Parallel RNA and DNA analysis after Deep-sequencing (PRDD-seq) reveals cell type-specific lineage patterns in human brain

August Yue Huang[a,b,c,1], Pengpeng Li[a,b,c,1], Rachel E. Rodin[a,b,c,d], Sonia N. Kim[a,b,c,d], Yanmei Dou[e], Connor J. Kenny[a,b,c], Shyam K. Akula[a,b,c,d], Rebecca D. Hodge[f], Trygve E. Bakken[f], Jeremy A. Miller[f], Ed S. Lein[f], Peter J. Park[e], Eunjung Alice Lee[a,b,c], Christopher A. Walsh[a,b,c,d,2]

[a]Division of Genetics and Genomics, Howard Hughes Medical Institute, and Manton Center for Orphan Disease Research, Boston Children's Hospital, Boston, MA, USA.

[b]Departments of Neurology and Pediatrics, Harvard Medical School, Boston, MA, USA.

[c]Broad Institute of MIT and Harvard, Cambridge, MA, USA.

[d]Program in Neuroscience and Harvard/MIT MD-PHD Program, Harvard Medical School, Boston, MA, USA.

[e]Department of Biomedical Informatics and Ludwig Center at Harvard, Harvard Medical School, Boston, MA, USA.

[f]Allen Institute for Brain Science, Seattle, WA, USA

[1]These authors contributed equally to this work.

[2]Corresponding Author

Christopher A. Walsh, MD. PhD

Division of Genetics and Genomics, Boston Children's Hospital

Center for Life Sciences 15062

3 Blackfan Circle, Boston, MA 02115

Phone: 617-919-2923

Email: Christopher.walsh@childrens.harvard.edu

Classification

Biological Sciences: Neuroscience

Keywords

PRDD-seq, single-cell MosaicHunter, birthdating, cortical layer, neurodevelopment

Author Contributions

A.Y.H. and P.L. conceived the project and C.A.W. supervised it. A.Y.H. and P.L. developed the scMH algorithm. P.L. developed PRDD-seq and performed experiments. A.Y.H. performed computational and statistical analyses. R.E.R., S.N.K., and Y.D. helped with validation of sSNVs. C.J.K. helped with and provided insight on the comparison of PRDD-seq and scWTA. S.K.A. assisted with interpretation of neurodevelopmental discoveries. R.D.H., T.E.B., J.M. and E.S.L. generated the MTG single-cell RNA sequencing data, and provided it prior to publication. E.A.L. and P.J.P. provided suggestions on computational analyses. A.Y.H. and P.L. wrote the manuscript supervised by C.A.W., with input from all other authors.

This PDF file includes:

Main Text

Figures 1 to 5

SI Appendix, Figure S1 to S5 and Tables S1 to S3

Abstract

Elucidating the lineage relationships among different cell types is key to understanding human brain development. Here we developed Parallel RNA and DNA analysis after Deep-sequencing (PRDD-seq), which combines RNA analysis of neuronal cell types with analysis of nested spontaneous DNA somatic mutations as cell lineage markers, identified from joint analysis of single cell and bulk DNA sequencing by single-cell MosaicHunter (scMH). PRDD-seq enables the first-ever simultaneous reconstruction of neuronal cell type, cell lineage, and sequential neuronal formation ("birthdate") in postmortem human cerebral cortex. Analysis of two human brains showed remarkable quantitative details that relate mutation mosaic frequency to clonal patterns, confirming an early divergence of precursors for excitatory and inhibitory neurons, and an "inside-out" layer formation of excitatory neurons as seen in other species. In addition our analysis allows the first estimate of excitatory neuron-restricted precursors (about 10) that generate the excitatory neurons within a cortical column. Inhibitory neurons showed complex, subtype-specific patterns of neurogenesis, including some patterns of development conserved relative to mouse, but also some aspects of primate cortical interneuron development not seen in mouse. PRDD-seq can be broadly applied to characterize cell identity and lineage from diverse archival samples with single-cell resolution and in potentially any developmental or disease condition.

66 **Significance Statement**

67 Stem cells and progenitors undergo a series of cell divisions to generate the neurons of the brain,
68 and understanding this sequence is critical to studying the mechanisms that control cell division
69 and migration in developing brain. Mutations that occur as cells divide are known as the basis of
70 cancer, but have more recently been shown to occur with normal cell divisions, creating a
71 permanent, forensic map of the clonal patterns that define the brain. Here we develop new
72 technology to analyze both DNA mutations and RNA gene expression patterns in single cells
73 from human postmortem brain, allowing us to define clonal patterns among different types of
74 human brain neurons, gaining the first direct insight into how they form.
75

Introduction

Although we have learned a great deal about development of the cerebral cortex from animal models, we have remarkably little direct information about how the human brain, which differs vastly in shape, size, and composition from the brains of non-primates, forms the neurons of its cerebral cortex (1-4). Recent studies defining the fundamental cell types of the adult and developing human cortex (5-7) form a foundation for understanding how these cell types develop, how the unique aspects of the human cortex come about, and how developmental brain disorders might alter patterns of cell lineage or cell type in human brain. However, whether individual neural progenitor cells (NPCs) in embryonic stages are restricted to produce certain subtypes of neurons, or multi-potential to generate all neuronal types, is still an open question even in model animal species, since making this distinction requires simultaneous identification of cell lineage and transcriptional analysis of cell type, which remains a technical challenge (8-12).

Somatic genetic mutations accumulate with each cell division during early development, when spontaneous DNA damage escapes the DNA repair machinery, with single-nucleotide variants (SNVs) being the most common mutation type (13-15). The timing of somatic mutations can be inferred by either the cell fraction that carries each mutation or the co-occurrence status of multiple mutations, in which early mutations should be shared by a large fraction of cells whereas later mutations should be present in nested subpopulations of cells (16). Previous study has shown the ability to use somatic SNVs as a rich internal lineage map to birthdate the developmental timing of each neurons differentiated from neuronal progenitor cells (14) but has not combined that with direct analysis of the subtypes of neurons, defined by morphology, location, physiology, or RNA transcription pattern.

Single-cell transcriptomes provide granular information about cell identity (5-7), but it cannot provide lineage maps as it fails to capture most somatic mutations, since somatic mutations occur throughout the genome, most often in intronic or intergenic regions (16, 17). Similarly, DNA-sequencing alone fails to provide information about cell identity, and so lineage mapping using only somatic mutations from DNA sequencing is unable to address questions about the lineage of specific cell identities in neurodevelopment. Somatic mutations in mitochondrial DNA have been recently suggested as potential lineage marks as well, but the modest target size of the mitochondrial genome, and the multiple diverse mitochondrial genomes in each cell, represent challenges to the use of mitochondrial mutations as a rich source of stable lineage markers (18).

To address this challenge, we developed Parallel RNA and DNA analysis after Deep-sequencing (PRDD-seq) that identifies somatic SNVs (sSNVs) from single cell and bulk whole-genome sequencing (WGS) data, with multiplexed detection of sSNVs and multiple RNA marker transcripts from single nuclei. We then benchmarked the performance of the DNA and RNA assays of PRDD-seq against bulk WGS and single-cell RNA sequencing (scRNAseq), respectively. Applying PRDD-seq to two postmortem brains of individuals without neurological disease allowed unprecedented quantitative analysis of cell lineage in the human brain. While revealing the expected patterns of divergence of excitatory and inhibitory lineages and “inside-out” generation of excitatory neurons, our PRDD-seq data also directly suggest complex patterns of interneuron formation in the human brain.

Results

Simultaneous cell type and lineage analysis of single-cells by PRDD-seq

The workflow of PRDD-seq is illustrated in Figure 1. Single NeuN+ cortical neuronal nuclei from prefrontal cortex (PFC) of postmortem human brain tissue were purified by fluorescence-activated nuclear sorting (FANS) (16) (Fig. 1A), and subjected to one-step RT-qPCR with target-specific primers for 1] cDNA specific for up to 30 marker genes of major neuronal cell types, and 2] specific genomic DNA (gDNA) loci representing identified somatic mutations (see below) as markers of cell lineage (Fig. 1B). Aliquots of the pre-amplified gDNA and cDNA libraries were analyzed for the presence of specific somatic mutations and transcripts by microfluidic genotyping and gene expression profiling, respectively, using the Fluidigm Biomark system (Fig. 1C). The somatic mutations used in PRDD-seq were identified by single-cell MosaicHunter (scMH), described below, a new bioinformatic tool to identify lineage-informative sSNVs, jointly considering WGS data from MDA-amplified single cells and matched deep (>200X) WGS from bulk DNA samples collected from the same brain region (Fig. 1D).

We first created a map of neuronal cell types by analyzing >25,000 single neuronal nuclei -- FANS-sorted based on NeuN immunoreactivity -- by scRNAseq from two different datasets, to create a cell type landscape onto which PRDD-seq analyzed neurons could be located. We performed 10X Genomics scRNAseq of 10,967 NeuN+ nuclei from the same PFC region of one of the brains from which DNA mutations were identified (Fig. 1E). t-SNE analysis of this dataset defined 21 transcriptionally distinct cell clusters, including 8 excitatory neuron clusters that further clustered into upper, middle, and lower layers, and 13 inhibitory neuron clusters that could be further classified into SST+, PV+, VIP+, and LAMP5+ subtypes (Fig. 1F and *SI*

Appendix, Fig. S1) (5, 7). A recently published scRNAseq dataset of 15,928 single neuronal nuclei from human middle temporal gyrus (MTG) (5), sorted by NeuN immunoreactivity following microdissection of cerebral cortical layers, provided additional direct information about layer location of neuronal types (Fig. 1G and *SI Appendix*, Fig. S2) and so was used for cell type mapping in parallel. PFC and MTG share relatively generic cerebral cortical architecture as "association" cortex, and clustering analysis of the two datasets (Fig. 1H) shows that they identified similar major cell types, with cells clustering by cell type rather than by platform, although the SMART-seq dataset from MTG defined finer subdivisions of cell type as expected because of its larger sample size and deeper sequence depth.

We jointly analyzed single PRDD-seq cells and scRNAseq cells and mapped each PRDD-seq cell onto the t-SNE maps of scRNAseq based on gene expression similarity (Fig. 1I, see Methods). The cell type and cortical layer information of each PRDD-seq cell was then imputed based on its assigned cluster in scRNAseq datasets. Finally, the combination of genotype and gene expression information of PRDD-seq cells allowed lineage and birthdate analysis of particular cell types (Fig. 1J), as well as analysis of cell type differentiation of particular lineages (Fig. 1K).

Discovery of lineage-informative sSNVs from bulk brain and single-neuron WGS data

The resolution of lineage reconstruction is dependent on having a comprehensive list of somatic mutations identified from the specific brain under analysis. Whereas deep WGS (e.g., 200-250X coverage) of "bulk" DNA, isolated from tissue, efficiently identifies sSNVs present in 4% or more cells (19), it is insensitive to detecting later-occurring sSNVs that mark late cell lineage events. On the other hand, WGS of DNA amplified from single neuronal nuclei (16)

identifies later-occurring sSNVs but is limited by cost and subject to artifacts during single-cell amplification. Therefore, we developed scMH, which incorporates a Bayesian graphic model (20, 21) that integrates analysis of bulk WGS and single-cell WGS data to distinguish somatic mutations from germline mutations and technical artifacts (Fig. 2A; see Methods). scMH first calculates the likelihood and mosaic fraction of candidate sSNVs from a bulk DNA sample, and then applies these values as the priors to genotype each candidate SNV across every single cell being analyzed. The shared presence of a given sSNV in bulk DNA and one or more single cells serves as validation of the sSNV. To expand the utility of scMH when a matched bulk sample is unavailable, we further designed a "bulk-free" mode that can utilize a "synthetic" bulk WGS dataset, generated by *in silico* merging of the many WGS datasets of multiple single-cells obtained from the same donor. We benchmarked scMH using 45X single-cell WGS of 24 neurons—22 of which were sequenced in previous studies (16, 17) —as well as ~200X bulk WGS of PFC (both from the brain of the same individual, UMB1465, who died at age 17 with no neurological diagnosis), against existing single-cell sSNV callers including Monovar (22), SCcaller (23), LiRA (24), and Conbase (25). Sensitivity and false discovery rate (FDR) were estimated based on experimentally validated mutations and clade annotations identified previously (16). With either PFC bulk or synthetic bulk, scMH outperformed the other tools and achieved ~70% sensitivity to detect lineage-informative mutations with < 5% FDR; combining both the default and "bulk-free" modes improved detection sensitivity to 93% without increasing the FDR, suggesting that the "bulk-free" mode of scMH can detect sSNVs that are present in multiple single-cells but may be undetectable in the bulk 200X WGS samples because of the low mosaic fraction of these late mutations (Fig. 2B).

Applying scMH to data from brains of three normal individuals (UMB1465, UMB4638, and UMB4643 (16, 17), identified and validated 42, 19, and 22 sSNVs, respectively (Fig. 2C-E, and *SI Appendix*, Table S1), with an overall validation rate of 74.8% determined by Sanger sequencing of independently sorted neurons from the same brain region. The number and validation rate of lineage-informative sSNVs detected by scMH dramatically increased from previous studies (16, 17). sSNVs identified from all three brains showed an enrichment in C>T mutations, especially in CpG sites (*SI Appendix*, Fig. S3), a pattern observed in other studies of embryonic mutations and cancer mutations (13, 26), since such C>T mutations appear to be caused by cytosine deamination that is replicated into a fixed SNV before it can be repaired (27). Unsupervised clustering analysis grouped the 24 sequenced neurons from UMB1465 into six different clades; no cells harbored mutations of multiple clades, suggesting the high accuracy of scMH for single-cell genotyping of sSNVs (Fig. 2C). In clades C and E, we observed neurons that shared early mutations but harbored different sets of later mutations, suggesting that they were derived from different branches of the same clades (Fig. 2C). Clustering of ten and nine sequenced neurons from UMB4638 and UMB4643—respectively by their sSNVs—demonstrated similar nested patterns forming three primary clades for each individual and also showed evidence for branches of these clades (Fig. 2D, E). The mosaic fraction of each sSNV in "bulk" DNA (Fig. 2C, D, E) was used as an additional indicator of the sequence in which sSNV occurred, since early sSNVs tend to be found in many single cells, as well as at higher mosaic fraction in bulk DNA, whereas later mutations appear in fewer cells and lower mosaic fraction in bulk DNA. These two findings correlated very strongly.

Lineage and cell type identity of single-neurons revealed by PRDD-seq

To assess the performance of PRDD-seq in capturing lineage and cell type information from single-cells, we applied PRDD-seq to 1,710 cortical neurons from UMB1465 PFC, using probes to detect 30 out of 42 validated sSNVs in UMB1465, for which we successfully designed highly specific and sensitive probes (*SI Appendix*, Table S1), along with 30 marker genes whose expression levels distinguish major inhibitory and excitatory neuronal subtypes and cortical layers identified in the scRNAseq datasets (5, 7) (*SI Appendix*, Table S2). Overall, PRDD-seq mapped 1,112/1,710 (65%) cortical neurons from UMB1465 PFC into 20 lineage branches and 6 major clades (Fig. 3A). For each major clade, birthdate-ordered lineage branches were inferred from the nested sSNVs, where earlier derived neurons contained fewer clonal mutations, and neurons generated later harbored additional mutations from subsequent cell divisions (16). The nested nature of sSNVs in clades allow cells to be placed into clades using multiple sSNVs, so that cells whose genomes were subject to allelic dropout—which is not uncommon when single cell DNA molecules are amplified—could still be placed into clades based on other sSNV from the same clade (Fig. 3A and *SI Appendix*, Table S1). On the other hand, only 71/1710 (4.2%) neurons contained sSNVs from multiple clades, suggesting a low rate of false positive amplification or sorting of multiple nuclei into single wells in the DNA assay of PRDD-seq (Fig. 3B, upper panel). 527/1710 (30.8%) neurons showed the absence of any sSNVs from the 6 clades; these neurons may be from other clades in which we did not discover sSNV markers (Fig. 3B, upper panel). In PRDD-seq cells, mosaic fractions of sSNVs correlated linearly with the fractions calculated from ~200X bulk WGS, indicating generally unbiased sSNV detection (Fig. 3B, lower panel and Fig. 3C), and allowing confident inference of the developmental sequence of sSNVs according to the nested pattern.

Among the 1,112 PRDD-seq cells that were successfully claded, we ran the RNA assay of PRDD-seq to measure the expression of 30 marker genes for each cell. Our evaluation using simulation data derived from our own and published scRNAseq datasets (see Methods) suggested that these 30 marker genes were sufficiently informative to infer many aspects of cell type and dissected layer annotation (Fig. 3D), with an average accuracy of 84% for cortical layer classification (within +/- one-layer difference) and 83% for inhibitory neuron subtype classification. We then utilized expression of these 30 makers to successfully classify 747/1,112 PRDD-seq neurons (67.2%) from UMB1465 into 3 excitatory subgroups—corresponding to upper, middle, or lower cortical layers—and 4 inhibitory subgroups: somatostatin positive (SST+), vasoactive intestinal peptide-positive (VIP+), lysosomal associated membrane protein 5-positive (LAMP5+), and putative parvalbumin-positive (putative PVALB+, or pPVALB+), since probes for PVALB were not always directly assayed (Fig. 3E). PRDD-seq cells assigned to upper, middle, and lower layers by the 10X PFC scRNAseq dataset were also enriched in L2-L3, L4-L5, and L6 markers according to the SMART-seq MTG scRNAseq dataset, respectively, indicating the similarity of the cell type compositions between PFC and MTG, the similarity of the results with the two RNAseq methods, as well as the robustness of the mapping algorithm (Fig. 3E, upper panel). Both our 10X scRNAseq dataset and PRDDseq analysis of UMB1465 and UMB4638 showed higher proportions of inhibitory neurons (43-47%) than reported with other methods, however this ratio was very similar between the three experiments, suggesting that the ratio reflects our particular NeuN+ sorting protocol rather than technical aspects of the cell typing methods (Fig. 3F upper panel). We observed remarkably similar layer and subtype distribution between PRDD-seq and scRNAseq cells for excitatory neurons (Chi-square test; Fig. 3F, middle panel). Among inhibitory neurons, pPVALB+ inhibitory neurons showed a higher

proportional representation in PRDD-seq than in scRNAseq, suggesting that a few neurons in this category might reflect amplification failure of the other inhibitory probes (SST, VIP, and LAMP5). In summary, our analysis suggests that PRDD-seq captures the major aspects of cell types, without systematic loss of any given cell type.

Early divergence of progenitors for excitatory and inhibitory neurons

The simultaneous analysis of lineage and gene expression from the same neurons enabled us to study the change of cell type contribution during early neurogenesis. Using PRDD-seq, we profiled >2700 neurons from two brains, UMB1465 and UMB4638, and successfully captured both lineage and cell type information from 747 and 480 neurons, respectively. In both UMB4638 and UMB1465, all lineage clades showed early sSNVs in both excitatory and inhibitory neurons, reflecting mutations occurring during early embryogenesis before the divergence of these cell types, whereas late SNVs show progressive restriction to one or the other cell type (Fig. 4A, B). Among the six major clades in UMB1465, clade C contained seven nested branches with mosaic fractions diminishing from 0.33 to 0.0067 (Figure 3A and *SI Appendix*, Table S1), with an increasing percentage of excitatory neurons containing mutations C1 to C5, and only excitatory neurons containing mutations C6 to C7 (Fig. 4A), while clade F showed similar progressive restriction. Similarly, both clade A and B in UMB4638 showed nested mutations that became progressively limited to excitatory neurons (Fig. 4B). Interestingly, the excitatory neurons appeared exclusively in branches with mosaic fraction below ~0.04 (Fig. 4A, B, and *SI Appendix*, Table S1), corresponding to a progenitor giving rise to about 4% of the total cells in that cortical sample. Considering that ~40% of cortical cells are excitatory neurons, with the remainder being glial cells or inhibitory neurons (28, 29), this observation suggests that

ten or more excitatory neuronal progenitor cells (NPCs) generate excitatory neurons in a given cortical area, or "column"; the fact that 6-7 (including a branched clade) excitatory precursors are explicitly marked by non-overlapping clades, and account for 60-70% of excitatory neurons in our sample, independently supports this estimate. On the other hand, two clades (clade A and B) from UMB1465 are statistically enriched for inhibitory neurons (two-sided one-proportion Z-test's $P < 0.05$), with the percentage of inhibitory neurons increasing from B1 to B2 (Fig. 4A). These results show that at least some human NPCs demonstrate restricted cell type output, supporting the model first established in mice (30-32) and strongly supported by conserved gene expression patterns in the ganglionic eminence between humans and non-humans (33, 34), that excitatory and inhibitory neurons are generated from distinct progenitor regions.

"Inside-out" order of cortical layer formation for excitatory neurons

Further sub-typing of excitatory neurons using laminar markers revealed layer-specific patterns of excitatory neuron neurogenesis. For example, in UMB1465, the percentage of lower layer neurons carrying a mutation decreased from mutations C1 to C4, and no deep-layer neurons were detected carrying C5 to C7, with the percentage of upper layer neurons increasing correspondingly from C1 to C7 (Pearson correlation's $P = 2.9\times10^{-3}$; Fig. 4C, upper panel). To gain more precise layer identities of PRDD-seq cells, we mapped them to the SMART-seq MTG scRNAseq dataset obtained after layer microdissection using the same methods as earlier (5), which generated similar "birthdate" patterns in clade C, with early lineage sSNVs present in all layers, and later sSNVs restricted to middle and upper layers (Pearson correlation's $P = 1.4\times10^{-3}$; Fig. 4C, lower panel). A similar trend was also observed in clades A and B in UMB4638. Mapping PRDD-seq cells of UMB4638 to both 10X PFC and SMART-seq MTG scRNAseq

datasets showed that cells with later lineage markers were restricted to middle and upper layers (Fig. 4D). These results together directly indicate that human cortical excitatory neurons are formed in "inside-out" sequence after preplate cells are born, similar to mouse and non-human primates (35-37). Furthermore, it suggests that neurons in lower cortical layers begin becoming postmitotic relatively quickly after progenitors are specialized for excitatory neuron production.

Diverse spatiotemporal patterns of development of inhibitory neuron subtypes

Mapping PRDD-seq cells onto two different scRNAseq datasets also allowed analysis of cortical inhibitory neurons, which originate from multiple developmentally transient structures of the ventral telencephalon, including the medial, lateral and caudal ganglionic eminences (MGE, LGE, and CGE), and migrate into dorsal cortex (30, 38). However, the highly dispersed nature of inhibitory neuron clones observed in animal models (39-41) suggests that sSNVs in the inhibitory lineage are likely to be present at exceedingly low allele frequencies in bulk DNA and tiny fractions of single cells, so that only sSNVs occurring relatively early in development have been analyzed so far. Inhibitory neurons derived from MGE and CGE can be distinguished by expression of specific markers (5, 6), and PRDD-seq analysis showed that interneurons with diverse marker genes were generated over the same developmental window (Fig. 5A, B). The analyzed sSNVs were shared by multiple inhibitory subtypes, with hints that late marks might be more limited to cell types, but no differences that reached statistical significance (FDR-adjusted Chi-square test's $P > 0.05$). Previous studies cataloging interneurons in mouse and human have suggested that MGE-derived inhibitory neuron subtypes (SST+ and PVALB+) are enriched in infragranular cortical layers, while CGE-derived interneuron subtypes (LAMP5/PAX6+, VIP+) tend to occupy upper cortical layers preferentially (5, 42, 43) and thus our mapping of PRDD-seq

cells onto scRNAseq reflected these patterns. Birthdating analyses in mice and non-human primates have reached contradictory conclusions about whether inhibitory neurons follow inside-out patterns of generation similar to excitatory neurons (44, 45), though recent analyses in mice suggest that previous contradictions may reflect the convolution of multiple patterns of generation that may be subtype specific (46). We found that MGE-derived pPVALB+ subtype neurons, enriched in layer IV-VI, showed if anything a trend for the latest-generated neurons to show markers of deeper layers (Fig. 5C, D). SST+ neurons, widely distributed in layer II-VI, similarly did not show an inside-out pattern detectable with the mutations and cells analyzed (Fig. 5C, D). We robustly detected SST+ neurons with expression of layer I markers in human PFC (SST-like subclass) (Fig. 5C, D), consistent with observations in MTG (5, 47) and in mice, where such layer I SST+ expressing cells are rare but present (43, 47). These upper layer, CGE-derived SST-like cells are a subclass of LAMP5+ interneurons that are more transcriptionally related to VIP neurons than MGE derived SST+ interneurons, though they lack VIP expression (5, 47). Our data further confirm that LAMP5+ interneurons express markers suggesting broad laminar location, but also did not reveal a simple inside-out progression of formation (5). Interestingly, we observed a substantial proportion of LAMP5+ inhibitory neurons, particularly the SST-like class, labeled by later mutations, indicating that this subtype may be generated later during development than other inhibitory cell types (Fig. 5C, D). Overall, our findings suggest little evidence of the inside-out patterns of neurogenesis demonstrated by excitatory neurons, but also show that detailed analysis of interneurons will likely require deep datasets of sSNV occurring at late stages of interneuron development, and higher-throughput methods of analysis.

Discussion

We have developed scMH and PRDD-seq that allowed us, to our knowledge, the first simultaneous analysis of cell lineage and transcriptional cell type in human brain—and potentially, any mammalian brain—through improved identification of sSNVs in deep bulk and single-cell sequencing data. Our analysis of a single cortical area (PFC) in two individual brains revealed some conserved patterns of cell lineage compared to nonhumans, including that inhibitory and excitatory neurons diverge early in humans, and that excitatory neurons form following a similar "inside-out" order as seen in the animal models. However, PRDD-seq also provides the first quantitative estimate in any species of number of progenitor cells (approximately 10) that generate the excitatory neurons in a given cortical area. Furthermore, PRDD-seq also provided some direct insight into inhibitory neuron development in humans, supporting parallel development of different subtypes of inhibitory neurons, with spatial and temporal associations specific only to some subtypes. Our data show that, as methods improve to capture sSNVs present in small numbers of cells, the natural occurrence of sSNVs with each cell division (13, 14, 17) is likely sufficient to provide a very rich map of cell lineage patterns in any given postmortem human brain.

The human cerebral cortex has been thought to contain approximately 80% excitatory glutamatergic neurons and 20% GABAergic interneurons (48), although recent scRNAseq studies have reported a somewhat lower ratio of about 70% excitatory neurons (*SI Appendix*, Table S3) (5, 49, 50). Although our PRDD-seq analysis showed 661 excitatory versus 566 inhibitory PRDD-seq cells in total for UMB1465 and UMB4638, which represents 54% excitatory neurons (*SI Appendix*, Table S3), this higher proportion of inhibitory neurons seems to reflect either aspects of the tissue (which was stored for long periods frozen), or our NeuN+- sorting method, since similar ratios are seen in 10X scRNAseq from the one brain analyzed (*SI*

Appendix, Table S3). On the other hand, PRDD-seq cells are studied as containing at least one sSNV identified from scMH using a small number of deeply sequenced neuronal nuclei isolated from the same region, and so do not represent an unbiased sampling of the human brain region. Nonetheless, the fact that we can assign 60-70% of all excitatory neurons to clades in UMB1465, and that neurons with identified SNVs represent most major neuronal types in scRNAseq (Fig. 3E), suggests that our sampling has captured the majority of the lineage of the cortical patch, although rare lineages are likely to be missed without much deeper sequencing. Moreover, the presence of 6-7 explicitly marked clades, and the ability to correlate the allele frequency of a sSNV to the excitatory-restriction of the cells carrying that sSNV, allows two independent quantitative assessments of how many progenitors (approximately 10) contribute to the neurons of the patch of cortex from which neurons were isolated, illustrating the remarkable quantitative potential of this approach.

Since occasional dropout of DNA marks and RNA markers in PRDD-seq is unavoidable, limited by the quality of isolated nuclei, we emphasize that our results are most robust when analyzing cells positive for both. The quality of postmortem brain tissues can influence the integrity of both genomic DNA and mRNA. Regarding DNA, since no whole-genome amplification is performed prior to targeted pre-amplification, only a single molecular copy of each allele is available for genotyping of each sSNV, so occasional dropout is inevitable. However, our lineage strategy is based not only on the presence of clade-specific sSNVs but also the absence of many sSNVs from other clades (Fig. 3A), so the chance for mis-assigning cells should be relatively small. Nevertheless, mapping our sSNVs onto our scRNAseq dataset suggests that lineage marks are present in the major neuronal subtypes, although rare neuronal types are likely to be missed given our modest sample size. Regarding RNA, single nuclei from

postmortem human brain contains only a small amount of mRNAs. Fluidigm Biomark assays are microfluidics-based qPCR assays that are sensitive to subtle changes of the input or environment. As a result, we observed a 30.4% dropout rate of DNA markers and similar level of dropout of RNA marker dropout. However, since PRDD-seq analyses excluded these dropout events, and were completely based on the relative cell type proportions across different stages within one lineage, we have no reason to think that the dropouts are systematic with respect to cell type with one exception: the relatively larger proportion of pPVALB+ neurons in PRDD-seq than scRNAseq, likely reflecting the failure of some probes for SST, VIP, and LAMP5. Better and richer probe sets are likely to be able to resolve this in the future.

There are limitations to our analysis, since we are analyzing a small sample of the vast size of the human brain, and PRDD-seq is relatively low-throughput and expensive, so our initial analysis only can make conclusions about relatively common cell types. The present analysis is somewhat limited in the analysis of late mutations present in 1% of cells, especially interneurons, since it is challenging to detect those mutations with great sensitivity, but will await single-cell studies on subtypes of neurons in the future. On the other hand, the combined analysis of sSNVs and cell types is archival and progressive. The vast size of the human brain means that each subsequent round of DNA sequencing—whether of bulk tissue or of single or pooled cells—adds to the total depth of sequence data, and provides progressively richer information about late sSNVs. Indeed, the likely dispersed nature of inhibitory clones suggests that analyzing one cortical region could provide sequence data useful in the analysis of a completely different cortical region for these cell types.

Overall, PRDD-seq has many advantages even beyond the quantitative analysis of lineages and mosaic fractions that we begin to illustrate here. Since the method uses sSNVs as

lineage marks, it is inherently genomic and so allows correlation not only of normal developmental patterns, but would immediately capture alterations to lineage patterns caused by function-altering germline or somatic mutations. In addition, since sSNVs serve as *in vivo* cellular markers for drawing a developmental lineage map without any transgenic manipulation as demonstrated in this study, the method promises to be applicable in principle to any species or human disease condition for which post-mortem brain is available.

Materials and Methods

Human tissues whole-genome sequencing

Frozen post-mortem tissues from three neurologically normal individuals, UMB1465 (a 17-year-old male), UMB4638 (a 15-year-old female), and UMB4643 (a 42-year-old female), were obtained from the NIH NeuroBioBank at the University of Maryland, and prepared according to a standardized protocol (http://medschool.umaryland.edu/btbank/method2.asp) under the supervision of the NIH NeuroBioBank ethical guidelines. UMB1465 and UMB4638 died of injuries sustained in motor vehicle accidents, while UMB4643 died of cardiovascular disease. Bulk DNA samples and single neuronal nuclei amplified by multiple displacement amplification (MDA) were prepared and whole-genome sequenced by Illumina HiSeq platforms as part of previous studies in our lab (16). The average sequencing depth was about 40X for single neurons and about 200X for bulk brain samples.

Estimation of cell-specific dropout rate and error rate

Germline heterozygous mutations were called by GATK HaplotypeCaller (51) from the whole-genome sequencing data from bulk brain DNA samples, and only common SNPs annotated in

the 1000 Genome Project (52) were considered to reduce false positive calls. To estimate cell-specific allele dropout rate, we calculate the proportion of germline heterozygous sites that were genotyped as reference- or alternative-homozygous in single-cell sequencing data. One neuron of UMB4643 with significantly lower allele dropout rate (Z-score < -2) was excluded from subsequent analyses, since it likely represented a doublet from FANS sorting. Similarly, we also extracted the reference-homozygous sites at the 3' adjacent position of each germline heterozygous mutation and calculate the proportion of heterozygous and alternative-homozygous genotypes to estimate the genome-wide error rate in each single-cell.

Framework of single-cell MosaicHunter

The overall framework of single-cell MosaicHunter (scMH) was illustrated in Fig. 2A. sSNV candidates were first called from the bulk sequencing data using a Bayesian graphical model (20, 21), in which the likelihoods of somatic mutation and three genotypes of inherited mutation were calculated with the consideration of binomial sampling variation and base-calling errors (Fig. 2A, left panel). The presence or absence of somatic mutation in each single-cell was then inferred by adapting the likelihood and allele fraction (*f*) of somatic mutation estimated from bulk sample as prior probability, after controlling the cell-specific allele dropout rate (*d*) and error rate (*e*) (Fig. 2A, right panel). Specifically, the transition matrix between bulk and single-cell genotypes was developed as below,

$$\mathrm{P}(G_{sc} \mid G_{bulk}) = \begin{pmatrix} 1 & 0 & 0 & 1-f \\ 0 & 1 & 0 & f \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

where each column denotes reference-homozygous, heterozygous, alternative-homozygous, and mosaic genotype for bulk sequencing, and each row denotes reference-homozygous,

460 heterozygous, and alternative-homozygous genotype for single-cell sequencing. The genotype
461 likelihoods in single-cells were further adjusted for allele dropout rate (*d*) and error rate (*e*) as
462 below,

463 $$\mathrm{P}(G_{post} \mid G_{pre}) = \begin{pmatrix} 1-2e+e^2 & d(1-d) & e^2 \\ 2e(1-e) & 1-2d+2d^2 & 2e(1-e) \\ e^2 & d(1-d) & 1-2e+e^2 \end{pmatrix}$$

464 where each column and row denotes reference-homozygous, heterozygous, and alternative-
465 homozygous genotype before and after adjustment for single-cell sequencing. Single-cell
466 genotypes were binarized as mutant or wildtype by comparing the posterior probability of
467 heterozygous genotype to an empirical threshold. For each candidate site, the proportion of
468 mutant cells was calculated to further filter out germline mutations. Candidate sites with >50%
469 cells showing aberrant single-cell allele fractions were also removed to exclude hotspots of
470 technical artifacts. In “bulk-free” mode with synthetic bulk generated from in silico merging
471 sequencing data from multiple single-cells, scMH would only consider sSNVs which were
472 shared by at least two single-cells.
473
474 **Somatic SNV calling and performance comparison**
475 Paired-end reads from bulk and single-cell whole-genome sequencing data were aligned to the
476 GRCh37 human reference genome by BWA (53), and then processed by GATK (51) and Picard
477 (http://broadinstitute.github.io/picard/) for the removal of duplicated and error-prone reads, indel
478 realignments, and base-quality recalibrations. sSNVs in neurons of UMB1465, UMB4638, and
479 UMB4643 were called by scMH and four other tools including Monovar (22), SCcaller (23),
480 LiRA (24), and Conbase (25). Sensitivity was estimated as the detected proportion of lineage-
481 informative sSNVs that had been previously identified and validated in these three brain samples

(16). False discovery rate (FDR) was measured as the proportion of lineage-informative mutations that were shared by cells from conflicting clades (16).

Validation of somatic SNVs

Validation of somatic SNVs called from scMH was performed using PCR of 200-500 bp amplicons including the mutated base, followed by Sanger sequencing. All variants were validated in independently sorted single neuronal nuclei amplified by MDA.

Generation of simulated single-cell whole-genome sequencing data

To estimate the sensitivity of scMH to detect lineage-informative sSNVs, bulk and single-cell sequencing data with varied somatic mutation rates was generated *in silico* (*SI Appendix*, Fig. S4A). First, we developed a simplistic model to mimic the process of early embryogenesis: 1) ten rounds of symmetric cell division was applied to generate 1024 (2^{10}) daughter cells derived from a single zygote, in which somatic mutations was randomly introduced at a rate of 1, 2, 5, or 10 mutations per round; 2) each daughter cells accumulated cell-specific somatic mutations for another ten rounds with the same mutation rate. Then, for each daughter cell, sequencing reads of chromosome 1 was generated at 40X by ART (54) with default parameters for Illumina platforms, and then germline mutations identified from NA12878 and somatic mutations generated by our model was introduced to the sequencing read using BAMSurgeon (55), with an allele dropout rate of 1×10^{-2} per base and MDA amplification rate of 1×10^{-7} per base that were estimated from real single-cell sequencing data. Finally, we randomly selected 80 cells (consistent with the detection threshold of scMH in real brain bulk samples) from the 1024 daughter cells and merged their sequencing data with a down-sampling of 200X to generate the

bulk sequencing data, and another 16 cells was randomly selected for benchmarking the performance of scMH. Our simulation data suggested that scMH was able to detect, on average, 67% and 86% of cell-shared sSNVs with PFC bulk or synthetic bulk, respectively (*SI Appendix*, Fig. S4B).

Design and selection of Taqman genotyping and gene expression probes

Taqman genotyping probes for all validated sSNVs were designed using custom Taqman assay design tool provided Thermo Fisher Scientific. Off-the-shelf Taqman gene expression probes were ordered from Thermo Fisher Scientific. All designed probes were tested by ddPCR using human genomic DNA (Human male, Promega) as a negative control. Gene expression probes were further tested by isolated bulk brain RNA as a positive control. Genotyping probes were also tested by comparing the detected mosaic fractions and the fractions calculated from bulk sequencing (Fig. 3C, D).

Parallel RNA and DNA analysis after Deep-sequencing (PRDD-seq)

Single nuclei from postmortem brain samples were isolated using fluorescence-activated nuclear sorting (FANS) for NeuN as described previously (56). Isolated single neuronal nuclei were directly sorted into CellsDirect One-Step qRT-PCR (Thermo Fisher Scientific) pre-amplification buffers containing 0.14x Taqman gene expression assays and SNP genotyping assays. Pre-amplification of all cDNA and genomic DNA amplicons were performed directly after the FANS sorting. Following pre-amplification, samples were diluted 10-fold and loaded onto 96.96 genotyping or 192.24 gene expression dynamic assay integrated fluidic circuits for standard amplification per manufacturer's instructions (Biomark, Fluidigm). Genotype and gene

528 expression were further determined by Biomark machine and analyzed by Biomark & EP1
529 software (Fluidigm).
530
531 **10X Genomics preparation and sequencing**
532 Standard 10X Genomics Chromium 3' (v2 chemistry) was carried out according to the
533 manufacturer's recommendation. Single nuclei from postmortem brain samples were isolated
534 using FANS for NeuN, and were loaded onto a 10X Genomics Chromium chip. Reverse
535 transcription and library preparation was performed using the 10X Genomics Single Cell v2 kit
536 following the 10X Genomics protocol. The library was then sequenced on one lane of Illumina
537 NextSeq-500 with a high-output kit.
538
539 **Single-cell RNA sequencing analysis**
540 The expression matrix of 10X Genomic single-cell RNA sequencing (scRNAseq) was generated
541 by Cell Ranger following the recommended protocols. The expression matrix and cell
542 annotations of SMART-seq-based scRNAseq for human MTG (5) was downloaded from the
543 website (https://celltypes.brain-map.org/rnaseq/). Variance normalization, clustering and
544 visualization were performed by Pagoda2 (57) using a similar protocol to Lake et al (7). Cell
545 clusters containing more than 50 cells were plotted on the t-SNE map, and the annotation of
546 cortical layer (upper, middle, lower) for excitatory neurons and subtypes for inhibitory neurons
547 was manually curated for each cluster according to the expression level of marker genes (*SI*
548 *Appendix*, Fig. S1 and S2). Considering that Layer 1 dissections of MTG nuclei included the
549 upper part of Layer 2 and the absence of excitatory neurons in the Layer 1 of MTG based on in
550 situ labeling (5), all the MTG Layer 1 excitatory neurons were re-annotated as Layer 2. To

551 further compare the expression profile of cells clusters between two scRNAseq datasets, we
552 calculated the cosine similarity of average expression level for marker genes (*S*) between any
553 two cell clusters. Cell clusters were then hierarchically clustered using the Ward's method with a
554 distance of 1 – *S*.
555
556 **Joint analysis of PRDD-seq and scRNAseq cells**
557 To understand the cell type and cell origin of PRDD-seq cells, we utilized their gene expression
558 profiles to map them onto the t-SNE maps of scRNAseq. PRDD-seq cells were firstly separated
559 into excitatory or inhibitory neurons according to the expression of excitatory or inhibitory
560 marker genes (*SI Appendix*, Table S2), and cells with no or conflicting expression of these
561 marker genes were excluded. For excitatory neurons, missing expression status for layer marker
562 genes (*SI Appendix*, Table S2) were inferred if any layer-specific genes for a given layer were
563 expressed. The cosine similarity matrix was then generated by comparing PRDD-seq cells
564 against scRNAseq cells. For each PRDD-seq cell, its cell cluster was determined by the majority
565 voting among its 25-nearest scRNAseq cells in cosine similarity (Fig. 1I), and the cell type and
566 cortical layer information of PRDD-seq cell was further annotated based on their assigned cell
567 cluster in scRNAseq datasets. To benchmark how accurately we could infer cell type and layer
568 annotation from the 30 marker genes profiled in PRDD-seq cells, we randomly sampled 200
569 scRNAseq cells from each of the seven cell types (upper, middle, lower layer excitatory neurons
570 and VIP+, SST+, LAMP5+, pPVALB+ inhibitory neurons) from 10X Genomic dataset and each
571 of the six dissected cortical layers from SMART-seq dataset, and only extracted the expression
572 profiles of 30 marker genes from each scRNAseq cell. Using the same majority voting strategy,
573 we assigned them back to cell clusters on the t-SNE map. As shown in Fig. 3D, the majority of

574 the randomly sampled cells can be correctly assigned to their original cell type and layer
575 annotation, suggesting the accuracy of our mapping strategy in PRDD-seq.
576
577 **Quantification and statistical analysis**
578 All data are reported as mean ± 95% confident interval (CI) unless mentioned otherwise. All of
579 the statistical details can be found in the figure legends, figures, and Results. Significance was
580 defined for p values smaller than 0.05. All tests were performed using the R software package
581 (version 3.5.0).
582
583 **Data and code availability**
584 Sequencing data was deposited in the NCBI SRA with accession numbers SRP041470 and
585 SRP061939. MosaicHunter is publicly available at http://mosaichunter.cbi.pku.edu.cn/. Config
586 files of single-cell MosaicHunter (scMH) and other scripts about PRDD-seq can be accessed at
587 https://github.com/AugustHuang/PRDD-seq.
588

References

1. P. Rakic, Evolution of the neocortex: a perspective from developmental biology. *Nat Rev Neurosci* **10**, 724-735 (2009).
2. D. H. Geschwind, P. Rakic, Cortical evolution: judge the brain by its cover. *Neuron* **80**, 633-647 (2013).
3. M. Heide, K. R. Long, W. B. Huttner, Novel gene function and regulation in neocortex expansion. *Curr Opin Cell Biol* **49**, 22-30 (2017).
4. C. S. Raju *et al.*, Secretagogin is Expressed by Developing Neocortical GABAergic Neurons in Humans but not Mice and Increases Neurite Arbor Size and Complexity. *Cereb Cortex* **28**, 1946-1958 (2018).
5. R. D. Hodge *et al.*, Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61-68 (2019).
6. S. Zhong *et al.*, A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524-528 (2018).
7. B. B. Lake *et al.*, Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36**, 70-80 (2018).
8. A. McKenna *et al.*, Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
9. K. L. Frieda *et al.*, Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107-111 (2017).
10. B. Raj *et al.*, Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat Biotechnol* **36**, 442-450 (2018).
11. B. Spanjaard *et al.*, Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat Biotechnol* **36**, 469-473 (2018).
12. A. Rodriguez-Meira *et al.*, Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. *Mol Cell* **73**, 1292-1305 e1298 (2019).
13. Y. S. Ju *et al.*, Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714-718 (2017).
14. T. Bae *et al.*, Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science* **359**, 550-555 (2018).
15. S. De, Somatic mosaicism in healthy human tissues. *Trends Genet* **27**, 217-223 (2011).
16. M. A. Lodato *et al.*, Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94-98 (2015).
17. M. A. Lodato *et al.*, Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555-559 (2018).
18. L. S. Ludwig *et al.*, Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* **176**, 1325-1339 e1322 (2019).
19. Y. Dou *et al.*, Accurate detection of mosaic variants in sequencing data without matched controls. *Nat Biotechnol* **38**, 314-319 (2020).
20. A. Y. Huang *et al.*, MosaicHunter: accurate detection of postzygotic single-nucleotide mosaicism through next-generation sequencing of unpaired, trio, and paired samples. *Nucleic Acids Res* **45**, e76 (2017).
21. A. Y. Huang *et al.*, Postzygotic single-nucleotide mosaicisms in whole-genome sequences of clinically unremarkable individuals. *Cell Res* **24**, 1311-1327 (2014).

22. H. Zafar, Y. Wang, L. Nakhleh, N. Navin, K. Chen, Monovar: single-nucleotide variant detection in single cells. *Nat Methods* **13**, 505-507 (2016).
23. X. Dong *et al.*, Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat Methods* **14**, 491-493 (2017).
24. C. L. Bohrson *et al.*, Linked-read analysis identifies mutations in single-cell DNA-sequencing data. *Nat Genet* **51**, 749-754 (2019).
25. J. Hard *et al.*, Conbase: a software for unsupervised discovery of clonal somatic mutations in single cells through read phasing. *Genome Biol* **20**, 68 (2019).
26. A. Y. Huang *et al.*, Distinctive types of postzygotic single-nucleotide mosaicisms in healthy individuals revealed by genome-wide profiling of multiple organs. *PLoS Genet* **14**, e1007395 (2018).
27. T. Helleday, S. Eshtad, S. Nik-Zainal, Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* **15**, 585-598 (2014).
28. C. S. von Bartheld, J. Bahney, S. Herculano-Houzel, The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting. *J Comp Neurol* **524**, 3865-3895 (2016).
29. H. Markram *et al.*, Interneurons of the neocortical inhibitory system. *Nat Rev Neurosci* **5**, 793-807 (2004).
30. S. A. Anderson, D. D. Eisenstat, L. Shi, J. L. Rubenstein, Interneuron migration from basal forebrain to neocortex: dependence on Dlx genes. *Science* **278**, 474-476 (1997).
31. G. Fishell, C. A. Mason, M. E. Hatten, Dispersion of neural progenitors within the germinal zones of the forebrain. *Nature* **362**, 636-638 (1993).
32. S. Anderson, M. Mione, K. Yun, J. L. Rubenstein, Differential origins of neocortical projection and local circuit neurons: role of Dlx genes in neocortical interneuronogenesis. *Cereb Cortex* **9**, 646-654 (1999).
33. T. Ma *et al.*, Subcortical origins of human and monkey neocortical interneurons. *Nat Neurosci* **16**, 1588-1597 (2013).
34. D. V. Hansen *et al.*, Non-epithelial stem cells and cortical interneuron production in the human ganglionic eminences. *Nat Neurosci* **16**, 1576-1587 (2013).
35. P. Rakic, Neurons in rhesus monkey visual cortex: systematic relation between time of origin and eventual disposition. *Science* **183**, 425-427 (1974).
36. J. B. Angevine, Jr., R. L. Sidman, Autoradiographic study of cell migration during histogenesis of cerebral cortex in the mouse. *Nature* **192**, 766-768 (1961).
37. P. Gao *et al.*, Deterministic progenitor behavior and unitary production of neurons in the neocortex. *Cell* **159**, 775-788 (2014).
38. C. Mayer *et al.*, Developmental diversification of cortical inhibitory interneurons. *Nature* **555**, 457-462 (2018).
39. M. Turrero Garcia, E. Mazzola, C. C. Harwell, Lineage Relationships Do Not Drive MGE/PoA-Derived Interneuron Clustering in the Brain. *Neuron* **92**, 52-58 (2016).
40. K. N. Brown *et al.*, Clonal production and organization of inhibitory interneurons in the neocortex. *Science* **334**, 480-486 (2011).
41. C. B. Reid, S. F. Tavazoie, C. A. Walsh, Clonal dispersion and evidence for asymmetric cell division in ferret cortex. *Development* **124**, 2441-2450 (1997).
42. M. J. Nigro, Y. Hashikawa-Yamasaki, B. Rudy, Diversity and Connectivity of Layer 5 Somatostatin-Expressing Interneurons in the Mouse Barrel Cortex. *J Neurosci* **38**, 1622-1633 (2018).

43. B. Rudy, G. Fishell, S. Lee, J. Hjerling-Leffler, Three groups of interneurons account for nearly 100% of neocortical GABAergic neurons. *Dev Neurobiol* **71**, 45-61 (2011).
44. E. S. Ang, Jr., T. F. Haydar, V. Gluncic, P. Rakic, Four-dimensional migratory coordinates of GABAergic interneurons in the developing mouse cortex. *J Neurosci* **23**, 5805-5815 (2003).
45. V. V. Rymar, A. F. Sadikot, Laminar fate of cortical GABAergic interneurons is dependent on both birthdate and phenotype. *J Comp Neurol* **501**, 369-380 (2007).
46. S. M. Kelly, R. Raudales, M. Moissidis, G. Kim, Z. J. Huang, Multipotent radial glia progenitors and fate-restricted intermediate progenitors sequentially generate diverse cortical interneuron types. *bioRxiv* 10.1101/735019 (2019).
47. E. Boldog *et al.*, Transcriptomic and morphophysiological evidence for a specialized human cortical GABAergic cell type. *Nat Neurosci* **21**, 1185-1195 (2018).
48. C. P. Wonders, S. A. Anderson, The origin and specification of cortical interneurons. *Nat Rev Neurosci* **7**, 687-696 (2006).
49. N. Habib *et al.*, Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat Methods* **14**, 955-958 (2017).
50. B. B. Lake *et al.*, Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586-1590 (2016).
51. M. A. DePristo *et al.*, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498 (2011).
52. C. Genomes Project *et al.*, A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
53. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
54. W. Huang, L. Li, J. R. Myers, G. T. Marth, ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593-594 (2012).
55. A. D. Ewing *et al.*, Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods* **12**, 623-630 (2015).
56. G. D. Evrony *et al.*, Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85**, 49-59 (2015).
57. J. Fan *et al.*, Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat Methods* **13**, 241-244 (2016).

Acknowledgements

We thank R. Mattieu, K. Brownstein, J. Li, Flow Cytometry Facility in Boston Children's Hospital, BCH IDDRC Molecular Genetics Core Facility, and the Research Computing group at Harvard Medical School for assistance. We thank G. Fishell, C. Harwell, and F. Vaccarino for comments on the manuscript. Human tissue was obtained from the NIH NeuroBioBank at the University of Maryland and Autism BrainNet, and we thank the donors and their families for their invaluable donations for the advancement of science. P.L is a Howard Hughes Medical Institute – Helen Hay Whitney Foundation Fellow. C.A.W. is supported by the Manton Center for Orphan Disease Research, the Allen Discovery Center program through The Paul G. Allen Frontiers Group, grants from the NINDS (R01NS032457 and U01MH106883), and grant U01MH106883 from the NIMH. C.A.W. is an Investigator of the Howard Hughes Medical Institute.

Figure 1. PRDD-seq enables simultaneous assessment of cell identity and lineage in single cells. A. Neuronal nuclei from postmortem human brain were based on NeuN+ immunoreactivity. B. Target-specific one-step RT-qPCR amplification of cDNA and gDNA fragments of interest. C. Single-cell MosaicHunter co-analysis of single-cell and bulk deep sequencing data to identify lineage-informative somatic SNVs. D. Multiplex analysis of the amplified cDNA and gDNA fragments to genotype the somatic SNVs and profile 30 cell type-specific markers of gene expression. E. 10X Genomics scRNAseq was performed on NeuN+ nuclei isolated from the same PFC region. F. 21 cell clusters were identified based on 10X Genomics gene expression data, and then divided into upper, middle, and lower layer of excitatory neurons and four subtypes of inhibitory neurons. G. A second scRNAseq dataset (5) performed on nuclei isolated from the MTG region of another post-mortem healthy human brain was also analyzed where layer information was identified based on layer micro-dissection. Cell types were identified based on gene expression data. H. Transcriptional clustering revealed similar single-cell expression profiles between 10X Genomics PFC and SMART-seq MTG scRNAseq datasets. Cell clusters were color-coded to denote different cell type annotation, and clusters derived from 10X Genomics PFC (triangle) and SMART-seq MTG (circle) in general clustered by cell type but not by platform. I. Each PRDD-seq cell was mapped to the t-SNE maps by the cosine similarity of gene expression to scRNAseq cells, and then assigned cell type and dissected layer accordingly by majority voting of 25 nearest neighbors. J-K. A combination of genotype and gene expression information of PRDD-seq cells allowed lineage and birthdate analysis of particular cell types/layers (J), and cell type differentiation analysis of particular lineage reconstructed by somatic mutations (K). Colored triangles in (I) indicate PRDD-seq cells. Gray

749 bars in (K) indicate occurrences of somatic mutations, whereas all cells in one corresponding

750 sub-clade share the same somatic mutation.

751

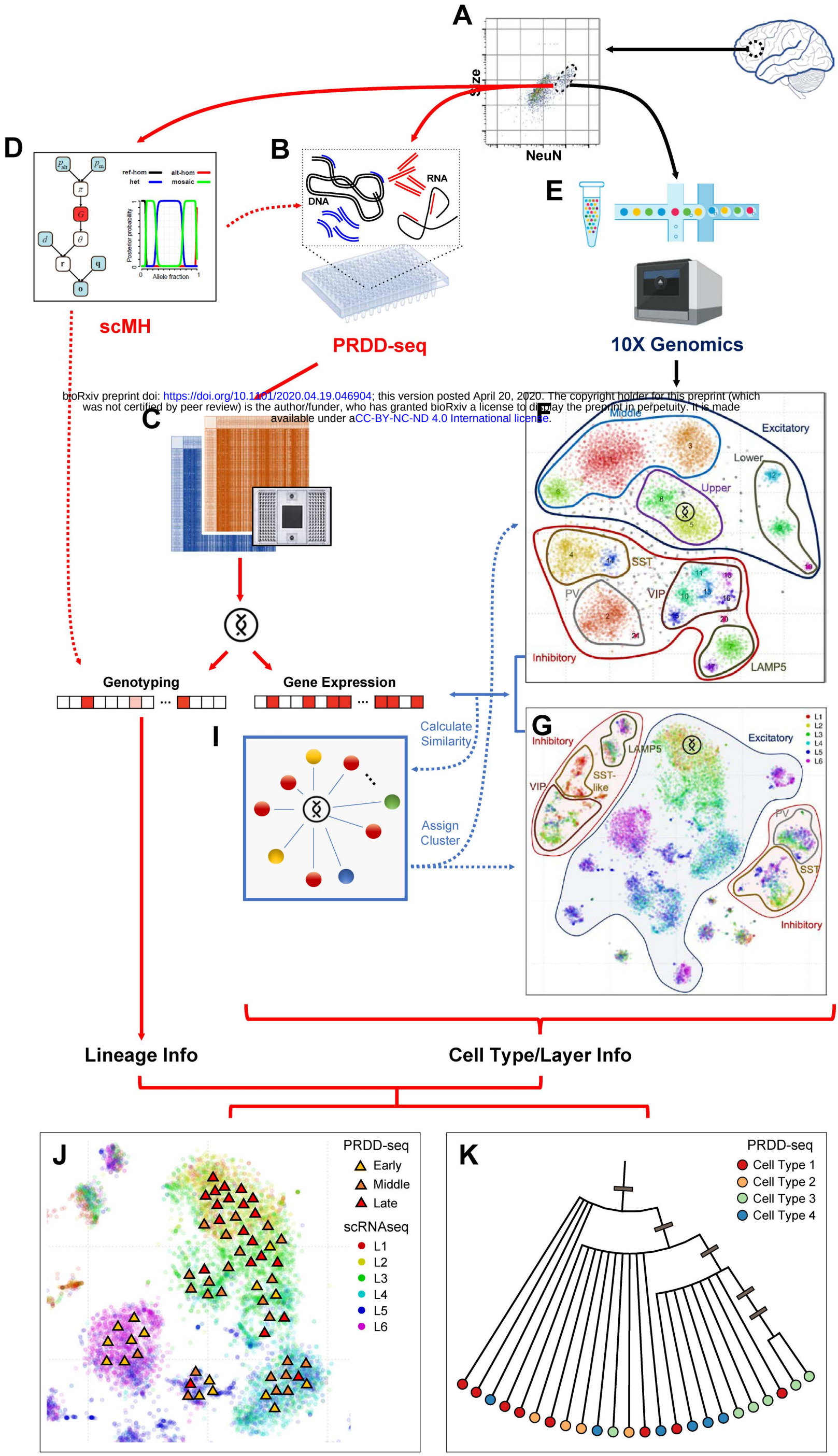
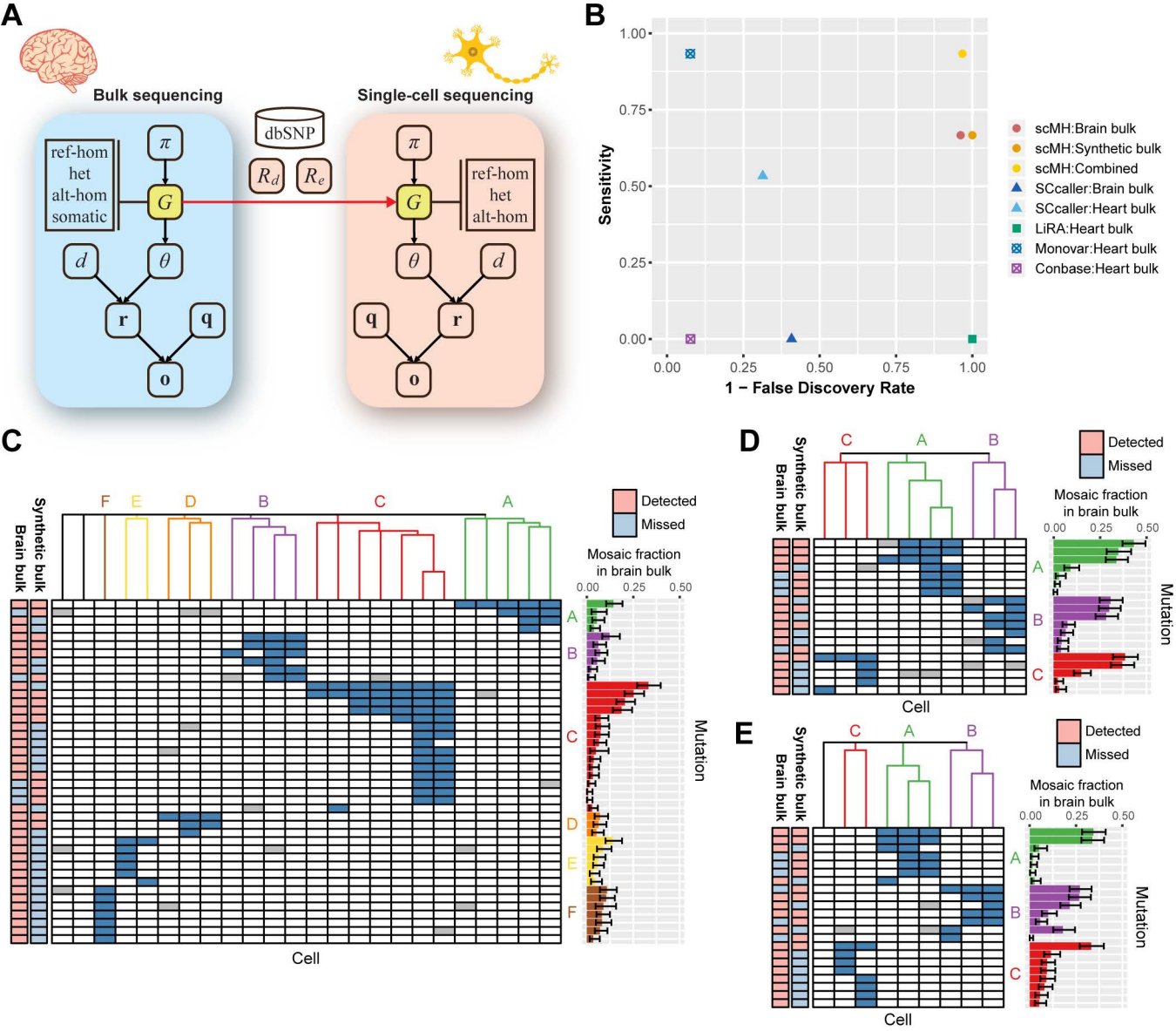


Figure 2. scMH identifies lineage-informative sSNVs from the joint analysis of bulk brain and single-neurons. A. Overview of the extended Bayesian model of scMH to use bulk sequencing data to facilitate sSNV calling in single cells. G denotes the genotype state, π denotes the prior probabilities of genotype, and d, $\mathbf{o}$, $\mathbf{q}$ denote the depth, observed bases, and their base qualities in bulk or single-cell sequencing data. B. Specificity and precision of identifying sSNVs using scMH and other published callers. scMH outcompeted other callers in both precision and sensitivity. C-E. Validated lineage-informative sSNVs identified by scMH in UMB1465 (C), UMB4638 (D), and UMB4643 (E). Heatmaps demonstrate the genotyping status of sSNVs; dark blue and white squares denote the presence or absence of sSNVs in a given cell, whereas grey squares denote unknown genotype due to locus dropout in single-cell WGS. Bar graphs show the mosaic fraction of each sSNV in WGS of bulk brain sample. Clade E in (C), and clade C in (E), represent likely branching clades where early shared mutations are present, while later sSNVs mark two branches with distinct mutations. Error bars reflect 95% confidence internals.


A
Bulk sequencing
Single-cell sequencing
dbSNP
ref-hom
het
alt-hom
somatic
ref-hom
het
alt-hom
B
Sensitivity
1 − False Discovery Rate
scMH:Brain bulk
scMH:Synthetic bulk
scMH:Combined
SCcaller:Brain bulk
SCcaller:Heart bulk
LiRA:Heart bulk
Monovar:Heart bulk
Conbase:Heart bulk
C
Synthetic bulk
Brain bulk
Detected
Missed
Mosaic fraction
in brain bulk
Mutation
Cell
D
E

766 **Figure 3. PRDD-seq profiles single-neurons with varied lineage markers and distinct cell**
767 **type identity.** A. Genotyping results of 30 sSNVs (by rows) from 20 lineages across PRDD-seq
768 cells (by columns) from UMB1465. Blue and white squares represent the presence or absence of
769 sSNV respectively, whereas light blue squares represent the sSNVs that were dropouts in PRDD-
770 seq assay but inferred by the presence of deeper mutations from the same clade. B. Clade
771 classification of PRDD-seq cells profiled in UMB1465. In upper panel, PRDD-seq cells which
772 contained sSNVs from multiple or no clades are labeled as “conflict” and “unknown”
773 respectively. C. Correlation of mosaic fractions from WGS and PRDD-seq (calculated as % of
774 assayed cells carrying a given sSNV) in UMB1465. Both methods showed significantly
775 concordant mosaic fractions (Pearson correlation’s $P < 0.001$). D. Accuracy of cell type (left
776 panel) and cortical layer (right panel) classification based on the expression profile of 30 marker
777 genes used in PRDD-seq. scRNAseq cells from each cell type (10X Genomics) and cortical layer
778 (SMART-seq) were randomly sampled and then re-assigned to clusters of t-SNE map using 30
779 marker genes under PRDD-seq mapping strategy. E. Taxonomy of 3 excitatory layers and 4
780 inhibitory subtypes based on average expression of 30 marker genes in PRDD-seq cells. Relative
781 density of cortical layers for each subgroup is also shown. pPVALB+ denotes PVALB+/SST-
782 VIP- LAMP5- subtype of inhibitory neurons. F. Relative ratio across different cell types of
783 excitatory and inhibitory neurons between PRDD-seq and 10X Genomics scRNAseq.
784

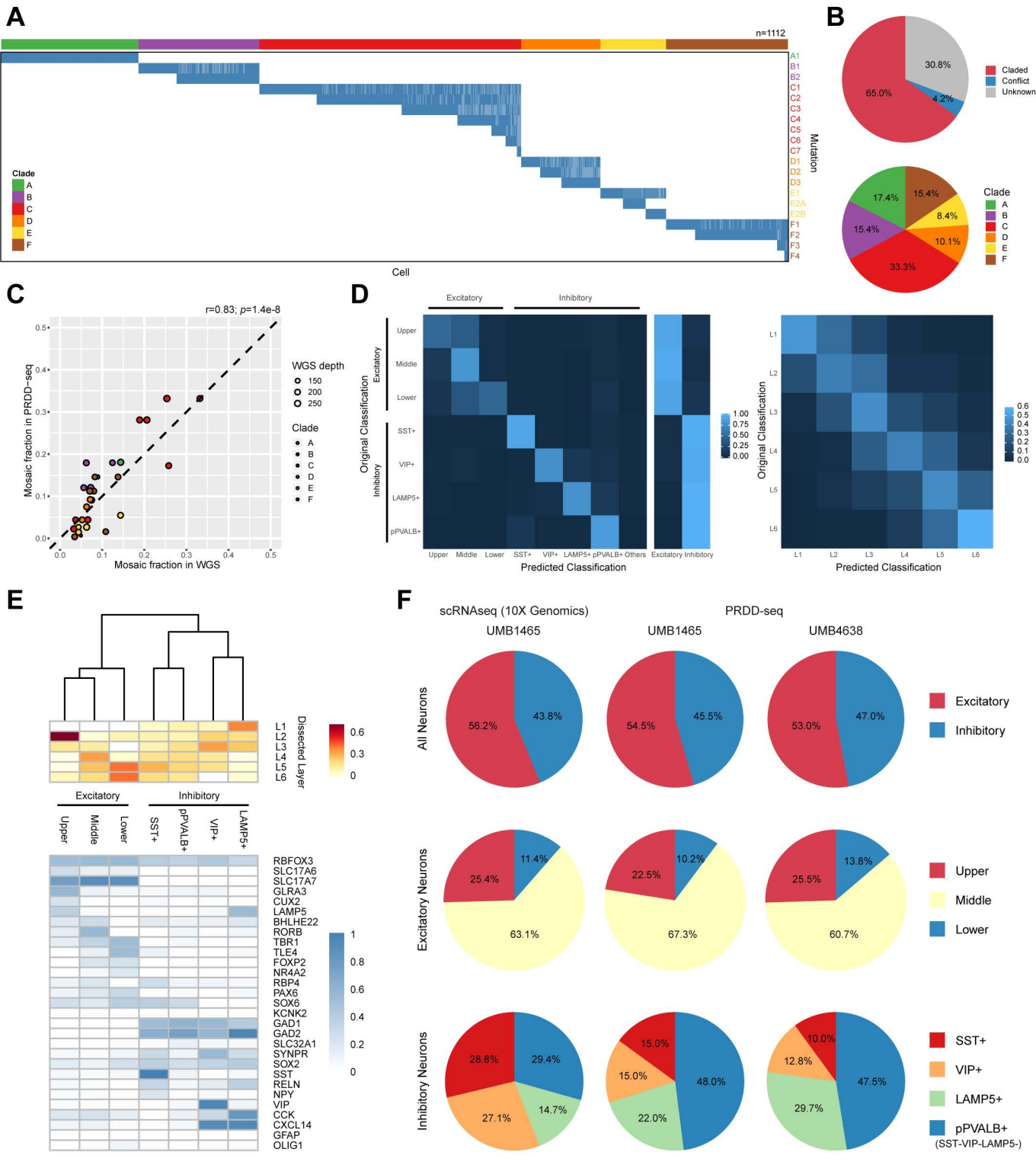

A
n=1112
Clade
A
B
C
D
E
F
Mutation
A1
B1
B2
C1
C2
C3
C4
C5
C6
C7
D1
D2
D3
E1
E2A
E2B
F1
F2
F3
F4
Cell
B
30.8%
65.0%
4.2%
Claded
Conflict
Unknown
17.4%
15.4%
8.4%
15.4%
10.1%
33.3%
Clade
A
B
C
D
E
F
C
r=0.83; p=1.4e-8
Mosaic fraction in PRDD−seq
Mosaic fraction in WGS
WGS depth
150
200
250
Clade
A
B
C
D
E
F
D
Excatory
Inhibitory
Original Classification
Upper
Middle
Lower
SST+
VIP+
LAMP5+
pPVALB+
Others
Predicted Classification
L1
L2
L3
L4
L5
L6
E
Dissected Layer
Excatory
Inhibitory
RBFOX3
SLC17A6
SLC17A7
GLRA3
CUX2
LAMP5
BHLHE22
RORB
TBR1
TLE4
FOXP2
NR4A2
RBP4
PAX6
SOX6
KCNK2
GAD1
GAD2
SLC32A1
SYNPR
SOX2
SST
RELN
NPY
VIP
CCK
CXCL14
GFAP
OLIG1
F
scRNAseq (10X Genomics)
PRDD-seq
UMB1465
UMB1465
UMB4638
All Neurons
56.2%
43.8%
54.5%
45.5%
53.0%
47.0%
Excitatory
Inhibitory
Excitatory Neurons
25.4%
11.4%
63.1%
22.5%
10.2%
67.3%
25.5%
13.8%
60.7%
Upper
Middle
Lower
Inhibitory Neurons
28.8%
29.4%
27.1%
14.7%
15.0%
15.0%
48.0%
22.0%
10.0%
12.8%
47.5%
29.7%
SST+
VIP+
LAMP5+
pPVALB+
(SST-VIP-LAMP5-)

Figure 4. PRDD-seq reveals distinct developmental sequence of excitatory neurons in different cortical layers. A-B. The total number (bar plot) and ratio (dot plot) of excitatory and inhibitory neurons in different lineage clades defined by one or more sSNVs in UMB1465 (A) and UMB4638 (B). Percentage of excitatory neurons increased in later lineage timepoints in clades C and F in UMB1465 and clades A and B in UMB4638. In Clade E of UMB1465, E1 branches into two subclades E2A and E2B. Dashed line: average excitatory neuron percentage. Asterisk denotes significantly different excitatory-inhibitory ratio from the average (two-sided one-proportion Z-test's $P < 0.05$). In clades C and F from UMB1465, and clades A and B from UMB4638, later mutations become progressively limited to excitatory neurons. C-D. Layer distributions of excitatory neurons in representative excitatory lineages in UMB1465 (C) and UMB4638 (D), respectively. Layers are determined by mapping PRDD-seq cells onto human PFC scRNAseq (upper panels) or human MTG scRNAseq (lower panels) based on the expression profile similarity of marker genes. In all three illustrated clades, the percentage of upper layer neurons increased while that of lower layer neurons decreased in cells containing sSNVs present at lower mosaic fraction. *P*-value was calculated by Pearson correlation with ordinal variables.

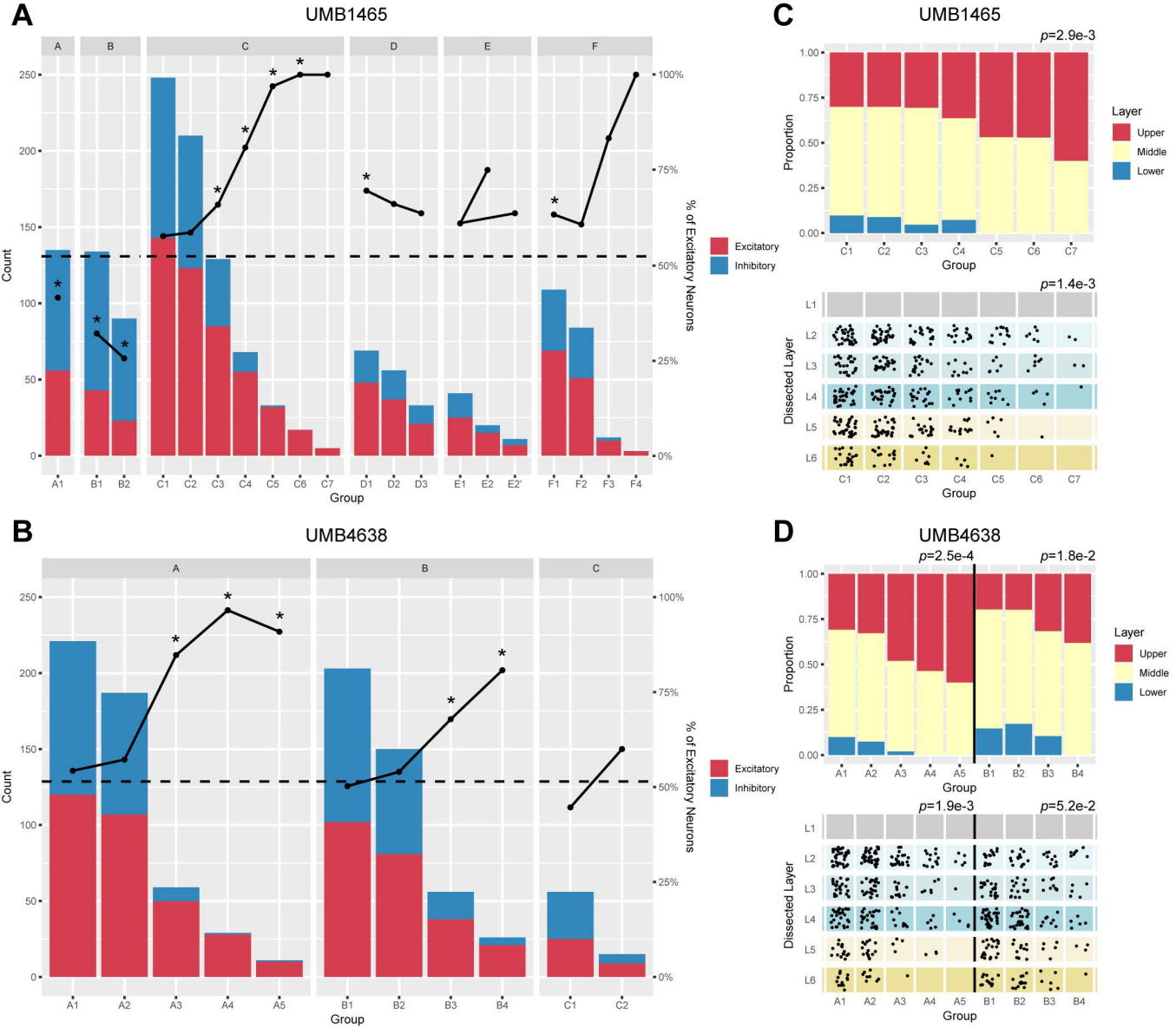
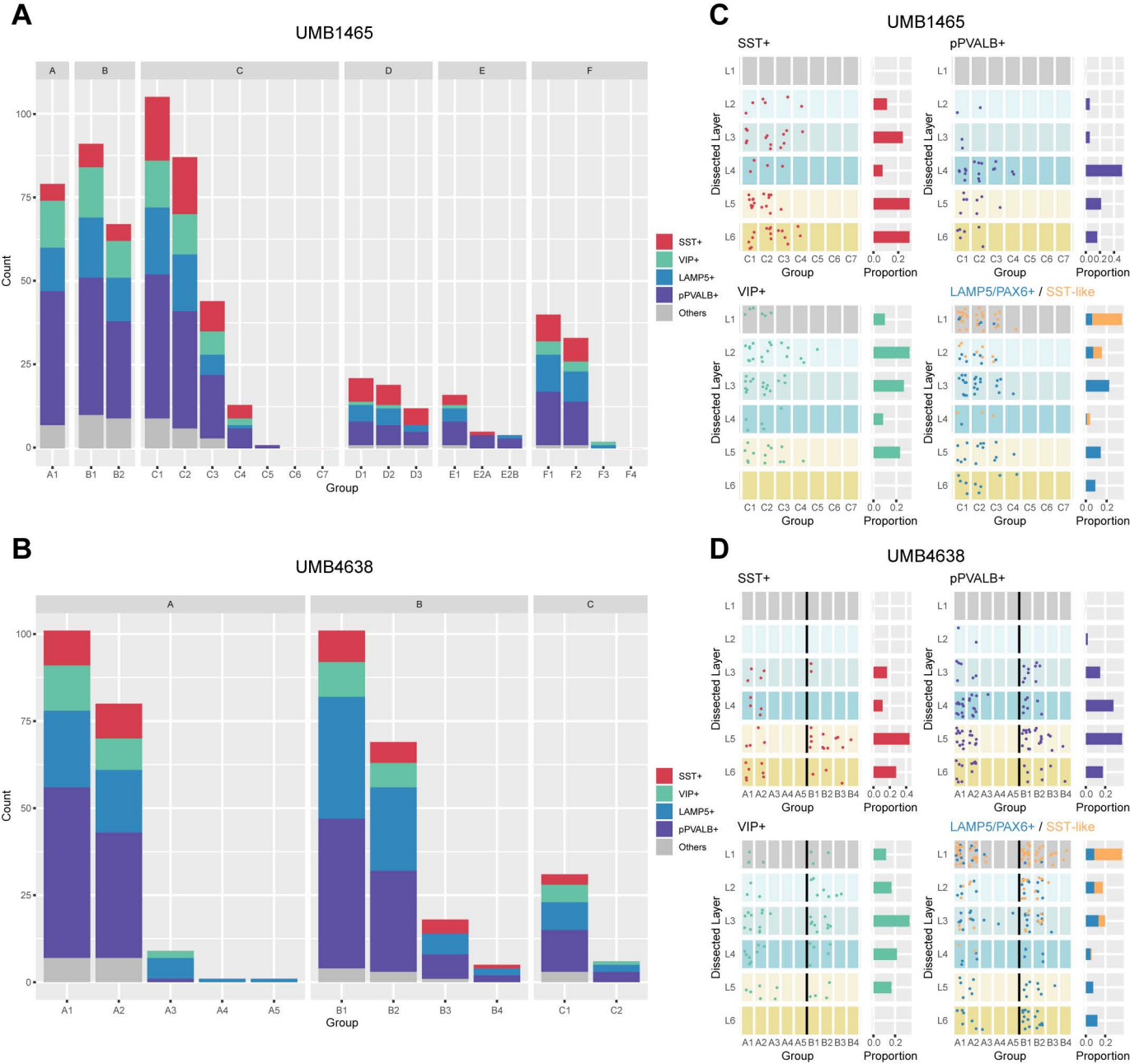

A
UMB1465
A
B
C
D
E
F
Count
% of Excitatory Neurons
Excitatory
Inhibitory
Group
A1
B1
B2
C1
C2
C3
C4
C5
C6
C7
D1
D2
D3
E1
E2
E2'
F1
F2
F3
F4
C
UMB1465
p=2.9e-3
Proportion
Layer
Upper
Middle
Lower
p=1.4e-3
Dissected Layer
L1
L2
L3
L4
L5
L6
B
UMB4638
A
B
C
A1
A2
A3
A4
A5
B1
B2
B3
B4
C1
C2
D
UMB4638
p=2.5e-4
p=1.8e-2
p=1.9e-3
p=5.2e-2

Figure 5. PRDD-seq reveals heterogeneous developmental process for inhibitory neurons.

A-B. Distribution of different subtypes of inhibitory neurons in different lineages in UMB1465 (A) and UMB4638 (B), respectively. Major subtypes of inhibitory neurons are widely distributed in different lineages. C-D. Layer distributions of inhibitory subtypes in representative lineages in UMB1465 (C) and UMB4638 (D), respectively. Bar graphs show the proportion of each subtype of neurons in different layers. MGE derived (SST+ and pPVALB+) and CGE derived (VIP+, LAMP5/PAX6+, and SST-like) interneurons showed similar mutation profiles, suggesting that the groups are produced simultaneously. pPVALB+ subtype neurons were enriched in layer IV-VI, while MGE-derived SST+ interneurons showed a similar laminar distribution as pPVALB+ interneurons, with no clear evidence of an “inside-out” birth dating pattern. CGE-derived interneurons were broadly distributed across cortical layers, with SST-like cells heavily favoring supragranular layers; LAMP5+, including SST-like cells, were enriched for later lineage marks, suggesting they may be produced later in development than other subtypes.


A
UMB1465
Count
Group
SST+
VIP+
LAMP5+
pPVALB+
Others
B
UMB4638
C
UMB1465
SST+
pPVALB+
VIP+
LAMP5/PAX6+ / SST-like
Dissected Layer
Proportion
D
UMB4638