

Inter-individual variation of cellular and gene-expression properties of the human striatum

Steven Burger^{1,2,*}, Olivia Yoo^{1,*}, James Nemes^{1,2,*}, Ezra Muratoglu¹, Charles Vanderburg¹, Jiayi Yuan^{1,2}, Khalid Shakir¹, Curtis J. Mello^{1,2}, Nirmala A. Rayan¹, Julianna Milidantri¹, Kathleen Kim¹, Sadie Drouin¹, Emily Finn¹, Haoyuan Gao¹, Nikita Budnik¹, Melissa Goldman^{1,2}, Haley Fritch¹, Giulio Genovese^{1,2}, Marina Hogan^{1,2}, Olivia Catalini¹, Seva Kashin^{1,2}, Nicole Rockweiler^{1,2}, Alec Wysoker^{1,2}, Lauren Macaisa¹, Lucas Reese¹, Katelyn Flowers¹, Andrew W. Kraft^{1,3}, Stephen J. Fleming¹, Madelynn Coe¹, Riyaan Gunaratne¹, Liv Spina^{1,2}, Catherine Crombie¹, Abir Mohsin¹, Nolan Kamitaki^{1,2}, Evan Z. Macosko^{1,4,5,**}, Kiku Ichihara^{1,2,**,+}, Steven A. McCarroll^{1,2,6,**,+,,+}

¹ Broad Institute of Harvard and MIT, Cambridge, MA, USA

² Department of Genetics, Harvard Medical School, Boston MA, USA

³ Department of Neurology, Massachusetts General Hospital, Boston MA, USA

⁴ Department of Psychiatry, Massachusetts General Hospital, Boston MA, USA

⁵ Department of Neurobiology, Harvard Medical School, Boston MA, USA

⁶ Howard Hughes Medical Institute, Boston MA, USA

*equal contribution; **corresponding authors; + co-senior authors; ++ lead contact

The human brain varies from person to person in ways that shape behaviors and vulnerabilities, yet the cellular and molecular bases for inter-individual variation are largely unknown. Here we describe an analysis of cellular and gene-expression variation in four key structures of the striatum complex – the caudate, putamen, nucleus accumbens, and internal capsule – as well as the prefrontal cortex, from single-nucleus RNA-seq analysis of 3.9 million nuclei from 178 adult brain donors. We found that people with more astrocytes in any one brain region tended to have this property in all brain regions sampled; the same was true of striatal interneurons, microglia, and oligodendrocyte precursor cells (OPCs). OPCs showed attrition with age, declining in numbers by approximately 40% between age 30 and age 80 in both gray matter and white matter regions. We identified thousands of age-associated (but few sex-associated) variations in gene expression; the vast majority of these effects of age were cell-type-specific. Aging most strongly affected gene expression in projection neurons – especially striatal medium spiny neurons (MSNs/SPNs) – and had a much smaller effect on gene expression in interneurons. Individuals' ages could be predicted to within about five years based on RNA-expression patterns from any of the striatal cell types. Common genetic variants detectably affected the expression levels of some ten thousand genes; the great majority of these effects were cell-type-specific. These data will provide a foundation for exploring natural inter-individual variation, aging, and tissue-based studies of human brain vulnerabilities.

Introduction

Human brains exhibit abundant biological variation – in size and activity patterns, in behavioral proclivities, and in vulnerabilities to illness. Several factors contribute to this biological variability. Humans, unlike isogenic laboratory animals, harbor much genetic variation: any two humans have tens of millions of genetic differences. Humans also live in diverse environments that shape brain development and function in many ways, through social context and through influences on nutrition and metabolism, stress-related neuroendocrine responses, and immune and inflammatory processes.

The striatum, as the principal input nucleus of the basal ganglia, is a key brain region in which biological variation may influence brain function. Striatal activity contributes to behavioral flexibility, reinforcement learning, decision making, motivation, and motor control¹⁻⁷. The striatum is strongly influenced by neuromodulatory systems that can both reflect and shape inter-individual differences in molecular state and cellular activity across striatal circuits. These networks are involved in many neurological and psychiatric disorders, including obsessive compulsive disorder⁸⁻¹⁰ and pronounced neurodegeneration in Huntington's disease^{11,12}.

Despite decades of anatomical and physiological investigation¹³, much of our current cellular and molecular understanding of the striatum is derived from rodent and non-human primate models, which have provided foundational insights into basal ganglia circuitry and cellular organization¹⁴⁻¹⁹ but do not address the ways in which such features might vary among individual humans. Significant work has been done to identify and characterize the distinct cell types within the striatum²⁰⁻²⁴, but there is still much to understand about how these cell types vary throughout the striatum, and from person to person. Recognizing and quantifying inter-individual variation in cell-type abundance and gene expression programs across the lifespan could in principle help provide a foundation for linking genetic variation to striatal circuit function and disease vulnerability. In this work, we sought to leverage advances in single-cell genomics to begin mapping some of the major axes along which the human striatum might commonly vary – including the cells that are present in the striatum, their relative abundances, and the genes they express. We also sought to identify anatomical features and age-related dynamics that reappear consistently across people.

Single-cell genomics generates detailed, multi-dimensional cellular and molecular profiles of any brain region or subregion; however, it has historically been challenging to determine which aspects of inter-individual variation are biological and which are technical. Small variations in how tissue specimens are subdissected, and how nuclei are extracted from tissue, can skew both the representations of different types of cells and the RNA transcripts that are ascertained for analysis. To address this, most studies have focused on how properties, including cell type composition and gene expression profiles, differ on average between two groups of individuals: for example, those with a specific clinical condition and those without. Less-emphasized – in part because it often involves an unknown mix of technical, statistical, and biological factors – is the large quantitative variation that appears within each such group.

We sought to address these challenges using two aspects of project design. The first was to analyze several brain regions in each of the individuals sampled; this allowed us to identify cellular and molecular features that were consistent properties of an individual in the sense that they appeared in most or in all brain regions under analysis. The second was to perform all experiments in a well-controlled "village" format²⁵, in which tissue samples from ~20 donors were processed together and later reassigned to their donor of origin using transcribed genetic variants. This approach made it straightforward to recognize and correct technical effects on the resulting data, since such effects were shared by all of the specimens in an experiment.

Here, and in concert with broader efforts by the BICAN consortium, we applied this approach to deeply analyze the main structures of the striatum – the caudate (CaH), putamen (Pu), and nucleus accumbens (NAC) – which share major cell types in common, but differ in functional connectivity^{26,27}, gene expression^{28,29}, and disease vulnerability^{12,30–33}. We also included the internal capsule (ic), a white matter tract between the caudate and putamen, and dorsolateral prefrontal cortex (DFC), a well-characterized cortical area, to help recognize properties that extend beyond the striatum. Neuroanatomical structures and region abbreviations are defined in the Allen Institute Developing Human Brain Atlas (DHBA) ontology³⁴ (RRID:SCR_027940).

We created a resource of single-nucleus RNA-seq profiles from 3.9 million nuclei sampled from 178 adult human brain donors. We identified many cellular features that show a surprising degree of quantitative variation from person to person; many of these features reappeared throughout a person's various brain regions or showed clear relationships to other variables (such as age and common genetic variation), suggesting that they embody true inter-individual variation in human neurobiology. We discuss below ways to use this data resource and the many patterns it contains, and provide a [web-based portal](#) (RRID:SCR_028065) to make the data simple to query and utilize in other work. This resource will enable a deep exploration of inter-individual variation and the ways it is shaped by age and genetics, and will serve as a resource for conceiving and designing future tissue-based studies of human neurobiology.

Results

A single-cell resource of inter-individual variation in the adult human striatum

We analyzed brain tissues that had been donated at the end of life by 178 persons. These included a core donor set ($n = 147$) and additional donors ($n = 31$) for whom one or more clinical or neuropathological findings became apparent in the course of the project (**Table S1**). Below we specify which analyses use the core set of donors and which use the extended (core + additional) set. Donors were selected based on donor characterization by brain bank staff, assessments of tissue quality (see **Methods**), and availability of tissues from the regions of interest. The donor age distribution spanned from 27 to 90+ years old (median = 62; precise ages >89 were unavailable to protect donor identity), and included 111 male and 67 female donors (**Figure S2B**, **Table S2**).

Variation between individuals is more subtle and quantitative than the categorical differences that distinguish cell types and species from one another. To better recognize and measure such variation, we implemented a "nuclei village" experimental framework²⁵. Systematic dissections of the striatum separated caudate, putamen, nucleus accumbens, and internal capsule (see **Methods**, **Figure S1**). Tissues from each subregion were then processed in pools of ~20 donors; each such pool ("village") was handled as a single sample through nuclei isolation and snRNA-seq (**Figure 1A**). Donor identity for each nucleus was recognized using combinations of thousands of transcribed sequence variants that appear on RNA transcripts, using a computational approach (Dropulation; RRID:SCR_018142) that we have previously described³⁵.

Gene-expression measurements were computationally adjusted for ambient RNA contamination using CellBender³⁶ (RRID:SCR_025990). The selection of cell barcodes that contained nuclei (as opposed to empty droplets and droplets containing cellular debris) was refined using DropSift³⁷ (RRID:SCR_028040; see **Methods**), a tool we recently developed that utilizes the fraction of reads that are intron-derived (alongside total UMI counts) to recognize the profiles of nuclei, in a data-driven manner that adapts to each dataset. Genotypic (two-donor) doublets were identified with Dropulation.

Using this approach, single-nucleus RNA-seq (snRNA-seq) data were generated from caudate head ($n = 175$), putamen ($n = 163$), nucleus accumbens ($n = 148$), internal capsule ($n = 98$), and dorsolateral prefrontal cortex ($n = 161$), with multiple regions sampled from the same donors where available (**Figure 1B**, **S2A**).

Unsupervised clustering of the single-nucleus RNA-expression profiles identified all major neuronal and non-neuronal cell classes across the regions sampled, including the diverse types of medium spiny neurons (MSNs, also known as striatal projection neurons or SPNs), the principal neuronal population of the striatum (**Figure 1C**, **S2D**). Manual cluster annotations utilized cell type labels from the consensus basal ganglia taxonomy³⁸ generated via MapMyCells³⁹ (RRID:SCR_024672) (**Figure S2E**). To identify the smaller number of remaining doublets (in which both nuclei are from the same donor), the far larger number of genotypic

doublets were used to recognize the gene expression patterns associated with doublets in general. This systematic removal of doublets facilitated the confident recognition of cell types with transcriptional profiles that happen to somewhat resemble mixtures of other prominent cell types — including D1-D2 hybrid medium spiny neurons (MSNs), striosome-matrix hybrid MSNs, and committed oligodendrocyte precursors (COPs), which express both OPC (oligodendrocyte precursor cells) and oligodendrocyte genes — when these were observed in clusters that contained few, if any, genotypic doublets (**Figure 1D**).

Following data quality filtering and removal of doublets, the final dataset comprised approximately 3.9 million nuclei. For each brain region, a small fraction of the subdissected specimens (generally about 15%) exhibited low ascertainment of nuclei and/or highly abnormal cell type proportions and were excluded from downstream analyses (see **Methods**). Expected neuronal and glial cell types were ascertained in tissue samples from all donors (**Figure 1E**). Following quality control of the data, we first looked at changes in cell type abundances across regions, then the variables that might relate to these changes (age, sex), and finally at effects of common genetic variation on gene expression in these striatal and cortical cell populations.

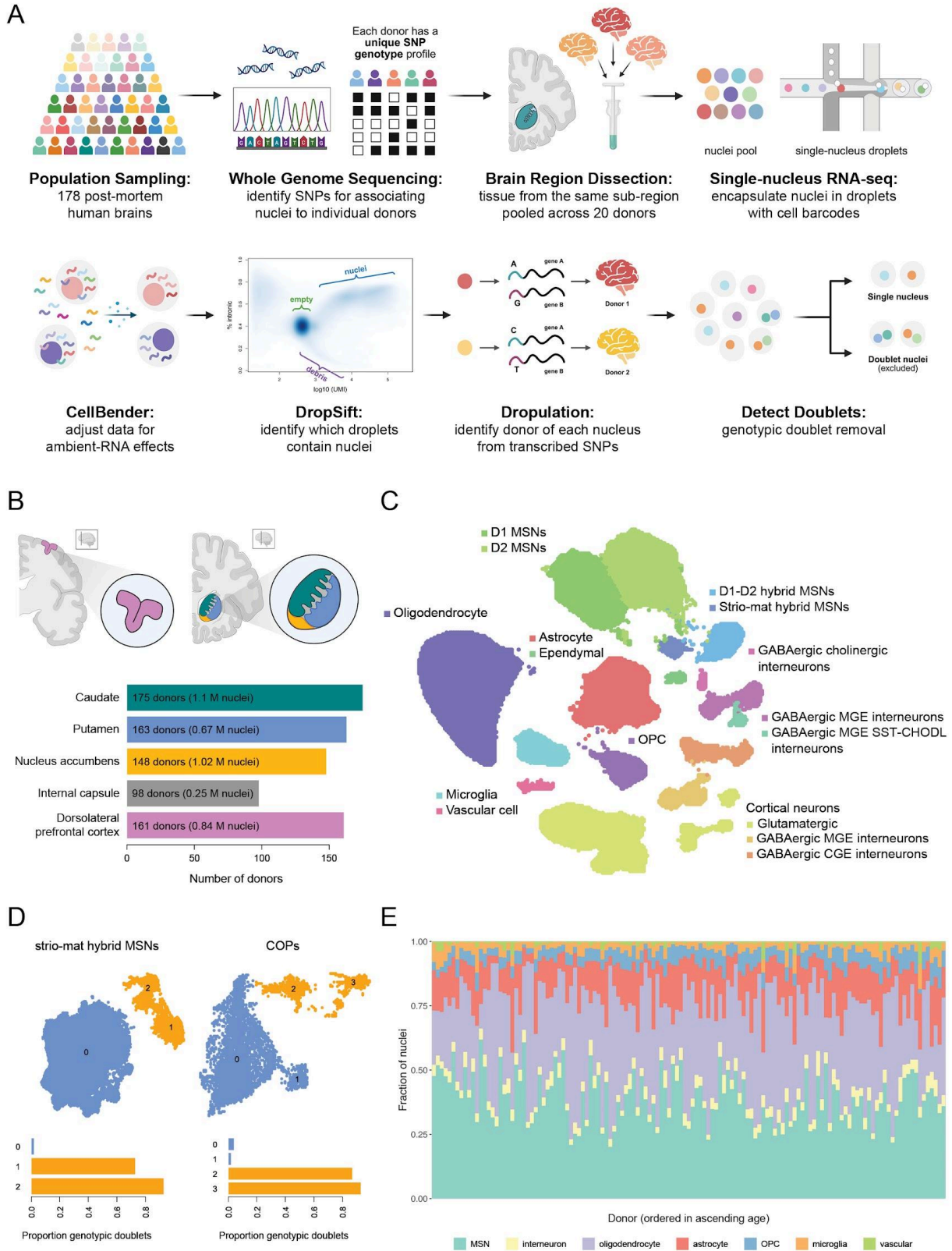


Figure 1. A single-cell resource for recognizing inter-individual variation in the adult human striatum.

A. Schematic of data generation and processing. Top: multi-donor (“village”) snRNA-seq experimental workflow. Bottom: computational data processing workflow. Diagrams created with BioRender (RRID:SCR_018361). **B.** Brain regions sampled. Left: dorsolateral prefrontal cortex (pink). Right: caudate (green), putamen (blue), nucleus accumbens (yellow), and internal capsule (gray). Diagrams created with BioRender. Barplot represents the total number of donors sampled for each region. **C.** UMAP of all 3.9M nuclei assigned to a single donor and cell type. **D.** Genotypic (two-donor) doublet calls identified by Dropulation help confidently identify clusters of cell types that resemble mixtures of other cell types. Left: UMAP of strio-mat hybrid MSNs; clusters 1 and 2 contain high proportions of genotypic doublets (and are excluded from downstream analyses) compared to cluster 0. Right: UMAP of committed oligodendrocyte precursor (COP) clusters; clusters 2 and 3 have high proportions of genotypic doublets (and are excluded from downstream analyses). **E.** Stacked bar plot shows representation of major cell types among the caudate samples from 131 donors. Donors are ordered from left to right by ascending age. Major cell classes were detected in all donors sampled.

Glial cell abundances vary from person to person

Cell type proportions exhibited extensive variation among the individual donors (**Figure 2A**). A key question involves whether such variation (at least in part) reflects true biological variation among the brain donors sampled, as opposed to experimental noise introduced by variation in tissue sampling, such as differences in microdissections and tissue micro-environments.

The analysis of microdissections from multiple brain regions in the same large panel of donors made it possible to ask whether this variability reflected intrinsic donor-specific features shared across brain regions. To exclude effects of clinically apparent conditions or conditions apparent to conventional neuropathological analyses of the specimens, we used the stringent donor group ($n = 147$) for these analyses.

We found that representations of the major glial cell types (astrocytes, microglia, OPCs, and oligodendrocytes) were significantly correlated from one striatal region to another: donors for whom the abundance of one cell class was high in the caudate tended to have this property in the putamen and nucleus accumbens as well (**Figure 2B, S3A**). These relationships arose from the entire set of donors, as opposed to a few outliers, and were highly significant even when assessed with conservative, non-parametric statistical tests (**Figure 2B, S3A**).

These relationships also extended beyond the striatum: the tendency of the individual donors to have relatively high or low abundances of astrocytes (or microglia, or OPCs) was also apparent in their internal capsule (**Figure S3B**) and even in their dorsolateral prefrontal cortex (**Figure 2C**). These correlations were similarly strong when the abundance of each cell type was calculated relative to the number of neurons rather than the total number of nuclei (to prevent the glial cell types from affecting one another's abundance estimates; **Figure S4**). These results suggest that a substantial component of inter-individual differences in glial abundance reflects biological properties of individuals, and that these properties also extend beyond the striatum.

Substantial fractions of inter-individual variation in the measured representations of OPCs, astrocytes, and microglia in the caudate could be predicted by measurements of the same properties in DFC (Pearson r^2 of 0.22 for OPCs, 0.18 for astrocytes, and 0.29 for microglia). Focusing on the component of inter-individual variation that was shared between DFC and caudate, we conservatively estimated that caudate samples from any two individuals will frequently (in at least half of comparisons) have at least a 1.2 fold difference in the representation of astrocytes and OPCs, and a 1.4 fold difference in the representation of microglia (see **Methods**).

The principal exception to this inter-region correlation pattern involved oligodendrocytes (Pearson $r^2 = 0.01$ between caudate and DFC). Across the individual donors, their abundances in the DFC (or ic) were not correlated with those in the striatum (CaH, Pu, and NAC) (**Figure 2C, S3B**), and thus appear to be shaped by differential sampling of white matter and/or axonal fiber bundles in sub-dissections.

These results suggest that, even among persons not known to have brain-related clinical conditions or neuropathological findings, the tendency of individual people to have higher-than-average or lower-than-average numbers of astrocytes, microglia, and/or OPCs is a consistent property of the individual that appears across multiple striatal and cortical regions.

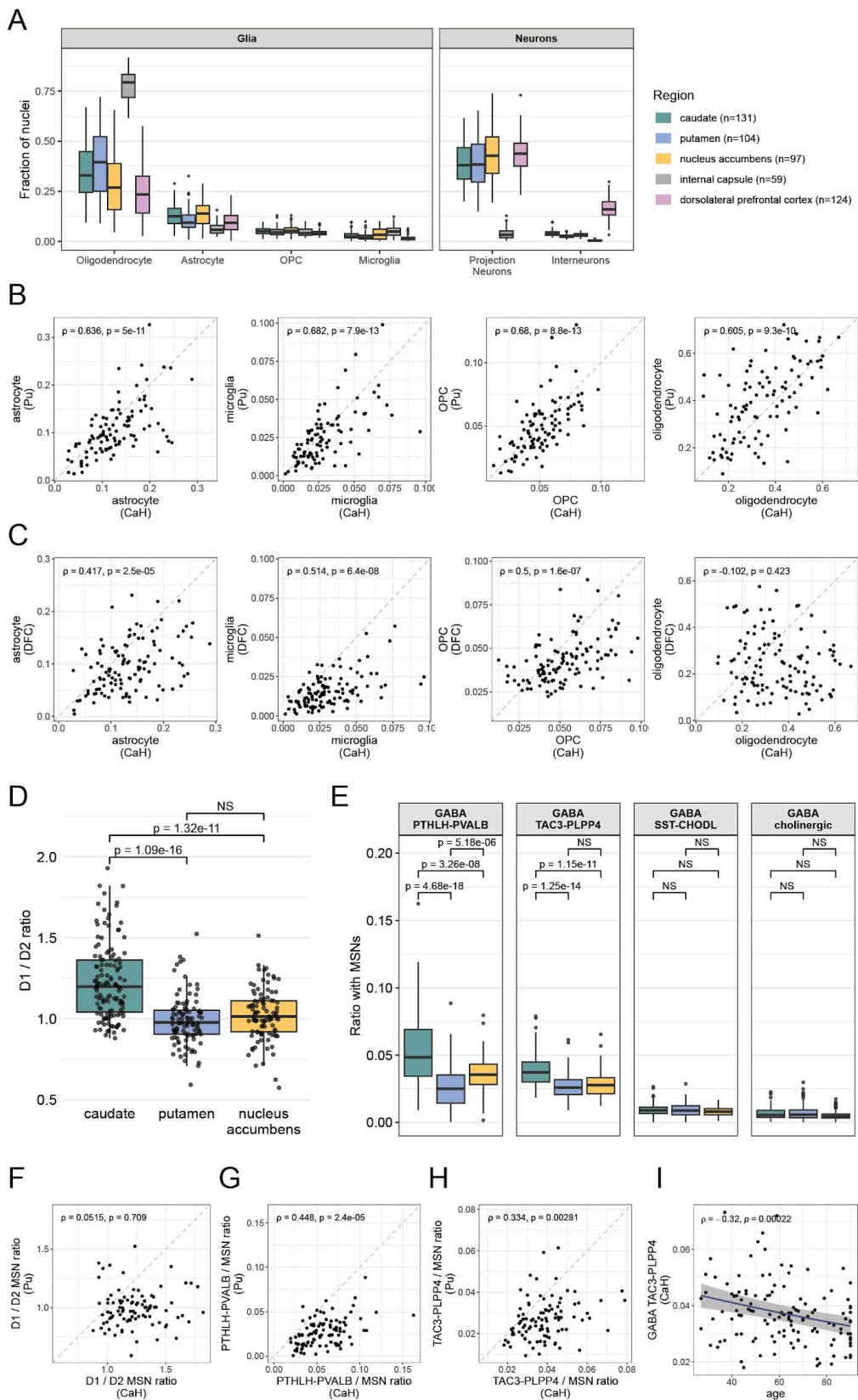


Figure 2. Variation in cell type abundances across individuals and between anatomical regions.

(Extended analyses are in **Figures S3-S7**). **A.** Representation of major cell types in the brain regions sampled. Boxes indicate the interquartile range across donors, center lines denote median values, and whiskers extend to the most extreme points within 1.5x the interquartile range. Relative to the gray matter regions, the internal capsule yielded more oligodendrocyte and fewer neuronal nuclei; DFC had a higher fraction of interneurons compared to striatal regions. **B.** Representation of glial cell type abundances (as a fraction of all nuclei sampled) in the caudate (x-axis) and putamen (y-axis) of 98 donors. Spearman correlation coefficient and corrected p-values are shown (Benjamini-Hochberg (BH) procedure, n=498 tests). **C.** Same as **B**, for 114 donors sampled in caudate (x-axis) and dorsolateral prefrontal cortex (y-axis). **D.** D1/D2 MSN ratio was significantly higher in CaH (n=131 donors) than in Pu (n=104 donors) and NAC (n=97 donors); nominal p-values are from the Wilcoxon rank-sum test. No significant difference was detected between Pu and NAC. **E.** Quantitative representation of striatal interneuron subtypes relative to the number of MSNs. Interneuron abundances are expressed relative to MSNs to mitigate the effects of inter-individual variation in glial abundances (panel **B**) on these calculations. Statistical significance was assessed using the Wilcoxon rank-sum test and corrected for multiple comparisons (BH procedure, 12 tests). **F.** Ratio of the number of D1 MSNs to D2 MSNs in CaH (x-axis) and Pu (y-axis) samples from 98 donors. Nearly all donors have a greater D1/D2 MSN ratio in CaH than Pu. This ratio was not correlated between CaH or Pu. Spearman rank correlation coefficient and corrected p-value shown (BH procedure, n=498 tests). **G.** Representation of GABAergic PTHLH-PVALB interneuron abundances (relative to MSNs) in the caudate (x-axis) and putamen (y-axis) of 98 donors. Spearman correlation coefficient and corrected p-values are shown (BH procedure, n=498 tests). **H.** Same as **G**, but for GABAergic TAC3-PLPP4 interneurons. **I.** Abundance of TAC3-PLPP4 interneurons declines with age across striatal regions, shown for CaH. Abundance is expressed as a fraction of all neurons sampled. Each point represents a donor; the blue lines indicate beta-binomial fit with 95% confidence intervals shown as gray ribbons. Abundance was negatively correlated with age for all brain regions (Spearman correlation coefficient and nominal p-value shown). Modeling with a beta-binomial regression confirms a significant decline with age ($\beta = -0.047$ per decade; 95% CI: -0.069 to -0.025 ; BH-adjusted p-value = 0.001; 44 tests; **Table S4**).

Neuronal proportions differ among striatal regions

Sampling from a large number of donors in multiple brain regions made it possible to recognize quantitative features of specific striatal regions that consistently differed from each other, in many different people. In order to minimize the effect of the large inter-individual variation in the glial cell types (**Figure 2**), we measured the abundances of neuronal types relative to those of other types of neurons.

MSNs are classically categorized based on their expression of either D1 or D2 dopamine receptors⁴⁰, corresponding to the direct (D1) and indirect (D2) output pathways to basal ganglia structures. D1 and D2 MSNs are further distributed across two anatomical compartments in the striatum: the matrix and striosomes^{41–43} (**Figure S5**). The observed ratio of D1 to D2 MSNs was 20% higher in caudate (mean 1.23) than in the putamen (mean 1.00) and nucleus accumbens (mean 1.02) (**Figure 2D**). This difference was consistent across both matrix and striosome MSNs (**Figure S6A**) and across donors (**Figure S6B**), and aligns with findings from mouse immunofluorescence studies^{44,45}. This finding also aligns with results from the contemporaneous discovery (from Slide-tags spatial transcriptomics data) that six discrete spatial zones underlie variation across the striatum: the two zones with the greatest representation of D1 MSNs overlap with the caudate⁴⁶. The ratio of striosome to matrix MSNs did not differ significantly between caudate and putamen, with striosome MSNs representing on average 17% of MSNs in both brain regions (**Figure S6C**).

The striatum also contains populations of non-canonical (also referred to as “hybrid” or “eccentric”) MSNs^{22,47}. The ratio of eccentric to canonical MSNs consistently distinguished the nucleus accumbens (in which approximately 19% of MSNs were of the eccentric subtypes) from the caudate and putamen (approximately 7%) (**Figure S6D**), largely due to the elevated abundance of D1 NUDAP (neurochemically unique domains of accumbens and putamen⁴⁸) MSNs (**Figure S5F,G**).

While MSNs are the principal neuronal population in the striatum, they are accompanied and modulated by a heterogeneous set of GABAergic interneurons with distinct structural and electrophysiological properties^{49–51}. The two most abundant striatal interneuron classes are the fast-spiking PTHLH-PVALB and the TAC3-PLPP4 types, the latter previously considered primate-specific¹⁷ but recently detected across several mammalian species⁵². PTHLH-PVALB interneurons were most numerous in caudate, followed by nucleus accumbens and then putamen, forming a medial-to-lateral gradient (**Figure 2E**). TAC3-PLPP4 interneurons were also most numerous in caudate but showed similar abundances in putamen and nucleus accumbens (**Figure 2E**).

We also looked for evidence of inter-individual variation in neuron abundances that was shared across striatal regions. Across individuals, the ratio of D1 to D2 MSNs was not significantly correlated between the caudate, putamen, and nucleus accumbens (**Figure 3F, S6B**). However, the abundances of interneuron subtypes exhibited strong evidence of inter-individual variation, in patterns that appeared consistently across the striatum. PTHLH-PVALB and TAC3-PLPP4

interneurons showed strong correlations in abundance across all pairs of striatal regions, particularly between caudate and putamen (**Figure 2G,H, S7A-C**). Abundances of different types of interneurons were also correlated with each other: donors with higher abundances of PTHLH-PVALB interneurons also had higher abundances of TAC3-PLPP4 interneurons (**Figure S7D**). We note that the observed correlations likely underestimate the strength of the underlying relationships, since striatal interneurons are sparse and our estimates of their abundance are therefore affected by substantial statistical sampling noise.

To evaluate the extent to which inter-individual variation in neuron abundances might reflect effects of age or sex, we applied a regression framework that incorporated age, sex, brain region, and additional donor- and sample-level covariates, while controlling for variability in nuclei counts per sample (see **Methods**). Canonical MSN subtypes, as a fraction of all MSNs, showed no significant associations with either age or sex (**Table S4**). Among interneurons (relative to all neurons), the TAC3-PLPP4 type had a statistically significant association with age ($\beta = -0.047$ per decade of age, 95% CI: $-0.069 - -0.025$, FDR-adjusted p-value = 0.001, **Figure 2I, S7E**). No type had a significant association with sex (**Table S4**). TAC3-PLPP4 interneurons appeared to decline modestly, by approximately 20% between ages 30 and 80, and at a similar rate across all three striatal regions (see **Methods**).

These results suggest that neuronal composition in the striatum reflects both consistent anatomical specialization between regions and inter-individual variability, the latter particularly evident for interneurons.

OPC proportions decline with age across brain regions

We next asked whether any other cell type abundances were shaped by age or sex. Among well-ascertained cell types (median > 1% of all nuclei across samples), and after correcting for multiple hypothesis testing, only OPC abundance was significantly associated with age ($\beta = -0.096$ per decade of age, 95% CI: $-0.12 - -0.07$, FDR-adjusted p-value = 4.70×10^{-12} , **Table S4**). Neither astrocytes nor microglia showed a statistically significant association of abundance with age (**Figure S8A,B**). No cell type showed a significant association with sex (**Table S4**).

OPCs (also referred to as polydendrocytes, in recognition of their many functions beyond serving as progenitors) remain proliferative in the adult brain and serve as the primary progenitors of myelinating oligodendrocytes; beyond this role, OPCs are increasingly recognized to have additional functions in the adult CNS, including through the synaptic inputs that they receive from neurons⁵³.

The presence of OPCs (relative to all nuclei sampled) declined with advancing age in all brain regions analyzed (**Figure 3A**). This attrition was remarkably similar in magnitude across these brain regions, in all of which OPCs declined by about 40% between age 30 and age 80 (see **Methods**). A declining fraction of OPCs with age has been previously observed in the human DFC⁵⁴ and in the mouse frontal cortex and striatum⁵⁵.

To see whether the presence of OPCs is affected by individually varying factors beyond chronological age, we calculated an OPC fraction residual, the difference between a sample's actual OPC fraction and the OPC fraction predicted from their age (and other model covariates, see **Methods**). These OPC fraction residuals were positively correlated between all pairs of brain regions (**Figure 3B, S8C**): donors with higher-than-expected OPC abundance for their age in one brain region tended to exhibit similarly elevated OPC abundance in other regions. The consistency across diverse brain regions of both the effect of age and the residual unexplained by age suggests a substantial role for genetics and/or environment.

To better characterize OPC attrition and OPC diversity across these brain regions, we sub-clustered the OPC gene expression profiles (from ~155K nuclei), identifying seven transcriptionally distinct clusters (**Figure S9**, see **Methods**). The two most abundant clusters closely resembled OPC populations previously described as “immature” (cluster 0, expressing *SHISA6*, *SGCZ*, *GPC6*) and “mature” (cluster 1, expressing *CDH19*, *SEMA3E*, *GRIA4*)⁵⁶. Among genes differentially expressed between these OPC populations, the most distinctive and interpretable signal for cluster 0 was the elevated expression of several genes that encode synaptic proteins (*GABRB2*, *KCNB2*, *SYN3*, *NCALD*, *SHISA6*) and specific transcription factors (*TCF7L1*, *PBX3*, *RFX4*). Cluster 1 was marked by higher levels of *CDH19*, which encodes a classic oligodendrocyte adhesion molecule. OPC clusters 0 and 1 were also distinguished by their expression levels of distinct sets of axon guidance genes (*CNTN5*, *EPHA5*, *ADGRV1* in cluster 0; *SEMA3E*, *PDZRN3*, *DACH2* in cluster 1) and extracellular matrix genes (*COL4A2*, *FBLN2*, *COL19A1* in cluster 0; *ADAMTSL1*, *COL20A1*, *CHRDL1* in cluster 1). These tissue-context-dependent expression profiles support the idea that OPCs have functions beyond serving as oligodendrocyte progenitors⁵⁷.

Both of these clusters were observed in all brain regions; however, the internal capsule (the only white matter region analyzed) had a far lower proportion of cluster 0 OPCs and greater proportion of cluster 1 OPCs (3.6 - 7.1 fold; DFC and CaH, respectively) (**Figure 3C, S9C,D**). To ask whether these clusters might vary in their spatial distributions (e.g., potentially between gray and white matter regions) we examined Slide-tags spatial transcriptomics data of the striatum generated concurrently⁴⁶ (**Figure 3D,E**). OPCs were evenly distributed throughout both the gray and white matter of the striatum (**Figure 3F**), consistent with reports in mice⁵⁵. However, when visualized spatially, marker genes for cluster 0 highlighted gray matter compartments (**Figure 3G, S9E**), while marker genes for cluster 1 highlighted white matter compartments (**Figure 3H, S9F**). The spatial localization of these OPC populations further supports the idea that the fundamental difference between these types of OPCs may be tissue context (gray matter vs. white matter).

Across all brain regions, the quantitative presence of both OPCs with a cluster 0 profile (gray matter-enriched) and OPCs with a cluster 1 profile (white matter-enriched) declined with age (**Figure 3I-J, S9G-H**), further suggesting that this age-associated attrition affects OPCs broadly and in diverse tissue contexts.

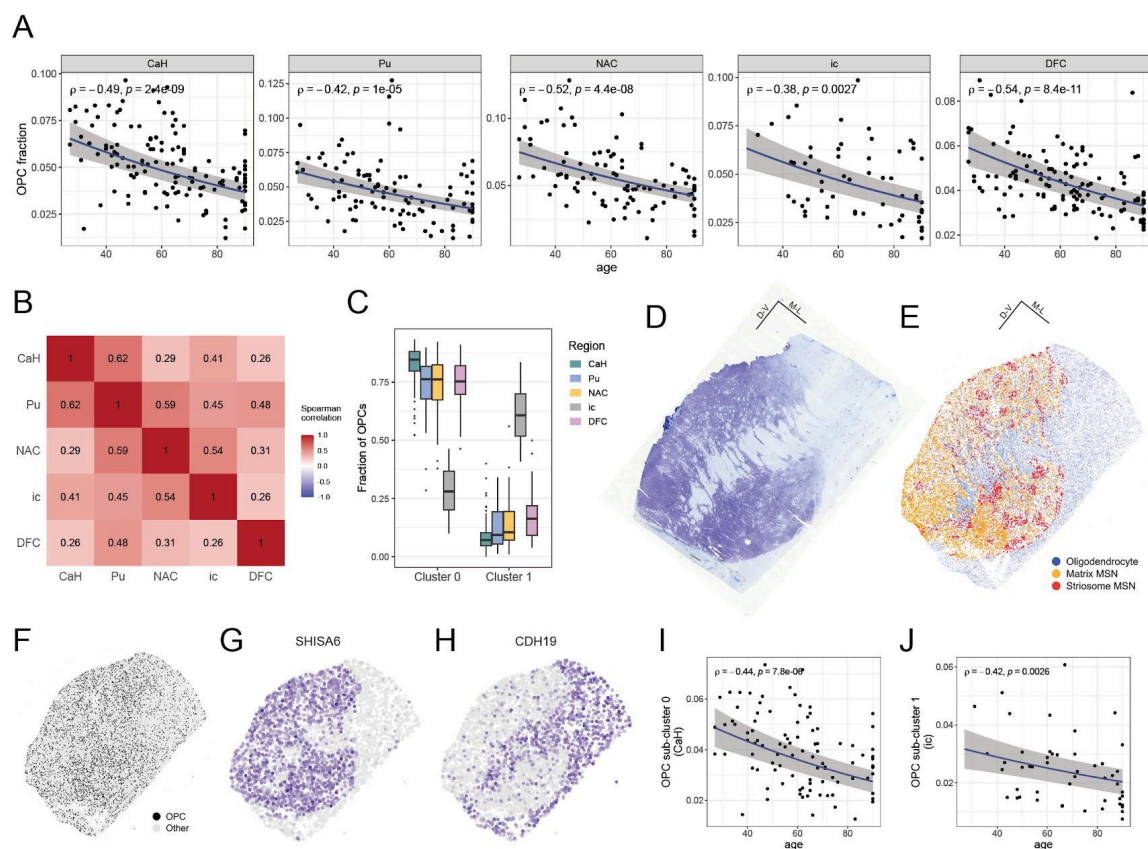


Figure 3. Decline in OPC populations in multiple brain regions with advancing age. (Extended analyses are in **Figure S9**). **A.** Attrition of OPCs with advancing age in five brain regions. In each sample, the abundance of OPCs was measured as a fraction of all nuclei sampled. Each point corresponds to a donor ($n=131$ in CaH; $n=104$ in Pu; $n=97$ in NAC; $n=59$ in ic; $n=124$ in DFC). Spearman rank correlation and nominal p-values are shown. Each blue line represents a beta-binomial fit of OPC abundance to age, with 95% confidence intervals shown by the gray ribbon. Modeling with a beta-binomial regression confirmed a significant decline of OPC fraction with age ($\beta = -0.096$ per decade of age; 95% CI: $-0.12 - -0.07$; Benjamini-Hochberg (BH) adjusted p-value = 4.70×10^{-12} ; $n=44$ tests; **Table S4**). **B.** Heatmap shows the cross-region correlation of “OPC fraction residual” – the observed OPC fraction minus the age-expected OPC fraction, where the age-expected fraction was calculated from the beta-binomial fit. Positive correlations are in shades of red, and negative correlations in blue. **C.** Representation (in each brain region) of cluster 0 and cluster 1 OPCs, calculated as a fraction of all OPCs. **D.** Nissl stain of the striatum (12 μ m section) from one donor. **E.** Spatial transcriptomics data of an adjacent section from the same donor, generated with Slide-tags. Each point shown is an individual nucleus corresponding to either oligodendrocytes (blue), matrix MSNs (orange), or striosome MSNs (red). **F.** Spatial arrangement of OPCs (black) vs. all other cells (gray). **G.** Expression of *SHISA6* (a cluster 0 marker gene) in OPCs highlights gray matter compartments. Color of the points represents log-normalized transcript counts (darker purple = greater expression; lighter gray = lower expression). **H.** Expression of *CDH19* (a cluster 1 marker gene) in OPCs highlights white matter compartments. **I.** OPC cluster 0 abundance, as a fraction of all nuclei sampled (y-axis), declines with age (x-axis) in CaH (Spearman correlation coefficient and nominal p-value shown). Modeling with a beta-binomial regression confirmed a significant decline with age ($\beta = -0.095$ per decade; 95% CI: -0.122 to -0.067 ; BH-adjusted p-value = 7.02×10^{-10} ; $n=44$ tests; **Table S4**). **J.** OPC cluster 1 abundance declines with age in ic. Modeling with a beta-binomial regression confirmed a significant decline with age ($\beta = -0.072$ per decade; 95% CI: -0.109 to -0.036 ; BH-adjusted p-value = 0.004 ; $n=44$ tests; **Table S4**).

Age-associated changes in cell-type-specific RNA expression

Cells of any one type also vary from person to person in quantitative levels of the expression of each gene. To recognize such variation and the effects of age and sex upon it, we used a regression framework that incorporated donor-, sample-, and cell-level covariates (see **Methods**).

A key question in this and other brain research involves what set of donors are most informative for characterizing normal variation. A longstanding practice involves narrow inclusion criteria that exclude a large set of common health circumstances; this practice tends to focus research utilization on a small fraction of brain-bank samples, and excludes the vast majority of donations from younger people, whose manner of death more frequently involves drug overdose or suicide. Our project provided an opportunity to relax assumptions and empirically evaluate the effects of exclusion criteria on statistical power for various kinds of analyses. We considered the effect of two kinds of exclusions: "metadata-driven" exclusions based on reported health circumstance (including major depressive disorder and substance dependence; see **Methods**), and "snRNA-seq-data-driven" exclusions for donors identified as clear outliers in gene expression or cell-type proportions (see **Methods**). For analyzing effects of age, sex and common genetic variants, we found that snRNA-seq-data-driven exclusions tended to preserve or increase statistical power without altering estimated effect sizes (**Figure S10, S15**). In contrast, metadata-driven exclusions (based on available health information) substantially reduced statistical power in these analyses, mainly by reducing sample size without any compensating benefit in reduced variance (**Figure S10, S15**), resulting in a reduced number of discoveries at any given false discovery rate. Therefore, our subsequent differential expression analyses (and the genetic analyses described in a subsequent section) implemented only snRNA-seq-data-driven donor exclusions – generally involving about 5% of the available samples (which we believe generally arise from peri-mortem circumstances and perhaps local subclinical conditions in specific brain regions).

Across the striatal and cortical regions we analyzed, associations of gene expression to chromosomal sex were largely limited to sex chromosome genes – primarily Y chromosome genes and *XIST* and *TSIX* on the X chromosome (**Figure 4A**). This pattern was shared across all cell types in the analysis.

In contrast to the very small effect of sex, age appeared to affect the expression of thousands of genes, on all chromosomes (**Figure 4B**) and in every cell type. Age-associated effects were highly correlated between closely related cell types (e.g., subtypes of MSNs) (**Figure 4C**) and between homologous cell types across regions (e.g., MSNs in different striatal regions; striatal versus cortical astrocytes) (**Figure 4D,E, S11B,C**), though the changes in MSNs appeared to proceed a bit more slowly in the nucleus accumbens than in the caudate (**Figure 4D**).

In contrast to the consistent effects of age on the same or similar cell types in different brain regions, age affected gene expression in different cell types (even in the same brain region) in very different ways (**Figure 4F**). Even when focusing on neurons, age affected gene expression

in different types of neurons in quite different ways and to quite different degrees (**Figure 4G,H,I**).

To better recognize the underlying patterns that shape global and cell-type-specific components of brain aging, we allowed the data to organize genes into clusters (using k-means clustering) based on their patterns of cell-type-specific age-associated expression changes (**Figure 4H, S11D**). A small subset of the age-affected genes (cluster 1 in **Figure 4H**) tended to increase modestly in expression with age across cell types. These included genes related to stress and glucocorticoid signaling (e.g., *FKBP5*; **Figure S12**), inflammatory pathways (*RSAD2*, *MGST2*, *CKLF*, *TNFSF13B*, *CD247*), DNA damage responses (*RAD51B*, *RPA1*, *BRIP1*, *FANCL*, *USP45*), and metabolism (*SLC25A24*, *ALDH3A2*, *PNPLA7*, *PTGR1/2*, *LSS*). Many of these kinds of changes have previously been recognized in analyses of age-associated gene-expression changes in "bulk" brain tissues, which have often aligned around functional themes such as metabolism, stress, inflammation, and DNA repair⁵⁸⁻⁶⁰.

Beyond this modest set of shared, pan-cell-type changes in gene expression, each type of cell exhibited a substantial set of cell-type-specific age-associated changes in RNA expression. For glial cell types, these changes replicated strongly between analyses of the striatal regions and analyses of the DFC (**Figure 4G, S11A**). Among neurons, changes were strongly shared only between certain specific (and closely related) neuronal types: canonical MSN sub-types (D1 and D2; matrix and striosome) and intratelencephalic-projecting (IT) glutamatergic neurons of different cortical layers. MSNs, interneurons, and cortical pyramidal neurons exhibited only slightly positive correlations (with one another) in their age-associated transcriptional changes (**Figure 4G, S11A**), with these changes arising from a small subset of the age-affected genes (**Figure 4H**). There was almost no correlation in the age-associated changes of the various glial cell types when compared with one another or with neurons (**Figure 4G,H**).

In principle, age might affect the biology of some brain cell types more than others. Quantifying and comparing the magnitude of transcriptional changes across cell types has been challenging due to the difference in power to recognize age-associated changes in different cell types, as very different numbers of cells and RNA transcripts (unique molecular identifiers, or UMIs) are sampled for each cell type, making the number of differentially expressed genes (whose identification is highly sensitive to statistical power) a misleading estimate of impact. To overcome this, we used transcriptome-wide analysis of differential expression (TRADE⁶¹), a statistical framework for recognizing the distribution of true differential expression effects while accounting for estimation error, to estimate the average transcriptome-wide impact of age upon each cell type (**Figure 4I**).

MSNs exhibited larger age-associated transcriptional changes than any other striatal or cortical cell type did (**Figure 4H,I**). More generally, projection neurons, including MSNs and cortical glutamatergic neurons, exhibited much larger age-associated changes than interneurons did (**Figure 4H,I**). Among glial cells, microglia and astrocytes exhibited the largest age-associated changes in gene expression, followed by oligodendrocytes and then OPCs (**Figure 4I**). Curiously, of all cell types in the analysis, OPCs exhibited the smallest age-associated effects

on RNA expression (**Figure 4I**), despite being the only glial cell type that exhibited substantial attrition with age (**Figure 3A**). Similarly, GABAergic TAC3-PLPP4 interneurons – whose abundance was also impacted by age (**Figure 2I, S7E**)– showed a relatively modest transcriptional impact of age, compared to projection neurons (**Figure 4I**).

To better characterize these age-associated gene expression changes in each cell type, we ranked genes by the age-associated differential expression test statistic and performed Gene Set Enrichment Analysis (GSEA) on these rankings⁶² (**Table S5**). MSNs, interneurons, and astrocytes all showed age-associated declines in expression of gene encoding cellular respiration enzymes (e.g., GO:0045333; p-value = 1e-10), a set of genes long observed to decline with expression with age in "bulk" tissue studies. However, this pattern was by no means universal to all cell types: microglia, for example, exhibited the opposite relationship, increasing expression of this same set of genes with age (e.g., GO:0045333; p-value = 4e-10). This suggests that these metabolic changes may themselves result from changes in cell-type-specific activities.

In neurons, this apparent average decline in cellular respiration activities was accompanied by a decline in expression of ribosomal/translation programs (e.g., GO:0003735; FDR = 8e-29), perhaps suggesting reduced commitment to (energetically expensive) turnover activities. In MSNs, the most prominent age-downregulated modules comprised terms related to synaptic function and signaling (e.g., GO:0097060; FDR = 2e-28), including reduced expression of *GPR158*, *RGS9*, *ADCY3*, *PDE10A*, *PKIA*, and *PPP1R1B* (additional genes represented in **Figure 5A**). GABAergic *PTHLH-PVALB* interneurons also showed a decline in expression of genes with synaptic functions (e.g., GO:0097060; FDR = 1e-28), with a pronounced (4:1), overall bias towards downregulation of these genes with age. Together, these neuronal signatures indicate a coordinated decline in bioenergetic activity and translation, in conjunction with reduced synthesis or turnover of synaptic molecular machinery.

GSEA analysis revealed the diverse cell-type-specific activities that appeared to be age-associated in the various glial cell types. In astrocytes, age was associated with increasing expression genes with annotated roles in immune-interface programs (e.g., GO:0002449; FDR = 9e-05) and complement activation (e.g., GO:0006958, FDR = 0.01) (including elevated expression of the complement gene *C3* with age). Microglia showed increased expression of genes with immune activation signatures (e.g., GO:0042611; FDR = 0.005) alongside reduced expression of TGF- β response pathways (e.g., GO:0071559; FDR = 1e-4). Oligodendrocytes exhibited reduced expression of genes associated with cell projection and cytoskeleton-associated components (e.g., GO:0000904; FDR = 7e-11), alongside increased expression of genes with roles in antigen processing and presentation (e.g., GO:0003823; FDR = 1e-05). Finally, the strongest signal in OPCs involved reduced expression of genes with roles in synapse organization and cell-cell communication (e.g., GO:0097060; FDR = 9e-15), consistent with loss, weakening, or reduced turnover of neuron-OPC synaptic contacts.

These findings point to a broad set of age-associated molecular changes in the striatum, largely cell-type-specific, and affecting MSNs in particularly strong ways.

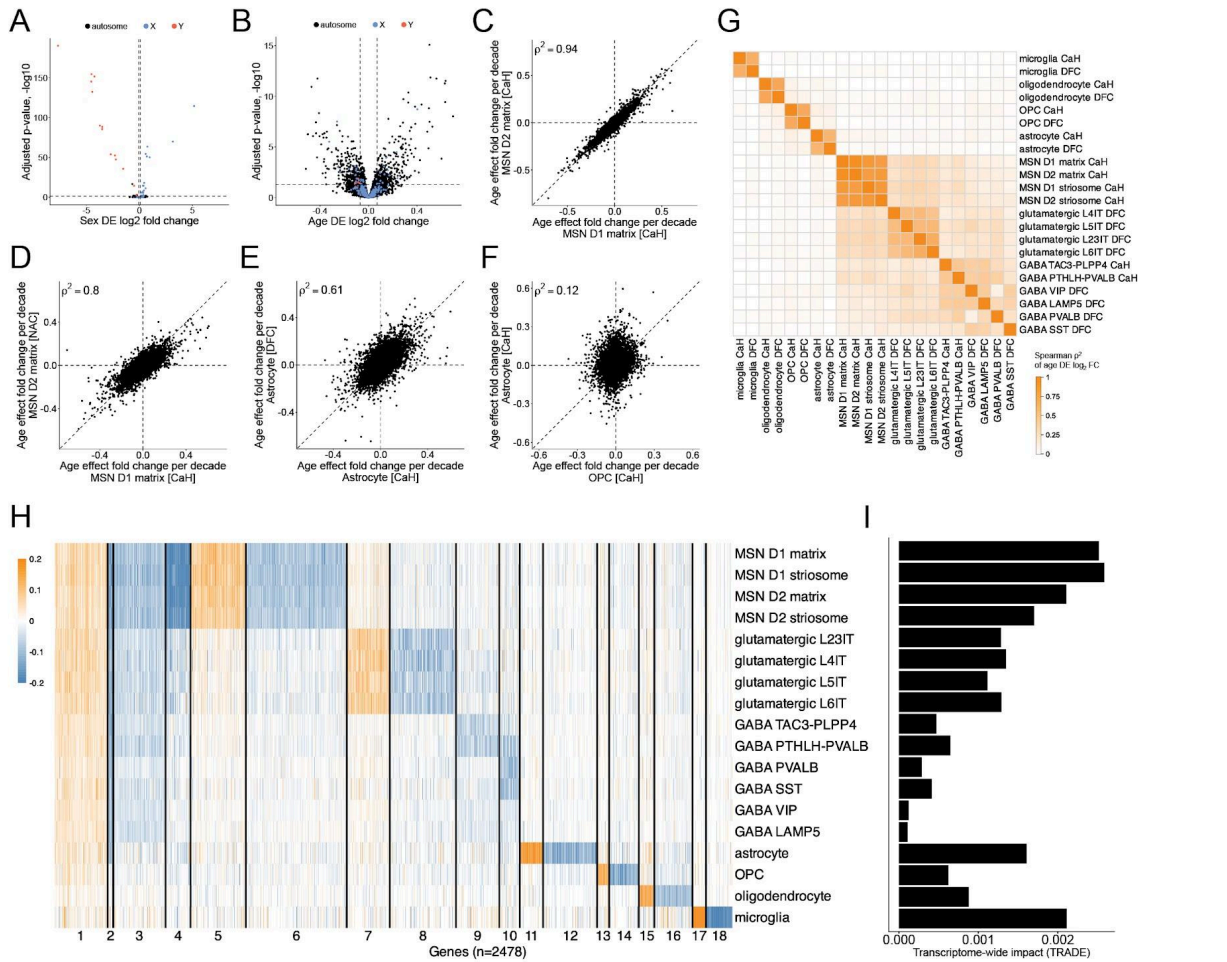


Figure 4. Age-associated changes in gene expression across brain cell types. (Extended analyses are in **Figure S11**). **A.** Analysis of sex differences in astrocyte gene expression (volcano plot). Each point represents one gene. Genes on the X chromosome, Y chromosome, and autosomes are shown in blue, red and black, respectively. **B.** Analysis of age-associated gene-expression changes in astrocytes (volcano plot). Effect sizes (x-axis) represent log₂ fold change per decade of age (see **Methods**). Chromosomal locations of genes are indicated as in panel **A**. **C.** Comparison of age-associated gene-expression changes between D1 MSNs (x-axis) and D2 MSNs (y-axis) (matrix subtype, sampled in the caudate; Spearman's ρ). **D.** Comparison of age-associated gene-expression changes between D1 MSNs (matrix subtype) sampled in the caudate (dorsal striatum) or the nucleus accumbens (ventral striatum). **E.** Comparison of age-associated gene-expression changes between striatal astrocytes (x-axis) and cortical astrocytes (y-axis). **F.** Comparison of age-associated gene-expression changes between OPCs (x-axis) and astrocytes (y-axis) in the caudate **G.** Heatmap shows correlation of age-associated gene-expression changes for each pair of cell types. Colors show Spearman's ρ^2 for correlations of gene-level log₂-fold-change per decade of age. An extended version of this analysis is in **Figure S11A**. **H.** Clustering of genes by their patterns of age-associated expression changes in the various cell types. Columns show genes (n = 2478) grouped by k-means clustering of their age-associated expression changes (log₂-fold-change per decade of age, shown in shades of orange and blue) across the various cell types. An extended version of this analysis is in **Figure S11D**. **I.** Magnitude of transcriptome-wide age-associated expression changes in each cell type, as estimated from these data using TRADE⁶¹.

Inter-individual variation in biological aging

Differential expression analyses identify genes whose expression tends to increase or decrease with advancing age. An interesting question involves the kinetics of these changes in individual persons, who could in principle exhibit faster-than-average or slower-than-average aging – either within specific cell types or across all cell types. Inspired by the informativeness of methylation-based aging clocks in analyzing blood samples⁶³, we sought to ask whether cell-type-specific RNA-expression measurements could be similarly informative about age-associated processes in brain cell types. To do this, we constructed an age prediction model for each cell type using a penalized regression model (elastic net⁶⁴, see **Methods**). This model generated, for each donor and cell type, both a predicted age and an age residual (defined as the difference between predicted and chronological age).

We trained separate models for each cell type (by region) and evaluated their ability to estimate each donor's age from RNA-expression data alone. These models were trained using the same cell-type-specific age-associated genes identified in the previous section. Across models, RNA-expression-predicted age was highly correlated with actual chronological age (**Figure 5C**). The median of the model-level mean absolute error was 6.6 years (range 5.1–14.4 years), with a corresponding median absolute deviation of 5.6 years (**Figure S13A**). Model accuracy depended strongly on the number of age-associated differentially expressed genes (adjusted $R^2 \approx 0.72$) and was thus greater for MSNs (**Figure S13B**). Predictions at younger and older ages were shifted toward the cohort mean (**Figure 5C**). As with overall prediction error, the magnitude of this shift tracked the number of age-associated differentially expressed genes available to the model (**Figure S13C**). Uncorrected residual age showed a shared age-dependent bias across models (**Figure S14A**). To account for this bias, we defined residual age relative to the fitted relationship between predicted and chronological age, which eliminated this age dependence and removed the correlation driven by this shared bias (**Figure S14B**).

For each cell type, the difference (residual) between chronological age and RNA-expression-predicted age could represent simple prediction error, or a difference between biological aging (which may vary across individuals and cell types) and simple chronological age. In support of the latter interpretation, we found that these residuals were positively correlated between all pairs of cell types in the analyses (**Figure 5D**), despite the fact that the genes used to estimate biological age in different cell types were almost completely distinct (Jaccard index < 0.1 for most cell-type pairs) (**Figure 5E**). This suggests that these cell-type-specific "aging clocks" exhibit a tendency to run faster or slower together in an individual person, a tendency that could in principle result from shared influences such as circulating factors, life history, or genetics. This correlation was only partial, though, and exhibited fluctuations that appeared to be biologically meaningful: for example, the correlations-of-residuals were generally larger when comparing the same cell type across brain regions than when comparing different cell types within a brain region (**Figure S14C,D**).

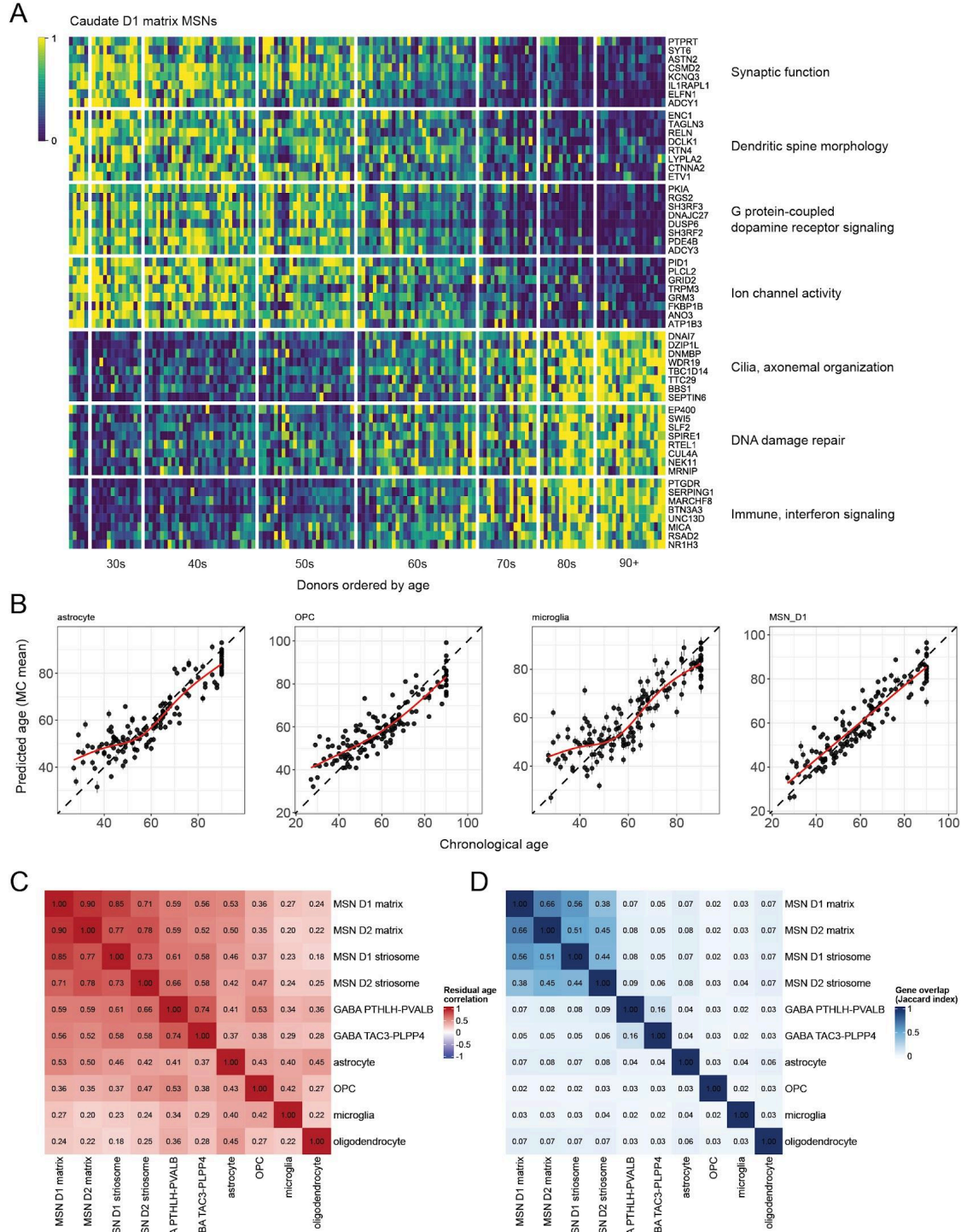


Figure 5. Chronological age can be predicted from gene expression in various cell types. (Extended analyses are in **Figures S13** and **S14**.) **A.** Age-associated changes in D1 matrix MSN gene expression, within the caudate. The 150 brain donors (columns) are ordered along the x-axis by ascending age (27 to 90+). Expression values (transcripts-per-million) for each gene (row) have been rescaled such that 0 (blue) represents the 10th percentile among all donors and 1 (yellow) represents the 90th percentile. **B.** RNA-expression-based age prediction in four representative caudate cell types. For each cell type, donor chronological age (x-axis) is plotted against RNA-expression-predicted age (y-axis), with predictions derived from an elastic net model. Each point represents one donor. The dashed line denotes the identity line (predicted = chronological age). Vertical error bars indicate plus or minus 1 SD around the predicted age. The red curve shows a Gaussian additive model (GAM) fit of predicted age as a function of chronological age. Age residuals (expression-predicted minus chronological age) are defined relative to this fitted relationship. **C.** Positive correlation of cell-type-specific RNA-expression "clocks", beyond the shared effects of chronological age. The heatmap shows pairwise correlations (for each pair of caudate cell types) of the GAM-corrected age residuals. **D.** Minimal overlap of the genes used to predict age in most cell types. The heatmap shows Jaccard indices quantifying overlap among age-associated genes used in each cell-type-specific model.

Effects of common genetic variation on gene expression

A pervasive source of biological diversity is the genetic variation that percolates through human populations. Any two people will generally have tens of millions of genetic differences. Most phenotypes are genetically complex – shaped by common and rare genetic variation at many loci, and thus requiring large samples for genetic mapping. However, samples of the current size (100 - 200 donors) are often sufficient for identifying strong effects of common genetic variants and haplotypes on expression levels of nearby genes (expression quantitative trait loci, eQTLs). To map common genetic effects on gene expression in each of the cell types of the striatum and cortex, we jointly analyzed the snRNA-seq and whole-genome sequence data from the brain donors.

We identified 9,899 genes with significant evidence (at a false discovery rate of 0.01) of cis-regulation ("eGenes") by a nearby SNP or haplotype in one or more cell types (three illustrative examples are shown in **Figure 6A**). As expected, more deeply-sampled cell populations such as D1 matrix MSNs yielded more eQTL discoveries ($n = 4,530$ unique eGenes detected in DFC and CaH) than sparser cell types such as microglia did ($n = 408$) (**Table S6**).

To systematically characterize patterns of cell-type specificity among these 9,899 eQTLs, we used k-means clustering to group eQTLs by shared patterns of "effect sizes" – the fractional change (in each cell type) in gene expression per allele inherited – across the full set of cortical and striatal cell types (**Figure 6B**, see **Methods**).

Some 16% of the eQTL effects we ascertained ($n = 1,540$ of 9,899; gene clusters 1 - 3 in **Figure 6B**) were shared across most of the cell types in the analysis, though often excluding microglia, which arise from a different developmental lineage⁶⁵. Such effects varied in magnitude across the cell types but almost never varied in direction (more versus less expression for a given allele). For example, expression of the *XRRA1* gene associated strongly with the same common haplotype (tagged by the SNP rs10899052) in every cell type of the striatum and cortex (**Figure 6A**, first row).

Another 16% of the eQTLs ($n = 1,559$ of 9,899; clusters 4 and 5 in **Figure 6B**) were shared across many neuronal classes (including projection neurons and interneurons), often with more-modest effects in one or more of the neuroglial cell types.

A large majority of eQTL effects, though, were specific to a cell type, almost always manifesting in that cell type in all of the brain regions in which that cell type was present (**Figure 6B, S16**). These included clusters of eQTLs specific to MSNs (clusters 6 and 7), interneurons (cluster 8), cortical pyramidal neurons (cluster 9), astrocytes (cluster 10), oligodendrocytes (cluster 11), OPCs (cluster 12), and microglia (cluster 13) (**Figure 6B**). Specific examples included a common genetic effect on *NPAS3* expression in microglia (**Figure 6A**, second row) and a common genetic effect on *CEP112* expression in MSNs of all types (D1 and D2; striosome and matrix) (**Figure 6A**, third row).

Notably, *NPAS3* illustrated a broader pattern in which cell-type-specific eQTLs often appeared in cell types that did not express the gene most strongly: microglia had the lowest *NPAS3* expression of all cell types in the analysis, yet were the only cells in which this genetic effect clearly manifested; and the genetic effect at *CEP112* tended to reduce expression in MSNs below the expression levels in other neurons and astrocytes (**Figure 6A**). In seven of eight cell-type-specific clusters (clusters 6 - 12), only 24 - 42% of eGenes had their highest median expression in the cell type in which they manifested cell-type-specific regulatory effects; for the microglia-specific cluster (cluster 13), this fraction was 62% (**Table S7**).

Overall, some 68% of the eQTL effects ($n = 6,800$ of 9,899; clusters 6 - 13) exhibited one of these forms of cell-type specificity – a number that is likely a lower bound on the true cell type specificity of brain eQTLs, since a cell-type-specific eQTL has fewer chances to be discovered at statistical significance.

These data also afforded the opportunity to compare cell-type specific eQTLs across striatum and cortex. Glial cell eQTLs were almost identical between cortex and striatum: nearly all genetic effects that manifested in striatal astrocytes were also apparent in independent analyses of cortical astrocytes, and vice versa; the same was also true for OPCs, oligodendrocytes, and microglia (**Figure 6B,C**). The various types of MSNs (D1 and D2 MSNs; and matrix and striosome MSNs) also tended to exhibit highly similar effects of common genetic variants (**Figure 6B,C**), as did MSNs from different striatal regions (**Figure S16**).

Neurons with more-distinct molecular identities, however – such as MSNs, interneurons, and cortical pyramidal neurons – tended to manifest highly divergent sets of genetic effects (**Figure 6B,C**). This may foreshadow the existence of a far-larger, heterogeneous set of common genetic effects that remain to be discovered in other neuronal types beyond the striatum and neocortex, especially among the astonishing variety of neuronal types that are found beyond the telencephalon^{66,67}.

Among the eQTLs, genes that are intolerant to loss-of-function mutations⁶⁸ tended to have smaller genetic effect sizes (fold-changes) than less-constrained genes did (**Figure S17**). This was true for the eQTLs as a whole ($p = 9.3e-19$), and for each cell type separately, and for each eQTL cluster (1 through 13) defined by patterns of cell-type specificity (**Figure S17**). This aligns with a recent finding about human blood eQTLs^{69,70} and suggests that eQTLs in the most functionally critical genes tend to involve more-modest quantitative effects – and thus require larger cohorts to recognize at genome-wide significance. This relationship may foreshadow that, as sample sizes for such analyses continue to expand, the number of findings that involve constrained and disease-affecting genes will grow disproportionately.

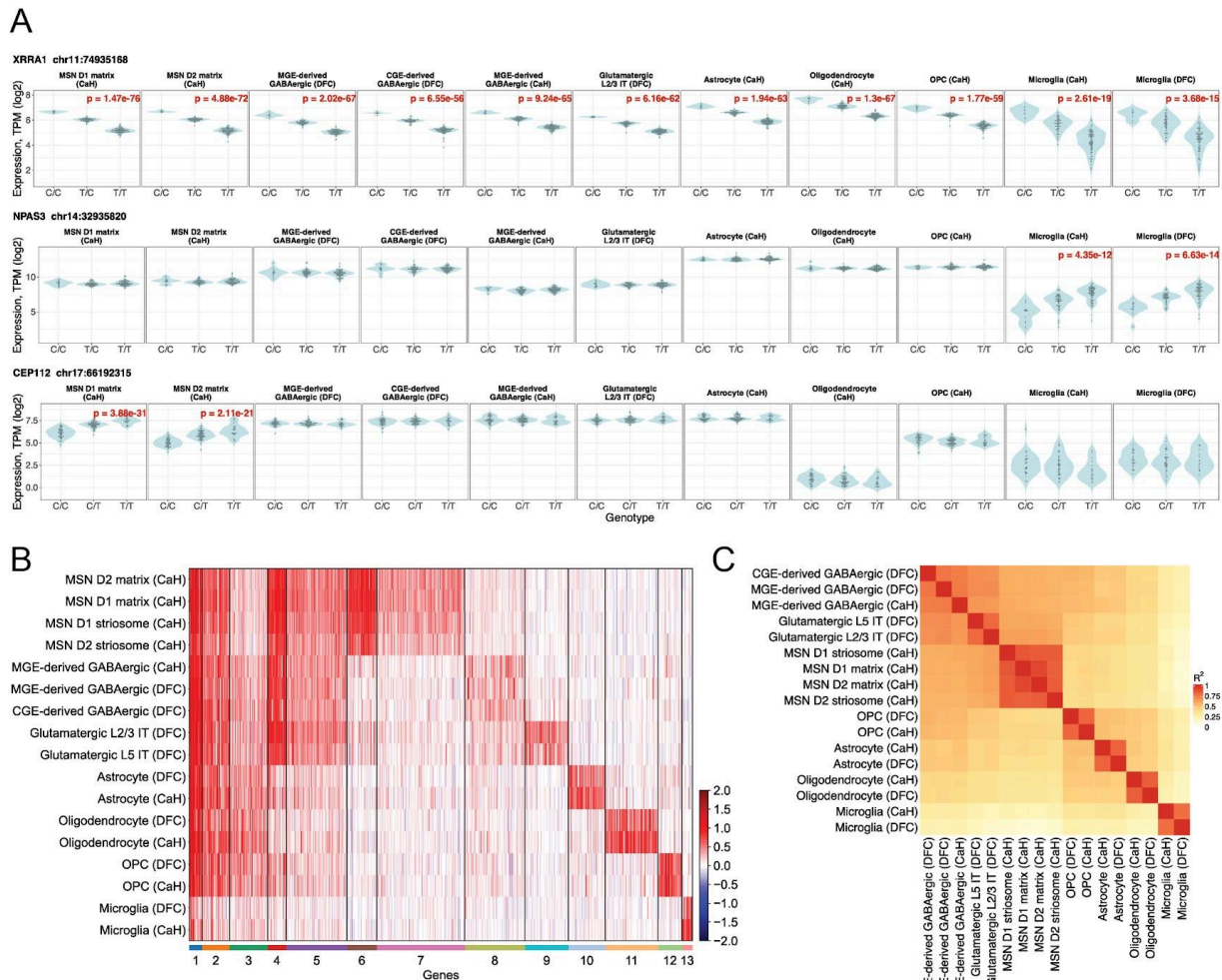


Figure 6: Effects of common genetic variation on gene expression. (Extended analyses including additional cell populations and brain regions are in **Figures S16** and **S17**.) **A.** Log₂-transformed expression levels of three genes (of 9,899) whose expression levels associated with common genetic variants: *XRRA1* (rs10899052), *NPAS3* (rs61980314), and *CEP112* (rs7209037). Brain donors' expression levels are shown for each genotype (as violin plots), for a variety of striatal and cortical cell types (red text indicates Benjamini–Hochberg adjusted p values). **B.** K-means clustering (K=13) of the 9,899 eQTLs based on their patterns of effect sizes on gene expression across the cell types in the analysis. Effect sizes represent standardized change in gene expression (inverse-normalized log₂-transformed values). For each effect, shades of red indicate the most common effect direction and shades of blue indicate any cell types with an opposite direction of association. **C.** Pairwise correlations (among cell types) of genome-wide sets of eQTL effects (corresponding to the rows in panel B). For each cell-type pair, Spearman's ρ^2 was calculated using eQTLs significant in at least one of the two cell types. Rows and columns are hierarchically clustered. The analyses in panels B and C are extended to additional brain regions in **Figure S16**.

Discussion

Decades of work by brain banks, and the generosity of brain donors and their families, is increasingly meeting new genomic analysis technologies that unlock the cellular and molecular information in these tissues. Human brain tissues offer unparalleled opportunities to investigate pathogenic processes, especially for the many human brain disorders that lack mechanism-based animal models.

At the same time, such work lacks the factors that have been the cornerstone of most laboratory studies and biological training: controlled laboratory conditions studying a single experimental variable in isolation. Human tissue studies operate in a context in which thousands of factors vary simultaneously – often without prior knowledge about the magnitude or independence of this variability. Better appreciating the patterns that underlie and structure human biological variation can support the substantial intellectual and resource investment involved in the design and interpretation of human brain studies, from selecting donors to appropriately calibrating statistical analyses and claims. We hope that the data generated by this project provide a useful resource for such studies.

One of the most remarkable observations from these data involved the degree to which cell type proportions vary from person to person (**Figure 2**). The quantitative representations of astrocytes, microglia, OPCs, and striatal interneurons — which might have been reasonably assumed to be optimized parameters of the human brain — were in fact highly variable (**Figure 3F,G,H**). This manifested in our data as consistent properties of individuals that re-appeared in all of the brain regions we analyzed from the same person (**Figure 2B,C,F-H**) – to an extent that was only partially (and for only certain cell types) explained by age effects (**Figure 3A,B**). This observation raises intriguing questions: what is the functional consequence of having a larger-than-average number of interneurons, OPCs, astrocytes, or microglia? How might such variation influence the biology of neural systems, the flexibility of cognitive processes, or the properties of neuronal circuits?

A fascinating question for much future work involves the causes and biological mechanisms that underlie this variation. For OPCs, age contributed to differences in abundance (in both gray- and white-matter tissue contexts, across striatum and cortex), but age alone explained only part of the interindividual variability that manifested across all of an individual's brain regions (**Figure 3B**). Sex had no apparent impact, despite ample statistical power to observe such relationships. Brain developmental processes seem likely to contribute, for example via heterochronic shifts in the timing of neurogenesis or gliogenesis, differences in migratory patterns, or variation in apoptotic pruning during development. Genetic factors seem likely to contribute as well, and as cohort sizes increase it will become possible to map such effects to specific genes.

Each cell type exhibited pronounced differences in gene expression across donors. Age, but not sex, had a strong relationship to gene expression in every cell type (**Figure 4A,B**) – and the vast majority of these effects were specific to each cell type (**Figure 4G,H**). We found that people's ages could be estimated with surprising accuracy (to about 5 years) based on only these RNA-expression patterns (**Figure 5B**) – but that, at the same time, there was clear

evidence that these same changes progress more slowly or rapidly in some individuals than in others, and (within individuals) in some cell types than others. In human brain tissue studies, analytically addressing such effects could in principle help to make other genetic and biological effects more clear. We hope that our data help provide a resource for such analyses and enable many new analytical approaches to be conceived, trained and critically evaluated.

Age appears to affect some brain cell types much more than others (**Figure 4H**). Of all the striatal and cortical cell types in our analyses, MSNs showed the largest effects of aging on gene expression, followed by cortical glutamatergic neurons (also a type of projection neuron); the various types of interneurons in these brain regions manifested much smaller changes. It is interesting to speculate whether this sensitivity arises in projection neurons' need to support spatially distant metabolic and physiological processes, or from their remarkable number of dendritic arbors and spines. The recognition of aging effects in a more-diverse set of types of neurons across the brain may well reveal clear patterns.

One way in which these results challenged our expectations was that the cell types with the greatest vulnerability to age-associated loss – OPCs and TAC3-PLPP4 interneurons – were not the cell types with the largest age-associated gene-expression changes (**Figure 4H**); in fact, OPCs showed the most modest changes among the glial cell types, and TAC3-PLPP4 interneurons showed the most modest expression changes among the neuronal cell types. In the case of OPCs, it may be that renewal from a stem population attenuates the effects of aging, even as the capacity of that stem population itself declines with age. In the case of these interneurons, one possibility is that age-associated changes lead to loss of the more severely affected interneurons (reducing their numbers, but also reducing the average amount of age-associated change among the surviving interneurons) but do not do this in projection neurons.

The data provided an opportunity to recognize common genetic effects on gene expression levels (eQTLs) in a wide variety of neuronal and glial cell types in multiple brain regions (**Figure 6**). We found that genetics shapes inter-individual variation in gene expression in ways that are remarkably cell type specific: a large majority of eQTLs identified predominantly affected expression in a particular neuronal or glial cell type (**Figure 6C**). Another clear pattern was that among cell-type specific genetic effects, the cell type in which a variant alters gene expression was often not the cell type with the highest baseline expression of that gene; simply knowing which cell type expresses a gene most strongly seems to provide a poor compass to recognizing where genetic effects manifest. Comparing cell-type-specific effects across striatum and cortex, we found that each glial cell type manifested very similar sets of genetic effects in cortex as in striatum. In contrast, sets of eQTLs for different types of neurons were quite distinct. This predicts that a vast set of neuron-specific genetic effects in other brain regions are yet to be discovered — among which MSNs and cortical projection neurons represent only a small fraction of total neuronal diversity.

Finally, we would like to highlight several observations that we hope could benefit the design and interpretation of research that uses human brain tissue.

The first involves the inclusion and exclusion criteria that are used to define "cases" and "controls" in disease-focused analyses of human tissue. It is often assumed that the key to successful analysis is to minimize within-group heterogeneity insofar as can be accomplished from available medical records. This often leads to the adoption of narrow inclusion criteria for affected individuals (to reduce disease heterogeneity) and to broad exclusion criteria for controls (e.g. excluding common health conditions), which can greatly reduce sample size. Our results suggest that even criteria such as those defining our core "normative" donor set admit abundant biological variation (**Figure 2,3**). The hoped-for reduction in variance obtained by narrow inclusion criteria and broad exclusion criteria may often be small compared to the biological variation that is present within each group. In the context of abundant inter-individual biological variation, it will usually be more informative to expand the sizes of case and control groups when possible, and to favor analyzing a larger number of donors over simple technical replication or deeper sequencing of cells from the same donors.

The second involves utilizing experimental approaches that minimize technical variability, recognizing that biological variation is ubiquitous yet often mixed with technical and statistical-sampling effects on variables of interest. Here we utilized an approach for isolating nuclei in large pools ("villages") of about 20 donors handled as a single sample and therefore exposed to identical handling conditions. Key factors such as age and (in some study designs) sex or diagnosis are ideally balanced across pools as much as possible. Other alternative experimental approaches could in principle be developed in the future. Any such approach should be evaluated empirically; reasonable metrics could include reductions in the variance (across donors) of quantitative measurements, and increased statistical power for recognizing already-established biological effects (such as the thousands of effects of age on gene expression described here).

The third involves the selection of appropriate statistical tests. The vast majority of the measurements we made – whether of cell-type proportions, or gene expression levels – did not exhibit "normal" (Gaussian) distributions. For this reason, the conventional statistical tests that are used in most biological studies – including t-tests, Pearson correlations, and conventional linear regressions (all of which assume such distributions) – will routinely overestimate the statistical significance of the inevitable average differences between experimental groups. We encourage researchers to use the data from this project to assess the ways in which biological variables of interest are empirically distributed – and to design studies, interpret data and perform statistical tests with these distributions in mind. This will often mean using non-parametric statistical frameworks and permutation tests, which are far less vulnerable to inflating the apparent significance of statistically unremarkable differences between experimental groups.

Our hope is that these data and analyses, and much other such work to come, will make natural human biological variation an ever-more-useful tool to neurobiology – a source not of unwelcome analytical complexity, but of new questions, new discoveries, and a deeper understanding of human biology.

Limitations of the study

Our analyses do not identify the biological mechanisms by which these sources of biological variation arise, nor (beyond the effects of common genetic variants on gene expression) make any particular claim about the relative contributions of genetics and environment. Such questions deserve extensive future work by a wide variety of scientific approaches. We hope our results inspire the conception of such studies and inform their design.

Humans have an enormous variety of subclinical biological conditions, as well as conditions that do not receive medical attention or a formal diagnosis. Our observations that certain kinds of biological variation appear frequently and in specific patterns make no claims about whether or not such biological variation is "healthy", affects brain functions, or contributes to vulnerabilities – in fact, we believe this is an important topic for future research.

Resource Availability

Lead Contact

Requests for further information and resources should be directed to the lead contact, Steven A. McCarroll (smccarro@broadinstitute.org)

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

In this study, we used data we generated for the BICAN (RRID:SCR_022794) which are accessed/deposited in the NeMO Archive (RRID:SCR_016152), under accession number: /nemo:prj-hqeuy8h.

Data may be accessed at: <https://assets.nemoarchive.org/grant/nemo:prj-hqeuy8h>

All original code has been deposited at GitHub (RRID:SCR_002630) and is publicly available as of the date of publication: <https://github.com/broadinstitute/bican-mccarroll-manuscript1> (RRID:SCR_028041).

We provide a publicly available web-based portal to query and visualize the data: <https://bican.mccarrolllab.org/app/bican>.

Acknowledgements

This publication was supported by and coordinated through the BRAIN Initiative Cell Atlas Network (BICAN) and funded by the National Institute of Mental Health (UM1MH130966). We thank the NIH NeuroBioBank network for providing tissue for this project. We also thank all of the brain donors and their families; this work would not be possible without their generous gifts

to science. We thank Sabina Berretta, Mark Daly, Elise Robinson, Marta Florio, and Avin Veerakumar for comments and suggestions on manuscript drafts.

Author Contributions

S.A.M., K.I., and E.Z.M. conceived the study. C.V., C.J.M., N.A.R., K.K., J.M., S.D., H.G., N.B., and L.M. generated the data, with assistance from O.C., M.C., R.G., L.S., and C.C. O.Y., S.B., J.N., E.M., and J.Y. conceived and pursued the analyses, with assistance from K.S., M.G., A.W., L.R., N.A.R., H.F., G.G., S.F., N.R., and S.K., and mentorship from S.A.M. J.N., K.S., and A.W. created computational infrastructure and workflows that supported a variety of analyses. S.B., O.Y., J.N., E.M., and J.Y. interpreted data, with assistance from N.A.R., H.G., H.F., A.W.K., N.K., and A.M., and mentorship from S.A.M., K.I., and E.Z.M. S.K. created a browser for interacting with the data. E.F., M.H., and K.F. provided project management support. K.I., S.B., O.Y., E.M., J.Y., and S.A.M. wrote the manuscript, with input from all authors.

Declaration of Interests

E.Z.M. is a founder of Curio Bioscience.

Declaration of Generative AI and AI-assisted Technologies

During the preparation of this work, the authors used GitHub Copilot, GitHub Copilot CLI, Claude, ChatGPT, Le Chat, Perplexity, Phind.com, Groq, and Gemini for searching for the exact syntax of existing utility methods such as filtering and sorting, searching for packaged Docker images, software installation instructions, debugging software errors, generating workflow code from templates, and potential software performance optimizations. Grammarly was used for spell checking and grammar suggestions. Additionally, during the preparation of this work the authors used ChatGPT for minor editing/proofreading of manuscript text. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Supplemental figures

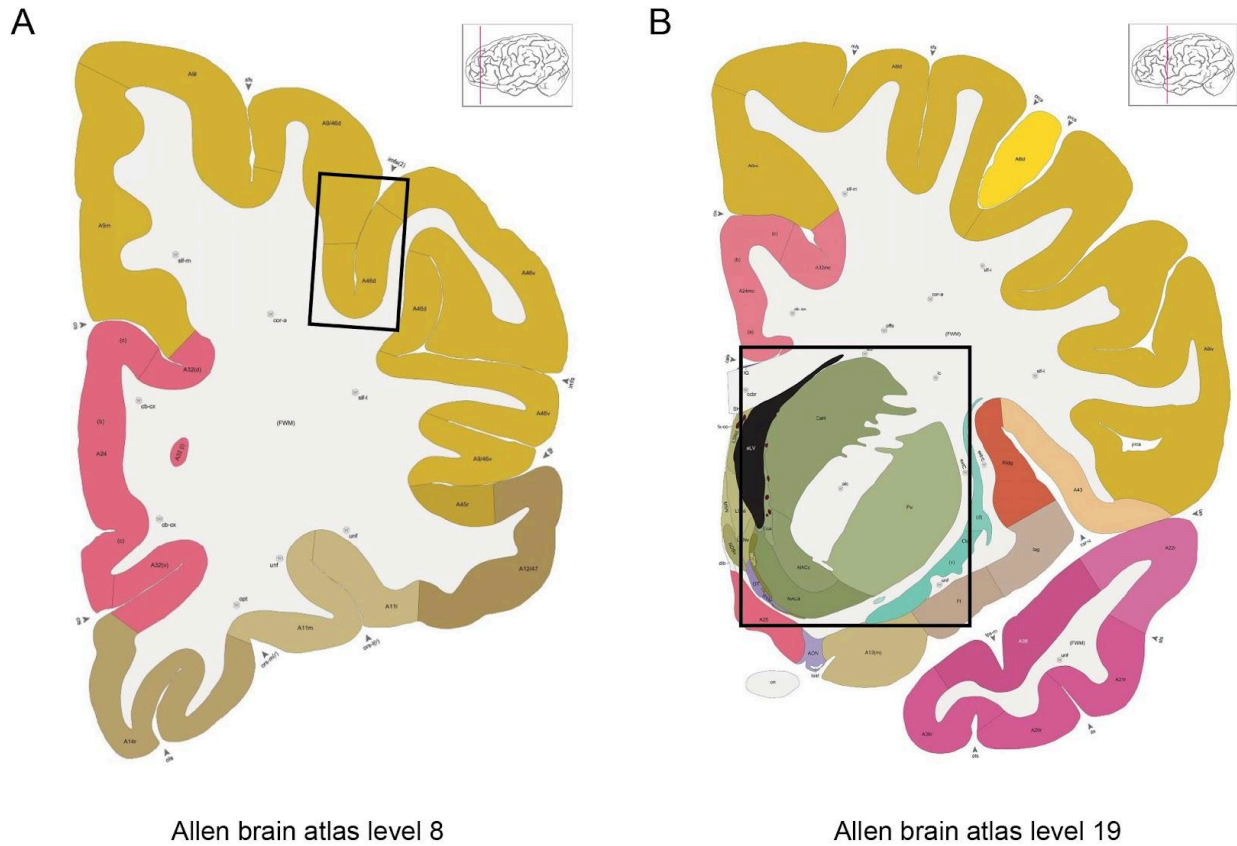


Figure S1. Tissue procurement and dissection reference for regions of interest (ROIs). Images from the Allen Institute adult human brain atlas, modified Brodmann annotation³⁴. Black box represents the requested ROIs from NBB. **A.** Atlas level for dorsolateral prefrontal cortex (DFC) samples. **B.** Atlas level for the striatum complex (caudate, putamen, nucleus accumbens, and internal capsule) samples.

Figure S2. Overview of donor sampling, metadata structure, and cell type annotations.

A. Donor sampling across brain regions. For each pair of brain regions, the heatmap shows the number of donors for which both regions were sampled. Values along the diagonal are the total donor counts for each region, as shown in **Figure 1B**. **B.** Counts and age distribution of female and male donors. **C.** Heatmap summarizing pairwise associations among demographic, genetic, technical, and sequencing-quality variables. Numeric–numeric pairs show Pearson correlations; factor–factor pairs show Cramér’s V ; and mixed numeric–factor pairs show the proportion of variation in the numeric variable explained by group differences. Pearson correlations range from -1 to 1 , whereas the other measures range from 0 to 1 . Circle size and color indicate the magnitude and direction of association. A complete description of variables is provided in **Table S3**. **D.** Cell-type specificity of gene expression measurements. Heatmap shows expression measurements for some of the genes (columns) whose expression levels distinguish major cell classes (rows). Colors represent normalized expression by column (across major cell classes) such that the highest expression is 1 (bright yellow) and lowest is 0 (dark purple). **E.** Confusion matrix comparing cell identities as inferred from two analytical approaches: MapMyCells label transfer from the consensus basal ganglia taxonomy (columns), and manual unsupervised clustering annotations (rows). Values (colors) represent fractions (of the total) in each row. **F.** Confusion matrix comparing cell identities as inferred from scPred label transfer from the Allen Brain human cortical neuron taxonomy (columns), and manual unsupervised clustering annotations (rows). Values (colors) represent fractions of the total in each row.

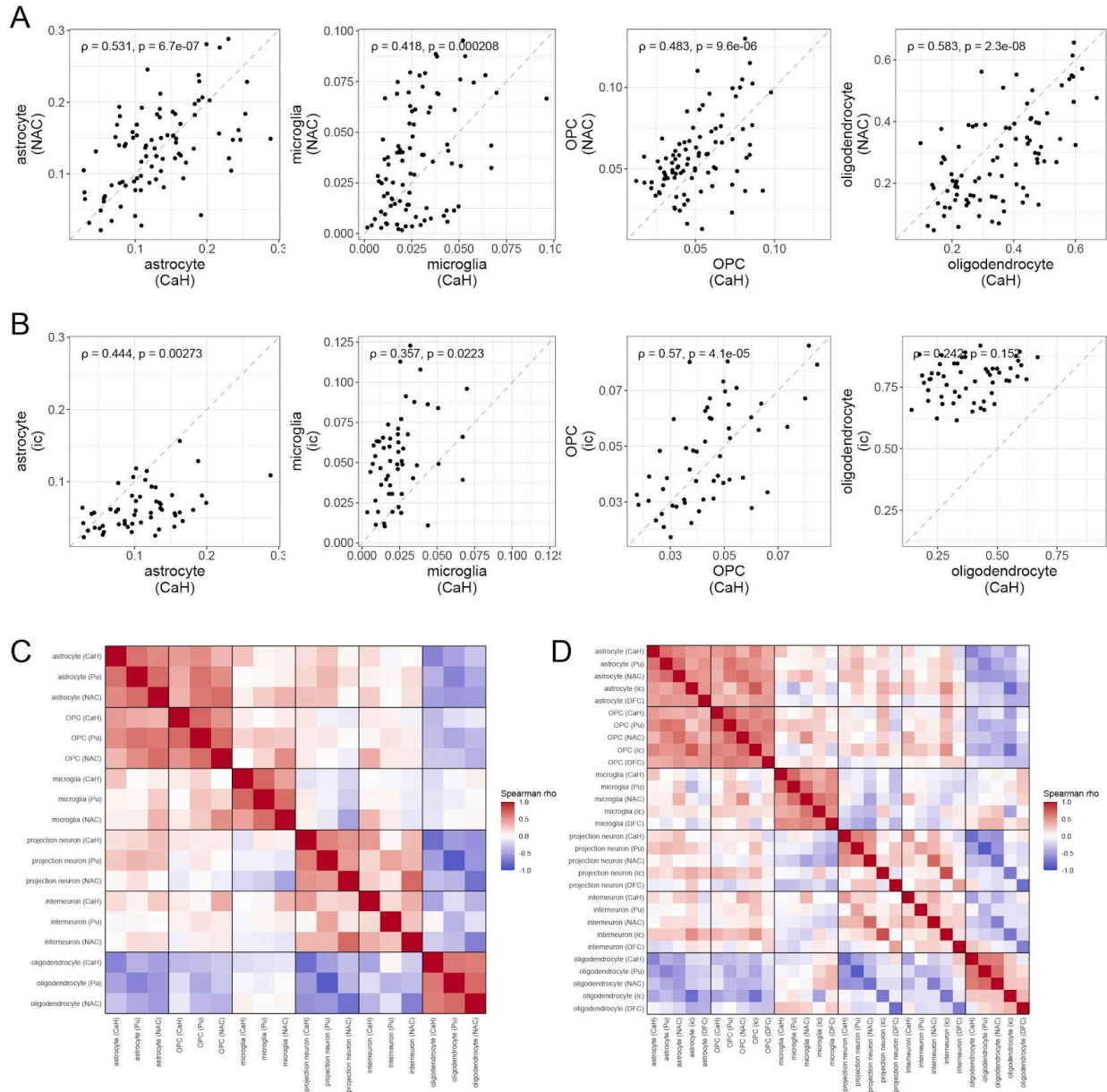


Figure S3. Cross-brain-region correlations in cell type abundances.

A. Representations of glial cell types (astrocytes, microglia, OPCs, oligodendrocytes) in the caudate (x-axis) and nucleus accumbens (y-axis) in the same 90 donors (points). Abundances are quantified as fractions of all nuclei ascertained in a donor. Spearman correlation coefficient and corrected p-value are shown for each cell type (Benjamini-Hochberg (BH) procedure, $n=498$ tests). **B.** Same as **A**, for the caudate (x-axis) and internal capsule (y-axis) across 54 donors. **C.** Pairwise correlations of cell type abundance measurements – quantified as fractions of all nuclei ascertained in a sample – in striatal regions (caudate, putamen, and nucleus accumbens). Spearman correlation coefficients are shown; positive correlations in red, negative in blue. Note the blocks of positive correlation (red) for specific cell types regardless of the brain regions in which they are measured. **D.** Same analysis as **C**, for all five brain regions surveyed.

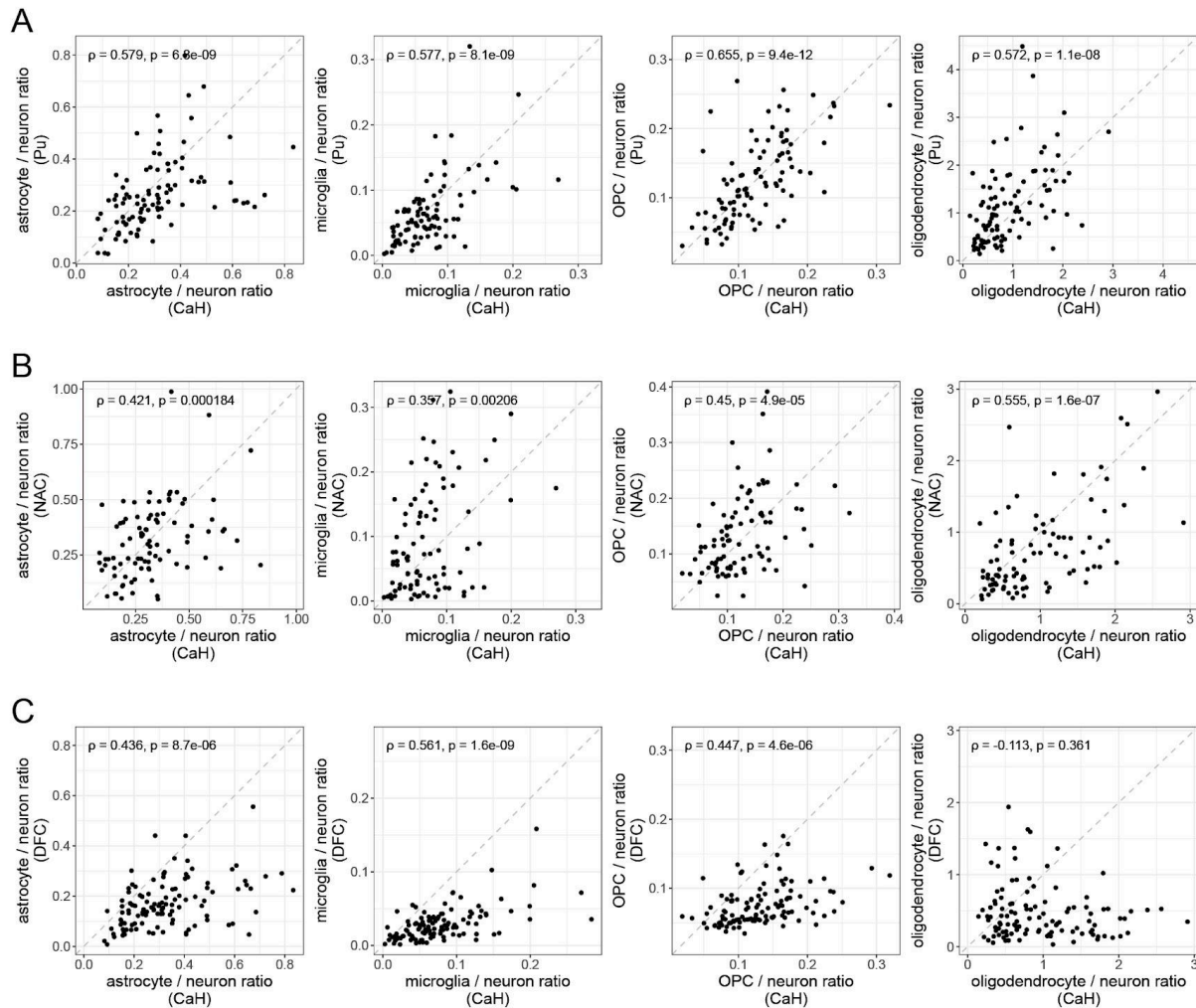


Figure S4. Cross-brain-region correlations in the abundances of glial cells.

A. Representations of glial cell types in the caudate (x-axis) and putamen (y-axis) across the same 98 donors. Here, abundances were quantified relative to the total number of neurons in the same sample, so that the glial cells do not influence one another's abundance estimates. Spearman correlation coefficient and corrected p-value are shown for each cell type (Benjamini-Hochberg (BH) procedure, $n=498$ tests). **B.** Same analysis as in **A**, focusing on correlation between caudate (x-axis) and nucleus accumbens (y-axis) across 90 donors. **C.** Same analysis as in **A**, comparing caudate (x-axis) and dorsolateral prefrontal cortex (y-axis) across 114 donors. Astrocyte, microglia, and OPC abundances are significantly correlated, indicating that these patterns extend beyond the striatum.

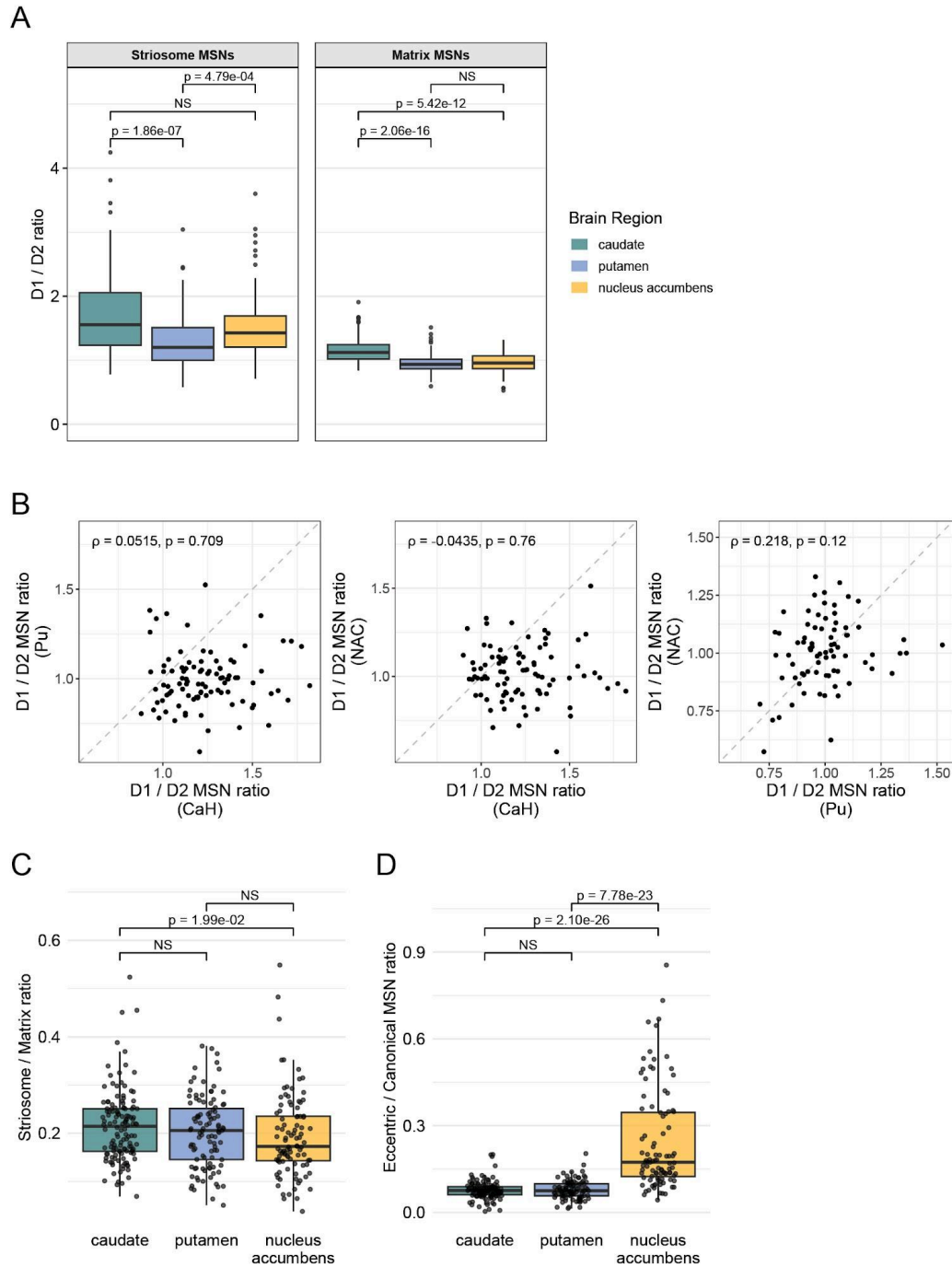


Figure S6. Representation of MSN subtypes across brain regions.

A. Comparison of D1 / D2 MSN ratios across CaH, Pu and NAC, stratified by striosome/matrix compartments and evaluated by Wilcoxon rank-sum test. **B:** Comparisons (and lack of significant correlations) between D1/D2 MSN ratios comparing different pairs of striatal regions. Spearman correlation coefficient and corrected p-values are shown (Benjamini-Hochberg (BH) procedure, $n=498$ tests). Left: caudate (x-axis) vs. putamen (y-axis), $n=98$ donors. Middle: caudate (x-axis) vs. nucleus accumbens (y-axis), $n=90$ donors. Right: putamen (x-axis) vs. nucleus accumbens (y-axis), $n=78$ donors. **C.** The striosome / matrix MSN ratio does not vary dramatically between regions, although there may be a slight striosome depletion in NAC, as evaluated by Wilcoxon rank-sum test. **D.** NAC is significantly enriched for eccentric MSNs relative to CaH and Pu, as evaluated by Wilcoxon rank-sum test.

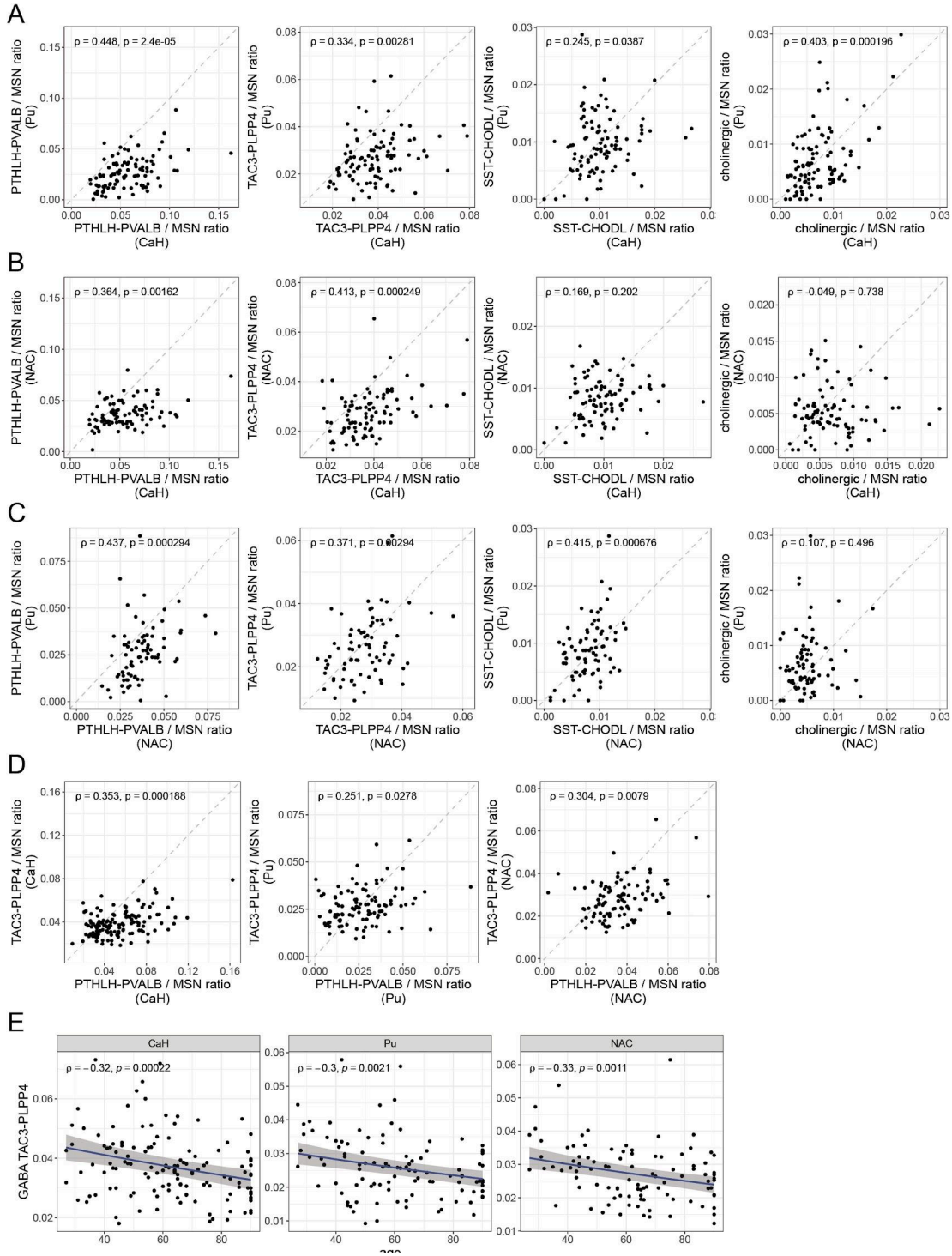


Figure S7. Relationships among the abundances of various types of striatal interneurons.

In all plots, abundances have been quantified relative to the number of MSNs (in the same tissue sample), to reduce dependence of these measurements on one another and to minimize the effects of variance in glial populations. **A.** Abundances of various types of interneurons in the caudate (x-axis) versus putamen (y-axis) across 98 donors (points). Spearman correlation coefficient and corrected p-values are shown (Benjamini-Hochberg (BH) procedure, $n=498$ tests). PTHLH-PVALB, TAC3-PLPP4, and cholinergic interneuron subtype abundances were significantly correlated between caudate and putamen. **B.** Same analysis as in **A**, comparing caudate (x-axis) and nucleus accumbens (y-axis) across 90 donors. PTHLH-PVALB and TAC3-PLPP4 interneuron subtype abundances are significantly correlated. **C.** Same analysis as in **A**, comparing nucleus accumbens (x-axis) and putamen (y-axis) across 78 donors. PTHLH-PVALB, TAC3-PLPP4, and SST-CHODL interneuron subtype abundances were significantly correlated. **D.** Within-sample comparison of PTHLH-PVALB and TAC3-PLPP4 abundances. Left: caudate ($n=131$ donors). Middle: putamen ($n=104$ donors). Right: nucleus accumbens ($n=97$ donors). Spearman correlation coefficient and corrected p-values are shown (BH-procedure, $n=498$ tests) **E.** Attrition of TAC3-PLPP4 interneurons with advancing age, in all striatal gray matter regions. Abundance is expressed relative to all neurons. Each point represents a donor, and blue lines indicate beta-binomial fits with 95% confidence intervals (gray ribbons). Abundance is negatively correlated with age for all brain regions (Spearman correlation coefficients and nominal p-values shown). Modeling with a beta-binomial regression confirmed a significant decline with age, shown with the regression line ($\beta = -0.047$ per decade; 95% CI: -0.069 to -0.025 ; BH-adjusted $p = 0.001$; $n=44$ tests; **Table S4**).

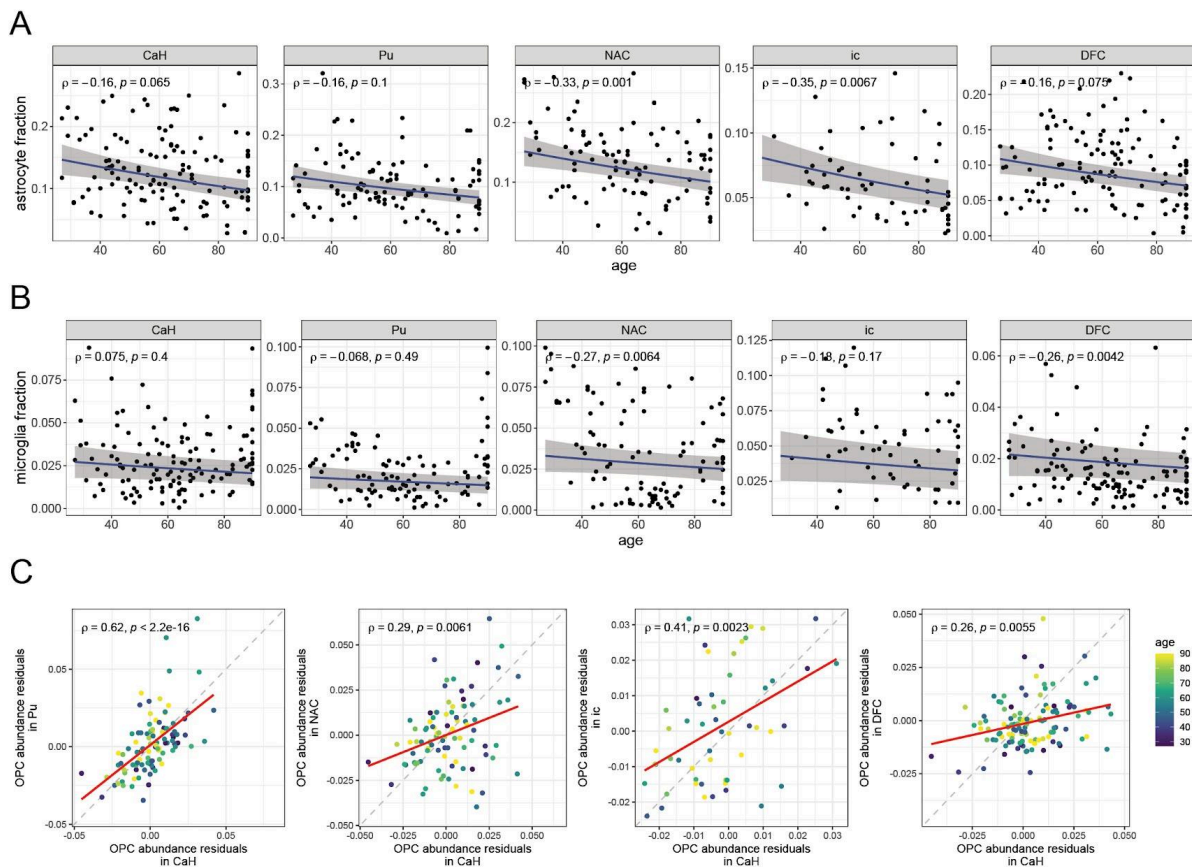


Figure S8. Additional analyses of glial cell abundances in relationship to age.

A. Abundance of astrocytes, as a fraction of all nuclei sampled. Each point represents a donor, and blue lines indicate beta-binomial fits with 95% confidence intervals (gray ribbons). While nominally significant negative Spearman correlations were observed in specific regions, the result of global modeling via beta-binomial regression ($\beta = -0.074$ per decade; 95% CI: -0.115 to -0.033 ; BH-adjusted $p = 0.018$; $n=44$ tests; **Table S4**) was not significant after correcting for multiple hypothesis testing; however, this is not a definitive negative result. **B.** Same as **A**, but for microglia abundance. Beta-binomial regression found no significant association between microglial abundance and age ($\beta = -0.045$ per decade; BH-adjusted $p = 1$; $n=44$ tests; **Table S4**). **C.** Comparison of OPC fraction residuals (observed minus age-predicted values from covariate-adjusted beta-binomial models) for the caudate compared to the other four brain regions analyzed. Each point represents a donor, colored by age; red lines indicate linear regression fits. The positive correlations in these residuals across all pairs of brain regions indicate that additional donor-level factors (beyond chronological age) influence OPC abundance in the striatum complex and DFC in a similar manner.

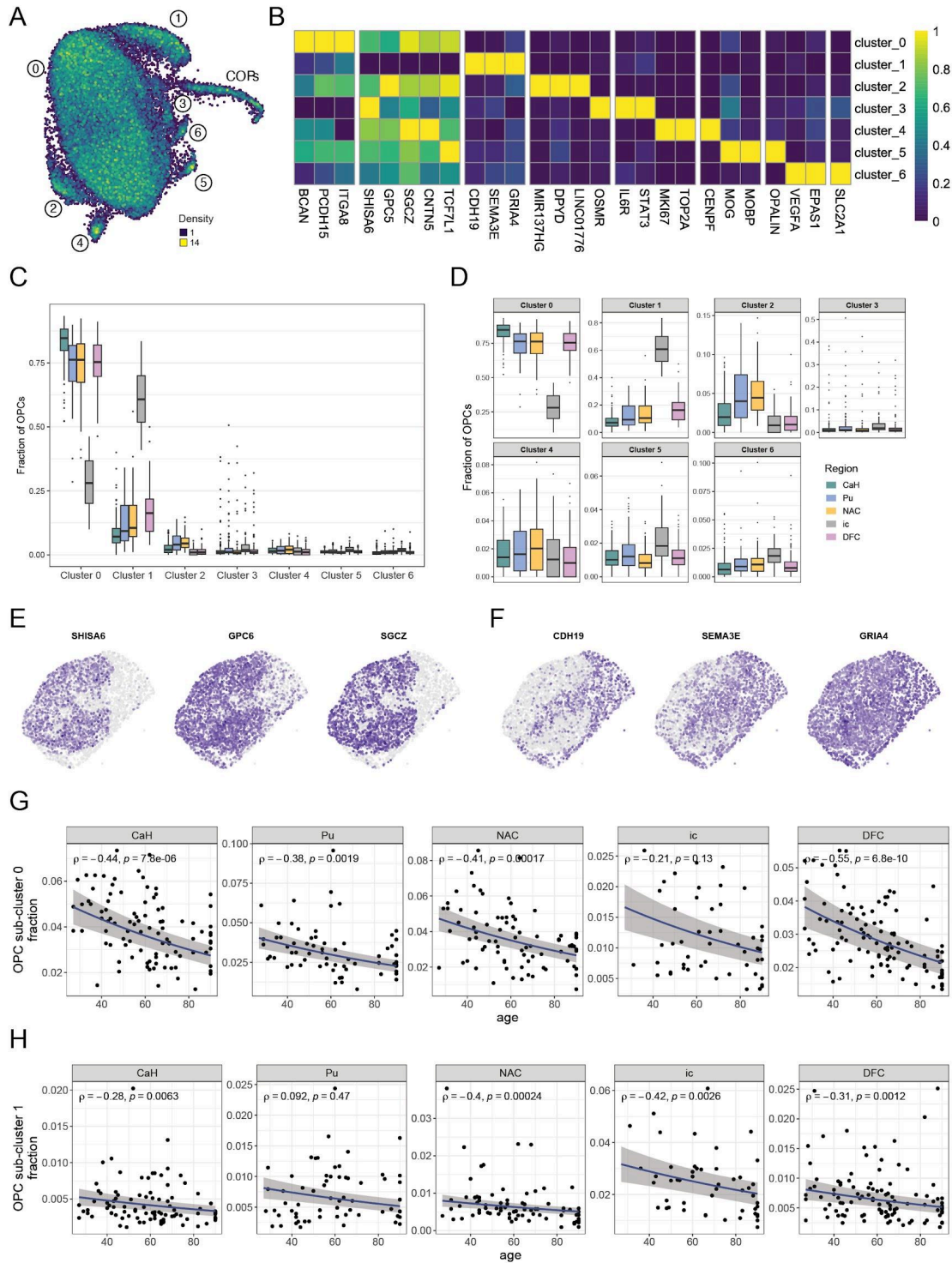


Figure S9. OPC diversity and attrition with age.

A. Unsupervised clustering assignments and UMAP representation of OPCs sampled from the caudate. Colors on the UMAP plot represent the density of nuclei in that region of the plot. **B.** Expression levels of specific marker genes (columns) for OPC sub-clusters (rows; note, COPs excluded). Colors represent normalized expression by column (across major cell classes) such that the highest expression is 1 (bright yellow) and lowest is 0 (dark purple). **C.** Distribution of OPC sub-cluster compositions across regions. Gray matter regions exhibit much lower numbers of cluster 1 OPCs and higher numbers of cluster 0 OPCs relative to the internal capsule. **D.** Distribution of OPC sub-cluster compositions across regions (free scales). **E,F.** Slide-tags spatial transcriptomics analysis of the striatum from a representative donor. Expression of genes whose expression defines OPC cluster 0 (e.g. *SHISA6*, *GPC6*, and *SGCZ*) highlights gray matter compartments. Color of the points represents log-normalized transcript counts for each gene (light grey lower expression, dark purple higher expression). Expression of genes whose expression defines OPC cluster 1 (*CDH19*, *SEMA3E*) highlights white matter compartments. *GRIA4* is expressed in all OPC sub-clusters, but most strongly in cluster 1, as seen in **B.** **G.** Decline in abundance of OPC cluster 0 with advancing age, across the brain regions sampled. Abundance is expressed relative to all nuclei sampled. Each point represents a donor, and blue lines indicate beta-binomial fits with 95% confidence intervals (gray ribbons). Spearman correlation coefficients and nominal p-values shown. Modeling with a beta-binomial regression confirmed a significant decline with age ($\beta = -0.095$ per decade; 95% CI: -0.122 to -0.067 ; BH-adjusted $p = 7.02e-10$; $n=44$ tests; **Table S4**). **H.** Same as **G**, but for OPC cluster 1. Modeling with a beta-binomial regression confirmed a significant decline with age ($\beta = -0.072$ per decade; 95% CI: -0.109 to -0.036 ; BH-adjusted $p = 0.004$; $n=44$ tests; **Table S4**).

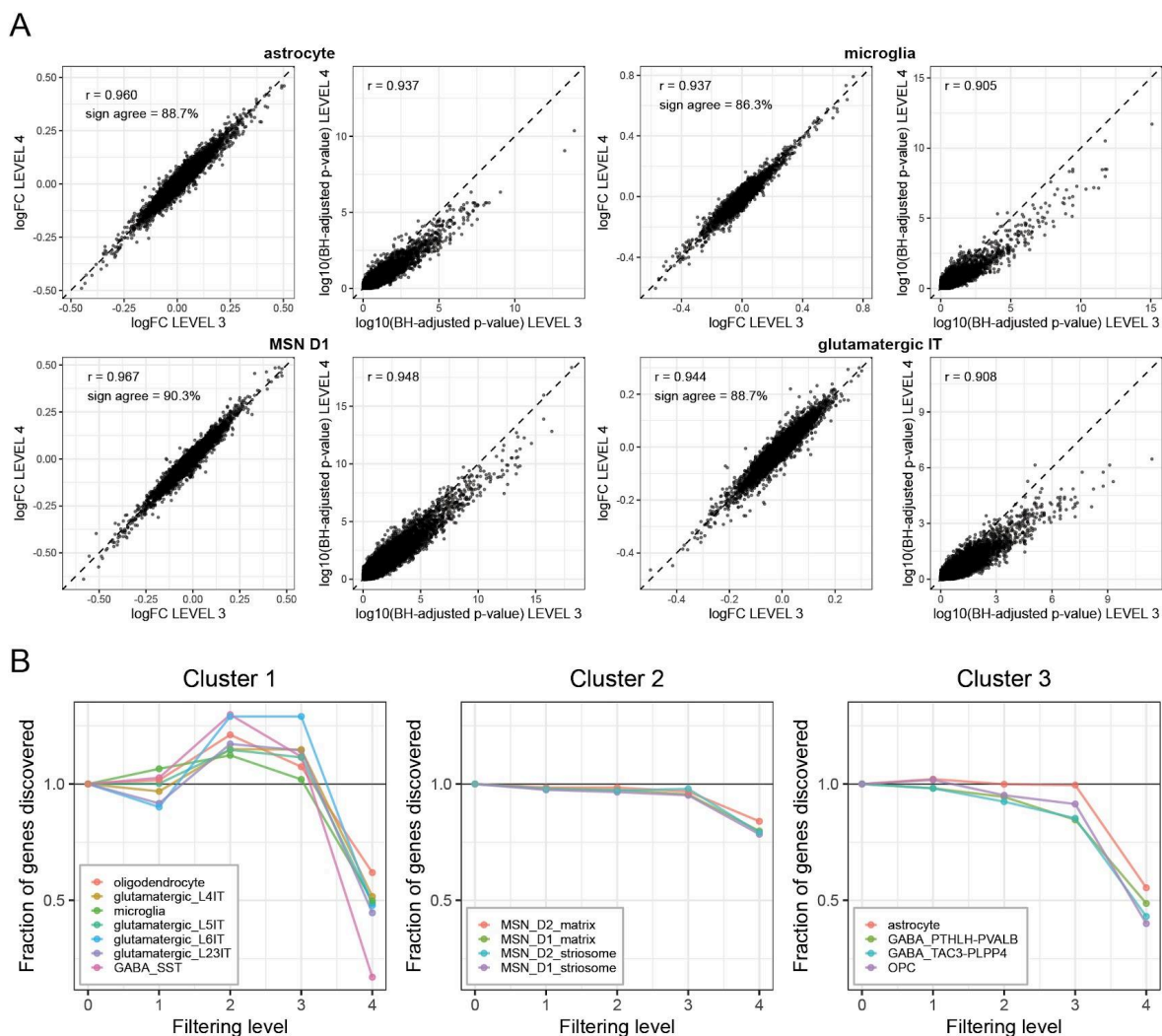


Figure S10. Effects of donor exclusion criteria on recognition and inference of age-associated changes.

A. Comparison, in multiple cell types, of gene-level log fold-change estimates and statistical significance ($-\log_{10}$ FDR-adjusted p-values) for age-associated genes as donor exclusion criteria are made more stringent. Stringency levels 0–3 represent progressively stricter snRNA-seq-data-driven exclusions; Level 4 additionally excludes donors based on metadata criteria from clinical records. Effect-size estimates remain highly concordant across filtering levels, and statistical significance is largely preserved or improved under snRNA-seq-data-driven filtering. In contrast, adding metadata-driven exclusions reduces statistical significance (and thus ability to distinguish real effects at genome-wide significance) while leaving effect-size estimates largely unchanged. **B.** Fraction of age-associated differentially expressed genes detected at each filtering level, normalized to the unfiltered dataset (Level 0), shown across representative cell types. Cell types were clustered using k-means clustering ($k=3$) based on the fraction of genes discovered at each filtering level to identify shared patterns. snRNAseq-data-driven filtering (Levels 1–3) preserved or increased the number of discoveries for nearly all cell types. Additional metadata-driven exclusions (Level 4) reduced the number of detected genes across cell types, indicating a loss of statistical power without changes in estimated effect sizes.

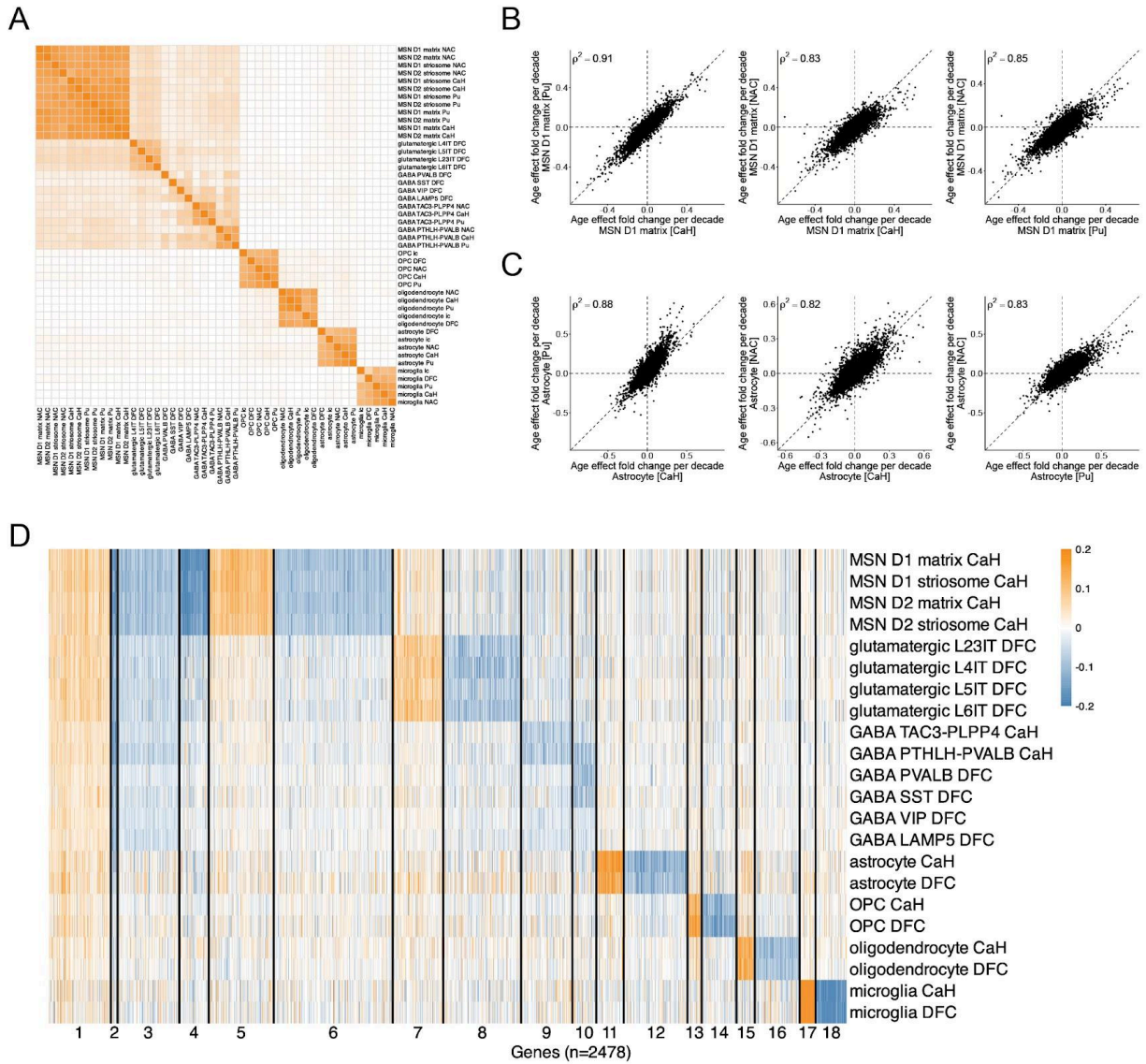


Figure S11. Extended analyses of aging effects. Same analyses as in **Figure 4**, but extended to additional cell populations. **A.** Heatmap shows correlation of age-associated gene-expression changes for each pair of cell types and brain regions. Colors show Spearman's ρ^2 for correlations of gene-level log₂-fold-change per decade of age. **B.** Comparison of D1 MSNs (matrix subtype) age-associated gene-expression changes between CaH and Pu, CaH and NAC, and Pu and NAC. **C.** Comparison of astrocyte age-associated gene-expression changes between CaH and Pu, CaH and NAC, and Pu and NAC. **D.** Clustering of genes by their patterns of age-associated expression changes in the various cell types and brain regions. Columns show genes ($n = 2478$) grouped by k-means clustering (same clustering plotted in **Figure 4H**) of their age-associated expression changes (log₂-fold-change per decade of age) across the various cell types

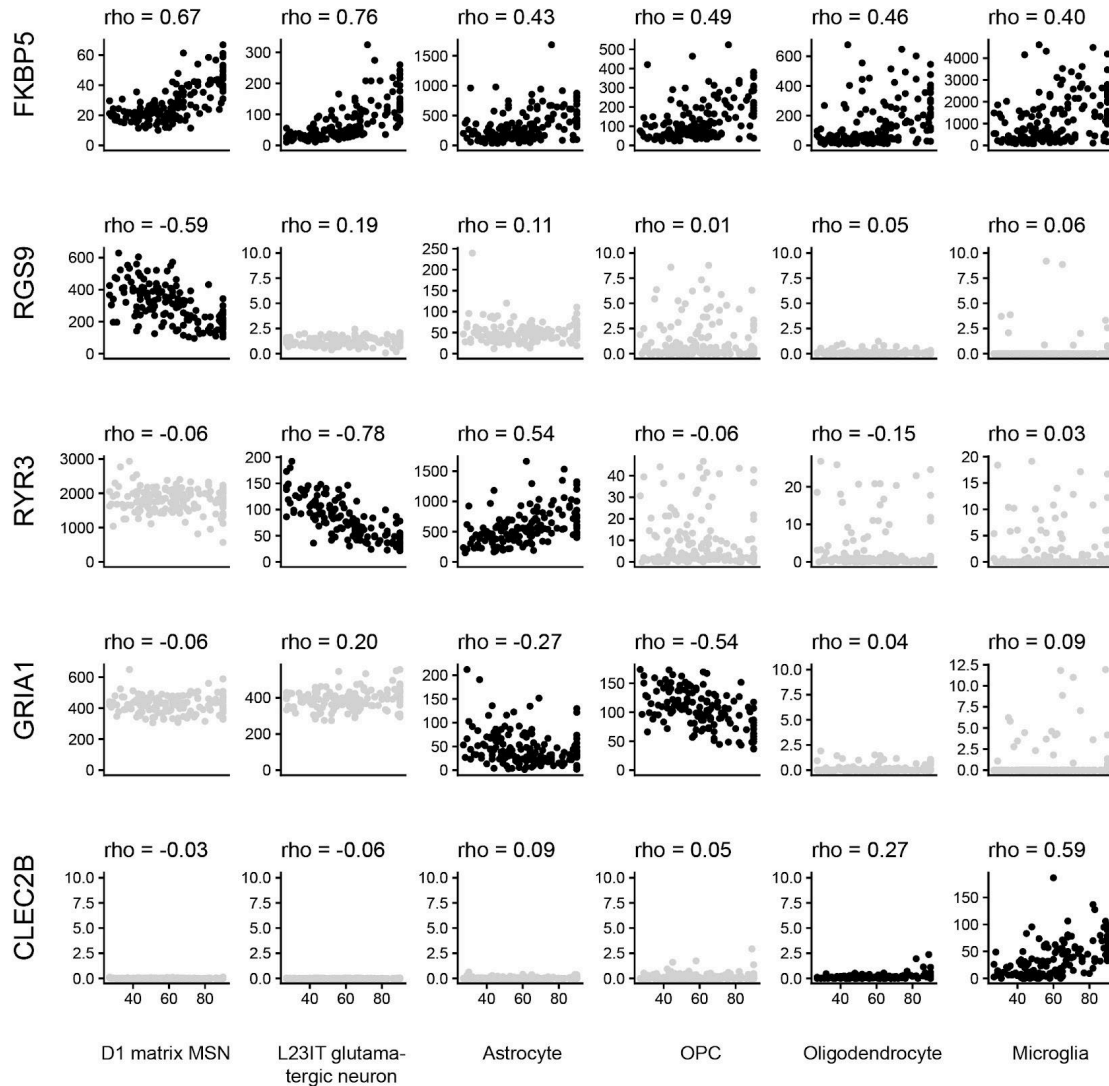


Figure S12. Age-associated effects on expression of specific genes.

Scatterplots show expression of specific genes in relationship to donor age (x-axis of each plot); expression levels have been normalized (transcripts per million; y-axis of each plot). Spearman's ρ , a non-parametric statistic assessing correlation between donor age and expression level (for the selected gene, cell type) is shown above each plot. For non-significant correlations (nominal p-value > 0.05) points are shown in grey.

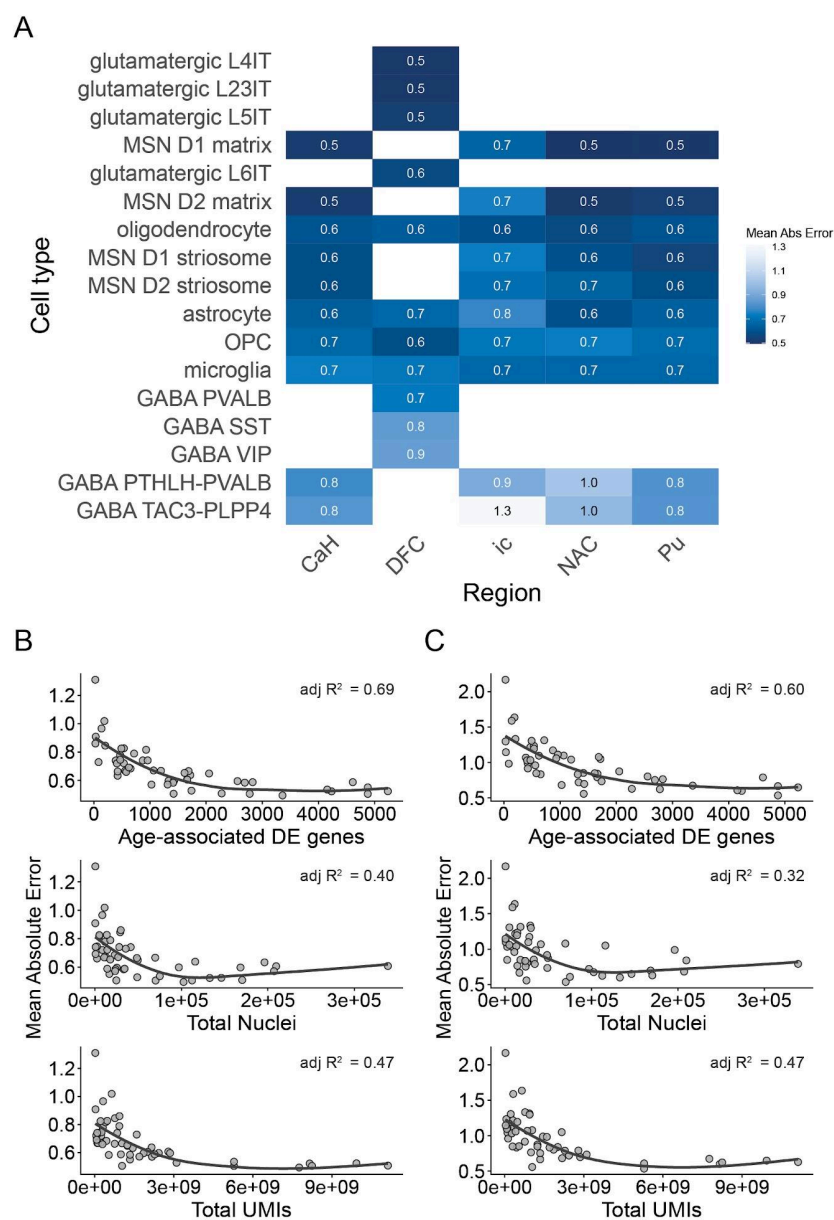


Figure S13. Mean absolute error for age prediction models.

A. Heatmap showing mean absolute error for each cell type–region age-prediction model. Cell-type/region combinations for which age-prediction models were built are shown in shade of blue, with the shade reporting the model error. **B.** Mean absolute error plotted against the number of age-associated differentially expressed genes, total nuclei, and total UMIs per model. Adjusted R^2 values are calculated separately for each predictor. A multivariable model including all three predictors explains variation in mean absolute error across models (adjusted $R^2 \approx 0.72$), indicating that prediction accuracy reflects both power to detect age-associated genes and cell-type-specific differences in the magnitude of age-related transcriptional effects. **C.** The analysis in B was repeated using only the youngest 20% of donors. Mean absolute error is higher than in the full cohort, reflecting regression-to-the-mean effects, but shows similar dependence on the predictors.

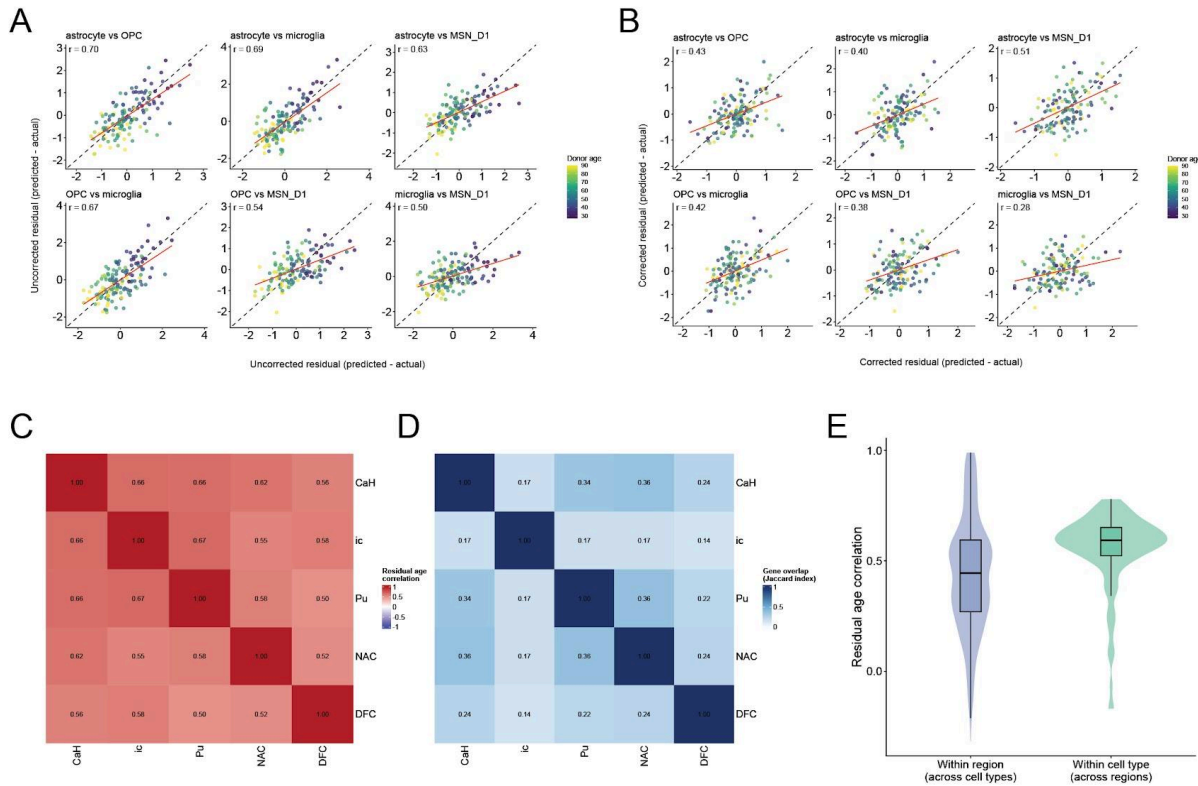
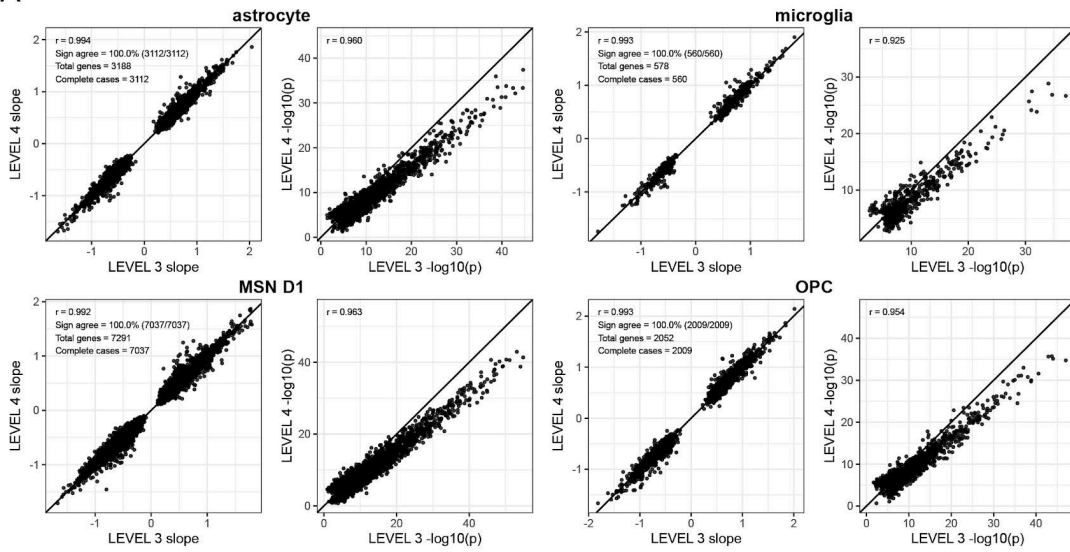


Figure S14. Correction and correlation of age prediction residuals.

A. Pairwise comparisons of uncorrected age residuals (predicted minus chronological age) between representative caudate cell types. Residuals were found to be positively correlated across cell types. Regression-to-the-mean effects can introduce age-dependent structure in these residuals, thereby inflating correlation across cell types. **B.** The same pairwise comparisons as **D** after GAM-based correction of predicted versus chronological age. Correlations are reduced relative to the uncorrected residuals, indicating removal of shared age-dependent bias. A non-zero correlation signal remains, consistent with shared inter-individual variation in biological aging beyond regression-to-the-mean effects. **C.** Positive correlation of cell-type-specific RNA-expression "clocks", beyond the shared effects of chronological age. The heatmap shows pairwise correlations (for each pair of caudate cell types) of the GAM-corrected age residuals. **D.** Minimal overlap of the genes used to predict age in most cell types. The heatmap shows Jaccard indices quantifying overlap among age-associated genes used in each cell-type-specific model. **E.** Pairwise correlations of corrected residual age were computed across cell types within each region and across regions within each cell type. On average, correlations were higher across regions for a given cell type than across cell types within a region, indicating that biological aging kinetics are more strongly shared within cell type (across brain regions) than within region (across cell types).

A



B

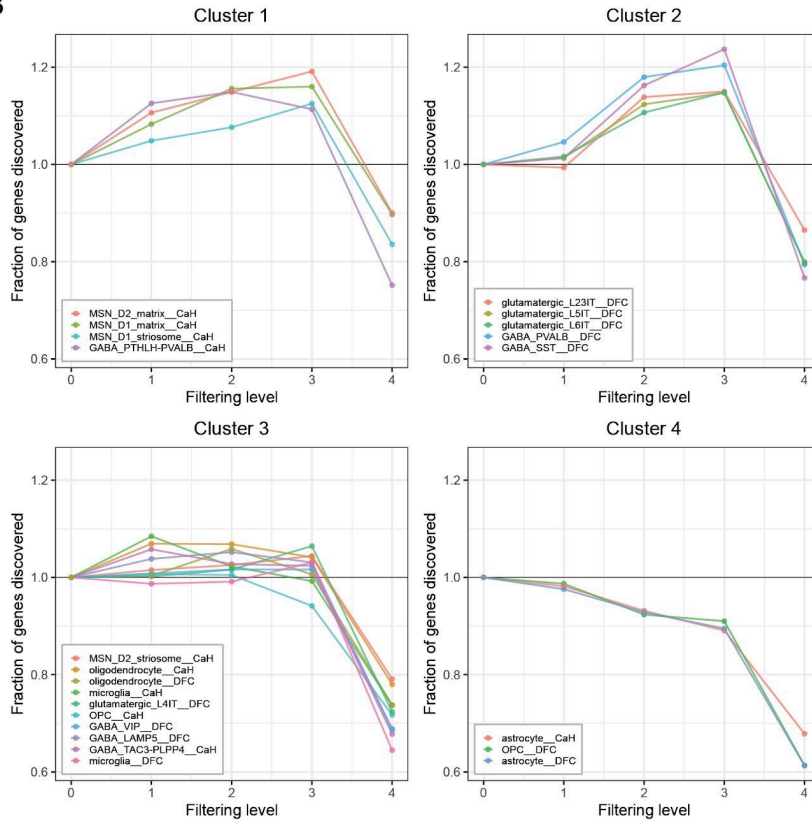


Figure S15. Effects of donor exclusion criteria on recognition and inference of eQTLs. A. Comparison of log fold-change estimates and statistical significance ($-\log_{10}$ empiric p-values) for expression quantitative trait loci (eQTL) discovery. As in the earlier analysis of age-associated effects (**Figure S10**), Levels 0–3 represent progressively stricter snRNA-seq-data-driven exclusions; Level 4 additionally excludes donors based on metadata criteria (findings in medical histories). Effect-size estimates remained highly concordant across filtering levels, and statistical significance is largely preserved or improved under snRNA-seq-data-driven filtering. In contrast, adding metadata-driven exclusions reduced statistical significance while effect-size estimates were largely unchanged. Annotated are the correlation of effect sizes, fraction of the time the direction of the eQTL effect sizes agree (sign test), total number of genes compared (union of eGenes), and total number of genes discovered in both data sets (intersect of eGenes.) **B.** Fractions of eGenes detected at each filtering level, normalized to the unfiltered dataset (Level 0), shown across representative cell types. Cell types were clustered using k-means clustering ($k=4$) based on the fraction of genes discovered at each filtering level to identify shared patterns. snRNA-seq-data-driven filtering (Levels 1–3) preserved or increased the number of discoveries. Additional metadata-driven exclusions (Level 4) reduced the number of detected genes across cell types, indicating a loss of statistical power without corresponding changes in estimated effect sizes.

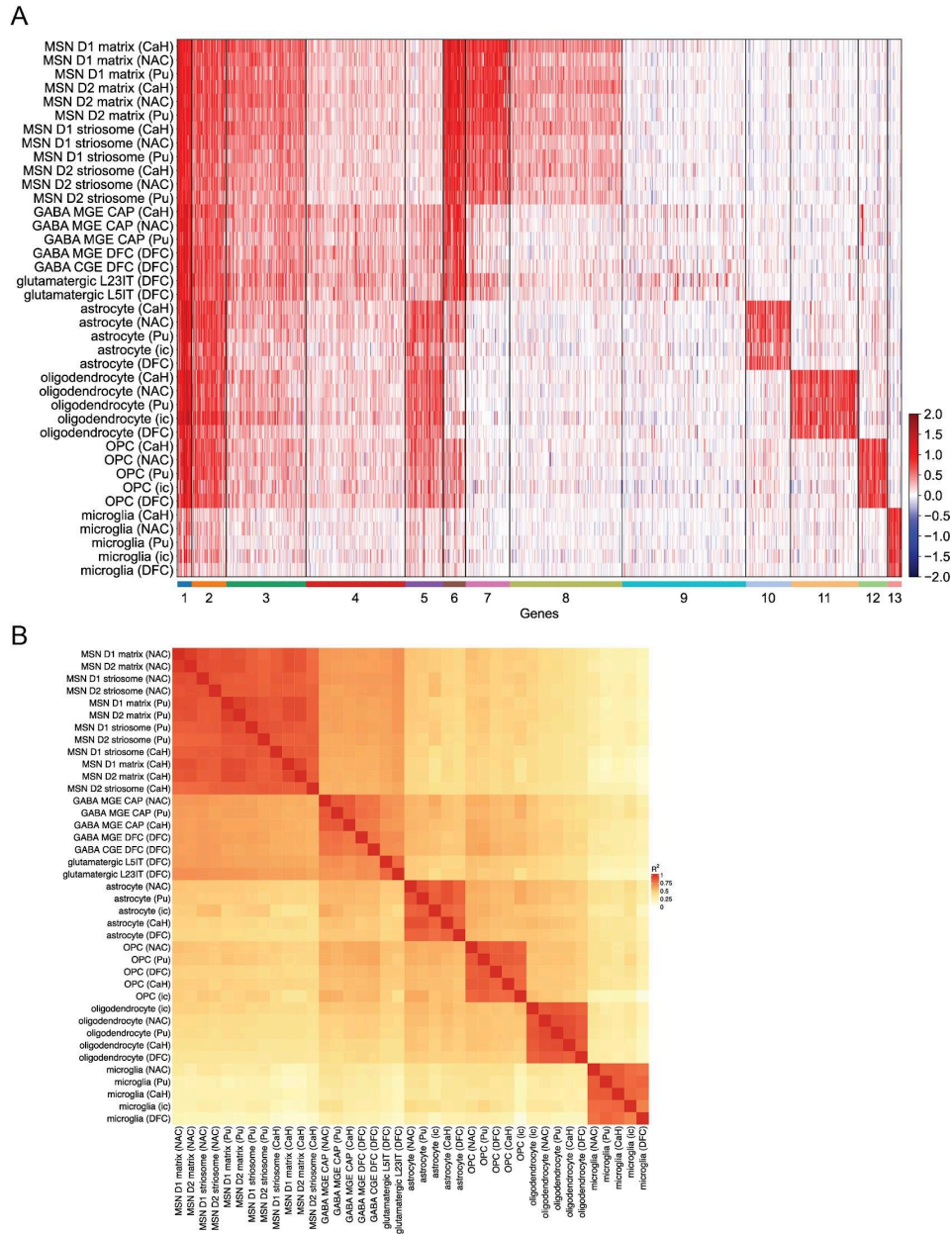
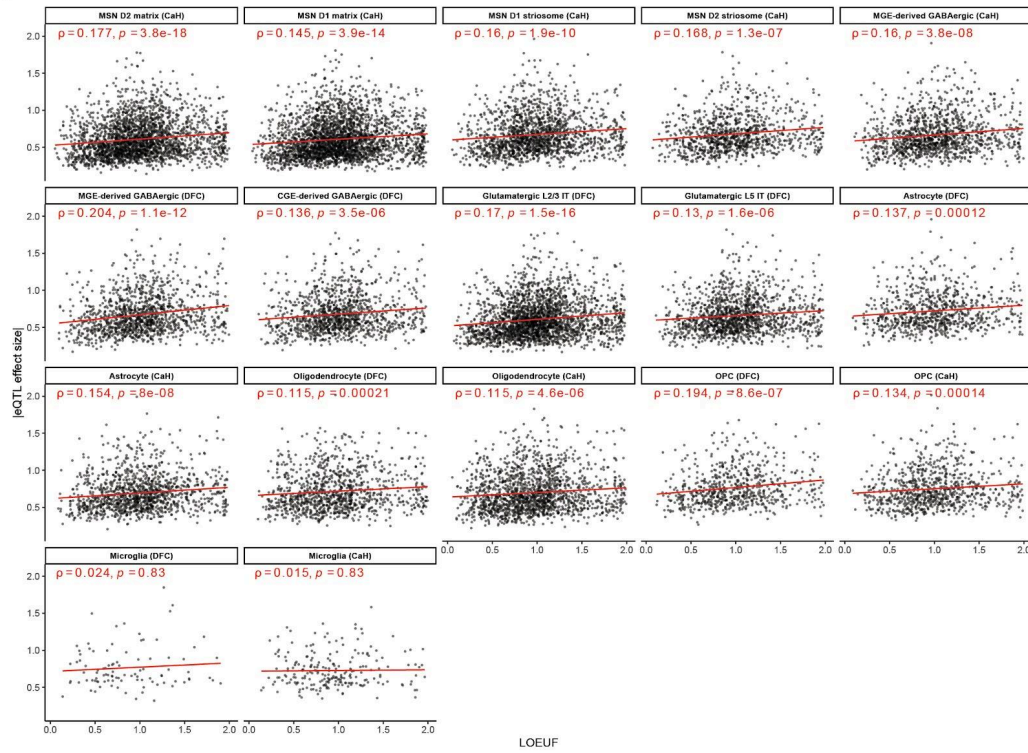


Figure S16. Cell-type specificity of expression QTLs. Extension of Figure 6 to additional brain regions.

A. K-means clustering ($K=13$) of the same 9,899 expression QTLs as shown in Fig. 6B (reaching significance in either CaH, DFC, or both regions), using effect sizes estimated across five brain regions (CaH, Pu, NAC, ic, DFC). Rows represent separate region-specific analyses of eQTLs in each cell type, and columns represent these eQTLs (gene-SNP pairs) grouped by cluster. Shades of red and blue indicate standardized effect sizes (inverse-normalized \log_2 -transformed expression), with red indicating the predominant direction of effect and blue indicating opposite-direction effects. **B.** Pairwise correlations (among cell types sampled in each brain region) of genome-wide sets eQTL effects (corresponding to the rows in panel A). For each cell-type pair, Spearman correlation (R^2) was calculated using eQTLs significant in at least one of the two cell types. Rows and columns were hierarchically clustered.

A



B

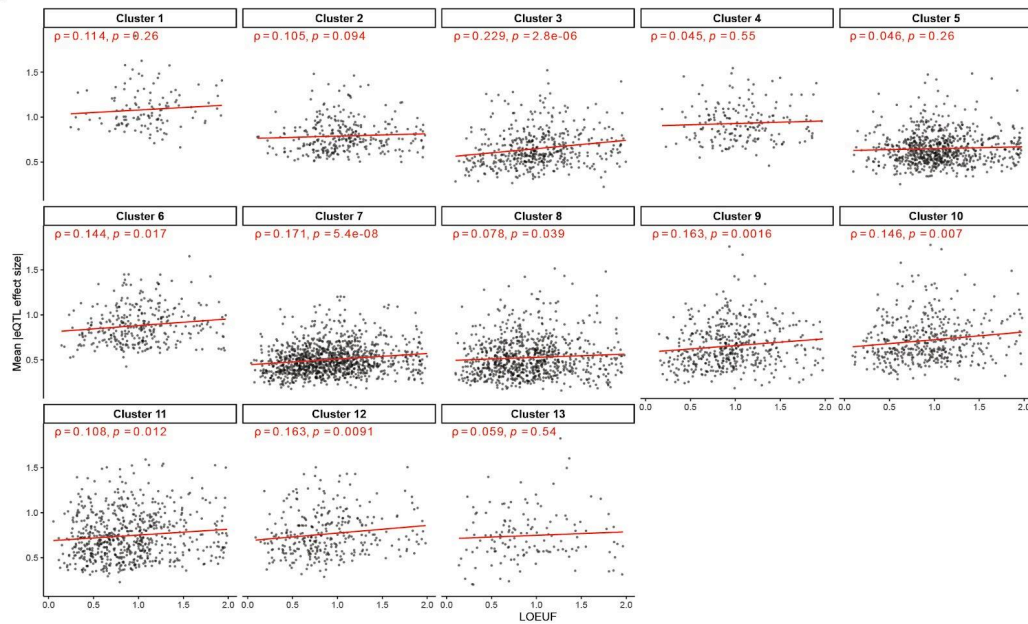


Figure S17: Relationship of eQTL effect sizes and gene function constraint.

eQTL effect sizes are measured as the change in gene expression per allele inherited; functional constraint is quantified by LOEUF scores (Loss-of-function Observed/Expected Upper bound Fraction). Genes' LOEUF scores have been measured in earlier work⁶⁸ and are based on the frequency (relative to chance expectation) with which genes are found to have loss-of-function mutations in human populations. Genes with low LOEUF scores have very few such mutations and are thus inferred to be under strong functional constraint (such as haploinsufficiency). **A.** Scatter plots showing the relationship of absolute eQTL effect sizes (y-axis) to LOEUF scores (x-axis) for each of 17 combinations of cell type and brain region for which eQTLs were recognized in independent analyses. **B.** Mean absolute eQTL effect sizes across statistically significant cell types for each K-means eQTL cluster (K=13). Analyses included all protein-coding eGenes ($q < 0.01$) with LOEUF scores from gnomAD v4.1. Higher LOEUF indicates less constraint (greater tolerance to loss-of-function). Red lines show linear regression fits; Spearman ρ and FDR-adjusted p-values are shown per cluster. Effect sizes were measured from tensorQTL slopes (relationship of inverse-normal-transformed expression level to genotype).

Supplemental tables

Table S1: Donor stratification and analysis inclusion by diagnostic "exclusion" category, age, and sex. Donor counts by clinical/neuropathological diagnostic "exclusion" categories, age groups (<50 years, >50 years), and sex. Diagnostic "exclusion" categories include overdose/substance dependence, psychiatric conditions, neurodegeneration, and cases with missing clinical brain diagnosis (CBD) or neuropathology (NP). The table also indicates the inclusion of each donor group in downstream analyses.

Table S2: Donor demographics. Comprehensive donor-level metadata, including demographic information, clinical and neuropathological diagnoses, tissue source, and cohort inclusion criteria. This table also indicates donor inclusion in downstream analyses.

Table S3: Sample covariates. Description of demographic, genetic, technical, and sequencing-quality variables.

Table S4: Beta-binomial regression results for age and sex effects on cell-type abundances. Effect sizes, standard errors, test statistics, and p-values for age and sex effects on cell-type abundances estimated using beta-binomial regression. Abundances are defined relative to the indicated denominator (e.g., all nuclei). Age is modeled per decade, and sex is coded as male versus female. Effect sizes are reported on the log-odds scale, with positive values indicating increased relative abundance and negative values indicating decreased relative abundance with the covariate. Nominal p-values and p-values adjusted for multiple comparisons using the Benjamini–Hochberg procedure are reported.

Table S5: Significant age-associated differential expression Gene Set Enrichment Analysis (GSEA) results by cell type. Counts of unique eGenes identified per cell type, defined as the union of eGenes detected in dorsolateral prefrontal cortex (DFC) and caudate head (CaH).

Table S6: Union of eGenes across DFC and CaH by cell type. Counts of unique eGenes identified per cell type, defined as the union of eGenes detected in dorsolateral prefrontal cortex (DFC) and caudate head (CaH).

Table S7: Fraction of cell-type-specific eGenes with maximum expression in corresponding cell type(s). For each eGene cluster (k-means; see **Figure 6**), the expected cell type(s) (by region), the number of eGenes with nonzero median expression across cell types, the number of eGenes whose highest median expression occurs in one of the corresponding cell types, and the proportion of eGenes whose maximum median expression falls in one of the expected cell type(s) (see **Methods**).

Methods

Donor selection and tissue procurement

Postmortem human brain tissue was obtained through the Human Brain Cell Atlas Collection (HBCAC), a dedicated program within the NIH NeuroBioBank (NBB; RRID:SCR_003131) supporting the NIH BRAIN Initiative Cell Atlas Network (BICAN). Brain and Tissue Repositories (BTRs) contributing to the NBB HBCAC followed harmonized protocols for donor evaluation, brain collection, and tissue processing and preparation such as photodocumentation, slab thickness, and uniform sectioning. Postmortem human brain tissue was also obtained from the

main NBB collection (“archival donors”). Archival donors appropriate for our study were selected by NBB, in accordance with HBCAC criteria. Postmortem tissue collection followed the provisions of the United States Uniform Anatomical Gift Act of 2006 described in the California Health and Safety Code section 7150 and other applicable state and federal laws and regulations. Informed consent for unrestricted research use and open data sharing was obtained by NBB from the legal next-of-kin. The Broad Institute Office of Research Subject Protection reviewed the use of postmortem brain tissue for research purposes and determined that the use of de-identified specimens from deceased individuals did not constitute human subjects research requiring IRB review (NHSR-8066) per federal regulation 45 CFR 46 and associated guidance.

Donor information is available in **Table S2**. Adult donors ages 27 to 89+ that were consented for open data sharing were considered for this study. Exact donor age at time of death for donors over 89 years old is not available as a subject’s age, in combination with other health information, could potentially be a unique identifier for individuals aged 89 and older. These subjects are listed as aged 89+ years, following HIPAA Privacy Rules. Donors with positive serology results for HIV-1/2, Hepatitis B, or Hepatitis C were not considered for this study.

We sampled from donors flagged by NBB as potentially meeting the HBCAC criteria at the time of donor intake at the brain bank. All donors were evaluated for major or persistent neurological or psychiatric conditions, such as Alzheimer’s disease and schizophrenia. Toxicology testing was conducted at each contributing brain bank to determine substance presence at the time of death and to interpret substance exposure alongside clinical brain diagnoses. All available donor metadata at the time of brain bank intake were comprehensively reviewed and used to guide donor selection and initial tissue sampling.

Donors that were later found to have major or persistent neurodegenerative or psychiatric conditions, or substance dependence (including alcohol, opioid, or cocaine) were not designated as healthy control donors (see **Methods**: Overview of Outlier Filtering, **Table S1, S2**). Additional exclusionary criteria for healthy control designation included clinical brain diagnoses of: recurrent major depressive disorder, and other psychiatric disorders (including post-traumatic stress disorder (PTSD), obsessive-compulsive disorder (OCD), and hoarding disorder). Exclusion criteria based on neuropathological evaluation included: Cerebral infarction in donors under 65 years of age⁷¹, Primary age-related tauopathy (PART) in donors under 65 years^{72,73}, Limbic-predominant age-related TDP-43 encephalopathy (LATE) in donors under 80 years⁷⁴. Age thresholds were informed by current literature characterizing the age-related onset of these pathologies. Additional exclusionary metadata was based on causes of death, specifically: cause of death of combined overdose (OD) that includes substances of ≥ 2 drug classes and cause of death due to chronic substance use/abuse/dependence. Minor neurological or psychiatric conditions deemed to have minimal functional impact (e.g., migraine, specific phobias) were permitted. Donors with historical diagnoses in clinical remission were eligible, provided there was no confirmatory evidence of the disorder at the time of death per clinical records. Findings associated with normal aging or incidental focal pathology (e.g., localized ischemic events) were not considered exclusionary.

To facilitate standardized tissue selection and annotation, brain slab images from HBCAC donors were uploaded to the custom-built Neuroanatomy-anchored Information Management Platform (NIMP; RRID:SCR_024684) developed by UTHealth Houston⁷⁵. Brain slab image collections, prepared under HBCAC protocols, consisted of an average of about 30 coronally sectioned slabs, each approximately 5 mm thick. Slabs were photographed with identification numbers and scale bars, enabling tissue requesters to estimate dimensions accurately. Neuroanatomists reviewed these slab collections and identified slabs that contained the relevant regions of interest (ROI). ROIs were drawn on the slab images to indicate the precise dissection sites and annotation pins were placed to link to neuroanatomical structures as defined in the DHBA ontology. To maintain experimental uniformity across donors and brain regions, ROIs were matched as closely as possible in anatomical position, minimizing variability caused by normal morphological differences between individuals.

Archival donors included in this study were selected by NBB staff based on the HBCAC criteria. Tissue was requested through the NBB request process and subdissections were performed by NBB in accordance with NBB's standard practice⁷⁶. Tissue preparation varied by brain bank, with tissue arriving as small blocks, chips, or pulverized (tissue samples that were pulverized are designated as such in **Table S2**). DFC, caudate, putamen, and nucleus accumbens were sampled from the archival donors. Internal capsule tissue was not sampled for the archival donors.

RNA quality scoring

Genomic DNA and total RNA were extracted from frozen cerebellar tissue from each individual donor and processed at Broad Clinical Laboratories. RNA quality was assessed using the LabChip GX system (PerkinElmer). RNA integrity was evaluated by microfluidic capillary electrophoresis, generating RNA Quality Scores (RQS) on a scale of 1–10, with higher values indicating greater integrity. DV200 values (proportion of RNA fragments >200 nucleotides; reported on a 0–1 scale) were calculated as an additional measure of RNA quality predictive of RNA-sequencing performance. RQS and DV200 served as quantitative quality control measures.

Tissue handling and dissections

The anatomical location of each ROI was confirmed by aligning slab images to the Ding atlas³⁴ coordinates, with “pinned” reference points marking the anterior–posterior, dorsal - ventral, and medial - lateral positions. Care was taken to match ROIs across donors, ensuring uniform anatomical coverage despite individual morphological variation. DFC sampling targeted the cortical layers anterior to the premotor cortex and dorsal to the inferior frontal sulcus (**Figure S1A**). Striatal dissections were guided by internal landmarks to isolate the entirety of caudate nucleus, putamen, nucleus accumbens core and shell (treated as a single region for downstream analysis for consistency with archival donor sampling), and segments of the internal capsule without contamination from neighboring territories (**Figure S1B**). Dissection was conducted in a cryostat (-25 °C) using pre-chilled razor blades. Slabs were kept on dry ice until processing and immediately returned to -80 °C storage post-sampling. All cuts were made

with care to avoid inclusion of unintended subregions and any adjacent white matter, particularly for small striatal nuclei. Subdissections were conducted by NBB staff for all donors designated as “archival”.

Whole genome sequencing

Whole-genome sequencing (WGS) was performed on genomic DNA from each donor. PCR-free libraries were prepared using the KAPA HyperPrep Library Construction Kit (Roche) with custom Broad indices. Libraries were sequenced on the NovaSeq X platform (Illumina; RRID:SCR_024569) to generate 150 bp paired-end reads to a target mean coverage of 30x. Reads were demultiplexed, aggregated, and aligned to the hg38 reference genome using Illumina DRAGEN processing. Data delivery included aggregated CRAM files, corresponding CRAI index files, and md5 checksum files.

Single-nucleus RNA-seq library preparation and sequencing

To enable direct comparative analysis of nuclei from multiple brain donors while minimizing technical variance, we processed frozen brain tissue specimens from multiple donors as a single pooled sample. Each village comprised nuclei isolated from an average of 19 donors (range: 9-26 donors). For each donor, equivalent amounts of tissue (30 ± 5 mg for most regions, ~ 15 mg for smaller anatomical regions) were collected. All pooled samples underwent nuclei isolation, droplet-based encapsulation, library preparation, and sequencing as a single batch. Between 2-8 encapsulation reactions were performed per village, depending on the desired sequencing depth. We used combinations of hundreds of transcribed single nucleotide polymorphisms (SNPs) in each cell's sequenced reads to assign each nucleus to its donor-of-origin, using the computational approach described below.

Nuclei were isolated from frozen brain tissue using a modified density gradient protocol⁷⁷. Briefly, tissue blocks were cryosectioned/microdissected at -25°C and transferred directly to pre-chilled Dounce homogenizers containing Nuclei EZ lysis buffer (MilliporeSigma, #NUC101) supplemented with 1 U/uL RiboLock RNase Inhibitor (Thermo #EO0382). Mechanical homogenization was performed with 10-20 strokes of each pestle size, followed by incubation on ice for 10 min. The lysate was passed through a 20 μm vacuum cell strainer (SCNY00020), centrifuged at $500 \times g$ for 5 min at 4°C , and the pellet resuspended in G30 solution (30% iodixanol (STEMCELL #07820), 3.4% sucrose, 20 mM Tricine, 25 mM KCl, 5 mM MgCl_2 , pH 7.8). The suspension was layered over a fresh G30 cushion in 1.5 mL volumes and centrifuged at $8000 \times g$ for 10 min at 4°C . Supernatants were discarded, and resulting nuclei pellets were resuspended in PBSAi buffer (composition as optimized in our protocol) and pooled into a single tube per village. Nuclei suspensions were brought to 1.5 mL total volume with PBSAi, centrifuged again ($500 \times g$, 5 min, 4°C), and the pellets resuspended in $\sim 100 \mu\text{L}$ of PBSAi. Nuclei were counted using a LUNA-FL Dual Fluorescence Cell Counter (Logos Biosystems, #L12005) with acridine orange/propidium iodide staining (Logos Biosystems #F23001) to assess concentration and viability.

Isolated nuclei were loaded onto a 10x Genomics Chromium instrument (RRID:SCR_024939) and processed using either Chromium NextGEM Single Cell 3' Reagent Kits V3.1 (CG00204, Rev D) or Chromium GEM-X Single Cell 3' Reagent Kits v4 (CG000731, Rev A) according to

the manufacturer's protocol, with minor modifications for nuclei input (19-22K loading for NextGEM, 29-32K loading for GEM-X). The majority of libraries were generated from small-scale (8-16 reactions) manual preparations using magnetic racks, while a subset of experiments were processed using automated high-throughput liquid-handling systems. Sequencing was performed on the Illumina NovaSeqX platform.

Sequence alignment and donor assignment

Sequence data was demultiplexed and aligned following the standard Drop-Seq protocol⁷⁸ and was aligned to the GRCh38 reference and GENCODE (RRID:SCR_014966) v43 gene models. The expression matrix was computed using Drop-Seq DigitalExpression program, and used two non-standard flags `READ_MQ=0` and `FUNCTIONAL_STRATEGY=STARSOLO` to replicate STARSolo expression quantification. Ambient/ background RNA was removed from digital gene expression (DGE) matrices using CellBender (v.0.3.2) `remove-background`³⁶. Nuclei selection was performed via DropSift (see below), then final DGE matrices were generated for those subsets of cell barcodes. These matrices were then used in combination with MapMyCells³⁹ to generate cell type labels. Donor assignment and doublet detection were performed by Dropulation³⁵.

Nucleus selection

Accurate nucleus selection is a critical step in single-nucleus RNA sequencing (snRNA-seq) analysis. Proper identification of nuclei ensures robust downstream expression quantification, cell type classification, and biological interpretation. However, distinguishing nuclei from empty droplets presents unique challenges, particularly in brain tissue, where high cellular diversity, ambient RNA contamination, and debris can complicate classification. To address these challenges, we developed DropSift, a supervised method that distinguishes nuclei from empty droplets in an experiment-specific manner.

Nucleus-containing droplets were distinguished from empty droplets using DropSift⁷⁹. DropSift is a supervised classifier that labels barcodes as nuclei or empty droplets using a support vector machine (SVM) trained separately for each experiment. The model uses cell-level summary metrics (including total UMI counts and the fraction of intronic UMIs), expression-based features derived from genes differentially expressed between nuclei and empty droplets, and, when available, CellBender-derived estimates of ambient RNA contamination as input features. For each dataset, DropSift first identifies exemplar nuclei and empty droplets using density-based thresholds in the joint distribution of total UMI counts and intronic fraction, with additional guidance from the CellBender contamination estimates. These exemplars are used to train an SVM with a radial basis function kernel to discriminate nuclei from empty droplets. The trained classifier is then applied to all barcodes above a minimal UMI threshold; very low-UMI barcodes (<20 transcripts) which cannot be reliably classified are excluded from downstream analysis.

By integrating multiple signals, DropSift effectively differentiates nuclei from empty droplets while adapting to variability in data structure, including differences in ambient RNA contamination, across experiments and species. Rather than relying on tissue- or organism-specific marker sets, it learns expression patterns that are specific to each dataset, which improves classification accuracy in heterogeneous brain tissue.

DropSift is implemented in R and released as an open-source package. The source code, documentation, and installation instructions are available from the project GitHub repository (<https://github.com/broadinstitute/DropSift>)

Unsupervised clustering

Following nuclei selection to establish a set of non-empty, non-debris libraries, we performed unsupervised clustering to identify cell-type structure in the dataset. All clustering analyses were performed in R using Seurat⁸⁰ (v4.4.3; RRID:SCR_007322), with analysis-specific parameters (e.g., number of highly variable genes, number of principal components, and clustering resolution) specified below.

Raw count matrices were preprocessed using Seurat's standard single-cell workflow. Gene expression values were log-normalized using `NormalizeData`. Highly variable genes (HVGs) were selected using `FindVariableFeatures`, modeling the gene mean-variance relationship, and HVG expression values were linearly scaled using `ScaleData` such that each gene had zero mean and unit variance across cells. Mitochondrial (MT-) genes were excluded from clustering analyses; no additional manual gene filtering or regression was applied. Principal component analysis was performed using `RunPCA` on the selected HVGs. Cell-cell neighbor graphs were constructed using `FindNeighbors`, followed by graph-based clustering using `FindClusters` (Leiden algorithm). Low-dimensional embeddings were computed using UMAP via `RunUMAP`. To encourage co-clustering of nuclei across donors, and in some analyses across brain regions, Harmony-based integration was applied using Seurat's Harmony wrapper (implementation specified per analysis).

Due to the large scale of the dataset (~5 million nuclei when including all donors, samples, and doublets), analyses were performed using `BPCells`⁸¹, an R package that enables memory-efficient storage and streaming of large single-cell expression matrices. `BPCells` was used in conjunction with Seurat to support normalization, dimensionality reduction, clustering, and visualization at scale.

Expression-based doublet removal

Donor multiplexing (experimental) with genotypic demultiplexing (computational) enabled identification of genotypic doublets, defined as nuclei containing genetic material from more than one donor. Experimental batches ("villages") typically included approximately 20 donors; under even donor representation, the expected proportion of heterotypic genotypic doublets is ~0.95.

To identify additional doublets based on transcriptional profiles, genotypic doublets were intentionally retained during initial clustering to serve as a reference set. Global clustering was first performed using 5,000 HVGs and 60 principal components, followed by iterative sub-clustering at higher resolution. Initial partitions separated major cell classes, including striatal neurons, cortical glutamatergic neurons, striatal GABAergic interneurons, cortical GABAergic interneurons, astrocytes / OPCs / microglia, oligodendrocytes, vascular cells,

ependymal cells, and lymphoid lineage immune cells. These groups were subsequently sub-clustered to generate terminal (“leaf”) clusters corresponding to finer cell-type structure.

For each leaf cluster, the proportion of nuclei labeled as genotypic doublets was computed. Clusters in which more than 85% of nuclei were genotypic doublets were designated as doublet-enriched clusters and removed in their entirety. All other clusters were retained, resulting in removal of both genotypic doublets and transcriptionally defined doublets not identified by genotypic criteria. No additional expression-based doublet detection algorithms were applied.

This procedure yielded a curated dataset with doublet annotations and an initial set of unsupervised cell-type clusters used in downstream analyses.

Label transfer from external taxonomies

To support initial annotation and evaluation of unsupervised clustering results, cell-type labels were transferred from external reference taxonomies. For striatal populations, the consensus basal ganglia taxonomy³⁸ was used as the primary reference. Label transfer was performed using the MapMyCells computational framework.

For each nucleus, best-match labels were computed at multiple hierarchical levels (neighborhood, class, subclass, group, and cluster), along with bootstrap-derived confidence probabilities. Rather than relying on nucleus-level assignments, label transfer results were summarized at the cluster level and used to assign initial putative labels to unsupervised clusters.

In some samples, particularly from nucleus accumbens dissections, sub-clusters were identified that corresponded to neuronal populations outside the target regions of analysis (e.g., globus pallidus GABAergic neurons). These populations formed distinct sub-clusters and were recorded as off-target dissection artifacts in sample-level metadata; downstream analyses focused on neuronal populations within the defined target regions (CaH, Pu, NAC, and ic).

Because the dataset also included cortical nuclei (DFC) as a comparator group, some non-basal-ganglia populations were not represented in the basal ganglia taxonomy. Although MapMyCells assigned labels to these populations, these assignments were not used in downstream analyses. For cortical neuronal populations, additional label transfer was performed using structured neuronal taxonomies from the Allen Brain Atlas (RRID:SCR_017001). Annotated snRNA-seq data from human postmortem dorsolateral frontal cortex (DFC)²⁵ were used to define exemplar populations. scPred⁸² models were trained on these reference datasets and applied to DFC-derived neuronal nuclei. Labels were assigned based on maximum predicted probability, with a minimum probability threshold of 0.8.

Sub-clustering of neuronal and glial populations

After removal of genotypic and expression-defined doublets and exclusion of donor- and sample-level CTP and GEX outliers (see **Methods**: Outlier identification and filtering), cell-type-specific sub-clustering was performed to achieve higher-resolution annotations.

The initial global clustering was sufficient to resolve certain populations (e.g., ependymal cells, B cells, and T cells), which were not further subdivided. Additional sub-clustering was performed for neuronal and major glial populations as described below. Input nuclei for sub-clustering analyses were filtered to libraries with a \log_{10} UMI count equal to or greater than two median absolute deviations (MADs) below the median \log_{10} UMI count for the cell type being clustered, calculated for each donor “village” to account for experimental batch effects influencing UMI counts.

Glial populations analyzed included astrocytes, oligodendrocyte precursor cells (OPCs), committed oligodendrocyte precursors (COPs), oligodendrocytes, microglia, and vascular cells. Vascular cells were sparsely sampled across donors and were excluded from downstream cell-type proportion and gene expression analyses.

Microglia were sub-clustered using 3,000 HVGs, 30 principal components, and Harmony integration by donor and brain region. A small sub-cluster of non-microglial myeloid cells annotated as border-associated macrophages (BAMs) by label transfer from the basal ganglia taxonomy was excluded from microglia analyses.

Oligodendrocytes were sub-clustered using 3,000 HVGs, 30 principal components, and Harmony integration by donor and brain region. Labels from the consensus basal ganglia taxonomy (Oligo OPALIN and Oligo PLEKHG1) were used to support annotation.

Striatal neuron sub-clustering

Striatal neuronal sub-clustering was performed to confirm representation of expected neuronal classes and to achieve high-resolution annotation. Canonical medium spiny neurons (MSNs), eccentric (“hybrid”) MSNs, and striatal interneurons were jointly sub-clustered using 5,000 HVGs, 30 principal components, and Harmony integration by donor.

Clustering was performed at resolution 0.25. Major striatal interneuron classes defined in the consensus basal ganglia taxonomy (PTHLH–PVALB GABA, TAC3–PLPP4 GABA, SST–CHODL GABA, and cholinergic GABA neurons) formed distinct clusters, supported by marker gene expression and label transfer.

Canonical MSNs separated into D1 and D2 populations. Within these populations, cells were further annotated along matrix-striosome and dorsal-ventral axes. For downstream analyses, matrix and striosome MSNs were defined by the intersection of unsupervised cluster membership (restricted to D1 or D2 clusters) and confident label transfer assignment (probability >0.8).

Astrocyte sub-clustering

Astrocytes were sub-clustered across brain regions using 3,000 HVGs, 30 principal components, and Harmony (RRID:SCR_022206) integration by donor and brain region. Clustering at resolution 0.25 yielded 11 astrocyte sub-clusters comprising ~315k nuclei. Two major clusters corresponding to striatal and cortical astrocytes collapsed at lower clustering resolutions and were combined for exploratory analyses (**Figure S18**).

As higher-resolution astrocyte annotations were not available in the consensus basal ganglia taxonomy, marker gene identification was performed using comparison of cluster pseudobulked expression profiles. For each sub-cluster, the top 20-50 genes with the largest fold changes relative to cluster 0 were selected as candidate markers. These gene sets were compared with published markers of astrocyte sub-types and states in the adult human brain^{83,84} (**Figure S18B**).

A single cluster (cluster 0) accounted for the majority of astrocytes (63%) and expressed canonical homeostatic astrocyte markers, including *SLC1A2*, *KCNJ10*, *FGFR3*, and *GRM3*, consistent with a population of homeostatic “protoplasmic” astrocytes. Although this population was present across all sampled regions, there was a pronounced enrichment in gray matter regions relative to ic (**Figure S18C,D**). Cluster 1 (14%) was characterized by elevated expression of *GFAP*, *VCAN*, *TNC*, *CD44*, and extracellular matrix-associated genes, and was strongly enriched in the ic (**Figure S18C,D**), consistent with white-matter-associated “fibrous” astrocytes.

Among lower-abundance clusters, we observed evidence of regionally enriched astrocyte populations (**Figure S18B-D**). Cluster 6 (2%) showed near-exclusive enrichment in DFC samples and expressed higher levels of *DPP6*, *DPP10*, *SEMA5A*, and *DSCAML1*. This population may correspond with “interlaminar” astrocytes localized to the cortex. Cluster 3 (4%) expressed *CHRD1*, *DIO2*, and *CALCRL* and was significantly enriched in putamen samples, potentially corresponding to a localized astrocyte population within the striatum, or astrocytes originating from an anatomically adjacent compartment resulting from off-target dissections. Cluster 5 (3%) was marked by expression of *SLC6A11* together with GABA-signaling-related genes (*GABBR2*, *RGS6*, *ADCY1*) and was enriched in NAC and ic samples. This population may reflect a recently reported *SLC6A11+* astrocyte population localized to the globus pallidus, captured in more-posterior samples targeting the NAC.

Cluster 9 (1%) was challenging to define. Elevated expression of *AQP4* would be consistent with perivascular astrocytes, but we did not identify a larger blood-brain-barrier associated program.

Two clusters seemed to reflect astrocyte states rather than stable subtypes.

Cluster 2 (6%) expressed inflammatory and cytokine-responsive genes (*C3*, *CHI3L1*, *SERPINA3*, *CCL2*, *CXCL10*, *SOCS3*). We noted a subset of donors with elevated proportions of

astrocytes in cluster 2. One plausible interpretation is that cluster 2 represents astrocytes upregulating a “reactive” cell-state program.

Cluster 7 (2%) was distinguished by robust induction of immediate early genes (*FOS*, *JUN*, *NR4A1/2/3*) and stress-response chaperones (*HSPA1A/B*), potentially indicating an activity- or stress-associated astrocyte state. Perimortem conditions have the potential to impact expression levels of activity / stress programs, so we interpret the presence of this cluster with caution.

Two clusters likely reflect technical artifacts. Cluster 4 (4%) expressed neuronal and synapse-associated transcripts (*RBFox1*, *SYT1*, *RIMS2*), raising the possibility of low-level neuronal RNA contamination or mixed nuclei, although a contribution from astrocytes with strong synaptic coupling cannot be excluded. Cluster 8 (2%) expressed oligodendrocyte-lineage markers (*MBP*, *PLP1*, *MOBP*, *BCAS1*) and likely represents oligodendrocyte or myelin-debris contamination.

OPC sub-clustering

OPCs were sub-clustered using 3,000 HVGs, 30 principal components, and Harmony integration by donor and brain region. Clustering at resolution 0.1 yielded eight sub-clusters, two of which collapsed at lower resolution and were combined, resulting in seven OPC sub-clusters (~155k nuclei; **Figure S9A**).

To confirm that regional integration did not obscure region-specific structure, region-specific clustering was also performed using 3,000 HVGs for CaH, Pu, and NAC and 2,000 HVGs for ic and DFC, with 30 principal components and Harmony integration by donor. The same seven OPC sub-types were recovered in each region-specific analysis. Marker genes were identified using the same pseudobulk-based approach as for astrocytes⁵⁶ (**Figure S9B**).

The resulting clusters exhibited highly skewed abundances (**Figure S9C,D**). One dominant cluster (cluster 0) accounted for approximately 84% of all OPCs, whereas the remaining clusters each comprised between 1 - 7% of the population. Despite these differences in abundance, all clusters were present to some degree across brain regions (in contrast with region-restricted sub-clusters like astrocyte cluster 6), with the possible exception of WM-localized OPCs noted below.

Cluster 0, the most abundant OPC population, expressed canonical OPC-associated genes including *BCAN*, *PCDH15*, and *ITGA8*, and lacked markers of proliferation, stress responses, or inflammatory signaling. The proportion of OPCs in this cluster was lower in the internal capsule relative to striatal GM regions; however, this pattern could be driven by the regional distribution of cluster 1 OPCs.

Cluster 1 (7%) was enriched in the internal capsule. We note several genes upregulated in cluster 1 (*CDH19*, *SEMA3E*, *GRIA4*), and several depleted in cluster 1 relative to cluster 0

(*SHISA6*, *GPC5*, *CNTN5*, *TCF7L1*). In spatial data, the former gene set marked OPCs localized to the ic and the latter marked OPCs within striatal GM regions. In contrast, genes upregulated in cluster 0 marked OPCs in all striatal regions, consistent with an OPC population localizing throughout GM and WM compartments. The ratio of cluster 1 to cluster 0 OPCs was also found to associate strongly with oligodendrocyte fraction in all regions.

Cluster 2 (3%) represented a smaller but distinct OPC population characterized by strong upregulation of *MIR137HG*, *DPYD*, and *LINC01776*, all located within the 1p21.3 genomic locus. This cluster was present across striatal and cortical brain regions, as well as across donors (**Figure S9C,D**).

Cluster 3 (2%) expressed inflammatory and cytokine-responsive genes including *OSMR*, *IL6R*, and *STAT3*. The proportion of cluster 3 OPCs varied markedly across donors and. These findings indicate that cluster 3 could represent a “reactive” OPC state linked to inflammatory signaling.

Cluster 4 (1%) was defined by expression of cell-cycle genes (*MKI67*, *TOP2A*, *CENPF*) indicating an actively mitotic OPC population. This cluster likely represents a transient cell-cycle state superimposed on other OPC identities.

Cluster 6 (1%) was marked by expression of hypoxia- and metabolism-related genes including *VEGFA*, *EPAS1*, and *SLC2A1*. Given its low abundance and transcriptional profile, this cluster may reflect a metabolic or hypoxic stress-associated OPC state, potentially influenced by local vascular environments or perimortem factors.

Finally, cluster 5 (1%) expressed mature oligodendrocyte and myelin genes (*MOG*, *MOBP*, *OPALIN*) and lacked canonical OPC markers, suggesting contamination by oligodendrocyte nuclei or ambient RNA. This cluster was excluded when interpreting OPC heterogeneity.

Visualizing expression of marker genes in Slide-tags data

To visualize the spatial distribution of expression of OPC subcluster marker genes, we utilized Slide-tags spatial transcriptomics data generated and analyzed by Kraft et al., 2026. This data is publicly available in <https://assets.nemoarchive.org/collection/nemo:dat-aqicdbo>. We chose to visualize gene expression in an exemplary sample (s5) that captured all of the striatal regions of interest.

Computing genetic principal components

Genetic principal components (PCs) were computed by co-embedding our samples with high-coverage whole-genome sequencing data from the 1000 Genomes Project⁸⁵ (RRID:SCR_006828). We first restricted the 1000 Genomes VCFs to variants detected in our dataset and filtered to retain only common variants (MAF > 0.05). Sex was inferred using PLINK^{86,87} (RRID:SCR_001757).

Prior to PCA, SNPs were pruned to ensure high-quality, independent markers. Specifically, we removed SNPs with extreme allele frequencies (MAF > 0.995), high missingness (>2%), excess heterozygosity (Hardy–Weinberg^{88,89} equilibrium p-value < 1×10^{-6}), and autosomal SNPs showing association with sex. Long-range linkage disequilibrium (LD) regions were also excluded. PCA was performed using EIGENSOFT^{90,91} smartpca (RRID:SCR_004965) on the 1000 Genomes reference panel; our samples were subsequently projected into this PC space.

Genetic ancestry proportions were estimated by fitting linear models for each of the five superpopulations as a function of the top 10 genetic principal components. Models were trained on individuals from the 1000 Genomes project with known ancestry, and predicted values with 95% confidence intervals were generated for all samples⁹².

Outlier identification and filtering

Throughout this paper, various levels of outlier filtering were used for analysis (see **Table S1**). All UMAP and subtype clustering was performed using all nuclei (no filtering). Cell type proportions (CTP) analyses, including exploratory analyses of cell type abundances (fractions, ratios) and regressions with metadata, were performed on the core donor set, excluding CTP outlier samples and donors. Cell type-specific analyses, including differential expression and downstream TRADE analysis, as well as RNA-based age modeling and expression quantitative trait locus (eQTL) analyses, were performed with CTP outlier samples, CTP outlier donors, gene expression (GEX) outlier samples, GEX-cell type outlier samples, and GEX-donor outlier samples removed.

Gene expression outliers

To identify gene expression (GEX) outliers, we analyzed pseudobulked expression profiles defined at the level of region × village × donor × cell type. Detecting outliers at the cell-type level is critical for downstream cell-type-specific analyses, including differential expression and eQTL mapping. Exclusion of outlier samples can improve statistical power by increasing signal-to-noise ratios, and help minimize spurious signals arising from a small subset of atypical samples. We did not distinguish between biological and technical sources of variation, as samples that deviate strongly from the population distribution may introduce noise regardless of origin.

For each region-village-donor-cell type combination, we generated pseudobulked metacells by aggregating single-cell UMI counts. Within each brain region and cell type, samples with log-transformed library sizes less than the regional mean minus 1.96 standard deviations were removed to exclude poorly ascertained samples. After filtering for library size, remaining metacells had their gene counts normalized to counts per million (CPM) and log-transformed to stabilize variance.

Pairwise Pearson correlations were then computed between all remaining samples within each region-cell-type. For each sample, the median Pearson correlation with all other samples in the same region-cell-type was defined as a conformity score, as described in *Ling et al.*²⁵.

Conformity scores were modified z-score-normalized within each region-cell type, and samples with conformity z-scores less than -3.75 were classified as GEX outliers.

Outlier detection was restricted to cell types with sufficient coverage for inclusion in differential expression and eQTL analyses. Glial cell types included oligodendrocytes, astrocytes, oligodendrocyte precursor cells (OPCs), and microglia. Neuronal cell types included medium spiny neurons (MSNs) in striatal gray matter regions; because conformity scores were highly correlated across MSN subtypes, MSN D1 matrix neurons were used as a proxy for all MSNs. In dorsolateral frontal cortex (DFC), GABAergic VIP interneurons, GABAergic PVALB interneurons, and glutamatergic L2/3 intratelencephalic (L23IT) projection neurons were included.

Donors for whom more than 50% of samples were flagged as GEX outliers (based on either low library size or low conformity score) were designated as donor-level GEX outliers and excluded from selected analyses. Additionally, donors for whom more than 50% of samples for a given cell type were flagged as GEX outliers were designated as donor-cell-type GEX outliers, and only samples of that specific cell type were excluded. Donor-level filtering was applied prior to donor-cell-type-level filtering.

Cell type proportion outliers

Cell type proportion outliers were determined using a sequential filtering approach similar to the gene expression approach described above. Cell type proportion vectors were generated by aggregating and normalizing cell type counts at the subclass level (annotation_sub_class_complete) for each distinct combination of donor and village. Donor × village samples with log-transformed nuclei counts less than the region mean minus 1.96 standard deviations were removed to exclude poorly ascertained samples.

Following removal of samples with low nuclei counts, gray matter samples (from CaH, Pu, NAC, and DFC) with unexpected neuron types contributing to more than 20% of the neuronal population were identified and removed in order to exclude potential misdissections. Since the abundance of neurons is much lower in white matter, this step was not performed for internal capsule samples. Unexpected neuron types for all brain regions included the neuronal subclusters identified from outside the target regions described previously. In the striatum, any cortical intratelencephalic neurons and cortical interneurons were deemed unexpected; likewise, all MSN types were deemed out-of-distribution for DFC.

For the remaining samples, cell type proportions were transformed with the arcsine square root transformation to stabilize their variance, and conformity scores were then computed as the modified z-score of the median (within-region) pairwise Pearson correlation for each sample. Samples with a conformity score less than -3.75 were labeled as cell type proportion outliers.

Whole donors were removed from some analyses on the basis of cell type proportions either if all, or if at least 3, of their samples were flagged as outliers.

15% of overall samples were flagged as cell type proportion outliers. This varied by region – 6% of DFC samples were cell type proportion outliers, 10% of caudate samples, and about 20% of NAC, Pu, and ic. Due to the nature of their position in the striatum, nucleus accumbens, putamen, and internal capsule were most likely to be affected by dissection artifacts from adjacent regions.

Cell type proportion analyses

For each donor and brain region, we aggregated counts of all single cells with cell type annotations. These aggregated counts were then used to calculate both the proportions of individual cell types relative to the total number of cells and ratios between cell types of interests. Donors with exclusionary metadata criteria were removed from these analyses (see **Methods**: Donor selection and tissue procurement; **Table S1,S2**).

Comparisons across brain regions

To examine differences in cell type composition across brain regions, we first calculated the proportion or ratio of each cell type for each donor-region combination. For the ratio-based neuron analyses, we removed outliers with z-scores greater than 5, where z-scores were calculated separately within each brain region. We performed pairwise comparisons between regions using the Wilcoxon rank-sum test.

To assess the consistency of cell type abundances and ratios across regions within individual donors, we calculated Spearman's rank correlation coefficient for each cell type abundance of interest across all pairs of regions (n=498 tests). P-values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg procedure⁹³. Selected region pairs were visualized using scatterplots and heatmaps.

Quantifying variability in glial cell type proportions

Variability of glial cell type proportions was assessed separately within each brain region. For each brain-region–cell-type combination, all pairwise fold ratios between samples were computed as the ratio of the larger to the smaller proportion. The median of these pairwise fold ratios was used as a summary metric of inter-donor variability for that cell type within the specified region. Under this definition, a median fold ratio of 1.2 indicates that, for a randomly selected pair of samples, the larger proportion is typically ~20% greater than the smaller proportion.

In the caudate, median fold ratios were found to be 1.5 for oligodendrocytes, 1.6 for astrocytes, 1.4 for OPCs, and 1.9 for microglia.

Because observed variability reflects both biological and technical sources, we estimated the fraction of variability attributable to donor-level biological differences by leveraging matched samples from dorsolateral prefrontal cortex (DFC). For each glial cell type, Pearson correlations were computed between donor-level proportions in caudate and DFC. The squared correlation

coefficient (r^2) represents the fraction of caudate variance that is shared with DFC across donors.

In caudate, the shared variance (r^2) with DFC was 0.01 for oligodendrocytes, 0.18 for astrocytes, 0.22 for OPCs, and 0.28 for microglia.

To estimate the magnitude of donor-intrinsic variability, observed median fold ratios were adjusted using the cross-region correlation. Specifically, corrected fold ratios were computed as

$$F_{corrected} = \exp(r \cdot \ln(F_{obs})),$$

where F_{obs} is the observed median fold ratio and r is the Pearson correlation between caudate and DFC proportions for the given cell type. This adjustment scales the observed variability by the component that is shared across brain regions, yielding a conservative lower bound on inter-individual biological variability.

The resulting corrected median fold ratios in caudate were 1.05 for oligodendrocytes, 1.2 for astrocytes, 1.2 for OPCs, and 1.4 for microglia.

Cell type proportion regressions

We sought to evaluate age and sex effects only on cell types that were well-ascertained, with a median fraction > 1% across all samples. We additionally computed neuronal subtype proportions with respect to the total number of neurons only. We limited analysis of neurons to gray matter regions only (ic excluded). Since not all samples had OPC subcluster labels, we restricted the analysis of subcluster 0 and 1 abundances to the n=397 samples that contained at least 10 nuclei from both cluster 0 and cluster 1.

As is common for postmortem human brain tissue, the total number of nuclei recovered per sample varied substantially. Because precision in estimated cell type proportions is related to the total counts, we sought a regression framework that explicitly accounted for the underlying counts, rather than treating each measurement as equally precise. We thus modeled cell type abundance using a beta-binomial regression, which operates directly on the observed counts (number of the cell type of interest vs. number of all other cell types). In addition, because cell type proportions are constrained to the unit interval [0,1], we applied a logit link function to allow modeling of the expected (transformed) proportions on an unbounded support, enabling standard regression inference.

Fixed-effect covariates considered for downstream analysis included imputed sex, age, five genetic principal components (PC1–PC5), 10x chemistry (variable “single_cell_assay”), brain region, and biobank. We additionally considered cell-type specific single-cell quality metrics per sample: percent of reads that are intronic (variable “pct_intronic”), the fraction of UMIs removed by CellBender (variable “frac_contamination”). Donor and village were included as random effects. We quantified the contribution of each of these covariates to variance in

(logit-transformed) cell type proportions using variancePartition⁹⁴ (RRID:SCR_019204) and selected covariates that explained substantial variation.

For each cell type of interest, we fit a beta-binomial regression mixed effects model using glmmTMB⁹⁵ (RRID:SCR_025512). Donor and village were modeled as random intercepts, while the selected covariates were incorporated as fixed effects. Analysis of cortical neurons was limited to DFC, so we omitted the fixed effect of brain region.

```
Shell
cell_type_glm <- glmmTMB(
  cbind(n_nuclei, n_other) ~ age_decades + imputed_sex +
  brain_region_abbreviation_simple + mean_frac_contamination + (1 |
  donor_external_id) + (1 | village)
  family=betabinomial(link="logit"),
  data = cell_type_df)
```

From these models, we extracted the coefficients corresponding to age and sex to quantify their effects on cell-type abundances while controlling for other technical and biological covariates (**Table S4**). Statistical significance was assessed using Wald tests, and p-values were adjusted for multiple comparisons across cell types and covariates of interest (age and sex; n=44 tests) using the Benjamini–Hochberg procedure, with a false discovery rate (FDR) threshold of 0.01.

Residuals were computed for each sample as the difference between the observed cell type proportion and the predicted proportion from the model based on fixed effects (age, sex, brain region, and mean contamination fraction) only.

To visualize the age effect on cell type abundance in the original proportion scale, we generated model-based predictions from the beta-binomial GLMM across the observed age range, for each brain region, while holding categorical covariates (sex) constant at the mode or continuous covariates (frac_contamination) fixed at the median values. Predictions were obtained on the response scale with random effects excluded to obtain the population-level fit, and plotted with 95% confidence intervals.

Population-level predicted cell type abundances were computed using emmeans⁹⁶ (RRID:SCR_018734). Predictions were obtained at selected ages (30 and 80 years) for each brain region, averaging over the levels of sex and fraction contamination, and 95% confidence intervals were derived on the response scale. Differences between ages were estimated using contrasts of the back-transformed predictions.

For OPC abundance (relative to all nuclei sampled), predicted proportions decreased significantly with age, from 0.0641 (95% CI: 0.0577–0.0705) at age 30 to 0.0406 (95% CI: 0.0376–0.0436) at age 80. This corresponds to an absolute difference of –0.0235 (95% CI: –0.0303 to –0.0166), an approximate 36.7% reduction in OPC abundance over fifty years.

GABAergic TAC3-PLPP4 interneuron abundance (relative to neurons only) also declined significantly with age, with predicted proportions decreasing from 0.0333 (95% CI: 0.0305–0.0361) at age 30 to 0.0265 (95% CI: 0.0249–0.0281) at age 80. This corresponds to an absolute difference of -0.0068 (95% CI: -0.0101 to -0.00351), an approximate 20.4% reduction over fifty years.

Differential expression

To detect sex-biased and age-dependent gene expression, we summed UMI counts of all assignable single cells for each distinct combination of donor, brain region, and 10x chemistry, generating a pseudobulk (observation-by-gene) matrix at the donor–region–chemistry level for each cell type. Fixed-effect covariates considered for downstream analysis included imputed sex, age, five genetic principal components (PC1–PC5), 10x chemistry (variable “single_cell_assay”), brain region, and biobank, as well as mean single-cell quality metrics per sample: percent of reads that are intronic (variable “pct_intronic”), the fraction of UMIs removed by CellBender (variable “frac_contamination”), and the z-score of the log10 number of nuclei captured (variable “z_log10_nuclei”). Additionally, donor and village are modeled as random effects.

We evaluated these potential covariates using two complementary approaches. First, we generated multidimensional scaling (MDS) plots in Glimma⁹⁷ (RRID:SCR_017389) for each cell type to visualize relationships among samples and identify variables explaining major axes of variation. This qualitative assessment highlighted covariates contributing visible structure in the data, such as donor identity, brain region, and 10x chemistry. Second, we quantified the contribution of each covariate to gene expression variance using the variancePartition⁹⁴ framework.

Based on these assessments, we selected covariates explaining substantial variation for inclusion in the dream⁹⁸ differential expression models. dream, an extension of limma-voom, fits linear mixed models with multiple random effects to account for repeated measurements. Donor and village were modeled as random intercepts, while the selected covariates were incorporated as fixed effects.

For each cell type, we filtered observations to remove donor–region–chemistry samples with low library size that were likely to contribute more noise than signal. Library size filtering was performed by computing the mean and standard deviation of total UMIs per sample and retaining only samples with at least mean $- 1.96$ SD total UMIs. Genes were then filtered to remove lowly expressed features. Counts were transformed to counts per million (CPM) using edgeR⁹⁹ (RRID:SCR_012802), and we retained genes where at least 10% of observations had $CPM \geq 1$.

Differential expression was tested at two levels of granularity. First, to estimate the average effect of sex or age on each cell type across brain regions, we fit a region-averaged model with age as a single fixed-effect term and region included as a categorical covariate. This model was specified in dream as:

None

```
~ 0 + age + z_log10_nuclei + PC1 + PC2 + PC3 + PC4 + PC5 +  
  pct_intronic + frac_contamination + imputed_sex +  
  single_cell_assay + region + biobank +  
  (1 | donor) + (1 | village)
```

To assess region-specific heterogeneity in age effects, we fit a second model in which age effects were allowed to vary by region using region-specific age terms. Using caudate as the reference region, region-specific age effects were encoded as separate fixed-effect terms (for example, age_regionCaH, age_regionDFC). The corresponding dream formula was:

None

```
~ 0 + z_log10_nuclei + PC1 + PC2 + PC3 + PC4 + PC5 +  
  pct_intronic + frac_contamination + imputed_sex +  
  single_cell_assay + region + biobank +  
  age_regionCaH + age_regionDFC + age_regionic +  
  age_regionNAC + age_regionPu +  
  (1 | donor) + (1 | village)
```

This formulation allowed us to estimate region-specific age effects from a single model fit and to compare the magnitude and direction of age-associated differential expression for each gene across brain regions.

Age-associated differential expression analyses and visualizations

Age-associated differential expression was assessed using precomputed results from linear modeling frameworks treating age as a continuous variable, as described in the section on differential regression framework. Results included gene-level effect size estimates, test statistics, and FDR-adjusted significance values. Effect size uncertainties derived from model outputs were used in downstream analyses requiring standard error estimates. Analyses were performed both with brain regions combined and with region-specific models, as noted in individual analyses below.

Cell-type aging effect-size correlations

Concordance of age-associated transcriptional effects across cell types and brain regions was evaluated using pairwise comparisons of gene-level effect sizes in R using cor.test. Genes were aligned across conditions, and comparisons were restricted to genes showing evidence of age association in at least one condition (FDR < 0.05). Spearman's rank correlation was used. Correlation strength and direction were summarized using signed metrics derived from correlation coefficients. Selected comparisons were visualized using scatterplots.

Gene k-means clustering across cell types

A matrix of gene-level age-associated effect sizes estimated across all five brain regions was constructed across analyzed cell types. To arrive at a set of informative genes for this analysis we filtered to genes expressed at 10 transcripts per million (for the cell type being analyzed; in either striatum or cortex) with an FDR < 0.01 and an absolute fold change (per decade) > 1.05 in at least one cell type. This left us with 2,478 input genes. These genes were clustered based on shared age-related expression patterns using k-means clustering (stats::kmeans) with Euclidean distance, a fixed random seed (42), 200 random starts, and a maximum of 20 iterations. Effect sizes were z-score-scaled across cell types prior to clustering. The number of clusters was guided by silhouette-based evaluation (evaluating k = 10-30). Clustered gene patterns were visualized as heatmaps.

TRADE analysis by cell type

Transcriptome-wide impact of aging was quantified using TRADE⁶¹, which integrates gene-level effect sizes and associated uncertainties to estimate global transcriptional perturbation. TRADE analyses were performed separately for each cell type using effect sizes estimated across all five brain regions. Autosomal genes were included in primary analyses, with sex chromosome genes analyzed separately. The primary outputs used in downstream analyses were the estimated transcriptome-wide impact values per cell type.

Gene set enrichment analyses

Rank-based gene set enrichment analysis was performed using fgsea (RRID:SCR_020938) in R. Genes were ranked by age-associated t-statistics. Curated gene sets from the MSigDB (RRID:SCR_016863) Human C5 collection (biological process, molecular function, and cellular compartment) were tested using permutation-based enrichment with multiple hypothesis correction. Gene sets used for visualization were selected from leadingEdge genes of statistically significant (padj < 0.01) GO categories related to the selected themes (e.g., GOBP_REGULATION_OF_SYNAPTIC_PLASTICITY, GOBP_DENDRITIC_SPINE_MORPHOGENESIS, GOBP_G_PROTEIN_COUPLED_DOPAMINE_RECEPTOR_SIGNALING_PATHWAY, GOBP_CALCIUM_ION_TRANSPORT; **Table S5**).

RNA based age prediction

Data preprocessing and feature selection

Single-cell RNA-sequencing data were aggregated to generate pseudobulk expression profiles defined at the donor × cell type × region level. For donors with multiple libraries or chemistries, raw UMI counts were summed across nuclei to yield a single expression profile per donor within each cell type and region. Only autosomal genes (chromosomes 1–22) were retained for

analysis; genes located on sex chromosomes or mitochondrial DNA were excluded prior to further filtering.

Samples with low library size were removed as described in the differential expression (DE) analysis. Genes were filtered using the same criteria applied in the DE framework, including removal of lowly expressed features based on counts per million (CPM). To restrict the feature space to age-informative genes, we retained only those genes identified as significantly associated with age ($FDR \leq 0.05$) in the corresponding cell type by region-specific DE analysis. This DE-based gene selection was performed once per cell type and region using all available donors prior to model fitting and served as a predefined feature set for age prediction.

Normalization and model specification

For each modeling iteration, raw count data were transformed to log counts per million (log-CPM) using edgeR. Gene expression features were standardized by z-score transformation, computed separately for each gene using the mean and standard deviation estimated from the training donors only. These scaling parameters were then applied unchanged to the held-out donors. Test samples were provided to the model as raw count data and transformed exclusively using parameters derived from the training set, mirroring the procedure that would be used for prediction in an external dataset.

Chronological age (expressed in decades) was predicted using ridge regression, implemented as elastic net regression with $\alpha = 0$ and a Gaussian error model. At each iteration, the regularization parameter λ was selected by internal 10-fold cross-validation within the training set, minimizing mean absolute error (MAE). Model coefficients and intercepts were estimated from the training donors, and the resulting model was used to generate predictions for the held-out donors.

Repeated age-stratified cross-validation

Predictive performance was assessed using repeated age-stratified 80/20 splits (Monte Carlo cross-validation). For each cell type by region model, donors were binned into five age quantiles, and approximately 20% of donors were withheld from each bin to form a test set, preserving the overall age distribution. This procedure was repeated 200 times, drawing a new stratified split at each iteration.

For each repeat, the model was trained on 80% of donors and used to predict age in the held-out 20%. Only out-of-fold predictions were retained. After 200 iterations, each donor had multiple independent out-of-fold predictions (approximately 40 on average). Final predicted age for each donor was defined as the mean of these out-of-fold predictions. Model performance metrics, including Pearson correlation and mean absolute error, were computed from these aggregated out-of-fold predictions.

Bias correction and residual age

Across models, predicted ages exhibited regression-to-the-mean shrinkage, with younger donors predicted older and older donors predicted younger. To account for this systematic age-dependent bias, we fit a generalized additive model (GAM) for each cell type by region using aggregated out-of-fold predictions as a smooth function of chronological age. The model used a cubic regression spline ($k = 5$) which was estimated by restricted maximum likelihood, and employed a scaled t distribution to provide robustness to outliers.

Bias-corrected predicted age was defined by removing the estimated age-dependent deviation from the identity relationship between predicted and chronological age, thereby eliminating regression-to-the-mean effects. Residual age was defined both as the raw difference between predicted and chronological age and as the bias-corrected difference. Both forms were used in downstream analyses, with bias-corrected residual age used to assess coordination of transcriptional aging across cell types and regions.

eQTL analysis

Discovery

To generate eQTLs, we used a workflow combining parts of previous pipelines^{25,100}. We summed UMI counts of all assignable single cells included in nucleus selection for each distinct combination of donor, brain region, and cell type to create pseudobulked (donor-by-gene) expression matrices at the region cell-type level. Pre-processing of expression data included summing the gene expression per gene defined by the GENCODE v43 gene models and removing the lower 50% of expressed genes, leaving 18,984 to 19,056 genes per brain region cell type matrix. Each matrix was then normalized per-donor using code derived from the pyQTL¹⁰¹ implementation of edgeR CPM, then standardized using the pyQTL implementation of an inverse normal transformation (INT).

The covariates from the differential expression were used for identifying fixed effects. Probabilistic Estimation of Expression Residuals (PEER)^{102,103} was used to generate 10 additional covariates for each brain region \times cell type analysis.

Variants included in the analysis were from chromosomes 1-22 and X using the GRCh38 reference. Sites with a minor allele frequency (MAF) less than 0.05 or more than 0.95 were excluded from the analysis. For each variant, 90% of the donors were required to have a genotype quality greater than 30, as measured by the GQ quality tag in the VCF. Sites where the HTSJDK¹⁰⁴ calculated a Hardy Weinberg equilibrium (HWE)^{88,89} p -value below $1e-4$ were excluded.

The variant genotypes, normalized gene expression phenotypes, and set of covariates were input into tensorQTL¹⁰⁵ to generate cis-QTLs using its cis mode with a seed of 777, a default genome-wide false discovery rate (FDR) setting of 0.05, using a setting of one megabase window before and after each transcription start site (TSS) as defined by the GENCODE v43 gene models. For cis-QTL results reaching genome-wide significance, the tensorQTL cis_independent mode was then used to generate conditionally independent cis-QTLs¹⁰⁶ with

the same seed, covariates, and TSS window. A list of all pairwise variant and gene expression summary results, without genomewide FDR calculation, was generated per matrix by using the same tensorQTL inputs with the `cis_nominal` mode.

Donor exclusion criteria were evaluated using CTP- and gene-expression-based outlier metrics, consistent with the approach used for differential expression analyses (**Figure S10**). Data-driven filtering preserved or improved discovery yield without materially altering estimated eQTL effect sizes, whereas metadata-driven exclusions substantially reduced power (**Figure S15**).

K-means clustering

We compiled the union of SNP–gene pairs that were significant ($q < 0.01$) in at least one cell type–region combination. For each gene, we then selected a single lead cis-eQTL variant, defined as the variant with the largest absolute effect size across cell type–region combinations. Effect directions were reoriented so that the strongest effect for each gene was positive. The effect sizes of these lead variants across 17 region-stratified cell-type profiles defined the cross-context effect profile for each gene. Missing effect sizes, arising from genes not tested in specific cell types, were imputed as zero. Genes were then clustered based on these effect-size profiles using k-means.

We selected $K = 13$ as the smallest value that resolved distinct, biologically interpretable effect-size patterns without merging them. For visualization, cell types were displayed in a fixed biologically informed order (for example, MSN subtypes adjacent to each other), and clusters were arranged to group related patterns. Within clusters, genes were ordered by hierarchical clustering of effect-size profiles (correlation distance, average linkage) to highlight internal structure.

Expression fraction

To quantify the extent to which cell-type-specific eQTLs is driven by cell-type-specific gene expression, we computed median CPM expression across donors for each of the 9,899 eGenes in each of the 17 cell type–region combinations. For each gene–cell type pair, the median was computed from donors with non-NA expression for that gene; gene–cell type pairs with no expressing donors were set to zero. For each of the eight cell-type-specific K-means clusters (clusters 6–13), we identified the cell type with the highest median expression for each eGene and computed the fraction of eGenes whose highest-expressing cell type matched the cell types(s) in which the eQTLs manifested cell-type-specific regulatory effects.

Pairwise correlation

To quantify similarity of eQTL architectures across cell types, we computed pairwise Spearman correlations of cis-eQTL effect sizes between all cell-type pairs. For each pair of cell types, we included all SNP–gene pairs significant ($q < 0.01$) in at least one of the two cell types.

Correlations were assembled into a symmetric matrix and reported as squared correlation coefficients (R^2) to facilitate interpretability.

The resulting R^2 matrix was visualized as a heatmap, with hierarchical clustering applied to both rows and columns to group cell types with similar effect-size profiles.

eQTL effect size and genetic constraint

To assess whether evolutionarily constrained genes tend to have smaller eQTL effect sizes, we correlated eQTL effect sizes with LOEUF (loss-of-function observed/expected upper bound fraction) scores from gnomAD⁶⁸ v4.1. Lower LOEUF indicates stronger selective constraint against loss-of-function variants. We restricted the analysis to protein-coding genes with LOEUF scores, using canonical MANE Select transcripts.

For the cluster-level analysis, we used the same lead SNP-gene pairs as the K-means clustering (see above). For each gene, we computed the mean absolute eQTL effect size across cell types in which the eQTL reached significance (nominal p-value below the per-gene empirical threshold from tensorQTL permutations). We then computed Spearman rank correlations between LOEUF and mean effect size within each K-means cluster ($K=13$), with p-values corrected for multiple testing using the Benjamini-Hochberg procedure. To obtain an overall measure of the LOEUF-effect size association while controlling for cluster membership, we fit a linear regression of mean effect size on LOEUF and cluster (as a categorical covariate) and report the p-value on the LOEUF coefficient.

For the cell-type-level analysis, we used the lead SNP nominated by tensorQTL for each eGene ($q < 0.01$) independently in each of the 17 cell type-region combinations and computed Spearman correlations between LOEUF and absolute effect size per cell type, with Benjamini-Hochberg correction across cell types.

Bibliography

1. Reinhold, K., Iadarola, M., Tang, S., Chang, A., Kuwamoto, W., Albanese, M.A., Sun, S., Hakim, R., Zimmer, J., Wang, W., et al. (2025). Striatum supports fast learning but not memory recall. *Nature* *643*, 458–467.
2. Yin, H.H., and Knowlton, B.J. (2006). The role of the basal ganglia in habit formation. *Nat. Rev. Neurosci.* *7*, 464–476.
3. Eshel, N., Toupouse, G.C., Wang, A.R., Osterman, A.K., Shank, A.N., Groome, A.M., Taniguchi, L., Cardozo Pinto, D.F., Tucciarone, J., Bentzley, B.S., et al. (2024). Striatal dopamine integrates cost, benefit, and motivation. *Neuron* *112*, 500–514.e5.
4. Cox, J., and Witten, I.B. (2019). Striatal circuits for reward learning and decision-making. *Nat. Rev. Neurosci.* *20*, 482–494.
5. Klaus, A., Alves da Silva, J., and Costa, R.M. (2019). What, if, and when to move: Basal ganglia circuits and self-paced action initiation. *Annu. Rev. Neurosci.* *42*, 459–483.
6. Vassiliadis, P., Beanato, E., Popa, T., Windel, F., Morishita, T., Neufeld, E., Duque, J., Derosiére, G., Wessel, M.J., and Hummel, F.C. (2024). Non-invasive stimulation of the human striatum disrupts reinforcement learning of motor skills. *Nat. Hum. Behav.* *8*, 1581–1598.
7. Kang, J., Kim, H., Hwang, S.H., Han, M., Lee, S.-H., and Kim, H.F. (2021). Primate ventral striatum maintains neural representations of the value of previously rewarded objects for habitual seeking. *Nat. Commun.* *12*, 2100.
8. Robinson, D., Wu, H., Munne, R.A., Ashtari, M., Alvir, J.M., Lerner, G., Koreen, A., Cole, K., and Bogerts, B. (1995). Reduced caudate nucleus volume in obsessive-compulsive disorder. *Arch. Gen. Psychiatry* *52*, 393–398.
9. Rangaprakash, D., Tadayonnejad, R., Deshpande, G., O’Neill, J., and Feusner, J.D. (2021). FMRI hemodynamic response function (HRF) as a novel marker of brain function: applications for understanding obsessive-compulsive disorder pathology and treatment response. *Brain Imaging Behav.* *15*, 1622–1640.
10. Saxena, S., Brody, A.L., Schwartz, J.M., and Baxter, L.R. (1998). Neuroimaging and frontal-subcortical circuitry in obsessive-compulsive disorder. *Br. J. Psychiatry Suppl.*, 26–37.
11. Reiner, A., and Deng, Y.-P. (2018). Disrupted striatal neuron inputs and outputs in Huntington’s disease. *CNS Neurosci. Ther.* *24*, 250–280.
12. Matsushima, A., Pineda, S.S., Crittenden, J.R., Lee, H., Galani, K., Mantero, J., Tombaugh, G., Kellis, M., Heiman, M., and Graybiel, A.M. (2023). Transcriptional vulnerabilities of striatal neurons in human and rodent models of Huntington’s disease. *Nat. Commun.* *14*, 282.
13. Parent, A. (2017). The history of the basal ganglia: The nuclei. In *Handbook of Behavioral Neuroscience Handbook of Behavioral Neuroscience*. (Elsevier), pp. 33–44.
14. Foster, N.N., Barry, J., Korobkova, L., Garcia, L., Gao, L., Becerra, M., Sherafat, Y., Peng, B., Li, X., Choi, J.-H., et al. (2021). The mouse cortico-basal ganglia-thalamic network. *Nature* *598*, 188–194.
15. Del Rey, N.L.-G., and García-Cabezas, M.Á. (2023). Cytology, architecture, development, and connections of the primate striatum: Hints for human pathology. *Neurobiol. Dis.* *176*, 105945.
16. Zhang, M., Pan, X., Jung, W., Halpern, A.R., Eichhorn, S.W., Lei, Z., Cohen, L., Smith, K.A., Tasic, B., Yao, Z., et al. (2023). Molecularly defined and spatially resolved cell atlas of the whole mouse brain. *Nature* *624*, 343–354.
17. Krienen, F.M., Goldman, M., Zhang, Q., C H Del Rosario, R., Florio, M., Machold, R., Saunders, A., Levandowski, K., Zaniewski, H., Schuman, B., et al. (2020). Innovations

- present in the primate interneuron repertoire. *Nature* 586, 262–269.
18. Langlieb, J., Sachdev, N.S., Balderrama, K.S., Nadaf, N.M., Raj, M., Murray, E., Webber, J.T., Vanderburg, C., Gazestani, V., Tward, D., et al. (2023). The molecular cytoarchitecture of the adult mouse brain. *Nature* 624, 333–342.
 19. Peters, A.J., Fabre, J.M.J., Steinmetz, N.A., Harris, K.D., and Carandini, M. (2021). Striatal activity topographically reflects cortical activity. *Nature* 591, 420–425.
 20. Ting, J.T., Johansen, N.J., Kalmbach, B.E., Taskin, N., Lee, B., Clark, J.K., Kendrick, R., Ng, L., Radaelli, C., Weed, N., et al. (2024). Distinctive physiology of molecularly identified medium spiny neurons in the macaque putamen. *Cell Rep.* 43, 114963.
 21. Planert, H., Berger, T.K., and Silberberg, G. (2013). Membrane properties of striatal direct and indirect pathway neurons in mouse and rat slices and their modulation by dopamine. *PLoS One* 8, e57054.
 22. He, J., Kleyman, M., Chen, J., Alikaya, A., Rothenhoefer, K.M., Ozturk, B.E., Wirthlin, M., Bostan, A.C., Fish, K., Byrne, L.C., et al. (2021). Transcriptional and anatomical diversity of medium spiny neurons in the primate striatum. *Curr. Biol.* 31, 5473–5486.e6.
 23. Kawaguchi, Y. (1993). Physiological, morphological, and histochemical characterization of three classes of interneurons in rat neostriatum. *J. Neurosci.* 13, 4908–4923.
 24. Tepper, J.M., and Bolam, J.P. (2004). Functional diversity and specificity of neostriatal interneurons. *Curr. Opin. Neurobiol.* 14, 685–692.
 25. Ling, E., Nemesh, J., Goldman, M., Kamitaki, N., Reed, N., Handsaker, R.E., Genovese, G., Vogelgsang, J.S., Gerges, S., Kashin, S., et al. A concerted neuron–astrocyte program declines in ageing and schizophrenia. *Nature*, 1–8.
 26. Harrison, B.J., Soriano-Mas, C., Pujol, J., Ortiz, H., López-Solà, M., Hernández-Ribas, R., Deus, J., Alonso, P., Yücel, M., Pantelis, C., et al. (2009). Altered corticostriatal functional connectivity in obsessive-compulsive disorder. *Arch. Gen. Psychiatry* 66, 1189–1200.
 27. Liu, X., Eickhoff, S.B., Caspers, S., Wu, J., Genon, S., Hoffstaedter, F., Mars, R.B., Sommer, I.E., Eickhoff, C.R., Chen, J., et al. (2021). Functional parcellation of human and macaque striatum reveals human-specific connectivity in the dorsal caudate. *Neuroimage* 235, 118006.
 28. Perez, M.S., Yin, R., Scott, M.R., Zong, W., Seney, M.L., Xue, X., Hildebrand, M.A., Shankar, V.G., Glausier, J.R., Lewis, D.A., et al. (2025). Sex and regional differences in gene expression across the striatum in psychosis. *Transl. Psychiatry* 15, 192.
 29. Parkes, L., Fulcher, B.D., Yücel, M., and Fornito, A. (2017). Transcriptional signatures of connectomic subregions of the human striatum. *Genes Brain Behav.* 16, 647–663.
 30. Ersche, K.D., Meng, C., Ziauddeen, H., Stochl, J., Williams, G.B., Bullmore, E.T., and Robbins, T.W. (2020). Brain networks underlying vulnerability and resilience to drug addiction. *Proc. Natl. Acad. Sci. U. S. A.* 117, 15253–15261.
 31. Hibar, D.P., Cheung, J.W., Medland, S.E., Mufford, M.S., Jahanshad, N., Dalvie, S., Ramesar, R., Stewart, E., van den Heuvel, O.A., Pauls, D.L., et al. (2018). Significant concordance of genetic variation that increases both the risk for obsessive-compulsive disorder and the volumes of the nucleus accumbens and putamen. *Br. J. Psychiatry* 213, 430–436.
 32. Horga, G., Cassidy, C.M., Xu, X., Moore, H., Slifstein, M., Van Snellenberg, J.X., and Abi-Dargham, A. (2016). Dopamine-related disruption of functional topography of striatal connections in unmedicated patients with schizophrenia. *JAMA Psychiatry* 73, 862–870.
 33. Han, I., You, Y., Kordower, J.H., Brady, S.T., and Morfini, G.A. (2010). Differential vulnerability of neurons in Huntington’s disease: the role of cell type-specific features. *J. Neurochem.* 113, 1073–1091.
 34. Ding, S.-L., Royall, J.J., Sunkin, S.M., Ng, L., Facer, B.A.C., Lesnar, P., Guillozet-Bongaarts, A., McMurray, B., Szafer, A., Dolbeare, T.A., et al. (2016). Comprehensive cellular-resolution atlas of the adult human brain. *J. Comp. Neurol.* 524,

- 3127–3481.
35. Wells, M.F., Nemesh, J., Ghosh, S., Mitchell, J.M., Salick, M.R., Mello, C.J., Meyer, D., Pietilainen, O., Piccioni, F., Guss, E.J., et al. (2023). Natural variation in gene expression and viral susceptibility revealed by neural progenitor cell villages. *Cell Stem Cell* *30*, 312–332.e13.
 36. Fleming, S.J., Chaffin, M.D., Arduini, A., Akkad, A.-D., Banks, E., Marioni, J.C., Philippakis, A.A., Ellinor, P.T., and Babadi, M. (2023). Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nat. Methods* *20*, 1323–1335.
 37. DropSift: cell caller for sn/snRNA sequence data (Github).
 38. Johansen, N.J., Fu, Y., Schmitz, M., Dubuc, A., Kempynck, N., Wirthlin, M., Garcia, A.D., Hewitt, M., Turner, M.A., Seeman, S.C., et al. (2025). Cross-species consensus atlas of the primate basal ganglia. *bioRxiv.org*. <https://doi.org/10.64898/2025.12.15.694496>.
 39. cell_type_mapper: Repository for storing prototype functionality implementations for the BKP (Github).
 40. Gerfen, C.R., and Surmeier, D.J. (2011). Modulation of striatal projection systems by dopamine. *Annu. Rev. Neurosci.* *34*, 441–466.
 41. Graybiel, A.M., and Ragsdale, C.W., Jr (1978). Histochemically distinct compartments in the striatum of human, monkeys, and cat demonstrated by acetylthiocholinesterase staining. *Proc. Natl. Acad. Sci. U. S. A.* *75*, 5723–5726.
 42. Gerfen, C.R. (1984). The neostriatal mosaic: compartmentalization of corticostriatal input and striatonigral output systems. *Nature* *311*, 461–464.
 43. Graybiel, A.M., and Matsushima, A. (2023). Striosomes and matrisomes: Scaffolds for dynamic coupling of volition and action. *Annu. Rev. Neurosci.* *46*, 359–380.
 44. Gangarossa, G., Espallergues, J., Mailly, P., De Bundel, D., de Kerchove d’Exaerde, A., Hervé, D., Girault, J.-A., Valjent, E., and Krieger, P. (2013). Spatial distribution of D1R- and D2R-expressing medium-sized spiny neurons differs along the rostro-caudal axis of the mouse dorsal striatum. *Front. Neural Circuits* *7*, 124.
 45. Ren, K., Guo, B., Dai, C., Yao, H., Sun, T., Liu, X., Bai, Z., Wang, W., and Wu, S. (2017). Striatal distribution and cytoarchitecture of dopamine receptor subtype 1 and 2: Evidence from double-labeling transgenic mice. *Front. Neural Circuits* *11*, 57.
 46. Kraft, A.W., Lee, M., Rayan, N., Gao, H., Milidantri, J., Vanderburg, C., Balderrama, K., Nadaf, N., Kumar, V., Flowers, K., et al. (2026). Mesoscale molecular architecture of the human striatum across cell types and lifespan. *bioRxiv*. <https://doi.org/10.64898/2026.03.04.709715>.
 47. Gayden, J., Puig, S., Srinivasan, C., Phan, B.N., Abdelhady, G., Buck, S.A., Gamble, M.C., Tejada, H.A., Dong, Y., Pfenning, A.R., et al. (2023). Integrative multi-dimensional characterization of striatal projection neuron heterogeneity in adult brain. *bioRxiv*. <https://doi.org/10.1101/2023.05.04.539488>.
 48. Voorn, P., Brady, L.S., Berendse, H.W., and Richfield, E.K. (1996). Densitometrical analysis of opioid receptor ligand binding in the human striatum--I. Distribution of mu opioid receptor defines shell and core of the ventral striatum. *Neuroscience* *75*, 777–792.
 49. Garma, L.D., Harder, L., Barba-Reyes, J.M., Marco Salas, S., Díez-Salguero, M., Nilsson, M., Serrano-Pozo, A., Hyman, B.T., and Muñoz-Manchado, A.B. (2024). Interneuron diversity in the human dorsal striatum. *Nat. Commun.* *15*, 6164.
 50. Kreitzer, A.C., and Malenka, R.C. (2008). Striatal plasticity and basal ganglia circuit function. *Neuron* *60*, 543–554.
 51. Liu, X.-P., Dalley, R., Johansen, N., Budzillo, A., Thijssen, J., Miller, J.A., Walling-Bell, S., Sawchuk, S., Alfiler, L., Andrade, J., et al. (2026). Morphoelectric diversity and specialization of neuronal cell types in the primate striatum. *bioRxiv*, 2026.02.26.708019. <https://doi.org/10.64898/2026.02.26.708019>.

52. Corrigan, E.K., DeBerardine, M., Poddar, A., Turrero García, M., de la O, S., He, S., Sen, H., Duhne, M., Lindberg, S., Song, M., et al. (2025). Conservation and alteration of mammalian striatal interneurons. *Nature* **647**, 187–193.
53. Bergles, D.E., Roberts, J.D., Somogyi, P., and Jahr, C.E. (2000). Glutamatergic synapses on oligodendrocyte precursor cells in the hippocampus. *Nature* **405**, 187–191.
54. Emani, P.S., Liu, J.J., Clarke, D., Jensen, M., Warrell, J., Gupta, C., Meng, R., Lee, C.Y., Xu, S., Dursun, C., et al. (2024). Single-cell genomics and regulatory networks for 388 human brains. *Science* **384**, eadi5199.
55. Allen, W.E., Blosser, T.R., Sullivan, Z.A., Dulac, C., and Zhuang, X. (2023). Molecular and spatial signatures of mouse brain aging at single-cell resolution. *Cell* **186**, 194–208.e18.
56. Barba-Reyes, J.M., Harder, L., Marco Salas, S., Jaisa-Aad, M., Muñoz-Castro, C., Garma, L.D., Rafati, N., Nilsson, M., Hyman, B.T., Serrano-Pozo, A., et al. (2025). Oligodendroglia vulnerability in the human dorsal striatum in Parkinson’s disease. *Acta Neuropathol.* **149**, 46.
57. Benarroch, E. (2023). What are the roles of oligodendrocyte precursor cells in normal and pathologic conditions? *Neurology* **101**, 958–965.
58. Kumar, A., Gibbs, J.R., Beilina, A., Dillman, A., Kumaran, R., Trabzuni, D., Ryten, M., Walker, R., Smith, C., Traynor, B.J., et al. (2013). Age-associated changes in gene expression in human brain and isolated neurons. *Neurobiol. Aging* **34**, 1199–1209.
59. Ham, S., and Lee, S.-J.V. (2020). Advances in transcriptome analysis of human brain aging. *Exp. Mol. Med.* **52**, 1787–1797.
60. Zannas, A.S., Jia, M., Hafner, K., Baumert, J., Wiechmann, T., Pape, J.C., Arloth, J., Ködel, M., Martinelli, S., Roitman, M., et al. (2019). Epigenetic upregulation of FKBP5 by aging and stress contributes to NF- κ B-driven inflammation and cardiovascular risk. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 11370–11379.
61. Nadig, A., Replogle, J.M., Pogson, A.N., Murthy, M., McCarroll, S.A., Weissman, J.S., Robinson, E.B., and O’Connor, L.J. (2025). Transcriptome-wide analysis of differential expression in perturbation atlases. *Nat. Genet.* **57**, 1228–1237.
62. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550.
63. Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115.
64. Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 301–320.
65. Ginhoux, F., Lim, S., Hoeffel, G., Low, D., and Huber, T. (2013). Origin and differentiation of microglia. *Front. Cell. Neurosci.* **7**, 45.
66. Siletti, K., Hodge, R., Mossi Albiach, A., Lee, K.W., Ding, S.-L., Hu, L., Lönnerberg, P., Bakken, T., Casper, T., Clark, M., et al. (2023). Transcriptomic diversity of cell types across the adult human brain. *Science* **382**, eadd7046.
67. Kozareva, V., Martin, C., Osorno, T., Rudolph, S., Guo, C., Vanderburg, C., Nadaf, N., Regev, A., Regehr, W.G., and Macosko, E. (2021). A transcriptomic atlas of mouse cerebellar cortex comprehensively defines cell types. *Nature* **598**, 214–219.
68. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443.
69. Kanai, M., Delorey, T.M., Honkanen, J., Rodosthenous, R.S., Juvila, J., Murphy, S., Teixeira-Soldano, I., Hwang, H.S., Karjalainen, J., Halonen, J., et al. (2025). Population-scale multiome immune cell atlas reveals complex disease drivers. *medRxiv*. <https://doi.org/10.1101/2025.11.25.25340489>.

70. Mostafavi, H., Spence, J.P., Naqvi, S., and Pritchard, J.K. (2023). Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat. Genet.* *55*, 1866–1875.
71. Fonarow, G.C., Reeves, M.J., Zhao, X., Olson, D.M., Smith, E.E., Saver, J.L., Schwamm, L.H., and Get With the Guidelines-Stroke Steering Committee and Investigators (2010). Age-related differences in characteristics, performance measures, treatment trends, and outcomes in patients with ischemic stroke. *Circulation* *121*, 879–891.
72. Kim, D., Kim, H.-S., Choi, S.-M., Kim, B.C., Lee, M.-C., Lee, K.-H., and Lee, J.-H. (2019). Primary age-related tauopathy: An elderly brain pathology frequently encountered during autopsy. *J. Pathol. Transl. Med.* *53*, 159–163.
73. Iida, M.A., Farrell, K., Walker, J.M., Richardson, T.E., Marx, G.A., Bryce, C.H., Purohit, D., Ayalon, G., Beach, T.G., Bigio, E.H., et al. (2021). Predictors of cognitive impairment in primary age-related tauopathy: an autopsy study. *Acta Neuropathol. Commun.* *9*, 134.
74. Nelson, P.T., Dickson, D.W., Trojanowski, J.Q., Jack, C.R., Boyle, P.A., Arfanakis, K., Rademakers, R., Alafuzoff, I., Attems, J., Brayne, C., et al. (2019). Limbic-predominant age-related TDP-43 encephalopathy (LATE): consensus working group report. *Brain* *142*, 1503–1527.
75. Tao, S., Li, X., Huang, Y., Abeysinghe, R., Tong, L., Lin, S., Chou, W.-C., Edgu-Fry, E., Smith, K.A., Ng, L., et al. (2025). Tissue-to-bytes: A catalytic digital twin platform for consortium-scale integration of single-cell omics data across the BRAIN Initiative Cell Atlas Network. *bioRxiv.org*. <https://doi.org/10.64898/2025.12.06.692733>.
76. NeuroBioBank best practices <https://neurobiobank.nih.gov/about-best-practices/>.
77. Spina, L., McDonald, T., Reed, N., and Lutservitz, A. (2023). Village Nuclei Isolation With Optiprep v1. <https://doi.org/10.17504/protocols.io.36wgg3bmxlk5/v1>.
78. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* *161*, 1202–1214.
79. Git-Forge HTML Preview (Github).
80. Hao, Y., Stuart, T., Kowalski, M.H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., et al. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* *42*, 293–304.
81. Parks, B., and Greenleaf, W. (2025). Scalable high-performance single cell data analysis with BPCells. *bioRxiv.org*. <https://doi.org/10.1101/2025.03.27.645853>.
82. Alquicira-Hernandez, J., Sathe, A., Ji, H.P., Nguyen, Q., and Powell, J.E. (2019). scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol.* *20*, 264.
83. Seeker, L.A., Bestard-Cuche, N., Jäkel, S., Kazakou, N.-L., Bøstrand, S.M.K., Wagstaff, L.J., Cholewa-Waclaw, J., Kilpatrick, A.M., Van Bruggen, D., Kabbe, M., et al. (2023). Brain matters: unveiling the distinct contributions of region, age, and sex to glia diversity and CNS function. *Acta Neuropathol. Commun.* *11*, 84.
84. Macnair, W., Calini, D., Agirre, E., Bryois, J., Jäkel, S., Smith, R.S., Kukanja, P., Stokar-Regenscheit, N., Ott, V., Foo, L.C., et al. (2025). snRNA-seq stratifies multiple sclerosis patients into distinct white matter glial responses. *Neuron* *113*, 396–410.e9.
85. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* *185*, 3426–3440.e19.
86. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* *4*, 7.
87. Shaun Purcell, C.C. PLINK 1.9.

88. Hardy, G.H. (1908). Mendelian proportions in a mixed population. *Science* 28, 49–50.
89. Weinberg, W. (1908). About proof of heredity in humans. Germany: The association for national natural history in württemberg 64, 368–382.
90. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 2, e190.
91. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909.
92. Genovese, G. (2019). *kgp2anc*.
93. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* 57, 289–300.
94. Hoffman, G.E., and Schadt, E.E. (2016). *variancePartition*: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* 17, 483.
95. McGillicuddy, M., Popovic, G., Bolker, B.M., and Warton, D.I. (2024). Parsimoniously fitting large multivariate random effects in *glmmTMB*. *arXiv [stat.ME]*.
96. Lenth, R.V., and Piaskowski, J. (2025). *emmeans*: Estimated Marginal Means, aka Least-Squares Means. Preprint.
97. Su, S., Law, C.W., Ah-Cann, C., Asselin-Labat, M.-L., Blewitt, M.E., and Ritchie, M.E. (2017). *Glimma*: interactive graphics for gene expression analysis. *Bioinformatics* 33, 2050–2052.
98. Hoffman, G.E., and Roussos, P. (2021). *Dream*: powerful differential expression analysis for repeated measures designs. *Bioinformatics* 37, 192–201.
99. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). *edgeR*: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140.
100. *eqtl_pipeline_terra*: scripts for eQTL pipeline in Terra (Github).
101. *pyqtl*: Collection of analysis tools for quantitative trait loci (Github).
102. Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* 6, e1000770.
103. Parts, L., Stegle, O., Winn, J., and Durbin, R. (2011). Joint genetic analysis of gene expression data with inferred cellular phenotypes. *PLoS Genet.* 7, e1001276.
104. *htsjdk*: A Java API for high-throughput sequencing data (HTS) formats (Github).
105. Taylor-Weiner, A., Aguet, F., Haradhvala, N.J., Gosai, S., Anand, S., Kim, J., Ardlie, K., Van Allen, E.M., and Getz, G. (2019). Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* 20, 228.
106. GTEx Consortium, Laboratory, Data Analysis & Coordinating Center (LDACC)-Analysis Working Group, Statistical Methods groups-Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site-NDRI, et al. (2017). Genetic effects on gene expression across human tissues. *Nature* 550, 204–213.