Single-cell RNA-seq of mouse 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

Genetic variation modulating risk of sporadic Parkinson's disease (PD) has been primarily explored through genome wide association studies (GWAS). However, like many other common genetic diseases, the impacted genes remain largely unknown. Here, we used single-cell RNA-seq to characterize dopaminergic (DA) neuron populations in the mouse brain at embryonic and early postnatal timepoints. These data facilitated unbiased identification of DA neuron subpopulations through their unique transcriptional profiles, including a novel postnatal neuroblast population and *substantia nigra* (SN) DA neurons. We use these population-specific data to develop a scoring system to prioritize candidate genes in all 49 GWAS intervals implicated in PD risk, including known PD genes and many with extensive supporting literature. As proof of principle, we confirm that the nigrostriatal pathway is compromised in *Cplx1* null mice. Ultimately, this systematic approach establishes biologically pertinent candidates and testable hypotheses for sporadic PD, informing a new era of PD genetic research.

31 The most commonly used genetic tool today for studying complex disease is the genome wide
32 association study (GWAS). As a strategy, GWAS was initially hailed for the insight it might
33 provide into the genetic architecture of common human disease risk. Indeed, the collective data
34 from GWAS since 2005 has revealed a trove of variants and genomic intervals associated with
35 an array of phenotypes[1]. The majority of variants identified in GWAS are located in non-coding
36 DNA[2] and are enriched for characteristics denoting regulatory DNA[2,3]. This regulatory variation
37 is expected to impact expression of a nearby gene, leading to disease susceptibility.
38
39 Traditionally, the gene closest to the lead SNP has been prioritized as the affected gene.
40 However, recent studies show that disease-associated variants can act on more distally located
41 genes, invalidating genes that were previously extensively studied[4,5]. The inability to
42 systematically connect common variation with the genes impacted limits our capacity to
43 elucidate potential therapeutic targets and can waste valuable research efforts.
44
45 Although GWAS is inherently agnostic to the context in which disease-risk variation acts, the
46 biological impact of common functional variation has been shown to be cell context
47 dependent[2,6]. Extending these observations, Pritchard and colleagues recently demonstrated that
48 although genes need only to be expressed in disease-relevant cell types to contribute to risk,
49 those expressed preferentially or exclusively therein contribute more per SNP[7]. Thus, accounting
50 for the cellular and gene regulatory network (GRN) contexts within which variation act may
51 better inform the identification of impacted genes. These principles have not yet been applied
52 systematically to many of the traits for which GWAS data exists. We have chosen Parkinson's

53 disease (PD) as a model complex disorder for which a significant body of GWAS data remains to
54 be explored biologically in a context dependent manner.
55
56 PD is the most common progressive neurodegenerative movement disorder. Incidence of PD
57 increases with age, affecting an estimated 1% worldwide beyond 70 years of age[8–10]. The genetic
58 underpinnings of non-familial or sporadic PD have been studied through the use of GWAS with
59 recent meta-analyses highlighting 49 loci associated with sporadic PD susceptibility[11,12]. While a
60 small fraction of PD GWAS loci contain genes known to be mutated in familial PD (*SNCA* and
61 *LRRK2*)[13,14], most indicted intervals do not contain a known causal gene or genes. Although PD
62 ultimately affects multiple neuronal centers, preferential degeneration of DA neurons in the SN
63 leads to functional collapse of the nigrostriatal pathway and loss of fine motor control. The
64 preferential degeneration of SN DA neurons in relation to other mesencephalic DA neurons has
65 driven research interest in the genetic basis of selective SN vulnerability in PD. Consequently,
66 one can reasonably assert that a significant fraction of PD-associated variation likely mediates its
67 influence specifically within the SN.
68
69 In an effort to illuminate a biological context in which PD GWAS results could be better
70 interpreted, we undertook single-cell RNA-seq (scRNA-seq) analyses of multiple DA neuronal
71 populations in the brain, including ventral midbrain DA neurons. This analysis defined the
72 heterogeneity of DA populations over developmental time in the brain, revealing gene
73 expression profiles specific to discrete DA neuron subtypes. These data further facilitated the
74 definition of GRNs active in DA neuron populations including the SN. With these data, we

75 establish a framework to systematically prioritize candidate genes in all 49 PD GWAS loci and
76 begin exploring their pathological significance.
77
78 **RESULTS**
79 *scRNA-seq characterization defines DA neuronal subpopulation heterogeneity*
80 In order to characterize DA neuron molecular phenotypes, we undertook scRNA-seq on cells
81 isolated from distinct anatomical locations of the mouse brain over developmental time. We used
82 fluorescence activated cell sorting (FACS) to retrieve single DA neurons from the Tg(Th-
83 EGFP)DJ76Gsat BAC transgenic mouse line, which expresses eGFP under the control of the
84 tyrosine hydroxylase (*Th*) locus[15]. We microdissected both MB and FB from E15.5 mice,
85 extending our analyses to MB, FB, and OB in P7 mice (Figure 1a). E15.5 and P7 time points
86 were chosen based on their representation of stable MB DA populations, either after neuron birth
87 (E15.5) or between periods of programmed cell death (P7) (Figure 1a)[16].
88
89 Quality control and outlier analysis identify 396 high quality cell transcriptomes to be used in
90 our analyses. We initially sequenced RNA from 473 single cells to an average depth of ~8 x 10^5
91 50 bp paired-end fragments per cell. Using Monocle 2, we converted normalized expression
92 estimates into estimates of RNA copies per cell[17]. Cells were filtered based on the distributions
93 of total mass, total number of mRNAs, and total number of expressed genes per cell (Figure 1 -
94 figure supplement 1a-1c; detailed in Methods). After QC, 410 out of 473 cells were retained.
95 Using principal component analysis (PCA) as part of the iterative analysis described below, we
96 identified and removed 14 outliers determined to be astrocytes, microglia, or oligodendrocytes

97 (Figure 1 - figure supplement 1e; Supplementary File 1), leaving 396 cells (~79 cells/timepoint-
98 region; Figure 1 - figure supplement 1d).
99
100 To confirm that our methods can discriminate between different populations of neurons, we first
101 explored differences between timepoints. Following a workflow similar to the recently described
102 "dpFeature" procedure[18], we identified genes with highly variable transcriptional profiles and
103 performed PCA. As anticipated, we observed that the greatest source of variation was between
104 developmental ages (Figure 1b). Genes associated with negative PC1 loadings (E15.5 cells) were
105 enriched for gene sets consistent with mitotically active neuronal, undifferentiated precursors
106 (Figure 1c). In contrast, genes associated with positive PC1 loadings (P7 cells) were enriched for
107 ontology terms associated with mature, post-mitotic neurons (Figure 1c). This initial analysis
108 establishes our capacity to discriminate among biological classes present in our data using PCA
109 as a foundation.
110
111 Further, we attempted to identify clusters of single cells between and within timepoints and
112 anatomical regions. In order to do this, we selected the PCs that described the most variance in
113 the data and used t-Stochastic Neighbor Embedding (t-SNE)[19] to further cluster cells in an
114 unsupervised manner (see Methods). Analysis of all cells revealed that the E15.5 cells from both
115 MB and FB cluster together (Figure 1d), supporting the notion that they are less differentiated.
116 By contrast, cells isolated at P7 mostly cluster by anatomical region, suggesting progressive
117 functional divergence with time (Figure 1d). We next applied this same scRNA-seq analysis
118 workflow (See Methods) in a recursive manner individually in all regions at both timepoints to
119 further explore heterogeneity. This revealed a total of 13 clusters (E15.5 FB.1-2, MB.1-2; P7

120 OB.1-3, FB.1-2, MB.1-4; Figure 1e), demonstrating the diversity of DA neuron subtypes and
121 providing a framework upon which to evaluate the biological context of genetic association
122 signals across closely-related cell types. Using known markers, we confirmed that all clusters
123 expressed high levels of pan-neuronal markers (*Snap25*, *Eno2*, and *Syt1*) (Figure 1 - figure
124 supplement 2a). In contrast, we observed scant evidence of astrocyte (*Aldh1l1*, *Slc1a3*, *Aqp4*, and
125 *Gfap*; Figure 1 - figure supplement 2a) or oligodendrocyte markers (*Mag*, *Mog*, and *Mbp*; Figure
126 1 - figure supplement 2a), thus confirming we successfully isolated our intended substrate, *Th*+
127 neurons.
128
129 *scRNA-seq revealed biologically and temporally discriminating transcriptional signatures*
130 With subpopulations of DA neurons defined in our data, we set out to assign a biological identity
131 to each cluster. Among the four clusters identified at E15.5, two were represented in t-SNE space
132 as a single large group that included cells from both MB and FB (E15.MB.1, E15.FB.1), leaving
133 two smaller clusters that were comprised solely of MB or FB cells (Figure 2 - figure supplement
134 1a). The latter MB cluster (E15.MB.2; Figure 2 - figure supplement 1a-1c) specifically expressed
135 *Foxa1*, *Lmx1a*, *Pitx3*, and *Nr4a2* and thus likely represents a post-mitotic DA neuron
136 population[20] (Supplementary File 2; Supplementary File 3). Similarly, the discrete E15.FB.2
137 cluster expressed markers of post-mitotic FB/hypothalamic neurons (Figure 2 - figure
138 supplement 1a-1b), including *Six3, Six3os1*, *Sst,* and *Npy* (Supplementary File 2; Supplementary
139 File 3). These embryonic data did not discriminate between cells populating known domains of
140 DA neurons, such as the SN.
141

142 By contrast, P7 cells mostly cluster by anatomical region and each region has defined subsets
143 (Figure 1d, 1e, 2a). Analysis of P7 FB revealed two distinct cell clusters (Figure 2b). Expression
144 of the neuropeptides *Gal* and *Ghrh* and the *Gsx1* transcription factor place P7.FB.1 cells in the
145 arcuate nucleus (Supplementary File 2; Supplementary File 3)[21–24]. The identity of P7.FB.2,
146 however, was less clear, although subsets of cells therein did express other arcuate nucleus
147 markers for $Th^{+}/Ghrh^{-}$ neuronal populations e.g. *Onecut2*, *Arx*, *Prlr*, *Slc6a3*, and *Sst* (Figure 2 -
148 figure supplement 1d; Supplementary File 3)[24]. All three identified OB clusters (Figure 2c)
149 express marker genes of OB DA neuronal development or survival (Supplementary File 2,
150 Supplementary File 3; Figure 2 - figure supplement 1e)[25]. It has previously been reported that
151 *Dcx* expression diminishes with neuronal maturation[26] and *Snap25* marks mature neurons[27]. We
152 observe that these OB clusters seem to reflect this continuum of maturation wherein expression
153 of *Dcx* diminishes and *Snap25* increases with progression from P7.OB1 to OB3 (Figure 2 -
154 figure supplement 1e). This pattern is mirrored by a concomitant increase in OB DA neuron fate
155 specification genes (Figure 2 - figure supplement 1e)[25,28]. In addition, we identified four P7 MB
156 DA subset clusters (Figure 2d). Marker gene analysis confirmed that three of the clusters
157 correspond to DA neurons from the VTA (*Otx2* and *Neurod6*; P7.MB.1)[29,30], the PAG (*Vip* and
158 *Pnoc*; P7.MB.3)[31,32], and the SN (*Sox6*, *Aldh1a7*, *Ndnf*, *Serpine2*, *Rbp4*, and *Fgf20*;
159 P7.MB.4)[29,33–35] (Supplementary File 2; Supplementary File 3). These data are consistent with
160 recent scRNA-seq studies of similar populations[34,36]. Through this marker gene analysis, we
161 successfully assigned a biological identity to 12/13 clusters.
162
163 The only cluster without a readily assigned identity was P7.MB.2. This population of P7 MB DA
164 neurons, P7.MB.2 (Figure 2d), is likely a progenitor-like population. Like the overlapping

165 E15.MB.1 and E15.FB.1 clusters (Figure 2 - figure supplement 1a), this cluster preferentially
166 expresses markers of neuronal precursors/differentiation/maturation (Supplementary File 2,
167 Supplementary File 3). In addition to sharing markers with the progenitor-like E15.MB.1 cluster,
168 P7.MB.2 exhibits gene expression consistent with embryonic mouse neuroblast populations[34],
169 cell division, and neuron development[37–41] (Supplementary File 2, Supplementary File 3).
170 Consistent with the hypothesis, this population displayed lower levels of both *Th* and *Slc6a3,*
171 markers of mature DA neurons, than the terminally differentiated and phenotypically discrete P7
172 MB DA neuron populations of the VTA, SN and PAG (Figure 2e).
173
174 With this hypothesis in mind, we sought to ascertain the spatial distribution of P7.MB.2 DA
175 neurons through multiplex, single molecule fluorescence *in situ* hybridization (smFISH) for *Th*
176 (pan-P7 MB DA neurons), *Slc6a3* (P7.MB.1, P7.MB.3, P7.MB.4), and one of the neuroblast
177 marker genes identified through our analysis, either *Lhx9* or *Ldb2* (P7.MB.2) (Figure 2e). In each
178 experiment, we scanned the ventral midbrain for cells that were Th+/Slc6a3- and positive for the
179 third gene. *Th*+/*Slc6a3*-/*Lhx9*+ cells were found scattered in the dorsal SN *pars compacta*
180 (SNpc) along with cells expressing *Lhx9* alone (Figure 2f, 2h). Expression of *Ldb2* was found to
181 have a similar pattern to *Lhx9*, with *Th*+/*Slc6a3*-/*Ldb2*+ cells found in the dorsal SNpc (Figure
182 2f, 2h). Expression of *Lhx9* and *Ldb2* was low or non-existent in *Th*+/*Slc6a3*+ cells in the SNpc
183 (Figure 2e, 2f). Importantly, cells expressing these markers express *Th* at lower levels than
184 *Th*+/*Slc6a3*+ neurons (Figure 2f, 2g), consistent with our scRNA-seq data (Figure 2e). Thus,
185 with the resolution of the spatial distribution of this novel neuroblast-like P7 MB DA population,
186 we assign biological identity to each defined brain DA subpopulation.
187

188 *Novel SN-specific transcriptional profiles and GRNs highlight its association with PD*

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

193

194 In order to look broadly at neuronal subtypes, we evaluated expression of canonical markers of
195 other neuronal subtypes in our *Th*+ neuron subpopulations. Interestingly, we observed
196 inconsistent detection of *Th* and eGFP in some E15.5 clusters (Figure 3 - figure supplement 1a).
197 This likely reflects lower *Th* transcript abundance at this developmental state, but sufficient
198 expression of the eGFP reporter to permit FACS collection (Figure 3 - figure supplement 1b).
199 The expression of other DA markers, *Ddc* and *Slc18a2*, mirror *Th* expression, while *Slc6a3*
200 expression is more spatially and temporally restricted (Figure 3 - figure supplement 1a). The SN
201 cluster displays robust expression of all canonical DA markers (Figure 3 - figure supplement 1a).
202 Multiple studies have demonstrated that *Th*+ neurons may also express markers characteristic of
203 other major neuronal subtypes[42–44]. We found that only the SN and PAG showed no expression
204 of either GABAergic (*Gad1*/*Gad2*/*Slc32a1*) or glutamatergic (*Slc17a6*) markers (Figure 3 -
205 figure supplement 1a). This neurotransmitter specificity is a potential avenue for exploring the
206 preferential vulnerability of the SN in PD.

207

208 Next, we postulated that genes whose expression defined the P7 SN DA neuron cluster might
209 illuminate their preferential vulnerability in PD. We identified 110 SN-specific genes, by first
210 finding all differentially expressed genes between P7 subset clusters and then using the Jensen-

211 Shannon distance to identify cluster specific genes (See Methods). Prior reports confirm the
212 expression of 49 of the 110 SN-specific genes (~45%) in postnatal SN (Supplementary File 4).
213 We then sought evidence to confirm or exclude SN expression for the remaining, novel 61 genes
214 (55%). Of these, 25/61 (~41%) were detected in adult SN neurons by *in situ* hybridization (ISH)
215 of coronal sections in adult (P56) mice (Allen Brain Atlas, ABA; http://developingmouse.brain-
216 map.org), including *Col25a1*, *Fam184a*, *Ankrd34b*, *Nwd2*, and *Cadps2* (Figure 3a,
217 Supplementary File 5). Only 4/61 genes, for which ISH data existed in the ABA, lacked clear
218 evidence of expression in the adult SN (Supplementary File 5). The ABA lacked coronal ISH
219 data on 32/61 genes, thus we were unable to confirm their presence in the SN. Collectively, we
220 identify 110 postnatal SN DA marker genes and confirm the expression of those genes in the
221 adult mouse SN for 74 (67%) of them, including 25 novel markers of this clinically relevant cell
222 population that we confirmed using the ABA image catalog.
223
224 We next asked whether we could identify significant relationships between cells defined as being
225 P7 SN DA neurons and distinctive transcriptional signatures in our data. We identify 16 co-
226 expressed gene modules by performing weighted gene co-expression network analysis
227 (WGCNA)[45,46] on all expressed genes of the P7 subset (Figure 3 - figure supplement 2;
228 Supplementary File 6). By calculating pairwise correlations between modules and P7 subset
229 clusters, we reveal that 7/16 modules are significantly and positively correlated (Bonferroni
230 corrected $p < 3.5e-04$) with at least one subset cluster (Figure 3c). We graphically represent the
231 eigenvalues for each module in each cell in P7 t-SNE space, confirming that a majority of these
232 significant modules (6/7) displayed robust spatial, isotype enrichment (Figure 3d).
233

234 In order identify the biological relevance of these modules, each module was tested for
235 enrichment for Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, Gene Ontology
236 (GO) gene sets, and Reactome gene sets. Two modules, the "brown" and "green" modules, were
237 significantly associated with the Parkinson's Disease KEGG pathway gene set (Figure 3c;
238 Supplementary File 7). Interestingly, the "brown" module was also significantly correlated with
239 the P7 VTA population (P7.MB.1) and enriched for addiction gene sets (Supplementary File 7)
240 highlighting the link between VTA DA neurons and addiction[47]. Strikingly, only the P7 SN
241 cluster was significantly correlated with both PD-enriched modules (Figure 3c). This specific
242 correlation suggests these gene modules may play a role in the preferential susceptibility of the
243 SN in PD.
244
245 *Integrating SN DA neuron specific data enables prioritization of genes within PD-associated*
246 *intervals*
247 With these context-specific data in hand, we posited that SN DA neuron-specific genes and the
248 broader gene co-expression networks that correlate with SN DA neurons might be used to
249 prioritize genes within loci identified in PD GWAS. Such a strategy would be agnostic to prior
250 biological evidence and independent of genic position relative to the lead SNP, the traditional
251 method used to prioritize causative genes.
252
253 To investigate pertinent genes within PD GWAS loci, we identified all human genes within
254 topologically associated domains (TADs) and a two megabase interval encompassing each PD-
255 associated lead SNP. TADs were chosen because regulatory DNA impacted by GWAS variation
256 is more likely to act on genes within their own TAD[48]. While topological data does not exist for

257 SN DA neurons, we use TAD boundaries from hESCs as a proxy, as TADs are generally
258 conserved across cell types[49]. To improve our analyses, we also selected +/- 1 megabase interval
259 around each lead SNP thus including the upper bounds of reported enhancer-promoter
260 interactions[50,51]. All PD GWAS SNPs interrogated were identified by the most recent meta-
261 analyses (49 SNPs in total)[11,12], implicating a total of 1751 unique genes. We then identified
262 corresponding one-to-one mouse to human homologs (1009/1751; ~58%), primarily through the
263 Mouse Genome Informatics (MGI) homology database.
264
265 To prioritize these genes in GWAS loci, we developed a gene-centric score that integrates our
266 data as well as data in the public domain. We began by intersecting the PD loci genes with our
267 scRNA-seq data as well as previously published SN DA expression data[34], identifying 430 genes
268 (430/1009; ~43%) with direct evidence of expression in SN DA neurons in at least one dataset.
269 Each PD-associated interval contained ≥1 SN-expressed gene (Supplementary File 8).
270 Emphasizing the need for a novel, systematic strategy, in 19/49 GWA intervals (~39%), the most
271 proximal gene to the lead SNP was not detectably expressed in mouse SN DA neuron
272 populations (Supplementary File 8; Supplementary File 9). Surprisingly, three loci contained
273 only one SN DA-expressed gene: *Mmp16* (*MMP16* locus, Figure 4a), *Tsnax* (*SIPA1L2* locus),
274 and *Satb1* (rs4073221 locus). The relevance of these candidate genes to neuronal
275 function/dysfunction is well supported[52–56]. This establishes gene expression in a relevant tissue
276 as a powerful tool in the identification of causal genes.
277
278 In order to prioritize likely diseases-associated genes in the remaining 46 loci, we scored genes
279 on three criteria: whether genes were identified as specific markers for the P7.MB.4 (SN) cluster

280 (Supplementary File 2), whether the genes were differentially expressed between all P7 DA
281 neuron populations, and whether the genes were included in PD gene set enriched and SN
282 correlated gene modules uncovered in WGCNA (Supplementary File 6). This strategy facilitated
283 further prioritization of a single gene in 22 additional loci including *SNCA*, *LRRK2*, and *GCH1*
284 loci (Figure 4a; Table 1; Supplementary File 9). Importantly, using this approach we indict the
285 familial PD gene encoding alpha-synuclein (*SNCA*), as responsible for the observed PD
286 association with rs356182 (Figure 4a, Table 1, Supplementary File 9). Thus, by using context-
287 specific data alone, we were able to prioritize a single candidate gene in roughly half (~49%) of
288 PD-GWAS associated loci.
289
290 Furthermore, at loci in which a single gene did not emerge, we identified dosage sensitive genes
291 by considering the probability of being loss-of-function (LoF) intolerant (pLI) metric from the
292 ExAC database[57,58]. Since most GWAS variation is predicted to impact regulatory DNA and in
293 turn impact gene expression, it follows that genes in GWAS loci that are more sensitive to
294 dosage levels may be more likely to be candidate genes. With that in mind, the pLI for each gene
295 was used to further "rank" the genes within loci where a single gene was not prioritized. For
296 those loci, including *MAPT* and *DDRGK1* loci (Figure 4a), we report a group of top scoring
297 candidate genes (Table 1, Supplementary File 9). Expression of prioritized genes in the adult SN
298 adds to the validity of the genes identified as possible candidates (Figure 4b).
299
300 Two interesting examples that emerge from this scoring are found at the *MAPT* and *TMEM175-*
301 *GAK-DGKQ* loci. Although *MAPT* has previously been implicated in multiple neurodegenerative
302 phenotypes, including PD (OMIM: 168600), we instead prioritize two genes before it (*CRHR1*

303 and *NSF*; Table 1). We detect *Mapt* and *Nsf* expression consistently across all assayed DA
304 neurons (Figure 4c). By contrast, expression of *Crhr1,* encoding the corticotropin releasing
305 hormone receptor 1*,* is restricted to P7 DA neurons in the SN and the more mature OB neuronal
306 populations (Figure 4c). Similarly, at the *TMEM175-GAK-DGKQ* locus, our data shows that
307 although all three proximal genes are expressed in the SN, the adjacent *CPLX1* was one of the
308 prioritized genes (Table 1, Supplementary File 9).
309
310 There are multiple lines of evidence that strengthen *CPLX1* as a candidate gene. Expression of
311 CPLX1 is elevated both in the brains of PD patients and the brains of mice overexpressing the
312 *SNCA* A53T PD mutation[59,60]. Additionally, mice deficient in *CPLX1* display an early-onset,
313 cerebellar ataxia along with prolonged motor and behavioral phenotypes[61,62]. However, the
314 impact of *Cplx1* deficiency on the integrity of the nigrostriatal pathway, to date, has not been
315 explored. In order to confirm *CPLX1* as a candidate gene, we performed immunohistochemistry
316 (IHC) for *Th* in the *Cplx1* knockout mouse model (Supplementary File 10, Supplementary File
317 11)[61–63]. We measured the density of *Th*+ innervation in the striatum of *Cplx1* -/- mice and
318 controls (Figure 4d, Supplementary File 11) and found that *Cplx1* -/- mice had significantly
319 lower *Th*+ staining in the striatum (p-value = 3.385e-08; Figure 4e). This indicates that *Cplx1*
320 KO mice have less *Th*+ fiber innervation and a compromised nigrostriatal pathway, supporting
321 its biological significance in MB DA populations and to PD.
322
323 The systematic identification of causal genes underlying GWAS signals is essential in order for
324 the scientific and medical communities to take full advantage of all the GWAS data published
325 over the last decade. Taken collectively, we demonstrate how scRNA-seq data from disease-

326 relevant populations can be leveraged to illuminate GWAS results, facilitate systematic
327 prioritization of GWAS loci implicated in PD, and can leads to the functional characterization of
328 previously underexplored candidate genes.
329
330 **DISCUSSION**
331 Midbrain DA neurons in the SN have been the subject of intense research since being
332 definitively linked to PD nearly 100 years ago[64]. While degeneration of SN DA neurons in PD is
333 well established, they represent only a subset of brain DA populations. It remains unknown why
334 nigral DA neurons are particularly vulnerable. We set out to explore this question using scRNA-
335 seq. Recently, others have used scRNA-seq to characterize the mouse MB, including DA
336 neurons[34]. Here, we extend these data significantly, extensively characterizing the transcriptomes
337 of multiple brain DA populations longitudinally and discovering GRNs associated with specific
338 populations.
339
340 Most importantly, our data facilitate the iterative and biologically informed prioritization of gene
341 candidates for all PD-associated genomic intervals. In practice, the gene closest to the lead SNP
342 identified within a GWAS locus is frequently treated as the prime candidate gene, often without
343 considering tissue-dependent context. Our study overcomes this by integrating genomic data
344 derived from specific cell contexts with analyses that are agnostic to one another. We posit that
345 genes pertinent to PD are likely expressed within SN DA neurons. This hypothesis is consistent
346 with the recent description of the "omnigenic" nature of common disease, wherein variation
347 impacting genes expressed in a disease tissue explain the vast majority of risk[7].
348

349 First, we identify intervals that reveal one primary candidate, i.e. those that harbor only one SN-
350 expressed gene. Next, we examine those intervals with many candidates, and prioritize based on
351 a cumulative body of biological evidence. In total, we prioritize 5 or fewer candidates in 47/49 (
352 ~96%) PD GWAS loci studied, identifying a single gene in twenty-four loci (24/49; ~49%) and
353 three or fewer genes in ~84% of loci (41/49). Ultimately this prioritization reduces the candidate
354 gene list for PD GWAS loci dramatically from 1751 genes to 111 genes.
355
356 The top genes we identify in three PD loci (*SNCA*, *FGF20*, *GCH1*) have been directly associated
357 with PD, MB DA development, and MB DA function[35] (OMIM: 163890, 128230). Furthermore,
358 our prioritization of *CPLX1* over other candidates in the *TMEM175-GAK-DGKQ* locus is
359 supported by multiple lines of evidence. Additionally, we demonstrate that the integrity of the
360 nigrostriatal pathway is disrupted in *Cplx1* knockout mice. Dysregulation of *CPLX1* RNA is also
361 a biomarker in individuals with pre-PD prodromal phenotypes harboring the *PARK4* mutation
362 (*SNCA* gene duplication)[65]. These results validate our approach and strengthen the argument for
363 the use of context specific data in pinpointing candidate genes in GWAS loci.
364
365 Many of the genes prioritized (Table 1) have been shown to have various mitochondrial
366 functions[66–72]. The identification of genes associated with mitochondrial functions is especially
367 interesting in light of the "omnigenic" hypothesis of complex traits[7]. Since mitochondrial
368 dysfunction has been extensively implicated in PD[73], the prioritized genes may represent "core"
369 genes that in turn can affect the larger mitochondrial-associated regulatory networks active in the
370 disease relevant cell-type (SN DA neurons). It is notable that one of these genes is the presenilin
371 associated rhomboid like gene or *PARL*. *PARL* cleaves *PINK1*, a gene extensively implicated in

372 PD pathology and recently a variant in *PARL* has been associated with early-onset PD (OMIM:
373 607858)[74–76].
374
375 While our method successfully prioritized one familial PD gene (*SNCA*), we do not prioritize
376 *LRRK2,* another familial PD gene harbored within a PD GWAS locus. *Lrrk2* is not prioritized
377 simply because it is not detectably expressed in our SN DA neuronal population. This is
378 expected as numerous studies have reported little to no *Lrrk2* expression in *Th*+ MB DA neurons
379 both in mice and humans[77,78]. Instead, our method prioritizes *PDZRN4*. This result does not
380 necessarily argue against the potential relevance of *LRRK2* but instead provides an additional
381 candidate that may contribute to PD susceptibility. The same logic should be noted for two other
382 PD-associated loci, wherein our scoring prioritizes different genes (*KCNN3* and *CRHR1*/*NSF*,
383 respectively) than one previously implicated in PD (*GBA* and *MAPT*) (OMIM: 168600). Notably,
384 *KCNN3*, *CRHR1*, and *NSF*, all have previous biological evidence making them plausible
385 candidates[79–81].
386
387 Studying disease-relevant tissue has proven to be essential for elucidating the genetic
388 architecture underlying GWA signals[2]; our scoring method relies upon data from the most
389 relevant cell-type to PD, SN DA neurons. While this study was under consideration for
390 publication, Chang and colleagues[12] endeavored to prioritize PD GWAS loci using publically
391 available data. Although their pipeline strives to be "neuro-centric," it is not predicated on the
392 biological relevance of candidates to SN DA neurons.
393

394 Through comparison of the two scoring paradigms, the methods agree on at least one gene in
395 17/44 (~39%) jointly scored loci, including *SNCA* (Supplementary File 12), bolstering the
396 evidence for those candidate genes. However, we see ~44% (31/71) of the genes prioritized by
397 Chang, *et al*, are not expressed in either of the SN DA expression data sets used in our scoring
398 scheme (Supplementary File 12), including *LRRK2* (addressed above). One prime example of
399 this discrepancy is the *MCCC1* locus. Chang, *et al*, identify the *MCCC1* gene to be the prime
400 candidate gene in the locus. However, we find that *MCCC1* is not expressed in SN DA neurons
401 (Supplementary File 8). Instead, we prioritize *PARL*, a gene with an established role in PD
402 pathogenesis[74–76].
403
404 Our focus on disease relevant cell-type data also leads us to identify genes previously implicated
405 in neurodegeneration, which make obvious candidates. For example, in the *TMEM175-GAK-*
406 *DGKQ* locus, we identify *CPLX1* and functionally confirm its relevance. We also identify *ATRN*
407 (attractin) as one of the candidate genes in the *DDRGK1* locus. Loss of *Atrn* has been shown to
408 cause age-related neurodegeneration of SN DA neurons in rats[82,83], making it an ideal candidate
409 in the *DDRGK1* locus. Neither gene is identified using other metrics[12] (Supplementary File 12).
410
411 Despite this success, we acknowledge several notable caveats. First, not all genes in PD-
412 associated human loci have identified mouse homologs. Thus, it remains possible that we may
413 have overlooked the contribution of some genes whose biology is not comprehensively queried
414 in this study. Secondly, we assume that identified genetic variation acts in a manner that is at
415 least preferential, if not exclusive, to SN DA neurons. Lastly, by prioritizing SN-expressed
416 genes, we assume that PD variation affects genes whose expression in the SN does not require

417 insult/stress. These caveats notwithstanding, our strategy sets the stage for a new generation of
418 independent and combinatorial functional evaluation of gene candidates for PD-associated
419 genomic intervals.

420 **MATERIALS AND METHODS**

421

422 **Data availability**

423 Raw data will be made available on Sequence Read Archive (SRA) and Gene Expression

424 Omnibus (GEO) prior to publication. Summary data is available where code is available below

425 (https://github.com/pwh124/DA_scRNA-seq).

426

427 **Code Availability**

428 Code for analysis, for the production of figures, and summary data is deposited at

429 https://github.com/pwh124/DA_scRNA-seq

430

431 **Animals.**

432 The Th:EGFP BAC transgenic mice (Tg(Th-EGFP)DJ76Gsat/Mmnc) used in this study were

433 generated by the GENSAT Project and were purchased through the Mutant Mouse Resource &

434 Research Centers (MMRRC) Repository (https://www.mmrrc.org/). Mice were maintained on a

435 Swiss Webster (SW) background with female SW mice obtained from Charles River

436 Laboratories (http://www.criver.com/). The Tg(Th-EGFP)DJ76Gsat/Mmnc line was primarily

437 maintained through matings between Th:EGFP positive, hemizygous male mice and wild-type

438 SW females (dams). Timed matings for cell isolation were similarly established between

439 hemizygous male mice and wild-type SW females. The observation of a vaginal plug was

440 defined as embryonic day 0.5 (E0.5). All work involving mice (husbandry, colony maintenance

441 and euthanasia) were reviewed and pre-approved by the institutional care and use committee.

442

443 *Cplx1* knockout mice and wild type littermates used for immunocytochemistry were taken from a
444 colony established in Cambridge using founder mice that were a kind gift of Drs K. Reim and N.
445 Brose (Gottingen, Germany). *Cplx1* mice in this colony have been backcrossed onto a C57/Bl6J
446 inbred background for at least 10 generations. All experimental procedures were licensed and
447 undertaken in accordance with the regulations of the UK Animals (Scientific Procedures) Act
448 1986. Housing, rearing and genotyping of mice has been described in detail previously[61,62].
449 Mice were housed in hard-bottomed polypropylene experimental cages in groups of 5-10 mice in
450 a housing facility was maintained at 21 – 23°C with relative humidity of 55 ± 10%. Mice had *ad*
451 *libitum* access to water and standard dry chow. Because homozygous knockout *Cplx1* mice have
452 ataxia, they have difficulty in reaching the hard pellets in the food hopper and drinking from the
453 water bottles. Lowered waterspouts were provided and access to normal laboratory chow was
454 improved by providing mash (made by soaking 100 g of chow pellets in 230 ml water for 60 min
455 until the pellets were soft and fully expanded) on the floor of the cage twice daily. *Cplx1*
456 genotyping to identify mice with a homozygous or heterozygous deletion of the *Cplx1* gene was
457 conducted as previously described[61], using DNA prepared from tail biopsies.

458 **Dissection of E15.5 brains.**

459 At 15.5 days after the timed mating, pregnant dams were euthanized and the entire litter of
460 embryonic day 15.5 (E15.5) embryos were dissected out of the mother and immediately placed
461 in chilled Eagle's Minimum Essential Media (EMEM). Individual embryos were then
462 decapitated and heads were placed in fresh EMEM on ice. Embryonic brains were then removed
463 and placed in Hank's Balanced Salt Solution (HBSS) without Mg^{2+} and Ca^{2+} and manipulated
464 while on ice. The brains were immediately observed under a fluorescent stereomicroscope and
465 $EGFP^{+}$ brains were selected. $EGFP^{+}$ regions of interest in the forebrain (hypothalamus) and the

466 midbrain were then dissected and placed in HBSS on ice. This process was repeated for each
467 $EGFP^+$ brain. Four $EGFP^+$ brain regions for each region studied were pooled together for
468 dissociation.
469
470 **Dissection of P7 brains.**
471 After matings, pregnant females were sorted into their own cages and checked daily for newly
472 born pups. The morning the pups were born was considered day P0. Once the mice were aged to
473 P7, all the mice from the litter were euthanized and the brains were then quickly dissected out of
474 the mice and placed in HBSS without Mg^{2+} and Ca^{2+} on ice. As before, the brains were then
475 observed under a fluorescent microscope, $EGFP^+$ status for P7 mice was determined, and $EGFP^+$
476 brains were retained. For each $EGFP^+$ brain, the entire olfactory bulb was first resected and
477 placed in HBSS on ice. Immediately thereafter, the $EGFP^+$ forebrain and midbrain regions for
478 each brain were resected and also placed in distinct containers of HBSS on ice. Five $EGFP^+$
479 brain regions for each region were pooled together for dissociation.
480
481 **Generation of single cell suspensions from brain tissue.**
482 Resected brain tissues were dissociated using papain (Papain Dissociation System, Worthington
483 Biochemical Corporation; Cat#: LK003150) following the trehalose-enhanced protocol reported
484 by Saxena, et. al, 2012[84] with the following modifications: The dissociation was carried out at
485 37^oC in a sterile tissue culture cabinet. During dissociation, all tissues at all time points were
486 triturated every 10 minutes using a sterile Pasteur pipette. For E15.5 tissues, this was continued
487 for no more than 40 minutes. For P7, this was continued for up to 1.5 hours or until the tissue
488 appeared to be completely dissociated.

489

490 Additionally, for P7 tissues, after dissociation but before cell sorting, the cell pellets were passed
491 through a discontinuous density gradient in order to remove cell debris that could impede cell
492 sorting. This gradient was adapted from the Worthington Papain Dissociation System kit.
493 Briefly, after completion of dissociation according to the Saxena protocol[84], the final cell pellet
494 was resuspended in DNase dilute albumin-inhibitor solution, layered on top of 5 mL of albumin-
495 inhibitor solution, and centrifuged at 70g for 6 minutes. The supernatant was then removed.

496

497 **FACS and single-cell collection.**

498 For each timepoint-region condition, pellets were resuspended in 200 μL of media without serum
499 comprised of DMEM/F12 without phenol red, 5% trehalose (w/v), 25 μM AP-V, 100 μM
500 kynurenic acid, and 10 μL of 40 U/μl RNase inhibitor (RNasin® Plus RNase Inhibitor, Promega)
501 at room temperature. The resuspended cells were then passed through a 40 uM filter and
502 introduced into a Fluorescence Assisted Cell Sorting (FACS) machine (Beckman Coulter MoFlo
503 Cell Sorter or Becton Dickinson FACSJazz). Viable cells were identified via propidium iodide
504 staining, and individual neurons were sorted based on their fluorescence (EGFP+ intensity, See
505 Figure 2 - supplement 2c) directly into lysis buffer in individual wells of 96-well plates for
506 single-cell sequencing (2 μL Smart-Seq2 lysis buffer + RNAase inhibitor, 1 μL oligo-dT primer,
507 and 1 μL dNTPs according to Picelli et al., 2014[85]. Blank wells were used as negative controls
508 for each plate collected. Upon completion of a sort, the plates were briefly spun in a tabletop
509 microcentrifuge and snap-frozen on dry ice. Single cell lysates were subsequently kept at -80°C
510 until cDNA conversion.

511

512 **Single-cell RT, library prep, and sequencing.**

513 Library preparation and amplification of single-cell samples were performed using a modified
514 version of the Smart-Seq2 protocol[85]. Briefly, 96-well plates of single cell lysates were thawed to
515 4°C, heated to 72°C for 3 minutes, then immediately placed on ice. Template switching first-
516 strand cDNA synthesis was performed as described above using a 5'-biotinylated TSO oligo.
517 cDNAs were amplified using 20 cycles of KAPA HiFi PCR and 5'-biotinylated ISPCR primer.
518 Amplified cDNA was cleaned with a 1:1 ratio of Ampure XP beads and approximately 200 pg
519 was used for a one-quarter standard sized Nextera XT tagmentation reaction. Tagmented
520 fragments were amplified for 14 cycles and dual indexes were added to each well to uniquely
521 label each library. Concentrations were assessed with Quant-iT PicoGreen dsDNA Reagent
522 (Invitrogen) and samples were diluted to ~2 nM and pooled. Pooled libraries were sequenced on
523 the Illumina HiSeq 2500 platform to a target mean depth of ~8.0×10^5 50bp paired-end
524 fragments per cell at the Hopkins Genetics Research Core Facility.

525

526 **RNA sequencing and alignment.**

527 For all libraries, paired-end reads were aligned to the mouse reference genome (mm10)
528 supplemented with the Th-EGFP$^+$ transgene contig, using HISAT2[86] with default parameters
529 except: -p 8. Aligned reads from individual samples were quantified against a reference
530 transcriptome (GENCODE vM8)[87] supplemented with the addition of the eGFP transcript.
531 Quantification was performed using cuffquant with default parameters and the following
532 additional arguments: --no-update-check –p 8. Normalized expression estimates across all
533 samples were obtained using cuffnorm[88] with default parameters.

534

535 **Single-cell RNA data analysis.**

536 *Expression estimates.*

537 Gene-level and isoform-level FPKM (Fragments Per Kilobase of transcript per Million) values
538 produced by cuffquant[88] and the normalized FPKM matrix from cuffnorm was used as input for
539 the Monocle 2 single cell RNA-seq framework[89] in R/Bioconductor[90]. Genes were annotated
540 using the Gencode vM8 release[87]. A CellDataSet was then created using Monocle (v2.2.0)[89]
541 containing the gene FPKM table, gene annotations, and all available metadata for the sorted
542 cells. All cells labeled as negative controls and empty wells were removed from the data.
543 Relative FPKM values for each cell were converted to estimates of absolute mRNA counts per
544 cell (RPC) using the Monocle 2 Census algorithm[17] using the Monocle function "relative2abs."
545 After RPCs were inferred, a new cds was created using the estimated RNA copy numbers with
546 the expression Family set to "negbinomial.size()" and a lower detection limit of 0.1 RPC.

547

548 *QC Filtering.*

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

552

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

557

558 ii. Cell Mass - Cells were then filtered based on the total mass of RNA in the cells calculated by
559 Monocle. Again, the total mass of the cell was plotted and mass thresholds were set based on
560 observations from each distribution. Only those cells with a total cell mass between 100,000 and
561 1,300,000 fragments mapped were retained.
562
563 iii. Total RNA copies per cell - Cells were then filtered based on the total number of RNA
564 transcripts estimated for each cell. Again, the total RNA copies per cell was plotted and RNA
565 transcript thresholds were set based on observations from each distribution. Only those cells with
566 a total mRNA count between 1,000 and 40,000 RPCs were retained.
567
568 A total of 410 individual cells passed these initial filters. Outliers found in subsequent, reiterative
569 analyses described below were analyzed and removed resulting a final cell number of 396. The
570 distributions for total mRNAs, total mass, and number of expressed, can be found in Figure 1 -
571 supplement 1a-1c.
572
573 *Log distribution QC.*
574 Analysis using Monocle relies on the assumption that the expression data being analyzed follows
575 a log-normal distribution. Comparison to this distribution was performed after initial filtering
576 prior to continuing with analysis and was observed to be well fit.
577
578 **Reiterative single-cell RNA data analysis.**

579 After initial filtering described above, the entire cds as well as subsets of the cds based on "age"
580 and "region" of cells were created for recursive analysis. Regardless of how the data was
581 subdivided, all data followed a similar downstream analysis workflow.
582
583 *Determining number of cells expressing each gene.*
584 The genes to be analyzed for each iteration were filtered based on the number of cells that
585 expressed each gene. Genes were retained if they were expressed in > 5% of the cells in the
586 dataset being analyzed. These are termed "expressed_genes." For example, when analyzing all
587 cells collected together (n = 410), a gene had to be expressed in 20.5 cells (410 x 0.05 = 20.5) to
588 be included in the analysis. Whereas when analyzing P7 MB cells (n = 80), a gene had to be
589 expressed in just 4 cells (80 x 0.05 = 4). This was done to include genes that may define rare
590 populations of cells that could be present in any given population.
591
592 *Monocle model preparation.*
593 The data was prepared for Monocle analysis by retaining only the expressed genes that passed
594 the filtering described above. Size factors were estimated using Monocle's
595 "estimateSizeFactors()" function. Dispersions were estimated using the "estimateDispersions()"
596 function.
597
598 *High variance gene selection.*
599 Genes that have a high biological coefficient of variation (BCV) were identified by first
600 calculating the BCV by dividing the standard deviation of expression for each expressed gene by
601 the mean expression of each expressed gene. A dispersion table was then extracted using the

602 dispersionTable() function from Monocle. Genes with a mean expression > 0.5 transcripts and a
603 "dispersion_empirical" >= 1.5*dispersion_fit or 2.0*dispersion_fit were identified as "high
604 variance genes."
605
606 *Principal component analysis (PCA).*
607 PCA was then run using the R "prcomp" function on the centered and scaled log2 expression
608 values of the "high variance genes." PC1 and PC2 were then visualized to scan the data for
609 obvious outliers as well as bias in the PCs for age, region, or plates on which the cells were
610 sequenced. If any visual outliers in the data was observed, those cells were removed from the
611 original subsetted cds and all filtering steps above were repeated. Once there were no obvious
612 visual outliers in PC1 or PC2, a screeplot was used plot the PCA results in order to determine the
613 number of PCs that contributed most significantly to the variation in the data. This was manually
614 determined by inspecting the screeplot and including only those PCs that occur before the
615 leveling-off of the plot.
616
617 *t-SNE and clustering.*
618 Once the number of significant PCs was determined, t-Distributed Stochastic Neighbor
619 Embedding (t-SNE)[19] was used to embed the significant PC dimensions in a 2-D space for
620 visualization. This was done using the "tsne" package available through R with "whiten =
621 FALSE." The parameters "perplexity" and "max_iter" were tested with various values and set
622 according what was deemed to give the cleanest clustering of the data.
623

624 After dimensionality reduction via t-SNE, the number of clusters was determined in an unbiased
625 manner by fitting multiple Gaussian distributions over the 2D t-SNE projection coordinates using
626 the R package ADPclust[91] and the t-SNE plots were visualized using a custom R script. The
627 number of genes expressed and the total mRNAs in each cluster were then compared.
628
629 **Differential expression Analyses.**
630 Since the greatest source of variation in the data was between ages (Figure 1), differential
631 expression analyses and downstream analyses were performed separately for each age.
632
633 In order to find differentially expressed genes between brain DA populations at each age, the
634 E15.5 and P7 datasets were annotated with regional cluster identity ("subset cluster").
635 Differential expression analysis was performed using the "differentialGeneTest" function from
636 Monocle that uses a likelihood ratio test to compare a vector generalized additive model
637 (VGAM) using a negative binomial family function to a reduced model in which one parameter
638 of interest has been removed. In practice, the following models were fit:
639
640 "~subset.cluster" for E15.5 or P7 dataset
641
642 Genes were called as significantly differentially expressed if they had a q-value (Benjamini-
643 Hochberg corrected p-value) < 0.05.
644
645 **Cluster specific marker genes.**

646 In order to identify differentially expressed genes that were "specifically" expressed in a
647 particular subset cluster, R code calculating the Jensen-Shannon based specificity score from the
648 R package cummeRbund[92] was used similar to what was described in Burns *et al*[93].
649
650 Briefly, the mean RPC within each cluster for each expressed gene as well as the percentage of
651 cells within each cluster that express each gene at a level > 1 transcript were calculated. The
652 ".specificity" function from the cummRbund package was then used to calculate and identify the
653 cluster with maximum specificity of each gene's expression. Details of this specificity metric can
654 be found in Molyneaux, *et al*[94].
655
656 To identify subset cluster specific genes, the distribution of specificity scores for each subset
657 cluster was plotted and a specificity cutoff was chosen so that only the "long right tail" of each
658 distribution was included (i.e. genes with a specificity score above the cutoff chosen). For each
659 iterative analysis, the same cutoff was used for each cluster or region (specificity $\geq$0.4). Once the
660 specificity cutoff was chosen, genes were further filtered by only retaining genes that were
661 expressed in >= 40% of cells within the subset cluster that the gene was determined to be
662 specific for.
663
664 **Gene Set Enrichment Analyses.**
665 Gene set enrichment analyses were performed in two separate ways depending upon the
666 situation. A Gene Set Enrichment Analysis (GSEA) PreRanked analysis was performed when a
667 ranked list (e.g. genes ranked by PC1 loadings) using GSEA software available from the Broad
668 Institute (v2.2.4)[95,96]. Ranked gene lists were uploaded to the GSEA software and a

669 “GSEAPreRanked” analysis was performed with the following settings: ‘Number of
670 Permutations’ = 1000, ‘Collapse dataset to gene symbols’ = true, ‘Chip platform(s)’ =
671 GENE_SYMBOL.chip, and ‘Enrichment statistic’ = weighted. Analysis was performed against
672 Gene Ontology (GO) collections from MSigDB, including c2.all.v5.2.symbols and
673 c5.all.v5.2.symbols. Top ten gene sets were reported for each analysis (Supplementary File 1).
674 Figures and tables displaying the results were produced using custom R scripts.
675
676 Unranked GSEA analyses for lists of genes was performed using hypergeometric tests from the
677 R package clusterProfiler implemented through the functions ‘enrichGO’, ‘enrichKEGG’, and
678 ‘enrichPathway’ with ‘pvalueCutoff’ set at 0.01, 0.1, 0.1, respectively with default settings[97].
679 These functions were implemented through the ‘compareCluster’ function when analyzing
680 WGCNA data.
681
682 **Weighted Gene Co-Expression Network Analysis (WGCNA).**
683 WGCNA was performed in R using the WGCNA package (v1.51)[45,46] following established
684 pipelines laid out by the packages authors (see
685 https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/ for more
686 detail). Briefly, an expression matrix for all P7 neurons containing all genes expressed in >= 20
687 cells (n = 12628) was used with expression counts in log2(Transcripts + 1). The data were
688 initially clustered in order to identify and remove outliers (n = 1) to leave 223 total cells (Figure
689 3 - supplement 1a). The soft threshold (power) for WGCNA was then determined by calculating
690 the scale free topology model fit for a range of powers (1:10, 12, 14, 16, 18, 20) using the
691 WGCNA function “pickSoftThreshold()” setting the networkType = “signed”. A power of 10

692 was then chosen based on the leveling-off of the resulting scale independence plot above 0.8
693 (Figure 3 - supplement 1b). Network adjacency was then calculated using the WGCNA function
694 "adjacency()" with the following settings: power = 10 and type = "signed." Adjacency
695 calculations were used to then calculate topological overlap using the WGCNA function
696 "TOMsimilarity()" with the following settings: TOMtype = "signed." Distance was then
697 calculated by subtracting the topological overlap from 1. Hierarchical clustering was then
698 performed on the distance matrix and modules were identified using the "cuttreeDynamic"
699 function from the dynamicTreeCut package[98] with the following settings: deepSplit = T;
700 pamRespectsDendro = FALSE, and minClusterSize = 20. This analysis initially identified 18
701 modules. Eigengenes for each module were then calculated using the "moduleEigengenes()"
702 function and each module was assigned a color. Two modules ("grey" and "turquoise") were
703 removed at this point. Turquoise was removed because it contained 11567 genes or all the genes
704 that could not be grouped with another module. Grey was removed because it only contained 4
705 genes, falling below the minimum set module size of 20. The remaining 16 modules were
706 clustered (Figure 3 - supplement 1c) and the correlation between module eigengenes and subset
707 cluster identity was calculated using custom R scripts. Significance of correlation was
708 determined by calculated the Student asymptotic p-value for correlations by using the WGCNA
709 "corPvalueStudent()" function. Gene set enrichments for modules were determined by using the
710 clusterProfiler R package[97]. The correlation between the t-SNE position of a cell and the module
711 eigengenes was calculated using custom R scripts.
712
713 **Prioritizing Genes in PD GWAS Loci.**
714 *Topologically Associated Domain (TAD) and Megabase Gene Data.*

715 The data for human TAD boundaries were obtained from human embryonic stem cell (hESC)
716 Hi-C data[49] and converted from human genome hg18 to hg38 using the liftOver tool from UCSC
717 Genome Browser (http://genome.ucsc.edu/). PD GWAS SNP locations in hg38 were intersected
718 with the TAD information to identify TADs containing a PD GWAS SNP. The data for +/- 1
719 megabase regions surrounding PD GWAS SNPs was obtained by taking PD GWAS SNP
720 locations in hg38 and adding or subtracting 1e+06 from each location. All hg38 Ensembl
721 (version 87) genes that fell within the TADs or megabase regions were then identified by using
722 the biomaRt R package[99,100]. All genes were then annotated with PD locus and SNP information.
723 Mouse homologs for all genes were identified using human to mouse homology data from
724 Mouse Genome Informatics (MGI)
725 (http://www.informatics.jax.org/downloads/reports/HOM_MouseHumanSequence.rpt; Date
726 accessed: 07/07/2017). Homologs of protein coding genes that did not have a mouse homolog in
727 the data above were manually curated by searching the human gene name in the MGI database
728 (http://www.informatics.jax.org/). Of the 742 genes with no mouse homologs, 92 (92/742,
729 ~12%) were annotated as protein coding genes (Figure 4 - supplement 1a). 24 loci include at
730 least one protein coding gene with no identified, one-to-one mouse homolog (Figure 4 -
731 supplement 1b). All 1009 genes with mouse homologs are annotated as "protein_coding." Gene
732 homologs were manually annotated using the MGI database if a homolog was found to exist. The
733 TAD and megabase tables were then combined to create a final PD GWAS locus-gene table.
734
735 *PD GWAS Loci Gene Scoring.*
736 Genes within PD GWAS loci were initially scored using two gene lists: Genes with an average
737 expression ≥0.5 transcripts in the SN cluster in our data (points = 1) and genes with an average

738 expression ≥0.5 transcripts in the SN population in La Manno, *et al*[34] (points = 1). Further
739 prioritization was accomplished by using three gene lists: genes that were differentially
740 expressed between subset clusters (points = 1); Genes found to be "specifically" expressed in the
741 P7 MB SN cluster (points = 1); Genes found in the WGCNA modules that are enriched for PD
742 (points = 1). Expression in the SN cluster was considered the most important feature and was
743 weighted as such through the use of two complementary datasets with genes found to be
744 expressed in both receiving priority. Furthermore, a piece of external data, pLI scores for each
745 gene from the ExAC database[58], was added to the scores in order to rank loci that were left with
746 ≥2 genes in the loci after the initial scoring. pLI scores
747 (fordist_cleaned_exac_r03_march16_z_pli_rec_null_data.txt) were obtained from
748 http://exac.broadinstitute.org/ (Date dowloaded: March 30, 2017).
749
750 **In situ hybridization.**
751 *In situ* hybridization data was downloaded from publically available data from the Allen Institute
752 through the Allen Brain Atlas (http://www.brain-map.org/). The image used in Figure 3a was
753 obtained from the Reference Atlas at the Allen Brain Atlas (http://mouse.brain-
754 map.org/static/atlas). URLs for all Allen Brain Atlas in situ data analyzed and downloaded for
755 SN marker genes (Figure 3b) are available in Supplementary File 6. Data for SN expression *in*
756 *situ* data for PD GWAS genes (Figure 4b) were obtained from the following experiments: 1056
757 (*Th*), 79908848 (*Snca*), 297 (*Crhr1*), 74047915 (*Atp6v1d*), 72129224 (*Mmp16*), and 414 (*Cntn1*).
758 Data accessed on 03/02/17.
759
760 **Single molecule in situ hybridization (smFISH).**

761 For *in situ* hybridization experiments, untimed pregnant Swiss Webster mice were ordered from
762 Charles River Laboratories (Crl:CFW(SW); http://www.criver.com/). Mice were maintained as
763 previously described. Pups were considered P0 on the day of birth. At P7, the pups were
764 decapitated, the brain was quickly removed, and the brain was then washed in 1x PBS. The intact
765 brain was then transferred to a vial containing freshly prepared 4% PFA in 1x PBS and incubated
766 at 4°C for 24 hours. After 24 hours, brains were removed from PFA and washed three times in 1x
767 PBS. The brains were then placed in a vial with 10% sucrose at 4°C until the brains sunk to the
768 bottom of the vial (usually ~1 hour). After sinking, brains were immediately placed in a vial
769 containing 30% sucrose at 4°C until once again sinking to the bottom of the vial (usually
770 overnight). After cryoprotection, the brains were quickly frozen in optimal cutting temperature
771 (O.C.T.) compound (Tissue-Tek) on dry ice and stored at -80°C until use. Brains were sectioned
772 at a thickness of 14 micrometers and mounted on Superfrost Plus microscope slides
773 (Fisherbrand, Cat. # 12-550-15) with two sections per slide. Sections were then dried at room
774 temperature for at least 30 minutes and then stored at -80°C until use.
775
776 RNAscope *in situ* hybridization (https://acdbio.com/) was used to detect single RNA transcripts.
777 RNAscope probes were used to detect *Th* (C1; Cat No. 317621, Lot: 17073A), *Slc6a3* (C2; Cat
778 No. 315441-C2, Lot: 17044A), *Lhx9* (C3; Cat No. 495431-C3, Lot: 17044A), and *Ldb2* (C3; Cat
779 No. 466061-C3, Lot: 17044A). The RNAscope Fluorescent Multiplex Detection kit (Cat No.
780 320851) and the associated protocol provided by the manufacturer were used. Briefly, frozen
781 tissues were removed from -80°C and equilibrated at room temperature for 5 minutes. Slides
782 were then washed at room temperature in 1x PBS for 3 minutes with agitation. Slides were then
783 immediately washed in 100% ethanol by moving the slides up and down 5-10 times. The slides

784 were then allowed to dry at room temperature and hydrophobic barriers were drawn using a
785 hydrophobic pen (ImmEdge Hydrophobic Barrier PAP Pen, Vector Laboratories, Cat. # H-4000)
786 around the tissue sections. The hydrophobic barrier was allowed to dry overnight. After drying,
787 the tissue sections were treated with RNAscope Protease IV at room temperature for 30 minutes
788 and then slides were washed in 1x PBS. Approximately 100 uL of multiplex probe mixtures (C1
789 - *Th*, C2 - *Slc6a3*, and C3 - one of *Lhx9* or *Ldb2*) containing either approximately 96 uL C1: 2 uL
790 C2: 2 uL C3 (*Th*:*Slc6a3*:*Lhx9*) or 96 uL C1: 0.6 uL C2: 2 uL C3 (*Th*:*Slc6a3*:*Ldb2*) were applied
791 to appropriate sections. Both mixtures provided adequate *in situ* signals. Sections were then
792 incubated at 40°C for 2 hours in the ACD HybEZ oven. Sections were then sequentially treated
793 with the RNAscope Multiplex Fluorescent Detection Reagents kit solutions AMP 1-FL, AMP 2-
794 FL, AMP 3-FL, and AMP 4 Alt B-FL, with washing in between each incubation, according to
795 manufacturer's recommendations. Sections were then treated with DAPI provided with the
796 RNAscope Multiplex Fluorescent Detection Reagents kit. One drop of Prolong Gold Antifade
797 Mountant (Invitrogen, Cat # P36930) was then applied to each section and a coverslip was then
798 placed on the slide. The slides were then stored in the dark at 4°C overnight before imaging.
799 Slides were further stored at 4°C throughout imaging. Manufacturer provided positive and
800 negative controls were also performed alongside experimental probe mixtures according to
801 manufacturer's protocols. Four sections that encompassed relevant populations in the P7 ventral
802 MB (SN, VTA, etc.) were chosen for each combination of RNAscope smFISH probes and
803 subsequent analyses.
804
805 **smFISH Confocal Microscopy.**

806 RNAscope fluorescent *in situ* experiments were analyzed using the Nikon A1 confocal system
807 equipped with a Nikon Eclipse Ti inverted microscope running Nikon NIS-Elements AR 4.10.01
808 64-bit software. Images were captured using a Nikon Plan Apo λ 60x/1.40 oil immersion lens
809 with a common pinhole size of 19.2 μM, a pixel dwell of 28.8 μs, and a pixel resolution of 1024
810 x 1024. DAPI, FITC, Cy3, and Cy5 channels were used to acquire RNAscope fluorescence.
811 Positive and negative control slides using probe sets provided by the manufacturer were used in
812 order to calibrate laser power, offset, and detector sensitivity, for all channels in all experiments
813 performed.
814
815 **smFISH image analysis and processing.**
816 Confocal images were saved as .nd2 files. Images were then processed in ImageJ as follows.
817 First, the .nd2 files were imported into ImageJ and images were rotated in order to reflect a
818 ventral MB orientation with the ventral side of the tissue at the bottom edge. Next the LUT
819 ranges were adjusted for the FITC (range: 0-2500), Cy3 (range: 0-2500), and Cy5 (range: 0-
820 1500) channels. All analyzed images were set to the same LUT ranges. Next, the channels were
821 split and merged back together to produce a "composite" image seen in Figure 2. Scale bars were
822 then added. Cells of interest were then demarcated, duplicated, and the channels were split.
823 These cells of interest were then displayed as the insets seen in Figure 2.
824
825 **Immunohistochemistry and quantification of *Th* striatum staining in *Cplx1* mice.**

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

829 peroxidase inactivated, and one in every 10 sections was processed immunohistochemically for
830 tyrosine hydroxylase. Sections were incubated in primary anti-tyrosine hydroxylase antibody
831 (AB152, Millipore) used at 1/2000 dilution in 1% normal goat serum in phosphate-buffered
832 saline and 0.2% Triton X-100 overnight at 4°C. Antigens were visualised using a horseradish
833 peroxidase-conjugated anti-rabbit second antibody (Vector, PI-1000, 1/2000 dilution) and
834 visualized using diaminobenzidine (DAB; Sigma). The slides were stored in the dark (in black
835 slide boxes) at room temperature (21 C).

836 Images of stained striatum were taken using a Nikon AZ100 microscope equipped with a 2x lens
837 (Nikon AZ Plan Fluor, NA 0.2, WD45), a Nikon DS-Fi2 camera, and NIS-Elements AR 4.5
838 software. Appropriate zoom and light exposure were determined before imaging and kept
839 constant for all slides and sections. Density of Th+ DAB staining was measured using ImageJ
840 software. Briefly, images were imported into ImageJ and the background was subtracted (default
841 50 pixels with "light background" selected). Next, images were converted to 8-bit and the image
842 was inverted. Five measurements of density were taken for each side of a striatum in a section
843 along with a density measurement from adjacent, unstained cortex. Striosomes were avoided
844 during measuring when possible. Striatal measurements had background (defined as staining in
845 the adjacent cortex in a section) subtracted. The mean section measurements (intensity/pixels
846 squared) for each brain were calculated and represented independent measurements of the same
847 brain. Variances were compared between the WT and KO populations. A two sample t-test was
848 then used to compare WT vs. *Cplx1* -/- section densities with the following parameters in R
849 using the "t.test" function: alternative = "two-sided", var.equal = "T".

850

851 **ACKNOWLEDGEMENTS**

852 The authors wish to thank Stephen M. Brown for implementation and optimization of smFISH.
853 Dr. Zhiguang Zheng and Mrs. Wendy Leavens for excellent technical support with the *Cplx1*
854 knockout mice and immunohistochemistry and Drs Kerstin Reim and Niels Brose for the gift of
855 the founder mice for the Cambridge Cplx1 knockout mice colony. This research was supported
856 in part by US National Institutes of Health grants R01 NS62972 and MH106522 to ASM and a
857 grant from CHDI *Inc*. to AJM.
858

859 **AUTHOR CONTRIBUTIONS**

860 PWH, ASM, and LAG designed the study and wrote the paper. PWH, SAM, WDL, GAC, and
861 AJM performed the experiments. PWH and LAG implemented the computational algorithms to
862 process the raw data and conduct analyses thereof. PWH, LAG, and ASM analyzed and
863 interpreted the resulting data. LAG contributed novel computational pipeline development.
864 Correspondence to ASM (andy@jhmi.edu) and LAG (loyalgoff@jhmi.edu).
865

866 **FINANCIAL INTERESTS STATEMENT**

867 The authors declare no competing financial interests.
868

869 **REFERENCES**

870 1. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *Am.*
871 *J. Hum. Genet.* **90,** 7–24 (2012).
872 2. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in
873 Regulatory DNA. *Science (80-.).* **337,** 1190–1195 (2012).
874 3. Farh, K. K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants.
875 *Nature* **518,** 337–343 (2015).
876 4. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections
877 with IRX3. (2014). doi:10.1038/nature13138
878 5. Gupta, R. M. *et al.* A Genetic Variant Associated with Five Vascular Diseases Is a Distal
879 Regulator of Endothelin-1 Gene Expression In Brief A common sequence variant that perturbs
880 long-range enhancer interactions mediates risk for different vascular diseases. A Genetic Variant
881 Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression.
882 *Cell* **170,** 522–533 (2017).
883 6. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nat.*
884 *Genet.* **47,** 955–61 (2015).
885 7. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic
886 to Omnigenic. *Cell* **169,** 1177–1186 (2017).
887 8. de Rijk, M. C. *et al.* Prevalence of parkinsonism and Parkinson's disease in Europe: the
888 EUROPARKINSON Collaborative Study. European Community Concerted Action on the
889 Epidemiology of Parkinson's disease. *J Neurol Neurosurg Psychiatry* **62,** 10–15 (1997).
890 9. Pringsheim, T., Jette, N., Frolkis, A. & Steeves, T. D. The prevalence of Parkinson's disease: a
891 systematic review and meta-analysis. *Mov Disord* **29,** 1583–1590 (2014).
892 10. Savitt, J. M., Dawson, V. L. & Dawson, T. M. Diagnosis and treatment of Parkinson disease:
893 molecules to medicine. *J Clin Invest* **116,** 1744–1754 (2006).
894 11. Nalls, M. a *et al.* Large-scale meta-analysis of genome-wide association data identifies six new

895 risk loci for Parkinson's disease. *Nat. Genet.* **56,** 1–7 (2014).
896 12. Chang, D. *et al.* A meta-analysis of genome-wide association studies identifies 17 new
897 Parkinson's disease risk loci. *Nat. Genet.* (2017). doi:10.1038/ng.3955
898 13. Puschmann, A. Monogenic Parkinson's disease and parkinsonism: clinical phenotypes and
899 frequencies of known mutations. *Park. Relat Disord* **19,** 407–415 (2013).
900 14. Klein, C. & Westenberger, A. Genetics of Parkinson's disease. *Cold Spring Harb Perspect Med* **2,**
901 a008888 (2012).
902 15. Heintz, N. Gene Expression Nervous System Atlas (GENSAT). *Nat. Neurosci.* **7,** 483–483 (2004).
903 16. Barallobre, M. J. *et al.* DYRK1A promotes dopaminergic neuron survival in the developing brain
904 and in a mouse model of Parkinson's disease. *Cell Death Dis.* **5,** e1289 (2014).
905 17. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat Methods*
906 **14,** 309–315 (2017).
907 18. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods*
908 (2017). doi:10.1038/nmeth.4402
909 19. Van Der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9,** 2579–
910 2605 (2008).
911 20. Arenas, E., Denham, M. & Villaescusa, J. C. How to make a midbrain dopaminergic neuron.
912 *Development* **142,** 1918–36 (2015).
913 21. Björklund, A. & Dunnett, S. B. Dopamine neuron systems in the brain: an update. *Trends*
914 *Neurosci* **30,** 194–202 (2007).
915 22. Li, H., Zeitler, P. S., Valerius, M. T., Small, K. & Potter, S. S. Gsh-1, an orphan Hox gene, is
916 required for normal pituitary development. *EMBO J* **15,** 714–724 (1996).
917 23. McNay, D. E., Pelling, M., Claxton, S., Guillemot, F. & Ang, S. L. Mash1 is required for generic
918 and subtype differentiation of hypothalamic neuroendocrine cells. *Mol Endocrinol* **20,** 1623–1632
919 (2006).
920 24. Campbell, J. N. *et al.* A molecular census of arcuate hypothalamus and median eminence cell

921 types. *Nat Neurosci* **20,** 484–496 (2017).

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

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

926 27. Gokce, O. *et al.* Cellular Taxonomy of the Mouse Striatum as Revealed by Single-Cell RNA-Seq.
927 *Cell Rep.* **16,** 1126–1137 (2016).

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

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

933 30. Viereckel, T. *et al.* Midbrain Gene Screening Identifies a New Mesoaccumbal Glutamatergic
934 Pathway and a Marker for Dopamine Cells Neuroprotected in Parkinson's Disease. *Sci Rep* **6,**
935 35203 (2016).

936 31. Kozicz, T., Vigh, S. & Arimura, A. The source of origin of PACAP- and VIP-immunoreactive
937 fibers in the laterodorsal division of the bed nucleus of the stria terminalis in the rat. *Brain Res.*
938 **810,** 211–219 (1998).

939 32. Darland, T., Heinricher, M. M. & Grandy, D. K. Orphanin FQ/nociceptin: A role in pain and
940 analgesia, but so much more. *Trends in Neurosciences* **21,** 215–221 (1998).

941 33. Cai, H., Liu, G., Sun, L. & Ding, J. Aldehyde Dehydrogenase 1 making molecular inroads into the
942 differential vulnerability of nigrostriatal dopaminergic neuron subtypes in Parkinson's disease.
943 *Transl. Neurodegener.* **3,** 27 (2014).

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

946 35. Itoh, N. & Ohta, H. Roles of FGF20 in dopaminergic neurons and Parkinson's disease. *Front Mol*

947 *Neurosci* **6,** 15 (2013).

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

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

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

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

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

958 41. Guttman, M. *et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation.
959 (2011). doi:10.1038/nature10398

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

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

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

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

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

972 47. Pascoli, V., Terrier, J., Hiver, A. & Lüscher, C. Sufficiency of Mesolimbic Dopamine Neuron

973 Stimulation for the Progression to Addiction. *Neuron* **88,** 1054–1066 (2015).

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

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

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

980 51. Benko, S. *et al.* Highly conserved non-coding elements on either side of SOX9 associated with
981 Pierre Robin sequence. *Nat. Genet.* **41,** 359–364 (2009).

982 52. Yong, V. W., Power, C., Forsyth, P. & Edwards, D. R. Metalloproteinases in biology and
983 pathology of the nervous system. *Nat Rev Neurosci* **2,** 502–511 (2001).

984 53. Li, Z., Wu, Y. & Baraban, J. M. The Translin/Trax RNA binding complex: clues to function in the
985 nervous system. *Biochim Biophys Acta* **1779,** 479–485 (2008).

986 54. Close, J. *et al.* Satb1 Is an Activity-Modulated Transcription Factor Required for the Terminal
987 Differentiation and Connectivity of Medial Ganglionic Eminence-Derived Cortical Interneurons.
988 *J. Neurosci.* **32,** 17690–17705 (2012).

989 55. Balamotis, M. a. *et al.* Satb1 Ablation Alters Temporal Expression of Immediate Early Genes and
990 Reduces Dendritic Spine Density during Postnatal Brain Development. *Mol. Cell. Biol.* **32,** 333–
991 347 (2012).

992 56. Brichta, L. *et al.* Identification of neurodegenerative factors using translatome-regulatory network
993 analysis. *Nat Neurosci* **18,** 1325–1333 (2015).

994 57. Doan, R. N. *et al.* Mutations in Human Accelerated Regions Disrupt Cognition and Social
995 Behavior. *Cell* **167,** 341–354.e12 (2016).

996 58. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–
997 291 (2016).

998 59. Basso, M. *et al.* Proteome analysis of human substantia nigra in Parkinson's disease. *Proteomics* **4,**

999 3943–3952 (2004).
1000 60. Gispert, S. *et al.* Complexin-1 and Foxp1 Expression Changes Are Novel Brain Effects of Alpha-
1001 Synuclein Pathology. *Mol. Neurobiol.* **52,** 57–63 (2015).
1002 61. Glynn, D., Drew, C. J., Reim, K., Brose, N. & Morton, A. J. Profound ataxia in complexin I
1003 knockout mice masks a complex phenotype that includes exploratory and habituation deficits.
1004 *Hum. Mol. Genet.* **14,** 2369–2385 (2005).
1005 62. Glynn, D., Sizemore, R. J. & Morton, A. J. Early motor development is abnormal in complexin 1
1006 knockout mice. *Neurobiol. Dis.* **25,** 483–495 (2007).
1007 63. Kielar, C., Sawiak, S. J., Negredo, P. N., Tse, D. H. Y. & Morton, A. J. Tensor-based
1008 morphometry and stereology reveal brain pathology in the complexin1 knockout mouse. *PLoS*
1009 *One* **7,** (2012).
1010 64. Parent, M. & Parent, A. Substantia nigra and Parkinson's disease: a brief history of their long and
1011 intimate relationship. *Can J Neurol Sci* **37,** 313–319 (2010).
1012 65. Lahut, S. *et al.* Blood RNA biomarkers in prodromal PARK4 and REM sleep behavior disorder
1013 show role of complexin-1 loss for risk of Parkinson's disease. *Dis. Model. Mech.* dmm.028035
1014 (2017). doi:10.1242/dmm.028035
1015 66. Hildick-Smith, G. J. *et al.* Macrocytic anemia and mitochondriopathy resulting from a defect in
1016 sideroflexin 4. *Am. J. Hum. Genet.* **93,** 906–914 (2013).
1017 67. Islam, M. M., Suzuki, H., Makoto, Y. & Tanaka, M. Primary structure of the smallest (6.4-kDa)
1018 subunit of human and bovine ubiquinol-cytochrome c reductase deduced from cDNA sequences.
1019 *Biochem Mol Biol Int.* **41,** 1109–1116 (1997).
1020 68. Swartz, D. A., Park, E. I., Visek, W. J. & Kaput, J. The e subunit gene of murine F1F0-ATP
1021 synthase. Genomic sequence, chromosomal mapping, and diet regulation. *J. Biol. Chem.* **271,**
1022 20942–20948 (1996).
1023 69. Tomar, D. *et al.* MCUR1 Is a Scaffold Factor for the MCU Complex Function and Promotes
1024 Mitochondrial Bioenergetics. *Cell Rep.* **15,** 1673–1685 (2016).

1025 70. Plovanich, M. *et al.* MICU2, a Paralog of MICU1, Resides within the Mitochondrial Uniporter
1026 Complex to Regulate Calcium Handling. *PLoS One* **8,** (2013).
1027 71. Wonsey, D. R., Zeller, K. I. & Dang, C. V. The c-Myc target gene PRDX3 is required for
1028 mitochondrial homeostasis and neoplastic transformation. *Proc. Natl. Acad. Sci. U. S. A.* **99,**
1029 6649–54 (2002).
1030 72. Curran, J. E. *et al.* Genetic variation in PARL influences mitochondrial content. *Hum. Genet.* **127,**
1031 183–190 (2010).
1032 73. Winklhofer, K. F. & Haass, C. Mitochondrial dysfunction in Parkinson's disease. *Biochim Biophys*
1033 *Acta* **1802,** 29–44 (2010).
1034 74. Shi, G. *et al.* Functional alteration of PARL contributes to mitochondrial dysregulation in
1035 Parkinson's disease. *Hum. Mol. Genet.* **20,** 1966–1974 (2011).
1036 75. Shi, G. & McQuibban, G. A. The Mitochondrial Rhomboid Protease PARL Is Regulated by PDK2
1037 to Integrate Mitochondrial Quality Control and Metabolism. *Cell Rep.* **18,** 1458–1472 (2017).
1038 76. Jin, S. M. *et al.* Mitochondrial membrane potential regulates PINK1 import and proteolytic
1039 destabilization by PARL. *J. Cell Biol.* **191,** 933–942 (2010).
1040 77. Galter, D. *et al.* LRRK2 expression linked to dopamine-innervated areas. *Ann Neurol* **59,** 714–719
1041 (2006).
1042 78. Higashi, S. *et al.* Expression and localization of Parkinson's disease-associated leucine-rich repeat
1043 kinase 2 in the mouse brain. *J Neurochem* **100,** 368–381 (2007).
1044 79. Soden, M. E. *et al.* Disruption of Dopamine Neuron Activity Pattern Regulation through Selective
1045 Expression of a Human KCNN3 Mutation. *Neuron* **80,** 997–1009 (2013).
1046 80. Abuirmeileh, A., Harkavyi, A., Kingsbury, A., Lever, R. & Whitton, P. S. The CRF-like peptide
1047 urocortin greatly attenuates loss of extracellular striatal dopamine in rat models of Parkinson's
1048 disease by activating CRF1 receptors. *Eur. J. Pharmacol.* **604,** 45–50 (2009).
1049 81. Simunovic, F. *et al.* Gene expression profiling of substantia nigra dopamine neurons: further
1050 insights into Parkinson's disease pathology. *Brain* **132,** 1795–1809 (2009).

1051 82. Ueda, S. *et al.* Age-related dopamine deficiency in the mesostriatal dopamine system of zitter
1052 mutant rats: Regional fiber vulnerability in the striatum and the olfactory tubercle. *Neuroscience*
1053 **95,** 389–398 (1999).
1054 83. Nakadate, K. *et al.* Progressive dopaminergic neurodegeneration of substantia nigra in the zitter
1055 mutant rat. *Acta Neuropathol.* **112,** 64–73 (2006).
1056 84. Saxena, A. *et al.* Trehalose-enhanced isolation of neuronal sub-types from adult mouse brain.
1057 *Biotechniques* **52,** 381–385 (2012).
1058 85. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9,** 171–181
1059 (2014).
1060 86. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory
1061 requirements. *Nat. Methods* **12,** 357–60 (2015).
1062 87. Mudge, J. M. & Harrow, J. Creating reference gene annotation for the mouse C57BL6/J genome
1063 assembly. *Mamm Genome* **26,** 366–378 (2015).
1064 88. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments
1065 with TopHat and Cufflinks. *Nat. Protoc.* **7,** 562–78 (2012).
1066 89. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by
1067 pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32,** 381–6 (2014).
1068 90. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*
1069 **12,** 115–121 (2015).
1070 91. Wang, X.-F. & Xu, Y. Fast clustering using adaptive density peak detection. *Stat. Methods Med.*
1071 *Res.* 1–14 (2015). doi:10.1177/0962280215609948
1072 92. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq.
1073 *Nat. Biotechnol.* **31,** 46–53 (2013).
1074 93. Burns, J. C., Kelly, M. C., Hoa, M., Morell, R. J. & Kelley, M. W. Single-cell RNA-Seq resolves
1075 cellular complexity in sensory organs from the neonatal inner ear. *Nat Commun* **6,** 8557 (2015).
1076 94. Molyneaux, B. J. *et al.* DeCoN: Genome-wide analysis of invivo transcriptional dynamics during

1077 pyramidal neuron fate selection in neocortex. *Neuron* **85,** 275–288 (2015).

1078 95. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting
1079 genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102,** 15545–15550 (2005).

1080 96. Mootha, V. K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are
1081 coordinately downregulated in human diabetes. *Nat Genet* **34,** 267–273 (2003).

1082 97. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological
1083 themes among gene clusters. *OMICS* **16,** 284–7 (2012).

1084 98. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the
1085 Dynamic Tree Cut package for R. *Bioinformatics* **24,** 719–720 (2008).

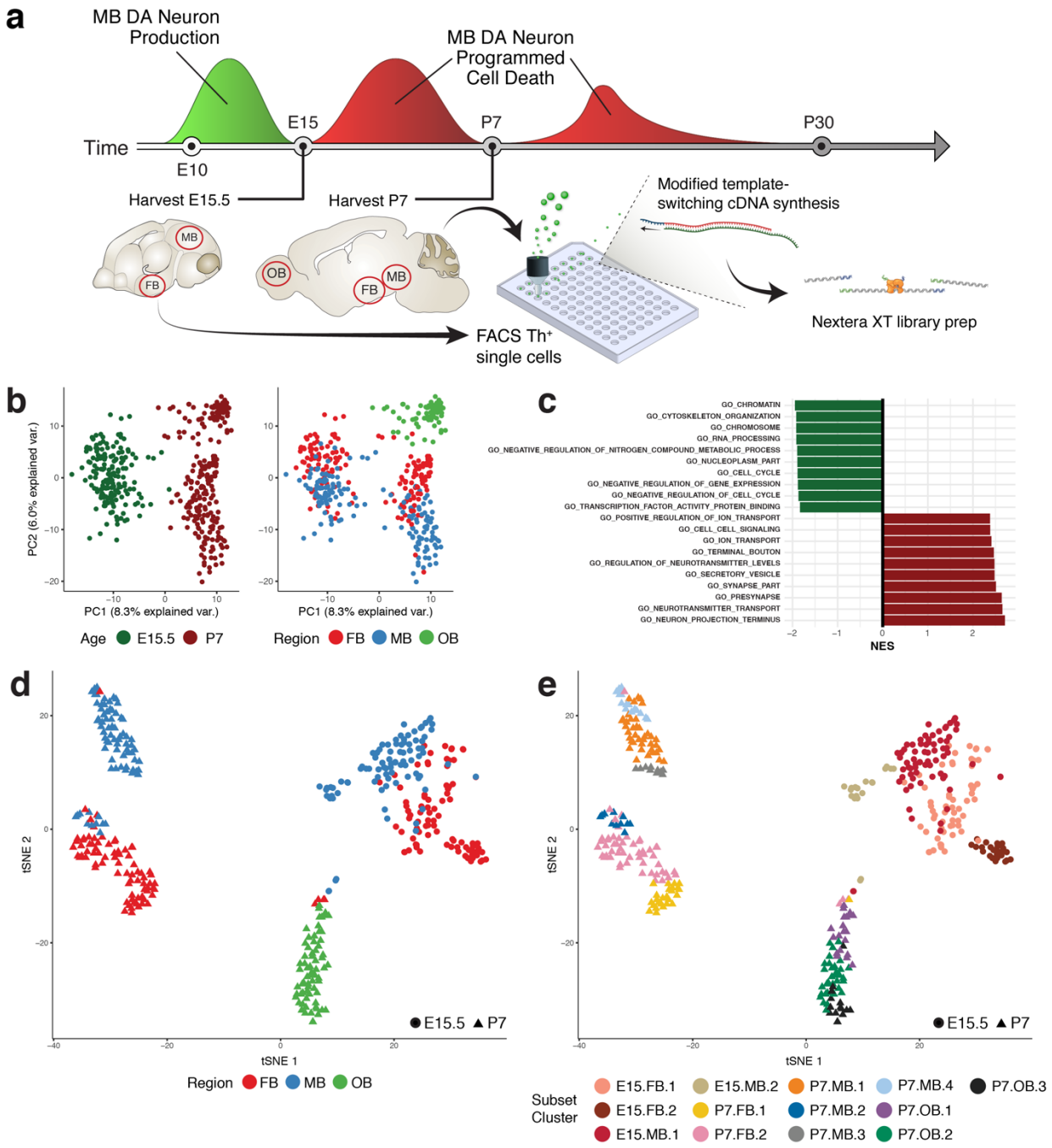
1086 99. Durinck, S. *et al.* BioMart and Bioconductor: A powerful link between biological databases and
1087 microarray data analysis. *Bioinformatics* **21,** 3439–3440 (2005).

1088 100. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of
1089 genomic datasets with the R/Bioconductor package biomaRt. (2009). doi:10.1038/nprot.2009.97

1090

1091 **FIGURES**

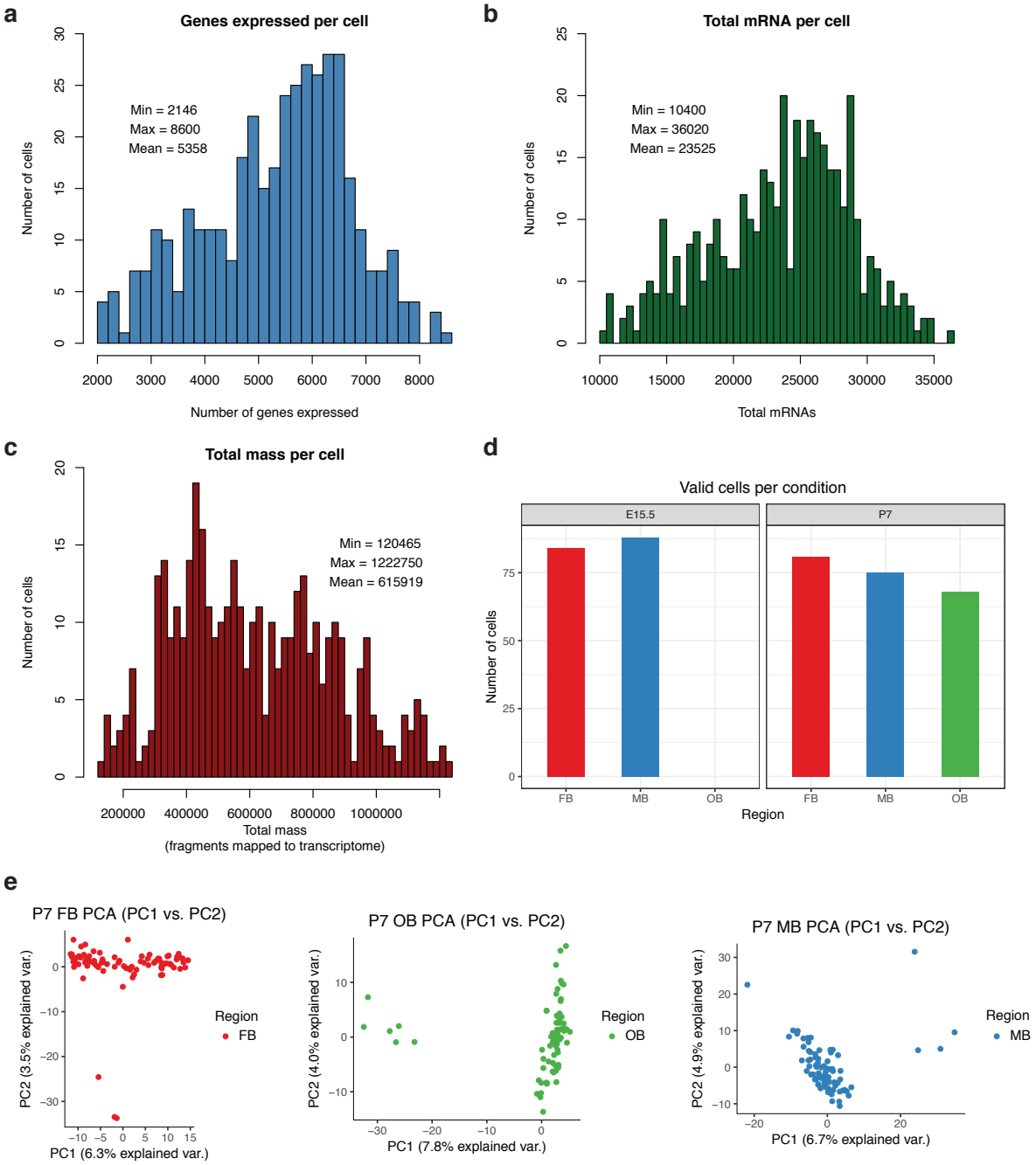
1092 **Figure 1. scRNA-seq analysis of isolated cells allows their separation by developmental time.**



1093

1094

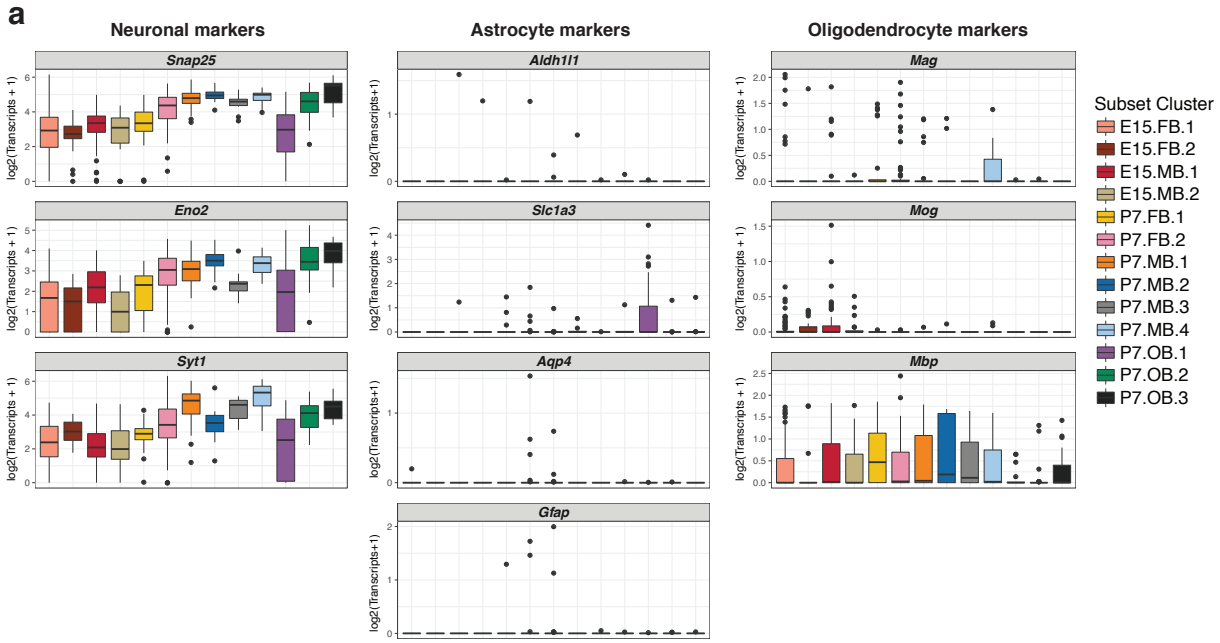
1095 **Figure 1 - supplement 1. Quality control used for filtering single-cell RNA-seq data**



1096

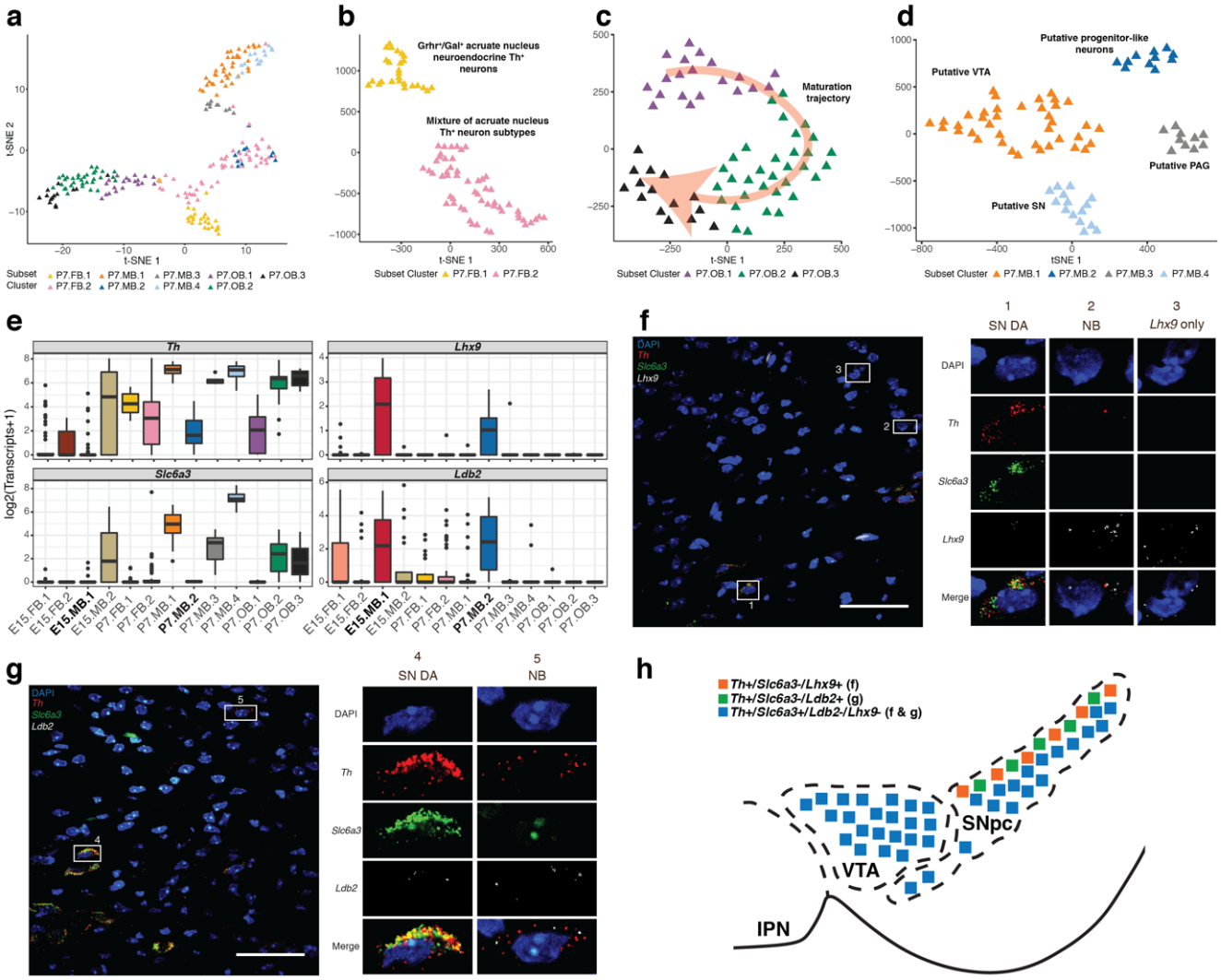
1097

1098 **Figure 1 - supplement 2. Expression of broad marker genes confirms successful isolation of neurons**



1099

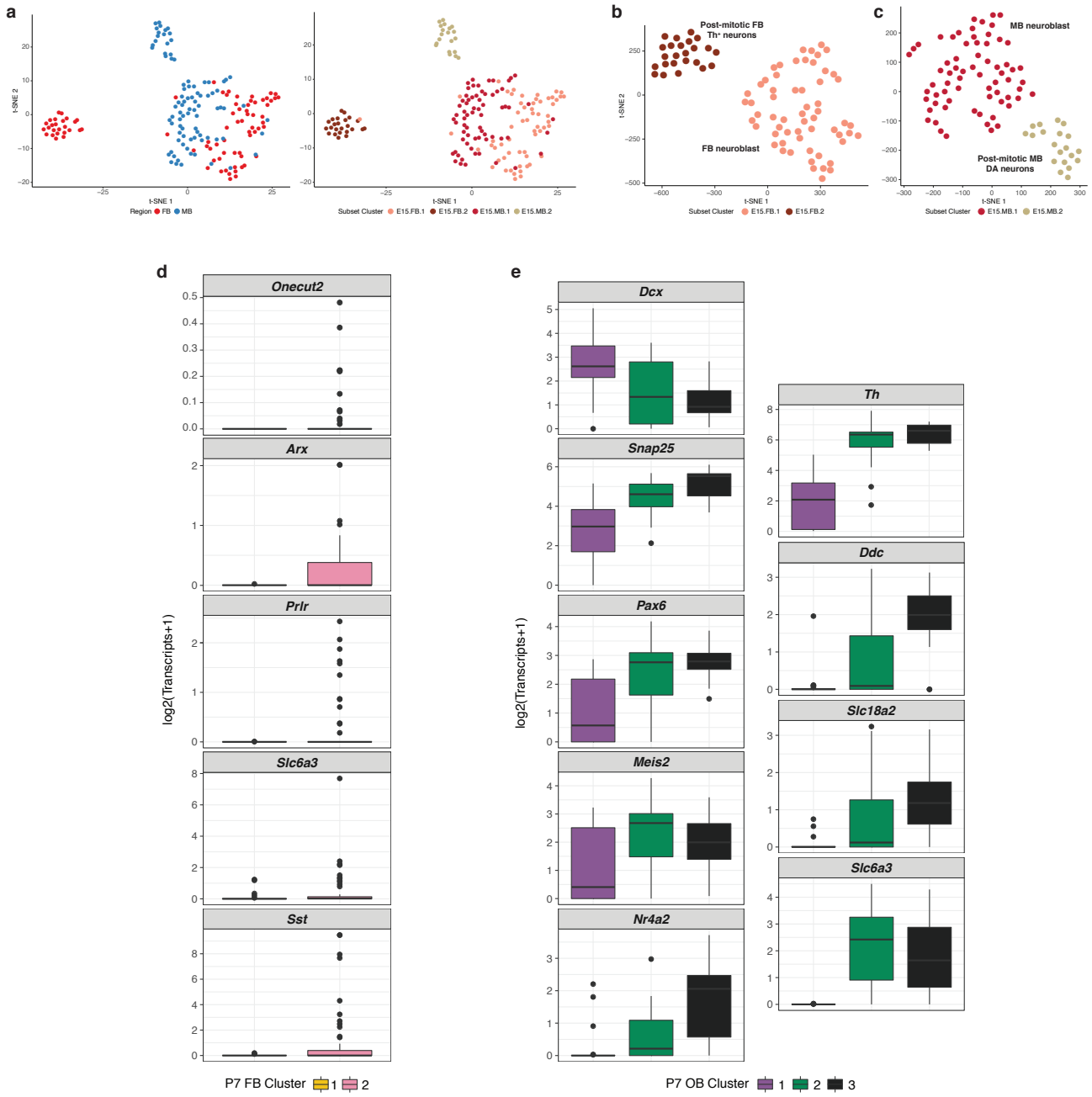
1100 **Figure 2. Subclusters of P7 *Th*+ neurons are identified based on marker gene analyses.**



1101

1102

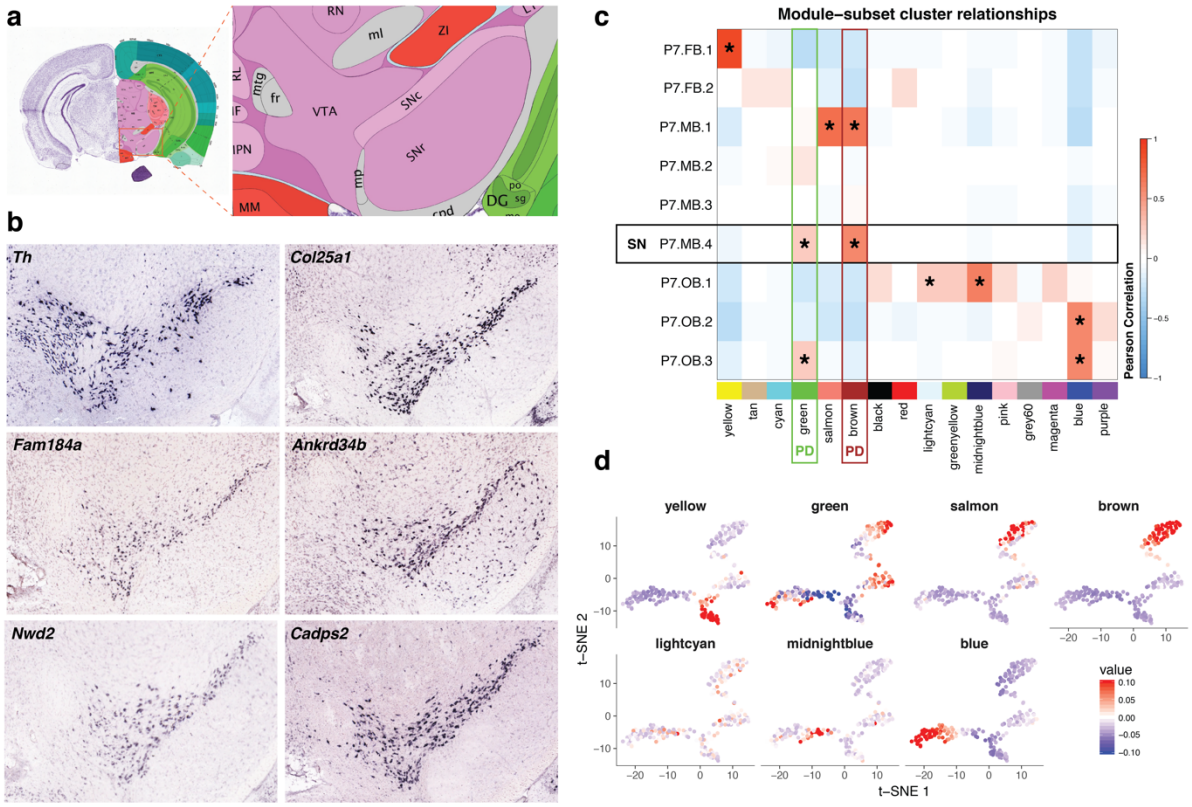
1103 **Figure 2 - supplement 1. Clusters of Th+ neurons are discovered through iterative, marker gene**
1104 **analysis.**
1105



1106

1107 **Figure 3. Novel markers and gene modules reveal context specific SN DA biology.**

1108

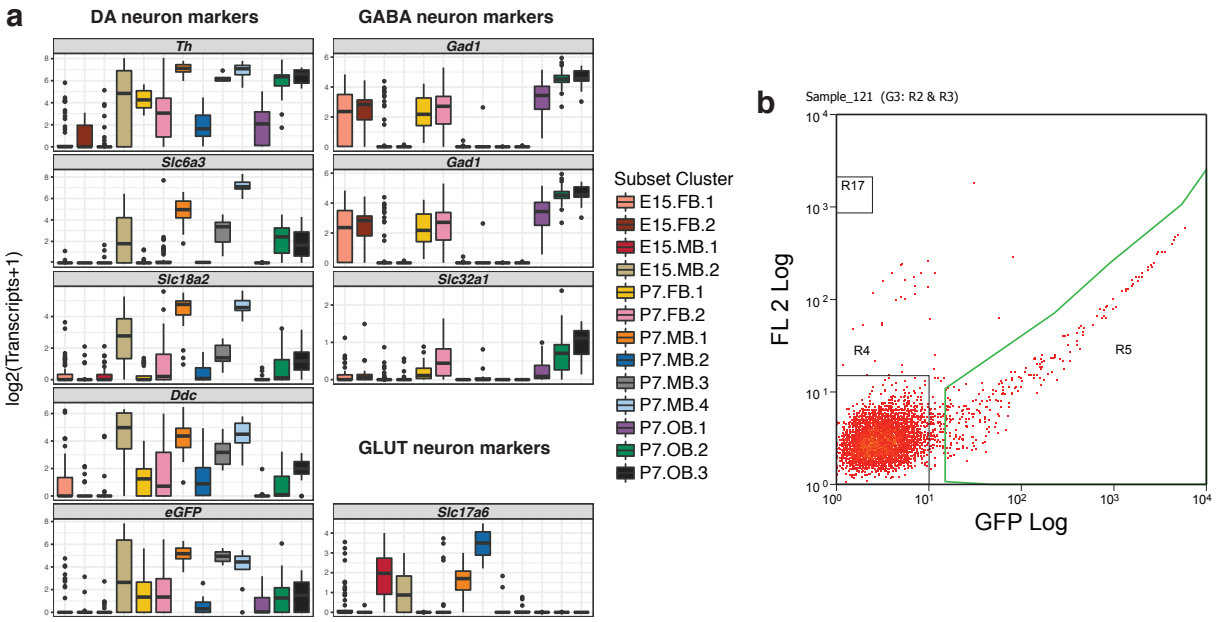


1109

1110

1111 **Figure 3 - supplement 1. Exploration of neuronal subtype markers in isolated DA neuron**

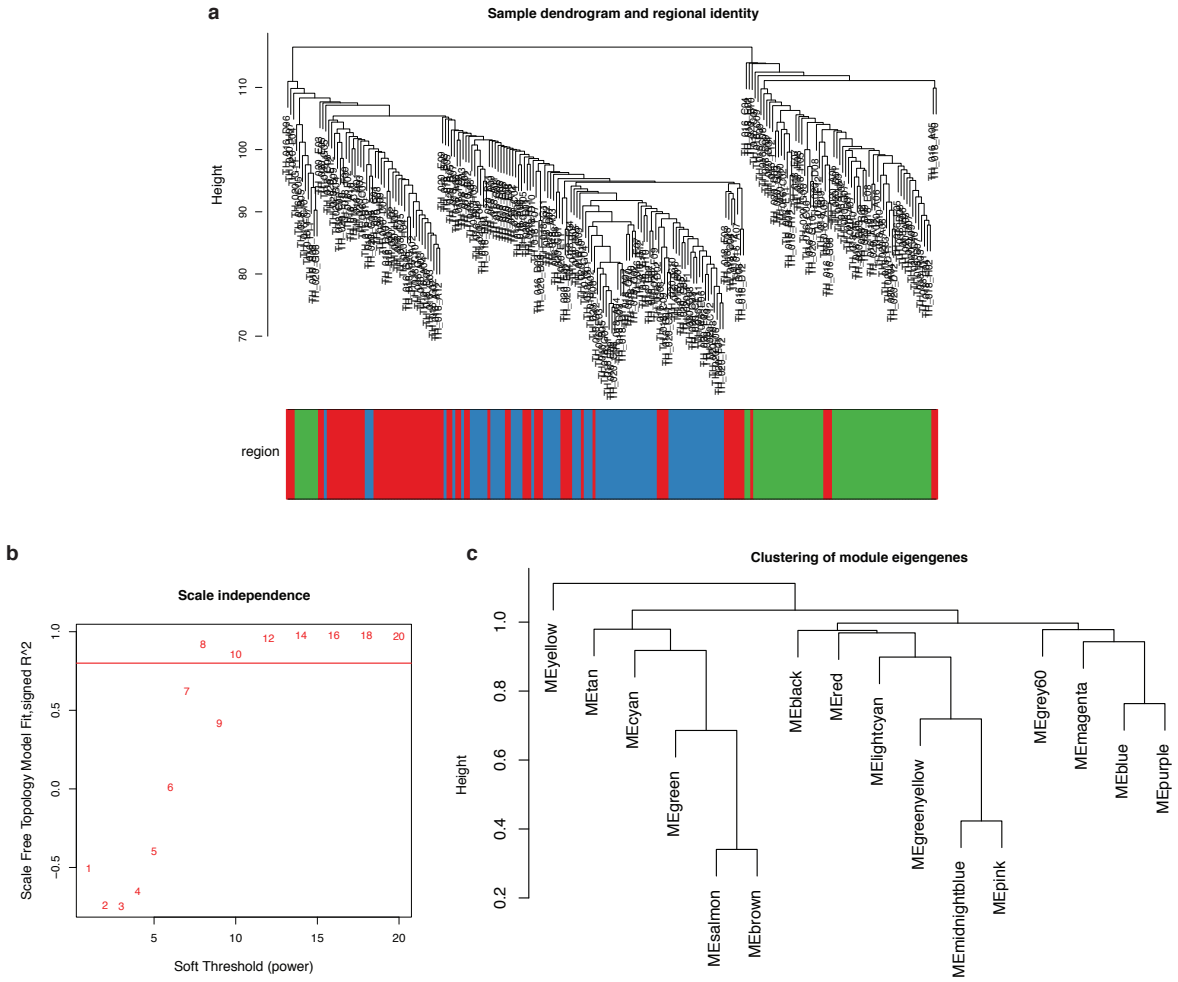
1112 **populations.**



1113

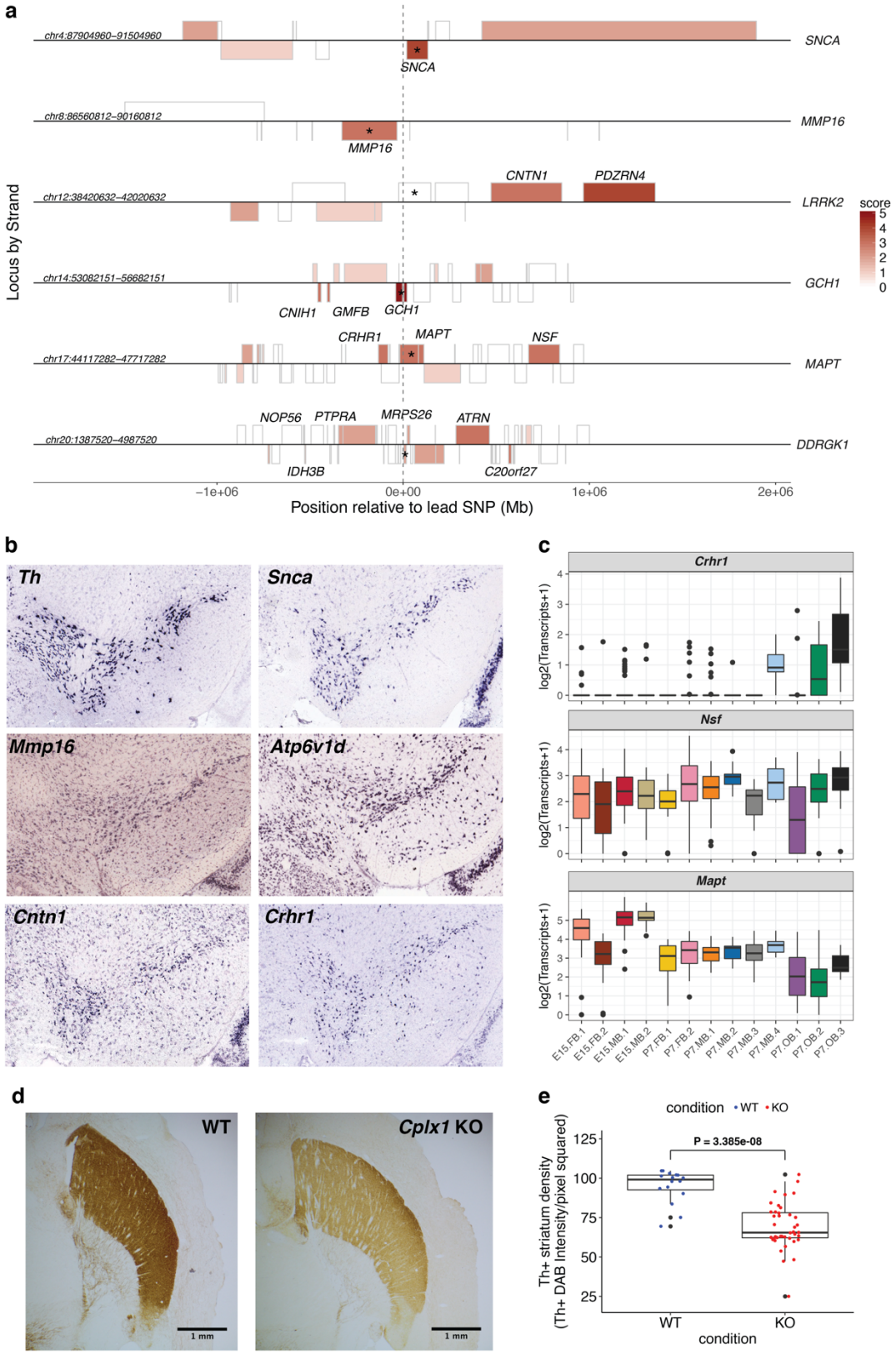
1114

1115 **Figure 3 - supplement 2. WGCNA analysis reveals 16 modules in P7 scRNA-seq data**



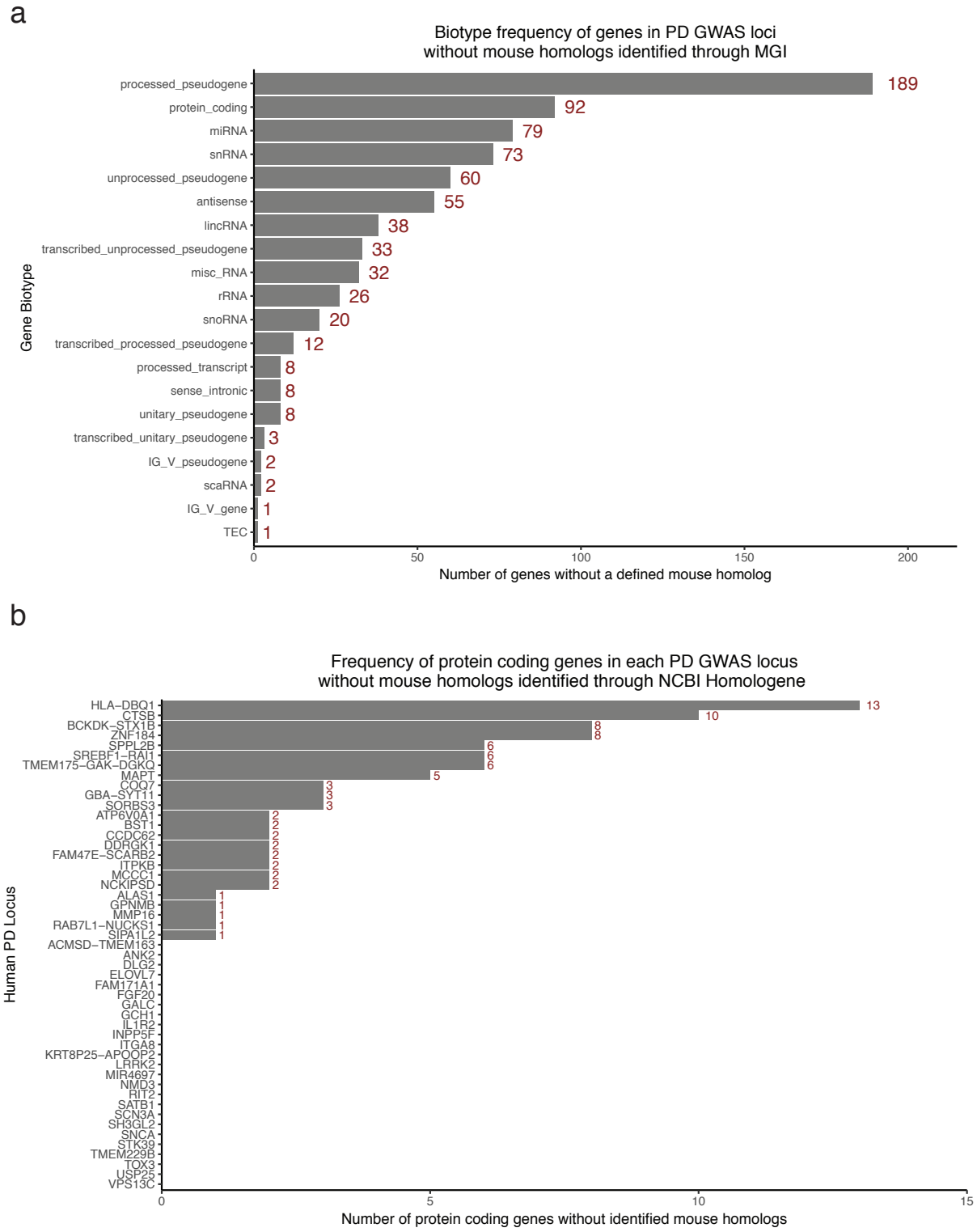
1116

1117 **Figure 4. Context specific SN DA data allows for the prioritization of genes in PD GWAS loci.**



1118

1119 **Figure 4 - supplement 1. The distribution of gene biotypes assigned to genes extracted from PD**
1120 **GWAS loci.**



1121

1122 **TABLES**

1123 **Table 1. Summary of the systematic scoring of genes in 49 GWAS loci associated with PD**

Lead SNP	Top Candidate Genes	Prioritized by
rs6430538	*UBXN4;CCNT2;R3HDM1;RAB3GAP1*	SN expression; pLI
rs14235	*MAPK3;VKORC1; BOLA2B*	SN expression; Differential expression; pLI
rs11724635	*CPEB2*	SN expression; pLI
rs11060180	*ARL6IP4*	SN expression; Differential expression
rs8118008	*ATRN; NOP56; MRPS26; C20orf27;IDH3B*	SN expression; Differential expression; pLI
rs3793947	*DLG2;CCDC90B*	SN expression; Differential expression; pLI
rs6812193	*G3BP2;CCNI;CDKL2*	SN expression; Differential expression; pLI
rs591323	*FGF20; ZDHHC2; TUSC3; MICU3; MTMR7*	SN expression; Differential expression; SN specific; pLI
rs35749011	*KCNN3*	SN expression; Differential expression; SN specific; WGCNA module
rs11158026	*GCH1*	SN expression; Differential expression; SN specific; WGCNA module
rs199347	*RAPGEF5*	SN expression; Differential expression
rs9275326	*ATP6V1G2*	SN expression; Differential expression; WGCNA module
rs117896735	*PRDX3;NANOS1;INPP5F;SFXN4*	SN expression; Differential expression; pLI
rs7077361	*FAM171A1*	SN expression; Differential expression
rs115185635	*CHMP2B*	SN expression; Differential expression
rs76904798	*PDZRN4*	SN expression; Differential expression; WGCNA module
rs17649553	*CRHR1; NSF; MAPT*	SN expression; Differential expression; pLI
rs12637471	*DCUN1D1; ABCC5; PARL*	SN expression; Differential expression; pLI
rs329648	*OPCML*	SN expression; Differential expression
rs60298754	*MMP16*	SN expression
rs34016896	*B3GALNT1*	SN expression; Differential expression
rs823118	*LRRN2; KLHDC8A; SRGAP2*	SN expression; Differential expression; pLI
rs12456492	*RIT2;SYT4*	SN expression; Differential expression; pLI
rs10797576	*TSNAX*	SN expression
rs356182	*SNCA*	SN expression; Differential expression; WGCNA module
rs62120679	*UQCR11*	SN expression; Differential expression; WGCNA module
rs11868035	*COPS3; NT5M*	SN expression; Differential expression; pLI
rs1474055	*STK39;B3GALT1*	SN expression; Differential expression; pLI

rs34311866	*MAEA; CPLX1; ATP5I; TMEM175*	SN expression; Differential expression; WGCNA module; pLI
rs1555399	*VTI1B; ATP6V1D*	SN expression; Differential expression; pLI
rs2823357	*HSPA13*	SN expression
rs2414739	*TLN2; RORA*	SN expression; pLI
rs143918452	*NISCH; PCBP4; SPCS1; SMIM4*	SN expression; Differential expression; pLI
rs78738012	*ANK2; CAMK2D*	SN expression; Differential expression; pLI
rs601999	*DNAJC7; ATP6V0A1; ACLY; PSME3; CNP; RPL27; VAT1; COA3; HAP1*	SN expression; Differential expression; pLI
rs11343	*SYT17*	SN expression; Differential expression; WGCNA module
rs2740594	*FAM167A*	SN expression; Differential expression; SN specific; WGCNA module
rs2694528	*NDUFAF2*	SN expression
rs10906923	*FAM171A1*	SN expression; Differential expression
rs8005172	*ZC3H14*	SN expression
rs34043159	*RPL31; CREG2*	SN expression; Differential expression; pLI
rs4653767	*SRP9; PSEN2; PARP1*	SN expression; pLI
rs12497850	*SMARCC1; PRKAR2A; RHOA; NICN1; UQCRC1; APEH; TCTA; TMA7; GPX1; IMPDH2; QARS; SHISA5; WDR6*	SN expression; Differential expression; pLI
rs4073221	*SATB1*	SN expression
rs353116	*SCN3A; CSRNP3*	SN expression; Differential expression; pLI
rs13294100	*BNC2*	SN expression; Differential expression; SN specific; WGCNA module
rs2280104	*CHMP7; DMTN*	SN expression; Differential expression; pLI
rs4784227	*TOX3; AKTIP*	SN expression; Differential expression; pLI
rs9468199	*ZSCAN26*	SN expression

1124

1125 **Figure and Table Titles and Legends**

1126

1127 **Figure 1. scRNA-seq analysis of isolated cells allows their separation by developmental time.**

1128 Figure 1. scRNA-seq analysis of isolated cells allows their separation by developmental time. a) Diagram
1129 of scRNA-seq experimental procedures for isolating and sequencing EGFP+ cells. Timeline adapted from
1130 Barallobre, et al., 2014a. b) Principal component analysis (PCA) on all cells collected using genes with
1131 highly variant transcriptional profiles. The greatest source of variation (PC1) is explained by the time
1132 point at which the cells were collected, not the region from which the cells were collected. c) The top ten
1133 Gene Ontology (GO) gene sets enriched in genes with positive (red) and negative (green) PC1 loadings.
1134 Genes with negative PC1 loadings and negative normalized enrichment scores (NES) were enriched for
1135 terms indicative of mitotically active cells. Genes with positive PC1 loadings and NES scores were
1136 enriched for terms expected of more mature neurons. d) A t-distributed Stochastic Neighbor Embedding
1137 (t-SNE) plot of all collected cells colored by regional identity. E15.5 cells cluster together while P7 cells
1138 cluster primarily by regional identity. e) A t-SNE plot of all collected cells colored by subset cluster
1139 identity. Through iterative analysis, timepoint-regions collected can be separated into multiple
1140 subpopulations (13 in total). Midbrain, Mb; Forebrain, FB; Olfactory bulb; OB; Fluorescence activated
1141 cell sorting; FACS.

1142

1143 **Figure 1 - supplement 1. Quality control used for filtering single-cell RNA-seq data**

1144 Figure 1 - supplement 1. Quality control used for filtering single-cell RNA-seq data. a) Histogram
1145 showing the final distribution of the number of genes expressed per cell (n cells = 396). b) Histogram
1146 showing the final distribution of the total mRNA per cell (n cells = 396). c) Histogram showing the final
1147 distribution of the total mass (fragments mapped to the transcriptome) per cell (n cells = 396). d) Barplot
1148 showing the number of cells in each timepoint-region. There were a mean of 79 cells/timepoint region. e)
1149 Principal component analysis (PCA) plots from the iterative analyses performed on P7 FB, P7 OB, and

1150 P7 MB cell populations. Initial analyses in these timepoint-regions revealed outliers that were
1151 subsequently removed.
1152
1153 **Figure 1 - supplement 2. Expression of broad marker genes confirms successful isolation of neurons**
1154 Figure 1 - supplement 2. Expression of broad marker genes confirms successful isolation of neurons. a)
1155 Boxplots showing the expression of pan-neuronal, pan-astrocyte, and pan-oligodendrocyte marker in all
1156 13 subpopulations. All subpopulations show robust expression of pan-neuronal markers. +/- 1.5x
1157 interquartile range is represented by the whiskers on the boxplots. Data points beyond 1.5x interquartile
1158 range are considered as outliers and plotted as black points.
1159
1160 **Figure 2. Subclusters of P7 *Th*+ neurons are identified based on marker gene analyses.**
1161 Figure 2. Subclusters of P7 *Th*+ neurons are identified based on marker gene analyses. a) A t-SNE plot of
1162 all P7 neurons collected using colored by subset cluster identity. The neurons mostly cluster by regional
1163 identity. b) t-SNE plot of P7 FB neurons. P7 FB neurons cluster into two distinct populations. c) t-SNE
1164 plot of P7 OB neurons. P7 OB neurons cluster into three populations. These populations represent a
1165 trajectory of Th+ OB maturation (Table S3) as indicated by the red arrow. d) A t-SNE plot of P7 MB
1166 neurons. P7 MB neurons cluster into four clusters: the *substantia nigra* (SN), the ventral tegmental area
1167 (VTA), the periaqueductal grey area (PAG), and a novel progenitor-like population. e) Boxplots
1168 displaying the expression of four genes (*Th*, *Slc6a3*, *Lhx9*, and *Ldb2*) across all subclusters identified. The
1169 novel P7 MB progenitor-like cluster (P7.MB.2) has a similar expression profile to E15.5 MB neuroblast
1170 population (E15.MB.1) (Table S2). +/- 1.5x interquartile range is represented by the whiskers on the
1171 boxplots. Data points beyond 1.5x interquartile range are considered as outliers and plotted as black
1172 points. f) Representative image of multiplex single molecule fluorescent *in situ* hybridization (smFISH)
1173 for *Th*, *Slc6a3*, and *Lhx9*, in the mouse ventral midbrain. Zoomed-in panels represent cell populations
1174 observed. Scale bar, 50 uM. g) Representative image of multiplex smFISH for *Th*, *Slc6a3*, and *Ldb2*, in
1175 the mouse ventral midbrain. Zoomed-in panels represent cell populations observed. h) Diagram of ventral

1176 midbrain summarizing the results of smFISH. Th+/Slc6a3-/Lhx9+ and Th+/Slc6a3-/Ldb2+ cells are both
1177 found in the dorsal SN. Scale bar, 50 uM. NB, neuroblast; SN, substantia nigra; VTA, ventral tegmental
1178 area; IPN, interpeduncular nucleus.
1179
1180 **Figure 2 - supplement 1. Clusters of Th+ neurons are discovered through iterative, marker gene**
1181 **analysis.**
1182 Figure 2 - supplement 2. Clusters of Th+ neurons are discovered through iterative, marker gene analysis.
1183 a) t-SNE plots of all E15.5 cells colored by regional identity and subset cluster assignment. b) t-SNE plot
1184 of FB E15.5 cells colored by subset cluster assignment. E15.5 FB cells cluster in two distinct populations.
1185 c) t-SNE plot of MB E15.5 cells colored by subset cluster assignment. E15.5 MB cells cluster in two
1186 distinct populations. d) Boxplots showing the expression of markers used to identify the P7.FB.2 cluster
1187 (Table S3). +/- 1.5x interquartile range is represented by the whiskers on the boxplots. Data points beyond
1188 1.5x interquartile range are considered as outliers and plotted as black points. e) Boxplots showing the
1189 expression of markers used to identify P7 olfactory bulb clusters (Table S3). +/- 1.5x interquartile range is
1190 represented by the whiskers on the boxplots. Data points beyond 1.5x interquartile range are considered
1191 as outliers and plotted as black points.
1192
1193 **Figure 3. Novel markers and gene modules reveal context specific SN DA biology.**
1194 Figure 3. Novel markers and gene modules reveal context specific SN DA biology. a) Reference atlas
1195 diagram from the Allen Brain Atlas (ABA; http://www.brain-map.org/) of the P56 mouse ventral
1196 midbrain. b) Confirmation of novel SN DA neuron marker genes through the use of ABA *in*
1197 *situ* hybridization data (http://www.brain-map.org/). Coronal, P56 mouse *in situ* data was explored in
1198 order to confirm the expression of 25 novel SN markers. *Th* expression in P56 mice was used as an
1199 anatomical reference during analysis. c) Correlation heatmap of the Pearson correlation between module
1200 eigengenes and P7 Th+ subset cluster identity. Modules are represented by their assigned colors at the
1201 bottom of the matrix. Modules that had a positive correlation with a subset cluster and had a correlation

1202 P-value less than the Bonferroni corrected significance level (P-value < 3.5e-04) contain an asterisk. SN
1203 cluster (P7.MB.4) identity is denoted by a black rectangle. Modules ("green" and "brown") that were
1204 enriched for the "Parkinson's Disease" KEGG gene set are labeled with "PD." d) The eigengene value for
1205 each P7 neuron in the seven WGCNA modules shown to be significantly positively associated with a
1206 subset cluster overlaid on the t-SNE plot of all P7 neurons (Figure 2a). Plotting of eigengenes confirms
1207 strict spatial restriction of module association. Only the "lightcyan" module does not seem to show robust
1208 spatial restriction.
1209
1210 **Figure 3 - supplement 1. Exploration of neuronal subtype markers in isolated DA neuron**
1211 **populations.**
1212 Figure 3 - supplement 1. Exploration of neuronal subtype markers in isolated DA neuron populations. a)
1213 Boxplots showing the expression of markers for dopaminergic (DA), GABAergic, or glutamatergic
1214 neurons. +/- 1.5x interquartile range is represented by the whiskers on the boxplots. Data points beyond
1215 1.5x interquartile range are considered as outliers and plotted as black points. b) Example of a
1216 fluorescence activated cell sorting (FACS) plot used to isolate EGFP+ cells. EGFP fluorescence levels are
1217 represented on the x-axis and RFP fluorescence levels are represented on the y-axis. Cells were collected
1218 that fell within the area outlined in green.
1219
1220 **Figure 3 - supplement 2. WGCNA analysis reveals 16 modules in P7 scRNA-seq data**
1221 Figure 3 - supplement 2. WGCNA analysis reveals 16 modules in P7 scRNA-seq data. a) A dendrogram
1222 of showing the relationship of P7 cells based on expressed genes. The cells are annotated by regional
1223 identity. b) Scale independence plot showing the scale free topology model fit for different levels of soft
1224 threshold power. This plot was used to determine the soft threshold that would be used for the rest of the
1225 analysis (soft threshold = 10). c) Hierarchical clustering shows the relationship between identified
1226 WGCNA modules.
1227

1228 **Figure 4. Context specific SN DA data allows for the prioritization of genes in PD GWAS loci.**

1229 Figure 4. Context specific SN DA data allows for the prioritization of genes in PD GWAS loci. a) A
1230 locus plot displaying four megabase regions in the human genome (hg38) centered on PD GWAS SNPs
1231 in six loci. Genes are displayed as boxes on their appropriate strand. Genes are shaded by their
1232 prioritization score and gene names are displayed for genes with a score of 3 or higher in each locus. b) *In*
1233 *situ* hybridization from the ABA (http://www.brain-map.org/) of five prioritized genes along with *Th* for
1234 an anatomical reference. Coronal, P56 mouse *in situ* data was used. c) Boxplots displaying expression of
1235 prioritized genes from the *MAPT* locus (Figure 4a; Table 1). +/- 1.5x interquartile range is represented by
1236 the whiskers on the boxplots. Data points beyond 1.5x interquartile range are considered as outliers and
1237 plotted as black points. d) Representative light microscopy images of *Th*+ innervation density in the
1238 striatum of WT and *Cplx1* knockout (KO) mice. Scale bar, 1 mm. e) Boxplots comparing the level of
1239 Th+ striatum innervation between WT and *Cplx1* KO mice. DAB staining density was measured in 35
1240 uM, horizontal sections in WT mice (mice = 3, sections = 16) and *Cplx1* KO mice (mice = 8, sections =
1241 40). Each point in the boxplot represents a stained, 35 uM section. Statistical analyses were performed
1242 between conditions with section averages in order to preserve observed variability (WT n = 16, Cplx1 KO
1243 n = 40). A two sample t-test revealed that Th+ innervation density was significantly lower in *Cplx1* KO
1244 mice (t = 6.4395, df = 54, p = 3.386e-08). Data points outside of 1.5x interquartile range, represented by
1245 the whiskers on the boxplots, are considered as outliers and plotted as black points.

1246

1247 **Figure 4 - supplement 1. The distribution of gene biotypes assigned to genes extracted from PD**
1248 **GWAS loci.**

1249 Figure 4 - supplement 1. The distribution of gene biotypes assigned to genes extracted from PD GWAS
1250 loci. a) Barplot displaying the frequency of gene biotypes in the 742 genes without mouse homologs
1251 identified in PD GWAS loci. Only 92/742 of those genes are annotated as protein coding. b) Barplot
1252 displaying the frequency of protein coding genes without mouse homologs in each PD GWAS locus
1253 studied. 24 loci include at least one protein coding gene without a mouse homolog.

1254

1255 **Table 1. Summary of the systematic scoring of genes in 49 GWAS loci associated with PD**

1256 Scoring was carried out at described in the Results and Methods. Candidate genes are presented for each
1257 of 49 PD GWAS loci analyzed. Information for each PD GWAS locus is presented including the lead
1258 SNP for each locus, the prioritized genes in each locus, and which data prioritized the top genes. Detailed
1259 scoring for each gene can be found in Supplementary File 9.

1260

1261 **Supplemental File Descriptions**

1262

1263 Supplementary File 1. A table with gene set enrichment analysis (GSEA) results for outliers removed
1264 during iterative analyses.

1265

1266 Supplementary File 2. A table with marker genes found for all 13 identified DA neuron populations.

1267

1268 Supplementary File 3. A table summarizing marker genes and observations that led to the biological
1269 classification of all 13 DA neuron populations

1270

1271 Supplementary File 4. A table showing marker genes of SN DA neurons with previous literature evidence
1272 of marking the SN.

1273

1274 Supplementary File 5. A table showing novel marker genes of SN DA neurons with summary of SN
1275 expression for each from Allen Brain Atlas (ABA) *in situ* data.

1276

1277 Supplementary File 6. A table showing all genes that comprise each identified WGCNA module.

1278 Supplementary File 7. A table with Gene Ontology, Reactome, and KEGG enrichment results for all
1279 WGCNA modules.

1280

1281 Supplementary File 8. A table with meta-data for each locus in Table 1. This includes the "Lead SNP"
1282 associated with each locus, the "Closest Genes" to the lead SNP, and whether or not the closest genes are
1283 expressed ("Closest Gene Expressed"). This also has meta-data for genes in each locus including: the
1284 number of human genes ("num_genes"), the number of genes expressed in either of the SN DA scRNA-
1285 seq datasets used in scoring ("num_expressed_either"), the number of genes expressed in both SN DA
1286 scRNA-seq datasets using in scoring ("num_expressed_both"), the number of genes that had a one-to-one
1287 mouse homolog ("num_homolog"), and the number of genes that did not have a one-to-one mouse
1288 homolog ("num_no_homolog").

1289

1290 Supplementary File 9. A table with detailed prioritization scoring for all genes within PD GWAS loci.

1291
1292 Supplementary File 10. A table summarizing information about *Cplx1* and WT mice used in this study
1293 including mouse name, age, genotype, the number of striatal sections measured, and the date
1294 immunohistochemistry was performed.
1295
1296 Supplementary File 11. A table showing all measurements taken for *Cplx1* and WT mice.
1297
1298 Supplementary File 12. A table summarizing the comparison of PD GWAS gene prioritization metrics
1299 found in this paper and in Chang, *et al* (2017).
1300

1301

1302