NeuroExpresso: A cross-laboratory database of brain cell-type expression profiles with applications to marker gene identification and bulk brain tissue transcriptome interpretation

B. Ogan Mancarci[1,2,3], Lilah Toker[2,3], Shreejoy Tripathy[2,3], Brenna Li[2,3], Brad Rocco[4,5], Etienne Sibille[4,5], Paul Pavlidis[2,3*]

[1]Graduate Program in Bioinformatics, University of British Columbia, Vancouver, Canada

[2]Department of Psychiatry, University of British Columbia, Vancouver, Canada

[3]Michael Smith Laboratories, University of British Columbia, Vancouver, Canada

[4]Campbell Family Mental Health Research Institute of CAMH

[5]Department of Psychiatry and the Department of Pharmacology and Toxicology, University of Toronto, Toronto, Canada.

Address correspondence to;

Paul Pavlidis, PhD
177 Michael Smith Laboratories 2185 East Mall
University of British Columbia Vancouver BC V6T1Z4
604 827 4157 paul@msl.ubc.ca

Ogan Mancarci: ogan.mancarci@msl.ubc.ca

Lilah Toker: ltoker@msl.ubc.ca

Shreejoy Tripathy: stripathy@msl.ubc.ca

Brenna Li: brenna.li@msl.ubc.ca

Brad Rocco: Brad.Rocco@camh.ca

Etienne Sibille: Etienne.Sibille@camh.ca

Abstract

The identification of cell type marker genes, genes highly enriched in specific cell types, plays an important role in the study of the nervous system. In particular, marker genes can be used to identify cell types to enable studies of their properties. Marker genes can also aid the interpretation of bulk tissue expression profiles by revealing cell type specific changes.

We assembled a database, NeuroExpresso, of publicly available mouse brain cell type-specific gene expression datasets. We then used stringent criteria to select marker genes highly expressed in individual cell types. We found a substantial number of novel markers previously unknown in the literature and validated a subset of them using in silico analyses and in situ hybridization. We next demonstrate the use of marker genes in analysis of whole tissue data by summarizing their expression into “cell type profiles” that can be thought of as surrogates for the relative abundance of the cell types across the samples studied.

Further analysis of our cell type-specific expression database confirms some recent findings about brain cell types along with revealing novel properties, such as Ddc expression in oligodendrocytes. To facilitate further use of this expanding database, we provide a user-friendly web interface for the visualization of expression data.

Significance Statement

Cell type markers are powerful tools in the study of the nervous system that help reveal properties of cell types and acquire additional information from large scale expression experiments. Despite their usefulness in the field, known marker genes for brain cell types are few in number. We present NeuroExpresso, a database of brain cell type specific gene expression profiles, and demonstrate the use of marker genes for acquiring cell type specific information from whole tissue expression. The database will prove itself as a useful resource for researchers aiming to reveal novel properties of the cell types and aid both laboratory and computational scientists to unravel the cell type specific components of brain disorders.

Introduction

Brain cells can be classified based on features such as their primary type (e.g. neurons vs. glia), location (e.g. cortex, hippocampus, cerebellum), electrophysiological properties (e.g. fast spiking vs. regular spiking), morphology (e.g. pyramidal cells, granule cells) or the neurotransmitter/neuromodulator they release (e.g. dopaminergic cells, serotonergic cells, GABAergic cells). Marker genes, genes that are expressed in a specific subset of cells, are often used in combination with other cellular features to define different types of cells (Hu et al., 2014; Margolis et al., 2006) and facilitate their characterization by tagging the cells of interest for further studies (Handley et al., 2015; Lobo et al., 2006; Tomomura et al., 2001). Marker genes have also found use in the analysis of whole tissue “bulk” gene expression profiling data, which are challenging to interpret due to the difficulty of determining the source cell type(s) of an observed expression change. For example, a decrease in a transcript can indicate a regulatory change in the expression level of the gene, a decrease in the number of cells expressing the gene, or both. To address this issue, computational methods have been proposed to estimate cell-type specific proportion changes based on expression patterns of known marker genes (Chikina et al., 2015; Westra et al., 2015; Xu et al., 2013). Finally, marker genes are obvious candidates for having cell-type specific functional roles.

An ideal cell type marker has a strongly enriched expression in a single cell type in the brain. However, this criterion can rarely be met, and for many purposes, cell type markers can be defined within the context of a certain brain region; thus, a useful marker may be specific for the cell type in one region but not necessarily in another region or brain-wide. For example, the calcium binding protein parvalbumin is a useful marker of both fast spiking interneurons in the cortex and Purkinje cells in the cerebellum (Celio and Heizmann, 1981; Kawaguchi et al., 1987). Whether the markers are defined brain-wide or in a region-specific context, the confidence in their specificity is established by testing their expression in as many different cell types as possible. This is important because a marker identified by comparing just two cell types might turn out to be expressed in a third, untested cell type, reducing its utility.

During the last decade, targeted cell specific purification followed by gene expression profiling has been applied to many cell types in the brain. Such studies, targeted towards well-characterized cell types, have greatly promoted our understanding of the functional and molecular diversity of these cells (Cahoy et al., 2008; Chung et al., 2005; Doyle et al., 2008). However, individual studies of this kind are limited in their ability to discover specific markers as they often analyse only a small subset of cell types (Shrestha et al., 2015; Okaty et al., 2009; Sugino et al., 2006) or have limited resolution as they group subtypes of cells together (Cahoy et al., 2008). Recently, advances in technology have enabled the use of single cell transcriptomics as a powerful tool to dissect neuronal diversity and derive novel molecular classifications of cells (Poulin et al., 2016). However, with single cell analysis the classification of cells to different types is generally done post-hoc, based on the clustering similarity in their gene expression patterns. These molecularly defined cell types are often uncharacterized otherwise (e.g. electrophysiologically, morphologically), challenging their identification outside of the original study and understanding their role in normal and pathological brain function. Regardless of the methodology used, technical biases and artefacts introduced during different stages of cellular purification (e.g. stress) and analysis (e.g. normalization) are known to have an impact on the obtained expression profiles (Handley et al., 2015; Okaty et al., 2011; Poulin et al., 2016). These effects can obscure true differences between two cell types or introduce disparities resulting from differences in cell state rather than cell type. Thus, individual studies may not be ideal for identification of cell type markers. We hypothesized that by aggregating cell-type specific studies, a more comprehensive data set more suitable for marker genes could be derived.

Here we report the analysis of an aggregated cross-laboratory data set of cell-type specific expression profiling experiments from mouse brain. We use this data set to identify sets of brain cell marker genes more comprehensive than any previously reported. Using homology relations, we also show how the markers can be applicable to human brain cell types. We validate the markers in external datasets and demonstrate their potential usage in analysis of whole tissue data via the summarization of marker gene expression into “marker gene profiles” (MGPs), which can be cautiously interpreted as correlates of cell type proportion. We have made the cell type expression profiles and marker sets available to the research community at neuroexpresso.org.

Methods

Figure 1A depicts the workflow and the major steps of this study. All analysis was performed in R version 3.1.2; the R code and data files are available from the authors.

Collection of cell type specific data sets:

We began with a collection of seven studies of isolated cell types from the brain, compiled by Okaty et. al. (2011). We expanded this by querying PubMed (http://www.ncbi.nlm.nih.gov/pubmed) and Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) (Barrett et al., 2013; Edgar et al., 2002) for cell type-specific expression datasets from the mouse brain that used Mouse Expression 430A Array (GPL339) or Mouse Genome 430 2.0 Array (GPL1261) platforms. These platforms were our focus as together, they are the most popular platform for analysis of mouse samples and are relatively comprehensive in gene coverage, and using a reduced range of platforms reduced technical issues in combining studies. Query terms included names of specific cell types (e.g. astrocytes, pyramidal cells) along with blanket terms such as “brain cell expression” and “purified brain cells”. Only samples derived from postnatal (> 14 days), wild type, untreated animals were included. Datasets obtained from cell cultures or cell lines were excluded due to the reported expression differences between cultured cells and primary cells (Cahoy et al., 2008; Halliwell, 2003; Januszyk et al., 2015)). We considered the use of single cell RNA sequencing data (Darmanis et al., 2015; Habib et al., 2016; Lake et al., 2016; Tasic et al., 2016; Zeisel et al., 2015). However, so far, these studies have been confined to a small number of brain regions and establishing the correspondence between the cell types as classified in these studies and the commonly defined cell types in the pooled cell data is not trivial. We thus leave this as a matter for future work. For the present study we instead used single cell data for validation (described below).

As a first step in the quality control of the data, we manually validated that each sample expressed the gene that was used as a marker for the purification step (expression greater than median expression among all gene signals in the dataset), along with other well established marker genes for the relevant cell type (eg. Pcp2 for Purkinje cells, Gad1 for GABAergic interneurons). We next excluded contaminated samples, namely, samples expressing established marker genes of nonrelated cell types in levels comparable to the cell type marker itself (for example neuronal samples expressing high levels of glial marker genes), which lead to the removal of 21 samples. The final database contains 30 major cell types compiled from 24 studies (summarized in **Table 1**); a complete list of all samples including those removed is available from the authors).

Grouping and re-assignment of cell-type samples:

When possible, samples were assigned to specific cell types based on the descriptions provided in their associated original publications. When expression profiles of closely related cell types were too similar to each other and we could not find differentiating marker genes meeting our criteria, they were grouped together into a single cell type ((e.g. A10 and A9 dopaminergic cells were grouped as “dopaminergic neurons”).

Because our focus was on finding markers specific to cell types within a single brain region, samples were grouped based on the brain region from which they were isolated, guided by the anatomical hierarchy of brain regions (Figure 1B). Brain subregions (e.g. locus coeruleus) were added to the hierarchy if there were multiple cell types represented in the subregion. An exception to the region assignment process is glial samples. Since these samples were only available only for cortex or cerebellum regions or extracted from whole brain, the following assignments were made: Cerebral cortex-derived astrocyte and oligodendrocyte samples were included in the analysis of other cerebral regions as well as thalamus, brainstem and spinal cord. Bergmann glia and cerebellum-derived oligodendrocytes were used in the analysis of cerebellum. The only microglia samples available had been isolated from whole brain homogenates and were included in the analysis of all brain regions.

Selection of cell type markers:

Cell type-enriched genes were selected for each brain region, based on fold change and clustering quality (see below). Since samples in the database originate from multiple independent studies, using different purification methods, they are expected to have study-specific effects that could potentially affect the gene selection process. The fold change between cell type of interest and the rest of the dataset was determined by taking the the median expression of all replicates and averaging the resulting values per cell type. The quality of clustering was determined by the mean silhouette coefficient (a measure of group dissimilarity ranged between -1 and 1(Rousseeuw, 1987)), assigning all samples from the same cell type to one cluster and all remaining cell types to another. Silhouette coefficients were calculated with the “silhouette” function from the “cluster” R package version 1.15.3 (Maechler et al., 2016), using the expression difference of the gene between samples as the distance metric.

Based on these metrics, genes were selected for each brain region using the following criteria (this process was run on each gene):

- A threshold expression level of 8 was selected to help ensure that the gene’s transcripts will be detectable in bulk tissue. Theoretically, if a gene has an expression level of 8 in a cell type, and the gene is specific to the cell type, an expression level of 6 would be observed if 1/8th of a bulk tissue is composed of the cell type. As many of the cell types in the database are likely to be as rare as or rarer than 1/8th, and 6 is generally close to background for these data, we picked 8 as a lower level of marker gene expression.
- If the $\log_2$ expression level in the cell type is above 9.5, there must be at least a 10-fold difference from the median expression level of the rest of the cells in the region.
- If $\log_2$ expression level in the cell type is below 9.5, the expression level must be higher than the expression level of every other cell type in that region. This condition was added because at 9.5 $\log_2$ expression level, for a 10-fold expression change to occur, the expression median should be lower than ~6 which is the median expression level of all genes in the dataset. Values below the median often represent the background signal that do not convey meaningful information but can prevent potentially useful marker genes from being selected.
- The mean silhouette coefficient for the gene must be higher than 0.5 for the associated cell type.
- The conditions above must be satisfied only by a single cell type in the region.

To prevent individual large studies from dominating the silhouette coefficient, we used the following randomization procedure, repeated 500 times. If studies representing same cell types did not have an equal number of samples, from each study, N samples are picked randomly. N being the number of samples in the smallest study that represent the same cell type. One-third (rounded up) of the remaining samples were randomly removed to ensure robustness against outlier samples resulting in the removal of 40%-65% of the samples depending on the brain region. A gene was retained as a marker if it satisfied the above criteria in more than 95% of the 500 resamplings. Human homologues of mouse genes were defined by NCBI HomoloGene (ftp://ftp.ncbi.nih.gov/pub/HomoloGene/build68/homologene.data).

Microglia enriched genes:

Microglia expression profiles differ significantly between activated and inactivated states and to our knowledge, the samples in our database represent only the inactive state (Holtman et al., 2015). In order to acquire marker genes with stable expression levels regardless of microglia activation state, we removed the genes differentially expressed in activated microglia based on Holtman et al. (2015). This step resulted in removal of 359 out of the original 629 microglial genes.

S100a10$^+$ pyramidal cell enriched genes:

The paper (Schmidt et al., 2012) describing the cortical S100a10$^+$ pyramidal cells emphasizes the existence of non-neuronal cells expressing S100a10$^+$. Schmidt et al. therefore limited their analysis to 7853 genes specifically expressed in neurons and advised third-party users of the data to do so as well. In compliance, we removed the selected marker genes for S100a10$^+$ pyramidal cells if they were not in the list of genes provided in Schmidt et al. (2012). Of note, we also removed S100a10 itself since based on the author’s description it was not specific to this cell type. In total, 40 of the 51 S100a10 pyramidal genes were removed in this manner.

In situ hybridization

Male C57BL/6J mice aged 13-15 weeks at time of sacrifice were used (n=5). Mice were euthanized by cervical dislocation and then the brain was quickly removed, frozen on dry ice, and stored at -80°C until sectioned via cryostat. Brain sections containing the sensorimotor cortex were cut along the rostral-caudal axis using a block advance of 14 μm, immediately mounted on glass slides and dried at room temperature (RT) for 10 minutes, and then stored at -80°C until processed using multi-label fluorescent in situ hybridization procedures.

Fluorescent in situ hybridization probes were designed by Advanced Cell Diagnostics, Inc. (Hayward, CA, USA) to detect mRNA encoding Cox6a2, Slc32a1, and Pvalb. Two sections per animal were processed using the RNAscope® 2.5 Assay as previously described (Wang et al., 2012). Briefly, tissue sections were incubated in a protease treatment for 30 minutes at RT and then the probes were hybridized to their target mRNAs for 2 hours at 40°C. The sections were exposed to a series of incubations at 40°C that amplifies the target probes, and then counterstained with NeuroTrace blue-fluorescent Nissl stain (1:50; Molecular Probes) for 20 minutes at RT. Cox6a2, Pvalb, and Slc32a1 were detected with Alexa Fluor® 488, Atto 550 and Atto 647, respectively.

Data were collected on an Olympus IX83 inverted microscope equipped with a Hamamatsu Orca-Flash4.0 V2 digital CMOS camera using a 60x 1.40 NA SC oil immersion objective. The equipment was controlled by cellSens (Olympus). 3D image stacks (2D images successively captured at intervals separated by 0.25 μm in the z-dimension) that are 1434 x 1434 pixels (155.35 μm x 155.35 μm) were acquired over the entire thickness of the tissue section. The stacks were collected using optimal exposure settings (i.e., those that yielded the greatest dynamic range with no saturated pixels), with differences in exposures normalized before analyses.

Laminar boundaries of the sensorimotor cortex were determined by cytoarchitectonic criteria using NeuroTrace labeling. Fifteen image stacks across the gray matter area spanning from layer 2 to 6 were systematic randomly sampled using a sampling grid of 220 x 220 μm^2, which yielded a total of 30 image stacks per animal. Every NeuroTrace labeled neuron within a 700 x 700 pixels counting frame was included for analyses; the counting frame was placed in the center of each image to ensure that the entire NeuroTrace labeled neuron was in the field of view. The percentage (± standard deviation) of NeuroTrace labeled cells containing Cox6a2 mRNA (Cox6a2+) and that did not contain Slc32a1 mRNA (Slc32a1-), that contained Slc32a1 but not Pvalb mRNA (Slc32a1+/Pvalb-), and that contained both Slc32a1 and Pvalb mRNAs (Slc32a1+/Pvalb+) were manually assessed.

Allen Brain Atlas in situ hybridization (ISH) data:

We downloaded in situ hybridization (ISH) images using the Allen Brain Atlas API (http://help.brain-map.org/display/mousebrain/API). Assessment of expression patterns was done by visual inspection.

Single cell analysis:

Mouse cortex single cell RNA sequencing data were acquired from Zeisel et al. (2015) (available from http://linnarssonlab.org/cortex/, GEO accession no: GSE60361,1691 cells) and Tasic et al. (2016) (GEO accession no: GSE71585, 1809 cells). Human single cell RNA sequencing data were acquired from Darmanis et al. (2015) (GEO accession no: GSE67835, 466 cells). For all studies, pre-processed expression data were encoded in a binary matrix with 1 representing any nonzero value. For all marker gene sets, Spearman's ρ was used to quantify internal correlation. A null distribution was estimated by calculating the internal correlation of 1000 randomly-selected prevalence-matched gene groups. Gene prevalence was defined as the total number of cells with a non-zero expression value for the gene. Prevalence matching was done by choosing a random gene with a prevalence of +/-2.5% of the prevalence of the marker gene. P-values were calculated by comparing the internal correlation of marker gene set to the internal correlations of random gene groups using Wilcoxon rank-sum test.

Pre-processing of microarray data:

All microarray data used in the study were pre-processed and normalized with the "rma" function of the "oligo" or "affy" (Carvalho and Irizarry, 2010) R packages. Probeset to gene annotations were obtained from Gemma (Zoubarev et al., 2012) (http://chibi.ubc.ca/microannots). Any probeset with maximal expression level lower than the median among all probeset signals was removed. Of the remaining probesets that represent the same gene, the one with the highest variance among different cell types was selected for further analysis.

Cell type specific samples that make up the NeuroExpresso database were processed by a modified version of the "rma" function that enabled collective processing of data from Mouse Expression 430A Array (GPL339) and Mouse Genome 430 2.0 Array (GPL1261) that shared 22690 of their probesets. As part of the rma function, the samples are quantile normalized at the probe level. However, possibly due to differences in the purification steps used by different studies (Okaty 2011), we still observed biases in signal distribution among samples originating from different studies. Thus, to increase the comparability across studies, we performed a second quantile normalization of the samples at a probeset level before selection of probes with the highest variance. After all processing the final data set included 11564 genes.

For comparison of marker gene profiles in white matter and frontal cortex, we acquired healthy donor brain region expression profiles from Trabzuni et al. (2013) (GEO accession no: GSE60862).

For estimation of dopaminergic marker gene profiles in Parkinson’s disease patients and healthy donors, we acquired substantia nigra expression profiles from Lesnick et al. (2007) (GSE7621) and Moran et al. (2006) (GSE8397).

Expression data for the Stanley Medical Research Institute (SMRI), which included post-mortem prefrontal cortex samples from bipolar disorder, major depression and schizophrenia patients along with healthy donors, was acquired through https://www.stanleygenomics.org/, study identifier 2.

Estimation of marker gene profiles (MGPs):

For each cell type relevant to the brain region analysed, we used the first principal component of the corresponding marker gene set expression as a surrogate for cell type proportions. Principal component analysis was performed using the “prcomp” function from the “stats” R package, using the “scale = TRUE” option. Since the sign of the loadings of the rotation matrix is arbitrary, we multiplied the loadings by -1 if a majority of the loadings were negative. It is plausible that some marker genes will be transcriptionally regulated under different conditions (e.g. disease state), reducing the correspondence between their expression level with the relative cell proportion. A gene that is regulated will have reduced correlation to the other marker genes with expression levels primarily dictated by cell type proportions which will reduce their loading in the first principal component. To reduce the impact of regulated genes on the estimation process, we removed marker genes from a given analysis if their loadings were negative when calculated based on all samples in the dataset. For visualization purposes, the scores were normalized to the range 0-1.

Two sided Wilcoxon rank-sum test (“wilcox.test” function from the “stats” package in R, default options) was used to compare between the different experimental conditions.

For estimations of cell type MGPs in samples from frontal cortex and white matter from the Trabzuni et al., (2013) study, results were subjected to multiple testing correction by the Benjamini & Hochberg method (Benjamini and Hochberg, 1995).

For the Parkinson’s disease datasets from Moran et al. (2006) and Lesnick et. al. (2007), we estimated MGPs for dopaminergic neuron markers in control and PD subjects. Moran et al. data included samples from two sub-regions of substantia nigra. Since some of the subjects were sampled in only one of the sub-regions while others in both, the two sub-regions were analysed separately.

The list of 22 PD genes, described as the “first PD expression signature”, were taken from Moran et al. (2006). Dopaminergic MGPs were calculated for the samples. Correlations of the PD gene expression as well as correlation of 1st principal component of PD gene expression to dopaminergic MGP was calculated using Spearman’s ρ.

For SMRI collection of psychiatric patients we estimated oligodendrocytes MGPs based on expression data available through SMRI website (as indicated above) and compared our results to experimental cell counts from the same cohort of subjects previously reported by Uranova et al. (2004). Figure 5B representing the oligodendrocyte cell counts in each disease group was adapted from Uranova et al. (2004). The data to re-create the figure was extracted using WebPlotDigitizer (http://arohatgi.info/WebPlotDigitizer/app/) by Ankit Rohatgi.

Results

Compilation of a brain cell type expression database

A key input to our search for marker genes is expression data from purified brain cell types. Expanding on work from Okaty et al., (2011), we assembled and curated a database of cell type-specific expression profiles from published data (see Methods). The database (“NeuroExpresso”) represents 30 major cell types from 12 brain regions (Sample count - number of samples that representing the cell type; Gene count) and a total of 263 samples from 23 published studies, and includes the expression profiles of 11509 genes. We used rigorous quality control steps to identify contaminated samples and outliers (see Methods). All cell types with the exception of ependymal cells are represented by at least 3 replicates and 8/30 cell types are represented by multiple independent studies (**Table 1**). The database is in constant growth as more cell type data becomes available. As more RNAseq data from pooled and single brain cell types will become available, these data will be added to our database. To facilitate access to the data and allow basic analysis we provide a simple search and visualization interface on the web, www.neuroexpresso.org (Figure 1C). The app provides means of visualising gene expression in different brain regions based on the cell type, study or methodology, as well as differential expression analysis between groups of selected samples.

Identification of cell type enriched marker gene sets

We used the NeuroExpresso data to identify marker gene sets (MGS) for each of the 30 cell types. An individual MGS is composed of genes highly enriched in a cell type in the context of a brain region (Figure 2A). Marker genes were selected based on a) fold of change relative to other cell types in the brain region and b) a lack of overlap of expression levels in other cell types (see Methods for details). This approach captured previously known marker genes (e.g. TH for dopaminergic cells (Pickel et al., 1976), Tmem119 for microglia (Bennett et al., 2016)) along with numerous new candidate markers such as Cox6a2 for fast spiking parvalbumin (PV)$^+$ interneurons. Some commonly used marker genes, as well as marker genes previously reported by individual studies included in the database, were not selected by our analysis. For example, we found that the commonly used oligodendrocyte marker Olig1 is also highly expressed in the NeuroExpresso astrocyte samples, and thus it was not selected as an oligodendrocyte marker (Figure 2B). Similarly, Fam114a1 (9130005N14Rik), identified as a marker of fast spiking basket cells by Sugino et al. (2006), is also highly expressed in oligodendrocytes and a subtype of pyramidal cells (Figure 2B). These cell types were not available in the Sugino et al. (2006) study, and thus the lack of specificity of Fam114a1 could not be observed by the authors. In total we identified 1149 marker genes, with 3-139 markers per cell type (**Table 1**). The next sections focus on verification and validation of our proposed markers, using multiple methodologies.

Verification of markers by in-situ hybridization

Two cell types in our database (Purkinje cells of the cerebellum and hippocampal dentate gyrus granule cells) are organized in well-defined anatomical structures readily identified in tissue sections. We exploited this fact to use in-situ hybridization data from the Allen Brain Atlas (http://mouse.brain-map.org) (Sunkin et al., 2013) to verify co-localization of known and novel markers for these two cell types. There was general agreement (19/26 of dentate granule cell markers 27/38 of Purkinje cell markers) that the markers were correctly localized to the corresponding brain structures. Figure 2C shows representative examples for the two cell types.

We independently verified Cox6a2 as a marker of cortical fast spiking PV+ interneurons using triple label in situ hybridization of mouse cortical sections for Cox6a2, Pvalb and Slc32a1 (a pan-GABAergic neuronal marker) transcripts. As expected, we found that approximately 25% of all identified neurons were GABAergic (that is, Slc32a1 positive), while 46% of all GABAergic neurons were also Pvalb positive. 80% of all Cox6a2+ neurons were Pvalb and Slc32a1 positive whereas Cox6a2 expression outside GABAergic cells was very low (1.65% of Cox6a2 positive cells), suggesting high specificity of Cox6a2 to PV+ GABAergic cells (Figure 3).

Further direct validation of individual markers was not feasible. Thus, we turned to additional approaches that support the overall validity of the MGS if not individual markers within them.

Verification of marker gene sets in single-cell RNAseq data

As a further independent validation of our marker gene signatures, we analysed their properties in recently published single cell RNAseq datasets derived from mouse cortex (Tasic et al., 2016; Zeisel et al., 2015) and human cortex (Darmanis et al., 2015). In these studies, the authors used post hoc clustering to propose novel cell types (Poulin et al., 2016). We could not directly compare our MGSs to markers of these cluster-defined cell types because their correspondence to the cell types in NeuroExpresso was not clear. However, since these studies analyse a large number of cells, they are likely to include cells that correspond to the cortical cell types in our database. Thus if our MGSs are cell type specific, and the corresponding cells are present in the single cell data sets, MGS should have a higher than random rate of being co-detected in the same cells, relative to non-marker genes. A weakness of this approach is that a failure to observe a correlation might be due to absence of the cell type in the data set rather than a true shortcoming of the markers.

Overall, all MGSs were successfully validated in at least one of the single cell datasets ($p<0.05$; Table 2). More MGSs were significantly coexpressed in mouse single cell datasets (Tasic et al – 10/10 MGSs; Zeisel et al - 9/10 MGSs) than in the human dataset (Darmanis et al – 5/10 MGSs). While this difference might indicate inter-specie differences, it might be merely an outcome of the substantially smaller number of cells in Darmanis et al dataset, where less abundant cell types may not have been adequately sampled (Table 2).

Neuroexpresso as a tool for understanding the biological diversity and similarity of brain cells

One of the applications of NeuroExpresso is an exploratory tool for exposing functional and biological properties of cell types. For example, high expression of genes involved in GABA synthesis and release (Gad1, Gad2 and Slc32a1) in forebrain cholinergic neurons suggests the capability of these cells to release GABA in addition to their cognate neurotransmitter acetylcholine (Figure 4A). Indeed, co-release of GABA and acetylcholine from forebrain cholinergic cells was recently demonstrated by Saunders et al. (2015). Similarly, the expression of the glutamate transporter Slc17a6, observed in thalamic (habenular) cholinergic cells suggests co-release of glutamate and acetylcholine from these cells, supported by experimental findings by Ren et al. (2011)) (Figure 4A).

A closer analysis of the data revealed interesting expression patterns suggesting previously unknown characteristics of several cell types. For example, one of the novel oligodendrocyte markers identified by our analysis is Ddc (encoding dopa decarboxylase, an enzyme involved in biosynthesis of monoamines) (Figure 4B). Expression of Ddc in oligodendrocytes is consistently high in three studies using different purification techniques and selection markers (Figure 4B), reducing the possibility that this finding results from contamination or a technical artefact. We also observed an unexpected bimodality of gene expression patterns of midbrain dopaminergic cells taken from two studies using similar purification method (Figure 4C), suggesting that the studies represent different sub-populations of this cell type.

Lastly, we found overlap between several markers of spinal cord and brainstem cholinergic cells, and midbrain noradrenergic cells, suggesting previously unknown functional similarity between cholinergic and noradrenergic cell types. The common markers included Chodl, Calca, Cda and Hspb8, shown to be expressed in brainstem cholinergic cells (Enjin et al., 2010) and Phox2b, a known marker of noradrenergic cells (Pattyn et al., 1997).

Marker Gene Profiles can be used to infer changes in cellular proportions in the brain

Marker genes are by definition cell type specific, and thus, changes in their expression observed in bulk tissue data might represent either changes in the number of cells or cell type specific transcriptional changes (or a combination). Expression profiles of marker genes (summarized as the first principal component of their expression, see Methods), can be treated as a surrogate for relative proportion of the corresponding cell type across the samples (Chikina et al., 2015; Westra et al., 2015; Xu et al., 2013). Marker genes of four major classes of brain cell types (namely neurons, astrocytes, oligodendrocytes and microglia) were previously used to gain cell type specific information from brain bulk tissue data (Bowling et al., 2016; Kuhn et al., 2011; Sibille et al., 2008; Skene and Grant, 2016; Tan et al., 2013b). Importantly, since marker genes are likely to correspond to the functional identities of the cell types, it is plausible for their expression to be conserved across different species.

In order to validate the use of MGPs as surrogates for relative cell type proportions, we used bulk tissue expression data from conditions with known changes in cellular proportions. Though we were not able to find a single human brain expression study with complementary cell count data, we were able to work with datasets with known cell type proportion differences.

Firstly, we calculated MGPs for human white matter and frontal cortex using data collected by Trabzuni et al., (2013). Comparing the MGPs in white vs. grey matter, we observed the expected increase in oligodendrocyte MGP, as well as increase in astrocyte and microglia MGPs, corroborating previously reported higher number of these cell types in white vs. grey matter (Ogura et al., 1994; Williams et al., 2013). We also observed decrease in MGPs of all neurons, corroborating the common knowledge of decreased neuronal density in white vs. grey matter (Figure 5A).

We then identified a gene expression dataset from brains of psychiatric patients (study 2 from SMRI microarray database, see methods section) and a separate study involving experimental cell counts of oligodendrocytes from similar brain region in the same cohort of subjects (Uranova et al., 2004). We thus calculated oligodendrocyte MGPs based on the expression data and compared the results to experimental cell counts from Uranova et. al. (2004). Our results demonstrate high similarity between oligodendrocyte MGPs and experimental cell counts at a group level (Figure 5B). Direct comparison between MGP and experimental cell count at a subject level was not possible, as Uranova et al. did not provide the subject identities corresponding to each of the cell count values.

To further assess and demonstrate the ability of MGPs to correctly represent cell type specific changes in neurological conditions, we calculated dopaminergic profiles of substantia nigra samples in two data sets of Parkinson's disease (PD) patients and controls from Moran et al. (2006) (GSE8397) and Lesnick et al. (2007) (GSE7621). We tested whether the well-known loss of dopaminergic cells in PD could be detected using our MGP approach. As expected, in both datasets MGP analysis correctly identified reduction in dopaminergic cells in substantia nigra of Parkinson's disease patients (Figure 5C).

Moran et al. (2006) identified 22 genes consistently differentially expressed in PD patients in their study and in earlier study of PD patients (Zhang et al., 2005), and suggested that these genes can be considered as a "PD expression signature". We hypothesized that differential expression of some of these genes might be better explained by a decrease in dopaminergic cells rather than intracellular regulatory changes. In order to address this issue we performed a principal component analysis of the 22 genes emphasized in Moran et al. (2006), in both the Moran and Lesnick datasets. We then calculated the correlation between the first principal component (PC1) of the 22 genes and dopaminergic MGPs, separately for control and PD subjects. Our results show that in both datasets PC1 of the proposed PD signature genes is highly correlated with dopaminergic

MGPs in both control (Moran: ρ = 0.90; Lesnick: ρ = 0.97) and PD (Moran: ρ = 0.63; Lesnic: ρ = 0.77) subjects (Figure 5D). Examination of individual expression patterns of the proposed signature genes revealed that in both datasets a majority of these genes is positively correlated to the dopaminergic MGPs, regardless of the disease state in both datasets (Figure 5E). These results suggest that majority of the genes identified by Moran et al. represent changes in dopaminergic cell number rather than disease-induced transcriptional changes. This further emphasizes the need to account for cellular changes in gene expression analyses.

Discussion

Cell type specific expression database as a resource for neuroscience

We present NeuroExpresso, a rigorously curated database of cell type specific gene expression data from pooled cell types (www.neuroexpresso.org). To our knowledge, this is the most comprehensive database of brain cell type expression data. The database will be expanded as more data become available and in the future will integrate RNAseq data from pooled and single cell samples.

NeuroExpresso allows simultaneous examination of gene expression numerous cell types across different brain regions. This approach promotes discovery of cellular properties that might have otherwise been unnoticed or overlooked when using gene-by-gene approaches or pathway enrichment analysis. For example, a simple examination of expression of genes involved in biosynthesis and secretion of GABA and glutamate, suggested the co-release of these neurotransmitters from forebrain and habenular cholinergic cells, respectively.

Studies that aim to identify novel properties of cell types can benefit from our database as an inexpensive and convenient way to seek novel patterns of gene expression. For instance, our database shows significant bimodality of gene expression in dopaminergic cell types from the midbrain (Figure 4C). The observed bimodality might indicate heterogeneity in the dopaminergic cell population, which could prove a fruitful avenue for future investigation. Another interesting finding from NeuroExpresso is the previously unknown overlap of several markers of motor cholinergic and noradrenergic cells. While the overlapping markers were previously shown to be expressed in spinal cholinergic cells, to our knowledge their expression in noradrenergic (as well as brain stem cholinergic) cells was previously unknown.

NeuroExpresso can be also used to facilitate interpretation of genomics and transcriptomics studies. Recently (Pantazatos et al., 2016) used an early release of the databases to interpret expression patterns in the cortex of suicide victims, suggesting involvement of microglia. Moreover, this database has further applications beyond the use of marker genes, such as interpreting other features of brain cell diversity, like in electrophysiological profiles (Tripathy et al., in preparation).

Importantly, NeuroExpresso is a cross-laboratory database. A consistent result observed across several studies raises the certainty that it represents a true biological finding rather than merely an artefact or contamination with other cell types. This is specifically important for findings such as the expression of Ddc in oligodendrocytes (Figure 4B). This surprising result is suggestive of a previously unknown ability of oligodendrocytes to produce monoamine neurotransmitters upon exposure to appropriate precursor, as previously reported for several populations of cells in the brain (Ren et al., 2016; Ugrumov, 2013). Alternatively, this finding might indicate a previously unknown function of Ddc.

Validation of cell type markers

In situ hybridization results that we generated and those made available by the Allen Institute for Brain Sciences (Sunkin et al., 2013) validated a subset of our results (Cox6a2 as a marker of fast spiking basket cells and Purkinje and dentate granule cell markers). Further validation was performed with computational methods in independent single cell datasets from mouse and human. This analysis revealed some limitations in generalizing mouse markers identified in mouse cell types to a human context. The differences between mouse and human datasets might indicate evolutionary changes in the function of the homologous genes or functional differences between the human and mouse cell types (Shay et al., 2013; Zhang et al., 2016). However, since most MGSs did validate between mouse and human data, it suggests that most marker genes preserve their specificity despite cross-species gene expression differences. An additional caveat with using single cell data sets for validation is that rare cell types may not have been adequately sampled. In particular, the dataset from Darmanis et al., representing single cell expression from human cortex, may be most prone to this problem of cell type rarity as the number of cells in the dataset are relatively low (466 cells vs 1809 cells from Tasic et al. and 1691 cells from Zeisel et al.). While there is no substantial difference in the total number of cells between the two mouse datasets, the cre-line based sampling approach used by Tasic et al was designed to target relatively rare neuron types (Tasic et al., 2016). As sample sizes for these single cells datasets continue to grow, and in particular those from human brain tissue, we expect that it will be easier to directly resolve this potential inconsistency.

Estimation of marker gene profiles in bulk tissue

Marker genes can assist with the interpretation of bulk tissue data in the form of marker gene profiles (MGPs). A parsimonious interpretation of a change in an MGP is a change in the relative abundance of the corresponding cell type, when compared to other samples. Because our approach focuses on the overall trend of MGS expression levels, it should be relatively insensitive to within-cell-type expression changes for a subset of genes. Still, we prefer to refer to "MGP expression" rather than "cell type proportions".

Our results show that MGPs based on NeuroExpresso marker gene sets (MGSs) can reliably recapitulate relative changes in cell type abundance across different conditions. Direct validation of cell count estimation based on MGSs in human brain was not feasible due to the unavailability of cell counts coupled with expression data. Instead, we compared oligodendrocyte MGPs based on a gene expression dataset available through the SMRI database to experimental cell counts taken from a separate study (Uranova et al., 2004) of the same cohort of subjects. While we were unable to match the samples one by one between the two studies, we show high similarity between oligodendrocyte MGPs and cell counts across the experimental groups.

Based on analysis of dopaminergic MGPs we were able to capture the well-known reduction in dopaminergic cell types in PD patients. Moreover, we show that multiple genes previously reported as differentially expressed in PD are highly correlated with dopaminergic MGPs in both control and PD subjects. This high correlation suggests that the observed differential expression of these genes might merely reflect changes in dopaminergic cell populations rather than PD related regulatory changes

Limitations and caveats

The NeuroExpresso database was assembled from multiple datasets, based on different mouse strains and cell type extraction methodologies. We attempted to address some of the inter-study variability by combined pre-processing of the raw data and quantile normalization of the probesets for each sample in the database. However, due to insufficient overlap between cell types represented by different studies, many of the potential confounding factors such as age, sex and methodology could not be explicitly corrected for. Thus, it is likely that some of the expression values in NeuroExpresso may be affected by confounding factors. While our confidence in the data is increased when expression signals are robust across multiple studies, many of the cell types in NeuroExpresso are represented by a single study. Hence, we advise that small differences in expression between cell types as well as previously unknown expression patterns based on a single data source should be treated with caution. In our analyses, we address these issues by enforcing a stringent set of criteria for the marker selection process, reducing the impact of outlier samples and ignoring small changes in gene expression.

An additional limitation of our study is that the representation for many of the brain cell types is still lacking in the NeuroExpresso database. Therefore, despite our considerable efforts to ensure cell type-specificity of the marker genes, we cannot rule out the possibility that some of them are also expressed in one or more of the non-represented cell types.

In summary, we believe that NeuroExpresso is an extremely valuable resource for neuroscientists. We identified numerous novel markers for 30 major cell types and used them to estimate cell type profiles in bulk tissue data, demonstrating high correlation between our estimates and experiment-based cell counts. This approach can be used to reveal cell type specific changes in whole tissue samples and to re-evaluate previous analyses on brain whole tissues that might be biased by cell type-specific changes. Information about cell type-specific changes is likely to be very valuable since conditions like neuron death, inflammation, and astrogliosis are common hallmarks of in neurological diseases.

Acknowledgements

This work is supported by a NeuroDevNet grant to PP, the UBC bioinformatics graduate training program (BOM), a CIHR post-doctoral fellowship to ST, by the Campbell Family Mental Health Research Institute of CAMH (ES and BR), NIH grants MH077159 to ES, and MH111099 and GM076990 to PP. We thank Ken Sugino for providing access to raw CEL files for Purkinje and TH$^+$ cells from locus coeruleus, Chee Yeun Chung for providing access to raw CEL dopaminergic cells, Dean Attali for providing insight on the usage of the Shiny platform, the Pavlidis lab for their inputs to the project during its development and Rosemary McCloskey for aid in editing the manuscript.

The authors declare no competing financial interests

References

Anandasabapathy, N., Victora, G.D., Meredith, M., Feder, R., Dong, B., Kluger, C., Yao, K., Dustin, M.L., Nussenzweig, M.C., Steinman, R.M., et al. (2011). Flt3L controls the development of radiosensitive dendritic cells in the meninges and choroid plexus of the steady-state mouse brain. J. Exp. Med. *208*, 1695–1705.

Beckervordersandforth, R., Tripathi, P., Ninkovic, J., Bayam, E., Lepier, A., Stempfhuber, B., Kirchhoff, F., Hirrlinger, J., Haslinger, A., Lie, D.C., et al. (2010). In Vivo Fate Mapping and Expression Analysis Reveals Molecular Hallmarks of Prospectively Isolated Adult Neural Stem Cells. Cell Stem Cell *7*, 744–758.

Bellesi, M., Pfister-Genskow, M., Maret, S., Keles, S., Tononi, G., and Cirelli, C. (2013). Effects of Sleep and Wake on Oligodendrocytes and Their Precursors. J. Neurosci. *33*, 14288–14300.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. *57*, 289–300.

Bennett, M.L., Bennett, F.C., Liddelow, S.A., Ajami, B., Zamanian, J.L., Fernhoff, N.B., Mulinyawe, S.B., Bohlen, C.J., Adil, A., Tucker, A., et al. (2016). New tools for studying microglia in the mouse and human CNS. Proc. Natl. Acad. Sci. U. S. A. *113*, E1738-1746.

Bowling, K., Ramaker, R.C., Lasseigne, B.N., Hagenauer, M., Hardigan, A., Davis, N., Gertz, J., Cartagena, P., Walsh, D., Vawter, M., et al. (2016). Post-mortem molecular profiling of three psychiatric disorders reveals widespread dysregulation of cell-type associated transcripts and refined disease-related transcription changes. bioRxiv 61416.

Cahoy, J.D., Emery, B., Kaushal, A., Foo, L.C., Zamanian, J.L., Christopherson, K.S., Xing, Y., Lubischer, J.L., Krieg, P.A., Krupenko, S.A., et al. (2008). A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function. J. Neurosci. *28*, 264–278.

Carvalho, B.S., and Irizarry, R.A. (2010). A framework for oligonucleotide microarray preprocessing. Bioinforma. Oxf. Engl. *26*, 2363–2367.

Celio, M.R., and Heizmann, C.W. (1981). Calcium-binding protein parvalbumin as a neuronal marker. Nature *293*, 300–302.

Chikina, M., Zaslavsky, E., and Sealfon, S.C. (2015). CellCODE: A robust latent variable approach to differential expression analysis for heterogeneous cell populations. Bioinformatics btv015.

Chung, C.Y., Seo, H., Sonntag, K.C., Brooks, A., Lin, L., and Isacson, O. (2005). Cell type-specific gene expression of midbrain dopaminergic neurons reveals molecules involved in their vulnerability and protection. Hum. Mol. Genet. *14*, 1709–1725.

Dalal, J., Roh, J.H., Maloney, S.E., Akuffo, A., Shah, S., Yuan, H., Wamsley, B., Jones, W.B., Strong, C. de G., Gray, P.A., et al. (2013). Translational profiling of hypocretin neurons identifies candidate molecules for sleep regulation. Genes Dev. *27*, 565–578.

Darmanis, S., Sloan, S.A., Zhang, Y., Enge, M., Caneda, C., Shuer, L.M., Gephart, M.G.H., Barres, B.A., and Quake, S.R. (2015). A survey of human brain transcriptome diversity at the single cell level. Proc. Natl. Acad. Sci. *112*, 7285–7290.

Dougherty, J.D., and Geschwind, D.H. (2005). Progress in Realizing the Promise of Microarrays in Systems Neurobiology. Neuron *45*, 183–185.

Dougherty, J.D., Maloney, S.E., Wozniak, D.F., Rieger, M.A., Sonnenblick, L., Coppola, G., Mahieu, N.G., Zhang, J., Cai, J., Patti, G.J., et al. (2013). The Disruption of Celf6, a Gene Identified by Translational Profiling of Serotonergic Neurons, Results in Autism-Related Behaviors. J. Neurosci. *33*, 2732–2753.

Doyle, J.P., Dougherty, J.D., Heiman, M., Schmidt, E.F., Stevens, T.R., Ma, G., Bupp, S., Shrestha, P., Shah, R.D., Doughty, M.L., et al. (2008). Application of a Translational Profiling Approach for the Comparative Analysis of CNS Cell Types. Cell *135*, 749–762.

Enjin, A., Rabe, N., Nakanishi, S.T., Vallstedt, A., Gezelius, H., Memic, F., Lind, M., Hjalt, T., Tourtellotte, W.G., Bruder, C., et al. (2010). Identification of novel spinal cholinergic genetic subtypes disclose Chodl and Pitx2 as markers for fast motor neurons and partition cells. J. Comp. Neurol. *518*, 2284–2304.

Fomchenko, E.I., Dougherty, J.D., Helmy, K.Y., Katz, A.M., Pietras, A., Brennan, C., Huse, J.T., Milosevic, A., and Holland, E.C. (2011). Recruited Cells Can Become Transformed and Overtake PDGF-Induced Murine Gliomas In Vivo during Tumor Progression. PLoS ONE *6*, e20605.

Galloway, J.N., Shaw, C., Yu, P., Parghi, D., Poidevin, M., Jin, P., and Nelson, D.L. (2014). CGG repeats in RNA modulate expression of TDP-43 in mouse and fly models of fragile X tremor ataxia syndrome. Hum. Mol. Genet. ddu314.

Görlich, A., Antolin-Fontes, B., Ables, J.L., Frahm, S., Ślimak, M.A., Dougherty, J.D., and Ibañez-Tallon, I. (2013). Reexposure to nicotine during withdrawal increases the pacemaking activity of cholinergic habenular neurons. Proc. Natl. Acad. Sci. *110*, 17077–17082.

Habib, N., Li, Y., Heidenreich, M., Swiech, L., Avraham-Davidi, I., Trombetta, J.J., Hession, C., Zhang, F., and Regev, A. (2016). Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. Science aad7038.

Halliwell, B. (2003). Oxidative stress in cell culture: an under-appreciated problem? FEBS Lett. *540*, 3–6.

Handley, A., Schauer, T., Ladurner, A.G., and Margulies, C.E. (2015). Designing Cell-Type-Specific Genome-wide Experiments. Mol. Cell *58*, 621–631.

Heiman, M., Heilbut, A., Francardo, V., Kulicke, R., Fenster, R.J., Kolaczyk, E.D., Mesirov, J.P., Surmeier, D.J., Cenci, M.A., and Greengard, P. (2014). Molecular adaptations of striatal spiny projection neurons during levodopa-induced dyskinesia. Proc. Natl. Acad. Sci. *111*, 4578–4583.

Holtman, I.R., Noback, M., Bijlsma, M., Duong, K.N., van der Geest, M.A., Ketelaars, P.T., Brouwer, N., Vainchtein, I.D., Eggen, B.J.L., and Boddeke, H.W.G.M. (2015). Glia Open Access Database (GOAD): A comprehensive gene expression encyclopedia of glia cells in health and disease. Glia n/a-n/a.

Hu, H., Gan, J., and Jonas, P. (2014). Fast-spiking, parvalbumin+ GABAergic interneurons: From cellular design to microcircuit function. Science *345*, 1255263.

Januszyk, M., Rennert, R.C., Sorkin, M., Maan, Z.N., Wong, L.K., Whittam, A.J., Whitmore, A., Duscher, D., and Gurtner, G.C. (2015). Evaluating the Effect of Cell Culture on Gene Expression in Primary Tissue Samples Using Microfluidic-Based Single Cell Transcriptional Analysis. Microarrays *4*, 540–550.

Kawaguchi, Y., Katsumaru, H., Kosaka, T., Heizmann, C.W., and Hama, K. (1987). Fast spiking cells in rat hippocampus (CA1 region) contain the calcium-binding protein parvalbumin. Brain Res. *416*, 369–374.

Kuhn, A., Thu, D., Waldvogel, H.J., Faull, R.L.M., and Luthi-Carter, R. (2011). Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. Nat. Methods *8*, 945–947.

Lake, B.B., Ai, R., Kaeser, G.E., Salathia, N.S., Yung, Y.C., Liu, R., Wildberg, A., Gao, D., Fung, H.-L., Chen, S., et al. (2016). Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science *352*, 1586–1590.

Lesnick, T.G., Papapetropoulos, S., Mash, D.C., Ffrench-Mullen, J., Shehadeh, L., de Andrade, M., Henley, J.R., Rocca, W.A., Ahlskog, J.E., and Maraganore, D.M. (2007). A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. PLoS Genet. *3*, e98.

Lobo, M.K., Karsten, S.L., Gray, M., Geschwind, D.H., and Yang, X.W. (2006). FACS-array profiling of striatal projection neuron subtypes in juvenile and adult mouse brains. Nat. Neurosci. *9*, 443–452.

Maechler, M., original), P.R. (Fortran, original), A.S. (S, original), M.H. (S, maintenance(1999-2000)), K.H. (port to R., Studer, M., and Roudier, P. (2016). cluster: “Finding Groups in Data”: Cluster Analysis Extended Rousseeuw et al.

Margolis, E.B., Lock, H., Hjelmstad, G.O., and Fields, H.L. (2006). The ventral tegmental area revisited: is there an electrophysiological marker for dopaminergic neurons? J. Physiol. *577*, 907–924.

Maze, I., Chaudhury, D., Dietz, D.M., Von Schimmelmann, M., Kennedy, P.J., Lobo, M.K., Sillivan, S.E., Miller, M.L., Bagot, R.C., Sun, H., et al. (2014). G9a influences neuronal subtype specification in striatum. Nat. Neurosci. *17*, 533–539.

Moran, L.B., Duke, D.C., Deprez, M., Dexter, D.T., Pearce, R.K.B., and Graeber, M.B. (2006). Whole genome expression profiling of the medial and lateral substantia nigra in Parkinson's disease. Neurogenetics *7*, 1–11.

Ogura, K., Ogawa, M., and Yoshida, M. (1994). Effects of ageing on microglia in the normal rat brain: immunohistochemical observations. Neuroreport *5*, 1224–1226.

Okaty, B.W., Miller, M.N., Sugino, K., Hempel, C.M., and Nelson, S.B. (2009). Transcriptional and electrophysiological maturation of neocortical fastspiking GABAergic interneurons. J. Neurosci. Off. J. Soc. Neurosci. *29*, 7040–7052.

Okaty, B.W., Sugino, K., and Nelson, S.B. (2011). A Quantitative Comparison of Cell-Type-Specific Microarray Gene Expression Profiling Methods in the Mouse Brain. PLoS ONE *6*, e16493.

Pantazatos, S.P., Huang, Y.-Y., Rosoklija, G.B., Dwork, A.J., Arango, V., and Mann, J.J. (2016). Whole-transcriptome brain expression and exon-usage profiling in major depression and suicide: evidence for altered glial, endothelial and ATPase activity. Mol. Psychiatry.

Pattyn, A., Morin, X., Cremer, H., Goridis, C., and Brunet, J.F. (1997). Expression and interactions of the two closely related homeobox genes Phox2a and Phox2b during neurogenesis. Dev. Camb. Engl. *124*, 4065–4075.

Paul, A., Cai, Y., Atwal, G.S., and Huang, Z.J. (2012). Developmental coordination of gene expression between synaptic partners during GABAergic circuit assembly in cerebellar cortex. Front. Neural Circuits *6*, 37.

Perrone-Bizzozero, N.I., Tanner, D.C., Mounce, J., and Bolognani, F. (2011). Increased Expression of Axogenesis-Related Genes and Mossy Fibre Length in Dentate Granule Cells from Adult HuD Overexpressor Mice. ASN Neuro *3*, AN20110015.

Petilla Interneuron Nomenclature Group, Ascoli, G.A., Alonso-Nanclares, L., Anderson, S.A., Barrionuevo, G., Benavides-Piccione, R., Burkhalter, A., Buzsáki, G., Cauli, B., Defelipe, J., et al. (2008). Petilla terminology: nomenclature of features of GABAergic interneurons of the cerebral cortex. Nat. Rev. Neurosci. *9*, 557–568.

Phani, S., Gonye, G., and Iacovitti, L. (2010). VTA neurons show a potentially protective transcriptional response to MPTP. Brain Res. *1343*, 1–13.

Pickel, V.M., Joh, T.H., and Reis, D.J. (1976). Monoamine-synthesizing enzymes in central dopaminergic, noradrenergic and serotonergic neurons. Immunocytochemical localization by light and electron microscopy. J. Histochem. Cytochem. *24*, 792–792.

Poulin, J.-F., Tasic, B., Hjerling-Leffler, J., Trimarchi, J.M., and Awatramani, R. (2016). Disentangling neural cell diversity using single-cell transcriptomics. Nat. Neurosci. *19*, 1131–1141.

Ren, J., Qin, C., Hu, F., Tan, J., Qiu, L., Zhao, S., Feng, G., and Luo, M. (2011). Habenula “Cholinergic” Neurons Corelease Glutamate and Acetylcholine and Activate Postsynaptic Neurons via Distinct Transmission Modes. Neuron *69*, 445–452.

Ren, L., Wienecke, J., Hultborn, H., and Zhang, M. (2016). Production of dopamine by aromatic L-amino acid decarboxylase cells after spinal cord injury. J. Neurotrauma.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

Rossner, M.J., Hirrlinger, J., Wichert, S.P., Boehm, C., Newrzella, D., Hiemisch, H., Eisenhardt, G., Stuenkel, C., Ahsen, O. von, and Nave, K.-A. (2006). Global Transcriptome Analysis of Genetically Identified Neurons in the Adult Cortex. J. Neurosci. *26*, 9956–9966.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. *20*, 53–65.

Saunders, A., Granger, A.J., and Sabatini, B.L. (2015). Corelease of acetylcholine and GABA from cholinergic forebrain neurons. eLife *4*.

Schmidt, E.F., Warner-Schmidt, J.L., Otopalik, B.G., Pickett, S.B., Greengard, P., and Heintz, N. (2012). Identification of the Cortical Neurons that Mediate Antidepressant Responses. Cell *149*, 1152–1163.

Shay, T., Jojic, V., Zuk, O., Rothamel, K., Puyraimond-Zemmour, D., Feng, T., Wakamatsu, E., Benoist, C., Koller, D., Regev, A., et al. (2013). Conservation and divergence in the transcriptional programs of the human and mouse immune systems. Proc. Natl. Acad. Sci. U. S. A. *110*, 2946–2951.

Shrestha, P., Mousa, A., and Heintz, N. (2015). Layer 2/3 pyramidal cells in the medial prefrontal cortex moderate stress induced depressive behaviors. eLife *4*.

Sibille, E., Arango, V., Joeyen-Waldorf, J., Wang, Y., Leman, S., Surget, A., Belzung, C., Mann, J.J., and Lewis, D.A. (2008). Large-scale estimates of cellular origins of mRNAs: enhancing the yield of transcriptome analyses. J. Neurosci. Methods *167*, 198–206.

Skene, N.G., and Grant, S.G.N. (2016). Identification of Vulnerable Cell Types in Major Brain Disorders Using Single Cell Transcriptomes and Expression Weighted Cell Type Enrichment. Front. Neurosci. *10*.

Sugino, K., Hempel, C.M., Miller, M.N., Hattox, A.M., Shapiro, P., Wu, C., Huang, Z.J., and Nelson, S.B. (2006). Molecular taxonomy of major neuronal classes in the adult mouse forebrain. Nat. Neurosci. *9*, 99–107.

Sugino, K., Hempel, C.M., Okaty, B.W., Arnson, H.A., Kato, S., Dani, V.S., and Nelson, S.B. (2014). Cell-Type-Specific Repression by Methyl-CpG-Binding Protein 2 Is Biased toward Long Genes. J. Neurosci. *34*, 12877–12883.

Sunkin, S.M., Ng, L., Lau, C., Dolbeare, T., Gilbert, T.L., Thompson, C.L., Hawrylycz, M., and

Dang, C. (2013). Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic Acids Res. *41*, D996–D1008.

Tan, C.L., Plotkin, J.L., Venø, M.T., Schimmelmann, M. von, Feinberg, P., Mann, S., Handler, A., Kjems, J., Surmeier, D.J., O'Carroll, D., et al. (2013a). MicroRNA-128 governs neuronal excitability and motor behavior in mice. Science *342*, 1254–1258.

Tan, P.P.C., French, L., and Pavlidis, P. (2013b). Neuron-Enriched Gene Expression Patterns are Regionally Anti-Correlated with Oligodendrocyte-Enriched Patterns in the Adult Mouse and Human Brain. Front. Neurosci. *7*, 5.

Tasic, B., Menon, V., Nguyen, T.N., Kim, T.K., Jarsky, T., Yao, Z., Levi, B., Gray, L.T., Sorensen, S.A., Dolbeare, T., et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat. Neurosci. *19*, 335–346.

Tomomura, M., Rice, D.S., Morgan, J.I., and Yuzaki, M. (2001). Purification of Purkinje cells by fluorescence-activated cell sorting from transgenic mice that express green fluorescent protein. Eur. J. Neurosci. *14*, 57–63.

Trabzuni, D., Ramasamy, A., Imran, S., Walker, R., Smith, C., Weale, M.E., Hardy, J., Ryten, M., and North American Brain Expression Consortium (2013). Widespread sex differences in gene expression and splicing in the adult human brain. Nat. Commun. *4*.

Ugrumov, M.V. (2013). Chapter Four - Brain Neurons Partly Expressing Dopaminergic Phenotype: Location, Development, Functional Significance, and Regulation. In Advances in Pharmacology, L.E. Eiden, ed. (Academic Press), pp. 37–91.

Uranova, N.A., Vostrikov, V.M., Orlovskaya, D.D., and Rachmanova, V.I. (2004). Oligodendroglial density in the prefrontal cortex in schizophrenia and mood disorders: a study from the Stanley Neuropathology Consortium. Schizophr. Res. *67*, 269–275.

Wang, Y., Winters, J., and Subramaniam, S. (2012). Functional classification of skeletal muscle networks. II. Applications to pathophysiology. J. Appl. Physiol. Bethesda Md 1985 *113*, 1902–1920.

Westra, H.-J., Arends, D., Esko, T., Peters, M.J., Schurmann, C., Schramm, K., Kettunen, J., Yaghootkar, H., Fairfax, B.P., Andiappan, A.K., et al. (2015). Cell Specific eQTL Analysis without Sorting Cells. PLoS Genet *11*, e1005223.

Williams, M.R., Hampton, T., Pearce, R.K.B., Hirsch, S.R., Ansorge, O., Thom, M., and Maier, M. (2013). Astrocyte decrease in the subgenual cingulate and callosal genu in schizophrenia. Eur. Arch. Psychiatry Clin. Neurosci. *263*, 41–52.

Xu, X., Nehorai, A., and Dougherty, J.D. (2013). Cell type-specific analysis of human brain transcriptome data to predict alterations in cellular composition. Syst. Biomed. *1*, 151–160.

Zamanian, J.L., Xu, L., Foo, L.C., Nouri, N., Zhou, L., Giffard, R.G., and Barres, B.A. (2012). Genomic Analysis of Reactive Astrogliosis. J. Neurosci. *32*, 6391–6410.

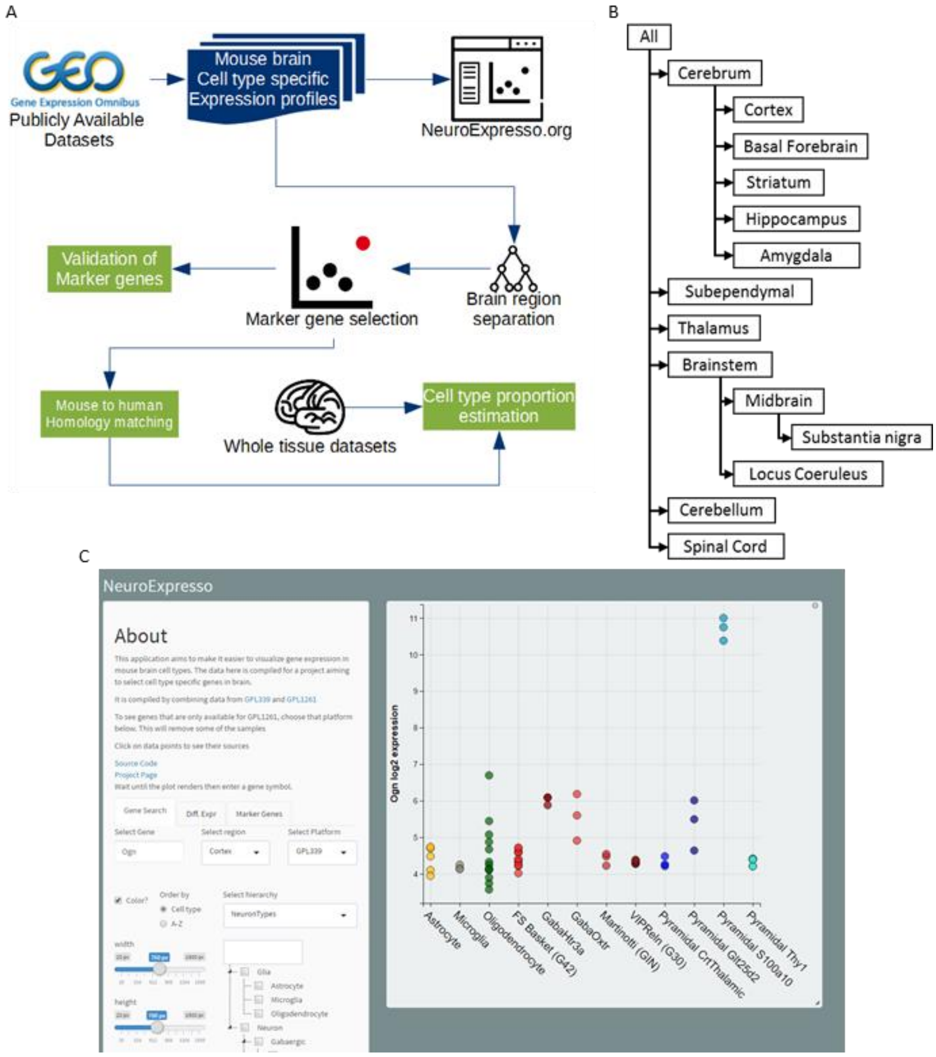
Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., Manno, G.L., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142.

Zhang, Y., James, M., Middleton, F.A., and Davis, R.L. (2005). Transcriptional analysis of multiple brain regions in Parkinson’s disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms. Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet. *137B*, 5–16.

Zhang, Y., Sloan, S.A., Clarke, L.E., Caneda, C., Plaza, C.A., Blumenthal, P.D., Vogel, H., Steinberg, G.K., Edwards, M.S.B., Li, G., et al. (2016). Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. Neuron *89*, 37–53.

Zoubarev, A., Hamer, K.M., Keshav, K.D., McCarthy, E.L., Santos, J.R.C., Van Rossum, T., McDonald, C., Hall, A., Wan, X., Lim, R., et al. (2012). Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. Bioinforma. Oxf. Engl. *28*, 2272–2273.

Figures



About

This application aims to make it easier to visualize gene expression in mouse brain cell types. The data here is compiled for a project aiming to select cell type specific genes in brain.

It is compiled by combining data from GPL339 and GPL1261

To see genes that are only available for GPL1261, choose that platform below. This will remove some of the samples

Click on data points to see their sources

Source Code
Project Page
Wait until the plot renders then enter a gene symbol.

Select Gene: Ogn
Select region: Cortex
Select Platform: GPL339
☑ Color?
Order by: ● Cell type ○ A-Z
Select hierarchy: NeuronTypes

Fig 1: Mouse brain cell type specific expression database compiled from publically available datasets. **(A)** Workflow of the study. Cell type specific expression profiles are collected from publically available datasets and personal communications. Acquired samples are grouped based on cell type and brain region. Marker genes are selected per brain region for all cell types. Marker genes are biologically and computationally validated and used in estimation of cell type proportions. **(B)** Brain region hierarchy used in the study. Samples included in a brain region based on the region they were extracted from. For instance, dopaminergic cells isolated from the midbrain were included when selecting marker genes in the context of brainstem and whole brain. Microglia extracted from whole brain isolates were added to all brain regions. **(C)** A web application is provided for easy visualization of the NeuroExpresso database. The application allows easy visualization of gene expression across cell types in brain regions. It also allows grouping of

related cell types and allows access to the original source of the samples. The application can be reached at www.neuroexpresso.org. Icons in A from www.flaticon.com

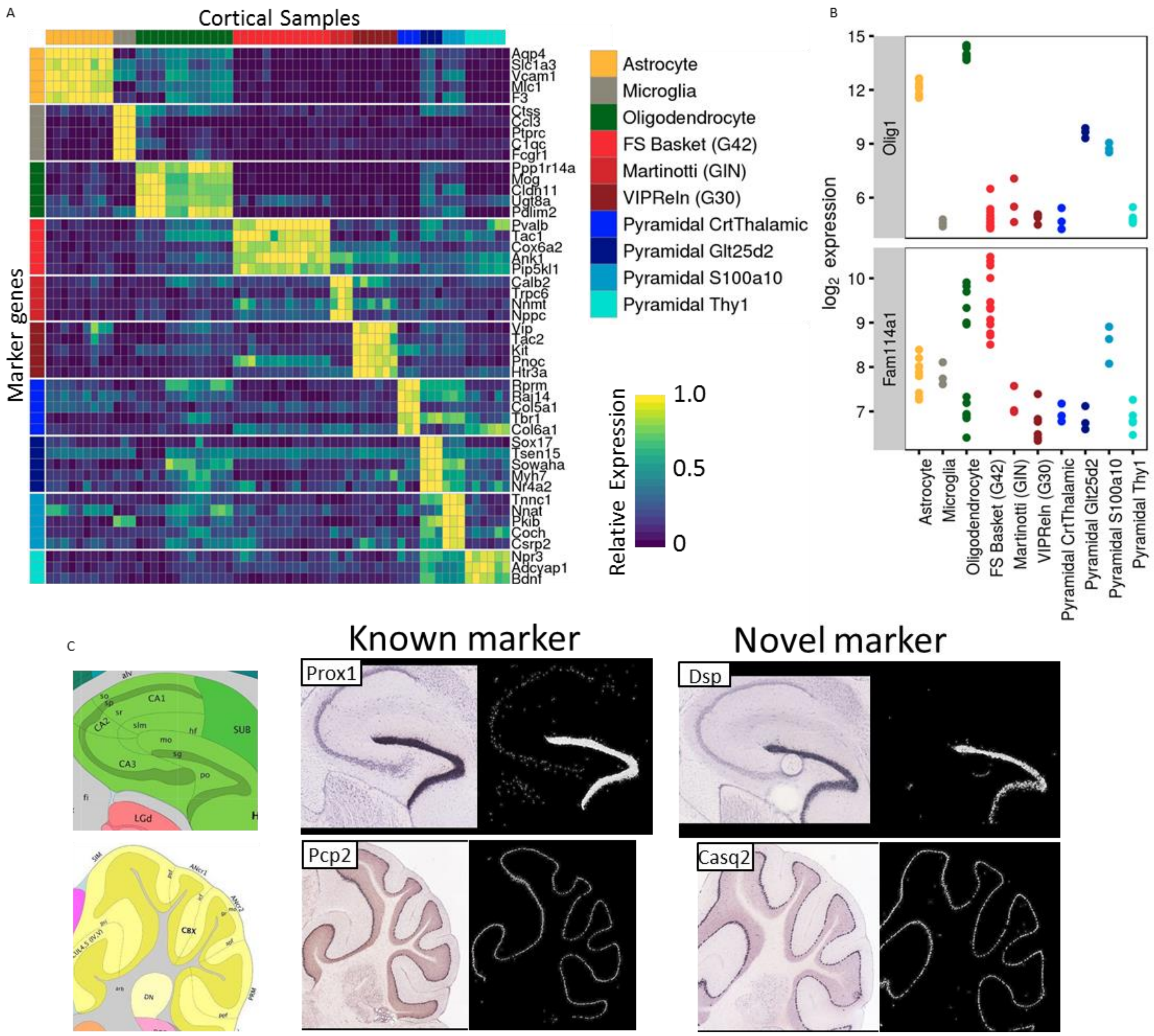


Fig 2: Marker genes are selected for mouse brain cell types and used to estimate cell type profiles. **(A)** Expression of top marker genes in the neo cortex. Expression levels are normalized per gene to be between 0-1. **(B)** Expression of Olig1 and Fam114a1 in frontal cortex. Olig1 is a commonly used oligodendrocyte marker. It was not selected with our method due to its high expression in astrocytes. Fam114a1 is a proposed fast spiking basket cell marker. It was not selected as a marker in this study due to its high expression in oligodendrocytes and S100a10 expressing

pyramidal cells that were both absent from the original study. **(C)** In situ hybridization images from the Allen Brain Atlas. Rightmost panels show the location of the image in the brain according to the Allen Brain mouse reference atlas. Panels on the left show the ISH image and normalized expression level of known and novel dentate granule (upper panels) and Purkinje cell (lower panels) markers.

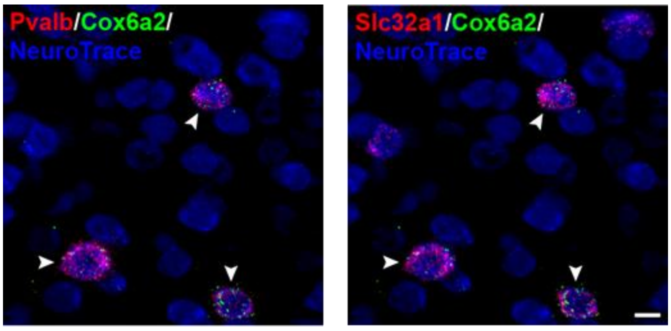


Fig 3: Single-plane image of mouse sensorimotor cortex labeled for Pvalb, Slc32a1, and Cox6a2 mRNAs and counterstained with NeuroTrace. Arrows indicate Cox6a2+ neurons. Bar = 10 µm.

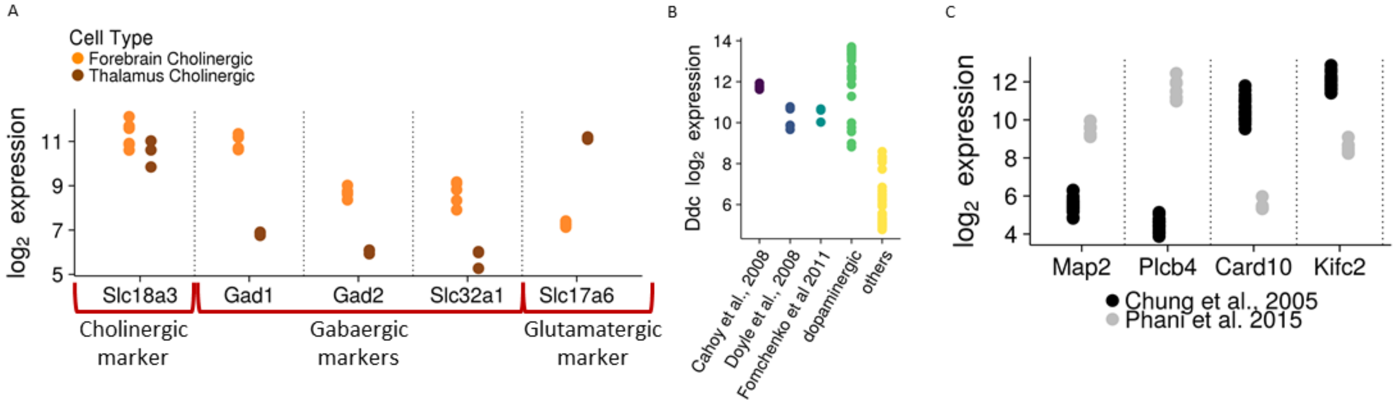


Fig 4: NeuroExpresso reveals gene expression patterns. **(A)** Expression of cholinergic, GABAergic and glutamatergic markers in cholinergic cells from forebrain and thalamus. Forebrain cholinergic neurons express GABAergic markers while thalamus (hubenular) cholinergic neurons express glutamatergic markers. **(B)** Expression of Ddc in oligodendrocyte samples from Cahoy et al., Doyle et al. and Fomchenko et al. datasets and in comparison to dopaminergic cells and other (non-oligodendrocyte) cell types from the frontal cortex. In all three datasets expression of Ddc in

oligodendrocytes is comparable to expression in dopaminergic cells and is higher than in any of the other cortical cells. Oligodendrocyte samples show higher than background levels of expression across datasets. **(C)** Bimodal gene expression in two dopaminergic cell isolates by different labs. Genes shown are labeled as marker genes in the context of midbrain if the two cell isolates are labeled as different cell types.

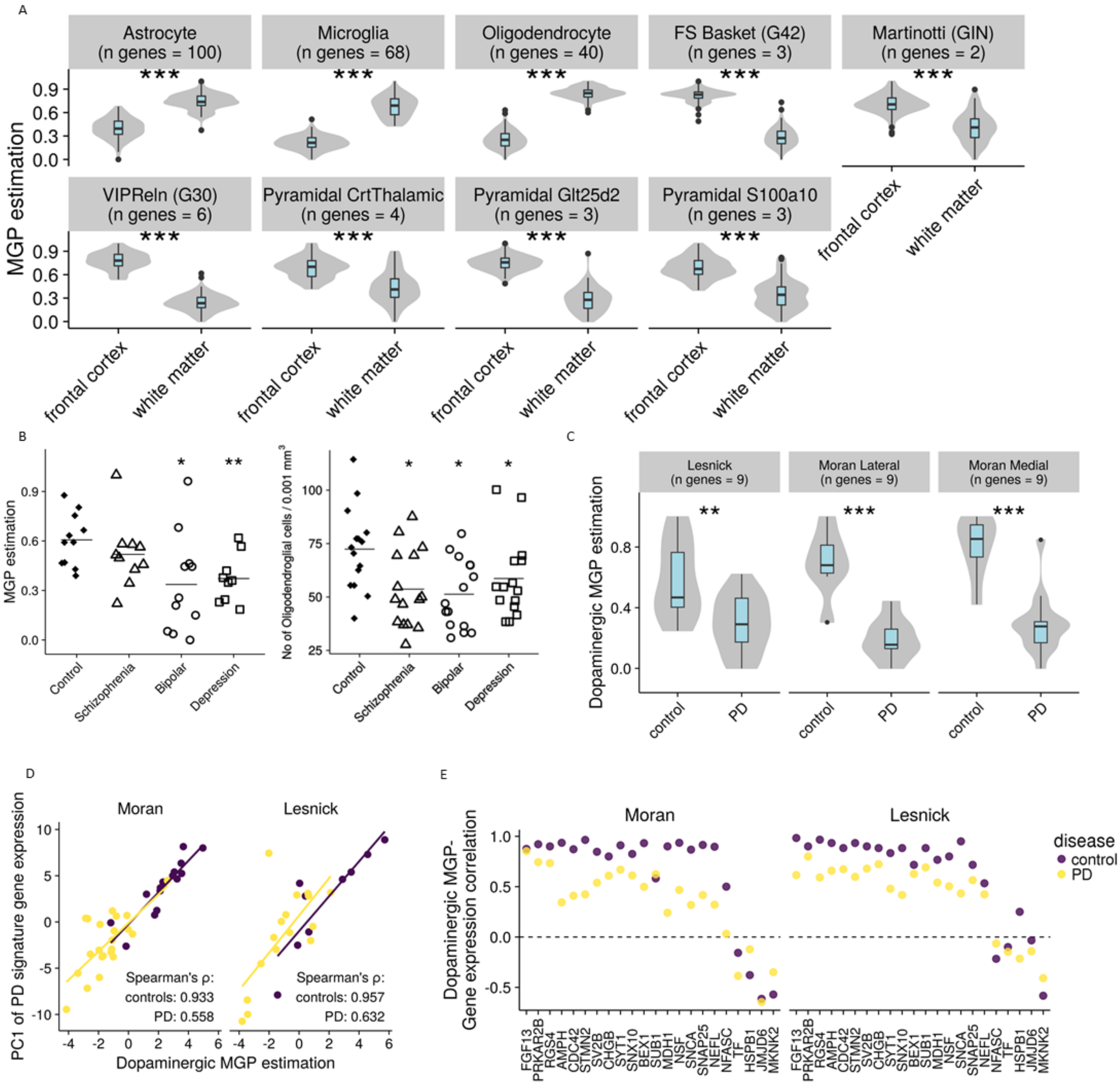


Fig 5: Marker gene profiles reveal cell type specific changes in whole tissue data. **(A)** Estimation of cell type profiles for cortical cells in frontal cortex and white matter. Values are normalized to be

between 0 and 1. (*p <0.05, **p<0.01, ***p<0.001). **(B)** Left: Oligodendrocyte MGPs in Stanley C cohort. Right: Morphology based oligodendrocyte counts of Stanley C cohort. Figure adapted from Uranova et. al. (2004). **(C)** Estimations of dopaminergic cell MGPs in substantia nigra of controls and Parkinson's disease patients. Values are relative and are normalized to be between 0 and 1 and are not reflective of absolute proportions. Dopaminergic cell loss is an expected consequence of Parkinson's Disease. (*p <0.05, ***p<0.001). **(D)** Correlation of first principle component of Parkinson's signature genes' (As published in Moran et al. (2006)) expression to dopaminergic MGPs in Parkinson's disease patients and healthy control in Moran et al. (2006) and Lesnick et al. (2007) datasets. **(E)** Correlation of expressions of Parkinson's signature genes to dopaminergic MGP.

Tables

Table 1. Cell types in the NeuroExpresso database

Cell Type	**Sample count**	**Marker gene count**	**Source**
Whole Brain			
Astrocyte	9	108*	Cahoy et al. (2008) (GSE9566), Zamanian et al. (2012)
Oligodendrocyte	25	27*	Bellesi et al., (2013) (GSE48369), Cahoy et al. (2008) (GSE9566), Doyle et al. (2008) (GSE13379), Fom-chenko et al. (2011) (GSE30016)
Microglia	3	139*	Anandasabapathy et al. (2011) (GSE29949)
Cortex			
FS Basket (G42)	13	6	Okaty et al. (2009) (GSE17806), Sugino et al. (2014) (GSE8720), Sugino et al. (2006) (GSE2882)
Martinotti (GIN)	3	4	Sugino et al. (2006) (GSE2882)
VIPReln (G30)	6	11	Sugino et al. (2006) (GSE2882)
Pyramidal CrtThalamic	3	6	Schmidt et al. (2012) (GSE2882)
Pyramidal Glt25d2	3	6	Schmidt et al. (2012) (GSE35758)
Pyramidal S100a10	3	9	Schmidt et al. (2012) (GSE35751)

Pyramidal Thy1	12	3	Sugino et al. (2006) (GSE2882)
BasalForebrain			
Forebrain Cholinergic	6	89	Doyle et al. (2008) (GSE13379)
Striatum			
Forebrain Cholinergic	6	45	Doyle et al. (2008) (GSE13379)
Spiny	39	78	Doyle et al. (2008) (GSE13379), Heiman et al. (2014) (GSE55096), Maze et al. (2014) (GSE54656), Tan et al. (2013) (GSE48813)
Amygdala			
Glutamatergic	3	11	Sugino et al. (2006) (GSE2882)
Pyramidal Thy1	12	22	Sugino et al. (2006) (GSE2882)
Hippocampus			
DentateGranule	3	32	Perrone-Bizzozero et al. (2011) (GSE11147)
GabaSSTReln	3	55	Sugino et al. (2006) (GSE2882)
Pyramidal Thy1	12	20	Sugino et al. (2006) (GSE2882)
Subependymal			
Ependymal	2	50	Beckervordersandforth et al. (2010) (GSE18765)
Thalamus			
GabaReln	3	53	Sugino et al. (2006) (GSE2882)
Hypocretinergic	4	34	Dalal et al. (2013) (GSE38668)
Thalamus Cholinergic	3	40	Görlich et al. (2013) (GSE43164)
Midbrain			
Midbrain Cholinergic	3	34	Doyle et al. (2008) (GSE13379)
Serotonergic	3	18	Dougherty and Geschwind (2005) (GSE36068)
Substantia nigra			
Dopaminergic	30	58*	Chung et al. (2005) (pc**), Phani et al. (2010) (GSE17542)
LocusCoeruleus			
Noradrenergic	9	52	Sugino et al. (2014) (GSE8720), Unpublished (pc**)
Serotonergic	3	29	Dougherty et al. (2013) (GSE36068)

Cerebellum			
Basket	16	7	Doyle et al. (2008) (GSE13379), Paul et al. (2012) (GSE37055)
Bergmann	3	52	Doyle et al. (2008) (GSE13379)
CerebGranule	3	11	Doyle et al. (2008) (GSE13379)
Golgi	3	27	Doyle et al. (2008) (GSE13379)
Purkinje	44	43	Doyle et al. (2008) (GSE13379), Galloway et al. (2014) (GSE57034), Paul et al. (2012) (GSE37055), Rossner et al. (2006) (pc**), Sugino et al. (2014) (GSE8720), Sugino unpublished (pc**)
SpinalCord			
SpinalCord Cholinergic	3	124	Doyle et al. (2008) (GSE13379)

Sample count - number of samples that representing the cell type; Gene count - number of marker genes detected for cell type.

* Marker genes for these cell types are identified in multiple regions displayed yet only the number of the genes that are found in the region specified on the table is shown for the sake of conservation of space. Astrocytes, microglia and oligodendrocyte markers are identified in the context of all other brain regions (accept cerebellum for astrocytes) and Dopaminergic markers are also identified for midbrain.

**pc - data acquired through personal communication

Table 2. Coexpression of cortical MGSs in single cell RNA-seq data.

	Tasic et al. (mouse)		**Zeisel et al. (mouse)**		**Darmanis et al.(human)**	
Cell Types	**p-value**	**Gene Count**	**p-value**	**Gene Count**	**p-value**	**Gene Count**
Astrocyte	p<0.001	159	p<0.001	161	p<0.001	156

	Tasic et al. (mouse)		Zeisel et al. (mouse)		Darmanis et al.(human)	
Cell Types	**p-value**	**Gene Count**	**p-value**	**Gene Count**	**p-value**	**Gene Count**
Microglia	p<0.001	152	p<0.001	155	p<0.001	141
Oligodendrocyte	p<0.001	49	p<0.001	50	p<0.001	50
FS Basket (G42)	p<0.001	6	0.007	6	0.006	6
Martinotti (GIN)	p<0.001	4	0.027	4	0.372	4
VIPReln (G30)	p<0.001	11	p<0.001	11	p<0.001	10
Pyramidal CrtThalamic	p<0.001	6	p<0.001	6	0.356	6
Pyramidal Glt25d2	0.006	6	0.2	6	0.457	6
Pyramidal S100a10	0.021	9	0.001	9	0.363	9
Pyramidal Thy1	0.016	3	0.002	3	0.372	3