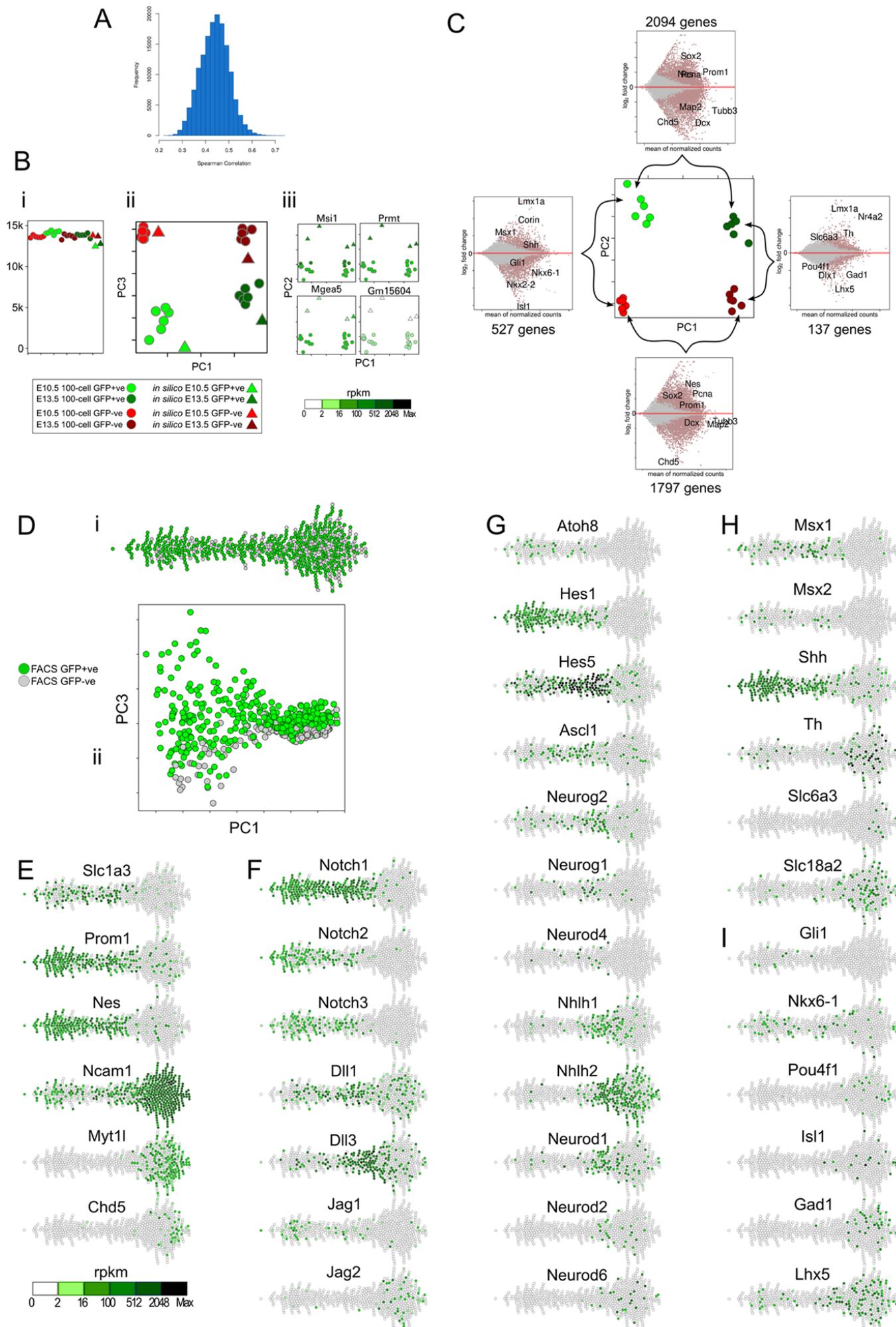


Supplemental Information

**Single-Cell Analysis Reveals a Close Relationship
between Differentiating Dopamine and Subthalamic
Nucleus Neuronal Lineages**

Nigel Kee, Nikolaos Volakakis, Agnete Kirkeby, Lina Dahl, Helena Storvall, Sara Nolbrant, Laura Lahti, Åsa K. Björklund, Linda Gillberg, Eliza Joodmardi, Rickard Sandberg, Malin Parmar, and Thomas Perlmann

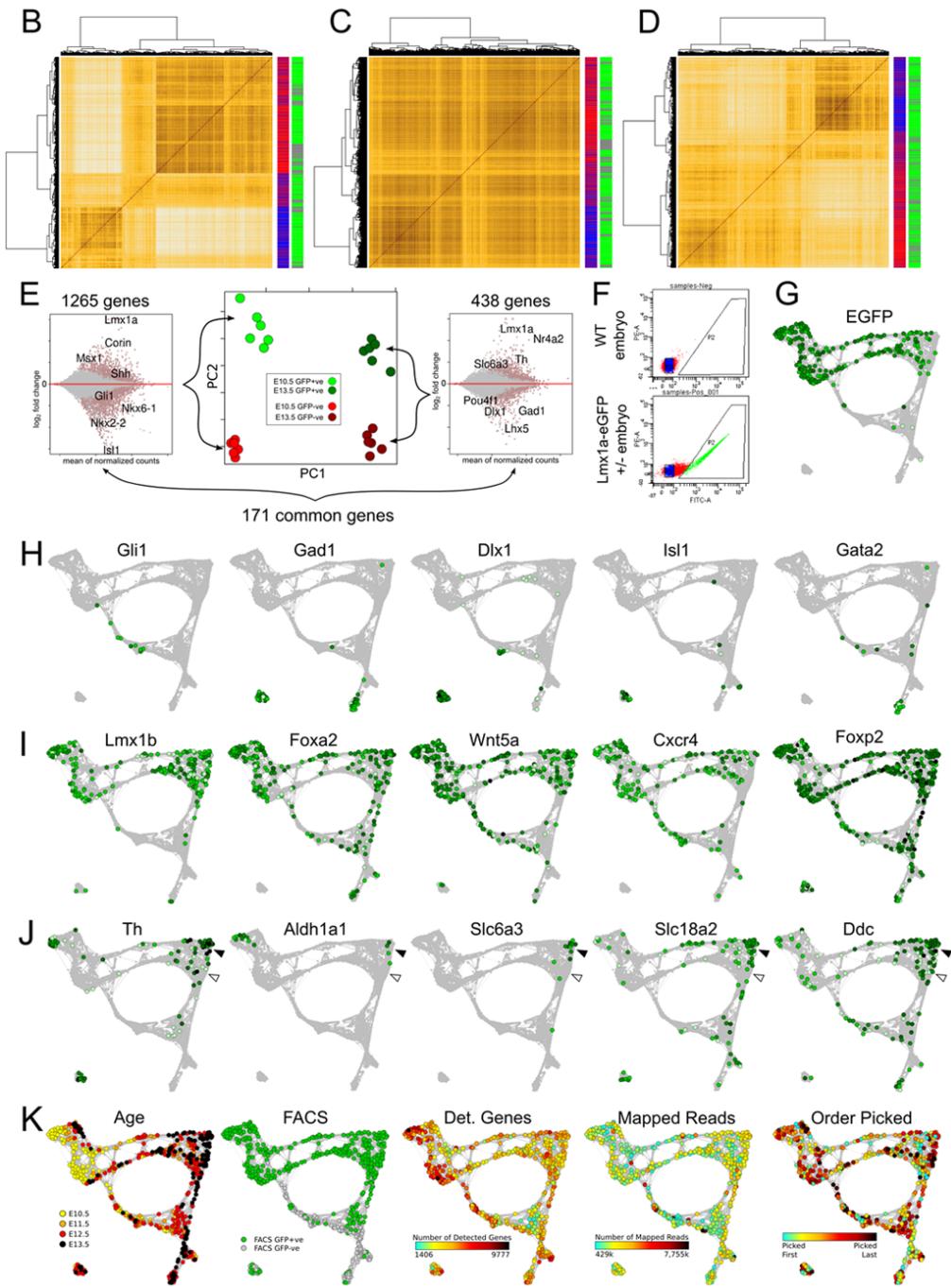
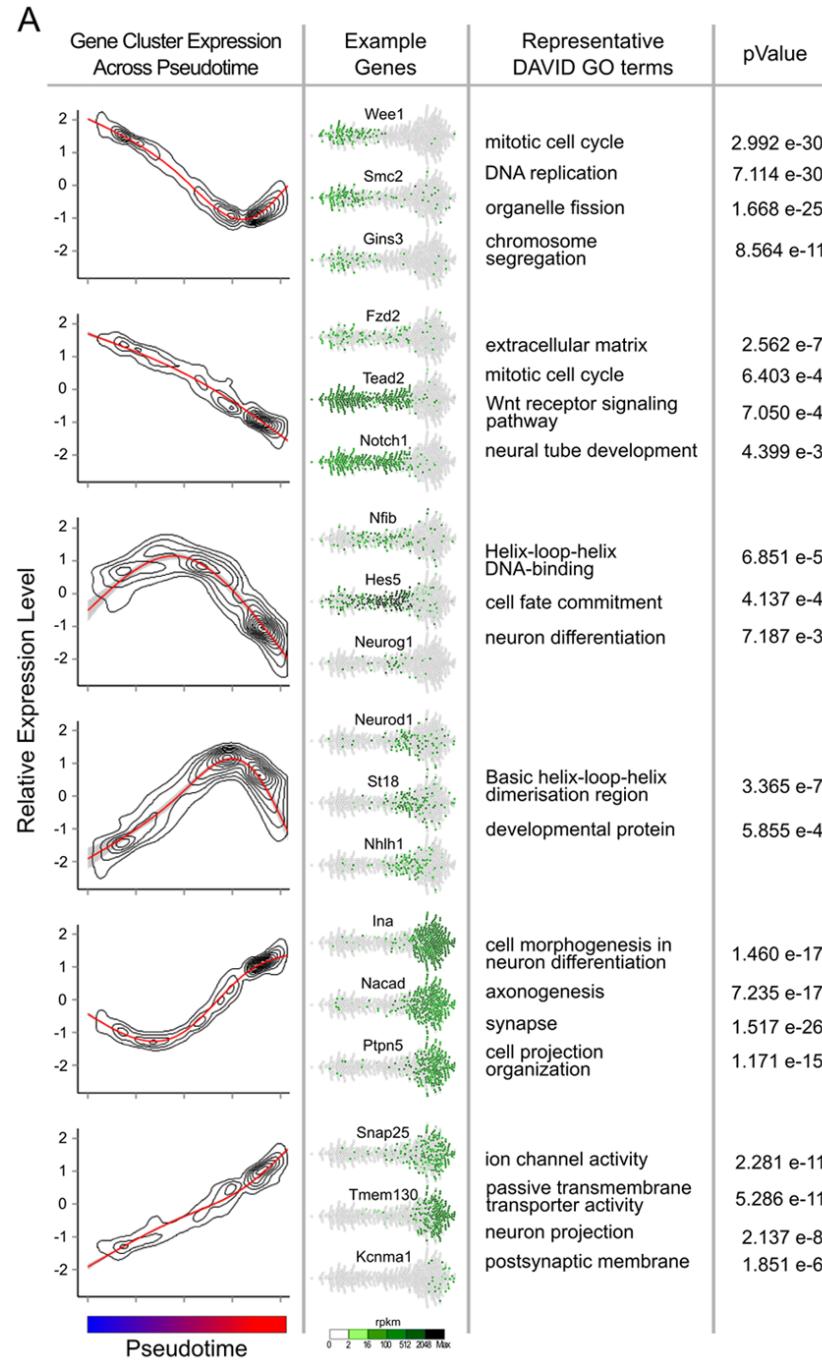
Figure S1. Related to Figure 1.



Deseq2 and PCA analysis of 550 single-cell libraries.

- (A) After de-multiplexing, read alignment, quality control filtering and rpkm analysis, 550 high quality single-cell RNAseq libraries showed an average Spearman correlation of roughly 0.45 considering all genes expressed above rpkm of one.
- (B) (i) Roughly the same number of genes (14k) were detected in *in silico* samples (triangles) derived from merging all E10.5 or E13.5, GFP positive or GFP negative single cells, when compared to 100-cell pool samples (circles). (ii) *in silico* samples cluster with their corresponding 100-cell pool samples after PCA analysis when considering PC1 and PC3. (iii) PC2 described genes that were enriched in either *in silico* (top two panels), or 100-cell pool (lower two panels) samples.
- (C) Deseq2 differential expression analysis on bulk RNA samples was used to derive an informative gene list for further principle component analysis. For each of the four pairwise comparisons, genes were required to satisfy two cutoffs, a padj value of less than 0.01, and a fold change of greater than four, totalling 3006 differentially expressed genes.
- (D) Upon PCA analysis, GFP positive and negative cells were randomly distributed along PC1 (i and ii - horizontal axis), while positive and negative cells weakly separated along PC2 (ii - vertical axis).
- (E) Along PC1 classic neural progenitor and postmitotic markers, notch pathway components (F), and bHLH proneural genes (G) displayed distinct, and expected, temporal distributions.
- (H) DA lineage specific markers, or non-DA lineage markers also expressed in the ventral mesodiencephalic region (I), displayed their appropriate early or late expression distributions.

Figure S2. Related to Figure 2.

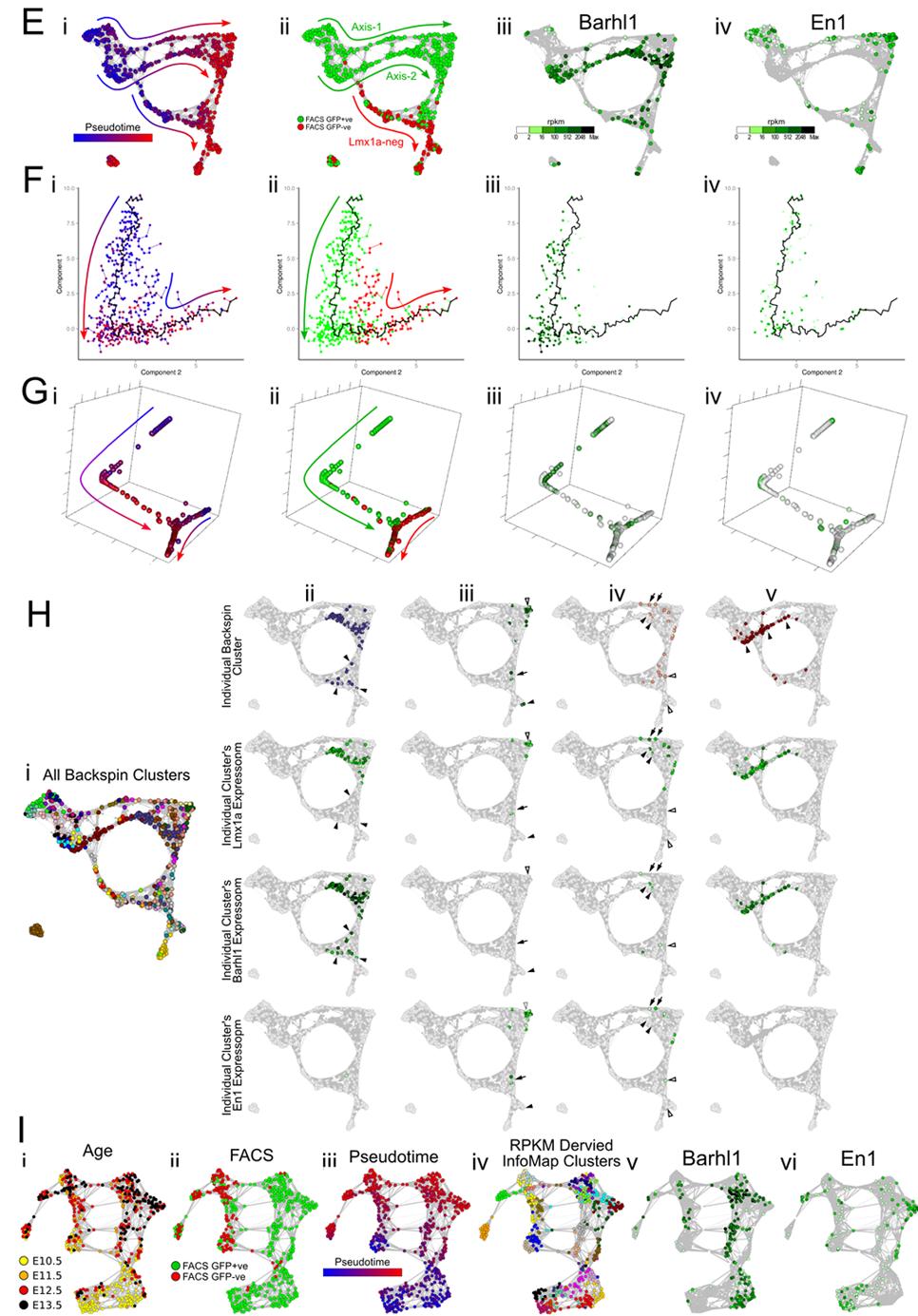
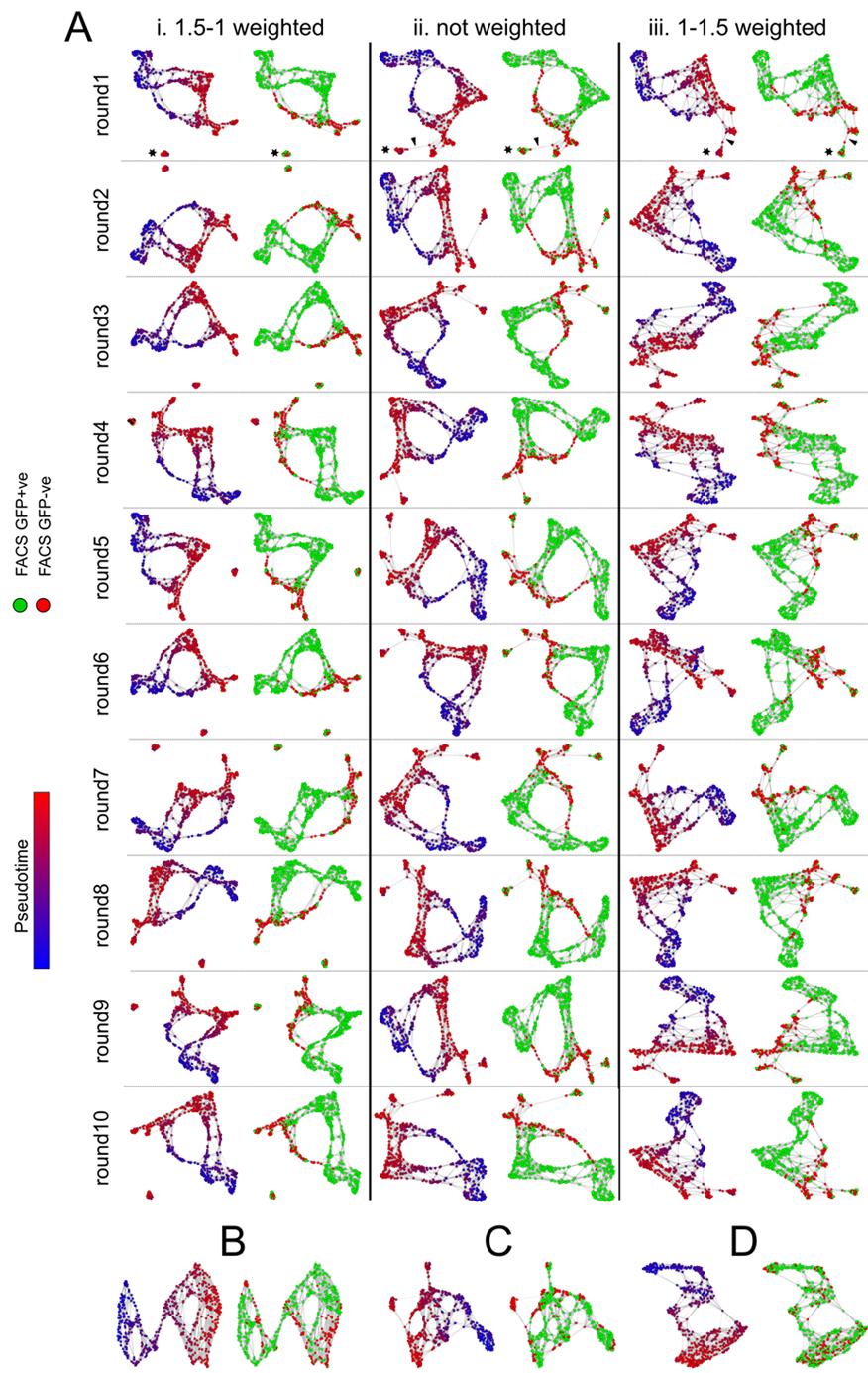


Whole transcriptome Monocle gene clusters, Spearman Hierarchical Clustering, and tSNE-based graphing using Lineage Intersect List.

- (A) Clusters of co-varying genes. Relative expression values (from -2 to +2) are displayed for each gene cluster along Monocle's pseudotime axis. Representative genes from each cluster are displayed with cells plotted as in Figure 1D. High scoring representative DAVID gene ontology terms for each cluster are indicated along with their corresponding p-value.
- (B) Spearman correlation heatmap of cells based on the original list of 3006 genes. Pseudotime (blue: early; red: late) and FACS-based (green: GFP-positive; grey: GFP-negative) classifications are depicted in the sidebars.
- (C) Spearman correlation heatmap of cells based on all expressed genes. Pseudotime and FACS-based classifications are depicted in the sidebars as in A.
- (D) Spearman correlation heatmap of cells based on the top 1000 variable genes. Pseudotime and FACS-based classifications are depicted in the sidebars as in A.
- (E) To select a gene list (termed the Lineage Intersect gene list) describing lineage irrespective of age, we considered genes differentially expressed between GFP positive and negative samples at both E10.5 and E13.5. Genes were required to satisfy a padj value of less than 0.01, and a foldchange of greater than two, and be common to both lists, leaving 171 genes.
- (F) Fluorescent gating chosen during FACS enrichment of E12.5 dissected vMB. A relaxed gating was intentionally chosen so as not to miss cells expressing low levels of GFP.
- (G) Rpkm expression plot of EGFP transgene levels.
- (H) Rpkm expression plots show non-DA markers expressed in the Lmx1a-neg axis.
- (I) Rpkm expression plots of known DA markers.
- (J) Dopaminergic neurotransmitter gene expression is absent or noticeably lower in the late pseudotime region of Axis-2 (white arrowhead), when compared to the late pseudotime region of Axis-1 (black arrowhead).
- (K) Cell age and FACS classification are consistent with Pseudotime and Lmx1a classifications (Figure 2C-D). The number of detected genes per cell appears highest in cycling progenitors and mature neurons,

and appears lowest in mid-pseudotime, which may reflect a smaller cell size during this transition phase. Neither the number of mapped reads per cell, nor the order that cells were picked on a given day, influenced the structure of the graph.

Figure S3. Related to Figure 2.



Weighted PCA input into tSNE based force directed graphing, and exploration of alternative bioinformatics approaches.

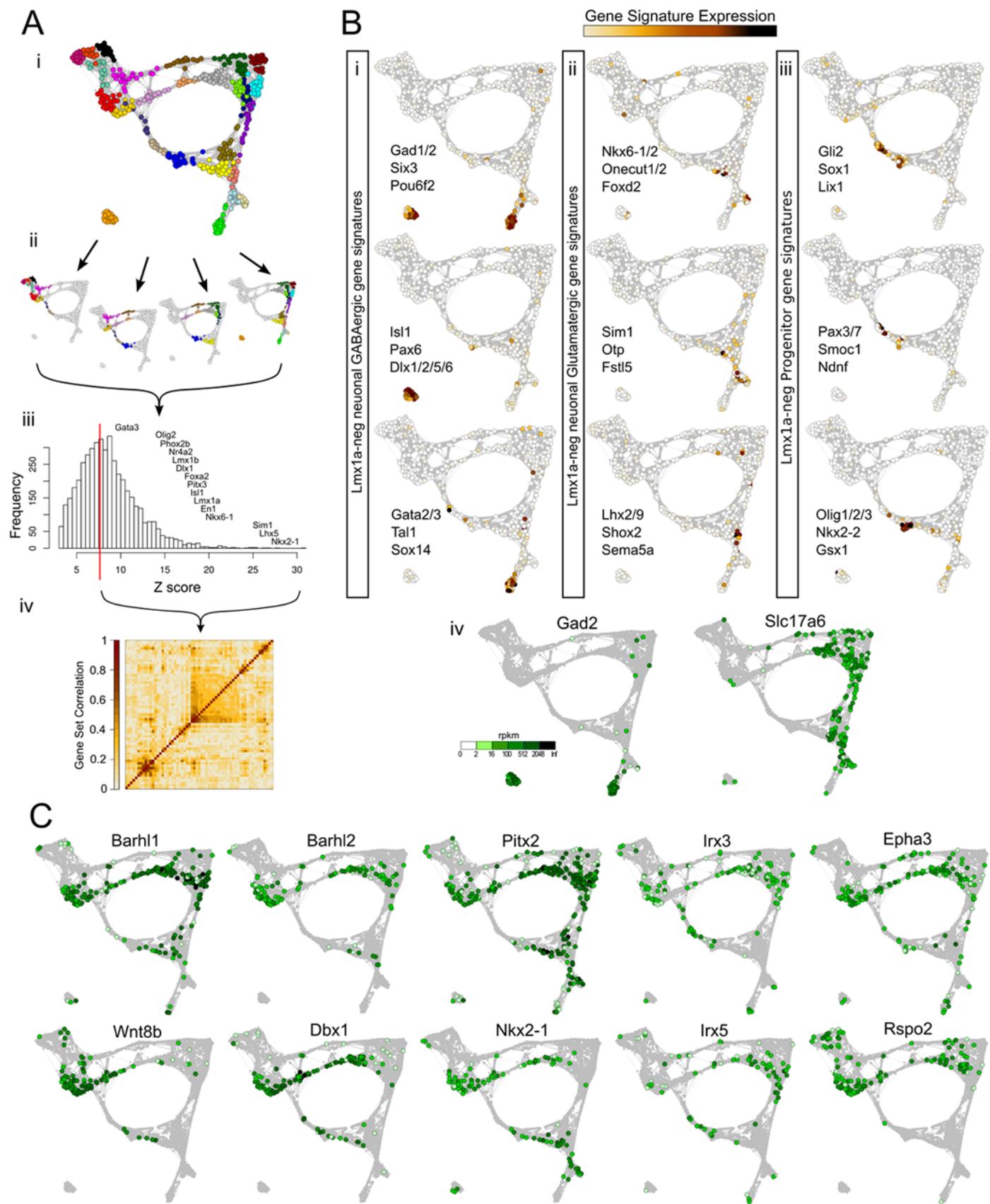
- (A) (i) PCA was performed considering the 171 genes on the Lineage Intersect list, which was weighted according to their Stouffers Z score (Experimental Procedures). Ten consecutive runs of tSNE-based graphing using this weighted PCA input produced ten graphs that almost always showed two distinct parallel Lmx1a positive pseudotime axes.
- (ii) Ten consecutive graphs using a non-weighted PCA input produced graphs with a less clear distinction between the two Lmx1a positive pseudotime axes.
- (iii) Reverse weighting of each gene's Z score produced graphs containing even weaker separation of the two Lmx1a positive pseudotime axis, and more inappropriate edges between disparate regions of the graph. Indicative of this are the *Dlx* expressing GABAergic cell cluster (Figure S5 Bi – middle panel), that is isolated from the body of the graph in weighted tSNE (black stars), but that acquires inappropriate edges in non-weighted or reverse weighted tSNE (black arrowheads).
- (B) tSNE based force directed graphing considering the original 3006 genes, all expressed genes (C), or the top 1000 variable genes (D) failed to resolve two parallel pseudotime axes within the Lmx1a expressing populations.
- (E) As described, the t-SNE network resolves multiple pseudotime axes (i), and two Lmx1a-expressing Axes (ii) that predominantly express either the STN transcription factor Barhl1 (iii), or the DA transcription factor En1 (i).
- (F) When considering the Lineage Intersect genelist, and when tasked to find three differentiation paths, Monocle roughly segregated early and late pseudotime cells (i, early; blue, late; red) and detected a single pseudotime axis (black line). However Barhl1 (iii) and En1(iv) expressing cells were intermingled along this single axis. Though not detected by Monocle, when considered with each cell's FACS classification, two axes can be appreciated, both a FACS-gfp positive and negative axes (ii, green; FACS-gfp positive, red; FACS-gfp negative).
- (G) We next tried Diffusion maps, using the Euclidean distance output from t-SNE when considering the Lineage Intersect gene list. Diffusion map analysis also segregated early and late pseudotime cells (i,

early; blue, late; red) that also roughly built FACS-gfp positive and negative axes (ii, green; FACS-gfp positive, red; FACS-gfp negative). However Barhl1 (iii) and En1(iv) expressing cells were again intermingled in this single FACS-gfp positive axis.

(H) Back-spin biclustering, considering the Lineage Intersect gene list, down five levels, assigned similar cells to one of 30 clusters (i) that generally agreed with our t-SNE network. However for many clusters a proportion of cells were found scattered across more than one of the pseudotime axes. For example, one cluster (ii) contained *Barhl1* expressing cells in both Axis-2 and Lmx1a-neg axis (arrowheads). Another cluster (iii) grouped together Axis-1 DA neurons expressing *Pitx3* (white arrowhead) with serotonergic neurons (black arrow) co-expressing *Fev*, *Tph2*, *Gata2* and *Gata3* and oculomotor neurons (black arrow head) co-expressing *Phox2b* and *Isl1* (see rshiny.nbis.se/shiny-server-apps/shiny-apps-scrnaseq/Kee_2016/). One cluster (iv) contained Axis-1 cells co-expressing *Lmx1a-En1* (black arrows), and Axis-2 cells co-expressing *Lmx1a-Barhl1* (black arrowheads), in addition to distantly located Lmx1a-negative axis cells (white arrowheads). Of note, when considering maturation, Backspin clusters were observed comprising cells at similar but clearly distinct maturation stages (v, arrowheads), information that is obscured when considering with Backspin, but that becomes immediately apparent upon observing a cell's positions along that axis' trajectory within the t-SNE cell-network. As expected, Backspin biclustering considering all genes, all variable genes, or the original 3006 Deseq2 genes, clustered according to differentiation stage only (data not shown).

(I) TPM normalisation, instead of RPKM normalisation, was used to generate a tSNE network using an identical workflow. This resulted in a network map that again reconstructed an age axis (i), GFP positive and negative axes (ii), and a Pseudotime axis (iii). Importantly, when RPKM derived Infomap clusters were plotted on to this TPM derived graph, cells of a given cluster continued to neighbor one another (iv). Additionally, Barhl1 (v) and En1(vi) expression also segregated between the two parrallel Lmx1a axes.

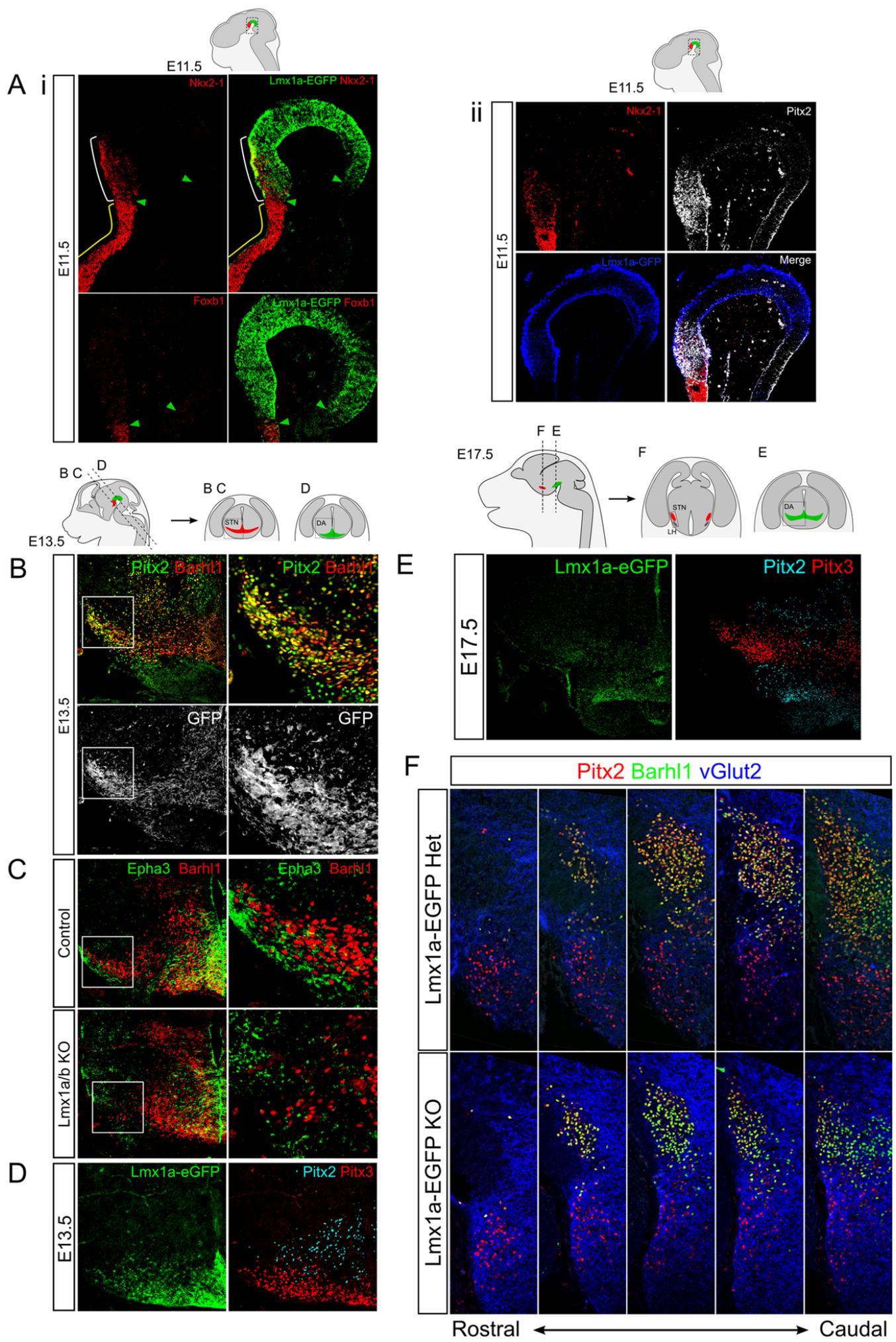
Figure S4. Related to Figure 3.



SCDE and WGCNA mediated gene filtering and clustering workflow.

- (A) (i) Each cell was, along with its similar neighbouring cells, assigned to one of 27 clusters within the graph using Infomap community detection.
- (ii, schematic) Cluster pairs whose mean pseudotime values fell within 10% of the maximum psuedotime were analysed using SCDE to find differentially expressed genes with a padj < 0.01 and a minimum fold change of two.
- (iii) Stouffers Z score method was used to combine pvalues for all pairwise comparisons. After ordering by Stouffers Z score the top ranked genes, including many classic transcriptional determinants, were passed to WGCNA.
- (iv) WGCNA clustered significant genes into similarly behaving gene signatures.
- (B) Each cell's correlation to Lmx1a-neg axis enriched gene signatures.
- (i) GABAergic gene signatures included both common genes (*Gad1/2-Six3*) and mutually exclusive rostral (*Dlx1/2/5/6-Pax6-Isl1*)(Marin et al., 2000; Yee et al., 2009) or lateral (*Gata2/3-Tall-Sox14*)(Lahti et al., 2013) populations developing in the vicinity of the Lmx1a-expressing domain.
- (ii) Several glutamatergic gene signatures were also enriched, including putative lateral mesodienchephalic (*Onecut1/2-Nkx6-1/2*)(Chakrabarty et al., 2012), and more rostral hypothalamic (*Lhx2/9-Shox2* or *Sim1-Otp*)(Balasubramanian et al., 2014; Michaud et al., 1998) populations.
- (iii) Additionally some gene signatures described distinct progenitor populations including lateral mes-dienchephalic (*Gli2-Sox1-Lix1*)(Hayes et al., 2011; Metzakopian et al., 2012), alar plate mesenchephalic (*Pax3/7-Smoc1*)(Matsunaga et al., 2001) and more rostral hypothalamic or telencephalic progenitors (*Olig1/2/3-Nkx2.2-Gsx1*)(Chapman et al., 2013; Inamura et al., 2011).
- (iv) Rpk expression values for the neurotransmitter genes *Gad2* and *vGlut2*.
- (C) Rpk expression values for Axis-2 genes.

Figure S5. Related to Figure 4.



Immunhistochemistry of WT, Lmx1a-EGFP KO, and Lmx1a/b conditional KO embryos.

(A) (i - top panel) At E11.5 Nkx2-1 immunostaining overlapped with rostral Lmx1a-eGFP expression (white bracket). Beyond the rostral boundary of Lmx1a-eGFP expression, Nkx2-1 was immediately expressed more strongly, with this strong expression extending into the ventral hypothalamus (yellow bracket). Rostral and caudal boundaries of Lmx1a-eGFP are shown with green arrowheads. (bottom panel) Foxb1 was expressed immediately rostral, and adjacent to, the Lmx1a-eGFP domain. Rostral and caudal boundaries of Lmx1a-eGFP are shown with green arrowheads. (ii) The Pitx2 expressing domain overlaps both Lmx1 and Nkx2-1 expression domains.

(B) In E13.5 WT embryos, co-expression of Lmx1a-eGFP, Barhl1 and Pitx2 can be seen in neural cells rostrotangentially migrating into the developing hypothalamus. Boxed area in left hand panels are magnified in the right hand panels.

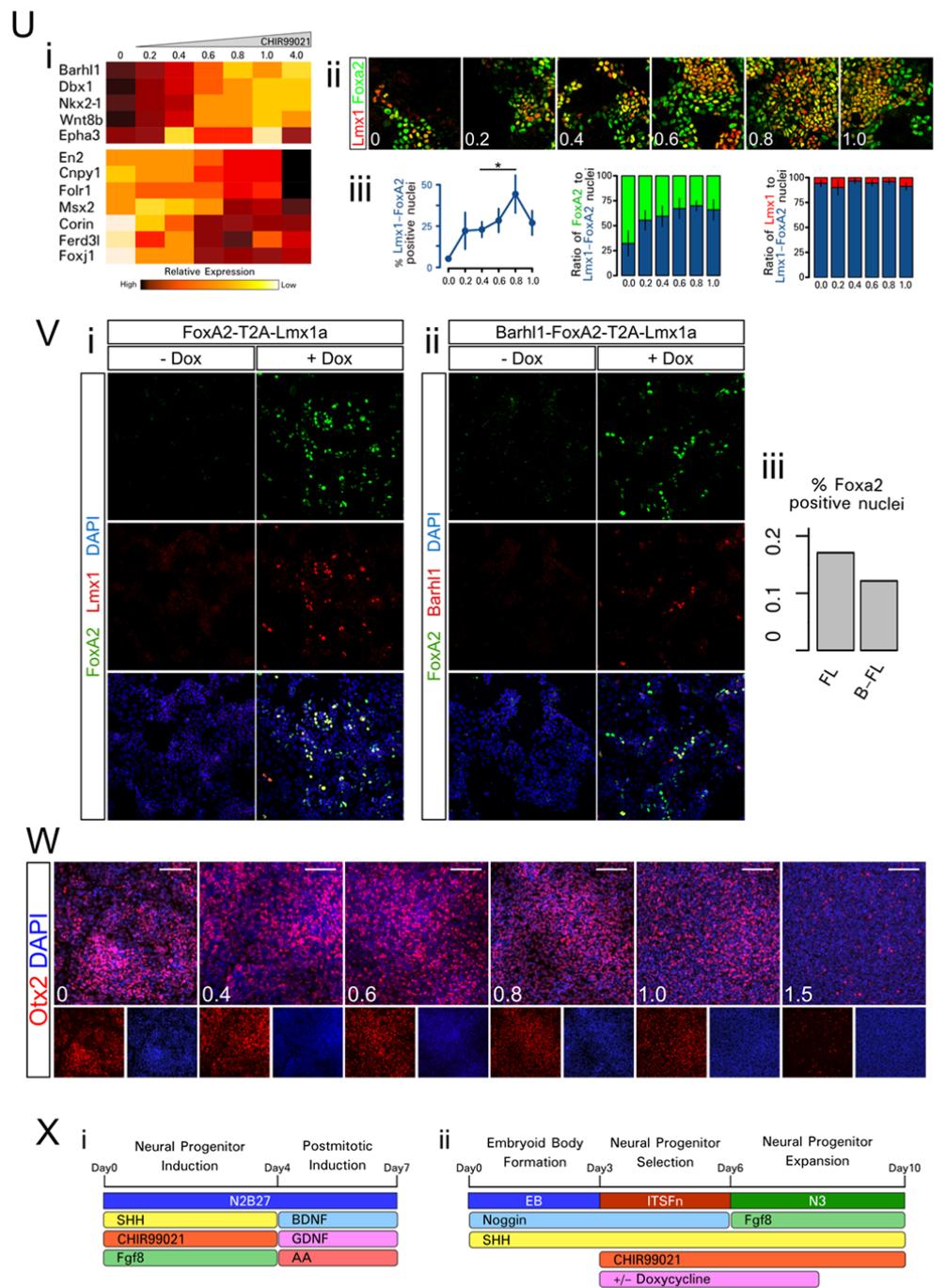
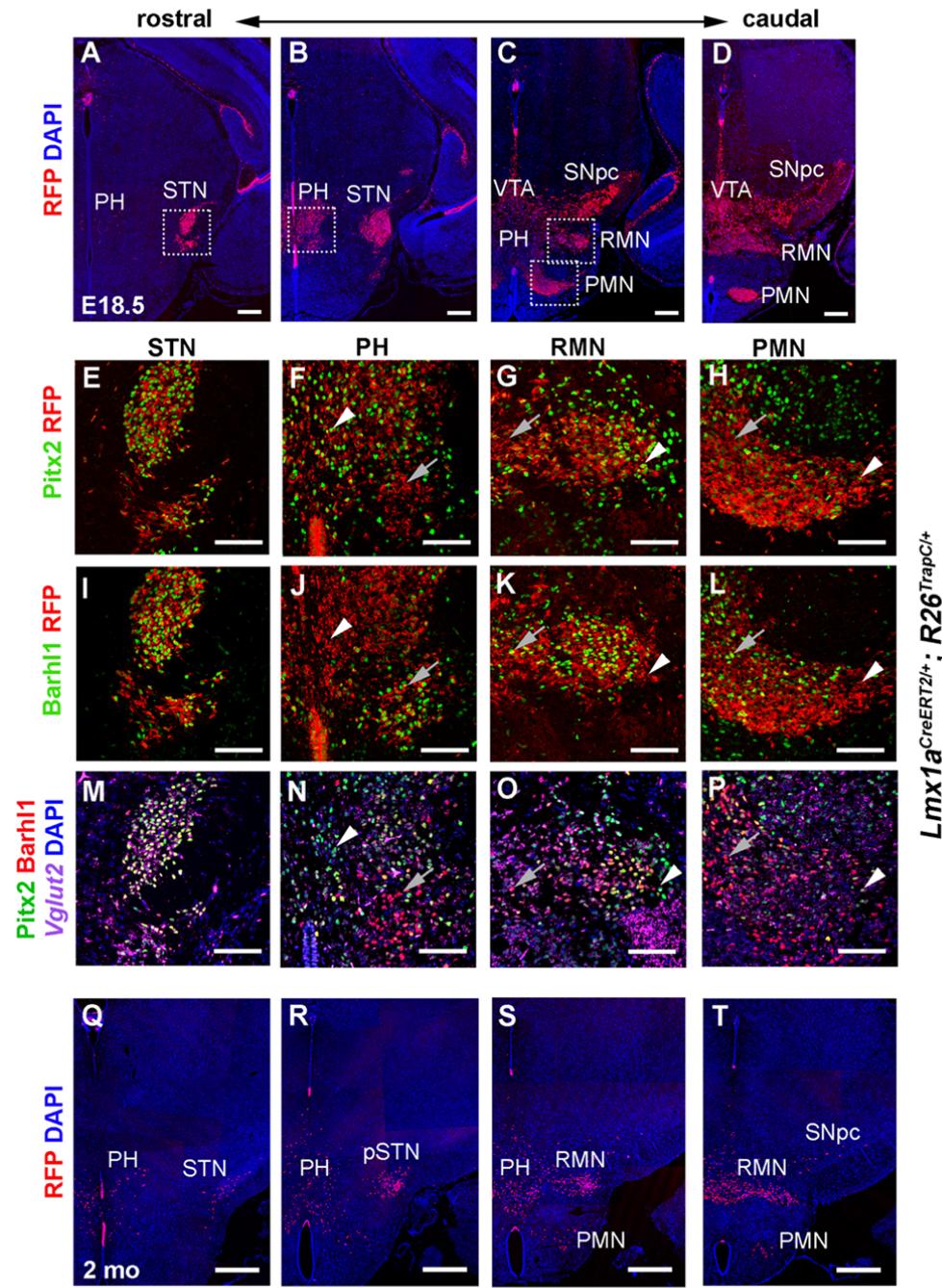
(C) E13.5 Lmx1a/b double conditional mutants display loss of rostrotangentially migrating Barhl1/Epha3 co-expressing STN neuroblasts (white box). Boxed area in left hand panels are magnified in the right hand panels.

(D) Pitx2 and Pitx3 expression is mutually exclusive at both E13.5, and E17.5 (E).

(F) E17.5 Lmx1a-eGFP homozygous knockout animals show a reduction in size of the STN, compared to heterozygous littermates.

(G) Representative Pitx2 immunostaining and quantitation (H) of lentivirus transduced primary vMB tissue showed no significant difference between control M2rtTA, and Lmx1a or Barhl1 overexpression.

Figure S6. Related to Figure 6.



Histological analysis of Lmx1a-CreERT2/R26-TrapC reporter mice, and mESC methods

(A-D) Embryos treated with tamoxifen at E9.5, and serial coronal sections collected at E18.5 from the hypothalamus through to the midbrain. RFP immunostaining of genetically labelled cells shows recombination in the subthalamic nucleus (STN), posterior hypothalamic (PH) area, retromammillary nucleus (RMN) and premammillary nucleus (PMN). As expected, labelled cells were also seen in the ventral tegmental area (VTA) and the substantia nigra pars compacta (SNpc). Images were compiled in Adobe Photoshop CS6 from a series of individual images, each taken using a 5x objective, using identical confocal settings.

(E-H) Higher magnification (20x) of each nuclei co-immunostained with Pitx2 and RFP. White arrowhead indicated cells positive for Pitx2 and Lmx1a. Grey arrows indicated cells positive only for Lmx1a.

(I-L) Higher magnification (20x) of each nuclei co-immunostained with Barhl1 and RFP. White arrowhead indicated cells positive for Barhl1 and Lmx1a. Grey arrows indicated cells positive only for Lmx1a.

(M-P) Higher magnification (20x) of each nuclei co-immunostained with Pitx2 and Barhl1, and *in situ* stained for *Vglut2* mRNA. White arrowhead indicated cells positive for Pitx2 but not Barhl1. Grey arrows indicated cells positive for Barhl1 but not Pitx2.

(Q-T) Embryos treated with tamoxifen at E14.5, and serial coronal sections collected at two months from the hypothalamus through to the midbrain. RFP immunostaining of genetically labelled cells shows recombination present, but to a lesser degree than (A-P), in the subthalamic nucleus (STN), parasubthalamic nucleus (pSTN), posterior hypothalamic (PH) area, retromammillary nucleus (RMN) and premammillary nucleus (PMN). Images were compiled in Adobe Photoshop CS6 from a series of individual images, each taken using a 5x objective, using identical confocal settings.

U) (i) mESCs were differentiated using the monolayer protocol (See Experimental Procedures) across a gradient of CHIR99021. qPCR analysis showed Axis-2 genes (Barhl1, Dbx1, Nkx2-1, Wnt8b and EphA3) were enriched at lower concentrations, while Axis-1 genes (En2, Cnpyp1, Fol1r, Msx2, Corin, Ferd3l and Foxj1) were enriched at higher concentrations. (ii) mESCs were differentiated using an adapted five-stage

protocol and differentiated across a 0-4M gradient of CHIR99021, and stained for Lmx1a/b (Lmx1) and Foxa2. (iii) Quantification of (ii) showing; Left panel, Lmx1/FoxA2- double positive cells increase up the CHIR gradient; Middle panel, ratio of Lmx1/FoxA2- double positive to FoxA2-single positive cells did not significantly change above a CHIR concentration of 0.2; Right panel, almost all Lmx1 expressing cells co-express FoxA2 at all concentrations along the CHIR gradient.

V) Immunostainings of Lentiviral transduction efficiency. Cells were transduced upon plating using the same cell to virus ratios as in Figure 6B, and cultured for 24hrs before addition of 1ug/ml DOX to the culture medium. Cells were then cultured for a further 48hrs before fixing and staining. FL Lentiviral particles (i) and B-FL Lentiviral particles (ii) showed specific expression of transgenes only after DOX addition, with a transduction efficiency of roughly 17 and 12 percent respectively.

W) Otx2 immunoreactivity is detected in a high proportion of cells at almost levels of the increasing gradient of CHIR99021. By 1.5uM CHIR99021, Otx2 expression begins to decrease.

X) Schematics of the monolayer (i) or adapted 5-stage (ii) protocols used for mESC differentiations in Figure S6U and Figure 6 respectively.

Table S1. Related to Experimental Procedures. Percentage of uniquely mapping reads, multimapping reads, and unmapped reads for each single cell library.

Table S2. Related to Figure 2. Gene clusters derived from Monocle's Pseudotime describing successive phases of embryonic neurogenesis.

Table S3. Related to Experimental Procedures. Mouse qPCR primer sequences.

RPL19	ribosomal proteinL19	GGTGACCTGGATGAGAAGGA TTCAGCTGTGGATGTGCTC CGCGGCACACGAGAGCTTCAG TCTCCTCCTCCTTGTCCGTGGC
Bahrl1	BarH-like homeobox 1	AACAACCCTCCCACGGGAGC CTACACACACCCTGCACGCGG
Dbx1	BarH-like homeobox 2	CCGCCTTACCAAGGACACCATGC GGCCCATGAAGCAGGGAGATGGC
Nkx2-1	NK2 homeobox 1	CGAACGCTAACCGGGAGACAG CCCCCAGTTGTCCATTTCGGGAG
Wnt8b	Wingless-type MMTV integration site family, member 8b	GGACTGCTGGCAGAAAGACAGG AGGTTTGATGGCCTTGCAGCCG
Epha3	EPH receptor A3	ATTCTGACCGGCCTTCTCAG GAAACTCAGCCTTGAGCCTCTG
En2	Engrailed 2	AGCTGGAGGACGACCCAGTGAC TGGCTTCGACGCGAATTCAAAG
Cnpy1	Canopy FGF signaling regulator1	GGCTGACGGCTCTAGAAGTCCCC GAGACAGCCGGCCGAGGTTAG
Folr1	folate receptor 1	CGCCGCCAGACATATGAGCC TTCTGGCGGAACTTGCGCTCC
Msx2	msh homeobox 2	AGCGTCCATCAGCTGGGAGTC CACAGGAGTCTGCAAGGCGGG
Corin	corin, serine peptidase	AGCTGCTTGGATGCTACCGTGC TCCCCAAAAGGGACCCCAGGTG
Ferd3l	Fer3-like	CTGAGCCAGGGAAAGAGTGTTCAG GCTCTCTGGCCCGCTGGTAAC
Foxj1	forkhead box J1	

Table S4. Related to Experimental Procedures. Human qPCR primer sequences.

Actb	beta-actin	CCTTGCACATGCCGGAG GCACAGAGCCTCGCCTT
GAPDH	glyceraldehyde-3-phosphate dehydrogenase	TTGAGGTCAATGAAGGGGTC GAAGGTGAAGGTGGAGTCA
En1	Engrailed 1	CGTGGCTTACTCCCCATTAA TCTCGCTGTCTCTCCCTCTC
En2	Engrailed 2	CCTCCTGCTCCTCCTTCTT GACGCAGACGATGTATGCAC
Wnt1	Wingless-type MMTV integration site family, member 1	GAGCCACGAGTTGGATGTT TGCAGGGAGAAAGGAGAGAA
CNPY1	Canopy FGF signaling regulator1	TTGGCCTCTCAAACACCATTCT GAGCGAAACAAACGCAATCAC
BARHL1	BarH-like homeobox 1	GTACCAGAACCGCAGGACTAAA AGAAATAAGGCACGGGAACAT
BARHL2	BarH-like homeobox 2	GGAGATTACGAGTAGCCGTGAG AAGCTACGCTCCAGTTGATTGA
Wnt8b	Wingless-type MMTV integration site family, member 8b	CTAGTGGGCATGACTTTCT TTCCTAGACCTCGGGGTATGT
Nkx2.1	NK2 homeobox 1	AGGGCGGGGCACAGATTGGA GCTGGCAGAGTGTGCCAGA
Nkx2.4	NK2 homeobox 4	AACTGCGATTCAAACGAACCG GCCTCGTGGCATAATGTTACAC
Dbx1	Developing brain homeobox 1	GAAGTTGGAGTGAACGCCATC GACCCTCGAAGTAGGGAAAGG

Supplemental Experimental Procedures

Single-cell Library Preparation and sequencing

For each given developmental stage (E10.5, E11.5, E12.5 and E13.5), between three to five heterozygous Lmx1a-EGFP embryos (C57/BL6 background) were collected, the vMB region (eGFP region and surrounding area) dissected out, then all vMBs were pooled together, and finally dissociated into single cells using a MACS Neural Tissue Dissociation kit (Miltenyi Biotec, 130-092-628). Three fire-polished glass needles were used, each one progressively thinner, to aid manual dissociation. Cell suspensions were then subject to FACS sorting into GFP positive and negative fractions. After this, GFP positive and negative cell suspensions were kept on ice and, using a pulled capillary glass needle, single cells were manually picked and transferred to a single tube from an eight strip PCR tube containing 2ul Smart-seq2 lysis buffer, and then frozen at -80C. Finally cells were processed using Smart-seq2 (Picelli et al., 2014) to generate cDNA libraries. While time consuming, manual picking eliminated cell-doublets or empty wells that can arise from FACS. On any given day, roughly 50-150 cells were collected from a given developmental stage, and the order that cells were picked did not influence the final structure of the tSNE graph (Figure S2J, right-most panel). cDNA Tagmentation was performed using Nextera XT DNA library preparation kits (FC-131-1024) using dual indexes (i5 + i7), and sequenced on Illumina HiSeq 2000, giving 43bp reads following de-multiplexing. Reads were aligned to the mm10 mouse assembly using default settings in STAR v2.3.0. Considering only uniquely mapping reads, expression values were calculated as reads per kilobase gene model and million mappable reads (RPKMs) for transcripts in Ensemble v75 using rpkmforgenes. To filter out low quality libraries the following quality control parameters were applied: > 50% uniquely mapping reads, > 45% exon mapping reads, < 18% mismatches per reads, at least 7% of all genes detected, > 400,000 normalization reads. Mapping statistics have been provided in Table S1. In total 550 cells (137 E10.5, 90 E11.5, 178 E12.5 and 145 E13.5 cells) passed these cutoffs and were thus considered for further analysis.

100-cell pool expression analysis

For E10.5 and E13.5 ages, in parallel to manual single cell picking, pools of 100 cells each were also manually picked from both the GFP positive and negative fractions and processed using Smart-seq2. For all steps up to and including cDNA synthesis, reaction volumes were increased ten fold to account for the increased starting material. Beyond this samples were treated identically to single cell samples.

Bulk 100-cell pool samples were compared to *in silico* derived bulk samples (where rpkm values were recalculated using all reads from E10.5 or E13.5, GFP positive or GFP negative cells accordingly). Roughly 14k genes were detected in both 100-cell and *in silico* samples (Figure S1B-i). PCA analysis was performed on all samples. PC1 described sample age, while PC3 described GFP positive or negative classification, and *in silico* samples clustered in together with their 100-cell pool counterparts (Figure S1B ii). PC2 contained genes that were selectively enriched in either *in silico* (Figure S1B iii, top two panels), or 100-cell pool (Figure S1B iii, two lower panles) samples.

For initial analyses Deseq2 differential expression (Love et al., 2014) was performed across four pair-wise comparisons of these bulk RNA samples (Figure S1C). For each comparison genes were required to satisfy both a padj value of less than 0.01, and a fold change cutoff of more than four. Many differentially expressed genes fell on more than one list such that combined, 3024 differentially expressed genes were detected across all four pairwise bulk RNA comparisons. 3006 of these genes were expressed in our single-cell dataset. For the Lineage Intersect gene list Deseq2 (Figure S2E) differential expression was performed between GFP positive and negative samples, at either E10.5 or E13.5. Genes were required to satisfy two cutoffs, a padj value of less than 0.01, and a foldchange of greater than two. Further, genes had to be common to both lists, leaving 171 common differentially expressed genes.

PCA analyses

The original list of 3006 genes was used for Principle Component analysis using the R package prcomp. Broadly speaking, the first three components described differentiation, the number of detected genes, and to a limited degree, Lmx1a-positive or negative status respectively. The R package beeswarm was used to generate plots with cells' ordered according to their positions along PC1.

Monocle

The same list of 3006 genes was used to derive a pseudotime axis using the R package monocle, faithfully recapitulating the age distribution along PC1(Trapnell et al., 2014). However this gene list did not contain known proneural genes important for the transition from a DA progenitor to a DA neuron. As we sought to resolve this transition, a new list was derived to capture such genes. Hierarchical Spearman clustering considering the original list of 3006 genes, yielded four cell clusters clearly segregating along PC1, and thus delineating clusters of different maturation stages. We then performed differential expression using SCDE (R package, scde) between all the cells of a given cluster versus the remaining cells. For this analysis, genes had to satisfy both a padj value of less than 0.01, and a cutoff of more than two-fold change. Additionally, since we wished to cluster genes changing across neurogenesis, and not ubiquitous or rarely expressed genes, only genes expressed in more than 15, and less than 400 cells were considered. Taken together, this gave a list of 2711 genes, now containing proneural genes such as *Ascl1*, and *Ngn2*.

We applied this new list to Monocle's gene clustering algorithm, which groups genes that behave similarly across pseudotime. For this approach, the number of desired clusters must be supplied. We tried a range from six to twelve total clusters, observing the array of clusters yielded each time. Some expression patterns were seen often, including genes high only in early cells, only in late cells, or only in mid-pseudotime. Often a cluster could be seen breaking into two similar, but distinct clusters (with distinct GO term enrichment), when more total clusters were allowed. After ten total clusters, no new expression patterns emerged. With this parameter two pairs of clusters contained near identical expression patterns and GO term enrichment, and were thus merged, leaving eight clusters. Two from these remaining clusters described genes expressed highly at early and late, but not mid-pseudotime, mirroring the number of detected genes across pseudotime (Figure S2K, middle panel). Additionally GO term enrichment for these two clusters included many ubiquitous biosynthetic processes seemingly unrelated to neuronal differentiation. These two clusters thus appear to reflect housekeeping genes, or genes affected

by sequencing depth and/or cell size and were therefore set aside. This left a final set of six gene clusters implicated in neural differentiation (Figure S2A, Table S2).

Hierachal Clustering

Distance between all cells was calculated from Spearman correlation across all cells considering either; all genes, the 3006 genes derived from 100-pool Deseq2 analysis, top 1000 variable genes or the Lineage Intersect gene list. This distance matrix was then clustered using ward's method (R parameter, `ward.D`).

tSNE nearest-neighbor cell network

We used a weighted PCA analysis considering only the 171 genes on the Lineage Intersect gene list as input into tSNE projection (R package, `Rtsne`) (Figure S3A-C). For each of the 171 genes the two pvalues from the individual Lineage lists were weighted by their associated fold change, and combined using the Stouffers Z-score method (code sourced from Wikipedia, https://en.wikipedia.org/wiki/Fisher's_method). Thus deriving a Z score for each gene. In this way genes that had a low pvalue, and a high fold change, across both of the individual Lineage lists, received a higher Z score. Example Z scores for DA genes including Lmx1a, Foxa1 and En1 were 28.68, 17.15 and 7.44 respectively. This range was then scaled to between 1.5-1 such that these same genes now had values of 1.500, 1.269 and 1.076 respectively. PCA was performed, and the resulting loadings for each gene were multiplied by that gene's scaled Stouffer Z score. This weighted PCA was then used as input into tSNE.

We considered the first 70 Principle Components as input into tSNE, and projected all cells down into five dimensions. Within this space the Euclidean distance between a cell and its twelve nearest neighbors was used to construct an adjacency matrix, where a directed edge from a given cell and its nearest neighbor was given a weight according to;

$$\text{Graph's Maximum Euclidian Distance} - \text{Pairwise Euclidean Distance}$$

$$\text{Graph's Maximum Euclidian Distance}$$

Also, a reciprocal reverse edge of the same weight was recorded. This directed weighted adjacency matrix was then graphed using force directed graphing (R package, igraph, fruchterman reingold layout) creating a network across all cells. In this network, cells inter-connected by multiple, strongly weighted edges, synergistically pull between one another, and coalesce upon one location in the graph. Cells with no common edges are indirectly repelled, and occupy disparate locations of the graph. In this way we resolved three parallel axis, as shown in Figure 2. Of note, when considering this same Lineage Intersect gene list, multiple 21iinformatics tools including; Monocle (Trapnell et al., 2014), Diffusions Maps (Haghverdi et al., 2015), and Backspin bi-clustering (Zeisel et al., 2015), could often segregate GFP-positive from GFP-negative differentiation axes. However using these tools we were unable to completely segregate the two continuous unbroken parallel GFP-positive pseudotime axes reconstructed by the tSNE nearest-neighbor cell-network (Figure S3E-I).

The tSNE projection initially requires a random seed to be set, such that the final relative position of all cells in five dimensional space, while very similar, can vary slightly from run to run. To account for this variability the adjacency matrix of four runs were averaged to give the final adjacency matrix and graph (Figure 2C-E), where random cell associations due to the random seed were minimized. To verify this, ten independent graphs were generated consecutively, showing minimal variation in the overall graph structure across all runs (Figure S3A).

When considering these ten consecutive runs, it became apparent that weighting the original PCA input from 1.5-1 produced a cleaner graph than with no weighting (Figure S3A-ii). More often the weighted graphs displayed two clearly distinct parallel pseudotime axes within the GFP positive population (Figure S3A-i), an expected pattern if sub-lineages of neurons exist within the Lmx1a lineage. By way of control, reversing the weighting from 1-1.5 (genes with a lower Z score are weighted higher) produced noisier graphs, with less clear pseudotime axes, and more inappropriate edges between seemingly unrelated cells (Figure S3A-iii). Such an observation suggests that genes important for defining distinct, but highly related lineages may be few, but can be identified and appropriately weighted using rational approaches that consider a gene's enrichment between biologically informative bulk RNA samples.

Isolation of lineage-defining genes - SCDE and WGCNA workflow

Many existing single-cell RNA-seq studies have stressed the requirement to filter out unwanted and thereby uninformative noise introducing genes, before attempting to cluster genes into gene sets (Langfelder and Horvath, 2008). Since we wished to isolate gene signatures describing lineage, we thus sought to deplete our gene list of differentiation-associated genes in an unbiased, data-driven fashion. We applied various steps to this effect, as outlined below.

To leverage the single-cell RNAseq data to find novel lineage related genes not found in the bulk RNA samples, we began by dividing the graph into 27 unsupervised clusters comprised of similar cells using Infomap community detection (Figure S4A-i)(R package, igraph) (Rosvall and Bergstrom, 2008). We then performed pairwise SCDE between clusters whose mean pseudotime was within 10% of the maximum pseudotime range, effectively comparing progenitor clusters with other progenitor clusters, or neuron clusters with other neuron clusters (Figure S4A-ii, schematic), thereby depleting the list of genes that change only over differentiation time. Further, only genes that satisfied a padjvalue < 0.01 and a fold change cutoff of two in at least one pairwise comparison were considered for downstream analysis, leaving 4748 genes.

Stouffers Z-score method was again used to combine the pvalues for each gene across all pairwise comparisons, after weighting for fold change, to give a Z score for each gene. When ordered by Z score, many classic transcriptional lineage determinants occupied the top ranks (Figure S4A-iii). As an additional filtering step before proceeding to WGCNA gene clustering, we considered only those genes with a Z score above the mode (Figure S5A-iii, red line), and a Stouffer padjval < 0.001, a total of 2156 genes.

WGCNA clustered these genes into sets according to their topological overlap across all cells, and also returned a padj value measuring how significantly a gene belonged to the gene set that it was assigned (R package, WGCNA). Known markers for dopamine neurons or progenitors were always among the top significant genes in their cluster. However, despite the previous filtering steps, many genes' padj values were greater than 0.01, indicating that they did not significantly fit into their assigned

gene cluster. Reasoning that an informative gene must fall significantly into it's cluster, we then removed 543 genes, each with a padjvalue > 0.01, and re-ran the WGCNA clustering. This yielded a new set of similar, more refined gene clusters. However in this subsequent clustering context, new genes now had a padjvalue > 0.01. We thus repeated this filtering step for seven additional cycles, removing respectively 205, 33, 8, 2, 5, 5, and finally 1 gene, in each cycle. All genes now had a padjval < 0.01.

This workflow resulted in 1354 differentially expressed lineage informative genes clustered into similarly expressed gene signatures. Using WGCNA, each gene signature was then simulated as an artificial transcriptome *in silico* and the Spearman correlation calculated against each cell. This correlation value was then plotted onto each cell on the tSNE graphs (Figure 3, Figure S4B).

Immunohistochemistry

Tissue was fixed in 4% PFA, embedded in OCT and sectioned using a cryostat. Sections were air dried for 30min and washed 3x5min in PBS/0.1% triton-x. When necessary, antigen retrieval was performed by heating sections for 20min at 90°C in Dako Target Retrieval Solution (S1699), followed by 3x5min washes in PBS/0.1% triton-x. Alternatively, for E17.5 sections antigen retrieval was performed using a 5min wash in PBS/1% SDS, followed by 3x5min washes in PBS/0.1%triton-x. Sections were then blocked in PBS/3% donkey serum/0.1% triton-x for one hour at room temperature, before application of primary antibodies overnight at 4°C or room temperature, including; Th (Pelfreeze, 1:1000, P40101-150 or P60101-150), EphA3 (1:100, R&Dsystems, AF640), Nkx2-1 (1:100, Abcam, ab-133737), Pitx2 (1:200, R&Dsystems, AF7388), Barhl1 (1:1000, Novusbio, NBP1-86513), Dbx1 (1:20, Sigma, HPA039802), GFP (1:200, Abcam, ab-13970), Corin (1:100, gift from Malin Parmer laboratory), Nr4a2 (1:1000, SantaCruz, sc-990), FoxA2 (1:100, SantaCruz, sc-6554), Pitx3 (1:8000, home made) , En1 (1:10, DSHB, 4G11), Otx2 (1:2000, Neuromics, GT15095), vGlut2 (1:1000, Synaptic Systems, 135-404) and Foxb1 (1:400, Abcam ab5274). Sections were washed 3x5min in PBS/0.1%triton-x before application of secondary antibodies for 35min at room temperature. Finally sections were washed 3x5min in PBS/0.1%triton-x and mounted.

Animals

Lmx1a/b double flox mice were crossed with CAGG-CreERT2. The generation of these mouse strains have been described previously (Deng et al., 2011; Hayashi and McMahon, 2002; Zhao et al., 2006). Pregnant females were treated with Tamoxifen by gavage at E10.5, embryos collected at E13.5 or E17.5, and processed for immunohistochemistry.

mESC differentiations

For the monolayer protocol, E14.1 mESCs were propagated on gelatinized culture dishes in DMEM (Invitrogen) supplemented with 2000 U/ml LIF (Chemicon), 10% KSR, 2% FCS, 0.1 mM nonessential amino acids, 1 mM pyruvate (Invitrogen), and 0.1 mM beta-mercaptoethanol (Sigma). Upon differentiation cells were plated in N2B27 media supplemented with Ag1.3 Patched agonist (0.1uM, Curtis), Fgf2 (20ng/ml, Invitrogen), Fgf8 (100ng/ml, Invitrogen) and CHIR99021 (0-4uM, Merkmillipore) for 3 days. Cells were then re-split, 1:3 for each condition, on Polyornithine- and laminin- (Sigma) coated 48-well plates in the same media. After one day (at day 4 of differentiation), the growth factors were removed and the cells were subsequently kept in N2B27 medium containing GDNF, BDNF and ascorbic acid for a further 3 days. cDNA libraries were generated using OligodT primers (LifeTechnologies) and micro RNEasy kits (Qiagen), and Sybergreen (LifeTechnologies) qPCR was performed on a ViiA7 system (annealing/elongation at 59C). For primer sequences see Table S3. For each gene, fold change was calculated using the $\Delta \Delta Ct$ method against a standard curve run in parallel from CD1 mouse vMB control cDNA, and normalized against L19 mRNA levels. Plots were generated in R, and colored for fold change normalized and centered across all to CHIR99021 concentrations.

For the adapted 5-stage protocol cells were initially plated for 3 days in EB medium containing 100ng/Noggin (R&D), Ag1.3 Patched agonist (0.1uM, Curtis) and Fgf2 (20ng/ml, Invitrogen). Subsequently, EBs were transferred to ITSFn media supplemented with 100ng/Noggin (R&D), Ag1.3 Patched agonist (0.1uM, Curtis), Fgf2 (20ng/ml, Invitrogen) and CHIR99021 (0-4uM, Millipore) in poly-O/laminin/fibronectin coated dishes. After three days, noggin was removed from this supplement cocktail

and replaced with Fgf8 (100ng/ml, Invitrogen) in N3 media, upon which the cells were cultured for a further four days. Cells were then fixed and processed for immunohistochemistry.

Lentivirus experiments

FoxA2-T2A-Lmx1a or Barhl1-FoxA2-T2A-Lmx1a ORFs were PCR'd from CD1 E11.5 vMB cDNA, cloned beside the pTRE3G-BI DOX inducible promoter (631332, Clontech), and inserted into a pRRL.SIN.cPPT.PGK-GFP.WPRE backbone (Addgene), in place of the PGK-GFP segment. Viral particles were produced as previously described (Panman et al., 2011), and mESCs were transduced upon initial seeding of embryoid bodies. Doxycycline was added to media between days 3-8. Upon which cells were washed 3x with PBS, before changing of fresh N3 media. After a further two days cells were fixed and processed for immunohistochemistry. Doxycycline was applied at three different concentrations, 100, 500 and 1000ng/ml, however no significant differences were seen between these groups, and thus the data were merged into a single treatment group. For each of FL and B-FL treatments, 60 images were taken across 12 wells each (5 images per well).

Human tissue experiments

Human fetal tissue was obtained from legally terminated embryos with approval of the Swedish National Board of Health and Welfare in accordance with existing guidelines including informed consent from women seeking elective abortions. The gestational age of each embryo was determined by measuring the crown-to-rump length and neck-to-rump length. For qPCRs, RNA was isolated from human fetal tissue with the Qiagen RNEasy Lipid Tissue kit, and 1ug of RNA was used for cDNA synthesis. Quantitative real-time PCR (qRT-PCR) was performed on cDNA using Sybr Green (Roche) on a LightCycler instrument with a 2-step protocol (annealing/elongation at 60C). Data was analysed using the $\Delta\Delta Ct$ method, presented as the average fold change relative to human fetal liver using two different housekeeping genes (ACTB and GAPDH). For primer sequences see Table S4. Tissue sections were processed with DAB immunohistochemistry using the previously described primary antibodies. For hESC differentiations, H9 cells (WiCell) were maintained on Lam-521 (Biolamina) in iPS Brew (Miltenyi

Biotec). For differentiation towards various rostro-caudal fates, we applied an embryoid body protocol as described previously (Kirkeby et al 2012) with various different concentrations of the GSK3 inhibitor CHIR99021 from day 0-9 of differentiation. The cells were simultaneously ventralised with 200 ng/ml SHH-C24II (R&D Systems) and 0.5 uM purmorphamine from day 0-9, and the cells were fixed on day 17 for immunocytochemistry for rostro-caudal markers. Cells were stained with the same antibodies as for mouse.

Supplemental References

- Balasubramanian, R., Bui, A., Xie, X., Deng, M., and Gan, L. (2014). Generation and characterization of Lhx9-GFPCreER(T2) knock-in mouse line. *Genesis* *52*, 827–832.
- Chakrabarty, K., Oerthel, von, L., Hellemons, A., Clotman, F., Espana, A., Groot Koerkamp, M., Holstege, F.C.P., Pasterkamp, R.J., and Smidt, M.P. (2012). Genome wide expression profiling of the mesodiencephalic region identifies novel factors involved in early and late dopaminergic development. *Biol Open* *1*, 693–704.
- Chapman, H., Waclaw, R.R., Pei, Z., Nakafuku, M., and Campbell, K. (2013). The homeobox gene Gsx2 controls the timing of oligodendroglial fate specification in mouse lateral ganglionic eminence progenitors. *Development* *140*, 2289–2298.
- Deng, Q., Andersson, E., Hedlund, E., Alekseenko, Z., Coppola, E., Panman, L., Millonig, J.H., Brunet, J.-F., Ericson, J., and Perlmann, T. (2011). Specific and integrated roles of Lmx1a, Lmx1b and Phox2a in ventral midbrain development. *Development* *138*, 3399–3408.
- Haghverdi, L., Buettner, F., and Theis, F.J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*.
- Hayashi, S., and McMahon, A.P. (2002). Efficient recombination in diverse tissues by a tamoxifen-inducible form of Cre: a tool for temporally regulated gene activation/inactivation in the mouse. *Dev Biol* *244*, 305–318.
- Hayes, L., Zhang, Z., Albert, P., Zervas, M., and Ahn, S. (2011). Timing of Sonic hedgehog and Gli1 expression segregates midbrain dopamine neurons. *J Comp Neurol* *519*, 3001–3018.
- Hockemeyer, D., Soldner, F., Cook, E.G., Gao, Q., Mitalipova, M., Jaenisch, R. (2008). A drug-inducible system for direct reprogramming of human somatic cells to pluripotency. *Cell Stem Cell* vol. 3 (3) pp. 346-53.
- Inamura, N., Ono, K., Takebayashi, H., Zalc, B., and Ikenaka, K. (2011). Olig2 lineage cells generate GABAergic neurons in the prethalamic nuclei, including the zona incerta, ventral lateral geniculate nucleus and reticular thalamic nucleus. *Dev. Neurosci.* *33*, 118–129.
- Lahti, L., Achim, K., and Partanen, J. (2013). Molecular regulation of GABAergic neuron differentiation and diversity in the developing midbrain. *Acta Physiol (Oxf)* *207*, 616–627.
- Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* *9*, 559.

- Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* *15*, 550.
- Marin, O., Anderson, S.A., and Rubenstein, J.L. (2000). Origin and molecular specification of striatal interneurons. *J Neurosci* *20*, 6063–6076.
- Matsunaga, E., Araki, I., and Nakamura, H. (2001). Role of Pax3/7 in the tectum regionalization. *Development* *128*, 4069–4077.
- Metzakopian, E., Lin, W., Salmon-Divon, M., Dvinge, H., Andersson, E., Ericson, J., Perlmann, T., Whitsett, J.A., Bertone, P., and Ang, S.-L. (2012). Genome-wide characterization of Foxa2 targets reveals upregulation of floor plate genes and repression of ventrolateral genes in midbrain dopaminergic progenitors. *Development* *139*, 2625–2634.
- Michaud, J.L., Rosenquist, T., May, N.R., and Fan, C.M. (1998). Development of neuroendocrine lineages requires the bHLH-PAS transcription factor SIM1. *Genes Dev* *12*, 3264–3275.
- Panman, L., Andersson, E., Alekseenko, Z., Hedlund, E., Kee, N., Mong, J., Uhde, C.W., Deng, Q., Sandberg, R., Stanton, L.W., Ericson, J., Perlmann, T. (2011). Transcription factor-induced lineage selection of stem-cell-derived neural progenitor cells. *Cell Stem Cell* vol. 8 (6) pp. 663-75.
- Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* *9*, 171–181.
- Rosvall, M., and Bergstrom, C.T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* *105*, 1118–1123.
- Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). the dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* *1*–11.
- Yee, C.L., Wang, Y., Anderson, S., Ekker, M., and Rubenstein, J.L.R. (2009). Arcuate nucleus expression of NKX2.1 and DLX and lineages expressing these transcription factors in neuropeptide Y(+), proopiomelanocortin(+), and tyrosine hydroxylase(+) neurons in neonatal and adult mice. *J Comp Neurol* *517*, 37–50.
- Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* *347*, 1138–1142.
- Zhao, Z.-Q., Scott, M., Chiechio, S., Wang, J.-S., Renner, K.J., Gereau, R.W., Johnson, R.L., Deneris, E.S., and Chen, Z.-F. (2006). Lmx1b is required for maintenance of central serotonergic neurons and mice lacking central serotonergic system exhibit normal locomotor activity. *J Neurosci* *26*, 12781–12788.