The American Journal of Human Genetics, Volume *102*

Supplemental Data

Single-Cell RNA-Seq of Mouse Dopaminergic Neurons Informs Candidate Gene Selection for Sporadic Parkinson Disease

Paul W. Hook, Sarah A. McClymont, Gabrielle H. Cannon, William D. Law, A. Jennifer Morton, Loyal A. Goff, and Andrew S. McCallion

Supplemental Figures

Figure S1. Quality control used for filtering single-cell RNA-seq data that led to a total dataset comprised of 396 cells

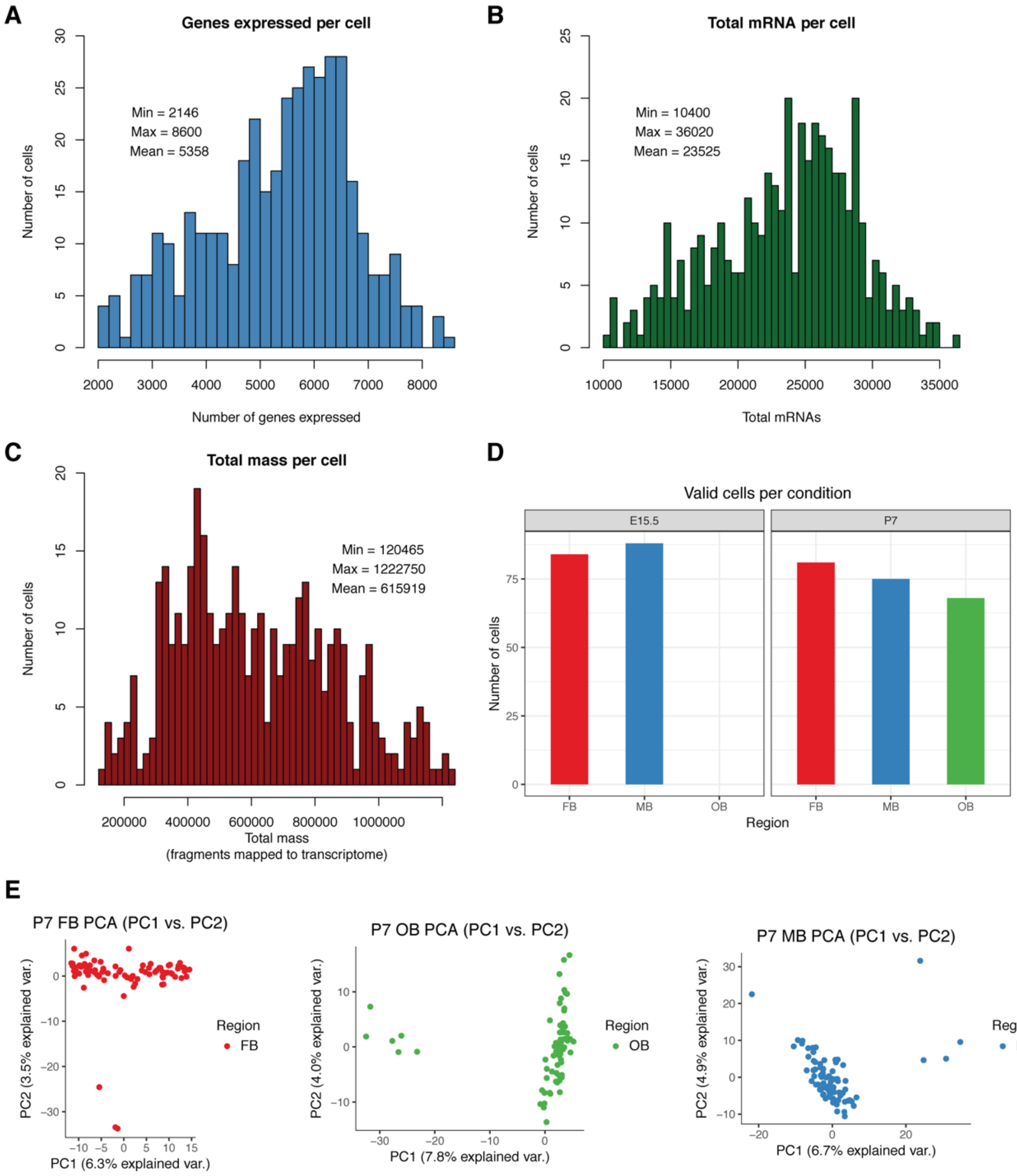

Figure S1. Quality control used for filtering single-cell RNA-seq data that led to a total dataset comprised of 396 cells. A) Histogram showing the final distribution of the number of genes expressed per cell (n cells = 396). B) Histogram showing the final distribution of the total mRNA per cell (n cells = 396). C) Histogram showing the final distribution of the total mass (fragments

mapped to the transcriptome) per cell (n cells = 396). D) Barplot showing the number of cells in each timepoint-region. There was a mean of 79 cells/timepoint region. E) Principal component analysis (PCA) plots from the iterative analyses performed on P7 FB, P7 OB, and P7 MB cell populations. Initial analyses in these timepoint-regions revealed outliers that were subsequently removed.

Figure S2. Expression of various marker genes confirms successful isolation of neurons

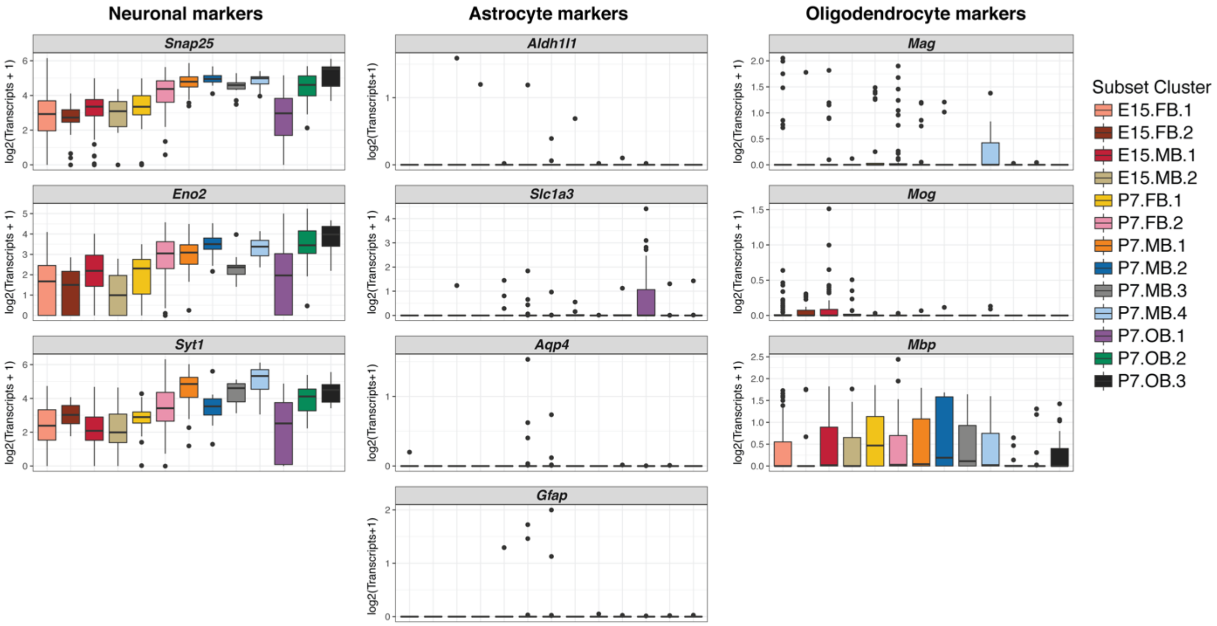

Figure S2. Expression of various marker genes confirms successful isolation of neurons. Included are boxplots showing the expression of pan-neuronal, pan-astrocyte, and pan-oligodendrocyte marker in all 13 subpopulations. All subpopulations show robust expression of pan-neuronal markers. +/- 1.5x interquartile range is represented by the whiskers on the boxplots. Data points beyond 1.5x interquartile range are considered as outliers and plotted as black points.

Figure S3. Clusters of Th^+ neurons are discovered through iterative, marker gene analysis.

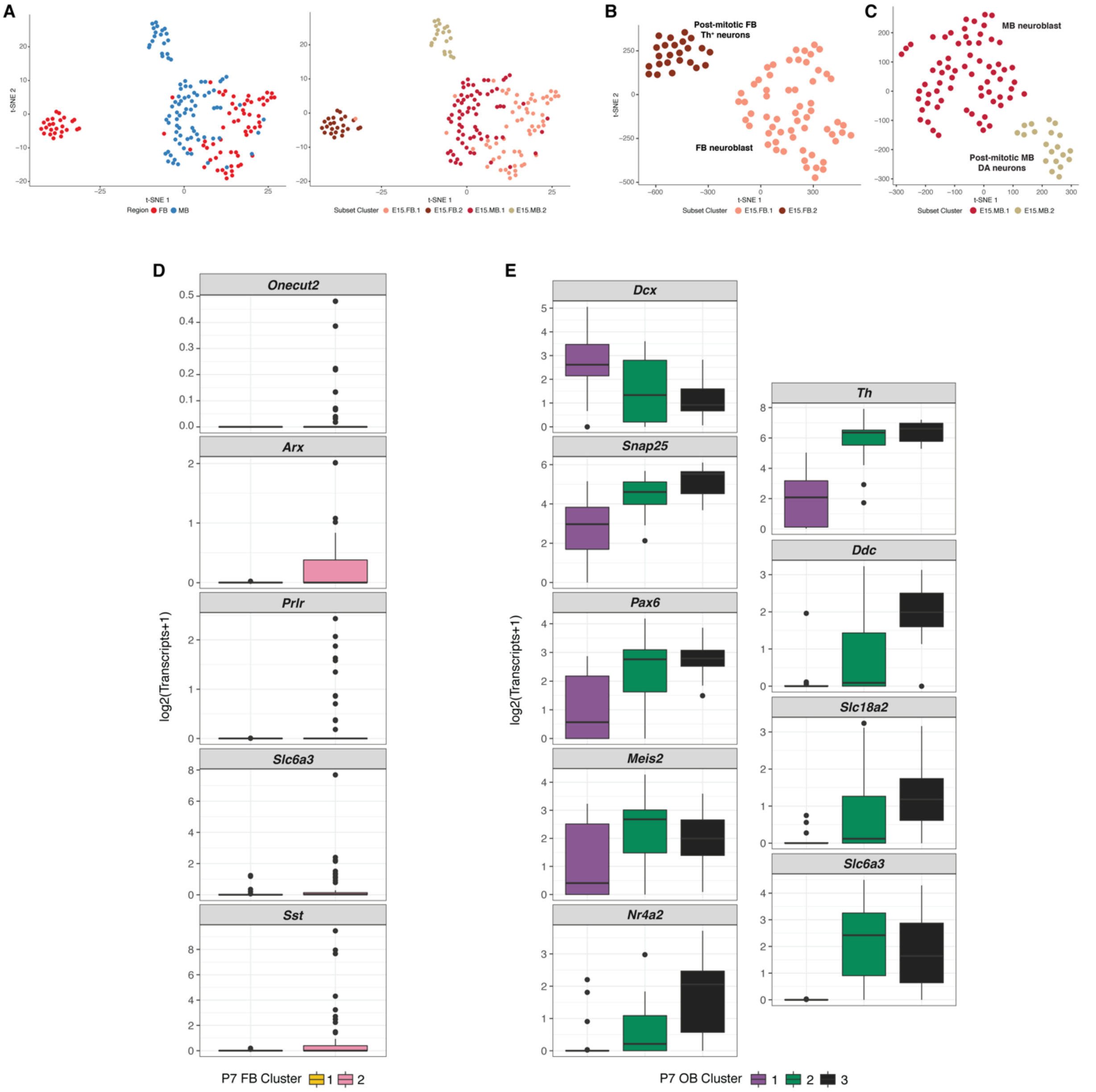

Figure S3. Clusters of Th^+ neurons are discovered through iterative, marker gene analysis. A) t-SNE plots of all E15.5 cells colored by regional identity and subset cluster assignment. B) t-SNE plot of FB E15.5 cells colored by subset cluster assignment. E15.5 FB cells cluster in two distinct populations. C) t-SNE plot of MB E15.5 cells colored by subset cluster assignment. E15.5 MB cells cluster in two distinct populations. D) Boxplots showing the expression of markers used to identify the P7.FB.2 cluster (Table S3). +/- 1.5x interquartile range is represented by the whiskers on the boxplots. Data points beyond 1.5x interquartile range are considered as outliers and plotted as black points. E) Boxplots showing the expression of markers used to identify P7 olfactory bulb clusters (Table S3). +/- 1.5x interquartile range is

represented by the whiskers on the boxplots. Data points beyond 1.5x interquartile range are considered as outliers and plotted as black points.

Figure S4. Expression of various marker genes confirms successful isolation of *Th*$^{+}$ neurons

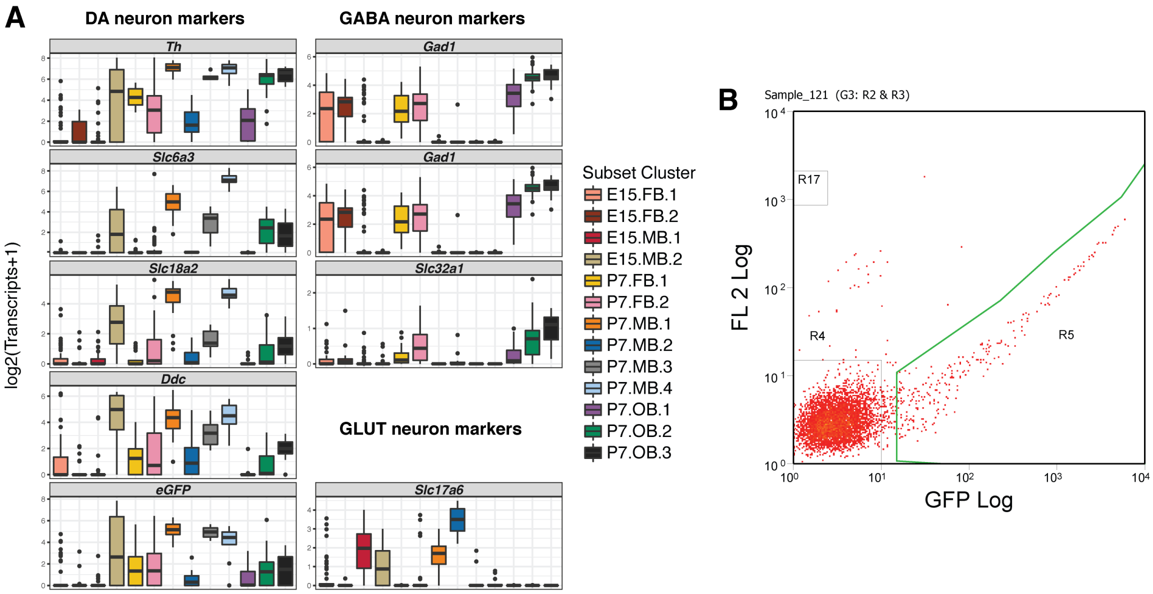

Figure S4. Expression of various marker genes confirms successful isolation of *Th*$^{+}$ neurons. A) Boxplots showing the expression of markers for dopaminergic (DA), GABAergic, or glutamatergic neurons. +/- 1.5x interquartile range is represented by the whiskers on the boxplots. Data points beyond 1.5x interquartile range are considered as outliers and plotted as black points. B) Representative example of fluorescence activated cell sorting (FACS) plot used to isolate E15.5 MB EGFP^{+} cells. EGFP fluorescence levels are represented on the x-axis and RFP fluorescence levels are represented on the y-axis. Cells were collected that fell within the area outlined in green.

Figure S5. WGCNA analysis reveals 16 modules in P7 scRNA-seq data

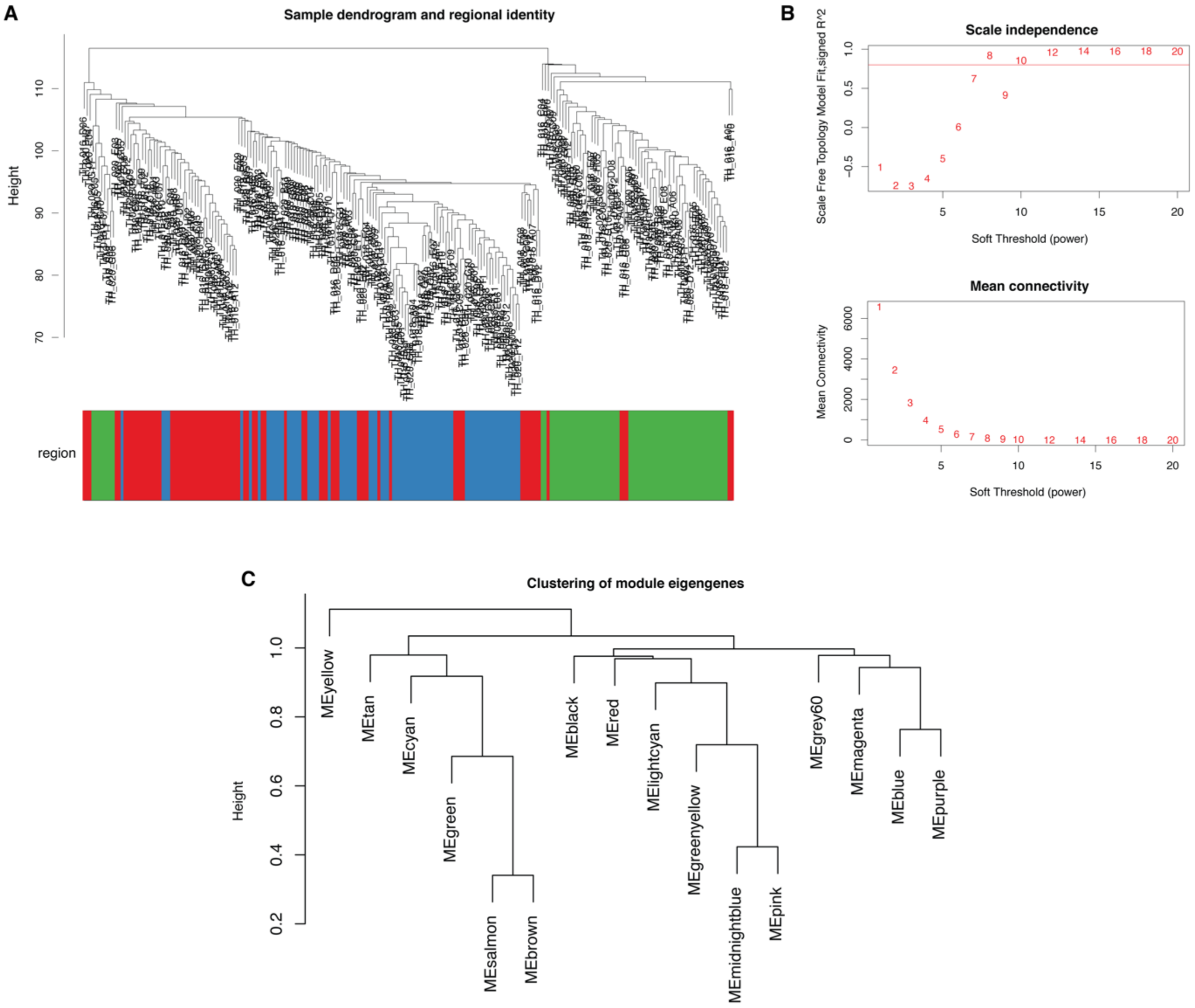

Figure S5. WGCNA analysis reveals 16 modules in P7 scRNA-seq data. A) A dendrogram of showing the relationship of P7 cells (n = 223) based on expressed genes. The cells are annotated by regional identity. B) Scale independence plot showing the scale free topology model fit for different levels of soft threshold power. This plot was used to determine the soft threshold that would be used for the rest of the analysis (soft threshold = 10). C) Hierarchical clustering shows the relationship between identified WGCNA modules.

Figure S6. Results from simulations involving National Human Genome Research Institute (NHGRI) - European Bioinformatics Institute (EBI) GWAS catalog loci.

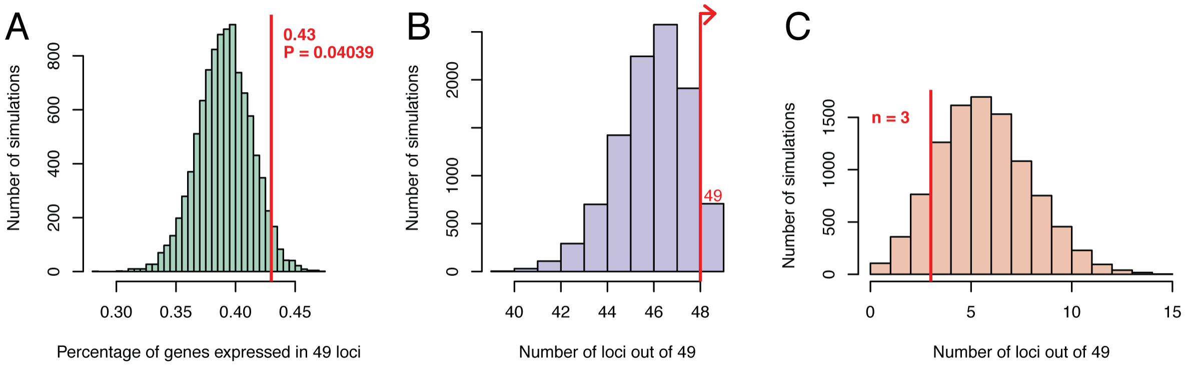

Figure S6. Results from simulations involving NHGRI-EBI GWAS catalog loci. Simulations were performed using all loci downloaded from the NHGRI-EBI GWAS catalog on November 27, 2017 (Web Resources). Only genes with a defined mouse homolog were included in the simulations. Simulations were performed using custom R scripts. Note that genes included in the simulations are those found within +/- 1 Mb of the lead SNP. A) Histogram showing the percentage of genes in 49 random GWAS loci that are expressed in SN DA neurons, simulated 10,000 times. This simulation showed that the percentage of genes expressed in SN DA neurons from PD GWAS loci (430/1009, ~43%; vertical red line) was significantly higher than what is expected from random 49 GWAS loci (one-tailed test applied to a normal distribution; P-value = 0.04039). Normality of data was confirmed by qqplot. B) Histogram showing the number of loci out of 49 random GWAS loci that contain at least one SN DA neuron expressed gene, simulated 10,000 times. All 49 PD GWAS loci analyzed have at least one SN DA expressed gene, which is slightly higher than what is expected from 49 random GWAS loci (right of the red, vertical line). C) Histogram showing the number of loci out of 49 random GWAS loci that contain only one SN DA neuron expressed gene, simulated 10,000 times. The number of PD GWAS loci that contain only one SN DA neuron expressed gene (n = 3; red, vertical line) is slightly less than what would be expected from 49 random GWAS loci.

Figure S7. The distribution of gene biotypes assigned to genes extracted from PD GWAS loci

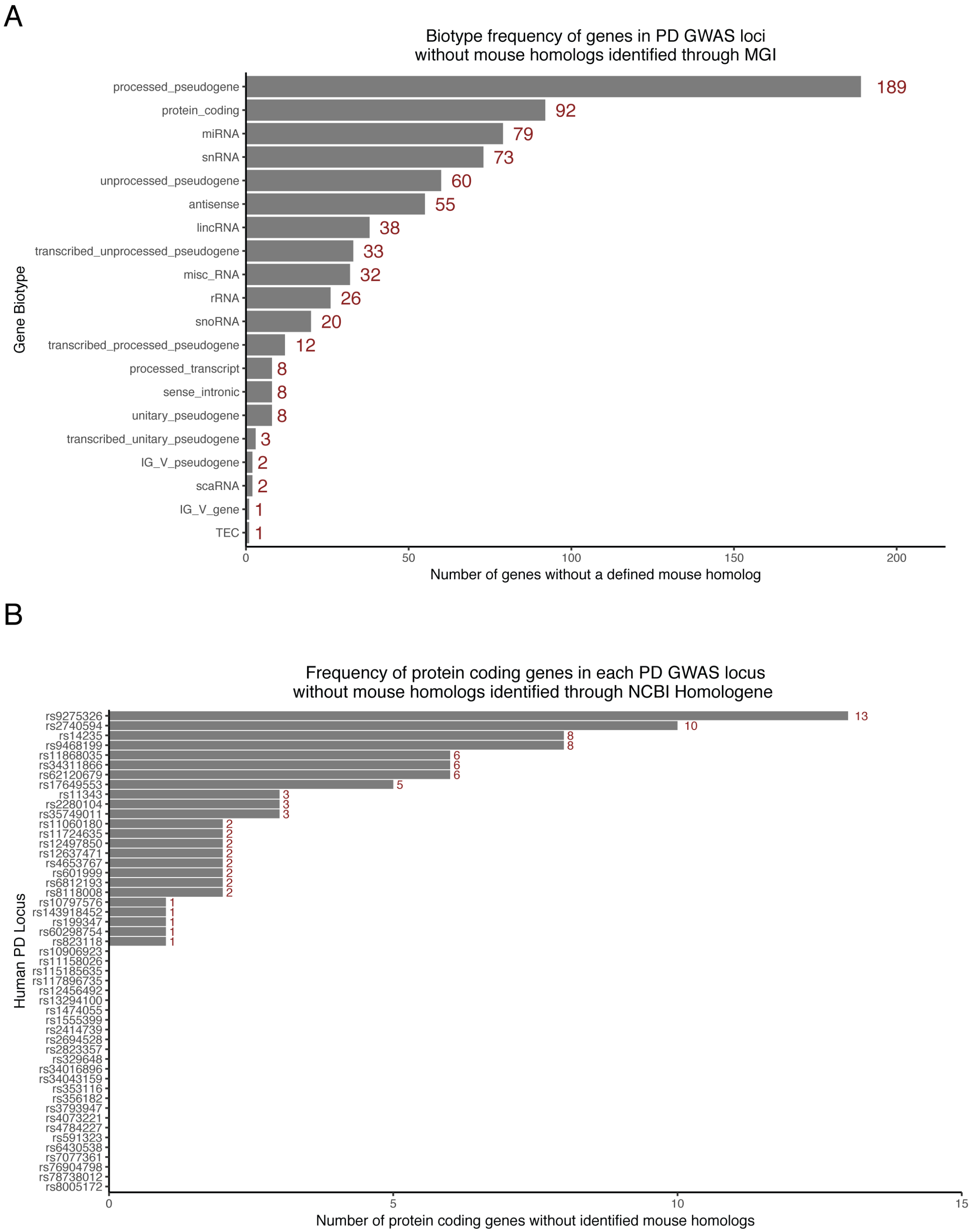

Figure S7. The distribution of gene biotypes assigned to genes extracted from PD GWAS loci. A) Barplot displaying the frequency of gene biotypes in the 742 genes without mouse homologs identified in PD GWAS loci. 92/742 (~12%) of those genes are annotated as protein coding. All 1009 genes with mouse homologs were annotated as “protein_coding.” B) Barplot displaying the frequency of protein coding genes without mouse homologs in each PD GWAS locus studied. 24 loci include at least one protein coding gene without a mouse homolog.

Supplemental Table Titles and Descriptions

Table S1. A table with gene set enrichment analysis (GSEA) results for outliers removed during iterative analyses.

Table S2. A table with marker genes found for all 13 identified DA neuron populations.

Table S3. A table summarizing marker genes and observations that led to the biological classification of all 13 DA neuron populations. Provides additional information for Table 1.

Table S4. A table showing marker genes of SN DA neurons with previous literature evidence of marking the SN.

Table S5. A table showing novel marker genes of SN DA neurons with summary of SN expression for each from Allen Brain Atlas (ABA) *in situ* data.

Table S6. A table showing all genes that comprise each identified WGCNA module.

Table S7. A table with Gene Ontology, Reactome, and KEGG enrichment results for all WGCNA modules.

Table S8. A table with meta-data for each locus in Table 1. This includes the "Lead SNP" associated with each locus, the "Closest Genes" to the lead SNP, and whether or not the closest genes are expressed ("Closest Gene Expressed"). This also has meta-data for genes in each locus including: the number of human genes ("num_genes"), the number of genes expressed in either of the SN DA scRNA-seq datasets used in scoring ("num_expressed_either"), the number of genes expressed in both SN DA scRNA-seq datasets using in scoring ("num_expressed_both"), the number of genes that had a one-to-one mouse homolog ("num_homolog"), and the number of genes that did not have a one-to-one mouse homolog ("num_no_homolog").

Table S9. A table with detailed prioritization scoring for all genes within PD GWAS loci.

Table S10. A table summarizing information about *Cplx1* and WT mice used in this study including mouse name, age, genotype, the number of striatal sections measured, and the date immunohistochemistry was performed.

Table S11. A table showing all measurements taken for *Cplx1* and WT mice.

Table S12. A table summarizing the comparison of PD GWAS gene prioritization metrics found in this paper and in Chang, *et al* (2017).