INFERENCE OF CELL TYPE COMPOSITION FROM HUMAN BRAIN TRANSCRIPTOMIC DATASETS ILLUMINATES THE EFFECTS OF AGE, MANNER OF DEATH, DISSECTION, AND PSYCHIATRIC DIAGNOSIS

*Megan Hastings Hagenauer, Ph.D.[1], Anton Schulmann, M.D.[2], Jun Z. Li, Ph.D.[3], Marquis P. Vawter, Ph.D.[4], David M. Walsh, Psy.D.[4], Robert C. Thompson, Ph.D.[1], Cortney A. Turner, Ph.D.[1], William E. Bunney, M.D.[4], Richard M. Myers, Ph.D.[5], Jack D. Barchas, M.D.[6], Alan F. Schatzberg, M.D.[7], Stanley J. Watson, M.D., Ph.D.[1], Huda Akil, Ph.D.[1]

[1]Mol. Behavioral Neurosci. Inst., Univ. of Michigan, Ann Arbor, MI, USA; [2] Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA, USA, [3]Genet., Univ. of Michigan, Ann Arbor, MI, USA; [4]Univ. of California, Irvine, CA; [5]HudsonAlpha Inst. for Biotech., Huntsville, AL, USA; [6]Stanford, Palo Alto, CA, [7]Cornell, New York, NY, USA

*Corresponding Author: Megan Hastings Hagenauer, Ph.D.

e-mail: hagenaue@umich.edu

Molecular Behavioral Neuroscience Institute (MBNI)

205 Zina Pitcher Pl.

Ann Arbor, MI 48109

Abstract

Psychiatric illness is unlikely to arise from pathology occurring uniformly across all cell types in affected brain regions. Despite this, transcriptomic analyses of the human brain have typically been conducted using macro-dissected tissue due to the difficulty of performing single-cell type analyses with donated post-mortem brains. To address this issue statistically, we compiled a database of several thousand transcripts that were specifically-enriched in one of 10 primary cortical cell types, as identified in previous publications. Using this database, we predicted the relative cell type composition for 833 human cortical samples using microarray or RNA-Seq data from the Pritzker Consortium (GSE92538) or publicly-available databases (GSE53987, GSE21935, GSE21138, CommonMind Consortium). These predictions were generated by averaging normalized expression levels across transcripts specific to each cell type using our R-package *BrainInABlender* (validated and publicly-released: https://github.com/hagenaue/BrainInABlender). Using this method, we found that the principal components of variation in the datasets were largely explained by the neuron to glia ratio of the samples. This variability was not simply due to dissection – the relative balance of brain cell types was influenced by a variety of demographic, pre- and post-mortem variables. Prolonged hypoxia around the time of death predicted increased astrocytic and endothelial content in the tissue, illustrating vascular upregulation. Aging was associated with decreased neuronal content. Red blood cell content was reduced in individuals who died following systemic blood loss. Subjects with Major Depressive Disorder had decreased astrocytic content, mirroring previous morphometric observations. Subjects with Schizophrenia had reduced red blood cell content, resembling the hypofrontality detected in fMRI experiments. Finally, in datasets containing samples with especially variable cell content, we found that controlling for predicted sample cell content while evaluating differential expression improved the detection of previously-identified psychiatric effects. We conclude that accounting for cell type can greatly improve the interpretability of microarray data.

1. Introduction

The human brain is a remarkable mosaic of diverse cell types stratified into rolling cortical layers, arching white matter highways, and interlocking deep nuclei. In the past decade, we have come to recognize the importance of this cellular diversity in even the most basic neural circuits. At the same time, we have developed the capability to comprehensively measure the thousands of molecules essential for cell function. These insights have provided conflicting priorities within the study of psychiatric illness: do we carefully examine individual molecules within their cellular and anatomical context or do we extract transcript or protein en masse to perform large-scale unbiased transcriptomic or proteomic analyses? In rodent models, researchers have escaped this dilemma by a boon of new technology: single cell laser capture, cell culture, and cell-sorting techniques that can provide sufficient extract for transcriptomic and proteomic analyses. However, single cell analyses of the human brain are far more challenging (1–3) – live tissue is only available in the rarest of circumstances (such as temporal lobe resection) and intact single cells are difficult to dissociate from post-mortem tissue without intensive procedures like laser capture microscopy.

Therefore, to date, the vast majority of unbiased transcriptomic analyses of the human brain have been conducted using macro-dissected, cell-type heterogeneous tissue. On Gene Expression Omnibus alone, there are at least 63[*] publicly-available macro-dissected post-mortem human brain tissue datasets, and many other macro-dissected human brain datasets are available to researchers via privately-funded portals (Stanley Medical Research Institute, Allen Brain Atlas, CommonMind Consortium). These datasets have provided us with novel hypotheses (e.g., (4,5)), but researchers who work with the data often report frustration with the relatively small number of candidate molecules that survive analyses using their painstakingly-collected samples, as well as the overwhelming challenge of interpreting molecular results in isolation from their respective cellular context. At the core of this issue is the inability to differentiate between (1) alterations in gene expression that reflect an overall disturbance in the relative

[*] As of 9-14-2017

ratio of the different cell types comprising the tissue sample, and (2) intrinsic dysregulation of one or more cell types, indicating perturbed biological function.

In this manuscript, we present results from an easily accessible solution to this problem that allows researchers to statistically estimate the relative number or transcriptional activity of particular cell types in macro-dissected human brain microarray data by tracking the collective rise and fall of previously identified cell type specific transcripts. Similar techniques have been used to successfully predict cell type content in human blood samples (6–9), as well as diseased and aged brain samples (10–12). Our method was specifically designed for application to large, highly-normalized human brain transcriptional profiling datasets, such as those commonly used by neuroscientific research bodies such as the Pritzker Neuropsychiatric Research Consortium and the Allen Brain Institute.

We took advantage of a series of newly available data sources depicting the transcriptome of known cell types, and applied them to infer the relative balance of cell types in our tissue samples in a semi-supervised fashion. We draw from seven large studies detailing cell-type specific gene expression in a wide variety of cells in the forebrain and cortex (2,13–18). Our analyses include all major categories of cortical cell types (17), including two overarching categories of neurons that have been implicated in psychiatric illness (19): projection neurons, which are large, pyramidal, and predominantly excitatory, and interneurons, which are small and predominantly inhibitory (20). These are accompanied by the three prevalent forms of glia that make up the majority of cells in the brain: oligodendrocytes, which provide the insulating myelin sheath that enhances electrical transmission in axons (21), astrocytes, which help create the blood-brain barrier and provide structural and metabolic support for neurons, including extracellular chemical and electrical homeostasis, signal propagation, and response to injury (21), and microglia, which serve as the brain's resident macrophages and provide an active immune response (21). We also incorporate structural and vascular cell types: endothelial cells, which line the interior surface of blood vessels, and mural cells (smooth muscle cells and pericytes), which regulate blood flow (22). Progenitor cells were also included in our analysis because they are widely regarded as important for the pathogenesis of mood disorders (23). Within the cortex, these cells mostly take the form of immature oligodendrocytes (17). Finally, the primary cells found in blood, erythrocytes or red blood cells (RBCs), carry essential oxygen throughout the brain. These cells do not contain a cell nucleus and do not generate new RNA, but still contain an existing, highly-specialized transcriptome (24). The relative presence of these cells could arguably represent overall blood flow, the functional marker of regional neural activity traditionally used in human imaging studies.

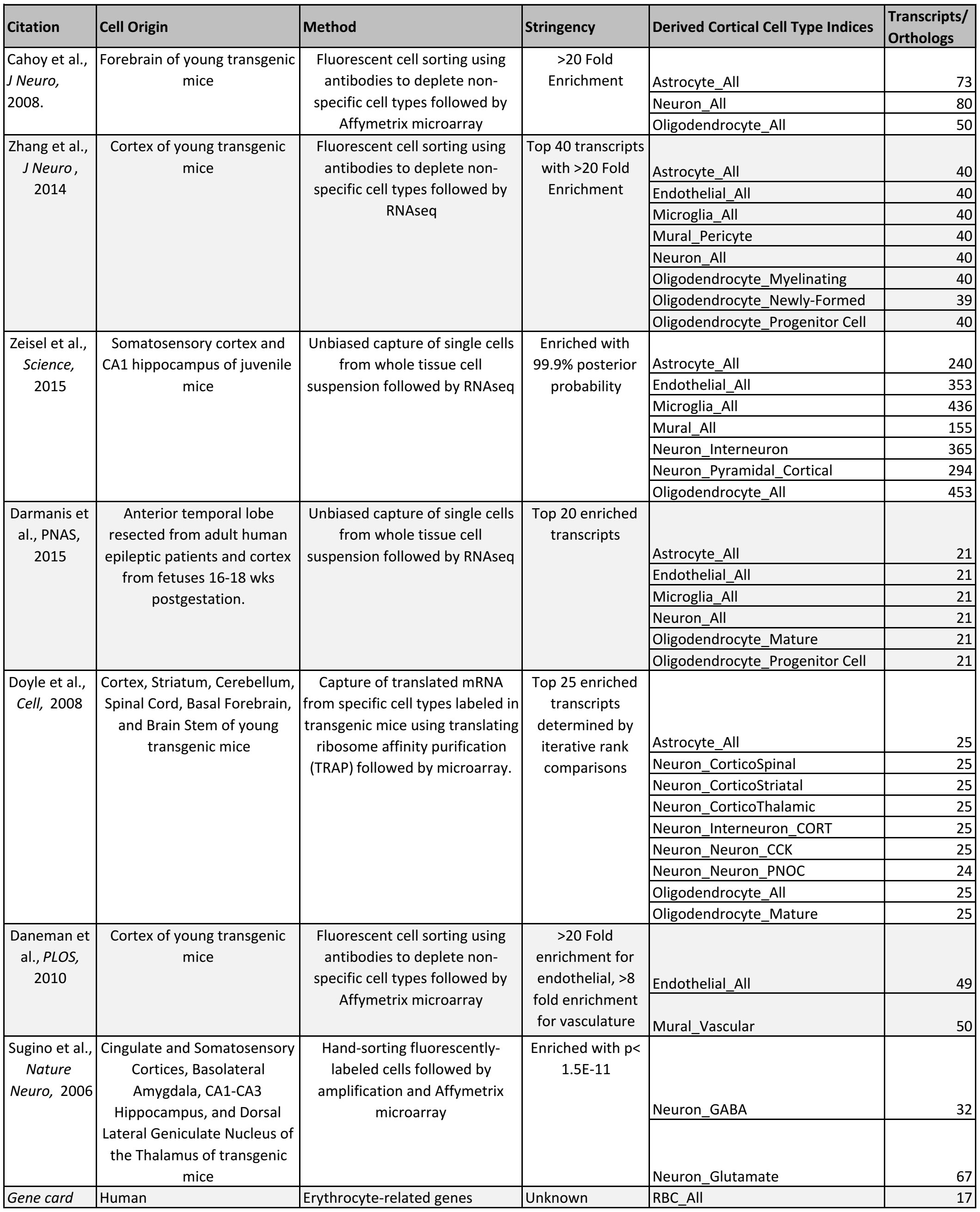
To characterize the balance of these cell types in psychiatric samples, we first compared the predictive value of cell type specific transcripts identified by diverse data sources and then summarized their collective predictions of relative cell type balance into covariates that could be used in larger linear regression models. We find that these “cell type indices” can successfully predict relative cell content in validation datasets, including *in vitro* and post-mortem datasets. We discover that the variability in the relative cell type balance of samples can explain a large percentage of the variation in macro-dissected human brain microarray and RNA-Seq datasets. This variability is driven by pre- and post-mortem subject variables, such as age, aerobic environment, and large scale blood loss, in addition to dissection. Finally, we demonstrate that this method enhances our ability to discover and interpret psychiatric effects in human brain microarray datasets, uncovering known changes in cell type balance in relationship to Major Depressive Disorder and Schizophrenia and potentially increasing our sensitivity to detect genes with previously-identified relationships to Bipolar Disorder and Schizophrenia in datasets that contain samples with highly-variable cell content.

2. Methods & Validation

2.1 Compiling a Database of Cell Type Specific Transcripts

To perform this analysis, we compiled a database of several thousand transcripts that were specifically-enriched in one of nine primary brain cell types within seven published single-cell or purified cell type transcriptomic experiments using mammalian brain tissues (2,13–18) (**Suppl. Table 1**). These primary brain cell types included six types of support cells: astrocytes, endothelial cells, mural cells, microglia, immature and mature oligodendrocytes, as well as two broad categories of neurons (interneurons and projection neurons). We also included a category for neurons that were generically extracted ("neuron_all"). The experimental and statistical methods for determining whether a transcript was enriched in a particular cell type varied by publication (**Figure 1**), and included both RNA-Seq and microarray datasets. We focused on cell-type specific transcripts identified using cortical or forebrain samples because the data available for these brain regions was more plentiful than for the deep nuclei or the cerebellum. In addition, we artificially generated a list of 17 transcripts specific to erythrocytes (red blood cells or RBC) by searching Gene Card for erythrocyte and hemoglobin-related genes (http://www.genecards.org/).

In all, we curated gene expression signatures for 10 cell types expected to account for most of the cells in the cortex. Our final database included 2499 unique human-derived or orthologous (as predicted by HCOP using 11 available databases: http://www.genenames.org/cgi-bin/hcop) transcripts, with a focus on coding varieties. We have made this database publicly accessible within our R package (https://github.com/hagenaue/BrainInABlender) and as a downloadable spreadsheet (https://sites.google.com/a/umich.edu/megan-hastings-hagenauer/home/cell-type-analysis).



Citation	Cell Origin	Method	Stringency	Derived Cortical Cell Type Indices	Transcripts/ Orthologs
Cahoy et al., *J Neuro,* 2008.	Forebrain of young transgenic mice	Fluorescent cell sorting using antibodies to deplete non-specific cell types followed by Affymetrix microarray	>20 Fold Enrichment	Astrocyte_All	73
				Neuron_All	80
				Oligodendrocyte_All	50
Zhang et al., *J Neuro* , 2014	Cortex of young transgenic mice	Fluorescent cell sorting using antibodies to deplete non-specific cell types followed by RNAseq	Top 40 transcripts with >20 Fold Enrichment	Astrocyte_All	40
				Endothelial_All	40
				Microglia_All	40
				Mural_Pericyte	40
				Neuron_All	40
				Oligodendrocyte_Myelinating	40
				Oligodendrocyte_Newly-Formed	39
				Oligodendrocyte_Progenitor Cell	40
Zeisel et al., *Science,* 2015	Somatosensory cortex and CA1 hippocampus of juvenile mice	Unbiased capture of single cells from whole tissue cell suspension followed by RNAseq	Enriched with 99.9% posterior probability	Astrocyte_All	240
				Endothelial_All	353
				Microglia_All	436
				Mural_All	155
				Neuron_Interneuron	365
				Neuron_Pyramidal_Cortical	294
				Oligodendrocyte_All	453
Darmanis et al., PNAS, 2015	Anterior temporal lobe resected from adult human epileptic patients and cortex from fetuses 16-18 wks postgestation.	Unbiased capture of single cells from whole tissue cell suspension followed by RNAseq	Top 20 enriched transcripts	Astrocyte_All	21
				Endothelial_All	21
				Microglia_All	21
				Neuron_All	21
				Oligodendrocyte_Mature	21
				Oligodendrocyte_Progenitor Cell	21
Doyle et al., *Cell,* 2008	Cortex, Striatum, Cerebellum, Spinal Cord, Basal Forebrain, and Brain Stem of young transgenic mice	Capture of translated mRNA from specific cell types labeled in transgenic mice using translating ribosome affinity purification (TRAP) followed by microarray.	Top 25 enriched transcripts determined by iterative rank comparisons	Astrocyte_All	25
				Neuron_CorticoSpinal	25
				Neuron_CorticoStriatal	25
				Neuron_CorticoThalamic	25
				Neuron_Interneuron_CORT	25
				Neuron_Neuron_CCK	25
				Neuron_Neuron_PNOC	24
				Oligodendrocyte_All	25
				Oligodendrocyte_Mature	25
Daneman et al., *PLOS,* 2010	Cortex of young transgenic mice	Fluorescent cell sorting using antibodies to deplete non-specific cell types followed by Affymetrix microarray	>20 Fold enrichment for endothelial, >8 fold enrichment for vasculature	Endothelial_All	49
				Mural_Vascular	50
Sugino et al., *Nature Neuro,* 2006	Cingulate and Somatosensory Cortices, Basolateral Amygdala, CA1-CA3 Hippocampus, and Dorsal Lateral Geniculate Nucleus of the Thalamus of transgenic mice	Hand-sorting fluorescently-labeled cells followed by amplification and Affymetrix microarray	Enriched with p< 1.5E-11	Neuron_GABA	32
				Neuron_Glutamate	67
Gene card	Human	Erythrocyte-related genes	Unknown	RBC_All	17

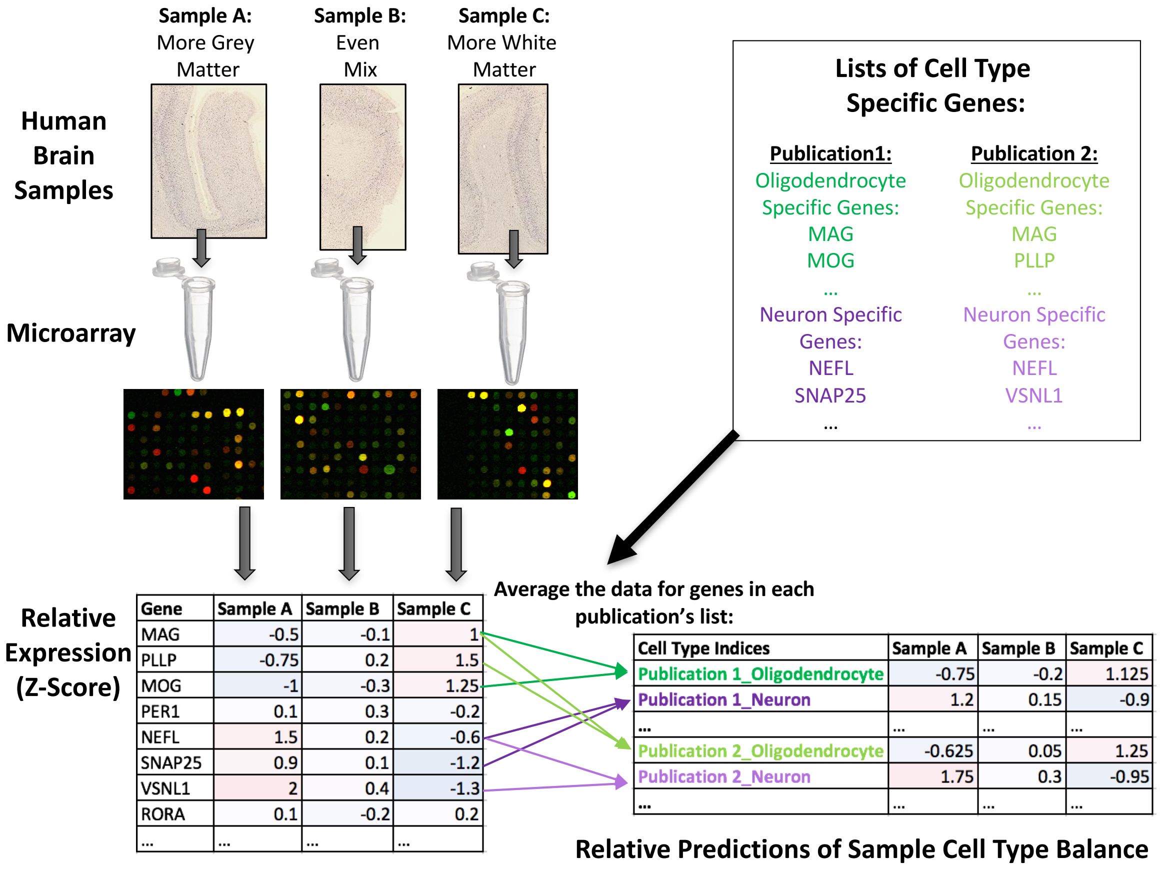
Figure 1. Thousands of transcripts have been identified as specifically-enriched in particular cortical cell types within published single-cell or purified cell type transcriptomic experiments ("reference datasets")*. The experimental and statistical methods for determining whether a transcript was enriched in a cell type varied by publication, and included both RNA-Seq and microarray datasets.*

2.2 “*BrainInABlender”:* Employing the Database of Cell Type Specific Transcripts to Predict Relative Cell Type Balance in Heterogenous Brain Samples

Next, we designed a method that uses the collective expression of cell type specific transcripts in brain tissue samples to predict the relative cell type balance of the samples (“BrainInABlender”). We specifically designed our method to be compatible with large, highly-normalized human brain transcriptional profiling datasets such as those used by our neuropsychiatric research consortium (Pritzker). We have made our method publicly-available in the form of a downloadable R package (https://github.com/hagenaue/BrainInABlender).

In brief, BrainInABlender extracts the data from any particular transcriptional profiling dataset (microarray, RNA-Seq) that represent genes identified in our database as having cell type specific expression in the brain (as curated by official gene symbol). The expression-level data for each of these transcripts (RNA-Seq: gene-level summary, microarray: probe or probeset summary) are then centered and scaled across samples (mean=0, sd=1) to prevent transcripts with more variable signal from exerting disproportionate influence on the results. Then, if necessary, the normalized data from all transcripts representing the same gene are averaged for each sample and re-normalized. Finally, for each sample, these values are averaged across the genes identified as having expression specific to a particular cell type in each publication included in the database of cell transcripts. This creates 38 cell type signatures derived from the cell type specific genes identified by the eight publications ("Cell Type Indices", **Figure 1**), each of which predicts the relative content for one of the 10 primary cell types in our brain samples (**Figure 2**).

Please note that our method was specifically designed to tackle challenges present in the Pritzker Consortium microarray data, but we later discovered that it bears some resemblance to the existing method of Population Specific Expression Analysis (PSEA, (10–12)). A more detailed discussion of the similarities and differences between the techniques can be found in **Section 7.2.**



Gene	Sample A	Sample B	Sample C
MAG	-0.5	-0.1	1
PLLP	-0.75	0.2	1.5
MOG	-1	-0.3	1.25
PER1	0.1	0.3	-0.2
NEFL	1.5	0.2	-0.6
SNAP25	0.9	0.1	-1.2
VSNL1	2	0.4	-1.3
RORA	0.1	-0.2	0.2
...	...	...	...

Cell Type Indices	Sample A	Sample B	Sample C
Publication 1_Oligodendrocyte	-0.75	-0.2	1.125
Publication 1_Neuron	1.2	0.15	-0.9
...	...	...	...
Publication 2_Oligodendrocyte	-0.625	0.05	1.25
Publication 2_Neuron	1.75	0.3	-0.95
...	...	...	...

Figure 2. Predicting the relative cell type balance in human brain samples using genes previously-identified as having cell type specific expression*. Within macro-dissected brain tissue samples, variable cell type balance is likely to influence the pattern of gene expression. To estimate this variability, we extracted the microarray data for probe sets representing genes that had been previously identified as having cell type specific expression in previous publications ("Lists of Cell Type Specific Genes",* ***Figure 1****) and then averaged across the transcripts identified as specific to a particular cell type in each publication to create 38 different "Cell Type Indices" that predicted relative cell content in each of the brain samples.*

2.3 Validation of Relative Cell Content Predictions Using Datasets Derived from Purified or Cultured Cells

We validated the method using publicly-available datasets from purified cell types and artificial cell mixtures (***Supplementary Methods and Results***). We found that the statistical cell type indices easily predicted the cell type identities of purified samples (datasets GSE52564 and GSE6783; (2,18); **Suppl. Figure 1, Suppl. Figure 2**). This was true regardless of the publication from which the cell type specific genes were derived: cell type specific gene lists derived from publications using different species (human vs. mouse), platforms (microarray vs. RNA-Seq), methodologies (florescent cell sorting vs. suspension), or statistical stringency all performed fairly equivalently, with some minor exception. Occasionally, we found that the cell type indices associated with cell type specific gene lists derived from TRAP methodology (15) did not properly predict the cell identity of the samples. In general the cell type indices associated with immature oligodendrocytes were somewhat inconsistent, most likely due to their dependency on developmental stage and experimental conditions.

Therefore, overall we found substantial support for simply averaging the individual publication-specific cell type indices within each of ten primary categories (astrocytes, endothelial cells, mural cells, microglia, immature and mature oligodendrocytes, red blood cells, interneurons, projection neurons, and indices derived from neurons in general) to produce ten consolidated primary cell-type indices for each sample. To perform this consolidation, we also removed any transcripts that were identified as “cell type specific” to multiple primary cell type categories (**Suppl. Figure 5**). These consolidated indices are included as a output from BrainInABlender.

Next, as further validation, we determined whether relative cell type balance could be accurately deciphered from microarray data for samples containing artificially-generated mixtures of cultured cells (GSE19380; (12)). We found that the consolidated cell type indices produced by BrainInABlender strongly correlated with the actual percentage of cells of a particular type included in the artificial mixtures (**Figure 3**, Neuron% vs. Neuron_All Index: R-squared=0.93, p=1.54e-15, Astrocyte % vs. Astrocyte Index: R-squared=0.77, p=5.05e-09, Microglia% vs. Microglia Index: R-Squared=0.64, p= 8.2e-07), although we found that the cell type index for immature oligodendrocytes better predicted the percentage of cultured oligodendrocytes in the samples than the cell type index for mature oligodendrocytes (Mature: R-squared=0.45, p=0.000179, Immature: R-squared=0.81, p= 4.14e-10). We believe this discrepancy is likely to reflect the specific cell culture conditions used in the original admixture experiment. In a follow-up analysis, artificial mixtures of cells produced *in silico* by averaging randomly-selected data from purified cell types similarly indicated that the cell type indices produced by

BrainInABlender follow a linear relationship with actual cell type balance in mixed samples, even for less prevalent cell varieties (endothelial, **Suppl. Figure 3**, **Suppl. Figure 4**).

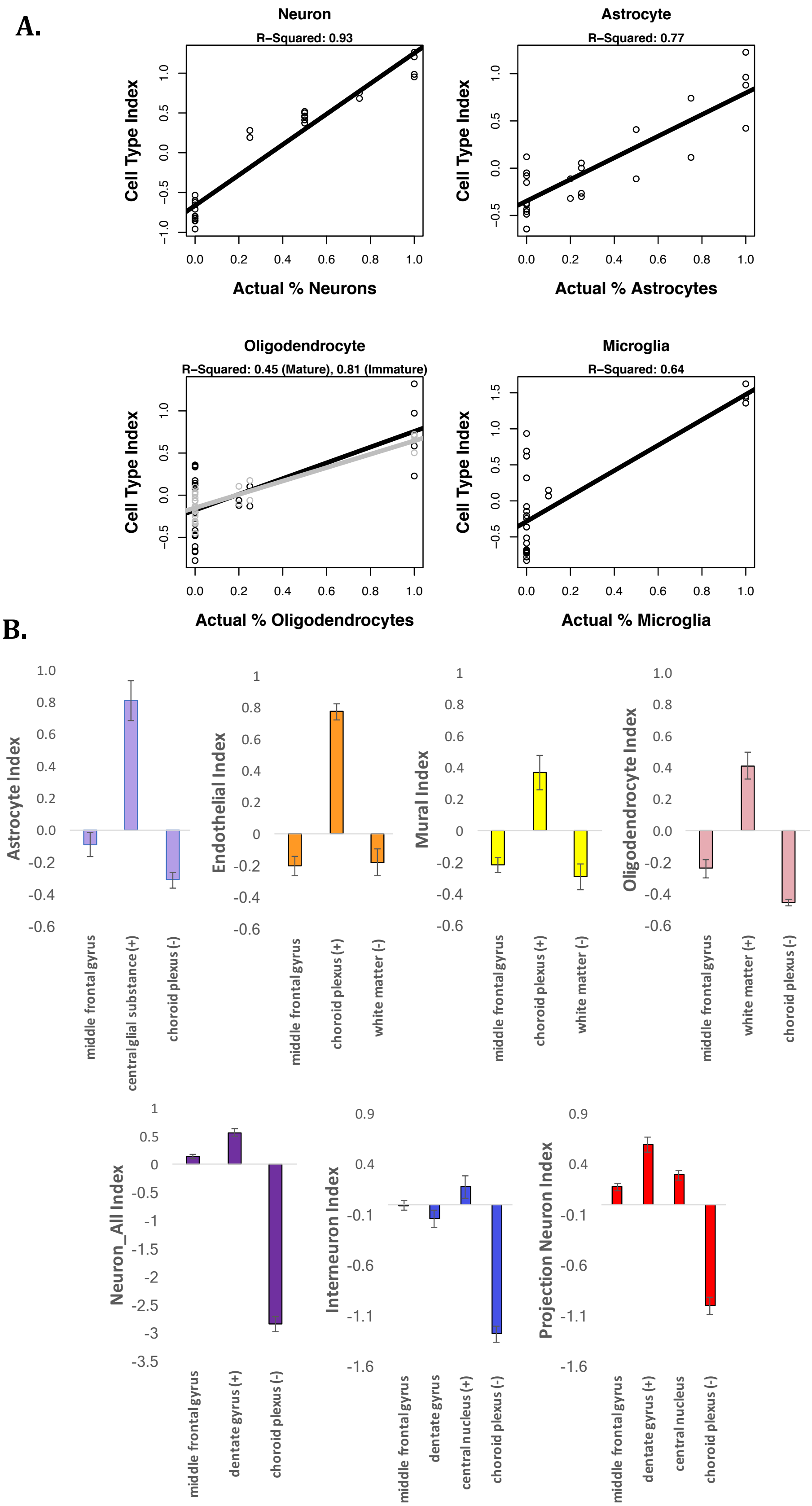

A.
Neuron
R–Squared: 0.93
Cell Type Index
Actual % Neurons
Astrocyte
R–Squared: 0.77
Cell Type Index
Actual % Astrocytes
Oligodendrocyte
R–Squared: 0.45 (Mature), 0.81 (Immature)
Cell Type Index
Actual % Oligodendrocytes
Microglia
R–Squared: 0.64
Cell Type Index
Actual % Microglia
B.
Astrocyte Index
middle frontal gyrus
central glial substance (+)
choroid plexus (-)
Endothelial Index
middle frontal gyrus
choroid plexus (+)
white matter (-)
Mural Index
middle frontal gyrus
choroid plexus (+)
white matter (-)
Oligodendrocyte Index
middle frontal gyrus
white matter (+)
choroid plexus (-)
Neuron_All Index
middle frontal gyrus
dentate gyrus (+)
choroid plexus (-)
Interneuron Index
middle frontal gyrus
dentate gyrus
central nucleus (+)
choroid plexus (-)
Projection Neuron Index
middle frontal gyrus
dentate gyrus (+)
central nucleus
choroid plexus (-)

Figure 3. Validation of Relative Cell Content Predictions. *A) Using a microarray dataset derived from samples that contained artificially-generated mixtures of cultured cells (GSE19380; (12)), we found that our relative cell content predictions ("cell type indices") closely reflected actual known content. B) Our cell type indices also easily differentiated human post-mortem samples derived from brain regions that are known to contain relatively more (+) or less (-) of the targeted the cell type of interest. Results from the middle frontal gyrus are included for comparison, since the rest of the paper primarily focuses on prefrontal cortical data. (Bars: average +/-SE).*

2.4 Validation of Relative Cell Content Predictions Using a Dataset Derived from Human Post-Mortem Tissue

Next, we wanted to see whether the cell content predictions produced by BrainInABlender correctly reflected relative cell type balance in human post-mortem samples. To test this, we applied our method to a large human post-mortem Agilent microarray dataset (841 samples) spanning 160 cortical and subcortical brain regions from the Allen Brain Atlas (**Suppl. Table 2**; (25)). This dataset was derived from high-quality tissue (absence of neuropathology, pH>6.7, post-mortem interval<31 hrs, RIN>5.5) from 6 human subjects (26). The tissue samples were collected using a mixture of block dissection and laser capture microscopy guided by adjacent tissue sections histologically stained to identify traditional anatomical boundaries (27). Prior to data release, the dataset had been subjected to a wide variety of normalization procedures to eliminate technical variation (28) which included log(base2) transformation, centering and scaling for each probe (http://human.brain-map.org/microarray/search, December 2015).

After applying BrainInABlender to the dataset, we extracted the results for a selection of brain regions that are known to contain relatively more (+) or less (-) of particular cell types (the results for the other brain regions can be found in **Suppl. Table 3**). The results clearly indicated that our cell type analyses could identify well-established differences in cell type balance across brain regions (**Figure 3**). Within the choroid plexus, which is a villous structure located in the ventricles made up of support cells (epithelium) and an extensive capillary network (29), there was an elevation of gene expression specific to vasculature (endothelial cells, mural cells). In the corpus callosum and cingulum bundle, which are large myelinated fiber tracts (29), there was an enrichment of oligodendrocytes- and microglia-specific gene expression. The central glial substance was enriched with gene expression specific to glia and support cells, with a particular emphasis on astrocytes. The dentate gyrus, which contains densely packed glutamatergic granule cells projecting to the mossy fibre pathway (30), was enriched for gene expression specific to projection neurons. The central nucleus of the amygdala, which includes a large number of GABA-ergic neurons (31), had a slight enrichment of gene expression specific to interneurons. These results provide fundamental validation that our methodology can accurately predict relative cell type balance in human post-mortem samples. Moreover, these results suggest that each of the consolidated cell type indices is capable of generally tracking their respective cell types in subcortical structures, despite the fact that our analysis method relies on cell type specific genes originally identified in the forebrain and cortex.

2.5 Using Cell Type Specific Transcripts to Predict Relative Cell Content in Transcriptomic Data from Macro-Dissected Human Cortical Tissue from Psychiatric Subjects

Next, we examined the collective variation in the levels of cell type specific transcripts in several large psychiatric human brain microarray datasets. The first was a large Pritzker Consortium Affymetrix U133A microarray dataset derived from high-quality human post-mortem dorsolateral prefrontal cortex samples (final sample size of 157 subjects, **Suppl. Table 14**), including tissue from subjects without a psychiatric or neurological diagnosis ("Controls", n=71), or diagnosed with Major Depressive Disorder ("MDD", n=40), Bipolar Disorder ("BP", n=24), or Schizophrenia ("Schiz", n= 22). The severity and duration of physiological stress at the time of death was estimated by calculating an agonal factor score for each subject (ranging from 0-4, with 4 representing severe physiological stress; (32,33)). Additionally, we measured the pH of cerebellar tissue as an indicator of the extent of oxygen deprivation experienced around the time of death (32,33) and calculated the interval between the estimated time of death and the freezing of the brain tissue (the postmortem interval or PMI) using coroner records. The transcriptional profiling of these samples had originally been processed in batches across multiple laboratories (1-5 replicates per sample). Before averaging the replicate samples for each subject, the data was highly normalized to correct for technical variation, including robust multi-array analysis (RMA) (34) and median-centering (detailed procedure: (35)). Our current analyses began with this subject-level summary gene expression data (GSE92538).

We determined the replicability of our results using three smaller publicly-available post-mortem human cortical Affymetrix U133Plus2 microarray datasets (GSE53987 (36), GSE21935 (37), GSE21138 (38), **Figure 4**). These datasets were selected because they included both psychiatric and control samples, and provided pH, PMI, age, and gender in the demographic information on the Gene Expression Omnibus website (https://www.ncbi.nlm.nih.gov/geo/). To control for technical variation, the sample processing batches were estimated using the microarray chip scan dates extracted from the .CEL files and RNA degradation was estimated using the R package AffyRNADegradation (39). Prior to running BrainInABlender, the probe-level signal data from each dataset was normalized using RMA (34), summarized using a custom .cdf (http://nmg-r.bioinformatics.nl/NuGO_R.html, “hgu133plus2hsentrezgcdf_19.0.0”), and cleaned of any samples that appeared low-quality or misidentified.

Finally, we also explored replicability within the recently-released large CommonMind Consortium (CMC) human dorsolateral prefrontal cortex RNA-seq dataset (603 individuals (40)). This dataset was downloaded from the CommonMind Consortium Knowledge Portal (https://www.synapse.org/CMC) and analyzed at UCI. The bam files were converted to fastq files and re-mapped to a more recent build of the human genome (GRCh38, (41)). The total reads mapping uniquely to exons (defined by Ensembl) were transformed into logCPM values (42). Prior to data upload, poor quality samples from the original dataset (40) were removed (<50 million reads, RIN<5.5) by the CMC and replaced with higher quality samples. We additionally excluded data from 10 replicates and 89 individuals with incomplete demographic data (missing pH), leaving a final sample size of 514 samples. We predicted the relative cell type content of these samples using a newer version of BrainInABlender (v2) which excluded a few of the weaker cell type specific gene sets (15). Later, the expression data were further filtered by expression threshold (CPM>1 in at least 50 individuals), leaving data from approximately 17,000 genes.

In general, the code for all analyses in the paper can be found at https://github.com/hagenaue/ or https://github.com/aschulmann/CMC_celltype_index.

Microarray:

GEO Accession #	Submitter & Date	Published?	Brain Bank	Brain Region	Sample Size (no outliers)	Subjects per group (no outliers)	# of replicates per sample	# of known Batches	AVE pH (+/- SD)	AVE Age (+/- SD)	AVE PMI (+/- SD)	% Female
GSE92538	Hagenauer (2016)	*current paper*	UC-Irvine	BA9/BA46	**337**	**157:** 71 CNTRL, 24 BP, 40 MDD, 22 SCHIZ	1-5 (AVE: 2)	15-36	**6.8** (+/-0.3)	**52** (+/-15)	**24** (+/-9)	**27%**
GSE53987	Lanz (2014)	Lanz et al. (2015)	PITT	BA46: *grey matter*	**66**	**66:** 18 CNTRL, 17 BP, 17 MDD, 14 SCHIZ	1	0	**6.6** (+/-0.3)	**46** (+/-10)	**20** (+/-6)	**45%**
GSE21935	Huxley-Jones (2011)	Barnes et al. (2011)	CCHPC	BA22	**42**	**42:** 19 CNTRL, 23 SCHIZ	1	2	**6.3** (+/-0.3)	**70** (+/-19)	**8** (+/-5)	**45%**
GSE21138	Thomas (2010)	Narayan et al. (2008)	MHRI	BA46: *grey matter*	**54**	**54:** 27 CNTRL, 27 SCHIZ	1	5	**6.3** (+/-0.2)	**45** (+/-17)	**40** (+/-13)	**17%**

RNA-Seq:

Public Data Release	Submitter & Date	Published?	Brain Bank	Brain Region	Sample Size (all)	Subjects per group (all)	# of replicates per sample	# of known Batches	AVE pH (+/- SD)	AVE Age (+/- SD)	AVE PMI (+/- SD)	% Female
Synapse.org	CMC (2016)	Fromer et al. (2016)	MSSM, PENN, PITT	BA9 / BA46 (PITT: *grey matter)*	**621**	**603:** 285 CNTRL, 263 SCZ, 47 BP, 8 AFF	1 (rarely 2)	491	**6.5** (+/- 0.3)	**65** (+/- 18) (binned 90+)	**17** +/- 11	**41%**

Figure 4. We examined the pattern of cell-type specific gene expression in five post-mortem human cortical tissue datasets that included samples from subjects with psychiatric illness. *Abbreviations: CTRL: control, BP: Bipolar Disorder, MDD: Major Depressive Disorder, SCHIZ: Schizophrenia, GEO: Gene Expression Omnibus, BA: Brodmann's Area, PMI: Post-mortem interval, SD: Standard Deviation, Brain Banks: UC-Irvine (University of California – Irvine), PITT (University of Pittsburgh), CCHPC (Charing Cross Hospital Prospective Collection), MSSM (Mount Sinai Icahn School of Medicine), MHRI (Mental Health Research Institute Australia), PENN (University of Pennsylvania)*

2.6 Does the Reference Dataset Matter? Cell Type Specific Transcripts Identified by Different Publications Produce Similar Predictions of Relative Cell Type Balance

We first confirmed that the predicted cell content for our post-mortem human cortical samples ("cell type indices") was similar regardless of the methodology used to generate the cell type specific gene lists used in the predictions. Within all five of the human cortical transcriptomic datasets, there was a strong positive correlation between cell type indices representing the same cell type, even when the predictions were derived using cell type specific gene lists from different species, cell type purification strategies, and platforms. Clustering within broad cell type categories was clear using visual inspection of the correlation matrices (**Suppl. Figure 11**, **Suppl. Figure 12**), hierarchical clustering, or consensus clustering (**Suppl. Figure 13**, ConsensusClusterPlus: (43), **Suppl. Figure 14**, **Suppl. Figure 16**). In some datasets, the cell type indices for support cell subcategories were nicely clustered and in others they were difficult to fully differentiate (**Suppl. Figure 11, Suppl. Figure 12**). Clustering was not able to reliably discern neuronal subcategories (interneurons, projection neurons) in any dataset. Similar to our previous validation analyses, oligodendrocyte progenitor cell indices derived from different publications did not strongly correlate with each other, perhaps due to heterogeneity in the progenitor cell types sampled by the original publications.

Therefore, for further analyses in the post-mortem human datasets, we consolidated the cell type indices using a procedure similar to our previous validation analyses. To do this, we averaged the 38 publication-specific cell type indices within each of ten primary categories: astrocytes, endothelial cells, mural cells, microglia, immature and mature oligodendrocytes, red blood cells, interneurons, projection neurons, and indices derived from neurons in general, with any transcripts that overlapped between categories removed (**Suppl. Figure 17**). This led to ten consolidated primary cell-type indices for each sample.

3. Results

3.1 Inferred Cell Type Composition Explains a Large Percentage of the Sample-Sample Variability in Microarray Data from Macro-Dissected Human Cortical Tissue

Using principal components analysis we found that the primary gradients of variation in all four of the cortical datasets strongly correlated with our estimates of cell type balance. For example, while analyzing the Pritzker dorsolateral prefrontal cortex microarray dataset, we found that the first principal component, which encompassed 23% of the variation in the dataset, spanned from samples with high support cell content to samples with high neuronal content. Therefore, a large percentage of the variation in PC1 (91%) was accounted for by an average of the astrocyte and endothelial indices (p<2.2e-82, with a respective r-squared of 0.80 and 0.75 for each index analyzed separately) or by the general neuron index (p<6.3e-32, r-squared=0.59; **Figure 5**). The second notable gradient in the dataset (PC2) encompassed 12% of the variation overall, and spanned samples with high projection neuron content to samples with high oligodendrocyte content (with a respective r-squared of 0.62 and 0.42, and respective p-values of p<8.5e-35 and p<8.7e-20).

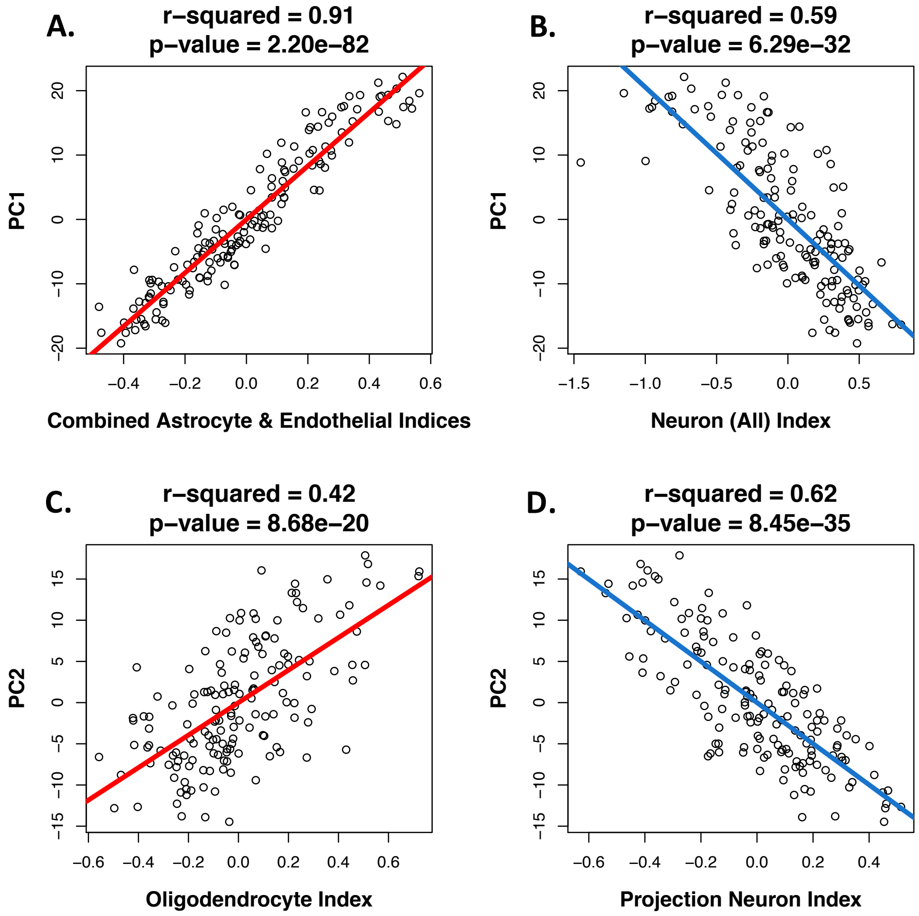


Figure 5. Cell content predictions explain a large percentage of the variability in microarray data derived from the human cortex. *As an example, within the Pritzker dataset the first principal component of variation (PC1) encompassed 23% of the variation in the dataset, and was A) positively correlated with predicted "support cell" content in the samples (a combination of the astrocyte and endothelial indices: r-squared: 0.91, p<2.2e-82) and B) negatively correlated with predicted neuronal content (r-squared=0.59, p<6.3e-32). The second principal component of variation (PC2) encompassed 12% of variation in the dataset, and was C) positively correlated with predicted oligodendrocyte content in the*

samples (r-squared: 0.42, p<8.7e-20) and D) negatively correlated with predicted projection neuron content (r-squared: 0.62, p<8.5e-35). Examples in other datasets can be found in ***Suppl. Figure 20****.*

When digging deeper, we found that none of the original 38 publication-specific cell type indices were noticeably superior to the consolidated indices when predicting the principal components of variation in the dataset. Human-derived indices did not outperform mouse-derived indices, and indices derived from studies using stricter definitions of cell type specificity (fold enrichment cut-off in **Figure 1,** e.g., (13) vs. (17)) did not outperform less strict indices.

Within the other four human cortical tissue datasets, the relationships between the top principal components of variation and the consolidated cell type indices were similarly strong (**Suppl. Figure 20**), despite the fact that these datasets had received less preprocessing to remove the effects of technical variation. Within the GSE21935 dataset (published in (37)) the first principal component of variation accounted for 37% of the variation in the dataset and similarly seemed to represent a gradient running from samples with high support cell content (PC1 vs. endothelial index: r-squared= 0.85, p<3.6e-18, PC1 vs. astrocyte index: r-squared= 0.67, p<3.6e-11) to samples with high neuronal content (PC1 vs. neuron_all index: r-squared= 0.85, p<3.9e-18). Within the GSE53987 dataset (submitted to GEO by Lanz, 2014), which had samples derived exclusively from gray-matter-only dissections, the first principal component of variation accounted for 13% of the variation in the dataset and was highly correlated with predicted astrocyte content (PC1 vs. astrocyte index: r-squared=0.80, p<4.6e-24). In GSE21138 (published in (39)), which also had samples derived exclusively from gray-matter-only dissections, the first principal component of variation accounted for 23% of the variation in the dataset and was strongly related to technical variation (batch), but the second principal component of variation, which accounted for 14% of the variation in the dataset, again represented a gradient from samples with high support cell content to high neuronal content (PC2 vs. astrocyte: r-squared=0.56, p<8.3e-11, PC2 vs. neuron_all: r-squared=0.54, p<2.3e-10). Finally, within the CMC RNA-Seq dataset, the first principal component of variation accounted for 16% of the variation in the dataset and was highly correlated with projection neuron content (PC1 vs. Neuron_Projection: r-squared=0.54, p=5.77e-104).

To confirm that the strong relationship between the top principal components of variation and our cell type composition indices did not originate artificially due to cell type specific genes representing a large percentage of the most highly variable transcripts in the dataset, we repeated the principal components analysis in the Pritzker dataset after excluding all cell type specific transcripts from the dataset and still found these strong correlations (**Suppl. Figure 21**). Indeed, individual cell type indices better accounted for the main principal components of variation in the microarray data than *all other major subject variables combined* (pH, Agonal Factor, PMI, Age, Gender, Diagnosis, Suicide; PC1: R-squared=0.4272, PC2: R-squared=0.2176). When examining the dataset as a whole, the six subject variables accounted for an average of only 12% of the variation for any particular probe (R-squared, Adj.R-squared=0.0715), whereas just the astrocyte and projection neuron indices alone were able to account for 17% (R-squared, Adj.R-squared=0.1601) and all 10 cell types accounted for an average of 31% (R-squared, Adj.R-squared=0.263), almost one third of the variation present in the data for any particular probe (**Suppl. Figure 22**).

These results indicated that accounting for cell type balance is important for the interpretation of post-mortem human brain microarray and RNA-Seq data and might improve the signal-to-noise ratio in analyses aimed at identifying psychiatric risk genes.

3.2 Cell Content Predictions Derived from Microarray Data Match Known Relationships Between Clinical/Biological Variables and Brain Tissue Cell Content

We next set out to observe the relationship between the predicted cell content of our samples and a variety of medically-relevant subject variables, including variables that had already been demonstrated to alter cell content in the brain in other paradigms or animal models. To perform this analysis, we first examined the relationship between seven relevant subject variables and each of the ten cell type indices in the Pritzker prefrontal cortex dataset using a linear model that allowed us to simultaneously control for other likely confounding variables in the dataset:

Equation 1:

$$\text{Cell Type Index} = \beta 0 + \beta 1^{*}(\text{Brain pH}) + \beta 2^{*}(\text{Agonal Factor}) + \beta 3^{*}(\text{PMI}) + \beta 4^{*}(\text{Age}) + \beta 5^{*}(\text{Sex}) + \beta 6^{*}(\text{Diagnosis}) + \beta 7^{*}(\text{Exsanguination}) + \varepsilon$$

We then examined the replicability of these relationships using data from the three smaller publicly-available human post-mortem microarray datasets (GSE53987, GSE21935, GSE21138). For these datasets, we initially lacked detailed information about manner of death (agonal factor and exsanguination), but were able to control for technical variation within the model using statistical estimates of RNA degradation and batch (scan date):

Equation 2:

$$\text{Cell Type Index} = \beta 0 + \beta 1^{*}(\text{Brain pH}) + \beta 2^{*}(\text{PMI}) + \beta 3^{*}(\text{Age}) + \beta 4^{*}(\text{Sex}) + \beta 5^{*}(\text{Diagnosis}) + \beta 6^{*}(\text{RNA Degradation}) + \beta 7^{*}(\text{Batch, } \textit{when applicable}) + \varepsilon$$

We evaluated the replicability of these relationships across the four microarray datasets by performing a meta-analysis for each variable and cell type combination. To do this, we applied random effects modeling to the respective betas and accompanying sampling variance derived from each dataset using the *rma.mv()* function within the *metafor* package (44). P-values were then corrected for multiple comparisons following the Benjamini-Hochberg method (*q-value*) using the *mt.rawp2adjp* function within the *multtest* package (45).

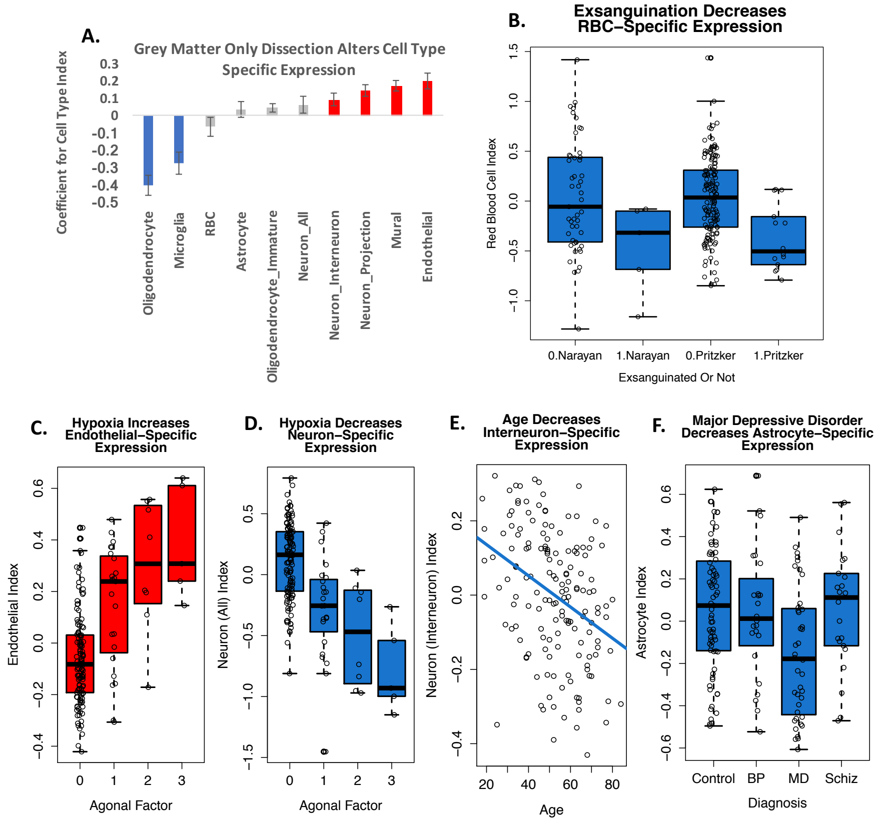
Finally, we characterized these relationships in the large CMC human post-mortem RNA-Seq dataset. For this dataset, we had some information about manner of death but lacked knowledge of agonal factor or exsanguination. We controlled for technical variation due to dissection site (institution) and RNA degradation (RIN):

Equation 3:

$$\text{Cell Type Index} = \beta 0 + \beta 1^{*}(\text{Brain pH}) + \beta 2^{*}(\text{PMI}) + \beta 3^{*}(\text{Age}) + \beta 4^{*}(\text{Sex}) + \beta 5^{*}(\text{Diagnosis}) + \beta 6^{*}(\text{RNA Degradation}) + \beta 7^{*}(\text{Institution}) + \beta 8^{*}(\text{MannerOfDeath}) + \varepsilon$$

This analysis uncovered many well-known relationships between brain tissue cell content and clinical or biological variables (**Figure 6, Suppl. Table 4**). First, as a proof of principle, we were able to clearly observe dissection differences between institutions within the large CMC RNA-Seq dataset, with samples from University of Pittsburgh having a predicted relative cell type balance that closely matched what would be expected due to their gray matter only dissection method (Oligodendrocyte: β =-0.404, p=2.42e-11, q=4.03e-10; Microglia: β =-0.274, p=3.06e-05, q=2.42e-04; Neuron_Interneuron: β=0.0916, p=0.0161, q=0.525; Neuron_Projection: β=0.145, p=2.31e-05, q=1.93e-04; Mural: β=0.170, p=2.14e-08, q=2.68e-07; Endothelial: β=0.200, p=1.12e-05, q=1.12e-04). Samples from University of Pennsylvania were associated with lower predicted cell content related to vasculature (Endothelial: β =-0.255, p=4.01-04, q=2.40e-03; Mural: β =-0.168, p=4.59e-04, q=2.59e-03; Astrocyte: β =-0.189, p=7.47e-03, q=0.0287).

Predicted cell type content was also closely related to manner of death. For example, within the Pritzker dataset we found that subjects who died in a manner that involved exsanguination had a notably low red blood cell index (β =-0.398; p=0.00056). Later, we were able replicate this result within GSE21138 using data from 5 subjects who we discovered were also likely to have died in a manner involving exsanguination (β =-0.516, p=0.052*trend,* manner of death reported in suppl. in *(38)*). The presence of prolonged hypoxia around the time of death, as indicated by either low brain pH or high agonal factor score within the Pritzker dataset, was associated with a large increase in the endothelial cell index (Agonal Factor: β=0.118 p=2.85e-07; Brain pH: β=-0.210, p= 0.0003) and astrocyte index (Brain pH: β=-0.437, p=2.26e-07; Agonal Factor: β=0.071, p=0.024), matching previous demonstrations of cerebral angiogenesis, endothelial and astrocyte activation and proliferation in low oxygen environments (46). Small increases were also seen in the mural index in response to low-oxygen (Mural vs. Agonal Factor: β= 0.0493493, p= 0.0286), most likely reflecting angiogenesis. In contrast, prolonged hypoxia was associated with a clear decrease in all of the neuronal indices (Neuron_All vs. Agonal Factor: β=-0.242, p=3.58e-09; Neuron_All vs. Brain pH: β=0.334, p=0.000982; Neuron_Interneuron vs. Agonal Factor: β=-0.078, p=4.13e-05; Neuron_Interneuron vs. Brain pH: β=0.102, p=0.034; Neuron_Projection vs. Agonal Factor: β=-0.096, p= 0.000188), mirroring the notorious vulnerability of neurons to low oxygen (e.g., (47)). These overall effects of hypoxia on cell type balance replicated in the smaller human microarray post-mortem datasets, despite lack of information about agonal factor (Astrocyte vs. Brain pH (meta-analysis: b= -0.459, p=2.59e-11, q=2.33e-09): Narayan et al. 2008: β= -0.856, p=0.00661, Lanz 2014: β=-0.461, p=0.00812, Neuron_All vs. Brain pH (meta-analysis: b= 0.245, p=7.72e-04, q=1.16e-02), Neuron_Interneuron vs. Brain pH (meta-analysis: b= 0.109, p=7.89e-03, q=5.52e-02**trend*): Narayan et al. 2008: β= 0.381134, p=0.0277) and partially replicated in the CMC human RNA-Seq dataset (Neuron_Interneuron vs. Brain pH: β=0.186, p=9.81e-05, q=6.69e-04). In several datasets, we also found that prolonged hypoxia correlated with fewer microglia (Microglia vs. Brain pH: Lanz 2014: β=0.462, p=0.00603; CMC: β=0.286, p=4.66e-04, q=2.59e-03), which may suggest that our microglia cell type index is specifically tracking ramified microglia, although we did not observe a relationship between microglia and death related to infection/parasitic disease (CMC: Microglia vs. CauseOfDeath(infection): β=0.231, p=0.121, q=0.256).



The most highly replicated effects across datasets:	Pritzker	Lanz 2014	Barnes et al. 2011	Narayan et al. 2008	Meta-Analysis: Nominal P-value	Meta-Analysis: BH Adjusted P-value	CMCC	CMCC: Nominal P-value	CMCC: BH Adjusted P-value	Replication in Microarray Meta-analysis and CMCC
Hypoxia:										
Astrocyte vs. Brain pH	**-5.4**	**-2.7**	-1.1	**-2.9**	2.59E-11	2.33E-09	**2.1**	3.25E-02	9.55E-02	
Neuron_All vs. Brain pH	**3.4**	0.1	1.3	1.9	7.72E-04	1.16E-02	0.6	5.60E-01	7.00E-01	*
Neuron_Interneuron vs. Brain pH	**2.1**	1.0	-0.5	**2.3**	7.89E-03	5.52E-02	**3.9**	9.81E-05	6.69E-04	**
Microglia vs. Brain pH	-1.0	**2.9**	-0.1	1.9	5.41E-01	7.91E-01	**3.5**	4.66E-04	2.59E-03	*
Oligodendrocyte_Immature vs. Brain pH	**-3.0**	0.4	-0.5	1.4	1.40E-01	3.71E-01	**3.0**	2.68E-03	1.22E-02	
Age										
Neuron_All vs. Age	-1.9	**-2.1**	-1.2	-1.2	1.57E-03	2.02E-02	**-4.3**	2.27E-05	1.93E-04	**
Neuron_Interneuron vs. Age	**-3.4**	-0.5	**-2.5**	**-2.5**	2.91E-06	6.56E-05	**-6.5**	2.10E-10	3.15E-09	**
Neuron_Projection vs. Age	**-2.8**	**-3.7**	-0.5	**-3.1**	1.61E-06	4.83E-05	**-7.5**	2.93E-13	7.33E-12	**
Oligodendrocyte vs. Age	1.8	1.2	0.3	**3.1**	2.74E-03	2.74E-02	1.6	1.02E-01	2.25E-01	*
Oligodendrocyte_Immature vs. Age	**-3.7**	**-4.7**	-0.2	**-4.5**	5.98E-11	2.69E-09	**-11.0**	3.32E-25	2.49E-23	**
PMI										
Oligodendrocyte vs. PMI	**-3.6**	**-3.6**	-1.6	-0.5	2.23E-05	4.02E-04	**-4.1**	4.70E-05	3.36E-04	**
Endothelial vs. PMI	-2.0	-0.8	0.4	-0.1	5.51E-02	2.36E-01	**-3.9**	1.32E-04	8.60E-04	*
Microglia vs. PMI	-1.0	-1.5	-1.2	-0.7	9.72E-02	3.05E-01	**-3.5**	5.15E-04	2.76E-03	*
Oligodendrocyte_Immature vs. PMI	**3.5**	1.0	1.6	-0.4	4.81E-03	4.33E-02	-0.4	6.86E-01	8.04E-01	
Neuron_Projection vs. PMI	**3.9**	1.6	-0.2	-1.2	2.28E-03	2.56E-02	**3.1**	1.97E-03	9.24E-03	**
Neuron_All vs. PMI	**2.5**	1.8	-0.4	0.0	1.74E-02	9.81E-02	**2.6**	1.10E-02	3.88E-02	*
Diagnosis:										
Astrocyte vs. Diagnosis_MDD	**-2.6**	-1.0			5.88E-03	4.81E-02				
Neuron_All vs. Diagnosis BP	-0.9	-0.7			2.46E-01	5.39E-01	**2.6**	8.44E-03	3.17E-02	
RBC vs. Diagnosis_Schiz	-1.2	0.1	-1.0	-0.5	2.04E-01	4.96E-01	**-2.5**	1.41E-02	4.71E-02	*
Gender:										
Neuron_Interneuron vs. GenderFemale	-0.9	0.0	0.7	0.3	6.65E-01	9.06E-01	**-2.5**	1.20E-02	4.09E-02	
Neuron_Projection vs. GenderFemale	1.0	-0.3	-0.4	1.6	3.60E-01	6.46E-01	**-2.5**	1.11E-02	3.88E-02	

Figure 6. Cell content predictions derived from microarray data match known relationships between subject variables and brain tissue cell content. *Boxplots represent the median and interquartile range, with whiskers illustrating either the full range of the data or 1.5x the interquartile range. A. Within the CMC dataset, cortical tissue samples that were dissected to only contain gray matter (PITT) show lower predicted oligodendrocyte and microglia content and more neurons and vasculature (bars: β+/- SE, red/blue: p<0.05). B. Subjects that died in a manner that involved exsanguination (n=14) had a notably low red blood cell index in both the Pritzker (p=0.00056) and Narayan et al. datasets (p=0.052**trend). *C. The presence of prolonged hypoxia around the time of death, as indicated by high agonal factor score, was associated with a large increase in the endothelial cell index (p=2.85e-07) matching previous demonstrations of cerebral angiogenesis, activation, and proliferation in low oxygen environments (46). D. High agonal factor was also associated with a clear decrease in neuronal indices (p=3.58e-09) mirroring the vulnerability of neurons to low oxygen (47). E. Age was associated with a decrease in the neuronal indices (p= 0.000956) which fits known decreases in gray matter density in the frontal cortex in aging humans (48). F. Major Depressive Disorder was associated with a moderate decrease in astrocyte index (p= 0.0118), which fits what has been observed morphometrically (49). G. The most highly-replicated relationships between subject variables and predicted cortical tissue cell content across all five of the post-mortem human datasets. Provided in the table are the T-stats for the effects (red=upregulation, blue=downregulation), derived from a larger linear model controlling for confounds* ***(Equation 1, Equation 2, Equation 3****), as well as the nominal p-values from the meta-analysis of the results across the four microarray studies, and p-values following multiple-comparisons correction (q-value). Only effects that had a q<0.05 in either our meta-analysis or the large CMC RNA-Seq dataset are included in the table. Asterisks denote effects that had consistent directionality in the meta-analysis and CMC dataset (*) or consistent directionality and q<0.05 in both datasets (**). Please note that lower pH and higher agonal factor are both indicators of greater hypoxia prior to death, but have an inverted relationship and therefore show opposing relationships with the cell type indices (e.g., when pH is low and agonal factor is high, support cell content is increased).*

In the Pritzker dataset, age was associated with a moderate decrease in two of the neuronal indices (Neuron_Interneuron vs. Age: β=- -0.00291, p= 0.000956; Neuron_Projection Neuron vs. Age: β=-0.00336, p=0.00505) and was strongly replicated in the large CMC RNA-Seq dataset (Neuron_All vs. Age: β=-0.00497, p=2.27e-05, q=1.93e-04; Neuron_Projection Neuron vs. Age: β=-0.00612, p=2.93e-13, q=7.33e-12; Neuron_Interneuron vs. Age: β=-0.00591, p=2.10e-10, q=3.15e-09). A similar decrease in predicted neuronal content was seen in all three of the smaller human post-mortem datasets (Neuron_All vs. Age (meta-analysis: b=-0.00415, p=1.57e-03, q=2.02e-02): Lanz 2014: β=-0.00722, p=0.0432, Neuron_Interneuron vs. Age (meta-analysis: b=-0.00335, p=2.91e-06, q=6.56e-05): Narayan et al. 2008: β=-0.00494, p=0.0173, Barnes et al. 2011: β=-0.00506, p=0.0172, Neuron_Projection vs. Age (meta-analysis: b=-0.00449, p=1.61e-06, q=4.83e-05): Lanz 2014: β=-0.0103, p=0.000497, Narayan et al. 2008: β=-0.00763, p=0.00386). This result fits with known decreases in gray matter density in the frontal cortex in aging humans (48), as well as age-related sub-region specific decreases in frontal neuron numbers in primates (50) and rats (51). There was also a consistent decrease in immature oligodendrocytes in relationship to age across datasets (Oligodendrocyte_Immature vs. Age (meta-analysis: b=-0.00514, p=5.98e-11, q=2.69e-09): Pritzker: β=-0.00432, p=0.000354, Narayan et al. 2008: β=-0.00721, p=5.73e-05, Lanz 2014: β=-0.00913, p=1.85e-05; CMCC: β=-0.00621, p=3.32e-25, q=2.49e-23), which seems intuitive, but actually contradicts animal studies on the topic (52). Since the validation of the Oligodendrocyte_Immature index was relatively weak, this result should perhaps be considered with caution.

Other non-canonical relationships between subject variables and predicted cell content can be found in the tables in **Figure 6.** In some datasets, there appears to be an increase in oligodendrocyte index with age (Oligodendrocyte vs. Age (meta-analysis: b=0.00343, p=2.74e-03, q=2.74e-02): Narayan et al. 2008, β= 0.00957, p=0.00349) which, at initial face value, seems to contrast with well-replicated observations that frontal white matter decreases with age in human imaging studies (48,53,54). However, it is worth noting that several histological studies in aging primates suggest that brain regions that are experiencing demyelination with age actually show an *increasing* number of oligodendrocytes, which is thought to be driven by the need for repair (52,55).

Another prominent unexpected effect was a large decrease in the oligodendrocyte index with longer post-mortem interval (Oligodendrocyte vs. PMI (meta-analysis: b=-0.00764, p=2.23e-05, 4.02e-04): Pritzker: β= -0.00749, p=0.000474, Lanz 2014: β= -0.0318, p=0.000749; CMC: β=-0.00759, p=4.70e-05, q=3.36e-04). Upon further investigation, we found a publication documenting a 52% decrease in the fractional anisotropy of white matter with 24 hrs post-mortem interval as detected by neuroimaging (56), but to our knowledge the topic is otherwise not well studied. These changes were paralleled by a decrease in endothelial cells (CMC: β=-0.00542, p=1.32e-04, q=8.60e-04) and microglia (CMC: β=-0.00710, p=5.15e-04, q=2.76e-03) and relative increase in immature oligodendrocytes (Oligodendrocyte_Immature vs. PMI (meta-analysis: b=0.00353, p=4.81e-03, q=4.33e-02): Pritzker: β= 0.00635, p= 0.000683) and neurons (Neuron_All vs. PMI: Pritzker: β= 0.006997, p= 0.000982; CMC: β=0.00386, p=0.0110, q=0.0388 ; Neuron_Projection vs. PMI (meta-analysis: b=0.00456, p=2.28e-03, q=2.56e-02): Pritzker:

β= 0.00708, p=1.64e-04; CMC: β=0.00331, p=0.00197, q=0.00924). This result could arise from the zero-sum nature of microarray analysis: due to the use of a standardized dissection size, RNA concentration, and data normalization, if there are large decreases in gene expression for one common variety of cell type in relationship to post-mortem interval (oligodendrocytes), then gene expression related to other cell types may appear increase.

Overall, these results indicate that statistical predictions of the cell content of samples effectively capture many known biological changes in cell type balance, and imply that within both chronic (age) and acute conditions (agonal, PMI, pH) there is substantial influences upon the relative representation of different cell types. Thus, when interpreting microarray data, it is as important to consider the chronic and acute demographic factors at the population level as well as cellular functional regulation.

3.3 Cell Type Balance Changes in Response to Psychiatric Diagnosis

Of most interest to us were potential changes in cell type balance in relation to psychiatric illness. In previous post-mortem morphometric studies, there was evidence of glial loss in the prefrontal cortex of subjects with Major Depressive Disorder, Bipolar Disorder, and Schizophrenia (reviewed in (57)). This decrease in glia, and particularly astrocytes, was replicated experimentally in animals exposed to chronic stress (58), and when induced pharmacologically, was capable of driving animals into a depressive-like condition (58). Replicating the results of (49), we observed a moderate decrease in astrocyte index in the prefrontal cortex of subjects with Major Depressive Disorder (meta-analysis: b= -0.132, p=5.88e-03, q=4.81e-02, Pritzker: β = -0.133, p= 0.0118, **Figure 6f**), but did not see similar changes in the brains of subjects with Bipolar Disorder or Schizophrenia. We also observed a decrease in red blood cell index in association with Schizophrenia (CMC: β=-0.104, p=0.0141, q=0.0471) which is tempting to ascribe to reduced blood flow due to hypofrontality (59). This decrease in red blood cell content could also arise due to psychiatric subjects having an increased probability of dying a violent death, but the effect remained present when we controlled for exsanguination, therefore the effect is likely to be genuinely tied to the illness itself.

3.4 Discriminating Between Changes in Cell Type Balance and Cell-Type Specific Function

Gray matter density has been shown to decrease in the frontal cortex in aging humans (48), and frontal neuron numbers decrease in specific subregions in aging primates (50) and rats (51). However, many scientists would argue that age-related decreases in gray matter are primarily driven by synaptic atrophy instead of decreased cell number (60). This raised the question of whether the decline that we saw in neuronal cell indices with age was being largely driven by the enrichment of genes related to synaptic function in the index. More generally, it raised the question of how well cell type indices could discriminate changes in cell number from changes in cell-type function.

We examined this question using two methods. First, we specifically examined the relationship between age and the functional annotation for genes found in the Neuron_All index in more depth. To do this, we evaluated the relationship between age and gene expression in the Pritzker dataset while controlling for likely confounds using the signal data for all probesets in the dataset:

Equation 4:

$$\text{Gene Expression (Probeset Signal)} = \beta0 + \beta1*(\text{Diagnosis})+\beta2*(\text{Brain pH})+\beta3*(\text{Agonal Factor})+\beta4*(\text{PMI})+\beta5*(\text{Age})+\beta6*(\text{Sex}) + \varepsilon$$

We used "DAVID: Functional Annotation Tool" (//david.ncifcrf.gov/summary.jsp, (61,62) to identify the functional clusters that were overrepresented by the genes included in our neuronal cell type indices (using the full HT-U133A chip as background), and then determined the average effect of age (beta) for the genes included in each of the 240 functional clusters (**Suppl. Table 5**). The vast majority of these functional clusters showed a negative relationship with age on average (**Suppl. Figure 13**). However, these functional clusters overrepresented dendritic/axonal related functions, so in a manner that was blind to the results, we identified 29 functional clusters that were clearly related to dendritic/axonal functions and 41 functional clusters that seemed distinctly unrelated to dendritic/axonal functions (**Suppl. Table 5**). Using this approach, we found that transcripts from both classifications of functional clusters showed an average decrease in expression with age (dendritic/axonal: T(28)=-4.5612, p= 9.197e-05, nondendritic/axonal: T(40)=-2.7566, p=0.008756), but the decrease was larger for transcripts associated with dendritic/axonal-related functions (T(50.082)=2.3385, p= 0.02339, **Suppl. Figure 23**). Based on this analysis, we conclude that synaptic atrophy could be partially driving age-related effects on neuronal cell type indices in the human prefrontal cortex dataset but are unlikely to fully explain the relationship.

Next, we decided to make the process of differentiating between altered cell type-specific functions and relative cell type balance more efficient. We used our cell type specific gene lists to construct gene sets in a file format (.gmt) compatible with the popular tool Gene Set Enrichment Analysis (63,64) and combined them with two other commonly-used gene set collections from the molecular signatures database (MSigDB: http://software.broadinstitute.org/gsea/msigdb/index.jsp, downloaded 09/2017, "C2: Curated Gene Sets" and "C5: GO Gene Sets", **Suppl. Table 6**). Then we tested the utility of incorporating our new gene sets into GSEA (fGSEA: (65)) using the ranked results (betas) for the relationship between each subject variable (**Equation 4**) and each probeset in the Pritzker dataset. Using this method, we could compare the enrichment of the effects of subject variables within gene sets defined by brain cell type to the enrichment seen within gene sets for other functional categories. In general, we found that gene sets for brain cell types tended to be the top result (most extreme normalized enrichment score, NES) for each of the subject variables that showed a strong relationship with cell type in our previous analyses (Agonal Factor vs. "Neuron_All_Cahoy_JNeuro_2008": NES=-2.46, p= 0.00098, q= 0.012, Brain pH vs. "Astrocyte_All_Cahoy_JNeuro_2008": NES=-2.48, p= 0.0011, q=0.014, MDD vs. "Astrocyte_All_Cahoy_JNeuro_2008": NES= -2.60, p= 0.0010, q= 0.017, PMI vs. "GO_OLIGODENDROCYTE_DIFFERENTIATION": NES=-2.42, p= 0.00078, q= 0.027; **Suppl. Table 7**). Similarly, the relationship between the effects of age and neuron-specific gene expression was ranked #4, following the gene sets "GO_SYNAPTIC_SIGNALING", "REACTOME_TRANSMISSION_ACROSS_CHEMICAL_SYNAPSES", "REACTOME_OPIOID_SIGNALLING", but each of them was assigned a similar p-value (p=0.001) and adjusted p-value (q=0.036). We conclude that it is important to consider cell type-specific expression during the analysis of macro-dissected brain microarray data above and beyond the consideration of specific functional pathways, and have submitted our .gmt files to the Broad Institute for potential addition to their curated gene sets in MSigDB to promote this form of analysis.

3.5 Including Cell Content Predictions in the Analysis of Microarray Data Improves Model Fit And Enhances the Detection of Diagnosis-Related Genes in Some Datasets

Over the years, many researchers have been concerned that transcriptomic and genomic analyses of psychiatric disease often produce non-replicable or contradictory results and, perhaps more disturbingly, are typically unable to replicate well-documented effects detected by other methods. We posited that this lack of sensitivity and replicability might be partially due to cell type variability in the samples, especially since such a large percentage of the principal components of variation in our samples were explained by neuron to glia ratio. Within the Pritzker dataset, we were particularly interested in controlling for cell type variability, because there were indications that dissection might have differed between technical batches that were unevenly distributed across diagnosis categories (**Figure 8, Suppl. Figure 10**). There was a similarly uneven distribution of dissection methods across diagnosis categories within the large CMC RNA-Seq dataset. In this dataset, the majority of the bipolar samples (75%) were collected by a brain bank that performed gray matter only dissections (PITT), whereas the control and schizophrenia samples were more evenly distributed across all three institutions (40).

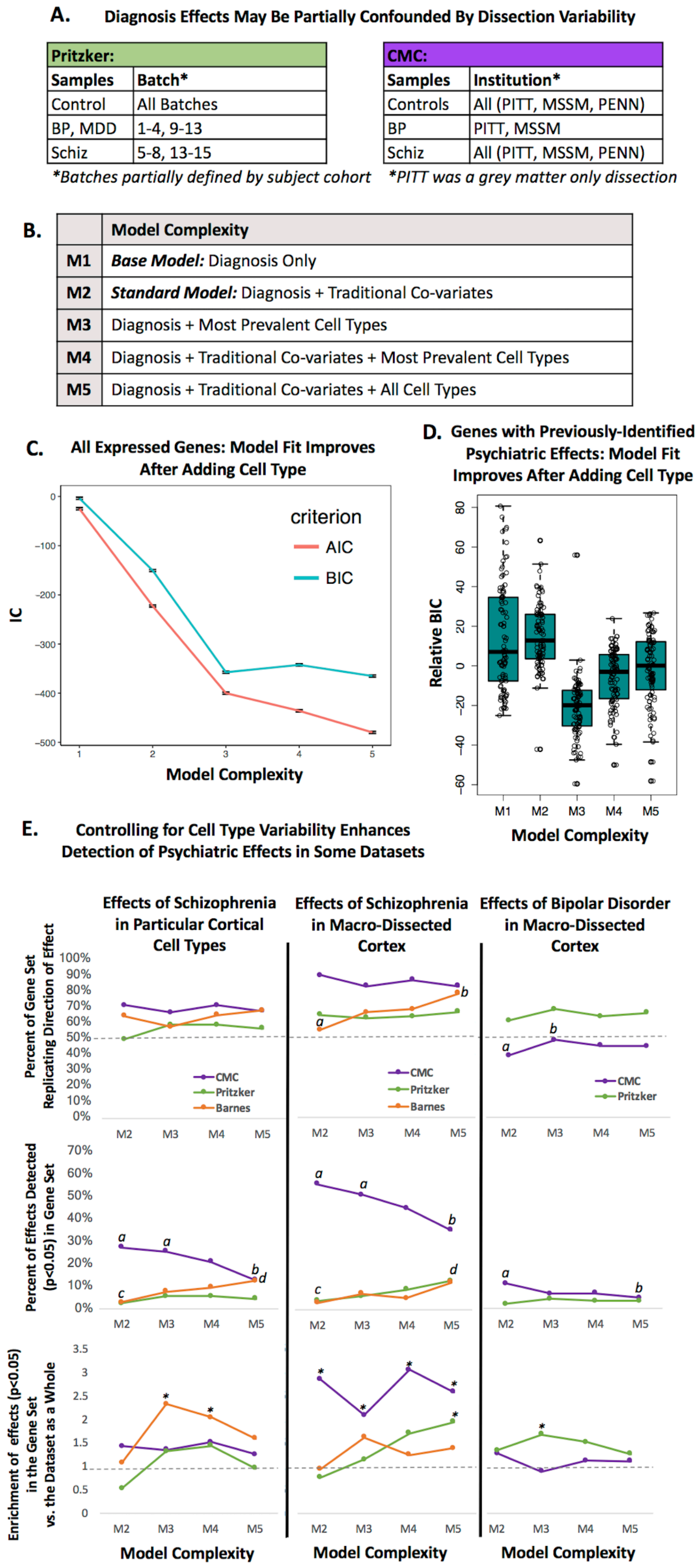
We hypothesized that controlling for cell type while performing differential expression analyses in these datasets would improve our ability to detect previously-documented psychiatric effects on gene expression, especially psychiatric effects on gene expression that were previously-identified within individual cells, since these effects on gene expression should not be mediated by psychiatric changes in overall cell type balance. To test the hypothesis, we first compiled a list of 130 strong, previously-documented relationships between Schizophrenia or Psychosis and gene expression in particular cell types in the human cortex, as detected by in situ hybridization or immunocytochemistry (reviewed further in (19); GAD1: (66–68); RELN:(66); SST: (69), SLC6A1 (GAT1): (70), PVALB:(67), *suicide:* HTR2A (71)), or by single-cell type laser capture microscopy (**Figure 7, Suppl. Table 8** (1,72,73)).

Validation Datasets:	# of Genes	Method:	Brain Bank:	# of Subjects	Brain Region	Co-Variates: Controlled?	Co-Variates: Balanced?	Statistical Stringency
Schizophrenia Effects In Particular Cortical Cell Types:								
Reviewed in Lewis & Sweet (2009)	7	ICC/in situ hybridization	Variable, often PITT	Variable	Prefrontal cortex	Variable	Variable	Variable
Arion et al. (2015)	41	LCM-Microarray: Pyramidal Neurons (Layers 3 &5)	PITT	72	BA9	Direction of effect evaluated, but covariates not included in final model.	Sex, Age, PMI, pH, RIN, tissue storage time, race	Top 40 (FDR<0.1 in both layers, Table 2A), Top 2 in Table 2B, FDR<10E-17 for Layer5)
Pietersen et al. (2014)	47	LCM-Microarray: PVALB Interneurons	HBTRC (MacLean)	16	BA42	Batch. Considered effects of Sex, Age, PMI but not included in final model.	Sex, Age, PMI, pH not significantly different	Top 47 (FDR<0.01, FC>2, Table 3)
Mauney et al. (2015)	35	LCM-Microarray: Oligodendrocyte Precursors	HBTRC (MacLean)	18	BA9	None	Sex, Age, PMI, *pH not reported*	Top 35 (FDR<0.001, Table S2)
Psychiatric Effects in Macro-dissected Prefrontal Cortex:								
Mistry et al. (2013)	126	Meta-analysis of microarray data: Schizophrenia effects	Stanley Foundation, HBTRC (MacLean), PITT, CCHPC , MSSM, MHRI	306	BA9, BA10, BA46	Model selection procedure included Batch, Age, pH, Study	Sex, PMI	FDR<0.1 (Table S2)
Choi et al. (2011)	367	Meta-analysis of microarray data: Bipolar effects	Stanley Foundation	83	BA46 (grey matter only)	Batch (Scan Date), pH, Psychosis, Medication at TOD	Age, BMI, *PMI not reported*	FDR<0.05, FC>1.3 (Table S1)

Figure 7. Gene lists used to assess whether controlling for cell type while performing differential expression analyses enhances the detection of previously-documented psychiatric effects on cortical gene expression. *These lists include genes with documented relationships to psychiatric illness in either 1) particular cortical cell types or 2) macro-dissected cortex. The full lists can be found in* ***Suppl. Table 8.*** *Abbreviations: LCM: Laser Capture Microscopy, PVALB: Parvalbumin, BA: Brodmann's Area, PMI: Post-mortem interval, FDR: False detection ratio (or q-value), Brain Banks: PITT (University of Pittsburgh), HBTRC (Harvard Brain Resource Tissue Center), CCHPC (Charing Cross Hospital Prospective Collection), MSSM (Mount Sinai Icahn School of Medicine), MHRI (Mental Health Research Institute Australia).*

As a comparison, we also considered lists of transcripts strongly-associated with Schizophrenia (74) and Bipolar Disorder (75) in meta-analyses of microarray data derived from human frontal cortical tissue (**Figure 7**). The effects of psychiatric illness on the expression of these transcripts could be mediated by either psychiatric effects on cell type balance or by effects within individual cells. Therefore, controlling for cell type balance while performing differential expression analyses could detract from the detection of some psychiatric effects, but perhaps also enhance the detection of other psychiatric effects by controlling for large, confounding sources of noise (*e.g.,* dissection variability).

Next, we examined our ability to detect these previously-documented psychiatric effects using regression models of increasing complexity (**Figure 8 B**), including a simple base model containing just the variable of interest ("Model 1"), a standard model controlling for traditional co-variates ("Model 2"), and a model controlling for traditional co-variates as well as each of the cell type indices ("Model 5": **Equation 5**). We also used two reduced models that only included the most prevalent cell types (Astrocyte, Microglia, Oligodendrocyte, Neuron_Interneuron, Neuron_Projection; (21)) to avoid issues with multicollinearity. The first of these models included traditional co-variates as well ("Model 4"), whereas the second model excluded them ("Model 3").



A. Diagnosis Effects May Be Partially Confounded By Dissection Variability

Pritzker:	
Samples	**Batch***
Control	All Batches
BP, MDD	1-4, 9-13
Schiz	5-8, 13-15

**Batches partially defined by subject cohort*

CMC:	
Samples	**Institution***
Controls	All (PITT, MSSM, PENN)
BP	PITT, MSSM
Schiz	All (PITT, MSSM, PENN)

**PITT was a grey matter only dissection*

B.

	Model Complexity
M1	***Base Model:*** Diagnosis Only
M2	***Standard Model:*** Diagnosis + Traditional Co-variates
M3	Diagnosis + Most Prevalent Cell Types
M4	Diagnosis + Traditional Co-variates + Most Prevalent Cell Types
M5	Diagnosis + Traditional Co-variates + All Cell Types

C.
All Expressed Genes: Model Fit Improves After Adding Cell Type
IC
0
−100
−200
−300
−400
−500
1
2
3
4
5
Model Complexity
criterion
AIC
BIC
D.
Genes with Previously-Identified Psychiatric Effects: Model Fit Improves After Adding Cell Type
Relative BIC
80
60
40
20
0
−20
−40
−60
M1
M2
M3
M4
M5
Model Complexity
E.
Controlling for Cell Type Variability Enhances Detection of Psychiatric Effects in Some Datasets
Effects of Schizophrenia in Particular Cortical Cell Types
Effects of Schizophrenia in Macro-Dissected Cortex
Effects of Bipolar Disorder in Macro-Dissected Cortex
Percent of Gene Set Replicating Direction of Effect
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
CMC
Pritzker
Barnes
a
b
M2
M3
M4
M5
Percent of Effects Detected (p<0.05) in Gene Set
70%
60%
50%
40%
30%
20%
10%
0%
a
b
c
d
Enrichment of effects (p<0.05) in the Gene Set vs. the Dataset as a Whole
3.5
3
2.5
2
1.5
1
0.5
0
*
Model Complexity

616

Figure 8. Including Cell Content Predictions in the Analysis of Microarray Data Improves Model Fit and Enhances the Detection of Previously-Identified Diagnosis-Related Genes in Some Datasets. A. *Diagnosis effects were likely to be partially confounded by dissection variability within the Pritzker and CMC datasets.* ***B:*** *We examined a series of differential expression models of increasing complexity, including a base model (M1), a standard model (M2), and three models that included cell type co-variates (M3-M5).* ***C.*** *Model fit improved with the addition of cell type (M1/M2 vs. M3-M5) when examining all expressed genes in the dataset (example from CMC: points= AVE +/-SE).* ***D.*** *Model fit improved with the addition of cell type (M1/M2 vs. M3-M5) when examining genes with previously-documented relationships with psychiatric illness in particular cell types (example from Pritzker: BIC values for all models for each gene were centered prior to analysis). Boxplots represent the median and interquartile range, with whiskers illustrating either the full range of the data or 1.5x the interquartile range.* ***E.*** *Evaluating the replication of previously-detected psychiatric effects (****Figure*** *7) in three datasets (Pritzker, CMC, and Barnes) using a standard differential expression model (M2) vs. models that include cell type co-variates (M3-5). Top graphs: The percentage of genes (y-axis: 0-1) replicating the direction of previously-documented psychiatric effects on cortical gene expression sometimes increases with the addition of cell type to the model (Barnes (effects of Schiz): M2 vs. M5, CMC (effects of Bipolar Disorder): M2 vs. M3). Middle graphs: The detection of previously-identified psychiatric effects on gene expression (p<0.05 & replicated direction of effect) increases with the addition of cell type to the model in some datasets (Barnes: M2 vs. M5, Pritzker: M2 vs. M5) but decreases in others (CMC: M2 vs. M5, M3 vs. M5). Bottom graphs: In some datasets we see an enrichment of psychiatric effects (p<0.05) in previously-identified psychiatric gene sets only after controlling for cell type (Barnes: M3, M4, Pritzker: M5, M3). For the CMC dataset, we see an enrichment using all models. The full results for all models can be found in* ***Suppl. Table 9****,* ***Suppl. Table 10****,* ***Suppl. Table 11****,* ***Suppl. Table 12****, and* ***Suppl. Table 13.***

Equation 5: A model of gene expression for each dataset, colored to illustrate the subcomponents evaluated during our model comparison (#M1-M5). *The base model (intercept and variable of interest) is presented in green, the typical subject variable covariates included in a standard model are blue, the cell type indices for the most prevalent cell types are colored red, and the remaining cell type indices are in purple. Model components unique to each dataset are underlined.*

The Pritzker microarray dataset:

Gene Expression (Probeset Signal) =
β0 + β1*(The variable of interest: Diagnosis)
+β2*(Brain pH)+ β3*(PMI)+ β4*(Age)+ β5*(Sex)+ β6*(Agonal Factor)+
+ β7*(Astrocyte)+β8*(Oligodendrocyte)+β9*(Microglia)+β10*(Interneuron)+β11*(ProjectionNeuron)
+β12*(Endothelial)+β13*(Neuron_All)+β14*(Oligodendrocyte_Immature)+β15*(Mural)+β16*(RBC)+ ε

The CMC RNA-Seq dataset:

Gene Expression (Probeset Signal) =
β0 + β1*(The variable of interest: Diagnosis)
+β2*(Brain pH)+β3*(PMI)+ β4*(Age)+ β5*(Sex)+ β6*(RIN)+β7*(Institution)+ β8*(CauseOfDeath)+
+ β9*(Astrocyte)+β10*(Oligodendrocyte)+β11*(Microglia)+β12*(Interneuron)+β13*(ProjectionNeuron)
+β14*(Endothelial)+β15*(Neuron_All)+β16*(Oligodendrocyte_Immature)+β17*(Mural)+β18*(RBC)+ ε

The smaller microarray datasets (GSE53987, GSE21935, GSE21138):

Gene Expression (Probeset Signal) =
β0 + β1*(The variable of interest: Diagnosis)
+β2*(Brain pH)+β3*(PMI)+ β4*(Age)+ β5*(Sex)+ β6*(RNADegradation)+
+ β7*(Astrocyte)+β8*(Oligodendrocyte)+β9*(Microglia)+β10*(Interneuron)+β11*(ProjectionNeuron)

+β12*(Endothelial)+β13*(Neuron_All)+β14*(Oligodendrocyte_Immature)+β15*(Mural)+β16*(RBC)+ ε

We found that including predictions of cell type balance in our models assessing the effect of diagnosis on gene expression dramatically improved model fit as assessed by Akaike's Information Criterion (AIC) or Bayesian Information Criterion (BIC) (**Figure 8**). These improvements were largest with the addition of the five most prevalent cell types to the model (M3, M4); the addition of less common cell types produced smaller gains (M5). These improvements were clear whether we considered the average model fit for all expressed genes (*e.g.,* **Figure 8B**) or just genes with previously-identified psychiatric effects (*e.g.,* **Figure 8C**).

However, models that included cell type were not necessarily superior at replicating previously-observed psychiatric effects on gene expression (**Figure 7**), even when examining psychiatric effects that were likely to be independent of changes in cell type balance. For each model, we quantified the percentage of genes replicating the previously-observed direction of effect in relationship to psychiatric illness, as well as the percentage of genes that replicated the effect using a common threshold for detection (p<0.05). Finally, we also looked at the enrichment of psychiatric effects (p<0.05) in each of the previously-documented psychiatric gene sets in comparison to the other genes in our datasets. For this analysis, to improve comparisons across datasets, we defined the statistical background for enrichment using genes universally represented in all three datasets (Pritzker, CMC, Barnes).

In general, we found that the two datasets that had the most variability in gene expression related to cell type (Pritzker, Barnes: **Results 3.1**) were more likely to replicate previously-documented psychiatric effects on gene expression when the differential expression model included cell type covariates. For example, in the Barnes dataset, adding cell type co-variates to the model increased our ability to detect effects of Schizophrenia that had been previously documented within particular cell types or macro-dissected tissue (**Figure 8E**, Fisher's exact test: M2 vs. M5, $p<0.05$ in both gene sets). Similarly, adding cell type co-variates to the model allowed us to see a significant enrichment of Schizophrenia effects (p<0.05) in genes with previously-documented psychiatric effects in particular cell types (Fisher's exact test p<0.05: M3 & M4). In the Pritzker dataset, we saw that adding cell type co-variates to the model increased our ability to detect previously-documented effects of Schizophrenia in macrodissected tissue (M2 vs. M5: p<0.05). Likewise, adding cell type co-variates to the model allowed us to see a significant enrichment of Schizophrenia and Bipolar effects (p<0.05) in genes with previously-documented psychiatric effects in macro-dissected tissue (Fisher's exact test p<0.05: Schizophrenia: M5, Bipolar: M3). This mirrored the results of another analysis that we had conducted suggesting that controlling for cell type increased the overlap between the top diagnosis results in the Pritzker dataset and previous findings in the literature as a whole (**Suppl. Figure 24**, **Suppl. Figure 25**).

In the large CMC RNA-Seq dataset, the rate of replication of previously-documented effects of Schizophrenia was already quite high using a standard differential expression model containing traditional co-variates (M2). Using a standard model, we could detect 27% of the previously-documented effects in cortical cell types and 55% of the previously-documented effects in macro-dissected tissue (with a replicated direction of effect and p<0.05). However, in contrast to what we had observed in the Pritzker and Barnes datasets, controlling for cell type seemed to actually *diminish* the ability to detect effects of Schizophrenia that had been previously-observed within particular cell types or macrodissected tissue in a manner that scaled with the number of co-variates included in the model (M2 or M3 vs. M5: p<0.05 for both gene sets), despite improvements in model fit parameters and a lack of significant relationship between Schizophrenia and any of the prevalent cell types (**Section 3.3**). Including cell type co-variates in the model did not improve our ability to observe a significant enrichment of Schizophrenia effects in genes with previously-documented psychiatric effects in macro-dissected tissue – this enrichment was present in the results from all differential expression models (Fisher's exact test p<0.05: M2-M5). In contrast, controlling for cell type slightly improved the replication of the direction of previously-documented Bipolar Disorder effects (Fisher's exact test: M2 vs. M3: p<0.05) in a manner that would seem appropriate due to the highly uneven distribution of bipolar samples across institutions and dissection methods, but even after this improvement the rate of replication was still no better than chance (48%), and, counterintuitively, the ability to successfully detect those effects still diminished in a manner that seemed to scale with the number of co-variates included in the model (Fisher's exact test: M2 vs. M5, p<0.05). In a preliminary analysis of the two smaller human microarray datasets that were derived from gray-matter only dissections (GSE53987, GSE21138), the addition of cell type co-variates to differential expression models clearly diminished both the percentage of genes replicating the previously-documented direction of effect of Schizophrenia in particular cell types (Fisher's exact test: Narayan et al.: M2 vs. M4 or M5: p<0.05, Lanz et al.: M2 vs. M4 or M5) and the ability to successfully detect previously-documented effects (Fisher's exact test: Narayan et al.: M2 vs. M4 or M5: p<0.05).

Therefore, we conclude that the addition of cell type covariates to differential expression models is only recommended when there is a particularly large amount of variability in the dataset associated with cell type balance. For public use we have released the full results for each dataset analyzed using the different models discussed above **(Suppl. Table 9**, **Suppl. Table 10**, **Suppl. Table 11**, **Suppl. Table 12**, and **Suppl. Table 13).**

4. Discussion

In this manuscript, we have demonstrated that the statistical cell type index is a relatively simple manner of interrogating cell-type specific expression in transcriptomic datasets from macro-dissected human brain tissue. We find that statistical estimations of cell type balance almost fully account for the top principal components of variation in microarray data derived from macro-dissected brain tissue samples, far surpassing the effects of traditional subject variables (post-mortem interval, hypoxia, age, gender). Indeed, our results suggest that many variables of medical interest are themselves accompanied by strong changes in cell type composition in naturally-observed human brains. We find that within both chronic (age, sex, diagnosis) and acute conditions (agonal, PMI, pH) there are substantial changes in the relative representation of different cell types. Thus, accounting for demography at the cellular population level is as important for the interpretation of microarray data as cell-level functional regulation. This form of data deconvolution was also useful for identifying the subtler effects of psychiatric illness within our samples, divulging the decrease in astrocytes that is known to occur in Major Depressive Disorder and the decrease in red blood cell content in the frontal cortex in Schizophrenia, resembling known fMRI hypofrontality. This form of data deconvolution may also aid in the detection of psychiatric effects while conducting differential expression analyses in datasets that have highly-variable cell content.

These results touch upon the fundamental question as to whether organ-level function responds to challenge by changing the biological states of individual cells (Lamarckian) or the life and death of different cell populations (Darwinian). To reach such a sweeping perspective in human brain tissue using classic cell biology methods would require epic efforts in labeling, cell sorting, and counting. We have demonstrated that you can approximate this vantage point using an elegant, supervised signal decomposition exploiting increasingly available genomic data. However, it should be noted that, similar to other forms of functional annotation, cell type indices are best treated as a hypothesis-generation tool instead of a final conclusion regarding tissue cell content. We have demonstrated the utility of cell type indices for detecting large-scale alterations in cell content in relationship with known subject variables in post-mortem tissue. We have not tested the sensitivity of the technique for detecting smaller effects or the validity under all circumstances or non-cortical tissue types. Likewise, while using this technique it is impossible to distinguish between alterations in cell type balance and cell-type specific transcriptional activity: when a sample shows a higher value of a particular cell type index, it could have a larger number of such cells, or each cell could have produced more of its unique group of transcripts, via a larger cell body, slower mRNA degradation, or an overall change in transcription rate. In this regard, the index that we calculate does not have a specific interpretation; rather it is a holistic property of the cell populations, the “neuron-ness” or “microglia-ness” of the sample. Such an abstract index represents the ecological shifts inferred from the pooled transcriptome. That said, unlike principal component scores or other associated techniques of removing unwanted variation from genomic data, our cell type indices do have real biological meaning - they can be interpreted in a known system of cell type taxonomy. When single-cell genomic data uncovers new cell types (e.g., the Allen Brain Atlas cellular taxonomy initiative (76)) or meta-analyses refine the list of genes defined as having cell-type specific expression (e.g., (77)), our indices will surely evolve with these new classification frameworks, but the power of the approach will remain, in that we can disentangle the intrinsic changes of individual genes from the population-level shifts of major cell types.

We found that many variables of medical interest are accompanied by strong changes in cell type composition in naturally-observed human brains. One result from this analysis seems particularly worth discussing in greater depth. It has been acknowledged for a long time that exposure to a hypoxic environment prior to death has a huge impact on gene expression in human post-mortem brains (e.g., (32,33,78–80)). This impact on gene expression is so large that up until recently the primary principal component of variation (PC1) in our Pritzker data was assumed to represent the degree of hypoxia, and was sometimes even systematically removed before performing diagnosis-related analyses (e.g., (35)). The strong relationship between hypoxia and gene expression in human post-mortem samples was hypothesized to be partially mediated by neuronal necrosis (81) and lactic acidosis (79). However, the magnitude of the effect of hypoxia on gene expression was still puzzling, especially when compared to the much more moderate effects of post-mortem interval, even when the intervals ranged from 8-40+ hrs. Our current analysis provides an explanation for this discrepancy, since it is clear from our results that the brains of our subjects are *actively compensating* for a hypoxic environment prior to death by altering the balance or overall transcriptional activity of support cells and neurons. The differential effects of hypoxia on neurons and glial cells have been studied since the 1960's (82), but to our knowledge this is the first time that anyone has related the large effects of hypoxia in post-mortem transcriptomic data to a corresponding upregulation in the transcriptional activity of vascular cell types (46).

This connection is important for understanding why results associating gene expression and psychiatric illness in human post-mortem tissue sometimes do not replicate. If a study contains mostly tissue from individuals who experienced greater hypoxia before death (e.g., hospital care with artificial respiration or drug overdose followed by coma), then the evaluation of the effect of neuropsychiatric illness is likely to inadvertently focus on differential expression in support cell types (astrocytes, endothelial cells), whereas a study that mostly contains tissue from individuals who died a fast death (e.g., car accident or myocardial infarction) will emphasize the effects of neuropsychiatric illness in neurons. That said, although both indicators of perimortem hypoxia (agonal factor and acidosis (pH)) showed similar strong relationships with cell type balance, we do recommend some caution when interpreting the relationship between pH and cell type in tissue from subjects with psychiatric disorders, as pH can indicate other biological changes besides hypoxia. For example, there are small consistent decreases in pH associated with Bipolar Disorder even in live subjects (83–85) and metabolic changes associated with pH are theorized to play an important role in Schizophrenia (80). Therefore, some of the relationship between pH and cell type balance may be driven by a third variable (psychiatric illness or psychiatric treatment). It is also possible that a change in the cell content of brain tissue could cause a change in pH (86).

We found that including cell type indices as co-variates while running differential expression analyses helped improve our ability to detect previously-documented relationships between psychiatric illness and gene expression in datasets that were particularly affected by variability in cell type balance. This improvement was not seen in datasets that were less affected by variability in cell type balance, despite improvements in model fit and a lack of strong multicollinearity between diagnosis and the cell type indices. This finding was initially surprising to us, but upon further consideration makes sense, as the cell type indices are multi-parameter gene expression variables. Therefore, there is increased risk of overfitting when modeling the data for any particular gene. We conclude that the addition of cell type covariates to differential expression models is only recommended when there is a particularly large amount of variability in the dataset associated with cell type balance, or when there is strong reason to believe that technical variation associated with cell type (such as dissection) may be highly confounding in the result. We strongly recommend that model selection while conducting differential expression analyses should be considered carefully, and evaluated not only in terms of fit parameters but also validity and interpretability.

Regarding the importance of model selection for interpretability, it is worth noting that an important difference between our final analysis methods and those used by some previous researchers (*e.g.,* 10–12) was the lack of cell type interaction terms included in our models (*e.g.,* Diagnosis*Astrocyte Index). Theoretically, the addition of cell type interaction terms should allow the researcher to statistically interrogate cell-type differentiated diagnosis effects because samples that contain more of a particular cell type should exhibit more of that cell type's respective diagnosis effect. Versions of this form of analysis have been successful in other investigations (e.g., (11,12,87)) but we were not able to validate the method using our database of previously-documented relationships with diagnosis in prefrontal cell types (**Figure 7**) and a variety of model specifications (*e.g.,* **Suppl. Figure 26**). Upon consideration, we realized that these negative results were difficult to interpret because significant diagnosis*cell type interactions should only become evident if the effect of diagnosis in a particular cell type is different from what is occurring in all cell types on average. For genes with expression that is reasonably specific to a particular cell type (*e.g.,* GAD1, PVALB), the overall average diagnosis effect may already largely reflect the effect within that cell type and the respective interaction term will not be significantly different, even though the disease effect is clearly tracking the balance of that cell population. In the end, we decided that the addition of interaction terms to our models was not demonstrably worth the associated decrease in overall model fit and statistical power.

Finally, our work drives home the fact that any comprehensive theory of psychiatric illness needs to account for the dichotomy between the health of individual cells and that of their ecosystem. We found that the functional changes accompanying psychiatric illness in the cortex occurred both at the level of cell population shifts (decreased astrocytic presence and red blood cell count) and at the level of intrinsic gene regulation not explained by population shifts. A similar conclusion regarding the importance of cell type balance in association with psychiatric illness was recently drawn by our collaborators (e.g.,(88)) using a similar technique to analyze RNA-Seq data from the anterior cingulate cortex. In the future, we plan to use our technique to re-analyze many of the other large microarray datasets existing within the Pritzker Neuropsychiatric Consortium with the hope of gaining better insight into psychiatric disease effects. This application of our technique seems particularly important in light of recent evidence linking disrupted neuroimmunity (89) and neuroglia (e.g., (49,58,90)) to psychiatric illness, as well as growing evidence that growth factors with cell type specific effects play an important role in depressive illness and emotional regulation (for a review see (23,91)).

In conclusion, we have found this method to be a valuable addition to traditional functional ontology tools as a manner of improving the interpretation of transcriptomic results. The capability to unravel alterations of cell type composition from modulation of cell state, even just probabilistically, is inherently useful for understanding the higher-level function of the brain as emergent properties of brain activity, such as emotion, cognition, memory, and addiction, usually involve ensembles of many cells. Facilitating the interpretation of gene activity data in macro-dissected tissue in light of both processes provides new opportunities to integrate results with findings from other approaches, such as electrophysiology analysis of brain circuits, brain imaging, optogenetic manipulations, and naturally occurring variation in response to injury and brain diseases.

For the benefit of other researchers, we have made our database of brain cell type specific genes (https://sites.google.com/a/umich.edu/megan-hastings-hagenauer/home/cell-type-analysis) and R code for conducting cell type analyses publicly available in the form of a downloadable R package (https://github.com/hagenaue/BrainInABlender) and we are happy to assist researchers in their usage for pursuing better insight into psychiatric illness and neurological disease.

5. Acknowledgements

We thank all the members of the Pritzker Consortium (especially the University of California, Irvine Brain Bank staff), Drs. Adriana Medina and David Krolewski for brain dissections and methodological input, and Dr. Simon Evans, Sharon Burke and Mary Hoverstein for their involvement in the initial mRNA extraction and microarrays. Grace Hsienyuan Chang, Jennifer Fitzpatrick, LeAnn Fitzpatrick, Jim Stewart, Tom Dixon, Doug Smith, Andy Lin, and Manhong Dai were invaluable for maintaining our databases of clinical information and biological specimens. We would also like to thank Drs. Elyse Aurbach, Katherine Prater, Kathryn Hilde, Fan Meng, Lilah Toker, and Mark Reimers for advice and feedback regarding the methodology or manuscript. We would also like to thank our undergraduate research assistants Isabelle Birt, Alek Pankonin, and Daniela Romero Vargas for their help compiling the Allen Brain Atlas data, annotating and uploading code, creating the BrainInABlender R package, and editorial assistance. Finally, we would like to thank our reviewers, whose insightful feedback helped inspire several particularly useful analyses, leading us to a stronger set of conclusions.

6. References

1. Arion D, Corradi JP, Tang S, Datta D, Boothe F, He A, et al. Distinctive transcriptome alterations of prefrontal pyramidal neurons in schizophrenia and schizoaffective disorder. Mol Psychiatry. 2015 Nov;20(11):1397–405.

2. Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, et al. A survey of human brain transcriptome diversity at the single cell level. Proc Natl Acad Sci U S A. 2015 Jun 9;112(23):7285–90.

3. Lake BB, Ai R, Kaeser GE, Salathia NS, Yung YC, Liu R, et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science. 2016 Jun 24;352(6293):1586–90.

4. Choi KH, Elashoff M, Higgs BW, Song J, Kim S, Sabunciyan S, et al. Putative psychosis genes in the prefrontal cortex: combined analysis of gene expression microarrays. BMC Psychiatry. 2008;8:87.

5. Evans SJ, Choudary PV, Neal CR, Li JZ, Vawter MP, Tomita H, et al. Dysregulation of the fibroblast growth factor system in major depression. Proc Natl Acad Sci U S A. 2004 Oct 26;101(43):15506–11.

6. 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. 2009;4(7):e6098.

7. Chikina M, Zaslavsky E, Sealfon SC. CellCODE: a robust latent variable approach to differential expression analysis for heterogeneous cell populations. Bioinforma Oxf Engl. 2015 May 15;31(10):1584–91.

8. Gaujoux R, Seoighe C. CellMix: a comprehensive toolbox for gene expression deconvolution. Bioinforma Oxf Engl. 2013 Sep 1;29(17):2211–2.

9. Shen-Orr SS, Gaujoux R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. Curr Opin Immunol. 2013 Oct;25(5):571–8.

10. Capurro A, Bodea L-G, Schaefer P, Luthi-Carter R, Perreau VM. Computational deconvolution of genome wide expression data from Parkinson's and Huntington's disease brain tissues using population-specific expression analysis. Front Neurosci. 2014;8:441.

11. Kuhn A, Kumar A, Beilina A, Dillman A, Cookson MR, Singleton AB. Cell population-specific expression analysis of human cerebellum. BMC Genomics. 2012;13:610.

12. Kuhn A, Thu D, Waldvogel HJ, Faull RLM, Luthi-Carter R. Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. Nat Methods. 2011 Nov;8(11):945–7.

13. Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, Christopherson KS, et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. J Neurosci Off J Soc Neurosci. 2008 Jan 2;28(1):264–78.

14. Daneman R, Zhou L, Agalliu D, Cahoy JD, Kaushal A, Barres BA. The mouse blood-brain barrier transcriptome: a new resource for understanding the development and function of brain endothelial cells. PloS One. 2010;5(10):e13741.

15. Doyle JP, Dougherty JD, Heiman M, Schmidt EF, Stevens TR, Ma G, et al. Application of a translational profiling approach for the comparative analysis of CNS cell types. Cell. 2008 Nov 14;135(4):749–62.

16. Sugino K, Hempel CM, Miller MN, Hattox AM, Shapiro P, Wu C, et al. Molecular taxonomy of major neuronal classes in the adult mouse forebrain. Nat Neurosci. 2006 Jan;9(1):99–107.

17. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015 Mar 6;347(6226):1138–42.

18. Zhang Y, Chen K, Sloan SA, Bennett ML, Scholze AR, O'Keeffe S, et al. An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. J Neurosci Off J Soc Neurosci. 2014 Sep 3;34(36):11929–47.

19. Lewis DA, Sweet RA. Schizophrenia from a neural circuitry perspective: advancing toward rational pharmacological therapies. J Clin Invest. 2009 Apr;119(4):706–16.

20. Lynch JC. The Cerebral Cortex. In: Fundamental Neuroscience. 2nd ed. Philadelphia: Churchill Livingstone; 2002. p. 505–20.

21. Hutchins DE, Naftel JP, Ard MD. The cell biology of neurons and glia. In: Fundamental Neuroscience. 2nd ed. Philadelphia: Churchill Livingstone; 2002. p. 15–36.

22. Bergers G, Song S. The role of pericytes in blood-vessel formation and maintenance. Neuro-Oncol. 2005 Oct;7(4):452–64.

23. Duman RS, Monteggia LM. A neurotrophic model for stress-related mood disorders. Biol Psychiatry. 2006 Jun 15;59(12):1116–27.

24. Doss JF, Corcoran DL, Jima DD, Telen MJ, Dave SS, Chi J-T. A comprehensive joint analysis of the long and short RNA transcriptomes of human erythrocytes. BMC Genomics. 2015;16(1):952.

25. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature. 2012 Sep 20;489(7416):391–9.

26. Allen Brain Atlas. Technical White Paper: Case qualification and donor profiles, v.7 [Internet]. 2013. Available from: help.brain-map.org

27. Allen Brain Atlas. Technical White Paper: Microarray Survey, v.7 [Internet]. 2013. Available from: help.brain-map.org

28. Allen Brain Atlas. Technical White Paper: Microarray Data Normalization, v.1 [Internet]. 2013. Available from: help.brain-map.org

29. Carpenter MB. Core Text of Neuroanatomy. 4th ed. Baltimore, MD: Williams & Wilkins; 1991.

30. Amaral DG, Scharfman HE, Lavenex P. The dentate gyrus: fundamental neuroanatomical organization (dentate gyrus for dummies). Prog Brain Res. 2007;163:3–22.

31. Sun N, Cassell MD. Intrinsic GABAergic neurons in the rat central extended amygdala. J Comp Neurol. 1993 Apr 15;330(3):381–404.

32. Li JZ, Vawter MP, Walsh DM, Tomita H, Evans SJ, Choudary PV, et al. Systematic changes in gene expression in postmortem human brains associated with tissue pH and terminal medical conditions. Hum Mol Genet. 2004 Mar 15;13(6):609–16.

33. Tomita H, Vawter MP, Walsh DM, Evans SJ, Choudary PV, Li J, et al. Effect of agonal and postmortem factors on gene expression profile: quality control in microarray analyses of postmortem human brain. Biol Psychiatry. 2004 Feb 15;55(4):346–52.

34. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostat Oxf Engl. 2003 Apr;4(2):249–64.

35. Li JZ, Bunney BG, Meng F, Hagenauer MH, Walsh DM, Vawter MP, et al. Circadian patterns of gene expression in the human brain and disruption in major depressive disorder. Proc Natl Acad Sci U S A. 2013 Jun 11;110(24):9950–5.

36. Lanz TA, Joshi JJ, Reinhart V, Johnson K, Grantham LE, Volfson D. STEP levels are unchanged in pre-frontal cortex and associative striatum in post-mortem human brain samples from subjects with schizophrenia, bipolar disorder and major depressive disorder. PloS One. 2015;10(3):e0121744.

37. Barnes MR, Huxley-Jones J, Maycox PR, Lennon M, Thornber A, Kelly F, et al. Transcription and pathway analysis of the superior temporal cortex and anterior prefrontal cortex in schizophrenia. J Neurosci Res. 2011 Aug;89(8):1218–27.

38. Narayan S, Tang B, Head SR, Gilmartin TJ, Sutcliffe JG, Dean B, et al. Molecular profiles of schizophrenia in the CNS at different stages of illness. Brain Res. 2008 Nov 6;1239:235–48.

39. Fasold M, Binder H. AffyRNADegradation: control and correction of RNA quality effects in GeneChip expression data. Bioinformatics. 2013 Jan;29(1):129–31.

40. Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nat Neurosci. 2016 Nov;19(11):1442–53.

41. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015 Apr;12(4):357–60.

42. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014 Feb 3;15(2):R29.

43. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinforma Oxf Engl. 2010 Jun 15;26(12):1572–3.

44. Viechtbauer W. Conducting Meta-Analyses in R with The metafor Package. J Stat Softw. 2010 Aug 1;36.

45. Pollard KS, Dudoit S, Laan MJ van der. Multiple Testing Procedures: the multtest Package and Applications to Genomics. In: Bioinformatics and Computational Biology Solutions Using R and Bioconductor [Internet]. Springer, New York, NY; 2005 [cited 2017 Oct 13]. p. 249–71. (Statistics for Biology and Health). Available from: https://link.springer.com/chapter/10.1007/0-387-29362-0_15

46. Li L, Welser JV, Dore-Duffy P, del Zoppo GJ, Lamanna JC, Milner R. In the hypoxic central nervous system, endothelial cell proliferation is followed by astrocyte activation, proliferation, and increased expression of the alpha 6 beta 4 integrin and dystroglycan. Glia. 2010 Aug;58(10):1157–67.

47. Banasiak KJ, Haddad GG. Hypoxia-induced apoptosis: effect of hypoxic severity and role of p53 in neuronal cell death. Brain Res. 1998 Jun 29;797(2):295–304.

48. Sowell ER, Peterson BS, Thompson PM, Welcome SE, Henkenius AL, Toga AW. Mapping cortical change across the human life span. Nat Neurosci. 2003 Mar;6(3):309–15.

49. Rajkowska G, Miguel-Hidalgo JJ, Wei J, Dilley G, Pittman SD, Meltzer HY, et al. Morphometric evidence for neuronal and glial prefrontal cell pathology in major depression. Biol Psychiatry. 1999 May 1;45(9):1085–98.

50. Smith DE, Rapp PR, McKay HM, Roberts JA, Tuszynski MH. Memory impairment in aged primates is associated with focal death of cortical neurons and atrophy of subcortical neurons. J Neurosci Off J Soc Neurosci. 2004 May 5;24(18):4373–81.

51. Stranahan AM, Jiam NT, Spiegel AM, Gallagher M. Aging reduces total neuron number in the dorsal component of the rodent prefrontal cortex. J Comp Neurol. 2012 Apr 15;520(6):1318–26.

52. Peters A, Sethares C. Oligodendrocytes, their progenitors and other neuroglial cells in the aging primate cerebral cortex. Cereb Cortex N Y N 1991. 2004 Sep;14(9):995–1007.

53. Resnick SM, Pham DL, Kraut MA, Zonderman AB, Davatzikos C. Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain. J Neurosci Off J Soc Neurosci. 2003 Apr 15;23(8):3295–301.

54. Salat DH, Buckner RL, Snyder AZ, Greve DN, Desikan RSR, Busa E, et al. Thinning of the cerebral cortex in aging. Cereb Cortex N Y N 1991. 2004 Jul;14(7):721–30.

55. Peters A, Sethares C, Moss MB. HOW THE PRIMATE FORNIX IS AFFECTED BY AGE. J Comp Neurol. 2010 Oct 1;518(19):3962–80.

56. Shepherd TM, Flint JJ, Thelwall PE, Stanisz GJ, Mareci TH, Yachnis AT, et al. Postmortem interval alters the water relaxation and diffusion properties of rat nervous tissue--implications for MRI studies of human autopsy samples. NeuroImage. 2009 Feb 1;44(3):820–6.

57. Cotter DR, Pariante CM, Everall IP. Glial cell abnormalities in major psychiatric disorders: The evidence and implications. Brain Res Bull. 2001 Jul 15;55(5):585–95.

58. Banasr M, Duman RS. Glial loss in the prefrontal cortex is sufficient to induce depressive-like behaviors. Biol Psychiatry. 2008 Nov 15;64(10):863–70.

59. RAGLAND JD, YOON J, MINZENBERG MJ, CARTER CS. Neuroimaging of cognitive disability in schizophrenia: Search for a pathophysiological mechanism. Int Rev Psychiatry Abingdon Engl. 2007 Aug;19(4):417–27.

60. Peters A, Sethares C, Luebke JI. Synapses are lost during aging in the primate prefrontal cortex. Neuroscience. 2008 Apr 9;152(4):970–81.

61. Huang DW, Sherman BT, Zheng X, Yang J, Imamichi T, Stephens R, et al. Extracting biological meaning from large gene lists with DAVID. Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al. 2009 Sep;Chapter 13:Unit 13.11.

62. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.

63. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545–50.

64. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003 Jul;34(3):267–73.

65. Sergushichev A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. bioRxiv [Internet]. 2016 Jun 20; Available from: http://biorxiv.org/content/early/2016/06/20/060012.abstract

66. Guidotti A, Auta J, Davis JM, Di-Giorgi-Gerevini V, Dwivedi Y, Grayson DR, et al. Decrease in reelin and glutamic acid decarboxylase67 (GAD67) expression in schizophrenia and bipolar disorder: a postmortem brain study. Arch Gen Psychiatry. 2000 Nov;57(11):1061–9.

67. Hashimoto T, Volk DW, Eggan SM, Mirnics K, Pierri JN, Sun Z, et al. Gene expression deficits in a subclass of GABA neurons in the prefrontal cortex of subjects with schizophrenia. J Neurosci Off J Soc Neurosci. 2003 Jul 16;23(15):6315–26.

68. Volk DW, Austin MC, Pierri JN, Sampson AR, Lewis DA. Decreased glutamic acid decarboxylase67 messenger RNA expression in a subset of prefrontal cortical gamma-aminobutyric acid neurons in subjects with schizophrenia. Arch Gen Psychiatry. 2000 Mar;57(3):237–45.

69. Morris HM, Hashimoto T, Lewis DA. Alterations in somatostatin mRNA expression in the dorsolateral prefrontal cortex of subjects with schizophrenia or schizoaffective disorder. Cereb Cortex N Y N 1991. 2008 Jul;18(7):1575–87.

70. Volk D, Austin M, Pierri J, Sampson A, Lewis D. GABA transporter-1 mRNA in the prefrontal cortex in schizophrenia: decreased expression in a subset of neurons. Am J Psychiatry. 2001 Feb;158(2):256–65.

71. Pandey GN, Dwivedi Y, Rizavi HS, Ren X, Pandey SC, Pesold C, et al. Higher expression of serotonin 5-HT(2A) receptors in the postmortem brains of teenage suicide victims. Am J Psychiatry. 2002 Mar;159(3):419–29.

72. Pietersen CY, Mauney SA, Kim SS, Passeri E, Lim MP, Rooney RJ, et al. Molecular profiles of parvalbumin-immunoreactive neurons in the superior temporal cortex in schizophrenia. J Neurogenet. 2014 Jun;28(1–2):70–85.

73. Mauney SA, Pietersen CY, Sonntag K-C, Woo T-UW. Differentiation of oligodendrocyte precursors is impaired in the prefrontal cortex in schizophrenia. Schizophr Res. 2015 Dec;169(1–3):374–80.

74. Mistry M, Gillis J, Pavlidis P. Genome-wide expression profiling of schizophrenia using a large combined cohort. Mol Psychiatry. 2013 Feb;18(2):215–25.

75. Choi KH, Higgs BW, Wendland JR, Song J, McMahon FJ, Webster MJ. Gene expression and genetic variation data implicate PCLO in bipolar disorder. Biol Psychiatry. 2011 Feb 15;69(4):353–9.

76. Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat Neurosci. 2016 Feb;19(2):335–46.

77. Mancarci O, Toker L, Tripathy S, Li B, Rocco B, Sibille E, 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 [Internet]. 2016 Nov 22; Available from: http://biorxiv.org/content/biorxiv/early/2016/11/22/089219.full.pdf

78. Atz M, Walsh D, Cartagena P, Li J, Evans S, Choudary P, et al. Methodological considerations for gene expression profiling of human brain. J Neurosci Methods. 2007 Jul 30;163(2):295–309.

79. Vawter MP, Tomita H, Meng F, Bolstad B, Li J, Evans S, et al. Mitochondrial-related gene expression changes are sensitive to agonal-pH state: implications for brain disorders. Mol Psychiatry. 2006 Jul;11(7):615, 663–79.

80. Sequeira PA, Martin MV, Vawter MP. The first decade and beyond of transcriptional profiling in schizophrenia. Neurobiol Dis. 2012 Jan 1;45(1):23–36.

81. Weis S, Llenos IC, Dulay JR, Elashoff M, Martínez-Murillo F, Miller CL. Quality control for microarray analysis of human brain samples: The impact of postmortem factors, RNA characteristics, and histopathology. J Neurosci Methods. 2007 Sep 30;165(2):198–209.

82. Hamberger A, Hyden H. Inverse enzymatic changes in neurons and glia during increased function and hypoxia. J Cell Biol. 1963 Mar;16:521–5.

83. Kato T, Murashita J, Kamiya A, Shioiri T, Kato N, Inubushi T. Decreased brain intracellular pH measured by P-31-MRS in bipolar disorder: a confirmation in drug-

free patients and correlation with white matter hyperintensity. Eur Arch Psychiatry Clin Neurosci. 1998 Dec;248(6):301–6.

84. Hamakawa H, Murashita J, Yamada N, Inubushi T, Kato N, Kato T. Reduced intracellular pH in the basal ganglia and whole brain measured by P-31-MRS in bipolar disorder. Psychiatry Clin Neurosci. 2004 Feb;58(1):82–8.

85. Johnson CP, Follmer RL, Oguz I, Warren LA, Christensen GE, Fiedorowicz JG, et al. Brain abnormalities in bipolar disorder detected by quantitative T1 rho mapping. Mol Psychiatry. 2015 Feb;20(2):201–6.

86. Chesler M, Kraig R. Intracellular Ph of Astrocytes Increases Rapidly with Cortical Stimulation. Am J Physiol. 1987 Oct;253(4):R666–70.

87. Montaño CM, Irizarry RA, Kaufmann WE, Talbot K, Gur RE, Feinberg AP, et al. Measuring cell-type specific differential methylation in human brain tissue. Genome Biol. 2013;14(8):R94.

88. Bowling K, Ramaker RC, Lasseigne BN, Hagenauer M, Hardigan A, Davis N, et al. Post-mortem molecular profiling of three psychiatric disorders reveals widespread dysregulation of cell-type associated transcripts and refined disease-related transcription changes. bioRxiv. 2016 Jun 29;061416.

89. Chase KA, Rosen C, Gin H, Bjorkquist O, Feiner B, Marvin R, et al. Metabolic and inflammatory genes in schizophrenia. Psychiatry Res. 2015 Jan 30;225(1–2):208–11.

90. Medina A, Watson SJ, Bunney W, Myers RM, Schatzberg A, Barchas J, et al. Evidence for alterations of the glial syncytial function in major depressive disorder. J Psychiatr Res. 2016 Jan;72:15–21.

91. Turner CA, Watson SJ, Akil H. The fibroblast growth factor family: neuromodulation of affective behavior. Neuron. 2012 Oct 4;76(1):160–74.

92. Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, et al. Functional organization of the transcriptome in human brain. Nat Neurosci. 2008 Nov;11(11):1271–82.

93. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR). 4th ed. Washington, D.C.: American Psychiatric Association; 2000.

94. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res. 2005;33(20):e175.

95. Coupel S, Moreau A, Hamidou M, Horejsi V, Soulillou J-P, Charreau B. Expression and release of soluble HLA-E is an immunoregulatory feature of endothelial cell activation. Blood. 2007 Apr 1;109(7):2806–14.

96. Tian H, McKnight SL, Russell DW. Endothelial PAS domain protein 1 (EPAS1), a transcription factor selectively expressed in endothelial cells. Genes Dev. 1997 Jan 1;11(1):72–82.

97. Tchorz JS, Tome M, Cloëtta D, Sivasankaran B, Grzmil M, Huber RM, et al. Constitutive Notch2 signaling in neural stem cells promotes tumorigenic features and astroglial lineage entry. Cell Death Dis. 2012;3:e325.

98. Boyles JK, Pitas RE, Wilson E, Mahley RW, Taylor JM. Apolipoprotein E associated with astrocytic glia of the central nervous system and with nonmyelinating glia of the peripheral nervous system. J Clin Invest. 1985 Oct;76(4):1501–13.

99. Fazzari P, Paternain AV, Valiente M, Pla R, Luján R, Lloyd K, et al. Control of cortical GABA circuitry development by Nrg1 and ErbB4 signalling. Nature. 2010 Apr 29;464(7293):1376–80.

100. Stephenson DT, Coskran TM, Kelly MP, Kleiman RJ, Morton D, O'Neill SM, et al. The distribution of phosphodiesterase 2A in the rat brain. Neuroscience. 2012 Dec 13;226:145–55.

101. Marszalek JR, Weiner JA, Farlow SJ, Chun J, Goldstein LS. Novel dendritic kinesin sorting identified by different process targeting of two related kinesins: KIF21A and KIF21B. J Cell Biol. 1999 May 3;145(3):469–79.

102. Rao JS, Harry GJ, Rapoport SI, Kim HW. Increased excitotoxicity and neuroinflammatory markers in postmortem frontal cortex from bipolar disorder patients. Mol Psychiatry. 2010 Apr;15(4):384–92.

103. Spiliotaki M, Salpeas V, Malitas P, Alevizos V, Moutsatsou P. Altered glucocorticoid receptor signaling cascade in lymphocytes of bipolar disorder patients. Psychoneuroendocrinology. 2006 Jul;31(6):748–60.

104. Le-Niculescu H, Patel SD, Bhat M, Kuczenski R, Faraone SV, Tsuang MT, et al. Convergent functional genomics of genome-wide association data for bipolar disorder: comprehensive identification of candidate genes, pathways and mechanisms. Am J Med Genet Part B Neuropsychiatr Genet Off Publ Int Soc Psychiatr Genet. 2009 Mar 5;150B(2):155–81.

105. Konopaske GT, Subburaju S, Coyle JT, Benes FM. Altered prefrontal cortical MARCKS and PPP1R9A mRNA expression in schizophrenia and bipolar disorder. Schizophr Res. 2015 May;164(1–3):100–8.

106. Glessner JT, Reilly MP, Kim CE, Takahashi N, Albano A, Hou C, et al. Strong synaptic transmission impact by copy number variations in schizophrenia. Proc Natl Acad Sci U S A. 2010 Jun 8;107(23):10584–9.

107. Ayoub MA, Angelicheva D, Vile D, Chandler D, Morar B, Cavanaugh JA, et al. Deleterious GRM1 mutations in schizophrenia. PloS One. 2012;7(3):e32849.

108. Frank RAW, McRae AF, Pocklington AJ, van de Lagemaat LN, Navarro P, Croning MDR, et al. Clustered coding variants in the glutamate receptor complexes of individuals with schizophrenia and bipolar disorder. PloS One. 2011;6(4):e19011.

109. Etain B, Dumaine A, Mathieu F, Chevalier F, Henry C, Kahn J-P, et al. A SNAP25 promoter variant is associated with early-onset bipolar disorder and a high expression level in brain. Mol Psychiatry. 2010 Jul;15(7):748–55.

110. Fatjó-Vilas M, Prats C, Pomarol-Clotet E, Lázaro L, Moreno C, González-Ortega I, et al. Involvement of NRN1 gene in schizophrenia-spectrum and bipolar disorders and its impact on age at onset and cognitive functioning. World J Biol Psychiatry Off J World Fed Soc Biol Psychiatry. 2016;17(2):129–39.

111. Volk DW, Chitrapu A, Edelson JR, Roman KM, Moroco AE, Lewis DA. Molecular mechanisms and timing of cortical immune activation in schizophrenia. Am J Psychiatry. 2015 Nov 1;172(11):1112–21.

112. Girgenti MJ, LoTurco JJ, Maher BJ. ZNF804a regulates expression of the schizophrenia-associated genes PRSS16, COMT, PDE4B, and DRD2. PloS One. 2012;7(2):e32404.

7. Supporting Information

Supplementary Material for:

INFERENCE OF CELL TYPE COMPOSITION FROM HUMAN BRAIN TRANSCRIPTOMIC DATASETS ILLUMINATES THE EFFECTS OF AGE, MANNER OF DEATH, DISSECTION, AND PSYCHIATRIC DIAGNOSIS

*Megan Hastings Hagenauer, Ph.D.[1], Anton Schulmann, M.D.[2], Jun Z. Li, Ph.D.[3], Marquis P. Vawter, Ph.D.[4], David M. Walsh, Psy.D.[4], Robert C. Thompson, Ph.D.[1], Cortney A. Turner, Ph.D.[1], William E. Bunney, M.D.[4], Richard M. Myers, Ph.D.[5], Jack D. Barchas, M.D.[6], Alan F. Schatzberg, M.D.[7], Stanley J. Watson, M.D., Ph.D.[1], Huda Akil, Ph.D.[1]

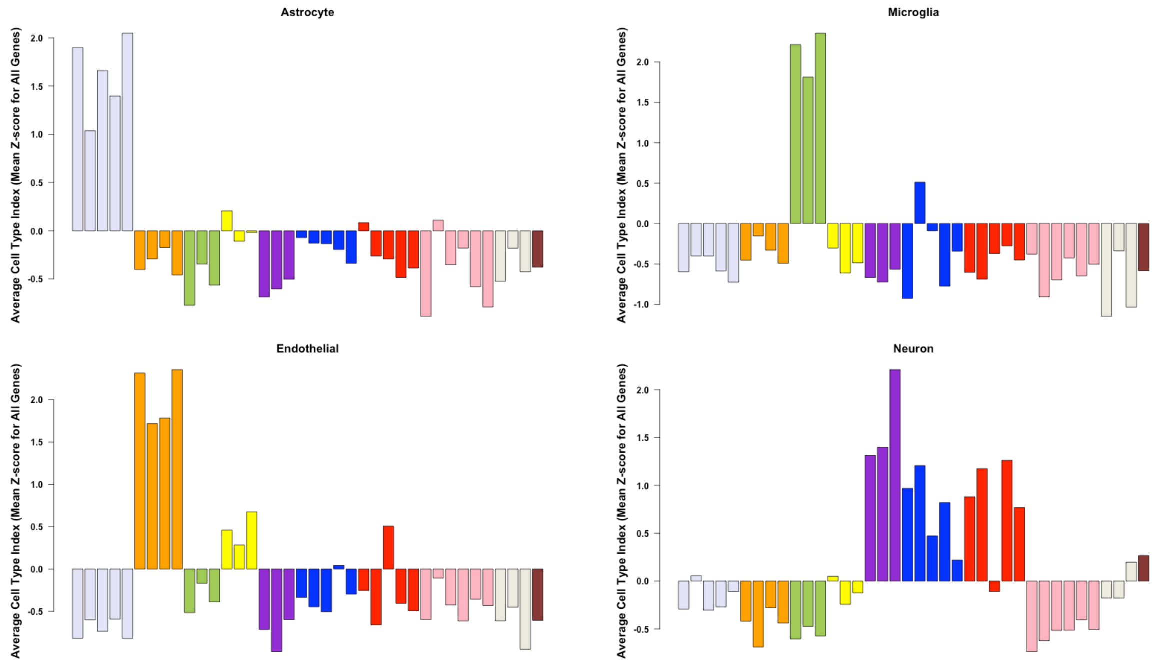
7.1 Detailed Methods and Results: Using Cell Type Specific Transcripts to Predict Relative Cell Content in Datasets from Purified Cells and Artificial Cell Mixtures

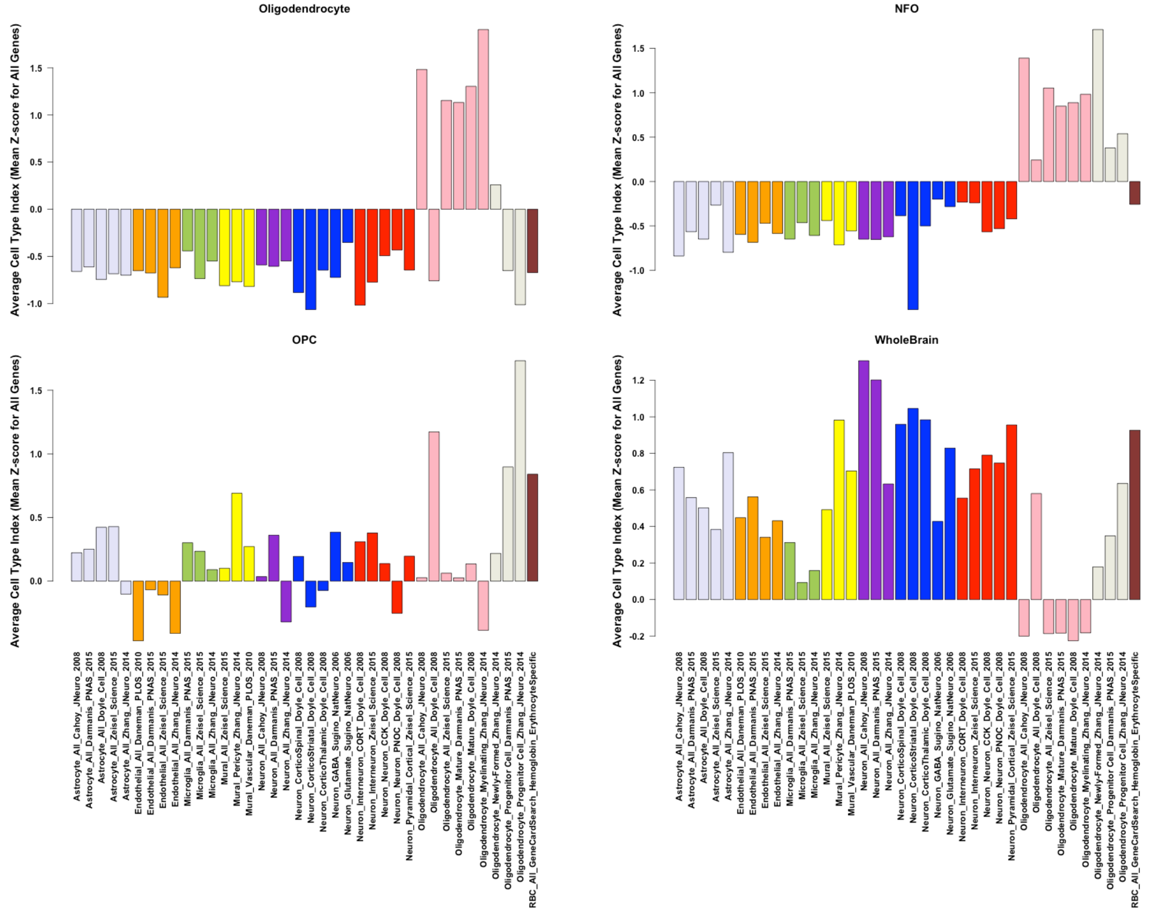
To validate our technique, we used the expression of the cell type specific transcripts included in our database to predict the relative balance of cell types in samples with known cell content (purified cells and artificial cell mixtures). To do this analysis, we used two RNA-Seq datasets: one derived from from purified cortical cell types in mice (n=17: two samples per purified cell type and 3 whole brain samples: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52564) (18), and one derived from 466 single-cells dissociated from freshly-resected human cortex (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67835) (2). To estimate the limitations and noise inherant in our technique, we also constructed *in silico* mixtures of 100 cells with known percentages of each cell type by randomly sampling from each dataset (with replacement).

The RNA-Seq data that we downloaded from GEO (Gene expression Omnibus) was already in the format of FPKM values (Fragments Per Kilobase of exon model per million mapped fragments) (18) or counts per gene (2). To stabilize the variance in the data, we used a log transformation (base 2), and then filtered out the data for any genes that completely lacked variation across samples (sd=0). Within the mouse dataset (18) this filtering decreased the dataset from 22462 genes to 17148 genes. Within the human dataset (2), this filtering decreased the dataset from 22085 genes to 21627 genes.

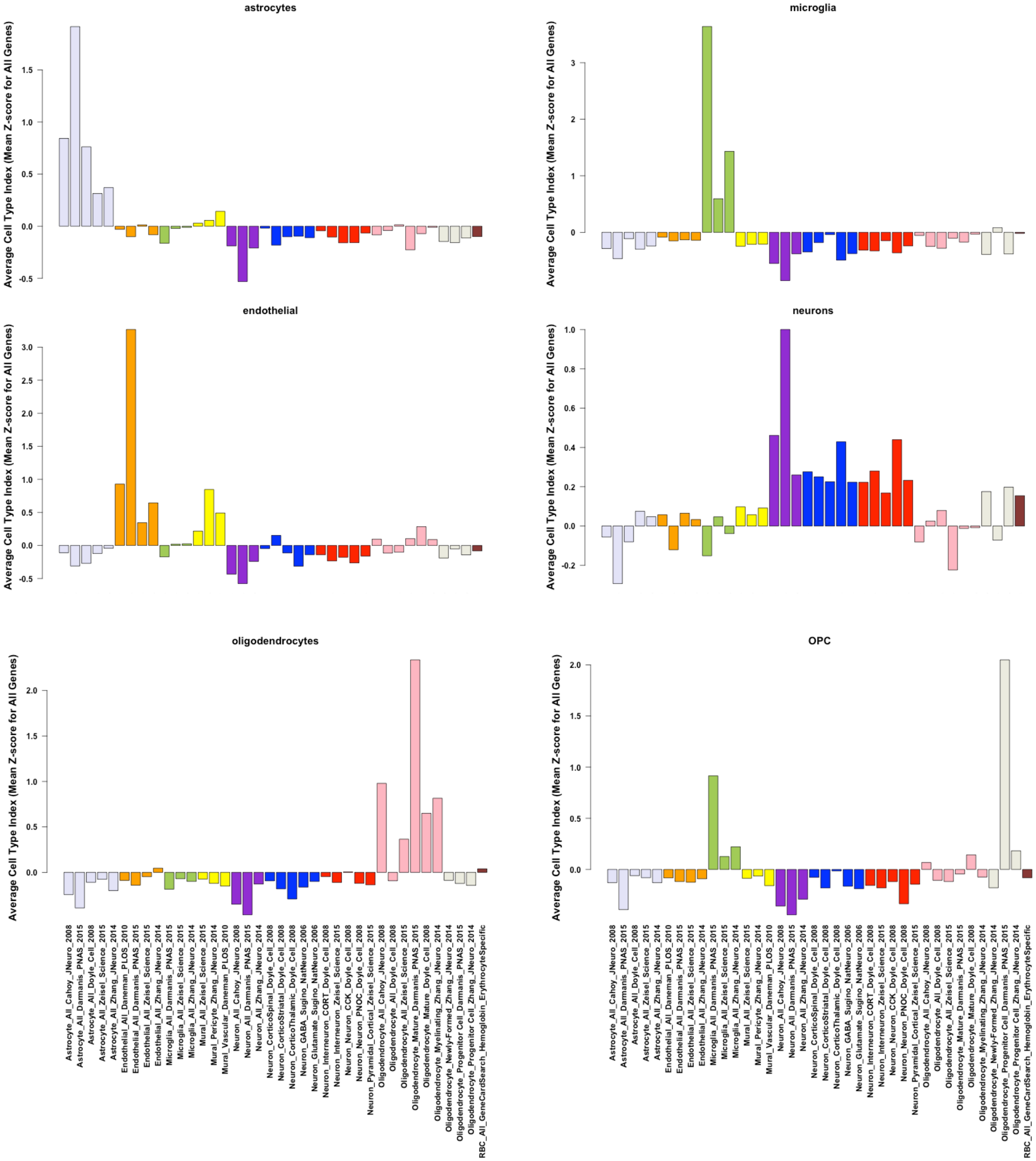
Then, using the methods now found in the BrainInABlender package, we extracted the data for genes previously identified as having cell type specific expression in our curated database. Within the mouse dataset, there were data from 2513 genes that aligned with 2914 entries in our database of cell type specific transcripts (as matched by official gene symbol). Within the human dataset, there were data from 2374 genes that aligned with 2882 entries in our database of cell type specific transcripts (as matched by gene symbol). We centered and scaled the expression levels for each gene across samples (mean=0, sd=1) to prevent genes with more variable signal from exerting disproportionate influence, and then, for each sample, averaged this value across the transcripts identified in each publication as specific to a particular cell type. This created 38 cell type signatures derived from the cell type specific genes identified by the eight publications (*"Cell Type Indices"*), each of which predicted the relative content for one of the 10 primary cell types in our cortical samples. All of the R script documenting these analyses can be found at https://github.com/hagenaue/CellTypeAnalyses_Darmanis and https://github.com/hagenaue/CellTypeAnalyses_Zhang.

We found that the statistical cell type indices easily predicted the cell type identities of the original samples (**Suppl. Figure 1, Suppl. Figure 2).** This was true regardless of the publication from which the cell type specific genes were derived: cell type specific gene lists derived from publications using different species (human vs. mouse), platforms (microarray vs. RNA-Seq), methodologies (florescent cell sorting vs. suspension), or statistical stringency all performed fairly equivalently, with some minor exception. Occassionally, we found that the cell type indices associated with cell type specific gene lists derived from TRAP methodology (15) did not properly predict the cell identity of the samples, and in general the cell type indices associated with immature oligodendrocytes were somewhat inconsistent. As would be expected, the cell type indices derived from cell type specific genes identified by the same publication that produced the test datasets (2,18) were (by definition) superb predictors of the sample cell identity in their own dataset, and were thus excluded from later validation analyses.


Astrocyte
Microglia
Endothelial
Neuron
Average Cell Type Index (Mean Z-score for All Genes)
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0

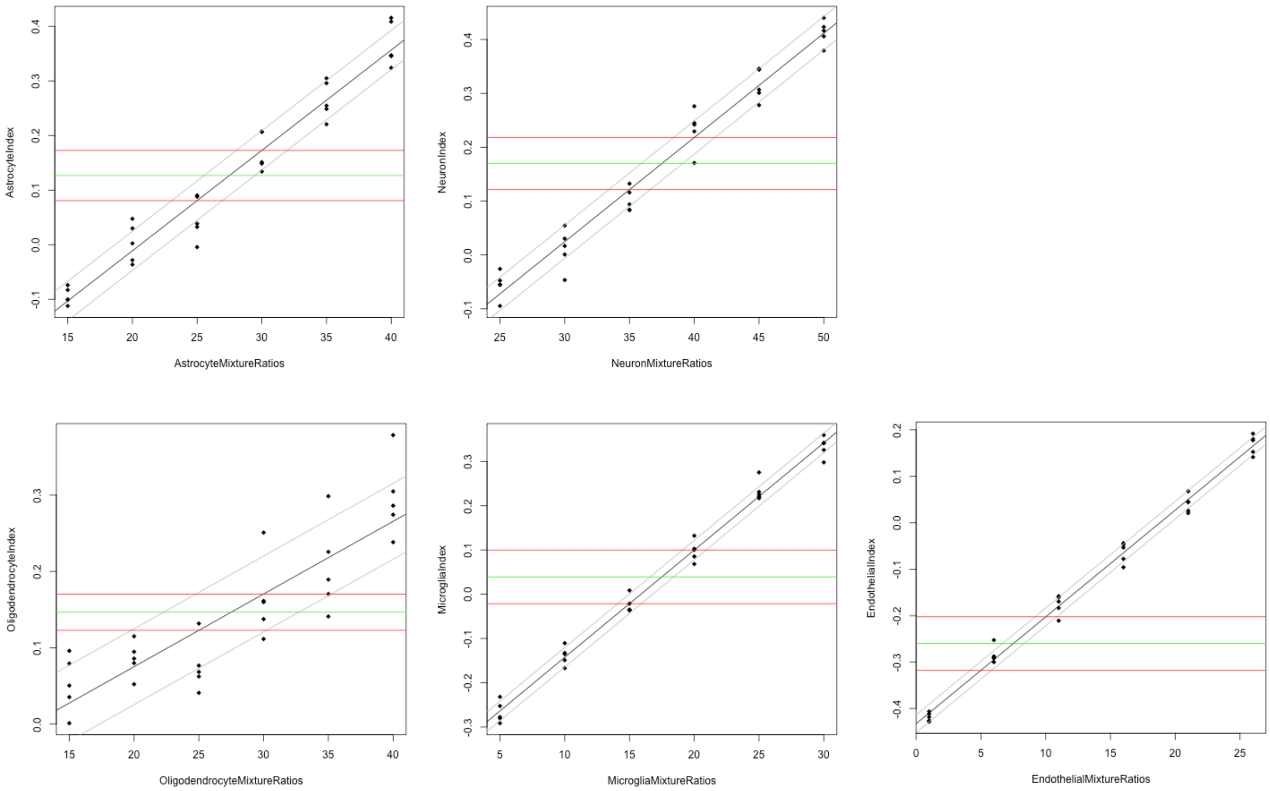

Oligodendrocyte
NFO
OPC
WholeBrain
Average Cell Type Index (Mean Z-score for All Genes)
1.5
1.2
1.0
0.8
0.6
0.5
0.4
0.2
0.0
-0.2
-0.5
-1.0
Astrocyte_All_Cahoy_JNeuro_2008
Astrocyte_All_Darmanis_PNAS_2015
Astrocyte_All_Doyle_Cell_2008
Astrocyte_All_Zeisel_Science_2015
Astrocyte_All_Zhang_JNeuro_2014
Endothelial_All_Daneman_PLOS_2010
Endothelial_All_Darmanis_PNAS_2015
Endothelial_All_Zeisel_Science_2015
Endothelial_All_Zhang_JNeuro_2014
Microglia_All_Darmanis_PNAS_2015
Microglia_All_Zeisel_Science_2015
Microglia_All_Zhang_JNeuro_2014
Mural_All_Zeisel_Science_2015
Mural_Pericyte_Zhang_JNeuro_2014
Mural_Vascular_Daneman_PLOS_2010
Neuron_All_Cahoy_JNeuro_2008
Neuron_All_Darmanis_PNAS_2015
Neuron_All_Zhang_JNeuro_2014
Neuron_CorticoSpinal_Doyle_Cell_2008
Neuron_CorticoStriatal_Doyle_Cell_2008
Neuron_CorticoThalamic_Doyle_Cell_2008
Neuron_GABA_Sugino_NatNeuro_2006
Neuron_Glutamate_Sugino_NatNeuro_2006
Neuron_Interneuron_CORT_Doyle_Cell_2008
Neuron_Interneuron_Zeisel_Science_2015
Neuron_Neuron_CCK_Doyle_Cell_2008
Neuron_Neuron_PNOC_Doyle_Cell_2008
Neuron_Pyramidal_Cortical_Zeisel_Science_2015
Oligodendrocyte_All_Cahoy_JNeuro_2008
Oligodendrocyte_All_Doyle_Cell_2008
Oligodendrocyte_All_Zeisel_Science_2015
Oligodendrocyte_Mature_Darmanis_PNAS_2015
Oligodendrocyte_Mature_Doyle_Cell_2008
Oligodendrocyte_Myelinating_Zhang_JNeuro_2014
Oligodendrocyte_Newly-Formed_Zhang_JNeuro_2014
Oligodendrocyte_Progenitor Cell_Darmanis_PNAS_2015
Oligodendrocyte_Progenitor Cell_Zhang_JNeuro_2014
RBC_All_GeneCardSearch_Hemoglobin_ErythrocyteSpecific

Suppl. Figure 1. The cell content predictions derived from cell type specific transcripts originating from different publications successfully predict sample cell type in mouse purified cell type RNA-Seq data. *The sample cell type in a mouse purified cell type RNA-Seq dataset (18) was predicted equally well by cell type indices derived from cell type specific transcripts originating from publications using different species, methodologies, and platforms. The actual sample cell type is indicated in the main heading above the plot (NFO: "newly-formed oligodendrocyte"), and each bar represents the average for two samples for each cell type index (identified by primary cell type, subtype, and publication on the x-axis). The cell type indices that fall within a particular primary category of cell are further identified by color (*lavender: *astrocytes,* orange: *endothelial,* green: *microglia,* yellow: *mural,* purple: *neuron_all,* blue: *neuron_projection,* red: *neuron_interneuron,* pink: *oligodendrocyte,* gray: *oligodendrocyte progenitor cell (OPC),* brown: *red blood cell (RBC)).*

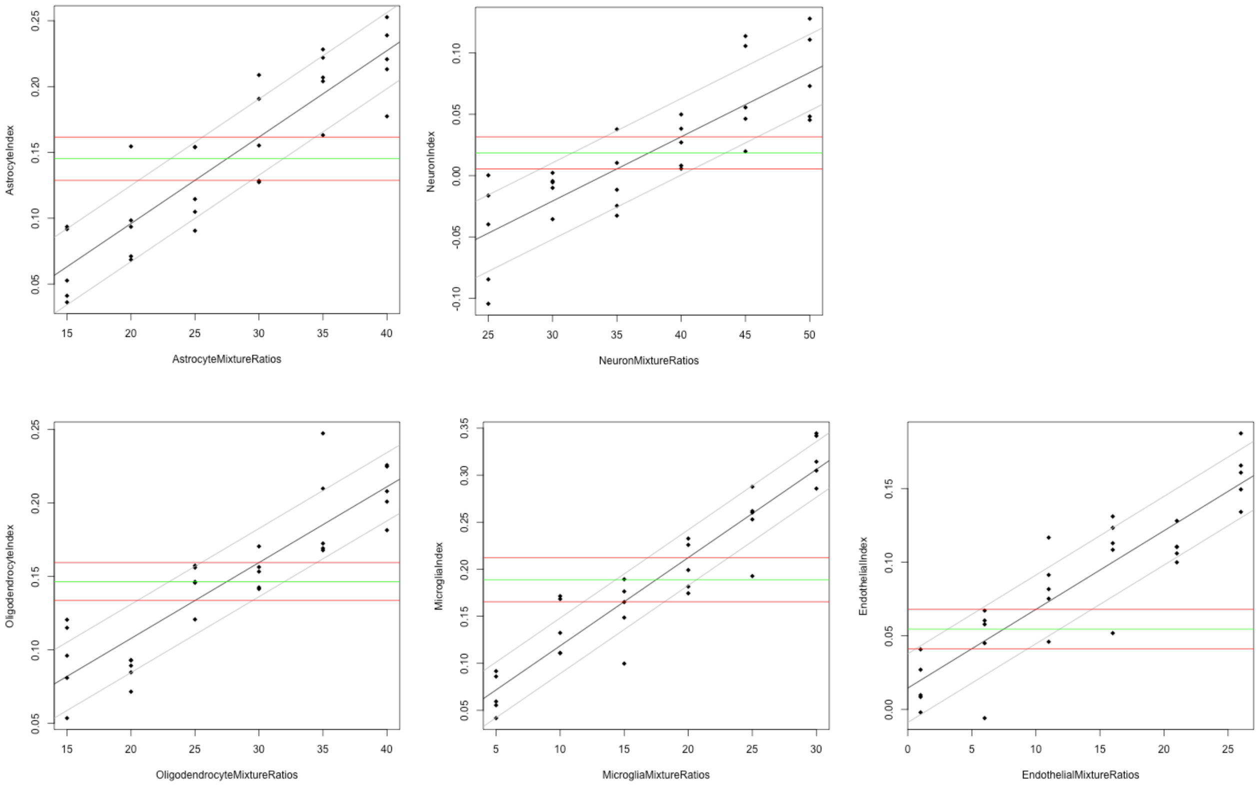

astrocytes
Average Cell Type Index (Mean Z-score for All Genes)
1.5
1.0
0.5
0.0
-0.5
microglia
Average Cell Type Index (Mean Z-score for All Genes)
3
2
1
0
endothelial
Average Cell Type Index (Mean Z-score for All Genes)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
neurons
Average Cell Type Index (Mean Z-score for All Genes)
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
oligodendrocytes
Average Cell Type Index (Mean Z-score for All Genes)
2.0
1.5
1.0
0.5
0.0
Astrocyte_All_Cahoy_JNeuro_2008
Astrocyte_All_Darmanis_PNAS_2015
Astrocyte_All_Doyle_Cell_2008
Astrocyte_All_Zeisel_Science_2015
Astrocyte_All_Zhang_JNeuro_2014
Endothelial_All_Daneman_PLOS_2010
Endothelial_All_Darmanis_PNAS_2015
Endothelial_All_Zeisel_Science_2015
Endothelial_All_Zhang_JNeuro_2014
Microglia_All_Darmanis_PNAS_2015
Microglia_All_Zeisel_Science_2015
Microglia_All_Zhang_JNeuro_2014
Mural_All_Zeisel_Science_2015
Mural_Pericyte_Zhang_JNeuro_2014
Mural_Vascular_Daneman_PLOS_2010
Neuron_All_Cahoy_JNeuro_2008
Neuron_All_Darmanis_PNAS_2015
Neuron_All_Zhang_JNeuro_2014
Neuron_CorticoSpinal_Doyle_Cell_2008
Neuron_CorticoStriatal_Doyle_Cell_2008
Neuron_CorticoThalamic_Doyle_Cell_2008
Neuron_GABA_Sugino_NatNeuro_2006
Neuron_Glutamate_Sugino_NatNeuro_2006
Neuron_Interneuron_CORT_Doyle_Cell_2008
Neuron_Interneuron_Zeisel_Science_2015
Neuron_Neuron_CCK_Doyle_Cell_2008
Neuron_Neuron_PNOC_Doyle_Cell_2008
Neuron_Pyramidal_Cortical_Zeisel_Science_2015
Oligodendrocyte_All_Cahoy_JNeuro_2008
Oligodendrocyte_All_Doyle_Cell_2008
Oligodendrocyte_All_Zeisel_Science_2015
Oligodendrocyte_Mature_Darmanis_PNAS_2015
Oligodendrocyte_Mature_Doyle_Cell_2008
Oligodendrocyte_Myelinating_Zhang_JNeuro_2014
Oligodendrocyte_Newly-Formed_Zhang_JNeuro_2014
Oligodendrocyte_Progenitor Cell_Darmanis_PNAS_2015
Oligodendrocyte_Progenitor Cell_Zhang_JNeuro_2014
RBC_All_GeneCardSearch_Hemoglobin_ErythrocyteSpecific
OPC
Average Cell Type Index (Mean Z-score for All Genes)
2.0
1.5
1.0
0.5
0.0
Astrocyte_All_Cahoy_JNeuro_2008
Astrocyte_All_Darmanis_PNAS_2015
Astrocyte_All_Doyle_Cell_2008
Astrocyte_All_Zeisel_Science_2015
Astrocyte_All_Zhang_JNeuro_2014
Endothelial_All_Daneman_PLOS_2010
Endothelial_All_Darmanis_PNAS_2015
Endothelial_All_Zeisel_Science_2015
Endothelial_All_Zhang_JNeuro_2014
Microglia_All_Darmanis_PNAS_2015
Microglia_All_Zeisel_Science_2015
Microglia_All_Zhang_JNeuro_2014
Mural_All_Zeisel_Science_2015
Mural_Pericyte_Zhang_JNeuro_2014
Mural_Vascular_Daneman_PLOS_2010
Neuron_All_Cahoy_JNeuro_2008
Neuron_All_Darmanis_PNAS_2015
Neuron_All_Zhang_JNeuro_2014
Neuron_CorticoSpinal_Doyle_Cell_2008
Neuron_CorticoStriatal_Doyle_Cell_2008
Neuron_CorticoThalamic_Doyle_Cell_2008
Neuron_GABA_Sugino_NatNeuro_2006
Neuron_Glutamate_Sugino_NatNeuro_2006
Neuron_Interneuron_CORT_Doyle_Cell_2008
Neuron_Interneuron_Zeisel_Science_2015
Neuron_Neuron_CCK_Doyle_Cell_2008
Neuron_Neuron_PNOC_Doyle_Cell_2008
Neuron_Pyramidal_Cortical_Zeisel_Science_2015
Oligodendrocyte_All_Cahoy_JNeuro_2008
Oligodendrocyte_All_Doyle_Cell_2008
Oligodendrocyte_All_Zeisel_Science_2015
Oligodendrocyte_Mature_Darmanis_PNAS_2015
Oligodendrocyte_Mature_Doyle_Cell_2008
Oligodendrocyte_Myelinating_Zhang_JNeuro_2014
Oligodendrocyte_Newly-Formed_Zhang_JNeuro_2014
Oligodendrocyte_Progenitor Cell_Darmanis_PNAS_2015
Oligodendrocyte_Progenitor Cell_Zhang_JNeuro_2014
RBC_All_GeneCardSearch_Hemoglobin_ErythrocyteSpecific

Suppl. Figure 2. The cell content predictions derived from cell type specific transcripts originating from different publications successfully predict sample cell type in human single cell RNA-Seq data. *The sample cell type in a human single cell RNA-Seq dataset (2) was predicted equally well by cell type indices derived from cell type specific transcripts originating from publications using different species, methodologies, and platforms. The sample cell type (as identified in the publication) is indicated in the main heading above the plot, and each bar represents the average cell type index (identified by primary cell type, subtype, and publication on the x-axis) for all samples of that cell type. The cell type indices that fall within a particular primary category of cell are further identified by color (*lavender: *astrocytes,* orange: *endothelial,* green: *microglia,* yellow: *mural,* purple: *neuron_all,* blue: *neuron_projection,* red: *neuron_interneuron,* pink: *oligodendrocyte,* gray: *oligodendrocyte progenitor cell (OPC),* brown: *red blood cell (RBC)).*

For further analyses, individual cell type indices were averaged within each of ten primary categories: astrocytes, endothelial cells, mural cells, microglia, immature and mature oligodendrocytes, red blood cells, interneurons, projection neurons, and indices derived from neurons in general, with any genes that were identifed as being specific to more than one category removed (*e.g.,* a gene identified as being specifically expressed in both microglia and endothelial cells). This led to ten consolidated primary cell-type indices for each sample. We then examined the relationship between these consolidated cell type indices and actual cell content in artificial mixtures of 100 cells generated *in silico* by randomly sampling from the purified cell datasets (with replacement). We found that the consolidated cell type indices strongly predicted the percentage of their respective cell type included in our artificial mixtures of 100 cells in a linear manner **(Suppl. Figure 3, Suppl. Figure 4)** across a range of values likely to encompass the true proportion of these cells in our cortical samples. The amount of noise present in these predictions varied by data type, with the predictions generated from single-cell data having substantially more noise than that generated from pooled, purified cells, but even the noiser data was associated with most of the data (+/- 1 stdev) falling within +/- 5% of the prediction. Therefore, we conclude that cell type indices are a relatively easy manner to estimate relative cell type balance across samples.



Suppl. Figure 3. Cell type indices successfully predict the percentage of cells of a particular type in artificial mixtures of 100 cells created using mouse purified cell type RNA-Seq data. *Depicted are the cell type indices (y-axis) calculated for mixed cell samples generated* in silico *using random sampling (with replacement) from a mouse purified cell type RNA-Seq dataset (18). Each sample contains 100 cells total, with a designated percentage of the cell type of interest (x-axis), with the percentages designed to roughly span what might be found in cortical tissue samples. The black best fit line (as defined by a linear model) is accompanied by the standard error of the regression (gray), and the green and red lines are visual guides to help illustrate a 5% increase in the cell type of interest.*

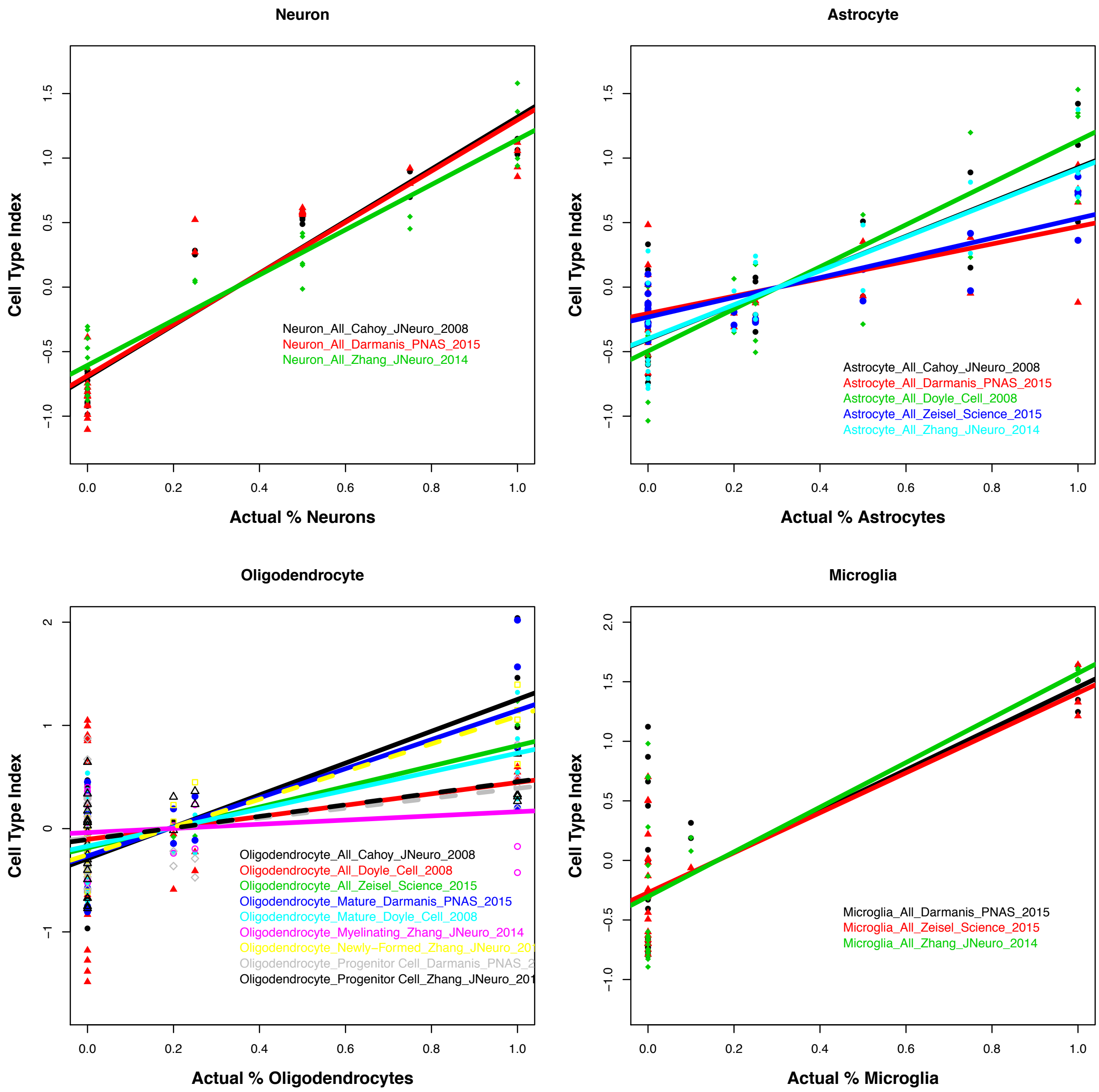


Suppl. Figure 4. Cell type indices successfully predict the percentage of cells of a particular type in artificial mixtures of 100 cells created using human single-cell RNA-Seq data. *Depicted are the cell type indices (y-axis) calculated for mixed cell samples generated* in silico *using random sampling (with replacement) from a human single cell RNA-Seq dataset (2). Each sample contains 100 cells total, with a designated percentage of the cell type of interest (x-axis), with the percentages designed to roughly span what might be found in cortical tissue samples. The black best fit line (as defined by a linear model) is accompanied by the standard error of the regression (gray), and the green and red lines are visual guides to help illustrate a 5% increase in the cell type of interest. Note the greater amount of variation present in the predictions for this dataset (based on single-cell data) versus the predictions based on mouse purified cell data (**Suppl. Figure 3**).*

As further validation, we determined whether relative cell type balance could be accurately deciphered from microarray data for samples containing artificially-generated mixtures of cultured cells (GSE19380; (12)). The cells used to make these mixtures were cultured from the cerebral cortices of rat P1 pups. The microarray profiling was then performed using a Affymetrix Rat Genome 230 2.0 Array, We downloaded the pre-processed gene expression dataset from GEO using the R package *GEOquery.* According to the methods published on GEO, this data had already undergone probe set summarization and normalization using robust multi-array averaging (RMA, *affy* package), including background subtraction, summarization by median polish, log (base 2) transformation, and quantile normalization. We used the R package *GEOquery* to extract out the description of the cell type mixture associated with each sample, and then used this data to construct a new matrix that contained the percent of each cell type (columns: neuron, astrocyte, oligodendrocyte, microglia) found in each sample. We then predicted the cell content of each sample from the microarray data using BrainInABlender, and plotted these predictions against the actual percent of each cell type found in the mixtures. We made these plots both for predictions derived from cell type specific gene lists from particular publications (**Suppl. Figure 6**) and after averaging these individual cell type indices within each of ten primary categories, with any genes that were identifed as being specific to more than one category removed (*e.g.,* a gene identified as being specifically expressed in both microglia and endothelial cells, **Suppl. Figure 5).** These results are included in main text of paper (**Figure 3**). The code for all of these analyses can be found at: https://github.com/hagenaue/CellTypeAnalyses_KuhnMixtures/tree/master.

Cell Type Index	Number of probesets in the microarray that represent cell type specific genes according to each publication	Percent that were truly specific to that cell type (not identified as "specific" to another category of cell type in a different publication)
Astrocyte_All_Cahoy_JNeuro_2008	47	87%
Astrocyte_All_Darmanis_PNAS_2015	14	93%
Astrocyte_All_Doyle_Cell_2008	12	100%
Astrocyte_All_Zeisel_Science_2015	181	88%
Astrocyte_All_Zhang_JNeuro_2014	32	78%
Endothelial_All_Daneman_PLOS_2010	34	76%
Endothelial_All_Darmanis_PNAS_2015	14	93%
Endothelial_All_Zeisel_Science_2015	261	90%
Endothelial_All_Zhang_JNeuro_2014	30	83%
Microglia_All_Darmanis_PNAS_2015	17	94%
Microglia_All_Zeisel_Science_2015	305	91%
Microglia_All_Zhang_JNeuro_2014	26	88%
Mural_All_Zeisel_Science_2015	114	93%
Mural_Pericyte_Zhang_JNeuro_2014	32	69%
Mural_Vascular_Daneman_PLOS_2010	36	64%
Neuron_All_Cahoy_JNeuro_2008	60	63%
Neuron_All_Darmanis_PNAS_2015	18	72%
Neuron_All_Zhang_JNeuro_2014	22	68%
Neuron_CorticoSpinal_Doyle_Cell_2008	17	59%
Neuron_CorticoStriatal_Doyle_Cell_2008	16	6%
Neuron_CorticoThalamic_Doyle_Cell_2008	14	64%
Neuron_GABA_Sugino_NatNeuro_2006	23	83%
Neuron_Glutamate_Sugino_NatNeuro_2006	48	81%
Neuron_Interneuron_CORT_Doyle_Cell_2008	13	77%
Neuron_Interneuron_Zeisel_Science_2015	259	90%
Neuron_Neuron_CCK_Doyle_Cell_2008	12	58%
Neuron_Neuron_PNOC_Doyle_Cell_2008	18	67%
Neuron_Pyramidal_Cortical_Zeisel_Science_2015	189	88%
Oligodendrocyte_All_Cahoy_JNeuro_2008	33	94%
Oligodendrocyte_All_Doyle_Cell_2008	19	74%
Oligodendrocyte_All_Zeisel_Science_2015	323	93%
Oligodendrocyte_Mature_Darmanis_PNAS_2015	15	100%
Oligodendrocyte_Mature_Doyle_Cell_2008	18	72%
Oligodendrocyte_Myelinating_Zhang_JNeuro_2014	34	100%
Oligodendrocyte_Newly-Formed_Zhang_JNeuro_2014	31	65%
Oligodendrocyte_Progenitor Cell_Darmanis_PNAS_2015	15	73%
Oligodendrocyte_Progenitor Cell_Zhang_JNeuro_2014	32	59%
RBC_All_GeneCardSearch_Hemoglobin_ErythrocyteSpecific	5	100%

Suppl. Figure 5. Identifying non-specific "cell-type specific genes": An example from dataset GSE19380 of the number of probesets that represented genes identified as cell type specific in each publication in our database vs. the percentage that were actually found to truly specific to that cell type (i.e., not identified as "specific" to another category of cell type in a different publication). *The data from genes that were identifed as being specific to more than one category of cell type (*e.g., *a gene identified as being specifically expressed in both microglia and endothelial cells) was removed before averaging the individual cell type indices within each of ten primary categories (astrocytes, endothelial cells, mural cells, microglia, immature and mature oligodendrocytes, red blood cells, interneurons, projection neurons, and indices derived from neurons in general) to create the ten consolidated primary cell-type indices used throughout our paper.*



Suppl. Figure 6. Validation of Relative Cell Content Predictions. *A) Using a microarray dataset derived from samples that contained artificially-generated mixtures of cultured cells (GSE19380; (12)), we found that our relative cell content predictions ("cell type indices") closely reflected actual known content, except that the percentage of cultured oligodendrocytes included in the mixtures was better predicted using cell type specific gene lists derived from immature oligodendrocytes instead of mature oligodendrocytes.*

7.2 Comparison of Our Method vs. PSEA: Predicting Cell Identity in a Human Single-Cell RNA-Seq Dataset

Although we generated our method independently to address microarray analysis questions that arose within the Pritzker Neuropsychiatric Consortium, we later discovered that it was quite similar to the technique of population-specific expression analysis (PSEA) introduced by (12) with several notable differences. Similar to our method, PSEA is a carefully-validated analysis method which aims to estimate cell type-differentiated disease effects from microarray data derived from brain tissue of heterogeneous composition and approaches this problem by including the averaged, normalized expression of cell type specific markers within a larger linear model that is used to estimate differential expression in microarray data (10–12). Analyses using PSEA similarly indicated that individual variability in neuronal, astrocytic, oligodendrocytic, and microglial cell content was sufficient to account for substantial variability in the vast majority of probe sets in microarray data from human brain samples, even within non-diseased samples (12). The differences between our techniques are mostly due to the recent growth of the literature documenting cell type specific expression in brain cell types. PSEA uses a very small set of markers (4-7) to represent each cell type, and screens these markers for tight co-expression within the dataset of interest, since co-expression networks have been previously demonstrated to often represent cell type signatures in the data (92). This is essential for the analysis of microarray data for brain regions that have not been well characterized for cell type specific expression (e.g., the substantia nigra), but risks the possibility of closely tracking variability in a particular cell function instead of cell content (as described in our results related to aging). Our analysis predominantly focused on the well-studied cortex, thus enabling us to expand our analysis to include hundreds of cell type specific markers derived from a variety of experimental techniques.

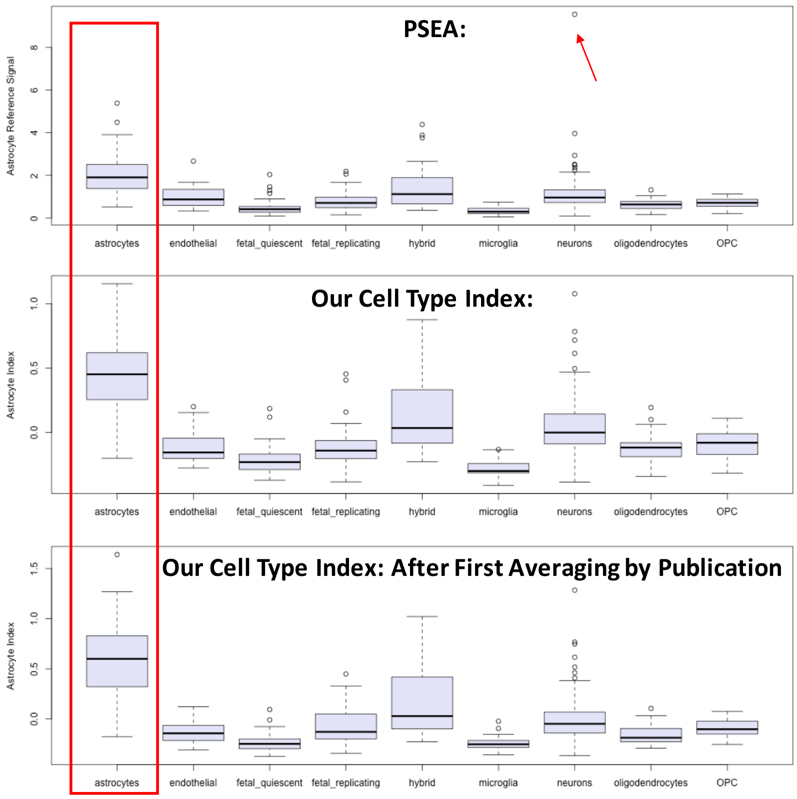
Our manner of normalizing data also differs: PSEA normalizes the expression values for each gene by dividing by the average expression of that gene across samples, whereas we use z-score normalization, both at the level of the individual transcript and later at the level of the gene level summary data. Due to the dependence of PSEA on ratios, genes that have average expression values that are close to zero can end up with normalized values that are extremely high for a handful of samples. For microarray data, this form of normalization should function well because log2 expression values rarely drop below 5. However, within RNA-Seq, counts of zero are quite common and therefore we suspected that the ratio-form of normalization used by PSEA might not function optimally for this data type.

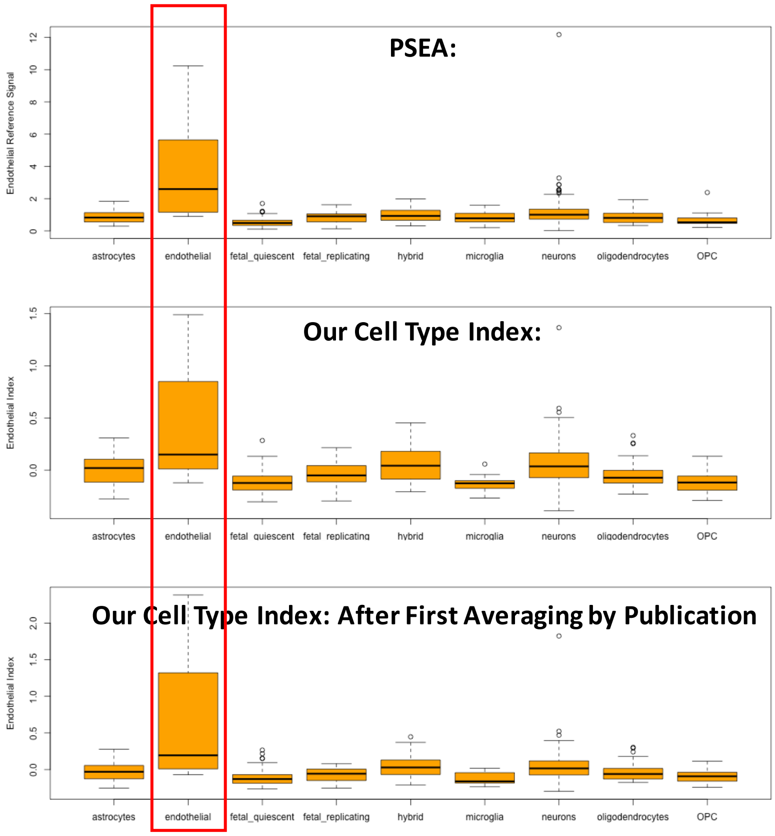
Therefore, we decided to run a head-to-head comparison of our method and PSEA using a single-cell RNA-Seq dataset derived from freshly-resected human cortex (2). To make the comparison as interpretable as possible, we used the same list of cell type specific genes for both methods: the cell type specific genes remaining in our database following the removal of all transcripts that were found to be "specifically expressed" in multiple categories of cell types (e.g., a transcript that is "specific" to both astrocytes and neurons). In order to avoid circular reasoning, we also did not include any cell type specific genes that had originally been identified by the publication currently used as the test dataset (2). Then we extracted the variance-stabilized and filtered data (see **Section 7.1)** for the cell type specific genes. For PSEA, we downloaded the PSEA package from Bioconductor (https://www.bioconductor.org/packages/release/bioc/vignettes/PSEA/inst/doc/PSEA.pdf) and used the marker() function to calculate the "Reference Signal" for the most common primary categories of cell types (astrocytes, endothelial cells, microglia, mature oligodendrocytes, and neurons in general). For our method, we used a procedure similar to that used in the manuscript. We applied a z-score transformation to the data for each gene (mean=0, sd=1), and then either averaged by the primary cell type category (to conduct an analysis most similar to PSEA), or averaged the data from the cell type specific genes identified by each publication, followed by averaging by primary cell type category (to create consolidated cell type indices similar to those used in most of our manuscript).

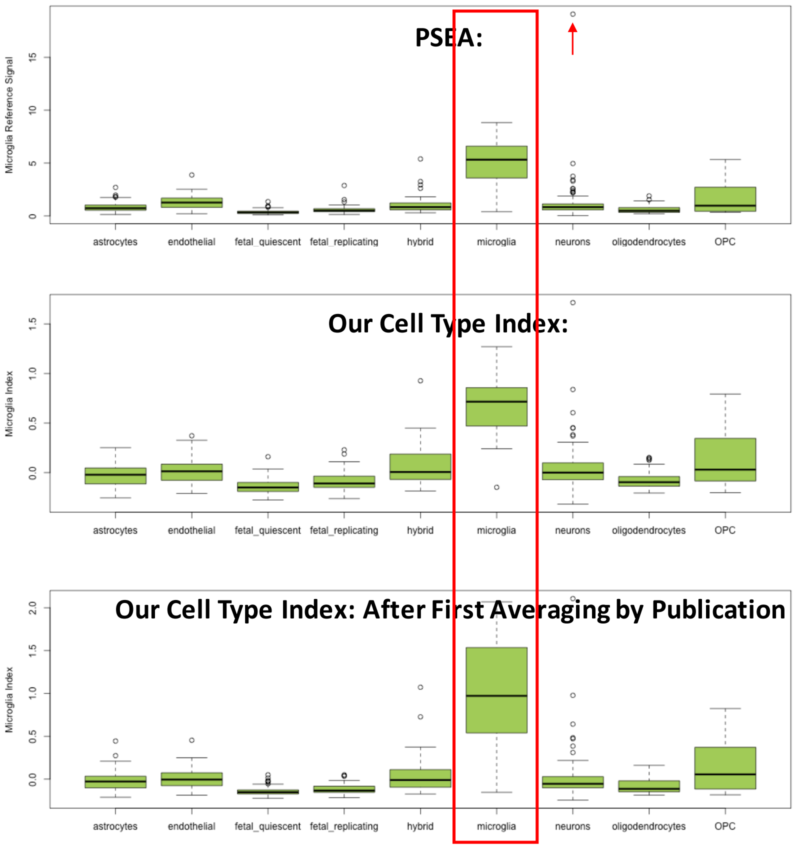
To compare the efficacy of each method, we ran a linear model to determine the percentage of variation in the population "reference signal" (PSEA) or "cell type index" (our method) accounted for by the cell type identities assigned to each cell in the original publication (2). We found that both the population reference signals (PSEA) and cell type indices (our method) for each cell were strongly related to their previously-assigned cell type identity, but in general the relationship was stronger when using our method: on average, 34% of the variation in the reference signal for each cell type was accounted for by cell identity, whereas an average of either 45% or 49% of the variation in our cell type indices was accounted for by cell identity using either the simplified or consolidated versions of our method, respectively (**Suppl. Figure 7**). An illustration of this improvement can be found in **Suppl. Figure 8**: note the presence of extreme outliers in the population reference signal when using the PSEA method. We conclude that the simple use of a different normalization method is sufficient to make our method a more effective manner of predicting cell type balance in some datasets. We also find that averaging the predictions drawn from the cell type specific genes identified by multiple publications into a consolidated index produces some additional improvement.

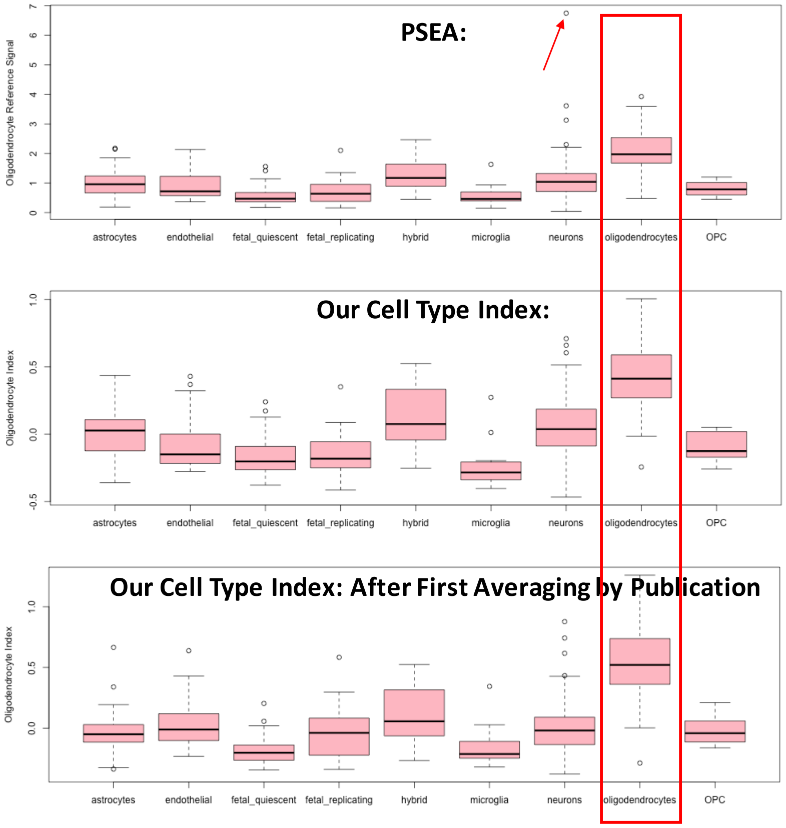
	Method of deriving a statistical cell type signal:		
Signal from cell type specific genes for:	**PSEA** (mean signal ratio average)	**Our Cell Type Indices** (z-score average)	**Our Cell Type Indices:** After first averaging by Publication
Astrocytes	34%	52%	57%
Oligodendrocytes	38%	45%	50%
Microglia	36%	42%	51%
Endothelial	30%	28%	33%
Neurons	33%	57%	53%

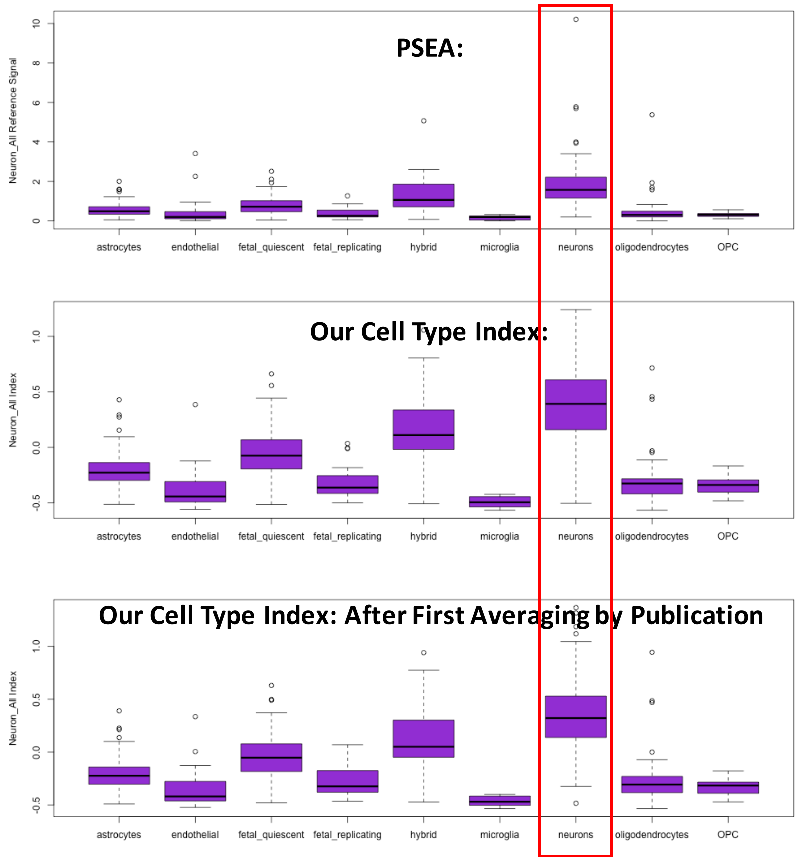
Suppl. Figure 7. The method for deriving predicted relative cell content determines the strength of the relationship with sample cell type. *Depicted below is a comparison of the efficacy of three different manners of predicting the relative cell content of samples (columns) in a human single-cell RNA-seq dataset (2)**:** 1) the "population reference signal" generated by PSEA, 2) a simplified version of our method that is meant to be relatively analogous to PSEA (a simple average of the z-score-transformed data for all genes specific to a particular cell type in our database), 3) the version of our method used in this manuscript, which consolidates the predictions derived from the cell type specific genes identified in different publications. For the predicted relative content of each of the major cell types (rows) derived using these different methods, the table provides the percentage of variation (r-squared) that is accounted for by the original cell type identities of the samples provided by the publication (2). Overall, there is a strong relationship between the predictions generated by all methods and sample cell type identity, but the method used in this manuscript produces predictions that best fit sample cell type.*


PSEA:
Astrocyte Reference Signal
0
2
4
6
8
Our Cell Type Index:
Astrocyte Index
0.0
0.5
1.0
Our Cell Type Index: After First Averaging by Publication
Astrocyte Index
0.0
0.5
1.0
1.5
astrocytes
endothelial
fetal_quiescent
fetal_replicating
hybrid
microglia
neurons
oligodendrocytes
OPC


PSEA:
Endothelial Reference Signal
0
2
4
6
8
10
12
Our Cell Type Index:
Endothelial Index
0.0
0.5
1.0
1.5
Our Cell Type Index: After First Averaging by Publication
Endothelial Index
0.0
0.5
1.0
1.5
2.0
astrocytes
endothelial
fetal_quiescent
fetal_replicating
hybrid
microglia
neurons
oligodendrocytes
OPC

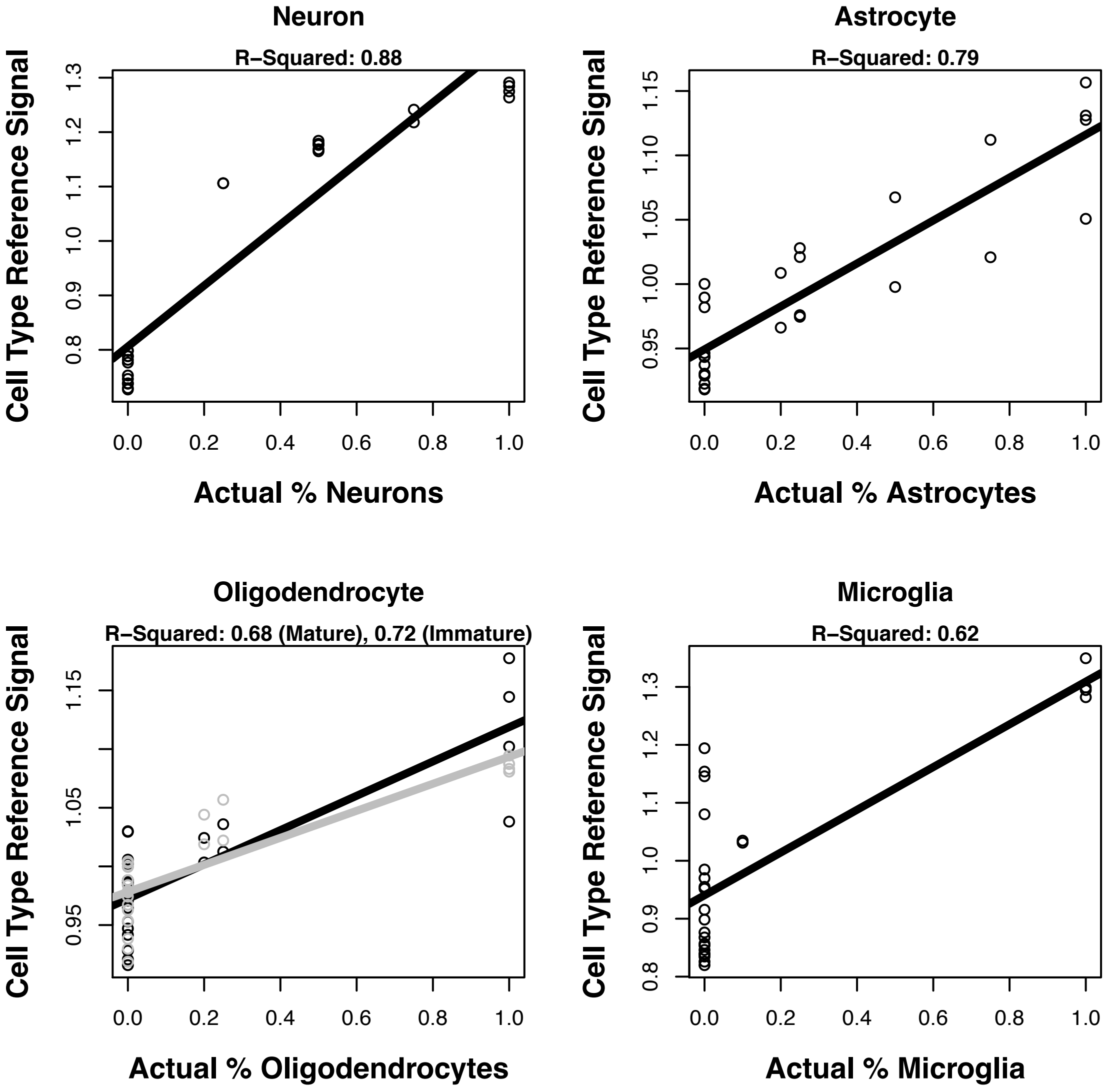

PSEA:
Microglia Reference Signal
0
5
10
15
Our Cell Type Index:
Microglia Index
0.0
0.5
1.0
1.5
Our Cell Type Index: After First Averaging by Publication
Microglia Index
0.0
0.5
1.0
1.5
2.0
astrocytes
endothelial
fetal_quiescent
fetal_replicating
hybrid
microglia
neurons
oligodendrocytes
OPC


PSEA:
Oligodendrocyte Reference Signal
0
1
2
3
4
5
6
7
Our Cell Type Index:
Oligodendrocyte Index
-0.5
0.0
0.5
1.0
Our Cell Type Index: After First Averaging by Publication
Oligodendrocyte Index
0.0
0.5
1.0
astrocytes
endothelial
fetal_quiescent
fetal_replicating
hybrid
microglia
neurons
oligodendrocytes
OPC



Suppl. Figure 8. The method for deriving predicted relative cell content determines the strength of the relationship with sample cell type. *Depicted below is a comparison of the efficacy of three different manners of predicting the relative cell content of samples (columns) in a human single-cell RNA-seq dataset (2)**:** 1) the "population reference signal" generated by PSEA, 2) a simplified version of our method that is meant to be relatively analogous to PSEA (a simple average of the z-score-transformed data for all genes specific to a particular cell type in our database), 3) the version of our method used in this manuscript, which consolidates the predictions derived from the cell type specific genes identified in different publications. Boxplots illustrate the distribution of the relative cell content predictions across samples identified as different cell types in the original publication (2).* Lavender: *astrocytes,* orange: *endothelial cells,* green: *microglia,* pink: *oligodendrocytes,* purple: *neurons. Note the presence of several extreme outliers (red) in the predictions produced by PSEA.*

Using similar methodology, we also calculated the population "reference signal" with PSEA for microarray data from artificially-created mixtures of cultured cells (GSE19380 – see discussion of data preprocessing in **Section 7.1**). The results strongly tracked the actual cell content of the mixed samples (**Suppl. Figure 9**) in a manner that was not strikingly better or worse than the predictions made using BrainInABlender for the same dataset (**Figure 3**). This again drives home the fact that the ratio-based normalization methods used in PSEA are particularly incompatible with low count data in RNA-Seq – results derived from microarray data are fine.



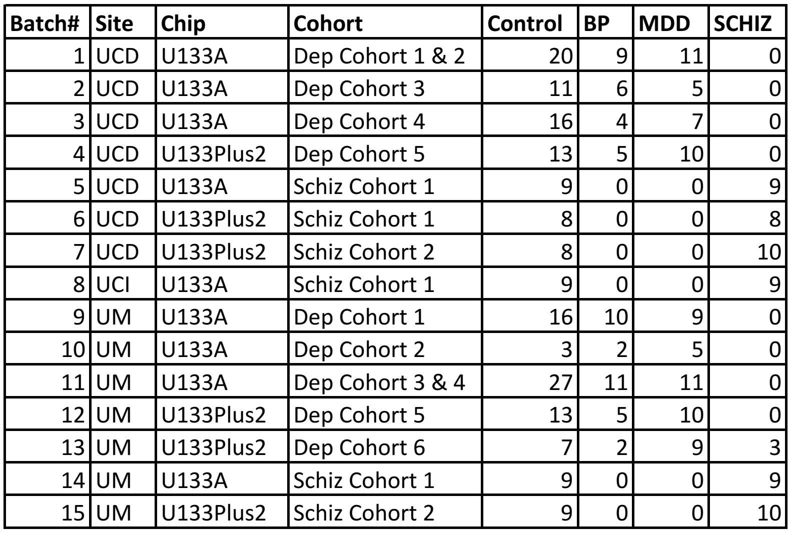
Suppl. Figure 9. Relative cell content predictions made using PSEA and our cell type specific gene lists. *Using a microarray dataset derived from samples that contained artificially-generated mixtures of cultured cells (GSE19380; (12)), we found that the relative cell content predictions ("cell type reference signal") produced by PSEA closely reflected actual known content, similar to the predictions made by BrainInABlender (**Figure 3**).*

7.3 Additional Detailed Preprocessing Methods for the Macro-Dissected Microarray Datasets

7.3.1 Pritzker Dorsolateral Prefrontal Cortex Microarray Dataset (GSE92538)

The original dataset included tissue from 172 high-quality human post-mortem brains donated to the Brain Donor Program at the University of California, Irvine with the consent of the next of kin. Frozen coronal slabs were macro-dissected to obtain dorsolateral prefrontal cortex samples. Clinical information was obtained from medical examiners, coroners' medical records, and a family member. Patients were diagnosed with either Major Depressive Disorder, Bipolar Disorder, or Schizophrenia by consensus based on criteria from the Diagnostic and Statistical Manual of Mental Disorders (93). Due to the extended nature of this study, this sample collection occurred in waves ("cohorts") over a period of many years. This research was overseen and approved by the University of Michigan Institutional Review Board (IRB # HUM00043530, Pritzker Neuropsychiatric Disorders Research Consortium (2001-0826)) and the University of California Irvine (UCI) Institutional Review Board (IRB# 1997-74).

As described previously (32,35), total RNA from these samples was then distributed to laboratories at three different institutions (University of Michigan (UM), University of California-Davis (UCD), University of California-Irvine (UCI)) to be hybridized to either Affymetrix HT-U133A or HT-U133Plus-v2 chips (1-5 replicates per sample, n=367). Before conducting the current analysis, the subset of probes found on both the Affymetrix HT-U133A and HT-U133Plus-v2 chips was extracted, reannotated for probe-to-transcript correspondance (94), summarized using robust multi-array analysis (RMA) (34), log (base 2)-transformed, quantile normalized, and gender-checked. Then, 15 batches of highly-correlated samples were identified that were defined a combination of cohort, chip, and laboratory (**Suppl. Figure 10**).



Batch#	Site	Chip	Cohort	Control	BP	MDD	SCHIZ
1	UCD	U133A	Dep Cohort 1 & 2	20	9	11	0
2	UCD	U133A	Dep Cohort 3	11	6	5	0
3	UCD	U133A	Dep Cohort 4	16	4	7	0
4	UCD	U133Plus2	Dep Cohort 5	13	5	10	0
5	UCD	U133A	Schiz Cohort 1	9	0	0	9
6	UCD	U133Plus2	Schiz Cohort 1	8	0	0	8
7	UCD	U133Plus2	Schiz Cohort 2	8	0	0	10
8	UCI	U133A	Schiz Cohort 1	9	0	0	9
9	UM	U133A	Dep Cohort 1	16	10	9	0
10	UM	U133A	Dep Cohort 2	3	2	5	0
11	UM	U133A	Dep Cohort 3 & 4	27	11	11	0
12	UM	U133Plus2	Dep Cohort 5	13	5	10	0
13	UM	U133Plus2	Dep Cohort 6	7	2	9	3
14	UM	U133A	Schiz Cohort 1	9	0	0	9
15	UM	U133Plus2	Schiz Cohort 2	9	0	0	10

Suppl. Figure 10. The number of microarray chips run in each batch, defined by processing site, Affymetrix chip type, and sample collection cohort. *Samples from the four diagnostic categories (Control, Bipolar Disorder, Major Depressive Disorder, Schizophrenia) were unevenly distributed across batches.*

Samples that exhibited markedly low average sample-sample correlation coefficients (<0.85: outliers) were removed from the dataset, including data from one batch that exhibited overall low sample-sample correlation coefficients with other batches and was a poor match with their duplicate microarrays run in a separate laboratory. The batch effects were then subtracted out using median-centering (detailed procedure: (35)) and the replicate samples were averaged for each subject. Our current analyses began with this sample-level summary gene expression data (publicly available in the Gene Expression Omnibus, GEO: GSE92538). We further removed data from any subjects lacking information regarding critical pre- or post-mortem variables necessary for our analysis, leaving a final sample size of n=157. All of the R script documenting these analyses can be found at https://github.com/hagenaue/CellTypeAnalyses_PritzkerAffyDLPFC.

7.3.2 Allen Brain Atlas Cross-Regional Microarray Dataset

The Allen Brain Atlas microarray data was downloaded from http://human.brain-map.org/microarray/search on December 2015. This microarray survey was performed in brain-specific batches, with multiple batches per subject. To remove technical variation across batches, a variety of normalization procedures had been performed by the original authors both within and across batches using internal controls, as well as across subjects (28). The dataset available for download had already been log-transformed (base 2) and converted to z-scores using the average and standard deviation for each probe. These normalization procedures were designed to remove technical artifacts while best preserving cross-regional effects in the data, but the full information about relative levels of expression within an individual sample were unavailable and the effects of subject-level variables (such as age and pH) were likely to be de-emphasized due to the inability to fully separate out subject and batch during the normalization process.

Prior to conducting other analyses, we averaged the expression level of the multiple probes that corresponded to the same gene, and re-scaled, so that the data associated with each gene symbol continued to be a z-score (mean=0, sd=1). The 30,000 probes mapped onto 18,787 unique genes (as determined by gene symbol). We then extracted the z-score data for the list of cell type specific genes derived from each publication (1608 total). Then, based on our results from analyzing the Pritzker dataset, we excluded the data for genes that were non-specific (i.e., included in a list of cell type specific genes from a different category of cells within any of the publications), and then averaged the data from the cell-type specific genes derived from each publication to predict the relative content of each of the 10 primary cell types in each sample. All of the R script documenting these analyses can be found at https://github.com/hagenaue/CellTypeAnalyses_AllenBrainAtlas.

7.3.3 Human Cortical Microarray Dataset GSE53987 (submitted to GEO by Lanz et al. (36))

The full publicly-available dataset GSE53987 (described in (36)) contained Affymetrix U133Plus2 microarray data from 205 post-mortem human brain samples from three brain regions: the DLPFC (Brodmann Area 46, focusing on gray matter only (Lanz T.A., *personal communication*)), the hippocampus, and the striatum. These samples were collected by the University of Pittsburgh brain bank. For the purposes of our current analysis, we only downloaded the microarray .CEL files for the dorsolateral prefrontal cortex samples. We summarized these data with robust multi-array analysis (RMA) (from the R package *affy* (34)) using a custom up-to-date chip definition file (.cdf) to define probe-to-transcript correspondence ("hgu133plus2hsentrezgcdf_19.0.0.tar.gz" from http://nmg-r.bioinformatics.nl/NuGO_R.html (94)). This process included background subtraction, log (base 2)-transformation, and quantile normalization. Gene Symbol annotation for probeset Entrez gene ids were provided by the R package *org.Hs.eg.db*. We extracted the sample characteristics from the GEO website using the R package *GEOquery.* To control for technical variation, the sample processing batches were estimated using the microarray chip scan dates extracted from the microarray .CEL files (using the function *protocolData* in the *GEOquery* package), but it appeared that all chips for the DLPFC were on the same date. RNA degradation was estimated using the R package AffyRNADegradation (39). During quality control, two samples were removed - GSM1304979 had a range of sample-sample correlations that was unusually low compared (median=0.978) compared to range for the dataset as a whole (median: 0.993) and GSM1304953 appeared to be falsely identified as female (signal for XIST<7). We then predicted the cell content of each sample from the microarray data using BrainInABlender. The code for all analyses can be found at:

https://github.com/hagenaue/CellTypeAnalyses_LanzHumanDLPFC/tree/master

7.3.4 Human Cortical Microarray Dataset GSE21138 (submitted to GEO by Narayan et al. (38))

The publicly-available dataset GSE21138 (described in (38))) contained Affymetrix U133Plus2 microarray data from 59 post-mortem human brain samples from the DLPFC (Brodmann Area 46, gray matter only (Thomas E.A., *personal communication*)) collected by the Mental Health Research Institute in Victoria, Australia. The procedures for data download and pre-processing were identical to those used above for GSE53987 with a few minor exceptions. In particular, there were six separate scan dates associated with the microarray .CEL files, but one of these scan dates was not included as a co-variate in our analyses because it had an n=1 ("06/14/06"). During quality control, the data for two subjects because they appeared to be falsely-identified as male (XIST>7, GSM528839 & GSM528840) , and one subject that appeared to be falsely-identified as female (XIST<7, GSM528880). Data for two more subjects were removed as outliers due to having an unsually low range of sample-sample correlations (GSM528866, GSM528873) as compared to the dataset as a whole. The code for all analyses can be found at: https://github.com/hagenaue/CellTypeAnalyses_NarayanHumanDLPFC.

7.3.5 Human Cortical Microarray Dataset GSE21935 (submitted to GEO by Barnes et al. (37))

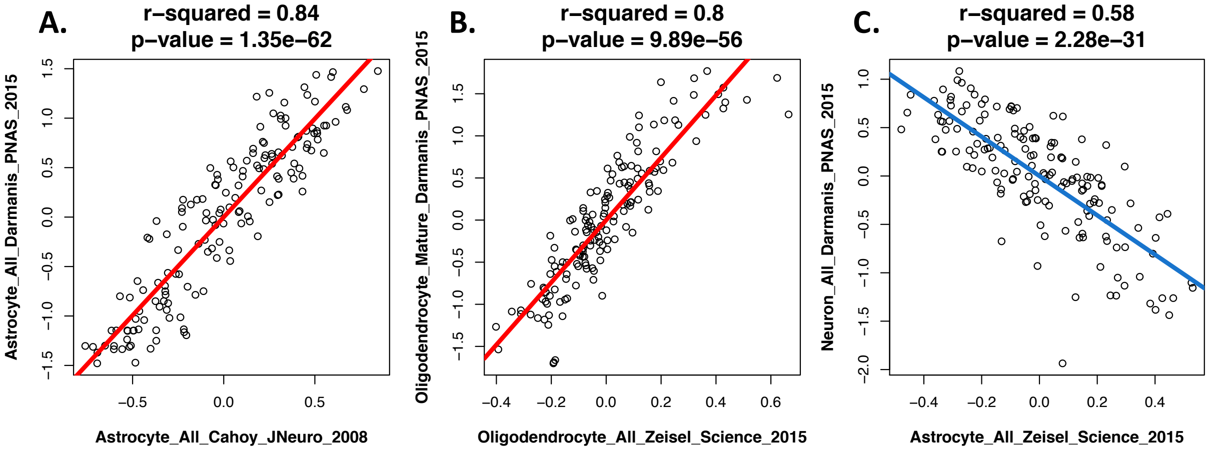
The publicly-available dataset GSE21935 (described in (37)) contained Affymetrix U133Plus2 microarray data from 42 post-mortem human brain samples from the temporal cortex (Brodmann Area 22) collected at the Charing Cross campus of the Imperial College of London. The procedures for data download and pre-processing were identical to those used above for GSE53987 with a few minor exceptions. In particular, there were two separate scan dates associated with the microarray .CEL files, but they were closely spaced (6/25/04 vs. 6/29/04) and we did not find any strong association between scan date and any of the top principal components of variation in the data, so we opted to not include scan date as a co-variate in our statistical models. Quality control did not identify any problematic samples. The code for all analyses can be found at: https://github.com/hagenaue/CellTypeAnalyses_BarnesHumanCortex/tree/master.

7.3.6 CommonMind Consortium Human Cortical RNA-Seq Dataset

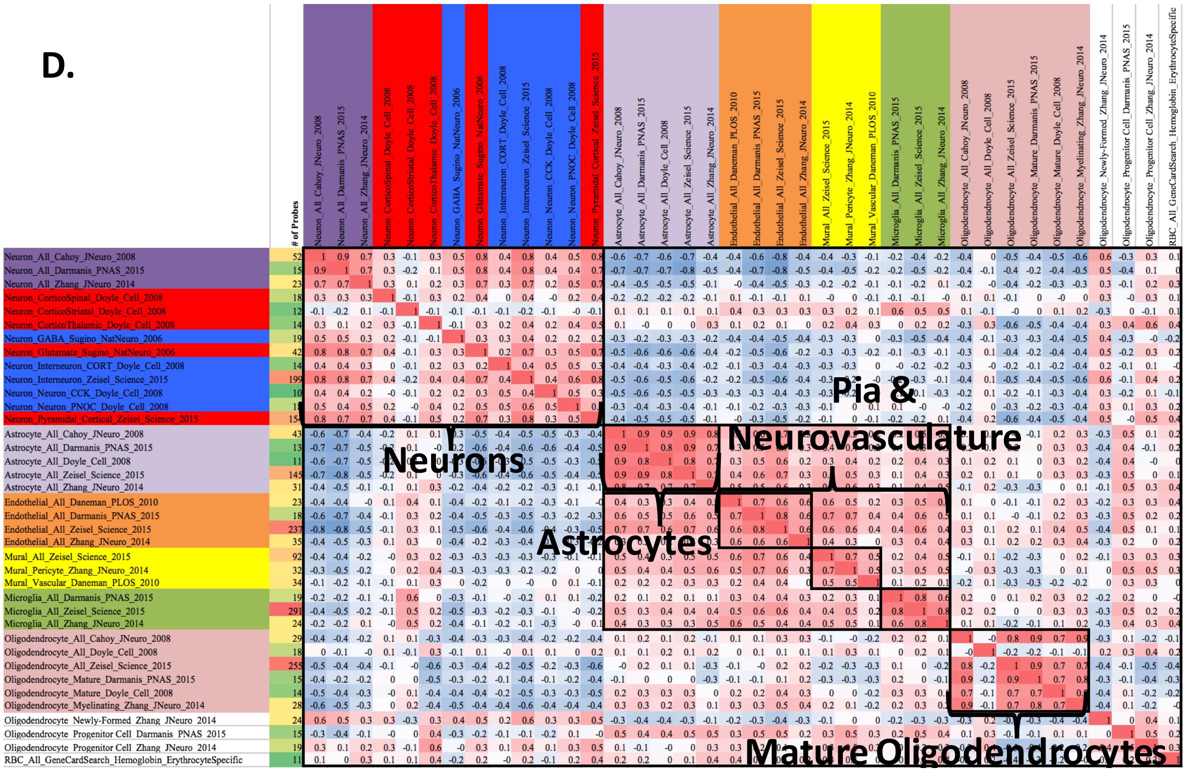
The CommonMind Consortium (CMC) RNA-seq dataset profiled prefrontal cortex samples from 603 individuals (40) collected at three brain banks: Mount Sinai School of Medicine, University of Pittsburgh, and University of Pennsylvania. This dataset was downloaded as GRCh37-aligned bam files from the CommonMind Consortium Knowledge Portal (https://www.synapse.org/CMC). Tophat-aligned bam files were converted back to fastq format and mapped to GRCh38 using HISAT2 (41) with default settings. Reads mapping uniquely to exons were then counted using subread featureCounts with ensembl transcript models. RNA-seq read counts were analyzed using limma/voom (42); cell type indices were calculated on logCPM values, and analysis of differential gene expression was performed using limma with observed precision weights in a weighted least squares linear regression. Prior to upload, poor quality samples from the original dataset (40) had already been removed (<50 million reads, RIN<5.5) and replaced with higher quality samples. We further excluded data from 10 replicates and 89 individuals with incomplete demographic data (missing pH), leaving a final sample size of 514 samples. The dataset was further filtered using an expression threshold (CPM>1 in at least 50 individuals) which reduced the dataset from including data from all annotated genes (about 60,000) to data from around 17.000 genes.

7.4 *Additional figures and results:* Does the Reference Dataset Matter? There is a Strong Convergence of Cell Content Predictions Derived from Cell Type Specific Transcripts Identified by Different Publications

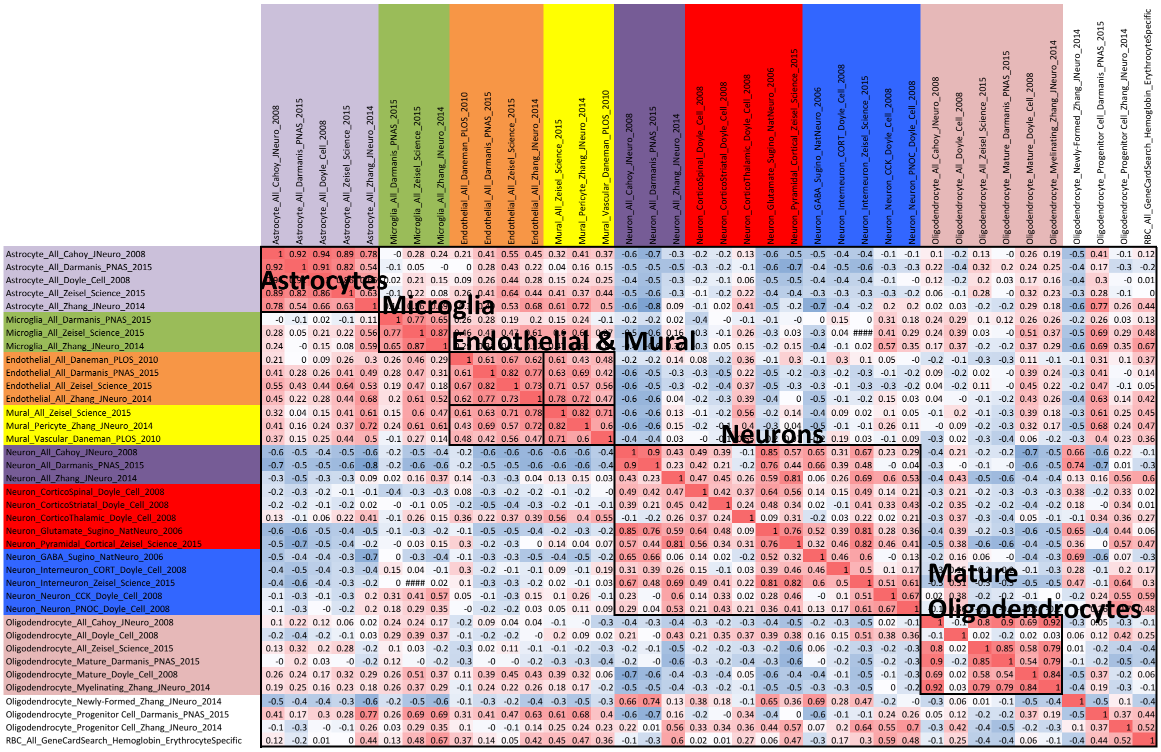
Similar to what we observed during our validation analyses using data from purified cell types, we found that the predicted cell content for our post-mortem human cortical samples (“cell type indices”) was similar regardless of the methodology used to generate the cell type specific gene lists used in the predictions. Within all four of the cortical microarray datasets, there was a strong positive correlation between cell type indices representing the same cell type, even when the predictions were derived using cell type specific gene lists from different species, cell type purification strategies, and platforms. In general, we found that the pattern of correlations between the 38 cell type indices clearly clustered within three large umbrella categories: neurons, oligodendrocytes, and support cells (astrocytes, microglia, and neurovasculature). This clustering was clear using visual inspection of the correlation matrices (**Suppl. Figure 11, Suppl. Figure 12**), hierarchical clustering, or consensus clustering (**Suppl. Figure 13**; ConsensusClusterPlus: (43)) and persisted even after removing data from genes identified as cell type specific in multiple publications (e.g., gene expression identified as astrocyte-expression in both Cahoy_Astrocyte and Zhang_Astrocyte; **Suppl. Figure 14, Suppl. Figure 16**). In some datasets, the cell type indices for support cell subcategories were also clearly clustered (**Suppl. Figure 11**). In contrast, clustering was not able to reliably discern neuronal subcategories (interneurons, projection neurons) in any dataset. Likewise, oligodendrocyte progenitor cell indices derived from different publications did not strongly correlate with each other, perhaps indicating a lack of significant presence of progenitor cells in the cortex of the primarily middle-aged subjects.



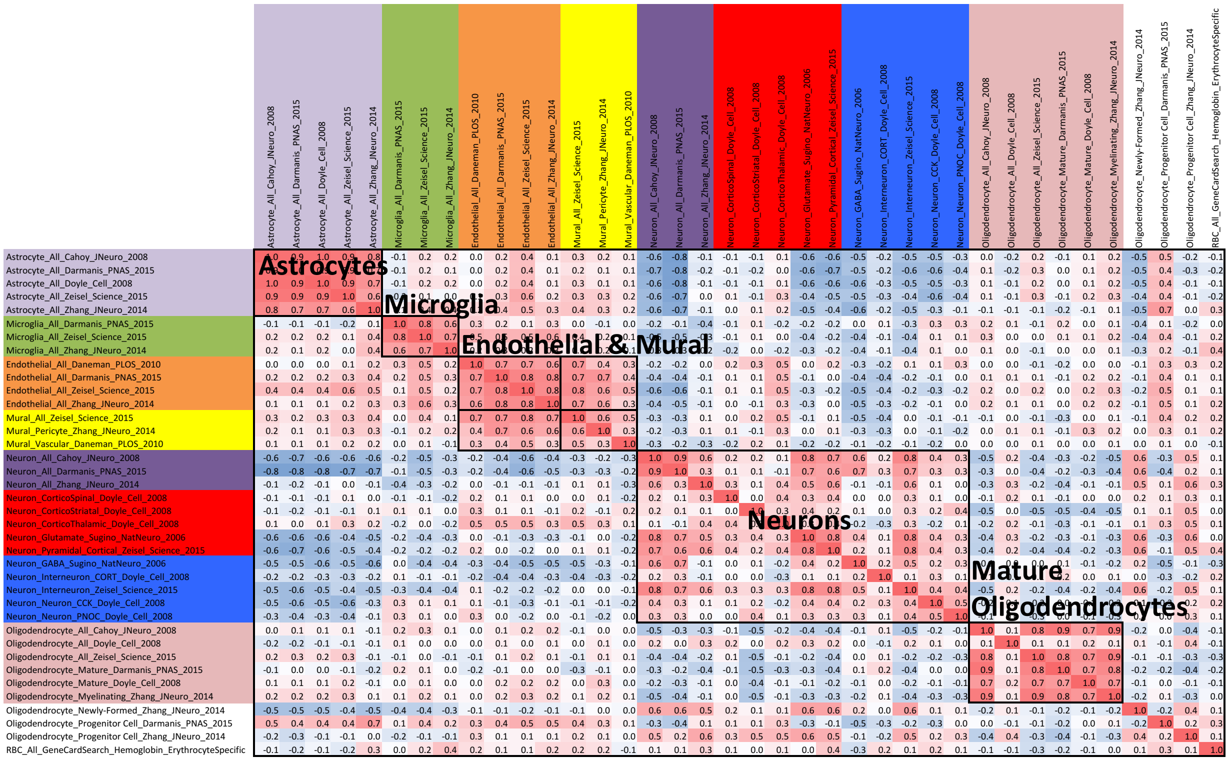
D.


# of Probes
Neurons
Astrocytes
Pia &
Neurovasculature
Mature Oligodendrocytes

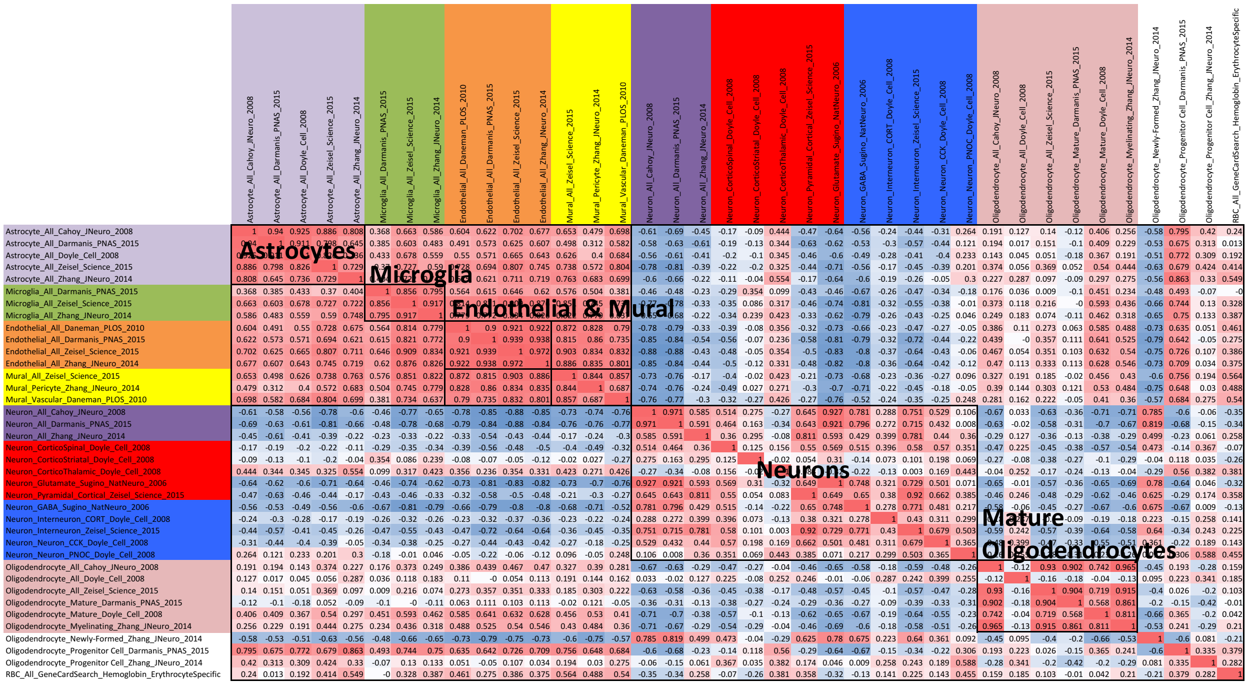
E.


Astrocytes
Microglia
Endothelial & Mural
Neurons
Mature
Oligodendrocytes

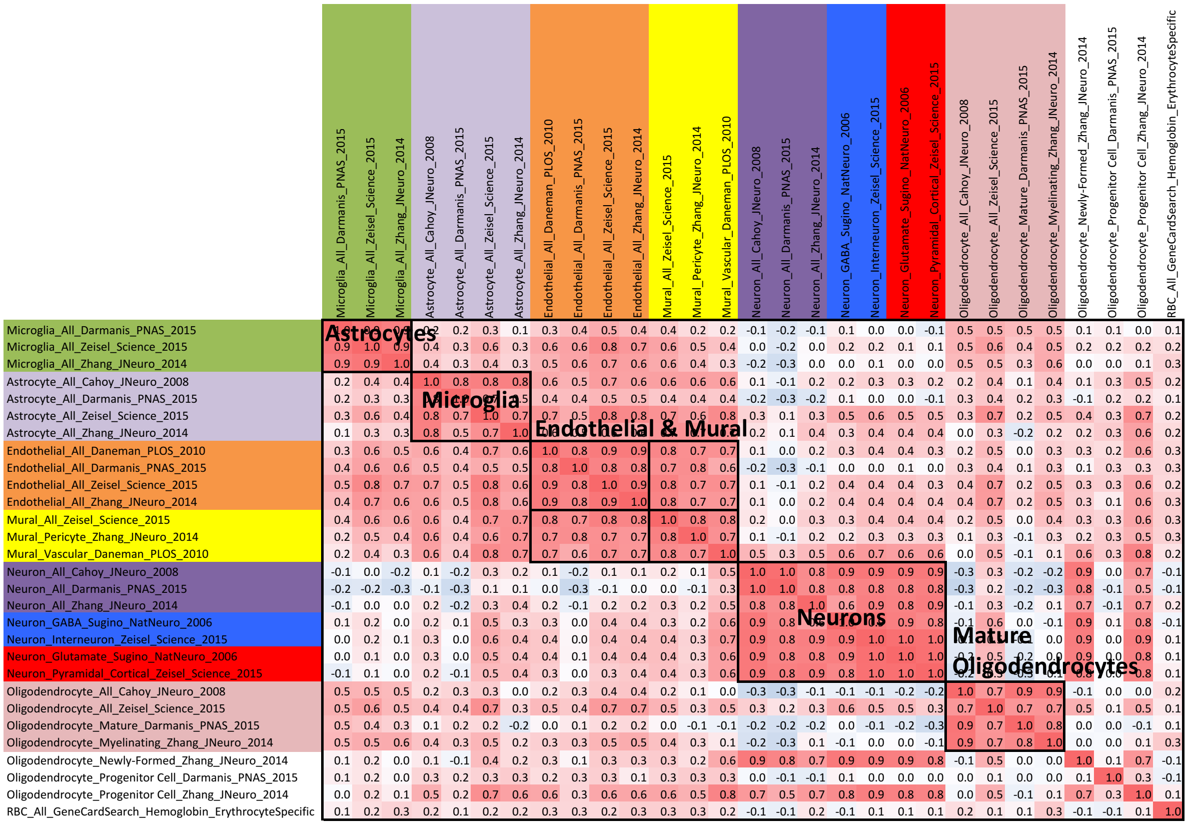
F.


Astrocytes
Microglia
Endothelial & Mural
Neurons
Mature
Oligodendrocytes
Astrocyte_All_Cahoy_JNeuro_2008
Astrocyte_All_Darmanis_PNAS_2015
Astrocyte_All_Doyle_Cell_2008
Astrocyte_All_Zeisel_Science_2015
Astrocyte_All_Zhang_JNeuro_2014
Microglia_All_Darmanis_PNAS_2015
Microglia_All_Zeisel_Science_2015
Microglia_All_Zhang_JNeuro_2014
Endothelial_All_Daneman_PLOS_2010
Endothelial_All_Darmanis_PNAS_2015
Endothelial_All_Zeisel_Science_2015
Endothelial_All_Zhang_JNeuro_2014
Mural_All_Zeisel_Science_2015
Mural_Pericyte_Zhang_JNeuro_2014
Mural_Vascular_Daneman_PLOS_2010
Neuron_All_Cahoy_JNeuro_2008
Neuron_All_Darmanis_PNAS_2015
Neuron_All_Zhang_JNeuro_2014
Neuron_CorticoSpinal_Doyle_Cell_2008
Neuron_CorticoStriatal_Doyle_Cell_2008
Neuron_CorticoThalamic_Doyle_Cell_2008
Neuron_Glutamate_Sugino_NatNeuro_2006
Neuron_Pyramidal_Cortical_Zeisel_Science_2015
Neuron_GABA_Sugino_NatNeuro_2006
Neuron_Interneuron_CORT_Doyle_Cell_2008
Neuron_Interneuron_Zeisel_Science_2015
Neuron_Neuron_CCK_Doyle_Cell_2008
Neuron_Neuron_PNOC_Doyle_Cell_2008
Oligodendrocyte_All_Cahoy_JNeuro_2008
Oligodendrocyte_All_Doyle_Cell_2008
Oligodendrocyte_All_Zeisel_Science_2015
Oligodendrocyte_Mature_Darmanis_PNAS_2015
Oligodendrocyte_Mature_Doyle_Cell_2008
Oligodendrocyte_Myelinating_Zhang_JNeuro_2014
Oligodendrocyte_Newly-Formed_Zhang_JNeuro_2014
Oligodendrocyte_Progenitor Cell_Darmanis_PNAS_2015
Oligodendrocyte_Progenitor Cell_Zhang_JNeuro_2014
RBC_All_GeneCardSearch_Hemoglobin_ErythrocyteSpecific

G.

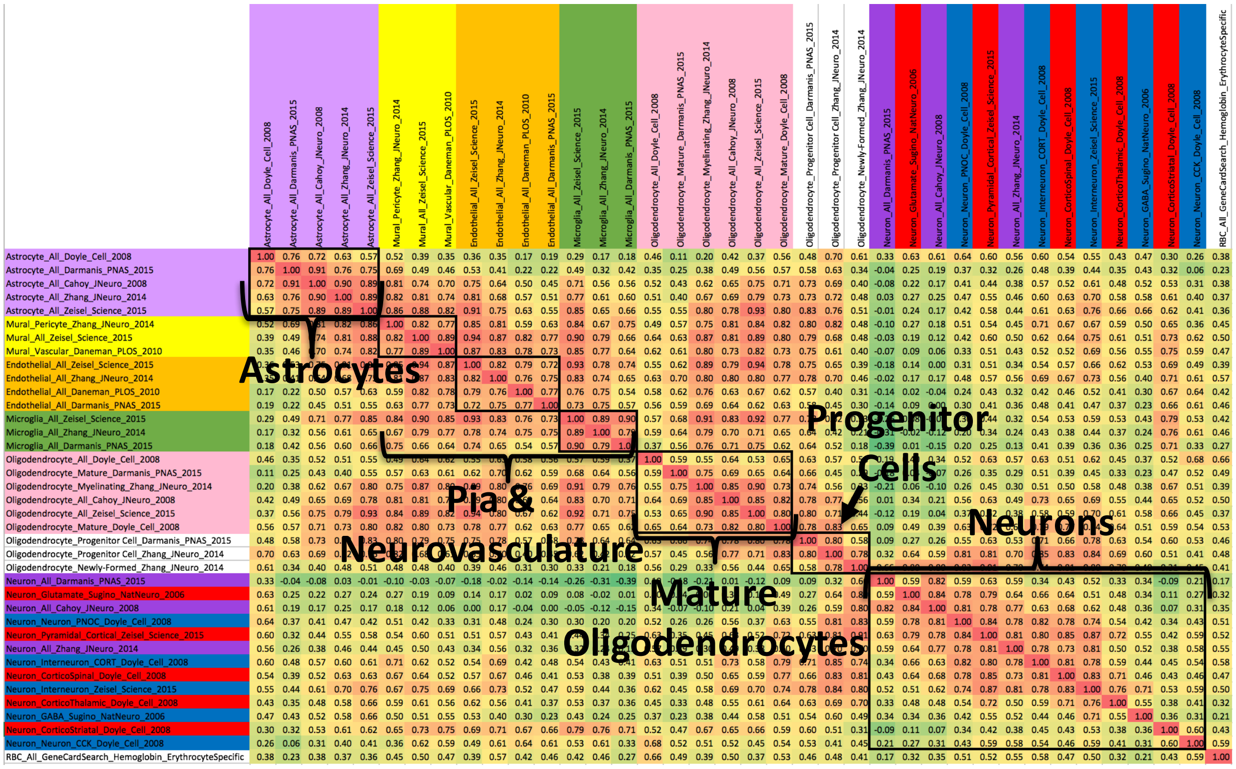

Astrocytes
Microglia
Endothelial & Mural
Neurons
Mature
Oligodendrocytes
Astrocyte_All_Cahoy_JNeuro_2008
Astrocyte_All_Darmanis_PNAS_2015
Astrocyte_All_Doyle_Cell_2008
Astrocyte_All_Zeisel_Science_2015
Astrocyte_All_Zhang_JNeuro_2014
Microglia_All_Darmanis_PNAS_2015
Microglia_All_Zeisel_Science_2015
Microglia_All_Zhang_JNeuro_2014
Endothelial_All_Daneman_PLOS_2010
Endothelial_All_Darmanis_PNAS_2015
Endothelial_All_Zeisel_Science_2015
Endothelial_All_Zhang_JNeuro_2014
Mural_All_Zeisel_Science_2015
Mural_Pericyte_Zhang_JNeuro_2014
Mural_Vascular_Daneman_PLOS_2010
Neuron_All_Cahoy_JNeuro_2008
Neuron_All_Darmanis_PNAS_2015
Neuron_All_Zhang_JNeuro_2014
Neuron_CorticoSpinal_Doyle_Cell_2008
Neuron_CorticoStriatal_Doyle_Cell_2008
Neuron_CorticoThalamic_Doyle_Cell_2008
Neuron_Glutamate_Sugino_NatNeuro_2006
Neuron_Pyramidal_Cortical_Zeisel_Science_2015
Neuron_GABA_Sugino_NatNeuro_2006
Neuron_Interneuron_CORT_Doyle_Cell_2008
Neuron_Interneuron_Zeisel_Science_2015
Neuron_Neuron_CCK_Doyle_Cell_2008
Neuron_Neuron_PNOC_Doyle_Cell_2008
Oligodendrocyte_All_Cahoy_JNeuro_2008
Oligodendrocyte_All_Doyle_Cell_2008
Oligodendrocyte_All_Zeisel_Science_2015
Oligodendrocyte_Mature_Darmanis_PNAS_2015
Oligodendrocyte_Mature_Doyle_Cell_2008
Oligodendrocyte_Myelinating_Zhang_JNeuro_2014
Oligodendrocyte_Newly-Formed_Zhang_JNeuro_2014
Oligodendrocyte_Progenitor Cell_Darmanis_PNAS_2015
Oligodendrocyte_Progenitor Cell_Zhang_JNeuro_2014
RBC_All_GeneCardSearch_Hemoglobin_ErythrocyteSpecific

H.

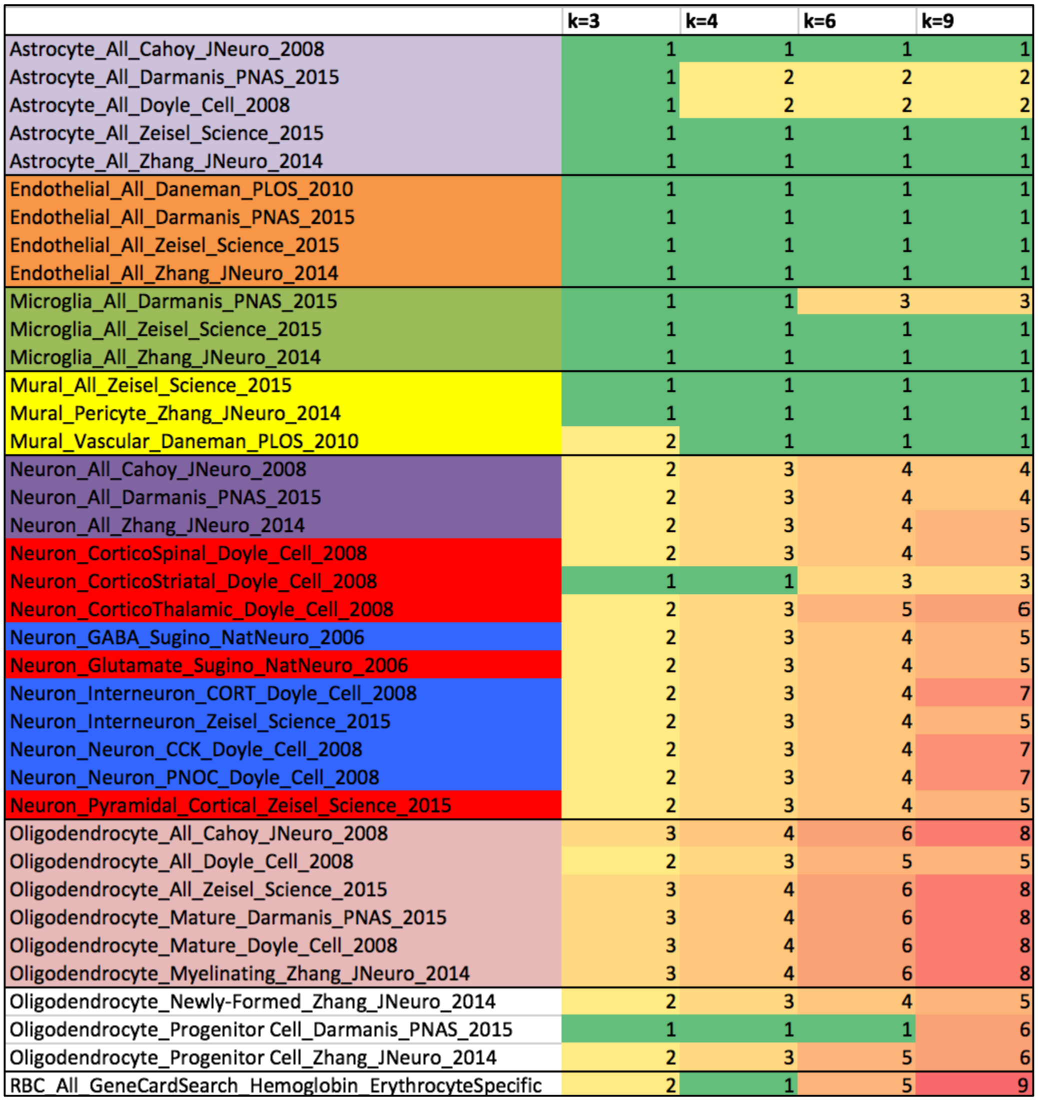


	Microglia_All_Darmanis_PNAS_2015	Microglia_All_Zeisel_Science_2015	Microglia_All_Zhang_JNeuro_2014	Astrocyte_All_Cahoy_JNeuro_2008	Astrocyte_All_Darmanis_PNAS_2015	Astrocyte_All_Zeisel_Science_2015	Astrocyte_All_Zhang_JNeuro_2014	Endothelial_All_Daneman_PLOS_2010	Endothelial_All_Darmanis_PNAS_2015	Endothelial_All_Zeisel_Science_2015	Endothelial_All_Zhang_JNeuro_2014	Mural_All_Zeisel_Science_2015	Mural_Pericyte_Zhang_JNeuro_2014	Mural_Vascular_Daneman_PLOS_2010	Neuron_All_Cahoy_JNeuro_2008	Neuron_All_Darmanis_PNAS_2015	Neuron_All_Zhang_JNeuro_2014	Neuron_GABA_Sugino_NatNeuro_2006	Neuron_Interneuron_Zeisel_Science_2015	Neuron_Glutamate_Sugino_NatNeuro_2006	Neuron_Pyramidal_Cortical_Zeisel_Science_2015	Oligodendrocyte_All_Cahoy_JNeuro_2008	Oligodendrocyte_All_Zeisel_Science_2015	Oligodendrocyte_Mature_Darmanis_PNAS_2015	Oligodendrocyte_Myelinating_Zhang_JNeuro_2014	Oligodendrocyte_Newly-Formed_Zhang_JNeuro_2014	Oligodendrocyte_Progenitor Cell_Darmanis_PNAS_2015	Oligodendrocyte_Progenitor Cell_Zhang_JNeuro_2014	RBC_All_GeneCardSearch_Hemoglobin_ErythrocyteSpecific
Microglia_All_Darmanis_PNAS_2015	[illegible]	[illegible]	[illegible]	0.2	0.2	0.3	0.1	0.3	0.4	0.5	0.4	0.4	0.2	0.2	-0.1	-0.2	-0.1	0.1	0.0	0.0	-0.1	0.5	0.5	0.5	0.5	0.1	0.1	0.0	0.1
Microglia_All_Zeisel_Science_2015	0.9	1.0	0.9	0.4	0.3	0.6	0.3	0.6	0.6	0.8	0.7	0.6	0.5	0.4	0.0	-0.2	0.0	0.2	0.2	0.1	0.1	0.5	0.6	0.4	0.5	0.2	0.2	0.2	0.2
Microglia_All_Zhang_JNeuro_2014	0.9	0.9	1.0	0.4	0.3	0.4	0.3	0.5	0.6	0.7	0.6	0.6	0.4	0.3	-0.2	-0.3	0.0	0.0	0.1	0.0	0.0	0.5	0.5	0.3	0.6	0.0	0.0	0.1	0.3
Astrocyte_All_Cahoy_JNeuro_2008	0.2	0.4	0.4	1.0	0.8	0.8	0.8	0.6	0.5	0.7	0.6	0.6	0.6	0.6	0.1	-0.1	0.2	0.2	0.3	0.3	0.2	0.2	0.4	0.1	0.4	0.1	0.3	0.5	0.2
Astrocyte_All_Darmanis_PNAS_2015	0.2	0.3	0.3	[illegible]	[illegible]	[illegible]	0.5	0.4	0.4	0.5	0.5	0.4	0.4	0.4	-0.2	-0.3	-0.2	0.1	0.0	0.0	-0.1	0.3	0.4	0.2	0.3	-0.1	0.2	0.2	0.1
Astrocyte_All_Zeisel_Science_2015	0.3	0.6	0.4	0.8	0.7	1.0	0.7	0.7	0.5	0.8	0.8	0.7	0.6	0.8	0.3	0.1	0.3	0.5	0.6	0.5	0.5	0.3	0.7	0.2	0.5	0.4	0.3	0.7	0.2
Astrocyte_All_Zhang_JNeuro_2014	0.1	0.3	0.3	0.8	0.5	0.7	1.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.2	0.1	0.4	0.3	0.4	0.4	0.4	0.0	0.3	-0.2	0.2	0.2	0.3	0.6	0.2
Endothelial_All_Daneman_PLOS_2010	0.3	0.6	0.5	0.6	0.4	0.7	0.6	1.0	0.8	0.9	0.9	0.8	0.7	0.7	0.1	0.0	0.2	0.3	0.4	0.4	0.3	0.2	0.5	0.0	0.3	0.3	0.2	0.6	0.3
Endothelial_All_Darmanis_PNAS_2015	0.4	0.6	0.6	0.5	0.4	0.5	0.5	0.8	1.0	0.8	0.8	0.7	0.8	0.6	-0.2	-0.3	-0.1	0.0	0.0	0.1	0.0	0.3	0.4	0.1	0.3	0.1	0.3	0.3	0.3
Endothelial_All_Zeisel_Science_2015	0.5	0.8	0.7	0.7	0.5	0.8	0.6	0.9	0.8	1.0	0.9	0.8	0.7	0.7	0.1	-0.1	0.2	0.4	0.4	0.4	0.3	0.4	0.7	0.2	0.5	0.3	0.3	0.6	0.2
Endothelial_All_Zhang_JNeuro_2014	0.4	0.7	0.6	0.6	0.5	0.8	0.6	0.9	0.8	0.9	1.0	0.8	0.7	0.7	0.1	0.0	0.2	0.4	0.4	0.4	0.4	0.4	0.7	0.2	0.5	0.3	0.1	0.6	0.3
Mural_All_Zeisel_Science_2015	0.4	0.6	0.6	0.6	0.4	0.7	0.7	0.8	0.7	0.8	0.8	1.0	0.8	0.8	0.2	0.0	0.3	0.3	0.4	0.4	0.4	0.2	0.5	0.0	0.4	0.3	0.3	0.6	0.3
Mural_Pericyte_Zhang_JNeuro_2014	0.2	0.5	0.4	0.6	0.4	0.6	0.7	0.7	0.8	0.7	0.7	0.8	1.0	0.7	0.1	-0.1	0.2	0.2	0.3	0.3	0.3	0.1	0.3	-0.1	0.3	0.2	0.3	0.5	0.3
Mural_Vascular_Daneman_PLOS_2010	0.2	0.4	0.3	0.6	0.4	0.8	0.7	0.7	0.6	0.7	0.7	0.8	0.7	1.0	0.5	0.3	0.5	0.6	0.7	0.6	0.6	0.0	0.5	-0.1	0.1	0.6	0.3	0.8	0.2
Neuron_All_Cahoy_JNeuro_2008	-0.1	0.0	-0.2	0.1	-0.2	0.3	0.2	0.1	-0.2	0.1	0.1	0.2	0.1	0.5	1.0	1.0	0.8	0.9	0.9	0.9	0.9	-0.3	0.3	-0.2	-0.2	0.9	0.0	0.7	-0.1
Neuron_All_Darmanis_PNAS_2015	-0.2	-0.2	-0.3	-0.1	-0.3	0.1	0.1	0.0	-0.3	-0.1	0.0	0.0	-0.1	0.3	1.0	1.0	0.8	0.8	0.8	0.8	0.8	-0.3	0.2	-0.2	-0.3	0.8	-0.1	0.5	-0.1
Neuron_All_Zhang_JNeuro_2014	-0.1	0.0	0.0	0.2	-0.2	0.3	0.4	0.2	-0.1	0.2	0.2	0.3	0.2	0.5	0.8	0.8	1.0	0.6	0.9	0.8	0.9	-0.1	0.3	-0.2	0.1	0.7	-0.1	0.7	0.2
Neuron_GABA_Sugino_NatNeuro_2006	0.1	0.2	0.0	0.2	0.1	0.5	0.3	0.3	0.0	0.4	0.4	0.3	0.2	0.6	0.9	0.8	[illegible]	[illegible]	[illegible]	0.9	0.8	-0.1	0.6	0.0	-0.1	0.9	0.1	0.8	-0.1
Neuron_Interneuron_Zeisel_Science_2015	0.0	0.2	0.1	0.3	0.0	0.6	0.4	0.4	0.0	0.4	0.4	0.4	0.3	0.7	0.9	0.8	0.9	0.9	1.0	1.0	1.0	[illegible]	[illegible]	[illegible]	0.0	0.9	0.0	0.9	0.1
Neuron_Glutamate_Sugino_NatNeuro_2006	0.0	0.1	0.0	0.3	0.0	0.5	0.4	0.4	0.1	0.4	0.4	0.4	0.3	0.6	0.9	0.8	0.8	0.9	1.0	1.0	1.0	-0.2	0.5	-0.2	0.0	0.9	0.0	0.8	0.1
Neuron_Pyramidal_Cortical_Zeisel_Science_2015	-0.1	0.1	0.0	0.2	-0.1	0.5	0.4	0.3	0.0	0.3	0.4	0.4	0.3	0.6	0.9	0.8	0.9	0.8	1.0	1.0	1.0	-0.2	[illegible]	[illegible]	0.1	0.8	0.0	0.8	0.1
Oligodendrocyte_All_Cahoy_JNeuro_2008	0.5	0.5	0.5	0.2	0.3	0.3	0.0	0.2	0.3	0.4	0.4	0.2	0.1	0.0	-0.3	-0.3	-0.1	-0.1	-0.1	-0.2	-0.2	1.0	0.7	0.9	0.9	-0.1	0.0	0.0	0.2
Oligodendrocyte_All_Zeisel_Science_2015	0.5	0.6	0.5	0.4	0.4	0.7	0.3	0.5	0.4	0.7	0.7	0.5	0.3	0.5	0.3	0.2	0.3	0.6	0.5	0.5	0.3	0.7	1.0	0.7	0.7	0.5	0.1	0.5	0.1
Oligodendrocyte_Mature_Darmanis_PNAS_2015	0.5	0.4	0.3	0.1	0.2	0.2	-0.2	0.0	0.1	0.2	0.2	0.0	-0.1	-0.1	-0.2	-0.2	-0.2	0.0	-0.1	-0.2	-0.3	0.9	0.7	1.0	0.8	0.0	0.0	-0.1	0.1
Oligodendrocyte_Myelinating_Zhang_JNeuro_2014	0.5	0.5	0.6	0.4	0.3	0.5	0.2	0.3	0.3	0.5	0.5	0.4	0.3	0.1	-0.2	-0.3	0.1	-0.1	0.0	0.0	-0.1	0.9	0.7	0.8	1.0	0.0	0.0	0.1	0.3
Oligodendrocyte_Newly-Formed_Zhang_JNeuro_2014	0.1	0.2	0.0	0.1	-0.1	0.4	0.2	0.3	0.1	0.3	0.3	0.3	0.2	0.6	0.9	0.8	0.7	0.9	0.9	0.9	0.8	-0.1	0.5	0.0	0.0	1.0	0.1	0.7	-0.1
Oligodendrocyte_Progenitor Cell_Darmanis_PNAS_2015	0.1	0.2	0.0	0.3	0.2	0.3	0.3	0.2	0.3	0.3	0.1	0.3	0.3	0.3	0.0	-0.1	-0.1	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	1.0	0.3	-0.1
Oligodendrocyte_Progenitor Cell_Zhang_JNeuro_2014	0.0	0.2	0.1	0.5	0.2	0.7	0.6	0.6	0.3	0.6	0.6	0.6	0.5	0.8	0.7	0.5	0.7	0.8	0.9	0.8	0.8	0.0	0.5	-0.1	0.1	0.7	0.3	1.0	0.1
RBC_All_GeneCardSearch_Hemoglobin_ErythrocyteSpecific	0.1	0.2	0.3	0.2	0.1	0.2	0.2	0.3	0.3	0.2	0.3	0.3	0.3	0.2	-0.1	-0.1	0.2	-0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.3	-0.1	-0.1	0.1	1.0

Suppl. Figure 11. There is a convergence of cell content predictions derived from cell type specific transcripts identified by different publications. ***A-B.*** *Predictions of the relative cell content of our human cortical samples ("cell type indices") for any particular cell type were strongly correlated, even when the predictions were based on cell type specific transcripts identified by experiments using very different methodology. The examples given above include predictions based on cell type specific transcripts originally identified in mouse (x-axis) vs. human (y-axis) tissue.* ***C.*** *In contrast, there was a strong negative correlation between the predictions for dissimilar cell types, such as neurons and astrocytes.* ***D.*** *The similarity of different cell type indices in the Pritzker cortical dataset can be visualized using a correlation matrix. Within this matrix, correlations can range from a strong negative correlation of -1 (blue) to a strong positive correlation of 1 (red), therefore a large block of pink/red correlations is indicative of cell type indices that tend to be enriched in the same samples. The axis labels for cell type indices representing the same category of cell are color-coded: general neuronal categories are dark purple, pyramidal neurons are red, inhibitory interneurons are dark blue, astrocytes are light purple, endothelial cells are orange, mural cells are yellow, microglia are green, mature oligodendrocytes are pink, and the remaining indices remain white to represent lack of coherent categorization. The number of probes included in each index is present in the far left column (also color-coded, with green indicating few probes and red indicating many probes).* ***E-H.*** *The cell type index correlation matrices for the replication cortical datasets:* ***E.*** *Narayan et al. (GSE21138),* ***F***. *Lanz et al. (GSE53987),* ***G.*** *Barnes et al. (GSE21935)* ***H.*** *CMC RNA-Seq*

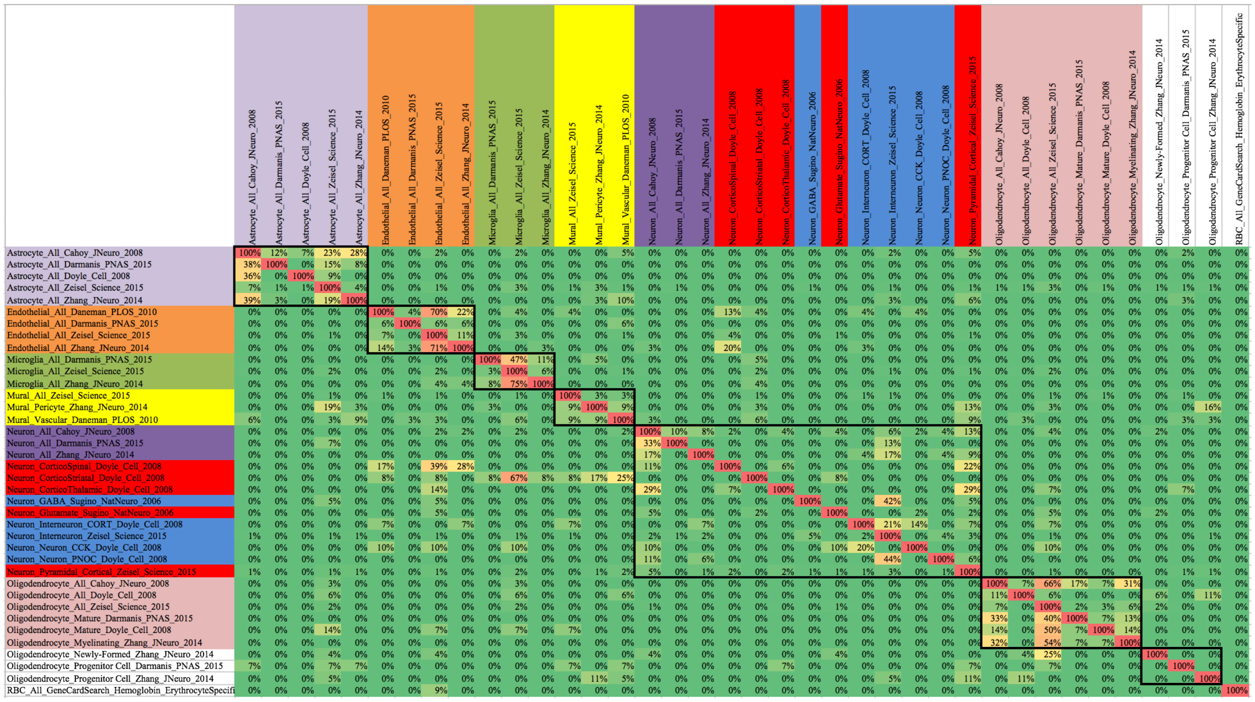


Suppl. Figure 12. The convergence of cell content predictions derived from cell type specific transcripts within the Allen Brain Atlas dataset. *Within the Allen Brain Atlas dataset, the main source of variation within the data (PC1) was negatively related to all cell types, with an especially strong relationship with mural cells (R^2=0.847), oligodendrocytes (R^2=0.808), and endothelial cells (R^2=0.775), suggesting that perhaps the main source of variation in the dataset (PC1) represented general tissue cell density instead of cell type balance per se. This causes the cell type indices for almost all cell types to be positively correlated (see below), and therefore this correlation matrix has a slightly different format than* ***Suppl. Figure 11****and* ***Suppl. Figure 16****. The cell types are still coded as in the previous figures, but the correlation coefficients (R) in the matrix are no longer color coded with blue indicating a negative correlation and red indicating a positive correlation. Instead, the green to red gradient indicates increasing percentile from most negative to most positive. The tightest correlations are red, with two obvious clusters: one cluster representing neurons and another representing glia and support cells. This differs from the Pritzker dorsolateral prefrontal cortex data, in which oligodendrocytes were found in their own cluster, perhaps due to a greater variation in the agonal conditions of the subjects (providing an impetus for the correlated upregulation of astrocytes and neurovasculature) or perhaps due to the spatial segregation of these cell types within the layered cortex (with an enrichment of vasculature and astrocytes at the surface of the cortex and an enrichment of white matter under the cortex). Also notable is the more coherent signature for progenitor cells within the Allen Brain Atlas dataset, perhaps due to the inclusion of tissue from neurogenic regions.*



	k=3	k=4	k=6	k=9
Astrocyte_All_Cahoy_JNeuro_2008	1	1	1	1
Astrocyte_All_Darmanis_PNAS_2015	1	2	2	2
Astrocyte_All_Doyle_Cell_2008	1	2	2	2
Astrocyte_All_Zeisel_Science_2015	1	1	1	1
Astrocyte_All_Zhang_JNeuro_2014	1	1	1	1
Endothelial_All_Daneman_PLOS_2010	1	1	1	1
Endothelial_All_Darmanis_PNAS_2015	1	1	1	1
Endothelial_All_Zeisel_Science_2015	1	1	1	1
Endothelial_All_Zhang_JNeuro_2014	1	1	1	1
Microglia_All_Darmanis_PNAS_2015	1	1	3	3
Microglia_All_Zeisel_Science_2015	1	1	1	1
Microglia_All_Zhang_JNeuro_2014	1	1	1	1
Mural_All_Zeisel_Science_2015	1	1	1	1
Mural_Pericyte_Zhang_JNeuro_2014	1	1	1	1
Mural_Vascular_Daneman_PLOS_2010	2	1	1	1
Neuron_All_Cahoy_JNeuro_2008	2	3	4	4
Neuron_All_Darmanis_PNAS_2015	2	3	4	4
Neuron_All_Zhang_JNeuro_2014	2	3	4	5
Neuron_CorticoSpinal_Doyle_Cell_2008	2	3	4	5
Neuron_CorticoStriatal_Doyle_Cell_2008	1	1	3	3
Neuron_CorticoThalamic_Doyle_Cell_2008	2	3	5	6
Neuron_GABA_Sugino_NatNeuro_2006	2	3	4	5
Neuron_Glutamate_Sugino_NatNeuro_2006	2	3	4	5
Neuron_Interneuron_CORT_Doyle_Cell_2008	2	3	4	7
Neuron_Interneuron_Zeisel_Science_2015	2	3	4	5
Neuron_Neuron_CCK_Doyle_Cell_2008	2	3	4	7
Neuron_Neuron_PNOC_Doyle_Cell_2008	2	3	4	7
Neuron_Pyramidal_Cortical_Zeisel_Science_2015	2	3	4	5
Oligodendrocyte_All_Cahoy_JNeuro_2008	3	4	6	8
Oligodendrocyte_All_Doyle_Cell_2008	2	3	5	5
Oligodendrocyte_All_Zeisel_Science_2015	3	4	6	8
Oligodendrocyte_Mature_Darmanis_PNAS_2015	3	4	6	8
Oligodendrocyte_Mature_Doyle_Cell_2008	3	4	6	8
Oligodendrocyte_Myelinating_Zhang_JNeuro_2014	3	4	6	8
Oligodendrocyte_Newly-Formed_Zhang_JNeuro_2014	2	3	4	5
Oligodendrocyte_Progenitor Cell_Darmanis_PNAS_2015	1	1	1	6
Oligodendrocyte_Progenitor Cell_Zhang_JNeuro_2014	2	3	5	6
RBC_All_GeneCardSearch_Hemoglobin_ErythrocyteSpecific	2	1	5	9

Suppl. Figure 13. Consensus clustering indicates the cell type indices clearly cluster into three large umbrella categories: neurons, oligodendrocytes, and support cells. *The cell type indices were developed using cell type specific genes identified by different publications, species, and methodologies, and are categorically color-coded in a manner similar to* ***Suppl. Figure 11.*** *Each column represents the numerical category (cluster) assigned to a cell type index in a k-means clustering algorithm with k number of clusters – for example, in the column "k=3", the algorithm sorted each of the cell type indices into 3 clusters based on similarity (defined by Euclidean distance). These 3 clusters are easily identifiable as neurons, oligodendrocytes, and support cells. Increasing the number of clusters (k) did not improve the ability of the algorithm to detect more specific neuronal subcategories (interneurons, projection neurons) or support cell subcategories (astrocytes, endothelial cells, mural cells, microglia), and the immature oligodendrocyte indices from different publications showed a notable lack of convergence. The consensus clustering was run using 50 bootstraps with a proportion of 0.8 item subsampling and 1.0 feature subsampling.*

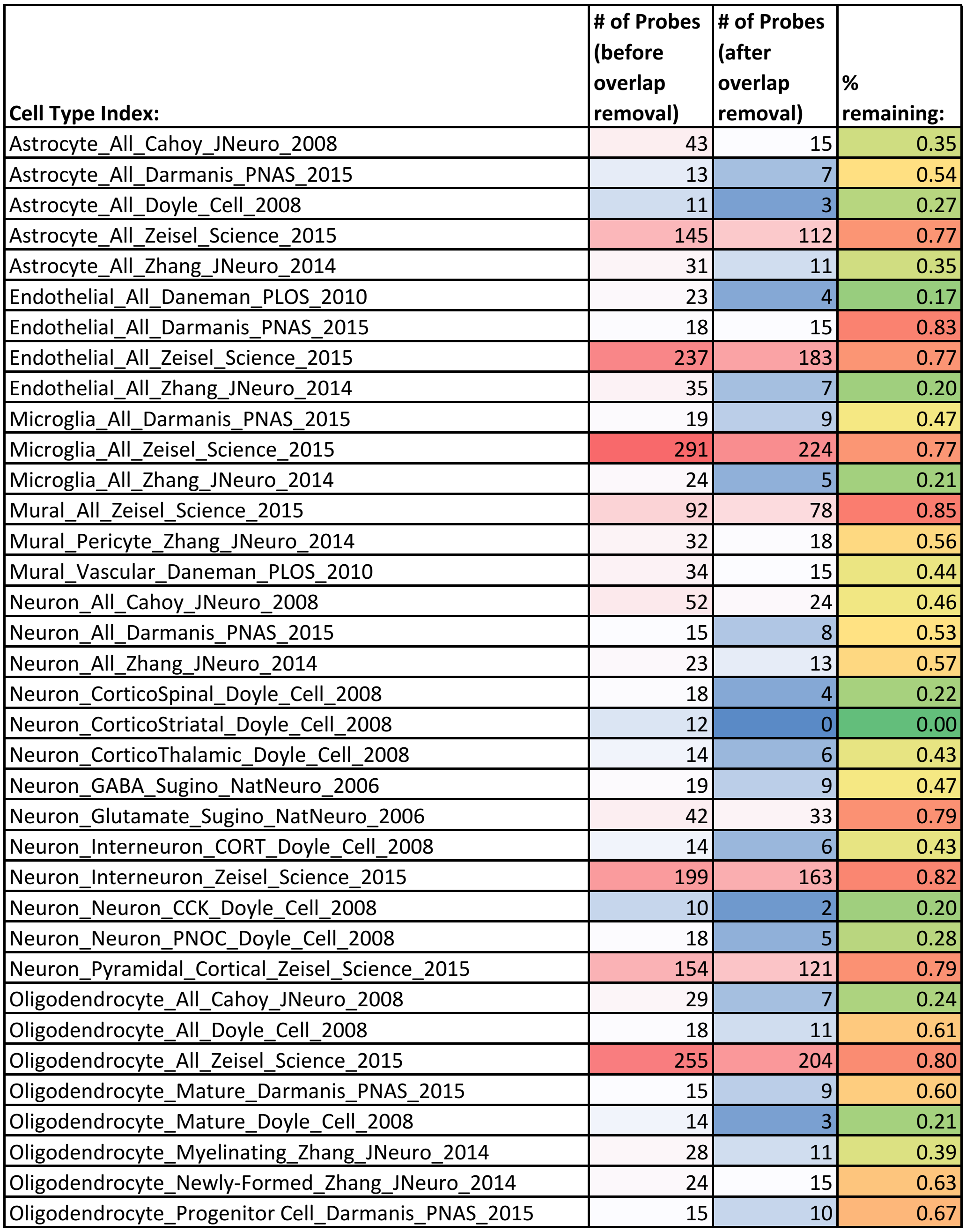


	Astrocyte_All_Cahoy_JNeuro_2008	Astrocyte_All_Darmanis_PNAS_2015	Astrocyte_All_Doyle_Cell_2008	Astrocyte_All_Zeisel_Science_2015	Astrocyte_All_Zhang_JNeuro_2014	Endothelial_All_Daneman_PLOS_2010	Endothelial_All_Darmanis_PNAS_2015	Endothelial_All_Zeisel_Science_2015	Endothelial_All_Zhang_JNeuro_2014	Microglia_All_Darmanis_PNAS_2015	Microglia_All_Zeisel_Science_2015	Microglia_All_Zhang_JNeuro_2014	Mural_All_Zeisel_Science_2015
Astrocyte_All_Cahoy_JNeuro_2008	100%	12%	7%	23%	28%	0%	0%	2%	0%	0%	2%	0%	0%
Astrocyte_All_Darmanis_PNAS_2015	38%	100%	0%	15%	8%	0%	0%	0%	0%	0%	0%	0%	0%
Astrocyte_All_Doyle_Cell_2008	36%	0%	100%	9%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Astrocyte_All_Zeisel_Science_2015	7%	1%	1%	100%	4%	0%	0%	1%	0%	0%	3%	0%	1%
Astrocyte_All_Zhang_JNeuro_2014	39%	3%	0%	19%	100%	0%	0%	0%	0%	0%	0%	0%	0%
Endothelial_All_Daneman_PLOS_2010	0%	0%	0%	0%	0%	100%	4%	70%	22%	0%	4%	0%	4%
Endothelial_All_Darmanis_PNAS_2015	0%	0%	0%	0%	0%	6%	100%	6%	6%	0%	0%	0%	0%
Endothelial_All_Zeisel_Science_2015	0%	0%	0%	1%	0%	7%	0%	100%	11%	0%	3%	0%	0%
Endothelial_All_Zhang_JNeuro_2014	0%	0%	0%	0%	0%	14%	3%	71%	100%	0%	0%	3%	0%
Microglia_All_Darmanis_PNAS_2015	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	47%	11%	0%
Microglia_All_Zeisel_Science_2015	0%	0%	0%	2%	0%	0%	0%	2%	0%	3%	100%	6%	0%
Microglia_All_Zhang_JNeuro_2014	0%	0%	0%	0%	0%	0%	0%	4%	4%	8%	75%	100%	0%
Mural_All_Zeisel_Science_2015	0%	0%	0%	1%	0%	1%	0%	1%	0%	0%	1%	0%	100%
Mural_Pericyte_Zhang_JNeuro_2014	0%	0%	0%	19%	3%	0%	0%	0%	0%	3%	0%	0%	9%
Mural_Vascular_Daneman_PLOS_2010	6%	0%	0%	3%	9%	0%	3%	3%	0%	0%	6%	0%	9%
Neuron_All_Cahoy_JNeuro_2008	0%	0%	0%	0%	0%	0%	0%	2%	2%	0%	2%	0%	0%
Neuron_All_Darmanis_PNAS_2015	0%	0%	0%	7%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Neuron_All_Zhang_JNeuro_2014	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Neuron_CorticoSpinal_Doyle_Cell_2008	0%	0%	0%	0%	0%	17%	0%	39%	28%	0%	0%	0%	0%
Neuron_CorticoStriatal_Doyle_Cell_2008	0%	0%	0%	0%	0%	8%	0%	8%	0%	8%	67%	8%	8%
Neuron_CorticoThalamic_Doyle_Cell_2008	0%	0%	0%	0%	0%	0%	0%	14%	0%	0%	0%	0%	0%
Neuron_GABA_Sugino_NatNeuro_2006	0%	0%	0%	5%	0%	0%	0%	5%	0%	0%	0%	0%	0%
Neuron_Glutamate_Sugino_NatNeuro_2006	0%	0%	0%	0%	0%	0%	0%	5%	0%	0%	0%	0%	0%
Neuron_Interneuron_CORT_Doyle_Cell_2008	0%	0%	0%	0%	0%	7%	0%	0%	7%	0%	0%	0%	7%
Neuron_Interneuron_Zeisel_Science_2015	1%	0%	0%	1%	1%	0%	0%	1%	0%	0%	1%	0%	1%
Neuron_Neuron_CCK_Doyle_Cell_2008	0%	0%	0%	0%	0%	10%	0%	10%	0%	0%	10%	0%	0%
Neuron_Neuron_PNOC_Doyle_Cell_2008	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Neuron_Pyramidal_Cortical_Zeisel_Science_2015	1%	0%	0%	1%	1%	0%	0%	1%	0%	0%	2%	0%	0%
Oligodendrocyte_All_Cahoy_JNeuro_2008	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	3%	0%	0%
Oligodendrocyte_All_Doyle_Cell_2008	0%	0%	0%	6%	0%	0%	0%	0%	0%	0%	6%	0%	0%
Oligodendrocyte_All_Zeisel_Science_2015	0%	0%	0%	2%	0%	0%	0%	0%	0%	0%	2%	0%	0%
Oligodendrocyte_Mature_Darmanis_PNAS_2015	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Oligodendrocyte_Mature_Doyle_Cell_2008	0%	0%	0%	14%	0%	0%	0%	7%	0%	0%	7%	0%	7%
Oligodendrocyte_Myelinating_Zhang_JNeuro_2014	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Oligodendrocyte_Newly-Formed_Zhang_JNeuro_2014	0%	0%	0%	4%	0%	0%	0%	4%	0%	0%	0%	0%	0%
Oligodendrocyte_Progenitor Cell_Darmanis_PNAS_2015	7%	0%	0%	7%	7%	0%	0%	0%	0%	0%	0%	0%	7%
Oligodendrocyte_Progenitor Cell_Zhang_JNeuro_2014	0%	0%	0%	5%	0%	0%	0%	0%	0%	0%	0%	0%	0%
RBC_All_GeneCardSearch_Hemoglobin_ErythrocyteSpecifi	0%	0%	0%	0%	0%	0%	0%	9%	0%	0%	0%	0%	0%

	Mural_Pericyte_Zhang_JNeuro_2014	Mural_Vascular_Daneman_PLOS_2010	Neuron_All_Cahoy_JNeuro_2008	Neuron_All_Darmanis_PNAS_2015	Neuron_All_Zhang_JNeuro_2014	Neuron_CorticoSpinal_Doyle_Cell_2008	Neuron_CorticoStriatal_Doyle_Cell_2008	Neuron_CorticoThalamic_Doyle_Cell_2008	Neuron_GABA_Sugino_NatNeuro_2006	Neuron_Glutamate_Sugino_NatNeuro_2006	Neuron_Interneuron_CORT_Doyle_Cell_2008	Neuron_Interneuron_Zeisel_Science_2015	Neuron_Neuron_CCK_Doyle_Cell_2008
Astrocyte_All_Cahoy_JNeuro_2008	0%	5%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	0%
Astrocyte_All_Darmanis_PNAS_2015	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Astrocyte_All_Doyle_Cell_2008	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Astrocyte_All_Zeisel_Science_2015	3%	1%	0%	1%	0%	0%	0%	0%	1%	0%	0%	1%	0%
Astrocyte_All_Zhang_JNeuro_2014	3%	10%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%
Endothelial_All_Daneman_PLOS_2010	0%	0%	0%	0%	0%	13%	4%	0%	0%	0%	4%	0%	4%
Endothelial_All_Darmanis_PNAS_2015	0%	6%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Endothelial_All_Zeisel_Science_2015	0%	1%	0%	0%	0%	4%	0%	0%	0%	1%	0%	0%	0%
Endothelial_All_Zhang_JNeuro_2014	0%	0%	3%	0%	0%	20%	0%	0%	0%	0%	3%	0%	0%
Microglia_All_Darmanis_PNAS_2015	5%	0%	0%	0%	0%	0%	5%	0%	0%	0%	0%	0%	0%
Microglia_All_Zeisel_Science_2015	0%	1%	0%	0%	0%	0%	2%	0%	0%	0%	0%	1%	0%
Microglia_All_Zhang_JNeuro_2014	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%
Mural_All_Zeisel_Science_2015	3%	3%	0%	0%	0%	0%	1%	0%	0%	0%	1%	1%	0%
Mural_Pericyte_Zhang_JNeuro_2014	100%	9%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%
Mural_Vascular_Daneman_PLOS_2010	9%	100%	3%	0%	0%	0%	6%	0%	0%	0%	0%	0%	0%
Neuron_All_Cahoy_JNeuro_2008	0%	2%	100%	10%	8%	2%	0%	4%	0%	4%	0%	6%	2%
Neuron_All_Darmanis_PNAS_2015	0%	0%	33%	100%	0%	0%	0%	0%	0%	0%	0%	13%	0%
Neuron_All_Zhang_JNeuro_2014	0%	0%	17%	0%	100%	0%	0%	0%	0%	0%	4%	17%	0%
Neuron_CorticoSpinal_Doyle_Cell_2008	0%	0%	11%	0%	0%	100%	0%	6%	0%	0%	0%	0%	0%
Neuron_CorticoStriatal_Doyle_Cell_2008	17%	25%	0%	0%	0%	0%	100%	0%	0%	8%	0%	0%	0%
Neuron_CorticoThalamic_Doyle_Cell_2008	0%	0%	29%	0%	0%	7%	0%	100%	0%	0%	0%	0%	0%
Neuron_GABA_Sugino_NatNeuro_2006	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	42%	0%
Neuron_Glutamate_Sugino_NatNeuro_2006	0%	0%	5%	0%	0%	0%	2%	0%	0%	100%	0%	0%	2%
Neuron_Interneuron_CORT_Doyle_Cell_2008	0%	0%	0%	0%	7%	0%	0%	0%	0%	0%	100%	21%	14%
Neuron_Interneuron_Zeisel_Science_2015	0%	0%	2%	1%	2%	0%	0%	0%	5%	0%	2%	100%	0%
Neuron_Neuron_CCK_Doyle_Cell_2008	0%	0%	10%	0%	0%	0%	0%	0%	0%	10%	20%	0%	100%
Neuron_Neuron_PNOC_Doyle_Cell_2008	0%	0%	11%	0%	6%	0%	0%	0%	0%	0%	0%	44%	0%
Neuron_Pyramidal_Cortical_Zeisel_Science_2015	1%	2%	5%	0%	1%	2%	0%	2%	1%	1%	1%	3%	0%
Oligodendrocyte_All_Cahoy_JNeuro_2008	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Oligodendrocyte_All_Doyle_Cell_2008	0%	6%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Oligodendrocyte_All_Zeisel_Science_2015	0%	0%	1%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%
Oligodendrocyte_Mature_Darmanis_PNAS_2015	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Oligodendrocyte_Mature_Doyle_Cell_2008	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Oligodendrocyte_Myelinating_Zhang_JNeuro_2014	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Oligodendrocyte_Newly-Formed_Zhang_JNeuro_2014	0%	0%	4%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%
Oligodendrocyte_Progenitor Cell_Darmanis_PNAS_2015	0%	7%	0%	0%	0%	0%	0%	7%	0%	0%	0%	0%	0%
Oligodendrocyte_Progenitor Cell_Zhang_JNeuro_2014	11%	5%	0%	0%	0%	0%	0%	0%	0%	0%	0%	5%	0%
RBC_All_GeneCardSearch_Hemoglobin_ErythrocyteSpecifi	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

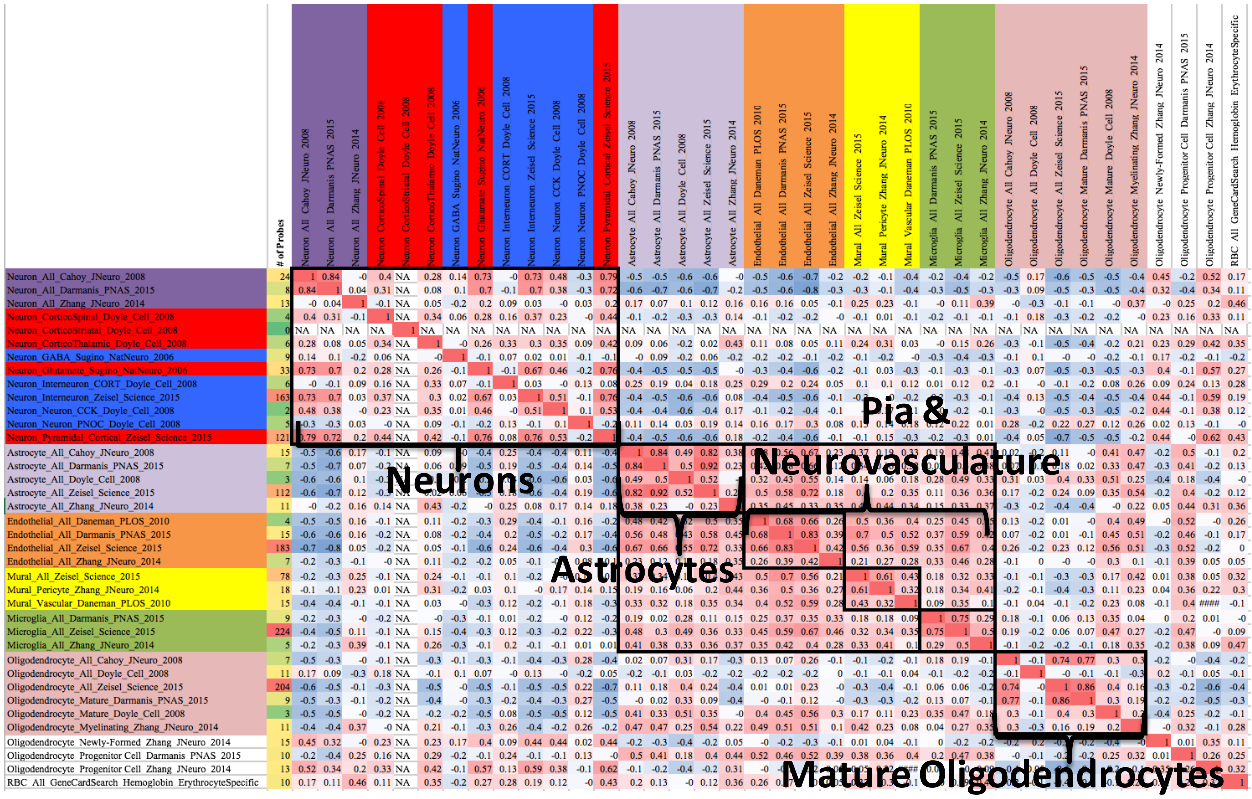
	Neuron_Neuron_PNOC_Doyle_Cell_2008	Neuron_Pyramidal_Cortical_Zeisel_Science_2015	Oligodendrocyte_All_Cahoy_JNeuro_2008	Oligodendrocyte_All_Doyle_Cell_2008	Oligodendrocyte_All_Zeisel_Science_2015	Oligodendrocyte_Mature_Darmanis_PNAS_2015	Oligodendrocyte_Mature_Doyle_Cell_2008	Oligodendrocyte_Myelinating_Zhang_JNeuro_2014	Oligodendrocyte_Newly-Formed_Zhang_JNeuro_2014	Oligodendrocyte_Progenitor Cell_Darmanis_PNAS_2015	Oligodendrocyte_Progenitor Cell_Zhang_JNeuro_2014	RBC_All_GeneCardSearch_Hemoglobin_ErythrocyteSpecific
Astrocyte_All_Cahoy_JNeuro_2008	0%	5%	0%	0%	0%	0%	0%	0%	0%	2%	0%	0%
Astrocyte_All_Darmanis_PNAS_2015	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Astrocyte_All_Doyle_Cell_2008	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Astrocyte_All_Zeisel_Science_2015	0%	1%	1%	1%	3%	0%	1%	0%	1%	1%	1%	0%
Astrocyte_All_Zhang_JNeuro_2014	0%	6%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%
Endothelial_All_Daneman_PLOS_2010	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Endothelial_All_Darmanis_PNAS_2015	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Endothelial_All_Zeisel_Science_2015	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Endothelial_All_Zhang_JNeuro_2014	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Microglia_All_Darmanis_PNAS_2015	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Microglia_All_Zeisel_Science_2015	0%	1%	0%	0%	2%	0%	0%	0%	0%	0%	0%	0%
Microglia_All_Zhang_JNeuro_2014	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Mural_All_Zeisel_Science_2015	0%	0%	0%	0%	1%	0%	1%	0%	0%	1%	0%	0%
Mural_Pericyte_Zhang_JNeuro_2014	0%	13%	0%	0%	3%	0%	0%	0%	0%	0%	16%	0%
Mural_Vascular_Daneman_PLOS_2010	0%	9%	0%	3%	0%	0%	0%	0%	0%	3%	3%	0%
Neuron_All_Cahoy_JNeuro_2008	4%	13%	0%	0%	4%	0%	0%	0%	2%	0%	0%	0%
Neuron_All_Darmanis_PNAS_2015	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Neuron_All_Zhang_JNeuro_2014	4%	9%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Neuron_CorticoSpinal_Doyle_Cell_2008	0%	22%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Neuron_CorticoStriatal_Doyle_Cell_2008	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Neuron_CorticoThalamic_Doyle_Cell_2008	0%	29%	0%	0%	7%	0%	0%	0%	0%	7%	0%	0%
Neuron_GABA_Sugino_NatNeuro_2006	0%	5%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Neuron_Glutamate_Sugino_NatNeuro_2006	0%	2%	0%	0%	5%	0%	0%	0%	2%	0%	0%	0%
Neuron_Interneuron_CORT_Doyle_Cell_2008	0%	7%	0%	0%	7%	0%	0%	0%	0%	0%	0%	0%
Neuron_Interneuron_Zeisel_Science_2015	4%	3%	0%	0%	1%	0%	0%	0%	0%	0%	1%	0%
Neuron_Neuron_CCK_Doyle_Cell_2008	0%	0%	0%	0%	10%	0%	0%	0%	0%	0%	0%	0%
Neuron_Neuron_PNOC_Doyle_Cell_2008	100%	6%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Neuron_Pyramidal_Cortical_Zeisel_Science_2015	1%	100%	0%	0%	1%	0%	0%	0%	0%	1%	1%	0%
Oligodendrocyte_All_Cahoy_JNeuro_2008	0%	0%	100%	7%	66%	17%	7%	31%	0%	0%	0%	0%
Oligodendrocyte_All_Doyle_Cell_2008	0%	0%	11%	100%	6%	0%	0%	0%	6%	0%	11%	0%
Oligodendrocyte_All_Zeisel_Science_2015	0%	0%	7%	0%	100%	2%	3%	6%	2%	0%	0%	0%
Oligodendrocyte_Mature_Darmanis_PNAS_2015	0%	0%	33%	0%	40%	100%	7%	13%	0%	0%	0%	0%
Oligodendrocyte_Mature_Doyle_Cell_2008	0%	0%	14%	0%	50%	7%	100%	14%	0%	0%	0%	0%
Oligodendrocyte_Myelinating_Zhang_JNeuro_2014	0%	0%	32%	0%	54%	7%	7%	100%	0%	0%	0%	0%
Oligodendrocyte_Newly-Formed_Zhang_JNeuro_2014	0%	0%	0%	4%	25%	0%	0%	0%	100%	0%	0%	0%
Oligodendrocyte_Progenitor Cell_Darmanis_PNAS_2015	0%	7%	0%	0%	7%	0%	0%	0%	0%	100%	0%	0%
Oligodendrocyte_Progenitor Cell_Zhang_JNeuro_2014	0%	11%	0%	11%	0%	0%	0%	0%	0%	0%	100%	0%
RBC_All_GeneCardSearch_Hemoglobin_ErythrocyteSpecifi	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%

Suppl. Figure 14. There was minimal overlap between the transcripts included within different cell type indices that did not fall under the same primary cell type category. *The cell type indices were developed using cell type specific genes identified by different publications, species, and methodologies, and are categorically color-coded in a manner similar to* ***Suppl. Figure 16.*** *Percentage overlap between indices is color-coded using a gradient from green (0% overlap) to red (100% overlap), with the denominator in the percentage overlap equation defined as the cell type index specified by the row. Notably, only cell type indices derived from (15) shows overlap of >20% with cell type indices not found within the same primary cell type category. This may be due to (15) using different methodology (TRAP: Translating Ribosome Affinity Purification) to define cell type specific transcripts than the other publications.*

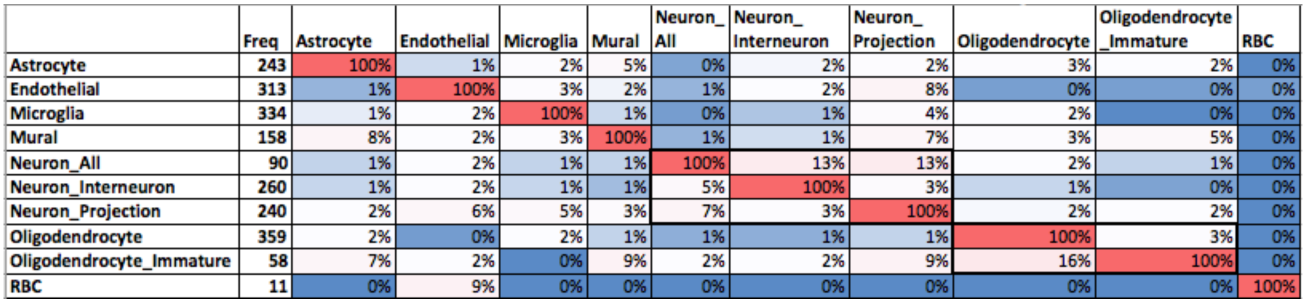


Cell Type Index:	# of Probes (before overlap removal)	# of Probes (after overlap removal)	% remaining:
Astrocyte_All_Cahoy_JNeuro_2008	43	15	0.35
Astrocyte_All_Darmanis_PNAS_2015	13	7	0.54
Astrocyte_All_Doyle_Cell_2008	11	3	0.27
Astrocyte_All_Zeisel_Science_2015	145	112	0.77
Astrocyte_All_Zhang_JNeuro_2014	31	11	0.35
Endothelial_All_Daneman_PLOS_2010	23	4	0.17
Endothelial_All_Darmanis_PNAS_2015	18	15	0.83
Endothelial_All_Zeisel_Science_2015	237	183	0.77
Endothelial_All_Zhang_JNeuro_2014	35	7	0.20
Microglia_All_Darmanis_PNAS_2015	19	9	0.47
Microglia_All_Zeisel_Science_2015	291	224	0.77
Microglia_All_Zhang_JNeuro_2014	24	5	0.21
Mural_All_Zeisel_Science_2015	92	78	0.85
Mural_Pericyte_Zhang_JNeuro_2014	32	18	0.56
Mural_Vascular_Daneman_PLOS_2010	34	15	0.44
Neuron_All_Cahoy_JNeuro_2008	52	24	0.46
Neuron_All_Darmanis_PNAS_2015	15	8	0.53
Neuron_All_Zhang_JNeuro_2014	23	13	0.57
Neuron_CorticoSpinal_Doyle_Cell_2008	18	4	0.22
Neuron_CorticoStriatal_Doyle_Cell_2008	12	0	0.00
Neuron_CorticoThalamic_Doyle_Cell_2008	14	6	0.43
Neuron_GABA_Sugino_NatNeuro_2006	19	9	0.47
Neuron_Glutamate_Sugino_NatNeuro_2006	42	33	0.79
Neuron_Interneuron_CORT_Doyle_Cell_2008	14	6	0.43
Neuron_Interneuron_Zeisel_Science_2015	199	163	0.82
Neuron_Neuron_CCK_Doyle_Cell_2008	10	2	0.20
Neuron_Neuron_PNOC_Doyle_Cell_2008	18	5	0.28
Neuron_Pyramidal_Cortical_Zeisel_Science_2015	154	121	0.79
Oligodendrocyte_All_Cahoy_JNeuro_2008	29	7	0.24
Oligodendrocyte_All_Doyle_Cell_2008	18	11	0.61
Oligodendrocyte_All_Zeisel_Science_2015	255	204	0.80
Oligodendrocyte_Mature_Darmanis_PNAS_2015	15	9	0.60
Oligodendrocyte_Mature_Doyle_Cell_2008	14	3	0.21
Oligodendrocyte_Myelinating_Zhang_JNeuro_2014	28	11	0.39
Oligodendrocyte_Newly-Formed_Zhang_JNeuro_2014	24	15	0.63
Oligodendrocyte_Progenitor Cell_Darmanis_PNAS_2015	15	10	0.67

Suppl. Figure 15. Removing probes that represent genes indentified as having cell type specific expression in multiple publications (preparation for the analysis in Suppl. Figure 16). *A summary of the full number of probes included in each cell type index before and after removal of the probes that overlapped with any other index (color-labeled with pink representing a large number of probes and blue representing fewer probes).The percentage of probes retained in the index after full overlap removal is also provided, with green indicating a small percentage of probes retained, and red indicating a large percentage.*



Suppl. Figure 16. The convergence of cell content predictions derived from cell type specific transcripts originating from different publications remains after removing overlapping transcripts. *This figure follows the format of* ***Suppl. Figure 11****(Pritzker cortical dataset), but uses cell type indices calculated following removal of any probes identified as present in more than one index (see* ***Suppl. Figure 15****).The similarity of different cell type indices can be visualized using a correlation matrix. Within this matrix, correlations can range from a strong negative correlation of -1 (blue) to a strong positive correlation of 1 (red), therefore a large block of pink/red correlations is indicative of cell type indices that tend to be enriched in the same samples. The labels for cell type indices representing the same category of cell are color-coded as in* ***Suppl. Figure 11.***



	Freq	Astrocyte	Endothelial	Microglia	Mural	Neuron_ All	Neuron_ Interneuron	Neuron_ Projection	Oligodendrocyte	Oligodendrocyte _Immature	RBC
Astrocyte	243	100%	1%	2%	5%	0%	2%	2%	3%	2%	0%
Endothelial	313	1%	100%	3%	2%	1%	2%	8%	0%	0%	0%
Microglia	334	1%	2%	100%	1%	0%	1%	4%	2%	0%	0%
Mural	158	8%	2%	3%	100%	1%	1%	7%	3%	5%	0%
Neuron_All	90	1%	2%	1%	1%	100%	13%	13%	2%	1%	0%
Neuron_Interneuron	260	1%	2%	1%	1%	5%	100%	3%	1%	0%	0%
Neuron_Projection	240	2%	6%	5%	3%	7%	3%	100%	2%	2%	0%
Oligodendrocyte	359	2%	0%	2%	1%	1%	1%	1%	100%	3%	0%
Oligodendrocyte_Immature	58	7%	2%	0%	9%	2%	2%	9%	16%	100%	0%
RBC	11	0%	9%	0%	0%	0%	0%	0%	0%	0%	100%

Suppl. Figure 17. For later analyses, individual cell type indices were averaged by primary category, with any transcripts that overlapped between categories removed*. The percent overlap between transcripts defined as specific to different categories of cell type is illustrated below, color-coded with a gradient from blue (indicating 0% overlap) to red (indicating 100% overlap). The denominator in the percentage overlap equation was defined as the cell type category specified by the row. The column on the far left provides the number of probes included in each cell type category.*

7.5 *Additional figures and results:* Cell Type Indices Predict Other Genes Known to Be Cell Type Enriched

To identify other transcripts important to cell type specific functions in the human cortex, we ran a linear model on the signal from each gene probeset in the Pritzker prefontal cortex microarray dataset that included each of the ten consolidated primary cell type indices as well as six co-variates traditionally included in the analysis of human brain gene expression data (pH, Agonal Factor, PMI, Age, Gender, Diagnosis; **Equation 5**). On average, this model explained 35% of the variation in the data (R^2). Shown in **Suppl. Figure 18** are the most significant 10 gene probe sets positively associated with each cell type while controlling for the other cell types and co-variates within the model. Additional gene probe sets and statistical details can be found in **Suppl. Table 15.**

Astrocyte	Endothelial	Microglia	Mural	Neuron_All	Neuron_ Projection	Neuron_ Interneuron	Mature Oligodendrocyte	Red Blood Cell (RBC)
NOTCH2	**HLA-E**	**AIF1**	**TAGLN**	**VSNL1**	**PDE2A**	**TAC3**	**KLK6**	**HBD**
SDC2	**EPAS1**	**LAPTM5**	**MYL9**	**SYT1**	USF2	SLC24A3	**UGT8**	**HBB**
NTRK2	CLCN7	**IRF8**	MYH11	**SYNGR3**	DGKZ	**GAD1**	**MAG**	**PKLR**
CLDN10	**CLDN5**	**FCER1G**	CNN1	**NEFL**	**NUAK1**	**KIT**	**ELOVL1**	PGC
FGFR3	PAK4	**PTPRC**	MGP	**NRXN1**	SLC38A7	**GAD2**	**EVI2A**	*NA*
APOE	MYOF	**LAIR1**	**ACTA2**	**SNAP25**	BEGAIN	**ERBB4**	**PLLP**	DKK4
EZR	**ICAM2**	**LY86**	TP53I11	BCL2L1	KIAA0182	**LHX6**	**MOG**	LIPE
SLC1A3	ABCB1	FPR1	**COL18A1**	**MAPK1**	**KIF21B**	**SLC6A1**	**ASPA**	SPDEF
CST3	**GPR116**	**C3**	**TPM2**	EEF1A2	**PLXNA1**	**RELN**	**TF**	C19orf57
MLC1	**SDPR**	**ALOX5AP**	CRABP1	MEF2C	SLC8A2	ARL4C	**MAL**	*NA*

Suppl. Figure 18. The top 10 transcripts associated with each cell type index include those previously-identified as cell type enriched in the literature*. Transcripts are identified by official gene symbol. Yellow labels identify transcripts included in the original cell type index, orange transcripts were previously-identified as cell type enriched in the literature but were not included in the original list of cell type specific transcripts used to create the index. Additional transcripts and statistical details can be found in* ***Suppl. Table 15.*** *Please note that not all of the genes listed in the top ten list associated wit the Red Blood Cell index would survive a traditional threshold for false detction threshold (q<0.05).*

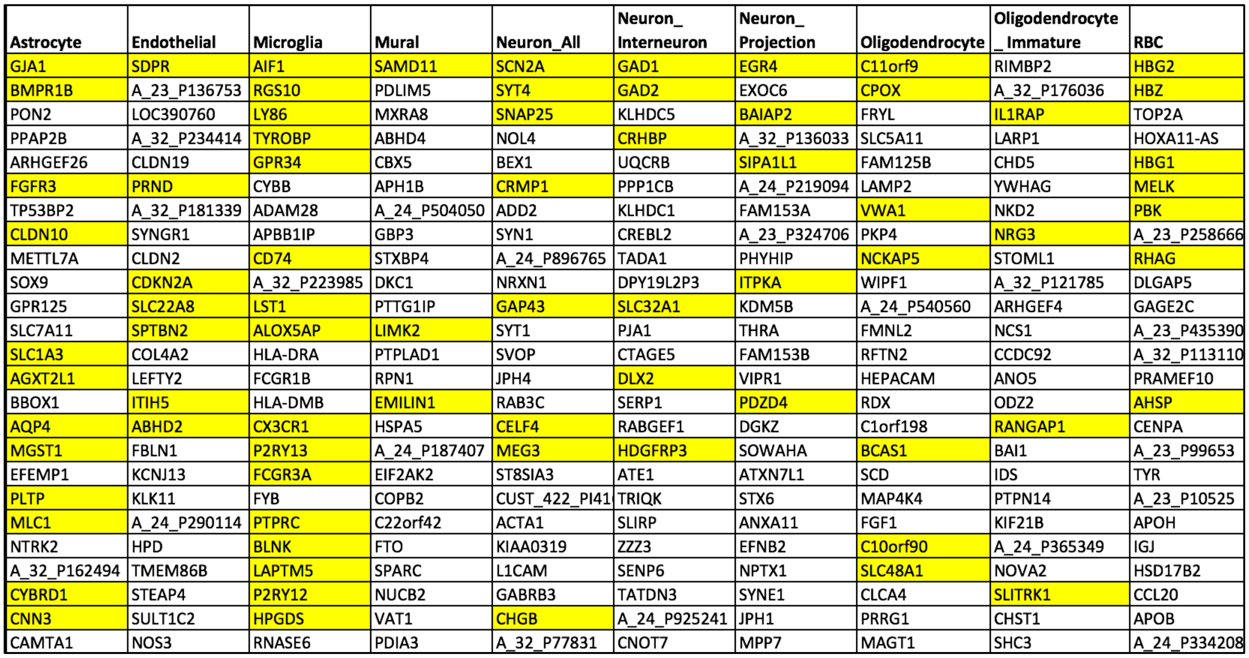
Many of the top gene probesets that we found to be related to each of the cell type indices are already known to be associated with that cell type in previous publications, validating our methodology. Importantly, this is true even when the genes were not included in the original list of cell type specific genes used to generate the index. For example, we found that HLA-E (Major Histocompatibility Complex, Class I, E) and EPAS1 (endothelial PAS domain protein 1) were both strongly associated with our endothelial index, and both are known to be involved in endothelial cell activation (HLA-E, in response to immune challenge: (95); EPAS1, in response to lack of oxygen: (96)). NOTCH2 (Notch 2), one of the top astrocyte-related genes, promotes astrocytic cell lineage (97), and APOE (Apolipoprotein E) is primarily secreted by astrocytes in the central nervous system (98). One of the top interneuron genes, LHX6 (LIM Homeobox 6), is specifically enriched in parvalbumin-containing interneurons in the human cortex (2). Another top interneuron gene, ERBB4 (Erb-B2 Receptor Tyrosine Kinase 4), controls the development of GABA circuitry in the cortex (99). The top neuron-related genes include several genes related to synaptic function (SYT1 (Synaptotagmin I), SYNGR3 (Synaptogyrin 3), NRXN1 (Neurexin 1); http://www.genecards.org/). The top projection neuron-related gene, PDE2A

(Phosphodiesterase 2A, CGMP-Stimulated), is preferentially expressed in cortical pyramidal neurons (100), and KIF21B (Kinesin Family Member 21B) is a kinesin that has been found in the dendrites of pyramidal neurons (101). We also rediscovered probesets representing genes that were listed as alternative orthologs to those included in our original cell type specific gene lists (oligodendrocytes: EVI2A vs.CTD-2370N5.3, microglia: LAIR1 vs. LAIR2, mural cells: COL18A1 vs. COL15A1, ACTA2 vs. ACTG1). Altogether, these results suggest that our cell type indices were associated with the variability of transcripts in the cortex that represented particular cell types and could re-identify known cell type specific markers.

As a follow-up analysis, we also outputted a table of the top genes associated with each cell type within the Allen Brain Atlas dataset (as assessed using the model in **Equation 6**). We found that the results similarly included a mixture of well-known cell type markers and novel findings (**Suppl. Figure 19, Suppl. Table 6**).

Equation 6: A model of gene expression for the Allen Brain Atlas dataset, colored to illustrate subcomponents. *The base model (intercept) is presented in green, the cell type indices for the most prevalent cell types are colored red, and the remaining cell type indices are in purple.*

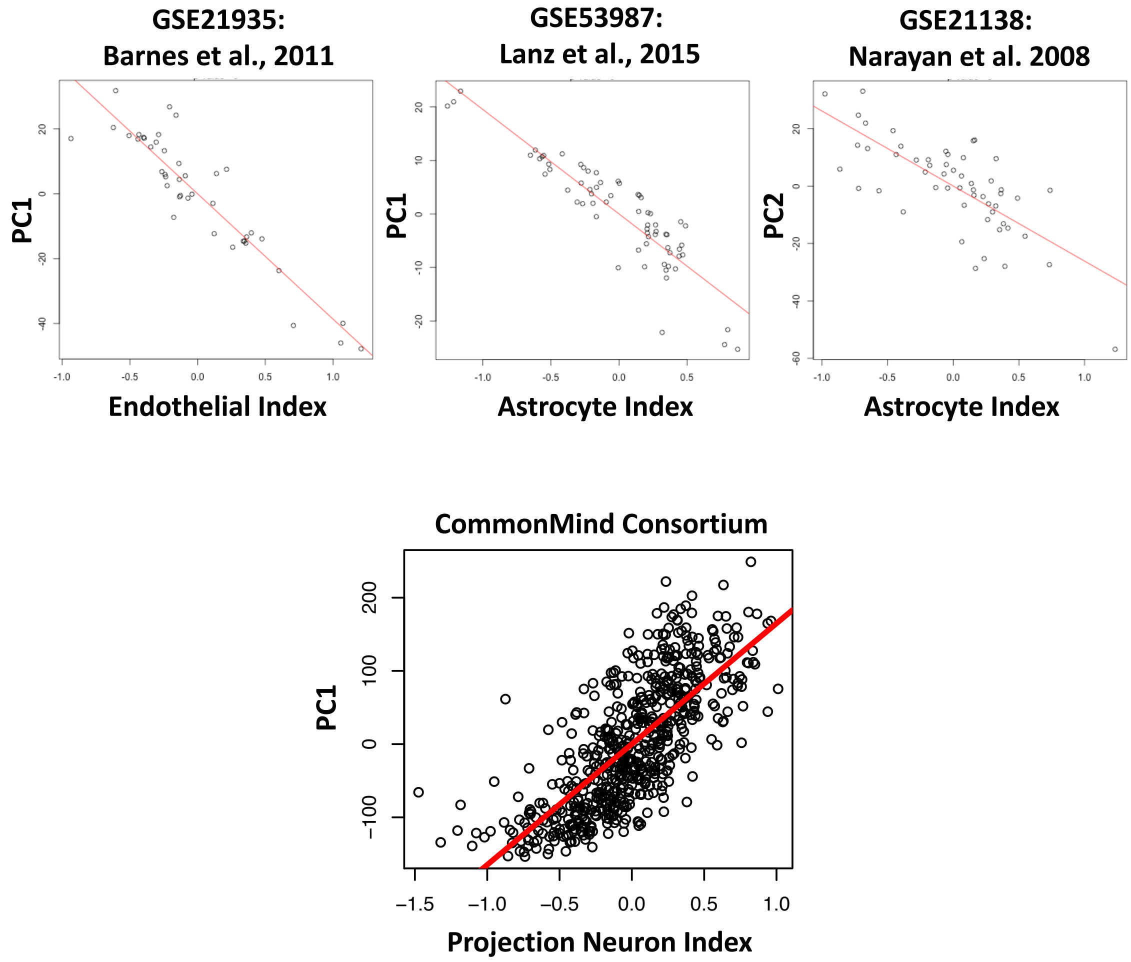
$$\text{Gene Expression (Probeset Signal)} = \beta0 + \beta1*(\text{Astrocyte})+\beta2*(\text{Oligodendrocyte})+\beta3*(\text{Microglia})+\beta4*(\text{Interneuron})+\beta5*(\text{ProjectionNeuron}) +\beta6*(\text{Endothelial})+\beta7*(\text{Neuron\_All})+\beta8*(\text{Oligodendrocyte\_Immature})+\beta9*(\text{Mural})+\beta10*(\text{RBC})+ \varepsilon$$



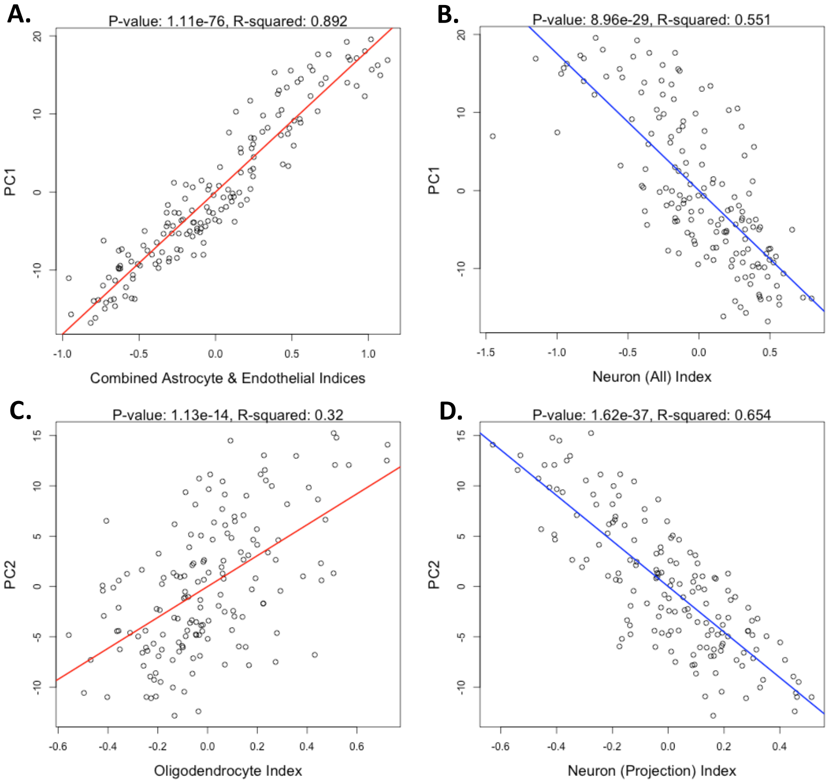
Astrocyte	Endothelial	Microglia	Mural	Neuron_All	Neuron_ Interneuron	Neuron_ Projection	Oligodendrocyte	Oligodendrocyte _ Immature	RBC
GJA1	SDPR	AIF1	SAMD11	SCN2A	GAD1	EGR4	C11orf9	RIMBP2	HBG2
BMPR1B	A_23_P136753	RGS10	PDLIM5	SYT4	GAD2	EXOC6	CPOX	A_32_P176036	HBZ
PON2	LOC390760	LY86	MXRA8	SNAP25	KLHDC5	BAIAP2	FRYL	IL1RAP	TOP2A
PPAP2B	A_32_P234414	TYROBP	ABHD4	NOL4	CRHBP	A_32_P136033	SLC5A11	LARP1	HOXA11-AS
ARHGEF26	CLDN19	GPR34	CBX5	BEX1	UQCRB	SIPA1L1	FAM125B	CHD5	HBG1
FGFR3	PRND	CYBB	APH1B	CRMP1	PPP1CB	A_24_P219094	LAMP2	YWHAG	MELK
TP53BP2	A_32_P181339	ADAM28	A_24_P504050	ADD2	KLHDC1	FAM153A	VWA1	NKD2	PBK
CLDN10	SYNGR1	APBB1IP	GBP3	SYN1	CREBL2	A_23_P324706	PKP4	NRG3	A_23_P258666
METTL7A	CLDN2	CD74	STXBP4	A_24_P896765	TADA1	PHYHIP	NCKAP5	STOML1	RHAG
SOX9	CDKN2A	A_32_P223985	DKC1	NRXN1	DPY19L2P3	ITPKA	WIPF1	A_32_P121785	DLGAP5
GPR125	SLC22A8	LST1	PTTG1IP	GAP43	SLC32A1	KDM5B	A_24_P540560	ARHGEF4	GAGE2C
SLC7A11	SPTBN2	ALOX5AP	LIMK2	SYT1	PJA1	THRA	FMNL2	NCS1	A_23_P435390
SLC1A3	COL4A2	HLA-DRA	PTPLAD1	SVOP	CTAGE5	FAM153B	RFTN2	CCDC92	A_32_P113110
AGXT2L1	LEFTY2	FCGR1B	RPN1	JPH4	DLX2	VIPR1	HEPACAM	ANO5	PRAMEF10
BBOX1	ITIH5	HLA-DMB	EMILIN1	RAB3C	SERP1	PDZD4	RDX	ODZ2	AHSP
AQP4	ABHD2	CX3CR1	HSPA5	CELF4	RABGEF1	DGKZ	C1orf198	RANGAP1	CENPA
MGST1	FBLN1	P2RY13	A_24_P187407	MEG3	HDGFRP3	SOWAHA	BCAS1	BAI1	A_23_P99653
EFEMP1	KCNJ13	FCGR3A	EIF2AK2	ST8SIA3	ATE1	ATXN7L1	SCD	IDS	TYR
PLTP	KLK11	FYB	COPB2	CUST_422_PI41	TRIQK	STX6	MAP4K4	PTPN14	A_23_P10525
MLC1	A_24_P290114	PTPRC	C22orf42	ACTA1	SLIRP	ANXA11	FGF1	KIF21B	APOH
NTRK2	HPD	BLNK	FTO	KIAA0319	ZZZ3	EFNB2	C10orf90	A_24_P365349	IGJ
A_32_P162494	TMEM86B	LAPTM5	SPARC	L1CAM	SENP6	NPTX1	SLC48A1	NOVA2	HSD17B2
CYBRD1	STEAP4	P2RY12	NUCB2	GABRB3	TATDN3	SYNE1	CLCA4	SLITRK1	CCL20
CNN3	SULT1C2	HPGDS	VAT1	CHGB	A_24_P925241	JPH1	PRRG1	CHST1	APOB
CAMTA1	NOS3	RNASE6	PDIA3	A_32_P77831	CNOT7	MPP7	MAGT1	SHC3	A_24_P334208

Suppl. Figure 19. The top 25 probes associated with each primary cell type index in the Allen Brain Atlas dataset. *Depicted are the top probes identified in association with each of the primary cell types as determined by a linear model that included indices for all 10 primary cell types. This model was run using samples from all 160 brain regions. Similar to the results for the Pritzker dorsolateral prefrontal cortex data, the genes identified in the Allen Brain Atlas data as having strong relationships with particular cell types include a mixture of well-known cell type markers and more novel findings.*

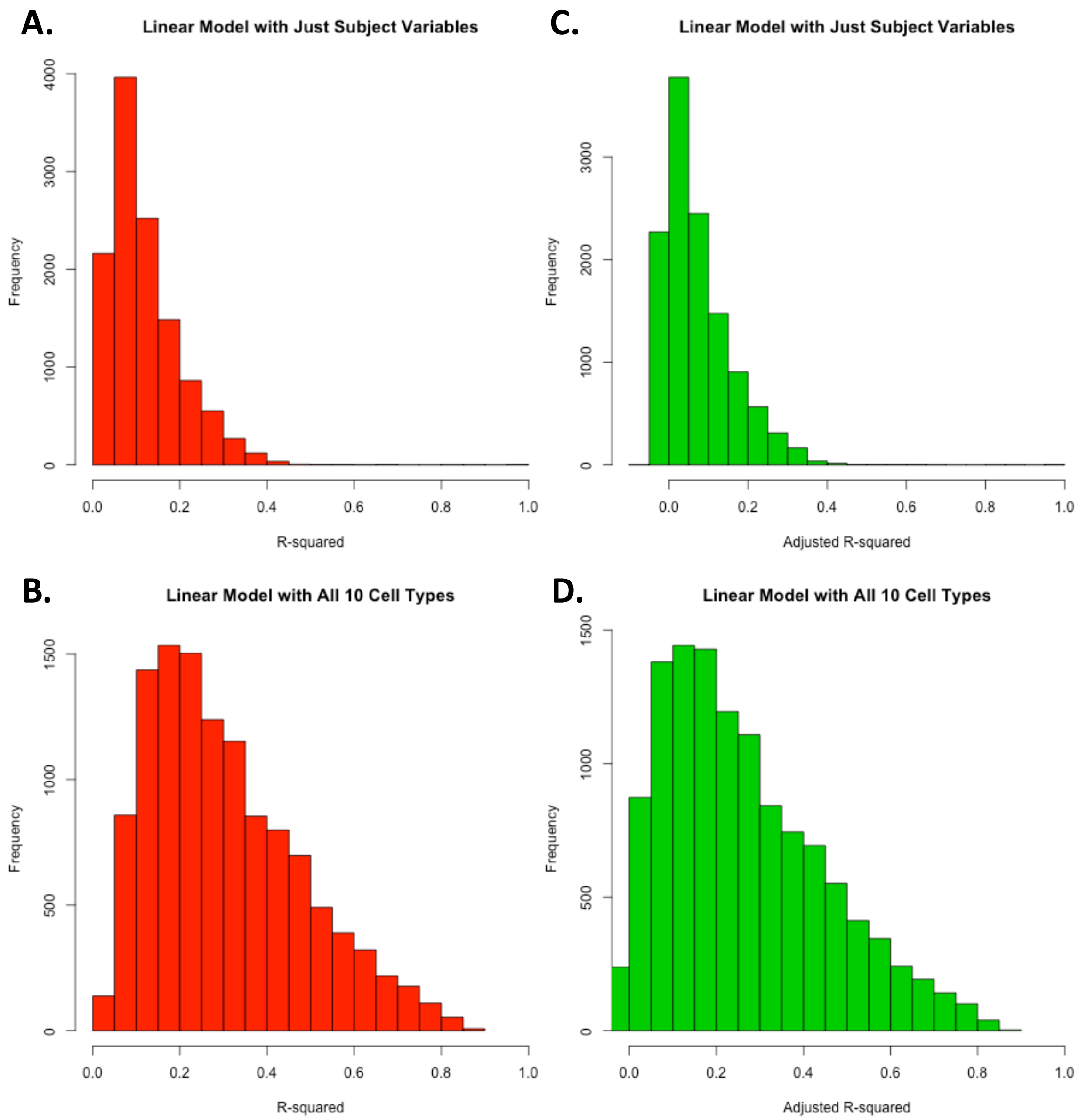
7.6 *Additional figures and results:* Inferred Cell Type Composition Explains a Large Percentage of the Sample-Sample Variability in Microarray Data from Macro-Dissected Human Cortical Tissue



Suppl. Figure 20. Replication: Cell content predictions explain a large percentage of the variability in microarray and RNA-Seq data derived from the human cortex. *The results shown above are from the four other human cortical datasets discussed in the paper. Within the cross-regional Allen Brain Atlas dataset, we also found that the top principal components of variation were overwhelmingly explained by predicted cell type balance, with a model that included all 10 cell type indices accounting for a larger percentage of the variation in the top 4 principal components (PC1: F(10, 830)=1051, R^2=0.927, p<2.2e-16; PC2: F(10, 830)=96.98, R^2=0.539, p<2.2e-16; PC3: F(10, 830)=133.2, R^2= 0.616, p<2.2e-16; PC4: F(10, 830)=121.3, R^2= 0.594, p<2.2e-16), although the specific relationships sometimes differed from what was seen in the prefrontal cortex.*

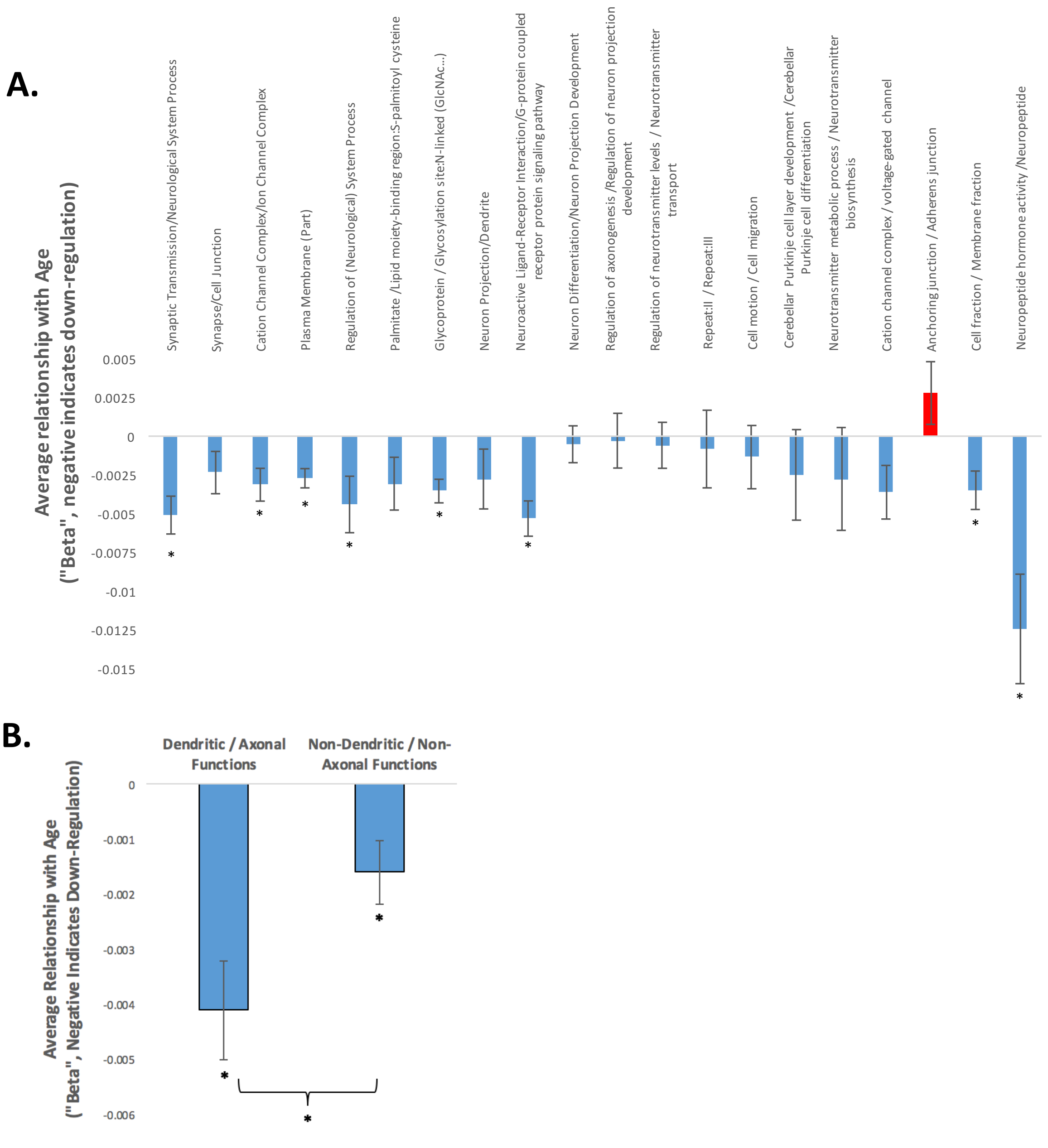


Suppl. Figure 21. Cell content predictions explain a large percentage of the variability in microarray data from non-cell type specific genes. *The results shown here look almost identical to those shown in* ***Figure 5 ,*** *except that the principal components analysis in this case was run while excluding all cell type specific genes from the dataset.*



Suppl. Figure 22. Predicted cell content accounts for a larger percentage of the variability in the signal from individual probesets than the most commonly examined subject variables. *Shown below are histograms illustrating the R-squared (A & B) and adjusted R-squared (C & D) for all 11979 probesets in the Pritzker dorsolateral prefrontal cortex dataset as fit using two linear models: (A & C) A model that includes diagnosis (MDD, BP, Schiz) and five subject variables commonly used as co-variates in the analysis of brain microarray data (Brain pH, agonal factor, age, gender, post-mortem interval), (B & D) A model that includes the consolidated indices for all 10 primary cell types.*

7.7 *Additional figures and results:* Discriminating Between Changes in Cell Type Balance and Cell-Type Specific Function



Suppl. Figure 23. The predicted decrease in neuronal cell content in relationship to age is unlikely to be fully explained by synaptic atrophy. *Within the list of neuron-specific genes, 240 functional clusters were identified using DAVID (using the full HT-U133A chip as background).* ***A)*** *The genes in 19 out of the top 20 functional clusters showed decreased expression with age on average, as determined within a linear model that controlled for known confounds. Depicted is the average effect of age +/-SE for each* *cluster (asterisks: p<0.05). Blue represents down-regulation, red is up-regulation. Overall, 76% of all 240 functional clusters showed a negative relationship with age on average (**Suppl. Table 5**). **B.)** We blindly chose 29 functional clusters that were clearly related to dendritic/axonal functions and 41 functional clusters that seemed distinctly unrelated to dendritic/axonal functions. Transcripts from both classifications showed an average decrease in expression with age (*p= *9.197e-05,* p=*0.008756, respectively), but the decrease was larger for transcripts associated with dendritic/axonal-related functions (p= 0.02339). Depicted is the average effect of age +/-SE for each classification of cluster.*

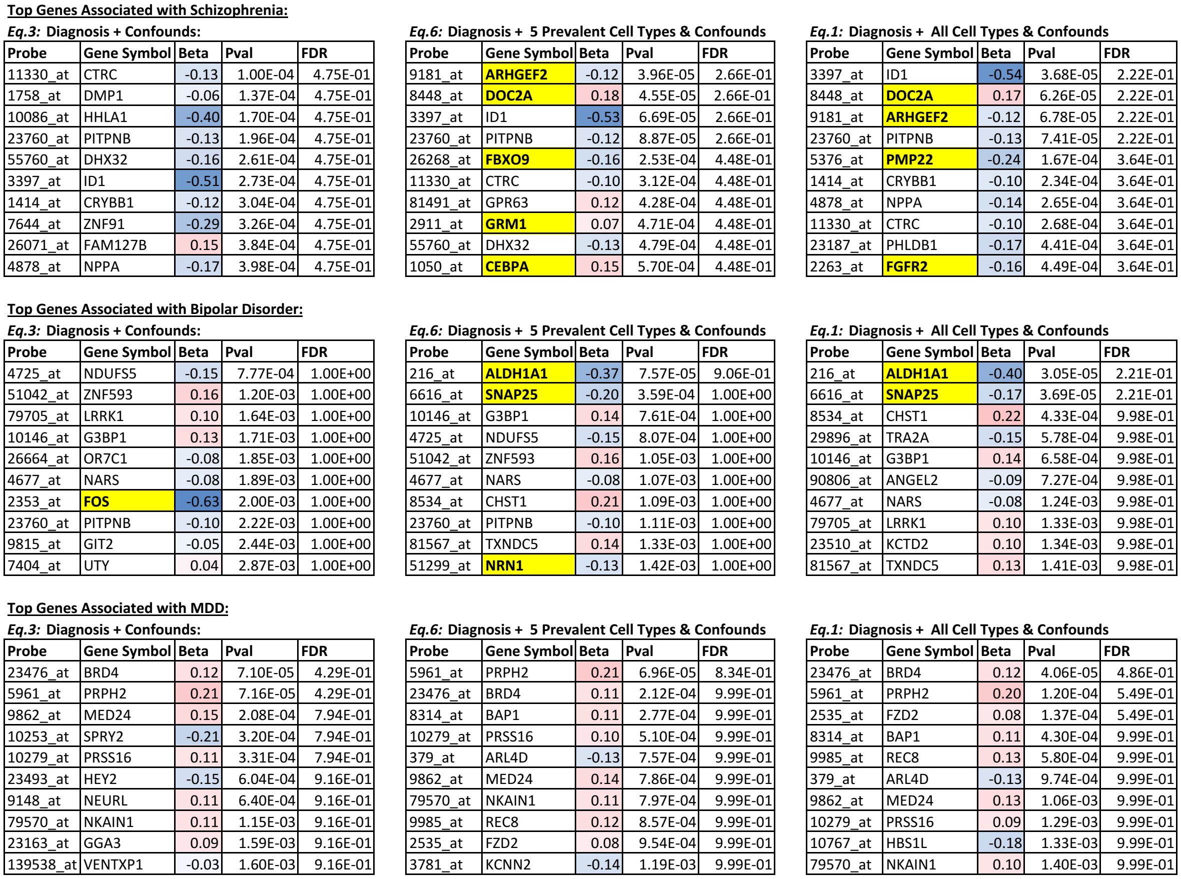
7.8 The Top Diagnosis-Related Genes Identified by Models that Include Cell Content Predictions Pinpoint Known Risk Candidates

Although the inclusion of predicted cell type balance in our model occasionally improved our ability to detect previously-identified relationships with diagnosis, most relationships still went undetected in the Pritzker dataset and none of the diagnosis relationships survived standard p-value corrections for multiple comparisons when included in a full microarray analysis. This could be due to a variety of factors, including microarray platform and probe sensitivity as well as the possibility that other cell types in the dataset are showing effects in a competing direction. Therefore, we decided to ask a complementary question: Of the top diagnosis relationships that we see in our dataset, how many have been previously observed in the literature? If including predicted cell type balance in our models improves the signal to noise ratio of our analyses, then we would expect that the top diagnosis-related genes in our dataset would be more likely to overlap with previous findings. In an attempt to perform this comparison in an unbiased and efficient manner, we limited our search to PubMed, using as search terms only the respective human gene symbol and diagnosis (“Schizophrenia”, “Bipolar”, or “Depression”). For the genes related to MDD in our dataset, we also expanded the search to include two highly-correlated traits that are more quantifiable and likely to have a genetic basis: “Anxiety” and “Suicide”. Then we narrowed our results only to studies using human subjects.

Before controlling for cell type, we found that only one of the top 10 genes related to diagnosis (FOS: (102,103)) or the presence or absence of psychiatric illness (ALDH1A1: (104)) had been previously noted in the human literature. In contrast, when we used a model that included the five most prevalent cortical cell types (Model#4), we found that five of the top 10 genes associated with

Schizophrenia had been previously identified in the literature (ARHGEF2: (105), DOC2A: (106), FBX09: (74), GRM1: (107,108); CEBPA: (89)), and three of the top 10 genes associated with Bipolar Disorder (ALDH1A1: (104), SNAP25: (109), NRN1:(110); **Suppl. Figure 24**, **Suppl. Table 9**). This was a significant enrichment in overlap with the literature when compared to the rate of overlap with the literature for 100 randomly-selected genes in the dataset subjected to the same protocol (Schizophrenia: 5/10 vs. 7/100, p=0.0012; Bipolar: 3/10 vs. 8/100, p=0.0610). Likewise, if we replaced diagnosis with a term representing the general presence or absence of a psychiatric illness, we found that four of the top 10 genes had been previously identified in the literature (ALDH1A1: (104); HBS1L: (4); HIVEP2: (111), FBX09: (74), **Suppl. Figure 25**, **Suppl. Table 9**), and 9/10 of the top genes were actually significant with an FDR<0.05 when using permutation based methods (using the R function lmp{lmPerm}, iterations=9999). The top 10 genes associated with psychiatric illness in models selected using forward/backward stepwise model selection (criterion=BIC) similarly included five that had been previously identified in the literature (PRSS16: (112), GRM1: (107,108); ALDH1A1: (104); SNAP25: (109); HIVEP2: (111), a significant improvement in overlap with the literature than what can be seen in 100 randomly-selected genes in the dataset subjected to the same protocol (Fisher's exact test: 5/10 vs.15/100, p=0.0168).

Together, we conclude that including cell content predictions in the analysis of macro-dissected microarray data can sometimes improve the sensitivity of the assay for detecting altered gene expression in relationship to psychiatric disease, especially if the dataset is confounded with dissection variation.



Top Genes Associated with Schizophrenia:

Eq.3: **Diagnosis + Confounds:**

Probe	Gene Symbol	Beta	Pval	FDR
11330_at	CTRC	-0.13	1.00E-04	4.75E-01
1758_at	DMP1	-0.06	1.37E-04	4.75E-01
10086_at	HHLA1	-0.40	1.70E-04	4.75E-01
23760_at	PITPNB	-0.13	1.96E-04	4.75E-01
55760_at	DHX32	-0.16	2.61E-04	4.75E-01
3397_at	ID1	-0.51	2.73E-04	4.75E-01
1414_at	CRYBB1	-0.12	3.04E-04	4.75E-01
7644_at	ZNF91	-0.29	3.26E-04	4.75E-01
26071_at	FAM127B	0.15	3.84E-04	4.75E-01
4878_at	NPPA	-0.17	3.98E-04	4.75E-01

Eq.6: **Diagnosis + 5 Prevalent Cell Types & Confounds**

Probe	Gene Symbol	Beta	Pval	FDR
9181_at	**ARHGEF2**	-0.12	3.96E-05	2.66E-01
8448_at	**DOC2A**	0.18	4.55E-05	2.66E-01
3397_at	ID1	-0.53	6.69E-05	2.66E-01
23760_at	PITPNB	-0.12	8.87E-05	2.66E-01
26268_at	**FBXO9**	-0.16	2.53E-04	4.48E-01
11330_at	CTRC	-0.10	3.12E-04	4.48E-01
81491_at	GPR63	0.12	4.28E-04	4.48E-01
2911_at	**GRM1**	0.07	4.71E-04	4.48E-01
55760_at	DHX32	-0.13	4.79E-04	4.48E-01
1050_at	**CEBPA**	0.15	5.70E-04	4.48E-01

Eq.1: **Diagnosis + All Cell Types & Confounds**

Probe	Gene Symbol	Beta	Pval	FDR
3397_at	ID1	-0.54	3.68E-05	2.22E-01
8448_at	**DOC2A**	0.17	6.26E-05	2.22E-01
9181_at	**ARHGEF2**	-0.12	6.78E-05	2.22E-01
23760_at	PITPNB	-0.13	7.41E-05	2.22E-01
5376_at	**PMP22**	-0.24	1.67E-04	3.64E-01
1414_at	CRYBB1	-0.10	2.34E-04	3.64E-01
4878_at	NPPA	-0.14	2.65E-04	3.64E-01
11330_at	CTRC	-0.10	2.68E-04	3.64E-01
23187_at	PHLDB1	-0.17	4.41E-04	3.64E-01
2263_at	**FGFR2**	-0.16	4.49E-04	3.64E-01

Top Genes Associated with Bipolar Disorder:

Eq.3: **Diagnosis + Confounds:**

Probe	Gene Symbol	Beta	Pval	FDR
4725_at	NDUFS5	-0.15	7.77E-04	1.00E+00
51042_at	ZNF593	0.16	1.20E-03	1.00E+00
79705_at	LRRK1	0.10	1.64E-03	1.00E+00
10146_at	G3BP1	0.13	1.71E-03	1.00E+00
26664_at	OR7C1	-0.08	1.85E-03	1.00E+00
4677_at	NARS	-0.08	1.89E-03	1.00E+00
2353_at	**FOS**	-0.63	2.00E-03	1.00E+00
23760_at	PITPNB	-0.10	2.22E-03	1.00E+00
9815_at	GIT2	-0.05	2.44E-03	1.00E+00
7404_at	UTY	0.04	2.87E-03	1.00E+00

Eq.6: **Diagnosis + 5 Prevalent Cell Types & Confounds**

Probe	Gene Symbol	Beta	Pval	FDR
216_at	**ALDH1A1**	-0.37	7.57E-05	9.06E-01
6616_at	**SNAP25**	-0.20	3.59E-04	1.00E+00
10146_at	G3BP1	0.14	7.61E-04	1.00E+00
4725_at	NDUFS5	-0.15	8.07E-04	1.00E+00
51042_at	ZNF593	0.16	1.05E-03	1.00E+00
4677_at	NARS	-0.08	1.07E-03	1.00E+00
8534_at	CHST1	0.21	1.09E-03	1.00E+00
23760_at	PITPNB	-0.10	1.11E-03	1.00E+00
81567_at	TXNDC5	0.14	1.33E-03	1.00E+00
51299_at	**NRN1**	-0.13	1.42E-03	1.00E+00

Eq.1: **Diagnosis + All Cell Types & Confounds**

Probe	Gene Symbol	Beta	Pval	FDR
216_at	**ALDH1A1**	-0.40	3.05E-05	2.21E-01
6616_at	**SNAP25**	-0.17	3.69E-05	2.21E-01
8534_at	CHST1	0.22	4.33E-04	9.98E-01
29896_at	TRA2A	-0.15	5.78E-04	9.98E-01
10146_at	G3BP1	0.14	6.58E-04	9.98E-01
90806_at	ANGEL2	-0.09	7.27E-04	9.98E-01
4677_at	NARS	-0.08	1.24E-03	9.98E-01
79705_at	LRRK1	0.10	1.33E-03	9.98E-01
23510_at	KCTD2	0.10	1.34E-03	9.98E-01
81567_at	TXNDC5	0.13	1.41E-03	9.98E-01

Top Genes Associated with MDD:

Eq.3: **Diagnosis + Confounds:**

Probe	Gene Symbol	Beta	Pval	FDR
23476_at	BRD4	0.12	7.10E-05	4.29E-01
5961_at	PRPH2	0.21	7.16E-05	4.29E-01
9862_at	MED24	0.15	2.08E-04	7.94E-01
10253_at	SPRY2	-0.21	3.20E-04	7.94E-01
10279_at	PRSS16	0.11	3.31E-04	7.94E-01
23493_at	HEY2	-0.15	6.04E-04	9.16E-01
9148_at	NEURL	0.11	6.40E-04	9.16E-01
79570_at	NKAIN1	0.11	1.15E-03	9.16E-01
23163_at	GGA3	0.09	1.59E-03	9.16E-01
139538_at	VENTXP1	-0.03	1.60E-03	9.16E-01

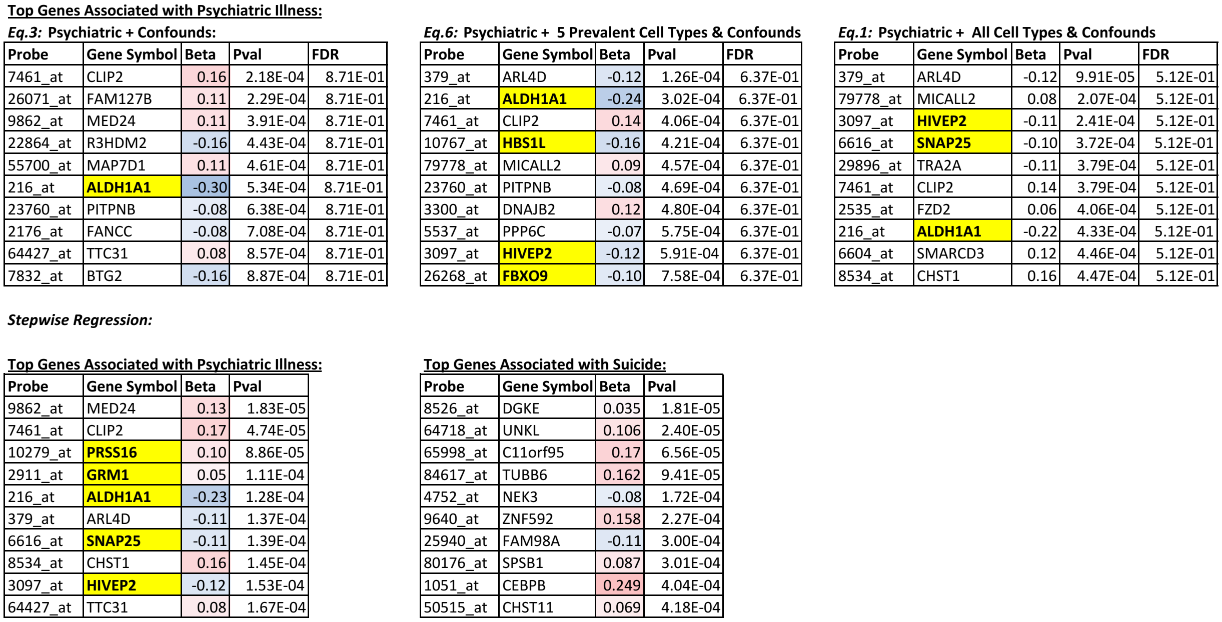
Eq.6: **Diagnosis + 5 Prevalent Cell Types & Confounds**

Probe	Gene Symbol	Beta	Pval	FDR
5961_at	PRPH2	0.21	6.96E-05	8.34E-01
23476_at	BRD4	0.11	2.12E-04	9.99E-01
8314_at	BAP1	0.11	2.77E-04	9.99E-01
10279_at	PRSS16	0.10	5.10E-04	9.99E-01
379_at	ARL4D	-0.13	7.57E-04	9.99E-01
9862_at	MED24	0.14	7.86E-04	9.99E-01
79570_at	NKAIN1	0.11	7.97E-04	9.99E-01
9985_at	REC8	0.12	8.57E-04	9.99E-01
2535_at	FZD2	0.08	9.54E-04	9.99E-01
3781_at	KCNN2	-0.14	1.19E-03	9.99E-01

Eq.1: **Diagnosis + All Cell Types & Confounds**

Probe	Gene Symbol	Beta	Pval	FDR
23476_at	BRD4	0.12	4.06E-05	4.86E-01
5961_at	PRPH2	0.20	1.20E-04	5.49E-01
2535_at	FZD2	0.08	1.37E-04	5.49E-01
8314_at	BAP1	0.11	4.30E-04	9.99E-01
9985_at	REC8	0.13	5.80E-04	9.99E-01
379_at	ARL4D	-0.13	9.74E-04	9.99E-01
9862_at	MED24	0.13	1.06E-03	9.99E-01
10279_at	PRSS16	0.09	1.29E-03	9.99E-01
10767_at	HBS1L	-0.18	1.33E-03	9.99E-01
79570_at	NKAIN1	0.10	1.40E-03	9.99E-01

Suppl. Figure 24. When analyzing the full dataset, the top genes associated with diagnosis in models that include cell content predictions include genes previously identified in the literature. *Depicted are the top 10 genes associated with diagnosis using three different models of increasing complexity, along with their β's (magnitude and direction of effect within the model – blue indicates downregulation, pink is upregulation), nominal p-values, and p-values that have been corrected for false detection rate using the Benjamini-Hochberg method. Gene symbols that are bolded and highlighted yellow have been previously detected in the human literature in association with their respective diagnosis in papers identified using the PubMed search terms "Schizophrenia" (Row 1) and "Bipolar" (Row 2). None of the top genes associated with major depressive disorder in any of the three models were found to be associated with "Depression", "Anxiety", or "Suicide" on PubMed (Row 3).*



Top Genes Associated with Psychiatric Illness:

Eq.3: Psychiatric + Confounds:

Probe	Gene Symbol	Beta	Pval	FDR
7461_at	CLIP2	0.16	2.18E-04	8.71E-01
26071_at	FAM127B	0.11	2.29E-04	8.71E-01
9862_at	MED24	0.11	3.91E-04	8.71E-01
22864_at	R3HDM2	-0.16	4.43E-04	8.71E-01
55700_at	MAP7D1	0.11	4.61E-04	8.71E-01
216_at	**ALDH1A1**	-0.30	5.34E-04	8.71E-01
23760_at	PITPNB	-0.08	6.38E-04	8.71E-01
2176_at	FANCC	-0.08	7.08E-04	8.71E-01
64427_at	TTC31	0.08	8.57E-04	8.71E-01
7832_at	BTG2	-0.16	8.87E-04	8.71E-01

Eq.6: Psychiatric + 5 Prevalent Cell Types & Confounds

Probe	Gene Symbol	Beta	Pval	FDR
379_at	ARL4D	-0.12	1.26E-04	6.37E-01
216_at	**ALDH1A1**	-0.24	3.02E-04	6.37E-01
7461_at	CLIP2	0.14	4.06E-04	6.37E-01
10767_at	**HBS1L**	-0.16	4.21E-04	6.37E-01
79778_at	MICALL2	0.09	4.57E-04	6.37E-01
23760_at	PITPNB	-0.08	4.69E-04	6.37E-01
3300_at	DNAJB2	0.12	4.80E-04	6.37E-01
5537_at	PPP6C	-0.07	5.75E-04	6.37E-01
3097_at	**HIVEP2**	-0.12	5.91E-04	6.37E-01
26268_at	**FBXO9**	-0.10	7.58E-04	6.37E-01

Eq.1: Psychiatric + All Cell Types & Confounds

Probe	Gene Symbol	Beta	Pval	FDR
379_at	ARL4D	-0.12	9.91E-05	5.12E-01
79778_at	MICALL2	0.08	2.07E-04	5.12E-01
3097_at	**HIVEP2**	-0.11	2.41E-04	5.12E-01
6616_at	**SNAP25**	-0.10	3.72E-04	5.12E-01
29896_at	TRA2A	-0.11	3.79E-04	5.12E-01
7461_at	CLIP2	0.14	3.79E-04	5.12E-01
2535_at	FZD2	0.06	4.06E-04	5.12E-01
216_at	**ALDH1A1**	-0.22	4.33E-04	5.12E-01
6604_at	SMARCD3	0.12	4.46E-04	5.12E-01
8534_at	CHST1	0.16	4.47E-04	5.12E-01

Stepwise Regression:

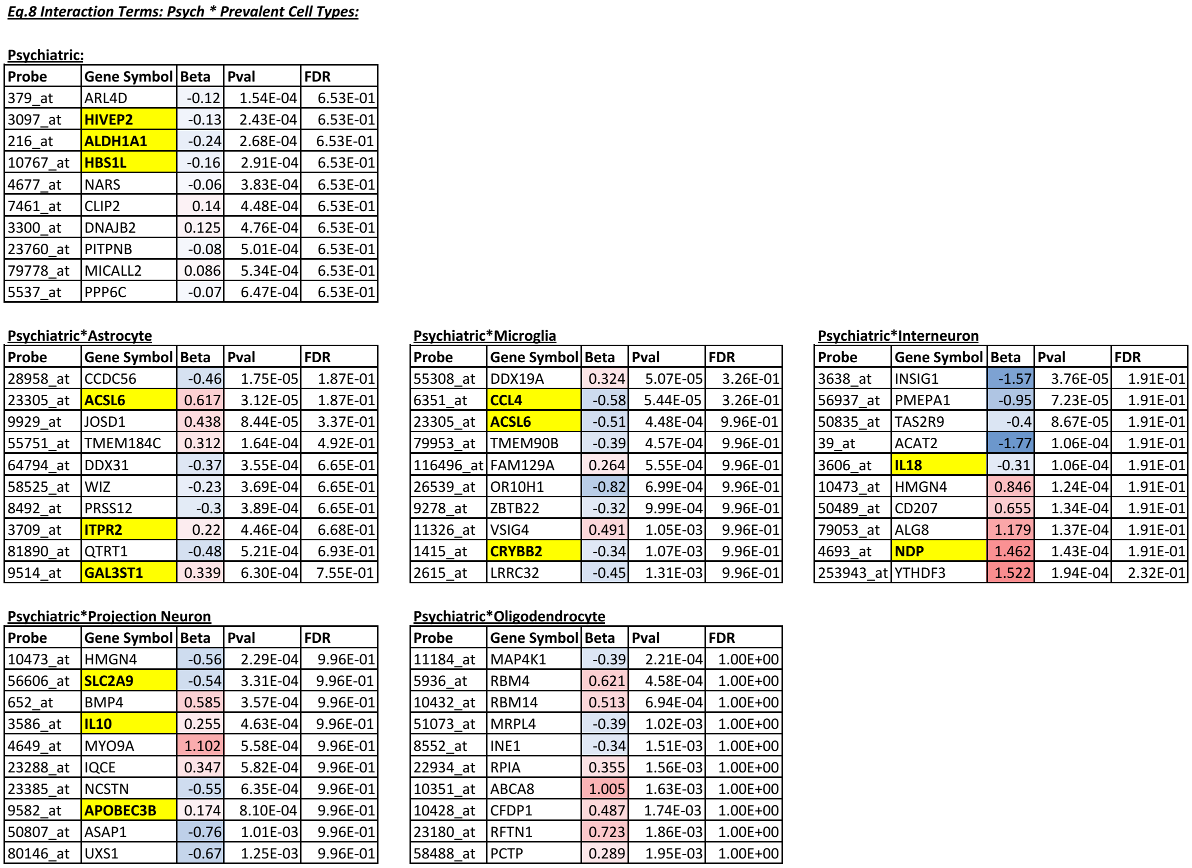
Top Genes Associated with Psychiatric Illness:

Probe	Gene Symbol	Beta	Pval
9862_at	MED24	0.13	1.83E-05
7461_at	CLIP2	0.17	4.74E-05
10279_at	**PRSS16**	0.10	8.86E-05
2911_at	**GRM1**	0.05	1.11E-04
216_at	**ALDH1A1**	-0.23	1.28E-04
379_at	ARL4D	-0.11	1.37E-04
6616_at	**SNAP25**	-0.11	1.39E-04
8534_at	CHST1	0.16	1.45E-04
3097_at	**HIVEP2**	-0.12	1.53E-04
64427_at	TTC31	0.08	1.67E-04

Top Genes Associated with Suicide:

Probe	Gene Symbol	Beta	Pval
8526_at	DGKE	0.035	1.81E-05
64718_at	UNKL	0.106	2.40E-05
65998_at	C11orf95	0.17	6.56E-05
84617_at	TUBB6	0.162	9.41E-05
4752_at	NEK3	-0.08	1.72E-04
9640_at	ZNF592	0.158	2.27E-04
25940_at	FAM98A	-0.11	3.00E-04
80176_at	SPSB1	0.087	3.01E-04
1051_at	CEBPB	0.249	4.04E-04
50515_at	CHST11	0.069	4.18E-04

Suppl. Figure 25. When analyzing the full dataset, the top genes associated with psychiatric illness in models that include cell content predictions include genes previously identified in the literature. *Depicted are the top 10 genes associated with psychiatric illness using three different models of increasing complexity, or associated with psychiatric illness or suicide in models chosen using stepwise regression. Notably, the results from stepwise regression for the diagnosis term are not included in this figure because the term was only included in the model for eight genes total (DHX32, ID1, CSRP1, AKR1B10, TBPL1, HIST1H4F, SETD3, GAL). Included are the β's (magnitude and direction of effect within the model – blue indicates downregulation, pink is upregulation), nominal p-values, and p-values that have been corrected for false detection rate using the Benjamini-Hochberg method. Note that the p-values associated with stepwise regression are likely to be optimistic due to overfitting. Gene symbols that are bolded and highlighted yellow have been previously detected in the human literature in association with their respective diagnosis in papers identified using the PubMed search terms "Schizophrenia", "Bipolar", "Depression", "Anxiety", or "Suicide".*



Eq.8 Interaction Terms: Psych * Prevalent Cell Types:

Psychiatric:

Probe	Gene Symbol	Beta	Pval	FDR
379_at	ARL4D	-0.12	1.54E-04	6.53E-01
3097_at	**HIVEP2**	-0.13	2.43E-04	6.53E-01
216_at	**ALDH1A1**	-0.24	2.68E-04	6.53E-01
10767_at	**HBS1L**	-0.16	2.91E-04	6.53E-01
4677_at	NARS	-0.06	3.83E-04	6.53E-01
7461_at	CLIP2	0.14	4.48E-04	6.53E-01
3300_at	DNAJB2	0.125	4.76E-04	6.53E-01
23760_at	PITPNB	-0.08	5.01E-04	6.53E-01
79778_at	MICALL2	0.086	5.34E-04	6.53E-01
5537_at	PPP6C	-0.07	6.47E-04	6.53E-01

Psychiatric*Astrocyte

Probe	Gene Symbol	Beta	Pval	FDR
28958_at	CCDC56	-0.46	1.75E-05	1.87E-01
23305_at	**ACSL6**	0.617	3.12E-05	1.87E-01
9929_at	JOSD1	0.438	8.44E-05	3.37E-01
55751_at	TMEM184C	0.312	1.64E-04	4.92E-01
64794_at	DDX31	-0.37	3.55E-04	6.65E-01
58525_at	WIZ	-0.23	3.69E-04	6.65E-01
8492_at	PRSS12	-0.3	3.89E-04	6.65E-01
3709_at	**ITPR2**	0.22	4.46E-04	6.68E-01
81890_at	QTRT1	-0.48	5.21E-04	6.93E-01
9514_at	**GAL3ST1**	0.339	6.30E-04	7.55E-01

Psychiatric*Microglia

Probe	Gene Symbol	Beta	Pval	FDR
55308_at	DDX19A	0.324	5.07E-05	3.26E-01
6351_at	**CCL4**	-0.58	5.44E-05	3.26E-01
23305_at	**ACSL6**	-0.51	4.48E-04	9.96E-01
79953_at	TMEM90B	-0.39	4.57E-04	9.96E-01
116496_at	FAM129A	0.264	5.55E-04	9.96E-01
26539_at	OR10H1	-0.82	6.99E-04	9.96E-01
9278_at	ZBTB22	-0.32	9.99E-04	9.96E-01
11326_at	VSIG4	0.491	1.05E-03	9.96E-01
1415_at	**CRYBB2**	-0.34	1.07E-03	9.96E-01
2615_at	LRRC32	-0.45	1.31E-03	9.96E-01

Psychiatric*Interneuron

Probe	Gene Symbol	Beta	Pval	FDR
3638_at	INSIG1	-1.57	3.76E-05	1.91E-01
56937_at	PMEPA1	-0.95	7.23E-05	1.91E-01
50835_at	TAS2R9	-0.4	8.67E-05	1.91E-01
39_at	ACAT2	-1.77	1.06E-04	1.91E-01
3606_at	**IL18**	-0.31	1.06E-04	1.91E-01
10473_at	HMGN4	0.846	1.24E-04	1.91E-01
50489_at	CD207	0.655	1.34E-04	1.91E-01
79053_at	ALG8	1.179	1.37E-04	1.91E-01
4693_at	**NDP**	1.462	1.43E-04	1.91E-01
253943_at	YTHDF3	1.522	1.94E-04	2.32E-01

Psychiatric*Projection Neuron

Probe	Gene Symbol	Beta	Pval	FDR
10473_at	HMGN4	-0.56	2.29E-04	9.96E-01
56606_at	**SLC2A9**	-0.54	3.31E-04	9.96E-01
652_at	BMP4	0.585	3.57E-04	9.96E-01
3586_at	**IL10**	0.255	4.63E-04	9.96E-01
4649_at	MYO9A	1.102	5.58E-04	9.96E-01
23288_at	IQCE	0.347	5.82E-04	9.96E-01
23385_at	NCSTN	-0.55	6.35E-04	9.96E-01
9582_at	**APOBEC3B**	0.174	8.10E-04	9.96E-01
50807_at	ASAP1	-0.76	1.01E-03	9.96E-01
80146_at	UXS1	-0.67	1.25E-03	9.96E-01

Psychiatric*Oligodendrocyte

Probe	Gene Symbol	Beta	Pval	FDR
11184_at	MAP4K1	-0.39	2.21E-04	1.00E+00
5936_at	RBM4	0.621	4.58E-04	1.00E+00
10432_at	RBM14	0.513	6.94E-04	1.00E+00
51073_at	MRPL4	-0.39	1.02E-03	1.00E+00
8552_at	INE1	-0.34	1.51E-03	1.00E+00
22934_at	RPIA	0.355	1.56E-03	1.00E+00
10351_at	ABCA8	1.005	1.63E-03	1.00E+00
10428_at	CFDP1	0.487	1.74E-03	1.00E+00
23180_at	RFTN1	0.723	1.86E-03	1.00E+00
58488_at	PCTP	0.289	1.95E-03	1.00E+00

Suppl. Figure 26. When analyzing the full dataset using a model that includes Psychiatric Illness*Cell Type interaction terms, the top genes associated with psychiatric illness include genes previously identified in the literature. *Depicted are the top 10 genes associated with psychiatric illness and its interaction with the five most prevalent cell types in the cortex using the model in* Equation 6*:*

***Equation** 7*:

$$Gene\ Expression = \beta 0 + \beta 1*(Astrocyte\ Index) + \\ +\beta 2*(Microglia\ Index) + \beta 3*(Neuron\_Interneuron\ Index) + \beta 4*(Neuron\_Projection\ Neuron\ Index) + \beta 5*(Oligodendrocyte\ Index) + \beta 6*(Brain\ pH) + \beta 7*(Agonal\ Factor) + \beta 8*(PMI) + \beta 9*(Age) + \beta 10*(Gender) + \beta 11*(Psychiatric\ Illness) + \beta 12*(Psychiatric\ Illness)*(Astrocyte\ Index) + \beta 13*(Psychiatric\ Illness)*(Microglia\ Index) + \beta 14*(Psychiatric\ Illness)*(Neuron\_Interneuron\ Index) + \beta 15*(Psychiatric\ Illness)*(Neuron\_Projection\ Neuron\ Index) + \beta 17*(Psychiatric\ Illness)*(Oligodendrocyte\ Index) + \varepsilon$$

Included are the β's (magnitude and direction of effect within the model – blue indicates downregulation, pink is upregulation), nominal p-values, and p-values that have been corrected for false detection rate using the Benjamini-Hochberg method. The number of top genes that were found to be previously-

identified in literature does not significantly surpass what was observed in a group of 100 randomly selected genes from our dataset (14/60 vs. 15/100).

8. Supplementary Tables

Suppl. Table 1. Master Database of Cortical Cell Type Specific Gene Expression. *The attached excel document contains a single spreadsheet listing the genes defined as having cell type specific expression in our manuscript, including the species, age of the subjects, and brain region from which the cells were purified, the platform used to measure transcript, the statistical criteria and comparison cell types used to define "cell type specific expression", the gene symbol or orthologous gene symbol in mouse/human (depending on the species used in the original experiment), and citation. If a gene was identified as having cell type specific expression in multiple experiments, there is an entry for each experiment – thus the full 3383 rows included in the spreadsheet do not represent 3383 individual cell type specific genes. A web-version of this spreadsheet kept interactively up-to-date can be found at https://sites.google.com/a/umich.edu/megan-hastings-hagenauer/home/cell-type-analysis.*

Suppl. Table 2. Microarray data spanning 160 human brain regions downloaded from the Allen Brain Atlas. *Included in this excel file are three worksheets. The first includes all of the sample information, including the subject identifier and brain region. The second includes all of the probe information. Finally, the third includes the relative expression for each probe for each sample (z-score), including the official gene symbol, Entrez gene ID, and gene name. Additional information about the human microarray dataset can be found on the Allen Brain Atlas website.*

Suppl. Table 3. The average cell type indices for all 160 brain regions included in the Allen Brain Atlas dataset. *This excel file contains two worksheets. The first includes the average cell type index for 10 primary cell types for all 160 brain regions included in the Allen Brain Atlas. More detail about those brain regions can be found in the first worksheet (Columns_Sample Info) in* ***Suppl. Table 2****. The second spreadsheet contains the standard error (SE) for the averages in the first worksheet.*

Suppl. Table 4. Output for the analyses of cell type vs. subject variables for all datasets. *The first spreadsheet provides the output from the meta-analysis for each cell type vs. subject variable combination ("b"= the estimated effect, provided in the units for the variable – e.g., the effect of one year of age, or the effect of one hour of PMI; "SE"= standard error,"p-value"= nominal p-value, "BH_adj_P-value (q-value)"= the p-value corrected for multiple comparisons). The second spreadsheet includes the T-statistics for all cell type vs. subject variable combinations for all datasets.*

Suppl. Table 5. Functions associated with genes identified as having neuron-specific expression. *The first column of the excel spreadsheet is a list of general physiological functions that were identified by DAVID as associated with our list of neuron-specific genes (relative to the full list of probesets included in the microarray). We used the functional cluster option in DAVID because it prevents multiple functions that share a large subset of overlapping genes from dominating the results. We named each cluster by the top two functions included in it. The second column of the spreadsheet indicates whether an experimenter blindly categorized the functional cluster as being clearly related or unrelated to synaptic function. The "Mean Fold Enrichment" column indicates how well on average each of the functions within that cluster were associated with our list of neuron-specific genes. The next three columns (Top p-value, Top Bonferronni-corrected p-value, and top BH (Benjamini-Hochberg)-corrected p-value) indicate the statistical strength of the association between the top function within that cluster and our list of neuron-specific genes. The number of genes from each functional cluster included in our results is listed in column G. The next few columns indicate the strength of the relationship between the functional cluster and age. Columns H-J indicate the mean, standard deviation, and standard error, for the betas for Age for each gene included in the cluster. The betas indicate the strength and direction of the association with Age as determined within a larger linear model controlling for known confounds (pH, PMI, gender, agonal factor). Columns K-M indicate whether, on average, the age-related betas for the genes in that cluster are statistically different from 0 as determined by a Welch's t-test (t-stat, df, p-value). The final* *column indicates what percentage of the genes included in the cluster have a negative relationship (β) with age.*

Suppl. Table 6. A .gmt file created using our database of cell type specific genes for use with Gene Set Enrichment Analysis (GSEA). *This file should be in the correct format for usage with either GSEA (http://software.broadinstitute.org/gsea/index.jsp) or fGSEA.*

Suppl. Table 7. Performing Gene Set Enrichment Analysis using a .gmt that includes traditional functional gene sets and cell type specific gene lists indicates that cell type specific gene sets are enriched for effects related to a wide variety of subject variables. *Gene set enrichment analysis was performed using the results from a differential expression analysis performed on the Pritzker dataset using a model that included diagnosis, pH, agonal factor, age, PMI, and sex. The gene set enrichment results for each variable is included as its own worksheet in the file.*

Suppl. Table 8. Previously-identified relationships between gene expression and psychiatric illness in the human cortex in either particular cell types or macro-dissected cortex. *We used this database of previously-identified effects to determine whether controlling for cell type while performing differential expression analyses increased our ability to observe previously-documented effects.*

Suppl. Table 9. The relationship between diagnosis and all probesets in the Pritzker Dorsolateral Prefrontal Cortex dataset as assessed using models of increasing complexity. *For all probesets in the dataset, the spreadsheets for Model #2 and Model#4 include the β for all variables in the model ("Beta": magnitude and direction of the association, with positive associations labeled pink and negative associations labeled blue), the p-value ("Pval_nominal") and the p-value adjusted for multiple comparisons using the Benjamini-Hochberg method ("BH_Adj"), both labeled with green indicating more significant relationships and red indicating less significant relationships. There are also summary spreadsheets that include just the results for Bipolar Disorder and Schizophrenia for Models#1-5. In these spreadsheets, the formatting is a little different: T-statistics are provided, the β is called "LogFC", the BH_Adj p-value is called"adj.P.Val".*

Suppl. Table 10. The relationship between diagnosis and all genes in the CMC RNA-Seq dataset as assessed using models of increasing complexity. *There are two summary spreadsheets that include the results for Bipolar Disorder and Schizophrenia for Models#1-5. For all genes in the dataset, each spreadsheet includes the β ("LogFC": magnitude and direction of the association), the T-statistic, the p-value ("Pval_nominal") and the p-value adjusted for multiple comparisons using the Benjamini-Hochberg method ("adj.P.Val") for the effect of diagnosis in each model (#M1-M5).*

Suppl. Table 11. The relationship between diagnosis and all probesets in the Barnes et al. microarray dataset as assessed using models of increasing complexity. *There is a summary spreadsheet that includes the results for Schizophrenia for Models#1-5. For all probesets in the dataset, each spreadsheet includes the β ("LogFC": magnitude and direction of the association), the T-statistic, the p-value ("Pval_nominal") and the p-value adjusted for multiple comparisons using the Benjamini-Hochberg method ("adj.P.Val") for the effect of diagnosis in each model (#M1-M5).*

Suppl. Table 12. The relationship between diagnosis and all probesets in the Lanz et al. microarray dataset as assessed using models of increasing complexity. *There are two summary spreadsheets that include the results for Bipolar Disorder and Schizophrenia for Models#1-5. For all probesets in the dataset, each spreadsheet includes the β ("LogFC": magnitude and direction of the association), the T-statistic, the p-value ("Pval_nominal") and the p-value adjusted for multiple comparisons using the Benjamini-Hochberg method ("adj.P.Val") for the effect of diagnosis in each model (#M1-M5).*

Suppl. Table 13. The relationship between diagnosis and all probesets in the Narayan et al. microarray dataset as assessed using models of increasing complexity. *There is a summary spreadsheet that includes the results for Schizophrenia for Models#1-5. For all probesets in the dataset, each spreadsheet includes the β ("LogFC": magnitude and direction of the association), the T-statistic, the p-value ("Pval_nominal") and the p-value adjusted for multiple comparisons using the Benjamini-Hochberg method ("adj.P.Val") for the effect of diagnosis in each model (#M1-M5).*

Suppl. Table 14. Sample demographics for the Pritzker Consortium Dorsolateral Prefrontal Cortex Affymetrix microarray data.

Suppl. Table 15. The relationship between each cell type index and all probes in the Pritzker Dorsolateral Prefrontal Cortex dataset. *The attached excel document (.xlsx) contains multiple spreadsheets. The first spreadsheet ("Methods") contains a brief summary of the methods used to evaluate the relationship between the cell type indices and expression of each probe in the dataset (also discussed in the body of the manuscript). The second spreadsheet ("GeneByCellType_DF") contains the statistical output associated with all cell type index terms in the linear model for all probes in the dataset, including the β ("Beta": magnitude and direction of the association, with positive associations labeled pink and negative associations labeled blue), the p-value from the original model ("Pval") and the p-value adjusted for multiple comparisons using the Benjamini-Hochberg method ("AdjP"), both labeled with green indicating more significant relationships and red indicating less significant relationships. All other spreadsheets contain the top 100 probes positively associated with each cell type index, including each of the statistical outputs presented in the full "GeneByCellType_DF" summary spreadsheet, as well as a column "CellTypeSpecific" which indicates whether the probe was included in one of the original cell type indices (1=included, 0=not included).*

Suppl. Table 16. The relationship between each cell type index and all probes in the Allen Brain Atlas dataset. *Depicted are the β (magnitude and direction) and p-values for the relationship between the expression for each probe and each primary cell type across samples from all 160 brain regions as determined in a large linear model that includes all 10 primary cell types. Please note that the p-values in this spreadsheet have not been corrected for multiple comparisons. Additional information about the probes can be found in* ***Suppl. Table 2****.*