bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Temporal and spatial variation among single dopaminergic neuron transcriptomes informs cellular phenotype diversity and Parkinson's Disease gene prioritization

Paul W. Hook[1], Sarah A. McClymont[1], Gabrielle H. Cannon[1], William D. Law[1], Loyal A. Goff[1,2*], Andrew S. McCallion[1,3,4*]

[1]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America
[2]Department of Neuroscience, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America
[3]Department of Comparative and Molecular Pathobiology, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America
[4]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America

*, To whom correspondence should be addressed: andy@jhmi.edu and loyalgoff@jhmi.edu

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT

Parkinson's Disease (PD) causes collapse of *substantia nigra* (SN) dopaminergic (DA) neurons of the midbrain (MB), while other DA populations are relatively spared. Here, we used single-cell RNA-seq (scRNA-seq) to characterize DA neuron populations in the mouse brain at embryonic and postnatal timepoints. These data allow for the discrimination between olfactory bulb (OB), forebrain (FB), and MB DA populations as well identification of subpopulations of DA neurons in each region. We observe a longitudinal axis of MB DA development, during which specialization and heterogeneity increases. We identify three distinct subpopulations of known MB DA neurons and provide evidence of a postnatal MB DA precursor, identifying novel markers for each subpopulation. Further, we discover gene regulatory networks (GRNs) that are significantly associated with neurodegenerative diseases and highly correlated with specific DA neuron subpopulations. By integrating these data with published genome-wide association studies (GWAS), we prioritize candidate genes in all 32 PD associated loci. Collectively, our data reveal genes and pathways that may begin to explain the selective vulnerability of SN DA neurons and allow for the systematic prioritization of genes in PD GWAS loci for functional evaluation.

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Parkinson's Disease (PD) is the most common progressive neurodegenerative movement disorder. Incidence of PD increases with age[1,2] affecting an estimated 1% worldwide beyond 70 years of age[3]. Although PD ultimately impacts multiple neuronal centers, preferential degeneration of the ventral midbrain (VM) dopamine (DA) neurons leading to collapse of the nigrostriatal pathway is a common theme.

Mesencephalic DA neurons and their efferent connections with the striatum are responsible for the acquisition and maintenance of fine motor control and reward pathways. In turn, motor control is largely dependent on DA neurons populating the *substantia nigra* (SN), whereas the ventral tegmental area (VTA) is responsible for reward based behaviors and satiety. Despite their shared neurotransmitter characteristic, PD compromises the viability of SN DA neurons preferentially. By contrast VTA and VM DA periaqueductal gray area (PAG) DA neurons are largely spared[4,5]. This fact has driven research interest in the genetic basis of SN vulnerability in PD compared with that of VTA/PAG DA neurons.

To date, of the more than 20 genes that have been implicated in familial PD, mutations in less than 10 have been robustly shown to explain disease expression[6,7]. Beyond rare familial cases, a recent meta-analysis of PD GWAS highlighted 32 loci associated with sporadic PD susceptibility[8]. While some GWAS loci contain genes known to be mutated in familial PD (*SNCA* and *LRRK2*)[6,7], most do not contain a known causal gene. The inability to systematically identify the causative gene/s within GWA loci establishes a roadblock to the translation of genetic findings to medical practice. This requires an understanding of the pathogenesis of the disease and a thorough characterization of the specific cell population/s affected. In PD, one can reasonably assert that a significant fraction of disease-associated variation likely mediates its influence specifically within the SN. The answers to the implicitly related question of what renders SN DA neurons more vulnerable than other DA neurons also depends on the impact of such variation on gene regulatory networks (GRNs) essential to their viability or function regardless of whether they are unique to SN or shared among DA neurons more widely.

In an effort to resolve heterogeneity among central nervous system (CNS) DA populations, we undertook single-cell RNA-seq (scRNA-seq) analyses of CNS DA neurons from discrete anatomical regions of both embryonic and postnatal mouse brains. We evaluated both MB and forebrain (FB) DA neurons at embryonic day 15.5 (E15.5) and expanded our analyses at postnatal day 7 (P7) to include DA neurons isolated from the olfactory bulb (OB), FB (posterior hypothalamus) and MB. Deeper analysis of the P7 MB allowed for refinement of DA neuronal composition including elucidation of transcriptomic differences and similarities of different anatomical DA populations, identification of novel genetic markers for the SN, and the characterization of modules of co-expressed genes in our data. The results of our analyses provide a framework within which we begin to prioritize and test hypotheses of the potential disease modulating role played by genes within PD GWAS loci.

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RESULTS

Temporal scRNA-seq characterization of DA neuronal populations reveals axis of DA neuron development

To characterize DA neuronal molecular phenotypes, we undertook scRNA-seq on cells isolated from distinct anatomical locations of the mouse brain, over developmental time. To obtain DA populations, we used the Tg(Th-EGFP)DJ76Gsat BAC transgenic mouse line, expressing EGFP under the control of the tyrosine hydroxylase (*Th*) locus. We microdissected both MB and FB from E15.5 mice, extending our analyses to MB, FB, and OB in P7 mice (Figure 1a). E15.5 and P7 time points were chosen based on their representation of stable MB DA populations, either after neuron birth (E15.5) or between periods of programmed cell death (P7) (Figure 1a)[9]. We used fluorescence activated cell sorting (FACS) to retrieve single eGFP$^+$ cells from enzymatically dissociated samples (Methods).

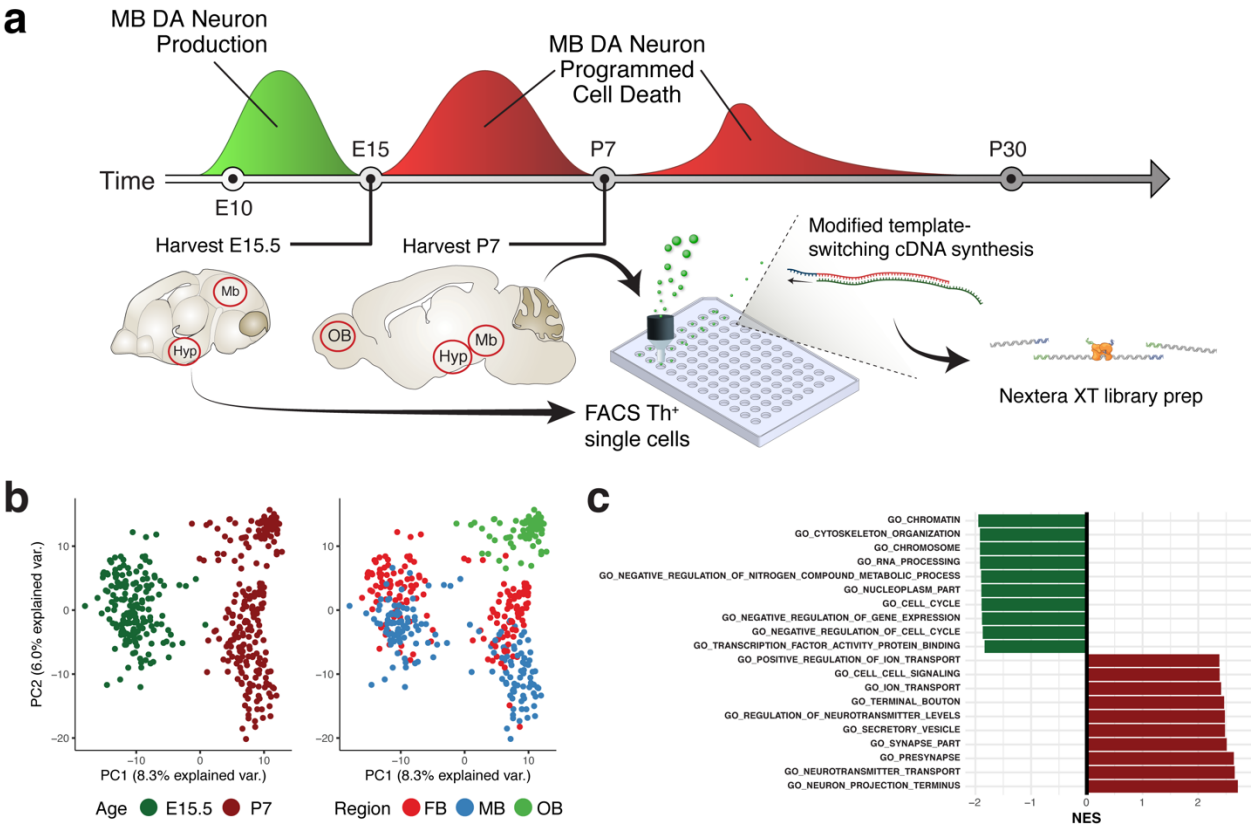

Figure 1. scRNA-seq analysis of isolated cells allows their separation by developmental time. a) Diagram of the experimental procedures. scRNA-seq of individual, GFP$^+$ cells was carried out through dissociation and FACS of GFP$^+$ cells from specific brain regions at E15.5 (MB, FB) and P7 (MB, FB, OB). GFP$^+$ cells were sorted into individual wells and cDNA was prepared via a modified Smart-Seq2 protocol and sequenced on a HiSeq 2500. E15.5 and P7 timepoints were chosen because they represent stable populations that are neither undergoing cell birth nor programmed cell death. b) Principal component analysis (PCA) on all collected cells using genes with a high biological coefficient of variation (BCV). The greatest source of variation (PC1) is explained by the time point at which the cells were collected, not the region from which the cells were collected. c) The top ten Gene Ontology

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(GO) gene sets enriched in genes with positive (red) and negative (green) PC1 loadings. Analysis was carried out using Gene Set Enrichment Analysis (GSEA) PreRanked analysis. Genes with negative PC1 loadings and negative normalized enrichment scores (NES) were enriched for terms indicative of mitotically active cells such as "cell cycle," whereas genes with positive PC1 loadings and NES scores were enriched for terms expected of a more mature neuron such as "neuron projection terminus."

We sequenced RNA from single cells to an average depth of ~8.0 x 10^5 50bp paired-end fragments per cell. Using Monocle[10], we converted normalized expression estimates into estimates of RNA copies per cell. Cells were filtered based on the distributions of total mass, total number of mRNAs, and total number of expressed genes per cell (Figure S1, detailed in Methods). After QC, 410 out of 473 cells were retained. Using principal component analysis (PCA), we identified and removed 14 outliers determined to be astrocytes, microglia, or oligodendrocytes (Figure S2; Table S1), leaving 396 cells (~79 cells/timepoint-region; Figure S1d).

Following a workflow similar to the recently described "dpFeature" procedure[11], we first identified highly variant genes within the data. We then selected the PCs that described the highest percentages of variance in the data using these to represent the cells in two dimensions using t-Stochastic Neighbor Embedding (t-SNE)[12]. We called clusters of related cells within the data in an unbiased manner (see Methods). As anticipated, we observed that the greatest source of variation was between timepoints (Figure 1b). Genes associated with negative PC1 loadings (E15.5 cells) were enriched for gene sets consistent with mitotically active neuronal precursors (Figure 1c). In contrast, genes associated with positive PC1 loadings (P7 cells) were enriched for GO terms associated with mature, post-mitotic neurons (Figure 1c).

Analyses by region and timepoint reveal additional novel neuronal diversity.

Consistent with the suggestion that the embryonic cells include a less diverse progenitor population, analysis of all cells revealed that the E15.5 cells from both MB and FB cluster together (Figure 2a). By contrast, cells isolated at P7 mostly cluster by anatomical region, suggesting progressive functional divergence with time (Figure 2a). We then applied the scRNA-seq analysis workflow in a recursive manner in all regions at both timepoints to further explore heterogeneity. This revealed a total of 13 clusters (E15.5 FB, 2; MB, 2; P7 OB, 3; FB, 2; MB, 4; Figure 2b). Using known markers, we established that all clusters expressed high levels of pan-neuronal markers (*Snap25*, *Eno2*, and *Syt1*) (Figure S3). In contrast, we found weak or no, evidence of astrocyte (*Aldh1l1*, *Slc1a3*, *Aqp4*, and *Gfap*) or oligodendrocyte markers (*Mag*, *Mog*, and *Mbp*; Figure S3).

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

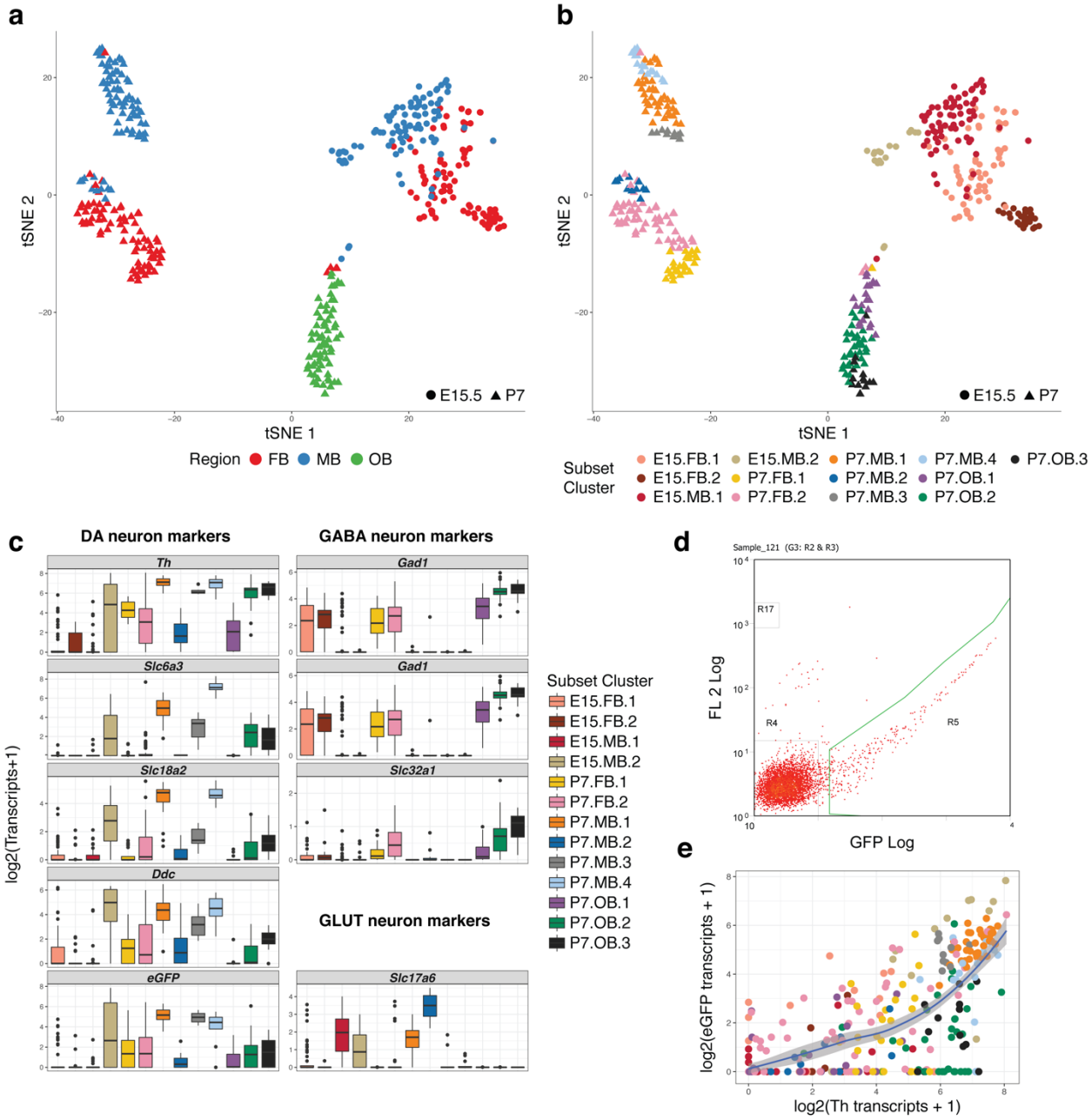

Figure 2. Subpopulations of isolated cells, identified through recursive analysis, express well-known neuronal markers. a) A t-distributed Stochastic Neighbor Embedding (t-SNE) plot of all collected cells colored by regional identity. It shows that E15.5 cells cluster together while P7 cells cluster primarily by regional identity. b) A t-SNE plot of all collected cells colored by subset cluster identity. Through iterative analysis, it was found that all timepoint-regions collected can be separated into multiple subpopulations (13 in total). Individual breakdowns of distinct subpopulations can bee seen in later figures c) Expression boxplots of well-established neuronal subtype markers including dopaminergic (DA) markers (*Th*, *Slc6a3*, *Slc18a2*, *Ddc*), GABAergic (GABA) markers (*Gad1*, *Gad2*, *Slc32a1*), and a glutamatergic (GLUT) maker (*Slc17a6*). The expression of eGFP in every cluster is also shown and mirrors the expression pattern of *Th*. d) A representative FACS plot showing the results of sorting E15.5 GFP^+ cells. A clear GFP^+ cell population was identified and that population has a wide range of GFP intensity. e) A plot of *Th* against eGFP log2 expression levels in all cells. A Loess fit regression line with a 95% confidence interval is shown, indicating that increasing levels of *Th* expression correspond to increasing levels of detected eGFP

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

transcripts. This plot also shows that *Th* transcript is more abundant than eGFP in the cells collected as indicated by the cells distributed along the x-axis with little to no eGFP expression.

We then evaluated the expression of known markers of DA neurons along with eGFP (Figure 2c). We detected consistently high levels of *Th* in cluster E15.MB.2 and all P7 clusters (Figure 2c) which correlated with eGFP expression (Figure 2c; Figure 2e). The inconsistent detection of *Th* and eGFP in other E15.5 clusters likely reflects their low transcript abundance at this time point, but sufficient expression of the eGFP reporter to permit FACS collection (Figure 2d). The expression of DA neuron markers *Ddc* and *Slc18a2* correlate with *Th* expression, while *Slc6a3* expression is more spatially and temporally restricted (Figure 2c).

Multiple studies have demonstrated that *Th*-expressing neurons may also express markers characteristic of other major neuronal subtypes[13–15]. Consequently, we evaluated expression of canonical markers of other neuronal subtypes in our DA neuron subpopulations. We found co-expression of *Th* with GABAergic (*Gad1*/*Gad2*/*Slc32a1*) or glutamtergic (*Slc17a6*) markers in 11/13 subset clusters (Figure 2c). The notable exception being two P7 MB DA neuron clusters (MB3 and MB4), which exclusively expressed DA markers (Figure 2c).

Analysis of E15.5 cells reveal regionally specialized maturing neurons

Recursive analysis of E15.5 DA neuron regions revealed two distinct populations in both the MB and FB. When analyzed collectively, we observe a major cluster, consisting of both MB and FB cells and two smaller clusters comprising solely MB or FB cells (Figure 3a). The discrete E15.5 MB cluster (E15.MB.2; Figure 3b) was highlighted by specific expression of genes known to mark mature MB DA neurons and have roles in MB DA neuron function (*Foxa1*, *Lmx1a*, *Pitx3*, and *Nr4a2*)[16] (Figure 3c; Table S2), suggesting that E15.MB.2 represents a post-mitotic, maturing DA neuron population. By contrast, E15.MB.1 neurons preferentially express genes including *Meis2*, *Lhx9*, *Id4*, *Ebf1*, *Pax5*, *Ephb1*, *Mir124-2hg*, and *Nrg1* (Table S2). All have established roles in neuronal precursors or neuronal differentiation/maturation[17–28]. Further, E15.MB.1 expresses *Slc17a6*, which along with *Meis2* and *Lhx9*, were recently used to identify embryonic DA neuroblasts[29] (Data accessed: 02/26/17; Table S11). Collectively these data support E15.MB.1 as a presumptive DA precursor population.

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

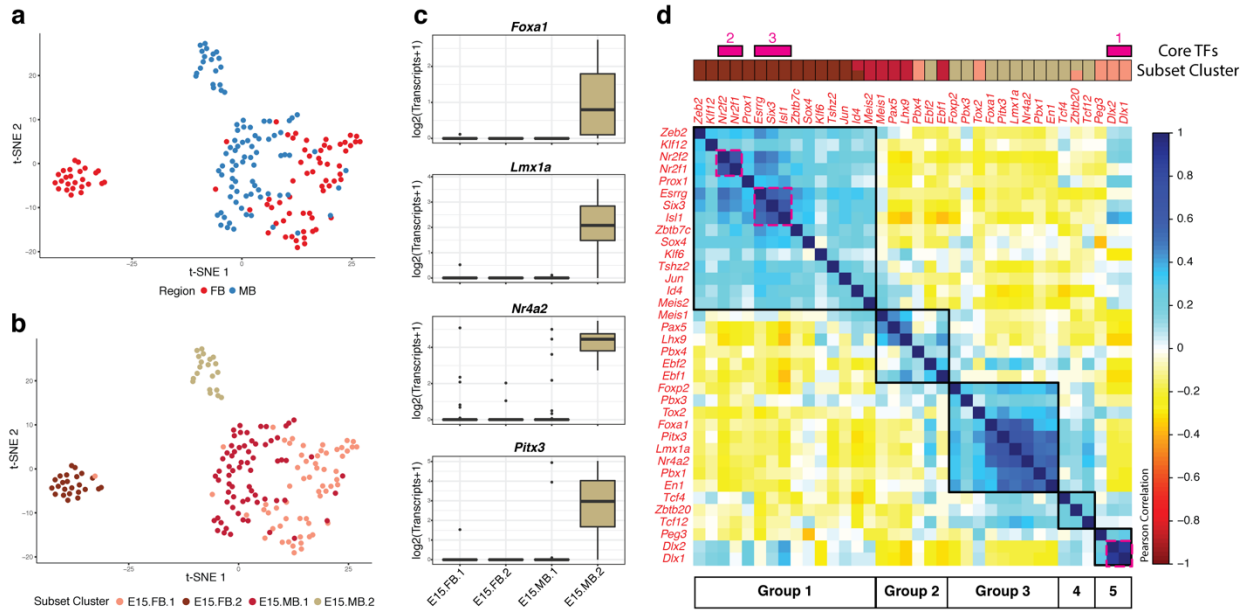

Figure 3. Analysis of E15.5 cells reveals clusters of neurons at different stages of maturation that may be defined by co-correlated transcription factor programs. a) A t-SNE plot of all neurons collected at E15.5 using high variance genes colored by regional identity. The t-SNE plot shows two distinct clusters containing only cells from one region and one cluster that contains a combination of both regions. b) The same t-SNE plot seen in Figure 3a, but colored by subset cluster identity. This plot shows that iterative analysis defines two clusters in both E15.5 MB and E15.5 FB. c) Boxplots showing log2 expression in all E15.5 clusters of four genes identified as being specifically expressed in the E15.MB.2 cluster. The expression of these genes (*Foxa1*, *Nr4a2*, *Lmx1a*, and *Pitx3*) as well as others indicate that cells belonging to the E15.MB.2 cluster are post-mitotic, maturing MB DA neurons. d) A correlation plot displaying the Pearson correlation between transcription factors (TFs) that were identified as being specifically expressed in any of the four E15.5 clusters identified through iterative analysis. Above the plot is a bar labeled "Subset Cluster" indicating (by color) the cluster or clusters in which each gene was determined to be specific. Five clusters (indicated by black boxes) are identified through hierarchical clustering that primarily groups TFs by which cluster they were identified in. Sets of "Core TFs" were identified in three groups through the use of bootstrap resampling of hierarchical clustering (see Methods). Sets of core TFs are indicated by pink boxes (3) within the plot and pink blocks above the plot labeled "Core TFs."

The markers identified for the discrete E15.FB.2 cluster, including *Six3* and *Six3os1*, are consistent with more mature FB/hypothalamic neurons[30–33]. This observation is supported by E15.5.FB.2 expression of *Sst* and *Npy*; both of which encode hormones indicative of specified, post-mitotic neurons[34]. E15.FB.1 clusters with E15.MB.1 (Figure 3b) potentially suggesting that it also represents an immature neuronal population. Indeed, the most specific marker for this population *Rnd3,* has been implicated in limiting the number of divisions of newly-fated neurons and in migration[35,36] (Table S2).

We further identified marker genes that are transcription factors (TFs) in order to define networks of TFs associated with these populations. Expression of these marker genes were correlated and hierarchical clustering was used to reveal five groups of TFs primarily defining

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the different E15.5 neuron populations (Figure 3d). Core groups of TFs were also identified within two of five groups (pink outline; Figure 3d; See Methods). Two sets of core TFs were identified within "Group 1" which was primarily composed of mature FB cluster (E15.FB.2) specific genes. Both core TF sets (2 and 3) contain genes that have been previously implicated in FB neuron development[30–32,37,38] while potentially implicating a new TF in *Esrrg*. Group 5, which defined the immature E15.5 FB cluster (E15.FB.1) contained one core TF set (set #1) containing *Dlx1* and *Dlx2*, both of which have established roles in early FB neuronal development[39].

P7 neurons display regionally discrete transcriptional signatures

In contrast to E15.5, DA neurons isolated at P7 mostly cluster by anatomical region (Figure 4a). We sought to identify genes displaying region-dependent expression, identifying 54, 14 and 85 genes that defined OB, FB and MB DA neurons, respectively (Table S2).

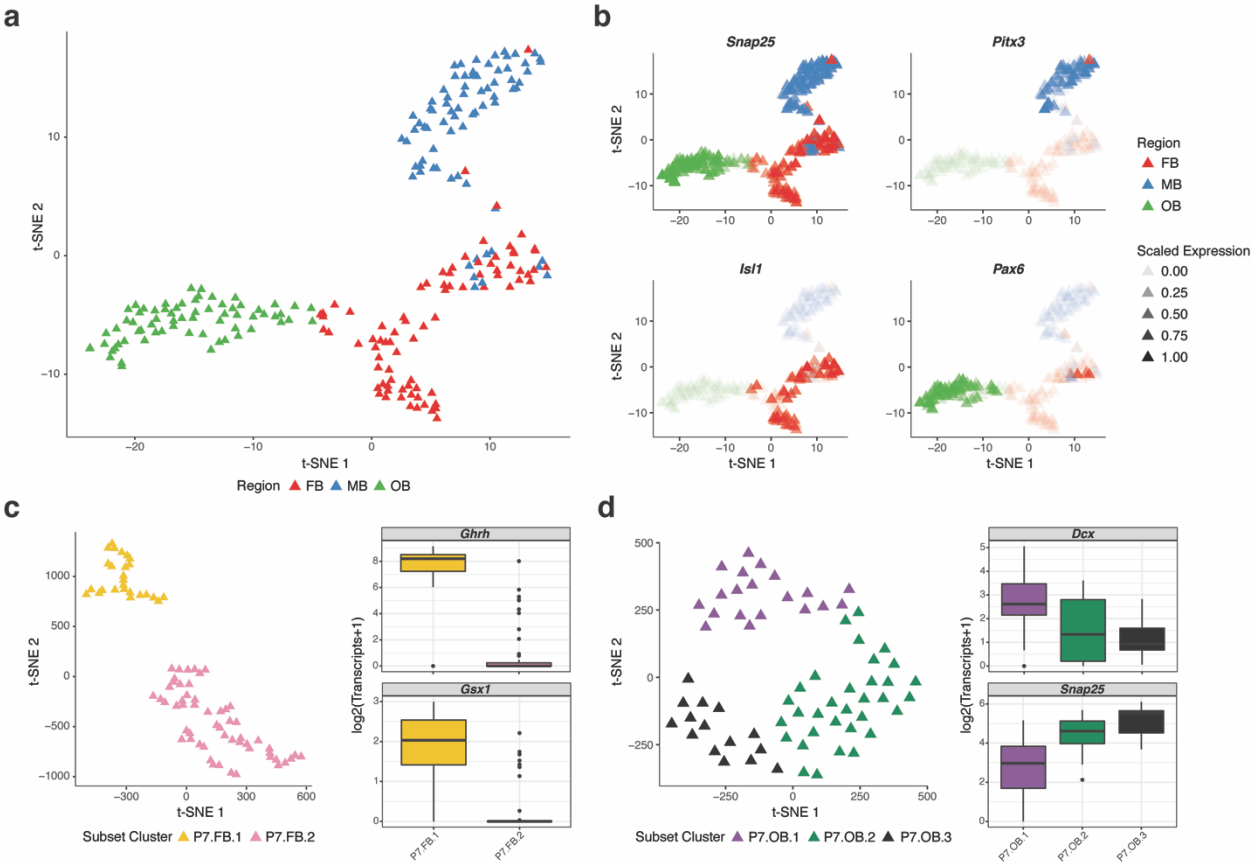

Figure 4. P7 *Th*+ neurons cluster by neuroanatomical domain, expressing markers characteristic of those regions, and further revealing regional subclusters based on iterative marker gene analyses. a) A t-SNE plot of all P7 neurons collected using high BCV genes, colored by regional identity. The neurons mostly cluster by regional identity. b) t-SNE plots showing expression scaled to max gene expression observed of a pan-neuronal gene (*Snap25*) and an example of a marker gene identified for each region at P7 (MB: *Pitx3*, FB: *Isl1*, OB: *Pax6*) in all P7 cells. c) t-SNE plot of P7 FB neurons plotted by using high BCV genes and colored by subcluster identity showing P7 FB neurons clustering into two distinct populations identified by ADPclust. Marker genes for the P7.FB.1 cluster

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

include *Ghrh* and *Gsx1*. Both of these genes as well as others indicate that this population of cells are tuberofundibular dopaminergic neurons located in the arcuate nucleus. d) A t-SNE plot of P7 OB neurons plotted by using high BCV genes and colored by subcluster identity. The plot shows that the P7 OB neurons cluster into three populations identified by ADPclust. Increase in expression of genes that indicate neuronal maturity (for example, *Snap25*) from P7.OB.1 to P7.OB.3 and decrease in expression of genes that indicate neuronal immaturity (for example, *Dcx*) in the same direction indicate that the clusters identified represent neuronal populations of increasing maturity.

The FB-restricted genes include markers associated with hypothalamic development and function e.g. *Isl1*[38] and *Asb4*[40] (Figure 4b; Table S2). Analyzing P7 FB Th^+ neurons alone revealed two distinct cell clusters (Figure 4c). P7.FB.1 specifically expressed the neuropeptides *Gal* and *Ghrh* and the *Gsx1* transcription factor (Figure 4c; Table S2). All three play roles in arcuate nucleus neurons[41–43] and were markers for a recently described $Th^+/Ghrh^+/Gal^+$ hypothalamic population[44]. By contrast, marker genes for P7.FB.2 did not reveal a signature or gene expression profile consistent with a known cellular phenotype (Table S2)[44,45]. However, several other arcuate nucleus markers for $Th^+/Ghrh^-$ neuronal populations were expressed in subsets of P7.FB.2 cells, including *Onecut2*, *Arx*, *Prlr*, *Slc6a3*, and *Sst* (Figure S4a)[44]. Thus, some Th^+ populations detected in other scRNA-seq analyses may be present within our data, but likely in insufficient numbers to facilitate classification here.

Of the genes whose expression defined OB Th^+ cells, many have established roles in the development or survival of OB DA neurons[46–51] (Table S2). Recursive analysis revealed three subset clusters in P7 OB (Figure 4d). In identifying marker genes for P7 OB subset clusters, we observed that P7.OB.1 expressed *Dcx* at significantly greater levels than P7.OB.2/P7.OB.3 and that *Dcx* levels decrease along a continuum towards the lowest expression in P7 OB3 (Figure 4d). *Dcx* expression diminishes with neuronal maturation with the lowest expression in adult neurons[52]. Consistent with this observation expression of the mature neuronal marker *Snap25* is anticorrelated with Dcx (Figure 4d), suggesting a progression in maturation from P7.OB.1 to P7.OB.3. This too is corroborated by concomitant increase in expression of DA neuron markers and OB DA neuron fate specification genes (Figure S4b)[53,54].

Many genes that define $eGFP^+$ MB neurons, including *Pitx3* (Figure 4b), have established roles in MB DA neuron development and biology[16]. We identified four P7 MB DA subset clusters within P7 MB DA neurons (Figure 5a). Marker gene analysis (Table S2) confirmed that three of the clusters correspond to DA neurons from the VTA (*Otx2* and *Neurod6*; P7.MB.1)[55,56], the PAG (*Vip* and *Pnoc*; P7.MB.3)[57,58], and the SN (*Sox6* and *Aldh1a7*; P7.MB.4)[55,59] (Figure 5b). These data are consistent with recent scRNA-seq studies of similar populations[29,60]. We further identify an as-yet undescribed population (P7.MB.2; Figure 5a) of Th^+ DA neurons. This population of cells display an expression signature consistent with a neuronal progenitor cell population. This postnatal population shares many markers with the progenitor-like E15.MB.1 cluster including *Fam19a2* and *Meis2* (Table S2; Figure 5b). Furthermore, P7.MB.2 markers

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Meis2, *Lhx9*, and *Ldb2* have been shown to mark embryonic mouse neuroblast populations[29]. Interestingly, this P7.MB.2 population clusters with P7 FB neurons in t-SNE space (Figure 2a; Figure 2b; Figure 4a). This may be driven by lower expression of key MB DA neuron genes compared to the other P7 MB clusters, resulting in a signature more similar to both P7 FB clusters (Figure 2c).

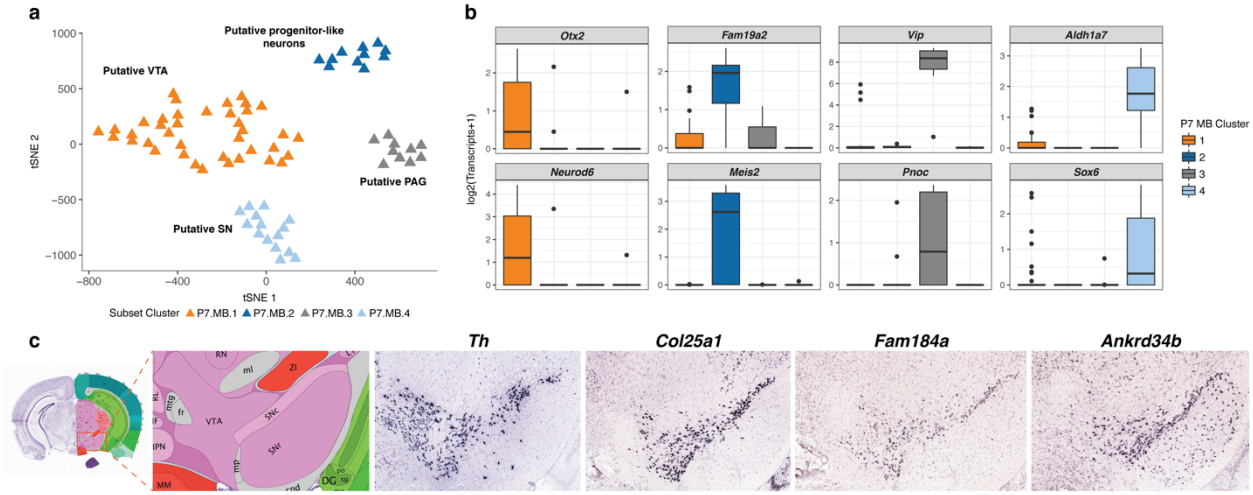

Figure 5. Identification of four distinct clusters of P7 MB *Th*$^+$ neurons based upon expression of established marker genes leads to the discovery of novel *substantia nigra* (SN) marker genes. a) t-SNE plot of P7 MB neurons colored by subcluster identity. P7 MB neurons separate into four clusters of which three can be defined using marker genes of known *Th*$^+$ populations: the ventral tegmental area (VTA), the *substantia nigra* (SN), and the periaqueductal grey area (PAG). The P7.MB.2 cluster was identified as a progenitor-like neuron primarily due to the overlap of specific markers with the immature E15.MB.1 cluster. b) Boxplots showing the expression of cluster specific genes used for identification of cluster populations. *Otx2* and *Neurod6* are established markers of VTA DA neurons. *Vip* and *Pnoc* are established markers of PAG neurons. *Aldh1a7* and *Sox6* are established markers of SN DA neurons. *Fam19a2* and *Meis2* are markers for P7.MB.2 that are shared with the E15.MB.1 cluster (See Table S2). c) Confirmation of novel SN DA neurons through the use of Allen Brain Atlas (ABA) *in situ* hybridization data (http://www.brain-map.org/). Coronal, P56 mouse *in situ* hybridization data was explored in order to confirm the expression of 37 novel SN markers. *Th* expression in P56 mice was used as an anatomical reference during analysis. Overall, 15/37 novel SN markers are confirmed to be expressed in the SN in adult mice, including *Col25a1*, *Fam184a*, and *Ankrd34b*.

We sought to ascertain the spatial distribution of P7.MB.2 DA neurons through multiplex, single molecule fluorescence *in situ* hybridization (smFISH) for *Th* (pan-P7 MB DA neurons), *Slc6a3* (P7.MB.1, P7.MB.3, P7.MB.4), and one of the marker genes identified through our analysis, either *Lhx9*/*Ldb2*/*Meis2* (P7.MB.2) (Figure 6). The ventral MB was scanned in each experiment for cells that were Th$^+$/Slc6a3$^-$ and positive for the third gene. *Th*$^+$/*Slc6a3*$^-$/*Lhx9*$^+$ cells were found scattered in the dorsal SN *pars compacta* (SNpc) along with cells expressing *Lhx9*$^+$ alone (Figure 6a, 6e). Expression of *Ldb2* was found to follow a similar pattern to *Lhx9*, with *Th*$^+$/*Slc6a3*$^-$/*Ldb2*$^+$ cells also found in the dorsal SNpc (Figure 6b, 6e). In the SNpc, *Meis2*$^+$ cells were common, however they did not display co-expression of *Th* (Figure 6d). Cells that were

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Th^{+}/*Slc6a3*$^{-}$/*Meis2*$^{+}$ were found in the interpeduncular nucleus (IPN) of the ventral MB (Figure 6d, 6e). Neither *Lhx9* nor *Ldb2* were detected in the IPN (data not shown). Expression of *Lhx9*, *Ldb2*, and *Meis2* was low or non-existent in *Th*$^{+}$/*Slc6a3*$^{+}$ cells in the SNpc (Figure 6a, 6b, 6c). Importantly, cells expressing these markers express *Th* at lower levels than *Th*$^{+}$/*Slc6a3*$^{+}$ neurons (Figure 6), consistent with our scRNA-seq data (Figure 2c).

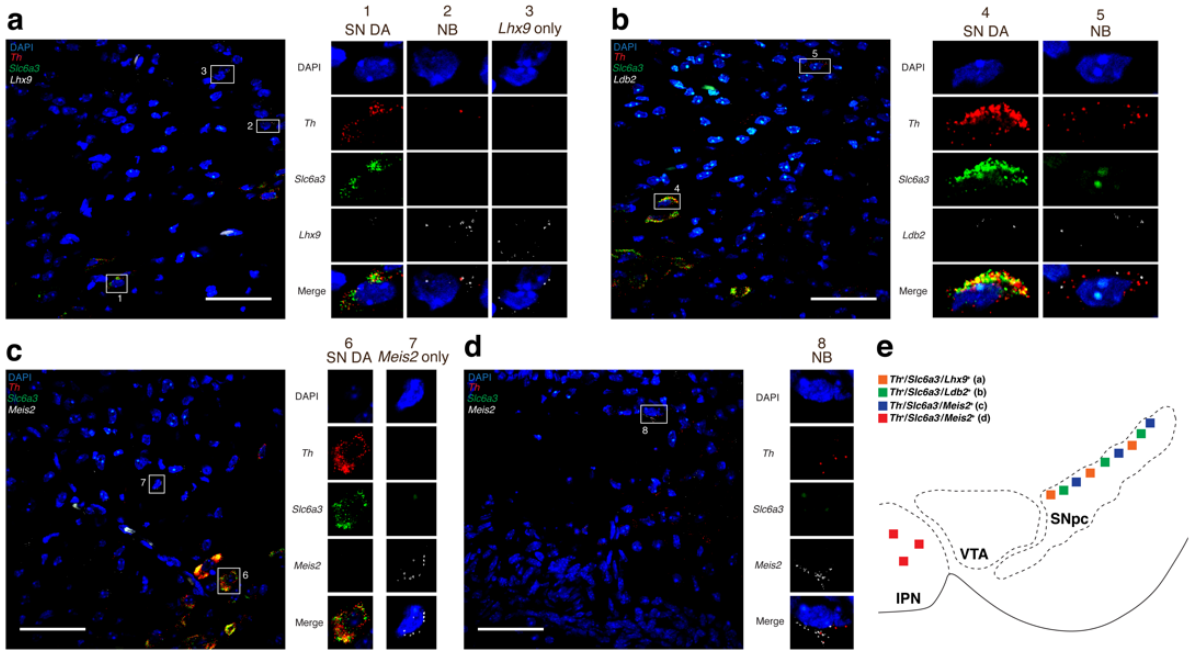

Figure 6. Single molecule fluorescent in situ hybridization (smFISH) validates the existence of a neuroblast-like P7.MB.2 cell population. Representative images of smFISH experiments analyzing four MB sections for each probe combination. a) smFISH performed with probes for *Th* (red), *Slc6a3* (green), and *Lhx9* (white). smFISH in the ventral MB reveals the presence of *Th*$^{+}$/*Slc6a3*$^{-}$/*Lhx9*$^{+}$ in the dorsal portion of the *substantia nigra* (SN) (inset 2), the presence of *Lhx9*$^{+}$ only cells in the dorsal SN (inset 3), and the lack of expression of *Lhx9* in SN DA neurons (inset 1). Scale bar, 50 uM. b) smFISH performed with probes for *Th* (red), *Slc6a3* (green), and *Ldb2* (white). smFISH in the ventral MB reveals the presence of *Th*$^{+}$/*Slc6a3*$^{-}$/*Ldb2*$^{+}$ in the dorsal portion of the SN (inset 5) with little to no expression of *Ldb2* in SN DA neurons (inset 1). Scale bar, 50 uM. c) smFISH performed with probes for *Th* (red), *Slc6a3* (green), and *Meis2* (white). smFISH in the ventral MB reveals the presence of *Meis2*$^{+}$ in the dorsal portion of the SN (inset 7) with little to no expression of *Meis2* in SN DA neurons (inset 6). Scale bar, 50 uM. d) smFISH performed with probes for *Th* (red), *Slc6a3* (green), and *Meis2* (white). smFISH in the ventral MB reveals the presence of *Th*$^{+}$/*Slc6a3*$^{-}$/*Meis2*$^{+}$ in the interpeduncular nucleus (IPN) (inset 8). Scale bar, 50 uM. e) Diagram summarizing locations of the P7 MB2 cells validated through smFISH. All images were taken at 60x magnification and processed using ImageJ (see Methods). NB, neuroblast; VTA, ventral tegmental area; SNpc, *substantia nigra* pars compacta; IPN, interpeduncular nucleus.

Furthermore, regional and subset cluster marker gene correlation analysis revealed four groups of TFs through hierarchical clustering with three groups clearly demarcating regions at P7 (Figure S5a) including seven core TF sets (three in OB, one in FB, and three in MB) (Figure S5a;

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S3). To expand upon these TF networks, we performed correlation analysis with all TFs that were found to be differentially expressed between regions at P7. Five out of seven (5/7) sets of core TFs found with P7 marker gene analysis were recovered (Figure S5b). These core groups of TFs were expanded through the addition of other differentially expressed TFs found to be highly correlated with the original core TFs (Figure S5b). Two other core groups ("FB program" and "SN program") were identified as containing TFs (*Isl1* and *Sox6*) known to be associated with P7 FB and P7 MB SN, respectively. Additional sets of "core" TFs were also identified (Table S3).

Identification of novel SN-specific DA Neuron marker genes

Motivated by the clinical relevance of SN DA neurons to PD, we set out to understand what makes them transcriptionally distinct from other MB DA neurons. We postulated that genes with specific expression in the P7 SN DA neuron cluster data might illuminate their preferential vulnerability in PD. By hypergeometric testing, the 67 SN marker genes are enriched for GO terms consistent with SN DA neuron biology (Table S4). Of the 67 SN-defining genes (Table S2), three (4.5%; *Ntf3*, *Chrna6*, and *Ntn1*) were shared with P7.MB.1 (VTA) and were excluded from subsequent analyses. Prior reports support the expression of 27/64 genes (42%) in postnatal SN (Table S5). We then sought evidence confirming SN expression for the remaining, novel 37 (58%) genes. Of these, 15/37 (~41%) were detected in adult SN neurons by *in situ* hybridization (ISH) from the Allen Brain Atlas (ABA) including *Col25a1*, *Fam184a*, and *Ankrd34b*. (Figure 5c, Table S6). The ABA lacks coronal ISH data on 20/37 genes; and for 2/37 genes ABA had relevant ISH data but lacked evidence of expression in the adult SN (*Tspan12*, *Igfbp3*) (Table S6). Collectively, we identify 64 postnatal SN DA marker genes and confirm the expression of those genes in the SN for 42 (65%) of them, included 15 previously undescribed markers.

Gene-coexpression modules are enriched for PD gene sets only in SN-derived data

In order to explore relationships between cellular subtype identity and transcriptional programs, we performed weighted gene co-expression network analysis (WGCNA)[61] on our P7 data. We used all expressed genes to establish 16 co-expressed gene modules (Figure 7; Figure S6; Table S7). We determined whether identified modules were enriched for Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, Gene Ontology (GO) gene sets, and Reactome gene sets[62]. Notably, "green" and "brown" modules were significantly associated with the Parkinson's Disease KEGG pathway gene set (Figure 7a) suggesting these two modules may specifically contribute to PD etiopathology. Further, the brown module was significantly associated with KEGG pathways that include "Cocaine addiction", while the green module was enriched for genes associated with additional neurodegenerative disorders including "Alzheimer's Disease" as well as "Oxidative "phosphorylation (Figure 7a). We also found the green module to be significantly associated with GO gene sets including select metabolic processes and mitochondrial function including sets related to the electron transport chain (Table S8-12).

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

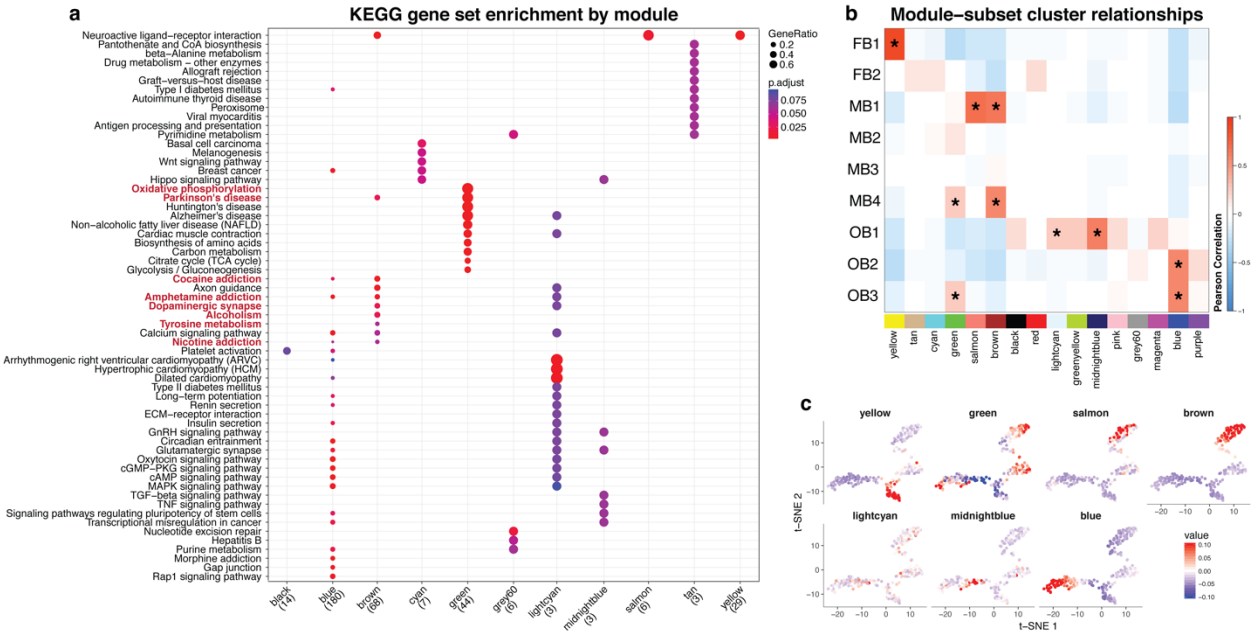

Figure 7. Weighted Gene Co-Expression Network Analysis (WGCNA) reveals gene modules that correlate with P7 *Th*$^+$ subset cluster identity and are enriched for PD gene sets. a) Enrichment for KEGG gene sets with gene modules identified through WGCNA analysis of P7 *Th*$^+$ neurons. Significance of enrichment was determined by hypergeometric test and an adjusted P-value (Benjamini & Hochberg correction) cutoff of 0.1 was used. The gene ratio (displayed by dot size) is the ratio of genes within a given module that are found in any given gene set. Notably, two identified modules ("green" and "brown") are enriched for the "Parkinson's Disease" KEGG gene set. Other notable gene sets are highlighted in red. b) Correlation heatmap of the Pearson correlation between module eigengenes and P7 *Th*$^+$ subset cluster identity. Modules are represented by their assigned colors at the bottom of the matrix. The significance (P-value) of the correlations was calculated using the Student's asymptotic test as implemented in the WGCNA R package with the function "corPvalueStudent()." Modules that had a positive correlation with a subset cluster and had a correlation P-value less than the Bonferroni corrected significance level (P-value = 4.386e-04) contain an asterisk. As shown, 7/16 modules display significant correlations with 6/9 of the P7 subset clusters including the VTA (P7.MB.1) and the *substantia nigra* (P7.MB.4). c) The eigengene value for each P7 neuron in the seven WGCNA modules shown to be significantly associated with a subset cluster overlaid on the t-SNE plot of all P7 neurons (Figure 4a). Most of these plots (6/7) show that positive eigengene values from these modules mark subsets of cells in t-SNE space corresponding to the subset clusters they are correlated with in Figure 7b. The "lightcyan" module does not seem to show the same spatial restriction, even though it is significantly correlated with P7.OB.1 identity (Figure 7b).

We next asked whether these biologically relevant gene expression modules were engaged by distinct P7 DA neurons subtypes. WGCNA analysis establishes eigengenes for each module that can facilitate their correlation with cellular traits. We calculated the pairwise correlations between module eigengenes and P7 subset clusters. This analysis revealed 7/16 modules significantly, positively correlated (Bonferroni corrected p < 3.5e-04) with at least one subset cluster, including the two gene sets (brown and green) enriched for PD (Figure 7b). The majority of these significant modules (6/7) displayed strict spatial enrichment in t-SNE space (Figure 7c)

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

confirming the correlations. Strikingly, the SN (P7.MB.4) was the only P7 subset cluster significantly associated with both modules enriched for PD gene sets (brown and green). The identification, and subtype-specific association of these modules, reinforces their significance in disease etiopathology and expands the scope of SN-associated genes identified above.

Integration of MB DA neuron subtype specificity enables prioritization of genes within PD-associated intervals

The capacity to make informed connections between GWAS loci and causal gene/s is often impeded by a paucity of expression data in biologically relevant cell populations. This is particularly true of disorders in which those affected populations are difficult to isolate, as in the mammalian CNS. We posited that SN DA neuron-specific genes and the broader gene co-expression networks that correlate with SN DA neurons might be used to help prioritize genes within loci identified in PD GWAS. Such a strategy would be unbiased and independent of genic position relative to the lead SNP or prior biological evidence.

To investigate pertinent genes within PD GWAS loci, we identified all human genes within a topologically association domain (TAD) encompassing each identified PD-associated lead SNP. We chose to use TAD boundaries because regulatory sequences preferentially interact with promoters in TADs[63]. Since data describing TAD structure of the cell types analyzed here does not exist, we also examined all the genes within +/- 1 megabase of a PD GWAS SNP. We selected this interval as it includes the upper bounds of reported enhancer-promoter interactions[64,65]. All PD GWAS SNPs interrogated were identified by the most recent meta-analysis (32 SNPs in total)[8], implicating a total of 966 unique genes. We then identified corresponding mouse homologs (673/966; ~70%), primarily through the NCBI Homologene database (Methods). Of the remaining 293 genes with no mouse homologs (Table S13), 62 (62/293, ~21%) are annotated as protein coding genes (Figure S7a). 17 loci include at least one protein coding gene with no identified mouse homolog (Figure S7b).

To prioritize the genes in all 32 loci, we developed a gene-centric score that integrates gene expression, differential gene expression, cluster specificity (Table S2), WGCNA module co-regulation (Table S7), and evolutionary mutation tolerance. We began by intersecting the PD loci genes with our scRNA-seq data, identifying 256 genes (256/673; 38%) with direct evidence of expression in SN DA neurons (P7.MB.4). Each PD-associated interval contained ≥1 SN-expressed gene (Table S14). Emphasizing the need for a systematic strategy, in 14/32 GWA intervals (~44%), the most proximal gene to the lead SNP was not detectably expressed in the mouse SN DA neuron population (Table S14; Figure S8).

Four loci (*MMP16*, *SIPA1L2, USP25*, *VPS13C*) contained only one SN-expressed gene (Figure 8a, Table S14): *Mmp16* (*MMP16* locus), *Tsnax* (*SIPA1L2* locus), *Hspa13* (*USP25* locus), and *Rora* (*VPS13C* locus). The relevance of these candidate genes is well supported[66–71].

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Furthermore, both *Mmp16* and *Rora* are detected in adult mouse SN (Figure 8c, ABA, Table S14).

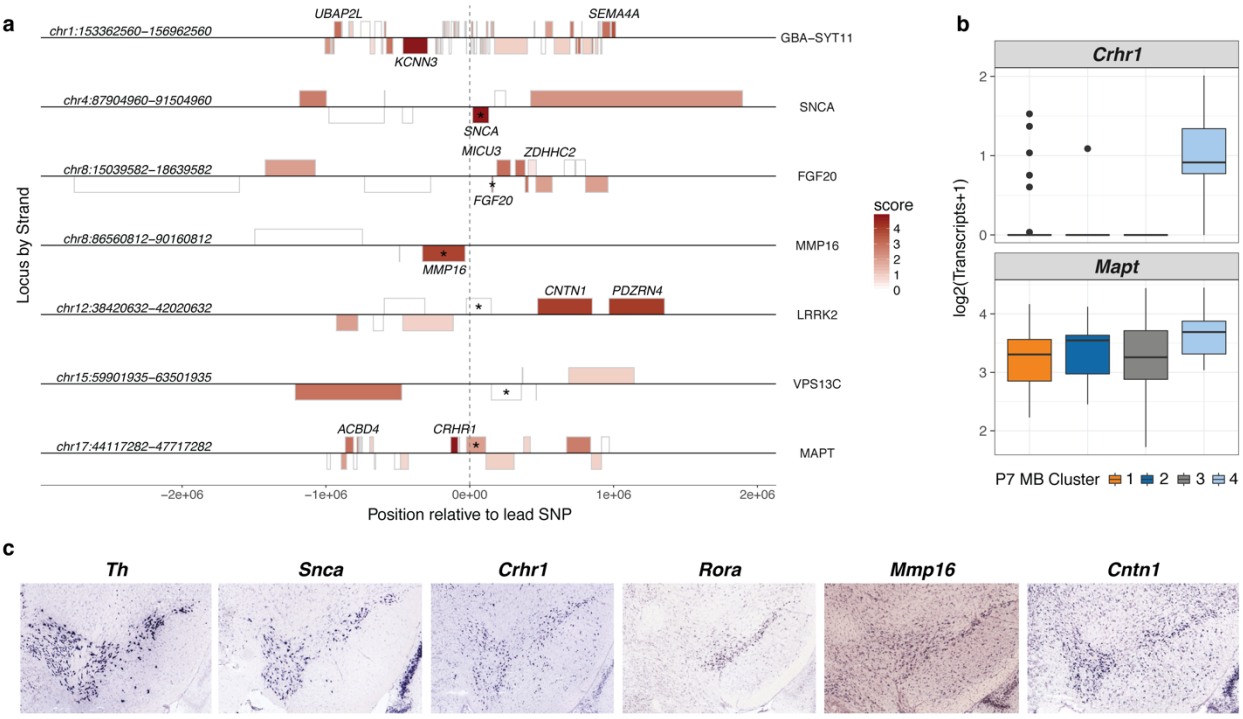

Figure 8. Potential candidate genes in PD GWAS loci can be identified and prioritized through the use of scRNA-seq data. a) A plot displaying the four megabase regions in the human genome (hg38) centered on PD GWAS SNPs in seven loci. Genes within +/- 1 Mb from the lead SNP in those loci are displayed as boxes on their appropriate strand. Genes are shaded by their prioritization score which is based on five factors: whether the gene is expressed in the P7 *substantia nigra* population (P7.MB.4); whether the gene was found to be a specific marker gene for P7.MB.4; whether the gene is differentially expressed between P7 MB clusters; whether the gene is found in any of the PD-associated WGCNA modules ("green" and "brown"); the pLI score of the gene in the ExAC database. The gene or genes with the highest scores in a given locus are considered the best candidate genes. The genes that are closest to the lead SNP are marked by an asterisk. This analysis allowed for the prioritization of all 32 PD GWAS loci in Nalls, et al[8]. b) Boxplots showing the expression of *Crhr1* and *Mapt* in identified P7 MB clusters. *Crhr1* shows specific expression in P7 SN *Th*$^+$ neurons while *Mapt* is expressed at equal levels in all P7 MB clusters. This specific expression of *Crhr1* but not *Mapt* along with *CRHR1* being identified as the highest scoring gene in the MAPT locus may help to explain why variation in this locus leads to preferential degeneration of SN *Th*$^+$ neurons in PD. c) *In situ* hybridization data from the Allen Brain Atlas (ABA) for some of the highest scoring genes in Figure 8a. The expression data for *Th* is used here for anatomical identification of the SN and for comparison purposes. As shown, the mouse homologs of high scoring genes in the *SNCA* (*Snca*), *MAPT* (*Crhr1*), VPS13C (*Rora*), MMP16 (*Mmp16*), and LRRK2 (*Cntn1*) loci show clear expression in SN DA neurons in P56 mice.

To further prioritize the remaining 28 loci, we scored on whether genes were differentially expressed between P7 MB *Th*$^+$ populations; whether they were identified as a marker gene for

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

P7.MB.4 (SN) cluster; and whether the genes were co-expressed with PD enriched gene modules uncovered in WGCNA. This strategy facilitated further prioritization of one or two genes in 16 additional loci (Table S14; Table S15). Importantly, we prioritize the major PD gene, *SNCA* in the *SNCA* locus (Figure 8a; Figure S8; Table S15). In three of these loci (*GBA-SYT11*, *LRRK2* and *MAPT*), our scoring prioritizes a different gene (*Kcnn3*, *Pdzrn4*, and *Crhr1*, respectively) than one previously implicated in PD phenotypes (Figure 8a, Table S15). This apparent conflict is well exemplified by our observations at the *MAPT* locus. Although *MAPT* is broadly implicated in neurodegeneration, we detect *Mapt* expression consistently across all assayed MB DA neurons (Figure 8b). By contrast *Crhr1,* encoding the corticotropin releasing hormone receptor 1, is specifically expressed in SN neurons (Figure 8b).

We then sought to further prioritize the SN-expressed genes in the remaining ten loci by integrating the probability of being loss-of-function (LoF) intolerant (pLI) metric from the ExAC database[72], due to a recent study using this metric to identify dosage sensitive genes[73]. Since most GWAS variation is predicted to impact regulatory DNA and in turn impact gene expression, it follows that genes in GWAS loci that are more sensitive to dosage levels may be more likely to be candidate genes. With that in mind, the pLI for each gene was used to further "rank" the genes within loci that were not previously prioritized. For those loci, we report a group of top scoring candidate genes (≤ 5) (Table S15). By integrating this step, we prioritized candidates for the remaining 10/32 (31%) loci. In total, we prioritize candidates in all 32 PD GWAS loci, establishing a systematic rationale for the identification of biologically pertinent candidates and testable hypotheses.

DISCUSSION

Midbrain DA neurons in the *substantia nigra* have been the subject of intense research since being definitively linked to PD nearly 100 years ago[74]. While degeneration of SN DA neurons in PD is well established, they represent only a subset of CNS DA populations. It remains unknown why nigral DA neurons are particularly vulnerable. We set out to explore this question using scRNA-seq to characterize the transcriptomes of DA neuron populations from distinct regions of the mouse brain over developmental time. Recently others have used scRNA-seq to characterize the mouse MB including DA neurons[29]. We undertake a highly complementary strategy, making several distinct and significant findings.

Previously unknown and unappreciated aspects of SN biology are revealed through scRNA-seq

By analyzing a broad array of Th^{+} neuronal populations (MB, FB, and OB), we reveal what underlies their functional diversity. Perhaps most intriguing, we demonstrate that SN DA neurons display no evidence of neurotransmitter or hormone co-transmission/release with dopamine, unique amongst the Th^{+} populations studied. Although, this observation is consistent with reports suggesting co-transmission in hypothalamus and olfactory bulb[41,75] as well VTA[13]

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DA neurons, we see no evidence supporting co-transmission in SN DA neurons. This result raises the question of whether this sole neurotransmitter phenotype of SN DA neurons may contribute to their selective vulnerability in PD.

We further reveal SN marker genes and GRN components that more fully characterize the unique biology of these neurons. Several genes, including SN marker gene *Prr16*, play roles in oxidative phosphorylation and mitochondrial function, consistent with established SN neuronal biology[76,77]. We also identify new SN marker genes that encode secreted proteins and cell-surface proteins that further define how SN DA neurons may interact with their environment. For example, we identify *Fam19a4* as being specifically expressed in the SN at P7. *FAM19A4* encodes a secreted protein that has been shown to be expressed in the brain and act as a chemo-attractant and activator of macrophages through the binding of *FPR1*[78,79]. *FAM19A4* expression has also been found to be upregulated in immune cells upon lipopolysaccharide (LPS) treatment, a model of neuroinflammation[79]. This finding potentially links SN DA specific gene expression and protein secretion to the role of inflammation in PD[80] and the specific vulnerability of SN DA neurons to degeneration caused by inflammation[81].

A novel postnatal MB *Th*$^+$ cell type is a putative progenitor-like MB DA neuron

Our analysis of embryonic and postnatal MB *Th*$^+$ neurons revealed a population of neurons, present at both embryonic and postnatal timepoints (E15.MB.2 and P7.MB.2), that share expressed genes indicative of MB DA neuron progenitors. While progenitor cell populations in the ventral MB have been previously characterized at embryonic timepoints[29], the existence of a postnatal MB progenitor neuron population has not been noted in previous scRNA-seq studies[29,60]. Notably, previous studies characterized postnatal neurons marked by transgenes under *Slc6a3* regulatory control. Given that we demonstrate this marker to be absent from P7.MB.2 cluster, it follows that this population would likely have been overlooked. By contrast, our use of *Th* left this population available for discovery. We show that specific markers for this population place it in the dorsal portion of the SN or the IPN at P7. The existence of markers of this population in different ventral MB sites potentially indicates that this cluster of cells represents a specific cell state reflecting neuronal immaturity instead of a reflection of spatial arrangement.

One may speculate regarding the function of a postnatal MB progenitor population. While beyond the scope of this paper, some clues may be found in the literature about *Th*$^+$ neuron development. Studies of SN DA neuron development in mice have shown that there are two periods of programmed cell death with peak apoptosis occurring at P2 and P14 (Figure 1a)[82]. Paradoxically, even though there are high levels of cell death at these points, the actual number of *Th*$^+$ neurons in the mouse SN does not decrease[82,83]. It has been shown that this can be explained by increasing levels of *Th* in cells over time, leading to "new" neurons appearing that are increasingly able to be immunostained[82]. These results have led to the suggestion that there is

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a "phenotypic maturation" of MB DA neurons during the early postnatal time period[82]. This very well may explain the presence of our "progenitor-like" MB DA neurons at P7, which display much lower levels of *Th* than other populations.

Prioritization of genes within PD GWAS loci identifies genes that may contribute to common PD susceptibility

The majority of variants identified in GWAS are located in non-coding DNA[84]. They are enriched for characteristics denoting regulatory DNA[84,85], and have been shown to modulate tissue-dependent elements[84–87]. Despite this evidence, in practice, the gene closest to the lead SNP identified within a GWAS locus is frequently treated as the prime candidate gene, often without considering tissue-dependent context. In an effort to more systematically identify and prioritize gene candidates from GWAS, our study integrates layers of orthogonal genetic and genomic data. We posit that genes pertinent to PD are likely expressed within MB DA neurons, specifically within the SN. We conservatively define an interval of interest (TAD boundary/2 Mb) around each lead SNP and ask which genes therein are expressed in the SN. We systematically move from intervals that reveal one primary candidate, by harboring only one SN-expressed gene, to those with many candidates, requiring a cumulative body of biological evidence to prioritize genes for functional inquiry.

Supporting our strategy, we prioritize one gene in each of three PD loci (*Snca*, *Fgf20*, *Gch1*), that have been directly associated with PD, MB DA development, and MB DA function. *SNCA* is mutated in autosomal dominant versions of PD; it is a pro-aggregation component of Lewy Bodies, a pathognomonic hallmark of PD neuronal degeneration (OMIM: 163890). *Fgf20* is expressed preferentially in the SN, contributes to DA neuron differentiation in cell culture, and protects against DA neuron degeneration[88]. Additionally, SNPs within *Fgf20* have been reported to modulate PD risk[88]. Finally, *Gch1* encodes the rate-limiting enzyme in tetrahydrobiopterin synthesis (GTP cyclohydrolase I). Tetrahydropterin is an important cofactor for many enzymes including *Th*, the rate limiting enzyme of dopamine synthesis. Consistent with these data, mutations in *GCH1* cause dopa-responsive dystonia that often presents with parkinsonian symptoms (OMIM: 128230).

While our method successfully prioritized one familial PD gene (*Snca*), we do not prioritize *Lrrk2,* another familial PD gene harbored within a PD GWAS locus. *Lrrk2* is not prioritized simply because it is not expressed in our SN DA neuronal population. This is expected as numerous studies have reported little to no *Lrrk2* expression in Th^{+} MB DA neurons both in mice and humans[89,90]. Instead, our method prioritizes *PDZRN4* within the *LRRK2* locus, based upon differential expression and the finding that it is co-expressed with identified PD gene modules. Whether *PDZRN4* should now be considered a novel alternative PD candidate independent of or in addition to *LRRK2* requires functional evaluation.

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

This strategy also reveals genes that may be biologically relevant but are overlooked due to the presence of prior candidates. *MAPT*, for example, is known to play a significant role in the broad neurodegenerative pathology of Alzheimer's disease, and has additionally been associated with susceptibility to PD (OMIM: 168600). Our data confirms that *Mapt* is expressed at consistent levels throughout MB DA neurons, including the SN. However, we prioritize *Crhr1* because it is specifically expressed in the SN compared to the other MB DA populations. Although prior data demonstrated *Crhr1* to be expressed in MB DA neurons[91], it is noteworthy that the MB DA neuroprotective activity of the urocortin (Ucn) neuropeptide in PD animal models is mediated through its interaction with Crhr1[92–95]. Recently, Ucn-Crhr1 binding was shown to improve DA neuron differentiation *in vitro,* data supported by reports linking it to a role in MB DA neuron development[96]. We do not believe that these results contradict the clear connection between genes in these loci and PD risk. Rather, we propose that other genes in these loci may also contribute to PD susceptibility, possibly in combination with other genes in the locus. These data set the stage for a new generation of independent and combinatorial functional evaluation.

By extending our ranking of candidate genes from exclusive or preferential expression in the SN to include, co-regulation with WGCNA identified modules implicated in PD and ultimately the inference of dosage sensitivity through the pLI (ExAC) metric, we establish a rank order of candidate genes within every one of 32 major GWAS-implicated PD loci.

Despite this success, we should acknowledge several notable caveats. First, not all genes in PD-associated human loci have identified mouse homologs. Thus, it remains possible that we overlooked genes whose biology is not comprehensively queried in this study. Secondly, we assume that identified genetic variation acts in a manner that is at least preferential, if not exclusive, to SN DA neurons. Lastly, by prioritizing expressed genes, we assume that PD variation affects genes that are normally expressed in the SN. We readily acknowledge that regulatory variation may require stress/insult to reveal its relevance.

CONCLUSIONS

In summary, our study of DA neurons in the developing mouse brain using scRNA-seq allowed for further definition of Th^{+} neuron signatures at both embryonic and postnatal ages. These data facilitated definition of a SN DA neuron signature as well as revealed previously undescribed markers of this important neuronal population. This data also provides the first demonstration of a postnatal progenitor-like MB neuron and its characteristic molecular signature. Finally, we use the totality of our data to provide the first comprehensive candidate gene prioritization of GWAS loci for a major common disease trait. Collectively these data establish a platform from which the next generation exploration of PD genetics can more effectively proceed.

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

METHODS

Data availability

Raw data will be made available on Sequence Read Archive (SRA) and Gene Expression Omnibus (GEO) prior to publication. Summary data is available where code is available below (https://github.com/pwh124/DA_scRNA-seq).

Code Availability

Code for analysis, for the production of figures, and summary data is deposited at https://github.com/pwh124/DA_scRNA-seq

Propagation of Th:GFP BAC transgenic mice.

The Th:EGFP BAC transgenic mice (Tg(Th-EGFP)DJ76Gsat/Mmnc) used in this study were generated by the GENSAT Project and were purchased through the Mutant Mouse Resource & Research Centers (MMRRC) Repository (https://www.mmrrc.org/). Mice were maintained on a Swiss Webster (SW) background with female SW mice obtained from Charles River Laboratories (http://www.criver.com/). All work involving mice (husbandry, colony maintenance and euthanasia) were reviewed and pre-approved by the institutional care and use committee.

The Tg(Th-EGFP)DJ76Gsat/Mmnc line was primarily maintained through matings between Th:EGFP positive, hemizygous male mice and wild-type SW females (dams). Timed matings for cell isolation were similarly established between hemizygous male mice and wild-type SW females. The observation of a vaginal plug was defined as embryonic day 0.5 (E0.5).

Dissection of E15.5 brains.

At 15.5 days after the timed mating, pregnant dams were euthanized and the entire litter of embryonic day 15.5 (E15.5) embryos were dissected out of the mother and immediately placed in chilled Eagle's Minimum Essential Media (EMEM). Individual embryos were then decapitated and heads were placed in fresh EMEM on ice. Embryonic brains were then removed and placed in Hank's Balanced Salt Solution (HBSS) without Mg^{2+} and Ca^{2+} and manipulated while on ice. The brains were immediately observed under a fluorescent stereomicroscope and $EGFP^{+}$ brains were selected. $EGFP^{+}$ regions of interest in the forebrain (hypothalamus) and the midbrain were then dissected and placed in HBSS on ice. This process was repeated for each $EGFP^{+}$ brain. Four $EGFP^{+}$ brain regions for each region studied were pooled together for dissociation.

Dissection of P7 brains.

After matings, pregnant females were sorted into their own cages and checked daily for newly born pups. The morning the pups were born was considered day P0. Once the mice were aged to

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

P7, all the mice from the litter were euthanized and the brains were then quickly dissected out of the mice and placed in HBSS without Mg^{2+} and Ca^{2+} on ice. As before, the brains were then observed under a fluorescent microscope, $EGFP^+$ status for P7 mice was determined, and $EGFP^+$ brains were retained. For each $EGFP^+$ brain, the entire olfactory bulb was first resected and placed in HBSS on ice. Immediately thereafter, the $EGFP^+$ forebrain and midbrain regions for each brain were resected and also placed in distinct containers of HBSS on ice. Five $EGFP^+$ brain regions for each region were pooled together for dissociation.

Generation of single cell suspensions from brain tissue.

Resected brain tissues were dissociated using papain (Papain Dissociation System, Worthington Biochemical Corporation; Cat#: LK003150) following the trehalose-enhanced protocol reported by Saxena, et. al, 2012[97] with the following modifications: The dissociation was carried out at 37°C in a sterile tissue culture cabinet. During dissociation, all tissues at all time points were triturated every 10 minutes using a sterile Pasteur pipette. For E15.5 tissues, this was continued for no more than 40 minutes. For P7, this was continued for up to 1.5 hours or until the tissue appeared to be completely dissociated.

Additionally, for P7 tissues, after dissociation but before cell sorting, the cell pellets were passed through a discontinuous density gradient in order to remove cell debris that could impede cell sorting. This gradient was adapted from the Worthington Papain Dissociation System kit. Briefly, after completion of dissociation according to the Saxena protocol[97], the final cell pellet was resuspended in DNase dilute albumin-inhibitor solution, layered on top of 5 mL of albumin-inhibitor solution, and centrifuged at 70g for 6 minutes. The supernatant was then removed.

FACS and single-cell collection.

For each timepoint-region condition, pellets were resuspended in 200 µL of media without serum comprised of DMEM/F12 without phenol red, 5% trehalose (w/v), 25 µM AP-V, 100 µM kynurenic acid, and 10 µL of 40 U/µl RNase inhibitor (RNasin® Plus RNase Inhibitor, Promega) at room temperature. The resuspended cells were then passed through a 40 uM filter and introduced into a Fluorescence Assisted Cell Sorting (FACS) machine (Beckman Coulter MoFlo Cell Sorter or Becton Dickinson FACSJazz). Viable cells were identified via propidium iodide staining, and individual neurons were sorted based on their fluorescence (EGFP+ intensity, See Figure 2d) directly into lysis buffer in individual wells of 96-well plates for single-cell sequencing (2 µL Smart-Seq2 lysis buffer + RNAase inhibitor, 1 µL oligo-dT primer, and 1 µL dNTPs according to Picelli et al., 2014[98]. Ninety-five cells of each type were collected along with a control blank well. Upon completion of a sort, the plates were briefly spun in a tabletop microcentrifuge and snap-frozen on dry ice. Single cell lysates were subsequently kept at -80°C until cDNA conversion.

Single-cell RT, library prep, and sequencing.

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Library preparation and amplification of single-cell samples were performed using a modified version of the Smart-Seq2 protocol[98]. Briefly, 96-well plates of single cell lysates were thawed to 4°C, heated to 72°C for 3 minutes, then immediately placed on ice. Template switching first-strand cDNA synthesis was performed as described above using a 5'-biotinylated TSO oligo. cDNAs were amplified using 20 cycles of KAPA HiFi PCR and 5'-biotinylated ISPCR primer. Amplified cDNA was cleaned with a 1:1 ratio of Ampure XP beads and approximately 200 pg was used for a one-quarter standard sized Nextera XT tagmentation reaction. Tagmented fragments were amplified for 14 cycles and dual indexes were added to each well to uniquely label each library. Concentrations were assessed with Quant-iT PicoGreen dsDNA Reagent (Invitrogen) and samples were diluted to ~2 nM and pooled. Pooled libraries were sequenced on the Illumina HiSeq 2500 platform to a target mean depth of ~8.0 x 10^5 50bp paired-end fragments per cell at the Hopkins Genetics Research Core Facility.

RNA sequencing and alignment.

For all libraries, paired-end reads were aligned to the mouse reference genome (mm10) supplemented with the Th-EGFP$^+$ transgene contig, using Hisat2[99] with default parameters except: -p 8. Aligned reads from individual samples were quantified against a reference transcriptome[100] (GENCODE vM8) supplemented with the addition of the eGFP transcript. Quantification was performed using cuffquant with default parameters and the following additional arguments: --no-update-check –p 8. Normalized expression estimates across all samples were obtained using cuffnorm[101] with default parameters.

Single-cell RNA data analysis.

Expression estimates.

Gene-level and isoform-level FPKM (Fragments Per Kilobase of transcript per Million) values produced by cuffquant[101] and the normalized FPKM matrix from cuffnorm was used as input for the Monocle2 single cell RNA-seq framework[102] in R/Bioconductor[103]. Genes were annotated using the Gencode vM8 release[100]. A CellDataSet was then created using Monocle (v2.2.0)[102] containing the gene FPKM table, gene annotations, and all available metadata for the sorted cells. All cells labeled as negative controls and empty wells were removed from the data. Relative FPKM values for each cell were converted to estimates of absolute mRNA counts per cell (RPC) using the Monocle2 Census algorithm[10] using the Monocle function "relative2abs." After RPCs were inferred, a new cds was created using the estimated RNA copy numbers with the expression Family set to "negbinomial.size()" and a lower detection limit of 0.1 RPC.

QC Filtering.

After expression estimates were inferred, the cds containing a total of 473 cells was run through Monocle's "detectGenes" function with the minimum expression level set at 0.1 transcripts. The following filtering criteria were then imposed on the entire data set:

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

i. Number of expressed genes - The number of expressed genes detected in each cell in the dataset was plotted and the high and low expressed gene thresholds were set based on observations of each distribution. Only those cells that expressed between 2,000 and 10,000 genes were retained.

ii. Cell Mass - Cells were then filtered based on the total mass of RNA in the cells calculated by Monocle. Again, the total mass of the cell was plotted and mass thresholds were set based on observations from each distribution. Only those cells with a total cell mass between 100,000 and 1,300,000 fragments mapped were retained.

iii. Total RNA copies per cell - Cells were then filtered based on the total number of RNA transcripts estimated for each cell. Again, the total RNA copies per cell was plotted and RNA transcript thresholds were set based on observations from each distribution. Only those cells with a total mRNA count between 1,000 and 40,000 RPCs were retained.

A total of 410 individual cells passed these initial filters. Outliers found in subsequent, reiterative analyses described below were analyzed and removed resulting a final cell number of 396. The distributions for total mRNAs, total mass, and number of expressed, can be found in Figure S1.

Log distribution QC.

Analysis using Monocle relies on the assumption that the expression data being analyzed follows a log-normal distribution. Comparison to this distribution was performed after initial filtering prior to continuing with analysis and was observed to be well fit.

Reiterative single-cell RNA data analysis.

After initial filtering described above, the cds was then broken into subsets based on "age" and "region" of cells for recursive analysis. Regardless of how the data was subdivided, all data followed a similar downstream analysis workflow.

Determining number of cells expressing each gene.

The genes to be analyzed for each subset iteration were filtered based on the number of cells that expressed each gene. Genes were retained if they were expressed in > 5% of the cells in the dataset being analyzed. These are termed "expressed_genes." For example, when analyzing all cells collected together (n = 410), a gene had to be expressed in 20.5 cells (410 x 0.05 = 20.5) to be included in the analysis. Whereas when analyzing P7 MB cells (n = 80), a gene had to be expressed in just 4 cells (80 x 0.05 = 4). This was done to allow include genes that may define rare populations of cells that could be present in any given population.

Monocle model preparation.

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The data was prepared for Monocle analysis by retaining only the expressed genes that passed the filtering described above. Size factors were estimated using Monocle's "estimateSizeFactors()" function. Dispersions were estimated using the "estimateDispersions()" function.

High variance gene selection.

Genes that have a high biological coefficient of variation (BCV) were identified by first calculating the BCV by dividing the standard deviation of expression for each expressed gene by the mean expression of each expressed gene. A dispersion table was then extracted using the dispersionTable() function from Monocle. Genes with a mean expression > 0.5 transcripts and a "dispersion_empirical" >= 1.5*dispersion_fit or 2.0*dispersion_fit were identified as "high variance genes."

Principal component analysis (PCA).

PCA was then run using the R prcomp function on the centered and scaled log2 expression values of the "high variance genes." PC1 and PC2 were then visualized to scan the data for obvious outliers as well as bias in the PCs for age, region, or plates on which the cells were sequenced. If any visual outliers in the data was observed, those cells were removed from the original subsetted cds and all filtering steps above were repeated. Once there were no obvious visual outliers in PC1 or PC2, a screeplot was used plot the PCA results in order to determine the number of PCs that contributed most significantly to the variation in the data. This was manually determined by inspecting the screeplot and including only those PCs that occur before the leveling-off of the plot.

t-SNE and clustering.

Once the number of significant PCs was determined, t-Distributed Stochastic Neighbor Embedding (t-SNE)[12] was used to embed the significant PC dimensions in a 2-D space for visualization. This was done using the "tsne" package available through R with "whiten = FALSE." The parameters "perplexity" and "max_iter" were tested with various values and set according what seemed to give the cleanest clustering of the data.

After dimensionality reduction via t-SNE, the number of clusters was determined in an unbiased manner by fitting multiple Gaussian distributions over the 2D t-SNE projection coordinates using the R package "ADPclust"[104] and the t-SNE plots were visualized using a custom R script. The number of genes expressed and the total mRNAs in each cluster were then compared.

Differential expression Analyses.

Differential expression analysis was performed using the "differentialGeneTest" function from Monocle that uses a likelihood ratio test to compare a vector generalized additive model

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(VGAM) using a negative binomial family function to a reduced model in which one parameter of interest has been removed. In practice, the following models were fit:

"~kmeans_tSNE_cluster" for timepoint-region datasets
"~region" for timepoint datasets

Genes were called as significantly differentially expressed if they had a q-value (Benjamini-Hochberg corrected p-value) < 0.05.

Cluster/Region Specific marker genes.
In order to identify differentially expressed genes that were "specifically" expressed in a particular cluster or region, R code calculating the Jensen-Shannon based specificity score from the R package cummerbund[105] was used similar to what was described in Kelly et. al[106].

Briefly, the mean RPC within each cluster for each expressed gene as well as the percentage of cells within each cluster that express each gene at a level > 1 transcript were calculated. The ".specificity" function from the cummRbund package was then used to calculate and identify the cluster with maximum specificity of each gene's expression. Details of this specificity metric can be found in Cabili, et al[107].

To identify cluster/region specific genes, the distribution of specificity scores for each region/cluster was plotted and a specificity cutoff was chosen so that only the "long right tail" of each distribution was included (i.e. genes with a specificity score above the cutoff chosen). For each iterative analysis, the same cutoff was used for each cluster or region. Once the specificity cutoff was chosen, genes were further filtered by only retaining genes that were expressed in >= 40% of cells within the cluster the gene was determined to be specific for.

Transcription Factor Correlation.
For transcription factor (TF) correlation analysis, aggregated lists of mouse genes (whether specific or differentially expressed) were intersected with the Animal Transcription Factor Database[108] (Data accessed: 04-04-2017; http://www.bioguo.org/AnimalTFDB/) in order to identify genes within those lists that were TFs. Pairwise correlation of log2(RPC + 1) for the TFs were calculated and plotted using the "corrplot" function from the "corrplot" R package with the following settings: order = "hclust", hclust.method = "ward.D2", cor.method = "pearson", method = "color". After plotting, the "corrplot" function option "addrect" was used to identify groups of TFs based on hierarchical clustering. "addrect" was set to a number that best fit the data. "Core transcription factors" within the larger groups were identified by using multiscale bootstrap resampling of hierarchical clustering using the R package 'pvclust' (v2.0-0)[109]. Again, log2(RPC + 1) for each group of TFs were used in these analyses. The analysis was carried out using the function 'pvclust()' with the following settings: nboot = 1000, method.dist =

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

"correlation"; method.hclust = "ward.D2"; and r = seq(0.5,1.4, by=.1). The distance metric and hclust method matched those used in the "corrplot" analysis described above. Significant clusters of TFs were identified using the 'pvpick()' function with the following settings: pv = 'au'; alpha = 0.90; max.only = F; and type = 'geq'. The clusters of TFs were deemed significant if the approximate unbiased (AU) p-value was greater than or equal to 90% (alpha = 0.90). Since "max.only" was set to FALSE, many smaller significant clusters were encompassed by larger clusters that were also significant. In those cases, the larger cluster was kept. Also, in the case where groups were identified by through the 'addrect' option of the 'corrplot' analysis above, if any significant cluster identified through bootstrap analysis was identical to the larger groups, it was not considered a "core" TF group.

Gene Set Enrichment Analyses.

Gene set enrichment analyses were performed in two separate ways depending upon the situation. A Gene Set Enrichment Analysis (GSEA) PreRanked analysis was performed when a ranked list (e.g. genes ranked by PC1 loadings) using GSEA software available from the Broad Institute (v2.2.4)[110,111]. Ranked gene lists were uploaded to the GSEA software and a "GSEAPreRanked" analysis was performed with the following settings: 'Number of Permutations' = 1000, 'Collapse dataset to gene symbols' = true, 'Chip platform(s)' = GENE_SYMBOL.chip, and 'Enrichment statistic' = weighted. Analysis was performed against Gene Ontology (GO) collections from MSigDB, including c2.all.v5.2.symbols and c5.all.v5.2.symbols. Top ten gene sets were reported for each analysis (Table S1). Figures and tables displaying the results were produced using custom R scripts.

Unranked GSEA analyses for lists of genes was performed using hypergeometric tests from the R package 'clusterProfiler' implemented through the functions 'enrichGO', 'enrichKEGG', and 'enrichPathway' with 'pvalueCutoff' set at 0.01, 0.1, 0.1, respectively with default settings[62]. These functions were implemented through the 'compareCluster' function when analyzing WGCNA data.

Weighted Gene Co-Expression Network Analysis (WGCNA).

WGCNA was performed us in R using the WGCNA package (v1.51)[112,113] following established pipelines laid out by the packages authors (see https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/ for more detail). Briefly, an expression matrix for all P7 neurons containing all genes expressed in >= 20 cells (n = 12628) was used with expression counts in log2(Transcripts + 1). The data were initially clustered in order to identify and remove outliers (n = 1) to leave 223 total cells (Figure S6). The soft threshold (power) for WGCNA was then determined by calculating the scale free topology model fit for a range of powers (1:10, 12, 14, 16, 18, 20) using the WGCNA function "pickSoftThreshold()" setting the networkType = "signed". A power of 10 was then chosen based on the leveling-off of the resulting scale independence plot above 0.8 (Figure S6).

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Network adjacency was then calculated using the WGCNA function "adjacency()" with the following settings: power = 10 and type = "signed." Adjacency calculations were used to then calculate topological overlap using the WGCNA function "TOMsimilarity()" with the following settings: TOMtype = "signed." Distance was then calculated by subtracting the topological overlap from 1. Hierarchical clustering was then performed on the distance matrix and modules were identified using the "cuttreeDynamic" function from the dynamicTreeCut package[114] with the following settings: deepSplit = T; pamRespectsDendro = FALSE, and minClusterSize = 20. This analysis initially identified 18 modules. Eigengenes for each module were then calculated using the "moduleEigengenes()" function and each module was assigned a color. Two modules ("grey" and "turquoise") were removed at this point. Turquoise was removed because it contained 11567 genes or all the genes that could not be grouped with another module. Grey was removed because it only contained 4 genes, falling below the minimum set module size of 20. The remaining 16 modules were clustered (Figure S6) and the correlation between module eigengenes and subset cluster identity was calculated using custom R scripts. Significance of correlation was determined by calculated the Student asymptotic p-value for correlations by using the WGCNA "corPvalueStudent()" function. Gene set enrichments for modules were determined by using the ClusterProfiler R package[62].The correlation between the t-SNE position of a cell and the module eigengenes was calculated using custom R scripts.

Prioritizing Genes in PD GWAS Loci.

Topologically Associated Domain (TAD) and Megabase Gene Data.

The data for human TAD boundaries were obtained from human embryonic stem cell (hESC) Hi-C data[115] and converted from human genome hg18 to hg38 using the liftOver tool from UCSC Genome Browser. PD GWAS SNP locations were then intersected with the TAD information to identify TADs containing a PD GWAS SNP. The data for +/- 1 megabase regions surrounding PD GWAS SNPs was obtained by taking PD GWAS SNP locations in hg38 and adding and subtracting 1e+06 from each location. All hg38 UCSC RefSeq genes that fall within the TADs or megabase regions were then identified by using the UCSC Table Browser. All genes were then annotated with PD locus and SNP information. Mouse homologs for all genes were identified using the NCBI Homologene database (Date accessed: 03/06/2017) and manual annotation. The TAD and megabase tables were then combined to create a final PD GWAS locus-gene table.

PD GWAS Loci Gene Scoring.

Genes within PD GWAS loci were initially scored using four gene lists: Genes with an average expression ≥0.5 transcripts in the SN cluster in our data (points = 2); Genes that were differentially expressed between P7 MB clusters (points = 1); Genes found to be "specifically" expressed in the P7 MB SN cluster (points = 1); Genes found in the WGCNA modules that are enriched for PD (points = 1). Expression in the SN cluster was considered the most important feature and was weighted as such. Furthermore, a piece of external data, pLI scores for each gene

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from the ExAC database[72], were added to the scores in order to rank all loci. pLI scores (fordist_cleaned_exac_r03_march16_z_pli_rec_null_data.txt) were obtained from http://exac.broadinstitute.org/ (Date dowloaded: March 30, 2017).

In situ hybridization.

In situ hybridization data was downloaded from publically available data from the Allen Institute through the Allen Brain Atlas (http://www.brain-map.org/). The image used in Figure 5 was obtained from the Reference Atlas at the Allen Brain Atlas (http://mouse.brain-map.org/static/atlas). URLs for all Allen Brain Atlas in situ data analyzed and downloaded for *substantia nigra* marker genes (Figure 5c) are available in Table S6. Data for *substantia nigra* expression *in situ* data for PD GWAS genes (Figure 8c) were obtained from the following experiments: 1056 (*Th*), 79908848 (*Snca*), 297 (*Crhr1*), 77371865 (*Rora*), 72129224 (*Mmp16*), and 414 (*Cntn1*). Data accessed on 03/02/17.

Single molecule in situ hybridization (smFISH).

For *in situ* hybridization experiments, untimed pregnant Swiss Webster mice were ordered from Charles River Laboratories (Crl:CFW(SW); http://www.criver.com/). Mice were maintained as previously described. Pups were considered P0 on the day of birth. At P7, the pups were decapitated, the brain was quickly removed, and the brain was then washed in 1x PBS. The intact brain was then transferred to a vial containing freshly prepared 4% PFA in 1x PBS and incubated at 4^{o}C for 24 hours. After 24 hours, brains were removed from PFA and washed three times in 1x PBS. The brains were then placed in a vial with 10% sucrose at 4^{o}C until the brains sunk to the bottom of the vial (usually ~1 hour). After sinking, brains were immediately placed in a vial containing 30% sucrose at 4^{o}C until once again sinking to the bottom of the vial (usually overnight). After cryoprotection, the brains were quickly frozen in optimal cutting temperature (O.C.T.) compound (Tissue-Tek) on dry ice and stored at -80^{o}C until use. Brains were sectioned at a thickness of 14 micrometers and mounted on Superfrost Plus microscope slides (Fisherbrand, Cat. # 12-550-15) with two sections per slide. Sections were then dried at room temperature for at least 30 minutes and then stored at -80^{o}C until use.

RNAscope *in situ* hybridization (https://acdbio.com/) was used to detect single RNA transcripts. RNAscope probes were used to detect *Th* (C1; Cat No. 317621), *Slc6a3* (C2; Cat No. 315441-C2), *Lhx9* (C3; Cat No. 495431-C3), *Ldb2* (C3; Cat No. 466061-C3), and *Meis2* (C3; Cat No. 436371-C3). The RNAscope Fluorescent Multiplex Detection kit (Cat No. 320851) and the associated protocol provided by the manufacturer were used. Briefly, frozen tissues were removed from -80^{o}C and equilibrated at room temperature for 5 minutes. Slides were then washed at room temperature in 1x PBS for 3 minutes with agitation. Slides were then immediately washed in 100% ethanol by moving the slides up and down 5-10 times. The slides were then allowed to dry at room temperature and hydrophobic barriers were drawn using a hydrophobic pen (ImmEdge Hydrophobic Barrier PAP Pen, Vector Laboratories, Cat. # H-4000)

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

around the tissue sections. The hydrophobic barrier was allowed to dry overnight. After drying, the tissue sections were treated with RNAscope Protease IV at room temperature for 30 minutes and then slides were washed in 1x PBS. Approximately 100 uL of multiplex probe mixtures (C1 - *Th*, C2 - *Slc6a3*, and C3 - one of *Lhx9*, *Ldb2*, or *Meis2*) containing either approximately 96 uL C1: 2 uL C2: 2 uL C3 (*Th*:*Slc6a3*:*Lhx9*) or 96 uL C1: 0.6 uL C2: 2 uL C3 (*Th*:*Slc6a3*:*Ldb2* or *Th*:*Slc6a3*:*Meis2*) were applied to appropriate sections. Both mixtures provided adequate *in situ* signals. Sections were then incubated at 40^{o}C for 2 hours in the ACD HybEZ oven. Sections were then sequentially treated with the RNAscope Multiplex Fluorescent Detection Reagents kit solutions AMP 1-FL, AMP 2-FL, AMP 3-FL, and AMP 4 Alt B-FL, with washing in between each incubation, according to manufacturer's recommendations. Sections were then treated with DAPI provided with the RNAscope Multiplex Fluorescent Detection Reagents kit. One drop of Prolong Gold Antifade Mountant (Invitrogen, Cat # P36930) was then applied to each section and a coverslip was then placed on the slide. The slides were then stored in the dark at 4^{o}C overnight before imaging. Slides were further stored at 4^{o}C throughout imaging. Manufacturer provided positive and negative controls were also performed along side experimental probe mixtures according to manufacturer's protocols. Four sections that encompassed relevant populations in the P7 ventral MB (SN, VTA, etc.) were chosen for each combination of RNAscope smFISH probes and subsequent analyses.

Confocal Microscopy.

RNAscope fluorescent *in situ* experiments were analyzed using the Nikon A1 confocal system equipped with a Nikon Eclipse Ti inverted microscope running Nikon NIS-Elements AR 4.10.01 64-bit software. Images were captured using a Nikon Plan Apo λ 60x/1.40 oil immersion lens with a common pinhole size of 19.2 μM, a pixel dwell of 28.8 μs, and a pixel resolution of 1024 x 1024. DAPI, FITC, Cy3, and Cy5 channels were used to acquire RNAscope fluorescence. Positive and negative control slides were used in order to calibrate laser power, offset, and detector sensitivity, for all channels in all experiments performed.

Image Analysis and processing.

Confocal images were saved as .nd2 files. Images were then processed in ImageJ as follows. First, the .nd2 files were imported into ImageJ and images were rotated in order to reflect a ventral MB orientation with the ventral side of the tissue at the bottom of the image. Next the LUT ranges were adjusted for the FITC (range: 0-2500), Cy3 (range: 0-2500), and Cy5 (range: 0-1500) channels. All analyzed images were set to the same LUT ranges. Next, the channels were split and merged back together to produce a "composite" image seen in Figure 6. Scale bars were then added. Cells of interest were then demarcated, duplicated, and the channels were split. These cells of interest were then displayed as the insets seen in Figure 6.

ACKNOWLEDGEMENTS

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The authors wish to thank Stephen M. Brown for implementation and optimization of smFISH. This research was supported in part by US National Institutes of Health grants R01 NS62972 and MH106522 to ASM.

AUTHOR CONTRIBUTIONS

PWH, ASM, and LAG designed the study and wrote the paper. PWH, SAM, WDL and GAC performed the experiments. PWH and LAG implemented the computational algorithms to process the raw data and conduct analyses thereof. PWH, LAG, and ASM analyzed and interpreted the resulting data. LAG contributed novel computational pipeline development. Correspondence to ASM (andy@jhmi.edu) and LAG (loyalgoff@jhmi.edu).

FINANCIAL INTERESTS STATEMENT

The authors declare no competing financial interests.

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

1. de Rijk, M. C. *et al.* Prevalence of parkinsonism and Parkinson's disease in Europe: the EUROPARKINSON Collaborative Study. European Community Concerted Action on the Epidemiology of Parkinson's disease. *J Neurol Neurosurg Psychiatry* **62,** 10–15 (1997).
2. Pringsheim, T., Jette, N., Frolkis, A. & Steeves, T. D. The prevalence of Parkinson's disease: a systematic review and meta-analysis. *Mov Disord* **29,** 1583–1590 (2014).
3. Savitt, J. M., Dawson, V. L. & Dawson, T. M. Diagnosis and treatment of Parkinson disease: molecules to medicine. *J Clin Invest* **116,** 1744–1754 (2006).
4. Paulus, W. & Jellinger, K. The neuropathologic basis of different clinical subgroups of Parkinson's disease. *J Neuropathol Exp Neurol* **50,** 743–755 (1991).
5. Brichta, L. *et al.* Identification of neurodegenerative factors using translatome-regulatory network analysis. *Nat. Neurosci.* **18,** 1325–33 (2015).
6. Puschmann, A. Monogenic Parkinson's disease and parkinsonism: clinical phenotypes and frequencies of known mutations. *Park. Relat Disord* **19,** 407–415 (2013).
7. Klein, C. & Westenberger, A. Genetics of Parkinson's disease. *Cold Spring Harb Perspect Med* **2,** a008888 (2012).
8. Nalls, M. a *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* **56,** 1–7 (2014).
9. Barallobre, M. J. *et al.* DYRK1A promotes dopaminergic neuron survival in the developing brain and in a mouse model of Parkinson's disease. *Cell Death Dis.* **5,** e1289 (2014).
10. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat Methods* **14,** 309–315 (2017).
11. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell developmental trajectories. *bioRxiv* (2017). doi:10.1101/110668
12. Van Der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9,** 2579–2605 (2008).
13. Morales, M. & Margolis, E. B. Ventral tegmental area: cellular heterogeneity, connectivity and behaviour. *Nat Rev Neurosci* **18,** 73–85 (2017).
14. Everitt, B. J., Hökfelt, T., Wu, J. Y. & Goldstein, M. Coexistence of tyrosine hydroxylase-like and gamma-aminobutyric acid-like immunoreactivities in neurons of the arcuate nucleus. *Neuroendocrinology* **39,** 189–191 (1984).
15. Asmus, S. E. *et al.* Increasing proportions of tyrosine hydroxylase-immunoreactive interneurons colocalize with choline acetyltransferase or vasoactive intestinal peptide in the developing rat cerebral cortex. *Brain Res* **1383,** 108–119 (2011).
16. Arenas, E., Denham, M. & Villaescusa, J. C. How to make a midbrain dopaminergic neuron. *Development* **142,** 1918–36 (2015).
17. Hou, P. S. *et al.* LHX2 regulates the neural differentiation of human embryonic stem cells via transcriptional modulation of PAX6 and CER1. *Nucleic Acids Res.* **41,** 7753–7770 (2013).
18. Agoston, Z. *et al.* Meis2 is a Pax6 co-factor in neurogenesis and dopaminergic periglomerular fate specification in the adult olfactory bulb. *Development* **141,** 28–38 (2014).
19. Agoston, Z. & Schulte, D. Meis2 competes with the Groucho co-repressor Tle4 for binding to Otx2 and specifies tectal fate without induction of a secondary midbrain-

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

hindbrain boundary organizer. *Development* **136,** 3311–3322 (2009).

20. Rétaux, S., Rogard, M., Bach, I., Failli, V. & Besson, M. J. Lhx9: a novel LIM-homeodomain gene expressed in the developing forebrain. *J. Neurosci.* **19,** 783–793 (1999).
21. Hoekstra, E. J. *et al.* Lmx1a Encodes a Rostral Set of Mesodiencephalic Dopaminergic Neurons Marked by the Wnt/B-Catenin Signaling Activator R-spondin 2. *PLoS One* **8,** 1–12 (2013).
22. Peukert, D., Weber, S., Lumsden, A. & Scholpp, S. Lhx2 and Lhx9 determine neuronal differentiation and compartition in the caudal forebrain by regulating Wnt signaling. *PLoS Biol.* **9,** (2011).
23. Yun, K., Mantani, A., Garel, S., Rubenstein, J. & Israel, M. a. Id4 regulates neural progenitor proliferation and differentiation in vivo. *Development* **131,** 5441–5448 (2004).
24. Bedford, L. *et al.* Id4 is required for the correct timing of neural differentiation. *Dev. Biol.* **280,** 386–395 (2005).
25. Yin, M. *et al.* Ventral mesencephalon-enriched genes that regulate the development of dopaminergic neurons in vivo. *J. Neurosci.* **29,** 5170–82 (2009).
26. Hegarty, S. V., Sullivan, A. M. & O'Keeffe, G. W. Midbrain dopaminergic neurons: A review of the molecular circuitry that regulates their development. *Developmental Biology* **379,** 123–138 (2013).
27. Makeyev, E. V., Zhang, J., Carrasco, M. A. & Maniatis, T. The MicroRNA miR-124 Promotes Neuronal Differentiation by Triggering Brain-Specific Alternative Pre-mRNA Splicing. *Mol. Cell* **27,** 435–448 (2007).
28. Mei, L. & Xiong, W. C. Neuregulin 1 in neural development, synaptic plasticity and schizophrenia. *Nat Rev Neurosci* **9,** 437–452 (2008).
29. La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* **167,** 566–580.e19 (2016).
30. Gestri, G. *et al.* Six3 functions in anterior neural plate specification by promoting cell proliferation and inhibiting Bmp4 expression. *Development* **132,** 2401–2413 (2005).
31. Lagutin, O. V *et al.* Six3 repression of Wnt signaling in the anterior neuroectoderm is essential for vertebrate forebrain development. *Genes Dev* **17,** 368–379 (2003).
32. Lavado, A., Lagutin, O. V & Oliver, G. Six3 inactivation causes progressive caudalization and aberrant patterning of the mammalian diencephalon. *Development* **135,** 441–450 (2008).
33. Geng, X., Lavado, A., Lagutin, O. V., Liu, W. & Oliver, G. Expression of Six3 Opposite Strand (Six3OS) during mouse embryonic development. *Gene Expr. Patterns* **7,** 252–257 (2007).
34. Kelsom, C. & Lu, W. Development and specification of GABAergic cortical interneurons. *Cell Biosci* **3,** 19 (2013).
35. Pacary, E. *et al.* Proneural transcription factors regulate different steps of cortical neuron migration through Rnd-mediated inhibition of RhoA signaling. *Neuron* **69,** 1069–1084 (2011).
36. Pacary, E., Azzarelli, R. & Guillemot, F. Rnd3 coordinates early steps of cortical neurogenesis through actin-dependent and -independent mechanisms. *Nat Commun* **4,** 1635 (2013).
37. Naka, H., Nakamura, S., Shimazaki, T. & Okano, H. Requirement for COUP-TFI and II in the temporal specification of neural stem cells in CNS development. *Nat Neurosci* **11,**

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1014–1023 (2008).

38. Lee, B., Lee, S., Lee, S.-K. & Lee, J. W. The LIM-homeobox transcription factor Isl1 plays critical roles in development of multiple arcuate nucleus neurons. *Development* (2016). doi:10.1242/dev.133967

39. Petryniak, M. A., Potter, G. B., Rowitch, D. H. & Rubenstein, J. L. Dlx1 and Dlx2 control neuronal versus oligodendroglial cell fate acquisition in the developing forebrain. *Neuron* **55,** 417–433 (2007).

40. Li, J. Y. *et al.* Arcuate nucleus transcriptome profiling identifies ankyrin repeat and suppressor of cytokine signalling box-containing protein 4 as a gene regulated by fasting in central nervous system feeding circuits. *J. Neuroendocrinol.* **17,** 394–404 (2005).

41. Björklund, A. & Dunnett, S. B. Dopamine neuron systems in the brain: an update. *Trends Neurosci* **30,** 194–202 (2007).

42. Li, H., Zeitler, P. S., Valerius, M. T., Small, K. & Potter, S. S. Gsh-1, an orphan Hox gene, is required for normal pituitary development. *EMBO J* **15,** 714–724 (1996).

43. McNay, D. E., Pelling, M., Claxton, S., Guillemot, F. & Ang, S. L. Mash1 is required for generic and subtype differentiation of hypothalamic neuroendocrine cells. *Mol Endocrinol* **20,** 1623–1632 (2006).

44. Campbell, J. N. *et al.* A molecular census of arcuate hypothalamus and median eminence cell types. *Nat Neurosci* **20,** 484–496 (2017).

45. Romanov, R. A. *et al.* Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat Neurosci* **20,** 176–188 (2017).

46. Kohwi, M., Osumi, N., Rubenstein, J. L. & Alvarez-Buylla, A. Pax6 is required for making specific subpopulations of granule and periglomerular neurons in the olfactory bulb. *J. Neurosci.* **25,** 6997–7003 (2005).

47. Hack, M. a *et al.* Neuronal fate determinants of adult olfactory bulb neurogenesis. *Nat. Neurosci.* **8,** 865–872 (2005).

48. Ninkovic, J. *et al.* The transcription factor Pax6 regulates survival of dopaminergic olfactory bulb neurons via crystallin ??A. *Neuron* **68,** 682–694 (2010).

49. Waclaw, R. R. *et al.* The zinc finger transcription factor Sp8 regulates the generation and diversity of olfactory bulb interneurons. *Neuron* **49,** 503–516 (2006).

50. Miller, T. E. *et al.* Lgr5 marks post-mitotic, lineage restricted cerebellar granule neurons during postnatal development. *PLoS One* **9,** (2014).

51. Chen, M. *et al.* Wnt-responsive $Lgr5^{+}$ globose basal cells function as multipotent olfactory epithelium progenitor cells. *J. Neurosci.* **34,** 8268–76 (2014).

52. Francis, F. *et al.* Doublecortin is a developmentally regulated, microtubule-associated protein expressed in migrating and differentiating neurons. *Neuron* **23,** 247–256 (1999).

53. Agoston, Z. *et al.* Meis2 is a Pax6 co-factor in neurogenesis and dopaminergic periglomerular fate specification in the adult olfactory bulb. *Development* **141,** 28–38 (2014).

54. Vergaño-Vera, E. *et al.* Nurr1 blocks the mitogenic effect of FGF-2 and EGF, inducing olfactory bulb neural stem cells to adopt dopaminergic and dopaminergic-GABAergic neuronal phenotypes. *Dev Neurobiol* **75,** 823–841 (2015).

55. Panman, L. *et al.* Sox6 and Otx2 control the specification of substantia nigra and ventral tegmental area dopamine neurons. *Cell Rep.* **8,** 1018–1025 (2014).

56. Viereckel, T. *et al.* Midbrain Gene Screening Identifies a New Mesoaccumbal Glutamatergic Pathway and a Marker for Dopamine Cells Neuroprotected in Parkinson's

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Disease. *Sci Rep* **6,** 35203 (2016).

57. Kozicz, T., Vigh, S. & Arimura, A. The source of origin of PACAP- and VIP-immunoreactive fibers in the laterodorsal division of the bed nucleus of the stria terminalis in the rat. *Brain Res.* **810,** 211–219 (1998).
58. Darland, T., Heinricher, M. M. & Grandy, D. K. Orphanin FQ/nociceptin: A role in pain and analgesia, but so much more. *Trends in Neurosciences* **21,** 215–221 (1998).
59. Cai, H., Liu, G., Sun, L. & Ding, J. Aldehyde Dehydrogenase 1 making molecular inroads into the differential vulnerability of nigrostriatal dopaminergic neuron subtypes in Parkinson's disease. *Transl. Neurodegener.* **3,** 27 (2014).
60. Poulin, J. F. *et al.* Defining midbrain dopaminergic neuron diversity by single-cell gene expression profiling. *Cell Rep* **9,** 930–943 (2014).
61. Waldegger, S., Barth, P., Raber, G. & Lang, F. Cloning and characterization of a putative human serine/threonine protein kinase transcriptionally modified during anisotonic and isotonic alterations of cell volume. *Proc Natl Acad Sci U S A* **94,** 4440–4445 (1997).
62. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16,** 284–7 (2012).
63. Krijger, P. H. *et al.* Cell-of-Origin-Specific 3D Genome Structure Acquired during Somatic Cell Reprogramming. *Cell Stem Cell* **18,** 597–610 (2016).
64. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12,** 1725–1735 (2003).
65. Benko, S. *et al.* Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nat. Genet.* **41,** 359–364 (2009).
66. Yong, V. W., Power, C., Forsyth, P. & Edwards, D. R. Metalloproteinases in biology and pathology of the nervous system. *Nat Rev Neurosci* **2,** 502–511 (2001).
67. Li, Z., Wu, Y. & Baraban, J. M. The Translin/Trax RNA binding complex: clues to function in the nervous system. *Biochim Biophys Acta* **1779,** 479–485 (2008).
68. Broer, L. *et al.* Association of heat shock proteins with Parkinson's disease. *Eur J Epidemiol* **26,** 933–935 (2011).
69. Zhu, Y., McAvoy, S., Kuhn, R. & Smith, D. I. RORA, a large common fragile site gene, is involved in cellular stress response. *Oncogene* **25,** 2901–2908 (2006).
70. Boukhtouche, F. *et al.* Human retinoic acid receptor-related orphan receptor alpha1 overexpression protects neurones against oxidative stress-induced apoptosis. *J Neurochem* **96,** 1778–1789 (2006).
71. Hamilton, B. A. *et al.* Disruption of the nuclear hormone receptor RORalpha in staggerer mice. *Nature* **379,** 736–739 (1996).
72. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
73. Doan, R. N. *et al.* Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. *Cell* **167,** 341–354.e12 (2016).
74. Parent, M. & Parent, A. Substantia nigra and Parkinson's disease: a brief history of their long and intimate relationship. *Can J Neurol Sci* **37,** 313–319 (2010).
75. Liu, S., Plachez, C., Shao, Z., Puche, A. & Shipley, M. T. Olfactory bulb short axon cell release of GABA and dopamine produces a temporally biphasic inhibition-excitation response in external tufted cells. *J Neurosci* **33,** 2916–2926 (2013).
76. Winklhofer, K. F. & Haass, C. Mitochondrial dysfunction in Parkinson's disease. *Biochim*

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Biophys Acta **1802,** 29–44 (2010).

77. Perier, C. & Vila, M. Mitochondrial biology and Parkinson's disease. *Cold Spring Harb Perspect Med* **2,** a009332 (2012).
78. Tom Tang, Y. *et al.* TAFA: a novel secreted family with conserved cysteine residues and restricted expression in the brain. *Genomics* **83,** 727–734 (2004).
79. Wang, W. *et al.* FAM19A4 is a novel cytokine ligand of formyl peptide receptor 1 (FPR1) and is able to promote the migration and phagocytosis of macrophages. *Cell Mol Immunol* **12,** 615–624 (2015).
80. Beal, M. F. Mitochondria, oxidative damage, and inflammation in Parkinson's disease. *Ann N Y Acad Sci* **991,** 120–131 (2003).
81. Kim, W. G. *et al.* Regional difference in susceptibility to lipopolysaccharide-induced neurotoxicity in the rat brain: role of microglia. *J Neurosci* **20,** 6309–6316 (2000).
82. Jackson-Lewis, V. *et al.* Developmental cell death in dopaminergic neurons of the substantia nigra of mice. *J Comp Neurol* **424,** 476–488 (2000).
83. Lieb, K. *et al.* Pre- and postnatal development of dopaminergic neuron numbers in the male and female mouse midbrain. *Brain Res Dev Brain Res* **94,** 37–43 (1996).
84. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science (80-.).* **337,** 1190–1195 (2012).
85. Farh, K. K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2015).
86. Emison, E. S. *et al.* A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature* **434,** 857–863 (2005).
87. Praetorius, C. *et al.* A polymorphism in IRF4 affects human pigmentation through a tyrosinase-dependent MITF/TFAP2A pathway. *Cell* **155,** 1022–1033 (2013).
88. Itoh, N. & Ohta, H. Roles of FGF20 in dopaminergic neurons and Parkinson's disease. *Front Mol Neurosci* **6,** 15 (2013).
89. Galter, D. *et al.* LRRK2 expression linked to dopamine-innervated areas. *Ann Neurol* **59,** 714–719 (2006).
90. Higashi, S. *et al.* Expression and localization of Parkinson's disease-associated leucine-rich repeat kinase 2 in the mouse brain. *J Neurochem* **100,** 368–381 (2007).
91. Sauvage, M. & Steckler, T. Detection of corticotropin-releasing hormone receptor 1 immunoreactivity in cholinergic, dopaminergic and noradrenergic neurons of the murine basal forebrain and brainstem nuclei - Potential implication for arousal and attention. *Neuroscience* **104,** 643–652 (2001).
92. Abuirmeileh, A., Harkavyi, A., Lever, R., Biggs, C. S. & Whitton, P. S. Urocortin, a CRF-like peptide, restores key indicators of damage in the substantia nigra in a neuroinflammatory model of Parkinson's disease. *J. Neuroinflammation* **4,** 19 (2007).
93. Abuirmeileh, A. *et al.* The corticotrophin-releasing factor-like peptide urocortin reverses key deficits in two rodent models of Parkinson's disease. *Eur J Neurosci* **26,** 417–423 (2007).
94. Abuirmeileh, A., Harkavyi, A., Kingsbury, A., Lever, R. & Whitton, P. S. The CRF-like peptide urocortin greatly attenuates loss of extracellular striatal dopamine in rat models of Parkinson's disease by activating CRF1 receptors. *Eur. J. Pharmacol.* **604,** 45–50 (2009).
95. Huang, H. Y. *et al.* Urocortin modulates dopaminergic neuronal survival via inhibition of glycogen synthase kinase-3β and histone deacetylase. *Neurobiol. Aging* **32,** 1662–1677 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/148049; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

96. Huang, H. Y. *et al.* Epigenetic regulation contributes to urocortin-enhanced midbrain dopaminergic neuron differentiation. *Stem Cells* **33,** 1601–1617 (2015).
97. Saxena, A. *et al.* Trehalose-enhanced isolation of neuronal sub-types from adult mouse brain. *Biotechniques* **52,** 381–385 (2012).
98. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9,** 171–181 (2014).
99. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12,** 357–60 (2015).
100. Mudge, J. M. & Harrow, J. Creating reference gene annotation for the mouse C57BL6/J genome assembly. *Mamm Genome* **26,** 366–378 (2015).
101. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7,** 562–78 (2012).
102. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32,** 381–6 (2014).
103. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12,** 115–121 (2015).
104. Wang, X. F. & Xu, Y. Fast clustering using adaptive density peak detection. *Stat Methods Med Res* (2015). doi:10.1177/0962280215609948
105. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31,** 46–53 (2013).
106. Burns, J. C., Kelly, M. C., Hoa, M., Morell, R. J. & Kelley, M. W. Single-cell RNA-Seq resolves cellular complexity in sensory organs from the neonatal inner ear. *Nat Commun* **6,** 8557 (2015).
107. Molyneaux, B. J. *et al.* DeCoN: genome-wide analysis of in vivo transcriptional dynamics during pyramidal neuron fate selection in neocortex. *Neuron* **85,** 275–288 (2015).
108. Zhang, H. M. *et al.* AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res* **40,** D144-9 (2012).
109. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22,** 1540–1542 (2006).
110. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102,** 15545–15550 (2005).
111. Mootha, V. K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34,** 267–273 (2003).
112. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).
113. Langfelder, P. & Horvath, S. Fast R Functions for Robust Correlations and Hierarchical Clustering. *J Stat Softw* **46,** (2012).
114. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24,** 719–720 (2008).
115. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).