Cell Systems, Volume *13*

Supplemental information

Simultaneous brain cell type and lineage determined by scRNA-seq reveals stereotyped cortical development

Donovan J. Anderson, Florian M. Pauler, Aaron McKenna, Jay Shendure, Simon Hippenmeyer, and Marshall S. Horwitz

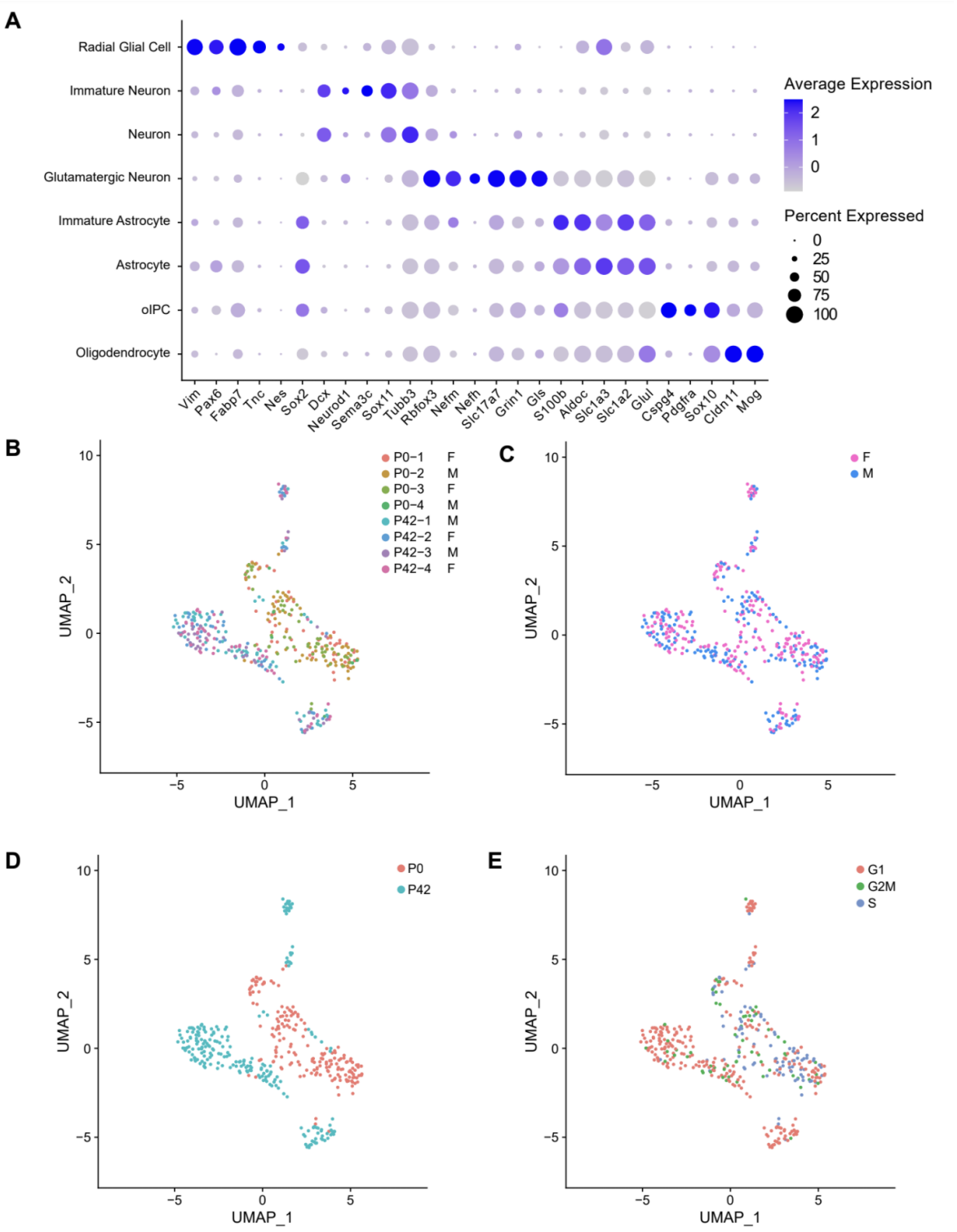

Figure S1. Identification and characterization of cell types and clusters, Related to Figure 1. (A) Average cluster expression of cortical cell marker genes. Scaled (Z-score) average cluster expression (color) is shown for each gene (x-axis) in each cluster (y-axis). Percent of cells in each cluster expressing the gene is indicated by dot size. Radial glial markers: *Vim*, *Pax6*, *Fabp7*, *Tnc*, *Nes*, *Sox2*, *Slc1a3*. Neuronal markers: *Dcx*, *Neurod1*, *Tubb3*, *Rbfox3*, *Nefm*, *Nefh*, *Slc17a7*, *Grin1*, *Gls*. Migration markers: *Sema3c*, *Sox11*. Astrocyte markers: *S100b*, *Aldoc*, *Slc1a3*, *Slc1a2*, *Glul*. Oligodendrocyte markers: *Cspg4*, *Pdgfra*, *Sox10*, *Cldn11*, *Mog*. oIPC = oligodendrocyte intermediate progenitor cell. (B-E) Distribution of 404 neocortical cells plotted in dimensionally reduced (UMAP) space based on gene expression. Colored by (B) mouse identity with corresponding gender indicated, (C) gender, (D) age, and (E) inferred cell cycle stage. UMAP visualizations are derived from the Seurat_CountMatrix Supplemental Data set.

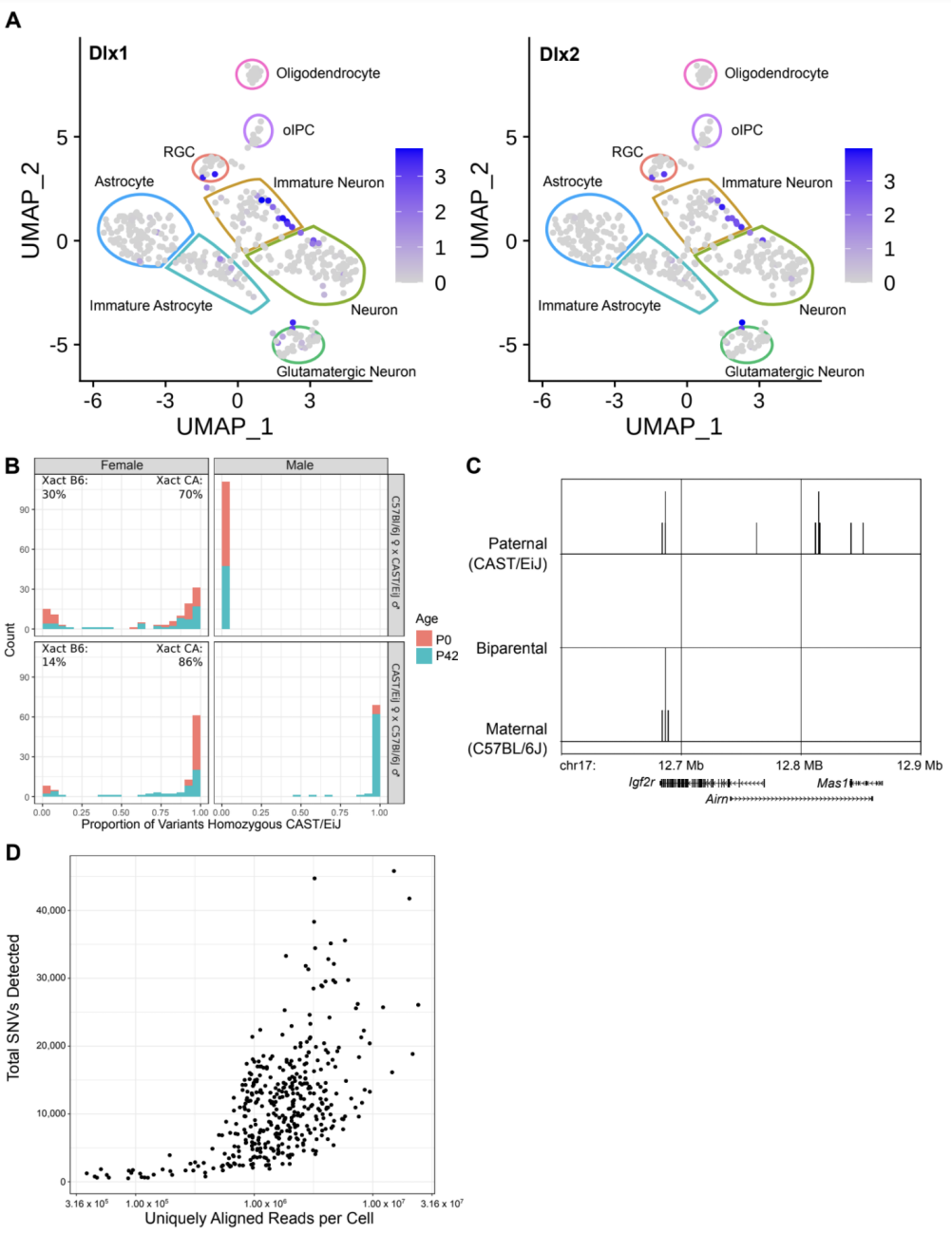

Figure S2. Overlap of neurogenesis in P0 and P42 mice, active X-chromosome skewing, imprinting, and relationship of unique mapped reads to detected variant loci, Related to Figure 1. In the context of Emx1^{+} cells isolated from the cortex, highlighting of *Dlx1* and *Dlx2* transcription shown (A) to mark neuroblasts that originate in the subventricular zone (SVZ) and that are migrating to the olfactory bulb via the rostral migratory stream. All 404 cells are shown as a UMAP projection with scaled normalized count data (ln(normalized counts + 1)) shown in purple. Cell clusters are shown as in Figure 1B, except bounded for clarity. (B) Histograms of individual cell X-activation (Xact) status from eight mice as measured by the proportion of observed single nucleotide variant positions homozygous for the CAST/EiJ variant. Cells are grouped by sex (column) and parental cross (row). Age of donor mice is shown by color. X-activation status is defined as B6 = variant proportion <0.33, CA = variant proportion >0.66. Proportions for each parental cross are different from 0.5:0.5, p <0.005, and different from each other, p <0.006 (two sample Z-test). (C) Relative histograms of observed SNVs from one mouse expressing maternal, paternal, or biparental variants at the *Igf2r* locus. RefSeq annotations for discussed genes are shown for clarity. (D) Scatter plot showing the relationship of unique aligned reads (x-axis) to total variant loci detected (y-axis). Each dot pertains to a single cell. 404 cells shown. UMAP visualizations are derived from the Seurat_CountMatrix Supplemental Data set.

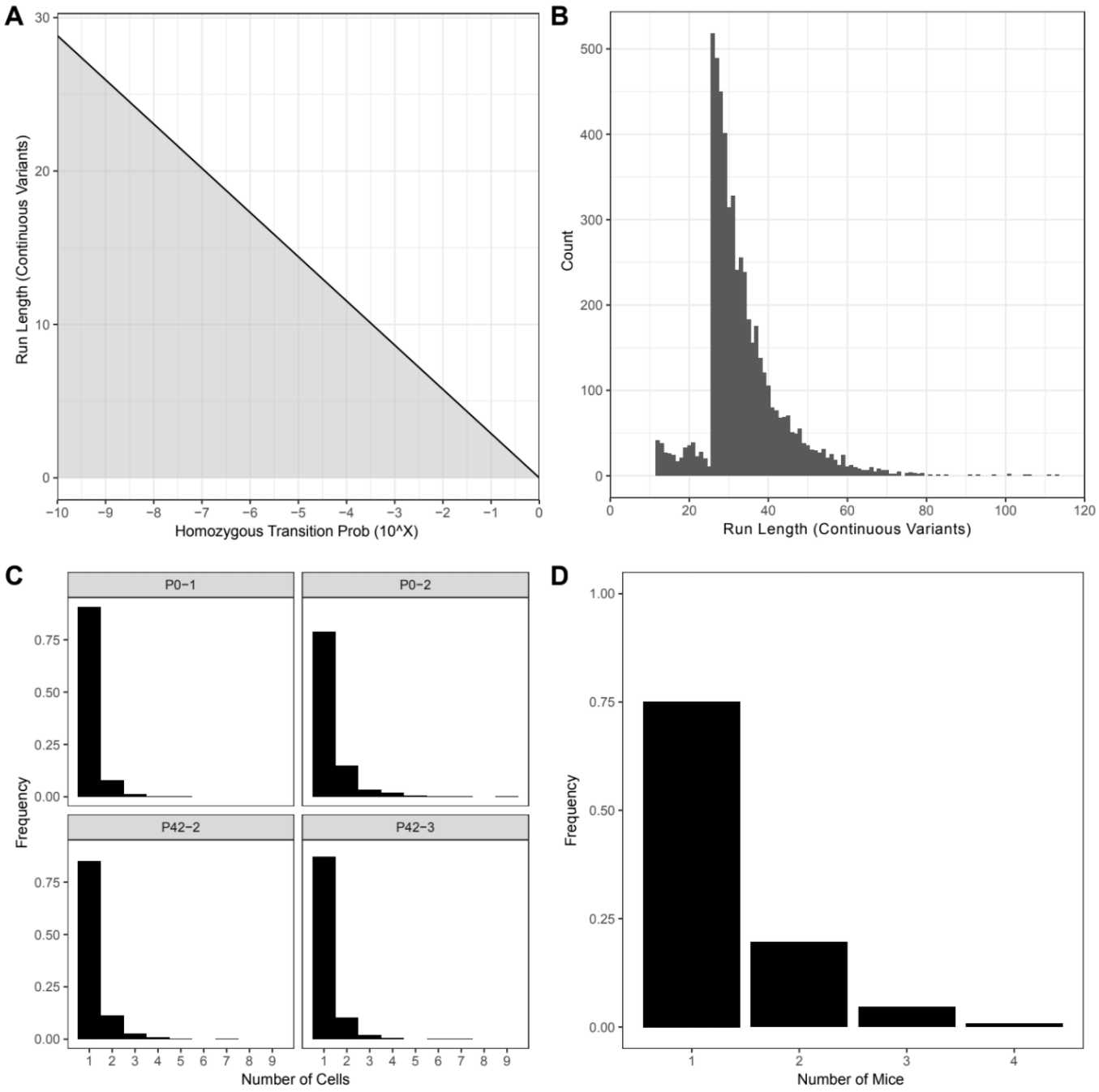

Figure S3. HMM conditions, continuous variant runs, and shared LOH alleles in four mice, Related to Figure 2. (A - B) The required number of continuous monoallelically expressed variants for a predicted LOH event is shaped by the homozygous state transition probability as well as the emission matrix. (A) The number of continuous SNV loci observed as monoallelic needed to trigger a homozygous hidden state change is shown as a function (black line) of the homozygous transition state probability (given the emission matrix described in the methods section). The gray region indicates the area where the hidden state remains heterozygous while the white region describes the conditions needed to trigger a hidden state change to homozygous for either allele, thus indicating an LOH event. (B) A histogram showing consecutive variant loci length of LOH events ≥1 Mb in size for all four mice analyzed. A predominant number of LOH events contain at least 26 consecutive variants. These events would still be called even if the homozygous transition probability was 10^{-8}. (C - D) Distribution of shared alleles within and across mice. (C) Frequency of shared LOH events (alleles) across cells within an individual mouse. Total number of alleles for each mouse; P0-1 = 557, P0-2 = 1544, P42-2 = 1243, P42-3 = 1080. (D) Frequency of shared LOH events (alleles) across four mice. Total number of alleles = 3380.

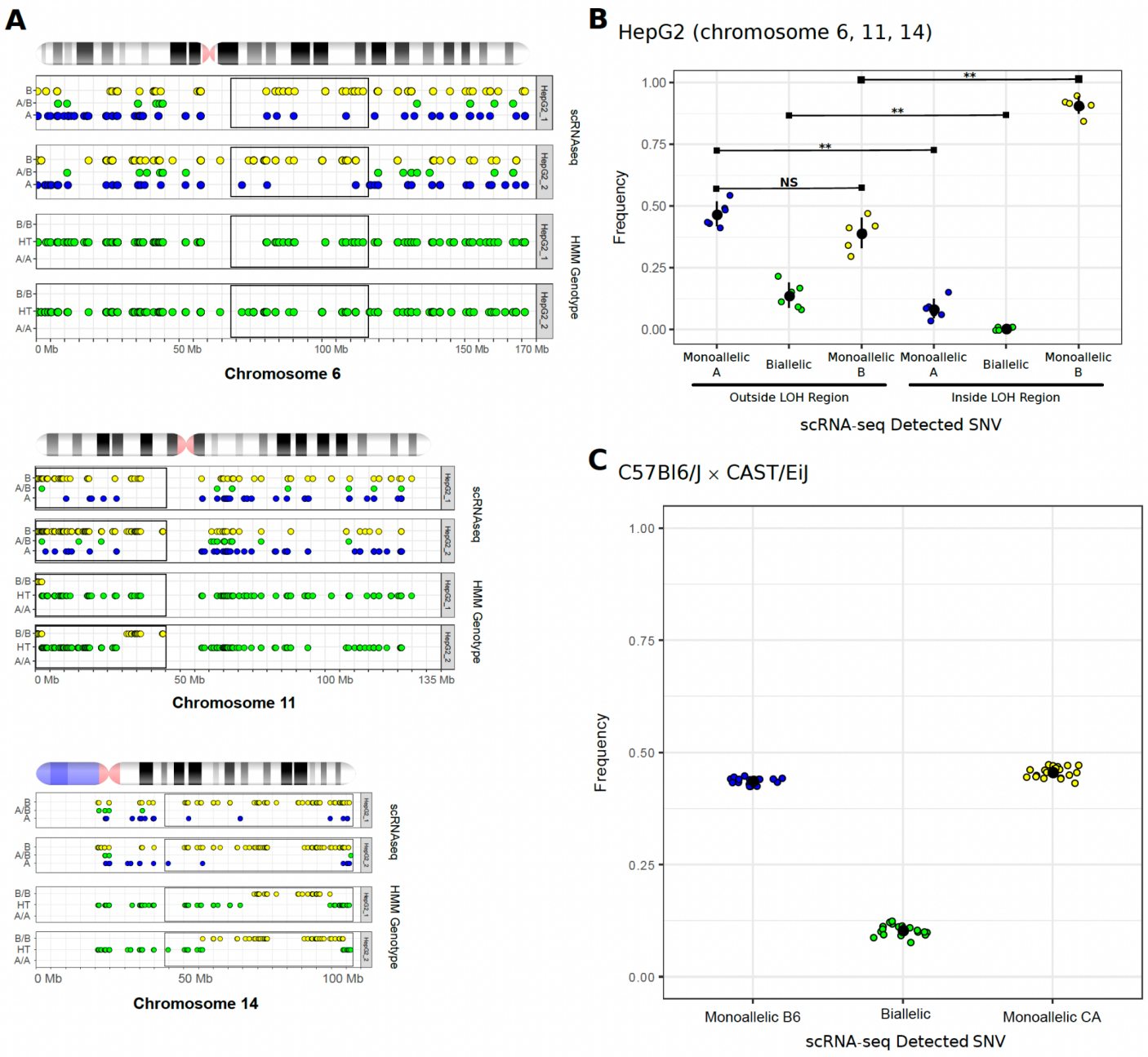

Figure S4. Detection of LOH in human HepG2 cells, Related to Figure 2. (A) scRNA-seq derived allele plots and HMM inferred genotype plots of two HepG2 cells for human chromosomes 6, 11, and 14. Previously reported LOH regions for these chromosomes are boxed in black. Ideograms for each chromosome are shown (red = centromere; purple = acrocentric p-arm, encoding negligible uniquely mapped transcripts). A = haplotype A, B = haplotype B, HT = heterozygous, blue = monoallelic/homozygous haplotype A, yellow = monoallelic/homozygous haplotype B, green = biallelic/heterozygous. (B) Frequency of scRNA-seq derived HepG2 SNV allele states by chromosome region. n = 6 for each region. **p <0.01, NS = not significant, ANOVA with Tukey's Honest Significant Difference test. (C) Average frequency of scRNA-seq derived B6 × CA F1 SNV allele states for mouse chromosomes 1 - 19, across 404 cells. B6 = C57Bl6/J, CA = CAST/EiJ.

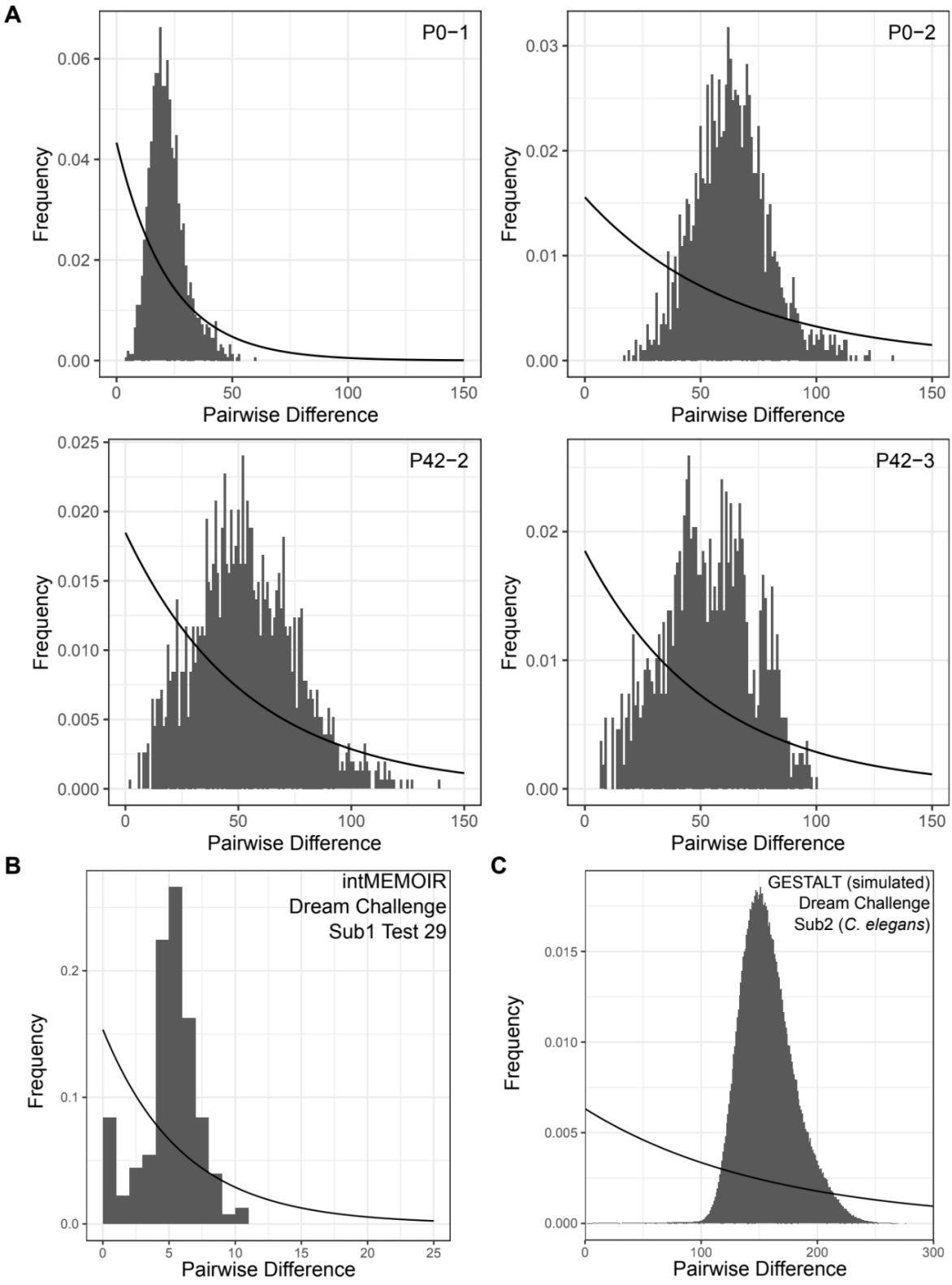

Figure S5. Evidence of expanding populations of cells, Related to Figure 3. (A) For each mouse, the distribution of pairwise allele differences between individual cells is shown. The solid line represents the expected distribution of differences in a population of constant size given the average pairwise distance for the sample. Populations undergoing exponential growth are predicted to have unimodal distributions of pairwise differences. (B) The same as (A) but derived from alleles generated in an *in vitro* expanding cell culture system using the intMEMOIR recording cassette. (C) The same as (A) but derived from simulated GESTALT alleles using the known *C. elegans* lineage as a template.

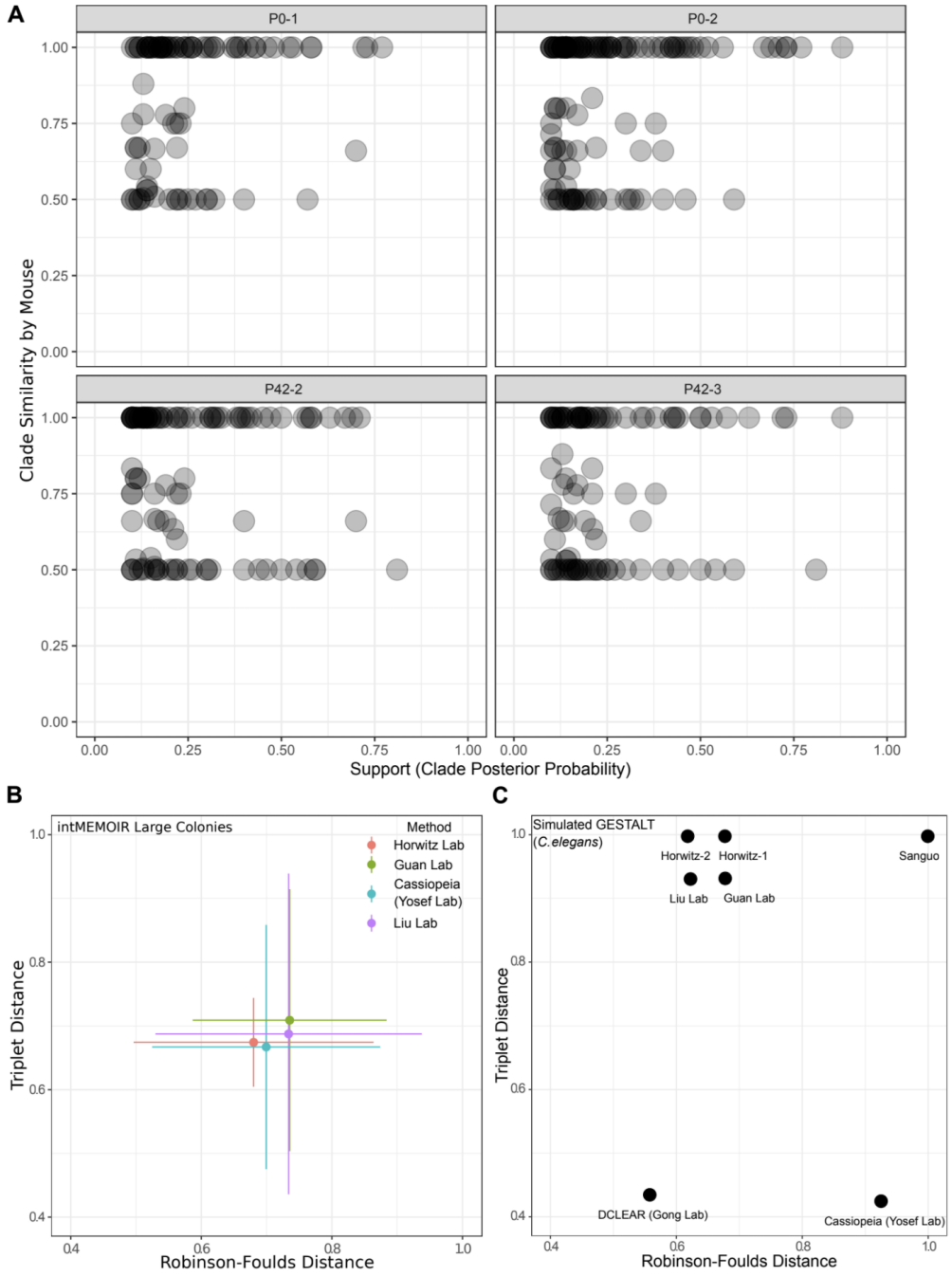

Figure S6. Evidence for clade support and Camin-Sokal Bayesian (CSB) method performance, Related to Figure 3. (A) Clade support and known cellular lineage. The support (clade posterior probability) of monophyletic groups in phylogenies of cells from pairwise comparisons of each mouse to the other three (individual plots) plotted (x-axis) against the group's composition (y-axis), as measured by similarity. A y-axis value of 1.0 indicates that all cells in the group are from the same mouse. Clades with a posterior probability of <0.1 are not shown. (B) Performance of the CSB model using *in vitro* cell colonies with the intMEMOIR recording cassette compared to top performers from the Allen Institute DREAM Challenge, sub challenge one). intMEMOIR generated barcodes from five large colonies (25 – 29 cells) were used to infer phylogenies using the CSB model and compared to observed ground truth lineages. The mean (dots) normalized Robinson-Foulds and triplet distances, along with standard deviation (lines), are plotted (by color) for each different method being compared. (C) Performance of the CSB model using simulated GESTALT recording data for the *C. elegans* lineage tree (Allen Institute DREAM Challenge, sub challenge two). As in (B), inferred phylogenies are compared to the ground truth tree using Robinson-Foulds and triplet distances. Our attempts at reconstruction (Horwitz-1 and Horwitz-2) are shown along with DREAM challenge submitted reconstructions.

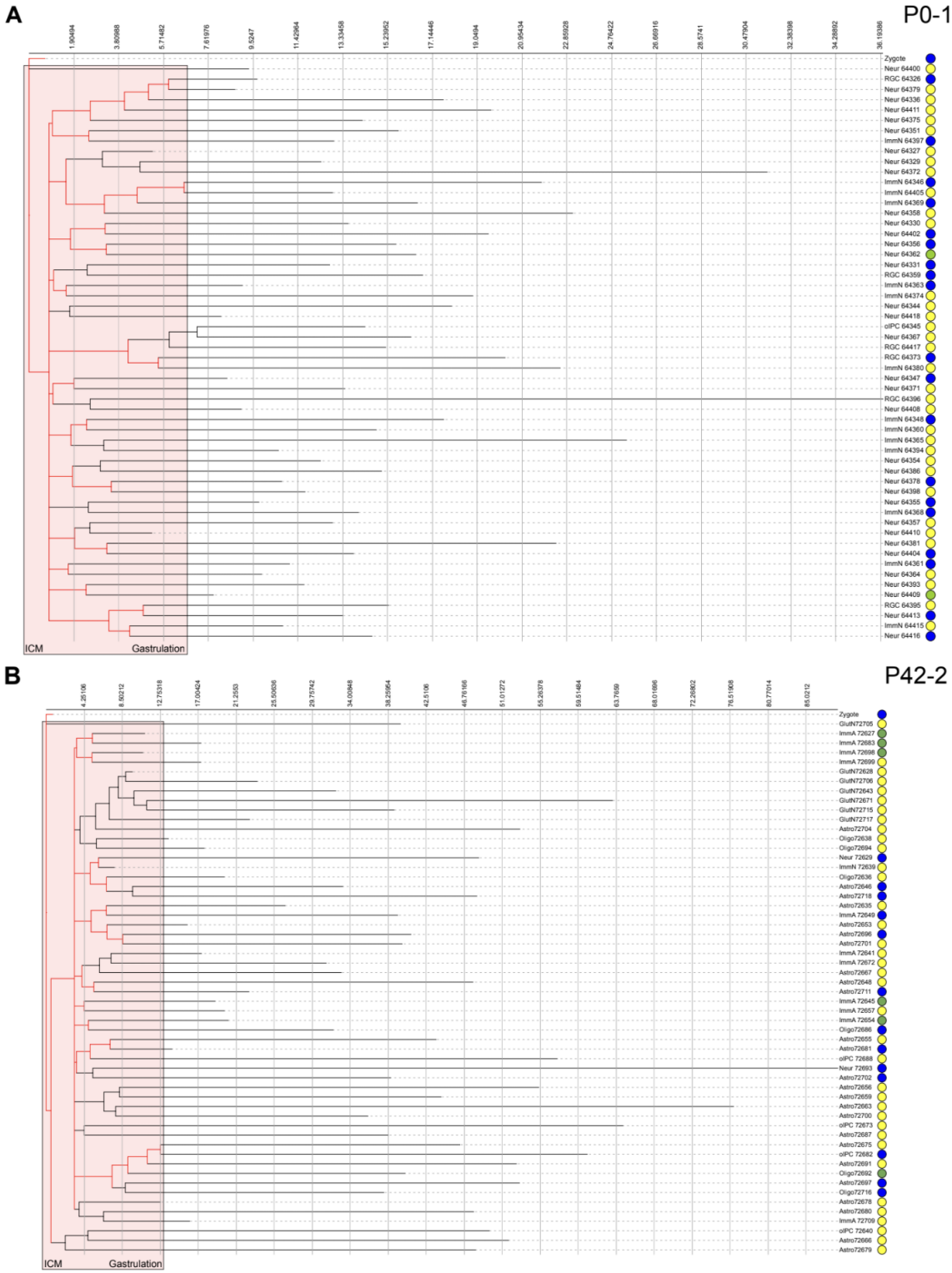

Figure S7. Registration of developmental timepoints to lineage, Related to Figure 4. Phylogenies of Emx1^{+} cortical cells plus inferred zygote. The active X-chromosome is overlaid onto cell lineages of two female mice at age P0 (A) and P42 (B). Cell type is indicated on the right along with the active X-chromosome, indicated by a colored circle, blue = B6, green = biparental/not called, yellow = CA. For mixed clades composed of cells with either active X-chromosome, the most recent common cell (internal nodes) are inferred to occur about or before the time of gastrulation (based on its temporal relationship with X-inactivation), and their respective preceding branches are colored red. The pink box represents branches of the lineage tree inferred to occur during expansion of the epiblast from the inner cell mass (ICM) to gastrulation. Branch length distance (x-axis) is shown in units of LOH events. RGC = radial glia cell, ImmN = immature neuron, Neur = neuron, GlutN = glutamatergic neuron, ImmA = immature astrocyte, Astro = astrocyte, oIPC = oligodendrocyte intermediate progenitor cell, Oligo = oligodendrocyte.