Single-cell RNA-seq of dopaminergic neurons informs candidate gene selection for sporadic Parkinson's disease

Paul W. Hook[1], Sarah A. McClymont[1], Gabrielle H. Cannon[1], William D. Law[1], A. Jennifer Morton[2], Loyal A. Goff[1,3*], Andrew S. McCallion[1,4,5*]

[1]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America

[2]Department of Physiology Development and Neuroscience, University of Cambridge, Cambridge, United Kingdom

[3]Department of Neuroscience, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America

[4]Department of Comparative and Molecular Pathobiology, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America

[5]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

ABSTRACT

Parkinson's disease (PD) is caused by the collapse of *substantia nigra* (SN) dopaminergic (DA) neurons of the midbrain (MB), while other DA populations remain relatively intact. Common variation influencing susceptibility to sporadic PD has been primarily identified through genome wide association studies (GWAS). However, like many other common genetic diseases, the genes impacted by common PD-associated variation remain to be elucidated. Here, we used single-cell RNA-seq to characterize DA neuron populations in the mouse brain at embryonic and early postnatal timepoints. These data allow for the unbiased identification of DA neuron subpopulations, including a novel postnatal neuroblast population and SN DA neurons. Comparison of SN DA neurons with other DA neurons populations in the brain reveals a unique transcriptional profile, novel marker genes, and specific gene regulatory networks. By integrating these cell population specific data with published GWAS, we develop a scoring system for prioritizing candidate genes in PD-associated loci. With this, we prioritize candidate genes in all 32 GWAS intervals implicated in sporadic PD risk, the first such systematically generated list. From this we confirm that the prioritized candidate gene *CPLX1* disrupts the nigrostriatal pathway when knocked out in mice. Ultimately, this systematic rationale leads to the identification of biologically pertinent candidates and testable hypotheses for sporadic PD that will inform a new era of PD genetic research.

The most commonly used genetic tool today for studying complex disease is the genome wide association study (GWAS). As a strategy, GWAS was initially hailed for the insight it might provide into the genetic architecture of common human disease risk. Indeed, the collective data from GWAS since 2005 has revealed a trove of variants and genomic intervals associated with an array of phenotypes[1]. The majority of variants identified in GWAS are located in non-coding DNA[2] and are enriched for characteristics denoting regulatory DNA[2,3]. This regulatory variation is expected to impact expression of a nearby gene, leading to disease susceptibility.

Traditionally, the gene closest to the lead SNP has been prioritized as the affected gene. However, recent studies show that disease-associated variants can act on more distally located genes, invalidating genes that were previously extensively studied[4,5]. The inability to systematically connect common variation with the genes impacted limits our capacity to elucidate potential therapeutic targets and can waste valuable research efforts.

Although GWAS is inherently agnostic to the context in which disease-risk variation acts, the biological impact of common functional variation has been shown to be cell context dependent[2,6]. Extending these observations, Pritchard and colleagues recently demonstrated that although genes need only to be expressed in disease-relevant cell types to contribute to risk, those expressed preferentially or exclusively therein contribute more per SNP[7]. Thus, accounting for the cellular and gene regulatory network (GRN) contexts within which variation act may better inform the identification of impacted genes. These principles have not yet been applied systematically to many of the traits for which GWAS data exists. We have chosen Parkinson's

disease (PD) as a model complex disorder for which a significant body of GWAS data remains to be explored biologically in a context dependent manner.

PD is the most common progressive neurodegenerative movement disorder. Incidence of PD increases with age, affecting an estimated 1% worldwide beyond 70 years of age[8–10]. The genetic underpinnings of non-familial or sporadic PD have been studied through the use of GWAS with a recent meta-analysis highlighting 32 loci associated with sporadic PD susceptibility[11]. While a small fraction of PD GWAS loci contain genes known to be mutated in familial PD (*SNCA* and *LRRK2*)[12,13], most indicted intervals do not contain a known causal gene or genes. Although PD ultimately affects multiple neuronal centers, preferential degeneration of DA neurons in the SN leads to functional collapse of the nigrostriatal pathway and loss of fine motor control. The preferential degeneration of SN DA neurons in relation to other mesencephalic DA neurons has driven research interest in the genetic basis of selective SN vulnerability in PD. Consequently, one can reasonably assert that a significant fraction of PD-associated variation likely mediates its influence specifically within the SN.

In an effort to illuminate a biological context in which PD GWAS results could be better interpreted, we undertook single-cell RNA-seq (scRNA-seq) analyses of multiple DA neuronal populations in the brain, including ventral midbrain DA neurons. This analysis defined the heterogeneity of DA populations over developmental time in the brain, revealing gene expression profiles specific to discrete DA neuron subtypes. These data further facilitated the definition of GRNs active in DA neuron populations including the SN. With these data, we

80 establish a framework to systematically prioritize candidate genes in all 32 PD GWAS loci and
81 begin exploring their pathological significance.
82
83 **RESULTS**
84 *scRNA-seq characterization defines DA neuronal subpopulation heterogeneity*
85 In order to characterize DA neuron molecular phenotypes, we undertook scRNA-seq on cells
86 isolated from distinct anatomical locations of the mouse brain over developmental time. We used
87 fluorescence activated cell sorting (FACS) to retrieve single DA neurons from the Tg(Th-
88 EGFP)DJ76Gsat BAC transgenic mouse line, which expresses eGFP under the control of the
89 tyrosine hydroxylase (*Th*) locus[14]. We microdissected both MB and FB from E15.5 mice,
90 extending our analyses to MB, FB, and OB in P7 mice (Figure 1a). E15.5 and P7 time points
91 were chosen based on their representation of stable MB DA populations, either after neuron birth
92 (E15.5) or between periods of programmed cell death (P7) (Figure 1a)[15].
93
94 Quality control and outlier analysis identify 396 high quality cell transcriptomes to be used in
95 our analyses. We initially sequenced RNA from 473 single cells to an average depth of ~8 x 10^5
96 50 bp paired-end fragments per cell. Using Monocle 2, we converted normalized expression
97 estimates into estimates of RNA copies per cell[16]. Cells were filtered based on the distributions
98 of total mass, total number of mRNAs, and total number of expressed genes per cell
99 (Supplementary Figure 1a, 1b, 1c; detailed in Methods). After QC, 410 out of 473 cells were
100 retained. Using principal component analysis (PCA) as part of the iterative analysis described
101 below, we identified and removed 14 outliers determined to be astrocytes, microglia, or

oligodendrocytes (Supplementary Figure 1e; Supplementary Table 1), leaving 396 cells (~79 cells/timepoint-region; Supplementary Figure 1d).

To confirm that our methods can discriminate between different populations of neurons, we first explored differences between timepoints. Following a workflow similar to the recently described "dpFeature" procedure[17], we identified genes with highly variable transcriptional profiles and performed PCA. As anticipated, we observed that the greatest source of variation was between developmental ages (Figure 1b). Genes associated with negative PC1 loadings (E15.5 cells) were enriched for gene sets consistent with mitotically active neuronal, undifferentiated precursors (Figure 1c). In contrast, genes associated with positive PC1 loadings (P7 cells) were enriched for ontology terms associated with mature, post-mitotic neurons (Figure 1c). This initial analysis establishes our capacity to discriminate among biological classes present in our data using PCA as a foundation.

Further, we attempted to identify clusters of single cells between and within timepoints and anatomical regions. In order to do this, we selected the PCs that described the most variance in the data and used t-Stochastic Neighbor Embedding (t-SNE)[18] to further cluster cells in an unsupervised manner (see Methods). Analysis of all cells revealed that the E15.5 cells from both MB and FB cluster together (Figure 1d), supporting the notion that they are less differentiated. By contrast, cells isolated at P7 mostly cluster by anatomical region, suggesting progressive functional divergence with time (Figure 1d). We next applied this same scRNA-seq analysis workflow (See Methods) in a recursive manner individually in all regions at both timepoints to further explore heterogeneity. This revealed a total of 13 clusters (E15.5 FB.1-2, MB.1-2; P7

OB.1-3, FB.1-2, MB.1-4; Figure 1e), demonstrating the diversity of DA neuron subtypes and providing a framework upon which to evaluate the biological context of genetic association signals across closely-related cell types. Using known markers, we confirmed that all clusters expressed high levels of pan-neuronal markers (*Snap25*, *Eno2*, and *Syt1*) (Supplementary Figure 2a). In contrast, we observed scant evidence of astrocyte (*Aldh1l1*, *Slc1a3*, *Aqp4*, and *Gfap*; Supplementary Figure 2a) or oligodendrocyte markers (*Mag*, *Mog*, and *Mbp*; Supplementary Figure 2a), thus confirming we successfully isolated our intended substrate, *Th*+ neurons.

scRNA-seq revealed biologically and temporally discriminating transcriptional signatures

With subpopulations of DA neurons defined in our data, we set out to assign a biological identity to each cluster. Among the four clusters identified at E15.5, two were represented in t-SNE space as a single large group that included cells from both MB and FB (E15.MB.1, E15.FB.1), leaving two smaller clusters that were comprised solely of MB or FB cells (Supplementary Figure 3a). The latter MB cluster (E15.MB.2; Supplementary Figure 3a, Supplementary Figure 3b) specifically expressed *Foxa1*, *Lmx1a*, *Pitx3*, and *Nr4a2* and thus likely represents a post-mitotic DA neuron population[19] (Supplementary Table 2; Supplementary Table 3). Similarly, the discrete E15.FB.2 cluster expressed markers of post-mitotic FB/hypothalamic neurons (Supplementary Figure 3b), including *Six3, Six3os1*, *Sst,* and *Npy* {Supplementary Table 2; Supplementary Table 3}. These embryonic data did not discriminate between cells populating known domains of DA neurons, such as the SN.

By contrast, P7 cells mostly cluster by anatomical region and each region has defined subsets (Figure 1d, 1e, 2a). Analysis of P7 FB revealed two distinct cell clusters (Figure 2b). Expression

of the neuropeptides *Gal* and *Ghrh* and the *Gsx1* transcription factor place P7.FB.1 cells in the arcuate nucleus (Supplementary Table 2; Supplementary Table 3)[20–23]. The identity of P7.FB.2, however, was less clear, although subsets of cells therein did express other arcuate nucleus markers for $Th^{+}/Ghrh^{-}$ neuronal populations e.g. *Onecut2*, *Arx*, *Prlr*, *Slc6a3*, and *Sst* (Supplementary Figure 3c; Supplementary Table 3)[23]. All three identified OB clusters (Figure 2c) express marker genes of OB DA neuronal development or survival (Supplementary Table 2, Supplementary Table 3; Supplementary Figure 3d)[24]. It has previously been reported that *Dcx* expression diminishes with neuronal maturation[25] and *Snap25* marks mature neurons[26]. We observe that these OB clusters seem to reflect this continuum of maturation wherein expression of *Dcx* diminishes and *Snap25* increases with progression from P7.OB1 to OB3 (Supplementary Figure 3d). This pattern is mirrored by a concomitant increase in OB DA neuron fate specification genes (Supplementary Figure 3d)[24,27]. In addition, we identified four P7 MB DA subset clusters (Figure 2d). Marker gene analysis confirmed that three of the clusters correspond to DA neurons from the VTA (*Otx2* and *Neurod6*; P7.MB.1)[28,29], the PAG (*Vip* and *Pnoc*; P7.MB.3)[30,31], and the SN (*Sox6*, *Aldh1a7*, *Ndnf*, *Serpine2*, *Rbp4*, and *Fgf20*; P7.MB.4)[28,32–34] (Supplementary Table 2; Supplementary Table 3). These data are consistent with recent scRNA-seq studies of similar populations[33,35]. Through this marker gene analysis, we successfully assigned a biological identity to 12/13 clusters.

The only cluster without a readily assigned identity was P7.MB.2. This population of P7 MB DA neurons, P7.MB.2 (Figure 2d), is likely a progenitor-like population. Like the overlapping E15.MB.1 and E15.FB.1 clusters (Supplementary Figure 3a), this cluster preferentially expresses markers of neuronal precursors/differentiation/maturation {Supplementary Table 2,

171 Supplementary Table 3}. In addition to sharing markers with the progenitor-like E15.MB.1
172 cluster, P7.MB.2 exhibits gene expression consistent with embryonic mouse neuroblast
173 populations[33], cell division, and neuron development[36–40] (Supplementary Table 2,
174 Supplementary Table 3). Consistent with the hypothesis, this population displayed lower levels
175 of both *Th* and *Slc6a3,* markers of mature DA neurons, than the terminally differentiated and
176 phenotypically discrete P7 MB DA neuron populations of the VTA, SN and PAG (Figure 2e).
177
178 With this hypothesis in mind, we sought to ascertain the spatial distribution of P7.MB.2 DA
179 neurons through multiplex, single molecule fluorescence *in situ* hybridization (smFISH) for *Th*
180 (pan-P7 MB DA neurons), *Slc6a3* (P7.MB.1, P7.MB.3, P7.MB.4), and one of the neuroblast
181 marker genes identified through our analysis, either *Lhx9* or *Ldb2* (P7.MB.2) (Figure 2e). In each
182 experiment, we scanned the ventral midbrain for cells that were Th+/Slc6a3- and positive for the
183 third gene. *Th*+/*Slc6a3*-/*Lhx9*+ cells were found scattered in the dorsal SN *pars compacta*
184 (SNpc) along with cells expressing *Lhx9* alone (Figure 2f, 2h). Expression of *Ldb2* was found to
185 have a similar pattern to *Lhx9*, with *Th*+/*Slc6a3*-/*Ldb2*+ cells found in the dorsal SNpc (Figure
186 2f, 2h). Expression of *Lhx9* and *Ldb2* was low or non-existent in *Th*+/*Slc6a3*+ cells in the SNpc
187 (Figure 2e, 2f). Importantly, cells expressing these markers express *Th* at lower levels than
188 *Th*+/*Slc6a3*+ neurons (Figure 2f, 2g), consistent with our scRNA-seq data (Figure 2e). Thus,
189 with the resolution of the spatial distribution of this novel neuroblast-like P7 MB DA population,
190 we assign biological identity to each defined brain DA subpopulation.
191
192 *Novel SN-specific transcriptional profiles and GRNs highlight its association with PD*

Overall our analyses above allowed us to successfully separate and identify 13 brain DA neuronal populations present at E15.5 and P7, including SN DA neurons. Motivated by the clinical relevance of SN DA neurons to PD, we set out to understand what makes them transcriptionally distinct from the other MB DA neuron populations.

In order to look broadly at neuronal subtypes, we evaluated expression of canonical markers of other neuronal subtypes in our *Th*+ neuron subpopulations. Interestingly, we observed inconsistent detection of *Th* and eGFP in some E15.5 clusters (Supplementary Figure 2b). This likely reflects lower *Th* transcript abundance at this developmental state, but sufficient expression of the eGFP reporter to permit FACS collection (Supplementary Figure 2c). The expression of other DA markers, *Ddc* and *Slc18a2*, mirror *Th* expression, while *Slc6a3* expression is more spatially and temporally restricted (Supplementary Figure 2b). The SN cluster displays robust expression of all canonical DA markers (Supplementary Figure 2b). Multiple studies have demonstrated that *Th*+ neurons may also express markers characteristic of other major neuronal subtypes[41–43]. We found that only the SN and PAG showed no expression of either GABAergic (*Gad1*/*Gad2*/*Slc32a1*) or glutamatergic (*Slc17a6*) markers (Supplementary Figure 2b). This neurotransmitter specificity is a potential avenue for exploring the preferential vulnerability of the SN in PD.

Next, we postulated that genes whose expression defined the P7 SN DA neuron cluster might illuminate their preferential vulnerability in PD. We identified 110 SN-specific genes, by first finding all differentially expressed genes between P7 subset clusters and then using the Jensen-Shannon distance to identify cluster specific genes (See Methods). Prior reports confirm the

expression of 49 of the 110 SN-specific genes (~45%) in postnatal SN (Supplementary Table 4). We then sought evidence to confirm or exclude SN expression for the remaining, novel 61 genes (55%). Of these, 25/61 (~41%) were detected in adult SN neurons by *in situ* hybridization (ISH) of coronal sections in adult (P56) mice (Allen Brain Atlas, ABA; http://developingmouse.brain-map.org), including *Col25a1*, *Fam184a*, *Ankrd34b*, *Nwd2*, and *Cadps2* (Figure 3, Supplementary Table 5). Only 4/61 genes, for which ISH data existed in the ABA, lacked clear evidence of expression in the adult SN (Supplementary Table 5). The ABA lacked coronal ISH data on 32/61 genes, thus we were unable to confirm their presence in the SN. Collectively, we identify 110 postnatal SN DA marker genes and confirm the expression of those genes in the adult mouse SN for 74 (67%) of them, including 25 novel markers of this clinically relevant cell population that we confirmed using the ABA image catalog.

We next asked whether we could identify significant relationships between cells defined as being P7 SN DA neurons and distinctive transcriptional signatures in our data. We identify 16 co-expressed gene modules by performing weighted gene co-expression network analysis (WGCNA)[44,45] on all expressed genes of the P7 subset (Supplementary Figure 4; Supplementary Table 6). By calculating pairwise correlations between modules and P7 subset clusters, we reveal that 7/16 modules are significantly and positively correlated (Bonferroni corrected $p < 3.5e-04$) with at least one subset cluster (Figure 3c). We graphically represent the eigenvalues for each module in each cell in P7 t-SNE space, confirming that a majority of these significant modules (6/7) displayed robust spatial, isotype enrichment (Figure 3d).

238 In order identify the biological relevance of these modules, each module was tested for
239 enrichment for Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, Gene Ontology
240 (GO) gene sets, and Reactome gene sets. Two modules, the "brown" and "green" modules, were
241 significantly associated with the Parkinson's Disease KEGG pathway gene set (Figure 3c;
242 Supplementary Table 7). Interestingly, the "brown" module was also significantly correlated
243 with the P7 VTA population (P7.MB.1) and enriched for addiction gene sets (Supplementary
244 Table 7) highlighting the link between VTA DA neurons and addiction[46]. Strikingly, only the P7
245 SN cluster was significantly correlated with both PD-enriched modules (Figure 3c). This specific
246 correlation suggests these gene modules may play a role in the preferential susceptibility of the
247 SN in PD.
248
249 *Integrating SN DA neuron specific data enables prioritization of genes within PD-associated*
250 *intervals*
251 With these context-specific data in hand, we posited that SN DA neuron-specific genes and the
252 broader gene co-expression networks that correlate with SN DA neurons might be used to
253 prioritize genes within loci identified in PD GWAS. Such a strategy would be agnostic to prior
254 biological evidence and independent of genic position relative to the lead SNP, the traditional
255 method used to prioritize causative genes.
256
257 To investigate pertinent genes within PD GWAS loci, we identified all human genes within
258 topologically associated domains (TADs) and a two megabase interval encompassing each PD-
259 associated lead SNP. TADs were chosen because regulatory DNA impacted by GWAS variation
260 is more likely to act on genes within their own TAD[47]. While topological data does not exist for

261 SN DA neurons, we use TAD boundaries from hESCs as a proxy, as TADs are generally
262 conserved across cell types[48]. To improve our analyses, we also selected +/- 1 megabase interval
263 around each lead SNP thus including the upper bounds of reported enhancer-promoter
264 interactions[49,50]. All PD GWAS SNPs interrogated were identified by the most recent meta-
265 analysis (32 SNPs in total)[11], implicating a total of 1132 unique genes. We then identified
266 corresponding one-to-one mouse to human homologs (670/1132; ~59%), primarily through the
267 Mouse Genome Informatics homology database (Methods).
268
269 To prioritize these genes in GWAS loci, we developed a gene-centric score that integrates our
270 data as well as data in the public domain. We began by intersecting the PD loci genes with our
271 scRNA-seq data as well as previously published SN DA expression data[33], identifying 285 genes
272 (285/670; ~43%) with direct evidence of expression in SN DA neurons in at least one dataset.
273 Each PD-associated interval contained ≥1 SN-expressed gene (Table 1; Supplementary Table 8).
274 Emphasizing the need for a novel, systematic strategy, in 13/32 GWA intervals (~41%), the most
275 proximal gene to the lead SNP was not detectably expressed in mouse SN DA neuron
276 populations (Table 1; Supplementary Table 8). Surprisingly, two loci contained only one SN
277 DA-expressed gene (Table 1): *Mmp16* (*MMP16* locus) and *Tsnax* (*SIPA1L2* locus) (Figure 4a).
278 The relevance of these candidate genes to neuronal function/dysfunction is well supported[51,52].
279 This establishes gene expression in a relevant tissue as a powerful tool in the identification of
280 causal genes.
281
282 In order to prioritize likely diseases-associated genes in the remaining 30 loci, we scored genes
283 on three criteria: whether genes were identified as specific markers for the P7.MB.4 (SN) cluster

(Supplementary Table 2), whether the genes were differentially expressed between all P7 DA neuron populations, and whether the genes were included in PD gene set enriched and SN correlated gene modules uncovered in WGCNA (Supplementary Table 6). This strategy facilitated further prioritization of a single gene in 14 additional loci including *SNCA*, *LRRK2*, and *GCH1* loci (Figure 4a; Table 1). Importantly, using this approach we indict the familial PD gene encoding alpha-synuclein (SNCA), as responsible for the observed PD association within 4q22.1 (Figure 4a, Table 1). Thus, by using context-specific data alone, we were able to prioritize a single candidate gene in exactly half of the PD-GWAS associated loci.

Furthermore, at loci in which a single gene did not emerge, we identified dosage sensitive genes by considering the probability of being loss-of-function (LoF) intolerant (pLI) metric from the ExAC database[53,54]. 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 where a single gene was not prioritized. For those loci, including *MAPT* and *DDRGK1* loci (Figure 4a), we report a group of top scoring candidate genes (≤ 5) (Table 1). Expression of prioritized genes in the adult SN adds to the validity of the genes identified as possible candidates (Figure 4b).

Two interesting examples that emerge from this scoring are found at the *MAPT* and *TMEM175-GAK-DGKQ* loci. Although *MAPT* has previously been implicated in multiple neurodegenerative phenotypes, including PD (OMIM: 168600), we instead prioritize two genes before it (*CRHR1* and *NSF*; Table 1). We detect *Mapt* and *Nsf* expression consistently across all assayed DA

neurons (Figure 4c). By contrast, expression of *Crhr1,* encoding the corticotropin releasing hormone receptor 1*,* is restricted to P7 DA neurons in the SN and the more mature OB neuronal populations (Figure 4c). Similarly, at the *TMEM175-GAK-DGKQ* locus, our data shows that although all three proximal genes are expressed in the SN, the adjacent *CPLX1* was one of the prioritized gene (Table 1). There are multiple lines of evidence that strengthen *CPLX1* as a candidate gene. Expression of CPLX1 is elevated both in the brains of PD patients and the brains of mice overexpressing the *SNCA* A53T PD mutation[55,56]. Additionally, mice deficient in *CPLX1* display an early-onset, cerebellar ataxia along with prolonged motor and behavioral phenotypes[57,58]. However, the impact of *Cplx1* deficiency on the integrity of the nigrostriatal pathway, to date, has not been explored. In order to confirm *CPLX1* as a candidate gene, we performed immunohistochemistry (IHC) for *Th* in the *Cplx1* knockout mouse model (Supplementary Table 9)[57–59]. We measured the density of *Th*+ innervation in the striatum of *Cplx1* -/- mice and controls (Figure 4d, Supplementary Table 10) and found that *Cplx1* -/- mice had significantly lower *Th*+ staining in the striatum (p-value = 3.385e-08; Figure 4e). This indicates that *Cplx1* KO mice have less *Th*+ fiber innervation and a compromised nigrostriatal pathway, supporting its biological significance in MB DA populations and to PD.

The systematic identification of causal genes underlying GWAS signals is essential in order for the scientific and medical communities to take full advantage of all the GWAS data published over the last decade. Taken collectively, we demonstrate how scRNA-seq data from disease-relevant populations can be leveraged to illuminate GWAS results, facilitate systematic prioritization of GWAS loci implicated in PD, and can leads to the functional characterization of previously underexplored candidate genes.

DISCUSSION

Midbrain DA neurons in the SN have been the subject of intense research since being definitively linked to PD nearly 100 years ago[60]. While degeneration of SN DA neurons in PD is well established, they represent only a subset of brain DA populations. It remains unknown why nigral DA neurons are particularly vulnerable. We set out to explore this question using scRNA-seq. Recently, others have used scRNA-seq to characterize the mouse MB, including DA neurons[33]. Here, we extend these data significantly, extensively characterizing the transcriptomes of multiple brain DA populations longitudinally and discovering GRNs associated with specific populations.

Most importantly, our data facilitate the iterative and biologically informed prioritization of gene candidates for all PD-associated genomic intervals, the first such systematically generated dataset. 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. Our study overcomes this by integrating genomic data derived from specific cell contexts with analyses that are agnostic to one another. We posit that genes pertinent to PD are likely expressed within SN DA neurons. This hypothesis is consistent with the recent description of the "omnigenic" nature of common disease, wherein variation impacting genes expressed in a disease tissue explain the vast majority of risk[7].

First, we identify intervals that reveal one primary candidate, i.e. those that harbor only one SN-expressed gene. Next, we examine those intervals with many candidates, and prioritize based on a cumulative body of biological evidence. In total, we prioritize 5 or fewer candidates in all 32

353 PD GWAS loci studied, identifying a single gene in sixteen loci (16/32; 50%) and three or fewer
354 genes in ~84% of loci (27/32). Ultimately this prioritization reduces the candidate gene list for
355 PD GWAS loci dramatically from 1132 genes to 65 genes.
356
357 The top genes we identify in three PD loci (*SNCA*, *FGF20*, *GCH1*) have been directly associated
358 with PD, MB DA development, and MB DA function[34] (OMIM: 163890, 128230). Furthermore,
359 our prioritization of *CPLX1* over other candidates in the *TMEM175-GAK-DGKQ* locus is
360 supported by multiple lines of evidence. Additionally, we demonstrate that the integrity of the
361 nigrostriatal pathway is disrupted in *Cplx1* knockout mice. Dysregulation of *CPLX1* RNA is also
362 a biomarker in individuals with pre-PD prodromal phenotypes harboring the *PARK4* mutation
363 (*SNCA* gene duplication)[61]. These results validate our approach and strengthen the argument for
364 the use of context specific data in pinpointing candidate genes in GWAS loci.
365
366 Many of the genes prioritized (Table 1) have been shown to have various mitochondrial
367 functions[62–68]. The identification of genes associated with mitochondrial functions is especially
368 interesting in light of the "omnigenic" hypothesis of complex traits[7]. Since mitochondrial
369 dysfunction has been extensively implicated in PD[69], the prioritized genes may represent "core"
370 genes that in turn can affect the larger mitochondrial-associated regulatory networks active in the
371 disease relevant cell-type (SN DA neurons). It is notable that one of these genes is the presenilin
372 associated rhomboid like gene or *PARL*. *PARL* cleaves *PINK1*, a gene extensively implicated in
373 PD pathology and recently a variant in *PARL* has been associated with early-onset PD (OMIM:
374 607858)[70–72].
375

376 While our method successfully prioritized one familial PD gene (*SNCA*), we do not prioritize
377 *LRRK2,* another familial PD gene harbored within a PD GWAS locus. *Lrrk2* is not prioritized
378 simply because it is not detectably expressed in our SN DA neuronal population. This is
379 expected as numerous studies have reported little to no *Lrrk2* expression in *Th*+ MB DA neurons
380 both in mice and humans[73,74]. Instead, our method prioritizes *PDZRN4*. This result does not
381 necessarily argue against the potential relevance of *LRRK2* but instead provides an additional
382 candidate that may contribute to PD susceptibility. The same logic should be noted for two other
383 PD-associated loci, wherein our scoring prioritizes different genes (*KCNN3* and *CRHR1*/*NSF*,
384 respectively) than one previously implicated in PD (*GBA* and *MAPT*) (OMIM: 168600). Notably,
385 *KCNN3*, *CRHR1*, and *NSF*, all have previous biological evidence making them plausible
386 candidates[75–77].
387
388 Despite this success, we acknowledge several notable caveats. First, not all genes in PD-
389 associated human loci have identified mouse homologs. Thus, it remains possible that we may
390 have overlooked the contribution of some genes whose biology is not comprehensively queried
391 in this study. Secondly, we assume that identified genetic variation acts in a manner that is at
392 least preferential, if not exclusive, to SN DA neurons. Lastly, by prioritizing SN-expressed
393 genes, we assume that PD variation affects genes whose expression in the SN does not require
394 insult/stress. These caveats notwithstanding, our strategy sets the stage for a new generation of
395 independent and combinatorial functional evaluation of gene candidates for PD-associated
396 genomic intervals.
397
398

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

Animals.

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/). 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). All work involving mice (husbandry, colony maintenance and euthanasia) were reviewed and pre-approved by the institutional care and use committee.

Cplx1 knockout mice and wild type littermates used for immunocytochemistry were taken from a colony established in Cambridge using founder mice that were a kind gift of Drs K. Reim and N. Brose (Gottingen, Germany). *Cplx1* mice in this colony have been backcrossed onto a C57/Bl6J inbred background for at least 10 generations. All experimental procedures were licensed and undertaken in accordance with the regulations of the UK Animals (Scientific Procedures) Act 1986. Housing, rearing and genotyping of mice has been described in detail previously[57,58]. Mice were housed in hard-bottomed polypropylene experimental cages in groups of 5-10 mice in a housing facility was maintained at 21 – 23^{o}C with relative humidity of 55 ± 10%. Mice had *ad libitum* access to water and standard dry chow. Because homozygous knockout *Cplx1* mice have ataxia, they have difficulty in reaching the hard pellets in the food hopper and drinking from the water bottles. Lowered waterspouts were provided and access to normal laboratory chow was improved by providing mash (made by soaking 100 g of chow pellets in 230 ml water for 60 min until the pellets were soft and fully expanded) on the floor of the cage twice daily. *Cplx1* genotyping to identify mice with a homozygous or heterozygous deletion of the *Cplx1* gene was conducted as previously described[57], using DNA prepared from tail biopsies.

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

445 midbrain were then dissected and placed in HBSS on ice. This process was repeated for each
446 $EGFP^+$ brain. Four $EGFP^+$ brain regions for each region studied were pooled together for
447 dissociation.
448
449 **Dissection of P7 brains.**
450 After matings, pregnant females were sorted into their own cages and checked daily for newly
451 born pups. The morning the pups were born was considered day P0. Once the mice were aged to
452 P7, all the mice from the litter were euthanized and the brains were then quickly dissected out of
453 the mice and placed in HBSS without Mg^{2+} and Ca^{2+} on ice. As before, the brains were then
454 observed under a fluorescent microscope, $EGFP^+$ status for P7 mice was determined, and $EGFP^+$
455 brains were retained. For each $EGFP^+$ brain, the entire olfactory bulb was first resected and
456 placed in HBSS on ice. Immediately thereafter, the $EGFP^+$ forebrain and midbrain regions for
457 each brain were resected and also placed in distinct containers of HBSS on ice. Five $EGFP^+$
458 brain regions for each region were pooled together for dissociation.
459
460 **Generation of single cell suspensions from brain tissue.**
461 Resected brain tissues were dissociated using papain (Papain Dissociation System, Worthington
462 Biochemical Corporation; Cat#: LK003150) following the trehalose-enhanced protocol reported
463 by Saxena, et. al, 2012[78] with the following modifications: The dissociation was carried out at
464 37°C in a sterile tissue culture cabinet. During dissociation, all tissues at all time points were
465 triturated every 10 minutes using a sterile Pasteur pipette. For E15.5 tissues, this was continued
466 for no more than 40 minutes. For P7, this was continued for up to 1.5 hours or until the tissue
467 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[78], 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 Supplementary Figure 2c) 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[79]. Blank wells were used as negative controls for each plate collected. 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.

Library preparation and amplification of single-cell samples were performed using a modified version of the Smart-Seq2 protocol[79]. 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[80] with default parameters except: -p 8. Aligned reads from individual samples were quantified against a reference transcriptome (GENCODE vM8)[81] 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[82] 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[82] and the normalized FPKM matrix from cuffnorm was used as input for the Monocle 2 single cell RNA-seq framework[83] in R/Bioconductor[84]. Genes were annotated using the Gencode vM8 release[81]. A CellDataSet was then created using Monocle (v2.2.0)[83] 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 Monocle 2 Census algorithm[16] 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:

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 Supplementary Figure 1.

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 entire cds as well as subsets of the cds based on "age" and "region" of cells were created 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 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 include genes that may define rare populations of cells that could be present in any given population.

Monocle model preparation.

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)[18] 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 was deemed 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[85] 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.

Since the greatest source of variation in the data was between ages (Figure 1), differential expression analyses and downstream analyses were performed separately for each age.

In order to find differentially expressed genes between brain DA populations at each age, the E15.5 and P7 datasets were annotated with regional cluster identity ("subset cluster"). Differential expression analysis was performed using the "differentialGeneTest" function from Monocle that uses a likelihood ratio test to compare a vector generalized additive model (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:

"~subset.cluster" for E15.5 or P7 dataset

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

Cluster specific marker genes.

In order to identify differentially expressed genes that were "specifically" expressed in a particular subset cluster, R code calculating the Jensen-Shannon based specificity score from the R package cummeRbund[86] was used similar to what was described in Burns *et al*[87].

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 Molyneaux, *et al*[88].

To identify subset cluster specific genes, the distribution of specificity scores for each subset 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 (specificity $\geq$0.4). Once the specificity cutoff was chosen, genes were further filtered by only retaining genes that were expressed in >= 40% of cells within the subset cluster that the gene was determined to be specific for.

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)[89,90]. 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 (Supplementary Table 1). 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[91]. These functions were implemented through the 'compareCluster' function when analyzing WGCNA data.

Weighted Gene Co-Expression Network Analysis (WGCNA).

WGCNA was performed in R using the WGCNA package (v1.51)[44,45] 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 (Supplementary Figure 4a). 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 (Supplementary Figure 4b). 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[92] 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 (Supplementary Figure 4c) 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[91]. 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.

694 The data for human TAD boundaries were obtained from human embryonic stem cell (hESC)
695 Hi-C data[48] and converted from human genome hg18 to hg38 using the liftOver tool from UCSC
696 Genome Browser (http://genome.ucsc.edu/). PD GWAS SNP locations in hg38 were intersected
697 with the TAD information to identify TADs containing a PD GWAS SNP. The data for +/- 1
698 megabase regions surrounding PD GWAS SNPs was obtained by taking PD GWAS SNP
699 locations in hg38 and adding or subtracting 1e+06 from each location. All hg38 Ensembl
700 (version 87) genes that fell within the TADs or megabase regions were then identified by using
701 the biomaRt R package[93,94]. All genes were then annotated with PD locus and SNP information.
702 Mouse homologs for all genes were identified using human to mouse homology data from
703 Mouse Genome Informatics (MGI)
704 (http://www.informatics.jax.org/downloads/reports/HOM_MouseHumanSequence.rpt; Date
705 accessed: 07/07/2017). Homologs of protein coding genes that did not have a mouse homolog in
706 the data above were manually curated by searching the human gene name in the MGI database
707 (http://www.informatics.jax.org/). Of the 462 genes with no mouse homologs, only 60 (60/462,
708 ~13%) were annotated as protein coding genes (Supplementary Figure 5). 17 loci include at least
709 one protein coding gene with no identified, one-to-one mouse homolog (Supplementary Figure
710 5). All 670 genes with mouse homologs are annotated as "protein_coding." Genes homologs
711 were manually annotated if a homolog was found to exist. The TAD and megabase tables were
712 then combined to create a final PD GWAS locus-gene table.
713
714 *PD GWAS Loci Gene Scoring.*
715 Genes within PD GWAS loci were initially scored using two gene lists: Genes with an average
716 expression ≥0.5 transcripts in the SN cluster in our data (points = 1) and genes with an average

717 expression ≥0.5 transcripts in the SN population in La Manno, *et al*[33] (points = 1). Further
718 prioritization was accomplished by using three gene lists: genes that were differentially
719 expressed between subset clusters (points = 1); Genes found to be "specifically" expressed in the
720 P7 MB SN cluster (points = 1); Genes found in the WGCNA modules that are enriched for PD
721 (points = 1). Expression in the SN cluster was considered the most important feature and was
722 weighted as such through the use of two complementary datasets with genes found to be
723 expressed in both receiving priority. Furthermore, a piece of external data, pLI scores for each
724 gene from the ExAC database[54], was added to the scores in order to rank loci that were left with
725 ≥2 genes in the loci after the initial scoring. pLI scores
726 (fordist_cleaned_exac_r03_march16_z_pli_rec_null_data.txt) were obtained from
727 http://exac.broadinstitute.org/ (Date dowloaded: March 30, 2017).
728
729 **In situ hybridization.**
730 *In situ* hybridization data was downloaded from publically available data from the Allen Institute
731 through the Allen Brain Atlas (http://www.brain-map.org/). The image used in Figure 3a was
732 obtained from the Reference Atlas at the Allen Brain Atlas (http://mouse.brain-
733 map.org/static/atlas). URLs for all Allen Brain Atlas in situ data analyzed and downloaded for
734 SN marker genes (Figure 3b) are available in Supplementary Table 6. Data for SN expression *in*
735 *situ* data for PD GWAS genes (Figure 4b) were obtained from the following experiments: 1056
736 (*Th*), 79908848 (*Snca*), 297 (*Crhr1*), 74047915 (*Atp6v1d*), 72129224 (*Mmp16*), and 414 (*Cntn1*).
737 Data accessed on 03/02/17.
738
739 **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, Lot: 17073A), *Slc6a3* (C2; Cat No. 315441-C2, Lot: 17044A), *Lhx9* (C3; Cat No. 495431-C3, Lot: 17044A), and *Ldb2* (C3; Cat No. 466061-C3, Lot: 17044A). 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) 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* or *Ldb2*) 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*) 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 alongside 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.

smFISH 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 using probe sets provided by the manufacturer were used in order to calibrate laser power, offset, and detector sensitivity, for all channels in all experiments performed.

smFISH 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 edge. 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 2. 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 2.

Immunohistochemistry and quantification of *Th* striatum staining in *Cplx1* mice.

Mice (N=8 Cplx $1^{-/-}$; N=3 WT littermates; ages between 4-7.5 weeks) were euthanized and their brains fresh-frozen on powdered dry ice. Brains were sectioned at 35 mm and sections and mounted onto Superfrost-plus glass slides (VWR International, Poole, UK). Sections were

808 peroxidase inactivated, and one in every 10 sections was processed immunohistochemically for
809 tyrosine hydroxylase. Sections were incubated in primary anti-tyrosine hydroxylase antibody
810 (AB152, Millipore) used at 1/2000 dilution in 1% normal goat serum in phosphate-buffered
811 saline and 0.2% Triton X-100 overnight at 4^{o}C. Antigens were visualised using a horseradish
812 peroxidase-conjugated anti-rabbit second antibody (Vector, PI-1000, 1/2000 dilution) and
813 visualized using diaminobenzidine (DAB; Sigma). The slides were stored in the dark (in black
814 slide boxes) at room temperature (21 C).

815 Images of stained striatum were taken using a Nikon AZ100 microscope equipped with a 2x lens
816 (Nikon AZ Plan Fluor, NA 0.2, WD45), a Nikon DS-Fi2 camera, and NIS-Elements AR 4.5
817 software. Appropriate zoom and light exposure were determined before imaging and kept
818 constant for all slides and sections. Density of Th+ DAB staining was measured using ImageJ
819 software. Briefly, images were imported into ImageJ and the background was subtracted (default
820 50 pixels with "light background" selected). Next, images were converted to 8-bit and the image
821 was inverted. Five measurements of density were taken for each side of a striatum in a section
822 along with a density measurement from adjacent, unstained cortex. Striosomes were avoided
823 during measuring when possible. Striatal measurements had background (defined as staining in
824 the adjacent cortex in a section) subtracted. The mean section measurements (intensity/pixels
825 squared) for each brain were calculated and represented independent measurements of the same
826 brain. Variances were compared between the WT and KO populations. A two sample t-test was
827 then used to compare WT vs. *Cplx1* -/- section densities with the following parameters in R
828 using the "t.test" function: alternative = "two-sided", var.equal = "T".

829

830 **ACKNOWLEDGEMENTS**

831 The authors wish to thank Stephen M. Brown for implementation and optimization of smFISH.
832 Dr. Zhiguang Zheng and Mrs. Wendy Leavens for excellent technical support with the *Cplx1*
833 knockout mice and immunohistochemistry and Drs Kerstin Reim and Niels Brose for the gift of
834 the founder mice for the Cambridge Cplx1 knockout mice colony. This research was supported
835 in part by US National Institutes of Health grants R01 NS62972 and MH106522 to ASM and a
836 grant from CHDI *Inc*. to AJM.
837

838 **AUTHOR CONTRIBUTIONS**

839 PWH, ASM, and LAG designed the study and wrote the paper. PWH, SAM, WDL, GAC, and
840 AJM performed the experiments. PWH and LAG implemented the computational algorithms to
841 process the raw data and conduct analyses thereof. PWH, LAG, and ASM analyzed and
842 interpreted the resulting data. LAG contributed novel computational pipeline development.
843 Correspondence to ASM (andy@jhmi.edu) and LAG (loyalgoff@jhmi.edu).
844

845 **FINANCIAL INTERESTS STATEMENT**

846 The authors declare no competing financial interests.
847

REFERENCES

1. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *Am. J. Hum. Genet.* **90,** 7–24 (2012).
2. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science (80-.).* **337,** 1190–1195 (2012).
3. Farh, K. K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2015).
4. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. (2014). doi:10.1038/nature13138
5. Gupta, R. M. *et al.* A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression In Brief A common sequence variant that perturbs long-range enhancer interactions mediates risk for different vascular diseases. A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. *Cell* **170,** 522–533 (2017).
6. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47,** 955–61 (2015).
7. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169,** 1177–1186 (2017).
8. 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).
9. 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).

10. 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).

11. 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).

12. Puschmann, A. Monogenic Parkinson's disease and parkinsonism: clinical phenotypes and frequencies of known mutations. *Park. Relat Disord* **19,** 407–415 (2013).

13. Klein, C. & Westenberger, A. Genetics of Parkinson's disease. *Cold Spring Harb Perspect Med* **2,** a008888 (2012).

14. Heintz, N. Gene Expression Nervous System Atlas (GENSAT). *Nat. Neurosci.* **7,** 483–483 (2004).

15. 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).

16. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat Methods* **14,** 309–315 (2017).

17. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* (2017). doi:10.1038/nmeth.4402

18. Van Der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9,** 2579–2605 (2008).

19. Arenas, E., Denham, M. & Villaescusa, J. C. How to make a midbrain dopaminergic neuron. *Development* **142,** 1918–36 (2015).

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

894 21. Li, H., Zeitler, P. S., Valerius, M. T., Small, K. & Potter, S. S. Gsh-1, an orphan Hox
895 gene, is required for normal pituitary development. *EMBO J* **15,** 714–724 (1996).
896 22. McNay, D. E., Pelling, M., Claxton, S., Guillemot, F. & Ang, S. L. Mash1 is required for
897 generic and subtype differentiation of hypothalamic neuroendocrine cells. *Mol Endocrinol*
898 **20,** 1623–1632 (2006).
899 23. Campbell, J. N. *et al.* A molecular census of arcuate hypothalamus and median eminence
900 cell types. *Nat Neurosci* **20,** 484–496 (2017).
901 24. Agoston, Z. *et al.* Meis2 is a Pax6 co-factor in neurogenesis and dopaminergic
902 periglomerular fate specification in the adult olfactory bulb. *Development* **141,** 28–38
903 (2014).
904 25. Francis, F. *et al.* Doublecortin is a developmentally regulated, microtubule-associated
905 protein expressed in migrating and differentiating neurons. *Neuron* **23,** 247–256 (1999).
906 26. Gokce, O. *et al.* Cellular Taxonomy of the Mouse Striatum as Revealed by Single-Cell
907 RNA-Seq. *Cell Rep.* **16,** 1126–1137 (2016).
908 27. Vergaño-Vera, E. *et al.* Nurr1 blocks the mitogenic effect of FGF-2 and EGF, inducing
909 olfactory bulb neural stem cells to adopt dopaminergic and dopaminergic-GABAergic
910 neuronal phenotypes. *Dev Neurobiol* **75,** 823–841 (2015).
911 28. Panman, L. *et al.* Sox6 and Otx2 control the specification of substantia nigra and ventral
912 tegmental area dopamine neurons. *Cell Rep.* **8,** 1018–1025 (2014).
913 29. Viereckel, T. *et al.* Midbrain Gene Screening Identifies a New Mesoaccumbal
914 Glutamatergic Pathway and a Marker for Dopamine Cells Neuroprotected in Parkinson's
915 Disease. *Sci Rep* **6,** 35203 (2016).
916 30. 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).

31. 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).

32. 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).

33. La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* **167,** 566–580.e19 (2016).

34. Itoh, N. & Ohta, H. Roles of FGF20 in dopaminergic neurons and Parkinson's disease. *Front Mol Neurosci* **6,** 15 (2013).

35. Poulin, J. F. *et al.* Defining midbrain dopaminergic neuron diversity by single-cell gene expression profiling. *Cell Rep.* **9,** 930–943 (2014).

36. Uhde, C. W., Vives, J., Jaeger, I. & Li, M. Rmst is a novel marker for the mouse ventral mesencephalic floor plate and the anterior dorsal midline cells. *PLoS One* **5,** (2010).

37. Ng, S. Y., Bogu, G. K., Soh, B. & Stanton, L. W. The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis. *Mol. Cell* **51,** 349–359 (2013).

38. Ellis, B. C., Molloy, P. L. & Graham, L. D. CRNDE: A long non-coding RNA involved in CanceR Neurobiology, and DEvelopment. *Frontiers in Genetics* **3,** 1–15 (2012).

39. Lin, M. *et al.* RNA-Seq of human neurons derived from iPS cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders. *PLoS One* **6,** (2011).

40. Guttman, M. *et al.* lincRNAs act in the circuitry controlling pluripotency and

940 differentiation. (2011). doi:10.1038/nature10398

941 41. Morales, M. & Margolis, E. B. Ventral tegmental area: cellular heterogeneity,
942 connectivity and behaviour. *Nat Rev Neurosci* **18,** 73–85 (2017).

943 42. Everitt, B. J., Hökfelt, T., Wu, J. Y. & Goldstein, M. Coexistence of tyrosine hydroxylase-
944 like and gamma-aminobutyric acid-like immunoreactivities in neurons of the arcuate
945 nucleus. *Neuroendocrinology* **39,** 189–191 (1984).

946 43. Asmus, S. E. *et al.* Increasing proportions of tyrosine hydroxylase-immunoreactive
947 interneurons colocalize with choline acetyltransferase or vasoactive intestinal peptide in
948 the developing rat cerebral cortex. *Brain Res* **1383,** 108–119 (2011).

949 44. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network
950 analysis. *BMC Bioinformatics* **9,** 559 (2008).

951 45. Langfelder, P. & Horvath, S. Fast R Functions for Robust Correlations and Hierarchical
952 Clustering. *J Stat Softw* **46,** (2012).

953 46. Pascoli, V., Terrier, J., Hiver, A. & Lüscher, C. Sufficiency of Mesolimbic Dopamine
954 Neuron Stimulation for the Progression to Addiction. *Neuron* **88,** 1054–1066 (2015).

955 47. Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional
956 organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* **14,**
957 390–403 (2013).

958 48. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of
959 chromatin interactions. *Nature* **485,** 376–380 (2012).

960 49. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing
961 limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12,** 1725–1735
962 (2003).

963 50. Benko, S. *et al.* Highly conserved non-coding elements on either side of SOX9 associated
964 with Pierre Robin sequence. *Nat. Genet.* **41,** 359–364 (2009).
965 51. Yong, V. W., Power, C., Forsyth, P. & Edwards, D. R. Metalloproteinases in biology and
966 pathology of the nervous system. *Nat Rev Neurosci* **2,** 502–511 (2001).
967 52. Li, Z., Wu, Y. & Baraban, J. M. The Translin/Trax RNA binding complex: clues to
968 function in the nervous system. *Biochim Biophys Acta* **1779,** 479–485 (2008).
969 53. Doan, R. N. *et al.* Mutations in Human Accelerated Regions Disrupt Cognition and Social
970 Behavior. *Cell* **167,** 341–354.e12 (2016).
971 54. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,**
972 285–291 (2016).
973 55. Basso, M. *et al.* Proteome analysis of human substantia nigra in Parkinson's disease.
974 *Proteomics* **4,** 3943–3952 (2004).
975 56. Gispert, S. *et al.* Complexin-1 and Foxp1 Expression Changes Are Novel Brain Effects of
976 Alpha-Synuclein Pathology. *Mol. Neurobiol.* **52,** 57–63 (2015).
977 57. Glynn, D., Drew, C. J., Reim, K., Brose, N. & Morton, A. J. Profound ataxia in complexin
978 I knockout mice masks a complex phenotype that includes exploratory and habituation
979 deficits. *Hum. Mol. Genet.* **14,** 2369–2385 (2005).
980 58. Glynn, D., Sizemore, R. J. & Morton, A. J. Early motor development is abnormal in
981 complexin 1 knockout mice. *Neurobiol. Dis.* **25,** 483–495 (2007).
982 59. Kielar, C., Sawiak, S. J., Negredo, P. N., Tse, D. H. Y. & Morton, A. J. Tensor-based
983 morphometry and stereology reveal brain pathology in the complexin1 knockout mouse.
984 *PLoS One* **7,** (2012).
985 60. 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).

61. Lahut, S. *et al.* Blood RNA biomarkers in prodromal PARK4 and REM sleep behavior disorder show role of complexin-1 loss for risk of Parkinson's disease. *Dis. Model. Mech.* dmm.028035 (2017). doi:10.1242/dmm.028035

62. Hildick-Smith, G. J. *et al.* Macrocytic anemia and mitochondriopathy resulting from a defect in sideroflexin 4. *Am. J. Hum. Genet.* **93,** 906–914 (2013).

63. Islam, M. M., Suzuki, H., Makoto, Y. & Tanaka, M. Primary structure of the smallest (6.4-kDa) subunit of human and bovine ubiquinol-cytochrome c reductase deduced from cDNA sequences. *Biochem Mol Biol Int.* **41,** 1109–1116 (1997).

64. Swartz, D. A., Park, E. I., Visek, W. J. & Kaput, J. The e subunit gene of murine F1F0-ATP synthase. Genomic sequence, chromosomal mapping, and diet regulation. *J. Biol. Chem.* **271,** 20942–20948 (1996).

65. Tomar, D. *et al.* MCUR1 Is a Scaffold Factor for the MCU Complex Function and Promotes Mitochondrial Bioenergetics. *Cell Rep.* **15,** 1673–1685 (2016).

66. Plovanich, M. *et al.* MICU2, a Paralog of MICU1, Resides within the Mitochondrial Uniporter Complex to Regulate Calcium Handling. *PLoS One* **8,** (2013).

67. Wonsey, D. R., Zeller, K. I. & Dang, C. V. The c-Myc target gene PRDX3 is required for mitochondrial homeostasis and neoplastic transformation. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 6649–54 (2002).

68. Curran, J. E. *et al.* Genetic variation in PARL influences mitochondrial content. *Hum. Genet.* **127,** 183–190 (2010).

69. Winklhofer, K. F. & Haass, C. Mitochondrial dysfunction in Parkinson's disease. *Biochim Biophys Acta* **1802,** 29–44 (2010).

1009 70. Shi, G. *et al.* Functional alteration of PARL contributes to mitochondrial dysregulation in
1010 Parkinson's disease. *Hum. Mol. Genet.* **20,** 1966–1974 (2011).
1011 71. Shi, G. & McQuibban, G. A. The Mitochondrial Rhomboid Protease PARL Is Regulated
1012 by PDK2 to Integrate Mitochondrial Quality Control and Metabolism. *Cell Rep.* **18,** 1458–
1013 1472 (2017).
1014 72. Jin, S. M. *et al.* Mitochondrial membrane potential regulates PINK1 import and
1015 proteolytic destabilization by PARL. *J. Cell Biol.* **191,** 933–942 (2010).
1016 73. Galter, D. *et al.* LRRK2 expression linked to dopamine-innervated areas. *Ann Neurol* **59,**
1017 714–719 (2006).
1018 74. Higashi, S. *et al.* Expression and localization of Parkinson's disease-associated leucine-
1019 rich repeat kinase 2 in the mouse brain. *J Neurochem* **100,** 368–381 (2007).
1020 75. Soden, M. E. *et al.* Disruption of Dopamine Neuron Activity Pattern Regulation through
1021 Selective Expression of a Human KCNN3 Mutation. *Neuron* **80,** 997–1009 (2013).
1022 76. Abuirmeileh, A., Harkavyi, A., Kingsbury, A., Lever, R. & Whitton, P. S. The CRF-like
1023 peptide urocortin greatly attenuates loss of extracellular striatal dopamine in rat models of
1024 Parkinson's disease by activating CRF1 receptors. *Eur. J. Pharmacol.* **604,** 45–50 (2009).
1025 77. Simunovic, F. *et al.* Gene expression profiling of substantia nigra dopamine neurons:
1026 further insights into Parkinson's disease pathology. *Brain* **132,** 1795–1809 (2009).
1027 78. Saxena, A. *et al.* Trehalose-enhanced isolation of neuronal sub-types from adult mouse
1028 brain. *Biotechniques* **52,** 381–385 (2012).
1029 79. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9,**
1030 171–181 (2014).
1031 80. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory

requirements. *Nat. Methods* **12,** 357–60 (2015).

81. Mudge, J. M. & Harrow, J. Creating reference gene annotation for the mouse C57BL6/J genome assembly. *Mamm Genome* **26,** 366–378 (2015).

82. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7,** 562–78 (2012).

83. 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).

84. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12,** 115–121 (2015).

85. Wang, X.-F. & Xu, Y. Fast clustering using adaptive density peak detection. *Stat. Methods Med. Res.* 1–14 (2015). doi:10.1177/0962280215609948

86. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31,** 46–53 (2013).

87. 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).

88. Molyneaux, B. J. *et al.* DeCoN: Genome-wide analysis of invivo transcriptional dynamics during pyramidal neuron fate selection in neocortex. *Neuron* **85,** 275–288 (2015).

89. 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).

90. 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).

1055 91. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing
1056 biological themes among gene clusters. *OMICS* **16,** 284–7 (2012).
1057 92. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree:
1058 the Dynamic Tree Cut package for R. *Bioinformatics* **24,** 719–720 (2008).
1059 93. Durinck, S. *et al.* BioMart and Bioconductor: A powerful link between biological
1060 databases and microarray data analysis. *Bioinformatics* **21,** 3439–3440 (2005).
1061 94. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the
1062 integration of genomic datasets with the R/Bioconductor package biomaRt. (2009).
1063 doi:10.1038/nprot.2009.97

FIGURES

Figure 1

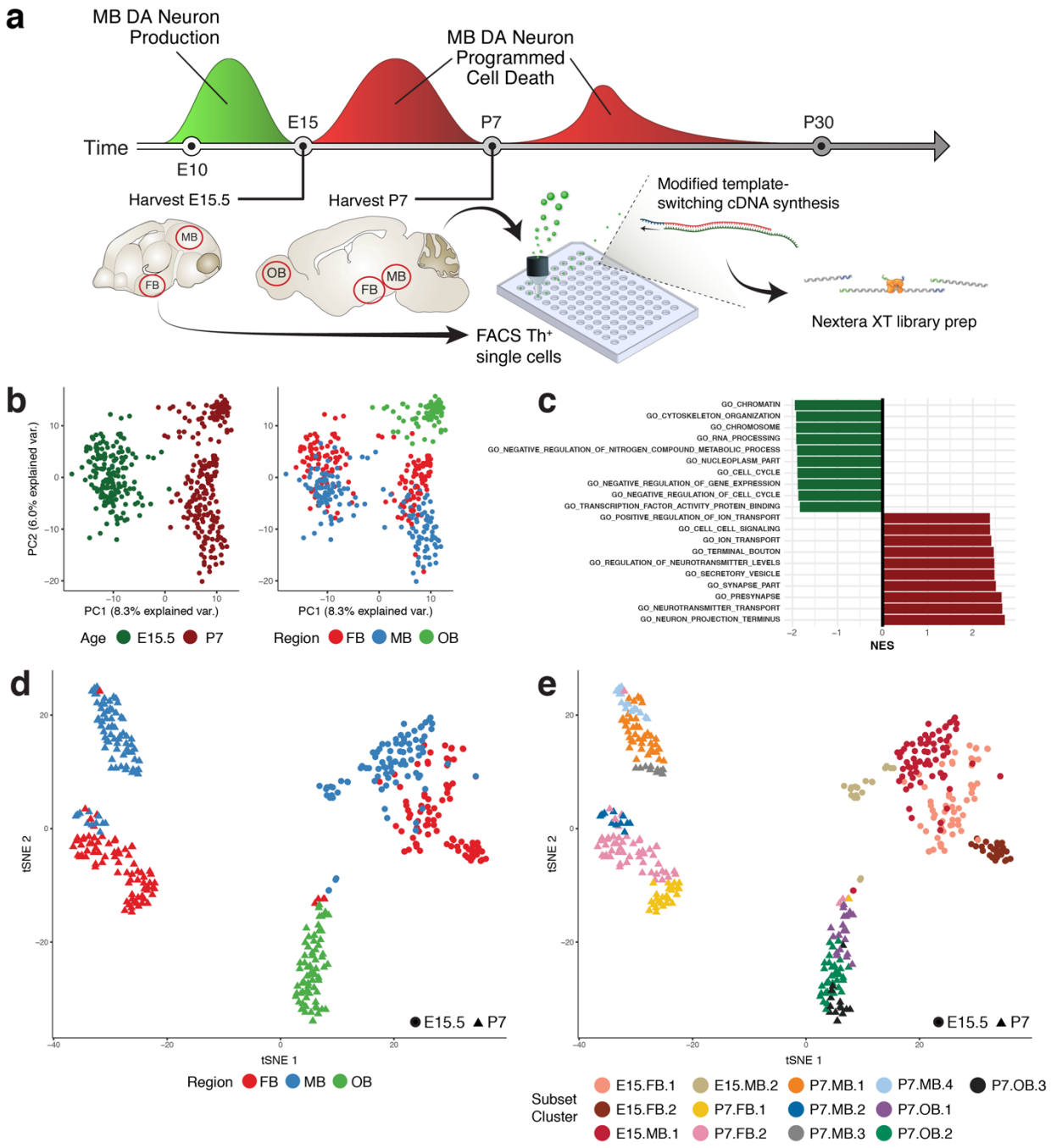


Figure 1. scRNA-seq analysis of isolated cells allows their separation by developmental time. a) Diagram of scRNA-seq experimental procedures for isolating and sequencing EGFP+ cells. b) Principal component analysis (PCA) on all cells collected using genes with highly variant transcriptional profiles. 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 (GO) gene sets enriched in genes with positive (red) and negative (green) PC1 loadings. Genes with negative PC1 loadings and negative normalized enrichment scores (NES) were enriched for terms indicative of mitotically active cells. Genes with positive PC1 loadings and NES scores were enriched for terms expected of more mature neurons. d) A t-distributed Stochastic Neighbor Embedding (t-SNE) plot of all collected cells colored by regional identity. E15.5 cells cluster together while P7 cells cluster primarily by regional identity. e) A t-SNE plot of all collected cells colored by subset cluster identity. Through iterative analysis, timepoint-regions collected can be separated into multiple subpopulations (13 in total). Midbrain, Mb; Forebrain, FB; Olfactory bulb; OB; Fluorescence activated cell sorting; FACS.

Figure 2

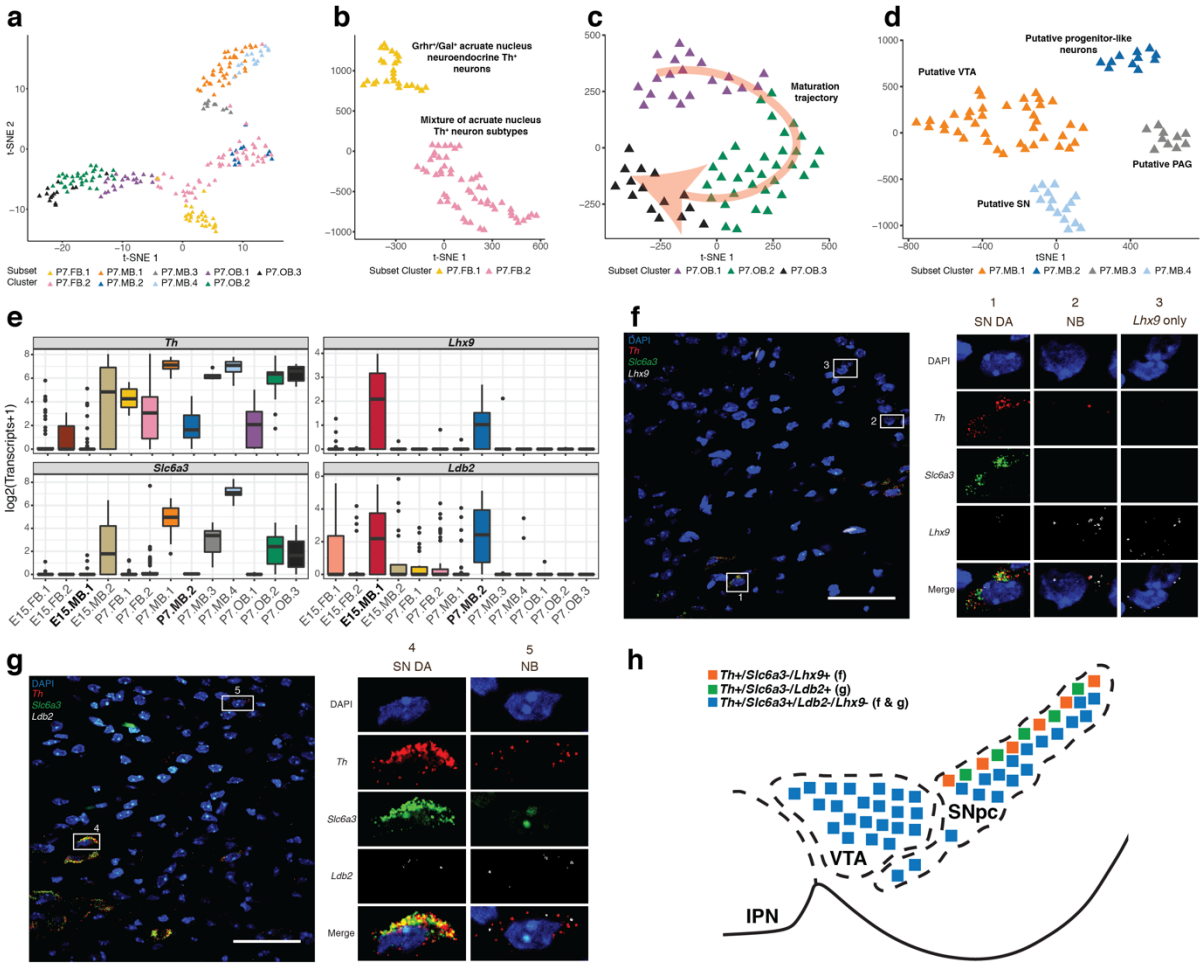


Figure 2. Subclusters of P7 *Th*+ neurons are identified based on marker gene analyses. a) A t-SNE plot of all P7 neurons collected using colored by subset cluster identity. The neurons mostly cluster by regional identity. b) t-SNE plot of P7 FB neurons. P7 FB neurons cluster into two distinct populations. c) t-SNE plot of P7 OB neurons. P7 OB neurons cluster into three populations. These populations represent a trajectory of Th+ OB maturation (Table S3) as indicated by the red arrow. d) A t-SNE plot of P7 MB neurons. P7 MB neurons cluster into four clusters: the *substantia nigra* (SN), the ventral tegmental area (VTA), the periaqueductal grey area (PAG), and a novel progenitor-like population. e) Boxplots displaying the expression of four genes (*Th*, *Slc6a3*, *Lhx9*, and *Ldb2*) across all subclusters identified. The novel P7 MB progenitor-like cluster (P7.MB.2) has a similar expression profile to E15.5 MB neuroblast population (E15.MB.1) (Table S2). +/- 1.5x interquartile range is represented by the whiskers on the boxplots. Data points beyond 1.5x interquartile range are considered as outliers and plotted as black points. f) Representative image of multiplex single molecule fluorescent *in situ* hybridization (smFISH) for *Th*, *Slc6a3*, and *Lhx9*, in the mouse ventral midbrain. Zoomed-in panels represent cell populations observed. Scale bar, 50 uM. g) Representative image of multiplex smFISH for *Th*, *Slc6a3*, and *Ldb2*, in the mouse ventral midbrain. Zoomed-in panels represent cell populations observed. h) Diagram of ventral midbrain summarizing the results of smFISH. Th+/Slc6a3-/Lhx9+ and Th+/Slc6a3-/Ldb2+ cells are both found in the dorsal SN. Scale bar, 50 uM. NB, neuroblast; SN, substantia nigra; VTA, ventral tegmental area; IPN, interpeduncular nucleus.

Figure 3

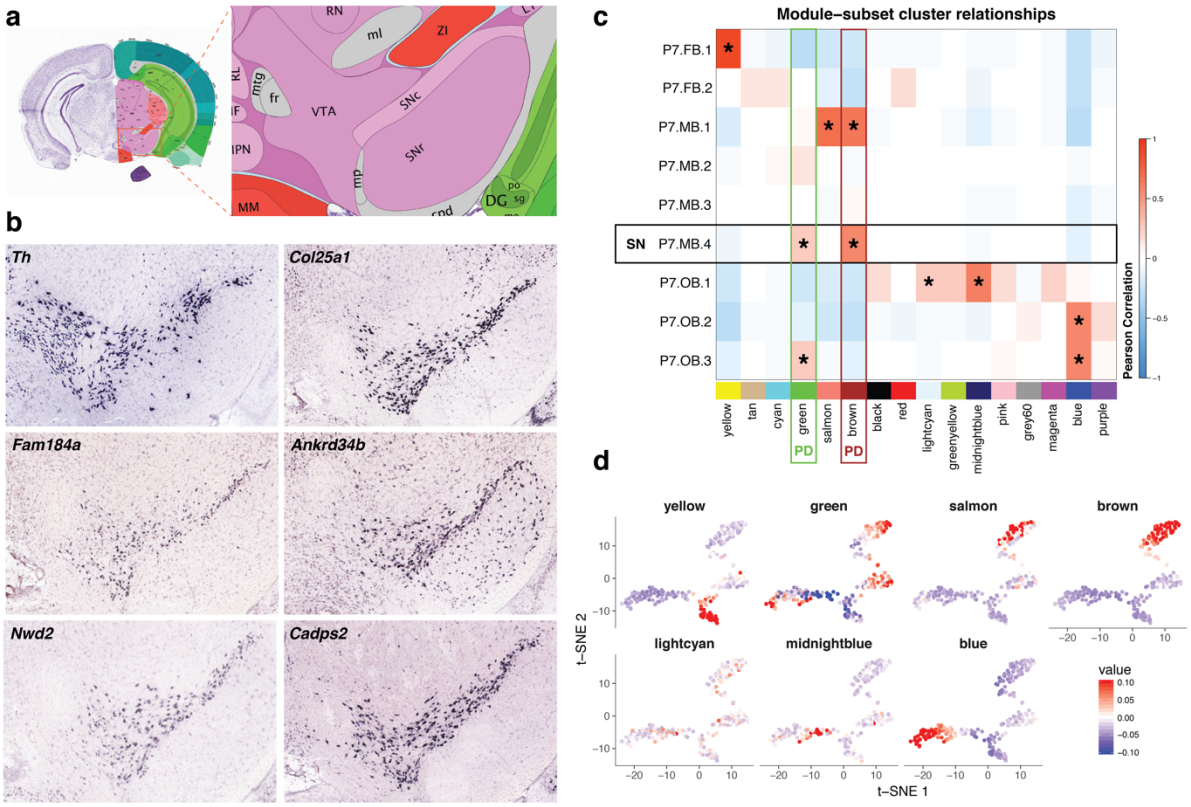


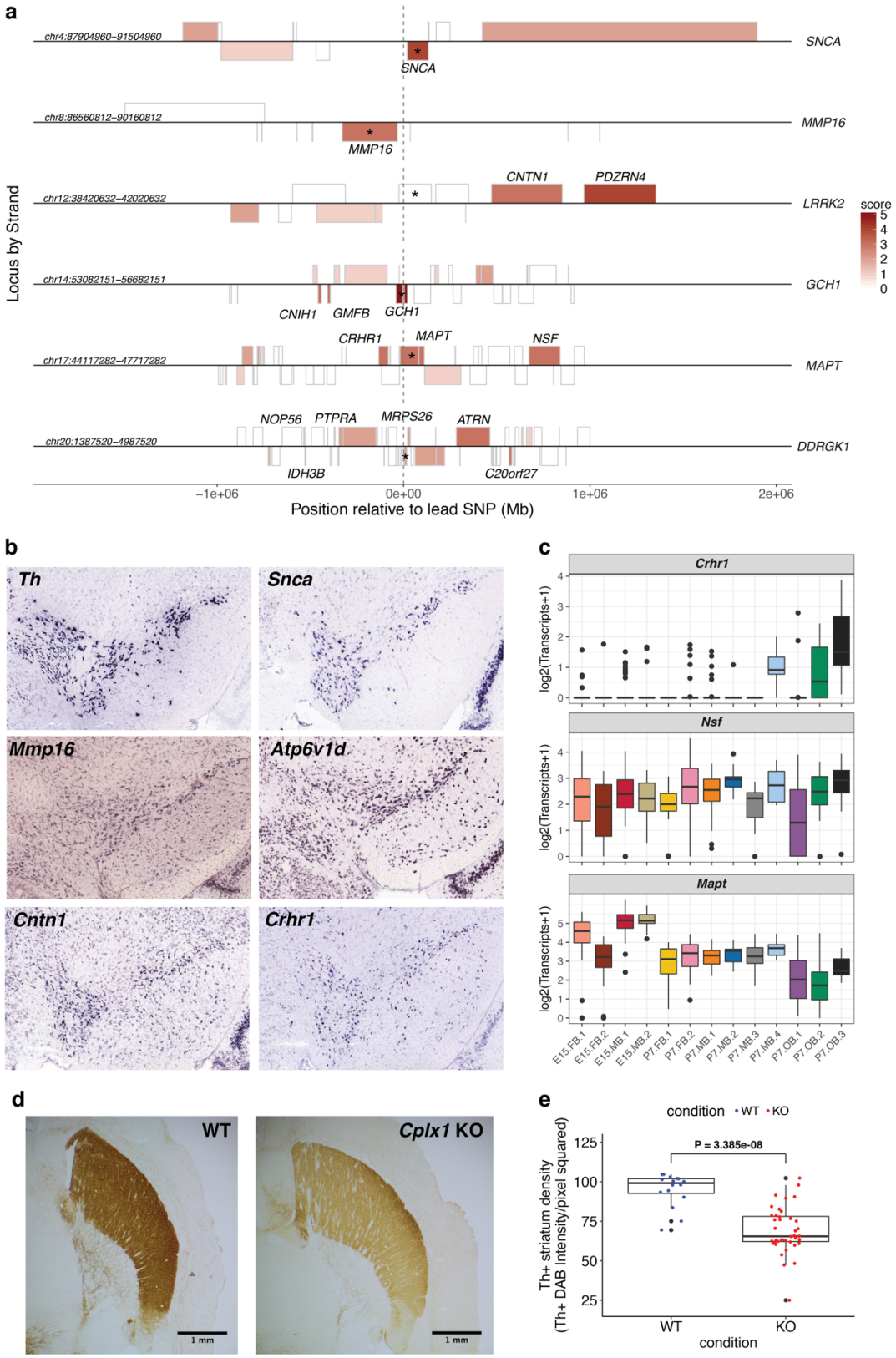
Figure 3. Novel markers and gene modules reveal context specific SN DA biology. a) Reference atlas diagram from the Allen Brain Atlas (ABA; http://www.brain-map.org/) of the P56 mouse ventral midbrain. b) Confirmation of novel SN DA neuron marker genes through the use of ABA *in situ* hybridization data (http://www.brain-map.org/). Coronal, P56 mouse *in situ* data was explored in order to confirm the expression of 25 novel SN markers. *Th* expression in P56 mice was used as an anatomical reference during analysis. c) 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. 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 < 3.5e-04) contain an asterisk. SN cluster (P7.MB.4) identity, denoted by a black rectangle, was found to be highly correlated with two modules ("green" and "brown") that were enriched for the "Parkinson's Disease" KEGG gene set (labeled with "PD"). d) The eigengene value for each P7 neuron in the seven WGCNA modules shown to be significantly positively associated with a subset cluster overlaid on the t-SNE plot of all P7 neurons (Figure 2a). Plotting of

1114 eigengenes confirms strict spatial restriction of module association. Only the “lightcyan” module does not seem to

1115 show robust spatial restriction.

1116

1117 **Figure 4**



1118

Figure 4. Context specific SN DA data allows for the prioritization of candidate genes in PD GWAS loci. a) A locus plot displaying four megabase regions in the human genome (hg38) centered on PD GWAS SNPs in six loci. Genes are displayed as boxes on their appropriate strand. Genes are shaded by their prioritization score and gene names are displayed for genes with a score of 3 or higher in each locus. b) *In situ* hybridization from the ABA (http://www.brain-map.org/) of five prioritized genes along with *Th* for an anatomical reference. Coronal, P56 mouse *in situ* data was used. c) Boxplots displaying expression of prioritized genes from the *MAPT* locus (Figure 4a; Table 1). +/- 1.5x interquartile range is represented by the whiskers on the boxplots. Data points beyond 1.5x interquartile range are considered as outliers and plotted as black points. d) Representative light microscopy images of *Th*+ innervation density in the striatum of WT and *Cplx1* knockout (KO) mice. Scale bar, 1 mm. e) Boxplots comparing the level of Th+ striatum innervation between WT and *Cplx1* KO mice. DAB staining density was measured in 35 uM, horizontal sections in WT mice (mice = 3, sections = 16) and *Cplx1* KO mice (mice = 8, sections = 40). Each point in the boxplot represents a stained, 35 uM section. Statistical analyses were performed between conditions with section averages in order to preserve observed variability (WT n = 16, Cplx1 KO n = 40). A two sample t-test revealed that Th+ innervation density was significantly lower in *Cplx1* KO mice ($t = 6.4395$, $df = 54$, $p = 3.386e-08$). Data points outside of 1.5x interquartile range, represented by the whiskers on the boxplots, are considered as outliers and plotted as black points.

TABLES

Table 1. Summary of the systematic scoring of genes in 32 GWAS loci associated with PD

Locus	Genes	Mouse Homologs	Expressed Genes	Top Candidate Genes	Prioritized by	Closest Gene Expressed
ACMSD-TMEM163	17	9	5	*UBXN4; CCNT2; R3HDM1; RAB3GAP1*	SN expression; pLI	No, neither
BCKDK-STX1B	90	54	19	*MAPK3; VKORC1; BOLA2B*	SN expression; Differential expression; pLI	Yes, both
BST1	19	10	3	*CPEB2*	SN expression; Differential expression	No
CCDC62	50	41	17	*ARL6IP4*	SN expression; Differential expression	No
DDRGK1	54	37	13	*ATRN; NOP56; MRPS26; C20orf27; IDH3B*	SN expression; Differential expression; pLI	Yes
DLG2	15	7	4	*DLG2; CCDC90B*	SN expression; Differential expression; pLI	Yes
FAM47E-SCARB2	33	22	10	*G3BP2; CCNI; CDKL2*	SN expression; Differential expression; pLI	Yes - *SCARB2*, No - *FAM47E*
FGF20	18	12	8	*FGF20; ZDHHC2; TUSC3; MICU3; MTMR7*	SN expression; Differential expression; SN specific; pLI	Yes
GBA-SYT11	84	56	36	*KCNN3*	SN expression; Differential expression; SN specific; WGCNA module	Yes, both
GCH1	29	15	7	*GCH1*	SN expression; Differential expression; SN specific; WGCNA module	Yes
GPNMB	28	13	5	*RAPGEF5*	SN expression; Differential expression	No
HLA-DBQ1	164	99	31	*ATP6V1G2*	SN expression; Differential expression; WGCNA module	No
INPP5F	29	13	7	*PRDX3; NANOS1; INPP5F; SFXN4*	SN expression; Differential expression; pLI	Yes
ITGA8	27	15	5	*FAM171A1*	SN expression; Differential expression	No
KRT8P25-APOOP2	17	7	2	*CHMP2B*	SN expression; Differential expression	No, neither are in mouse
LRRK2	10	7	4	*PDZRN4*	SN expression; Differential expression; WGCNA module	No
MAPT	40	20	8	*CRHR1; NSF; MAPT*	SN expression; Differential expression; pLI	Yes
MCCC1	25	11	5	*DCUN1D1; ABCC5; PARL*	SN expression; Differential expression; pLI	No
MIR4697	14	11	6	*OPCML*	SN expression; Differential expression	Not in mouse
MMP16	9	2	1	*MMP16*	SN expression	Yes
NMD3	20	10	3	*B3GALNT1*	SN expression; Differential expression	Yes
RAB7L1-NUCKS1	42	31	11	*LRRN2; KLHDC8A; SRGAP2*	SN expression; Differential expression; pLI	Yes - *NUCKS1*, No - *RAB7L1* (*Rab29*)
RIT2	6	3	3	*RIT2; SYT4*	SN expression; Differential expression; pLI	Yes
SIPA1L2	15	6	1	*TSNAX*	SN expression	No
SNCA	11	7	4	*SNCA*	SN expression; Differential expression; WGCNA module	Yes
SPPL2B	80	65	29	*UQCR11*	SN expression; Differential expression; WGCNA module	Yes
SREBF1-RAI1	67	26	12	*COPS3; NT5M*	SN expression; Differential expression; pLI	No, neither

STK39	17	10	4	*STK39; B3GALT1*	SN expression; Differential expression; pLI	Yes
TMEM175-GAK-DGKQ	40	25	10	*MAEA; CPLX1; ATP5I; TMEM175*	SN expression; Differential expression; WGCNA module; pLI	Yes, all three
TMEM229B	24	15	8	*VTI1B; ATP6V1D*	SN expression; Differential expression; pLI	Yes
USP25	27	6	2	*HSPA13*	SN expression	Yes
VPS13C	11	5	2	*TLN2; RORA*	SN expression; pLI	No

Scoring was carried out at described in the Results and Methods. Candidate genes are presented for each of 32 PD GWAS loci (identified by Nalls, *et al*[11]. Additional information for each PD GWAS locus is presented including the number of unique genes scored, the number genes with a mouse homolog, the number of genes expressed in mouse SN DA neurons, which data prioritized the top genes, and whether the closest gene to the lead SNP is expressed. Detailed scoring for each gene can be found in Supplementary Table 8.