DEVELOPMENT

Landmarks of human embryonic development inscribed in somatic mutations

Sara Bizzotto[1,2,3]*, Yanmei Dou[4]*, Javier Ganz[1,2,3]*, Ryan N. Doan[1], Minseok Kwon[4], Craig L. Bohrson[4], Sonia N. Kim[1,2,3,5], Taejeong Bae[6], Alexej Abyzov[6], NIMH Brain Somatic Mosaicism Network†, Peter J. Park[4,7]‡, Christopher A. Walsh[1,2,3]‡

Although cell lineage information is fundamental to understanding organismal development, very little direct information is available for humans. We performed high-depth (250×) whole-genome sequencing of multiple tissues from three individuals to identify hundreds of somatic single-nucleotide variants (sSNVs). Using these variants as "endogenous barcodes" in single cells, we reconstructed early embryonic cell divisions. Targeted sequencing of clonal sSNVs in different organs (about 25,000×) and in more than 1000 cortical single cells, as well as single-nucleus RNA sequencing and single-nucleus assay for transposase-accessible chromatin sequencing of ~100,000 cortical single cells, demonstrated asymmetric contributions of early progenitors to extraembryonic tissues, distinct germ layers, and organs. Our data suggest onset of gastrulation at an effective progenitor pool of about 170 cells and about 50 to 100 founders for the forebrain. Thus, mosaic mutations provide a permanent record of human embryonic development at very high resolution.

Although recent strategies involving DNA editing have used molecular barcodes as clonal markers to map the developmental processes of proliferation, migration, and tissue formation (*1*), such methods are not applicable to understanding human development. Single-cell RNA-sequencing (RNA-seq) methods have been used to analyze transcriptional changes and cell differentiation during human development (*2*), but they are inadequate for lineage tracing, leaving global lineage patterns in humans still largely unexplored. Here, to examine developmental ancestries and clonal composition across the body, we characterized somatic single-nucleotide variants (sSNVs), which are suitable as lineage markers because they accumulate with each cell division (*3*) and most mutations are predicted to be functionally silent (*4*, *5*).

High-depth whole-genome sequencing (WGS; >250× per sample) was performed for five bulk DNA samples from a 17-year-old male (ID: UMB1465) who died with no medical diagnosis [prefrontal cortex (PFC) section 2 gray matter (GM) and white matter (WM), heart, spleen, and liver (>1250× total); Fig. 1A and table S1]. Similarly, >250× WGS was also performed for PFC and two visual cortex samples [Brodmann area (BA)17 and BA18] from two additional individuals who also died with no medical diagnosis, a 15-year-old female (ID: UMB4638) and a 42-year-old female (ID: UMB4643). Applying MosaicForecast, a machine-learning algorithm (*4*), to bulk data and integrating with previously published single-cell WGS (*6*, *7*), we identified 516 total sSNVs (*8*) (table S2). Among the 297 sSNVs detected in UMB1465, 65 (22%) were found across all tissues and 181 (61%) in at least two (Fig. 1B and table S2). All 65 widely shared sSNVs showed alternate allele frequency (AAF) >1%, with 38 (58%) showing >3% (Fig. 1B and table S2). Sensitivity estimates suggest that our approach achieved nearly 100% sensitivity for detecting sSNVs of 3 to 30% AAF (*8*) (Fig. 1C and fig. S1, A to C). Most sSNVs were predicted to be functionally neutral (only two of 297 sSNVs in UMB1465 were exonic; table S3) and thus represent unbiased lineage markers.

Clonal sSNVs in all organs showed similar base substitution patterns, with 55% being C>T substitutions (Fig. 1D and fig. S1, D and E). The trinucleotide context resembled that of sSNVs seen in proliferating tissues and cancer, e.g., clock-like Signature 1 in the COSMIC catalog (*9*), which likely reflects faulty repair of cytosine deamination in cycling cells (*5*, *7*). Liver-specific variants were more common than heart- or brain-specific variants (n = 57, 33, and 19, respectively), consistent with known patterns of clonal amplification and replacement of hepatic units from resident stem cells (*10*), whereas spleen-specific variants were the least common (Fig. 1B and table S2). Amplicon-based targeted sequencing (~25,000× on average) of 94 samples from 17 organs (Fig. 1A and

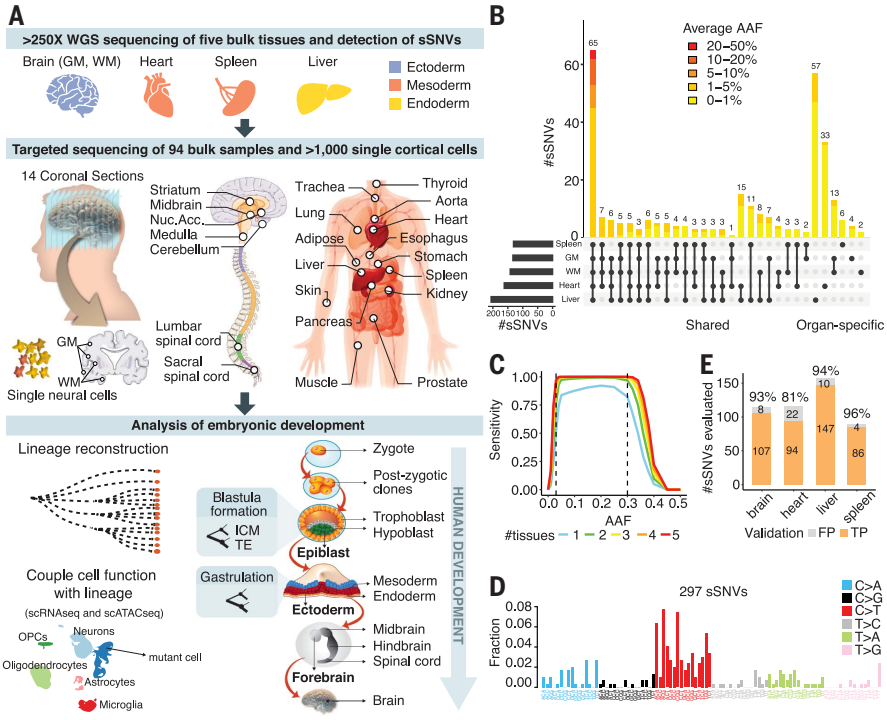


Fig. 1. Mosaic events of human development. (**A**) Schematic of the workflow for individual UMB1465. (**B**) Number and AAF of sSNVs detected across samples from individual UMB1465. (**C**) Sensitivity of MosaicForecast in detecting sSNVs from five 250× WGS data. (**D**) Trinucleotide context profile of the identified sSNVs. (**E**) Numbers of true-positive (TP) and false-positive (FP) sSNVs present in the WGS data validated by deep-amplicon sequencing.

[1]Division of Genetics and Genomics, Manton Center for Orphan Disease Research, Department of Pediatrics, and Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA 02115, USA. [2]Departments of Pediatrics and Neurology, Harvard Medical School, Boston, MA 02115, USA. [3]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA. [5]PhD Program in Biological and Biomedical Sciences, Harvard University, Boston, MA 02115, USA. [6]Department of Health Sciences Research, Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA. [7]Division of Genetics, Brigham and Women's Hospital, Boston, MA 02115, USA.
*These authors contributed equally to this work. †NIMH Brain Somatic Mosaicism Network members and affiliations are listed in the supplementary materials.
‡Corresponding author. Email: christopher.walsh@childrens.harvard.edu (C.A.W.); peter_park@hms.harvard.edu (P.J.P.)

table S1) reidentified most sSNVs (>93%) when the same biopsy used for WGS was profiled (table S1); it identified slightly fewer when distinct tissue biopsies were profiled (81%), and, overall, 196 of 229 (86%) of targeted variants were validated (Fig. 1E, fig. S1F, and table S4).

Single-cell WGS data of 20 single neurons (*6*, *7*) from UMB1465 resolved 82 of 297 sSNVs into branching clades or clones, producing a lineage tree that spans early postzygotic cell generations and traces the origin of each mutation back to the embryo (Fig. 2A, fig. S2A, and tables S2 and S5). As expected, earlier sSNVs showed higher mosaic fractions (MFs), which are the fractions of cells carrying the variant, defined as 2×bulk AAF for autosomal SNV, with the MFs from daughter clades summing to that of the mother clone. Similar patterns of early lineage were also identified in the two additional individuals based on bulk WGS and single-cell (*7*, *11*) analysis (Fig. 2, B and C; fig. S2, B and C; and table S5). In UMB1465, we identified the first eight postzygotic progenitors corresponding to the third-cell generation (c1 to c8, with c5 to c6 not fully resolved and annotated as a second-generation clone), with the MFs of c1 to c8 summing to ≈100%, suggesting that all major early lineages were captured; we then traced their relative contributions to each organ (Fig. 2D and fig. S2D) (*8*). Contributions of c1 to c8 were highly unequal across organs, with c4 undetected in heart and spleen, and c3 and c8 together contributed >50% of the cellular content (Fig. 2D).

Changes in MFs across cell generations suggest highly asymmetrical segregation of the earliest progenitors between embryonic and extraembryonic tissues and in the several germ layers within the embryo. Instead of the expected twofold reduction of MFs with cell division, observed MFs for one branch (c8) barely decreased (30, 26, and 24%; $P < 10^{-6}$, $P < 10^{-22}$, and $P < 10^{-56}$, respectively; two-tailed binomial test); deviations from twofold reduction were also observed in other branches (Fig. 2, A and E, and fig. S2A) and in the two additional individuals (Fig. 2, B and C, and fig. S2, B and C). This pattern suggests unequal clonal partitioning during blastula formation, when extraembryonic tissues separate from embryonic tissue lineages (Fig. 1A). The observed MF asymmetries indicate that lineage segregation in human embryo might happen as early as the two-cell stage, as suggested in the mouse (*12*–*14*). To further test this hypothesis, we analyzed published (*11*) bulk WGS data (250×)

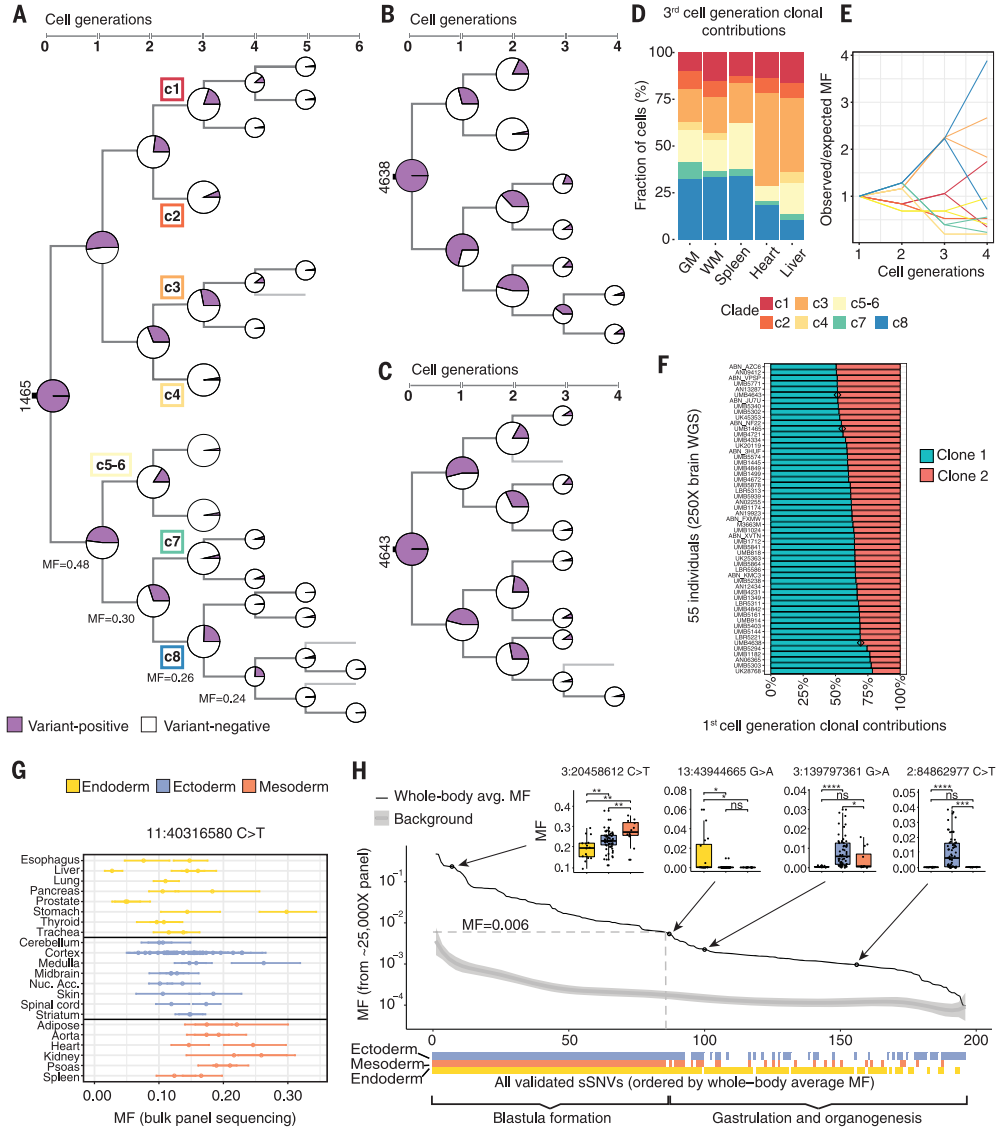


Fig. 2. Asymmetric contribution of early embryonic clones to the human body. (**A** to **C**) Phylogenetic trees of individuals UMB1465 (A), UMB4638 (B), and UMB4643 (C). The cell-generation numbers for later sSNVs (fifth and sixth) are likely to be underestimates because of the limited number of cells used for lineage reconstruction and the reduced power of detecting very-low-MF sSNVs. (**D**) Third-cell-generation clones (c1 to c8) of UMB1465 showing unequal contributions to specific organs ($P < 10^{-15}$, chi-square test), with the fraction of cells in each tissue contributed by clones c1 to c8 normalized by summing to 100% (see fig. S2D for non-normalized values). (**E**) Observed whole-body MFs for sSNVs from clades c1 to c8 across the two- to four-cell generations strongly deviate from expected values based on a symmetrical model of development. The 95% confidence intervals (95% CIs) calculated with binomial sampling are reported in table S2. (**F**) First-cell-generation clonal contributions are asymmetric and variable across 55 individuals ($P < 10^{-13}$, Kolmogorov-Smirnov test for the null hypothesis of symmetry). Individuals UMB1465, UMB4638, and UMB4643 are marked with diamond symbols. (**G**) High intra-organ fluctuation of MFs for early-embryonic mosaic variants illustrated for chr11:40316580 C>T. (**H**) sSNVs restricted to one or two germ layers mark the beginning of gastrulation. A total of 196 validated sSNVs are ordered "chronologically" by their whole-body MFs (*8*). MFs in different germ layers are compared in four examples (two-tailed Wilcoxon rank sum test; ns, nonsignificant; *$P \le 0.05$; **$P \le 0.01$; ***$P \le 0.001$; ****$P \le 0.0001$).

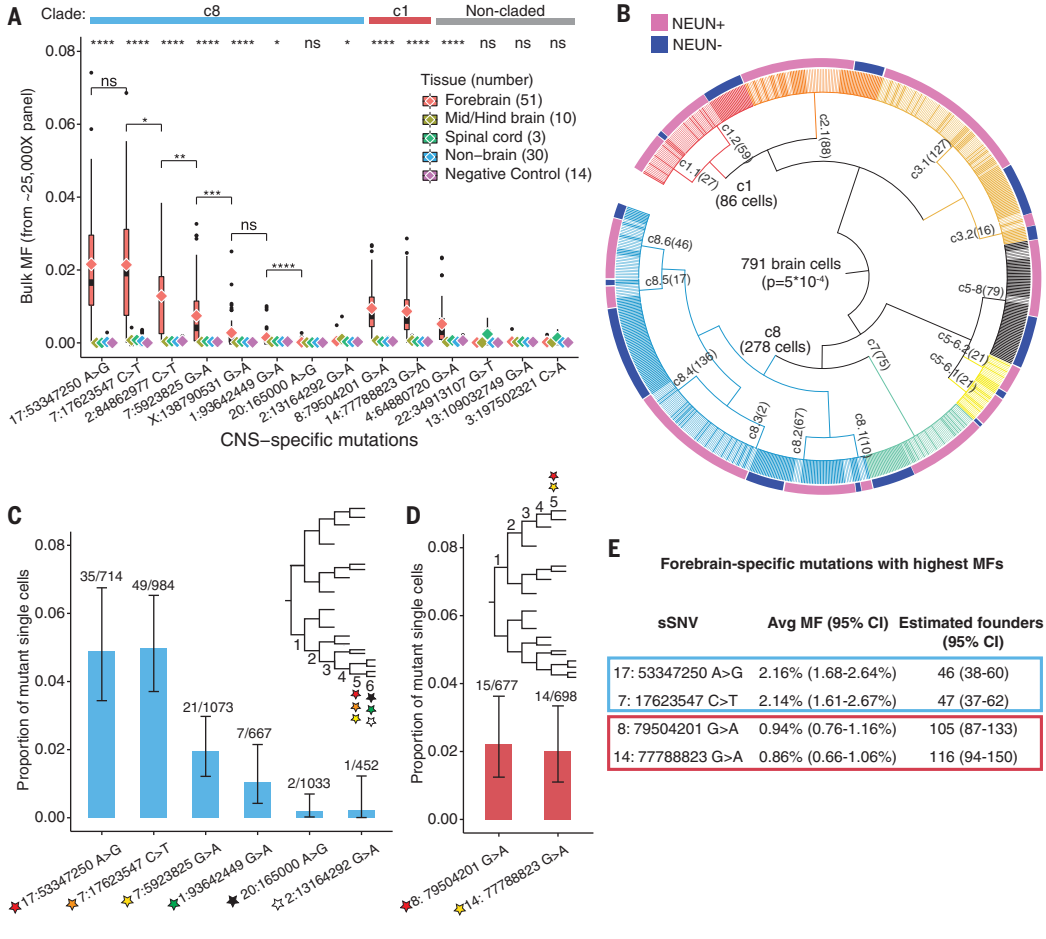
from 74 individuals. Our maximum likelihood estimates (*8*) indicate overall asymmetric contributions of the first-cell-generation clones to the human body with strong interindividual variability, from a 50:50 symmetry in some individuals to a 20:80 asymmetry and potentially higher (Fig. 2F and table S6).

MFs of 196 sSNVs across 94 biopsies from 17 different organs (table S1) from UMB1465 also revealed asymmetric contributions of early lineages to embryonic germ layers during gastrulation (Fig. 1A; fig. S3, A to C; and table S4) (*8*). The relative contributions of several clades to organs of endoderm, ectoderm, and mesoderm varied up to several fold (fig. S3, B and C). Furthermore, multiple biopsies from the same organ showed noticeable intra-organ MF differences (Fig. 2G and fig. S3D). For example, MFs for sSNV chr11:40316580 (C>T) ranged from 5 to 26% across cerebral cortex samples, suggesting highly variable local clonal amplification in all tissues (Fig. 2G).

The tissue distribution of sSNVs identified the effective progenitor pool size at the onset of gastrulation. sSNVs with higher MFs were found in all organs and germ layers (*8*) (Fig. 2H, fig. S3E, and tables S4 and S7), but as MFs decreased past ~0.6%, many sSNVs became undetectable in one or two germ layers (Fig. 2H, fig. S3E, and table S7), reflecting lineage divergence during gastrulation. The effective cell number at the time of mutation occurrence can be inferred as ~1/MF; therefore, 0.6% MF corresponds to ~170 epiblast cells. Despite the asymmetries of clonal contributions to various tissues, multiple germ layer–restricted variants gave similar estimates (Fig. 2H), and our in vivo estimates are consistent with counts from cultured human embryos (*15*).

The earliest brain-specific sSNVs provide similar estimates for the number of brain founder cells. Fourteen sSNVs were present in at least one of 64 central nervous system (CNS) samples but not in 30 non-CNS samples (Fig. 3A and tables S1 and S8), with 10 sSNVs showing significantly higher MFs in the forebrain than in other CNS regions [Fig. 3A and table S8, e.g. $P \leq 0.0001$ for variant chr17:53347250 (A>G)]. The earliest-occurring sSNVs were confirmed from analysis of 1228 single cortical cells [88% were from PFC section 2, so forebrain MFs estimated from single cells may be biased (*8*); table S9], of which 791 were successfully placed in a lineage tree (Fig. 3B, figs. S4 and S5, and table S9), with the neuronal and non-neuronal cells differentially distributed across the clades. The two earliest sSNVs showed wider presence in single cells (Fig. 3C and fig. S5) and a higher overall bulk MF (~2.2%) than other CNS-specific mutations from the same c8 branch (Fig. 3A). We also examined CNS-specific sSNVs with the highest bulk MF (~1%) in clade c1 (Fig. 3, A and D, and fig. S5). These early variants showed wide distribution across the forebrain (fig. S6, A and B) at relatively high MFs (table S8), but were undetectable in most other samples. These variants therefore serve as markers of the first forebrain progenitors and, based on their average bulk MFs, the number of forebrain founder cells is estimated to be ~50 to 100 of an estimated 600 to 1300 epiblasts (Fig. 3E and fig. S6C).

Analysis of sSNVs in 47 DNA samples spanning the rostrocaudal extent of the cerebral cortex (Fig. 1A and table S1) confirmed previous descriptions of widespread clonal distribution at low MFs (*6*, *16*), and suggested broadly definable topographic variation between the frontal (sections 1 to 7) and posterior cortex (sections 8 to 14) (*8*) (Fig. 4A and table S8). Early (first- to fourth-cell-generation) sSNVs were found in all rostrocaudal sections



sSNV	Avg MF (95% CI)	Estimated founders (95% CI)
17: 53347250 A>G	2.16% (1.68-2.64%)	46 (38-60)
7: 17623547 C>T	2.14% (1.61-2.67%)	47 (37-62)
8: 79504201 G>A	0.94% (0.76-1.16%)	105 (87-133)
14: 77788823 G>A	0.86% (0.66-1.06%)	116 (94-150)

Fig. 3. Brain-specific sSNVs estimate the number of forebrain founder cells. (**A**) MFs of 14 CNS-restricted sSNVs showing significant enrichment of some variants in forebrain-derived samples (two-tailed Wilcoxon rank sum test; significance levels are shown at the top). c8 and c1 (Fig. 2A) and noncladed variants are indicated. chr17:53347250 A>G and chr7:17623547 C>T are the earliest brain-specific sSNVs in c8, based on average forebrain MFs (diamond symbols). The forebrain MFs between sSNVs were compared with the estimate of the likelihood that they arose at the same generation (two-tailed Wilcoxon rank sum test). (**B**) A total of 791 single cells (of 1228) were successfully assigned to lineage clades upon targeted sequencing of 37 sSNVs (*8*). $NEUN^+$ and $NEUN^-$ cells are differentially distributed across clades (two-tailed Fisher's exact test). (**C**) chr17:53347250 A>G and chr7:17623547 C>T were confirmed as the earliest lineage markers within c8 by single-cell genotyping (shown are the number of mutant cells over the number of cells with >10× coverage at the position). (**D**) Same as (C) but for c1. (**E**) Estimates of forebrain founder cells based on average MFs (25,000× sequencing).

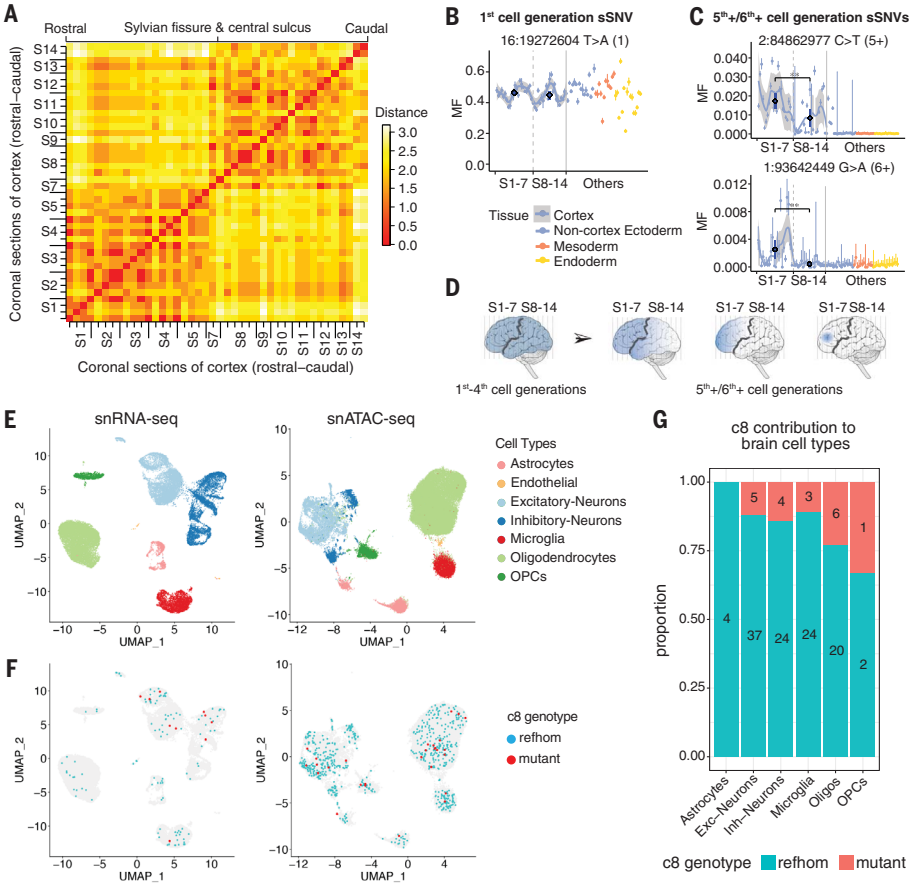


Fig. 4. Topographic patterns and function of embryonic clones in the rostrocaudal cerebral cortex. (**A**) Frontal regions (sections 1 to 7) and posterior regions (sections 8 to 14) form two broadly definable lineage clusters. Euclidian distances were computed based on the presence (score = 1) or absence (score = 0) of sSNVs. (*8*). (**B**) Earlier clones from the first- to the fourth-cell generations contribute to all rostrocaudal sections, as illustrated by an sSNV from the first-cell generation (fig. S6A). The AAFs across sequential sections of cortex are shown with a confidence band. The average MFs (dark blue) in the two regions are compared using a Wilcoxon rank sum test. (**C**) Fifth- and sixth-(or later)cell-generation clones from the lineage tree showing restriction in the frontal cortex regions (fig. S6B). (**D**) Successive subclones from the first- to the sixth-(or later)cell generations showing progressive restriction to frontal cortical areas separated by Sylvian fissure and central sulcus (black line). (**E**) Clusters of major brain cell types identified by PFC snRNA-seq and snATAC-seq. (**F**) Distribution of reference homozygous ("refhom") and mutant cells for clade c8 markers with >0 coverage across cell types. (**G**) Proportions of refhom cells and mutant cells for fourth-cell-generation clade c8 markers across brain cell types. P = 0.58, Fisher's exact test (see also table S10).

(*8*) (Fig. 4B and fig. S6, A and B), although their widely varying mosaic fractions highlighted unexpectedly large local nonuniformities in clonal amplification (Fig. 4B and fig. S6, A and B). Later (fifth- and sixth-(or later)cell-generation) sSNVs showed progressive restriction to the frontal cortex (Fig. 4C and fig. S6, A and B) and finally the PFC, where they were discovered. Thus, whereas founder clones of the cortex show little topographic restriction for MFs of ~1% or higher, lower MF clones show evidence of broad differences in distribution from frontal to posterior regions, separated approximately by the Sylvian fissure and the central sulcus (Fig. 4D).

Single-nucleus RNA-seq (snRNA-seq) and single-nucleus assay for transposase-accessible chromatin sequencing (snATAC-seq) data reveal cell-type classification, but the clusters can also be linked to genotypes. Although limited by the per-cell coverage sparsity, snATAC-seq reads were more uniformly distributed across the genome compared with snRNA-seq reads (fig. S7A), suggesting that snATAC-seq may be better suited to detecting sSNVs genome wide (fig. S7). At the 297 sSNV positions, 5.6% of snRNA-seq cells (1933 of 34,325) and 12.8% of snATAC-seq cells (8356 of 65,199) obtained coverage over at least one of the 297 sSNV loci (table S10). To link cell lineage information with cell types, we classified all ~100,000 cells into seven groups (Fig. 4E and figs. S8 and S9) (*8*, *17*) and checked cells with at least one lineage marker from fig. S2A (Fig. 4F; fig. S7, B to F; and table S10). The sparse coverage of late-occurring variants generally prevents observations of lineage divergence with this approach, although a few trends of c8 contributions to distinct cell types were seen (Fig. 4, E to G, and fig. S10). Our data point to the potential of newer methods for combining analysis of DNA and RNA (*18*, *19*) at high throughput to systematically analyze the formation of distinct cell types at scale in humans.

Our analysis shows that hundreds of sSNVs occurring over several postzygotic cell divisions mark the landmarks of embryonic human development and inform the patterns of clonal distribution within and between organs and tissues. Although analysis of peripheral blood DNA had suggested asymmetries in the contribution of early postzygotic clones to embryonic tissues (*5*), we show here sequential asymmetries and variabilities in clonal proliferation at later steps during gastrulation and organogenesis. The high intra-organ fluctuation of MFs (Fig. 2G and fig. S3D) highlights a stochastic clonal pattern within and across all tissues examined.

We found that clones generated by brain-specific progenitors have average MFs <2.2% across the cortex, underscoring the need for single-cell sequencing for their identification. Regional restrictions of sSNVs to the frontal lobe are seen at even lower MFs (≤0.6%). The observed dispersion of founder clones is consistent with previous estimates (*19*) that a given zone of the human cerebral cortex is formed from ~10 progenitors specified to form excitatory neurons that intermingle widely over a broad region of the cortex (*6*, *16*, *19*). Given the growing list of conditions associated with somatic mutations (*20*, *21*), a deeper understanding of the patterns of cell lineage described here coupled with functional information will help to elucidate the origin and consequence of mosaicism in these diseases.

REFERENCES AND NOTES

1. R. Kalhor *et al.*, *Science* **361**, eaat9804 (2018).
2. X. Han *et al.*, *Nature* **581**, 303–309 (2020).
3. R. E. Rodin *et al.*, *Nat. Neurosci.* **24**, 176–185 (2021).
4. Y. Dou *et al.*, *Nat. Biotechnol.* **38**, 314–319 (2020).
5. Y. S. Ju *et al.*, *Nature* **543**, 714–718 (2017).
6. M. A. Lodato *et al.*, *Science* **350**, 94–98 (2015).
7. M. A. Lodato *et al.*, *Science* **359**, 555–559 (2018).
8. Materials and methods are available as supplementary materials.
9. J. G. Tate *et al.*, *Nucleic Acids Res.* **47** (D1), D941–D947 (2019).
10. R. R. Zhang *et al.*, *Stem Cell Res. Ther.* **9**, 29 (2018).
11. R. E. Rodin *et al.*, The landscape of mutational mosaicism in autistic and normal human cerebral cortex. *bioRxiv*, 2020.2002.2011.944413 (2020).
12. A. Hupalowska *et al.*, *Cell* **175**, 1902–1916.e13 (2018).
13. M. D. White *et al.*, *Cell* **165**, 75–87 (2016).
14. K. Piotrowska, M. Zernicka-Goetz, *Nature* **409**, 517–521 (2001).
15. L. Xiang *et al.*, *Nature* **577**, 537–542 (2020).
16. G. D. Evrony *et al.*, *Neuron* **85**, 49–59 (2015).
17. T. Stuart *et al.*, *Cell* **177**, 1888–1902.e21 (2019).
18. A. S. Nam *et al.*, *Nature* **571**, 355–360 (2019).
19. A. Y. Huang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 13886–13895 (2020).

20. H. Y. Koh, J. H. Lee, *Mol. Cells* **41**, 881–888 (2018).
21. S. Baldassari *et al.*, *Acta Neuropathol.* **138**, 885–900 (2019).
22. S. Bizzotto *et al.*, Landmarks of human embryonic development inscribed in somatic mutations, NIMH Data Archive (2021); https://doi.org/10.15154/1503337.

ACKNOWLEDGMENTS

We thank R. S. Hill, J. E. Neil, D. Gonzalez, S. Yip, and M. Chin for assistance; S. R. Ehmsen for help with graphics; H. Gold, E. Maury, and T. Shin for help with data analysis; A. Y. Huang and P. Li for sharing their snRNA-seq data; Walsh and Park laboratory members, especially R. E. Andersen, C. M. Dias, M. B. Miller, and V. V. Viswanadham, for discussions; the Boston Children's Hospital Flow Cytometry Core and IDDRC Molecular Genetics Core; the Biopolymers Facility and Research Computing at HMS; and the donors and their families for human tissues obtained from the NIH NeuroBioBank at the University of Maryland. **Funding:** This work was supported by National Institute of Mental Health (NIMH Brain Somatic Mosaicism Network grant U01MH106883 to C.A.W. and P.J.P.); NINDS (grant R01NS032457 to C.A.W. and P.J.P.); and the Allen Discovery Center program, a Paul G. Allen Frontiers Group advised program of the Paul G. Allen Family Foundation. Boston Children's Hospital Intellectual and Developmental Disabilities Research Center is funded by NIH grant U54HD090255. S.B. was supported by the Manton Center for Orphan Disease Research at Boston Children's Hospital. J.G. was supported by a Basic Research Fellowship from the American Brain Tumor Association (BRF1900016) and by Brain SPORE grant P50CA165952. S.N.K. is a Stuart H.Q. & Victoria Quan fellow at Harvard Medical School. C.A.W. is an investigator of the Howard Hughes Medical Institute. **Author contributions:** S.B., Y.D., and J.G. conceived the study. S.B. and J.G. performed the experiments. Y.D. led bioinformatics analysis and performed WGS and amplicon-sequencing data analysis. S.B. and Y.D. performed snRNA-seq and snATAC-seq data analyses. R.N.D. designed the targeted-sequencing protocol and contributed to analysis. M.K., C.L.B., T.B., and A.A. contributed to variant analysis. S.N.K. contributed additional brain WGS data. S.B. and Y.D. wrote the manuscript, greatly helped by J.G. C.A.W. and P.J.P. directed the research. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All genomic data are available from dbGaP under accession number phs001485.v2.p1 and from the NIMH Data Archive (*22*). Other materials are available from the authors upon reasonable request.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/371/6535/1249/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S10
References (*23*–*42*)
Tables S1 to S10
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

10 August 2020; accepted 9 February 2021
10.1126/science.abe1544