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], Jun Z. Li, Ph.D.[2], David M. Walsh, Psy.D.[3], Marquis P. Vawter, Ph.D.[3] Robert C. Thompson, Ph.D.[1], Cortney A. Turner, Ph.D.[1], William E. Bunney, M.D.[3], Richard M. Myers, Ph.D.[4], Jack D. Barchas, M.D.[5], Alan F. Schatzberg, M.D.[6], 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]Genet., Univ. of Michigan, Ann Arbor, MI, USA; [3]Univ. of California, Irvine, CA; [4]HudsonAlpha Inst. for Biotech., Huntsville, AL, USA; [5]Stanford, Palo Alto, CA, [6]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

Most neuroscientists would agree that psychiatric illness is unlikely to arise from pathological changes that occur uniformly across all cells in a given brain region. Despite this fact, the majority of transcriptomic analyses of the human brain to date are conducted using macro-dissected tissue due to the difficulty of conducting single-cell level analyses on 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 brain cell types identified in published single cell type transcriptomic experiments. Using this database, we predicted the relative cell type composition for 157 human dorsolateral prefrontal cortex samples using Affymetrix microarray data collected by the Pritzker Neuropsychiatric Consortium, as well as for 841 samples spanning 160 brain regions included in an Agilent microarray dataset collected by the Allen Brain Atlas. These predictions were generated by averaging normalized expression levels across the transcripts specific to each primary cell type to create a "cell type index". Using this method, we determined that the expression of cell type specific transcripts identified by different experiments, methodologies, and species clustered into three main cell type groups: neurons, oligodendrocytes, and astrocytes/support cells. Overall, the principal components of variation in the data were largely explained by the neuron to glia ratio of the samples. When comparing across brain regions, we were able to statistically identify canonical cell type signatures – increased endothelial cells and vasculature in the choroid plexus, oligodendrocytes in the corpus callosum, astrocytes in the central glial substance, neurons and immature cells in the dentate gyrus, and oligodendrocytes and interneurons in the globus pallidus.

The relative balance of these cell types was influenced by a variety of demographic, pre- and post-mortem variables. Age and prolonged hypoxia around the time of death were associated with decreased neuronal content and increased astrocytic and endothelial content in the tissue, replicating the known higher vulnerability of neurons to adverse conditions and illustrating the proliferation of vasculature in a hypoxic environment. We also found that the red blood cell content was reduced in individuals who died in a manner that involved systemic blood loss. Finally, statistically accounting for cell type improved both the sensitivity and interpretability of diagnosis effects within the data. We were able to observe a decrease in astrocytic content in subjects with Major Depressive Disorder, mirroring what had been previously observed morphometrically. By including a set of “cell type indices” in a larger model examining the relationship between gene expression and neuropsychiatric illness, we were able to successfully detect almost twice as many genes with previously-identified relationships to bipolar disorder and schizophrenia than using more traditional analysis methods.

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 dissect larger tissue samples in order to extract sufficient transcript or protein to perform full 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 high quality post-mortem tissue is precious, especially tissue donated by the families of individuals with rare psychiatric or neurological disorders.

Therefore, to date, the vast majority of unbiased transcriptomic analyses of the human brain have been conducted using macro-dissected, cell-type heterogeneous tissue. They 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 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 may be less prevalent in the aging human brain, but are widely regarded as important for the pathogenesis of mood disorders (23), and thus were also included in our analysis. 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.

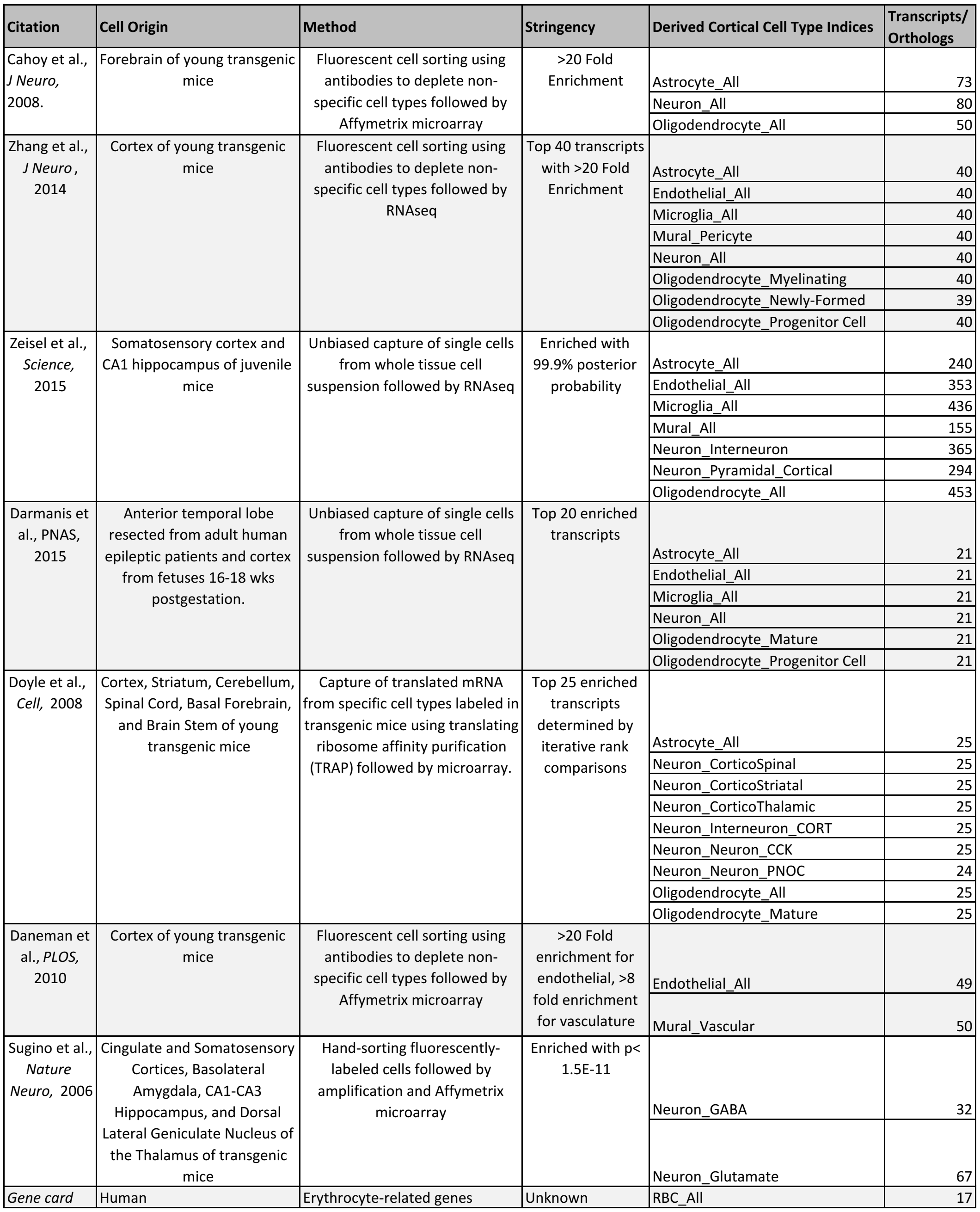
To characterize the balance of these cell types in psychiatric samples, we first compare the predictive value of cell type specific transcripts identified by diverse data sources and then summarize their collective predictions of relative cell type balance into covariates that can be used in larger linear regression models. We demonstrate that statistically estimating the relative cell type balance of samples can explain a large percentage of the variation in human brain microarray datasets. We also find that the incorporation of a set of “cell type indices” into a larger regression model can successfully predict other cell type-enriched gene expression as well as known changes in cell type balance in response to age, aerobic environment, large scale blood loss, and 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 increasing our sensitivity to detect genes with previously-identified relationships to bipolar disorder and schizophrenia.

2. Results

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 for 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) and neurons in general. 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 brain.

Most of the cell-type specific transcripts were derived from microarray experiments using cDNA extracted from laboratory mice, therefore in order to use this information for the analysis of human microarray data it was necessary to identify the respective orthologs for the cell type specific transcripts in humans using HCOP: Orthology Prediction Search (http://www.genenames.org/cgi-bin/hcop). Our final database included 2499 unique human-derived or orthologous transcripts, with a focus on coding varieties**.**



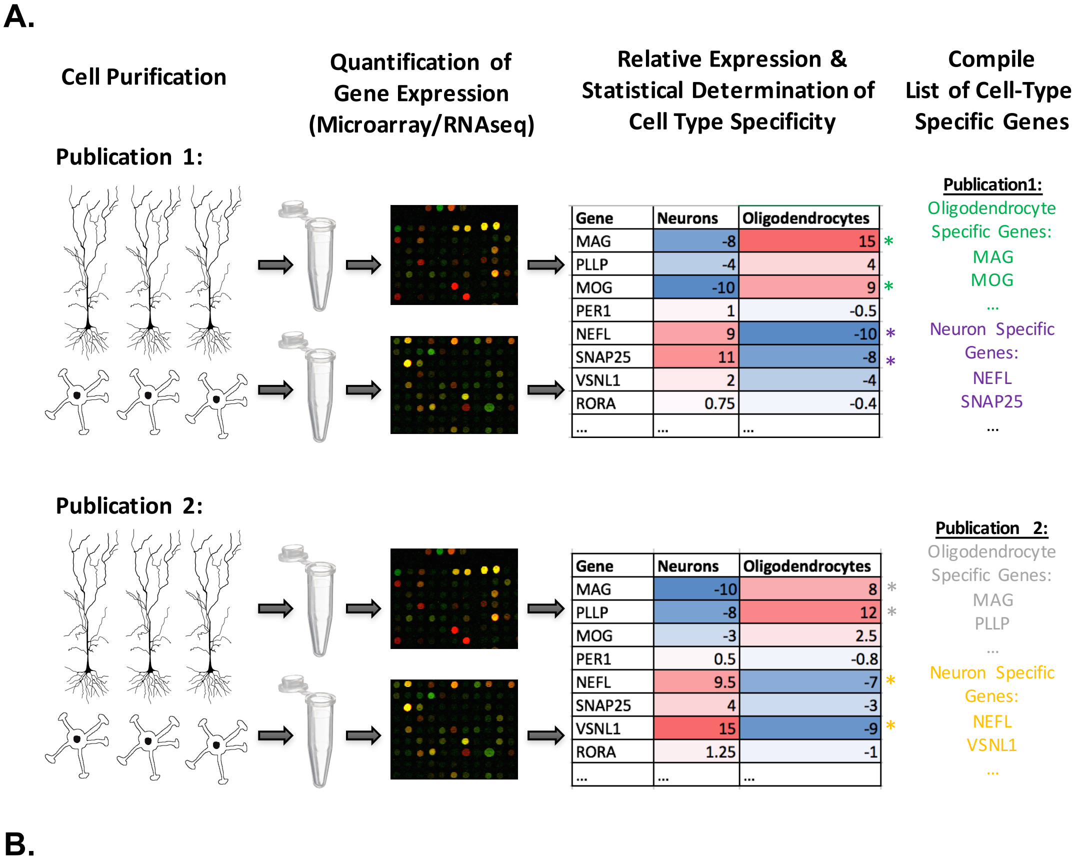
Citation	Cell Origin	Method	Stringency	Derived Cortical Cell Type Indices	Transcripts/ Orthologs
Cahoy et al., *J Neuro,* 2008.	Forebrain of young transgenic mice	Fluorescent cell sorting using antibodies to deplete non-specific cell types followed by Affymetrix microarray	>20 Fold Enrichment	Astrocyte_All	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. 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 Using Cell Type Specific Transcripts to Predict Relative Cell Content in Microarray Data from Macro-Dissected Human Dorsolateral Prefrontal Cortex Tissue

Next, we examined the collective variation in the levels of cell type specific transcripts in an Affymetrix microarray dataset from 157 high-quality human post-mortem dorsolateral prefrontal cortex samples (**Suppl. Table 2**), 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; (25,26)). Additionally, we measured the pH of cerebellar tissue as an indicator of the extent of oxygen deprivation experienced around the time of death (25,26) 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.

To predict the relative cell content in each of the samples, we used a technique validated using datasets from purified cell types and artificial cell mixtures (**Supplementary Methods and Results, Suppl. Figs 1-4**). We identified 2678 gene probe sets in the Affymetrix dataset that were found in our curated database of cell type specific transcripts as matched by official gene symbol. We centered and scaled the expression level of each gene probeset across samples (mean=0, sd=1) to prevent probe sets with more variable hybridization 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"* , Figure 1)**, each of which predicted the relative content for one of the 10 primary cell types in our cortical samples (**Figure 2).**


A.
Cell Purification
Quantification of Gene Expression (Microarray/RNAseq)
Relative Expression & Statistical Determination of Cell Type Specificity
Compile List of Cell-Type Specific Genes
Publication 1:
Publication 2:
B.

Publication 1:

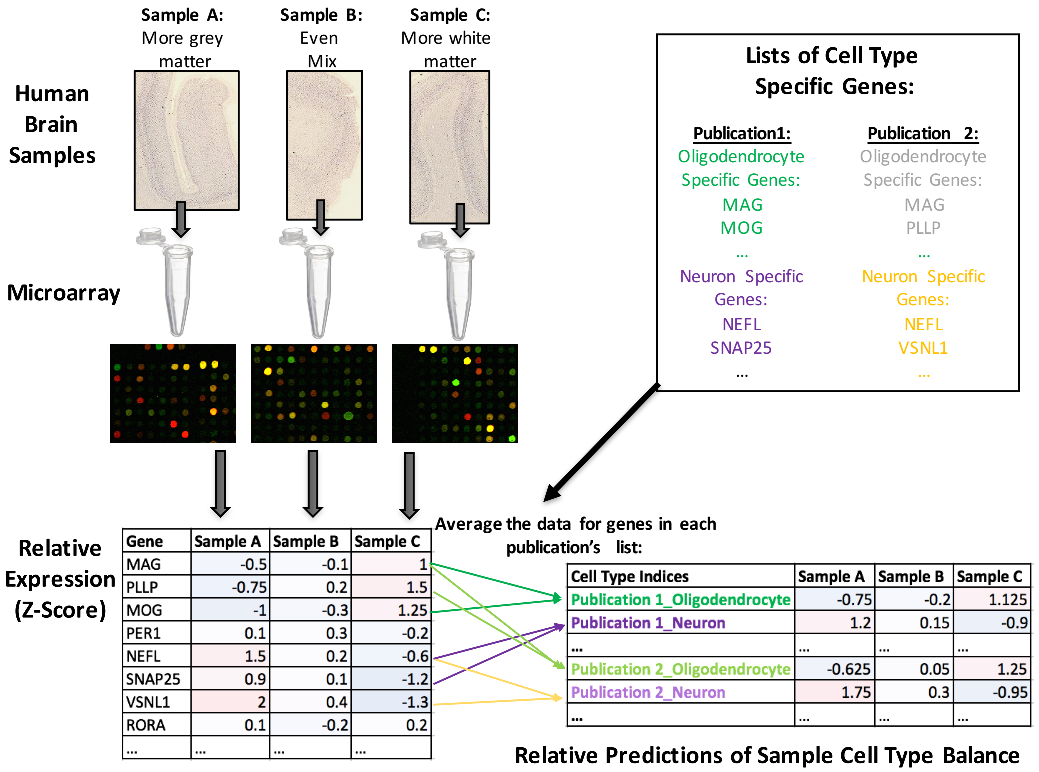
Gene	Neurons	Oligodendrocytes	
MAG	-8	15	*
PLLP	-4	4	
MOG	-10	9	*
PER1	1	-0.5	
NEFL	9	-10	*
SNAP25	11	-8	*
VSNL1	2	-4	
RORA	0.75	-0.4	
...	...	...	

Publication1:
Oligodendrocyte Specific Genes:
MAG
MOG
...
Neuron Specific Genes:
NEFL
SNAP25
...

Publication 2:

Gene	Neurons	Oligodendrocytes	
MAG	-10	8	*
PLLP	-8	12	*
MOG	-3	2.5	
PER1	0.5	-0.8	
NEFL	9.5	-7	*
SNAP25	4	-3	
VSNL1	15	-9	*
RORA	1.25	-1	
...	...	...	

Publication 2:
Oligodendrocyte Specific Genes:
MAG
PLLP
...
Neuron Specific Genes:
NEFL
VSNL1
...


Sample A:
More grey matter
Sample B:
Even Mix
Sample C:
More white matter
Human Brain Samples
Microarray
Lists of Cell Type Specific Genes:
Publication1:
Oligodendrocyte Specific Genes:
MAG
MOG
...
Neuron Specific Genes:
NEFL
SNAP25
...
Publication 2:
Oligodendrocyte Specific Genes:
MAG
PLLP
...
Neuron Specific Genes:
NEFL
VSNL1
...
Relative Expression (Z-Score)
Average the data for genes in each publication's list:
Relative Predictions of Sample Cell Type Balance

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. A. We compiled a database of genes that were identified in previous publications as specifically enriched in particular forebrain or cortical cell types. Within these publications, researchers purified particular brain cell types or individual brain cells and transcriptionally profiled them using microarray or RNA-Seq. They then performed statistical comparisons across the purified brain cell types to determine which transcripts showed relatively elevated expression in each cell type. **B.** Within the Pritzker brain tissue samples, we expected variable cell type balance that would influence the pattern of gene expression measured by microarray. To estimate this variability, we pulled out the microarray data for probe sets representing genes that had been previously identified as having cell type specific expression and then averaged

across the transcripts identified in each publication **(Figure 1)** as specific to a particular cell type to create 38 different ***"Cell Type Indices"*** that predicted relative cell content in each of the cortical samples.

2.3 There is a Strong Convergence of Cell Content Predictions Derived from Cell Type Specific Transcripts Originating from Different Publications

We found that the predicted cell content of the prefrontal cortex samples was relatively similar regardless of the origin of the cell type specific gene lists used to create the predictions. When comparing the pattern of correlations between the 38 cell type indices, they clearly cluster into three large umbrella categories: Neurons, Oligodendrocytes, and Support Cells (Astrocytes, Microglia, and Neurovasculature*)* even when the cell type signatures were derived from cell type specific gene lists from different source publications, species, and methodologies. This clustering was clear using visual inspection of the correlation matrix (**Figure 3)**, hierarchical clustering, or consensus clustering (**Suppl. Figure 5;** ConsensusClusterPlus: (27))**.** Moreover, the clustering was not due to the different publications identifying a similar subset of cell-type specific genes, because the clustering persisted in a follow-up analysis in which data from genes identified as cell type specific in multiple publications (e.g., Cahoy_Astrocyte and Zhang_Astrocyte) were removed list wise from the dataset **(Suppl. Figure 6 & 7).** Clustering was not able to reliably discern neuronal subcategories (interneurons, projection neurons) or support cell subcategories. Oligodendrocyte progenitor cell indices derived from different publications did not strongly correlate with each other, which may indicate a lack of significant presence of progenitor cells in the cortex of our primarily middle-aged subjects.

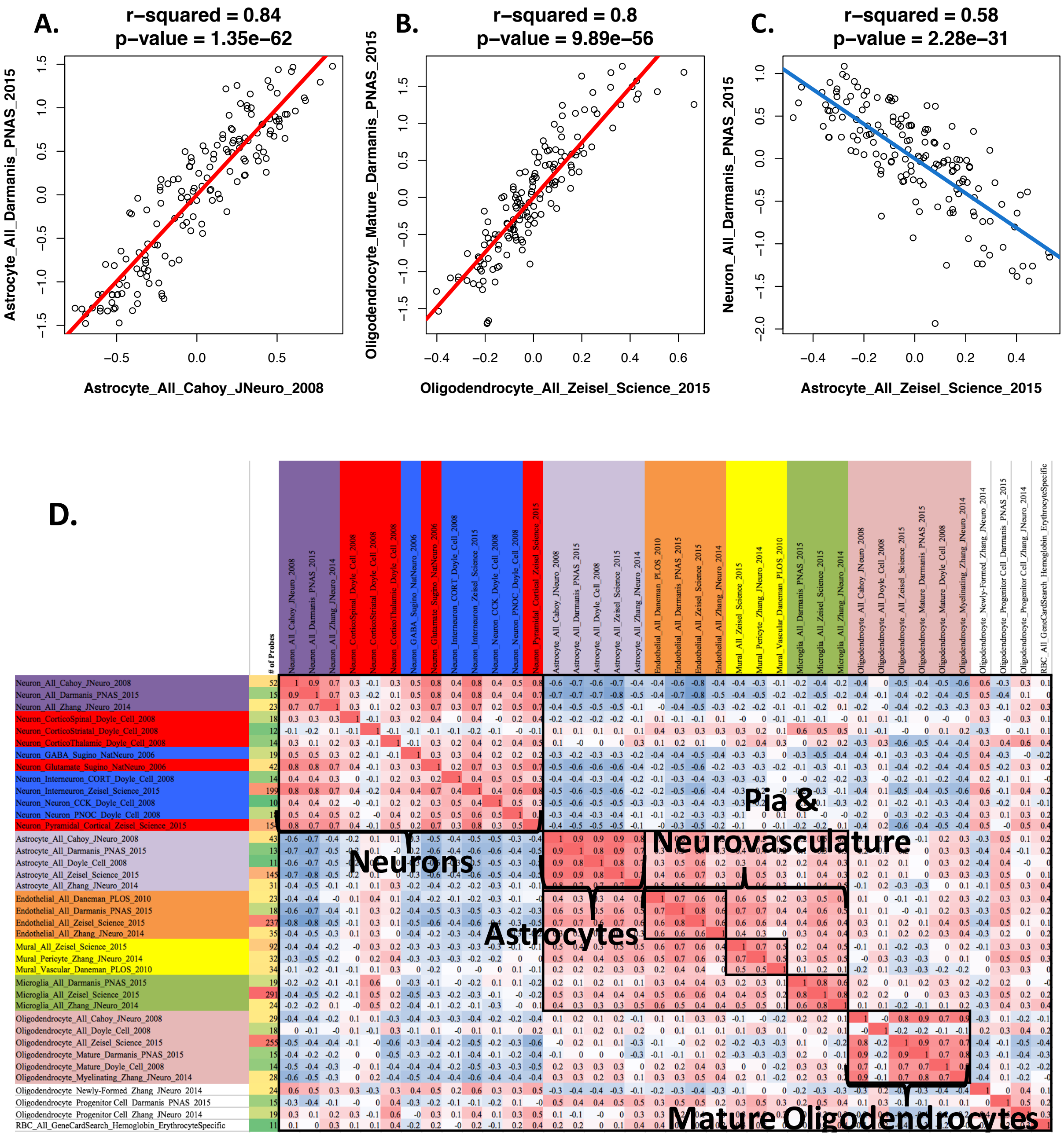

A.
r–squared = 0.84
p–value = 1.35e–62
Astrocyte_All_Darmanis_PNAS_2015
Astrocyte_All_Cahoy_JNeuro_2008
B.
r–squared = 0.8
p–value = 9.89e–56
Oligodendrocyte_Mature_Darmanis_PNAS_2015
Oligodendrocyte_All_Zeisel_Science_2015
C.
r–squared = 0.58
p–value = 2.28e–31
Neuron_All_Darmanis_PNAS_2015
Astrocyte_All_Zeisel_Science_2015
D.
# of Probes
Neurons
Astrocytes
Pia & Neurovasculature
Mature Oligodendrocytes
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
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
Mural_All_Zeisel_Science_2015
Mural_Pericyte_Zhang_JNeuro_2014
Mural_Vascular_Daneman_PLOS_2010
Microglia_All_Darmanis_PNAS_2015
Microglia_All_Zeisel_Science_2015
Microglia_All_Zhang_JNeuro_2014
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

Figure 3. There is a convergence of cell content predictions derived from cell type specific transcripts originating from different publications. A-B. There was a strong positive correlation between predictions originating from *"cell type indices"* representing the same cell type (e.g., neurons vs. neurons, oligodendrocytes vs. oligodendrocytes…) in our dorsolateral prefrontal cortical samples, even when the predictions were based on cell type specific transcripts identified by experiments using very different methodology. The examples given below include predictions based on cell type specific transcripts identified in the mouse (x-axis) vs. human (y-axis). **C.** There was a strong negative correlation between predictions originating from *"cell type indices"* representing very dissimilar cell types, such as neurons and astrocytes. **D.** 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 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).

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

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 transcripts that overlapped between categories removed **(Suppl. Figure 8).** This led to ten consolidated primary cell-type indices for each sample. Using these consolidated cell type indices and principal components analysis, we found that the first principal component, which encompassed 23% of the variation in the full Pritzker dorsolateral prefrontal cortex microarray 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 4**). 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). In general, none of the original 38 individual cell type indices were noticeably superior to the indices that were averaged by primary cell type for predicting the principal components of variation in the dataset, although the variation in PC1 was slightly better accounted for by the general neuron index derived from ((13), r-squared=0.62) and the variation in PC2 was best accounted for by the cortical pyramidal neuron index (r-squared=0.65) and oligodendrocyte index (r-squared=0.57) derived from (17). 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.

To investigate whether the strong relationship between the top principal components of variation in our dataset and cell type composition indices originated artificially due to cell type specific genes representing a large percentage of the most highly variable transcripts in the dataset, we performed principal components analysis after excluding all cell type specific transcripts from the dataset and still found these strong correlations (**Suppl. Figure 9**). 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 10**). These results suggested that accounting for cell type balance was highly important for the interpretation of microarray data and could improve the signal-to-noise ratio in analyses aimed at identifying psychiatric risk genes.

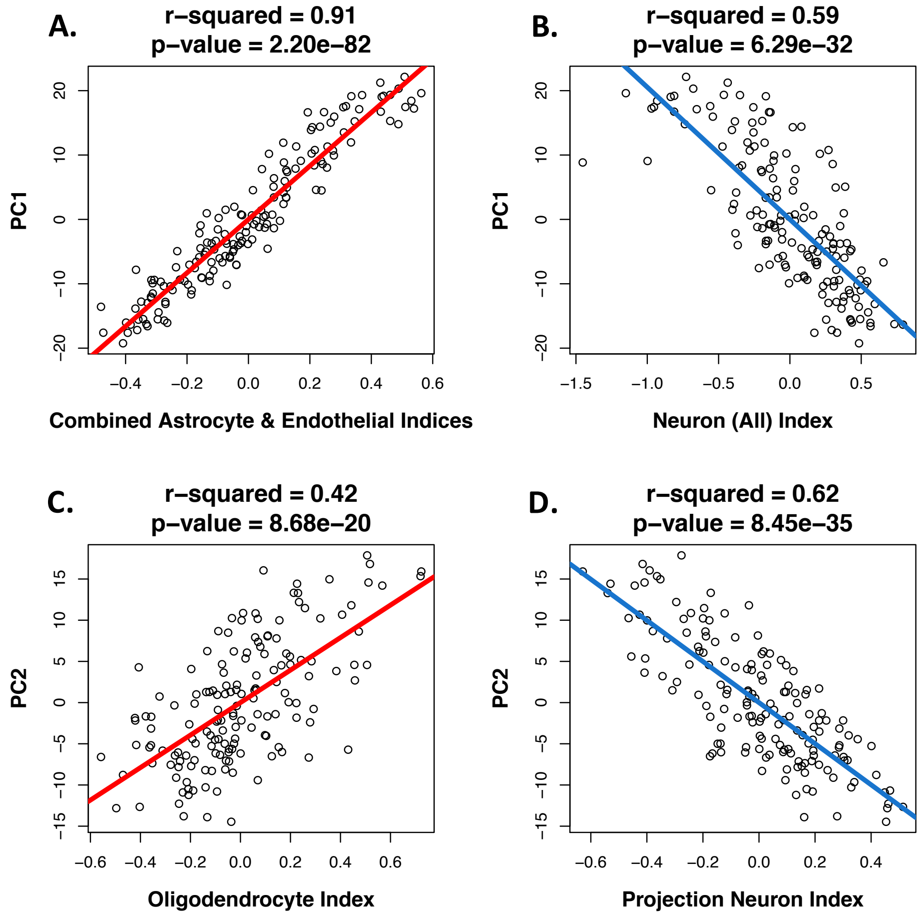
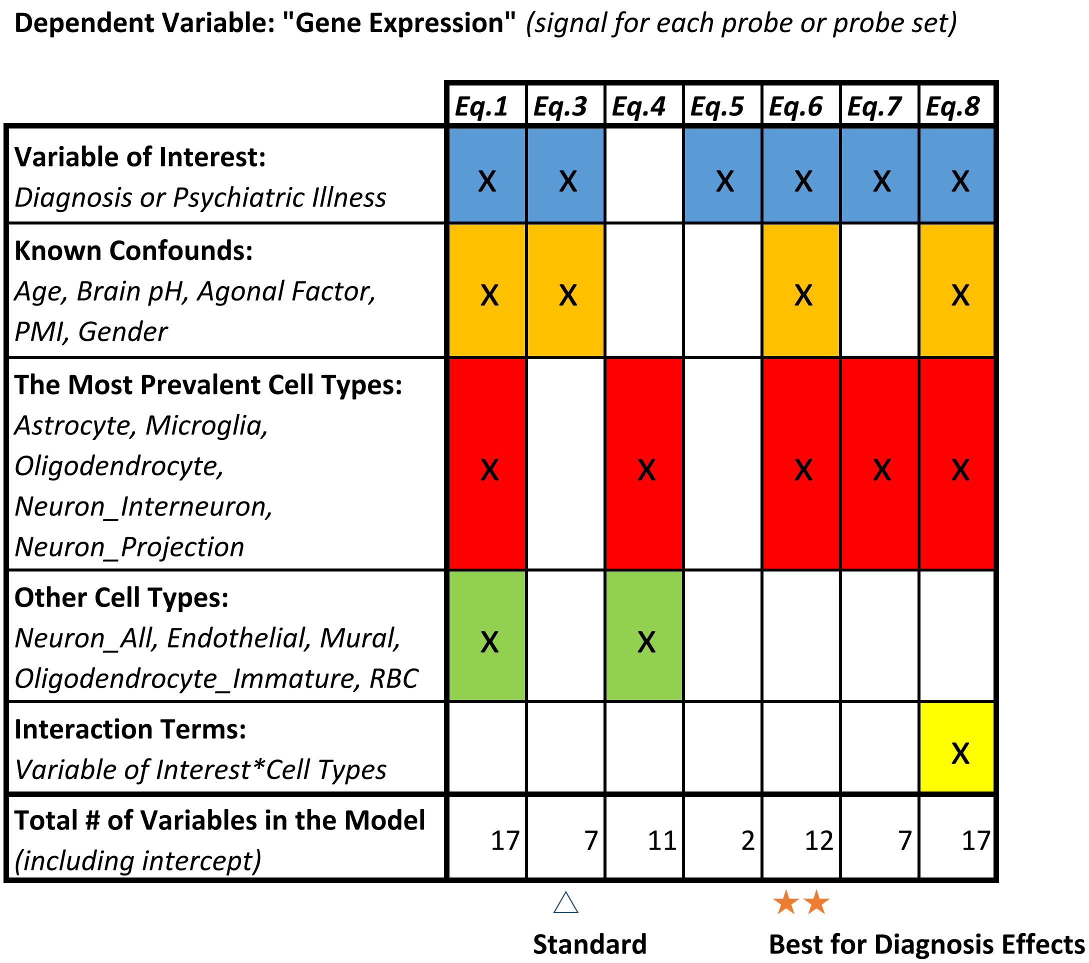


Figure 4. Cell content predictions explain a large percentage of the variability in microarray data derived from the human cortex. The first principal component of variation (PC1) encompassed 23% of the variation in the dataset, and was **A)** positively correlated with predicted “support cell” content 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).

2.5 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 1* in **Figure 5**). On average, this model explained 35% of the variation in the data (R^2). Shown in **Figure 6** 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 4.**



Dependent Variable: "Gene Expression" *(signal for each probe or probe set)*

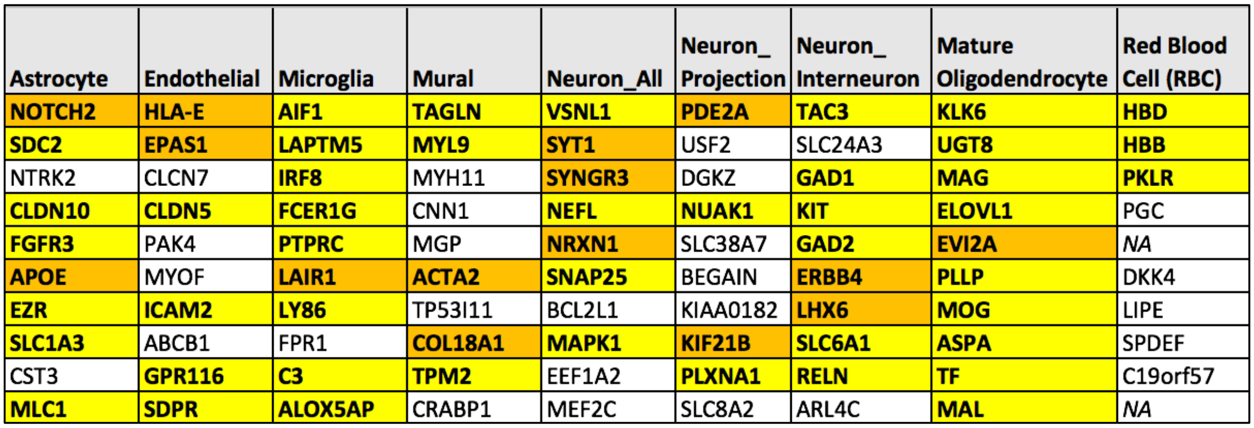
	Eq.1	*Eq.3*	*Eq.4*	*Eq.5*	*Eq.6*	*Eq.7*	*Eq.8*
Variable of Interest: *Diagnosis or Psychiatric Illness*	X	X		X	X	X	X
Known Confounds: *Age, Brain pH, Agonal Factor, PMI, Gender*	X	X			X		X
The Most Prevalent Cell Types: *Astrocyte, Microglia, Oligodendrocyte, Neuron_Interneuron, Neuron_Projection*	X		X		X	X	X
Other Cell Types: *Neuron_All, Endothelial, Mural, Oligodendrocyte_Immature, RBC*	X		X				
Interaction Terms: *Variable of Interest*Cell Types*							X
Total # of Variables in the Model *(including intercept)*	17	7	11	2	12	7	17
		△			★★		

Standard **Best for Diagnosis Effects**

Figure 5: Overview of the variables included in each model of gene expression (Eq.1, Eq.3-Eq.8).

General Linear Regression Model Format:

$$\text{Gene Expression} = \beta 0 + \beta 1 * (\text{Variable of Interest}) + \beta 2 * (\text{Variable 2}) + \beta 3 * (\text{Variable 3}) \ldots + \varepsilon$$



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*

Figure 6. 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 **Supplementary Table 4.**

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: (28); *EPAS1*, in response to lack of oxygen: (29)).

NOTCH2 (*Notch 2*), one of the top astrocyte-related genes, promotes astrocytic cell lineage (30), and *APOE* (*Apolipoprotein E*) is primarily secreted by astrocytes in the central nervous system (31). 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 (32). 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 (33), and *KIF21B* (*Kinesin Family Member 21B*) is a kinesin that has been found in the dendrites of pyramidal neurons (34). 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.

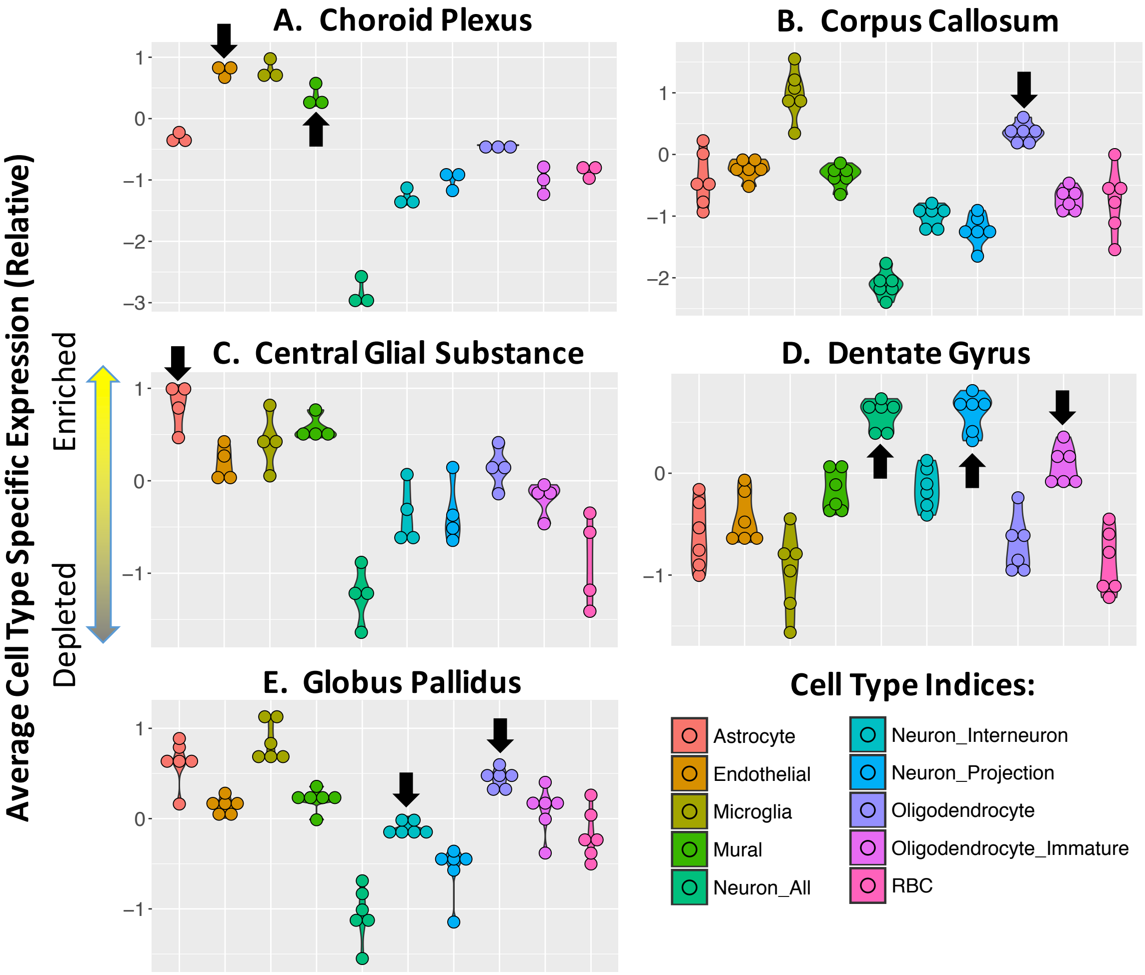
2.6 Using Cell Type Specific Transcripts to Predict Cell Content in Microarray Data for >840 Samples from 160 Human Brain Regions

For validation, we decided to also apply our cell type analysis to a large Agilent microarray dataset (841 samples) spanning 160 cortical and subcortical brain regions from the Allen Brain Atlas (**Suppl. Table 3;** (35)). This dataset included high-quality tissue (absence of neuropathology, pH>6.7, PMI<31 hrs, RIN>5.5) from 6 human subjects (36). 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 (37).

The 30,000 probes mapped onto 18,787 unique genes (as determined by gene symbol). We found that 1608 of these genes were identified as having cell type specific expression within our database. Then, using a procedure similar to that used for the Pritzker prefrontal cortex dataset, we 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.

2.7 Predicted Cell Content Accurately Reflects Regional Differences in Cell Type Balance

To explore the generalizability of our method to non-cortical samples, we examined the relative balance of each of the 10 primary cell types in all 160 brain regions included in the Allen Brain Atlas microarray dataset. To do this, we used violin plots, which are preferable for visualizing trends in data with small sample sizes (1-6 subjects per region). The results clearly indicated that our cell type analyses could identify well-established differences in cell type balance across brain regions (**Figure 7)**. Within the choroid plexus, which is a villous structure located in the ventricles made up of support cells (epithelium) and an extensive capillary network (38), there is an enrichment of cells related to vasculature (endothelial cells, mural cells) and immunity (microglia). In the corpus callosum, which is the primary myelinated fiber tract connecting the cerebral hemispheres (38), there is an enrichment of oligodendrocytes and microglia. The central glial substance is enriched with glia and support cells, with a particular emphasis on astrocytes. The dentate gyrus, which is one of the only neurogenic regions in the adult brain (39) and which contains the predominantly glutamatergic granule cells projecting into the mossy fibre pathway (40), has an enrichment of both immature-like cells and projection neurons. The internal segment of the globus pallidus, which is highly GABA-ergic and named after its white matter intrusions (38), was enriched for oligodendrocytes, astrocytes, and microglia, as well as a prominent subset of interneurons. The relative cell content predictions for the other brain regions can be found in **Suppl. Table 5.** Even though this analysis was based on cell type specific genes identified in the forebrain and cortex, these results provide fundamental validation that each of primary consolidated cell type indices is generally tracking their respective cell type in subcortical structures.



346

347 **Figure 7. Cell content predictions derived from microarray data capture canonical regional differences in cell type balance.** As an additional form of validation, we applied our cell type analysis to a large Agilent microarray dataset (841 samples) spanning 160 brain regions from the Allen Brain Atlas. Depicted below is the relative average cell type specific expression for each subject (dot) for each of the 10 primary cell type indices. A color-coded region encompasses the data cloud for each cell type. Within the example brain regions, we are able to clearly identify canonical differences in cell type balance, with black arrows indicating cell types that are known to be enriched in that respective region. **A.** Within the choroid plexus there is an enrichment of cells

related to vasculature (endothelial cells, mural cells) and immunity (microglia). **B.** In the corpus callosum there is an enrichment of oligodendrocytes and microglia. **C.** The central glial substance is enriched with glia and support cells, with a particular emphasis on astrocytes. **D.** The dentate gyrus has an enrichment of both immature cells and projection neurons. **E.** The internal segment of the globus pallidus was enriched for oligodendrocytes, astrocytes, and microglia, as well as a prominent subset of interneurons. The relative cell content predictions for the other brain regions can be found in **Suppl. Table 5.**

Similar to the Pritzker dataset, we outputted a table of the top genes associated with each cell type (as assessed using the model in *Equation 4*, **Figure 5**). We found that the results included a mixture of well-known cell type markers and novel findings **(Suppl. Figure 11; Suppl. Table 6)**. When this model was applied to the principal components of variation in the dataset instead of the data for individual genes, we again found that the main sources of variation in the dataset could be overwhelmingly accounted for by cell type balance (*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 12**). Overall, these results indicate that our method for statistically predicting cell content can be a useful addition to the analysis of non-cortical as well as cortical data sets.

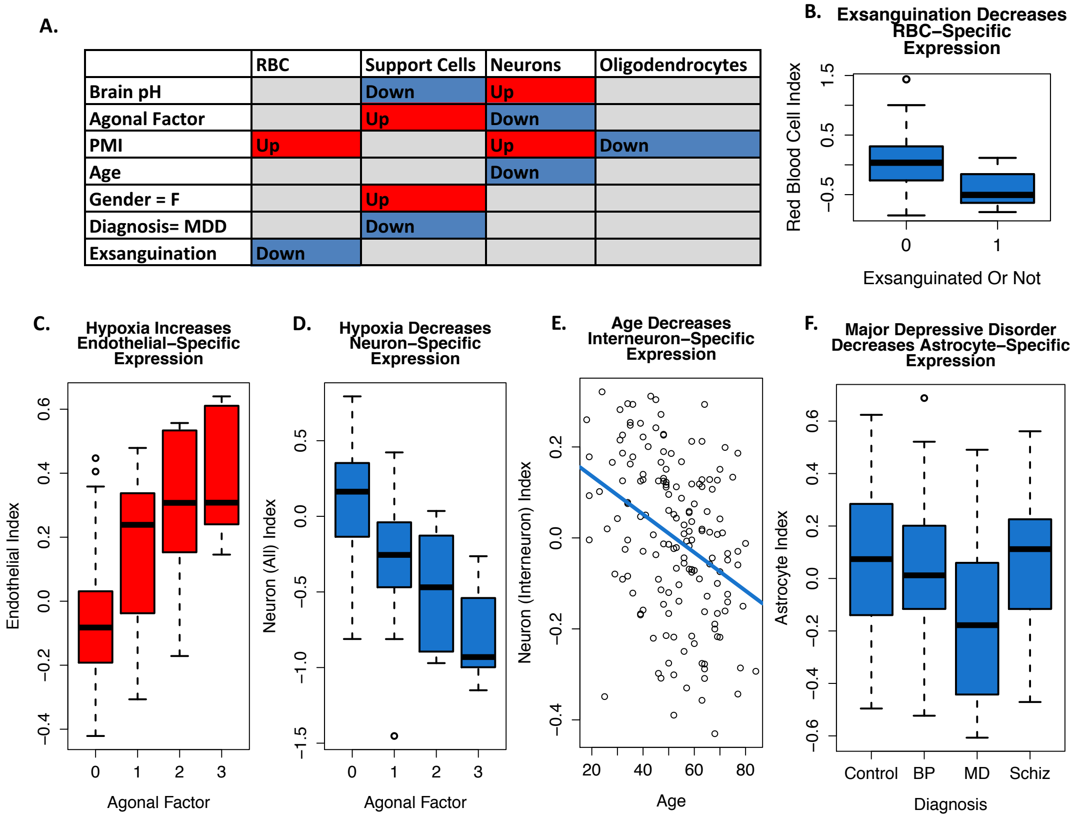
2.8 Cell Content Predictions Derived from Microarray Data Match Known Relationships Between Clinical/Biological Variables and Brain Tissue Cell Content

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 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 2:

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

This analysis uncovered many well-known relationships between brain tissue cell content and clinical or biological variables (**Figure 8)**. For example, 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; **Figure 8b**). The presence of prolonged hypoxia around the time of death, as indicated by either low brain pH or high agonal factor score, 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; **Figure 8c**) 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 (41)**.** 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., (42); **Figure 8d**). Finally, we saw a prominent increase in the microglia index in response to low oxygen (Microglia vs. Agonal Factor: β= 0.122096, p= 0.0000181), paralleling known activation of microglia in response to hypoxia (43,44), although we could find little evidence in the literature for actual proliferation under hypoxic events (unlike other injury). This lead us to wonder whether our microglial indices might largely reflect reactive (vs. ramified) microglia since they were typically derived from experiments performed on microglia in dissociated conditions. This possibility was at least partially supported by the presence of many immune-related molecules in the original microglial indices, including many of the interleukins, chemokines, and tumor necrosis factor.



A.

	RBC	Support Cells	Neurons	Oligodendrocytes
Brain pH		Down	Up	
Agonal Factor		Up	Down	
PMI	Up		Up	Down
Age			Down	
Gender = F		Up		
Diagnosis= MDD		Down		
Exsanguination	Down			

Figure 8. 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.** A table illustrating the relationship between cell types and subject variables (p<0.05). General cell type categories showing similar effects have been summarized together ("Support Cells": Astrocytes, Endothelial Cells, Mural Cells, Microglia; "Neurons": Neuron_All, Neuron_Interneuron, Neuron_Projection). 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). A more detailed table illustrating the relationship between subject variables and each specific cell type can be

found in **Supplementary Figure 14. B.** Subjects that died in a manner that involved exsanguination (n=14) had a notably low red blood cell index (*β* =-0.398; *p*=0.00056). **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 (Agonal Factor: *β*=0.118 *p*=2.85e-07; Brain pH: *β*=-0.210, *p*= 0.0003) matching previous demonstrations of cerebral angiogenesis, activation, and proliferation in low oxygen environments (41). **D.** High agonal factor was also associated with a clear decrease in neuronal indices (Neuron_All vs. Agonal Factor: *β*=-0.242, p=3.58e-09) mirroring the vulnerability of neurons to low oxygen (42). **E.** Age was associated with a moderate decrease in the neuronal indices (Neuron_Interneuron vs. Age: *β*=- -0.00291, p= 0.000956) which fits known decreases in grey matter density in the frontal cortex in aging humans (45). **F.** Major Depressive Disorder was associated with a moderate decrease in astrocyte index (*β* = -0.1326572, p= 0.0118), which fits what has been observed morphometrically (46).

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; **Figure 8e**), which fits known decreases in gray matter density in the frontal cortex in aging humans (45), as well as age-related sub-region specific decreases in frontal neuron numbers in primates (47) and rats (48). However, in some regions of the prefrontal cortex, age-related decreases in grey matter are primarily driven by synaptic atrophy instead of decreased cell number (49). This raised the question of whether the decline in our neuronal cell indices with age was being largely driven by the enrichment of genes related to synaptic function in the index.

To explore this possibility, we first evaluated the relationship between age and gene expression while controlling for likely confounds using the signal data for all probesets in the dataset (*Equation 3*, **Figure 5).** We used "DAVID: Functional Annotation Tool" (//david.ncifcrf.gov/summary.jsp, (50,51) 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 7).** 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 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 **(Suppl. Table 7).** Using this approach, we found that transcripts from both classifications of functional clusters showed an average decrease in expression with age (T(28)=-4.5612, *p*= 9.197e-05, T(40)=-2.7566, *p*=0.008756, respectively), but the decrease was larger for transcripts associated with dendritic/axonal-related functions (T(50.082)=2.3385, p= 0.02339, **Suppl. Figure 13**). 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.

Non-canonical relationships between subject variables and predicted cell content can be found in **Figure 8a** and **Suppl. Figure 14.** One of the more prominent unexpected effects was a large decrease in the oligodendrocyte index with longer post-mortem interval (β= - 0.00749, p=0.000474). 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 (52), but to our knowledge the topic is otherwise not well studied. This effect was accompanied by an increase in two of the neuron indices (Neuron_All vs. PMI: β = 0.006997, p= 0.013509; Neuron_Projection Neuron vs. PMI: β = 0.0070766, p=0.000164), and RBC index (β = 0.009612, p= 0.00721), for which we have no good explanation. We also saw an increased mural index (β = 0.0950444, p= 0.00635) and endothelial index (β = 0.06917, p= 0.042738) in females, which, combined with a trend towards increased RBC index (p=0.08) seemed to suggest increased vascularization or meninges, but we could not find any existing support for the hypothesis in the literature.

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

2.9 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 (53)). This decrease in glia, and particularly astrocytes, was replicated experimentally in animals exposed to chronic stress (54), and when induced pharmacologically, was capable of driving animals into a depressive-like condition (54). Replicating the results of (46), we observed a moderate decrease in astrocyte index in the prefrontal cortex of subjects with Major Depressive Disorder (β = -0.1326572, p= 0.0118), but did not see similar changes in the brains of subjects with Bipolar Disorder or Schizophrenia **(Figure 8f).** We did not see significant changes in any of the other cell type indices in relationship to diagnosis.

2.10 Including Cell Content Predictions in the Analysis of Microarray Data Improves the Detection of Diagnosis-Related Genes

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 are explained by neuron to glia ratio. Therefore, we compiled a list of genes that had previously documented relationships with psychiatric illness in particular cell types in the human prefrontal cortex, as detected using in situ hybridization, immunocytochemistry, or single-cell laser capture microscopy. These included several genes with a well-documented downregulation in interneurons in relationship to schizophrenia or psychosis (reviewed further in (19)); *GAD1*: (55–57); *RELN*:(55); *SST*: (58), *SLC6A1*

(*GAT1*): (59), *PVALB*:(56)), and 25 genes recently shown to have highly altered expression in pyramidal neurons in cortical layers 3 and 5 of subjects with schizophrenia using single cell laser capture and microarray (1). We also considered *SYP*, which encodes a protein decreased in projection neurons in subjects with schizophrenia (reviewed further in (19); (60)) and *HTR2A*, which encodes a protein increased in projection neurons in subjects who committed suicide (61). As further validation, it seemed prudent to include genes that were known to have differential expression in relationship with non-psychiatric variables in specific cells within the prefrontal cortex as well. These included *CALB1* and *CALB2,* both of which encode proteins in neurons that decrease with age (62).

We then examined our ability to detect these known relationships using models of increasing complexity (**Figure 5**), including a simple base model containing just the variable of interest (*Equation 5,* **Figure 5),** a model controlling for known confounds in the dataset (pH, agonal factor, age, post-mortem interval, and sex, *Equation 3,* **Figure 5**) and a model controlling for known confounds as well as each of the 10 cell type indices (*Equation 1*, **Figure 5**). Due to the multicollinearity present between the variables included in *Equation 1*, we also used two models that only included the most prevalent cell types (21) and avoided highly correlated categories. The first of these models (*Equation 6,* **Figure 5**) included other major confounds as well, whereas the second model excluded them (*Equation 7,* **Figure 5**).

We found that including predictions of cell type balance in our models assessing the effect of diagnosis or age on the expression of our validation genes dramatically improved model fit as assessed by Akaike's Information Criterion (AIC) or Bayesian

Information Criterion (BIC), and a 27% reduction in residual standard error (**Figure 9, Suppl. Figure 15**). These improvements were largest with the addition of the five most prevalent cell types to the model; the addition of less common cell types produced smaller gains. We also tried replacing the diagnosis term in our models with a more general term representing presence or absence of a psychiatric condition because we had found in the past that many of the genes that were associated with diagnosis in our samples were altered across diagnostic categories. This replacement slightly improved model fit in all versions of the analysis (Eq. 1, 3, 5-7).

Overall we found that adding predictions of cell type balance to our models improved our ability to detect previously-documented relationships with diagnosis in the Pritzker dataset (**Figure 9**). Prior to the addition of cell type to the model, we found that only one of 32 genes with a previously documented relationship to diagnosis in individual cells in the prefrontal cortex showed that relationship with a nominal p<0.05 in our dataset (*Eq. 5: HTR2A, Eq.1: SLC6A1).* After including cell type balance in the model, the relationship of three genes with diagnosis was now detectable (*SLC6A1, SST, COX7B*; **Suppl. Figure 16)**. Overall, the number of validation genes showing the same direction of effect as previously documented increased from 56% (18/32) to 68-72% (23/32). Models that included a more general term for presence or absence of a psychiatric condition performed even better (**Suppl. Figure 17)**. When using a basic model (Eq. 5) or when controlling for known confounds (Eq. 3) only one out of 32 validation genes were associated with psychiatric illness (*SLC6A1*). However, once we included predicted cell type balance (Eq. 1) five of the 32 validation genes showed a diagnosis relationship (p<0.05, *SST, PVALB, LGALS1, MGST3, ACTR10*), and the percentage of validation genes showing the same direction of effect as previously documented increased from 34% (11/32) to 78% (25/32), a significant improvement as indicated by Fisher's exact test (p=0.036). The use of forward/backward stepwise model selection (using the R function stepwise{Rcmdr}, criterion=BIC) drawing from a pool of variables that included diagnosis, general presence of a psychiatric illness, suicide, known confounds, and all 10 cell types, was also successful at detecting several genes in association with psychiatric illness (*SST, SLC6A1, MGST3, PCSK1*) and suicide (*LGALS1*), but these results should be viewed more cautiously due to the known presence of overfitting in stepwise procedures producing overly optimistic p-values. Backward/forward stepwise selection was noticeably less sensitive and included multiple false positives (incorrect direction of effect with a p<0.05). Both genes with a previously documented relationship to age (*CALB1, CALB2*) had such strong age-related effects in our dataset (*p*=4.84E-23, *p*=8.33E-08, respectively) that model specification had little impact on their results (**Suppl. Figure 18)**.

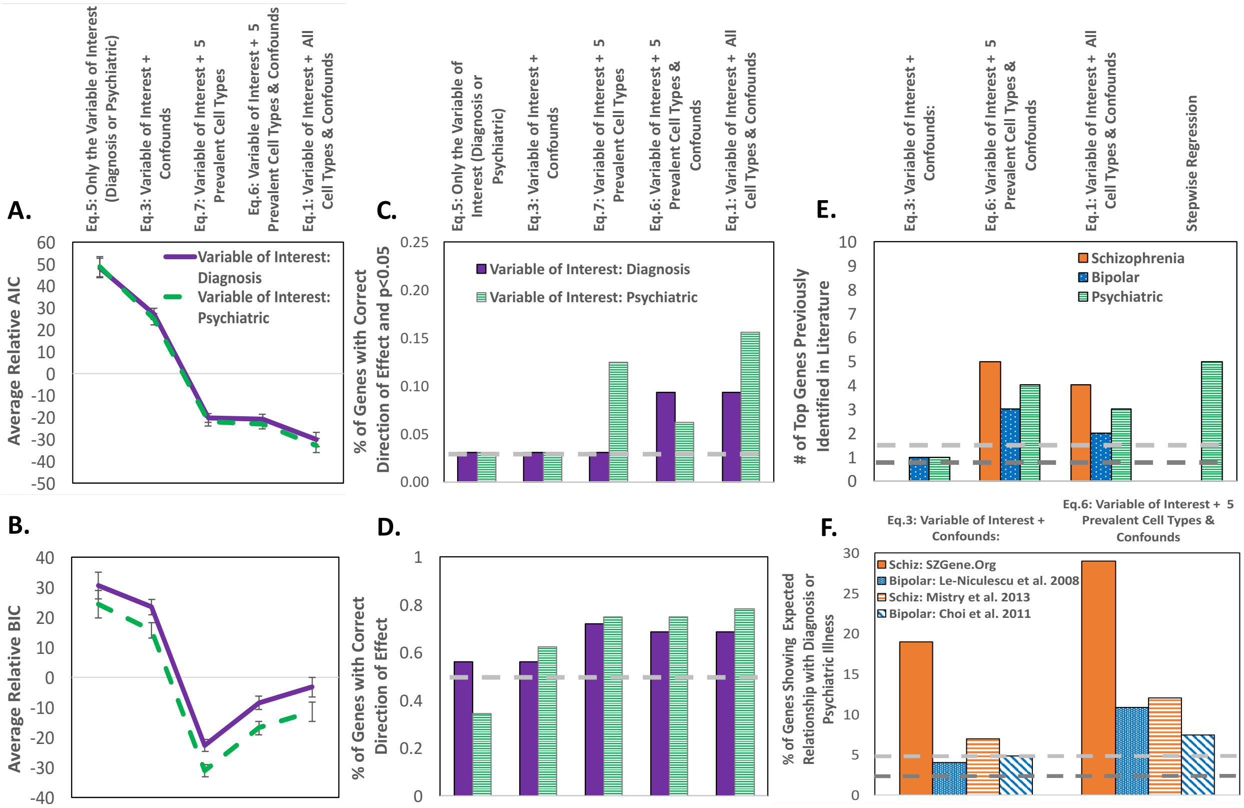


Figure 9. More previously-identified diagnosis-related transcripts are successfully detected by including cell content predictions in the analysis of microarray data. **A-D.** We assessed the fit and sensitivity of five different types of models (arranged in order of complexity) to signal data from 32 genes previously demonstrated to have altered expression or protein content in particular cell types within the dorsolateral prefrontal cortex in relationship to diagnosis (1,19,55–61). Improvements in model fit were largest with the addition of the five most prevalent cell types to the model; the addition of less common cell types produced smaller gains as demonstrated by **A.** Diminishing reductions in AIC with additional model terms (AIC relative to the average for each gene across models, +/-SE), **B.** Lowest BIC for a model that only included psychiatric illness and the five most prevalent cortical cell types (BIC relative to the average for each gene across models, +/-SE). **C.**The addition of cell type to the model

increased the percentage of genes showing the correct relationship with diagnosis or psychiatric illness (with a nominal p<0.05) so that it now surpasses what would be expected by random chance (dotted line: 0.05 (nominal p-value)*0.5 (chance of correct direction of effect)). **D.** The addition of cell type to the model increased the percentage of genes showing the correct direction of effect in relationship with diagnosis and psychiatric illness so that it now surpasses what would be expected by random chance (dotted line: 0.5). **E.** When analyzing the full dataset, the top 10 genes associated with diagnosis or psychiatric illness in models that include cell content predictions include genes previously identified in the literature indexed on PubMed. The dotted lines illustrate the rate of overlap with the literature for 100 randomly selected genes from our dataset for Schizophrenia or Bipolar Disorder (grey: 7/100 and 8/100, respectively), or for psychiatric illness in general (dark grey: 15/100, search terms: “Schizophrenia”, “Bipolar”, “Depression”, “Anxiety”, “Suicide”). **F.** The addition of cell type to the model increased the percentage of genes showing the correct relationship with diagnosis or psychiatric illness (with a nominal p<0.05) for top genes associated with diagnosis in other types of studies: genome-wide association (63), convergent functional genomics (64), and meta-analyses of macro-dissected microarray (65,66), so that the number of genes showing the expected effects now surpasses what would be expected by random chance (for microarray data: dark grey dotted line: the presence of the relationship and correct direction of effect: 0.05 (nominal p-value)*0.5 (chance of correct direction of effect), for other studies: light grey dotted line: 0.05 for presence of a relationship).

We found that adding predictions of cell type balance to our models also improved our ability to detect altered gene expression associated with known genetic risk loci as well as candidate genes identified by convergent functional genomics, and even enhanced our ability to replicate previous findings from macro-dissected microarray. For example, SZGene.org identified 38 top genes associated with genetic risk loci for Schizophrenia (as reported in (63)), 31 of which were represented in our dataset. Of these, six (19%) were found to have a significant relationship (p<0.05) with either Schizophrenia or psychiatric illness in our dataset when controlling for known confounds (Eq.3), whereas nine (29%) were related to either Schizophrenia or psychiatric illness when controlling for known confounds and the most prevalent cell types (*Eq. 6*; Fisher's exact test p=0.5541, **Suppl. Fig 19**). Similarly, out of the top 114 genes associated with Bipolar disorder using convergent functional genomics (64), 101 were represented in our dataset. Of these, only four (4%) had a significant relationship (p<0.05) with either Bipolar Disorder or psychiatric illness in our dataset when controlling for known confounds, whereas 11 (10.9%) were related to either Bipolar Disorder or psychiatric illness in a model controlling for known confounds and the most prevalent cell types (Fisher's exact test p=0.1047, **Suppl. Fig 20**). We had less success identifying altered gene expression in association with the top MDD risk loci identified by (67): out of the 18 genes associated with the 17 SNPs that reached genome-wide significance in their joint analysis of three large sequencing datasets, 12 were represented by probesets in our dataset. Only one of these was found to have a significant relationship with psychiatric illness while accounting for confounds

(*SLC6A15*, ß=0.11708, p=0.0302), and no relationships were found when considering confounds and prevalent cell types.

We expected that controlling for cell type would weaken our ability to replicate the diagnosis effects observed in other microarray experiments performed on macro-dissected prefrontal tissue, since any changes in cell type balance due to psychiatric illness would be selectively ignored by our analysis. The opposite turned out to be true. For example, (65) found that Bipolar disorder was strongly related to gene expression in the dorsolateral prefrontal cortex data for 400 probes (FDR<0.05), 326 of which represented genes included in our dataset. Of those, only 16 (4.9%) showed the same direction of effect and a p<0.05 in a model including either diagnosis or psychiatric illness and known confounds, whereas 24 (7.4%) showed the same direction of effect and a p<0.05 if the model also included prevalent cell types (Fisher's exact test p=0.2531, **Suppl. Fig 21).** Likewise, (66) found that 125 probes were consistently associated with Schizophrenia in a large meta-analysis of microarray data derived from macro-dissected prefrontal tissue, of which 111 were represented our data set. Of these, eight (7%) showed the same direction of effect and p<0.05 in a model that included either diagnosis or psychiatric illness and known confounds, whereas 13 (12%) showed the same direction of effect and p<0.05 if the model also included the most prevalent cell types (Fisher's exact test= 0.3593, **Suppl. Fig 22**). There was an increase in the number of genes showing the correct direction of effect as well: 50% (psychiatric illness) or 62% (diagnosis) when considering confounds vs. 68% (psychiatric illness) or 62% (diagnosis) when considering confounds and prevalent cell types. Altogether, including the most prevalent cell types in our model significantly enhanced our ability to detect relationships between gene expression and diagnosis-related genes identified by a variety of techniques (Fisher's exact test: p=0.0221, **Figure 9F**).

2.11 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 improved our ability to detect previously-identified relationships with diagnosis, most relationships still went undetected 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.

We found that only one of the top 10 diagnosis-related genes detected using a model that included diagnosis and known confounds (*Equation 3)* was previously noted in the human literature *(FOS:* (68,69)). The same was true if we replaced diagnosis with a term representing the general presence or absence of a psychiatric illness (*ALDH1A1:* (64)). In contrast, when we used a model that included diagnosis, known confounds, and predictions for the balance of the five most prevalent cortical cell types (*Equation 6*), we found that five of the top 10 genes associated with Schizophrenia had been previously identified in the literature (*ARHGEF2:* (70)**,** *DOC2A:* (71), *FBX09:* (66), *GRM1:* (72,73); *CEBPA:* (74)), and three of the top 10 genes associated with Bipolar Disorder (*ALDH1A1:* (64)**,** *SNAP25:* (75), *NRN1:*(76)**; Suppl. Figure 23, Suppl. Table 8, Suppl. Table 10**). This was a significant enrichment in overlap with the literature as indicated by a Fisher's exact test across all three diagnosis groups (1/30 vs. 8/30 overlap with the literature, p=0.0257, **Figure 9E**) or when comparing the results for the schizophrenia group 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*: (64); *HBS1L:* (4); *HIVEP2:* (77), *FBX09*: (66), **Suppl. Figure 24, Suppl. Table 9, Suppl. Table 11**), 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:* (63), *GRM1:* (72,73); *ALDH1A1:* (64); *SNAP25:* (75); *HIVEP2:* (77), 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 improves the sensitivity of the assay for detecting altered gene expression in relationship to psychiatric disease.

3. 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 principal components of variation in microarray data derived from macrodissected brain tissue samples, far surpassing the importance of other 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 is substantial turbulence 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 particularly 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 doubling the sensitivity of our assay to detect previously-identified diagnosis-related genes.

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 strong effects in a microarray dataset, including other genes with highly cell-type specific expression and large-scale alterations in cell content in relationship with known subject variables. We have not tested the sensitivity of the technique for detecting smaller effects or parsing effects for genes related to multiple cell types, 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 (78)) or meta-analyses refine the list of genes defined as having cell-type specific expression (e.g., (79)), our indices will surely evolve with these new classification frameworks, but the power of our approach will remain, in that we can disentangle the intrinsic changes of individual genes from the population-level shifts of major cell types. The same approach can be extended to studying other structurally complex organs that involve the concerted function of many cell types.

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 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. Likewise, using PSEA, (12) also found 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, even within non-diseased samples. Most importantly, the PSEA technique has been carefully validated: PSEA was found to successfully predict the content of RNA-mixing experiments (12), cellular expression data from *in situ* hybridization or laser-capture microdissection experiments (11), and neuron-specific neurodegenerative effects found with laser-capture microdissection (10).The differences between our techniques are mostly due to our access to a large sample size and 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 (80). 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. Likewise, PSEA was designed for use with small microarray datasets, and thus depends on a variety of model selection techniques to minimize the number of terms included in the linear model. Although necessary, this step introduces the risk of mis-assigning effects associated with correlated cell types. Using a large dataset gave us the opportunity to include terms for all major cell types in the analysis, as well as terms representing a number of important identified confounds (age, pH, PMI, gender). Due to these analytical differences, we are able to effectively characterize gene expression associated with less prevalent cell types (*e.g.,* endothelial cells) and compare the utility of cell type specific markers derived from a variety of species and experimental techniques.

There was one seemingly-small difference between our method and PSEA that actually turned out to produce a large difference in efficacy: normalization of the original gene expression data using a z-score instead of a ratio of the mean (81). As part of a set of later validation analyses (**Suppl. Methods and Results, Suppl. Figures 25-26**), we performed a head-to-head comparison of our method and PSEA using a single-cell RNA-Seq dataset and the same database of cell type specific genes. Both methods strongly predicted cell identity, but on average we found that one third of the variation in the predictions of relative cell content derived from PSEA ("population reference signal") were related to the cell identity of the samples versus almost half of the variation in our consolidated cell type indices. We conclude that our method may be a more effective manner of predicting cell type balance in some datasets.

Another notable difference between our final analysis methods and those used by PSEA (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., (11,12,82)) but we were not able to validate the method using our database of previously-documented relationships with diagnosis in prefrontal cell types and a variety of model specifications (e.g., **Suppl. Figure 27)**. 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), 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.

One result from our 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., (25,26,83,84)). 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., (85)). However, the magnitude of the effect of hypoxia was 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. Although the differential effects of hypoxia on neurons and glial cells have been studied since the 1960's (86), to our knowledge this is the first time that anyone has related the large effects of hypoxia in post-mortem transcriptomic data to alterations in cell type balance in the samples. 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.

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 dorsolateral prefrontal cortex occurred both at the level of cell population shifts (decreased astrocytic presence) 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.,*(87)) 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 (74) and neuroglia (*e.g.,* (46,54,88)) 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 (*e.g.,* Brain-Derived Neurotrophic Factor (*BDNF*), the Fibroblast Growth Factor (*FGF*) family, Glial-cell derived neurotrophic factor (*GDNF*), Vascular Endothelial Growth Factor (*VEGF*); for a review, see (23,89)).

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 as well as removing unwanted noise due to variations in cell content caused by dissection variability. 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 publically 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.

4. Materials and Methods

4.1 Ortholog Prediction: The gene symbols for the cell type specific transcripts derived from mouse datasets were fed into HCOP: Orthology Prediction Search (http://www.genenames.org/cgi-bin/hcop). We selected the ortholog for each transcript that was most commonly identified amongst the 11 available databases: EggNOG, Ensembl, HGNC, HomoloGene, Inparanoid, OMA, OrthoDB, OrthoMCL, Panther, PhylomeDB, and TreeFam.

4.2 Pritzker Dorsolateral Prefrontal Cortex Microarray Dataset:

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 and total RNA was extracted and hybridized to Affymetrix HT-U133A or HT-U133Plus-v2 chips in duplicate or triplicate at different laboratories using procedures described previously (25,85). 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 (90). Data from any subjects lacking information regarding critical pre- or post-mortem variables were removed from the analysis, leaving a final sample size of *n=157.* For detailed data collection methodology, see (85). 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).

Before conducting the current analysis, the microarray dataset was reannotated for probe-to-transcript correspondance (91), summarized using robust multi-array analysis (RMA) (92), log (base 2)-transformed, quantile normalized, gender-checked, median centered to remove batch effects, and the replicate microarrays for each subject were averaged (for a more detailed description of data preprocessing see (85)). Samples that exhibited markedly low average sample-sample correlation coefficients prior to median centering (<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 poor match with their duplicate microarrays run in a separate laboratory.

The data from control subjects is publically available in the Gene Expression Omnibus (GEO: Accession Number GSE6306) and the data for all subjects has been submitted and should be available shortly (GEO: *curation pending*) . All of the R script documenting these analyses can be found at https://github.com/hagenaue/CellTypeAnalyses_PritzkerAffyDLPFC.

4.3 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 (93). 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). We then extracted the z-score data for the list of cell type specific genes derived from each publication. 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.

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, and Mark Reimers for advice and feedback regarding the methodology or manuscript. Finally, 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.

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. 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.

26. 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.

27. 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.

28. 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.

29. 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.

30. 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.

31. 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.

32. 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.

33. 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.

34. 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.

35. 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.

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

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

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

39. Altman J, Das GD. Autoradiographic and histological evidence of postnatal hippocampal neurogenesis in rats. J Comp Neurol. 1965 Jun;124(3):319–35.

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

41. 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.

42. 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.

43. Lu D-Y, Liou H-C, Tang C-H, Fu W-M. Hypoxia-induced iNOS expression in microglia is regulated by the PI3-kinase/Akt/mTOR signaling pathway and activation of hypoxia inducible factor-1alpha. Biochem Pharmacol. 2006 Oct 16;72(8):992–1000.

44. Tadmouri A, Champagnat J, Morin-Surun MP. Activation of microglia and astrocytes in the nucleus tractus solitarius during ventilatory acclimatization to 10% hypoxia in unanesthetized mice. J Neurosci Res. 2014 May;92(5):627–33.

45. 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.

46. 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.

47. 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.

48. 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.

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

50. 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.

51. 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.

52. 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.

53. 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.

54. 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.

55. 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.

56. 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.

57. 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.

58. 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.

59. 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.

60. Glantz LA, Lewis DA. Reduction of synaptophysin immunoreactivity in the prefrontal cortex of subjects with schizophrenia. Regional and diagnostic specificity. Arch Gen Psychiatry. 1997 Jul;54(7):660–9.

61. 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.

62. Bu J, Sathyendra V, Nagykery N, Geula C. Age-related changes in calbindin-D28k, calretinin, and parvalbumin-immunoreactive neurons in the human cerebral cortex. Exp Neurol. 2003 Jul;182(1):220–31.

63. 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.

64. 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.

65. 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.

66. 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.

67. Hyde CL, Nagle MW, Tian C, Chen X, Paciga SA, Wendland JR, et al. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. Nat Genet. 2016 Sep;48(9):1031–6.

68. 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.

69. 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.

70. 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.

71. 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.

72. 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.

73. 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.

74. 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.

75. 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.

76. 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.

77. 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.

78. 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.

79. 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

80. 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.

81. Cheadle C, Cho-Chung YS, Becker KG, Vawter MP. Application of z-score transformation to Affymetrix data. Appl Bioinformatics. 2003;2(4):209–17.

82. 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.

83. 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.

84. 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.

85. 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.

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

87. 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.

88. 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.

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

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

91. 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.

92. 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.

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

7. Supporting Information Captions

S1 Text. Supplementary Methods & Results

S2 Text. Supplementary Figures and Figure Legends.

S1 Table. 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.

S2 Table. Sample demographics for the Pritzker Consortium Dorsolateral Prefrontal Cortex Affymetrix microarray data.

S3 Table. 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.

S4 Table. 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).

S5 Table. 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 **Supplementary Table 3**. The second spreadsheet contains the standard error (SE) for the averages in the first worksheet.

S6 Table. 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 **Supplementary Table 3**.

S7 Table. 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.

S8 Table. The relationship between diagnosis and all probes in the Pritzker Dorsolateral Prefrontal Cortex dataset using a traditional model that controls for standard confounds (*Equation 3* in Figure 5). For all probes in the dataset, the spreadsheet includes 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_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.

S9 Table. The relationship between psychiatric illness and all probes in the Pritzker Dorsolateral Prefrontal Cortex dataset using a traditional model that controls for standard confounds (*Equation 3* in Figure 5). For all probes in the dataset, the spreadsheet includes 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_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.

S10 Table. The relationship between diagnosis and all probes in the Pritzker Dorsolateral Prefrontal Cortex dataset using a model that controls for standard confounds and the five most prevalent cortical cell types (*Equation 6* in Figure 5). For all probes in the dataset, the spreadsheet includes 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_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.

S11 Table. The relationship between psychiatric illness and all probes in the Pritzker Dorsolateral Prefrontal Cortex dataset using a model that controls for standard confounds and the five most prevalent cortical cell types (*Equation 6* in Figure 5). For all probes in the dataset, the spreadsheet includes 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_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.