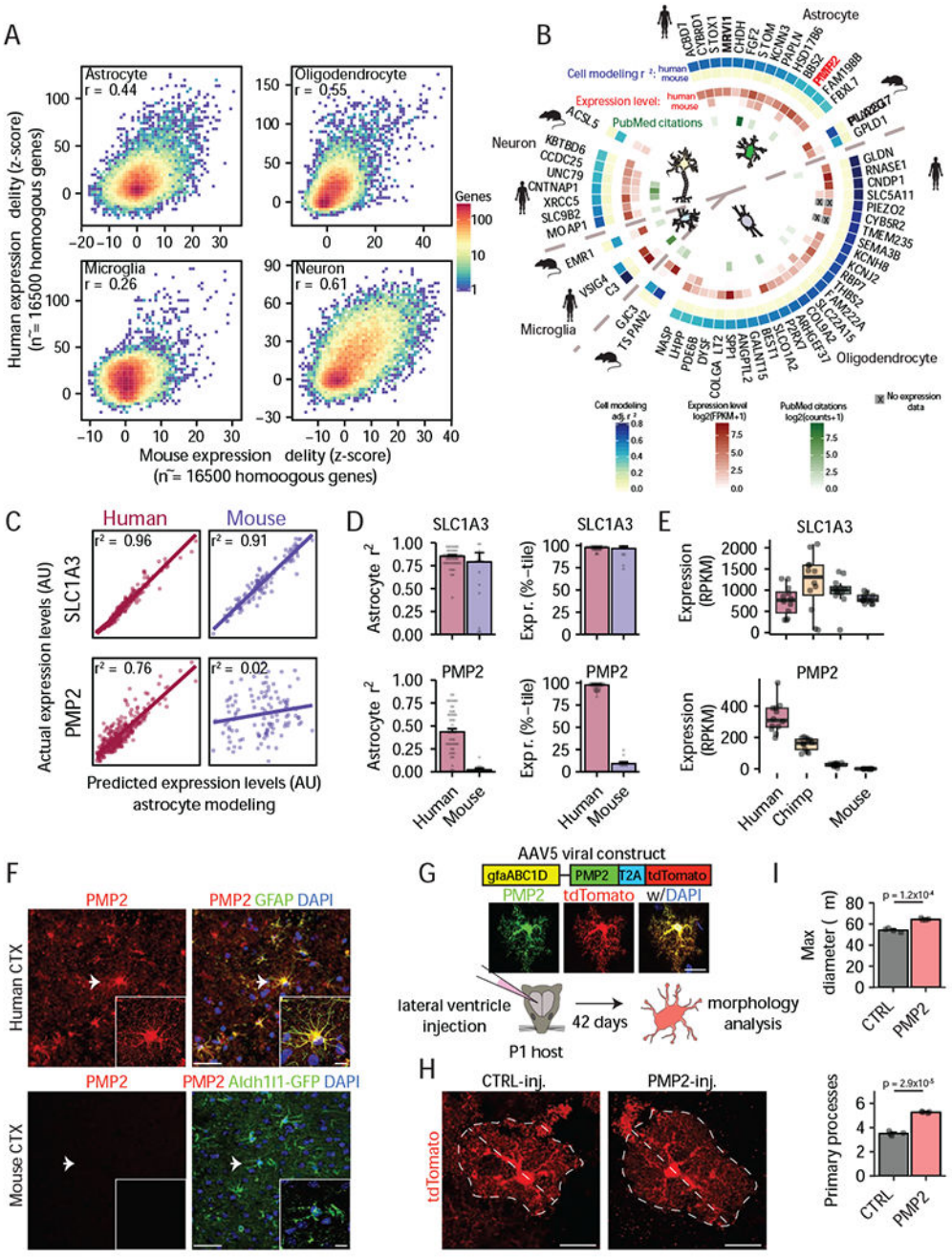

Fig. 8 |. Gene expression modeling identifies cell-class-specific transcriptional differences between humans and mice.

A) Comparisons of gene expression fidelity for homologous genes from humans and mice for each cell class. Pearson correlations are shown. **B)** Predicted cell-class-specific transcriptional differences between humans and mice. Expression levels are from independent datasets[3, 47] that were not used to predict species differences. PubMed citations were obtained by queries with gene symbol and cell type (e.g. gene symbol and ‘astrocyte’). **C)** Examples of linear regression modeling results in humans (Hs.PCX.ABI) and mice

(Ms.GSE64398) (Table S1). *SLC1A3* is predicted to be expressed by astrocytes in both species and *PMP2* is predicted to be expressed by astrocytes in humans but not mice. **D)** Astrocyte modeling results and mean expression percentiles for genes in (**C**) from all datasets (see Table S3 and Table S7 for the number of datasets and samples for each gene and species). Barplots: mean values; error bars: s.e.m. **E)** *SLC1A3* and *PMP2* expression in an independent gene expression dataset from human, chimpanzee, macaque, and mouse prefrontal and visual cortex[45] (n=12 independent tissue samples for each species). **F)** Immunostaining for PMP2 in adult human DLPFC and P42 mouse neocortex. GFAP and Aldh1l1 label astrocytes. Arrowheads: cells in insets. Scale bar: 40μm; inset scale bar: 10μm. Immunostaining was repeated at least twice on independent samples with similar results. **G)** Experimental strategy for studying *PMP2* effects on mouse astrocytes. Scale bar: 20μm. **H)** Representative examples of CTRL and *PMP2*-infected astrocytes in mouse neocortex. Dashed lines outline the cell and the maximum diameter through the nucleus. Scale bar: 20μm. **I)** Quantification of maximum diameter and the number of primary processes in CTRL and *PMP2*-infected astrocytes. n=4 animals per group, n=100 CTRL and 110 *PMP2*-infected astrocytes, bars denote mean ± s.e.m., with significance determined by a one-sided Welch's *t*-test on animal averages.