The Foundational data initiative for Parkinson's disease (FOUNDIN-PD): enabling efficient translation from genetic maps to mechanism

Elisangela Bressan 1*, Xylena Reed 2*, Vikas Bansal 1*, Elizabeth Hutchins 3, Melanie M. Cobb 4, Michelle G Webb 5, Eric Alsop 3, Francis P. Grenn 2, Anastasia Illarionova 1, Natalia Savytska 1, Ivo Violich 5, Stefanie Broeer 1, Noémia Fernandes 1, Ramiyapriya Sivakumar 5, Alexandra Beilina 2, Kimberley Billingsley 2, Joos Berghausen 2, Caroline B. Pantazis 2, Bessie Meechoovet 3, Rebecca Reiman 3, Amanda Courtright-Lim 3, Amber Logemann 3, Jerry Antone 3, Mariya Barch 4, Robert Kitchen 6, Yan Li 7, Clifton L. Dalgard 8,9, The American Genome Center, Patrizia Rizzu 1, Dena G Hernandez 2, Brooke E. Hjelm 5, Mike Nalls 2, 10, 11, J. Raphael Gibbs 2, Steven Finkbeiner 4,12^, Mark R Cookson 2^, Kendall Van Keuren-Jensen 3^, David W Craig 5^, Andrew B Singleton 2,10^, Peter Heutink 1^, Cornelis Blauwendraat 2^
*joint first authors
^shared last and corresponding

1) German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany
2) Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA
3) Division of Neurogenomics, The Translational Genomics Research Institute, Phoenix, AZ, USA
4) Center for Systems and Therapeutics, Gladstone Institutes, San Francisco, CA, USA
5) Department of Translational Genomics, Keck School of Medicine, University of Southern California, 1450 Biggy Street, Los Angeles, CA, USA
6) Massachusetts General Hospital, Cardiovascular Research Center, Charlestown, MA, USA
7) Protein/peptide Sequencing facility, National Institute of Neurological, Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA
8) The American Genome Center, Uniformed Services University of the Health Sciences, Bethesda, MD, USA
9) Department of Anatomy, Physiology & Genetics, Uniformed Services University of the Health Sciences, Bethesda, MD, USA
10) Center for Alzheimer's and Related Dementias, National Institutes of Health, Bethesda, MD, USA
11) Data Tecnica International, Glen Echo, MD, USA
12) Departments of Neurology and Physiology, University of California, San Francisco, San Francisco, CA, USA.

Corresponding authors:
Steven Finkbeiner (Gladstone), steve.finkbeiner@gladstone.ucsf.edu
Mark R Cookson (NIH-NIA), cookson@mail.nih.gov
Kendall Van Keuren-Jensen (TGEN), kjensen@tgen.org
David W Craig (USC), davidwcr@usc.edu
Andrew B Singleton (NIH-NIA), singleta@mail.nih.gov
Peter Heutink (DZNE), Peter.Heutink@dzne.de
Cornelis Blauwendraat (NIH-NIA), cornelis.blauwendraat@nih.gov

Keywords:
Parkinson's disease, genetic risk, induced pluripotent stem cell, dopaminergic neurons, omics

Abstract

In the FOUNdational Data INitiative for Parkinson's Disease (FOUNDIN-PD) we sought to produce a multi-layered molecular dataset in a large cohort of 95 Induced pluripotent stem cells (iPSC) lines at multiple timepoints during differentiation to dopaminergic (DA) neurons, a major affected cell type in Parkinson's Disease (PD). The lines are derived from the Parkinson's Progression Markers Initiative (PPMI) study that includes both people with PD and unaffected individuals across a wide range of polygenic risk scores (PRS) with both risk variants identified by genome-wide association studies (GWAS), and monogenic causal alleles. We generated genetic, epigenetic, regulatory, transcriptomic, proteomic, and longitudinal cellular imaging data from iPSC-derived DA neurons to understand key molecular relationships between disease associated genetic variation and proximate molecular events in a PD relevant cell-type. Analyses of all data modalities collected in FOUNDIN-PD suggest that the differentiation to DA neurons, while not fully mature, was successful and robust. Interrogation of PD genetic risk in this relevant cellular context may elucidate the functional effects of some of these risk variants alone or in combination with other variants. These data reveal that DA neurons derived from human iPSC provide a valuable cellular context and foundational atlas for modeling PD-related genetic risk. In addition to making the data and analyses for this molecular atlas readily available, we have integrated these data into the browsable FOUNDIN-PD data portal (https://www.foundinpd.org) to be used as a resource for understanding the molecular pathogenesis of PD.

Introduction

Our understanding of the genetic architecture of Parkinson's disease (PD) has expanded considerably over the last decade. Investigations of rare monogenic forms of PD and parkinsonism have revealed multiple genes that contain disease-causing mutations [1]. Iterative application of genome wide association studies (GWAS) in increasingly large sample sizes has identified 90 independent risk variants for PD, which cumulatively contribute a significant amount of risk for disease [2]. A number of the genetic risk factors identified by GWAS also influence overall PD risk in carriers of *LRRK2* or *GBA* mutations [3,4], which are the most common genetic causes of PD. In addition, multiple GWAS-nominated loci include genes implicated in monogenic forms of PD, highlighting a clear etiologic link between familial and sporadic disease. Thus, understanding the molecular mechanisms of the currently known genetic risk factors and mutations will provide compelling actionable insight on the biology of disease risk, onset, progression, and modifiers of disease.

While the pace of discovery for genetics has increased dramatically in recent years, our ability to characterize the associated function and dysfunction of nominated genes and risk loci has not matched this progress. Research centered on the biology of genes that contain rare disease-causing mutations has revealed important insights into the molecular mechanisms leading to disease; however, it is more challenging to demonstrate how risk loci identified by GWAS may lead to disease. Generally these loci nominate genomic regions, not specific genes, disease effect sizes are modest, the cellular context is unknown, and the genetic mediator is unlikely to be protein coding. These factors combine to make studying the biology of GWAS loci in traditional cellular and animal models extremely challenging. This challenge, coupled with the sheer number of known and to-be-discovered risk loci, require an alternative strategy to understand the biology underlying risk variants. The development of cellular models based on human induced pluripotent stem cells (iPSC) provides a unique opportunity to address the collective impact of genetic risk factors, provides a platform to define the relevant cellular context for modeling these variants, and a space to investigate their proximal molecular effects at scale.

Since the effect sizes of individual risk factors are generally moderate and our knowledge of the function and interactions of many genes at GWAS loci is often limited, it is essential to study their effects in an unbiased fashion. Molecular, cellular and genomic methods that can quantify epigenetic, regulatory, transcriptomic, proteomic and cellular alterations have the potential to provide us with an atlas that describes coordinated molecular and cellular changes. When such maps are generated across cells from varied genetic backgrounds, they can reveal the consequences of genetic variation on complex processes and how these consequences are interrelated. Combining iPSC approaches with quantitative molecular assays provides the capacity to assess genes of interest and risk loci at scale within a disease relevant cellular context and provides an unprecedented opportunity for insights into the pathogenesis of PD.

We therefore created the FOUNdational Data INitiative for Parkinson's Disease (FOUNDIN-PD; https://www.foundinpd.org/). One of the main pathological hallmarks of PD is the progressive degeneration of dopaminergic neurons in the substantia nigra and the accumulation of alpha-synuclein protein aggregates, known as Lewy bodies and Lewy neurites [5]. Additionally, our and others' work have highlighted that genetic risk in PD is likely to be

active in a dopaminergic (DA) neuronal context [2,6]. Thus, we focused on the production of a large series of iPSC lines, driven to a dopaminergic neuronal cell type, from which a host of genetic, epigenetic, regulatory, transcriptomic, and cellular data were collected (Figure 1). All iPSC lines are derived from subjects within the Parkinson's Progression Marker Initiative (PPMI; https://www.ppmi-info.org/) [7]. We describe here the production and characterization of the first release of the FOUNDIN-PD data, the creation of a portal for data access and analysis, and provide evidence that this system represents a relevant cellular context to investigate PD related risk alleles. This represents the largest multi-omic iPSC-derived DA neuron dataset to date. Lastly, we discuss what opportunities and challenges these data have revealed for the next stages of FOUNDIN-PD.

Results

FOUNDIN-PD overview

The basis of FOUNDIN-PD is the generation of molecular readouts from 95 iPSC lines driven to a dopaminergic neuronal state (Figure 1, Supplemental Table 1). These lines were available as a part of PPMI. The PPMI study is a landmark longitudinal study that has collected data from more than 1,400 individuals at 33 sites in 11 countries; the study contains a wealth of clinical, imaging, and biomarker data. From the PPMI iPSC collection we included lines derived from healthy controls, idiopathic PD patients, and individuals carrying known disease-linked mutations.

Full genome sequence data were available for all donors, thus we were able to not only identify subjects with damaging mutations in *LRRK2* (p.G2019S, n=25 and p.R1441G, n=1), *GBA* (p.N370S, n=20), and *SNCA* (p.A53T, n=4) (hereafter refer to these variants as LRRK2+, GBA+, and SNCA+, respectively) (Table 1), but also those with high and low polygenic risk scores (Supplemental Figure 1, Supplemental Table 1). These 95 iPSC lines were differentiated into dopaminergic (DA) neurons using a well-established protocol [8], with minor modifications Figure 2A [9], and an automated cell culture system [10].

Small molecule-based differentiation protocols produce an enriched culture of neurons with variable numbers of TH positive (DA neuron) cells depending in part on the iPSC line used as the starting point. To estimate the total number of neurons and the fraction of DA neurons produced by the 95 iPSC lines included in this study, we performed immunocytochemistry (ICC) using neuron (MAP2, Microtubule-Associated Protein 2) and DA neuron (TH, Tyrosine Hydroxylase) markers followed by high-throughput imaging and quantification of the percentages of positive cells at day 65 of the neuronal differentiation. Up to 30 fields were imaged and at least 1000 cells per iPSC line were analysed. Quantification of MAP2 and TH positive cells revealed that on average 80% (range 52-93%) of the cells were converted to neurons and 20% of the cells (range 4-42%) expressed TH (Figure 2B; Supplemental Figure 2). The average proportion of TH positive cells in the iPSC lines were highly similar when assessed by two independent TH antibodies (Supplemental Figure 3).

In order to measure how robust and reproducible the iPSC to neuron differentiation protocol was using our automated system, we included a control line in each batch as a technical replicate (n=5). The percentage of MAP2 and TH positive cells obtained from the control cell line using ICC across all five differentiation batches was very consistent (Supplemental Figure

4), and no significant differences in total neurons or TH and MAP2 positive neurons between batches were identified (p>0.2 for both).

Quantifying gene expression in FOUNDIN-PD data using RNA sequencing and proteomics

To further characterize the iPSC-derived cells, we generated a wealth of data types including genetic, epigenetic, regulatory, transcriptomic, proteomic and cellular imaging data (Figure 1). To characterize the identity of the cells generated by the iPSC differentiation protocol used in the present study, we performed single-cell RNA sequencing (scRNA-seq) on the majority of the day 65 cell lines (n=79+4 control replicates, 84% of total included iPSC). In total, 416,216 high quality cells were retained with an average of 5,015 cells per sample (range 584 to 9,640). After evaluation, seven distinct broad cell types were identified across all samples including: early neuron progenitors (131,251 cells, 32% of total), late neuron progenitors (113,425 total, 27% of total), dopaminergic (DA) neurons (96,623 total, 23% of total), immature DA neurons (41,267 total, 10% of total), proliferating floor plate progenitors (18,984 total, 5% of total), neuroepithelial-like cells (8,979 total, 2% of total) and ependymal-like cells (5,687 total, 1% of total) (Figure 2C, Supplemental Table 2). Expression of *TH*, *MAP2* and *SNCA* was clearly enriched in the neuronal cell types (Figure 2D). A number of well-known marker genes were used to annotate the cell types identified using the scRNA-seq data (Supplemental Figure 5).

Next, we compared our identified cell type populations to public datasets from human post mortem substantia nigra [11] and human iPSC-derived DA neuron subtypes using a slightly modified DA neurons differentiation protocol and distinct set of iPSC cell lines [12]. The DA neuron population identified in our data showed the highest correlation (Spearman's R=0.69) with the TH positive neuron cluster found in the human substantia nigra, as represented by both groups clustering together (Figure 2E and Supplemental Figure 6A-B). The second highest correlation was observed between our immature DA neurons and the TH positive neuronal cluster from Agarwal et al (Spearman's R=0.67) Supplemental Figure 6A-B). Both immature and DA neurons were highly correlated to the iPSC-derived dopaminergic neuron subtypes (DAn1-4) identified by Fernandes and collaborators [12] (Supplemental Figure 6C), which were produced with a similar iPSC to DA neuron differentiation protocol. Another similarity detected between both iPSC-derived neurons datasets was the expression of serotoninergic markers in immature DA neurons (FOUNDIN-PD, Supplemental Table 2) and DAn2 [12]. Importantly, combining ICC-based estimates of neurons (MAP2 positive cells) and DA neurons (TH positive cells) with the percentage of positive cells obtained from scRNA-seq data, showed high correlations (Pearson correlation of R=0.8562, P<0.0001 and R=0.8916, P<0.0001 for TH (Pel-Freeze) and MAP2, respectively (Figure 2F-G). Similar results were obtained with a second TH (Millipore) antibody (Supplemental Figure 7). Although the differentiation efficiency (percentage of each cell type) varied between cell lines (Figure 2H), no consistent cell type enrichment could be identified based on batch, phenotype, recruitment category, genetic sex and PD-linked genotype (GBA+, LRRK2+, SNCA+) (Supplemental Figure 8). Additionally, very high correlation was observed (R>0.9) in the technical replicates (n=4) using gene-level scRNA-seq data of the identified DA neuron cluster (Supplemental Figure 9) and in total TH and MAP2 levels (Supplemental Figure 4), suggesting that while there is a variation in differentiation efficiency, this is likely not caused by the robustness of the differentiation protocol but seems to be partially due to an inherent characteristic of each line.

Bulk RNA-seq was generated at day 0 (n=99), day 25 (n=98) and day 65 (n=96) with each timepoint including five technical replicates. A principal component analysis (PCA) of bulk RNA-seq showed clear clustering by timepoint (Figure 3A). Similarly as in the scRNA-seq data, we observed a very high correlation (R>0.9) in the gene level correlations of each timepoint of the bulk RNA-seq (n=5) (Supplemental Figure 10A-C). Evaluation of the bulk RNA-seq further for overall transcription differences across all timepoints shows clear enrichment signatures that correlate with neuron-like features including synapse assembly, neurotransmitter transport and action potential (Supplemental Table 4) at timepoints day 25 and day 65. Additionally, specific genes of interest such as *MAP2* and *TH* (Figure 3B-C), *GBA*, *SNCA*, *LRRK2* and *SYN1* (Supplemental Figure 11A-D) increase their levels of expression as cells are differentiated. Concurrently, iPSC-associated genes such as *POU5F1* (Figure 3D), *NANOG* and *TDGF1* (Supplemental Figure 11E-F) showed significant reductions in expression at later timepoints relative to day 0, which correlated with a decrease in pathway signatures of iPSC differentiation and growth, including somatic stem cell population maintenance and positive regulation of cell population proliferation (Supplemental Table 5).

Next, we used the day 0 bulk RNA-seq gene expression data to predict DA neuronal differentiation efficiency. We defined DA neuronal differentiation efficiency as the fraction of DA neurons in our scRNA-seq datasets at day 65 using a similar method described by Jerber and collaborators [13]. Using logistic regression, ten genes were identified that have a non-zero coefficient and predict good neuronal differentiation efficiency with an area under the curve (AUC) = 0.93 and 0.87 accuracy (95% confidence interval [0.78, 0.93]) (Supplemental Figure 12A-C). Repeated 5-fold cross-validation achieved a mean AUC of 0.85 with 0.03 standard deviation. Out of these ten genes that have a non-zero coefficient, five were significantly correlated with neuronal differentiation efficiency (FDR < 1%, Figure 3E-I and Supplemental Figure 12D).Three (*HNRNPH3*, *SRSF5* and *HSD17B6*) of these associated genes are positively correlated with neuronal differentiation efficiency. Moreover, the expression of these genes is significantly reduced as iPSC are differentiated to DA neurons (adjusted p-value < 0.05 from day 0 to day 65, Figure 3J), suggesting that their high expression in iPSC may represent a higher differentiation potential. *SRSF5* has previously been associated with neuronal differentiation efficiency (R=0.25, adjusted p-value of 0.013 [13]. A separate study found that SRSF5 binds to pluripotency-specific transcripts and positively correlates with the cytoplasmic mRNA levels of pluripotency-specific factors [14]. The remaining two associated genes (*ZSWIM8* and *ARSA*) are negatively correlated with neuronal differentiation efficiency and their overall expression is significantly increased during differentiation (adjusted p-value < 0.05 from day 0 to day 65, Figure 3J).

We also performed proteomics profiling in 10 iPSC lines differentiated into DA neurons, as it provides unique insight into which transcripts are translated and allows for relative quantitation of proteins between samples. Proteomics was generated only for batch 1 iPSC lines at day 0 (n=10) and 65 (n=10) due to the large number of cells required as input for this assay. Data were generated with three replicates per sample at each timepoint and two separate runs (day 0 and day 65). MAP2 expression was robustly detected at day 65, however TH and other makers used here in scRNA-seq analysis (*ATP1A3*, *ZCCHC12*, *SYT1* and *SNAP25*) were not detected, likely due to technical limitations of the assay. The sparsity of the data, and the fact that day 0 and day 65 samples were run in separate experiments, made timepoint comparisons computationally difficult. When comparing the MAP2 protein expression data

with bulk RNA-seq *MAP2* expression from the same cells, we see a significant correlation across the ten lines (Supplemental Figure 13).

Establishing regulatory maps of iPSC-derived dopaminergic (DA) neurons

To identify epigenetic and regulatory features of genes in iPSC and differentiated DA neurons, we generated DNA methylation, ATAC-seq (both bulk and single-cell), HiC sequencing and small RNA-seq data across several timepoints. DNA methylation data from bulk cultures were generated at day 0 (n=97 after QC, including five technical replicates) and day 65 (n=82 after QC, including three technical replicates). These data were generated to assess changes in epigenetic patterns that potentially regulate gene transcription. The methylation data showed clear separation between both timepoints (Supplemental Figure 14). Additionally, marker genes such as *MAP2* and *TH* showed a significant reduction in methylation at their promoter regions from iPSC at day 0 to DA neurons at day 65 (Supplemental Figure 15A-E).

ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is a commonly used technique to assess genome-wide chromatin accessibility. ATAC-seq was generated from bulk cultures at day 0 (n=99), day 25 (n=97) and day 65 (n=94) with each timepoint including five technical replicates. As with the other assays, PCA across all samples shows clustering of samples by timepoint (Figure 4A). Peak sets merged from all samples at each timepoint showed an enrichment in open chromatin near promoters (0-3000 bp from the transcription start site) and a corresponding reduction in the proportion of peaks in distal intergenic regions by analysis with Cistrome [15] (Supplemental Figure 16A). Interestingly, we observed an increase in evolutionary sequence conservation at merged ATAC-seq peak sets in more differentiated cells, where the lowest PhastCons score [16] across all peak sets was at day 0 and the highest at day 65 (Supplemental Figure 16B).

To provide a cell type specific view of chromatin accessibility in our differentiated cells, we generated single cell ATAC-seq (scATAC-seq) at day 65 for a subset of the samples (n=27+2 replicates). Following quality control, 139,659 cells were retained with an average of 4,816 cells per sample (range 944 to 11,649). We identified similar broad cell types as in the scRNA-seq data (Figure 4B), however, the percentage of immature DA neurons and progenitor cell-types was significantly different between the two single- cell methods. Cell type specific chromatin accessibility was observed at particular genes of interest. For example, a distinct peak was identified at the promoter of *TH* in bulk ATAC-seq at days 25 and 65 that when examined in scATAC-seq only appears in DA neurons (Figure 4C). Overall, ATAC-seq reads are enriched at the promoter of expressed genes, but it is important to note that not all genes in this region have peaks at their promoters in either bulk or scATAC-seq reflecting cell-type specificity. ATAC-seq is also known to mark cell-type specific intergenic regulatory regions, reflecting this we see ATAC-seq peaks at putative regulatory regions upstream of *TH* that are restricted to the progenitor and DA neuron populations, suggesting that these sequences may play a role in priming *TH* expression. The peak in bulk ATAC-seq at *MAP2* also appears at day 25 and day 65, but as a broader neuronal marker is found in all cell types identified in scATAC-seq, except for the neuroepithelial-like cells, a non-neuronal cell type (Supplemental Figure 17).

HiC sequencing (HiC-seq) is a method used to identify three dimensional chromosome interactions (chromosome loops). These loops are known to be involved in regulating gene transcription by enabling physical interaction of enhancers with their cognate promoters [17];

[18]. These data can be particularly useful for linking distal risk loci/variants with regulatory regions and genes. HiC-seq data was generated for a subset of batch 1 at day 0 (n=9) and day 65 (n=8) due to the large number of cells required as input for this assay. The HiC chromosome loops show clear separation of both timepoints and marker gene *MAP2* shows distinct differences in HiC loop patterns (Supplemental Figure 18A-B).

To complement the other gene expression and regulatory data we performed small RNA-seq to investigate various classes of small RNAs including microRNAs (miRNAs) and other noncoding RNAs less than 35bp which are often involved in gene silencing and post-transcriptional regulation of gene expression. Small RNA-seq was generated at day 0 (n=99), day 25 (n=98) and day 65 (n=96) with each timepoint including five technical replicates. Separation was seen between all timepoints (Supplemental Figure 19) and enrichment in expression of miRNAs known to be expressed higher in the brain was observed across timepoints (Supplemental Figure 20).

Longitudinal imaging of iPSC-derived dopaminergic (DA) neurons

To assess the relationship between the various molecular readouts described above and neuronal phenotypes, we performed longitudinal single cell analysis. Cell-based imaging can be a valuable complementary approach to molecular analyses for characterizing phenotypes. Knowledge of all critical disease-relevant molecular changes remains incomplete, and it remains difficult to accurately predict effects on neuronal phenotype from genotypic data alone, which can involve complex non-linear interactions amongst molecular species. By studying intact cells alone or in tissues, it is possible to study disease-relevant molecular variation in context and as a system and understand better relationships to cell function. To perform longitudinal single cell analysis, 10 out of 95 iPSC lines differentiated into dopaminergic (DA) neurons (batch 1) were frozen on day 25 of differentiation. Frozen neurons were thawed, plated in 96-well dishes, matured for additional 25 days and transduced with a lentivirus for expression of GFP under control of a synapsin I (SYN1) promoter on day 50. In our experience, lentiviral methods for transduction tend to be highly efficient, even in difficult-to-transfect DA neurons, and lead to useful levels of expression of the ectopic gene. They also have fewer unwanted effects on cell morphology, especially fine neuritic processes, and viability compared with lipofectamine- or nucleofection-based methods. As shown by the scRNA-seq data, these cultures contain DA neurons, but they also contain other cell types and less mature neuron precursors. In an effort to focus our analysis on subpopulations of cells perceived to be most relevant to PD, we expressed the marker GFP from a SYN1 promoter because we have seen it is the best for restricting marker gene expression in relatively mature neurons, avoiding expression in non-neuronal cells or immature precursor cells, and generating a moderate non-toxic level of expression that easily detectable with standard imaging techniques. Fluorescence became visible within a day of transfection, and robotic microscopy [19] was used to image cells every 24 hours for approximately 10 days. Cells exhibiting GFP fluorescence in these studies had the characteristic morphological features of relatively mature DA neurons (Figure 5A). The GFP morphology signal was used to unambiguously identify individual neurons, and to track each cell from one imaging timepoint until the next. By virtue of its ability to track individual cells, robotic microscopy is able to monitor whether and how phenotypes change over time and obtain a cumulative measure of phenotypic endpoints that better controls for variability and increases the sensitivity of

comparisons of phenotypes between cohorts. Live neurons could be followed throughout all experimental time. Neuron survival over 6 days is shown in Figure 5A (green arrowhead). Cell death was detected as an abrupt loss of signal, indicative of a loss of membrane integrity (Figure 5A, red arrow). In total, 2,992 cells were analyzed across the 10 lines. The time required for complete loss of signal (time of death) from hundreds of neurons was analyzed with the Kaplan-Meier survival model [20], and cumulative risk of death curves were generated (Figure 5B).

Testing the contextual fit of iPSC-derived DA neurons for modeling PD related genetic risk

There is a wide genetic risk spectrum that is identified across the iPSC lines that we studied (Table1, Supplemental Figure 1). In addition to the contribution of genetic risk from known damaging variants in *GBA*, *LRRK2* and *SNCA,* there is a substantial common risk element that can be quantified by polygenic risk score, as previously shown using GWAS [2]. One limitation of GWAS is that it often cannot identify the causal variants, genes and relevant cell-type for each locus without additional expression or functional data. A currently commonly used method to infer cell type relevance based on GWAS statistics is MAGMA (Multi-marker Analysis of GenoMic Annotation). This powerful method relies on the convergence of unbiased genetic risk maps with unbiased single cell expression data; the enrichment of expression of genes from risk loci in individual cell types acts as a powerful indicator of cell type relevance [21]. Previous analysis using brain expression data derived from mouse and human has shown that dopaminergic neurons are a critical cellular context for PD related genetic risk [2,6]. Here, the FOUNDIN-PD scRNA-seq expression data revealed a dramatic enrichment of expression of genes within PD linked risk loci in the two identified dopaminergic cell types (immature- and DA neurons) relative to the other cell types (Figure 6A and Supplemental Table 6). In combination with the comparisons detailed above these data reveal that this model resembles human brain neurons and provides a cellular context that is appropriate for modeling complex genetic risk in PD.

In an effort to identify the causal gene and potential molecular mechanism tagged by each GWAS locus, we combined whole genome sequencing from PPMI donors with our scRNA-seq expression data in differentiated cells to identify expression quantitative trait loci (eQTL) in each broadly defined cell type. Using this approach, we replicated known eQTL in the *KANSL1* and *LRRC37A* region reflecting the H1/H2 *MAPT* haplotypes (Supplemental Figure 21A-D). When exploring the eQTL results further we specifically focused on the 90 risk variants from the most recent GWAS in PD [2]. Of these 90 risk variants, multiple showed significant eQTL associations in at least one of the defined cell-types (Supplemental Table 7). For example, the locus with rs11950533 as the lead variant harbors at least 25 genes (Figure 6B) and based on the PD GWAS browser prioritization tool [22], four (*CAMLG, JADE2, TXNDC15* and *SAR1B*) were prioritized based on their high correlation between cortical brain eQTL data [23] and PD GWAS signal (Supplemental Figure 22A-D). In the current FOUNDIN-PD scRNA-seq expression data, an eQTL for *CAMLG* was identified (Figure 6C) which shows high correlation with the PD GWAS signal (Figure 6D). However, no eQTL signals were identified for *JADE2*, *SAR1B* or *TXNDC15* (Supplemental Figure 23A-C), despite all genes being expressed in our DA neurons (Supplemental Figure 24). Inspection of the *CAMLG* bulk RNA-seq eQTL signal and the PD risk signal intersection revealed that this eQTL was not detected at the iPSC state at day 0 but became significant and detectable at day 65 (Supplemental Figure 25). This may suggest that the regulatory effect signal or trajectory of *CAMLG* expression corresponds with

differentiation to DA neurons. Therefore, based on FOUNDIN-PD data, *CAMLG* can be prioritized further and would be a good candidate for functional follow-up to confirm the association between *CAMLG* and PD risk.

The FOUNDIN-PD Data Portal

To allow rapid and easy data access to researchers, gene and region-level views of data are available through a web-based portal (https://www.foundinpd.org) integrating the multiple omic data types together (Figure 1). All users can access summary level data for a region (<5Mb) or gene by registering with a single sign-in Google account. A single-integrated view allows for visualization of genomic data by genomic coordinates with tracks available for scRNA-seq, scATAC-seq, bulk RNA-seq, bulk ATAC-seq, methylation arrays, HiC-seq, small RNA-seq, among others. The portal is fully interactive, allowing the dynamic ability to view facets/partitions of data split by *LRRK2*/*GBA*/*SNCA* status, sex, and diagnosis. The tracks are fully responsive for dynamic zooming and panning by touch or mouse, can be reordered or hidden from views. Users can download data backing the graphs via CSVs and export screen snapshots. For users that are authenticated for access to individual-level data via PPMI-INFO, the other ability is provided to visualize individual-level data. Additional phenotypic detail is available, and users can, for example, dynamically plot expression versus SNP genotype or many other variables available on subjects. The portal contains links to several additional access points, including PPMI-INFO for individual-level data and a GitHub site (https://github.com/FOUNDINPD) with analysis and standard operating procedures (SOPs). Finally, a specific single-cell view of the data is available via an embedded cellxgene instance [24], providing UMAP and PCA views. Through this interface, users can view identified clusters, genes, and gene families across profiled cells.

Data availability statement

All iPSC lines used in this study are available upon request at https://www.ppmi-info.org/access-data-specimens/request-cell-lines/. Extensive protocols and all data generated is available at https://www.ppmi-info.org/ => Access-data-specimens / Download-data / Genetic data / FOUNDIN-PD. All code used is available at https://github.com/FOUNDINPD. All data is available in the FOUNDIN-PD data browser located at https://www.foundinpd.org.

Discussion

Genetic understanding of disease is the first step on the rational path that leads through biological insight, target identification, and to development of mechanistic-based treatment. However, in order to translate genetics to biology we require an ability to model the influence of genetic risk in a contextually appropriate system and to generate replicable disease relevant readouts. The rapidly growing number of genetic risk variants and mutations associated with PD offers considerable challenges because modeling tens or hundreds of genetic factors cannot be sustainably achieved using traditional reductionist approaches. Moreover, this problem becomes more complex when considering risk variants in combination. However, this expanding risk landscape also offers opportunities. The more disease-linked genetic insight that can be modeled in a system, the more complete our understanding of disease biology will be and, as the molecular consequences of modeling risk coalesce, the more certainty we can have that these resulting events are disease-related. The application of large scale iPSC models, with robust and reproducible molecular readouts,

offers us for the first time the ability to assess the biological consequences of genetic risk factors in a disease appropriate cellular context.

Here we generated genetic, epigenetic, regulatory, proteomic, cellular imaging, and transcriptomic data for 95 iPSC lines from the PPMI study. These samples include healthy controls, PD patients with damaging mutations in *LRRK2*, *GBA* and *SNCA* as well as unaffected mutation carriers and individuals with idiopathic PD. Notably, there exists extensive biologic, clinical, and imaging data on each of the subjects from whom lines were derived. Thus, the data described in the current study can be combined and compared with data collected on these subjects including longitudinal blood RNA-seq [25], CSF markers [26], clinical data [27] and many more (https://www.ppmi-info.org/access-data-specimens/). Although we generated very large datasets totalling over ~20TB of primary data, we have sought to make these data available and accessible through the deposition of processed datasets, detailed experimental procedures, and data processing pipelines. These data have been deposited within the LONI website for PPMI (https://www.ppmi-info.org/). In addition we have created a dynamic browser (https://www.foundinpd.org) that allows users to interact with the data and to examine the features captured by FOUNDIN-PD at loci of interest and in genetic, phenotypic, and cellular subsets.

In characterizing the first data release from the FOUNDIN-PD resource, we show that the large scale differentiation process is robust and reliable. Molecular characterization of the differentiation process and of the terminally differentiated cells reveals transcriptional and epigenetic changes in line with neuronal development. Further, our data reveal that, in the context of transcriptional profiles, the DA neurons created here closely model those from the adult human brain. Our work, combining previously published unbiased GWA-derived association loci with scRNA-seq data from FOUNDIN-PD, showed that the DA neurons generated here are an appropriate cellular context to model complex genetic risk. We believe that these data will also begin to offer insights into the mechanisms of disease-related loci by providing regulatory and expression information that has not been previously available.

During the course of this resource generating study, some important lessons have been learned. Although the differentiation of multiple iPSC lines using a small-molecule approach produced a highly enriched neuronal culture (up to 93% MAP2+), a variable amount of DA neurons (5-42% TH+) mixed with a small percentage of non-neuronal cell types (2%) was detected. This variation was not related to batch, genetic sex, or the robustness of the differentiation protocol as the technical replicates showed a very high correlation in subsequent rounds of differentiation. We were able to show some early predictive markers of each line with potential to generate high levels of DA neurons and it is tempting to speculate that sorting iPSCs based on a high expression of *SRSF5* may improve differentiation efficiency. While the emergence of single cell molecular methods relieves some concerns regarding cellular heterogeneity, improving differentiation consistency line to line would be of benefit.

FOUNDIN-PD represents the largest collection of omics data generated in iPSC-derived DA neurons, and we know of no other study that has analyzed as many LRRK2+, GBA+, or idiopathic PD lines alone or in combination. However, we acknowledge that this dataset is underpowered to reveal all but the strongest of mechanisms associated with complex disease

risk loci. While the number of lines required to generate insights at the remaining loci will vary from risk allele to risk allele, we believe that the next stage of FOUNDIN-PD should include a significant increase in scale. Notably, as initiatives such as the Global Parkinson's Genetics Program (GP2) focus on diversifying the ancestral basis of our genetic understanding in PD [28], efforts such as FOUNDIN-PD should also prioritize the generation of reference data in well-powered ancestrally diverse systems. We also see the value in diversifying our terminal differentiation target to include other cell types potentially relevant for PD.

The inclusion of single cell methods, which emerged into general usage during the execution of this study, have clearly been of great benefit to FOUNDIN-PD. These data help overcome the cell type heterogeneity of differentiated cultures, provide a cellular context for genetic risk and also have the capacity to reveal transcriptomic and regulatory features specific to that disease-relevant cellular context. Based on our observations thus far, the expansion of these methods to include combined single cell transcriptomics and ATAC-seq, the emergence of single cell HiC, single cell chromatin immunoprecipitation methods to reveal transcriptional binding factor targets, and the future of single cell proteomics promise to add more resolution to the FOUNDIN-PD study and more disease-relevant insights. Inclusion of these single cell data will be a key part of the next stage of FOUNDIN-PD.

Lastly, we believe that longitudinal imaging of intact cells can complement the molecular analyses and add significantly to the characterization of patient-derived iPSC lines and to a goal of FOUNDIN-PD to conduct functional genomics for PD [29]. FOUNDIN-PD includes an extensive set of molecular analyses, but we recognize that some potentially important classes of bioactive molecules (e.g., lipids, metabolites) and functions (e.g., electrical activity) were not measured. For some assays, important subcellular spatial relationships of the macromolecules are necessarily lost during sample preparation. Imaging provides a method to study cells as intact living systems, preserving critical components, their spatial relationships *in situ*, and enabling functional measurements relevant to PD that would be difficult to infer from reductionist molecular analyses. As noted above, there are inherent challenges associated with understanding how genetic variants implicated in PD contribute to disease. The effect size of individual variants is often small, making functional effects hard to detect, and it may be the case that substantial disease risk for an individual is conferred through the combined non-linear effects of multiple variants. If so, then combining imaging with molecular analyses may be particularly helpful because it provides an approach to study the integrated effect of genetic variants on specific cell functions relevant to disease. Finally, imaging data are especially amenable to powerful machine learning types of analyses, which can be used to discover biological insights from images that elude the human eye [30] and provide a computational framework for integrating other data types, including types of multi-omic data produced by FOUNDIN-PD.

Overall, we present here the first data release of the FOUNDIN-PD project which includes multi-omics and imaging data on iPSC differentiated to DA neurons of 95 PPMI participants harboring a wide range of genetic risk. We believe the FOUNDIN-PD data will serve as a foundational resource for PD research with easily accessible data and browsers designed for basic scientists. This dataset will help the community to better understand the mechanisms of PD, identify new disease-relevant targets and potentially impact the development of novel therapeutic strategies.

Methods

Induced pluripotent stem cell lines

The induced pluripotent stem cell (iPSC) lines (n=134) used were obtained from the Parkinson's Progression Marker Initiative (PPMI; https://www.ppmi-info.org/). Each cell line vial was identified with an unique barcode and accompanied by a quality control certificate for showing normal karyotype, pluripotency and a negative test for mycoplasma. Frozen cell line stocks were quickly thawed at 37 °C, washed once with DMEM/F-12 (Gibco) to remove cryopreservation medium, resuspended in Essential 8 Flex (E8) or Essential 6 (E6) media (both Gibco) supplemented with 10 µM Y-27632 and plated on matrigel (Corning)-coated plates. E8 and E6 media were supplemented with growth factors to become equivalent in composition. Cells were kept in culture for about one month (5 passages) to allow recovery from thawing and to obtain enough cells for differentiation and assays on day 0 (iPSC state).

Sample selection and batching

Upon receiving, all cell lines were NeuroChip array genotyped [31] to confirm sample origin and to assess if large genomic events occurred during reprogramming, iPSC culture and differentiation. The data were compared to donor (blood derived) whole-genome sequencing (WGS) to identify large genomic abnormalities. Of 134 subjects, 80 are males and 54 females. The cell line collection included healthy controls, PD cases without mutations in PD related genes and affected and unaffected individuals harboring damaging point mutations including *SNCA* p.A53T, *LRRK2* p.G2019S, *LRRK2* p.R1441G, *GBA* p.E326K, *GBA* p.T369M and *GBA* p.N370S. Note that one iPSC line carries both *LRRK2* p.G2019S and *GBA* p.N370S and another iPSC line carries both *LRRK2* p.G2019S and *GBA* p.T369M. Given the much larger effect size on PD risk of *LRRK2* p.G2019S, these lines were annotated as LRRK2+, but with a comment that they also carry a *GBA* mutation. From the 134 cell lines, 95 passed QC and were selected for DA neuron differentiation and split into five batches (Table 1). One control cell line was included in each batch as a technical replicate (n=5) totaling 99 samples (Supplementary Table 1).

Table 1. Overview of PPMI iPSC lines included in FOUNDIN-PD. LRRK2+ denotes p.G2019S (n=25) or p.R1441G (n=1), GBA+ denotes: p.N370S (n=19), p.T369M (n=1) or p.E326K (n=1), SNCA+ denotes: p.A53T (n=4). Note that one individual was carrying both *LRRK2* p.G2019S and *GBA* p.N370S and one individual was carrying both *LRRK2* p.G2019S and *GBA* p.T369M. Given the much larger effect size on PD risk of *LRRK2* p.G2019S, these lines were annotated as LRRK2+, but with a comment that they also carry a *GBA* mutation. "No mutation" means that no reported pathogenic mutation in *GBA*, *LRRK2*, *SNCA* or any other known PD gene was identified.

Group	All	No mutation	LRRK2+	GBA+	SNCA+
PD	56	32	13	8	3
Unaffected carrier	26	0	13	12	1

Healthy control	**9**	**9**	**0**	**0**	**0**
Prodromal/SWEDD	**4**	**4**	**0**	**0**	**0**
Total	**95**	**45**	**26**	**20**	**4**

Differentiation of iPSC into dopaminergic (DA) neurons

The PPMI iPSC lines were thawed and grown on matrigel (Corning)-coated plates with Essential 8 Flex (E8, Batches 1, 2 and 3) or Essential 6 (E6, Batches 4 and 5) media (both Gibco) for about one month (5 passages). Essential 6 medium was supplemented with growth factors to become equivalent in composition to Essential 8. Upon reaching 70-80% confluency, iPSC lines were dissociated into a single cell suspension with Accutase (Gibco) and plated at 200,000 cells/cm^2 on matrigel-coated one-well plates (barcoded, Greiner) suitable for automated cell culture. Cells were grown until they covered the plate surface, usually 24-48 hours after single cell plating. The time required to reach confluence was variable and dependent on the growth rate of each iPSC line. The DA differentiation protocol was adapted from Kriks and collaborators [8] with minor modifications [9]. Differentiations were carried out in an automated cell culture system [10] with manual replatings on day 25 and 32 for final differentiation and immunocytochemistry (ICC), respectively [10]. Samples for assays were collected on days 0 (iPSC), 25 (mid-point) and 65 (DA neurons). For DNA assays, cells were dissociated with Accutase, washed once with PBS and spun down at 200x g. The cell pellet was snap frozen or processed according to assay protocols. Most of the RNA assays required snap frozen cells collected by scraping the plate surface with PBS or lysis buffer. Single-cell (sc) RNA-seq and scATAC-seq assays required a single cell suspension prepared in 0.04% human serum albumin (HSA)/PBS. All samples were stored at -80°C until further processing. For cryopreservation, day 25 DA neuron precursors were dissociated with Accutase, washed once with neurobasal medium (Gibco), resuspended in cold Synth-a-Freeze cryopreservation medium (Gibco) supplemented with 10µM Y-27632 and aliquoted into barcoded cryovials (NovaStora) at 10x10^6 cells/ml/vial (on ice). The cryovials were placed in CoolCell cell freezing containers (Biocision), kept overnight at -80°C and transferred to liquid nitrogen for long term storage.

Immunocytochemistry (ICC) and image analysis

Cells were differentiated until day 65, fixed in 4% PFA, washed 3x5 min in PBS and blocked in 5% goat serum/1% bovine serum albumin (BSA)/0.1% Triton-X100/PBS for 1 h at room temperature (RT). Primary antibodies were applied overnight at 4°C and included TH (Pel-Freez Biologicals #P40101 and Merck Millipore #AB9702, both at 1:750 dilution) and MAP2 (Santa Cruz #sc-74421, 1:100). After incubation with primary antibodies, cells were washed 3x5 min in PBS. Cells were incubated with second antibodies (AF488 and AF594, Invitrogen, 1:1000) for 2 h at RT followed by nuclear counterstaining with Hoechst 33342 (Invitrogen, 1:8000) for 30 min at RT. Finally, cells were washed 3x5 min in PBS and imaged with a CellVoyager 7000 (Yokogawa) confocal microscope and 20x objective. Images were analysed on Columbus (PerkinElmer) as described previously [10]. The total number of TH (DA neuron) and MAP2 (neuron) positive cells was estimated and normalized to the total number of nuclei. Data is represented as the percentage of positive cells per 30 fields.

Longitudinal image analysis of iPSC DA neurons

Frozen day 25 DA neuron precursors were thawed and replated at a density of approximately 450,000 cells/cm^2 onto dishes coated with 0.1mg/ml poly-l-ornithine (PLO), 5µg/ml laminin, and 5µg/ml fibronectin in NB/B27 medium prepared as described [9] with the addition of 10µM ROCKi and 100µg/ml matrigel (Corning). The media was changed 4 h later to remove ROCKi. DA neurons were matured in NB/B27 medium, then replated into 96-well plates on day 49. On day 50, cells were transduced with synapsin-driven GFP via lentivirus (SignaGen), followed by a media change the next day. Cells were imaged daily from approximately day 54 through 66 using robotic microscopy, a previously described automated imaging platform [19,20]. Images obtained from 8 consecutive days were processed using custom programs in Galaxy [32; 33] to assemble arrays of images into montages representing each well, and to stack montages across timepoints. Neuron survival was analyzed using a custom program written in MATLAB. Live neurons expressing GFP were selected for analysis only if they had extended processes at the first timepoint. Neurons were tracked longitudinally across timepoints until death, and survival time was defined as the last timepoint a neuron was seen alive. The survival package for R statistical software was used to construct Kaplan-Meier curves from the survival data[20]. Survival functions were fitted to these curves to derive cumulative risk-of-death curves.

Methylation library preparation and data-processing

DNA was extracted from each timepoint using standard phenol:chloroform extraction. DNA from day 0 and day 65 underwent Bi-sulfite conversion using the EZ-96 DNA Methylation Kit (Zymo Research). Bisulfite converted DNA was then put through the standard Infinium HD array based methylation assay (Illumina) with Illumina Infinium HumanMethylation EPIC BeadChips. Raw signal intensity data were processed from raw idat files through a standard pipeline using Meffil [34]. A number of standard quality control steps were performed to these data prior to normalization including: sample origin confirmation based on SNP presence on array, sex concordance check, methylated versus unmethylated ratio, low bead numbers, control probes quality and, finally, general outlier samples were identified using principal component analysis and excluded. Subsequently, the quality controlled data was normalized using quantile normalization. The analysis pipeline can be found here: https://github.com/FOUNDINPD/METH.

Quantitative tandem mass tag proteomics preparation, mass spectrometry and data-processing

Proteomics data were generated only for batch 1 (n=10) lines. Initially, only proteins from day 65 were collected but iPSCs were rethawed at a later date to profile proteins at day 0. Cell pellets were lysed in buffer containing 0.3M HEPES (Sigma), 2% CHAPS (Sigma), 10mM EDTA, 1x HALT phosphatase inhibitor cocktail and 1x HALT protease inhibitor cocktail (both Thermo Scientific). Lysates were transferred to Protein LoBind tubes (Eppendorf) and briefly sonicated (2x 15 seconds, Amp=50) in a cup-water sonicator with ice. Complete lysis was achieved by incubating samples for an additional 20 min on ice and insoluble membranes were removed by centrifugation. Sample concentrations were determined using manufacturer instructions with the Pierce 660nm protein assay kit (Thermo Scientific) and samples were normalized to the lowest concentration using an additional lysis buffer. Lysates were separated into three technical replicates and pooled samples were made for each replicate. Proteins were digested with trypsin and labeled as directed by the TMT10plex Isobaric label reagent set plus TMT11-

131C label reagent user manual (Thermo Fisher). Peptides were desalted on a Waters Oasis HLB 200mg cartridge (Waters) using standard vendor procedures. Samples were then dried in a speedvac and resuspended in Hydrophilic Interaction Liquid Chromatography loading buffer (HILIC, 98% Acetonitrile (ACN), 0.1% Trifluoroacetic acid (TFA)) and loaded into and Agilent 1290 Infinity High Performance Liquid Chromatography (HPLC) instrument with fraction collector. Peptides were separated into 14 HILIC fractions with a 32 min gradient increasing Mobile Phase B (MPB; 2% to 60%; MPA 98% ACN, 0.1% TFA; MPB: 98% H2O, 0.1% TFA) on a TSK gel amide-80 HR 5µm column (Sigma). The 14 fractions were then merged into 10 based on the predicted complexity of the peptides from the original fractions. A Fusion Lumos Mass Spectrometer coupled to a 3000 Ultimate HPLC (Thermo Fisher) was then used to run one 75 minute LC-MS/MS for each LC-MS fraction. Proteome Discoverer version 2.2 was used to calculate relative ratios of each sample to the TMT-126 pool sample for each replicate. Proteins were assigned using the NCBI protein database. A csv file with annotations and relative ratios was created for each replicate run to be analyzed in Rstudio. The csv was loaded into RStudio as a matrix and proteins with more than 50% missingness across all samples were removed. Extreme values were filtered out and the remaining protein means were scaled across samples. Finally, regression was used to remove batch effects. The analysis pipeline can be found here: https://github.com/FOUNDINPD/Proteomics/.

Bulk ATAC sequencing library preparation, sequencing and data-processing

Bulk ATAC-seq data was generated from all batches at all timepoints. Cells at each timepoint were collected using Accutase (Gibco) to make a single cell suspension and 75,000 cells per sample were aliquoted for bulk ATAC-seq. Standard procedures with slight modifications were used [35]. In brief, cells were lysed (10mM Tris-HCl, 10mM NaCl, 3mM $MgCl_2$, 0.1% (v/v) NP-40), nuclei were then spun down, resuspended in transposition buffer (TD buffer, Tn5 Transposase from the Illumina Tagment DNA Enzyme and Buffer kit) and incubated for 30 min at 37°C. After incubation, DNA was isolated using Qiagen MinElute Reaction Cleanup Kit (Qiagen) according to manufacturer's recommendations. DNA was eluted in 10µl of EB buffer (10mM Tris-Cl, pH 8.5) and then frozen at -80°C.

Libraries were prepared by combining transposed DNA with NEBNext High-Fidelity 2X PCR Master Mix (New England Biolabs) and 1.25µM indexing primers (Ad1_noMX primer and Ad2.x indexing primer [35] or IDT for Illumina dual index primers (Illumina, Nextera DNA UD Indexes Set A Ref 20025019). Standard PCR conditions for NEBNext High-Fidelity 2X PCR Master Mix were used with 10-12 cycles completed on each library. Libraries were purified using AMPure XP (Beckman Coulter) beads with the manufacturer's protocol for double-sided purification. Quality assessment of libraries was measured on an Agilent High Sensitivity DNA analysis chip (Agilent) to determine average library size and concentration. The concentration of each library was verified by Qubit Fluorometric Quantification (Thermo Scientific) before sequencing. Batch 1 libraries were sequenced at the NIH Intramural Sequencing Center (NISC) on an Illumina NovaSeq, with 50bp paired-end (PE) reads. Batches 2-5 were sequenced at The American Genome Center (TAGC) at the Uniformed Services University on an Illumina NovaSeq with 100bp PE reads. Fastq files for each sample were assessed using FastQC (v0.11.9) and reads were aligned to GRCh38 using Bowtie2 (v2.4.1; Langmead and Salzberg, 2012) in local mode. Reads mapping to ChrM and ChrUn were filtered out and samples with less than 20 million PE reads remaining were removed from analysis. MACS2 was used to call peaks [36]. The full analysis pipeline can be found here: https://github.com/FOUNDINPD/ATACseq_bulk.

HiC sequencing library preparation, sequencing and data-processing

HiC sequencing data were generated from batch 1 day 0 and day 65 samples. Library preparation was performed by Phase Genomics (https://phasegenomics.com/) using their standard protocol. Fastq files from each lane were merged to give each sample two read fastqs. Fastqc was run on all sample fastq files before further analysis. The Juicer pipeline was used to obtain high resolution contact maps and loop regions for each sample [37]. Preliminary testing indicated excessive mitochondrial data in samples, so the pipeline was altered to remove mitochondrial reads after mapping. The Juicer pipeline incorporates the Burrows-Wheeler Aligner (BWA) to map fastqs to a reference genome [38]. Loop regions in samples were detected using the HiCCUPs algorithm included in the Juicer pipeline. These regions were saved in .bedpe files and used for further analysis. Loop region overlap was calculated between samples and with public PsychENCODE data [39]. The HiCCUPSDiff tool was used to detect differential loops between day 0 and day 65. Heatmaps were generated for each sample and each chromosome to visualize chromatin interactions using the HiCExplorer tool [40]. The analysis pipeline can be found here: https://github.com/FOUNDINPD/HiC_Pipelines.

Bulk RNA sequencing library preparation, sequencing and data-processing

Bulk RNA sequencing data was generated from all batches and all timepoints. RNA was isolated using Qiagen's "Purification of miRNA from animal cells using the RNeasy Plus Mini Kit and RNeasy MinElute Cleanup Kit" using protocol 1 to purify total RNA containing miRNA. Briefly, cells were lysed with Guanidine-isothiocyanate and homogenized with QIAshredder, then passed through a gDNA Eliminator spin column. The lysate was combined with ethanol to bind RNA to the spin column while contaminants are washed away. Samples were separated into 4 different RNA isolation protocols dependent on the sample's cell counts (target of 1.3-4 million cells per column). Samples with 1.33-4 million cells/vial were isolated using 1 column. Samples with 4.61-7.86 million cells/vial were isolated on 2 columns with 2.3-3.93 million cells/column. Samples with 8.17-12 million cells/vial were isolated on 3 columns with 2.72-4.0 million cells/column. Samples with 12.75-52 million cells/vial were isolated on 3 columns with 4 million cells/column and the leftover lysate was stored. High quality total RNA (containing miRNA) was then eluted and used for either bulk RNA sequencing or small RNA sequencing library preparation. Libraries were prepared using the SMARTer Stranded Total RNA Sample Prep Kit - HI Mammalian (Takara Bio USA, Inc.), which incorporates both RiboGone and SMART (Switching Mechanism At 5' end of RNA Template) technologies to deplete nuclear rRNA and synthesize first-strand cDNA. This along with PCR amplification and AMPure Bead Purification generates Illumina-compatible libraries. Using the total RNA stock concentration, we determined the volume needed for 1ug RNA input. Samples were concentrated by SpeedVac or diluted with nuclease-free water to obtain a volume of 9µl per sample. Addition of buffers and enzymes including RNase H, DNase I, and 10X RNase H Buffer along with three PCR reactions and a 1.8X bead purification removed specific rRNA sequences (5S, 5.8S, 12S, 18S, and 28S). rRNA depleted RNA was fragmented at 94°C for 3 min and immediately placed on ice. Master mix containing reverse transcriptase and an oligonucleotide added to samples and incubated in a preheated thermal cycler to convert RNA to single-stranded cDNA. cDNA was purified at 1X ratio with AMPure beads. Unique dual-indexed PCR primers (allowing for multiplexing) combined with SeqAmp DNA Polymerase were added to each first-strand cDNA. Using 12 cycles on a preheated thermal cycler, cDNA was amplified into RNAseq libraries. AMPure Bead purification (1X), 80% ethanol wash, and

elution of 34µl with nuclease-free water generated final libraries ready for Illumina sequencing. 2µl of cDNA library placed on a well plate with 2µl Sample Buffer and analyzed on 4200 TapeStation to determine peak range (bp). Concentration of libraries determined by 40K, 80K dilutions on Kapa SYBR Fast qPCR (Roche). Libraries were pooled together into 2 pools with a concentration of 60pM and volume of 100µl and sequenced on an iSeq 100 300-cycle flow cell. Libraries were normalized based on these results. Libraries were re-pooled together with a final concentration of 5000pM and final volume of 220µl, concentration obtained by QuantStudio. Pool was run on a NovaSeq 6000 S1 200-cycle flow cell with a loading concentration of 1,500pM and volume of 100µl with a 20% PHiX spike-in, with the following parameters: 100 x 9 x 9 (+7 dark cycles) x 100. The sequencing depth was a minimum of 30M read pairs per sample. The bcl files were demultiplexed using bcl2fastq v2.19.1.403 (Illumina) using default parameters. Reads were trimmed with cutadapt v2.7 [41] to remove the first three nucleotides of the first sequencing read (Read 1), which are derived from the template-switching oligo. Trimmed reads were aligned to the GRCh38 genome using STAR v2.6.1d [42]. Following genome alignment, reads were counted with featureCounts v1.6.4 [43], (part of the subread package) using a non-redundant genome annotation combined from GENCODE 29 and LNCipedia5.2 [44] (https://github.com/FOUNDINPD/annotation-RNA). Count data was loaded into R v3.6.3 for analysis. Normalized counts, variance stabilizing transformation, and differential expression analysis were performed using DESeq2 v1.26.0 [45], and CPM values were generated using edgeR v3.28.1 [46]. Heatmaps were created using the pheatmap v1.0.12 package in R. Trimmed fastq files were also quasi-mapped to the same annotation using salmon quant v1.2.2 [47]. In order to identify upregulated and downregulated genes from day 0 to day 65, differentially expressed genes (defined as baseMean > 100, adjusted $p < 0.01$, and absolute value of the log2 fold change > 1) were further filtered using a general linearized model, retaining genes that have a slope > 0.05 for upregulated genes and a slope < -0.05 for downregulated genes. Gene ontology analysis was performed on these upregulated and downregulated genes with GOfuncR 1.6.1, using the refine function with a FWER = 0.1, and GOxploreR 1.1.0 [48]was used to remove redundant GO terms Parameters used for genome alignment, annotation, and quasi-mapping are described on GitHub. The analysis pipeline can be found here: https://github.com/FOUNDINPD/bulk_RNASeq.

Small RNA sequencing library preparation, sequencing and data-processing

Small RNA sequencing data were generated from all batches and all timepoints. RNA was isolated in the same manner as for bulk RNA sequencing, using Qiagen's "Purification of miRNA from animal cells using the RNeasy Plus Mini Kit and RNeasy MinElute Cleanup Kit" using protocol 1. Small RNA libraries were made using the NEXTFLEX Small RNA v3 kit (PerkinElmer), followed by 3' adapter ligation and excess 3' adapter removal according to manufacturer's protocol. Excess adapter inactivation was not performed. 2µl of the inactivation ligation buffer were used without enzyme in lieu of the inactivation step. 5' adapter ligation and reverse transcription was performed per manufacturer's protocol. 62.5µl of cDNA, beads, and isopropanol solution was transferred instead of 70µl to help reduce adapter dimer moving forward to PCR. Libraries of appropriate size were collected using gel purification. Purified libraries were quantified using the high sensitivity DNA kit on the Bioanalyzer (Agilent). Equimolar pools were made and sequenced on a Hiseq 2500 at 8pM. The bcl files were demultiplexing using bcl2fastq. Small RNA sequencing reads (fastq files) were processed using the exceRpt pipeline. The pipeline was run using the RANDOM_BARCODE_LENGTH=4 parameter to trim off the random 4-bp ends in NEXTFLEX

sequencing data along with the Illumina (TruSeq) smallRNA adapters. All other parameters were set to defaults. Pipeline was run using a custom transcriptome database composed of human sequences from mirBase 22, gencode 28, piRBase and tRNAscan-SE. Following the pipeline run on each sample an R summary script (mergePipelineRuns.R) was run which generates raw read alignment counts, RPMs and QC metrics for all small RNA species across all samples. Expression of small RNAs that were consistently increasing over time points were investigated for their expression patterns using data from a small RNA tissue atlas (Aslop et al, in preparation). The analysis pipeline can be found here: https://github.com/FOUNDINPD/exceRpt_smallRNAseq.

Single-cell (ATAC and RNA) library preparation, sequencing and data-processing

Cells harvested on day 65 of differentiation were processed following the 10x Genomics single-cell (sc) RNA and ATAC sequencing protocols to generate DNA libraries. To note, batch 1 cells processed for scRNA-seq were generated from a second run of differentiation, since this assay was included later in the study. Additionally, scATAC-seq was performed only for cells from batches 4 and 5. For scRNA-seq, the libraries comprised standard Illumina paired-end constructs which begin with P5 and end with P7. The 16bp 10X barcodes are encoded at the start of TruSeq Read 1, while 8bp sample index sequences are incorporated as the i7 index read. TruSeq Read 1 and Read 2 are standard Illumina sequencing primer sites used in paired-end sequencing. TruSeq Read 1 is used to sequence 16bp 10x barcodes (cell identifier) and 12bp UMI (transcript identifier). scATAC-seq libraries compatible with Illumina sequencing were generated by adding a P7 and a sample index via PCR. Sequencing was performed on Illumina NovaSeq. Libraries were sequenced at a minimum depth of 20,000 read pairs per cell for scRNA-seq and 25,000 read pairs per nucleus for scATAC-seq.

scRNA-seq: The BCL files obtained after sequencing were demultiplexed into FASTQ files using the cellranger "*mkfastq*" software and unique molecular identifier (UMI) gene counts were calculated by cellranger "*count*" software (v3.1.0) [49]. UMI gene counts for each sample were merged into a table and imported into R (v3.6.0). We used Seurat (v3.1.1) [50]within the R environment for filtering, normalization, integration of multiple single-cell samples, unsupervised clustering, visualization and differential expression analyses. The following data processing was done: (1) Filtering. Cells with less than 1,000 and more than 9,000 genes expressed (≥1 count) were filtered out, and only genes that were expressed in at least 100 cells were kept. Moreover, cells with more than 20% of counts in mitochondrial genes were filtered out. After filtering, there were 34,960 genes in 416,216 cells; (2) Data normalization and integration. Gene UMI counts for each cell were normalized using the "*SCTransform*" function in Seurat. Integration of scRNA-seq data from multiple samples was performed using top 3000 variable features and top 3 samples as reference with the highest number of cells; (3) Clustering and visualization. Clustering was performed using "*FindClusters*" function with default parameters except resolution was set to 0.5 and first 30 PCA dimensions were used in the construction of the shared-nearest neighbor (SNN) graph and to generate 2-dimensional embeddings for data visualization using UMAP; (4) Differential expression analyses: We used "*FindAllMarkers*" function with default parameters and only tested genes that are detected in a minimum of 40% of cells in either of the two clusters. Genes with an adjusted p value <0.05 were considered to be differentially expressed. The pipelines used in this study are available at https://github.com/FOUNDINPD/FOUNDIN_scRNA.

scATAC-seq: The BCL files obtained after sequencing were demultiplexed into FASTQ files using the cellranger-atac "mkfastq" software and unique molecular identifier (UMI) counts

were calculated by cellranger-atac "count" software (v1.2.0) [51]. Peaks for each sample were merged into a table and imported into R (v3.6.0). We used Seurat (v3.2.0), Signac (v1.1.0) [52] and Harmony (v1.0) [53] within the R environment for filtering, normalization, integration of multiple single-cell samples, unsupervised clustering, visualization and predicting the cell-types. The following data processing was done: (1) Filtering. We kept the cells with minimum 1,000 peaks (≥ 1 count), respectively and the peaks that were called in at least 100 cells. Moreover, cells with more than 20% of counts in mitochondrial DNA were filtered out. After filtering, there were 459,495 peaks in 139,659 cells; (2) Data normalization and integration. Peak counts for each cell were normalized using the "*RUNTFIDF*" function in Signac that performs term frequency-inverse document frequency normalization followed by SVD decomposition to generate latent semantic indexing (LSI). Integration of scATAC-seq data from multiple samples was performed using the "*RUNHarmony*" function with LSI reduction; (3) Clustering and visualization. Clustering was performed using the "*FindClusters*" function with default parameters except resolution was set to 0.1 or 0.2. First 30 harmony dimensions were used to generate 2-dimensional embeddings for data visualization using UMAP; (4) Predicting cell-types: Fragments in the genes (extended 2kb upstream) were calculated for each cell to generate a gene activity matrix and normalized the data using the "*LogNormalize*" method. Cell-types were predicted using scRNA-seq data as a reference and scATAC-seq data as a query for "*FindTransferAnchors*" and "*TransferData*" functions. Prediction often results in heterogeneous cell-type annotation with-in the same cluster. We assigned the cell-type to a cluster with the maximum occurrence. The neuroepithelial-like cluster was separated using 0.2 resolution. The pipelines used in this study are available at https://github.com/FOUNDINPD/FOUNDIN_scATAC.

Prediction of neuronal differentiation efficiency using bulk RNA-seq data at day 0

To test the predictive value of the genic expression profile in iPSC for neuronal differentiation efficiency, we performed supervised machine learning (logistic regression) on the DESeq2 v1.26.0 [45] normalized count expression values for genes at day 0 and estimated DA neurons fractions from the differentiated cell lines at day 65. DA neuron fractions were calculated from scRNA-seq data, based on the total number of cells and the number of cells in the 'Dopaminergic Neurons' cluster (see the Method section for scRNA-seq). Cell lines were classified into high (n=62) and low (n=21) differentiation efficiency classes based on the relative abundance of the DA neurons at day 65; as a threshold for classification, we used first quartile value of cell percentages (Q25=15.7%), as it was best separating the two observed distribution peaks of DA neuron counts across the cell lines.

To reduce possible bias in the predictive model, we used a full set of reliably expressed genes (threshold of inclusion mean normalized count >=50). As we did expect a significant number of genes to be highly correlated to one another in their expression, with multitude of them being possibly relevant for prediction, and the total number of relevant features for our model is unknown, we resolved to using elastic net regularization approach, which combines both lasso regression (shrinking less important features and pruning some) and ridge regression (assigns proportional coefficients to highly correlated possibly relevant features and prevents model overfitting) to equal degree (alpha=0.5) with a penalty lambda equal to 0.22. To further control for possible overfitting, repeated (100 times) 5-fold cross-validation was performed using the "*cv.glmnet*" function. Data preprocessing and logistic regression was executed in R (v3.6.3) [54], using packages caret (v6.0-86) [55] for model training and glmnet (v4.0) [56] for elastic net regularization of the model and repeated cross-validation. As the sample size is small and

imbalanced, we directly tested the relation of the resulting predictive candidate genes' expression to the percentage of DA neurons in each cell line. We performed Spearman's rank correlation test, using R package stats (v3.6.3) [54]. Benjamini & Hochberg procedure was used for multiple testing corrections of p-value.

MAGMA to identify causative cell types

Expression gene profiles obtained from the scRNA-seq dataset were used to test for a cell-type association with PD. We used the R package MAGMA_Celltyping (v1.0.0, https://github.com/NathanSkene/MAGMA_Celltyping), which utilizes MAGMA [21] software package as a backend, to identify cell types positively associated with the common-variant genetic hits from the most recent PD GWAS [2]. LD regions were calculated using the European panel of 1000 Genomes Project Phase 3 [57]. The cell-type enrichment analysis was performed on 5000 subsampled cells from each scRNA-seq cluster.

Single cell expression quantitative trait loci analysis

Variants from the whole genome sequencing data were correlated with normalized average gene expression levels per cell cluster by performing single cell expression quantitative trait loci analysis. After quality control, 77 samples were included for analysis and expression data were filtered for 0.025 average expression in all samples. Then genes were removed with zero expression in 15 or more samples resulting in expression of 1256 genes across 90 risk variants. eQTL analysis was performed using MatrixEQTL [58] including variants with minor allele frequency >5% and using the following covariates: batch, sex, age of donor, *GBA*, *SNCA*, *LRRK2*, phenotype, TH+ levels, MAP2+ levels, number of cells, reads per cell, total genes detected, and median UMI counts per cell. Overlap between eQTL variants and GWAS was determined using the most recent PD GWAS [2]. For GWAS loci of interest, violin plots were generated to visualize the correlation between genotype and gene expression. Additionally, LocusZoom [59] and LocusCompare [60] plots were generated to visualize correlations between GWAS signal and eQTL signal and the PD GWAS locus browser was used for loci numbering and prioritization [22]. The analysis pipeline can be found here: https://github.com/FOUNDINPD/SCRN_EQTL_v2. Bulk eQTL analysis was performed separately on day 0, 25, and 65 data using tensorQTL [61] and included estimated cell fractions as covariates. The estimated cell fractions were generated using the Scaden [62] deconvolution tool trained on the day 65 single-cell data.

FOUNDIN-PD browser

Architecturally, the FOUNDIN-PD portal is a single-page application (SPA) framework where a public javascript application interacts with a secured JSON API to build the user DOM within the user browser. The client-side nature of the application allows for dynamic interactions with the user with low latency and high scalability, leveraging the fact that many users will leverage modern computing and browsing capabilities. At a granular level, the FOUNDIN-PD application is based on JavaScript ECMAScript 2016 and builds upon Vega.js (vega.github.io; version 5.22) visualization grammar and D3.js (https://d3js.org; version 6) for dynamic responsive graphing. The API is within a sharded MongoDB 4.2 (https://mongodb.com) framework on a CentOS8 cloud server using an NGINX (https://nginx.org; 1.18) proxy, NodeJS 12 middleware (https://nodejs.org), to provide a protected JSON API. API data is secured using JSON/JWT authentication via Auth0 (https://auth0.com) and Google OAUTH 2.0 (https://oauth.net/2/) for the identification of users.

Acknowledgements
We would like to thank all of the subjects who donated their time and biological samples to PPMI, without whom we could not have done this study. This work is supported by the Michael J. Fox Foundation for Parkinson's Disease Research and is part of the PD Pathogenesis consortium. Cell lines used in the analyses presented in this article were obtained from the Golub Capital iPSC Parkinson's Progression Markers Initiative (PPMI) Sub-study (www.ppmi-info.org/cell-lines). Data used in the preparation of this article were obtained from the PPMI database (www.ppmi-info.org/data). As such, the investigators within PPMI contributed to the design and implementation of PPMI and/or provided data and collected samples but did not participate in the analysis or writing of this report. For up-to-date information on the study, visit www.ppmi-info.org. PPMI – a public-private partnership – is funded by The Michael J. Fox Foundation for Parkinson's Research and corporate sponsors, including: Abbvie, AcureX Therapeutics, Allergan, Amathus therapeutics, Avid Radiopharmaceuticals, BIAL Biotech, Biogen, Biolegend, Briston-Myers Squibb, Calico, Celgene, Denali, 4D Pharma, GE Healthcare, Genentech, GlaxoSmithKline, Golub Capital, Handl Therapeutics, Insitro, Janssen neuroscience, Lilly, Lundbeck, Merck, Meso Scale Discovery, Neurocrine Biosciences, Pfizer, Piramal, Prevail Therapeutics, Roche, Sanofi Genzyme, Servier, Takeda, Teva, UCB, Verily and Voyager Therapeutics. An up to date list of all PPMI Industry Partners can be found at http://www.ppmi-info.org/about-ppmi/who-we-are/study-sponsors/. This work is supported in part by the Intramural Research Program of the National Institute on Aging, National Institutes of Health, part of the Department of Health and Human Services; project ZO1 AG000949. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD, USA. (http://biowulf.nih.gov). Additional support include ERACoSysMed2; PD-Strat. Multi-dimensional stratification of Parkinson's disease patients for personal interventions (FKZ 031L0137A). VB is supported by a Career Development Fellowship at DZNE Tuebingen. The authors would like to thank the NIH Intramural Sequencing Center (NISC) and the American Genome Center for sequencing services and Phase Genomics for HiC library construction. Additional support for this work came from RF1 AG1058476, U54 NS191046 and R37 NS101966 to SF.

Author contributions
Study design: All
Funding acquisition: ABS, PH, MRC, CB, JRB, SF, KVKJ, DWC
Data analysis: VB, IV, DWC, CB, AI, NS, EB, JRG, MC, SF, XR, MRC, FPG, EH, EA
Statistical analysis: VB, IV, DWC, CB, AI, NS, EB, JRG, FPG
Manuscript drafting: CB, EB, XR, VB, DWC, PH, MMC, SF, ABS, MRC
Manuscript revision: All
DA neuron culture: EB, SB, MMC
Assay preparation and processing: EB (ICC, scRNA-seq, scATAC-seq, HiC-seq), SB (ATAC-seq), NF & PR (scRNA-seq), XR (Genotyping, Methylation, ATAC-seq, Proteomics), CB (Genotyping, Methylation, Proteomics, HiC-seq), MMC (Imaging), CD (ATAC-seq, HiC-seq), JB (Genotyping, Methylation), YL (Proteomics), FPG (Methylation, HiC-seq), BM, RR, AL, JA, AC (RNA isolation for small and long RNA library preparation; library preparation and sequencing)
SOPs: CB, IV, EB, PR, SB, XR, MMC, VB
Browser: DWC, MW, RS, IV

Main Figure Legends

Figure 1. Graphical overview of the FOUNdational Data INitiative for Parkinson's Disease (FOUNDIN-PD). Classes of assays, timepoints and number (n) of samples included in each assay are shown. Blue icons represent assays that are included in the initial data release and light blue icons represent assays that are ongoing and will be released at a later stage.

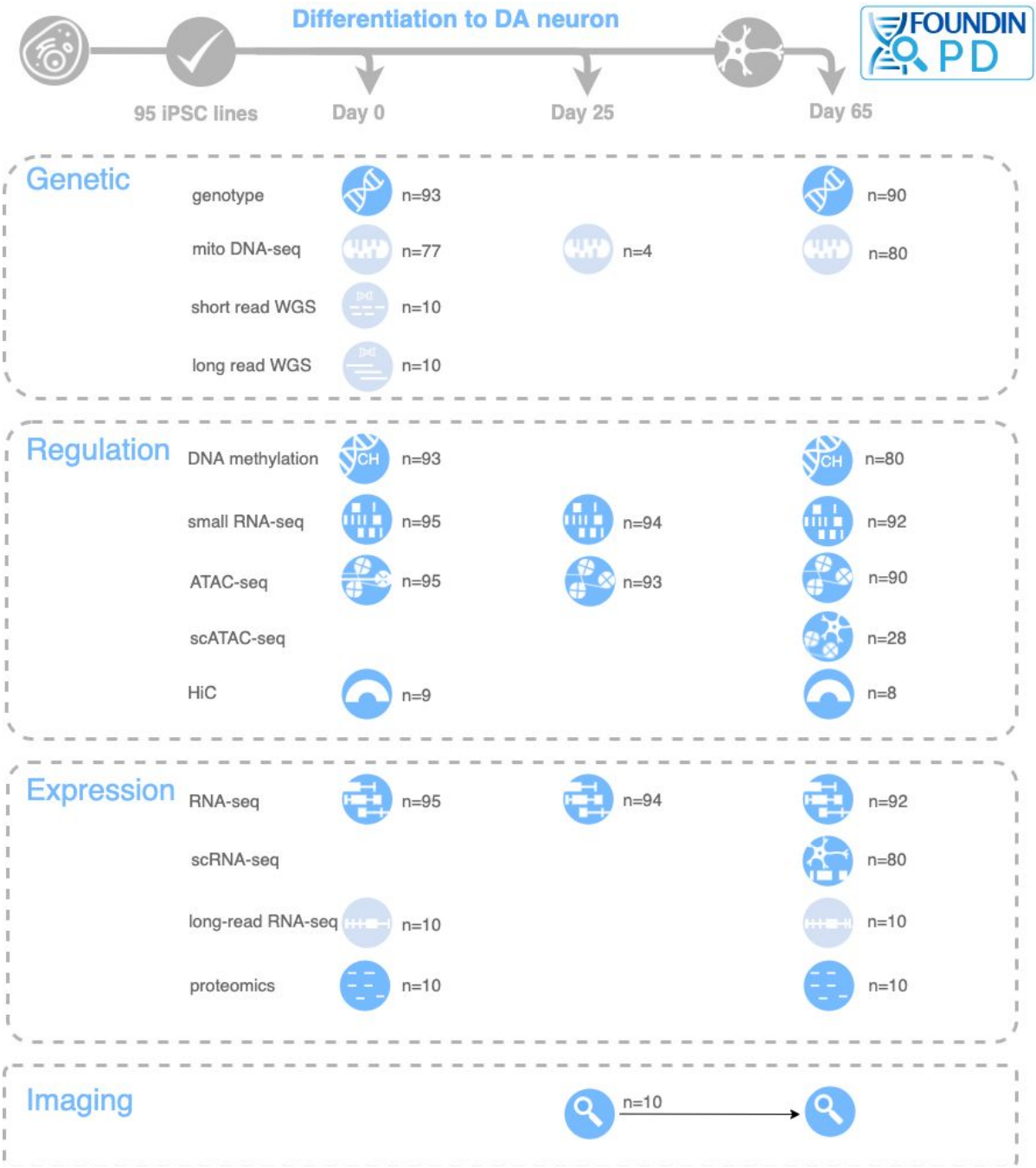


Figure 2. Quality control and scRNA-seq on day 65. A, Schematic overview of the differentiation protocol. **B**, Left: Representative ICC image showing TH (DA neurons) and MAP2 (neuron) positive cells co-stained with DAPI (nuclei). Scale bar: 50 µm; Right: Quantification of the percent TH (Pel-Freez) and MAP2 positive cells detected by ICC. Each dot represents one cell line (n=95). **C**, Uniform Manifold Approximation and Projection (UMAP) illustrates cell clusters identified at day 65 (n=416,216 single cells, n=79+4 control replicate cell lines). Cell types with their respective percentages are indicated. **D**, Percentage of cells and average expression level of *TH*, *MAP2* and *SNCA* for each cell type. The dot color scale from blue to red, corresponds to lower and higher expression, respectively. The size of the dot is directly proportional to the percentage of cells expressing the gene in a given cell type. **E**, Spearman's correlation test showing high correlation of gene expression across FOUNDIN-PD DA neuronal types and post mortem substantia nigra human brain. ODCs, oligodendrocytes; OPCs, oligodendrocyte precursor cells. See Supplemental Figure 6A for UMAP of cell types identified by using Agarwal and collaborators data. **F-G**, Correlation between percentages of TH (Pel-Freez) and MAP2 positive cells in ICC and scRNA-seq (R, Pearson correlation coefficient; $p<0.0001$). Each dot represents one cell line (n=83). **H**, Variability in differentiation efficiency across the iPSC lines. Each color represents the cell types annotated in scRNA-seq UMAP. HC, Healthy Control; iPD, idiopathic PD.

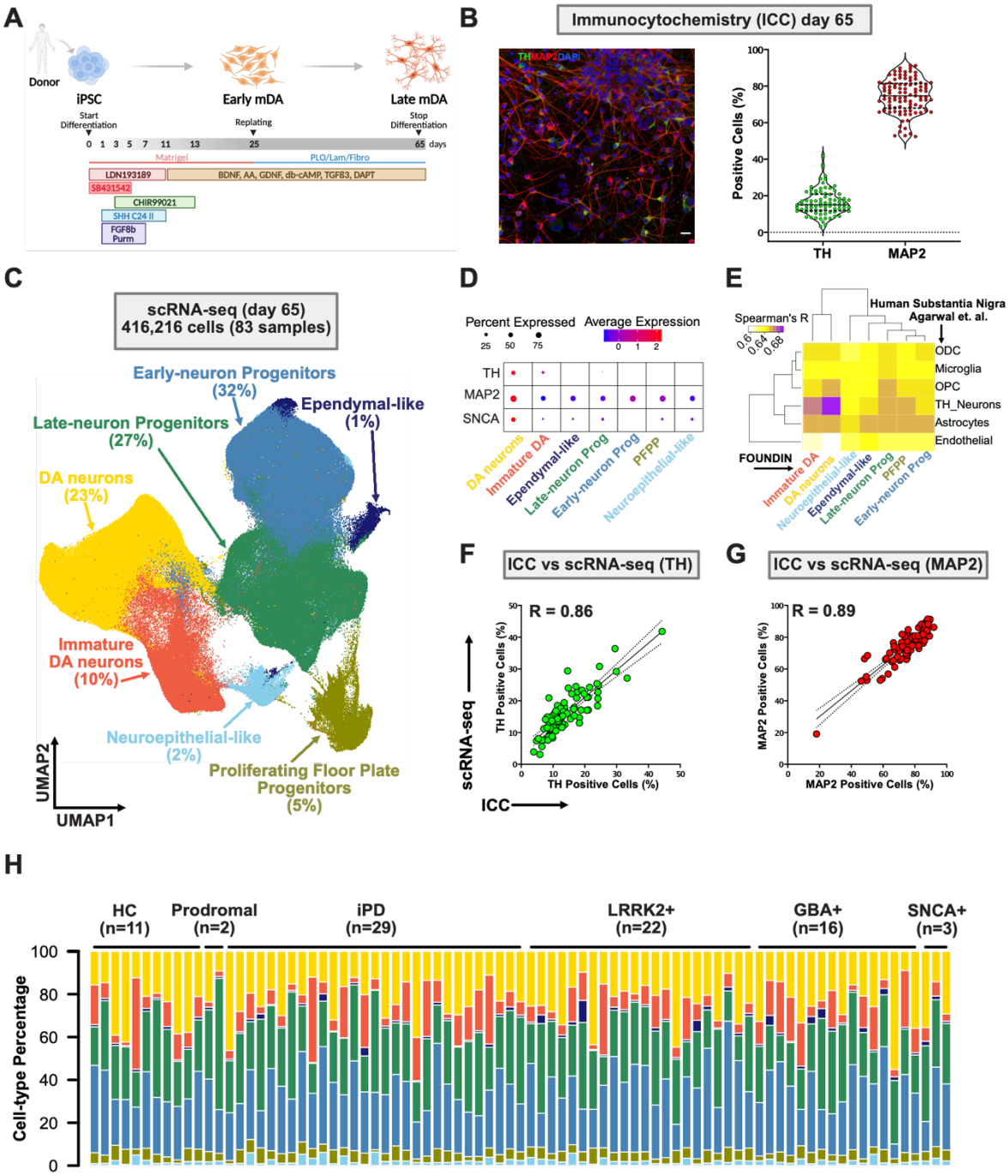

A
Donor
iPSC
Early mDA
Late mDA
Start Differentiation
Replating
Stop Differentiation
0 1 3 5 7 11 13 25 65 days
Matrigel
PLO/Lam/Fibro
LDN193189
BDNF, AA, GDNF, db-cAMP, TGFB3, DAPT
SB431542
CHIR99021
SHH C24 II
FGF8b
Purm
B
Immunocytochemistry (ICC) day 65
TH MAP2 DAPI
Positive Cells (%)
TH
MAP2
C
scRNA-seq (day 65)
416,216 cells (83 samples)
Early-neuron Progenitors (32%)
Ependymal-like (1%)
Late-neuron Progenitors (27%)
DA neurons (23%)
Immature DA neurons (10%)
Neuroepithelial-like (2%)
Proliferating Floor Plate Progenitors (5%)
UMAP2
UMAP1
D
Percent Expressed
Average Expression
TH
MAP2
SNCA
DA neurons
Immature DA
Ependymal-like
Late-neuron Prog
Early-neuron Prog
PFPP
Neuroepithelial-like
E
Human Substantia Nigra
Agarwal et. al.
Spearman's R
ODC
Microglia
OPC
TH_Neurons
Astrocytes
Endothelial
FOUNDIN
F
ICC vs scRNA-seq (TH)
R = 0.86
TH Positive Cells (%)
scRNA-seq
ICC
G
ICC vs scRNA-seq (MAP2)
R = 0.89
MAP2 Positive Cells (%)
H
HC (n=11)
Prodromal (n=2)
iPD (n=29)
LRRK2+ (n=22)
GBA+ (n=16)
SNCA+ (n=3)
Cell-type Percentage

Figure 3. Bulk RNA-seq and neuronal differentiation efficiency prediction. A, Principal component analysis (PCA) of bulk RNA-seq showing clustering by timepoint (days 0, 25 and 65). **B-D**, Changes in expression of neuronal (*MAP2*), DA (*TH*) and iPSC (*POU5F1*) genes from day 0 to day 65. **E-I**, Genes significantly correlated with neuronal differentiation efficiency. *HNRNPH3*, *SRSF5* and *HSD17B6* show positive and ZSWIM8 and *ARSA* negative correlation. **J**, Expression levels of genes associated with neuronal differentiation efficiency. All five genes are significantly differentially expressed between day 0 and day 65 (adjusted p-value < 0.05).

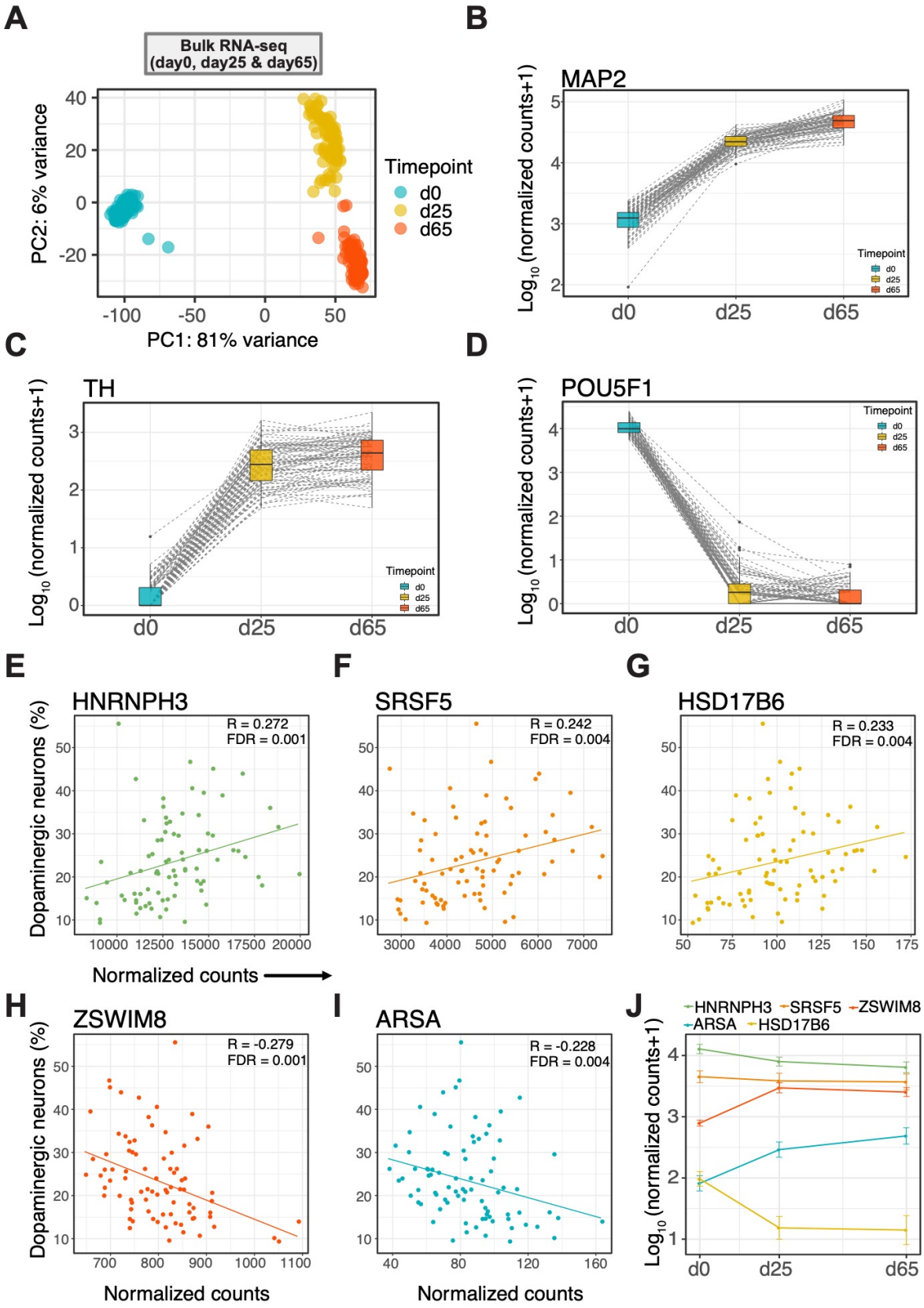

A
Bulk RNA-seq
(day0, day25 & day65)
PC2: 6% variance
PC1: 81% variance
Timepoint
d0
d25
d65
B
MAP2
Log10 (normalized counts+1)
C
TH
D
POU5F1
E
HNRNPH3
R = 0.272
FDR = 0.001
Dopaminergic neurons (%)
Normalized counts
F
SRSF5
R = 0.242
FDR = 0.004
G
HSD17B6
R = 0.233
FDR = 0.004
H
ZSWIM8
R = -0.279
FDR = 0.001
I
ARSA
R = -0.228
FDR = 0.004
J
HNRNPH3
SRSF5
ZSWIM8
ARSA
HSD17B6

Figure 4. Chromatin accessibility in iPSC-derived neurons on day 65. A, PCA across all bulk ATAC-seq samples showing clustering by timepoint. **B**, UMAP of scATAC-seq data at day 65 showing the clustering of 139,659 cells (from 29 samples) and similar broad cell types as in scRNA-seq (Figure 2C). **C**, chromatin accessibility data at the *TH* locus showing timepoint specific peaks identified in bulk ATAC-seq at day 25 and day 65 and cell-type specific peaks in scATAC-seq at day 65. This figure was generated using the FOUNDIN-PD browser (https://www.foundinpd.org).

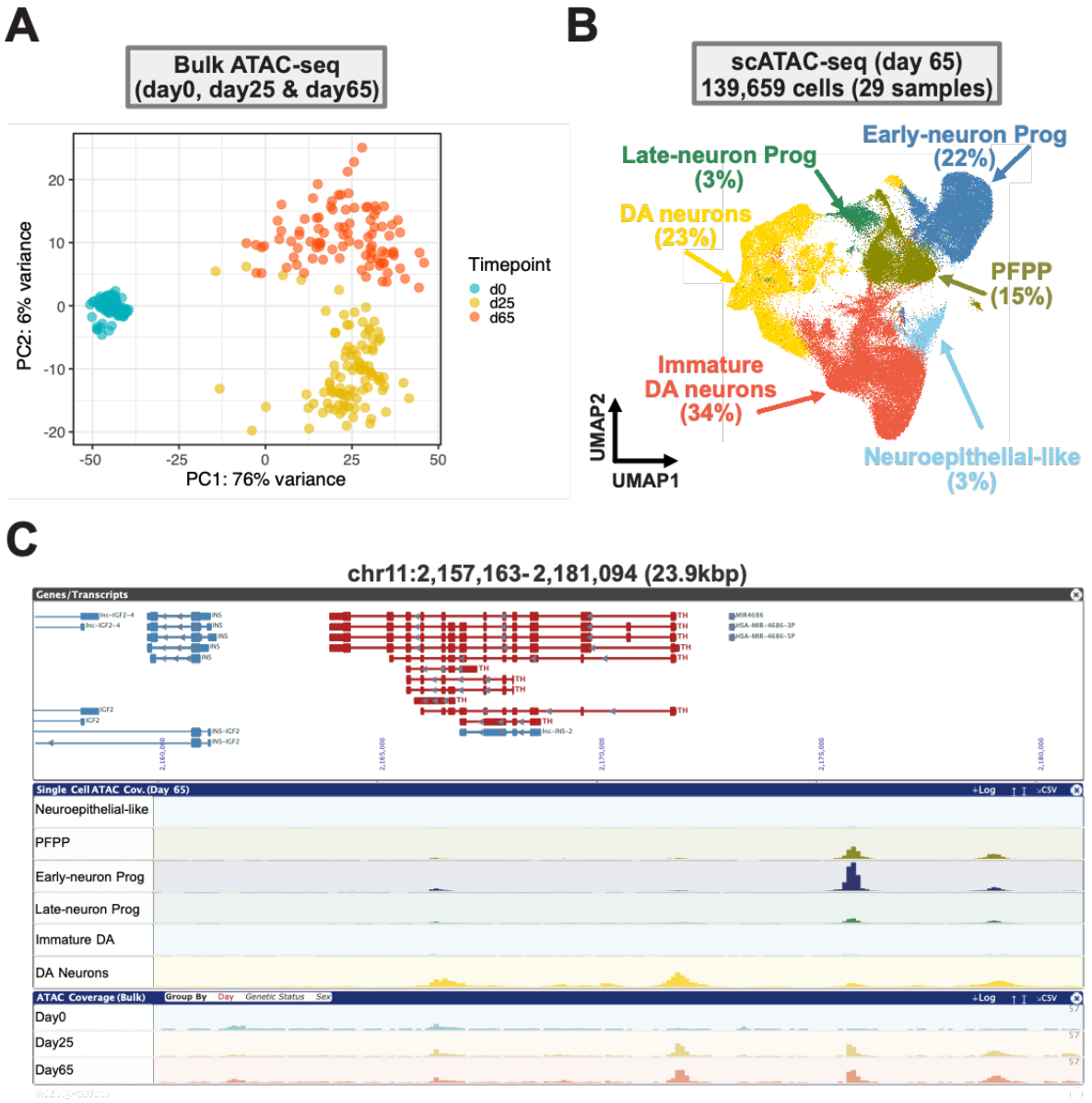


Figure 5. Automated longitudinal imaging of DA neurons. A, Time-lapse imaging of DA neurons (FOUNDIN_5136) expressing synapsin-I driven GFP. Analysis started on day 55-56 of differentiation. One neuron (green arrowhead) survives the entire duration of imaging. A second neuron (red arrow) dies at 96h. Scale bar: 60 µm. **B**, Cumulative risk of death curves showing the neuronal survival from all batch 1 lines over 8 days of automated imaging.

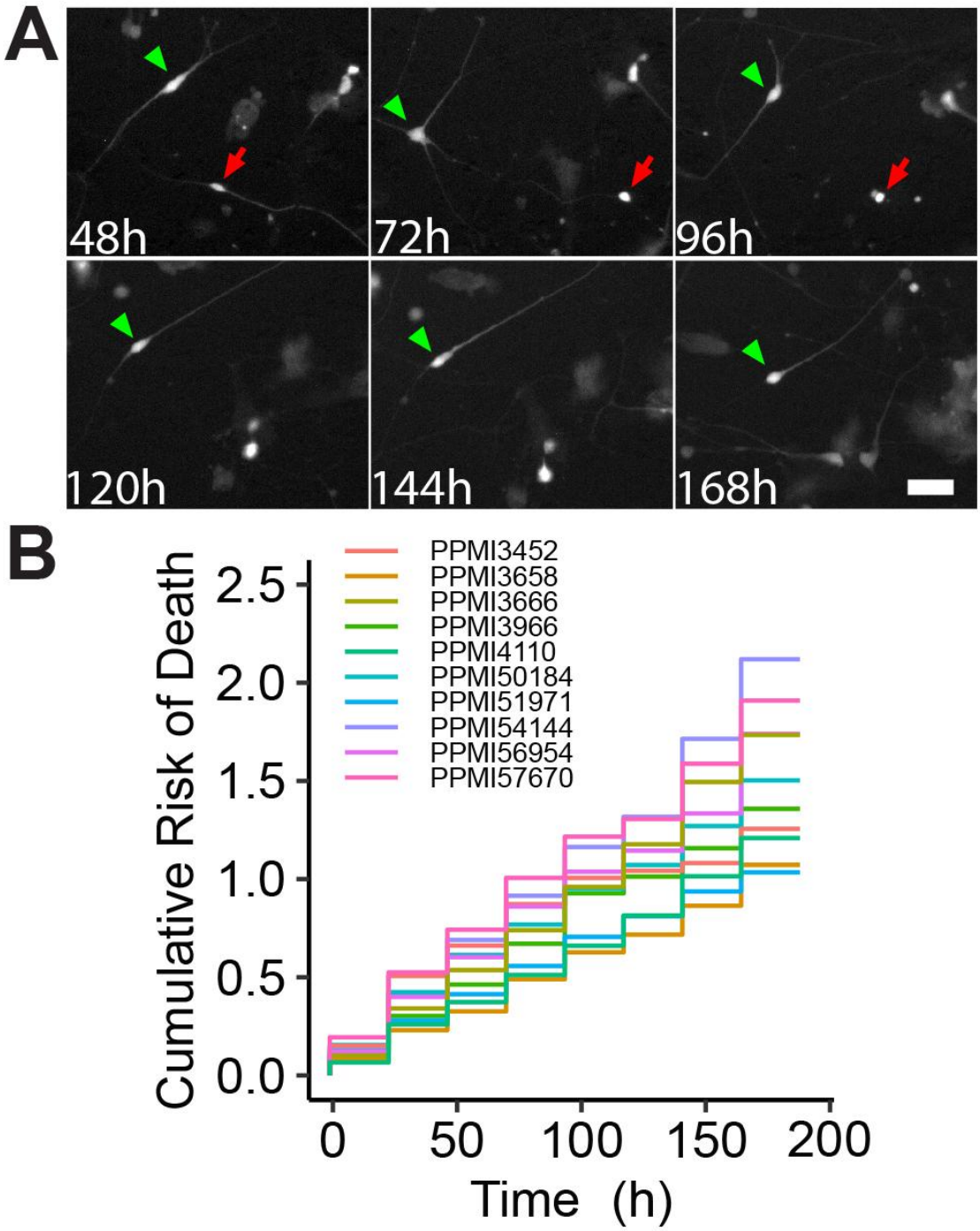
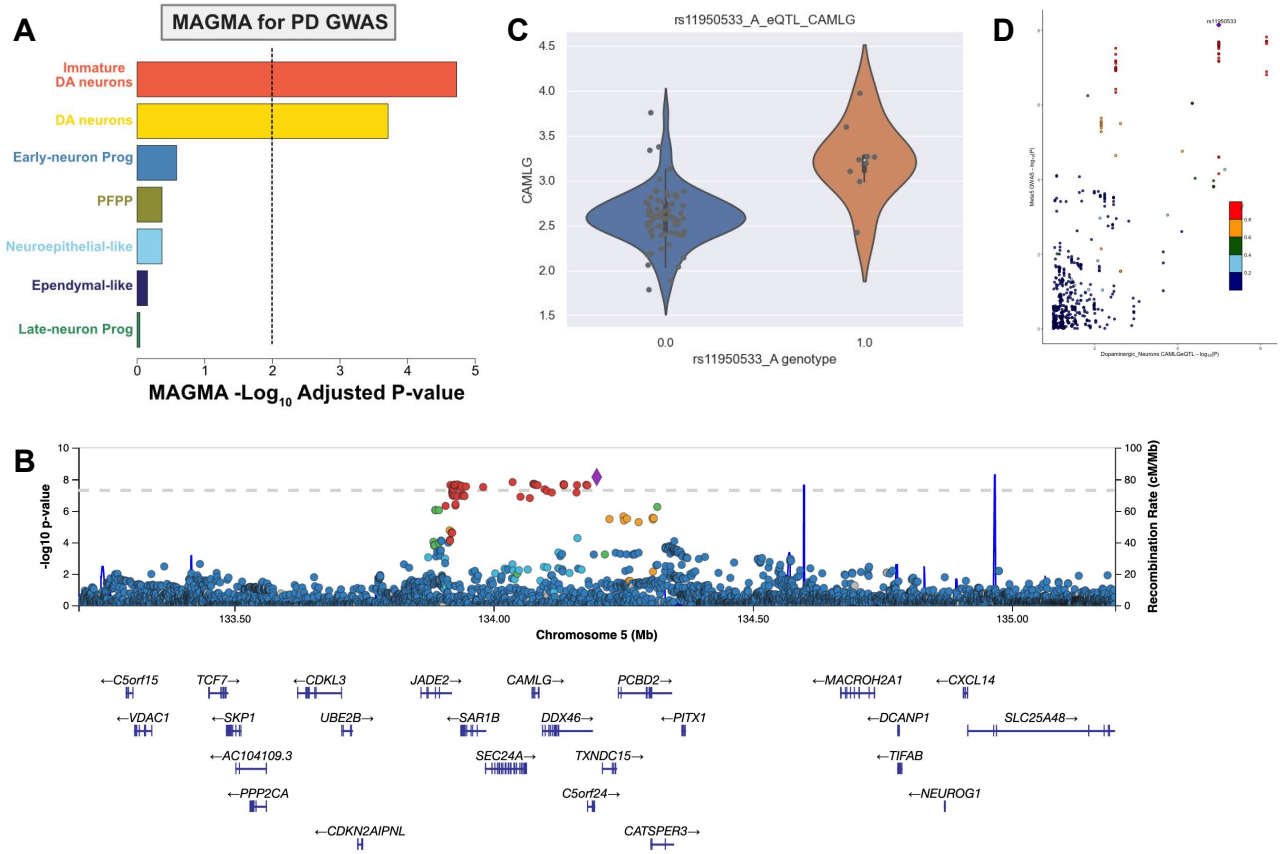


Figure 6. Using scRNAseq expression data to dissect genetic risk. A, MAGMA gene-set enrichment based on the scRNA-seq data showed significant associations with both DA neuron cell clusters. Colors represent the same cell types as in Figure 2C. **B**, LocusZoom plot of locus 28 with rs11950533 as the index variant. Association data is derived from the most recent PD GWAS [2]. **C**, Violin plot showing correlation between the genotype at rs11950533 and expression of *CAMLG* in the DA neuron cell cluster. **D**, LocusCompare plot of the correlation between the PD GWAS [2] association results and the scRNA-seq expression quantitative trait analysis.



Main Table Legends

Table 1. Overview of PPMI iPSC lines included in FOUNDIN-PD. LRRK2+ denotes p.G2019S (n=25) or p.R1441G (n=1), GBA+ denotes: p.N370S (n=19), p.T369M (n=1) or p.E326K (n=1), SNCA+ denotes: p.A53T (n=4). Note that one individual was carrying both *LRRK2* p.G2019S and *GBA* p.N370S and one individual was carrying both *LRRK2* p.G2019S and *GBA* p.T369M. Given the much larger effect size on PD risk of *LRRK2* p.G2019S, these lines were annotated as LRRK2+, but with a comment that they also carry a *GBA* mutation. “No mutation” means that no reported pathogenic mutation in *GBA*, *LRRK2*, *SNCA* or any other known PD gene was identified.

Supplemental Figure Legends

Supplemental Figure 1. Parkinson's disease (PD) genetic risk score based on the effect sizes of the most recent PD GWAS [2]. Groups are split by genetic status. "Other" includes prodromal (n=3) and SWEDD (n=1).

Supplemental Figure 2. Percentage of TH (Millipore) and MAP2 positive cells at day 65. Each dot represents one cell line (n=95).

Supplemental Figure 3. Comparison of the percentages of positive cells co-stained with the same MAP2 (Santa Cruz) but two independent TH antibodies (M, Millipore; PF, Pel-Freeze) by ICC at day 65. The barcharts represent the mean±SD of 95 lines.

Supplemental Figure 4. Percentages of positive cells identified in the control line replicates for one MAP2 and two independent TH antibodies in ICC and scRNA-seq data (SCRN). The barcharts represent the mean±SD of 83 lines that were included in both ICC and scRNA-seq assays. Each symbol represents one replicate of the control cell line.

Supplemental Figure 5. Heatmap showing expression of marker genes used in the cell-type assignment of the single cell RNA-seq data.

Supplemental Figure 6. A, UMAP of cell types identified by Agarwal and collaborators in the post mortem substantia nigra human brain (n=4,781 single cells, n=5+2 replicates). The raw data were analyzed using the FOUNDIN-PD GTF annotation file. 'TH_Neurons' were annotated because this cluster showed up-regulation of *TH*. ODCs, oligodendrocytes; OPCs, oligodendrocyte precursor cells. **B**, Dendrogram showing clustering of FOUNDIN-PD cell types with Agarwal and collaborators' dataset [11] using ClusterMap [63]. **C**, Dendrogram showing clustering of FOUNDIN-PD cell types with Fernandes and collaborators dataset [12] using the ClusterMap tool.

Supplemental Figure 7. A, Correlation between percentages of TH (Millipore) **B**, and MAP2 positive cells in ICC and scRNA-seq (R, Pearson correlation coefficient; $p<0.0001$). Each dot represents one cell line (n=83).

Supplemental Figure 8. Variability in differentiation efficiency across the iPSC lines depicted as Heatmap plot. The number shows the percentage of cell-type for a particular sample and clustering was performed using default settings in ComplexHeatmap [64]. For each sample, annoations are shown at the top depicting batch, phenotype, recruitment category, genetic sex and PD-linked genotype (GBA+, LRRK2+, SNCA+) information.

Supplemental Figure 9. Gene level expression correlation between four technical replicates of the control cell line using scRNA-seq dopaminergic neuron data. Only genes with average expression greater than 0.01 TPM were included. Average expression values were log normalized and compared with other replicates in scatter plots in the lower triangle. Distribution of log normalized values is shown on the diagonal. Pearson's r correlation is displayed in the upper triangle.

Supplemental Figure 10. Gene level expression correlation between five technical replicates of the control cell line using bulk RNA-seq data. A, timepoint day 0; B, timepoint day 25; C, timepoint day 65. Only genes with normalized counts greater than five were included. Counts were log normalized and compared with other replicates in scatter plots in the lower triangle. Distribution of log normalized values is shown on the diagonal. Pearson's r correlation is displayed in the upper triangle.

Supplemental Figure 11. A-D, Expression of monogenic PD genes (*GBA*, *SNCA*, *LRRK2)* and neuronal markers (SYN1) goes up across timepoints in bulk RNA-seq. **E-F**, Stem cell marker gene (*NANOG* and *TDGF1)* expression goes down across timepoints.

Supplemental Figure 12. A, Ten genes identified with non-zero coefficients using elastic net regularization approach (see Methods) to predict neuronal differentiation efficiency. **B**, Confusion matrix results are shown as binary scatter plots. Predicted probability is on the y-axis and initially assigned labels are on the x-axis (see Methods). Each dot represents a cell line. Blue and red color denotes truly and falsely predicted labels by the trained model. Dashed line represents the optimal probability cut-off. **C**, Area under the curve (AUC) using ten genes and all samples for training. **D**, Correlation analysis of these ten genes with neuronal differentiation efficiency. Numbers in the bar are adjusted p-values. * represents significant associations (FDR < 1%).

Supplemental Figure 13. Correlation of MAP2 protein expression from TMT proteomics ratios and bulk RNA-seq *MAP2* expression from day 65 in batch 1.

Supplemental Figure 14. Uniform Manifold Approximation and Projection (UMAP) analysis of methylation data shows clear clustering by timepoints (day 0 vs day 65).

Supplemental Figure 15. A-B, Methylation probes in close proximity to the *TH* transcription start sites show a clear decrease in methylation levels between timepoints (day 0 vs day 65). **C-E**, Methylation probes across the *MAP2* (gene body, 5' UTR and 3' UTR) locus show a decrease in methylation levels between timepoints.

Supplemental Figure 16. A, Cistrome analysis showing the percent of bulk ATAC-seq peaks across genomic regions in a merged peak set at each timepoint (day 0 = red, day 25 = green, day 65 = blue) relative to the background genomic space (black). **B**, The average genetic conservation across species, PhastCons score, across merged ATAC-seq peak sets as cells are differentiated (day 0 = red, day 25 = green, day 65 = blue).

Supplemental Figure 17. Open chromatin at the *MAP2* locus shows differential peaks across timepoints (bulk ATAC-seq) and cell types (scATAC-seq).

Supplemental Figure 18. A, Principal component analysis (PCA) of HiC-seq data shows clustering by timepoints (day 0 vs day 65). **B**, HiC loop structure at the *MAP2* locus shows clear differences between timepoints (day 0 vs day 65).

Supplemental Figure 19. Principal component analysis (PCA) of small RNA-seq data shows clear clustering of expression by timepoint (day 0 vs day 25 vs day 65).

Supplemental Figure 20. Heatmap of small RNA expression at day 65 shows enrichment for small RNAs that have high expression in the brain.

Supplemental Figure 21. Expression quantitative trait loci (eQTL) results of rs62053943 and genes expressed at the *MAPT* region (17q21-31) showing clear eQTLs likely representing the H1/H2 MAPT haplotype status.

Supplemental Figure 22. LocusCompare plots of the correlation between the most recent PD GWAS association results [2] and cortical brain eQTL data [23]. **A-D**, *TXNDC15*, *JADE2*, *SAR1B* and *CAMLG*.

Supplemental Figure 23. Absence of significant expression quantitative trait loci for **A**, *TXNDC15;* **B**, *JADE2* and **C**, *SAR1B* in DA neuron scRNA-seq data.

Supplemental Figure 24. Overview of scRNA-seq data gene level expression in the DA neuron cell cluster at the selected PD GWAS locus.

Supplemental Figure 25. Intersection of the *CAMLG* bulk RNA-seq eQTL signal and the PD risk signal at day 0 (blue), day 25 (red) and day 65 (green).

Supplemental Table Legends

Supplemental Tables 1. Included iPSC overview with meta-data.

Supplemental Tables 2. Single cell RNA-seq cell type cell marker gene overview.

Supplemental Tables 3. Single cell RNA-seq cell type percentage per iPSC line.

Supplemental Tables 4. Bulk RNA-seq gene-set enrichment across timepoints (upregulated).

Supplemental Tables 5. Bulk RNA-seq gene-set enrichment across timepoints (downregulated).

Supplemental Tables 6. MAGMA cell type enrichment results using scRNA-seq cell types.

Supplemental Tables 7. FOUNDIN-PD expression quantitative trait loci results using most recent PD GWAS index variants.

References

1. Blauwendraat, C., Nalls, M. A. & Singleton, A. B. The genetic architecture of Parkinson's disease. *Lancet Neurol.* **19**, 170–178 (2020).

2. Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).

3. Blauwendraat, C. *et al.* Genetic modifiers of risk and age at onset in GBA associated Parkinson's disease and Lewy body dementia. *Brain* **143**, 234–248 (2020).

4. Iwaki, H. *et al.* Penetrance of Parkinson's Disease in LRRK2 p.G2019S Carriers Is Modified by a Polygenic Risk Score. *Mov. Disord.* (2020) doi:10.1002/mds.27974.

5. O'Keeffe, G. W. & Sullivan, A. M. Evidence for dopaminergic axonal degeneration as an early pathological process in Parkinson's disease. *Parkinsonism & Related Disorders* vol. 56 9–15 (2018).

6. Bryois, J. *et al.* Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson's disease. *Nat. Genet.* **52**, 482–493 (2020).

7. Parkinson Progression Marker Initiative. The Parkinson Progression Marker Initiative (PPMI). *Prog. Neurobiol.* **95**, 629–635 (2011).

8. Kriks, S. *et al.* Dopamine neurons derived from human ES cells efficiently engraft in animal models of Parkinson's disease. *Nature* **480**, 547–551 (2011).

9. Bressan, E., Cobb, M. & On behalf of the Foundational Data Initiative for Parkinson's Disease (FOUNDIN-PD). Differentiation of iPSC into dopaminergic neurons. *protocols.io* doi:10.17504/protocols.io.bfpzjmp6.

10. Dhingra, A. *et al.* Automated Production of Human Induced Pluripotent Stem Cell-Derived Cortical and Dopaminergic Neurons with Integrated Live-Cell Monitoring. *J. Vis. Exp.* (2020) doi:10.3791/61525.

11. Agarwal, D. *et al.* A single-cell atlas of the human substantia nigra reveals cell-specific pathways associated with neurological disorders. *Nat. Commun.* **11**, 4183 (2020).

12. Fernandes, H. J. R. *et al.* Single-Cell Transcriptomics of Parkinson's Disease Human In Vitro Models Reveals Dopamine Neuron-Specific Stress Responses. *Cell Rep.* **33**, 108263 (2020).

13. Jerber, J. *et al.* Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nat. Genet.* **53**, 304–312 (2021).

14. Botti, V. *et al.* Cellular differentiation state modulates the mRNA export activity of SR proteins. *J. Cell Biol.* **216**, 1993–2009 (2017).

15. Liu, T. *et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.* **12**, R83 (2011).

16. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).

17. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).

18. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).

19. Arrasate, M., Mitra, S., Schweitzer, E. S., Segal, M. R. & Finkbeiner, S. Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. *Nature* **431**, 805–810 (2004).

20. Arrasate, M. & Finkbeiner, S. Automated microscope system for determining factors that predict neuronal fate. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 3840–3845 (2005).

21. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).

22. Grenn, F. P. *et al.* The Parkinson's Disease Genome-Wide Association Study Locus Browser. *Mov. Disord.* **35**, 2056–2067 (2020).

23. Sieberts, S. K. *et al.* Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions. *Sci Data* **7**, 340 (2020).

24. Li, K. *et al.* cellxgene VIP unleashes full power of interactive visualization, plotting and analysis of scRNA-seq data in the scale of millions of cells. *bioRxiv* 2020.08.28.270652 (2020) doi:10.1101/2020.08.28.270652.

25. Hutchins, E. *et al.* Quality Control Metrics for Whole Blood Transcriptome Analysis in the Parkinson's Progression Markers Initiative (PPMI). doi:10.1101/2021.01.05.21249278.

26. Mollenhauer, B. *et al.* Longitudinal analyses of cerebrospinal fluid α-Synuclein in prodromal and early Parkinson's disease. *Mov. Disord.* **34**, 1354–1364 (2019).

27. Simuni, T. *et al.* Longitudinal Change of Clinical and Biological Measures in Early Parkinson's Disease: Parkinson's Progression Markers Initiative Cohort. *Mov. Disord.* **33**, 771–782 (2018).

28. Global Parkinson's Genetics Program. GP2: The Global Parkinson's Genetics Program. *Mov. Disord.* **36**, 842–851 (2021).

29. Finkbeiner, S. Functional genomics, genetic risk profiling and cell phenotypes in neurodegenerative disease. *Neurobiol. Dis.* **146**, 105088 (2020).

30. Christiansen, E. M. *et al.* In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images. *Cell* **173**, 792–803.e19 (2018).

31. Blauwendraat, C. *et al.* NeuroChip, an updated version of the NeuroX genotyping platform to rapidly screen for variants associated with neurological diseases. *Neurobiol. Aging* **57**, 247.e9–247.e13 (2017).

32. Linsley, J. W. *et al.* Automated four-dimensional long term imaging enables single cell tracking within organotypic brain slices to study neurodevelopment and degeneration.

Commun Biol **2**, 155 (2019).

33. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **44**, W3–W10 (2016).
34. Min, J. L., Hemani, G., Davey Smith, G., Relton, C. & Suderman, M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics* **34**, 3983–3989 (2018).
35. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
36. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
37. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).
38. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
39. PsychENCODE Consortium *et al.* The PsychENCODE project. *Nat. Neurosci.* **18**, 1707–1712 (2015).
40. Wolff, J. *et al.* Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **46**, W11–W16 (2018).
41. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. **17**, 10–12 (2011).
42. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
43. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

44. Volders, P.-J. *et al.* LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Res.* **47**, D135–D139 (2019).

45. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

46. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

47. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

48. Manjang, K., Tripathi, S., Yli-Harja, O., Dehmer, M. & Emmert-Streib, F. Graph-based exploitation of gene ontology using GOxploreR for scrutinizing biological significance. *Sci. Rep.* **10**, 16672 (2020).

49. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

50. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).

51. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).

52. Stuart, T., Srivastava, A., Lareau, C. & Satija, R. Multimodal single-cell chromatin analysis with Signac. doi:10.1101/2020.11.09.373613.

53. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

54. R Core Team. R Core Team (2021). R Core Team. (2021). R: A Language and

Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/.

55. Kuhn, M. Building Predictive Models inRUsing thecaretPackage. *J. Stat. Softw.* **28**, (2008).

56. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).

57. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

58. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).

59. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).

60. Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768–769 (2019).

61. Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (2019).

62. Menden, K. *et al.* Deep learning-based cell composition analysis from tissue expression profiles. *Sci Adv* **6**, eaba2619 (2020).

63. Gao, X., Hu, D., Gogol, M. & Li, H. ClusterMap: compare multiple single cell RNA-Seq datasets across different experimental conditions. *Bioinformatics* **35**, 3038–3045 (2019).

64. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).