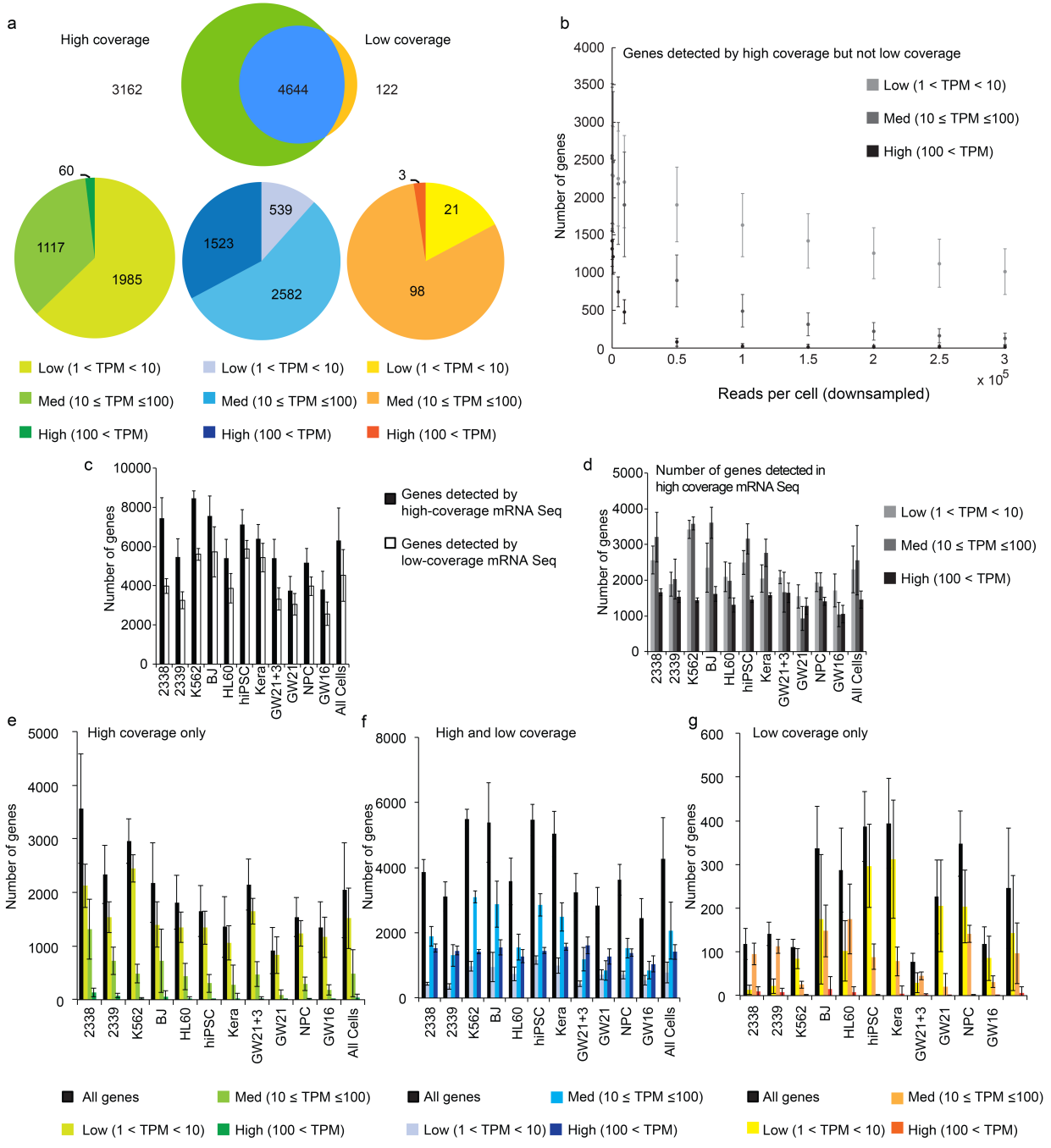
Supplementary Information

Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex

Alex A Pollen[1,2,4], Tomasz J Nowakowski[1,2,4], Joe Shuga[3,4], Xiaohui Wang[3,4], Anne A Leyrat[3], Jan H Lui[1,2], Nianzhen Li[3], Lukasz Szpankowski[3], Brian Fowler[3], Peilin Chen[3], Naveen Ramalingam[3], Gang Sun[3], Myo Thu[3], Michael Norris[3], Ronald Lebofsky[3], Dominique Toppani[3], Darnell Kemp[3], Michael Wong[3], Barry Clerkson[3], Brittnee N Jones[3], Shiquan Wu[3], Lawrence Knutsson[3], Beatriz Alvarado[3], Jing Wang[3], Lesley S Weaver[3], Andrew P May[3], Robert C Jones[3], Marc A Unger[3], Arnold R Kriegstein[1,2] & Jay AA West[3]

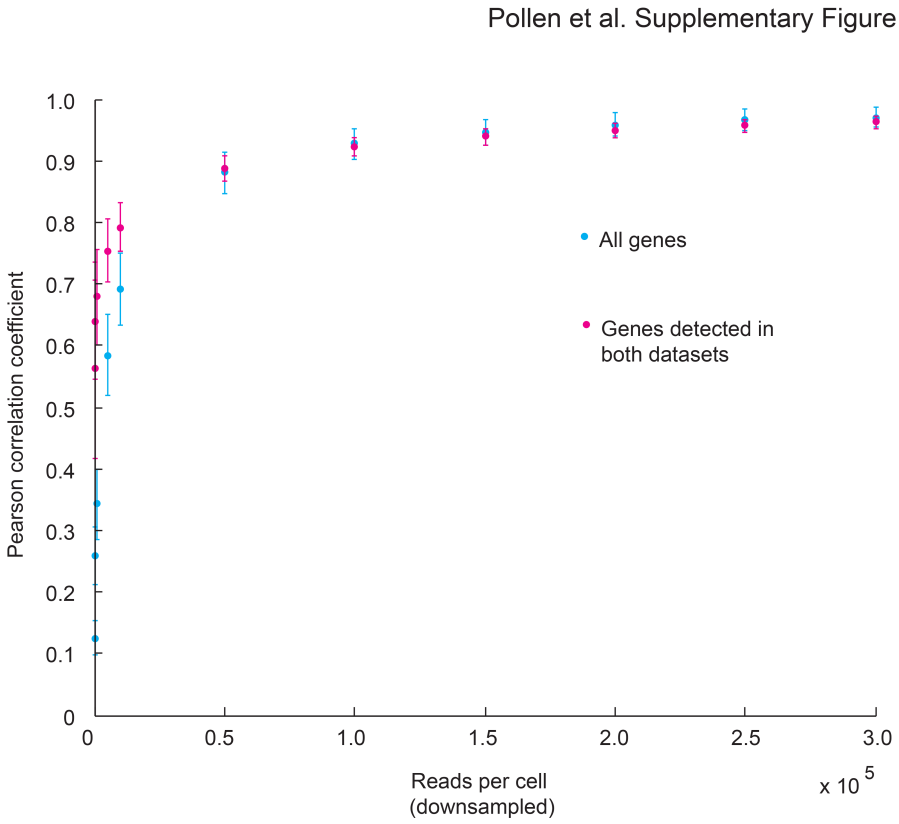
1. Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California, San Francisco, San Francisco, California, USA
2. Department of Neurology, University of California, San Francisco, San Francisco, California, USA
3. Fluidigm Corporation, South San Francisco, California, USA
4. These authors contributed equally to this work

Correspondence should be addressed to A.A.P. (alex.pollen@gmail.com) or A.R.K. (KriegsteinA@stemcell.ucsf.edu).

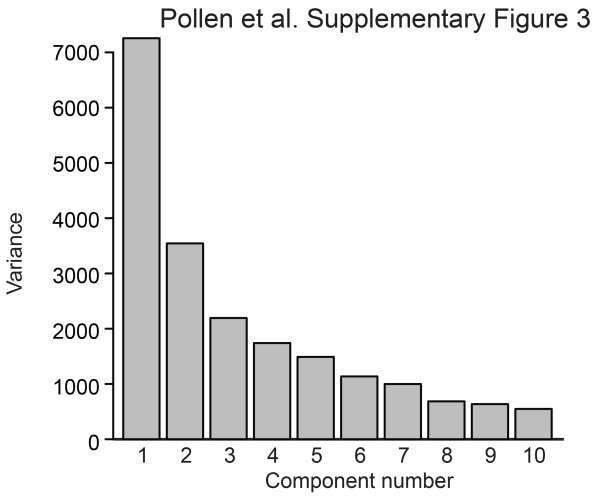


Supplementary Figure 1 Genes with high- and medium-expression levels are robustly detected by low- and high- coverage sequencing. (a) Comparison of genes detected by low- and high-coverage sequencing of a cDNA library from a representative single cell. Total

numbers of genes detected above the threshold (TPM>1) are indicated in shaded circles. Venn diagram shows the overlap between sets of detected genes. Genes in each category (detected only by low- or high-coverage sequencing or by both) were further subdivided based on their expression levels and are shown in the pie charts of corresponding colors. Green pie chart- of the 3162 genes that were detected in the high-coverage data but not detected in the low-coverage data, only 60 (1.9%) were found to be high-abundance genes (TPM>100). In contrast, 1985 of the detected in the high-coverage data but not detected in the low-depth data (62.8%) were found to be low-abundance genes (1<TPM<10). Orange pie chart- A low percentage (1.5%) of total genes detected above 1 TPM were only detected by low-coverage sequencing. These transcripts were captured when sequencing the libraries by MiSeq®, possibly due to sampling error during cluster generation, but were not sufficiently abundant to be detected when sequencing the libraries by HiSeq®. Of these 122 genes, only 3 (2.5%) were found to be high-abundance genes (TPM>100), while 98 (80.3%) were detected at medium expression levels (10≤TPM≤100) in the low-coverage data. Blue pie chart- in contrast with both of the dropout data sets, the 4644 genes that were detected in both the low- and high-coverage data were more evenly distributed and were composed of 1523 (32.8%) high-abundance genes, 2582 (55.6%) medium-abundance genes, and 539 (11.6%) low-abundance genes, based on the TPM levels determined from the high-coverage data. **(b)** Graph showing average number of genes detected by high-coverage sequencing that are no longer detectable across downsampled sequencing depths. **(c)** Graph showing average total number of genes detected above the threshold (TPM>1) in high- and low-coverage data across cells of a given type. **(d)** Graph (same as Fig 1c) showing the average number of genes expressed at various levels detected by high coverage sequencing in each cell type. **(e-g)** Graphs show the average numbers of genes at various expression levels detected only in high-coverage data set (e), in both data sets (f), and only in low-coverage data set across all cells (g). Results represent mean ± s.d.

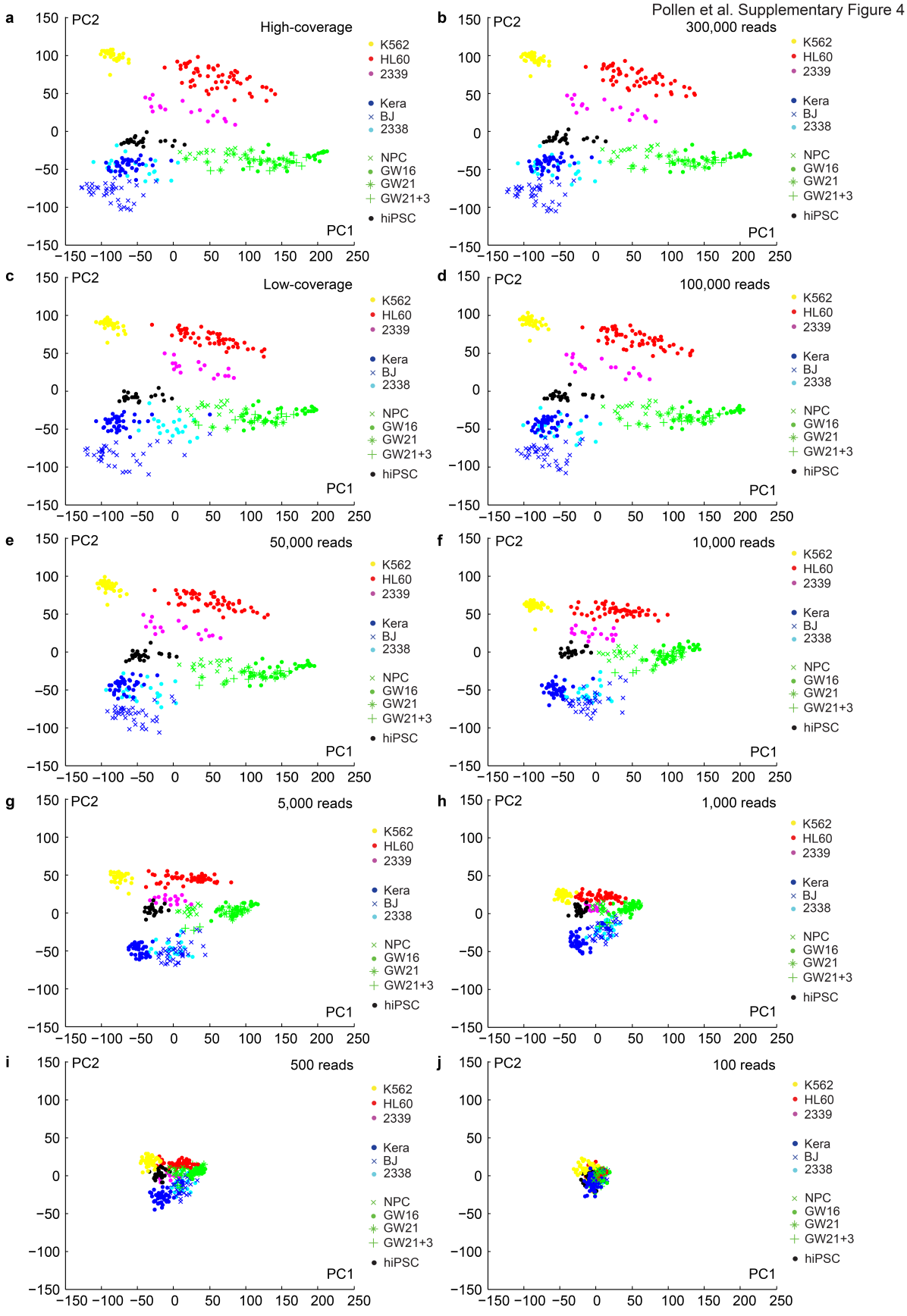


Supplementary Figure 2 Estimates of gene expression levels between high-coverage mRNA Seq and downsampled datasets. Graph showing correlation coefficients between gene expression levels derived from high-coverage mRNA Seq and data derived from downsampled reads. Cyan data points indicate correlation for all genes and magenta data points indicate the correlations calculated only for genes detected in both datasets.

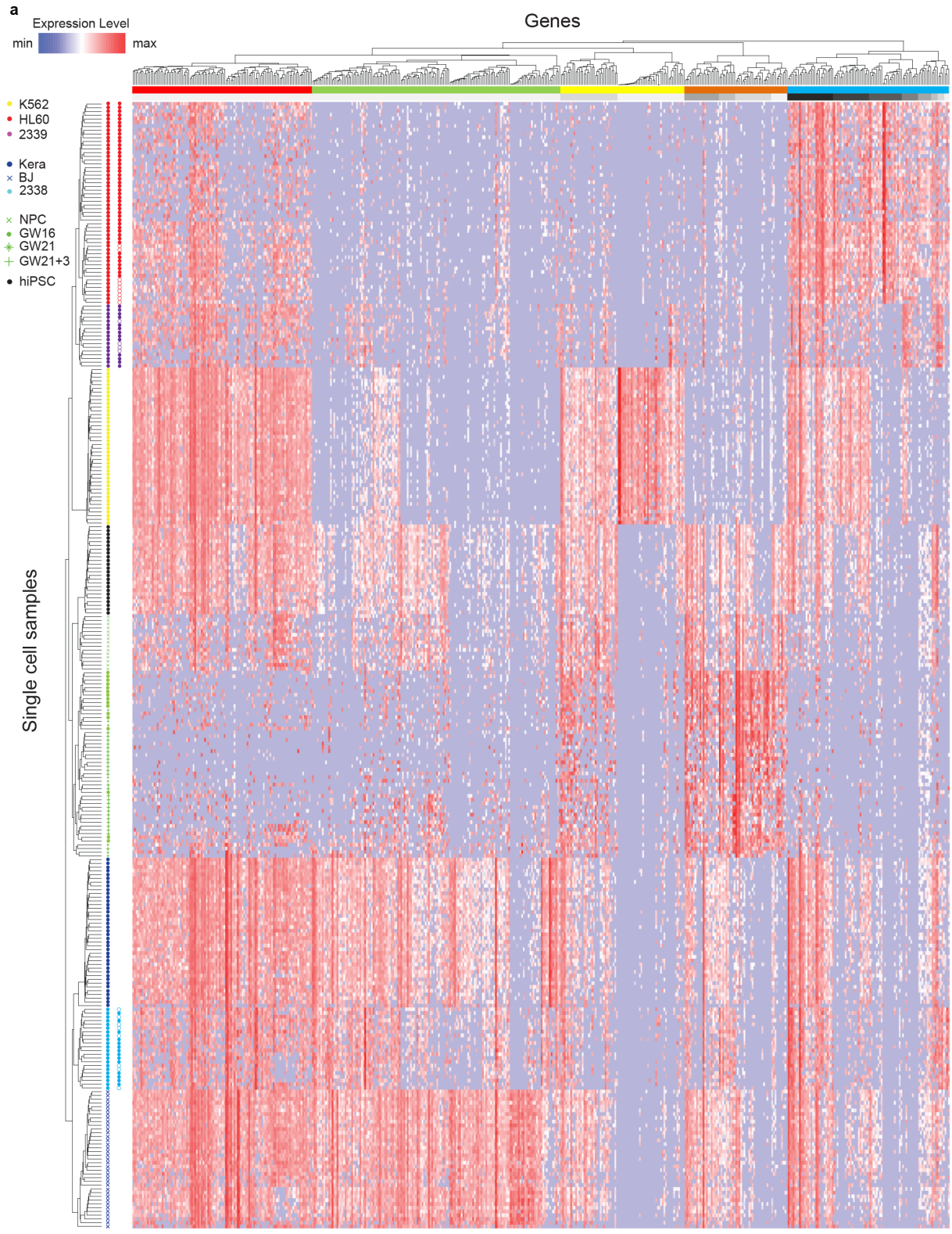


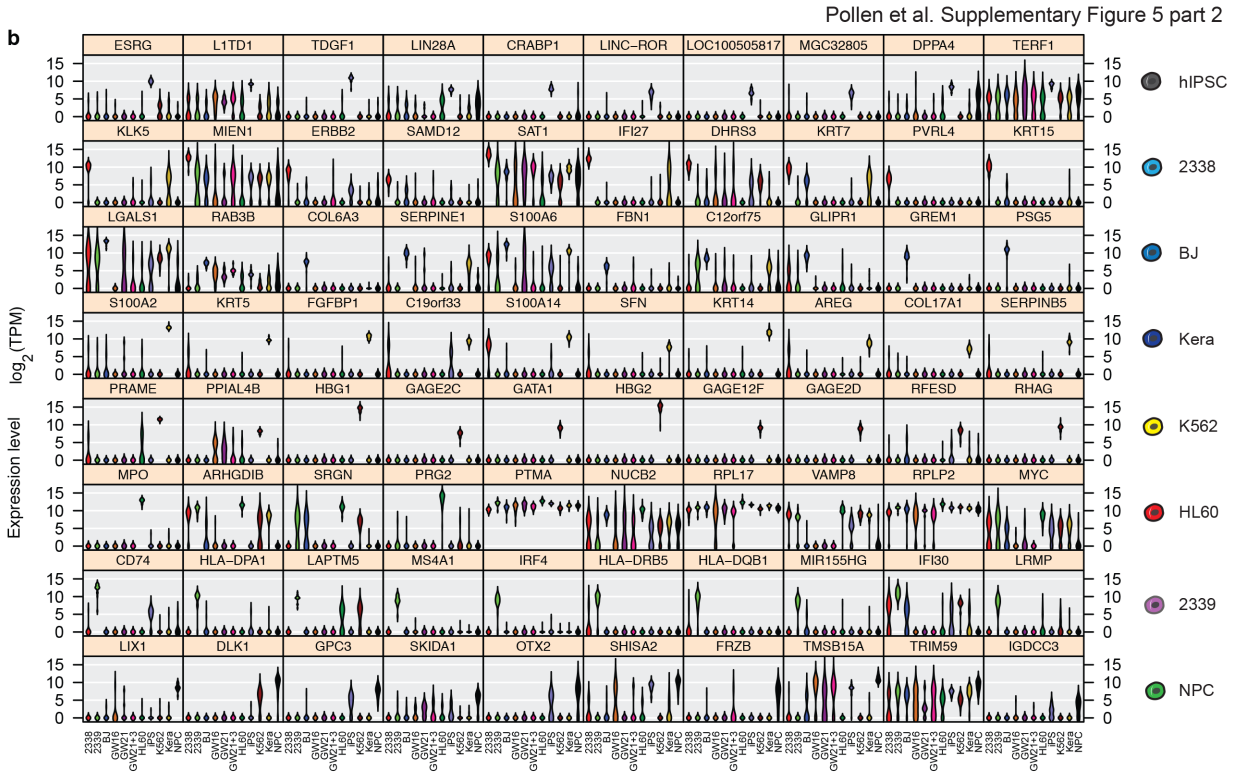
Supplementary Figure 3 PCA scree plot for low-coverage sequencing data for all analyzed cells. Genes driving PC1 and PC2 have the greatest relative contribution to the variation between analyzed cells.

Pollen et al. Supplementary Figure 4



Supplementary Figure 4 Principal component analysis distinguishes cell populations even at very low sequencing depths. Plots of sample scores in PC1 and PC2 for all 301 cells derived from (a) high-coverage data, (b) high-coverage data downsampled to 300,000 reads per cell, (c) low-coverage data (averaging 270,000 reads per cell), high-coverage data further downsampled to: (d) 100,000 reads, (e) 50,000 reads, (f) 10,000 reads, (g) 5,000 reads, (h) 1,000 reads, (i) 500 reads, and (j) 100 reads per cell. Notably, the broad distinctions between diverse cell types are still visible with fewer than 10,000 sequencing reads per cell.

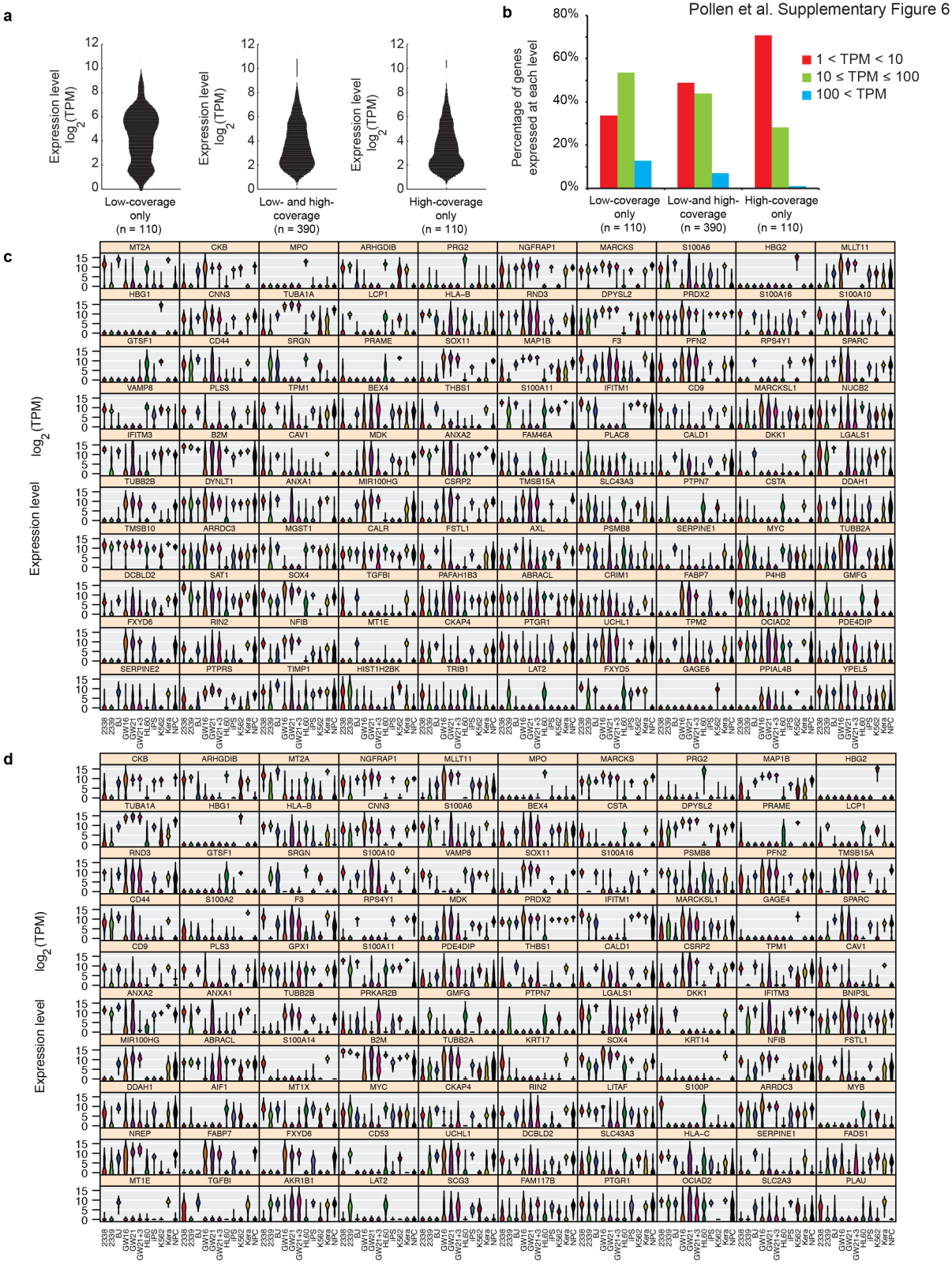

a
Expression Level
min
max
Genes
K562
HL60
2339
Kera
BJ
2338
NPC
GW16
GW21
GW21+3
hiPSC
Single cell samples



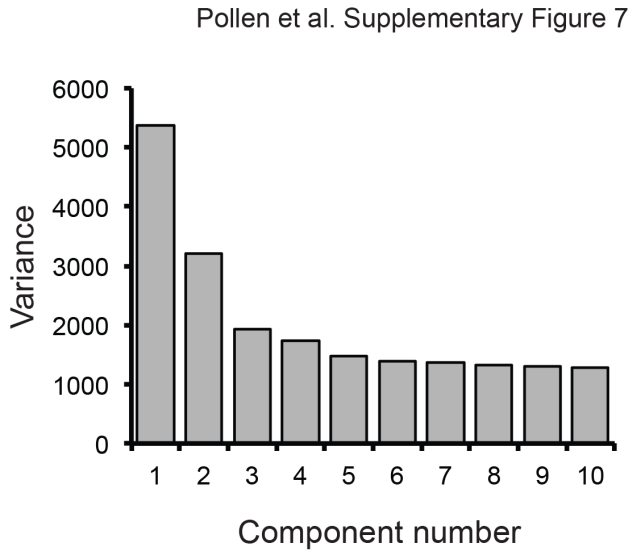
Supplementary Figure 5 Hierarchical clustering reveals specific gene expression profiles of different cell types. (a) Heatmap of expression data for 301 cells clustered using the 500 genes with the strongest PC1-3 gene loading scores derived from low-coverage mRNA sequencing data. Rows correspond to individual cell samples and columns correspond to genes. Five major gene clusters indicated with the colors: red, green, yellow, orange and cyan are revealed. Cells from three populations: HL60 cells, 2339 cells, and 2338 cells were captured using independent chips on different days and are highlighted in the right column of symbols next to the dendrogram with filled versus open circles. For 2338 cells and HL60 cells, the independent chips also included different capture site geometries (Supplementary Table 1). In all three cases, cells of the same origin collected using different chips clustered together. Within the HL60 group, the cells captured on a different day using a smaller capture site clustered more closely on the dendrogram (open red circles), while cells from different chips were largely interspersed along the dendrogram within their respective source groups (2338 cells - blue

circles and 2339 cells - purple circles), suggesting that plate-to-plate variation does not have a major impact on cell type classification. **(b)** Violin plots display the distribution of gene expression levels in each population studied for the 10 genes most strongly correlated with group memberships. Biomarkers defining groups from heterogeneous primary neural tissue are further examined in Supplementary Fig. 11.

Pollen et al. Supplementary Figure 6

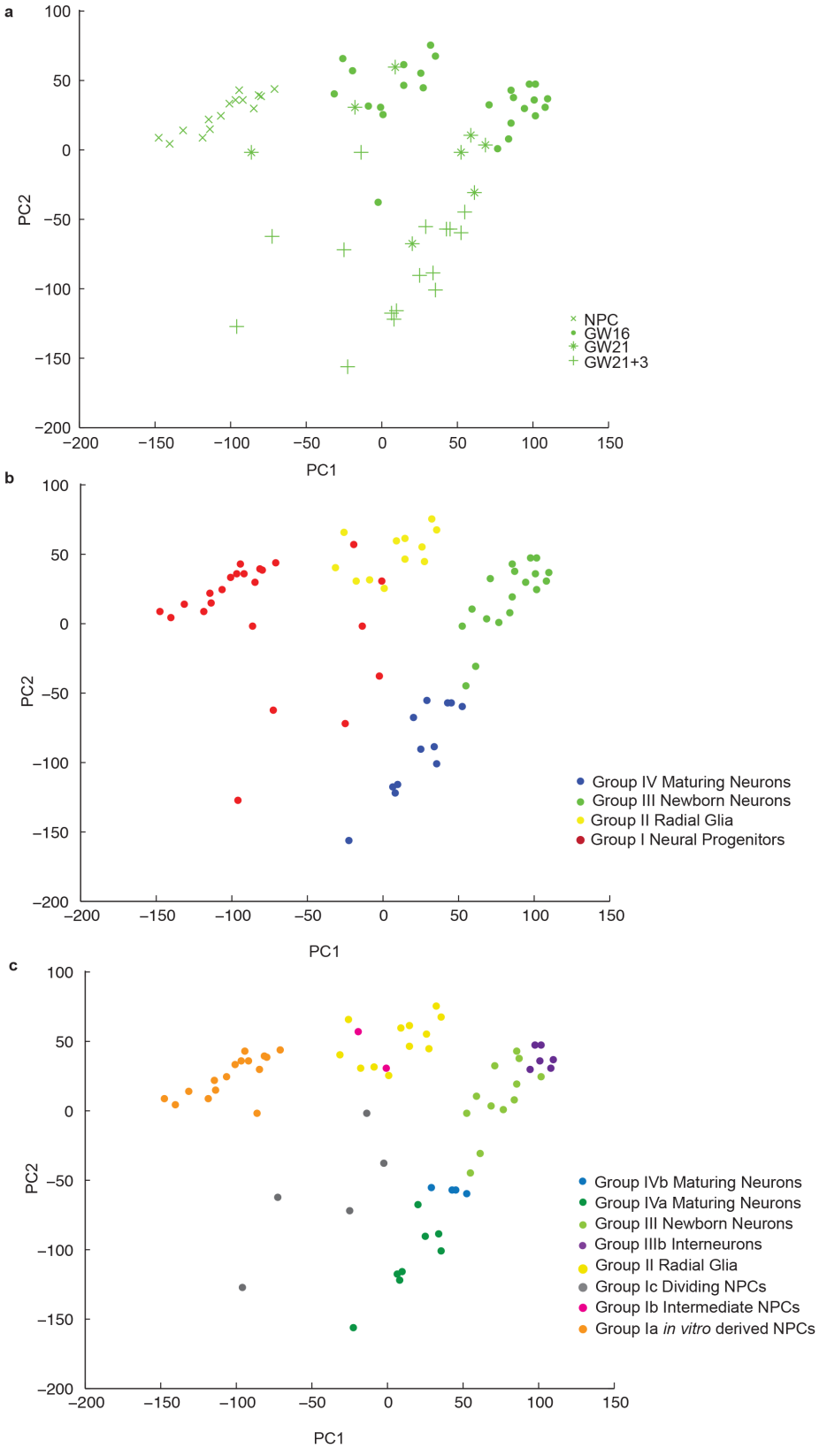


Supplementary Figure 6 Distribution of expression levels for genes best explaining the variation across single cells from distinct populations. (a-b) Distribution of expression levels for the 500 genes with the strongest PC1-3 gene loading scores in low-coverage and high-coverage datasets. Distributions are presented as violin plots (a) and histograms (b) that represent expression levels as determined by high-coverage sequencing. The majority (78%) of the top 500 PCA genes derived from high- and low-coverage datasets overlap (middle plots in a, b). The distribution of gene expression levels differs for the 22% of PCA genes that explain variation exclusively in the low-coverage data (left plots: a, b) and in the high-coverage data (right plots: a,b). Specifically, the high-coverage sequencing is more sensitive to detecting low-abundance PCA genes, while the low-coverage data relies on a greater proportion of a high-abundance genes that may explain less variation in the more sensitive high-coverage dataset. Although the majority of PCA genes are shared across depths, this analysis demonstrates a limitation for shallow sequencing to detect a fraction of the low-abundance genes important for variation across cells. **(c-d)** Violin plots representing the distribution of expression levels across each cell type are displayed for the top 100 PCA genes derived from analysis of low-coverage (c) and high-coverage (d) sequencing results. Some genes that appear to be low-abundance by averaging all 301 cells may actually be expressed at very high-levels in a single population (e.g., *HBG1, HBG2* in K562 erythroleukemia cells, *MPO* in HL60 promyelocytic leukemia cells, and *KRT14* in keratinocytes, violin plots).



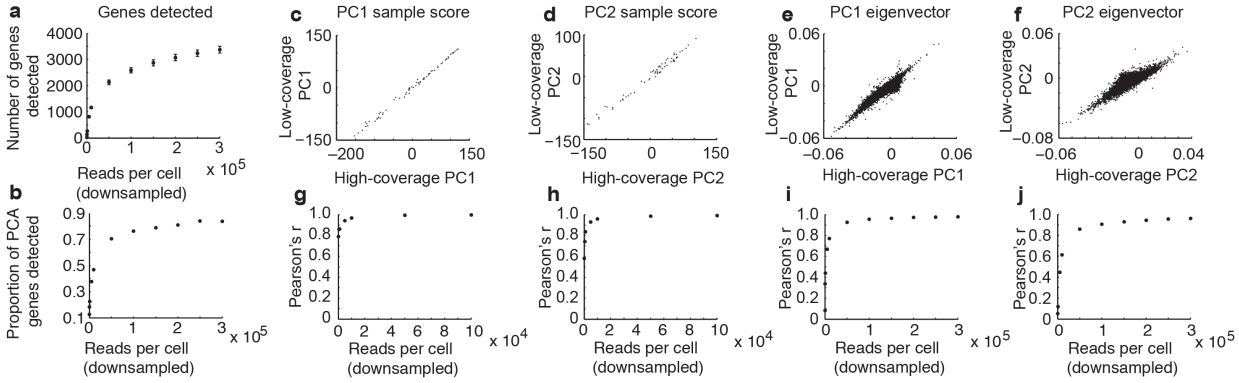
Supplementary Figure 7 The PCA scree plot for the low-coverage mRNA Seq of 65 single cells in the neural development lineage. Genes driving PC1 and PC2 have the greatest relative contribution to the variation between analyzed cells.

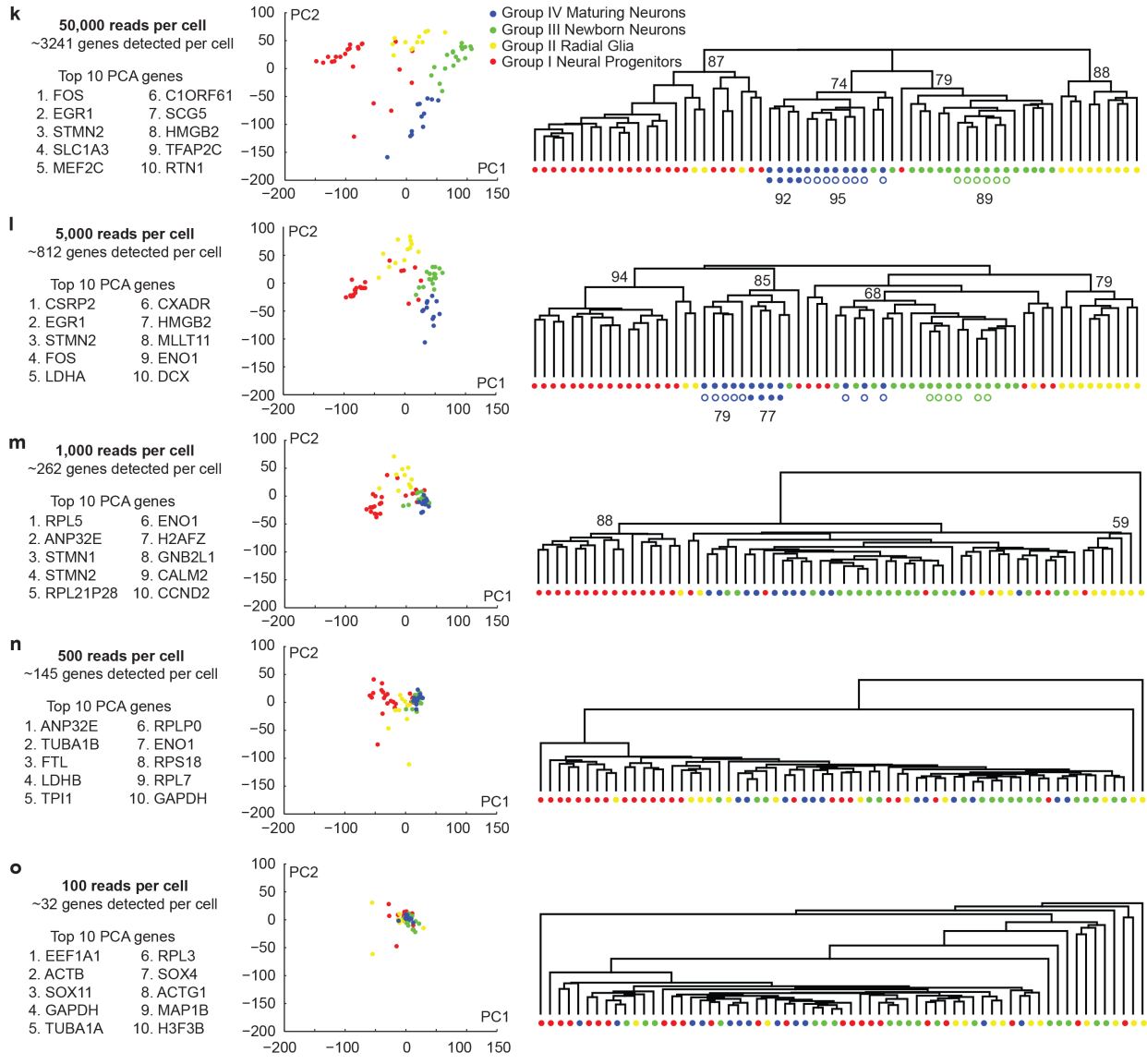
Pollen et al. Supplementary Figure 8



Supplementary Figure 8 Cells derived from samples representing various stages of neural development can be separated using low-coverage single-cell mRNA Seq. (a) PCA of low-coverage mRNA Seq data from 65 cells representing various stages of neural development. Cells are labeled based on source: NPCs, GW16 cortex, GW21 cortex, and GW21 cortex cultured under neurogenic conditions for three weeks (GW21+3) and are schematized in Fig. 3a. Based on this analysis, the 500 genes with highest PCA loading scores were used for hierarchical clustering of single cells in Fig. 3b. **(b)** Re-labeling of samples using the broad cell type classifications from Fig 3c. **(c)** Re-labeling of samples based on further analysis of subgroups in hierarchical clustering discussed in Supplementary Fig. 10.

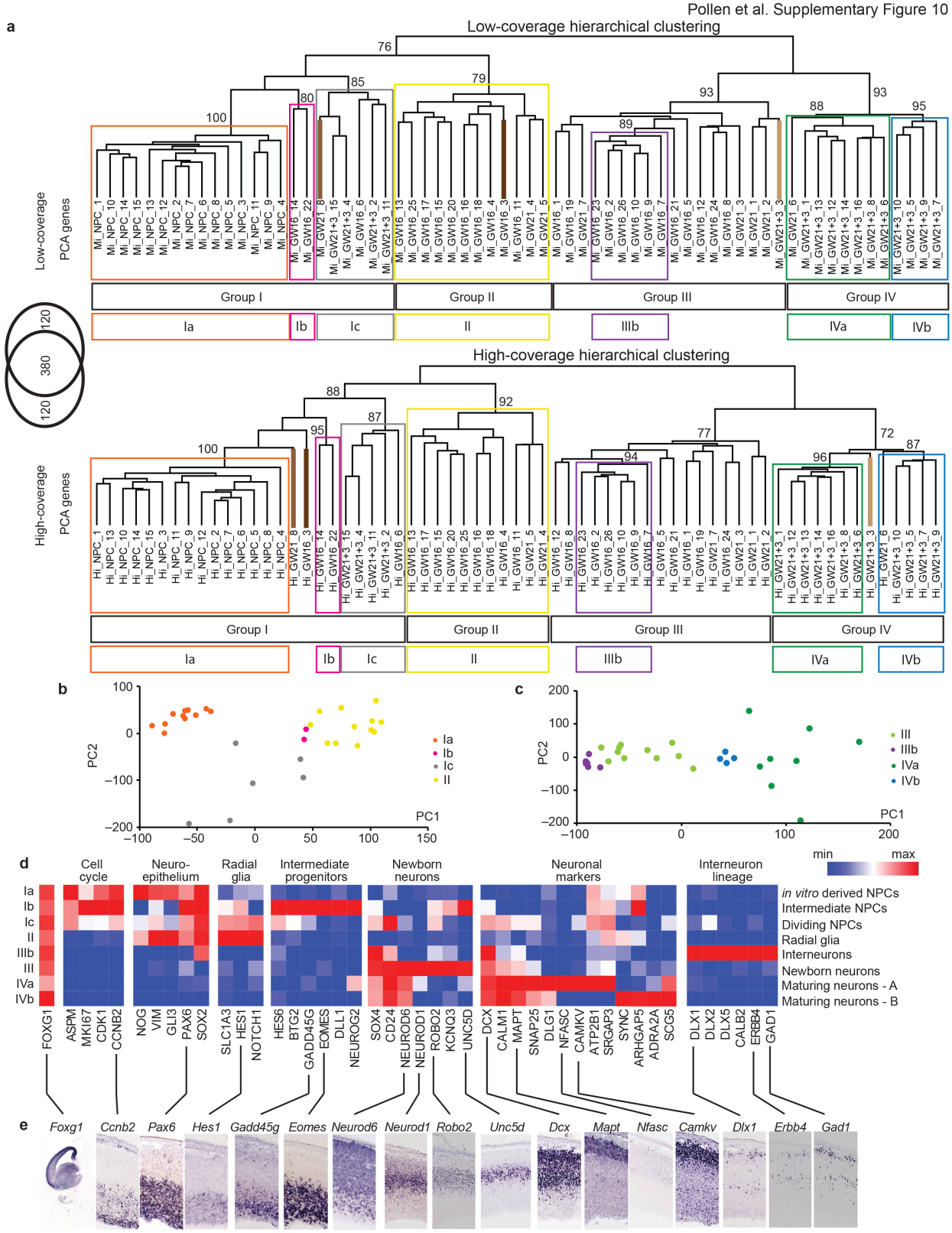
Pollen et al. Supplementary Figure 9





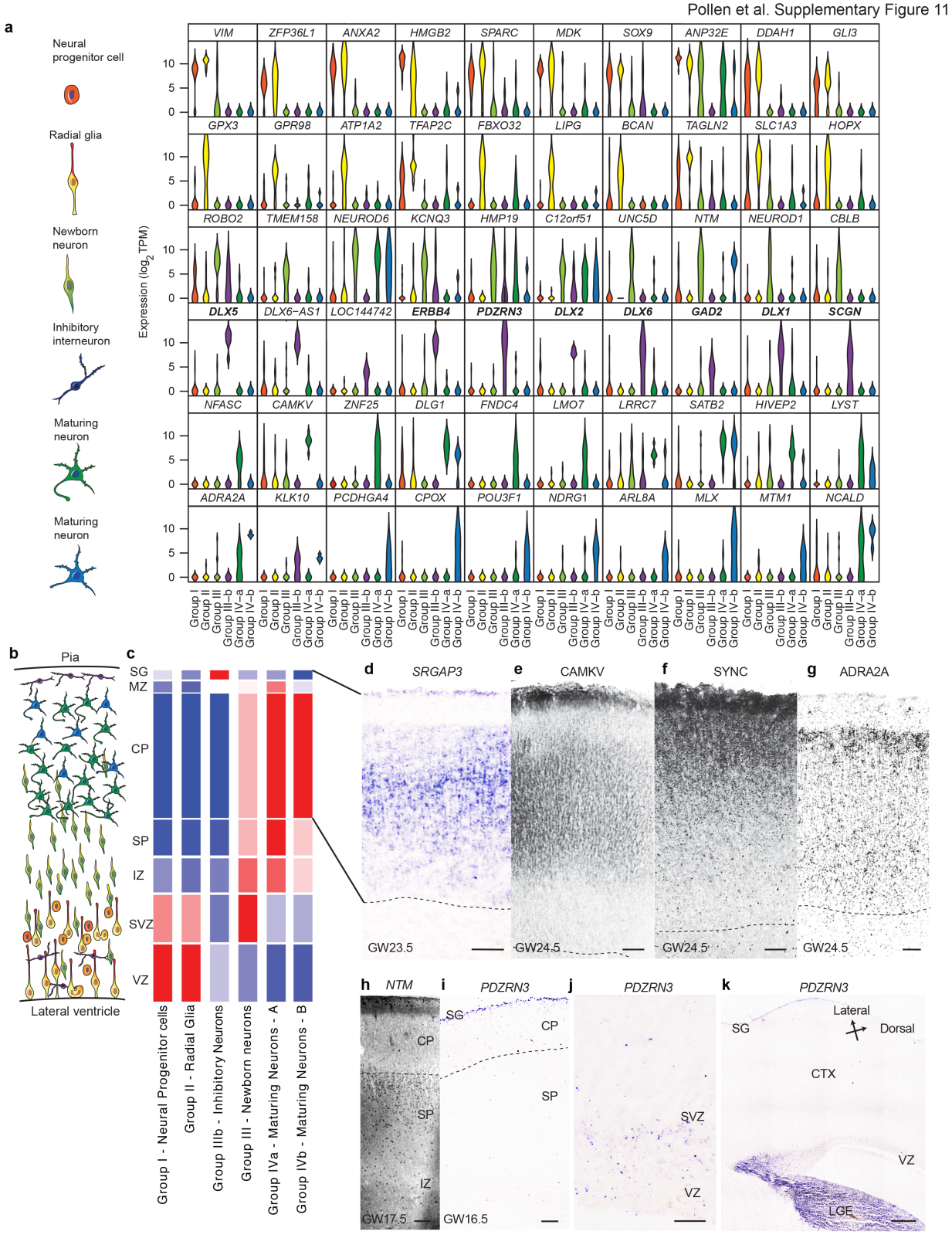
Supplementary Fig. 9 Neural cell type classification is possible at ultra low sequencing depths. (a) Plot showing the average total number of genes detected per cell at different downsampled sequencing depths across 65 neural cells. Error bars represent standard error of the mean. **(b)** Plot showing the fraction of the top 500 PCA genes identified in high-coverage data that are detected at each downsampled sequencing depth. PCA at 10,000 reads per cell identifies nearly half (46.8%) of the top 500 PCA genes derived from high-coverage data even though only about one quarter (25.9%) of all genes detected by high-coverage data are observed at this low depth. **(c-j)** PCA of high-coverage mRNA Seq data was compared with low-coverage PCA (c-f) and with PCA conducted on downsampled datasets (g-j). **(c-f)** Sample scores derived from low- and high-coverage data were calculated using eigenvectors from high-coverage data and correlated strongly across all 65 neural cells for (c) PC1 (r = 0.998) and (d) PC2 (r = 0.993). Results from low- and high-coverage datasets also correlate strongly for the eigenvectors defining (e) PC1 (r = 0.95) and (f) PC2 (r = 0.9). **(g-j)** Plots showing Pearson correlation coefficient between sample scores (g-h) and eigenvectors (i-j) derived from the PCA conducted on downsampled datasets and from the PCA conducted on high coverage mRNA Seq data. At 5,000 reads per cell, the PCA sample scores for PC1 (g) and PC2 (h) are strongly (r > 0.9) correlated with high-coverage PCA sample scores, indicating that cell type distinction is possible even at very low sequencing depths. Eigenvectors derived from downsampled data strongly correlate with the eigenvectors derived from high-coverage data at 50,000 reads (PC1: r = 0.93; PC2: r = 0.86). **(k-o)** Panels displaying (left to right) top 10 PCA genes, PCA sample plots and hierarchical clustering dendrograms for downsampled datasets corresponding to 50,000 reads per cell (k), 5,000 reads per cell (l), 1,000 reads per cell (m), 500 reads per cell (n), and 100 reads per cell (o). Samples in the dendrogram are colored according to cell type assignments in Figure 3b, and numbers indicate the approximately unbiased p-values calculated using pvclust for clusters of cells corresponding to groups I-IV and subgroups described in Supplementary Figure 10 (see Methods). **(k)** At 50,000 reads, finer distinctions

within newborn and maturing neuron categories observed at high-coverage remained detectable: immature inhibitory neurons (open green circle), maturing neuron-A (open blue circle), maturing neurons-B (filled blue circle). **(l)** At 5,000 reads, the finer distinctions begin to blur, but the four major groups of cell types remain visible. Notably, *EGR1* and *FOS*, which are the top-ranked PCA genes in the PCA at 50,000 reads per cell, are still detected in among top 10 PCA genes. **(m)** At 1,000 reads, the distinction between newborn and maturing neurons is lost, but distinct progenitor, radial glia, and neuronal groups are visible. Unbiased identification of finer distinctions between cell types is also likely to be influenced by the number of cells surveyed from each population and the heterogeneity within each category. The ability to classify cells at ultra low depths anticipates the utility of cellular barcoding and cell capture strategies that retain cell-of-origin information for transcripts sampled from tens of thousands of cells derived from heterogeneous tissue.


a
Low-coverage hierarchical clustering
Low-coverage PCA genes
Group I
Group II
Group III
Group IV
Ia
Ib
Ic
II
IIIb
IVa
IVb
High-coverage hierarchical clustering
High-coverage PCA genes
b
PC1
PC2
c
d
Cell cycle
Neuro-epithelium
Radial glia
Intermediate progenitors
Newborn neurons
Neuronal markers
Interneuron lineage
min
max
in vitro derived NPCs
Intermediate NPCs
Dividing NPCs
Radial glia
Interneurons
Newborn neurons
Maturing neurons - A
Maturing neurons - B
FOXG1
ASPM
MKI67
CDK1
CCNB2
NOG
VIM
GLI3
PAX6
SOX2
SLC1A3
HES1
NOTCH1
HES6
BTG2
GADD45G
EOMES
DLL1
NEUROG2
SOX4
CD24
NEUROD6
NEUROD1
ROBO2
KCNQ3
UNC5D
DCX
CALM1
MAPT
SNAP25
DLG1
NFASC
CAMKV
ATP2B1
SRGAP3
SYNC
ARHGAP5
ADRA2A
SCG5
DLX1
DLX2
DLX5
CALB2
ERBB4
GAD1
e
Foxg1
Ccnb2
Pax6
Hes1
Gadd45g
Eomes
Neurod6
Neurod1
Robo2
Unc5d
Dcx
Mapt
Nfasc
Camkv
Dlx1
Erbb4
Gad1

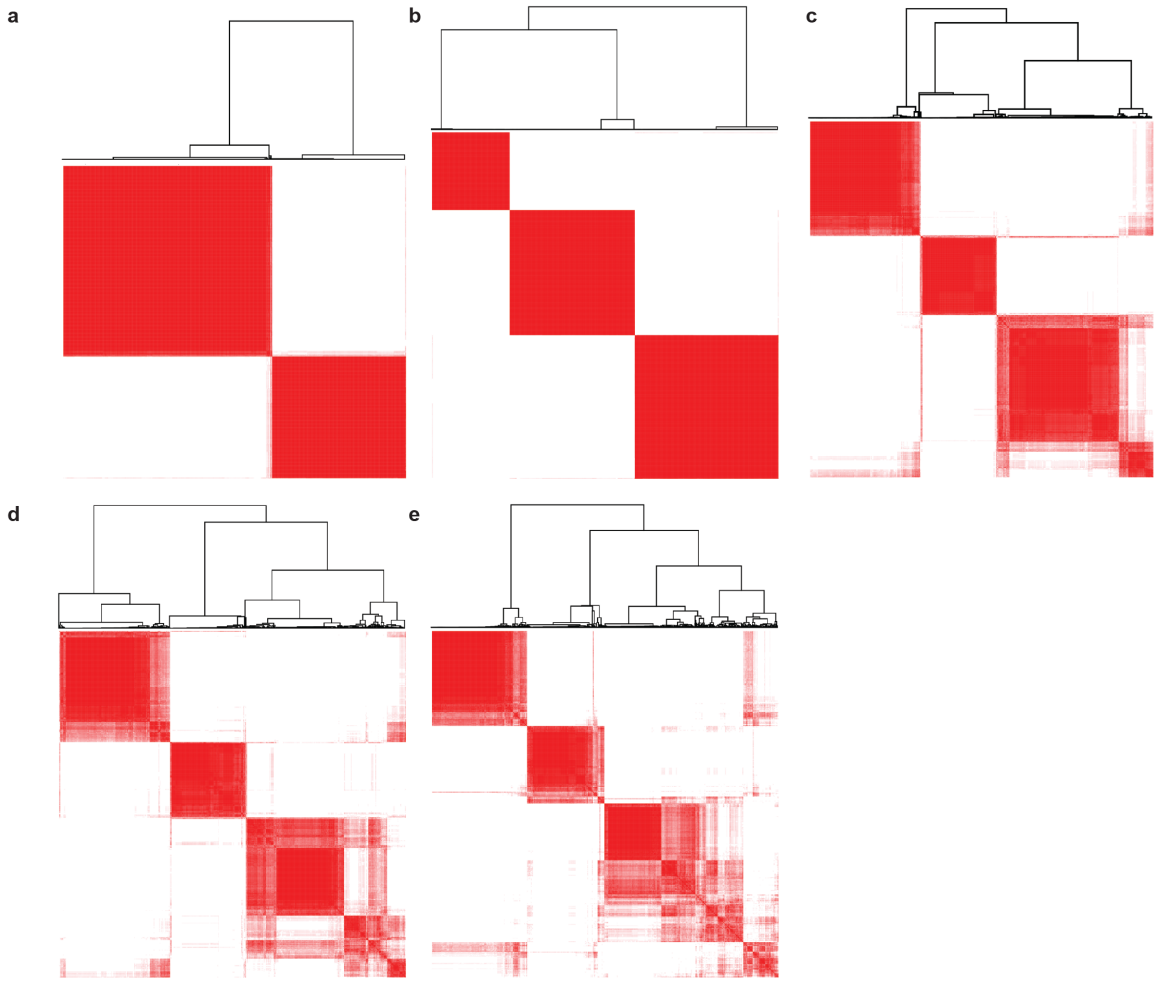
Supplementary Figure 10 Classification of 65 neural cells reveals additional cellular diversity within broad categories. (a) The sets of genes that explain variation between the same 65 neural cells sequenced at low- and high-coverage are largely overlapping: 380 of the 500 genes with the strongest loading scores in PC1-3 are shared (Venn diagram). Hierarchical clustering of 65 neural cells across the top 500 PCA genes derived from low-coverage mRNA sequencing (top dendrogram) reveals distinct groups of cells that are largely similar to the groups derived from clustering cells according to the gene expression values and top 500 PCA genes derived from high-coverage mRNA sequencing (bottom dendrogram). Sample names and cell type assignments are listed below the dendrogram, and three cells that transition between groups are highlighted in brown. Based on the expression of known marker genes, groups I-IV correspond to major classes of cells in the developing neocortex (Fig. 3b-c). However, additional subgroups highlighted with colored rectangles in dendrogram may correspond to meaningful distinctions of additional cellular diversity. Numbers above nodes of the dendrograms indicate approximately unbiased p-values calculated using pvclust (see Methods). **(b)** PCA plots of low-coverage mRNA Seq data from 35 cells expressing markers of neural progenitor cells and radial glia (Fig. 3a-c) that formed groups I and II and subgroups highlighted in the dendrogram in a. **(c)** PCA plot of low-coverage mRNA Seq data from 30 cells that expressed markers of newborn and maturing neurons (Fig. 3a-c) and formed groups III, IIIb, IVa, and IVb. **(d)** Heatmap of gene expression levels averaged across cells of each group (identified in a) for marker genes selected to examine additional cellular diversity and for genes distinguishing cell groups. Bottom row shows select *in situ* hybridization from developing mouse cortex to illustrate the distinct and overlapping expression of select marker genes[1]. All cell groups expressed *FOXG1* consistent with their telencephalic origin[2]. Group I cells strongly expressed genes important for cell division[3-6]. Cells from group Ia and group II shared strong expression for neuroepithelial (*NOG*, *SOX2*, *VIM*) [7-9] and cortical (*PAX6* and *GLI3*) [10, 11] markers, but only group II cells expressed mature radial glia markers (*SLC1A3*, *HES1,* and *NOTCH1*) [12-

[14]. Although group Ib consisted of only two cells, these cells clustered together in analysis of low- and high coverage data (a), and these cells expressed markers of intermediate neural progenitors (*HES6*, *BTG2*, *GADD45G*, *EOMES*, *DLL1,* and *NEUROG2*) [15, 16], which are dividing cells that are born from radial glia and amplify their neuronal output by undergoing divisions in the SVZ. Group Ic contained cells that expressed markers of cell division but also expressed early markers of neuronal fate including CD24 and SOX4. Group III contained 18 cells that expressed markers of newborn neurons with minimal expression of progenitor markers. Within group III, six cells clustered together in both dendrograms (a), and in downsampled dataset at 50,000 reads per cell (Supplementary Figure 9k), and in PCA space in c. These cells expressed multiple markers of inhibitory interneurons, which are produced by the germinal zone of the ventral telencephalon and migrate tangentially to populate the cortex[17]. The remaining 12 neurons in group III are likely to be newborn excitatory neurons and were largely captured from the GW16 germinal zone. These cells were characterized by the expression of proneural factors such as *NEUROD1*, *NEUROD6*, and *SOX4,* as well as *UNC5D* a gene transiently upregulated in postmitotic newborn neurons that regulates the earliest phase of neuronal migration[18], and the gene *ROBO2* which may be involved in regulating radial migration of cortical neurons[19, 20]. *NEUROD1, ROBO2,* and *UNC5D* are expressed in the intermediate zone of the mouse cortex but not the cortical plate (*in situs,* bottom row) suggesting that these cells correspond to transient developmental intermediates in neurogenesis with distinct transcriptional profiles. Group IV cells express markers of young neurons such as *DCX,* but also early markers of neuronal maturation including *MAPT*, *SNAP25*, and *MEF2C* [21-25]. Interestingly, two putative subgroups of maturing neurons classified in both low- and high-coverage datasets differentially expressed *CAMKV* and *ADRA2A*, which have not been previously described in the context of developmental human neuronal diversity.

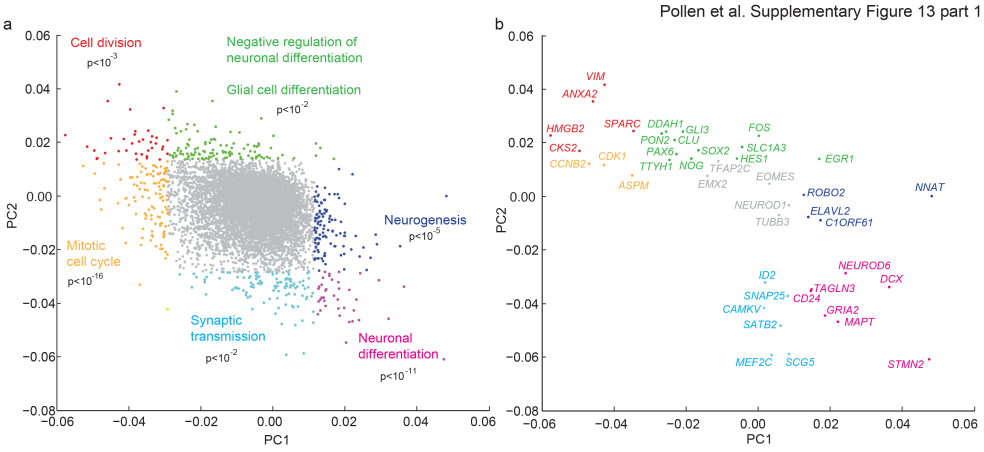

a
Neural progenitor cell
Radial glia
Newborn neuron
Inhibitory interneuron
Maturing neuron
Maturing neuron
Expression (log2 TPM)
VIM
ZFP36L1
ANXA2
HMGB2
SPARC
MDK
SOX9
ANP32E
DDAH1
GLI3
GPX3
GPR98
ATP1A2
TFAP2C
FBXO32
LIPG
BCAN
TAGLN2
SLC1A3
HOPX
ROBO2
TMEM158
NEUROD6
KCNQ3
HMP19
C12orf51
UNC5D
NTM
NEUROD1
CBLB
DLX5
DLX6−AS1
LOC144742
ERBB4
PDZRN3
DLX2
DLX6
GAD2
DLX1
SCGN
NFASC
CAMKV
ZNF25
DLG1
FNDC4
LMO7
LRRC7
SATB2
HIVEP2
LYST
ADRA2A
KLK10
PCDHGA4
CPOX
POU3F1
NDRG1
ARL8A
MLX
MTM1
NCALD
Group I
Group II
Group III
Group IIIb
Group IVa
Group IVb
b
Pia
Lateral ventricle
c
SG
MZ
CP
SP
IZ
SVZ
VZ
Group I - Neural Progenitor cells
Group II - Radial Glia
Group IIIb - Inhibitory Neurons
Group III - Newborn neurons
Group IVa - Maturing Neurons - A
Group IVb - Maturing Neurons - B
d
SRGAP3
GW23.5
e
CAMKV
GW24.5
f
SYNC
GW24.5
g
ADRA2A
GW24.5
h
NTM
SG
CP
SP
IZ
GW17.5
i
PDZRN3
SG
CP
SP
GW16.5
j
PDZRN3
SVZ
VZ
k
PDZRN3
Lateral
SG
Dorsal
CTX
VZ
LGE

Supplementary Figure 11 Expression patterns of biomarkers for neural cell types identified by single cell analysis. (a) Additional cellular diversity recognized by interpretation of subgroups in Supplementary Fig. 10a. Violin plots display the distribution of gene expression levels in each interpreted cell type for the 10 genes most strongly correlated with group memberships. Notably, 8/10 biomarkers for the six cells classified as immature inhibitory neurons in group IIIb (bold gene names above violin plots) overlap with the 15 top-ranked interneuron markers identified by gene coexpression analysis in a recent large-scale survey of hundreds of heterogeneous regions of the prenatal human neocortex[26] (module C31). (b) Schematic of the position of cell types in developing human cortex. **(c)** This large-scale survey of gene expression from precisely microdissected tissue samples[26] provides the opportunity to examine the expression patterns of candidate cell-type specific biomarkers in distinct regions of developing human cortex. Heatmap represents the average expression levels of 20 genes most strongly correlated with each group membership across 695 regions of the developing human cortex from seven distinct laminae: VZ – ventricular zone, SVZ – subventricular zone, IZ – intermediate zone, SP – subplate, CP – cortical plate, MZ– marginal zone, SG – subpial granular layer. Biomarkers for neural progenitors (group I) and radial glia (group II) show strongest expression in the germinal zones, where these cells are common. Interestingly, progenitor cell biomarkers also show moderate expression in the subpial granular layer. Biomarkers for immature inhibitory neurons (group IIIb) show strongest expression in the subpial granular layer, likely corresponding to a population of interneurons that migrates from the ganglionic eminences. Additionally, the interneuron biomarkers show moderate expression in the ventricular zone, corresponding to the region where the interneurons in this study were captured. Biomarkers for the remaining cells in group III show strongest expression in the subventricular zone and intermediate zone where newborn neurons begin their earliest phases of radial migration towards the cortical plate. Maturing neuron (group IVa, IVb) biomarkers show strongest expression outside of the germinal zone, but strong expression of group IVb

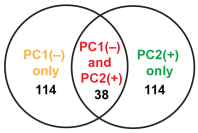
biomarkers is more restricted to the cortical plate. **(d-k)** Candidate biomarkers show complementary expression patterns in developing human cortex. **(d)** *In situ* hybridization in GW23.5 cortical plate for *SRGAP3*, a gene with strong expression in group IVa, but not group IVb (Supplementary Fig. 10, heatmap), shows strongest expression in the middle of the cortical plate (layers II-V). **(e)** Similarly, immunostaining for *CAMKV*, a group IVa-specific marker (violin plots) shows strongest expression in deep cortical layers. **(f)** In contrast, immunostaining for *SYNC,* a gene with strong expression in group IVb, but not group IVa (Supplementary Fig. 10, heatmap), shows strongest expression in upper cortical layers I-III. **(g)** Similarly, immunostaining for *ADRA2A,* a gene with expression across most mature neurons, but with strongest expression in group IVb (violin plots), shows strongest expression in upper cortical layers II-III, but also moderate expression across all cortical layers, including layer VI. **(h)** Immunostaining for NTM, a candidate marker of newborn neurons, reveals specific expression in a subset of cells in regions of cortical neuron migration: the intermediate zone and subplate, but limited expression in the cortical plate. **(i-k)** *In situ* hybridization for *PDZRN3*, a candidate marker of immature interneurons, specifically labels a subset of cells in the subpial granular layer (i) and the ventricular zone, where cells in this study were captured (j), with broad staining in the lateral GE, a zone of interneuron production (k).



Supplementary Figure 12 Consensus clustering of top 500 genes explaining variation between neural cells optimally identifies three gene clusters. Consensus matrices clustering the 500 PCA genes best explaining the variation between 65 cells in the neural lineage were generated using k-means clustering for (a) 2, (b) 3, (c) 4, (d) 5 (e) 6 clusters. After visual inspection of the consensus matrices we concluded that three gene clusters optimally group the top 500 PCA genes.



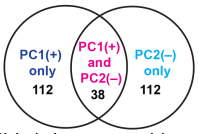
c



Biological processes enrichment

Name and GO ID	p-value	Name and GO ID	p-value	Name and GO ID	p-value
Mitotic cell cycle GO:0000278	6×10^{-17}	Cell division GO:0051301	2×10^{-4}	Negative regulation of canonical Wnt receptor signaling pathway GO:0090090	2×10^{-3}
Mitotic cell cycle checkpoint GO:0007093	5×10^{-15}	Mitotic cell cycle GO:0000278	6×10^{-3}	Negative regulation of neuron differentiation GO:0045665	5×10^{-3}
Cell cycle process GO:0022402	2×10^{-14}	Nuclear division GO:0000280	9×10^{-3}	Oligodendrocyte differentiation GO:0048709	8×10^{-3}
Negative regulation of mitotic cell cycle phase transition GO:1901991	6×10^{-14}	Mitosis GO:0007067	9×10^{-3}	Glial cell differentiation GO:0010001	1×10^{-2}

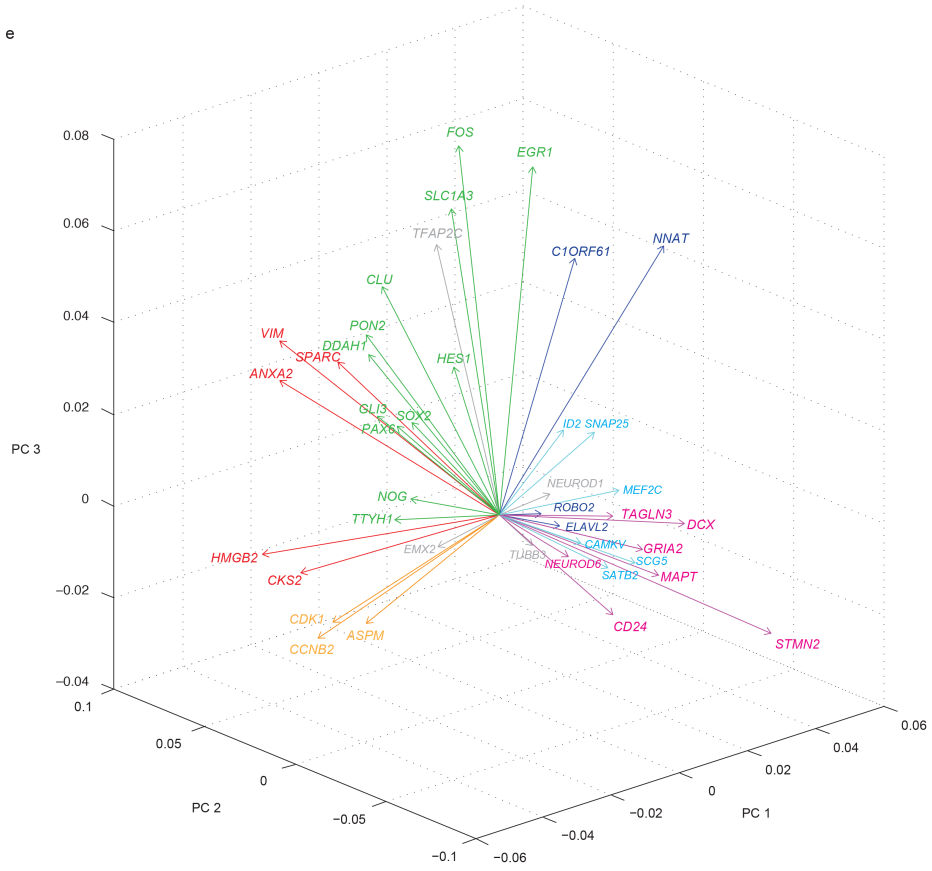
d



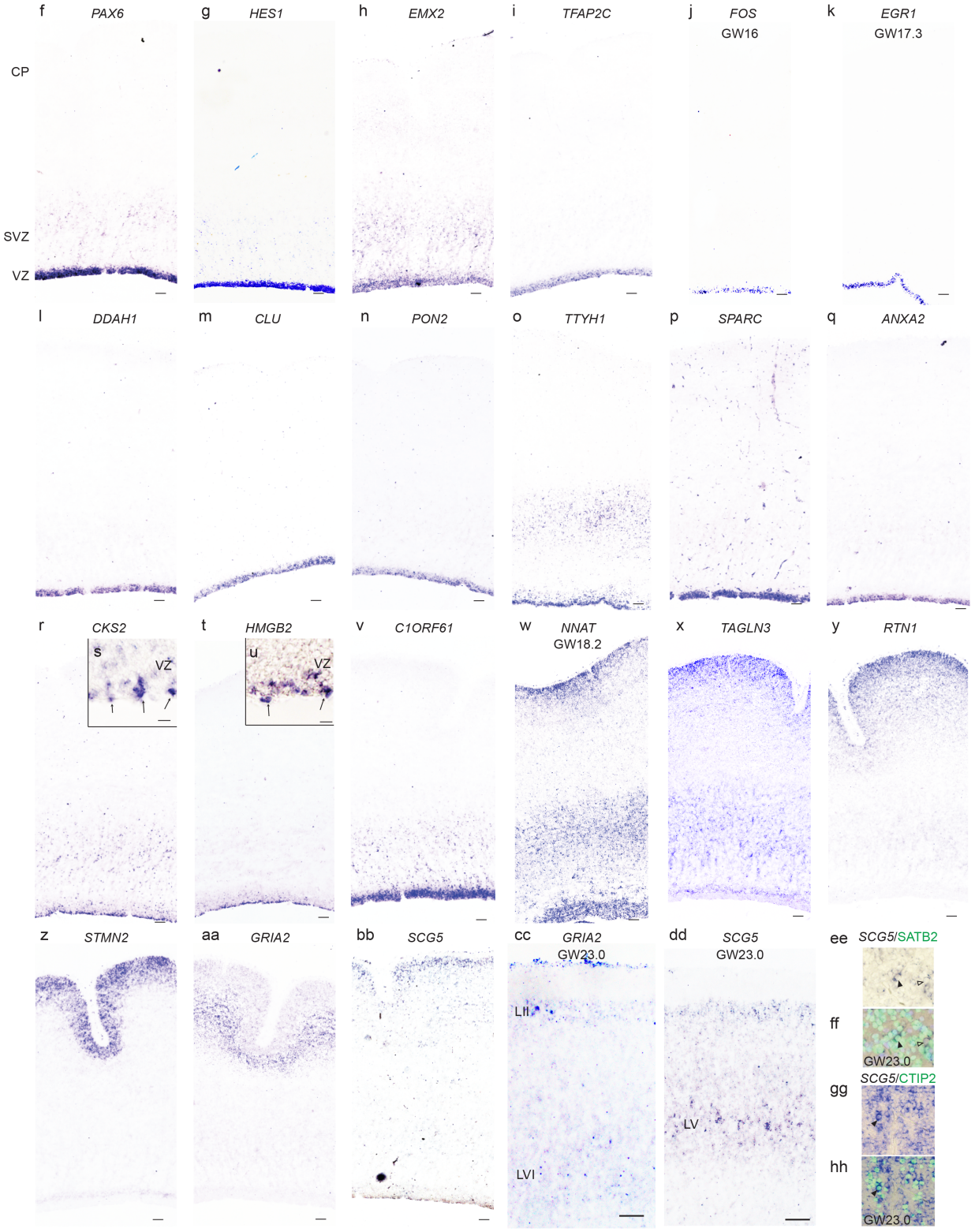
Biological processes enrichment

Name and GO ID	p-value	Name and GO ID	p-value	Name and GO ID	p-value
Neuron differentiation GO:0030182	1×10^{-6}	Neuron differentiation GO:0030182	2×10^{-12}	Synaptic transmission GO:0007268	3×10^{-3}
Neurogenesis GO:0022008	4×10^{-6}	Generation of neurons GO:0048699	9×10^{-12}	Multicellular organismal signaling GO:0035637	5×10^{-3}
Generation of neurons GO:0048699	8×10^{-6}	Neuron development GO:0048666	1×10^{-11}	Cell-cell signaling GO:0007267	1×10^{-2}
Neuron projection development GO:0031175	8×10^{-6}	Neuron projection development GO:0031175	2×10^{-11}	Transmission of nerve impulse GO:0019226	2×10^{-2}

e



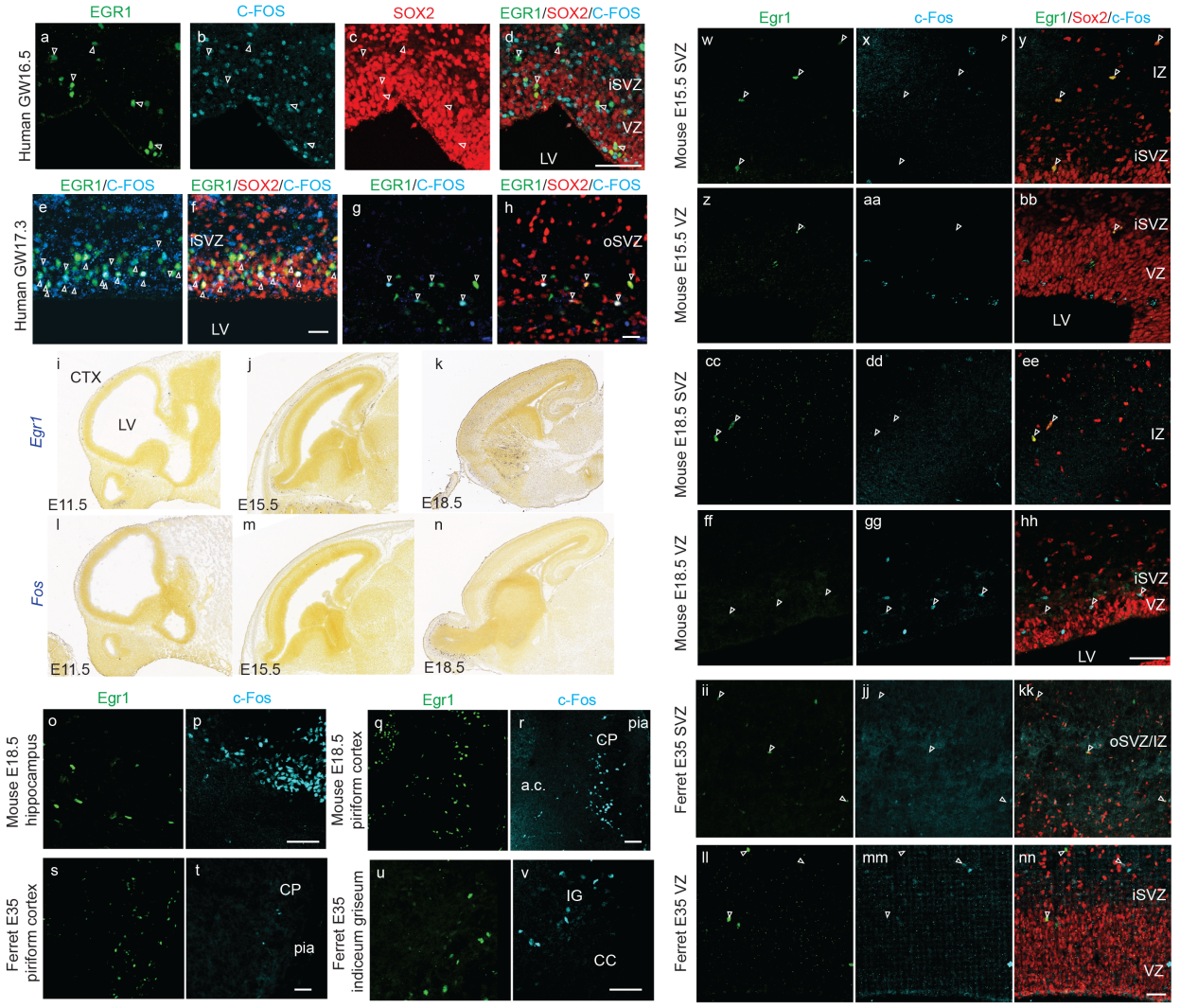
Pollen et al. Supplementary Figure 13 part 2



Supplementary Figure 13 PCA gene loading scores reveal important markers of cell types and states in the developing human cortex. (a) Plot of the PCA gene loading scores for all genes along PC1 and PC2 based on analysis of low-coverage mRNA Seq data. The 150 genes with the greatest gene loading scores are highlighted separately for positive and negative PC1 and PC2 values in colors based on the sign of the gene loading score, and the corresponding top biological process enrichment determined by ToppGene Suite[27] are listed along with Bonferroni corrected p-values. **(b)** Selected markers of neuroepithelial and radial glial cells (*PAX6*, *GLI3*, *SOX2*, *HES1*, *SLC1A3*, *VIM*, *EMX2*), dividing cells (*ASPM*, *CDK1*, *CCNB2*), intermediate progenitors (*EOMES*), early neuronal commitment (*TUBB3*, *DCX*, *CD24, ELAVL2*) and maturing neurons (*MAPT*, *SATB2*, *SNAP25*, *MEF2C*) are highlighted in the PCA gene loading plot along with a subset of nearby genes. Colors correspond to gene groups highlighted in a. **(c-d)** Venn diagrams represent overlap between the 150 genes with the most negative PC1 loading values and the 150 genes with the most positive PC2 loading values (c) and between the 150 genes with the most negative PC2 values and most positive PC1 values (d). Tables in c and d represent the most significant biological process gene ontology enrichments determined by ToppGene Suite. Notably, PC1 appears to correspond to an axis from mitotic cell cycle (negative loading) to neuronal differentiation (positive loading), while PC2 corresponds to an axis from radial glia markers (positive loading) to neuronal maturation (negative loading). **(e)** Three dimensional plot of PCA gene loading scores for selected genes along PC1, PC2, and PC3. Notably, PC3 further distinguishes a subset of known and novel radial glia markers including *FOS* and *EGR1* from other genes. **(f-hh)** Expression of candidate novel markers related to categories identified in a–e was analyzed in the developing human cortex by *in situ* hybridization. Classical markers of radial glia including *PAX6* (f), *HES1* (g), and *EMX2* (h) are strongly expressed in the germinal regions, the ventricular zone (VZ) and the subventricular zone (SVZ), where radial glia are abundant. Genes not previously characterized in the developing human cortex located nearby these classical radial glia markers in the PC plot show

similar expression patterns: *TFAP2C* (i), *FOS* (j), *EGR1* (k), *DDAH1* (l), *CLU* (m), *PON2* (n) and *TTYH1* (o). Interestingly, many of the novel radial glia markers were more restricted to the VZ, which was recently observed for other novel biomarkers of radial glia[26]. **(p-q)** Both *SPARC* (p) and *ANXA2* (q) are highly expressed in the VZ and occupy a position in PCA space nearby the neuroepithelial marker *VIM,* with a negative value for PC1 loading and a positive value for PC2 and PC3 loading (e). **(r-u)** In contrast, *CKS2* (r) and *HMGB2* (t) share strongly negative PC1 loading scores, but also have low scores for PC2 and PC3 and are nearby numerous markers of cell division in the PC plot, including *CDK1*, *CCNB2*, and *ASPM.* Both of these candidate mitotic markers are expressed in a narrow domain at the edge of the ventricular surface (see higher magnification inset images s for *CKS2* and u for *HMGB2*; arrows mark dividing cells), where ventricular radial glia undergo mitosis. **(v-w)** In contrast to many genes with positive PC1 loading scores, *C1ORF61* (v) and *NNAT* (w) have neutral PC2 loading scores and strongly positive PC3 scores, falling closer in PCA space to radial glia genes. Interestingly, the expression of both genes is enriched in the germinal zones, but *NNAT* which also has the strongest positive PC1 loading score of all genes (toward the neurogenesis enriched quadrant) is also detected in the intermediate zone and the cortical plate (CP), zones of newborn neuron migration and early maturation. **(x-z)** Genes nearby known neural differentiation markers with a negative PC2 loading score and a positive or neutral PC1 score are expressed most strongly in the CP: *TAGLN3* (x), *RTN1* (y), *STMN2* (z). **(aa-bb)** At GW14.5, corresponding to peak layer V neurogenesis[28], we detected complementary expression of *SCG5* (aa) and *GRIA2* (bb) in upper and lower portions of the cortical plate. **(cc-hh)** In the cortical plate towards the end of cortical neurogenesis, *GRIA2* (cc) and *SCG5* (dd) were both detected in layer II, but *GRIA2* was also expressed in layer VI, while *SCG5* was expressed in columnar patterns within layer V. Pyramidal neurons of layer V project to diverse targets. We observed that a fraction of cells immunoreactive for SATB2, which marks callosal projection neurons in the mouse cortex also expressed *SCG5* (ee-ff), while numerous cells immunoreactive for CTIP2, a marker of

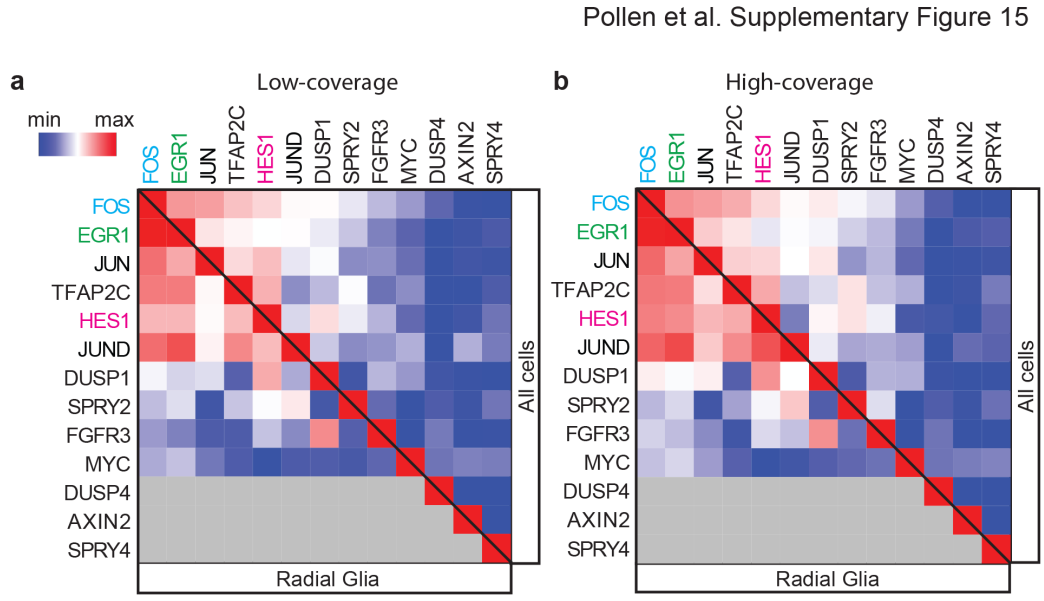
subcortical projections, expressed *SCG5* (gg-hh), highlighting the additional cellular and molecular diversity that emerges in the developing brain and will likely be amenable to efficient large-scale surveys of single cell gene expression. Scale bar is 100 μm except for s and u where scale bar is 25 μm.



Supplementary Figure 14 Expression domains of EGR1 and FOS in the developing cortex across species. (a-h) Numerous SOX2 positive cells in the germinal zones of the human cortex express EGR1 and C-FOS. **(a-d)** Composite channels of the image of human GW16.5 ventricular zone (VZ) from Fig. 4b showing staining for EGR1 (a), C-FOS (b), SOX2 (c), and an overlay image (d); LV - lateral ventricle. **(e-h)** Images of the VZ (e-f) and subventricular zone (SVZ, g-h) of GW17.3 human fetal cortex showing double labeling for EGR1 and C-FOS (e, g) and overlay images with staining for SOX2 (f, h). Arrows indicate triple-labeled cells in panel a-h. We note that there are many uncontrolled variables in the preparation of

primary human tissues compared with the acute preparation of samples from model organisms. Some of these variables could influence gene expression. Therefore, we reproduced the staining patterns for EGR1 and C-FOS immunohistochemistry,(a-h, Fig. 4e) as well as *EGR1* and *FOS in situ* hybridization experiments (Fig. 4a-d, Supplementary Figure 13j-k), in five independent samples, including two acutely fixed samples (data not shown). **(i-n)** Because expression of EGR1 and C-FOS in radial glia had not been previously reported, we examined the expression of these genes in publicly available mouse *in situ* hybridzation data across phases of neurogenesis (Allen Developing Mouse Brain Atlas: http://developingmouse.brain-map.org). *Egr1* (i-k) and *Fos* (l-n) were not readily detectable in the germinal zones of the mouse cortex during neurogenesis suggesting that these genes are unlikely to be involved in mouse neocortical progenitor cell development. **(o-nn)** We further examined the expression of Egr1 and c-Fos in the developing mouse and ferret brain by immunohistochemistry. **(o-v)** We found that the antibodies used in human also detected specific cell populations of Egr1 (o, q) and c-Fos (p, r) immunoreactive cells in the developing mouse hippocampus (o-p) and piriform cortex (q-r) at E18.5, consistent with previous reports about the roles of these proteins in development of neuronal circuitry in these regions[29, 30]: a.c. - anterior commissure. Similarly, in the developing ferret brain at E35 we found numerous Egr1 (s, u) and c-Fos (t, v) expressing cells in the piriform cortex (s, t) and indiceum griseum (IG, u, v): cc - corpus callosum. **(w-nn)** In contrast, Sox2-expressing cells in mouse or ferret were rarely immunoreactive for Egr1 and c-Fos during phases of neurogenesis. **(w-bb)** Images of the SVZ (w-y) and VZ (z-bb) of E15.5 mouse cortex immunostained for Egr1 (w, z), c-Fos (x, aa) and overlay images with Sox2 expression in red (y, bb). Open white arrows indicate sporadic examples of Sox2 expressing cells that are immunopositive for Egr1, but do not stain for c-Fos. **(cc-ee)** Immunostaining of mouse E18.5 subventricular zone / intermediate zone (SVZ/IZ) for Egr1 (cc), c-Fos (dd), and overlay with Sox2 (ee) showing a subset of Sox2 positive cells expressing Egr1 (open white arrows). **(ff-hh)** Representative images of mouse E18.5 VZ showing immunostaining for Egr1

(ff), c-Fos (gg) and overlay image with immunostaining for Sox2 (hh). Open white arrows indicate examples of Sox2 expressing cells that express c-Fos but not Egr1 **(ii-nn)** Images of the ferret cortical outer SVZ (ii) showing immunostaining for Egr1 (nn, ll), c-Fos (jj,mm), and an overlay image with Sox2 (kk,nn). Sporadic Egr1 positive Sox2 expressing cells are highlighted with white open arrows (ii-kk). Similar sparse patterns of Egr1 (ll) and c-Fos (mm) immunoreactivity can be detected in the VZ/SVZ of the ferret cortex at this age with a subset of Sox2 immunoreactive cells expressing either Egr1 or c-Fos, but not both (white arrows, ll-nn).



Supplementary Figure 15 Correlation of *EGR1* and *FOS* mRNA levels with canonical effectors of key signaling pathways regulating radial glia development using low- and high-coverage data. Heatmaps represent correlation coefficients between mRNA levels for *EGR1, FOS,* other immediate early genes, and canonical effectors of FGF, Notch and Wnt signaling pathways across all 65 neural cells (above diagonal) and within radial glia (below diagonal). Correlations were calculated based on expression levels detected in low-coverage (a, same heatmap as in Figure 4l) and high coverage data (b).

Supplementary Table 1 Overview of cell types studied, sequencing depth, and alignment rates. The cell types studied, integrated fluidic circuit (IFC) chips used, the single cell capture efficiency, the percentage of outliers following low-coverage sequencing, and the number of cells used in this study for comparison between high- and low-coverage sequencing depths are listed in the left panel of the table. The middle panel and right panel provide details on the number of total and aligned reads per a cell from each source for low-coverage and high-coverage sequencing respectively.

Supplementary Table 2 Single cell capture efficiency data. The single cell capture efficiency of the C_1^{TM} System was examined across a range of primary cells and cell lines loading a range of cell numbers onto the chip. Columns describe, cell types, integrated fluidic circuit (IFC) chips used, the average number of cells captured per 96 capture sites on chip at a given cell input number, and the average diameter of cells from each population.

Supplementary Table 3 Biological properties of PCA gene clusters. Enrichment analysis of genes belonging to individual clusters using ToppGene Suite (http://toppgene.cchmc.org/) provides insights into the possible functions of genes that define various cell types. Genes belonging to the green cluster are most highly expressed in skin cells and are highly enriched for genes belonging to the extracellular matrix organization (GO:0030198, $p<10^{-15}$). Genes belonging to the orange cluster are most highly expressed in the neural cells, and are enriched for neurogenesis (GO:0022008, $p<10^{-11}$). Finally, the cyan gene cluster is primarily expressed in K562 erythroleukemia cells, promyeloblast HL60 cells, and B lymphoblast 2339 cells, and this cluster was found to be most highly enriched for the immune response category (GO:0006955, $p<10^{-5}$).

Supplementary Table 4 Comparison of the top 500 PCA genes identified in low- and high-coverage mRNA Seq data across 301 cells. Table of the 500 genes with the greatest

contribution to PC1-3 derived from high-coverage and separately from low-coverage mRNA sequencing data, including the rank and gene loading score for PC1 and PC2.

Supplementary Table 5 Top 500 PCA genes explaining variation across 65 neural cells. Table of 500 genes with the greatest contribution to PC1-3 across 65 neural cells presented in Fig. 2a. Colors correspond to gene clusters highlighted in Fig. 2a', b' and were identified by consensus clustering in Supplementary Fig. 10.

Supplementary References

1. Diez-Roux, G. et al. A high-resolution anatomical atlas of the transcriptome in the mouse embryo. *PLoS biology* **9**, e1000582 (2011).
2. Tao, W. & Lai, E. Telencephalon-restricted expression of BF-1, a new member of the HNF-3/fork head gene family, in the developing rat brain. *Neuron* **8**, 957-966 (1992).
3. Kouprina, N. et al. The microcephaly ASPM gene is expressed in proliferating tissues and encodes for a mitotic spindle protein. *Human molecular genetics* **14**, 2155-2165 (2005).
4. Sjostrom, S.K., Finn, G., Hahn, W.C., Rowitch, D.H. & Kenney, A.M. The Cdk1 complex plays a prime role in regulating N-myc phosphorylation and turnover in neural precursors. *Developmental cell* **9**, 327-338 (2005).
5. Bolognese, F. et al. The cyclin B2 promoter depends on NF-Y, a trimer whose CCAAT-binding activity is cell-cycle regulated. *Oncogene* **18**, 1845-1853 (1999).
6. Verheijen, R. et al. Ki-67 detects a nuclear matrix-associated proliferation-related antigen. I. Intracellular localization during interphase. *Journal of cell science* **92 (Pt 1)**, 123-130 (1989).
7. Sasai, Y., Lu, B., Steinbeisser, H. & De Robertis, E.M. Regulation of neural induction by the Chd and Bmp-4 antagonistic patterning signals in Xenopus. *Nature* **377**, 757 (1995).
8. Collignon, J. et al. A comparison of the properties of Sox-3 with Sry and two related genes, Sox-1 and Sox-2. *Development* **122**, 509-520 (1996).
9. Stagaard, M. & Mollgard, K. The developing neuroepithelium in human embryonic and fetal brain studied with vimentin-immunocytochemistry. *Anatomy and embryology* **180**, 17-28 (1989).
10. Stoykova, A., Fritsch, R., Walther, C. & Gruss, P. Forebrain patterning defects in Small eye mutant mice. *Development* **122**, 3453-3465 (1996).
11. Theil, T., Alvarez-Bolado, G., Walter, A. & Ruther, U. Gli3 is required for Emx gene expression during dorsal telencephalon development. *Development* **126**, 3561-3571 (1999).
12. Furuta, A. et al. Expression of glutamate transporter subtypes during normal human corticogenesis and type II lissencephaly. *Brain research. Developmental brain research* **155**, 155-164 (2005).
13. Hatakeyama, J. et al. Hes genes regulate size, shape and histogenesis of the nervous system by control of the timing of neural stem cell differentiation. *Development* **131**, 5539-5550 (2004).

14. Gaiano, N., Nye, J.S. & Fishell, G. Radial glial identity is promoted by Notch1 signaling in the murine forebrain. *Neuron* **26**, 395-404 (2000).
15. Kowalczyk, T. et al. Intermediate neuronal progenitors (basal progenitors) produce pyramidal-projection neurons for all layers of cerebral cortex. *Cerebral cortex* **19**, 2439-2450 (2009).
16. Kawaguchi, A. et al. Single-cell gene profiling defines differential progenitor subclasses in mammalian neurogenesis. *Development* **135**, 3113-3124 (2008).
17. Hansen, D.V. et al. Non-epithelial stem cells and cortical interneuron production in the human ganglionic eminences. *Nature neuroscience* **16**, 1576-1587 (2013).
18. Miyoshi, G. & Fishell, G. Dynamic FoxG1 expression coordinates the integration of multipolar pyramidal neuron precursors into the cortical plate. *Neuron* **74**, 1045-1058 (2012).
19. Zheng, W., Geng, A.Q., Li, P.F., Wang, Y. & Yuan, X.B. Robo4 regulates the radial migration of newborn neurons in developing neocortex. *Cerebral cortex* **22**, 2587-2601 (2012).
20. Gonda, Y. et al. Robo1 regulates the migration and laminar distribution of upper-layer pyramidal neurons of the cerebral cortex. *Cerebral cortex* **23**, 1495-1508 (2013).
21. Hernandez, F., Perez, M., de Barreda, E.G., Goni-Oliver, P. & Avila, J. Tau as a molecular marker of development, aging and neurodegenerative disorders. *Current aging science* **1**, 56-61 (2008).
22. Englund, C. et al. Pax6, Tbr2, and Tbr1 are expressed sequentially by radial glia, intermediate progenitor cells, and postmitotic neurons in developing neocortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **25**, 247-251 (2005).
23. Bailey, J.A. & Lahiri, D.K. Neuronal differentiation is accompanied by increased levels of SNAP-25 protein in fetal rat primary cortical neurons: implications in neuronal plasticity and Alzheimer's disease. *Annals of the New York Academy of Sciences* **1086**, 54-65 (2006).
24. Britanova, O. et al. Satb2 is a postmitotic determinant for upper-layer neuron specification in the neocortex. *Neuron* **57**, 378-392 (2008).
25. Lyons, G.E., Micales, B.K., Schwarz, J., Martin, J.F. & Olson, E.N. Expression of mef2 genes in the mouse central nervous system suggests a role in neuronal maturation. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **15**, 5727-5738 (1995).
26. Miller, J.A. et al. Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199-206 (2014).
27. Chen, J., Bardes, E.E., Aronow, B.J. & Jegga, A.G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research* **37**, W305-311 (2009).
28. Clancy, B. et al. Web-based method for translating neurodevelopment from laboratory species to humans. *Neuroinformatics* **5**, 79-94 (2007).
29. Veyrac, A. et al. Zif268/egr1 gene controls the selection, maturation and functional integration of adult hippocampal newborn neurons by learning. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 7062-7067 (2013).
30. Herms, J., Zurmohle, U., Schlingensiepen, R., Brysch, W. & Schlingensiepen, K.H. Developmental expression of the transcription factor zif268 in rat brain. *Neuroscience letters* **165**, 171-174 (1994).