# *A single-cell atlas of the human substantia nigra reveals cell-specific pathways associated with neurological disorders*

**Agarwal** *et al.*

**Supplementary Information**

## **Supplementary Note 1**

*Single-cell atlas of the substantia nigra:* all cell-types were mostly contributed to by all samples (**Supplementary Figure 1, Supplementary Figure 5**). Some sample-specific clustering and thus batch-specific effects can also be seen for reactive astrocyte-2 (N2B only, Day2)  $\&$  ODC-3 (predominantly N2B) and this may be associated with heterogeneity between samples (**Supplementary Figure 2**, **Supplementary Table 3** and **Supplementary Table 1**). We found that neuronal populations have a higher UMI depth (median nUMI: DaNs = 8,419; GABA neurons  $=6,512$ ) and gene coverage range (median nGene: DaNs=  $3,579$ ; GABA neurons=  $3,529$ ) than non-neuronal populations (median nUMI= 2,822;median nGene = 1,863)(**Supplementary Figure 1**).

*Single-cell atlas of the cortex:* the cortex cell-types were broadly classified into the major cell groups with the help of enriched marker genes and we resolved clusters as astrocytes (*AQP4+, GFAP+*), **Excitatory neuron**s (**Ex**)(*SATB2+,SLC17A7+*), Inhibitory neurons (In) (*GAD1+,GAD2+*) (see below), microglia(*CSF3R+,PLXDC2+*), oligodendrocytes cells (ODCs) (*MOG+,MOBP+*) and oligo precursor cells (OPCs) (*OLIG1+,VCAN+*) (**Supplementary Figure 4**, **Supplementary Data 3**). Similar to the SN, the major cell-types were contributed to by all samples, except for ODCs for which some sample-specific clustering (C1B & C3) can be seen (**Supplementary Figure 3**, **Supplementary Table 3**). Moreover, a uniform UMI depth, number of genes and cell-cycle phase trends were also observed across the cortex for most cell-types, except for the Ex neurons (median nGene= 3,559; nUMI=7,332) and microglia (median nGene= 3,759, nUMI=7,330) showing higher UMI depth and gene coverage than other cell-types (median nGene = 2,096; nUMI=3,599) (**Supplementary Figure 3**).

### **Supplementary Note 2**

Joint clustering analysis of SN and cortex samples: Similarly using Seurat<sup>2</sup>, a joint clustering of the CCA aligned cortex and SN nuclei was performed using the **transcripts per million** (**TPM**) expression matrix of only proteincoding genes which were log-transformed and linearly regressed for nUMI. As before we used the "FindVariableGenes" function, a minimum log(variance/mean) cutoff of 0.5 and maximum cut-off of 9 on the mean-variability plot to detect outliers and identify 4221 HVGs. A **principal component analysis** (**PCA**) was then performed across all nuclei only using the HVGs. The standard deviations of the PCs were plotted on to a scree plot to identify the first 25 PCs as significant that were used as input to identify cell clusters by the **shared nearest neighbor** (**SNN**) based graph approach at a resolution of 0.4 through the "FindClusters" function in Seurat. The identified clusters were then visualised as a **Uniform Manifold Approximation and Projection** (**UMAP**) plot with significant PCAs as input to understand the global relationship between similar cell-types across the cortex and SN (**Figure 1d**). We found that changing the number of significant PCs in the range of 15- 30 and the clustering resolution in the range of 0.4-0.8 did not identify very different clusters.

## **Supplementary Note 3**

*Cell-type association methods:* To reduce the chance of spurious conclusions, we performed cell-type association analyses with two polygenic approaches, stratified LD score regression (LDSC) method<sup>15</sup> and the Multi**marker Analysis of GenoMic Annotation** (MAGMA) gene set analysis method<sup>16</sup>. The stratified LD score regression partitions the heritability from genome-wide association studies. (GWAS) summary statistics (**Supplementary Table 4**) according to different sets of genes specifically expressed in cell-types to identify disease-relevant cell-types. MAGMA performs gene-set enrichment analysis based on GWAS summary statistics while accounting for LD structure between SNPs. A competitive gene-set analysis linear regression model was performed using MAGMA to test the hypothesis that a cell-type-specific gene set has a greater mean association with the complex trait than the genes not present in the gene set. A previous study comparing these two methods has shown that MAGMA can identify more significant results or false positives than LDSC, which the authors believe could be due to uncorrected genomic confounding that can be corrected by including gene-level covariates15. While both methods are based on different assumptions and algorithms and could have different results for certain traits, in this study, both methods include gene-level covariates such as gene size, gene sample size and linkage disequilibrium in the regression models to correct for confounding biases. We showed here that globally, there is agreement between the results given by both methods. Based on the Spearman rank correlations of cell-type association strength (- $log_{10}$  Pvalue) between LDSC & MAGMA for each complex trait, we observed a very similar overall ranking of the cell-types across the SN (R=0.55, p<10<sup>-16</sup>) and the cortex (R= 0.53, p<10<sup>-14</sup>) (**Supplementary Figure 13**) for most brain-related complex traits such as schizophrenia, major depression disorder and bipolar disease. Moreover, non-brain-related traits (e.g. Height) also, show high positive correlations across the two methods.

.

*Cell-type-specific gene sets:* To define cell-type-specific gene sets from the TPM expression matrix and the brain cell-types identified, we followed the approach of Finucane *et al.* 201815. Here, for each gene a t-statistic is calculated between the expression of that gene in a given cell population as compared to its expression in all other cells, with cell-type-specific gene sets defined as the 10% of genes with the highest t-statistic in that cell-type (**Supplementary Data 4 and Supplementary Data 5**). Note that genes are not necessarily exclusive to a single cell-type. For the GTEx-brain tissue gene set, we used the GTEx brain gene set defined by Finucane et *al.* 201815 and downloaded the following file:

#### https://data.broadinstitute.org/alkesgroup/LDSCORE/LDSC\_SEG\_ldscores/GTEx\_brain\_1000Gv3\_ldscores.tgz

*Locus definition:* On either side of the transcribed region of each gene in the set of cell-type specific genes, we extended the locus to define cell-(or tissue-) specific genomic region, aiming to capture regulatory elements that could affect the expression of that gene and mediate any GWAS variant effect on a trait. To define the size of this extension, we exploited the robust association between schizophrenia risk loci and genes specifically expressed in Ex neurons<sup>17</sup>. Thus, we ran LDSC <sup>15</sup> cell-type analysis with different window sizes (5kb, 10 kb, 25 kb, 50kb,75 kb and 100 kb) to evaluate the association between schizophrenia risk loci and Ex neuron gene expression, and found that extending 25k around the transcribed region of each gene produced the most significant p-value for identifying Ex enrichment for schizophrenia (**Supplementary Figure 14**). We followed a similar approach for  $MAGMA<sup>16</sup>$  and ran the gene-level based trait analysis with multiple window sizes for schizophrenia GWAS<sup>18</sup>, and evaluated the p-value for cell-type-specific association of cortical Ex and In neuron populations from our study with schizophrenia common genetic risk (**Supplementary Figure 15**). As with LDSC above, the most significant p-value for both Ex and In neurons was found to be 25 kb upstream and downstream of each gene and thus we used a 25kb window size for MAGMA analyses for all traits. Furthermore, we evaluated how the samples displaying amyloid angiopathy may affect the results of cell-type association analysis (individual 3, **Supplementary Table 1**) and found consistent results in the LDSC association for all different traits with and without nuclei of this individual: **Supplementary Figure 16** (SN: R=0.82 (p <10<sup>-16</sup>) Cortex: R=0.93(p< p <10<sup>-16</sup>)  $^{16}$ )).

*Conditional analysis for multiple cell-types within and across regions:* Where LDSC identified multiple celltypes associated with the same trait, we used conditional analyses to evaluate whether it was the same genetic variants acting in different cell-types or distinct sets of genetic variants suggesting different cellular aetiologies. To test the association of cell-type A conditioned on cell-type B, we replaced the control gene set of all proteincoding genes expressed in both the cortex and nigral cell atlases by the cell-type B gene set and re-ran LDSC. When this conditional test produced a p-value associated with an LDSC Coefficient  $\leq 0.05$ , we considered that both cell-types could be involved independently in the same trait.

*Correction for the cell-heterogeneity between cortex and SN:* In the case of traits for which genetic variants were simultaneously associated with SN and cortical cell-type, potential bias was that the different proportion of glial versus neuronal cell-types between these two brain regions may lead to the identification of different gene markers for the same cell-type (e.g. microglia) and suggested erroneously two distinct cell-type signals. To discard this bias for the conditional analysis, we created SN and cortex cellular atlas including the same number of astrocytes, microglia/OPC/ODC/neurons (398 Astrocytes, 72 DaNs or Ex neurons, 123 GABA neurons or In neurons, 325 microglia cells, 160 oligodendrocyte precursor cells and 147 oligodendrocytes cells) by random sampling cells in the original cellular atlas and performed the same LDSC analyses as with the original cellular atlases. As the endothelial cells were only identified in SN and as we found no evidence of co-association for the endothelial cells with a cortical cell-type for no trait, we discarded the endothelial cells in the creation of these homogenous cellular atlases. Globally, we noted a good agreement with LDSC done with the original cell atlas (**Supplementary Figure 6**) (SN R=0.83 ( $p < 10^{-16}$ ), cortex R=0.72 ( $p < 10^{-16}$ )), nevertheless we noted a p-value inflation in the microglia cells identified in the cortex for SCZ and bipolar disorder. We then repeated the conditional LDSC analysis for traits with SN and cortical cell-type enrichment (**Supplementary Table 8**, column P-value (LDSC) (Matched Cellular Atlas)).

An RMarkdown document, including a version with R code to generate these gene sets and perform cell-type association, can be found here: https://github.com/csandorfr/SN\_Atlas.

## **Supplementary Note 4**

*Functional Analysis - Protein Protein Interaction network:* A combined protein-protein interaction network was created based on diverse resources: BioGRID 3.4 (accessed on September 2017)<sup>19</sup>, HitPredict (accessed on September 2017)<sup>20</sup>, IntAct (accessed on September 2017)<sup>21</sup>, STRING (accessed on September 2017, restricted to Homo sapiens and experimental scores higher than zero)<sup>22</sup>, CORUM (accessed on September 2017)<sup>23</sup> and Reactome (accessed on September 2017)<sup>24</sup>. The combined network consisted of a total of 20,591 genes and 1,973,967 interactions.

## **Supplementary Note 5**

*Nuclei extraction:* nuclei extraction was based on the protocol described by Krishnaswami *et al. 1* with some modifications. Briefly, 3mm<sup>2</sup> pieces of fresh-frozen brain tissue were placed in homogenization buffer (250 mM sucrose, 25 mM KCl, 5 mM MgCl<sub>2</sub>, 10 mM Tris Buffer pH 8.0, 1 mM DTT, 1X protease inhibitor, 0.4 U ul<sup>-1</sup> RNaseIN, 0.2 U ul<sup>-1</sup> Superasin, 0.1% v/v Triton X-100) for 10 minutes prior to mechanical disruption with a dounce homogenizer. Homogenate was filtered through a 35 mM cell strainer before concentrating nuclei by centrifugation 400 g 5 minutes. Nuclei were resuspended in FACS buffer (1 XPBS, 1 RNase-Free BSA, 0.2 U  $\mu$ <sup>-1</sup> of RNasin Plus RNase inhibitor, 10 ng m<sup>1-1</sup> Hoechst 33342) prior to FACS to obtain a population of nuclei free cellular debris.

*Single nuclei processing and 10x genomics library preparation:* nuclei were washed once in PBS and resuspended in 0.04% BSA in PBS for loading on the 10x Chromium 3' chip. Cell capturing, and library preparation was carried as per kit instructions (Chromium Single Cell Kit [v2 chemistry]). 3000 nuclei were targeted for capture per sample, after cDNA synthesis, 12-14 cycles were used for library amplification. The resultant libraries were size selected, pooled and sequenced using 75bp paired-end sequencing protocol on an Illumina HiSeq 4000 instrument. We used 12 samples for the 10X emulsions, with matched cortex and SN samples from 5 individuals with biological replicates for **substantia nigra** (**SN**) sample 4 (N4, N4B) and 5 (N5, N5B) run on two separate days (**Supplementary Table1**).

*10x sequencing data processing***:** reads were processed and mapped to the Human Genome (GRCh38.84 premrna) with Cell Ranger 2.1.1. We used default mapping arguments for cellranger "count" function to merge multiple sequencing runs for the same sample libraries. In order to have counts from both unspliced/pre-mRNAs and mature mRNAs, a custom human (GRCh38.84) "pre-mRNA" reference was created, whereby each gene transcript was annotated as an exon, so that pre-mRNA intronic reads can also be included in the UMI counts for each gene and barcode. The filtered gene/barcode matrices containing only the detected cell-associated barcodes and UMIs produced as a result were then further analysed.

## **Supplementary Note 6**

*Dimensionality reduction and cell clustering analysis:* Seurat R package (v2.3.4) was used to integrate different samples for the cortex and SN and then to perform a dimensionality reduction analysis. **Highly variable genes** (**HVGs**) were identified from a mean variability plot (average expression versus dispersion (variance/mean) assigned to 20 bins based on average expression) using a log(variance/mean) cutoff of 0.5 and maximum cut-off of 3.5 to identify outliers for each sample. Within each bin, a z-score of log-transformed dispersion measure (variance/mean) was calculated. A z-score cutoff of 0.5 was applied to identify the highly variable genes for each sample and the union of top 3000 of these from each sample resulted in 1031 and 2243 genes in the SN and cortex, respectively. A **canonical correlation analysis** (CCA)  $^2$  was performed and after discarding rare – nonoverlapping cells, the top numbers of CCA dimensions to align were identified by examining the correlation strength of all vectors through the biweight midcorrelation, a median based similarity metric saturation plot. This enabled to identify 25 and 42 canonical correlation vector dimensions to align for the SN and cortex.

The shared-nearest neighbour graph was constructed on a cell-to-cell distance matrix firstly by calculating the knearest neighbours with k=30 from the aligned CCA vectors. These were then used as input to the Louvain algorithm <sup>3</sup> with different clustering resolutions for the SN and cortex to identify cell-type clusters. We performed the analysis with different resolutions in Seurat by sequentially increasing the resolution from 0.4 to 2.4 and

assessing the clusters by constructing a phylogenetic tree for the averaged cluster populations at each resolution based on the HVG genes. For each branch of the subsequent tree a random forest classification error was calculated and an accuracy threshold of 80% was used to assess tree splits and cluster validity. This resulted in an optimal resolution parameter for the SN and cortex to be 0.4 and 0.8, respectively.

## **Supplementary Note 7**

*Cortex neuronal subtypes:* Lake et al<sup>4,5</sup> have previously identified subtypes of Ex and In neurons based on specific markers that we used here to further, identify nine subtypes of Ex neurons and ten subtypes of In neurons drawn from across different cortical layers.

Neuronal subtypes were resolved into Ex and In neurons based on known marker genes *SATB2, SLC17A7, GAD1 and GAD2* (**Supplementary Figure 4**, **Supplementary Data 3**). The Ex neurons could be further classified based on their layer positions and known marker genes. A population of Ex2 subtype expressing the L2/3/ *CUX2* & L4 specific (*RORB+* ) markers and *COL5A2+* was the most prevalent in our dataset (19%, **Supplementary Table 3**), followed by the second most prevalent, L2-3 subpopulation (*LAMP5+*), Ex1 (14%, **Supplementary Table 3**) showing *CBLN2+, CUX2+* expression. While we did identify layer 4 subpopulations, Ex3 (*RORB+, COL5A2-– ,IL1RAPL2+)* and Ex4 (*RORB+,TPBG+* ) like Lake *et al* <sup>5</sup> , (**Supplementary Figure 4**), we were not able to further distinctly resolve this sub-population into Ex3 (a-d) subtypes. We also, identified distinct subpopulations located in cortical L5*(PCP4+, ETV1+*), Ex5a (*PCP4+,HS3ST5-–* ), Ex5b (*PCP4-– ,HS3ST5+*) and Ex6a ( *HTR2C+,PCP4+,TLE4+*), with the last Ex6b (*HTR2C-–* , *TLE4+*,*SEMA3D+*) bordering on cortical L6 /6b or white matter(*NR4A2+*) (**Supplementary Figure 5**). Lastly, the least prevalent (0.59%) cortical neuronal subpopulation we identified was Ex8 (*TLE4+,NTNG2+)* localised to L6.

In neuron, the subtypes were discriminated from the other cell-types by expression of *GAD1* and *GAD2* marker genes (**Supplementary Figure 4**) and further resolved into 10 subpopulations based on enriched marker gene expression that showed distinct profiles of established In neuron marker genes (**Supplementary Data 3**). We resolved spatially distinct *CNR1* expressing L1-2 subpopulations as In1a (*RELN+*), In1b (*THSD7B+*), In1c ( *VIP+,TAC3+)* and In3 (*VIP+,TSHZ2+*), but not the upper layer In1d and In2 subtypes (**Supplementary Figure 4**). We further identified distinct *SV2C* expressing L2-3 (*LAMP5+*) subtypes as In4a (*COL5A2+*), In4b (*COL5A2-– ,EYA4+*) and In5 (*EYA4+ NOS1+*) (**Supplementary Figure 4**); layer 4/5 (*SULF1+*) located subtypes expressing *PVALB*, In6a (*RYR1<sup>+</sup>*)<sup>5</sup> and the more peripheral and the most prevalent (8.65%) In neuron subtype in our cortical cell atlas In6b (*TAC1+*) ; and finally *SST* expressing L6 (*SYNPR+*) subtype In8 (*STXBP6+*) but not the other *SST*  positive In7 sub-population <sup>5</sup> (**Supplementary Figure 4**).

## **Supplementary Note 8**

*Laser- capture microscopy cell-types in the SN:* Microarray profiles of laser-captured human dopaminergic neuron from control and PD patients were obtained from GEO (GSE24378)<sup>6</sup>. We downloaded the raw intensity values and performed a **Robust Multi-array Average** (**RMA**) transformation on probe-level with function RMA of R library oligo<sup>7</sup>, and computed the expression by gene by using the mean of expression probe measures, which were then log-transformed. Only the 9 control samples were used for the comparison.

Fastq files for RNA-seq data of **laser- capture microscopy** (**LCM**) astrocytes and microglia for a total of 18 samples for each cell-type in the SN<sup>8</sup> were downloaded from **National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site** (**NIAGADS**) (NG00057). The paired-end sequencing data for each sample were mapped to the human reference transcriptome (GRCh38; Ensembl release 91) and quantified transcript abundance counts were obtained with the tool Salmon (v  $0.11.0$ )<sup>9</sup>, using mapping based mode with default parameters, automatic library type inferring and additional options --seqBias,--posBias & --gcBias to account for sequence-specific biases, fragment -level GC biases and to account for 5' or 3' positional biases in the data. The transcript abundances were imported and summarized to gene-level using the R library tximport10. Only the protein-coding genes were used and data were filtered to include genes with >20 counts across all samples. The data were then normalised using the **variance stabilizing transformation** (**VST**) with vst function in the R library DESeq2<sup>11</sup>. Sample outliers were removed after visualising VST transformed data as PCA plots. Technical effects associated with the total number of genes were regressed out by using the removeBatchEffect function in the R library limma<sup>12</sup>. Finally, vst-transformed data for the control samples  $(n=9)$  for the LCM astrocytes and microglia were taken forward for comparison with the averaged log-transformed TPM 10x SN celltype expression data. We discarded protein-coding genes that were not expressed in all three datasets(10x SN, LCM DaN & LCM astrocytes and microglia) used to perform these comparisons: 11,703 protein-coding genes were used in the comparisons for 28 samples/cells.

Observing that the distributions of gene expression levels were different between the datasets even after the ComBat batch correction of the R library sva<sup>13</sup>, we applied a non-parametric approach by ranking the genes according to their expression level in each tissue/cell. From this rank matrix, we computed pairwise spearman correlations between LCM cell-types/cell and performed clustering analyses by using Ward's hierarchical approach and visualised the resulting matrix as a heatmap (**Supplementary Figure 12**).

*Single cell -RNA sequencing Temporal cortex:* For comparison with single-cell RNA-seq data from human **temporal cortex** (**TC**) 14, we obtained gene count data from GEO (GSE67835) and converted these to TPM counts and restricted the analysis to protein-coding genes ( $n= 15,081$ ) and cells ( $n= 465$ ), maintaining the cell-type annotations from the original publication<sup>14</sup>. The expression matrix was then normalized and linearly regressed for batch identity with Seurat as described above, using a minimum cutoff of 1,000 protein-coding genes detected in at least three cells. HVG were identified from a mean variability plot (average expression versus dispersion (variance/mean) assigned to 20 bins based on average expression) using a log(variance/mean) cutoff of 0.5 to identify 3,662 genes. Moreover, differentially expressed genes associated with each cell-type (excluding the hybrid cell-type) were identified by Seurat, using parameters and thresholds as above. Transcriptional Pearson correlations were calculated in a pairwise manner across the averaged cell-type populations from the external and 10x datasets, grouped via Ward's hierarchical clustering and visualised as a heatmap (**Supplementary Figure 12)**. The final comparison was restricted to log-transformed averaged cell-type TPM expression values of 6,446 genes, composed of HVG (protein-coding only) from both datasets along with the top 5 enriched genes (logFC > 2) in each cell-type.

*Single nuclei -DropSeq visual and frontal cortex:* We obtained the UMI counts for the **visual cortex** (**VC**) and **frontal cortex** (**FC**) single nuclei drop-seq data from GEO (GSE97930) and restricted the analysis to proteincoding genes ( $n= 10,919$ ) and annotated nuclei ( $n= 29,441$ ) maintaining cell-type annotations from the original publication<sup>5</sup>. As above, the expression matrix was then loaded into Seurat and nuclei with fewer than 300 UMI counts or more than 5000 UMI counts were omitted and further filtering to only retain nuclei with a minimum of 200 genes detected in at least 1% of cells was also, performed. We normalized and linearly regressed for technical effects associated UMI coverage and batch identity like the original publication with Seurat as above. HVGs were identified from a mean variability plot using a log(variance/mean) cutoff of 0.5 to identify 2,181 genes. Moreover, differentially expressed genes associated with each cell-type were identified by Seurat, using parameters and thresholds as above. Transcriptional Pearson correlations were calculated in a pairwise manner across the averaged log-transformed cell-type and subtype populations expression values from the external and 10x cortex dataset, grouped via Ward's hierarchical clustering and visualised as a heatmap (**Supplementary Figure 12**). The analysis was restricted to 3,614 genes composed of HVG (protein-coding only) from both datasets and top 5 differentially expressed genes ( $logFC > 2$ ) defining each cell-type.

## **Supplementary Figures**



**Supplementary Figure 1: Covariate distributions in the substantia nigra (SN) cell atlas.**

**a)** T-distributed stochastic neighbor embedding (tSNE) plot for all non-neuronal (astrocytes, endothelial, microglia, ODC and OPC) and neuronal (DaNs and GABA) cell-types or subtypes across the SN**. b)** barplot plot showing batch proportions per level-2 cell-type**. c)** barplot as in (b) showing sample proportions per level-2 celltype. **d)** Violin plot showing unique molecular identifier (UMI) depth across all level-2 cell-types. **e)** Violin plot as in (d) showing gene coverage across all level-2 cell-types. In (d) and (e) the black dot denotes the median UMI and genes for the cell-type. DaNs: Dopaminergic neurons; GABA: GABAergic neurons; ODC: Oligodendrocytes; OPC: Oligo-precursor-cells.



### **Supplementary Figure 2: Violin plots of DaN and GABA neuron markers in the SN neurons.**

Violin plots of expression values (log10 TPM values) for known DaN and GABA markers. Differential expression p-values after Bonferroni correction(p.adj) using the two-sided negative binomial test in Seurat were calculated between the two neuronal populations (Methods & data S3). In each of the violin plots the black dot denotes the mean expression for the marker genes for the cell-type. DaNs: Dopaminergic neurons; GABA: GABAergic neurons.



**Supplementary Figure 3: Covariate distributions in the cortex cell atlas.** 

**a)** T-distributed stochastic neighbor embedding (**tSNE**) plot for all non-neuronal (astrocytes, microglia, ODC and OPC), Ex and In neuron cell-types or subtypes across the cortex. **b)** barplot plot showing batch proportions per level-1 cell-type. **c)** barplot as in **(b)** showing sample proportions per level-1 cell-type. **d)** Violin plot showing unique molecular identifier (**UMI**) depth across all level-1 cell-types**. e)** Violin plot as in **(d)** showing gene coverage across all level-1 cell-types. In **(d)** and **(e)** the black dot denotes the median UMI and genes for the celltype. Ex:Excitatory;In:Inhibitory ;ODC: Oligodendrocytes;OPC: Oligo-precursor-cells.



#### **Supplementary Figure 4: Cortex cell-types:**

(a) Violin plots of expression values (log10 TPM values) of cell-type specific markers for the cell-types in the cortex. (b) Violin plots of expression values (log10 TPM values) of cell-type and cortical layer-specific enriched marker gene profiles<sup>6</sup> for the Excitatory neurons subtypes. (c) Violin plots of expression values (log10 TPM values) of cell-type and cortical layer-specific enriched marker gene profiles for the Inhibitory neuron subtypes. Ex: Excitatory; In: Inhibitory, ODC: Oligodendrocytes; OPC: Oligo-precursor cells, L1-2; Layer 1/2, L2-3; Layer 2/3,Layer 2-4; Layer 2/3/4, L2 ; Layer2, L3; Layer 3, L4; Layer 4, L4-5; Layer 4/5, L6; Layer 6; Layer 6-6b; Layer 6/6b



**Supplementary Figure 5: Only the minor cell populations show difference in the frequency between the individuals.**

Cell composition (left) & contribution (right) of each individual within SN (top) and cortex (bottom) single nuclei atlas. (Left graph) within each atlas, we calculated the Pearson residual reflecting an under (red) or overrepresentation (blue) of a specific cell-type for an individual. (Right graph) we calculated the % represented by chi2 residual reflecting whether an individual contributes more to a cell-family category than another individual. The size of the circle is proportional to the absolute value of the Pearson residual or the % represented by chi<sup>2</sup> residual.



**Supplementary Figure 6: Comparison of cell-type enrichment analyses performed with LDSC for different traits between the cellular atlases with matched proportions of cell types between cortex and SN and the original cellular atlases**

From the original cellular atlas, we created SN (A) and cortex (B) cellular atlas including the same number of astrocytes microglia/OPC/ODC/Neurons by random sampling cells in the original cellular atlas and performed the same LDSC analyses as with original cellular atlases (Methods). For a specific cell-type, the x-axis and y-axis of each plot represent the -log 10 p-value for enrichment of genetic variants associated with human trait by using the original cellular atlases and the matched cell family proportion cellular atlases respectively.



**Supplementary Figure 7 : Expression of known Parkinson's disease genes in SN level-1 cell-types.** 

Violin plots of expression values (log10 TPM values) of Parkinson's Disease (PD) risk genes<sup>26</sup> for the cell-types in the SN.



**Supplementary Figure 8: The microglia-specific signals for the genetic variants associated with HDL cholesterol level and AD correspond to distinct genetic aetiologies.**

LDSC analysis by using the AD and HDL GWA summary statistic adjusted (green) or not (red) for HLD and AD risk effect respectively to evaluate the cell-type association within microglia in the cortex (top) and SN (bottom). Each bar represents the -log 10 p-value of enrichment genetic variants associated (p-value associated with an LDSC Coefficient) with AD or HDL (adjusted or not for the other trait) for microglia cell-type identified in the cortex or SN.



**Supplementary Figure 9: Convergence of PD-risk (a), SCZ-risk (b) and BP-risk (c) across SN and cortex cell-type specific gene modules.**

We assessed SCZ-, PD- and BP- risk enrichment in cell-type specific gene modules from cortex and SN through MAGMA gene set analysis (one -sided positive two -sample t-test). Barplots represent -log<sub>10</sub> p-values of enrichment, vertical lines correspond to p-value=0.05 and star symbols (\*) indicate significant q value (Bonferroni correction for the number of cell type PPI modules tested). Only the associations to modules within cell-types that showed significance from the more general analysis (Figure 2) were discussed further.



#### GTEx brain tissue associations for complex traits

#### **Supplementary Figure 10: Tissue-trait associations across diverse human complex traits using GTEx bulk brain tissue transcriptomic data.**

Two classical approaches: stratified LD score regression (LDSC) (p-value associated with an LDSC Coefficient) and the MAGMA gene set analysis (one -sided positive two -sample t-test) have been used to identify the association between multiple complex traits genome wide association signals within 13 brain regions from bulk GTEx transcriptomic data. The heatmap colour indicates if the p-value of cell-type enrichment for a trait is significantly associated with both methods or either methods after Bonferroni multiple test correction(\*\*) and also, if nominally significant ( $p \le 0.05$ ) in the other method (\*). Bonferroni threshold was P-value/No.of.celltypes. The different traits have been clustered by category: Cognitive phenotype (Cog.), Autoimmune disease (Immune), Metabolic, Cardiovascular and Anthropometric Traits (Metabolic/Cardio/Anthropometric) Neurological disorder, Psychiatric disorder.



**Supplementary Figure 11: Evaluation of the similarity of Substantia nigra (SN) biological replicates from the same individuals.** 

a) T-distributed stochastic neighbor embedding (tSNE) plot for all non-neuronal (astrocytes, microglia, ODC and OPC) and neuronal (DaNs and GABA) cell-types or subtypes across biological replicates N5 and N5B in the SN. b) t-SNE plot as in (a) showing sample identity N5 and N5B (**Supplementary Table 1**). c) t-SNE as in (a) showing cell-types for N4 and N4B SN biological replicates (**Supplementary Table 1**). d) t-SNE as in (c) showing sample identity N4 and N4B (**Supplementary Table 1**). All major cell-types were identified in both sets of replicates from each individual. In both sets of replicates a distinct cluster of endothelial cell-type was not identified and in N5&N5B a small cluster of N=3 was not attributed to any particular cell-type. DaNs: Dopaminergic neurons; GABA: GABAergic neurons; ODC: Oligodendrocytes; OPC: Oligo-precursor-cells ; N4: Nigra 4;N4B;Nigra 4B;N5:Nigra 5; N5B: Nigra 5B.



#### **Supplementary Figure 12: Comparison of Substantia nigra and cortex (10x) cell-types with public transcriptomic datasets**

a) SN laser- capture microscopy (LCM) Cell-types. A Spearman correlation heatmap comparing averaged SN level-2 cell-type populations with LCM SN Dopaminergic neurons (LCM\_DaN\_Ctrl) <sup>6</sup>, microglia (Ctrl\_Mic) and astrocytes (Ctrl\_Ast) <sup>8</sup> from control samples. b) Cortex single cell RNA-seq data. Pairwise Pearson correlation comparison of the averaged 10x cortex and SN neuronal and non-neuronal subtype populations with the averaged temporal cortex (TC) single cell RNA-seq type populations c) Cortex single nuclei Drop-seq data. Pairwise Pearson correlation comparison of the averaged 10x cortex neuronal and non -neuronal subtype populations with the averaged cell-type populations in the frontal cortex (FC) and visual cortex (VC) of this single nuclei dataset.



#### **Supplementary Figure 13: LDSC vs MAGMA spearman rank correlation plot for cell-type associations ( log10 pvalues) in the SN and cortex cell atlas.**

The clustered Spearman rank correlations between the rank orders of cell-type associations(-log10 Pvalue) for LDSC and MAGMA across traits for the a) SN and b) cortex cell-types are shown. A moderate to high positive correlation was observed for neurological disorders such as PD and ALS, for psychiatric disorders such as Schizophrenia, ADHD, Autism Disorder, Major Depression and Insomnia. Moreover non-brain metabolic traits such as Height, BMI and Cholesterol (HDL & LDL) and immune traits (Crohn's disease) were also, highly concordant between the two methods.



#### **Supplementary Figure 14: Selection of window size parameters around each transcribed region in the LDSC cell-type analysis.**

Results of the LDSC cell-type analysis with different windows size (5kb, 10 kb, 25 kb, 50kb,75 kb and 100 kb) to evaluate the association between schizophrenia and Excitatory neurons and Inhibitory neurons in the cortex. The x-axis represents the windows size added on either side of each gene, while the y-axis represents -log10 pvalue associated with LDSC Coefficient.



#### **Supplementary Figure 15: Selection of window size parameters around each gene region for the MAGMA gene set based cell-type analysis.**

Heatmap showing the -log<sub>10</sub> p value of the one-sided positive; two -sample t-test based cell-type association analysis for Schizophrenia for the adult cortex single nuclei cell-types identified in this study at different incremental window sizes (5kb,10kb,25kb,50kb,75kb & 100kb). The window size was benchmarked on the schizophrenia p-value for Excitatory neurons (Ex) and Inhibitory neurons (In). Each cell has been labelled with an  $*$  if the -log10 p-value is  $>=$  to Bonferroni significance threshold (-log<sub>10</sub>( 0.05/No of Cell-types)). SCZ: Schizophrenia.



**Supplementary Figure 16: Comparison of cell-type enrichment (-log10 pvalues) for different traits between the original cell atlas and the reduced cell atlas (no Sample 3).**

The plot compares the cell-type association enrichments  $(-log10(pvalue))$  between the cell atlases with and without the individual with amyloidosis (Sample 3, Supplementary Table1) in the a) SN and b) cortex with the LDSC method in each cell-type. For a specific cell-type, the x-axis and y-axis of each plot represent the -log 10 p-value of enrichment genetic variants associated with human trait by using the original cellular atlases and the reduce cell cellular atlases respectively.

## **Supplementary Tables**

# **Supplementary Table 1: Single- nuclei RNA sequencing experimental design and sample information**

![](_page_22_Picture_530.jpeg)

**\*PMI = Post-mortem interval**

**Supplementary Table 2: Gene Ontology pathway analysis of two distinct sub-populations of astrocytes in the Substantia nigra**

![](_page_23_Picture_373.jpeg)

**B: Astrocyte -2 :Top GO Terms: Biological process for Up regulated genes in Astrocytes -2 vs all other cell types** 

![](_page_23_Picture_374.jpeg)

 $\bar{1}$  $\bar{1}$  $\bar{1}$ 

 $\bar{1}$ 

## **Supplementary Table 3: Substantia nigra and cortex cell population proportions**

**A: Substantia nigra**

![](_page_24_Picture_683.jpeg)

**B: Cortex**

![](_page_24_Picture_684.jpeg)

 $\overline{\phantom{a}}$ 

![](_page_25_Picture_341.jpeg)

**Supplementary Table 4: Source of genome-wide association studies meta-analysis summary statistics used in this study**

#### **Supplementary Table 5 : Risk association p-values for 30 brain and non-brain-related traits with cell-types from the Substantia nigra (SN) using LDSC & MAGMA**

![](_page_26_Picture_1254.jpeg)

![](_page_27_Picture_1342.jpeg)

![](_page_28_Picture_1387.jpeg)

![](_page_29_Picture_1371.jpeg)

![](_page_30_Picture_691.jpeg)

![](_page_31_Picture_297.jpeg)

**Supplementary Table 6: Conditional cell-type analyses to evaluate whether multiple significant Substantia nigra cell-type signatures associated with the same trait coincide or not with the same genetic variants.** 

#### **Supplementary Table 7: Risk association p-values for 30 brain and non-brain-related traits with cell-types from the Cortex using LDSC & MAGMA**

![](_page_32_Picture_817.jpeg)

![](_page_33_Picture_882.jpeg)

![](_page_34_Picture_882.jpeg)

![](_page_35_Picture_882.jpeg)

### *Single Nuclei SN Cortex*

![](_page_36_Picture_141.jpeg)

**Supplementary Table 8: Conditional cell-type analysis to evaluate whether a significant Microglia/OPC/ODC/Neurons detected simultaneously in the SN & cortex with the same trait coincide or not with the same genetic variants.**

![](_page_37_Picture_318.jpeg)

#### **Supplementary Table 9: Risk association p-value for 30 brain and non-brain-related traits with GTEx brain region tissue using LDSC & MAGMA.**

![](_page_38_Picture_798.jpeg)

![](_page_39_Picture_841.jpeg)

![](_page_40_Picture_841.jpeg)

![](_page_41_Picture_841.jpeg)

![](_page_42_Picture_841.jpeg)

![](_page_43_Picture_841.jpeg)

![](_page_44_Picture_841.jpeg)

![](_page_45_Picture_841.jpeg)

![](_page_46_Picture_264.jpeg)

## **Supplementary References**

- 1. Krishnaswami SR*, et al.* Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. *Nat Protoc* **11**, 499-524 (2016).
- 2. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).
- 3. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks %U http://stacks.iop.org/1742-5468/2008/i=10/a=P10008. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008).
- 4. Lake BB*, et al.* Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586-1590 (2016).
- 5. Lake BB*, et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36**, 70-80 (2018).
- 6. Zheng B*, et al.* PGC-1alpha, a potential therapeutic target for early intervention in Parkinson's disease. *Sci Transl Med* **2**, 52ra73 (2010).
- 7. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* **26**, 2363-2367 (2010).
- 8. Mastroeni D*, et al.* Laser-captured microglia in the Alzheimer's and Parkinson's brain reveal unique regional expression profiles and suggest a potential role for hepatitis B in the Alzheimer's brain. *Neurobiol Aging* **63**, 12-21 (2018).
- 9. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419 (2017).
- 10. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521 (2015).
- 11. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
- 12. Ritchie ME*, et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
- 13. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882-883 (2012).
- 14. Darmanis S*, et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci U S A* **112**, 7285-7290 (2015).
- 15. Finucane HK*, et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* **50**, 621-629 (2018).
- 16. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* **11**, e1004219 (2015).
- 17. Skene NG*, et al.* Genetic identification of brain cell types underlying schizophrenia. *Nat Genet* **50**, 825- 833 (2018).
- 18. Pardinas AF*, et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet* **50**, 381-389 (2018).
- 19. Chatr-Aryamontri A*, et al.* The BioGRID interaction database: 2017 update. *Nucleic Acids Res* **45**, D369- D379 (2017).
- 20. Lopez Y, Nakai K, Patil A. HitPredict version 4: comprehensive reliability scoring of physical proteinprotein interactions from more than 100 species. *Database (Oxford)* **2015**, (2015).
- 21. Orchard S*, et al.* The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* **42**, D358-363 (2014).
- 22. Szklarczyk D*, et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**, D447-452 (2015).
- 23. Ruepp A*, et al.* CORUM: the comprehensive resource of mammalian protein complexes--2009. *Nucleic Acids Res* **38**, D497-501 (2010).
- 24. Fabregat A*, et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res* **44**, D481-487 (2016).