

**TRANSCRIPTIONAL REGULATION AND DISRUPTION
IN PARKINSON DISEASE**

by
Sarah McClymont

A dissertation submitted to Johns Hopkins University
in conformity with the requirements for the degree of
Doctor of Philosophy

Baltimore, Maryland
December 2019

© 2019 Sarah McClymont
All Rights Reserved

Abstract

The characteristic motor symptoms of Parkinson disease (PD) are primarily the result of the progressive loss of dopaminergic (DA) neurons of the substantia nigra (SN). The common genetic variants responsible for sporadic PD, the mechanisms by which these variants exert their effects, and the origins of the preferential degeneration of SN DA neurons remain largely unknown.

We first aimed to identify common, non-coding PD-associated variants. We examined the chromatin of embryonic midbrain DA neurons and identified >100,000 regions of open chromatin, many of which are transcriptional enhancers, as demonstrated by a series of transgenic reporter assays. Among these enhancers, one, in intron 4 of the familial PD gene *SNCA*, directed reporter expression in catecholaminergic neurons from transgenic mice and zebrafish. Sequencing this enhancer in 986 PD patients and 992 controls revealed two common variants, rs2583959 and rs2737024, associated with elevated PD risk.

We next assessed how these and other common disease-associated variants influence transcription and disease risk. We developed a regulatory vocabulary of midbrain DA neurons to identify key transcription factors (TFs) and ranked >7,000 disease-associated variants on their capacity to disrupt TF binding. We tested ~20 prioritized variants using *in vitro* luciferase reporter assays. Established neuronal cell culture models are not valid cellular surrogates for embryonic DA neurons and resisted *in vitro* validation. Instead, we employed alternative strategies for evaluating TF and protein binding disruption, like protein binding arrays. Additionally, we

characterized an embryonic mouse-derived SN DA neuron cell culture model, SN4741, and evaluated its potential as a cellular surrogate for testing our variant predictions.

Finally, we queried the developmental origins of the preferential degeneration of SN DA neurons in PD. We examined single-cell transcriptomes from ~1,200 midbrain DA neurons in a mouse model of PD and characterized subpopulations of neurons, genes, and pathways disrupted in PD at an early post-natal time point. The PD mutation induces precocious maturation of neuroblasts into mature SN neurons, accompanied by striking dysregulation of genes involved in mitochondrial function. We propose a model suggesting a developmental origin of PD involving disrupted mitochondrial dynamics altering neuron maturation in key populations of DA neurons.

Advisor: Andrew McCallion, PhD

Reader: Jeremy Nathans, MD, PhD

Preface

This PhD would have been impossible without the support of so many people. It is hard to convey just how thankful I am to everyone who helped me along the way – but I will do my best.

To Dr. Andy McCallion, just thank you. Thank you for being the mentor I needed. Thank you for your training. Thank you for your encouragement. Thank you for providing me so many opportunities for growth. Thank you for your patience. Thank you for teaching me how to be a scientist.

To Dr. Manish Raizada, thank you for taking a risk on a high school student who had never held a pipette before. Thank you for providing the best first lab experience anyone could ask for. Thank you for your kind and caring mentorship. Thank you for your trust.

To Dr. Amelie Gaudin, thank you for teaching me so much about experimental design and creative problem solving. Thank you for allowing me to help; your confidence in my (untested) abilities was so important to me. Thank you for showing me how to be a PhD student.

To Drs. Alireza Navabi and Weilong Xie, thank you for providing me a second lab home. Thank you for teaching me so much about organizing experiments and data. Thank you for showing me that science is sometimes brute force rather than clever experiments.

To my thesis committee members, Drs. Jeremy Nathans, Dimitri Avramopoulos, Loyal Goff, and Valina Dawson, thank you for your advice and mentorship. Thank you for your questions and direction.

To Dr. Valle, thank you for building this amazing training program and the incredible environment of the IGM. To Ms. Sandy Muscelli, thank you for everything that you do – the program would fall apart without you.

To my lab mates, thank you for creating this little family with me. Thank you for providing me with so much support, thoughtfulness, fun, and friendship.

To Xylena, thank you for developing this project that I inherited. Thank you for teaching me so much. Thank you for not yelling at me when I messed up as a rotation student.

To Courtney, thank you for making this lab so welcoming and fun. Thank you for your unwavering support and friendship – you are the best cheerleader I could ask for.

To Becca, thank you for managing this lab so well. Thank you for all your help on my projects and your eagerness to learn and dive into new techniques. Thank you for all the friendship and birthday cakes.

To Paul, thank you for working right alongside me all these years. Thank you for your curiosity, enthusiasm, and imagination – I am so glad to have had the opportunity to work with you, not just as a friend, but also as somebody I respect as a scientist. Thank you for letting me rant at you and letting me hear your rants.

To Bill, thank you for being such a dependable lab mate and friend. Thank you for being so considerate and thoughtful. Thank you for marrying Ashley and bringing her to Baltimore so that I could become friends with her too.

To Nicole, Joey, Eric, Sam, Jay: thank you for your hard work and dedication. Thank you for your care in doing science. Thank you for putting up with me and my fledgling attempts at mentorship.

Thank you all for being not just my lab mates but also my friends.

To my classmates, thank you for being the first friends I made in Baltimore. Thank you for learning alongside me. Thank you for your willingness to work together. Thank you for your continued friendship.

To Laura and Patti, thank you for teaching me the importance of saying “I don’t know” and the willingness to ask questions. Thank you for sticking with me all these years, even when I’m not the greatest at texting back.

To Hannah, thank you for being my friend, classmate, lab mate, and collaborator – you are everything rolled into one amazing person.

To Genay, thank you for being my best friend. Thank you for everything – the last 6 years (and all the ones to come) are unfathomable without you in them.

To my family, thank you for your unconditional love and support. Thank you for encouraging my curiosity and enthusiasm for learning throughout my childhood. Thank you for letting me weigh down your pockets with “flossils” and for taking me to the used bookstore to feed my seemingly endless reading habit. Thank you for not flipping out when I moved to another country for six years of grad school. To my parents, thank you for your steadfast support and encouragement. Thank you for going on so many adventures with me. To Jenna, thank you for being the best sister I could dream of. Thank you for visiting me so much and your enthusiasm for exploring Baltimore. To my grandparents, thank you for your encouragement and love. I love you all.

Table of Contents

Abstract	ii
Preface	iv
Table of Contents	vi
List of Tables	viii
List of Figures	ix
Chapter 1: Introduction	x
1.1 A model complex disease: Parkinson disease	1
1.2 Enhancers and transcriptional regulation	3
1.3 Regulatory variation and human phenotypes	5
1.4 Identifying and validating enhancers.....	6
1.5 Evaluating gene expression	10
1.6 Applying scRNA-seq to biological questions.....	12
1.7 The transcriptional origins of sporadic Parkinson disease	15
1.8 Figures.....	17
Chapter 2: Chromatin-based analyses of dopaminergic neurons	21
2.1 Applying chromatin data to investigate non-coding variation in Parkinson disease	21
2.2 ATAC-seq identifies open chromatin in midbrain and forebrain dopaminergic neurons.....	23
2.3 Candidate regulatory regions are capable of directing expression <i>in vivo</i>	25
2.4 A midbrain-specific enhancer directs expression in catecholaminergic neuron populations.....	27
2.5 Enhancer variants are significantly associated with Parkinson disease risk	30
2.6 Deleting the disease-associated enhancer	31
2.7 Discussion	33
2.8 Methods	38
2.9 Figures and supplementary materials	52

Chapter 3: Transcription factors and non-coding variants that disrupt their binding in dopaminergic neurons.....	76
3.1 Interrogating non-coding variants and their role in disease	76
3.2 Candidate regulatory elements are enriched for transcription factor motifs active in dopaminergic neurons.....	77
3.3 Predicting and testing the effects of regulatory variants	80
3.4 Protein binding arrays are a viable alternative validation strategy	83
3.5 The suitability of <i>in vitro</i> dopaminergic neuron surrogates: the SN4741 cell line.....	85
3.6 Discussion	86
3.7 Methods.....	89
3.8 Figures and supplementary materials	97
Chapter 4: The developmental origins of Parkinson disease.....	111
4.1 Investigating the pathogenesis of Parkinson disease by single-cell RNA-seq	111
4.2 Characterizing E15.5 and P7 dopaminergic neurons in wildtype mice.....	112
4.3 Characterizing P7 dopaminergic neurons in a mouse model of Parkinson disease.....	115
4.4 The Parkinson disease mutation alters cell maturation, gene expression and mitochondrial dynamics.....	118
4.5 Discussion	121
4.6 Methods.....	125
4.7 Figures and supplementary materials	139
Chapter 5: Conclusions.....	168
5.1 Summary of significant findings.....	168
5.2 Future directions.....	171
Chapter 6: Appendices.....	174
6.1 The transcriptional targets of SOX9 in Type II Diabetes	174
6.2 References	203
6.3 Curriculum Vitae	226

List of Tables

Table 2.1: Two tightly linked SNPs within the enhancer are significantly associated with PD risk	62
Table 2.2: A single haplotype, containing the minor alleles of the implicated SNPs, is significantly associated with PD risk	62
Table 4.1: Marker gene expression analysis is used to assign identities to each of the clusters identified at E15.5 and P7	152
Table 4.2: Top six marker genes by foldchange difference per cluster.....	153
Table 4.3: Genes up- and downregulated globally and in a cluster-specific manner	154

List of Supplementary Tables

Supplementary Table 2.1: Summary of counts and percent overlap with the VISTA enhancer browser, related to Figure 2.2A-C	71
Supplementary Table 2.2: Allele and genotype counts and frequencies in PD cases and controls of all variants identified by sequencing within the intronic <i>SNCA</i> enhancer	72
Supplementary Table 2.3: r^2 values measuring linkage disequilibrium between <i>SNCA</i> variants in controls.....	73
Supplementary Table 2.4: Allele and genotype counts and frequencies in PD cases and controls of all variants genotyped from the Guella <i>et al.</i> panel.....	74
Supplementary Table 2.5 : Primer sequences used for qPCR and cloning for in vivo reporter assays	75
Supplementary Table 3.1: Summarizing the disease-associated variants collected for scoring by deltaSVM.....	108
Supplementary Table 3.2: Primer sequences used to clone the constructs for luciferase assay (bold: BP Gateway arms)	109
Supplementary Table 3.3: Primer sequences used for protein binding assays....	110
Supplementary Table 3.4: RT-qPCR primers for testing expression of dopaminergic neuron markers.....	110
Supplementary Table 4.1: Experiment and image identifiers for the Allen Mouse Brain Atlas <i>in situ</i> hybridization slides used in assigning cluster six as granule neurons.....	167

List of Figures

Figure 1.1: The pathological hallmarks of Parkinson disease	17
Figure 1.2: Genomic annotations can be used to prioritize GWAS-implicated variants	18
Figure 1.3: Enhancers regulate transcription	20
Figure 2.1: Preliminary validation of ATAC-seq catalogues generated from <i>ex vivo</i> DA neurons	52
Figure 2.2: Validation of the putative CRE catalogues <i>in vivo</i>	54
Figure 2.3: A MB-specific enhancer directs expression in catecholaminergic populations of neurons known to Parkinson disease biology.....	56
Figure 2.4: A schematic of the chromatin interactions, LD structure, variation, and open chromatin at the <i>SNCA</i> locus	58
Figure 2.5: Enhancer deletion experiments and proposed phenotyping.....	60
Figure 3.1: Identification of transcription factors (TFs) important to DA neurons..	97
Figure 3.2: Investigating rs4988232 for an effect on enhancer activity	99
Figure 3.3: Investigating the top variants falling in midbrain open chromatin regions for their effect on enhancer activity.....	101
Figure 3.4: Identification of proteins whose binding is impacted by the implicated PD-risk SNPs	103
Figure 3.5: Initial characterization of the SN4741 cell line suggests it is an unstable, triploid cell line that express markers of differentiation but not of dopaminergic neurons.....	105
Figure 4.1: scRNA-seq identifies subpopulations of dopaminergic neurons at E15.5 and P7	139
Figure 4.2: Seven transcriptionally distinct clusters of cells are identified	141
Figure 4.3: All clusters are assigned a biological identity and marker genes of each are identified for functional validation by single-molecule RNA FISH (smFISH)...	142
Figure 4.4: The VTA and two SN clusters are transcriptionally related and smFISH will be performed to assess their spatial relationships.....	144
Figure 4.5: The mutant α -synuclein transgene is expressed at post-natal day 7 and alters dopaminergic neuron population proportions.....	146
Figure 4.6: There is cluster-specific gene dysregulation of genes previously implicated in neurodegeneration.....	147
Figure 4.7: There is striking and extensive transcriptional dysregulation of the oxidative phosphorylation pathway in the mutant dopaminergic neurons.....	149
Figure 4.8: A mitochondrial fission or fusion defect could underlie the observed precocious maturation phenotype involving the neuroblasts and substantia nigra populations.....	151

List of Supplementary Figures

Supplementary Figure 2.1: RT-qPCR of key DA neuron markers.....	63
Supplementary Figure 2.2: <i>in silico</i> quality control metrics for the ATAC-seq libraries	64
Supplementary Figure 2.3: Correlation analysis of all ATAC-seq libraries.....	66
Supplementary Figure 2.4 Relating RNA-seq and ATAC-seq data	67
Supplementary Figure 2.5: All <i>lacZ</i> reporter mice and the mouse genomic locations of the putative CREs	69
Supplementary Figure 3.1: Motif analysis identifies transcription factors (TFs) important specifically for MB regulatory potential	107
Supplementary Figure 4.1: Mating scheme to generate litters with fluorescently-labelled dopaminergic neurons and all three A53T genotypes.....	155
Supplementary Figure 4.2: Quality control metrics used to filter scRNA-seq data	156
Supplementary Figure 4.3: Replicates and genotypes are well distributed across PC1 and PC2	157
Supplementary Figure 4.4: Replicates and genotypes are present in each cluster	158
Supplementary Figure 4.5: Assessing the clusters for bias in percent mitochondrial reads, number of reads and number of genes expressed per cell.....	159
Supplementary Figure 4.6: Cluster 5 displays markers of support cells, like oligodendrocytes and astrocytes	160
Supplementary Figure 4.7: Cluster 6 is not dopaminergic and likely represents a contaminating cell type.....	162
Supplementary Figure 4.8: The contaminating cluster 6 is likely to be granule cells.....	164
Supplementary Figure 4.9: Expression of previously established marker genes was used to assign biological identities to each of the remaining clusters.....	165

Chapter 1: Introduction

1.1 A model complex disease: Parkinson disease

Parkinson disease is the second most common neurodegenerative disease, affecting 1% of the population over the age of 60¹. It is characterized by progressive motor symptoms, like tremor, bradykinesia, rigidity, and postural instability². It is often preceded, over the course of decades, by a variety of non-motor phenotypes like insomnia, REM sleep disorders, depression and anxiety, anosmia, and constipation³. These motor symptoms are primarily the result of the degeneration of dopaminergic (DA) neurons, particularly in the *substantia nigra*⁴. This degeneration of nigral DA neurons and the presence of protein aggregates of α -synuclein, called Lewy bodies, are the pathological hallmarks of Parkinson disease (**Figure 1.1**).

Parkinson disease is a complex disease, likely arising from a combination of environmental and genetic factors. In terms of the genetic causes, both familial and sporadic forms exist, with familial cases representing less than 15% of cases⁵. Of these familial cases, 30% can be attributed to highly penetrant, monogenic mutations⁶. Both autosomal dominant (e.g.: *SNCA*⁷, *LRRK2*^{8,9}, and *VPS35*¹⁰) and autosomal recessive (e.g.: *PARKIN*¹¹, *DJ-1*¹², *PINK1*¹³) mutations have been identified. Interestingly, at the *SNCA* locus, in addition to six autosomal dominant missense mutations causing disease^{7,14–18}, duplications¹⁹ and triplications²⁰ of the gene results in disease, with the severity of the phenotype and age of onset commensurate with the number of gene copies²¹.

Sporadic Parkinson disease is more common than familial forms of the disease and often exhibits a later onset of phenotypes (familial age: 59.3 ± 11.3 vs sporadic

age: 66.5 ± 11.8)²². For sporadic cases, there exists a genetic component of risk, with a narrow-sense heritability of ~ 0.209 (95% confidence interval [CI]: 0.148-0.271)²³. Genome-wide association studies (GWASs) have been applied to elucidate the genetic basis of sporadic Parkinson disease and has implicated approximately 90 loci in conferring risk, finding both novel loci but those containing many of the familial Parkinson disease genes²⁴. For example, the most significant signal in sporadic Parkinson disease has consistently been located at the *SNCA* locus (odds ratio (OR) = 0.760, p-value = 4.16×10^{-73})²⁵, which given its role in Lewy bodies and the pathogenesis of Parkinson disease suggests common mechanisms are fundamental to both familial and sporadic forms.

Unfortunately, given the reliance of GWAS on linkage disequilibrium (LD), the causative variants at the implicated loci and the genes they disrupt are obscured. Further confounding the study of these loci, the majority of GWAS-implicated variants fall in non-coding sequences²⁶, where the mechanism by which they confer risk is less straightforward than in coding variants. To begin prioritizing non-coding variants at GWAS implicated loci, functional fine-mapping has often been employed²⁷. Here, the lead variant indicted by GWAS plus those in high LD are intersected with genomic annotations, like ChIP-seq, to identify variants that are more likely to be functional (**Figure 1.2**). Underlying this technique is the assumption that these non-coding variants confer risk by altering gene expression through disruption of a functional non-coding element, particularly enhancers.

1.2 Enhancers and transcriptional regulation

Enhancers are a cis-acting, non-protein coding DNA element that regulate gene transcription. Enhancers were first described in 1981, when a fragment of non-coding DNA from the SV40 viral genome was observed to increase transcription of a reporter gene in HeLa cells by several hundred-fold²⁸. Early studies into the general properties of enhancers found them to boost transcription of target genes in a tissue-specific manner that is irrespective of orientation (e.g.: forward strand or reverse strand), distance, or position relative to the target gene²⁸. Gene regulation by enhancers is highly dynamic across cell state (e.g.: cell type, developmental time, environmental or genetic perturbation)^{29,30} and each gene is often under the regulation of multiple enhancers^{31,32}, each at various strengths of activity. This combinatorial control allows for the intricacy and complexity of gene expression patterns.

Enhancers achieve this intricate transcriptional regulation largely through three mechanisms: transcription factor binding, DNA looping, and chromatin accessibility changes.

At the sequence level, enhancers contain transcription factor binding sites³³. The number of sites³⁴, the affinity of transcription factors for that sequence³⁵, and the availability of transcription factors to bind these motifs all modulate an enhancer's activity³⁶. Classically, it was believed that the transcription factors recruited to the enhancer would interact with RNA polymerase II at the target gene promoter to initiate gene transcription^{37,38}. Recently, it has been suggested that the polymerase initiates transcription in the absence of the transcription factors or enhancer, but

instead is paused and requires the transcription factors and enhancer to promote transcriptional elongation^{39,40}.

Regardless the model, there is a requirement for the enhancer to interact with its cognate gene promoter. This is achieved through DNA looping^{41,42}. There is a higher order 3D structure of DNA that regulates gene expression through the formation of loops in which an enhancer is brought into close physical proximity to its target promoter⁴³. This DNA looping structure is formed and maintained by the cohesin complex^{44,45} and the contacts between the enhancer and the promoter are likely bridged by the Mediator complex⁴⁶. Alterations in the chromatin conformation, either by the formation or dissolution of different complements of DNA loops, allow for the dynamic regulation of gene expression^{42,47}.

The compaction of the chromatin is also an important factor in regulating transcription. DNA is packaged tightly around histone octamers at regular 146bp intervals to form nucleosomes⁴⁸. This packaging not only compacts the DNA into the nucleus but it also regulates gene activity by restricting transcription factor binding and therefore, transcription. For an enhancer to function, the underlying DNA containing the transcription factor binding sites must be made accessible^{49,50}. Chromatin accessibility can be modulated by pioneering transcription factors and DNA remodelling proteins in a cell type specific manner⁵¹. Pioneer transcription factors are a class of proteins that have a strong DNA binding affinity which are able to recognize their motifs even when the sequence is partially occluded by the histone packaging^{52,53}. Once bound, pioneer transcription factors can recruit chromatin remodelers or other transcription factors to promote DNA accessibility⁵⁴⁻⁵⁶. Chromatin remodelers are ATP-dependent proteins that shuffle or remove histones to establish

nucleosome-free DNA that is accessible to non-pioneering transcription factors or other DNA binding proteins, like RNA polymerase II⁵⁷. Only once the DNA is accessible, looped such that open enhancers are brought into proximity with their target promoters, and bound by transcription factors and other complexes, effectively regulated transcription can be realized (summarized in **Figure 1.3**).

1.3 Regulatory variation and human phenotypes

While protein-coding mutations have traditionally been the focus of searches for disease-associated variation, non-coding, regulatory variation is frequently the culprit in both rare, Mendelian and common, complex diseases.

An early example of disease-causing mutations disrupting enhancer element activity is in β -thalassaemia, where translocations and deletions of a series of enhancers upstream of the β -globin gene cluster (the locus control region) ablates β -globin gene expression, leading to disease⁵⁸. We see another classic example of rare enhancer variants leading to disease in autosomal dominant pre-axial polydactyly. Here, point mutations in an enhancer of *SHH*, located over 1Mb away in an intron of a neighbouring gene⁵⁹, disrupt the expression of *SHH* in the developing limb bud, disrupting the normal morphogen gradient specifying digit location along the antero-posterior axis⁶⁰. Rare, highly penetrant point mutations in enhancers have been identified in other disorders like mutations in *PTF1A* in pancreatic agenesis⁶¹, *de novo* mutations in *PAX6* in aniridia^{62,63}, and *TBX5* in congenital heart defects⁶⁴.

There are also examples of common variants impacting variants and contributing to disease, as in a common variant occurring in an enhancer in intron 1 of *RET* that increases risk for Hirschsprung disease in a sex-specific manner⁶⁵. Many

common enhancer variants associated with common disease have been identified by GWAS. GWAS have identified thousands of risk loci, many of which do not contain protein-coding causal variants^{26,66}, suggesting a role for regulatory variants in conferring disease risk. These non-coding variants can be prioritized through intersections with disease-relevant, tissue-specific enhancer catalogues combined with mechanistic studies to identify the causative SNP at a locus. By these methods, common obesity-associated enhancer variants were found to regulate *IRX3* expression^{67,68}. Similarly, genetic fine-mapping studies in combination with genetic editing experiments identified an enhancer variant of *EDN1* that is associated with five vascular diseases, as identified by GWAS⁶⁹.

While these success stories exist, our ability to prioritize variants through intersections with enhancer catalogues are limited by our ability to identify enhancers; the specificity of enhancers to tissue types, developmental time, pathological status, and environmental conditions, remain barriers to prioritizing the causative non-coding variant at many GWAS-identified loci.

1.4 Identifying and validating enhancers

Unlike with coding sequences, we cannot reliably predict enhancers from DNA sequence alone. Machine learning algorithms have recently been employed to attempt to learn enhancer sequence composition in order to predict enhancers from sequence alone⁷⁰⁻⁷⁴. These algorithms are still limited in their predictions, especially given the lack of quality training sets and the complexity of the sequence encoded by degenerate and combinatorial transcription factor binding site sequences. Sequence conservation has also been used to predict enhancers under the expectation that enhancers are

functional elements and therefore will be conserved across species at higher rates than background, non-functional, non-coding DNA⁷⁵⁻⁷⁹. This approach has been effective in finding a variety of enhancers however, fails to identify a large proportion of functional non-coding sequences and also misses enhancers that are recent in the human evolutionary history⁸⁰⁻⁸².

Most enhancer searches have shifted away from sequence analysis and have instead focused on molecular methods to identifying enhancers. These methods generally rely on exploiting the properties of enhancers to aid their searches.

Exploiting that enhancers are bound by transcription factors, ChIP-seq (chromatin immunoprecipitation) is used to assay transcription factor binding sites genome-wide, a proportion of which will be enhancers⁸³. Often ChIP-seq is also used to pull down against histone marks characteristic of enhancers (e.g.: H3K27Ac and H3K4me1)⁸⁴⁻⁸⁶ or enhancer associated proteins, like EP300, a transcriptional activator^{87,88}.

We can also exploit the requirement of enhancers to physically interact with target promoters as a means to identify enhancers. Using chromatin conformation capture techniques, we can identify sequences that physically interact with a promoter. Using 4C-seq, we can use a promoter of interest as the anchor/viewpoint and identify all sequences that interact with it, likely including enhancers⁸⁹. On a larger scale, we can perform promoter-capture HiC to identify all sequences interacting with all promoters^{90,91}, under the assumption that a large fraction of non-coding interacting sequences are likely to be enhancers.

The increased chromatin accessibility in active enhancers is the main feature currently being exploited to identify enhancers genome-wide. Classic methods of probing DNA accessibility include DNase I hypersensitivity site sequencing (DNaseI-seq)^{92,93} and FAIRE-seq (formaldehyde-assisted isolation of regulatory elements)⁹⁴. DNaseI-seq relies on open chromatin being more susceptible to degradation by DNaseI. FAIRE-seq relies on open chromatin being more readily solubilized after crosslinking and sonication.

Both of these methods have successfully identified extensive catalogues of enhancers however, they both rely on assaying a large number of cells (on the order of tens of millions). This cell number requirement often precludes the study of rare populations of cells or in studying specific cell types rather than a whole tissue. As such, these techniques are often limited to use in cell culture models, in order to meet the input requirements, prohibiting the study of enhancers in their *in vivo* context, or in assaying across developmental time or in an appropriate or perturbed environment.

To overcome these limitations, the assay of transposase accessible chromatin with sequencing, ATAC-seq, has recently been developed⁹⁵. Like DNaseI-seq and FAIRE-seq, ATAC-seq relies on a specific property of open chromatin for assay, specifically that open chromatin is more susceptible to fragmentation by a hyperactive transposase. With this method, the cell input requirements are orders of magnitude less, requiring 500 to 50,000 cells, with the ability to assay open chromatin in single-cells also now a possibility.

None of these methods for identifying enhancers solely identify enhancers – other sequences of the genome may interact with transcription factors or DNA-binding

proteins, or are in contact with a promoter, or reside in open chromatin. Once we have generated these catalogues, there remains the burden of proof in verifying these sequences as capable of regulating transcription.

The main methods to validating enhancer activity generally rely on assaying reporter activity, in which the candidate enhancer sequence is placed upstream of a minimal reporter that directs expression of a reporter gene. In cell culture, this is often a luciferase assay. Luciferase assays use the firefly luciferase and renilla genes to quantitatively measure fluorescence activity being directed by a putative enhancer⁹⁶. This is a medium throughput assay with dozens of enhancers routinely tested at once. Scaling this up, massively parallel reporter assays (MPRAs) are a related, higher throughput assay, where instead of measuring luciferase fluorescent activity, the transcription of a reporter gene is measured by RNA-seq, allowing the assay of tens of thousands of sequences simultaneously^{97–100}.

Both of these methods are *in vitro*, transient expression assays. As such, they are unable to assess the cell type specificity of enhancer activity, are limited to assaying a single cell type and cannot assess the impact of developmental time or environmental perturbation on enhancer activity. Reporter assays in transgenic animals overcome these issues. These assays usually involve random integration of an enhancer-promoter-reporter construct into the animal's genome. Often, the reporters used in *in vivo* reporter models are *lacZ* expression with β -gal staining ("blue mice")^{101,102} or fluorescence activity – often used in zebrafish given their relative translucency throughout development¹⁰³.

These assays allow for the measurement of enhancer activity across developmental time, tissues, and perturbation. These assays are informative for the deep understanding of an enhancer but do not scale well to assaying large numbers of enhancers simultaneously. Unfortunately, neither the *in vitro* or *in vivo* reporter assays generally measure enhancer activity using the endogenous locus or promoter and as such do not preserve one of the core principles enhancers use in regulating gene expression: chromatin looping.

A recent method in probing enhancer activity seeks to overcome these limitations through genome editing at the endogenous locus, usually by CRISPR-Cas9^{104–106}. Here, the native locus is edited to disrupt an enhancer, either by modifying the chromatin state¹⁰⁷, introducing a variant^{67,108,109} or through its wholesale deletion^{110,111}, and changes to gene expression and chromatin conformation are measured. If the editing occurs in an *in vivo* model, the animal model can be examined for disease-related phenotypes and disease susceptibility can be tested. This method of validating enhancer activity remains low throughput but is commonly used in confirming enhancer activity or in assessing the effect of an enhancer variant on the cognate gene expression *in vivo*.

1.5 Evaluating gene expression

Enhancers regulate gene expression. Our ability to measure gene expression is fundamental to our understanding of how DNA gives rise to phenotypes. To measure gene expression, we can examine either the ultimate gene product, proteins, or we can consider the intermediary product, RNA.

Proteins are the culmination of all regulatory steps from transcription to translation and are the ultimate agents conferring phenotypes. Common strategies for measuring protein abundance include antibody-based methods, including western blot^{112,113} and enzyme-linked immunosorbent assay (ELISA)¹¹⁴, or spectrometry-based methods, like liquid chromatography mass spectrometry (LC/MS)¹¹⁵. While these methods are able to measure protein abundance with high sensitivity, they can be low throughput, rely on the availability and specificity of an appropriate antibody against the protein of interest (western, ELISA), are technically demanding (LC/MS), or require a large amount of sample (e.g.: western blots can require ~10-100µg of total protein¹¹⁶) precluding measurement of protein levels for rare populations of cells¹¹⁷.

Measuring RNA levels avoid these issues but relies on the assumption that transcript levels correlate with protein levels, which has been called into question^{118,119}. However, measuring the transcriptome is the favoured method for evaluating gene expression due to the relative ease of the protocols, the scale of data generated, and the low input requirements (i.e.: as low as single cells). Methods for measuring RNA levels, from low- to high-throughput, include northern blot¹²⁰, RT-qPCR¹²¹⁻¹²³, microarrays¹²⁴⁻¹²⁶, and RNA-seq¹²⁷⁻¹³¹. RNA-seq is a relatively new technology that generally involves RNA conversion to cDNA and ligation of sequencing adapters followed by next generation sequencing. The reads are aligned to either the transcriptome or genome and gene expression is quantified, novel splicing events or isoforms are uncovered, and non-coding RNAs (e.g.: miRNAs, lncRNAs, eRNAs) are identified¹³². RNA-seq allows for the quantification and comparison of transcripts across cell types, tissues, individuals, time points, disease state or, environmental perturbation with high sensitivity. However, bulk RNA-seq is

limited to measuring average gene expression across pools of cells, potentially masking subpopulations of cells. In response to this shortcoming, new protocols for quantifying RNA expression in single cells were developed.

The invention of single-cell RNA-seq (scRNA-seq)¹³³⁻¹³⁷ has allowed for the measurement of transcripts from hundreds to thousands of single cells simultaneously, enabling the in-depth analysis of heterogeneous populations of cells. scRNA-seq involves the isolation of single cells (either through cell sorting¹³⁸, combinatorial indexing¹³⁹, or droplet-based techniques¹³⁷), reverse transcription of extracted RNA into cDNA, and the production of sequencing-ready libraries with unique barcodes for each cell for deconvolution post-sequencing. Each cell's sequences are aligned and quantified and sparse matrices of cells by gene expression are generated to summarize the expression data. The gene expression data is sparse given the low efficiency of RNA capture coupled with the low expression of many genes¹⁴⁰. This scarcity of data is mitigated by the collection of a large number of cells; there is a tradeoff in analysing sparse data from many cells versus analysing very deep data from just a few samples. Once the expression is quantified across all cells, cells are clustered using dimensionality-reduction algorithms (e.g.: PCA¹⁴¹, tSNE¹⁴², UMAP¹⁴³) into subpopulations of related cells. From these clusters, marker genes of each can be identified, subpopulations of cells can be assigned to biological functions or anatomical regions, and genes differentially expressed between clusters are identified.

1.6 Applying scRNA-seq to biological questions

With this technology in hand, scRNA-seq has been rapidly applied to a variety of biological questions. scRNA-seq has been used to uncover tissue and cellular

heterogeneity, identify and profile rare cell types, and compare healthy and diseased tissues.

In one of the first demonstrations of scRNA-seq, the fledgling technology was applied to examining cancer, focusing on circulating tumour cells in a melanoma patient¹³⁵. This marked the beginning of a slew of studies into tumoural heterogeneity and microenvironment, markers of disease progression or metastasis, and response to therapy using scRNA-seq. For example, common brain tumours, glioblastoma and medulloblastoma, were examined for intratumoural heterogeneity. Researchers identified four cellular subtypes in glioblastoma tumours and three clusters of distinct cells in medulloblastoma tumours^{144,145}. Each tumour is composed of each of the subtypes and depending on the composition and degree of heterogeneity within a tumour, survival was influenced¹⁴⁶. Like tumours, brains are incredibly complex and highly heterogeneous; even within a brain region, there is a large diversity of cell types, both neuronal and non-neuronal. scRNA-seq has been used to identify cellular subtypes in a variety of brain regions, like the cortex and hippocampus where researchers identified 47 clusters¹⁴⁷, the temporal lobe¹⁴⁸, and the midbrain¹⁴⁹. Specific types of neurons and support cells have been assayed as well to find subtypes within these restricted populations. For example subsets of oligodendrocytes¹⁵⁰ and dopaminergic neurons¹⁵¹ have been identified.

These studies have used scRNA-seq to better understand cell type composition within a tissue. Even when assaying tissues that have been studied extensively, scRNA-seq can identify novel cell populations. In the case of blood dendritic cells and monocytes, which had been canonically classified into six subtypes, scRNA-seq suggests there are instead ten subtypes¹⁵². One of these subtypes was a novel

progenitor population, making up just 0.02% of cells assayed, which while rare, could be exploited as a new therapeutic target. Similarly, in the lung epithelia, a novel, rare population of cells, composed of pulmonary ionocytes, was identified by scRNA-seq^{153,154}. Interestingly, this population of cells, representing less than 1% of airway epithelial cells, express high levels of *CFTR*, suggesting that this rare, previously unidentified cell population may be involved in cystic fibrosis pathogenesis.

Clearly, scRNA-seq is an important technology in understanding disease. Many studies have used scRNA-seq to compare the diseased and healthy state in order to understand the molecular mechanisms of disease pathogenesis and progression. In one example, single cells were isolated from the lungs of control and influenza-treated mice and the host and viral transcriptomes were measured¹⁵⁵. Doing so, nine clusters of cells, corresponding to the major classes of cells expected, were observed. Unexpectedly, each cluster of cells contained cells with high rates of viral infection, where previously it had been believed that only certain subpopulations of cells carried the burden of infection^{156,157}. Another example comparing healthy and diseased tissues involved the comparison of failing and non-failing heart transcriptomes¹⁵⁸. This comparison identified subpopulations of cardiomyocytes that execute a dedifferentiation transcriptomic profile in the disease-state, suggesting an effort for these cells to regenerate following stress. A subsequent paper compared cardiomyocytes in failing hearts and in hearts that are stressed but not yet failing, to examine the molecular basis of that transition¹⁵⁹. These two studies indicate the value in using scRNA-seq to identify populations of cells implicated in disease.

Bringing together all of these use cases for scRNA-seq, we can turn to a study that used scRNA-seq to characterize the heterogeneity of cell types in the entorhinal

cortex in the context of Alzheimer disease (AD)¹⁶⁰. In this study, they identify subpopulations of microglia, astrocytes, neurons, oligodendrocytes, oligodendrocyte progenitors, and endothelial cells in the entorhinal cortex. Some of these clusters were novel and originated exclusively from AD brains, suggesting extensive transcriptional changes as a result of the disease. Examining these disease-specific clusters, groups of dysregulated genes, particularly transcription factor networks, are implicated in regulating cell fate transitions between healthy and AD subpopulations. This study exemplifies the power of scRNA-seq in defining the cellular heterogeneity of tissues, identifying new and rare cell subtypes, in the context of disease to better understand the mechanisms underlying pathogenesis and disease progression.

1.7 The transcriptional origins of sporadic Parkinson disease

We can consider sporadic Parkinson disease to be the result of variation leading to aberrant transcription ultimately leading to the preferential degeneration of midbrain dopaminergic neurons. To investigate this process, I set out to investigate each of these steps in depth.

I was motivated to identify non-coding variants that are responsible for elevating Parkinson disease risk. In Chapter 2, I identify enhancers in midbrain dopaminergic neurons and focus on a novel midbrain-specific enhancer of *SNCA* and within it, reveal two common Parkinson disease-associated variants.

I also wanted to investigate how disease-associated non-coding variants alter transcription in modulating risk. In Chapter 3, I identify important transcription factors for midbrain dopaminergic neurons and predict how thousands of variants associated with risk for a variety of neurodegenerative and neuropsychiatric disorders

alter their binding. In evaluating these predictions, I explore the suitability of a variety of *in vitro* cellular surrogates, particularly focusing on how appropriate an immortalized mouse substantia nigra cell line is as a surrogate for *in vitro* assays.

Finally, I was curious about the predisposition of dopaminergic neurons of the substantia nigra to preferentially degenerate in Parkinson disease. In Chapter 4, I evaluate the transcriptomes of developing dopaminergic neurons of the midbrain in the context of a mouse model of Parkinson disease. I identify subpopulations of neurons of early postnatal dopaminergic neurons, including a novel population of substantia nigra neurons. Additionally, in comparing the diseased and healthy states, I observe striking dysregulation of mitochondria that we hypothesize underlies the precocious differentiation of neuroblasts into mature substantia nigra neurons.

Overall, I examine how alterations in DNA sequence, transcription factor binding, and transcription culminate in risk for Parkinson disease.

1.8 Figures

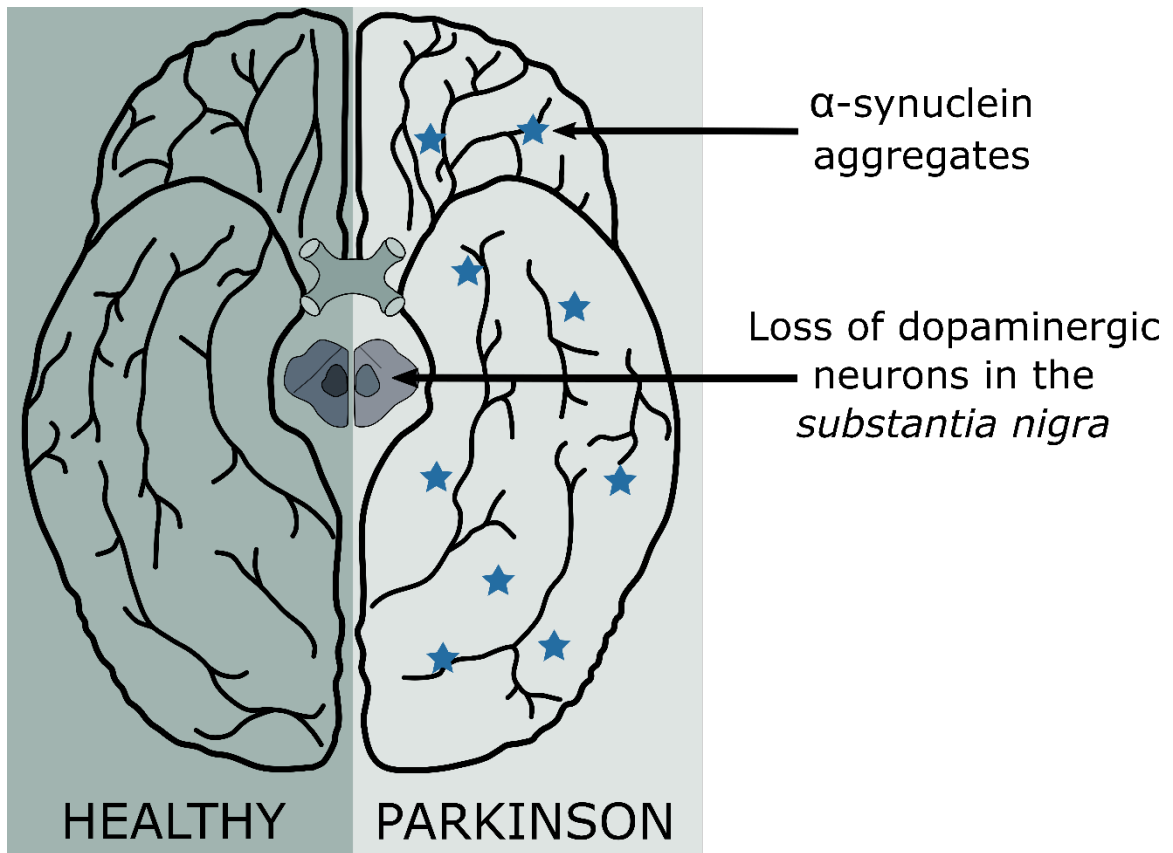
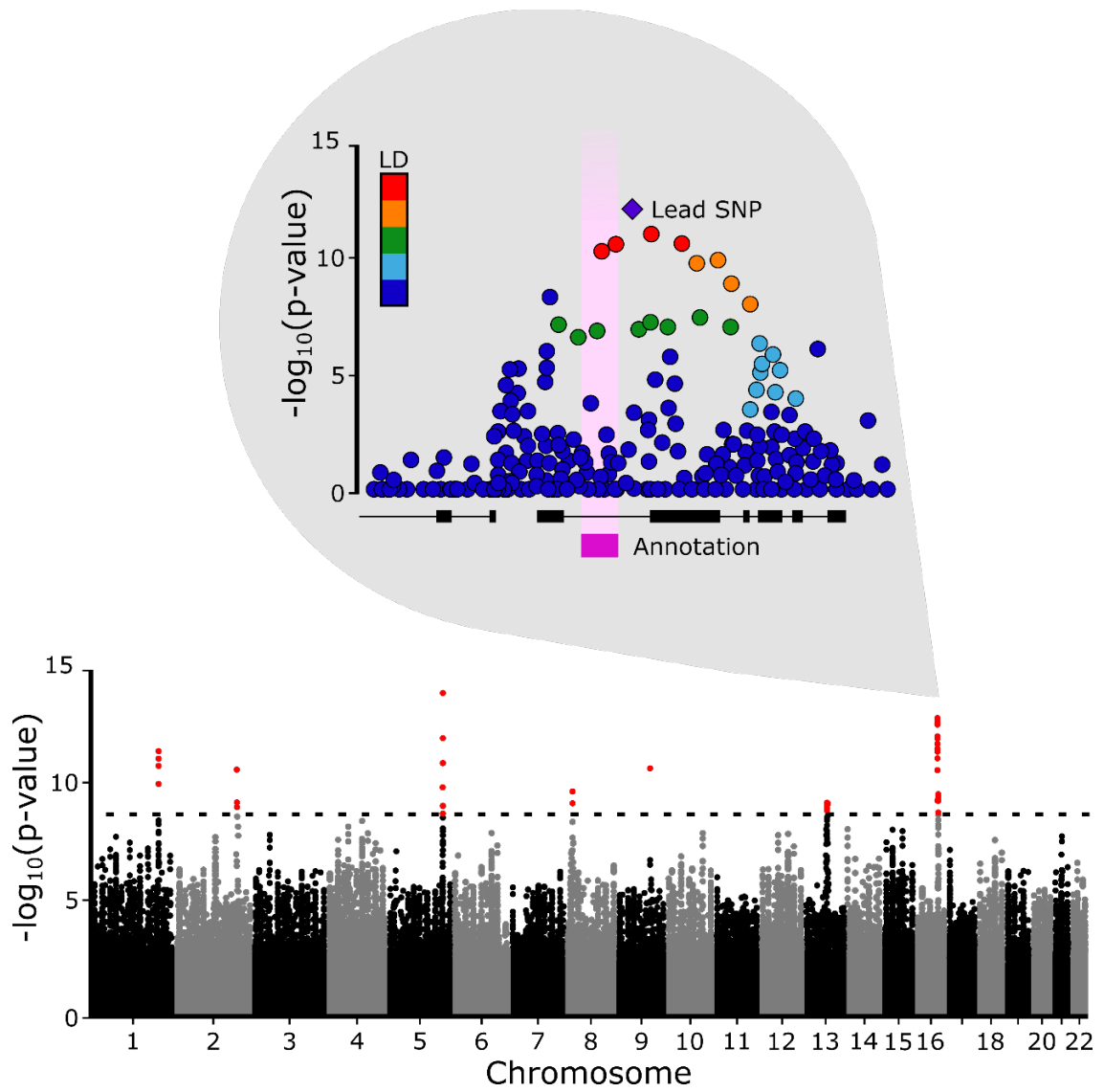


Figure 1.1: The pathological hallmarks of Parkinson disease

There are two pathological hallmarks of Parkinson disease primarily responsible for the motor symptoms: the selective degeneration of midbrain dopaminergic neurons in the substantia nigra and the appearance of protein aggregates of alpha-synuclein, called Lewy bodies.

Figure 1.2: Genomic annotations can be used to prioritize GWAS-implicated variants

After a GWAS is performed, the results can be summarized on a Manhattan plot, where each variant evaluated is plotted against the significance of its association with the trait. Examining each significant locus more closely, the magnitude of association for each SNP in the genomic region and the LD relationship with the lead SNP is examined and statistically likely causative variants are identified. This analysis can be augmented by the inclusion of functional annotations. In this example, the lead SNP (purple diamond) and four others in high LD (red dots) are the most statistically significant variants however, the two variants to the left also overlap a functional element (pink box) and are therefore more likely to be causative and are prioritized for functional follow-up.



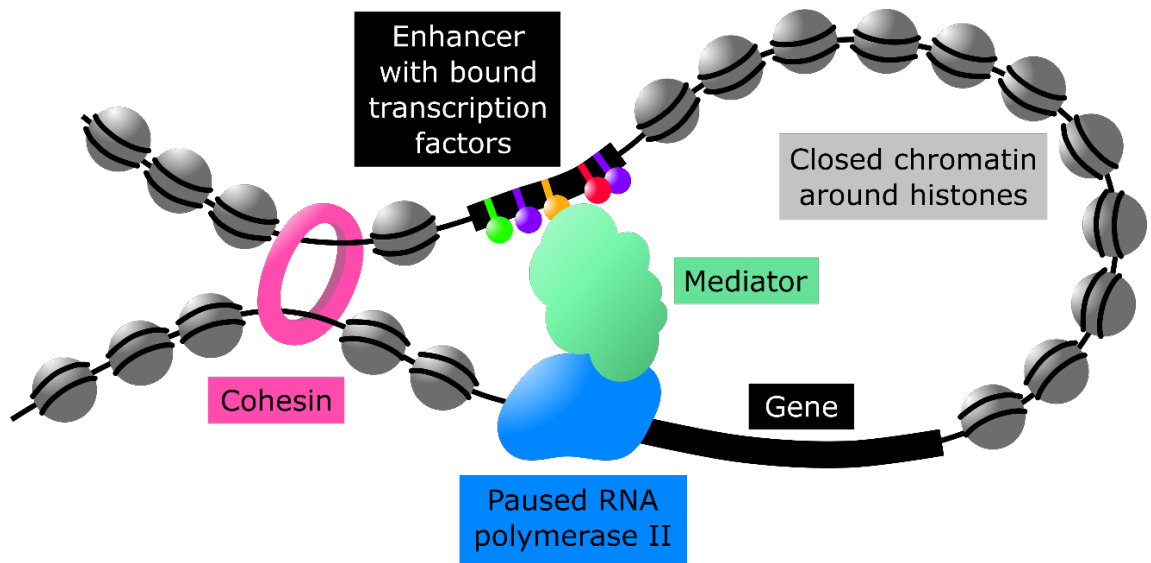


Figure 1.3: Enhancers regulate transcription

Chromatin accessibility, conformation, and transcription factor binding are factors influencing enhancer activity. In its inactive state, chromatin is tightly compacted around nucleosomes (grey circles), and as such is inaccessible for protein binding. For enhancers to be active, nucleosomes must be disassembled or shuffled out of the region such that the DNA is accessible. Through the actions of cohesin (pink hoop) and mediator (green cloud), an enhancer is brought into proximity with its target promoter through chromatin looping. To form this bridge between the enhancer and promoter, the mediator complex interacts with both the paused RNA polymerase II and transcription factors (green, purple, orange, and red circles) that are bound to short sequence motifs at the enhancer (stripes in the black box). Once in proximity, current models suggest that the enhancer promotes gene expression by signaling RNA polymerase II (blue), already bound at the gene's promoter, to exit its paused state and begin transcriptional elongation.

Chapter 2: Chromatin-based analyses of dopaminergic neurons¹

2.1 Applying chromatin data to investigate non-coding variation in Parkinson disease

Parkinson disease (PD) is a common progressive neurodegenerative disorder characterized by preferential and extensive degeneration of dopaminergic (DA) neurons in the substantia nigra^{4,161}. This loss of midbrain (MB) DA neurons disrupts the nigrostriatal pathway and results in the movement phenotypes observed in PD. While this disorder affects approximately 1% of people over 70 years old worldwide¹⁶², the mechanisms underlying genetic risk of sporadic PD in the population remains largely unknown. Familial cases of PD with known pathogenic mutations are better understood but account for $\leq 10\%$ of PD cases¹⁶³.

The α -synuclein gene (*SNCA*) is commonly disrupted in familial PD through missense mutations predicted to promote misfolding^{7,15,16} or genomic multiplications, resulting in an over-expression paradigm²⁰. The *SNCA* locus has also been shown by genome-wide association studies (GWAS) to harbour common variation modulating risk of sporadic PD²⁵. In the same way, common variation at over 40 additional loci have been implicated in PD²³, but the genes modulated and causal variants responsible for elevating risk remain largely undetermined.

¹ This work has been published in the American Journal of Human Genetics and adapted for use in this thesis. McClymont, S.A. et al. (2018). Parkinson-associated *SNCA* enhancer variants revealed by open chromatin in mouse dopamine neurons. *The American Journal of Human Genetics*, 103:874-892²³⁷.

That most GWAS-implicated variants are non-coding²⁶ is a major source of this uncertainty, obstructing the identification of: 1) the causative variant at a locus; 2) the context in which the variation is acting and; 3) the mechanism by which a variant asserts its effect on disease risk.

GWAS are inherently biologically agnostic and their exploitation of linkage disequilibrium (LD) structure frequently results in many variants being implicated at any one locus, with no one variant prioritized over those in LD. One method to prioritize non-coding variants is to examine the chromatin status at that locus^{26,164,165}. Accessible chromatin is more likely to be functional and variants therein may impact that activity, more so than those variants residing in inaccessible chromatin. Recent studies have prioritized neuropsychiatric variants through examination of the chromatin status of iPSC-derived neurons or post-mortem whole brain tissues^{166,167}. However, chromatin accessibility is dynamic, often varying across cell types and developmental time, therefore understanding and isolating the *in vivo* cellular context in which variation acts is critical to increase our ability to prioritize variants and query their methods of action^{26,168–170}.

Exploiting the preferential vulnerability of MB DA neurons in PD, we have prioritized DA neurons as the biological context in which a fraction of PD-associated variation likely acts. DA neurons in other brain regions, such as the forebrain (FB), provide a related substrate that is less vulnerable to loss in PD. We sought to use chromatin data from *ex vivo* populations of DA neurons to investigate the contributions of non-coding variation to PD risk. To maximize the specificity of the biological context, we generated chromatin signatures of purified mouse MB and FB DA neurons. We examined the resulting regulatory regions for their ability to direct

in vivo reporter expression and developed a regulatory sequence vocabulary specific to DA neurons. In doing so, we identified a novel MB DA regulatory element that falls within intron 4 of *SNCA* and demonstrate its ability to direct reporter expression in catecholaminergic neurons of transgenic mice and zebrafish. Furthermore, this enhancer harbours two common variants falling in a haplotype that we determine to be associated with PD risk. We demonstrate these enhancer variants impact protein binding and we propose a model for how the variants and the haplotype at large contribute to *SNCA* regulatory control. This work illustrates the power of cell context-dependent guided searches for the identification of disease associated and functional non-coding variation.

2.2 ATAC-seq identifies open chromatin in midbrain and forebrain dopaminergic neurons

To identify open chromatin regions (OCRs) in DA neurons, we performed ATAC-seq⁹⁵ on ~50,000 fluorescent-activated cell sorting (FACS)-isolated cells (per replicate) from microdissected regions of the MB and FB of embryonic day 15.5 (E15.5) Tg(*Th*-EGFP)DJ76Gsat BAC transgenic mice¹⁷¹ (**Figure 2.1A**). This mouse line expresses EGFP under the control of the tyrosine hydroxylase (*Th*) locus, labeling catecholaminergic neurons (i.e.: DA, noradrenergic, and adrenergic neurons). To confirm capture of the corresponding catecholaminergic neurons, we performed RT-qPCR on the isolated reporter-labeled cells, establishing them to be enriched for DA neuronal markers relative to unlabeled populations from the same dissected tissues (**Supplementary Figure 2.1**).

To evaluate the ATAC-seq libraries, we examined *in silico* quality control measures (**Supplementary Figure 2.2**), evaluated the called peaks and read pile-ups with the Integrative Genomics Viewer (IGV)^{172,173}, and quantified the correlation between brain regions and within replicates. A representative browser trace at the *Th* locus in both MB and FB libraries is presented in **Figure 2.1B**. Replicates are well correlated: MB library replicates have an average correlation of 0.72 (**Figure 2.1C**), and FB replicates are more correlated at $r = 0.86$ (**Figure 2.1D**). Given the robust correlation between replicates, we pooled all reads from the same brain region and called peaks on this unified set to increase our power to detect regions of open chromatin. As a result, we identified 104,217 regions of open chromatin in the MB DA neurons and 87,862 regions in the FB. MB and FB libraries are moderately well correlated (average $r = 0.64$; **Supplementary Figure 2.3**), with approximately 60% of MB OCRs also represented in the FB libraries.

To assess these catalogues for characteristics of functionality, we examined the sequence constraint underlying the called regions of open chromatin, excluding peaks that overlap promoters. Promoters are typically accessible⁹³ and thus, we aimed to reduce the inflation of sequence conservation due to highly conserved promoter-overlapping ATAC-seq peaks. Despite removal of these highly conserved peaks, we observed a high degree of sequence constraint underlying open chromatin peaks compared to background (**Figure 2.1E**). The fact that elements in these libraries of putative cis-regulatory elements (CREs) are constrained, highlights their likely functional significance.

To further examine the OCR catalogues for biological relevance, we explored the gene ontology (GO) terms of nearby genes. While CREs are not restricted to acting

solely on the nearest gene, this restriction is often used as a proxy in the absence of other information. To bolster our predictions, we have also generated bulk RNA-seq data on these same populations of sorted cells (**Supplementary Figure 2.4**) and used these data to examine the GO terms of the nearest expressed gene (RPKM ≥ 1). While still imperfect, implementing this as a proxy for function results in GO terms enriched for neuronal functions in both MB and FB OCR catalogues (**Figure 2.1F, G**). Thus, we establish these OCR catalogues are enriched for putative CREs likely directing the expression of genes with key roles in neuronal biology.

2.3 Candidate regulatory regions are capable of directing expression *in vivo*

Although our OCR catalogues appear to be enriched for functional elements on the basis of sequence conservation and GO, both of these metrics are indirect surrogates for true measures of function. To more directly measure the biological relevance of the catalogues and to identify enhancers, we assessed the capability of the candidate CREs to direct expression *in vivo*.

We took advantage of the large repository of elements that have already been tested in *lacZ* reporter assays *in vivo* and catalogued in the VISTA enhancer browser¹⁷⁴ (accessed September 4, 2016). Overlap between our catalogues and all 2,387 elements in the VISTA enhancer browser, which were scored for their ability to direct *lacZ* reporter expression in E11.5 mice, was quantified (**Supplementary Table 2.1**). Of the 1,264 elements in VISTA identified as enhancers, 786 were present in the MB catalogue and 719 were present in the FB catalogue (**Figure 2.2A**). Examining the overlap of the FB and MB catalogues with enhancers demonstrated to direct

expression in either non-neuronal or neuronal tissues, we observed that 42-47% of enhancers reported to direct expression in non-neuronal tissues are present in the catalogues. By contrast, 71-76% of enhancers that direct expression in one or more regions of the brain overlapped the FB and MB catalogues (**Figure 2.2B**) confirming an abundance of brain enhancers in our catalogues. Stratifying these confirmed neuronal enhancers on the basis of their expression patterns in VISTA, we observed an abundance of MB-specific enhancers in our MB catalogue, and an abundance of FB-specific enhancers in our FB catalogue, with 77% of MB- and FB-specific enhancers in VISTA captured in our MB and FB catalogues, respectively (**Figure 2.2C**). Collectively, these data establish that our region-specific OCR catalogues capture region-specific, active CREs with high efficiency.

To extend our assessment of the biological activity of sequences within these OCR catalogues, we focused on an additional five candidate CREs not already tested in the VISTA browser and evaluated their ability to act as enhancers in *lacZ* reporter mice and in transgenic zebrafish TdTomato reporter assays. All five regions were represented by robust peaks in both the MB and FB catalogues (**Supplementary Figure 2.5**). Two regions, one in the first intron of *Kcnq3* and the other downstream of *Foxg1*, were additionally prioritized using H3K27Ac ChIP-seq from a variety of tissues from E11.5 and E15.5 embryonic mice, seeking to limit our selection to candidate enhancers predicted to have neuronal-specific activity. The remaining three candidate CREs were selected on their proximity to genes important in DA neuron biology. We selected sequences at *Foxa2* and *Nr4a2*, as both are key transcription factors (TFs) in the development and maintenance of DA neurons¹⁷⁵⁻¹⁷⁸. The final region, located in an intron of *Crhr1*, was selected as this locus has been implicated in

PD by GWAS²⁵ and our group has recently prioritized this gene as a candidate for PD risk¹⁷⁹. All selected sequences were lifted over to hg19, facilitating the identification and assay of their corresponding human sequence intervals.

When tested in transgenic reporter mice at E11.5 (**Supplementary Figure 2.5**), two of the five regions (those near *KCNQ3* and *FOXA2*) were validated as enhancers (**Figure 2.2D, H**). Recognizing that a disparity exists between the developmental time at which we generated the catalogues (E15.5) and when the mice were assayed (E11.5), which may compromise validation rates, we also assayed each sequence across multiple time points in zebrafish. All assayed regions except that at *KCNQ3* directed reporter expression in mosaic transgenic zebrafish (**Figure 2.2E, F, G, H**). All five regions displayed enhancer activity *in vivo* in neuronal tissues in one or both transgenic assays. Our transgenic animal experiments corroborate the results of the retrospective VISTA enhancer browser intersection; implying that our OCR catalogues are biologically active and enriched for sequences capable of driving neural expression *in vivo*.

2.4 A midbrain-specific enhancer directs expression in catecholaminergic neuron populations

Having established the biological robustness of the OCR catalogue, we moved to exploit these data to investigate how non-coding variation therein may be contributing to PD risk. We established two complementary strategies. First we sought globally to examine the overlap of PD GWAS SNPs^{23,25} and those in LD ($r^2 > 0.8$) with our DA OCR catalogues. In doing so, we identify 129 unique PD-associated variants at 20 GWAS associated loci that are present in one or both of our OCR

catalogues (34 specifically overlap the MB catalogue, 14 specifically overlap the FB catalogue, and 81 overlap both).

Second, we examined the chromatin landscape surrounding familial PD genes, focusing on those with no obvious overlaps in the first strategy. In this, we turned our attention to the *SNCA* locus. Despite this locus being the most significant hit in PD GWASs^{23,25}, the LD structure surrounding the lead SNP (rs356182) is such that no variants in LD are apparent at our r^2 cut-off and the lead SNP itself is not overlapped by either our MB or FB catalogue. Given α -synuclein's established role in PD pathogenesis and the strength of GWAS signal at *SNCA*, we prioritized this locus for a closer, more targeted, inspection.

We first noted that *Snca* expression differs significantly between the MB and FB DA neurons in our bulk RNA-seq (**Figure 2.3B**). Examining the chromatin accessibility at the *Snca* locus, the MB and FB are largely the same with the exception of one robust peak in intron 4 (mm9: chr6:60,742,503-60,744,726) that is present in the MB and completely absent in the FB (**Figure 2.3A**). DNase hypersensitivity site (DHS) linkage^{93,180} suggests that this putative CRE interacts with the *SNCA* promoter. Given the MB-specificity of this putative CRE and indications that it interacts with the *SNCA* promoter, we anticipated that this region may be a driving force behind the MB-specific expression of *Snca*.

To test this hypothesis, we assayed whether the central portion of this putative CRE, when lifted over to hg19 (chr4:90,721,063-90,722,122), is capable of directing appropriate reporter expression in transgenic zebrafish and mouse reporter assays. Stable transgenesis of zebrafish indicates that this CRE directs reporter expression

at 72 hours post fertilization in the locus coeruleus, a key population of catecholaminergic neurons preferentially degenerated in PD¹⁸¹, and along the catecholaminergic tract through the hindbrain, which is largely composed of DA neurons¹⁸² (**Figure 2.3C**). Additionally, we observe reporter expression throughout the diencephalic catecholaminergic cluster with projections to the subpallium, which is analogous to mammalian dopaminergic projections from the ventral midbrain to the striatum¹⁸³. Reporter expression in these transgenic zebrafish is largely consistent with an enhancer active in catecholaminergic populations.

To further evaluate this CRE in a mammalian system, we generated *lacZ* reporter mice and examined reporter activity across developmental time. Whole mount E12.5 reporter mice indicate this enhancer directs exquisitely restricted expression in Th+ populations, including the dorsal root ganglia, extending into the sympathetic chain, and throughout the cranial nerves (particularly the trigeminal). Additional diffuse staining is noted throughout the MB and FB (**Figure 2.3D**). Specifically examining the brains of *lacZ* animals at E15.5, reporter expression is identified in the MB and hypothalamus, with strong expression through the amygdala/piriform cortex and along the anterior portion of the sympathetic chain (**Figure 2.3E**); similar reporter patterns are seen at P7 (**Figure 2.3F**). At P30, we detect reporter activity in the amygdala, hypothalamus, thalamus, periaqueductal grey area, brain stem, and importantly, in the *substantia nigra* and ventral tegmental area (**Figure 2.3G**). By contrast, in aged *lacZ* reporter mice (574 days old, ~19 months), we only detect strong reporter expression in the brain stem and observe weak reporter expression in the amygdala (**Figure 2.3H**). Collectively, the regions in which we detect reporter activity reflect those compromised in PD; Lewy bodies

(aggregates of α -synuclein) have been detected in the locus coeruleus, sympathetic chain, amygdala, hypothalamus, ventral tegmental area, periaqueductal grey area of PD patients^{184–188}, and critically the preferential degradation of the *substantia nigra* is the pathological hallmark of PD progression⁴. This enhancer directs region-specific appropriate expression throughout development in key locations concordant with SNCA activity in PD pathogenesis.

2.5 Enhancer variants are significantly associated with Parkinson disease risk

Following confirmation of this CRE's regulatory activity in brain regions associated with PD, we next inspected this sequence for PD-associated variation. We sequenced across this interval in 986 PD patients and 992 controls and identified 14 variants (***Supplementary Table 2.2***), 4 of which were common and present in both cases and controls with a minor allele frequency greater than 5%. Of these, two tightly linked variants ($r^2 = 0.934$; ***Supplementary Table 2.3***), rs2737024 (OR = 1.25, 95% CI = 1.09-1.44, p-value = 0.002) and rs2583959 (OR = 1.22, 95% CI = 1.06-1.40, p-value = 0.005), were significantly associated with PD (***Table 2.1***). These data support a role for variation within the enhancer in conferring PD risk.

Finally, we set out to refine the haplotype structure and understand how this identified variation may be interacting with other variants at this locus. A panel of common variants had previously been genotyped across *SNCA* and PD-associated haplotypes were identified¹⁸⁹. After genotyping our patients and controls for a subset of this panel of variants in addition to all enhancer-associated variants identified by sequencing (***Supplementary Table 2.4***), we identified a single haplotype that was

significantly associated with PD (p-value = 0.003), with a higher observed frequency in PD patients (28.3%) compared to controls (23.4%; **Table 2.2**). This haplotype implicates some of the same variants as in Guella *et al.*¹⁸⁹ (rs356220, rs737029) but also implicates rs356225 and rs356168, and the two enhancer-associated variants. Additionally, within the 1000 Genomes data, we observe that moderate LD structure exists between the lead GWAS variant (rs356182) and the enhancer variants ($r^2 = 0.418$, $D' = 0.745$) in the general European population. Despite the moderate LD, the risk allele of rs356182 falls in the PD-associated haplotype that we identify 94% of the time. Thus, it is likely that at least part of the risk captured by rs356182 can be attributed to these enhancer variants and the implicated haplotype reported here. Further, this does not preclude additional variants from being present and contributing to the risk captured by the lead SNP, as the rs356182 risk allele can occur in the absence of the enhancer associated variants (i.e.: ~31% of EUR individuals with the rs356182 risk allele do not carry the risk alleles of the PD-associated enhancer variants). A schematic of the variants, open chromatin regions, chromatin interactions^{93,180}, and LD structure at this locus is presented in **Figure 2.4**. Of the variants whose minor alleles define this PD-associated haplotype, including rs356182, only the two enhancer associated variants and rs2737029 are identified as eQTLs for *SNCA* expression in any tissue in the GTEx database. Collectively, these data identify a catecholaminergic enhancer harbouring common variation that is part of a larger haplotype associated with PD risk, likely by modulating *SNCA* activity.

2.6 Deleting the disease-associated enhancer

To confirm that this enhancer governs expression of *SNCA*, we deleted the enhancer in the SK-N-SH human neuroblastoma cell line, which is often used as an

in vitro model of DA neurons¹⁹⁰, with the goal of quantifying expression changes of nearby genes following the enhancer deletion. We designed four guide RNA pairs on either side of the enhancer, along with a repair template for co-transfection to allow for screening and selection of enhancer-deletion clones. The repair template contained a fluorescent marker, mCherry, for screening, and a selectable blasticidin-resistance cassette, all bound by *loxP* sites and arms of homology, to promote homology-directed repair following double-strand break by CRISPR-Cas9 (**Figure 2.5A**).

Unfortunately, upon transfection of the necessary plasmids for the enhancer deletion, the majority of transfected cells died within 24 hours, while control cells did not (mock transfection, transfection of the repair template alone). We considered the possibility that the strong SV40 promoter directing expression of the blasticidin-resistance gene might be interfering with transcription of *SNCA* at this locus and over-expressing toxic *SNCA* products; as such, we also generated an empty repair template, containing just the *loxP* sites and arms of homology for co-transfection with our guide RNA combinations (**Figure 2.5A**). Under these conditions, we continue to observe massive cell death within 24 hours of transfection of the editing components. These experiments indicate either a technical issue with editing SK-N-SH cells, which we have tried to control for, or that this locus is required for cell viability in cell culture.

To address this possibility, we have turned to an *in vivo* model and have generated enhancer deletion mouse lines, where we designed a pair of guide RNAs against the endogenous mouse *Snca* enhancer locus (**Figure 2.5B**). These edited mice appear to be viable and healthy. With this model, we will extract MB DA neurons and assess the impact of the enhancer deletion on *Snca* expression. Additionally, with this

in vivo model in hand, we can assay more complex phenotypes and examine the mice for PD risk and susceptibility. We will be treating the mice with lipopolysaccharide, an inflammatory agent known to induce Parkinson-like phenotypes in mice^{191,192}, and assessing the mice for a variety of movement phenotypes, using the rotarod or pole-descent tests, and evaluating by histology the degree of degeneration of MB DA neurons as compared to non-enhancer deleted mice (**Figure 2.5C**). These experiments are designed to illuminate the role of this disease-associated enhancer in both regulating cognate gene expression and in modulating the risk of Parkinson disease.

2.7 Discussion

The identification and prioritization of biologically pertinent non-coding variation associated with disease remains challenging. Recent studies by our and other groups have emphasized the importance of cellular context in the identification of sequences harbouring biologically pertinent variation and the genes they regulate. To this end, we used chromatin signatures from *ex vivo* isolated DA neurons to reveal biologically active sequences that harbour non-coding variation contributing to PD risk. We generated robust OCR catalogues for both MB and FB DA neurons, confirmed their capacity to act as enhancers, and notably, identified two variants located within a MB-specific enhancer that are associated with an increase in PD risk.

In contrast to strategies predicated solely on dissection of post-mortem tissues or on the differentiation of cultured cells, we leveraged the use of transgenic reporter mice to specifically isolate *Th*-expressing neurons from discrete neuroanatomical (FB and MB) domains. While our approach assays a more refined population of DA neurons than would be achieved via gross dissection, recent single-cell RNA-seq

analyses of these same cells make clear that even within these highly restricted MB and FB populations there exist two primary cellular phenotypes¹⁷⁹. The “homogenous” MB and FB populations each are comprised of an immature neuroblast population and a more mature, domain specific, post-mitotic population of DA neurons. As such, our OCR catalogues capture the chromatin accessibility from both of these states. These catalogues are demonstrably biologically relevant for our purposes, but future studies requiring even greater homogeneity may wish to consider single-cell ATAC-seq to refine these domains further¹⁹³.

In our *in silico* validation of the catalogues, we established them to be enriched for both sequence constraint and biological relevance in a manner consistent with function and their FB/MB origin. Furthermore, these sequences are frequently domain appropriate enhancers, with each catalogue capturing a large fraction (77%) of previously validated MB and FB enhancers. Although an abundance of regions are shown to direct neuronal expression compared to those annotated as negative or non-neuronal, it is interesting to note that almost half of the sequences previously documented not to direct expression *in vivo* are also represented in one or both of our catalogues.

Given the frequently dynamic nature of CRE activity, this overlap with negative regions likely results from temporal differences in these assays. Our data indicates these regions are accessible at E15.5 but the *lacZ* reporter assays were carried out at E11.5; regions that have been annotated as negative at E11.5 may be active at later time points and, as such, appear in our catalogues. As we moved from these unbiased functional comparisons to more highly selected ones, the potential impact of temporal differences became more pronounced. In mouse transgenic

reporter assays, two of five assayed putative CREs direct detectable expression of *lacZ* in neuronal populations. Consistent with the temporally dynamic nature of CREs, when these same regions are tested in zebrafish across multiple developmental time points, we observe four of the five sequences to act as neuronal enhancers.

Taken collectively, these data establish a robust biological platform in which PD-associated variation can be evaluated. To this end, an obvious candidate to interrogate was an apparent MB-specific open chromatin domain within intron 4 of the known PD-associated gene, *SNCA*. We assayed the activity of this putative CRE in zebrafish and across the life course of mice and found it to be active in key catecholaminergic structures injured in PD (e.g.: the *substantia nigra* and locus coeruleus), from mid-gestation until at least P30. Thereafter, the utilization of this enhancer in the brain is diminished and by late life appears restricted to the brainstem and amygdala. By the time of clinical presentation, PD patients have already lost a significant proportion ($\geq 30\%$) of their nigral DA neurons^{4,194}; the observed biology of this CRE is consistent with a progressive pathogenic influence acting early in life, rendering these populations preferentially vulnerable to loss over an extensive period of time.

Sequencing this interval in PD cases and controls revealed two common variants (rs2737024 and rs2583959) therein, individually associated with an increased risk of PD. Furthermore, we identify a larger haplotype containing these variants, also significantly associated with PD risk. While none of the other SNPs in this haplotype overlap with CREs identified in the DA neuron catalogues, variant rs356168 has significant functional evidence of its activity and contribution to PD risk¹⁰⁹. The same DHS correlation analysis^{93,180} that suggests an interaction between

the *SNCA* promoter and our identified CRE, also suggests an interaction between the *SNCA* promoter and the rs356168 variant. Additionally, ChIA-PET data^{180,195} indicates that sequence encompassing this variant may interact with our enhancer, suggesting a potential co-operative mode of action; a paradigm recently proposed by Gupta and colleagues⁶⁹ at the *EDNI* locus. We propose that the variants within the enhancer, independently or in concert with other variation within the identified haplotype, may act throughout the lifespan to render key populations of catecholaminergic neurons vulnerable, thus increasing PD risk in individuals harbouring this variation.

This work emphasizes the value of biologically informed, cell context-dependent guided searches for the identification of disease associated and functional non-coding variation. Given the extent of non-coding GWAS-identified variation, the need for strategies to prioritize variants for functional follow-up is greater than ever. Here, we generate chromatin accessibility data from purified populations of DA neurons to generate catalogues of putative CREs. We have demonstrated how these data can be used to reveal non-coding variation contributing to PD risk; focusing on a single region of open chromatin at the *SNCA* locus, we uncover PD-associated variation therein and propose a model through which this sequence can contribute to normal DA neuronal biology and PD risk. There remains a plethora of information still to be explored in these catalogues, either through further single locus investigations or through massively parallel assays. For example, our MB DA neuron OCR catalogue overlaps variants at 20 of 49 (41%) PD-associated loci^{23,25}, all of which can be investigated further for their mechanisms by which they impact PD risk. Our work establishes a powerful paradigm, leveraging transgenic model systems to

systematically generate cell type specific chromatin accessibility data and reveal disease-associated variation, in a manner that can be progressively guided by improved biological understanding.

2.8 Methods

Animal husbandry

Tg(Th-EGFP)DJ76Gsat mice (Th-EGFP) were generated by the GENSAT Project¹⁹⁶ and purchased through the Mutant Mouse Resource and Research Centers Repository. Colony maintenance matings were between hemizygous male Th-EGFP mice and female Swiss Webster (SW) mice, obtained from Charles River Laboratories. This same mating scheme was used to establish timed matings, generating litters for assay; day on which vaginal plug is observed, E0.5. Adult AB zebrafish lines were maintained in system water according to standard methods¹⁹⁷. All work involving mice and zebrafish (husbandry, colony maintenance, procedures, and euthanasia) were reviewed and pre-approved by the institutional care and use committee.

Neural dissociation and FACS

Pregnant SW mice were euthanized at E15.5 and the embryos were removed and immediately placed in chilled Eagle's Minimum Essential Media (EMEM) on ice. Embryos were decapitated and brains were removed into Hank's Balanced Salt Solution without Mg²⁺ and Ca²⁺ (HBSS w/o) on ice. Under a fluorescent microscope, EGFP+ brains were identified and microdissected to yield the desired forebrain (FB) and midbrain (MB) regions desired. Microdissected regions were placed in fresh HBSS w/o on ice, and pooled per litter for dissociation.

Pooled brain regions were dissociated using the Papain Dissociation System (Worthington Biochemical Corporation). The tissue was dissociated in the papain solution for 30 minutes at 37°C, with gentle trituration every 10 minutes using a sterile Pasteur pipette. Following dissociation, cells were passed through a 40µm cell

strainer into a 50mL conical, centrifuged for 5 minutes at 300g, resuspended in albumin-inhibitor solution containing DNase, applied to a discontinuous density gradient, and centrifuged for 6 minutes at 70g. The resulting cell pellet was resuspended in HBSS with Mg^{2+} and Ca^{2+} and submitted to fluorescent-activated cell sorting (FACS). Aliquots of 50,000 EGFP+ cells were sorted directly into 300 μ L HBSS with Mg^{2+} and Ca^{2+} with 10% FBS for ATAC-seq. Aliquots containing $\geq 50,000$ EGFP+ cells were sorted into kit-provided lysis buffer for RNA-seq. This procedure was repeated such that a single aliquot of cells from each region per litter were submitted to either ATAC-seq or bulk RNA-seq three times over for each region.

ATAC-seq library preparation and quantification

ATAC-seq library preparation generally follows the steps as set out in the original ATAC-seq paper⁹⁵ with minor modifications. Aliquots of 50,000 EGFP+ cells were centrifuged for 5 minutes at 4°C and 500g, washed with 50 μ L of chilled PBS and centrifuged again for 5 minutes at 4°C and 500g. The cell pellet was resuspended in lysis buffer, as set out in the protocol, and cells were left to lyse for 5 minutes at 4°C before being centrifuged for 10 minutes at 4°C at 500g. The resulting nuclei pellet was tagmented, as written, using the transposase from the Nextera DNA Library Preparation Kit. Following transposition, DNA was purified with the MinElute Reaction Clean-up Kit (Qiagen) and eluted in 10 μ L elution buffer.

Libraries were amplified according to the original ATAC-seq protocol⁹⁵. The qPCR surveillance steps were modified such that the additional number of cycles of amplification were calculated as $\frac{1}{4}$ maximum intensity, so as to limit PCR duplication rates in the final libraries. Amplified libraries were purified with Ampure XP beads

(Beckman Coulter) following the Nextera DNA Library Prep Protocol Guide. Libraries were quantified using the Qubit dsDNA High Sensitivity Assay (Invitrogen) in combination with the Agilent 2100 Bioanalyzer using the High Sensitivity DNA Assay (Agilent).

ATAC-seq sequencing, alignment, and peak calling

Individual ATAC-seq libraries were sequenced on the Illumina MiSeq to a minimum depth of 15 million, 2x75bp reads per library.

Quality of sequencing was evaluated using FastQC (v0.11.2). Reads were aligned to mm9 using Bowtie2¹⁹⁸ (v2.2.5), under --local mode. Reads aligning to the mitochondrial genome, unknown and random chromosomes, and PCR duplicates were removed prior to peak calling (SAMtools¹⁹⁹). Peaks were called on individual libraries and on a concatenated file combining all MB or all FB libraries (“Joint”) using MACS2²⁰⁰ (v2.1.1.20160309) “callpeak” with options: --nomodel --nolambda -B -f BAMPE --gsize mm --keep-dup all. Peaks overlapping blacklisted regions called by ENCODE and in the original ATAC-seq paper were removed^{95,195}. Peaks were examined for their genomic distribution using CEAS in the Cistrome pipeline^{201,202}. The fragment lengths were extracted from the SAM files and plotted using a custom script. Mouse (mm9) transcriptional start site (TSS) co-ordinates were extracted from the UCSC Genome Browser²⁰³ and deepTools²⁰⁴ was used to quantify the pileup of reads over TSSs.

RNA-seq library preparation and quantification

Total RNA was extracted using the Purelink RNA Micro Kit (Invitrogen). Following FACS isolation into kit-provided lysis buffer, samples were homogenized

and RNA extraction proceeded using manufacturer's recommendations. Total RNA integrity was determined using the Agilent 2100 Bioanalyzer using the RNA Pico Kit (Agilent). RNA samples were sent to the Sidney Kimmel Comprehensive Cancer Center Next Generation Sequencing Core at Johns Hopkins for library preparation, using the Ovation RNA-Seq System V2 (Nugen), and sequencing.

RNA-seq sequencing, alignment, and transcript quantification

Libraries were pooled and sequenced on Illumina's HiSeq 2500 in Rapid Run mode with 2x100bp reads to an average depth of >90 million reads per library. Quality of sequencing was evaluated using FastQC. FASTQ files were aligned to mm9 using HISAT2²⁰⁵ (v2.0.1-beta) with --dta specified.

Aligned reads from individual samples were quantified against a reference transcriptome using the Rsubread package^{206–208} (v1.22.3) function "featureCounts" with the following options: isPairedEnd = TRUE, requireBothEndsMapped = TRUE, isGTFAnnotationFile = TRUE, useMetaFeature = TRUE. The GENCODE vM9 GTF was downloaded²⁰⁹ (date: March 30, 2016) and lifted over from the mm10 to the mm9 genome using CrossMap (v0.2.2) with default parameters²¹⁰. This was used for quantification, in which gene-level raw counts were converted to RPKM values and means for each region were calculated.

RNA-seq and ATAC-seq relationship

The highest 1,000 expressed genes and the lowest 1,000 expressed genes (RPKM \geq 1) in both the MB and FB were identified and their transcriptional start sites (Ensembl) extracted from the UCSC Table Browser²⁰³. Intervals of 1, 10, and 100kb surrounding these TSSs were intersected with the ATAC-seq libraries, and the

overlap quantified²¹¹ and plotted. These same TSSs were provided to deepTools²⁰⁴, and the ATAC-seq signal over these highest and lowest expressed genes was quantified and plotted. Additionally, the 1,000 highest and lowest ATAC-seq peaks (by q-value) were extracted and the expression of the nearest gene was quantified and plotted as a final metric to relate the RNA-seq and ATAC-seq datasets.

cDNA synthesis and RT-qPCR for DA neuron markers

RNA was extracted using the RNeasy Mini Kit (Qiagen), after sorting 50,000 cells directly into Buffer RLT. Aliquots of 50,000 non-fluorescing cells were also collected and processed in parallel. 100ng of each RNA sample was submitted to first strand cDNA synthesis using the SuperScript III First-Strand Synthesis System for RT-PCR (Invitrogen), following the Oligo(dT) method.

Primers (*Supplementary Table 2.5*) were designed using Primer-BLAST²¹² under default parameters with the requirement for exon-exon junction spanning specified. qPCR was performed using Power SYBR Green Master Mix (Applied Biosystems). Reactions were run in triplicate, following default SYBR Green Standard cycle specifications on the Viia7 Real-Time PCR System (Applied Biosystems). Relative quantification followed the $2^{-\Delta\Delta CT}$ method, normalizing results to *Actb* in the EGFP- aliquot of cells for each region, respectively.

Correlation analysis between regions and within replicates

Peaks from all six ATAC-seq libraries and the two “Joint” ATAC-seq libraries were concatenated together, sorted on the basis of chromosomal location, merged into a unified peak set²¹¹, and converted to Simplified Annotation Format. Reads from each BAM file overlapping this unified peak set were quantified with the Rsubread package

“featureCounts” command, with the following options: isPairedEnd = TRUE, requireBothEndsMapped = FALSE. Read counts were normalized for each library using conditional quantile normalization²¹³, accounting for library size, peak length, and peak GC content. Pearson correlation co-efficients were calculated from this normalized count matrix and visualized using corrplot²¹⁴ and RColorBrewer²¹⁵ and LSD²¹⁶.

Sequence constraint analysis

Average phastCons²¹⁷ were calculated for the “Joint” peak file for both the MB and FB libraries using Cistrome²⁰². Beforehand, peaks with overlap of exons or promoters (defined here as +/-2,000bp from the transcriptional start site) were removed. The exon and promoter BED files were downloaded from the UCSC table browser²⁰³ (Mouse genome; mm9 assembly; Genes and Gene Predictions; RefSeq Genes track using the table refGene).

Gene ontology of nearest expressed gene

The Genomic Regions Enrichment of Annotations Tool²¹⁸ (GREAT; v3.0.0) predicted the GO term enrichment in the catalogues. Beforehand, peaks were processed to: a) remove peaks overlapping commonly open regions; b) select the top 20,000 peaks and; c) overlap the nearest expressed gene’s transcriptional start site (TSS).

First, regions that are commonly open were defined as those regions of the genome that are open in >30% of ENCODE DNase hypersensitivity site (DHS) assays in mouse tissues. These ubiquitously open regions were removed from the peak files. Next, to limit the number of regions submitted to GREAT such that the binomial

distribution for calculating fold enrichment values was still valid, peak files were limited to the top 20,000 peaks on the basis of q-value.

Finally, in order to limit ourselves to the nearest expressed gene, we supplied a list of the TSSs of the nearest expressed gene that are in the GREAT database. Only genes that are in this list with RPKM > 1 were considered as expressed. The nearest expressed gene to each of the top 20,000 peaks was identified. Each peak is associated with its nearest expressed gene and to ensure that GREAT only considered these nearest genes for analysis, we submitted these nearest expressed gene's TSSs as a proxy for each peak. These proxy peaks were submitted to GREAT using the NCBI build 37 (mm9) assembly, under whole genome background regions, with the single nearest gene as the association rule, including curated regulatory domains.

Quantification of overlap between CRE catalogues and the VISTA Enhancer Browser

All elements tested *in vivo* were downloaded from the VISTA Enhancer Browser on September 4, 2016. These regions were stratified into those annotated as positive or negative. BED co-ordinates of these regions were extracted and intersected with the ATAC-seq catalogues. Positive regions were further stratified into those with annotations for only forebrain, only midbrain, only hindbrain, combinations of regions ("Multiple regions"), all three regions ("Whole brain"), summing to the "Neuronal" category, or were annotated as positive but driving expression in none of those three regions ("Non-neuronal").

Testing five putative CREs for *in vivo* reporter activity

Prioritized regions were PCR amplified (*Supplementary Table 2.5*) from human gDNA and cloned into either pENTR for mouse *lacZ* assays (Invitrogen) or pDONR221 for zebrafish assays (Invitrogen). Regions were sequence validated and LR cloned (Invitrogen) into either an *hsp68-lacZ* vector or pXIG vector, with a TdTomato cassette in place of GFP.

Generation of transgenic mice and E11.5 embryo staining was performed as previously described^{78,101,219} using FVB strain mice. Embryos expressing the *lacZ* reporter gene were scored and annotated for their expression patterns by multiple curators. For a construct to be considered positive, a minimum of three embryos per construct were required to demonstrate reporter activity in the same tissue. Mouse transient transgenic assays were approved by the Lawrence Berkeley National Laboratory Animal Welfare and Research Committee.

Generation of transgenic zebrafish was performed as previously described¹⁰³ in AB zebrafish. At 3 days post fertilization (dpf) and 5dpf, reporter expression patterns were evaluated. For a construct to be considered as positive, $\geq 25\%$ of mosaic embryos had to display reporter activity in one or more anatomical structures. Positive zebrafish were quantified for reporter activity in five anatomical regions (forebrain, midbrain, hindbrain, amacrine cells, spinal cord).

Intersection of CRE catalogues and PD-associated GWAS variants

Lead SNPs from the two most recent meta-analyses^{23,25} were submitted to rAggr and SNPs in LD were identified (1000 Genomes, Phase 3, EUR populations; minimum MAF = 0.05, $r^2 \geq 0.8$; maximum distance 5,000kb). These variants were

intersected²¹¹ with the CRE catalogues, after they were lifted over to hg19 coordinates, and the overlap extracted and quantified.

***In vivo* validation of the MB-specific enhancer**

The MB-specific peak was PCR amplified (***Supplementary Table 2.5***) from human genomic DNA and TA cloned into pCR8 (Invitrogen). Regions were sequence validated and LR cloned (Invitrogen) into either an *hsp68-lacZ* vector or a modified pXIG vector, with a TdTomato cassette in place of GFP.

For zebrafish transgenesis, the modified pXIG vector was injected into 1-2 cell stage embryos as previously described¹⁰³ in AB zebrafish. TdTomato reporter expression was assayed at 72 hours post fertilization (hpf) and 5dpf; mosaic embryos positive for TdTomato expression were selected and raised to adulthood and founders were identified. Progeny of founders were screened at 72hpf for reporter activity.

For mouse transgenesis, the generated *hsp68-lacZ* vector was purified in a double CsCl gradient (Lofstrand Labs Ltd) and stable mouse transgenesis was performed in C57BL/6 mice by Cyagen Biosciences Inc. Multiple founder lines were generated. For *lacZ* staining, embryos were collected at E12.5, and mouse brains were isolated at E15.5, P7, P30, and P574. Brains were roughly sectioned in 1mm sections at P7 and P30 and animals were perfused at P574 and fixed brains were sectioned (200 μ m) with a vibratome. Specimens were subsequently fixed for 2 hours on ice in 1% formaldehyde, 0.2% glutaraldehyde, 0.02% Igepal CA-630 in PBS. Following fixation, tissues were permeabilized over 3x15 minute washes in 2mM MgCl₂ and 0.02% Igepal CA-630 in PBS at room temperature. Embryos/tissues were incubated overnight at 37°C in staining solution, containing 320 μ g/mL X-Gal in N,N-dimethyl

formamide, 12mM K-ferricyanide, 12mM K-ferrocyanide, 0.002% Igepal CA-630, 4mM MgCl₂ in PBS. Specimens were washed in 0.2% Igepal CA-630 in PBS over 2x30 minutes and finally stored in 4% formaldehyde, 100mM sodium phosphate, and 10% methanol.

Sequencing and genotyping PD patients and controls at SNCA

A total of 986 individuals with PD and 992 controls who were seen at the Mayo Clinic in Jacksonville, FL were sequenced across the putative enhancer and genotyped for 25 variants across the *SNCA* locus. The variants chosen for genotyping were confirming those identified by sequencing of the enhancer as well as assessing those identified in Guella *et al*¹⁸⁹. For PD cases, median age at blood draw was 69 years (Range: 28-97 years), median age at PD onset was 67 years (Range: 28-97 years), and 631 cases (64.0%) were male. Median age at blood draw in controls was 67 years (Range: 18-92 years) and 415 subjects (41.8%) were male. Individuals with PD were diagnosed using standard clinical criteria²²⁰. All subjects are unrelated non-Hispanic Caucasians of European descent. The Mayo Clinic Institutional Review Board approved the study and all subjects provided written informed consent.

Genomic DNA was extracted from whole blood using the Autogen FlexStar. Sanger sequencing of the enhancer region was performed bidirectionally using the ABI 3730xl DNA analyzer (Applied Biosystems) standard protocol. Sequence data was analyzed using SeqScape (v2.5; Applied Biosystems). Statistical analyses were performed using both SAS and R²²¹. Of the variants identified within the enhancer, only those with minor allele frequency greater than 5% were evaluated for association with PD in single-variant analysis. Associations between individual variants and PD

were evaluated using logistic regression models, adjusted for age at blood draw and sex, and where variants were considered, under an additive model (i.e. effect of each additional minor allele). Odds ratios and 95% confidence intervals were estimated and a Bonferroni correction for multiple testing, due to the four common variants that were evaluated for association with PD, was utilized in single-variant analysis, after which p-values ≤ 0.0125 were considered as statistically significant.

Genotyping the 25 SNPs across the *SNCA* locus was performed using the iPLEX Gold protocol on the MassARRAY System and analysed with TYPER 4.0 software (Agena Bioscience). For the 25 SNPs genotyped across the *SNCA* locus, all genotype call rates were $>95\%$ and there was no evidence for departure from Hardy-Weinberg equilibrium (all χ^2 p-values > 0.05 after Bonferroni correction). Haplotype frequencies in cases and controls was estimated using the haplo.stats package²²² function “haplo.group”. Associations between haplotypes and risk of PD were evaluated using score tests of association²²³ using the “haplo.score” function. Tests were adjusted for age at blood draw and sex, haplotypes occurring in less than 1% of subjects were excluded, and only individuals with no missing genotype calls for any variants were included. A Bonferroni correction for multiple testing was applied, after which p-values ≤ 0.0042 were considered as statistically significant, due to the 12 different common haplotypes that were observed and tested for association with PD risk.

LD structure/ r^2 values at the *SNCA* locus in the 1000 Genomes EUR population were extracted from LDlink²²⁴ using the LDmatrix tool and plotted using R. The chromatin structure at *SNCA* was extracted from the 3D Genome Browser¹⁸⁰, querying POLR2A binding in MCF-7 cells at the *SNCA* promoter.

Generating enhancer knock-out cell lines and mice

SNCA deletion cell line

SK-N-SH neuroblastoma cells were grown in high glucose Dulbecco's Modified Eagle Medium (DMEM) supplemented with 1X penicillin and streptomycin, and 10% fetal bovine serum. Cells were maintained at 37°C, 5% CO₂.

Guide RNA sequences flanking the lifted over enhancer (hg19) were designed with CHOPCHOP²²⁵⁻²²⁷ and double stranded oligonucleotides with BbsI/BsaI compatible overhangs were synthesized by Integrated DNA Technologies (1F – GAAGGGACTCCTTGCTTGA; 3F – GTTGAAATCAAAGTAGTAGT; 1R – CTGGGAGCACAATTGGCCC; 2R – GAGCTGTGATAACCACTAA; 3R – TGGATTAGAACCACTGCTA; 4R - ATAACCACTAATGTTCCCT). Guides were cloned in pairs (1F-2R, 1F-4R, 3F-1R, 3F-3R) into a modified PX458 vector²²⁸ (pSpCas9(BB)-2A-GFP (PX458), a gift from Feng Zhang; Addgene plasmid #: 48138), in which a second guide RNA scaffold sequence with BsaI restriction sites was inserted into PX458 by restriction digest with KpnI and XbaI.

A repair template in a pUC57 backbone was ordered from GENEWIZ, containing arms of homology matching hg19: chr4:90,720,863-90,721,062 and chr4:90,722,123-90,722,322, two *loxP* sites and a custom multiple cloning site. This represents the “empty” repair template. This was modified by standard Gibson assembly²²⁹ and restriction cloning to further contain a blasticidin resistance cassette (pCMV/Bsd; Invitrogen #V51020), a T2A site (CMV-Cas9-2A-RFP; Sigma Aldrich #CAS9RFPP), and an mCherry fluorescence cassette (pmCherry; Takara #632522).

SK-N-SH neuroblastoma cells were plated at 300,000 cells/well in a 6-well plate. The cells were lipofected with the guide RNA plasmid along with a linearized repair template (SapI) the following day using the standard Lipofectamine 2000 protocol (Invitrogen). Each well was transfected with 3.0µg of DNA – 1.5µg of the guide/Cas9 plasmid and 1.5µg of either repair template. Matching repair template only control and no DNA control wells were also included.

SNCA deletion mice

Enhancer deletion mice were generated as previously described by Watkins-Chow et al.²³⁰, with minor modifications. RNA oligonucleotide guides corresponding to flanking region of interest were synthesized (Integrated DNA Technologies) (3F – TAATTTCTACTCTTG TAGATTGTTATTTAAAAGACATGTTTCT and 4R - TAATTTCTACTCTTG TAGATCAGTGCCTATAAAGGGACTACTC). Guides and CPF1 protein (Integrated DNA Technologies) were diluted in Opti-MEM media (Thermo-Fisher Scientific) to a final concentration of 2µM (each guide) and 5ng/µl CPF1 protein. A total of 50µl of guide-CPF1 solution was electroporated into roughly 150 C57BL/6J × FVB/N F1 hybrid zygotes, using a Nepa21 electroporator (Nepa Gene Co., Ltd., Japan) using manufacture's recommended pulse conditions. Hybrid zygotes were subsequently allowed to rest for 30-60min at 5%CO₂, and washed in M2 media (Sigma Aldrich) prior to being implanted using standard embryo transfer surgery protocols. Resulting founder mice were screened for deletion alleles using the flanking PCR primers TTGCAGTGCTGACAATAGGC and CTGGAGCCTGAGAGAAGTGT.

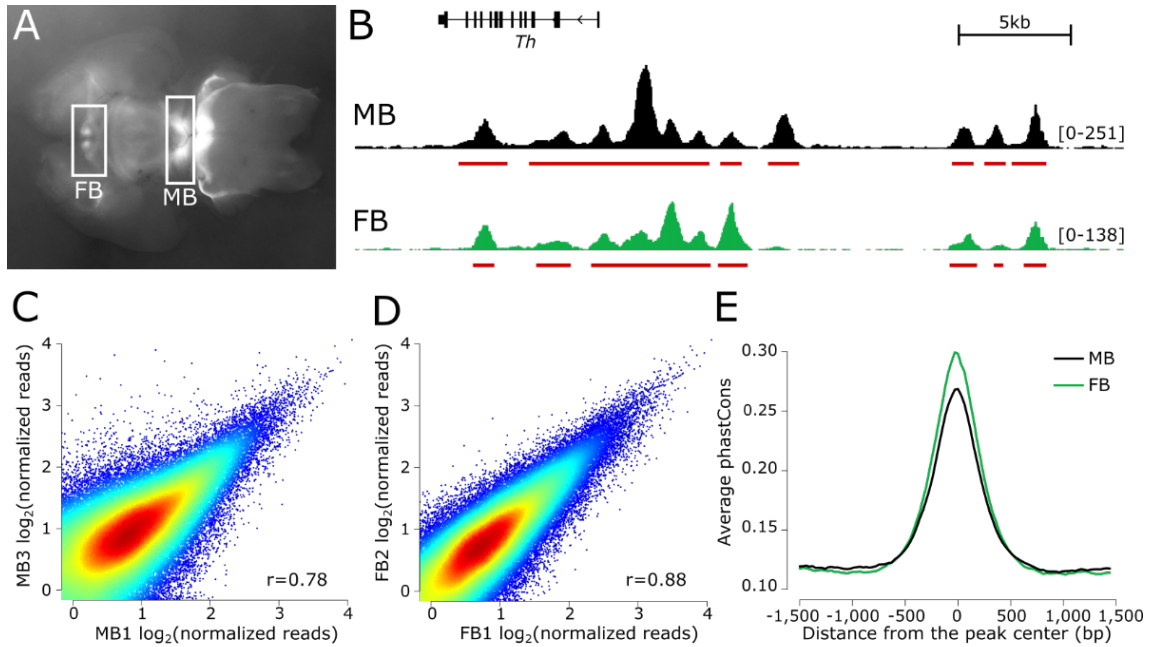
Data sharing and accessibility

ATAC-sequencing, RNA-sequencing and related data is available at the Gene Expression Omnibus (GEO) under the accession number GSE122450.

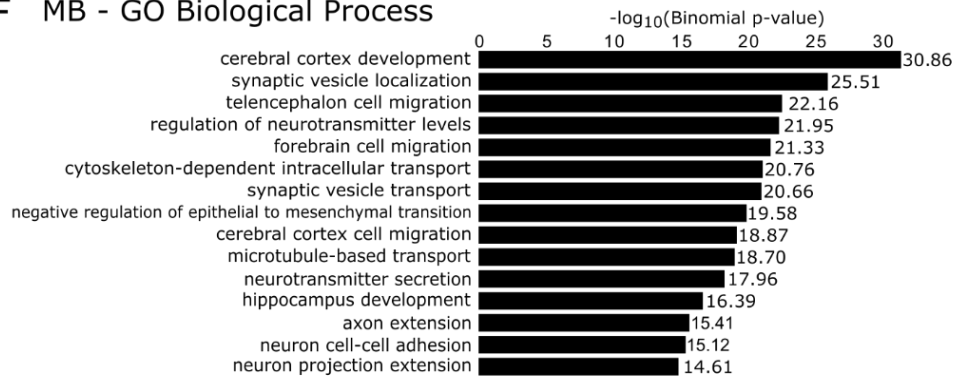
2.9 Figures and supplementary materials

Figure 2.1: Preliminary validation of ATAC-seq catalogues generated from *ex vivo* DA neurons

(A) The midbrain (MB) and forebrain (FB) of E15.5 brains from Tg(Th-EGFP)DJ76Gsat mice are microdissected, dissociated, and isolated by FACS. (B) Read pile-up and called peaks for the MB and FB libraries at the *Th* locus. (C, D) Chromatin accessibility, genome-wide, is correlated between replicates. (E) The sequences underlying MB and FB peaks display a high degree of evolutionary sequence constraint as measured by PhastCon scores. (F, G) Gene ontology terms of the nearest expressed genes to all peaks in both the MB and FB reflect the neuronal origin and function of these catalogues.



F MB - GO Biological Process



G FB - GO Biological Process

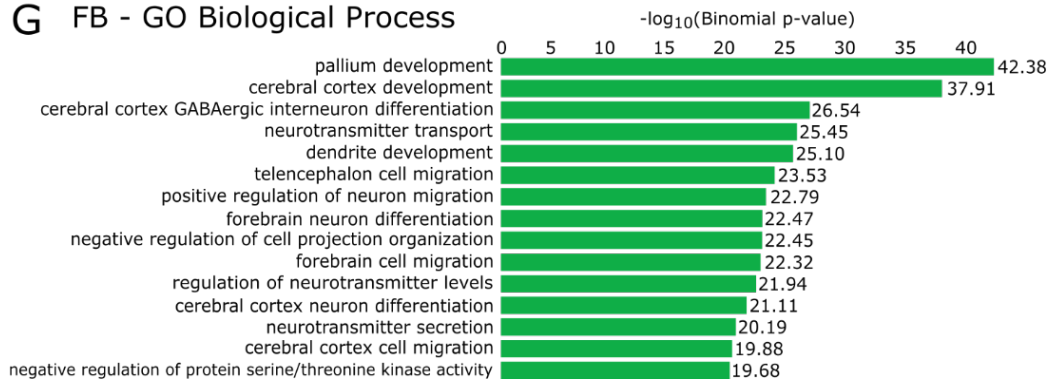


Figure 2.2: Validation of the putative CRE catalogues *in vivo*

(A) Of the elements annotated in VISTA as having enhancer activity, 62% and 56% of these are represented in the MB and FB catalogues, respectively. (B) An abundance of open chromatin regions in the MB and FB catalogues overlap confirmed neuronal enhancers ($\geq 70\%$). (C) Stratifying neuronal enhancers, MB- and FB-specific enhancers are enriched in our MB and FB catalogues, respectively. (D-H) Testing five prioritized putative CREs *in vivo* identifies five neuronal enhancers. (D) A putative CRE in intron 1 of *KCNQ3* directs expression in the midbrain, hindbrain, and neural tube of E11.5 *lacZ* reporter mice. It fails to direct expression in a transgenic zebrafish assay at either 3 or 5 days post fertilization (dpf); reporter expression present in $\leq 25\%$ of mosaics. (E, F, G) Putative CREs downstream of *FOXP1*, upstream of *NR4A2*, and in an intron of *CRHR1* fail to direct expression in transgenic mice, however, they direct robust neuronal appropriate expression in transgenic zebrafish reporter assays (scored for expression in MB, FB, amacrine cells (AC), hindbrain (HB), spinal cord (SC)). (H) A putative CRE downstream of *FOXA2* directs neuronal expression in both transgenic mice and zebrafish assays. N mosaic zebrafish scored: ≥ 141 for 3dpf, ≥ 119 for 5dpf. All constructs have since been deposited in the VISTA database, under the hs numbers supplied.

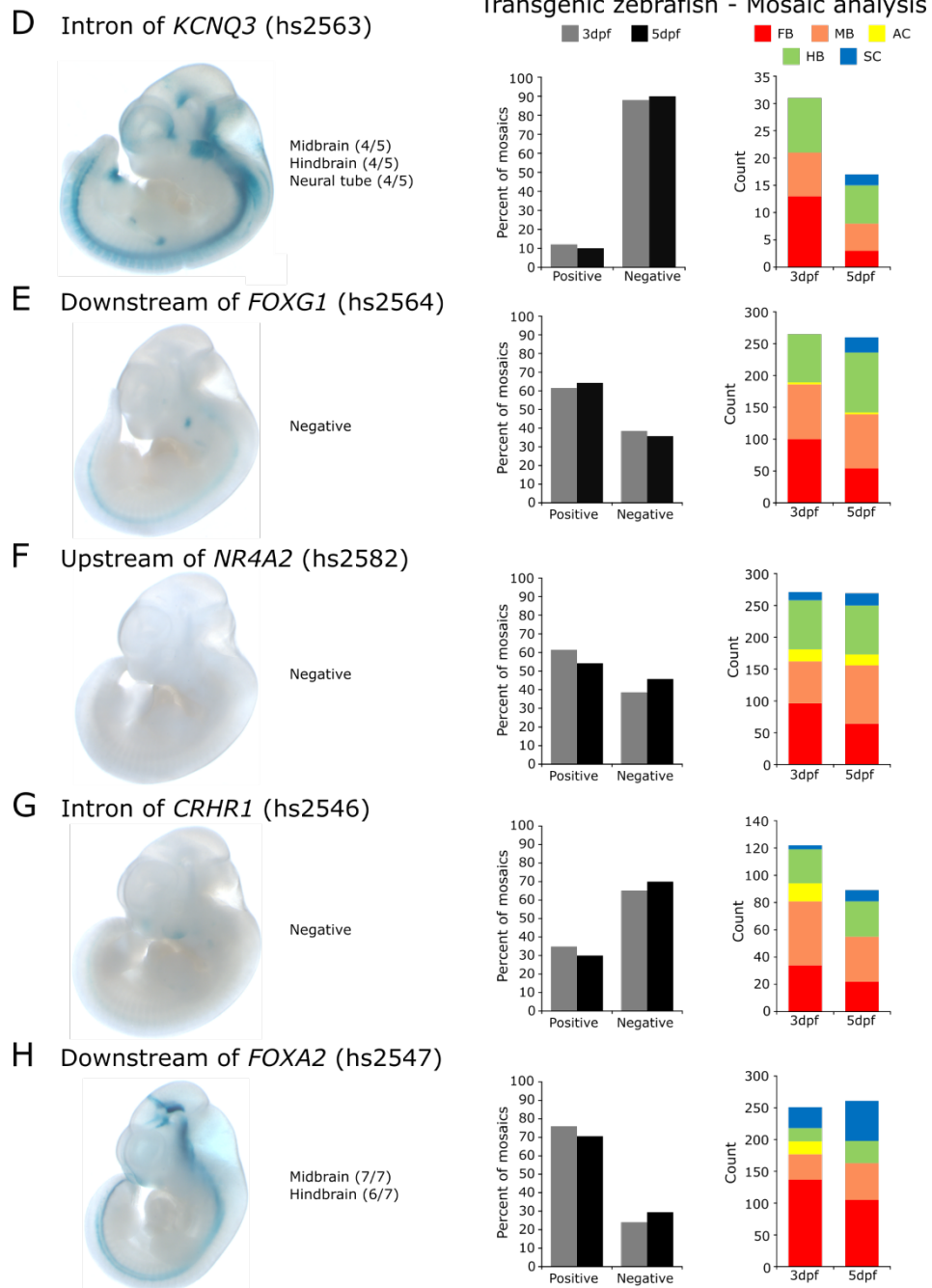
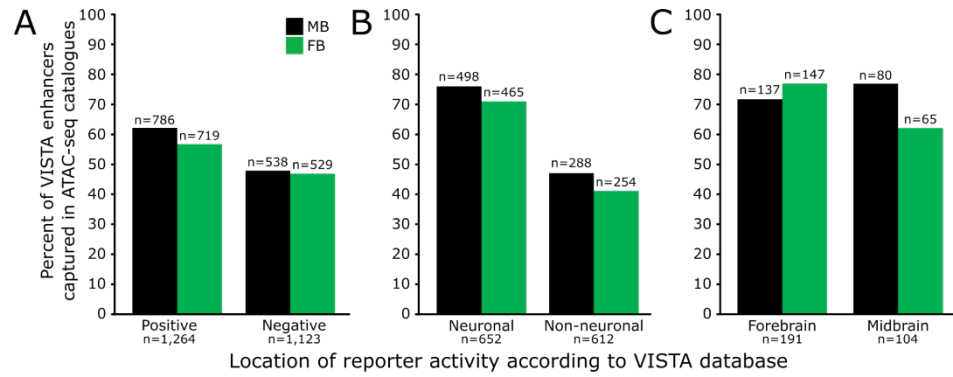


Figure 2.3: A MB-specific enhancer directs expression in catecholaminergic populations of neurons known to Parkinson disease biology

(A) IGV track indicating the location of the MB-specific region of open chromatin, located in intron 4 of *Snca*. (B) *Snca* is differentially expressed between the MB and FB DA neurons. Red bar is the mean expression of the four replicates (black dots). (C) At 72 hours post fertilization (hpf), stable transgenic zebrafish reporter assays indicate this putative CRE is capable of directing reporter expression in key catecholaminergic neuronal populations, including the locus coeruleus (LC), the catecholaminergic tract (CT) of the hindbrain, and the diencephalic cluster (DC) with projections to the subpallium (SP). (D-G) Further studies in *lacZ* reporter assays in embryonic (E) and post-natal (P) mice indicate dynamic enhancer usage across developmental time. (D) This enhancer directs expression throughout the MB, FB, dorsal root ganglia (DRG), sympathetic chain (SC), and cranial nerves (CN) of E12.5 mice. (E) By E15.5, reporter expression is observed in the amygdala and/or piriform cortex (AM/PC), sympathetic chain, MB, and hypothalamus (Hyp). (F) Patterns of reporter expression at P7 reflect those seen at E15.5. (G) Reporter activity is observed at P30 in the amygdala, hypothalamus and thalamus (Thal), brain stem (BS), *substantia nigra* (SN), ventral tegmental area (VTA), and the periaqueductal grey area (PAG). (H) In aged mice (P574), reporter expression is detected robustly in the brain stem and faintly in the amygdala.

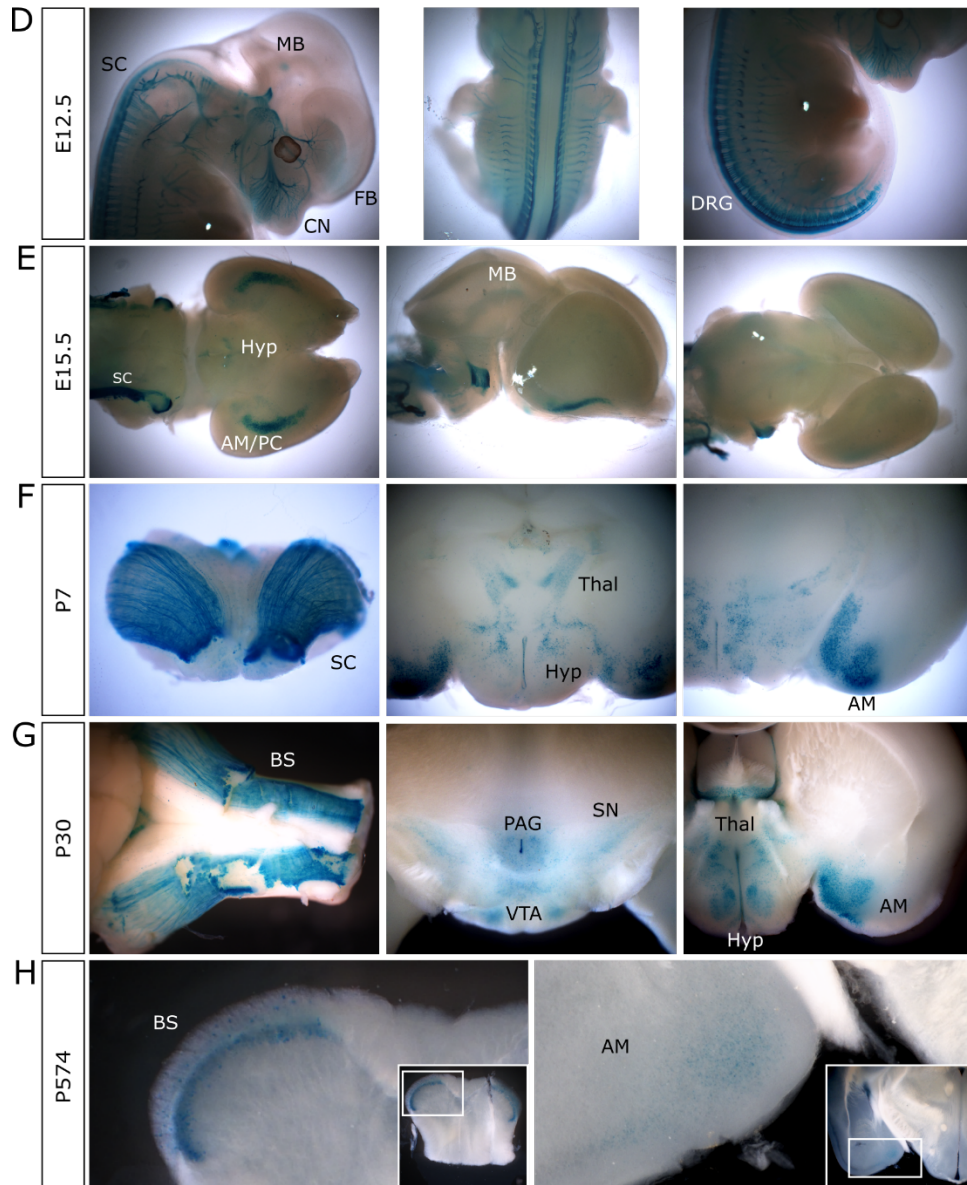
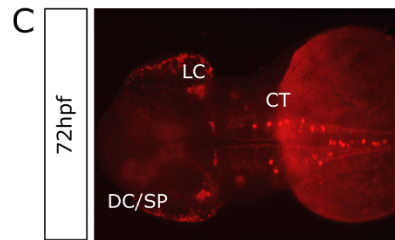
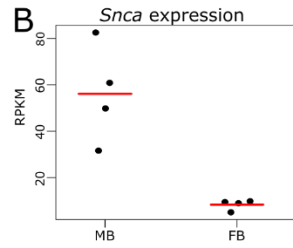
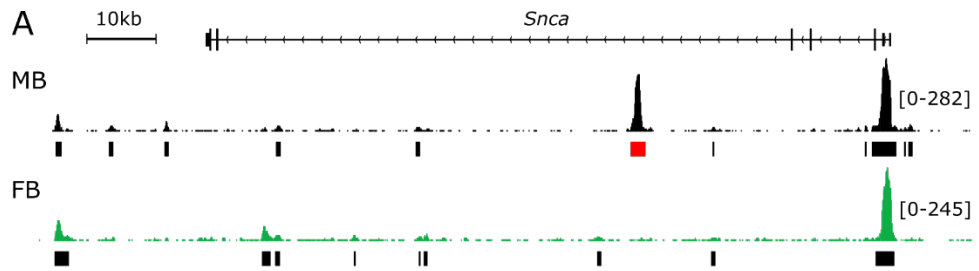


Figure 2.4: A schematic of the chromatin interactions, LD structure, variation, and open chromatin at the *SNCA* locus

Publicly available DNase hypersensitivity site (DHS) linkage analysis suggest that the promoter of *SNCA* possibly interacts with the identified MB-specific enhancer, the lead GWAS variant (rs356182), and a previously functionally validated variant (rs356168). ChIA-PET data suggests the MB-specific enhancer may interact with variant rs356168. Open chromatin data from DA neurons do not overlap with any variants at this locus/haplotype other than at the MB-specific enhancer. LD analysis at this locus indicates that despite the low LD structure between the lead GWAS variant (rs356182) and the enhancer associated variants (rs2737024 and rs2583959), the variants are in the same haplotype. As such the GWAS signal may, at least in part, be flagging the identified enhancer-associated variants.

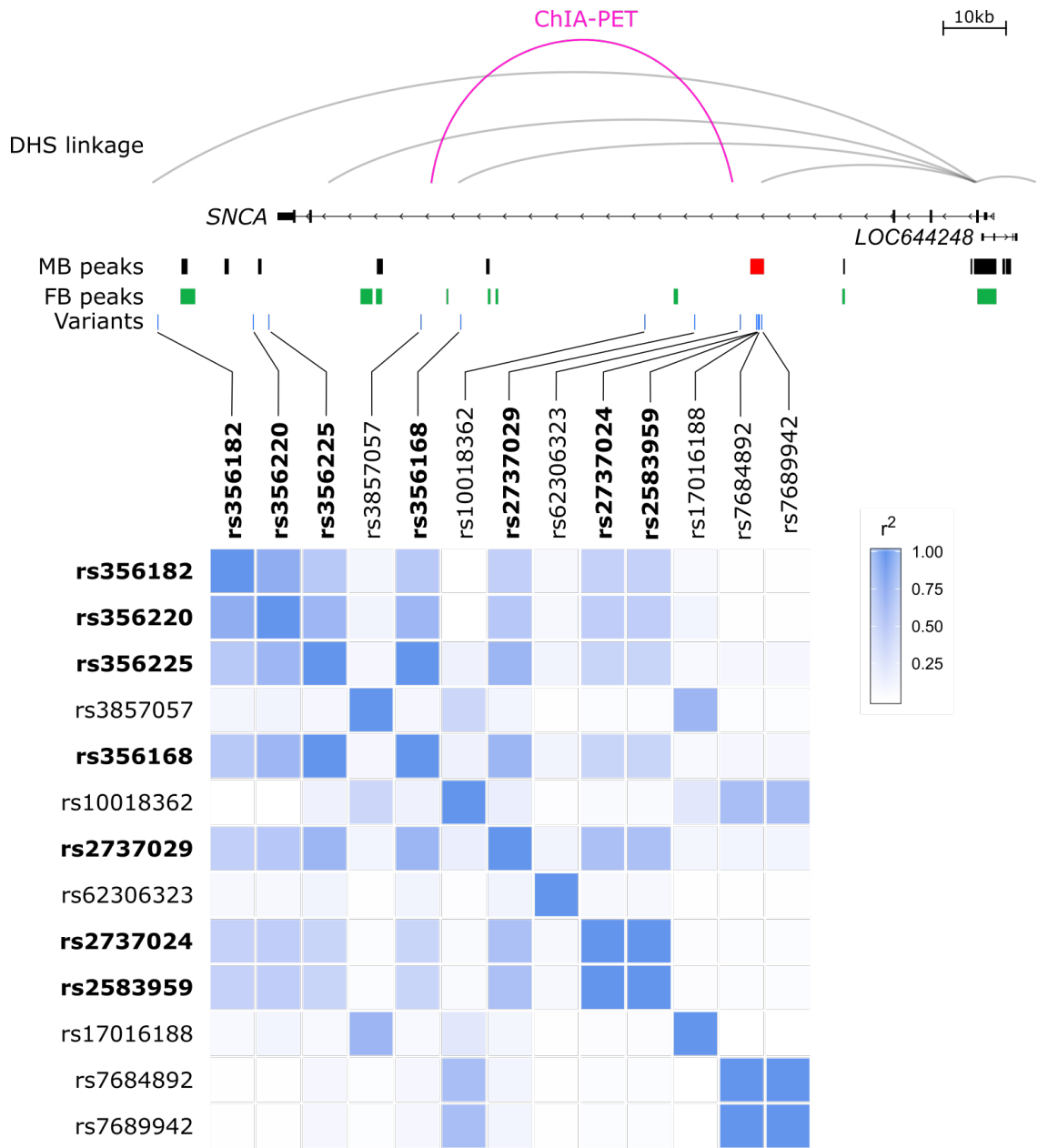


Figure 2.5: Enhancer deletion experiments and proposed phenotyping

(A) In SK-N-SH cells, the enhancer was modified by one of four guide RNA combinations (arrows) and exploiting homology directed repair, was replaced with one of two repair templates. The first repair template contained a blasticidin-resistance cassette (Blast^R) as a selectable marker and a fluorescent marker (mCherry) for screening. These were flanked by *loxP* sites (purple triangles) and arms of homology (dashed crosses). The alternative repair template contained solely the arms of homology and the *loxP* sites. These deletions resulted in cell death. (B) We also deleted the enhancer *in vivo* at the endogenous mouse *Snca* locus. (C) These enhancer deletion mice and matched non-deleted controls will be subjected to an LPS injection paradigm, where movement phenotypes will be assessed at the day of injection and at one month, two months, and 6 months. At six months, the mice will be sacrificed and their brains will be examined by histology for degeneration of midbrain dopaminergic neurons.

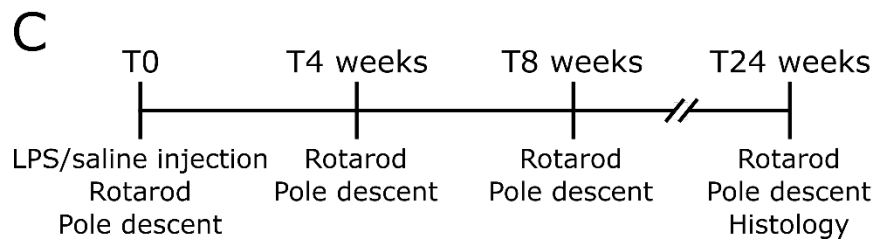
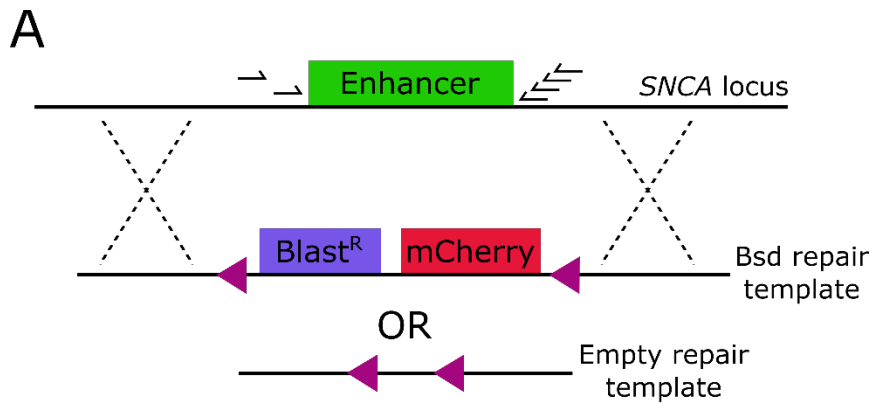


Table 2.1: Two tightly linked SNPs within the enhancer are significantly associated with PD risk

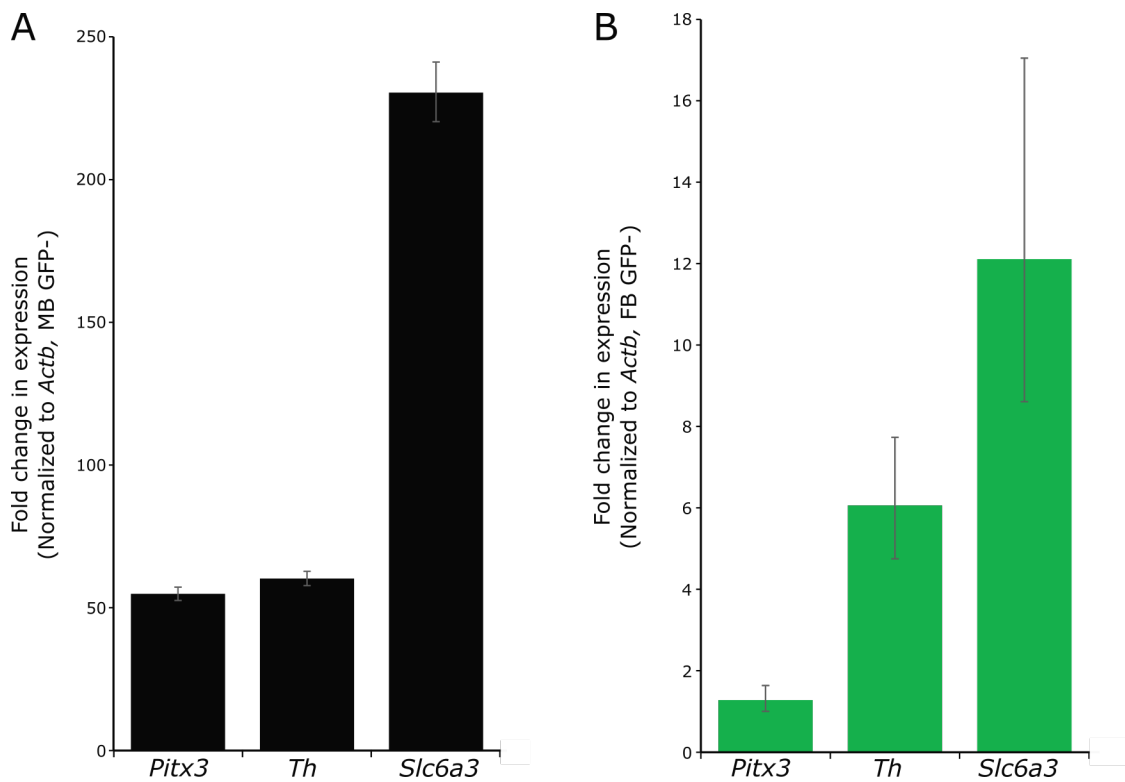
Variant	MA	MAF in PD cases (N=986)	MAF in controls (N=992)	Association with PD	
				OR (95% CI)	p-value
rs7684892	A	0.063	0.069	0.93 (0.72-1.20)	0.562
rs17016188	C	0.082	0.061	1.35 (1.04-1.75)	0.023
rs2583959	G	0.317	0.271	1.22 (1.06-1.40)	0.005 *
rs2737024	G	0.319	0.270	1.25 (1.09-1.44)	0.002 *

MA = minor allele; MAF = minor allele frequency; OR = odds ratio; CI = confidence interval
Only variants with MAF > 0.05 were considered
ORs, 95% CIs, and p-values result from additive logistic regression models adjusted for age at blood draw and sex
p-values ≤ 0.0125 were considered as statistically significant after applying a Bonferroni correction for multiple testing (*)

Table 2.2: A single haplotype, containing the minor alleles of the implicated SNPs, is significantly associated with PD risk

		rs356220	rs356225	rs3857057	rs356168	rs10018362	rs2737029	rs62306323	rs2737024	rs2583959	rs17016188	rs7684892	rs7689942	Frequency in PD cases	Frequency in controls	p-value
		Haplotypes spanning the SNCA	1	C	C	A	G	T	T	C	A	C	T	G	C	0.015
2	C		G	A	A	T	T	T	A	C	T	G	C	0.092	0.110	0.029
3	C		C	A	G	C	C	C	A	C	T	A	T	0.039	0.048	0.184
4	T		C	A	G	T	T	C	A	C	T	G	C	0.037	0.040	0.480
5	C		G	A	A	T	T	C	A	C	T	G	C	0.380	0.397	0.593
6	T		C	A	G	T	T	T	A	C	T	G	C	0.010	0.011	0.698
7	C		G	A	A	T	C	C	G	G	T	G	C	0.009	0.012	0.944
8	C		C	A	G	T	C	C	G	G	T	G	C	0.016	0.015	0.768
9	T		C	G	G	C	C	C	A	C	T	A	T	0.021	0.017	0.360
10	T		C	G	G	T	C	C	A	C	C	G	C	0.014	0.009	0.189
11	T		C	G	G	C	C	C	A	C	C	G	C	0.057	0.044	0.124
12	T		C	A	G	T	C	C	G	G	T	G	C	0.283	0.234	0.003 *

Only haplotypes with frequency ≥ 0.01 were considered
Black boxes indicate the minor allele in Europeans
p-values result from score tests for association, performed under an additive model, adjusted for age at blood draw and sex
p-values ≤ 0.0042 were considered as statistically significant after applying a Bonferroni correction for multiple testing (*)



Supplementary Figure 2.1: RT-qPCR of key DA neuron markers

Expression of key DA neuron markers (*Pitx3*, *Th*, *Slc6a3*) in MB FACS-isolated (**A**) and FB FACS-isolated (**B**) cells confirms isolation of purified MB and FB DA neurons. Error bars represent the fold change range after incorporation of the standard deviation values (n = 3 technical replicates).

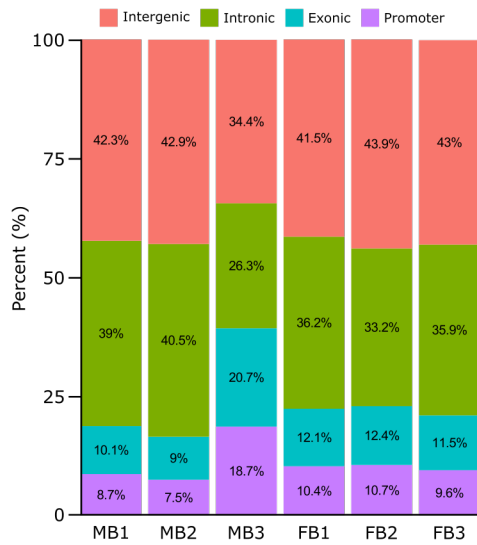
Supplementary Figure 2.2: *in silico* quality control metrics for the ATAC-seq libraries

(A) Sequencing statistics for the ATAC-seq libraries indicate all six libraries are of sufficient quality. (B) The genomic distribution of ATAC-seq peaks indicate a preference for promoters and intergenic regions. (C) The fragment length distribution of the ATAC-seq libraries indicate the presence of a nucleosome ladder (with one nucleosome fragments, perhaps, being selected against in the bead clean-up). (D) All ATAC-seq libraries display an abundance of reads overlapping gene promoters, genome-wide.

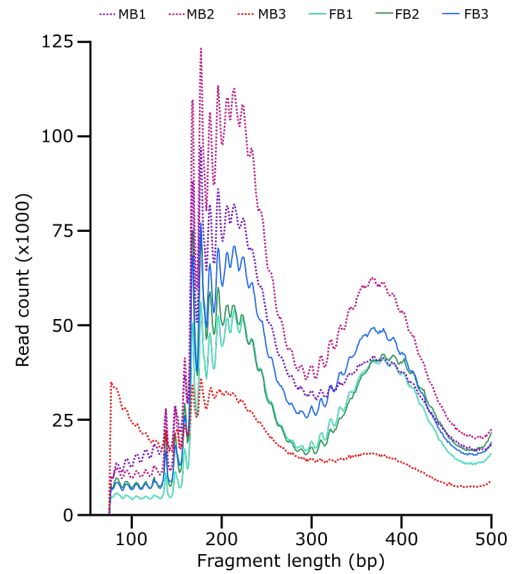
A

	MB1	MB2	MB3	FB1	FB2	FB3
Reads sequenced	27,464,563	28,786,058	42,760,942	16,849,958	50,580,772	21,497,236
% aligned	97.39%	97.28%	90.89%	97.97%	96.06%	96.68%
% duplicate	25.04%	9.79%	39.34%	5.46%	16.84%	8.46%
% mitochondrial	7.48%	5.80%	1.79%	3.32%	4.93%	3.51%
Peaks called	62,646	74,018	22,858	48,525	45,222	52,493
% blacklisted	1.48%	1.16%	4.77%	1.84%	1.91%	1.66%
Fraction of reads in peaks	30.65%	37.62%	10.49%	25.04%	22.16%	25.06%

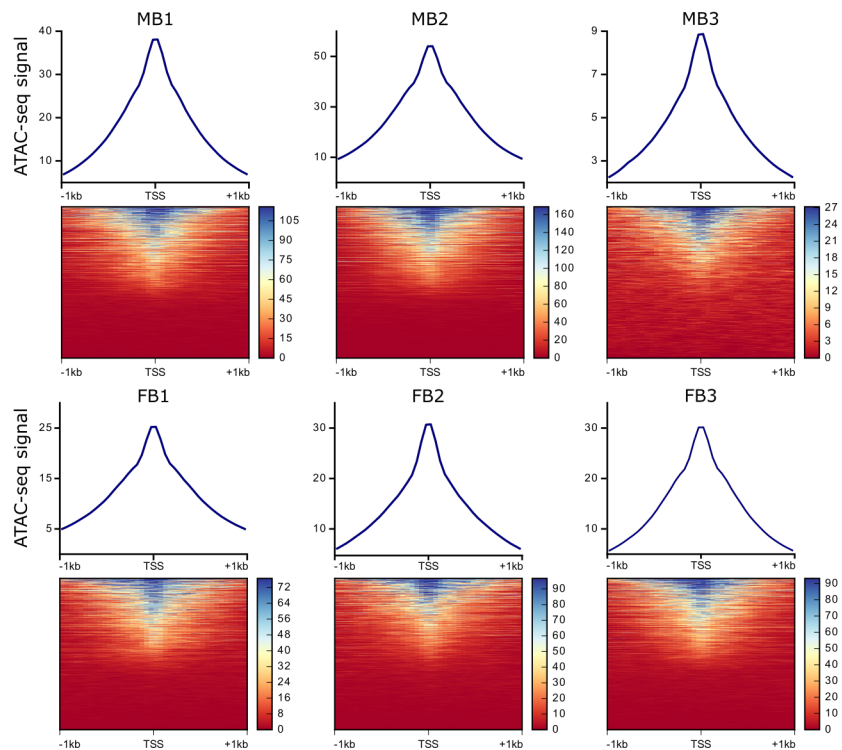
B Genomic distribution of ATAC-seq peaks

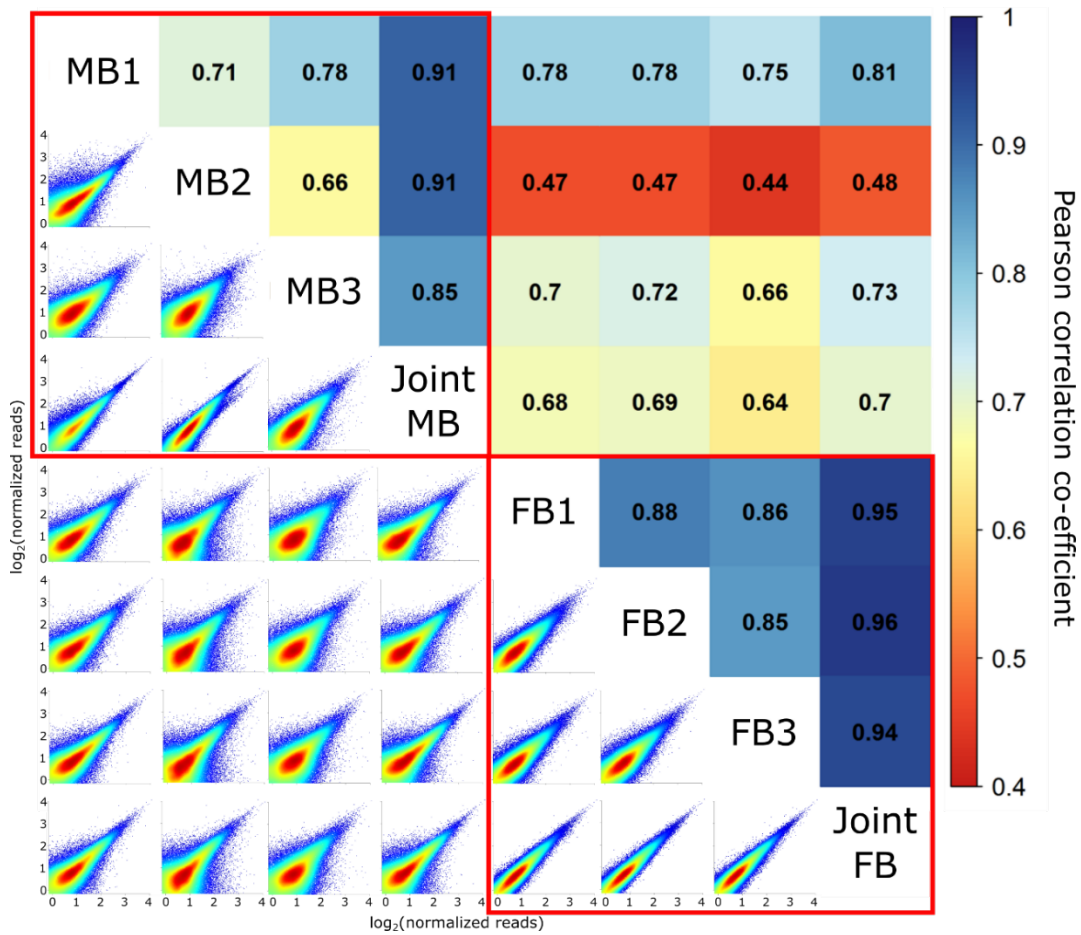


C Fragment length distribution



D ATAC-seq signal at transcription start sites





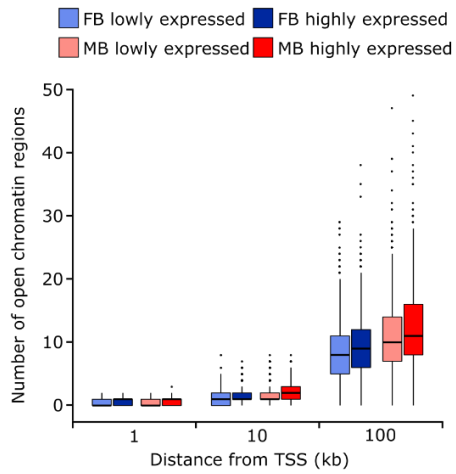
Supplementary Figure 2.3: Correlation analysis of all ATAC-seq libraries

Genome-wide correlation within replicates (red boxed areas) and between brain regions indicate there is strong correlation within a brain region across replicates, with correlation to a lesser extent between brain regions.

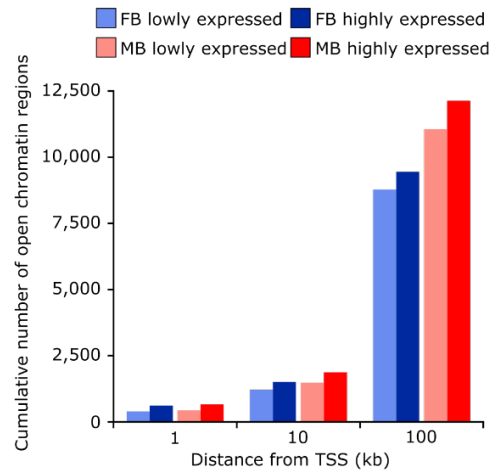
Supplementary Figure 2.4 Relating RNA-seq and ATAC-seq data

Broad analyses indicate that highly expressed genes are under greater regulatory control, in that there are more proximal regulatory elements (**A**, **B**) and their promoters are more open (**C**) compared to lowly expressed genes. (**D**) Additionally, the genes closest to the strongest ATAC-seq peaks are more highly expressed than those adjacent the weakest peaks.

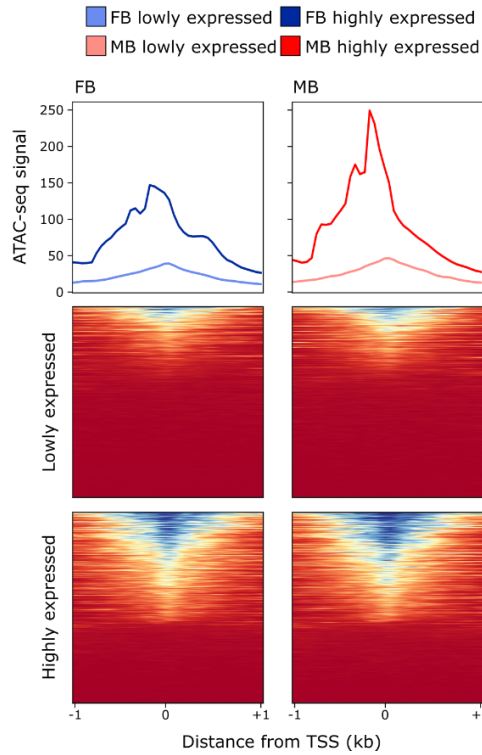
A Number of open chromatin regions per highly and lowly expressed genes



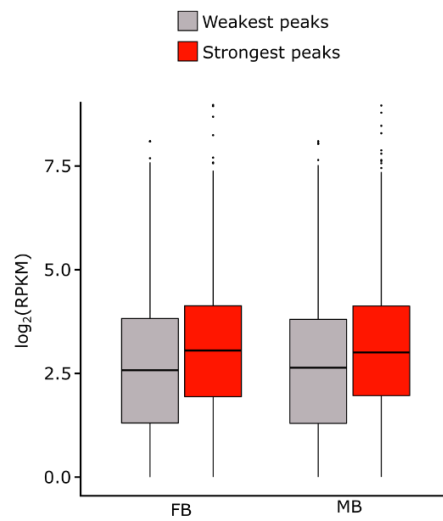
B Number of open chromatin regions adjacent to highly and lowly expressed genes



C ATAC-seq signal at promoters of highly and lowly expressed genes

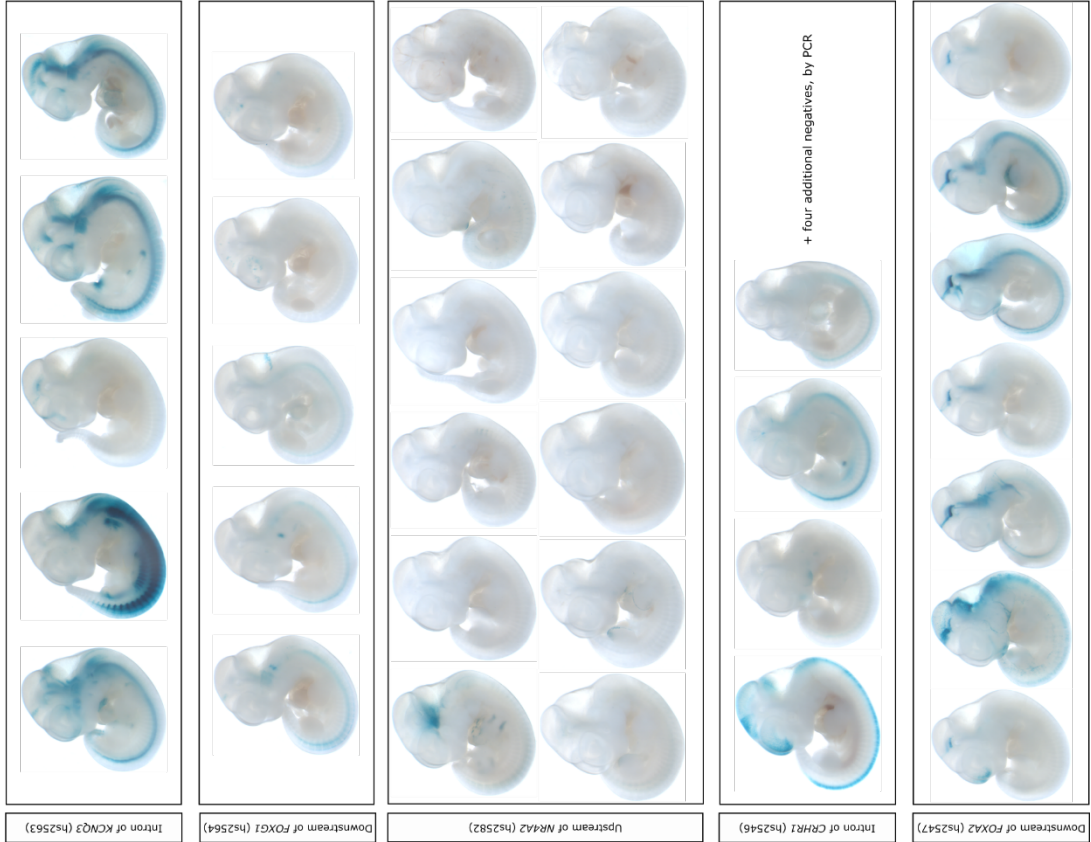
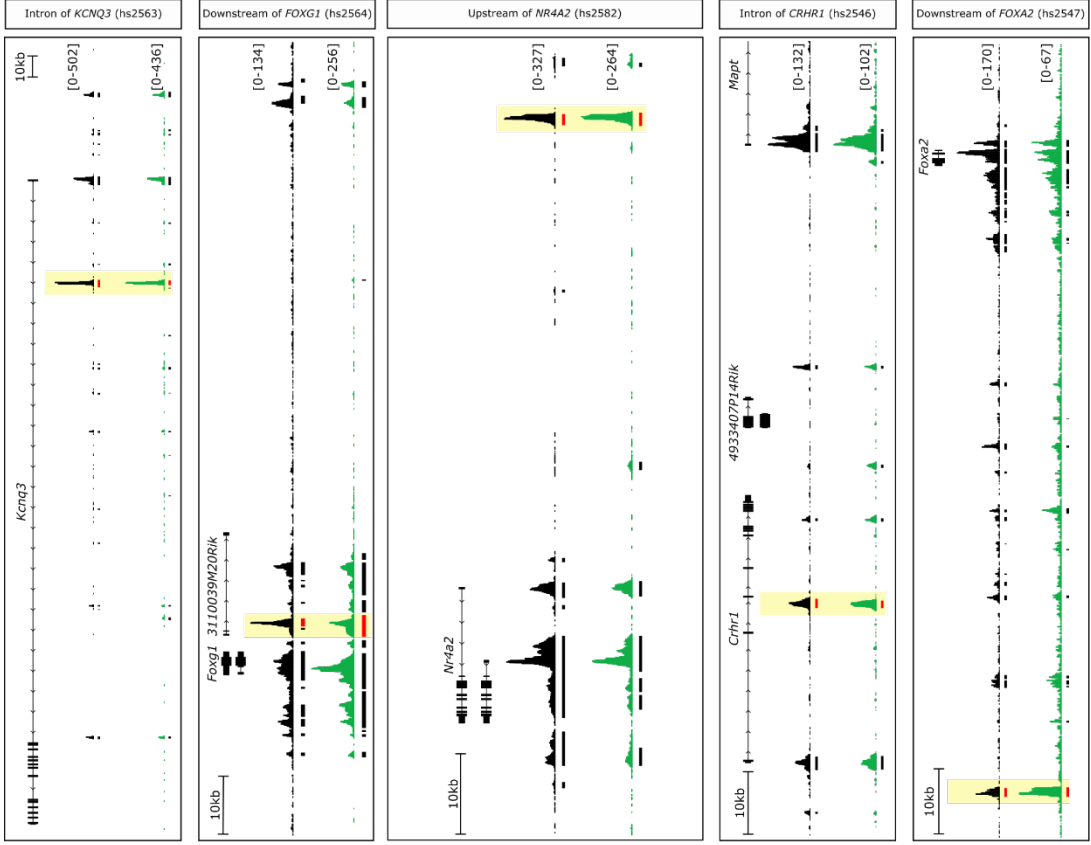


D Expression of genes closest to strong and weak peaks



Supplementary Figure 2.5: All *lacZ* reporter mice and the mouse genomic locations of the putative CREs

All transgenic mouse embryos assayed for *lacZ* reporter activity for each of the five putative CREs tested *in vivo* (left) and the genomic location and context of those putative CREs (right). MB: Black track, FB: Green track. Red peaks in yellow boxes: The putative CREs that were lifted over to hg19 and tested *in vivo*.



Supplementary Table 2.1: Summary of counts and percent overlap with the VISTA enhancer browser, related to **Figure 2.2A-C**

	Counts			Percentage	
	VISTA	MB	FB	MB	FB
Positive	1264	786	719	62	57
<i>Neuronal</i>	<i>652</i>	<i>498</i>	<i>465</i>	<i>76</i>	<i>71</i>
Forebrain	191	137	147	72	77
Midbrain	104	80	65	77	63
Hindbrain	94	66	58	70	62
Multiple regions	156	126	112	81	72
Whole brain	107	89	83	83	78
<i>Non-neuronal</i>	<i>612</i>	<i>288</i>	<i>254</i>	<i>47</i>	<i>42</i>
Negative	1123	538	529	48	47
TOTAL	2387	1324	1248	55	52

Supplementary Table 2.2: Allele and genotype counts and frequencies in PD cases and controls of all variants identified by sequencing within the intronic *SNCA* enhancer

Variant	MA	Population	Allele counts (frequency)		Genotype counts (frequency)		
			Minor allele	Major allele	Homozygous Minor	Heterozygous	Homozygous Major
rs537518252	A	Control	0 (0%)	1910 (100%)	0 (0%)	0 (0%)	955 (100%)
		PD	1 (0.1%)	1909 (99.9%)	0 (0%)	1 (0.1%)	954 (99.9%)
rs78789649	A	Control	0 (0%)	1910 (100%)	0 (0%)	0 (0%)	955 (100%)
		PD	1 (0.1%)	1909 (99.9%)	0 (0%)	1 (0.1%)	954 (99.9%)
rs112174335	C	Control	2 (0.1%)	1908 (99.9%)	0 (0%)	2 (0.2%)	953 (99.8%)
		PD	0 (0%)	1910 (100%)	0 (0%)	0 (0%)	955 (100%)
rs28720123	T	Control	4 (0.2%)	1906 (99.8%)	0 (0%)	4 (0.4%)	951 (99.6%)
		PD	1 (0.1%)	1909 (99.9%)	0 (0%)	1 (0.1%)	954 (99.9%)
rs2737024	G	Control	515 (27%)	1395 (73%)	76 (8%)	363 (38%)	516 (54%)
		PD	609 (31.9%)	1301 (68.1%)	105 (11%)	399 (41.8%)	451 (47.2%)
chr4:90721581 T>C	C	Control	1 (0.1%)	1909 (99.9%)	0 (0%)	1 (0.1%)	954 (99.9%)
		PD	0 (0%)	1910 (100%)	0 (0%)	0 (0%)	955 (100%)
rs2583959	G	Control	518 (27.1%)	1390 (72.9%)	89 (9.3%)	340 (35.6%)	525 (55%)
		PD	606 (31.7%)	1304 (68.3%)	105 (11%)	396 (41.5%)	454 (47.5%)
chr4:90721702 G>A	T	Control	0 (0%)	1910 (100%)	0 (0%)	0 (0%)	955 (100%)
		PD	1 (0.1%)	1909 (99.9%)	0 (0%)	1 (0.1%)	954 (99.9%)
chr4:90721760 T>-	-	Control	0 (0%)	1910 (100%)	0 (0%)	0 (0%)	955 (100%)
		PD	1 (0.1%)	1909 (99.9%)	0 (0%)	1 (0.1%)	954 (99.9%)
rs189903574	A	Control	0 (0%)	1910 (100%)	0 (0%)	0 (0%)	955 (100%)
		PD	1 (0.1%)	1909 (99.9%)	0 (0%)	1 (0.1%)	954 (99.9%)
rs17016188	C	Control	116 (6.1%)	1794 (93.9%)	4 (0.4%)	108 (11.3%)	843 (88.3%)
		PD	156 (8.2%)	1754 (91.8%)	5 (0.5%)	146 (15.3%)	804 (84.2%)
rs28536191	G	Control	2 (0.1%)	1908 (99.9%)	0 (0%)	2 (0.2%)	953 (99.8%)
		PD	0 (0%)	1910 (100%)	0 (0%)	0 (0%)	955 (100%)
chr4:90721974 T>A	A	Control	0 (0%)	1910 (100%)	0 (0%)	0 (0%)	955 (100%)
		PD	1 (0.1%)	1909 (99.9%)	0 (0%)	1 (0.1%)	954 (99.9%)
rs7684892	A	Control	131 (6.9%)	1777 (93.1%)	9 (0.9%)	113 (11.8%)	832 (87.2%)
		PD	121 (6.3%)	1789 (93.7%)	3 (0.3%)	115 (12%)	837 (87.6%)

Supplementary Table 2.3: r^2 values measuring linkage disequilibrium between *SNCA* variants in controls

	rs2737029	rs356168	rs356220	rs356225	rs3857057	rs62306323	rs7689942	rs7684892	rs28536191	rs17016188	rs2583959	chr4:90721581 T>C	rs2737024	rs28720123	rs112174335
rs10018362	0.185	0.14	0.02	0.137	0.417	0.018	0.547	0.534	0.008	0.242	0.039	0.004	0.046	0.017	<0.001
rs2737029	---	0.663	0.523	0.672	0.101	0.07	0.103	0.096	0.002	0.096	0.498	0.001	0.542	0.003	0.001
rs356168	---	---	0.675	0.985	0.085	0.063	0.078	0.078	0.001	0.072	0.307	0.001	0.341	0.002	0.001
rs356220	---	---	---	0.686	0.12	0.03	0.002	0.002	0.002	0.101	0.373	<0.001	0.411	0.003	0.002
rs356225	---	---	---	---	0.082	0.062	0.078	0.078	0.001	0.073	0.31	0.001	0.344	0.002	0.001
rs3857057	---	---	---	---	---	0.007	0.023	0.023	0.014	0.583	0.025	<0.001	0.028	0.027	<0.001
rs62306323	---	---	---	---	---	---	0.01	0.006	<0.001	0.008	0.047	<0.001	0.05	<0.001	0.001
rs7689942	---	---	---	---	---	---	---	0.974	<0.001	0.005	0.022	0.007	0.026	<0.001	<0.001
rs7684892	---	---	---	---	---	---	---	---	<0.001	0.005	0.021	0.007	0.023	<0.001	<0.001
rs28536191	---	---	---	---	---	---	---	---	---	0.016	<0.001	<0.001	<0.001	0.5	<0.001
rs17016188	---	---	---	---	---	---	---	---	---	---	0.019	<0.001	0.024	0.033	<0.001
rs2583959	---	---	---	---	---	---	---	---	---	---	---	<0.001	0.934	0.001	<0.001
chr4:90721581 T>C	---	---	---	---	---	---	---	---	---	---	---	---	<0.001	<0.001	<0.001
rs2737024	---	---	---	---	---	---	---	---	---	---	---	---	---	0.001	<0.001
rs28720123	---	---	---	---	---	---	---	---	---	---	---	---	---	---	<0.001
rs112174335	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Supplementary Table 2.4: Allele and genotype counts and frequencies in PD cases and controls of all variants genotyped from the Guella *et al.* panel

Variant	MA	Population	Allele counts (frequency)		Genotype counts (frequency)		
			Minor allele	Major allele	Homozygous Minor	Heterozygous	Homozygous Major
rs10018362	C	Control	213 (11.1%)	709 (88.9%)	9 (0.9%)	195 (20.3%)	757 (78.8%)
		PD	228 (12.1%)	656 (87.9%)	7 (0.7%)	214 (22.7%)	721 (76.5%)
rs2737029	C	Control	864 (41.1%)	240 (58.9%)	161 (15.3%)	542 (51.5%)	349 (33.2%)
		PD	869 (46.2%)	011 (53.8%)	208 (22.1%)	453 (48.2%)	279 (29.7%)
rs356168	G	Control	910 (47.2%)	016 (52.8%)	227 (23.6%)	456 (47.4%)	280 (29.1%)
		PD	965 (51.2%)	919 (48.8%)	246 (26.1%)	473 (50.2%)	223 (23.7%)
rs356220	T	Control	734 (38.1%)	190 (61.9%)	159 (16.5%)	416 (43.2%)	387 (40.2%)
		PD	827 (43.9%)	055 (56.1%)	190 (20.2%)	447 (47.5%)	304 (32.3%)
rs356225	C	Control	904 (47%)	1018 (53%)	223 (23.2%)	458 (47.7%)	280 (29.1%)
		PD	966 (51.3%)	918 (48.7%)	246 (26.1%)	474 (50.3%)	222 (23.6%)
rs3857057	G	Control	137 (7.1%)	791 (92.9%)	7 (0.7%)	123 (12.8%)	834 (86.5%)
		PD	179 (9.5%)	705 (90.5%)	4 (0.4%)	171 (18.2%)	767 (81.4%)
rs62306323	T	Control	241 (12.6%)	667 (87.4%)	19 (2%)	203 (21.3%)	732 (76.7%)
		PD	205 (10.9%)	679 (89.1%)	10 (1.1%)	185 (19.6%)	747 (79.3%)
rs7689942	T	Control	125 (6.5%)	801 (93.5%)	5 (0.5%)	115 (11.9%)	843 (87.5%)
		PD	117 (6.2%)	767 (93.8%)	2 (0.2%)	113 (12%)	827 (87.8%)

Supplementary Table 2.5 : Primer sequences used for qPCR and cloning for in vivo reporter assays

Characterizing mouse DA neurons by qPCR

	Forward	Reverse	Expected amplicon	mm9 co-ordinates
<i>Pitx3</i>	ACGCACTAGACCTCCCTCCAT	GCTTCTTCTTCAGAGAGCCGT	203	<u>Pitx3 exons 1, 2, 3</u>
<i>Th</i>	CTGTCCACGTCCCCAAGTTCA	CAATGGGTTCCAGGTTCCG	147	<u>Th exons 1, 2</u>
<i>Slc6a3</i>	GAGGCCCCGATAAGAGCTCAAG	CCTTCTTCTTCGACTGCCTCC	111	<u>Slc6a3 exons 1, 2</u>
<i>Actb</i>	TGGCTCCTAGCACCATGAAG	AGCTCAGTAACAGTCCGCCTA	188	<u>Actb exons 5, 6</u>

Testing five putative CREs in in vivo reporter assays

	Forward	Reverse	Expected amplicon	hg19 co-ordinates
<i>KCNQ3</i>	ATAAAGCAAGTGACCGGGGA	GGCTGCTCTTGAGACATTCG	2744	<u>chr8:133425146-133427889</u>
<i>FOXP1</i>	CGGCAAAGGAACATGGAGAG	TCACATCCAGGGCCAAGAAT	2188	<u>chr14:29242870-29245057</u>
<i>NR4A2</i>	ATCAGCCTGTGTCCTGTTCT	AAGGAAGGGGCAGCTTAGAG	2447	<u>chr2:157255824-157258270</u>
<i>CRHR1</i>	CAGGACTATGACGGCTGACT	GGAACACACCCTCTCCATCA	1691	<u>chr17:43889821-43891511</u>
<i>FOXA2</i>	GTCTGATGTTGTTCAACCAG	GCCGTTTTAAGCATTGGGAA	3288	<u>chr20:22382513-22385800</u>

Testing the SNCA enhancer in in vivo reporter assays

	Forward	Reverse	Expected amplicon	hg19 co-ordinates
<i>SNCA</i>	GGACTCCTTGCTTGAAGGAAAAAT	AGACAAAAGGAGTGCATTGATGT	1060	<u>chr4:90,721,063-90,722,122</u>

Chapter 3: Transcription factors and non-coding variants that disrupt their binding in dopaminergic neurons

3.1 Interrogating non-coding variants and their role in disease

The majority of variants implicated in disease by genome-wide association studies are non-coding^{26,66}. These variants are predicted to impact regulatory functioning of non-coding elements, like enhancers, likely through disrupting the binding sites of transcription factors (TFs) and other DNA-binding proteins. Learning the TFs active in a cell type and examining how disease-associated non-coding variants disrupt their activity are important for improving our understanding of disease.

Unlike with coding mutations, our ability to predict the consequences of non-coding variation is limited. There are a variety of algorithms and software available to predict how coding variants will impact the structure of a protein (like CADD, PolyPhen, and SIFT²³¹⁻²³³), as we have an accurate understanding of the genetic code (“vocabulary”) underlying protein specification. We do not have these vocabularies for regulatory element function. As such, there have been many efforts to learn the sequence control underlying regulatory element functioning, especially enhancers²³⁴⁻²³⁶.

One such effort, gapped-kmer support-vector machine (gkm-SVM⁷³), has been developed by our group to begin elucidating the sequence basis of regulatory control, using a machine-learning algorithm. This algorithm and a related method, deltaSVM¹⁶⁹, can be used in combination to predict TFs important to regulatory

control and anticipate how variants may impact those TFs' binding. Functional validation of the predictions of variant effect is commonly performed using *in vitro* reporter assays, like luciferase assay.

To begin predict how Parkinson disease (PD) associated variants might alter transcription and confer risk for PD, we first developed a regulatory vocabulary for midbrain (MB) dopaminergic (DA) neurons and identified four TFs conferring regulatory activity in these cells. With this vocabulary, we scored >7,000 variants associated with neurodegenerative and neuropsychiatric traits for their capacity to disrupt TF binding. We selected >20 of these sequences for validation by luciferase assay however, this has proved a challenge considering the cell-specificity of enhancer activity and the lack of DA neuronal cell culture models. We turn to a cell type agnostic approach, protein-binding arrays, as an alternative to luciferase for evaluating how variants impact protein binding. In a preliminary test of this technique, we identified four proteins, NOVA1, PEG10, SNRPA, and CHMP5, whose binding may be impacted by the *SNCA* enhancer variants. Finally, we have performed preliminary characterization of a possible *in vitro* cell surrogate, SN4741 cells, using karyotyping, and single-cell RNA-seq.

3.2 Candidate regulatory elements are enriched for transcription factor motifs active in dopaminergic neurons²

To identify sequence modules (kmers) predicted to contribute regulatory activity in MB and forebrain (FB) DA neurons, we applied the machine learning

² This section and associated methods have been published in the American Journal of Human Genetics and adapted for use in this thesis. McClymont, S.A. et al. (2018). Parkinson-associated *SNCA* enhancer variants revealed by open chromatin in mouse dopamine neurons. *The American Journal of Human Genetics*, 103:874-892²³⁷.

algorithm, gkm-SVM⁷³ to the MB DA neuron ATAC-seq catalogues²³⁷ generated in **Chapter 2**. The resulting regulatory vocabularies of kmers had high predictive power (auROC_{MB} = 0.915, auROC_{FB} = 0.927). We rank ordered and collapsed related kmers to reveal motifs enriched in the open chromatin regions (OCRs) and their corresponding TFs (**Figure 3.1A, E, I, M**). In the MB, the four most enriched motifs correspond to Rfx1, Foxa2, Ascl2, and Nr4a2.

Given the degeneracy of binding motifs within TF families, we consulted the bulk RNA-seq data for each of the implicated TF families and examined the relative expression levels to prioritize which TFs are most likely producing the observed motif enrichments (**Figure 3.1B, F, J, N**). For example, the reported DNA binding domain is highly conserved between RFX family members and as a result the predicted sequence motif for each is highly similar^{238,239}, thus we must use other means to identify which family member is likely acting in these cells. While no member of *the Rfx family* has been canonically associated with MB DA neurons, we anticipate *Rfx3* and *Rfx7*, as the two highest expressed *Rfx* genes, to likely be active in MB DA neurons and driving this motif enrichment (**Figure 3.1B**). Foxa1, and more specifically, Foxa2 are both known to DA neuron biology^{175,240} and both are highly expressed in the MB DA neurons (**Figure 3.1F**). Regarding enrichment for the Ascl family, Ascl1 is known to be involved in DA neuron biogenesis²⁴¹ and is more highly expressed than any other TF in the family (**Figure 3.1J**). Finally, Nr4a2 is both canonically associated with DA neurons and required for their development¹⁷⁸; we observe it to be highly expressed in MB DA neurons (**Figure 3.1N**). Examining the sequences underlying the OCR catalogues, we identified TF families known and unknown to DA neuron biology and further refined the TF associations using expression data.

We also examined the qualities that differentiate MB OCRs from FB OCRs by examining the sequences underlying MB-specific and FB-specific regions. We developed a vocabulary that discriminates MB and FB regions with high predictive power (auROC = 0.926) and identified kmers enriched in MB-specific peaks where the top corresponding TFs are Foxa1/2 and Nr4a2 (**Supplementary Figure 3.1**). We confirmed this MB bias by again considering the bulk RNA-seq for these genes. As expected, these TFs are more highly expressed in the MB where *Nr4a2* is present at 12-fold higher levels in the MB (135 RPKM in the MB vs 11 RPKM in the FB) and *Foxa1/2* are not expressed in the FB, but are present in the MB (*Foxa1*: 28 RPKM, *Foxa2*: 7 RPKM). Not only do we identify Foxa1/2 and Nr4a2 as more active in MB DA neurons than in the FB, we did so solely by comparing their role in the vocabulary of MB-specific OCRs versus FB-specific OCRs.

In a parallel strategy to identify TFs actively engaging the DNA in MB DA neurons, we performed TF footprinting in a single deeply sequenced MB ATAC-seq library. Doing so, we confirm that two of the TFs prioritized by gkm-SVM leave robust footprints. The motif corresponding to Rfx-binding results in a dearth of cuts directly over predicted binding sites (**Figure 3.1C**). The same can be seen to a lesser extent for the motif corresponding to Foxa1/2 (**Figure 3.1G**). By contrast, motifs corresponding to Ascl1 or Nr4a2 fail to leave a robust mark on the chromatin availability (**Figure 3.1K, O**). It has been noted that nuclear receptors, like Nr4a2, only transiently interact with DNA²⁴², and as a result, it may be that the short DNA residence time fails to result in a robust footprint detectable by transposition. These footprinting data substantiate the claim that the Rfx family of TFs and Foxa1/2 are active in MB DA neuron CREs.

We confirmed that these sequences are indeed enriched in the catalogues by examining the pileup of reads overlapping all genome-wide predicted motif binding sites for each motif identified by gkm-SVM. We see an abundance of reads over predicted binding sites of all four motifs (**Figure 3.1D, H, L, P**), with the strongest enrichment overlapping Rfx and Ascl1 motif sites (**Figure 3.1D, L**). Despite the less robust footprint generated at the Ascl1, this TF clearly underlies a larger than expected proportion of OCRs in the MB catalogue.

The integration of a support vector machine learning algorithm as applied to the sequences underlying OCRs with footprinting analysis in the same chromatin substrate powerfully identifies TFs that are important for DA neuron biology and suggests the Rfx family of TFs, Foxa1/2, Ascl1, and Nr4a2 are actively influencing gene expression in the MB DA neurons.

3.3 Predicting and testing the effects of regulatory variants

Next, we sought to use this vocabulary to predict the effect of disease-associated variants on enhancer functioning. We collected the lead SNPs plus those in high linkage disequilibrium (LD; $r^2 \geq 0.8$) from five genome-wide association studies (GWAS) for neurodegenerative and neuropsychiatric traits: Alzheimer disease²⁴³, epilepsy²⁴⁴, PD²⁵, progressive supranuclear palsy²⁴⁵, and schizophrenia²⁴⁶ (summarized in **Supplementary Table 3.1**). These variants were filtered for duplicates and those that are unable to be scored by deltaSVM (eg: multiallelic SNPs), after which a total of 7,719 variants were scored by deltaSVM for their impact based on the MB and FB DA neuron vocabularies. This calculation compares the gkm-SVM scores of the alternative and reference alleles and sums across the surrounding

sequence to predict the consequence of the variant on regulatory activity. The distribution of these scores are shown in **Figure 3.2A**. We observe a long right tail, representing variants that are predicted to be highly influential in TF binding.

The top variant identified by deltaSVM is rs1498232, associated with schizophrenia, which is predicted to be highly damaging to TF binding in both the MB and FB vocabularies (deltaSVM_{MB}: -22.347, deltaSVM_{FB}: -19.506). How deltaSVM sums the variant impact using the MB vocabulary is demonstrated in **Figure 3.2B**. This variant is predicted to be highly damaging to a RFX binding site (**Figure 3.2C**). There are ten SNPs in high LD ($r^2 > 0.8$), none of which are predicted by deltaSVM to strongly alter regulatory function (**Figure 3.2D**). Using other vocabularies our group has generated to assess the cell-type specificity of this variants effects, we scored these same variants for disrupting activity in a lymphoblastoid cell line²⁴⁷ (GM12878) and melanocytes²⁴⁸ (melanA). These variants are not scored as altering regulatory function (**Figure 3.2E, F**).

We sought to test these variants using luciferase in the SK-N-SH neuroblastoma cell line. We cloned 500bp fragments centred on the variant to be tested (**Figure 3.2G**) upstream of a minimal E1B promoter in a luciferase expression vector⁷⁷, containing either the reference or alternate allele. Just one of the constructs, and not that predicted by deltaSVM, exceeded background levels of regulatory activity (**Figure 3.2H**). The one construct exceeding background levels did not validate in a replication experiment (data not shown). This was disappointing but perhaps unsurprising given that these variants fall in a gene desert and none overlap an OCR in either MB or FB DA neurons. To increase the *a priori* probability of the variants

falling in an active regulatory region, we next restricted ourselves to testing variants that are highly rated as damaging that fall in ATAC-seq peaks.

Of the 7,719 variants scored by deltaSVM, 275 variants overlap either a FB or MB peak. Many of these variants overlap with promoters and were removed from our analysis, leaving 137 variants overlapping an OCR. The distribution of these scores (**Figure 3.3A**) largely reflects the distribution of the variants as a whole and remains skewed. We selected ten peaks containing the top 11 scored disease-associated variants (**Figure 3.3B**). A schematic of the peak locations, the intersecting variants, and the nearest gene is indicated in **Figure 3.3C**. Centering on the variants, we cloned 500bp fragments containing the risk and non-risk alleles for these regions and performed luciferase reporter assay. Again, no construct exceeded background levels of activity (**Figure 3.3D**).

While unlikely, it could be that none of the selected sequences direct enhancer activity. To test this, we assayed the OCR in intron 4 of *SNCA* that we had already demonstrated to have enhancer activity *in vivo* in **Chapter 2**. We cloned the exact enhancer region that was tested *in vivo*, containing the risk and non-risk haplotypes. After performing luciferase assay on these constructs, again we observe no enhancer activity (**Figure 3E**). This result indicated to us that there is likely an underlying issue with the *in vitro* reporter assay – we observe no enhancer activity in a sequence that is demonstrably an enhancer *in vivo*.

To address this, we have performed a variety of troubleshooting steps, like changing the promoters upstream of the luciferase reporter (E1B, TATA, SV40, SYN1), the promoters upstream of the renilla reporter (CMV, SV40), the lipofection

method (lipofectamine 3000, 2000), how the vectors are cloned (Gateway cloning, restriction cloning), the orientation and size of the constructs (forward and reverse, full enhancer and restricted to a central portion), and importantly, the cell line in which the construct is being tested (SK-N-SH, SH-SY-5Y, Neuro2A). No change has reliably demonstrated the *SNCA* sequence to drive expression *in vitro* (data not shown).

We believe these results represent a fundamental problem with these *in vitro* assays – the cells in which we are testing are not the correct cellular surrogate for embryonic MB DA neurons. We observed the exquisite restriction of the enhancer activity of the *SNCA* sequence in **Chapter 2**, it is likely that the immortalized cell lines are poor proxies and do not represent the restricted cell types that this enhancer is active in.

3.4 Protein binding arrays are a viable alternative validation strategy³

We turned to parallel strategies for assaying the effects of variants on enhancer activity that do not rely on cell surrogates. For a cell agnostic assay, we turned to protein binding arrays²⁴⁹. To verify the utility of this assay, we examined the proteins whose binding is disrupted by the *SNCA* enhancer variants.

We assayed differential protein binding at the PD-associated variants at *SNCA* for >16,000 proteins²⁴⁹. In doing so, we identify five proteins whose binding is robustly

³ This section and associated methods have been published in the American Journal of Human Genetics and adapted for use in this thesis. McClymont, S.A. et al. (2018). Parkinson-associated *SNCA* enhancer variants revealed by open chromatin in mouse dopamine neurons. *The American Journal of Human Genetics*, 103:874-892²³⁷.

impacted by these implicated variants: NOVA1, APOBEC3C, PEG10, SNRPA, and CHMP5 (**Figure 3.4A, B, C**). Of these, all are expressed at appreciable levels in both MB and FB DA neurons (**Figure 3.4D**), excluding APOBEC3C (RPKM ≤ 1). Of the remaining four proteins, three (PEG10, SNRPA, and CHMP5) demonstrate an increased binding affinity for the minor risk allele over the major allele; this direction of effect is consistent with the over-expression paradigm by which *SNCA* confers PD risk²⁰. Interestingly, CHMP5 is the sole protein we identify whose binding affinity is impacted by variant rs2583959, and our group has recently implicated one of its family members, CHMP7, in conferring PD risk¹⁷⁹, perhaps indicating a role for this family of proteins in PD. Although no single protein stands out, the increased affinity for the risk alleles of the identified enhancer variants by proteins expressed in DA neurons is consistent with a potential mechanistic contribution to *SNCA* expression and therefore, PD risk.

This was a promising result indicating to us that protein binding arrays are appropriate as an alternative strategy for luciferase to assess how enhancer variants affect TF binding and thus alter enhancer activity and gene expression. However, these assays are inherently synthetic as they occur in the absence of a cellular environment and only test small fragments of DNA not in their larger endogenous sequence context. While this method has been demonstrably successful and we will continue to use protein binding arrays to assay enhancer variants in the absence of a better test, we continue to search for a more appropriate cellular surrogate in which to perform our *in vitro* reporter assays.

3.5 The suitability of *in vitro* dopaminergic neuron surrogates: the SN4741 cell line

We have identified the cell line SN4741 as a likely *in vitro* cell surrogate. SN4741 cells are derived from E13.5 mouse DA neurons from the substantia nigra and contain a SV40Tag that directs the differentiation when the cell culture is shifted to 39°C²⁵⁰. We performed preliminary expression analysis by RT-qPCR to confirm expression of a variety of DA neuron markers. We observed increases of these markers under the higher temperature condition, indicating the cells are indeed differentiating towards a more DA state (**Figure 3.5A**).

With this promising result, we moved forward to more deeply characterize these cells to assess their suitability as *in vitro* surrogates for *in vivo* MB DA neurons. First, we performed karyotyping analysis on 20 cells to assess the chromatin complement of these cells (representative karyogram in **Figure 3.5B**). Interestingly, SN4741 cells appear to be an unstable triploid line (**Figure 3.5C**), with a variety of marker chromosomes. None of the 20 cells assessed had the same chromosome complement. This is concerning for the viability of these cells as a surrogate for a variety of reasons. The biggest being if the cells are genetically unstable, there may be large experimental batch effects as the cell populations shift across divisions.

To assess the consistency of the differentiation protocol, we compared the transcriptomes from $\geq 17,000$ cells in the permissive (37°C) and non-permissive (39°C) states. Cluster analysis indicates a separation of the cells at each temperature (**Figure 3.5D**). This separation of cells by temperature is accompanied by changes to the cell cycle, with cells at the permissive (37°C) temperature containing cells in either

G2M or S phase, while cells at the non-permissive temperature (39°C) are mostly differentiated and in G1 phase (**Figure 3.5E, F**). In expression analysis, we observe that markers of proliferation, like *Ki-67*, are exclusively expressed in cells at the permissive temperature (**Figure 3.5G**), corroborating the cell cycle analysis. SN4741 cells, when shifted to the non-permissive temperature, appear to robustly differentiate.

However, in examining expression of a variety of DA neuron markers, we fail to detect expression in either the permissive or non-permissive temperature. Markers, including *Th*, *Nr4a2*, and *Slc6a3* have few to no reads assigned to them (**Figure 3.5H, I, J**). It appears that while these cells are differentiating when shifted to the non-permissive temperature, we are unable to confirm these cells are entering a DA trajectory when doing so.

3.6 Discussion

As GWASs are applied to ever more common diseases, methods to identify non-coding variants contributing to risk through altering regulatory activity are needed. Machine learning algorithms, like the one we apply here, are being explored as a method for predicting TFs active in a cell type and how variants might impact their binding. Here, we use the machine learning algorithm, gkm-SVM, to predict TFs active in MB DA neurons and use this vocabulary to rank >7,000 variants for their capacity to alter regulatory functioning. Our ability to validate these predictions is limited as a result of the cell type specific action of enhancers, as demonstrated by our luciferase assay experiments. We explored cell type agnostic approaches, like protein binding arrays, to great success. In testing the *SNCA* enhancer variants for their

effect on protein binding, we identified five proteins whose binding is affected, three of which, PEG10, SNRPA, and CHMP5, display greater affinity for the risk allele. Finally, we explored the suitability of the SN4741 cell line as a surrogate for *in vivo* embryonic MB DA neurons.

In examining the sequence composition underlying the ATAC-seq peaks, we illuminate powerful vocabularies for both FB and MB DA neuron transcriptional regulatory control. Machine learning using gkm-SVM prioritizes four transcription factor families (Rfx, Foxa1/2, Nr4a2, Ascl1/2) as those conveying significant regulatory potential in the CRE catalogues. Of these, the Rfx family had not previously been implicated in DA neuron biology. Although several of the Rfx family members have been annotated as having expression in the cerebellum or fetal brain²³⁸, a role specifically in MB DA neurons has not previously been appreciated. By contrast, Nr4a2 is canonically associated with MB DA neurons^{177,178}, is highly expressed in this population (139 RPKM), and was prioritized as a TF conferring regulatory potential in these cells; however, TF footprinting fails to provide evidence supporting its activity. We postulate that this lack of footprint may reflect the transient DNA binding dynamics of Nr4a2. Transcription factors with short DNA residence times often fail to reveal footprints, and nuclear receptors, such as Nr4a2, have markedly transient DNA interactions²⁴².

We apply this vocabulary to predict how variants identified by GWASs conferring risk for Alzheimer disease, epilepsy, PD, progressive supranuclear palsy, or schizophrenia might impact regulatory activity. We prioritized >20 variants for validation using two strategies for prioritization; first, we examined the highest ranked variant and ten variants in LD and next, we examined the regulatory

functioning of ten MB OCRs containing highly ranked disrupting variants. Despite the variety of variants tested, none exceeded background levels of activity. We hypothesize this is the result of highly restricted enhancer activity of the regions being assayed and supporting this hypothesis, the sequence at *SNCA* that we previously identified as an enhancer *in vivo*, fails to direct reporter activity *in vitro*. While we have explored alternative strategies for interrogating enhancer variant functioning, ideally an *in vitro* cellular surrogate will be identified.

Towards this, we have begun characterizing the embryonic MB DA-derived neuronal cell line, SN4741. While initial qPCR results were promising, deeper characterization suggests these cells are an unstable triploid cell line that are perhaps not as DA as hoped. scRNA-seq in these cells fails to detect activity of key DA neuron markers, especially *Th*. However, scRNA-seq collects very sparse matrices of information and it may be that these markers are lowly expressed and escape detection by scRNA-seq. RT-qPCR is far more sensitive and those results do demonstrate a shift towards a DA trajectory with a shift in temperature. To assess how well these cells might match the *ex vivo* MB DA neurons, we have performed bulk RNA-seq and ATAC-seq at both temperatures. We have begun to explore changes induced by the temperature shift, but importantly, we will be focusing on comparing the transcriptomes and chromatin landscape of the *ex vivo* MB DA neurons to this cell line.

3.7 Methods

Regulatory vocabulary development

We applied the machine learning algorithm gkm-SVM⁷³ to the MB and FB catalogues generated in *Chapter 2*, under default settings. We trained on the sequences underlying the summits ± 250 bp of non-ubiquitously open, top 10,000 peaks by signal intensity, versus five negative sets, matched for GC content, length, and repeat content. Weights across all five tests were averaged for all 10-mers.

All 10-mers with weight ≥ 1.50 were clustered on sequence similarity using Starcode²⁵¹, using sphere clustering with distance set to 3. clustalOmega²⁵² aligned the sequences within these clusters and MEME²⁵³, under default parameters, excepting `-dna -maxw 12`, generated position weight matrices (PWMs) of these aligned clusters. Tomtom²⁵⁴, querying the Jolma 2013, JASPAR Core 2014, and Uniprobe mouse databases, identified the top transcription factors corresponding to these PWMs, under default parameters excepting `-no-ssc -min-overlap 5 -evaluate -thresh 10.0`.

The same procedure was used to identify transcription factors specifically conveying regulatory potential in the MB library relative to the FB library, except during gkm-SVM training, the positive set was specified as the top 10,000 non-ubiquitously open MB summits and the negative set was specified to be the top 10,000 non-ubiquitously open FB summits, both ± 250 bp.

Transcription factor footprinting

A single MB ATAC-seq library was sequenced on the Illumina HiSeq in Rapid Run mode with 2x100bp reads, to a depth of ≥ 350 million paired-end reads. Analysis was performed as described in *Chapter 2*. CENTIPEDE²⁵⁵ was used to identify footprints. Sequences underlying the deeply sequenced MB library peaks, less those ubiquitously open, were extracted. FIMO²⁵⁶, with options --text --parse-genomic-coord, identified all locations underlying ATAC-seq peaks of the motifs identified above. Additionally, conservation data from 30-way vertebrate phastCons was considered in the CENTIPEDE calculations; for each PWM site, those with mean conservation score greater than 0.9 were considered. Finally, the BAM file read end co-ordinates were adjusted in response to the shift in co-ordinates due to the transposase insertion²⁵⁷. As such, following the original ATAC-seq method⁹⁵, reads were adjusted +4bp on the positive strand and -5bp on the negative strand.

Genome-wide read pileup over predicted motif sites

FIMO, as above, was used to identify all co-ordinates genome wide of the identified motifs. deepTools²⁰⁴ “bamCoverage” tool was run under default conditions, to convert the deeply sequenced MB library BAM to bigwig format. Following this, a matrix file was generated with “computeMatrix”, with options --referencePoint center -b 1000 -a 1000 -bs 50 specified. Finally, “plotHeatmap” was used to generate plots indicating ATAC-seq read pileup over predicted motif sites.

deltaSVM predictions of variant effect

Selecting variants

Lead variants from GWASs on Alzheimer disease²⁴³, epilepsy²⁴⁴, PD²⁵, progressive supranuclear palsy²⁴⁵, and schizophrenia²⁴⁶ were downloaded from the NHGRI-EBI GWAS Catalog²⁵⁸. Proxy variants in high LD ($r^2 \geq 0.8$) were collected with rAggr, querying: 1000 Genomes phase 3 (Oct 2014) CEU+FIN+GBR+IBS+TSI populations for variants with minimum minor allele frequency of 0.001 and a maximum distance of 500kb from the lead SNP.

Scoring variants

Identified variants were submitted to the deltaSVM perl script¹⁶⁹. FASTA files were generated containing the 19bp sequence context centred on the variant, either containing the reference or alternate allele. The MB DA neuron vocabulary was used to score the variants. Vocabularies from GM12878 lymphoblastoid cells²⁴⁷ and melana cells²⁴⁸ were also used to score variants, as negative controls.

Motif analysis

The damaging effect of the variants on TF binding was predicted with motifbreakR²⁵⁹. Querying the MotifDb collection of protein-DNA binding motifs²⁶⁰, variants with a predicted damaging effect with p-value $\leq 1 \times 10^{-8}$ were considered.

Validation of deltaSVM predictions of effect

Cloning of prioritized sequences

Primers (**Supplementary Table 3.2**) were designed using Primer-BLAST²¹² under default parameters with the requirement that the primers fall within 400bp on either side of the variant. Sequences were PCR amplified from human genomic DNA extracted from lymphoblastoid cell lines that were heterozygous for each variant, identified from the 1000 Genomes Project, provided by the Coriell Institute (HG03832, HG01883, HG00187, HG03193, and HG03690). PCR amplicons were BP cloned (Gateway; Invitrogen #11789020) into pDONR221 (Invitrogen) and transformed. Plasmid sequences were confirmed by diagnostic digest (BsrGI) and Sanger sequencing. QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent #210518) was used to remove unwanted variation and, when necessary, to induce the alternative alleles. Sequences were LR cloned (Gateway; Invitrogen #11791020) into the E1B promoter luciferase vector⁷⁷. Final vectors were quantified with the Qubit dsDNA Broad Range Assay (Invitrogen) and standardized to 500ng/ μ L.

Cell culture of SK-N-SH cells

SK-N-SH cells were cultured in EMEM with 10% FBS and 1X penicillin and streptomycin for at least three passages before assay. For luciferase assay, cells were plated at a density of 100,000 cells per well in a 24-well plate and lipofected the following day.

Lipofection

Each well was lipofected with 500ng of the luciferase reporter plasmid containing the putative enhancers and 10ng of CMV-RL renilla expression vector (Promega #E2261) using 0.75µl Lipofectamine 3000 per well (Life Technologies #L3000015). Three to four biological replicates and two to four technical replicates were performed for each construct.

Luciferase activity assay

48 hours post-lipofection, cells were lysed and lysates were collected for assay with the Dual-Luciferase Reporter Assay System (Promega # E1960). Luciferase and renilla activity was measured with the Tecan GENiosPro Microplate Reader (Tecan Group Ltd.) luminometer with automatic injectors. Within samples, the luciferase activity was normalized to renilla activity. Across samples, luciferase activity was further normalized to the average luciferase activity of the empty pE1B plasmids (containing a single basepair between the Gateway arms).

Protein array testing differential binding

HuProt v3.1 human proteome microarrays printed on the PATH surface containing >16,000 unique proteins representing 12,586 genes (CDI laboratories)²⁴⁹ were blocked with 25mM HEPES pH 8.0, 50mM potassium glutamate, 8mM MgCl₂, 3mM DTT, 10% glycerol, 0.1% Triton X-100, 3% BSA on an orbital shaker at 4°C for ≥3 hours. Allele specific protein-DNA binding interactions were identified through dye-swap competition of major and minor alleles labeled with either Cy3 or Cy5. DNA fragments for rs2737024 and rs2583959 were synthesized with the SNP for each allele

flanked by 15 nucleotides of the upstream and downstream sequence and a common priming site at the 3' end (*Supplementary Table 3.3*).

The dsDNA fragments were created by separately annealing a primer containing a Cy3 or Cy5 label and adding Klenow (NEB) with dNTP to fill-in the complementary strand for each allele²⁶¹. Cy3 labeled major allele was mixed with Cy5 labeled minor allele (each at 40nM) in 1x hybridization buffer (10mM Tris-Cl pH 8.0, 50mM KCl, 1mM MgCl₂, 1mM DTT, 5% glycerol, 10μM ZnCl₂, 3mg/mL BSA) and added to an array, dyes were then swapped for each allele and the mixture was then added to a second array. DNA was allowed to bind overnight at 4°C on an orbital shaker with protection from light. Chips were washed once with cold 1xTBST (0.1% Triton X-100) for 5 minutes at 4°C, rinsed, and dried in the centrifuge. Cy5 and Cy3 images were taken separately on a Genepix 4000B scanner and, after alignment to the GAL file, individual spot intensities were extracted using the Genepix Pro software.

Allele specific interactions were identified through dye swap analysis. The ratio of major/minor allele binding was calculated using the duplicate spot average median foreground signal for each protein according to the following equation:

$$\log_2 \sqrt{\frac{Cy3_{major} * Cy5_{major}}{Cy3_{minor} * Cy5_{minor}}}$$

Mean intensity was calculated by averaging the foreground signal for the Cy3 and Cy5 channels of the major and minor alleles. MA plots were made for each allele using the calculated mean intensity and the log ratio of the major/minor allele.

SN4741 characterization

Cell culture of SN4741 cells

SN4741 cells were cultured in high glucose DMEM with 10% FBS and 1X penicillin and streptomycin at 37°C. To induce differentiation, 24 hours after the cells were passaged, the flask was moved to differentiation media (high glucose DMEM with 0.5% FBS) and were cultured at 39°C for 48 hours.

cDNA synthesis and RT-qPCR for DA neuron markers

RNA was extracted using the RNeasy Mini Kit (Qiagen). Each RNA sample was submitted to first strand cDNA synthesis using the SuperScript III First-Strand Synthesis System for RT-PCR (Invitrogen), following the Oligo(dT) method.

Primers (***Supplementary Table 3.4***) were designed using Primer-BLAST²¹² under default parameters with the requirement for exon-exon junction spanning specified. qPCR was performed using Power SYBR Green Master Mix (Applied Biosystems). Reactions were run in triplicate, following default SYBR Green Standard cycle specifications on the Viia7 Real-Time PCR System (Applied Biosystems). Relative quantification followed the $2^{-\Delta\Delta CT}$ method, normalizing results to *Actb* in the samples at the non-permissive temperature.

Karyotyping

G-banded karyotyping was performed by WiCell on twenty cells. Resulting karyograms were quantified and the karyotypes were summarized and plotted.

Single-cell RNA-seq

Cells at both the permissive and non-permissive temperatures were trypsinized and scRNA-seq libraries were generated following the Chromium 10X pipeline. Four replicates at each temperature across 17,000 cells were assayed.

scRNA-seq analysis was performed with Seurat. Cells were filtered to remove stressed/dying cells (% of reads mapping to the mitochondria > 15%) and empty droplets and doublets (number of unique genes detected less than 200 or greater than 6,000). Cells were normalized using ``SCTransform()`` and the effect of percent mitochondrial reads and the sequence depth was corrected for. Principal component (PC) analysis was performed and a PC cut-off was identified using ``JackStraw()`` and ``ElbowPlot()``. UMAP clustering using this PC cutoff and a minimum distance of 0.001 was used for dimensionality reduction.

Cells were scored for their stage in the cell cycle using ``CellCycleScoring()`` on cell cycle genes provided by Seurat (``cc.genes``). Expression was plotted on a log scale with ``VlnPlot()`` for a variety of proliferation and DA neuron markers (shown: *Mki67*, *Th*, *Nr4a2*, *Slc6a3*).

3.8 Figures and supplementary materials

Figure 3.1: Identification of transcription factors (TFs) important to DA neurons

(A) The kmer predicted to have the greatest regulatory potential underlying MB ATAC-seq peaks corresponds to the Rfx family of TFs. (B) RNA-seq quantification in these same cells indicates this enrichment is likely due to Rfx3 or Rfx7 activity. Examining the ATAC-seq signal over predicted binding sites reveals a robust TF footprint (C) and a general enrichment of reads overlapping Rfx sites genome-wide (D). (E-H) Similarly, a kmer corresponding to the TFs Foxa1/2 have similar evidence for their activity. (I-J) The third ranked motif likely corresponds to Ascl1, and while it fails to leave a robust TF footprint (K), there is clear enrichment of ATAC-seq signal overlapping genome-wide predicted Ascl1 binding sites (L). (M, N) Nr4a2, canonically associated with DA neuron biology, is identified as a highly expressed TF likely contributing to the regulatory potential of the putative CREs however, it fails to leave a TF footprint in the cut-site patterns around predicted motif sites (O) and is only mildly enriched for ATAC-seq reads over its predicted binding sites (P).

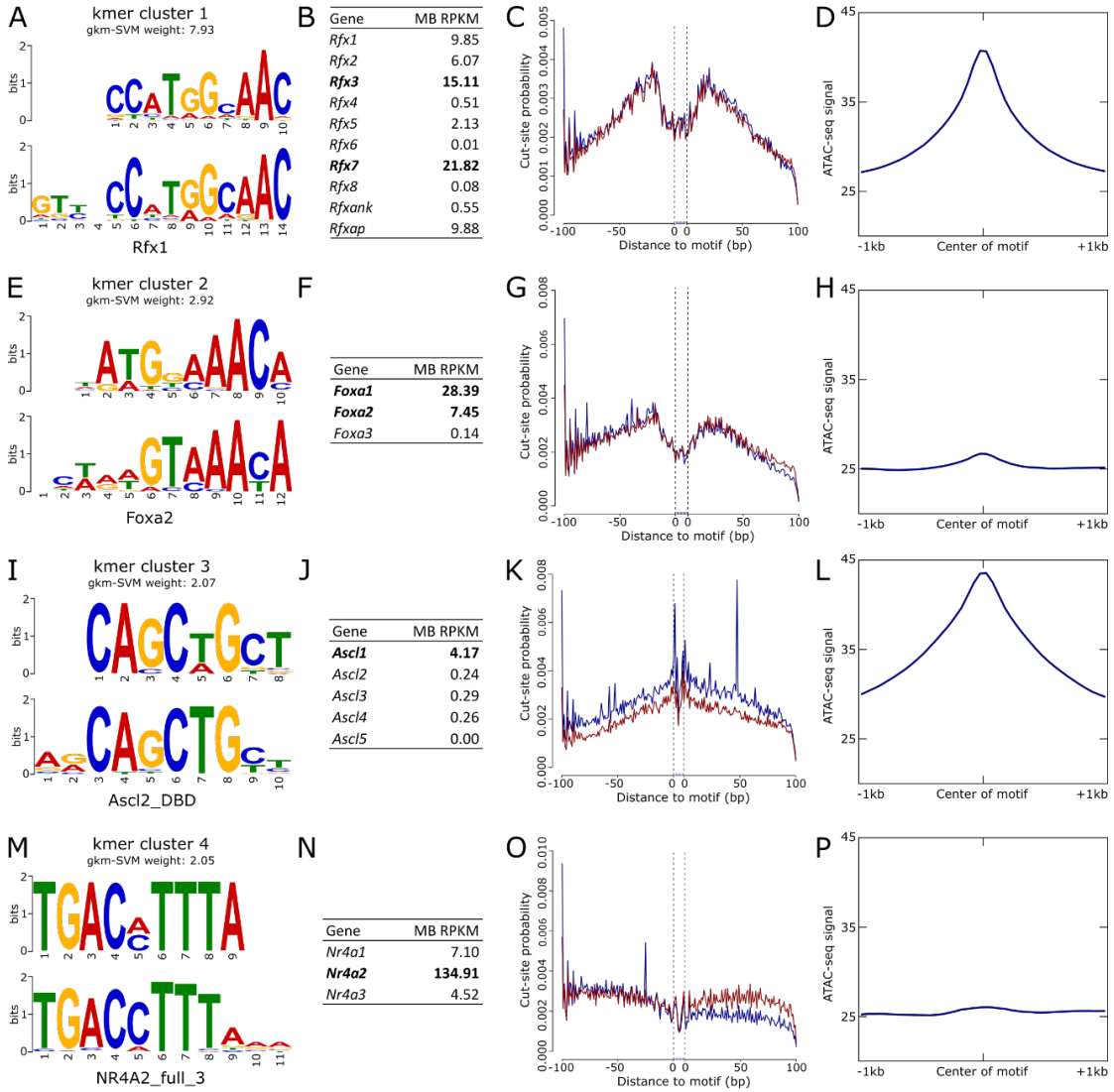
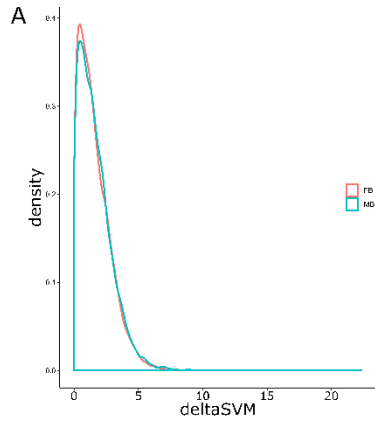


Figure 3.2: Investigating rs4988232 for an effect on enhancer activity

(A) The distribution of gkm-SVM scores for all disease-associated variants scored in the midbrain and forebrain. (B) The schizophrenia-associated variant, rs4988232, is the top variant ranked by deltaSVM as highly damaging. (C) This variant is predicted to disrupt an RFX binding site. (D) None of the other variants in high LD ($r^2 \geq 0.8$) are predicted to be nearly as damaging as rs4988232. (E, F) When these variants are scored with deltaSVM using a different cell type vocabulary, this locus is not predicted to be damaging to transcription factor binding. (G) 500bp centred on each variant was cloned for luciferase assay (red: top ranked variant). (H) Luciferase activity for the reference and alternative alleles of each variant. The top predicted variant does not validate, and nine of the ten constructs do not appear to direct reporter activity above background levels.



B Top scoring SNP: rs1498232, a schizophrenia related variant
 Reference allele: CCGTTTCCA**T**GGCAACCAG
 Alternate allele: CCGTTTCCA**C**GGCAACCAG

Reference 10mer	Weight	Alternate 10mer	Weight	Difference
CCGTTTCCA T	1.27	CCGTTTCCA C	0.25	-1.02
CGTTTCCA T G	2.07	CGTTTCCA C G	0.75	-1.32
GTTTCCA T GG	6.08	GTTTCCA C GG	2.20	-3.89
TTTCCA T GGG	2.86	TTTCCA C GGC	1.01	-1.85
TCCA T CCGA	2.15	TCCA C CCGA	0.76	-1.39
TCCATGGCAA	8.06	TCC C ATGGCAA	1.82	-2.85
CCATGGCAA C	7.93	CC C ATGGCAA	2.84	-5.09
CATCCCAACC	4.31	C C ATCCCAACC	1.61	-2.70
ATGCCAACC A	2.45	ACGGCAACC C	0.90	-1.55
TGGCAACC A G	1.96	CGGCAACC C A	0.77	-1.19
				-22.95

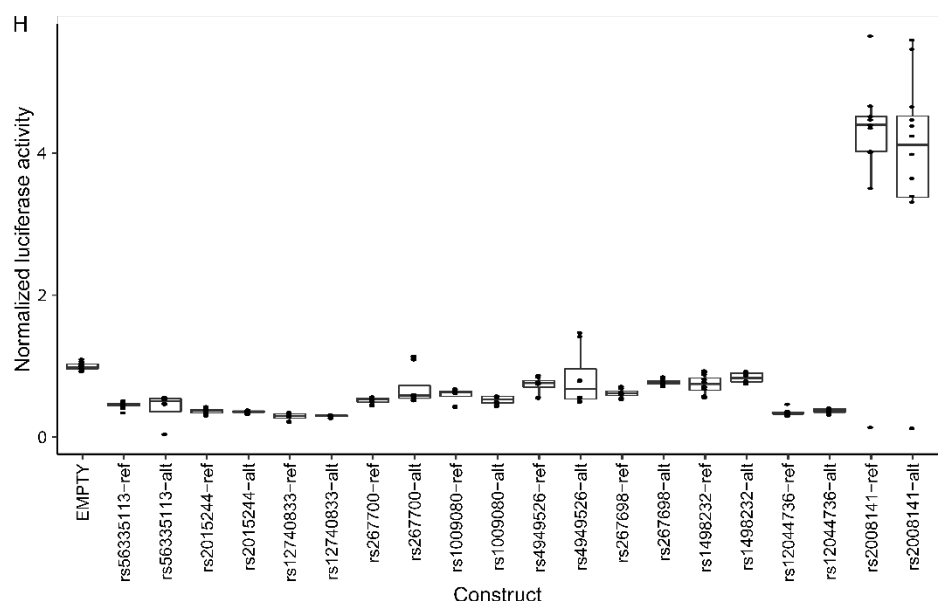
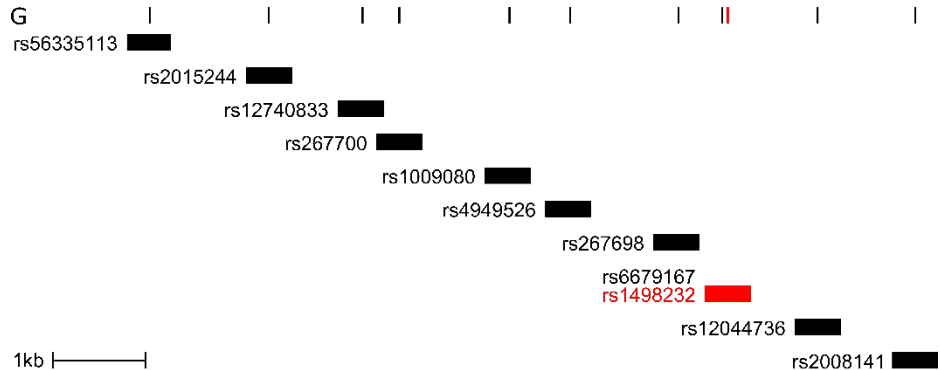
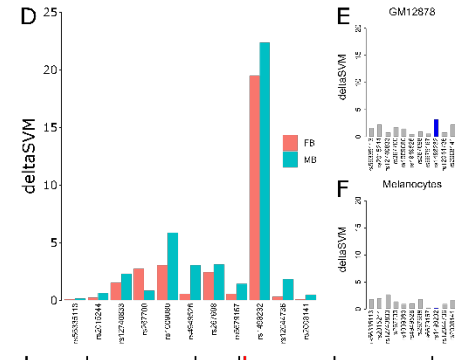
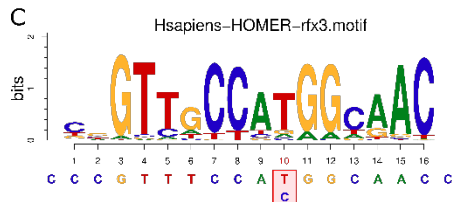


Figure 3.3: Investigating the top variants falling in midbrain open chromatin regions for their effect on enhancer activity

(A) The distribution of gkm-SVM scores for all disease-associated variants that overlap an open chromatin region in the midbrain. (B) Summary of the deltaSVM scores for the top eleven variants that fall in midbrain open chromatin regions. (C) Schematic of the variants, the open chromatin regions and genes that they overlap. (D) None of the tested variants, either as the reference or alternate allele, exceed background levels of reporter activity (“EMPTY”). (E) Even the enhancer at *SNCA*, containing either the reference sequence (non-risk) or the haplotype containing the two risk variants, fails to direct robust reporter activity in the SK-N-SH cell line.

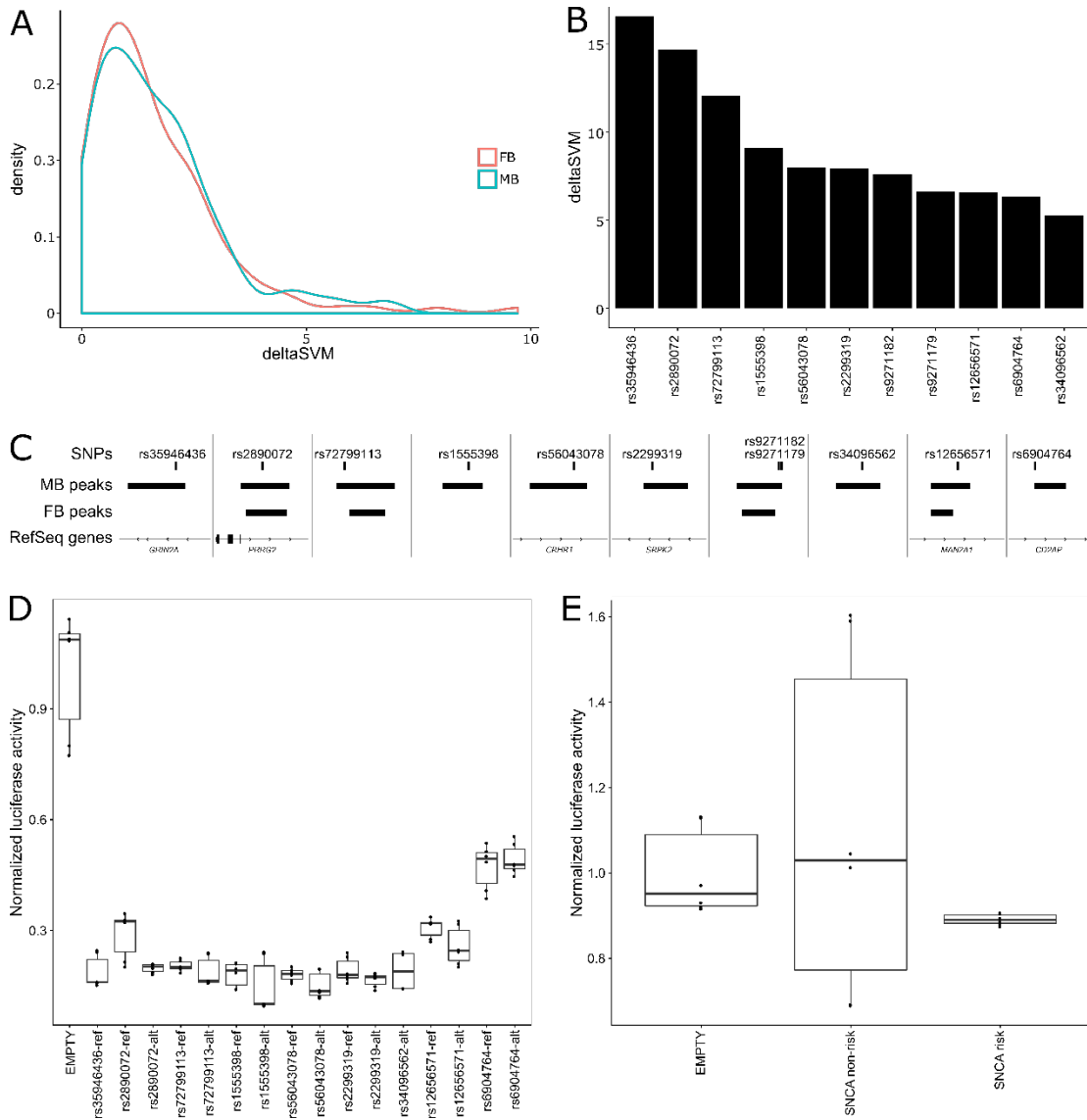


Figure 3.4: Identification of proteins whose binding is impacted by the implicated PD-risk SNPs

(A, B) MA plots for both rs2737024 and rs2583959 indicating the magnitude of the effect of the minor and major allele on binding. Cut-off for differential binding: $\log_2(\text{major}/\text{minor}) \geq 1.5$ or ≤ -1.5 . (A) NOVA1 and APOBEC3C (green circles) bind at rs2737024 with greater affinity for the major allele, while PEG10 and SNRPA (red circles) have a greater affinity for the minor allele. (B) CHMP5 (red circle) has a greater affinity for the minor allele of rs2583959. (C) Representative images of the protein binding for each of the differentially bound proteins. (D) Expression analysis in the MB and FB DA neurons for each of the differentially bound proteins indicate *Nova1*, *Peg10*, *Snrpa*, and *Chmp5* to be highly expressed in these populations, while none of the *Apobec* family member genes are expressed (RPKMs ≤ 1 , data not shown). Red bar is the mean expression of the four replicates (black dots).

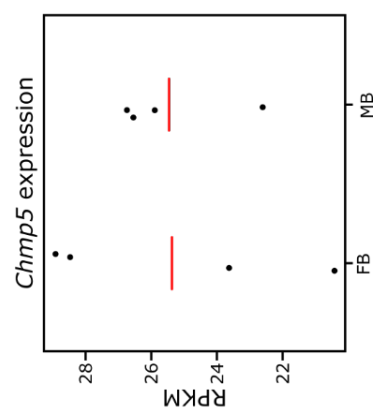
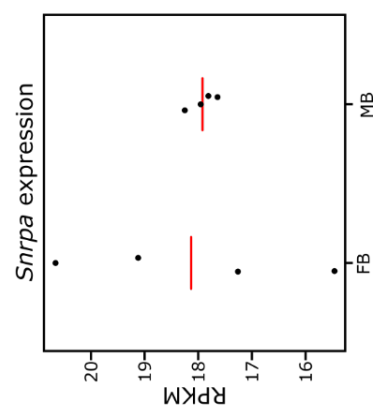
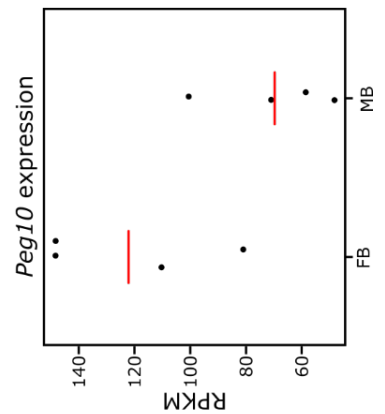
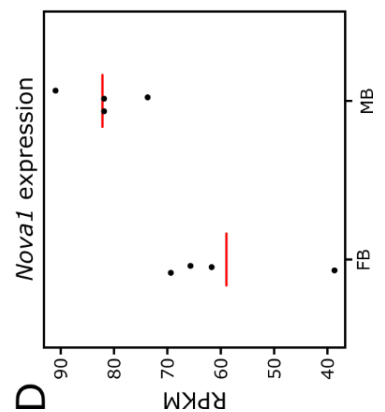
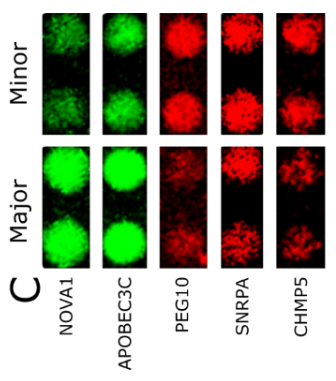
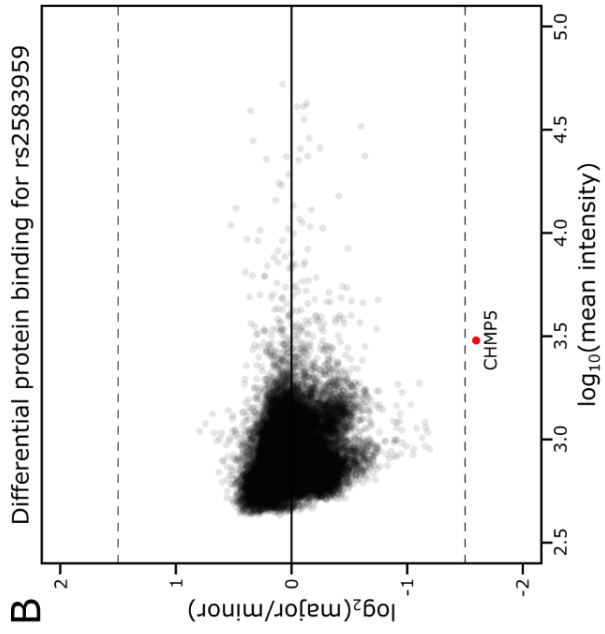
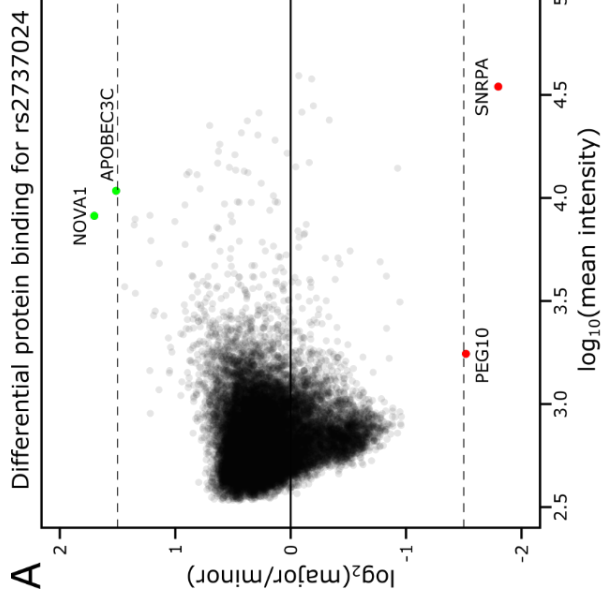
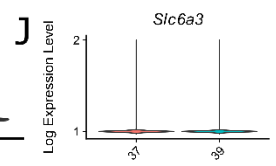
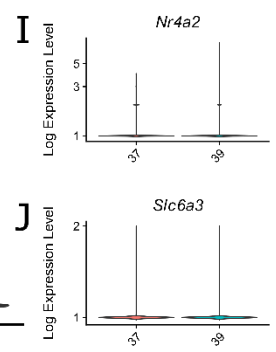
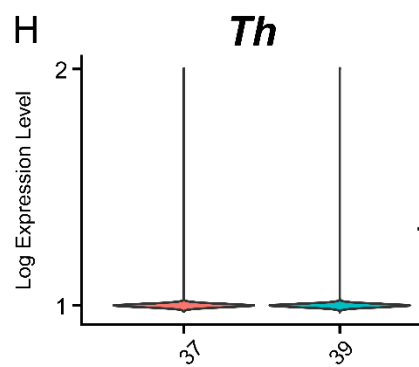
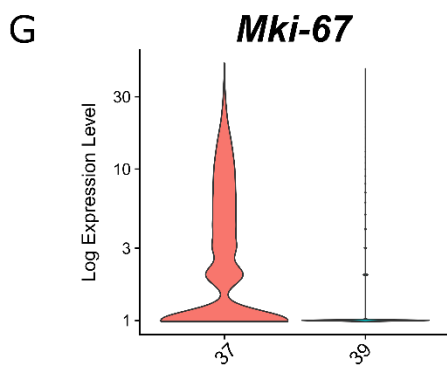
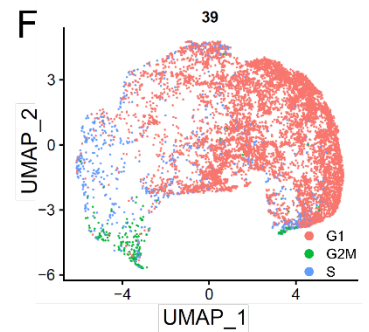
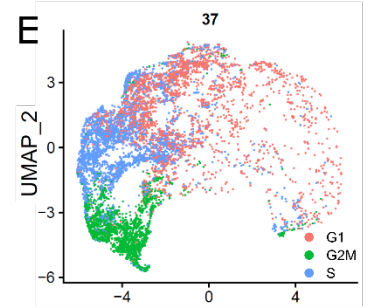
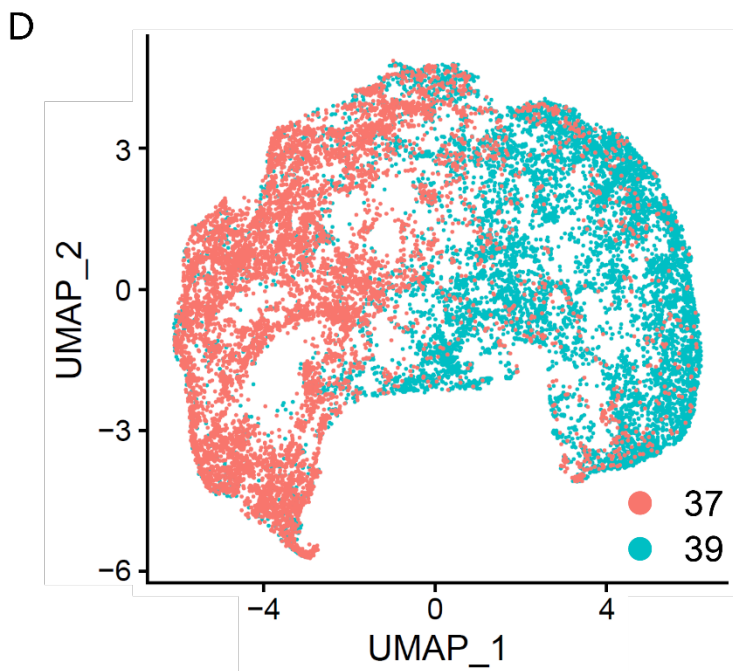
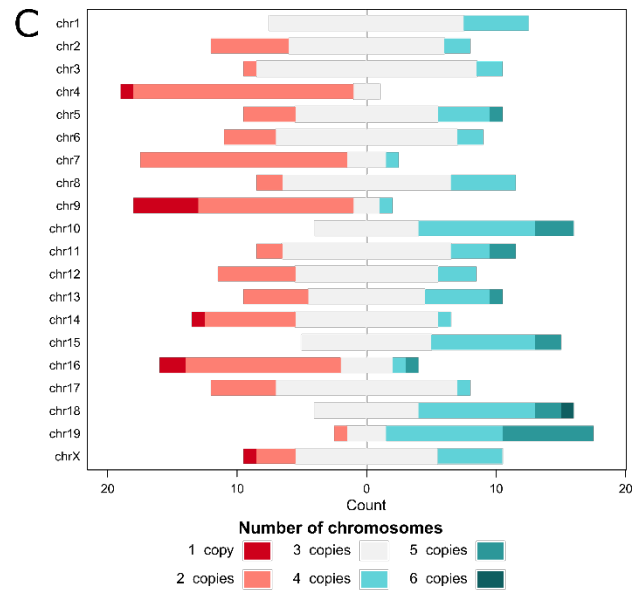
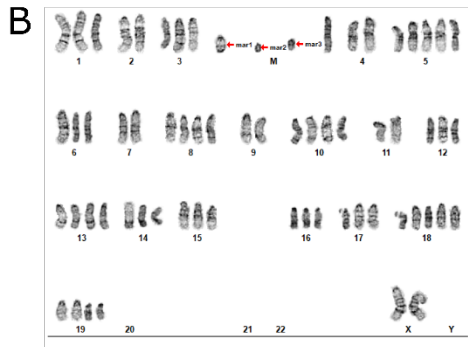
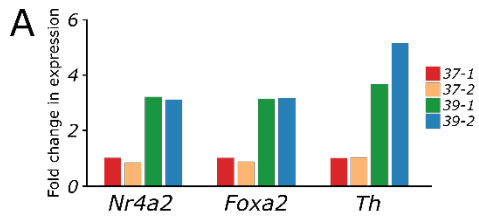
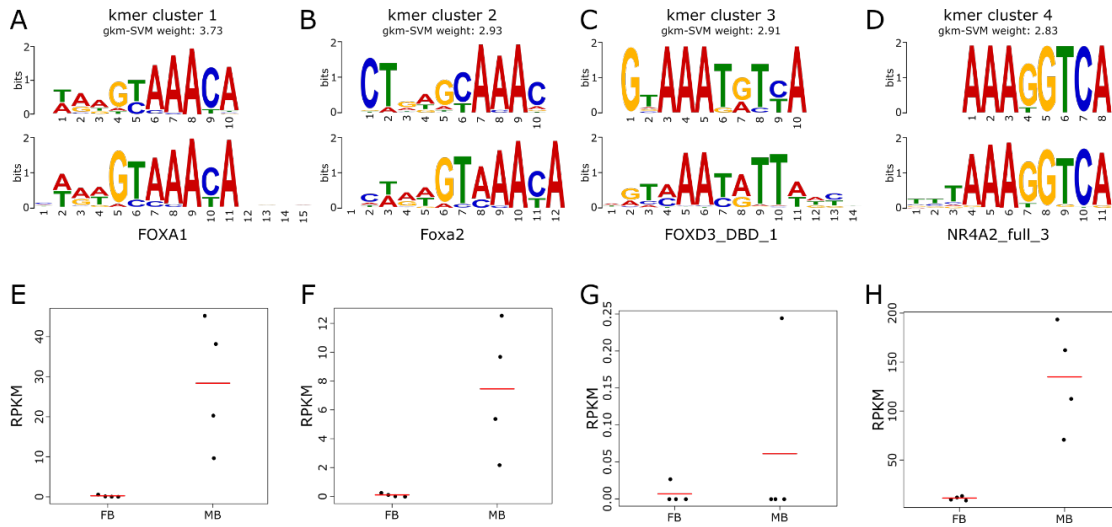


Figure 3.5: Initial characterization of the SN4741 cell line suggests it is an unstable, triploid cell line that express markers of differentiation but not of dopaminergic neurons

(A) RT-qPCR indicates the cells, when shifted a non-permissive temperature, begin to express markers of dopaminergic neurons at high levels. (B) A representative karyogram of SN4741 cells, indicating the likely triploid nature of the cells. Also shown are the marker chromosomes. (C) Summary of the karyotypes of 20 SN4741 cells demonstrating the instability of aneuploidy in these cells. (D) scRNA-seq at the permissive and non-permissive temperature indicates the cells at each temperature are transcriptionally distinct. (E, F) Shifting the cells to the non-permissive temperature is accompanied by a shift in cell cycle stage from G2M and S phase to primarily G1 phase. (G) A marker of proliferation, *Mki-67*, is only expressed at the permissive temperature, corroborating the cell cycle analysis shift. (H) There is no detected expression of tyrosine hydroxylase, *Th*, at either temperature. (I, J) Other markers of DA neurons, *Nr4a2* (I) and *Slc6a3* (J) are also not detected, calling into question the dopaminergic nature of these cells.





Supplementary Figure 3.1: Motif analysis identifies transcription factors (TFs) important specifically for MB regulatory potential

(A-D) The motifs with the greatest regulatory potential specific to the MB and the potential TF matching that motif were identified. (E-H) Expression analysis of these identified TFs confirm the sequence based analysis for Foxa1 (E), Foxa2 (F), and Nr4a2 (H). Foxd3 (G), while prioritized on the basis of sequence composition, is not expressed in MB or FB DA neurons (≤ 1 PKM) and was likely identified as a consequence of the sequence degeneracy within TF families.

Supplementary Table 3.1: Summarizing the disease-associated variants collected for scoring by deltaSVM

Disorder	# of lead SNPs	# of LD friend SNPs	PMID
Alzheimer disease	19	916	24162737
Epilepsy	4	317	25087078
Parkinson disease	26	3,700	25064009
Progressive supranuclear palsy	7	3,304	21685912
Schizophrenia	98	3,785	25056061

Supplementary Table 3.2: Primer sequences used to clone the constructs for luciferase assay (bold: BP Gateway arms)

	Variant	Primer sequence	Product size
Top variant scored by deltaSVM plus those in high LD	rs56335113	F GGGGACAAGTTTGTACAAAAAGCAGGCTAGTGATCATAGGGATTCTGAGCCC	471
		R GGGGACCACTTTGTACAAGAAAGCTGGGTGAGGGAGTTTACCCTGAAGGAG	
	rs2015244	F GGGGACAAGTTTGTACAAAAAGCAGGCTTCCCACCAGCTCTTCTGTTC	552
		R GGGGACCACTTTGTACAAGAAAGCTGGGTGGATTTCAGTGTGAGGGATGG	
	rs12740833	F GGGGACAAGTTTGTACAAAAAGCAGGCTGGAGGAAGCAGGAAGTCTCAC	532
		R GGGGACCACTTTGTACAAGAAAGCTGGGTCCATCTGTCTCCGGTTTCAT	
	rs267700	F GGGGACAAGTTTGTACAAAAAGCAGGCTCCTGGAACACCTAGATTCTTGAG	471
		R GGGGACCACTTTGTACAAGAAAGCTGGGTCTGAAGGCTGGTTTGTTCATTC	
	rs1009080	F GGGGACAAGTTTGTACAAAAAGCAGGCTGAGCTCCACGTGTTCTTGGT	506
		R GGGGACCACTTTGTACAAGAAAGCTGGGTCCAAATGCTGGAACGCTGAG	
rs4949526	F GGGGACAAGTTTGTACAAAAAGCAGGCTCCCAAGTGCCTGATGGTTTA	578	
	R GGGGACCACTTTGTACAAGAAAGCTGGGTGCACTAGAAGCAGGGCCATT		
rs267698	F GGGGACAAGTTTGTACAAAAAGCAGGCTGGGAGGGACCATTCCCAAAG	474	
	R GGGGACCACTTTGTACAAGAAAGCTGGGTAGGACTGAGCTTGTGATGGC		
rs1498232	F GGGGACAAGTTTGTACAAAAAGCAGGCTGCGGGACCTTGGTAGTGAA	494	
	R GGGGACCACTTTGTACAAGAAAGCTGGGTACGCCTCCCTGATTTTCGTTA		
rs12044736	F GGGGACAAGTTTGTACAAAAAGCAGGCTTCACTCTGTCATGCCCTACT	472	
	R GGGGACCACTTTGTACAAGAAAGCTGGGTACCAGCTGATTATGCAGTGTG		
rs2008141	F GGGGACAAGTTTGTACAAAAAGCAGGCTTAATCAGCCACCACTGTCTGC	628	
	R GGGGACCACTTTGTACAAGAAAGCTGGGTGACATGGCTAGAGGACATGC		
Top variants by deltaSVM in MB open chromatin regions	rs35946436	F GGGGACAAGTTTGTACAAAAAGCAGGCTGGGCTGGGGAAGGTTAAACA	517
		R GGGGACCACTTTGTACAAGAAAGCTGGGTGTTACACACTCAGCCCCACA	
	rs2890072	F GGGGACAAGTTTGTACAAAAAGCAGGCTAGGAGTTCCTTCCCTTAAGAC	702
		R GGGGACCACTTTGTACAAGAAAGCTGGGTCCCTTCCCTAGGACACAGGACTT	
	rs72799113	F GGGGACAAGTTTGTACAAAAAGCAGGCTATGAGGTCAGCAGATGTCAGAA	587
		R GGGGACCACTTTGTACAAGAAAGCTGGGTCAAACATCCTCGTTTCGCTTGT	
	rs1555398	F GGGGACAAGTTTGTACAAAAAGCAGGCTAATGGGCTGGAAAGGCTTCG	614
		R GGGGACCACTTTGTACAAGAAAGCTGGGTCCCTGTGCACCCTGATCTT	
	rs56043078	F GGGGACAAGTTTGTACAAAAAGCAGGCTTGAAGGAGGGACTCGGGAAG	655
		R GGGGACCACTTTGTACAAGAAAGCTGGGTAACCGTTTTCTGTGGAACCC	
	rs2299319	F GGGGACAAGTTTGTACAAAAAGCAGGCTTAGCTTCTGACCAAACCTGTGC	561
		R GGGGACCACTTTGTACAAGAAAGCTGGGTCCCTATCCCATTCCGGTCTTT	
	rs9271182/ rs9271179	F GGGGACAAGTTTGTACAAAAAGCAGGCTGTACTTAGTATTGGTGGGGAGAA	488
		R GGGGACCACTTTGTACAAGAAAGCTGGGTGGGCTTTGAATTTAGGCAGAAC	
rs34096562	F GGGGACAAGTTTGTACAAAAAGCAGGCTAAATTCCCCCTAAACAAACACCG	632	
	R GGGGACCACTTTGTACAAGAAAGCTGGGTACAGTCTCACTGTGGGAAAAAT		
rs12656571	F GGGGACAAGTTTGTACAAAAAGCAGGCTAGGCAGCTGAAACATTAGCCT	513	
	R GGGGACCACTTTGTACAAGAAAGCTGGGTAGAAAACCCAAGATGCACATTAGT		
rs6904764	F GGGGACAAGTTTGTACAAAAAGCAGGCTGGGGAAACCTCTACTTTTGGAT	529	
	R GGGGACCACTTTGTACAAGAAAGCTGGGTCCAGTGTCTTACGAAAGGC		

Supplementary Table 3.3: Primer sequences used for protein binding assays

rs2737024-maj	acatcacattgtcctAttacattccttgcccaACCCTATAGTGAGTGCTATTA
rs2737024-min	acatcacattgtcctGttacattccttgcccaACCCTATAGTGAGTGCTATTA
rs2583959-maj	ctttgttaataaatcCttgtataaaccccaACCCTATAGTGAGTGCTATT
rs2583959-min	ctttgttaataaatcGttgtataaaccccaACCCTATAGTGAGTGCTATT

Supplementary Table 3.4: RT-qPCR primers for testing expression of dopaminergic neuron markers

Marker	Forward	Reverse
<i>Th</i>	CTGTCCACGTCCCCAAGGTTCA	CAATGGGTTCCTCCAGGTTCCG
<i>Actb</i>	TGGCTCCTAGCACCATGAAG	AGCTCAGTAACAGTCCGCCTA
<i>Foxa2</i>	CCCTACGCCAAATGAACTCG	GTTCTGCCGGTAGAAAGGGA
<i>Nr4a2</i>	GTGTTCAGGCGCAGTATGG	TGGCAGTAATTTTCAGTGTGGT

Chapter 4: The developmental origins of Parkinson disease

4.1 Investigating the pathogenesis of Parkinson disease by single-cell RNA-seq

Parkinson disease (PD) is the second most common neurodegenerative disease and is characterized by progressive motor phenotypes, including resting tremor, bradykinesia, rigidity, and postural instability². These symptoms are largely the effect of substantial degeneration of midbrain dopaminergic (DA) neurons, specifically those in the substantia nigra⁴. This preferential degeneration and the appearance of protein aggregates of α -synuclein, in the form of Lewy bodies, are the pathological hallmarks of PD. How these protein aggregates result in cell death and why these processes might preferentially affect the DA neurons of the substantia nigra is not well understood.

PD can also present with a variety of non-motor phenotypes, including anosmia, constipation, and REM sleep disorders, which often precede the motor phenotypes by years to decades^{3,262}. This prodromal phase of PD suggests that neurodegeneration is occurring long before the clinical presentation of PD. Supporting this, by the time that movement phenotypes become apparent and a clinical diagnosis is made, up to 40-60% of DA neurons of the substantia nigra have already been lost⁴. It is therefore important to examine early time points, before the neurodegeneration has become apparent, in order to elucidate the mechanisms that lead to the preferential loss of nigral DA neurons.

Using a mouse model of PD, we set out to characterize populations of midbrain DA neurons at an early post-natal time point and identify subpopulations, genes, and pathways disrupted in PD. First, to examine the early development of DA neuron cell type subpopulations, we performed single-cell RNA-seq (scRNA-seq) on populations of wildtype DA neurons at both embryonic day E15.5 and post-natal day P7. This analysis defined the heterogeneity of DA populations over developmental time in the brain, revealing the development of distinct populations of mature DA neurons by P7. Next, at this early post-natal time point, we compared midbrain DA neurons in a mouse model of PD and identify differences. We find that, even this early in development, there are significant alterations in neural differentiation, gene expression, and mitochondrial function. Finally, we present a model to unify these observations, suggesting PD is a disorder that begins to manifest during early neuron development and maturation.

4.2 Characterizing E15.5 and P7 dopaminergic neurons in wildtype mice⁴

In order to characterize DA neurons, we performed scRNA-seq on cells isolated from distinct anatomical locations of the mouse brain over developmental time. We used fluorescence-activated cell sorting (FACS) to retrieve single DA neurons from the Tg(Th-EGFP)DJ76Gsat BAC transgenic mouse line (Th:eGFP), which expresses EGFP under the control of the tyrosine hydroxylase locus¹⁷¹. We microdissected both midbrain (MB) and forebrain (FB) from E15.5 mice, extending our analyses to MB,

⁴ This section and associated methods have been published in the American Journal of Human Genetics and adapted for use in this thesis. Hook, P. W. *et al.* Single-Cell RNA-Seq of Mouse Dopaminergic Neurons Informs Candidate Gene Selection for Sporadic Parkinson Disease. *Am. J. Hum. Genet.* 102, 427–446 (2018)¹⁵¹.

FB, and olfactory bulb (OB) in P7 mice (**Figure 4.1A**). Brains from four and five mice were pooled for E15.5 and P7, respectively. E15.5 and P7 time points were chosen based on their representation of stable MB DA populations, either after neuron birth (E15.5) or between periods of programmed cell death (P7)²⁶³ (**Figure 4.1A**).

To assess the transcriptomes of these tissues and time points, we sequenced ~95 cells from each, using the Smart-Seq2 protocol¹³⁸, for a total of 473 single cells sequenced to an average depth of $\sim 8 \times 10^5$ 50-bp paired-end fragments per cell. We performed standard quality control and filtered cells for total mass, total number of mRNAs, and total number of expressed genes per cell. We then used principle component analysis (PCA) in order to identify and remove cells representing oligodendrocytes and other support cells. After this filtering, 396 cells remained for downstream analysis.

With this data, we set out to identify clusters of single cells within time points and anatomical regions. First, to compare embryonic and post-natal DA neuron development, we analysed all cells collected. Doing so, we find that E15.5 cells from both MB and FB largely cluster together (**Figure 4.1B**), supporting the notion that they are less differentiated. By contrast, cells isolated at P7 mostly cluster by anatomical region, suggesting progressive functional divergence with time (**Figure 4.1B**). We performed recursive analysis of each of these clusters to identify subpopulations therein. This revealed a total of 13 clusters (E15.5 FB1-2, MB1-2; P7 OB1-3, FB1-2, MB1-4; **Figure 4.1C**), demonstrating the diversity of DA neuron subtypes and their progressive differentiation across developmental time.

With subpopulations of DA neurons defined by our data, we set out to assign a biological identity to each cluster. To do this, we identified differentially expressed genes between clusters within each time point, then identified marker genes for each cluster within each time point (summarized in **Table 4.1**). Among the four clusters identified at E15.5, two were represented in t-SNE space as a single large group that included cells from both MB and FB (E15.MB1, E15.FB1), leaving two smaller clusters that were comprised solely of MB or FB cells (**Figure 4.1D**). Both E15.MB1 and E15.FB1 show markers consistent with neuroblast populations^{264–272}. The isolated MB cluster (E15.MB2) specifically expressed *Foxa1*, *Lmx1a*, *Pitx3*, and *Nr4a2* and thus likely represents a post-mitotic DA neuron population²⁷³. Similarly, the discrete E15.FB2 cluster expressed markers of post-mitotic FB/hypothalamic neurons, including *Six3*, *Six3os1*, *Sst*, and *Npy*^{274–278} (**Figure 4.1E**). These embryonic data did not discriminate between cells populating known domains of DA neurons, such as the substantia nigra (SN) or ventral tegmental area (VTA).

By contrast, P7 cells mostly cluster by anatomical region and each region has defined subsets (**Figure 4.1F**). Focusing specifically on the midbrain DA neurons, we identified four P7 MB DA subset clusters (**Figure 4.1G**). Marker gene analysis confirmed that three of the clusters correspond to DA neurons from the VTA (*Otx2* and *Neurod6*^{279,280}; P7.MB1), the periaqueductal gray area (PAG; *Vip* and *Pnoc*^{149,281–283}; P7.MB3), and the SN (*Sox6*, *Aldh1a7*, *Ndnf*, *Serpine2*, *Rbp4*, and *Fgf20*^{149,279,284,285}; P7.MB4). These data are consistent with recent scRNA-seq studies of similar populations^{149,283}. The only cluster without a readily assigned identity was P7.MB2. This population of P7 MB DA neurons is likely a neuroblast-like population based on marker gene analysis^{149,286–290}. Like the overlapping E15.MB1 and E15.FB1 clusters,

this cluster preferentially expresses markers of neuronal precursors, differentiation, or maturation. Additionally, this cluster exhibits gene expression consistent with embryonic mouse neuroblast populations¹⁴⁹ as well as cell division and neuron development^{286–290}.

These results indicate a unique opportunity; at post-natal day 7 there are DA neurons that are mature and present in their ultimate anatomical locations, but there also remains a group of still developing neuroblasts. Further investigation at this time point would allow us to compare the impact of PD mutations on the mature DA neurons and structures that persist through to adulthood, but also consider how the PD mutation might impact DA neural development.

4.3 Characterizing P7 dopaminergic neurons in a mouse model of Parkinson disease

To assess the early developmental origins of PD, we performed scRNA-seq on midbrain DA neurons in a mouse model of PD. The A53T mouse model (B6;C3-Tg(Prnp-SNCA*A53T)83Vle/J; JAX: 004479)²⁹¹ expresses mutant human *SNCA* under the direction of the mouse prion protein promoter. This mouse model recapitulates human PD phenotypes, like progressive movement impairments and α -synuclein aggregates) and demonstrates an allele-dosage effect, with the age of onset earlier for homozygotes (14-15 months) than for hemizygotes (22-28 months)²⁹¹. We bred this allele onto the Th:eGFP background to allow retrieval of DA neurons. On a wildtype, hemizygous, and homozygous A53T background (***Supplementary Figure 4.1***), we microdissected the midbrain from two litters of P7 pups and collected 13,518 single cells by FACS for scRNA-seq.

Cells from each genotype were processed with the 10X Genomics pipeline for cDNA generation and library preparation. Following sequencing, it became apparent that our cells had not made it through this process completely intact and exhibited high levels of ambient RNA. As a consequence of either FACS, the 10X platform, or a combination of subjecting our neurons to both, our libraries exhibited a low fraction of reads in cells (~30%). Following sequencing, we captured 2,430 cells, with ~170,000 reads per cell. We further filtered these cells in order to remove empty droplets, doublets and multiplets, and stressed/dying cells by considering the number of unique genes, read depth, and percent of reads mapping to the mitochondria. Following filtering, 1,357 cells remained for analysis.

We performed clustering analysis on these cells to identify subpopulations present in our samples. Doing so, we identify seven clusters (**Figure 4.2A**). Preliminary expression analysis suggested two of these clusters (clusters 5 and 6) display little evidence of *Th* or eGFP expression (**Figure 4.2B**, **Supplementary Figure 4.7**) and likely consist of contaminating cell types, with cluster 5 expressing high levels of oligodendrocyte and other support cell markers^{292–297} (**Figure 4.2C**, **Supplementary Figure 4.6**). Marker analysis of this contaminating cell cluster identifies genes enriched in this cluster (eg: *Zic1*, *Nfib*, *Id2*, *Tcf4*; **Figure 4.2D**)^{298–303}, which in consultation with ISH slides from the Allen Mouse Brain Atlas³⁰⁴, suggest these are granule cells, likely from either the hippocampus or cerebellum (**Supplementary Figure 4.8**). Consequently, this cluster and the contaminating oligodendrocyte cluster were removed from further analysis.

Considering only the remaining dopaminergic neurons, we sought to assign each cluster to a biological identity or anatomical region. We performed differential

gene expression analysis between clusters to identify marker genes for each cluster (**Figure 4.2E-I, Table 4.2**). Comparing these markers with marker genes established from the literature^{149,279–287} and our work on wildtype P7 DA neurons¹⁵¹, we were able to assign a cell type to each cluster, identifying clusters corresponding to: the VTA, a post-natal neuroblast population, two SN clusters (SN1 and SN2), and the PAG (**Figure 4.3A, Supplementary Figure 4.9**). We have begun to confirm these cluster assignments with single molecular RNA fluorescence *in situ* hybridization (smFISH) on brain sections from wildtype P7 mice, using the genes indicated in **Figure 4.3B-I**.

Clusters 0, 2, and 4 – tentatively the VTA, SN1 and SN2 – are transcriptionally related to each other, sharing high expression of many marker genes (eg: *Aldh1a1* and *Ddc*, **Figure 4.4A**; extended in **Supplementary Figure 4.9**). Clusters 2 and 4 especially, both assigned to the SN, are highly similar in their expression patterns. We had not previously observed a separation of SN DA neurons in our study of wildtype P7 MB DA neurons, perhaps due to the relatively few cells sequenced (n = 80 MB neurons). With a combined 337 cells assigned to the SN clusters here, resolving nuanced differences becomes possible. Here, we observe cluster 2 (SN1) to exhibit higher expression of DA neuron markers, like *Th* and *Slc6a3* (**Figure 4.4A, Supplementary Figure 4.9**), while cluster 4 (SN2) has restricted expression of DA neuron markers, like *Pitx3* and *Foxa1* (**Supplementary Figure 4.9**)^{175,177}. We will use smFISH against the genes indicated in **Figure 4.4B-D** to confirm these subtle transcriptional differences between clusters 0, 2, and 4, paying particular focus to differences in spatial distribution of the two SN clusters.

4.4 The Parkinson disease mutation alters cell maturation, gene expression and mitochondrial dynamics

Ultimately, our goal was to examine the effects of the A53T PD mutation on the transcriptomes of developing MB DA neurons. Towards this, first, we confirmed that the transgene is expressed at P7. Given the transgene incorporates the mouse *Prnp* 3' and 5' UTRs²⁹¹ and the 3' sequencing method employed by 10X Genomics¹³⁷, reads belonging to the endogenous *Prnp* gene and the A53T transgene are indistinguishable. Thus, if the mutant transgene is active, we expect to see an apparent increase in *Prnp* expression in hemizygote and homozygote cells over wildtype. Indeed, we observe an increase in the number of cells with detectable *Prnp* expression between homozygote cells (79.2%) and wildtype cells (40.2%), with homozygotic cells expressing significantly higher levels (~3.3 fold increase) of *Prnp* (**Figure 4.5A**).

Next, we examined how the presence of the A53T mutation impacts the distribution of cells among the subpopulations (**Figure 4.5B**), particularly focusing on alterations to the SN clusters, given the preferential degeneration of these neurons in PD. We observe approximately double the number of mutant cells in the two SN clusters and interestingly, given the physical interspersed nature of the post-natal neuroblast population with the SN¹⁵¹, we see approximately half as many neuroblast cells present in homozygous mutant cells. This alteration could indicate a precocious differentiation phenotype, with mutant neurons exiting the neuroblast stage and maturing into SN cells, a substantial difference to observe as early as P7. Additionally, we observe no apparent changes to the VTA cluster proportions by genotype, but do see an unexpected halving of cells in the PAG in homozygotes compared to wildtypes.

Quantifying cell proportion differences from scRNA-seq data is fraught with false positives, so we will validate these cell population differences by smFISH and quantify the number of cells in each cluster, comparing the three genotypes.

Finally, we sought to identify transcriptional differences between the genotypes resulting from the mutation on a cluster-specific and global basis. The two SN clusters had no significantly differentially expressed genes by genotype, however the neuroblast population had four genes downregulated, all mitochondrial (*mt-Co3*, *mt-Nd1*, *mt-Atp6*, and *mt-Cytb*; **Figure 4.6A**), and a single gene, *Rtn3*, upregulated ($\ln\text{FC} = 0.88$, adjusted p-value = 0.016; **Figure 4.6A, Table 4.3**). *Reticulon-3*, *Rtn3*, is involved in the ER/secretory pathway and is likely involved in ER stress-mediated apoptosis via caspase signaling to the mitochondria. It has previously been implicated in Alzheimer disease, and it has been demonstrated that knockdown of *RTN3* increases β -amyloid levels while overexpression decreases β -amyloid levels³⁰⁵.

The VTA, while appearing unaffected by the mutation in terms of cellular proportions, has 232 genes dysregulated, 27 down- and 185 upregulated (**Table 4.3**). Of the top genes upregulated in the mutant condition, we find many genes that have previously been implicated in PD or other neurodegenerative disorders; for example, the top upregulated gene is *Cck* ($\ln\text{FC} = 1.931$, adjusted p-value = 3.74×10^{-3} ; **Figure 4.6B**), whose expression has been associated with the visual hallucinations that sometimes co-occur with PD^{306,307}. *Resp18* is also significantly upregulated ($\ln\text{FC} = 1.203$, adjusted p-value = 8.43×10^{-6} ; **Figure 4.6B**), and its expression has previously been associated with PD – downregulation of *RESP18* protected DA neurons against damage by MPTP, a chemical used in the modeling of PD, and its overexpression aggravated cell death in similar models^{308,309}. Of the 27 genes downregulated, one,

Wsb1 ($\lnFC = -0.857$, adjusted p-value = 1.89×10^{-4} ; **Figure 4.6B**) is present in Lewy bodies in *LRRK2*-associated PD and is involved in the ubiquitination and aggregation of *LRRK2*³¹⁰. Another interesting downregulated gene is *Gria2* ($\lnFC = -0.492$, adjusted p-value = 1.53×10^{-4} ; **Figure 4.6B**), a subunit of an AMPA receptor, whose expression is associated with ALS, neurodevelopmental disorders, and addiction behaviours^{311,312}. There are also 11 genes involved in the mitochondria and the electron transport chain that are significantly downregulated, which includes those genes that were downregulated in the neuroblast populations.

To assess whether this mitochondrial dysregulation is a common phenomenon in the mutant cells and to increase our power to detect differences between the genotypes, we explored global transcriptomic differences by genotype, regardless of cluster. Doing so, we identify 388 dysregulated genes, with 91 downregulated genes and 297 genes that are upregulated (**Table 4.3**). Pathway analysis³¹³ of these genes identifies the oxidative phosphorylation pathway as dysregulated with 37 differentially expressed genes involved in this process. Interestingly, 12 of the dysregulated genes in the pathway are downregulated while the other 25 genes are upregulated (**Figure 4.7A**) and strikingly, whether the gene is encoded on the mitochondrial or nuclear genome correlates with the direction of dysregulation. Genes that are encoded on the mitochondrial genome are downregulated and those that are encoded on the nuclear genome are upregulated in the mutant cells. To our knowledge, this pattern of expression has not been observed in PD models before. Opposite to what we observe here, significant downregulation of nuclear-encoded genes, but not mitochondrial-encoded genes, has previously been observed in Alzheimer disease³¹⁴. We will confirm these gene expression changes and those that are cluster-specific

through smFISH and will further examine mitochondrial function in the mutant mice. We will assess the mtDNA copy number to look for biogenesis defects, and use MitoTracker assays and cryoEM to examine mitochondrial size, number, location, and assess gross morphological characteristics.

4.5 Discussion

Understanding the timing and predisposing causes of the degeneration of DA neurons of the substantia nigra remains challenging. Examination of developing or early post-natal neurons could illuminate processes occurring far before the neurodegeneration becomes apparent. Understanding these developmental origins of PD could yield new insights into PD pathogenesis and help inform early interventions for the management or prevention of PD. To this end, we assayed wildtype and mutant DA neurons in a mouse model of PD for transcriptional differences occurring at post-natal day 7. Despite this early developmental stage, we identified a variety of subpopulations of midbrain DA neurons, discovered a potential maturation/differentiation defect, and uncovered extensive gene expression alterations implicating mitochondrial dysfunction.

Our strategy focused on assaying specifically the DA neurons of the midbrain, given their known role in the movement phenotypes of PD, in a mouse model of familial PD using FACS. However, it is not just the DA neurons of the substantia nigra that are degenerated in PD – many of the prodromal symptoms of PD are likely the result of degeneration of other classes of neurons (eg: serotonergic, GABAergic, glutamatergic) in structures throughout the brain^{315–318}. Additionally, the degeneration of substantia nigra neurons in PD may not be a cell-autonomous process;

LRRK2, the most commonly mutated gene in PD, is expressed in microglia as well as DA neurons³¹⁹⁻³²¹. Thus, to get a more rich view of the cells that may be dysregulated in PD, performing scRNA-seq in unsorted, gross-dissected regions of the brain could yield greater insights into the mechanisms by which PD-predisposing mutations exert their effects on disease pathogenesis. This strategy would also eliminate the need to subject the cells to FACS, which could reduce the ambient RNA background that our study design suffered from, improving the yield of cells assessed.

In examining the proportions of cells in each cluster stratified by genotype, we observed an approximate doubling of cells present in the PAG and the two SN clusters, and a halving of cells present in the post-natal neuroblast cluster. It is interesting to observe an increase in SN neurons in the mutant mice, given PD is a neurodegenerative disease and by the time of presentation, patients have fewer DA neurons in the SN. In the development of DA neurons in the mouse, there still remains a wave of programmed cell death occurring between P10 and P30²⁶³. It may be that this precocious differentiation is depleting the neuroblast pool for recovery following this wave of cell death. We will be testing this hypothesis by inspecting the various cell proportions during and following this programmed cell death, and comparing the genotypes for alterations in subpopulation proportions.

Given the observed dysregulation of cellular proportions, we expected there to be gross changes in gene expression by genotype in a cluster specific- and global manner. For example, we see specific upregulation of *Cck* in mutant DA neurons in the VTA cluster and altered expression of a variety of other genes including several others already known to PD biology like *Resp18* and *Wsb1*. Despite identifying no differentially expressed genes in the SN clusters, we do observe a handful of genes

that are up- and downregulated in the precursor neuroblast population, including several mitochondrial genes. This dysregulation is reflected in the mutant cells as a whole: there is global dysregulation of genes involved in the oxidative phosphorylation pathway.

Mitochondrial dysfunction has been extensively associated with sporadic, familial and environmentally-induced PD³²²⁻³²⁹. A long standing hypothesis implicates aberrant levels of reactive oxygen species with PD, focusing on the high energy demands of the substantia nigra as the source of the preferential degeneration of these neurons. In particular, disruptions to complex I of the electron transport chain have been associated with PD^{325,330,331}. Here, at a very early developmental age, we observe a striking pattern of dysregulation of electron transport chain associated genes, involving all five complexes, in which the encoding genome determines the direction of dysregulation. This is novel, unusual expression pattern that could suggest a disconnect in communication between the mitochondria and the cell. One mechanism we are particularly interested in is mitochondrial fission and fusion. In a variety of models, α -synuclein over-expression has recently been demonstrated to produce fragmented mitochondria through disrupted fission and fusion processes³³²⁻³³⁵. Importantly, in the A53T mouse model of PD that we use here, fusion and fission defects have been previously observed, but only in late stages of the disease (i.e.: not detected at 6 months but present at 12 months)³³⁶.

A mitochondrial fission or fusion problem during DA neuron development leads us to a particularly compelling hypothesis to unify our observations of mitochondrial dysregulation and precocious/aberrant maturation of neuroblasts into substantia nigra neurons. In the process of neuronal differentiation, there is a switch from

glycolysis to oxidative phosphorylation for energy production and this shift is accompanied by changes in mitochondrial morphology, dependent on fission and fusion (**Figure 4.8**)^{337–339}. Our observations would then suggest that mutant DA neurons early in development suffer from defects in mitochondrial dynamics that ultimately result in the precocious maturation of neuroblasts into SN neurons. We are actively assaying mitochondria and their dynamics in mutant DA neurons to pursue this hypothesis. If this fission/fusion mechanism is at the root of the early transcriptional, cellular and maturation defects we observe, it presents a lucrative target for modulation in early therapeutics for the prevention of PD.

That any differences are observed between PD mutant mice and their wildtype littermates at just post-natal day 7 is perhaps the most surprising result of these experiments. Despite months remaining before these mice begin to develop PD-related phenotypes, there are stark differences already present in midbrain DA neurons. These neurons will persist into adulthood, at which point the differences will become apparent and manifest in movement phenotypes. We suggest here a possible developmental origin for PD, a disease classically considered as a disease of ageing.

4.6 Methods

Mouse husbandry

E15.5 and P7 wildtype mice

The Th:EGFP BAC transgenic mice (Tg(Th-EGFP)DJ76Gsat) used in this study were generated by the GENSAT Project¹⁷¹ and were purchased through the Mutant Mouse Resource & Research Centers (MMRRC) Repository. Mice were maintained on a Swiss Webster (SW) background with female SW mice obtained from Charles River Laboratories. The Tg(Th-EGFP)DJ76Gsat line was primarily maintained through matings between Th:EGFP-positive, hemizygous male mice and wild-type SW females (dams). Timed matings for cell isolation were similarly established between hemizygous male mice and wild-type SW females. The observation of a vaginal plug was defined as embryonic day 0.5 (E0.5).

Mutant A53T mice

B6;C3-Tg(Prnp-SNCA*A53T)83Vle/J mice (A53T; JAX strain: 004479)²⁹¹ were obtained from Jackson Labs. The colony was maintained with hemizygous-hemizygous matings. To generate the mouse litters for assay, a series of matings were performed. Hemizygous A53T mice were crossed with hemizygous DA neuron reporter mice, Tg(Th-EGFP)DJ76Gsat mice (Th-EGFP)¹⁷¹, and pups hemizygous for both alleles were retained. These mice were crossed with a homozygous Th-EGFP mouse, and male mice hemizygous for the A53T allele and homozygous for the Th-EGFP allele were selected. Finally, these male mice were crossed with a female A53T hemizygote, such that all pups will be hemizygous for the Th-EGFP allele and all three possible combinations of A53T alleles will be present (25% wildtype, 50% hemizygous, 25%

homozygous). This mating scheme is summarized in *Supplementary Figure 4.1*. All mouse procedures and husbandry were reviewed and approved by the institutional care and use committee.

Neural dissociation and fluorescence-activated cell sorting (FACS)

E15.5 and P7 wildtype mice

At 15.5 days after the timed mating, pregnant dams were euthanized and the entire litter of E15.5 embryos were dissected out of the mother and immediately placed in chilled Eagle's Minimum Essential Media (EMEM). Individual embryos were then decapitated and heads were placed in fresh EMEM on ice. Embryonic brains were removed and placed in Hank's Balanced Salt Solution (HBSS) without Mg^{2+} and Ca^{2+} and manipulated while on ice. For P7 samples, the morning the pups were born was considered postnatal day 0 (P0). Once the mice were aged to P7, all the mice from the litter were euthanized and the brains were then quickly dissected and placed in HBSS without Mg^{2+} and Ca^{2+} on ice. The brains were immediately observed under a fluorescent stereomicroscope and EGFP+ brains were selected. EGFP+ regions of interest in the forebrain (hypothalamus) and the midbrain for both time points, and the olfactory bulbs at P7 were then dissected and placed in HBSS on ice. This process was repeated for each EGFP+ brain. Brain regions from four EGFP+ E15.5 mouse pups and from five EGFP+ P7 mice were pooled together for dissociation.

Resected brain tissues were dissociated using papain (Papain Dissociation System, Worthington Biochemical Corporation; Cat#: LK003150) following the trehalose-enhanced protocol reported by Saxena *et al.*³⁴⁰ with the following modifications. The dissociation was carried out at 37°C in a sterile tissue culture

cabinet and RNase inhibitor was added to all solutions. During dissociation, all tissues at all time points were triturated every 10 min using a sterile Pasteur pipette. For E15.5 tissues, this was continued for no more than 40 min. For P7, this was continued for up to 1.5 hours or until the tissue appeared to be completely dissociated.

Additionally, for P7 tissues, after dissociation but before cell sorting, the cell pellets were passed through a discontinuous density gradient in order to remove cell debris that could impede cell sorting. This gradient was adapted from the Worthington Papain Dissociation System kit. Briefly, after completion of dissociation according to the Saxena protocol³⁴⁰, the final cell pellet was resuspended in DNase dilute albumin-inhibitor solution, layered on top of 5mL of albumin-inhibitor solution, and centrifuged at $70 \times g$ for 6 min. The supernatant was then removed.

For each time point-region condition, pellets were resuspended in 200 μ L of media without serum comprised of DMEM/F12 without phenol red, 5% trehalose (w/v), 25 μ M AP-V, 100 μ M kynurenic acid, and 10 μ L of 40U/ μ L RNase inhibitor (RNasin Plus RNase Inhibitor, Promega) at room temperature. The resuspended cells were then passed through a 40 μ M filter and introduced into a FACS machine (Beckman Coulter MoFlo Cell Sorter or Becton Dickinson FACSJazz). Viable cells were identified via propidium iodide staining, and individual neurons were sorted based on their fluorescence directly into lysis buffer in individual wells of 96-well plates for single-cell sequencing (2 μ L Smart-Seq2 lysis buffer + RNase inhibitor, 1 μ L oligo-dT primer, and 1 μ L dNTPs) according to Picelli *et al*¹³⁸. Blank wells were used as negative controls for each plate collected. Upon completion of a sort, the plates were briefly spun in a tabletop microcentrifuge and snap-frozen on dry ice. Single-cell lysates were subsequently kept at -80°C until cDNA conversion.

Mutant A53T mice

Neural dissociation was carried out using a modified Worthington dissociation system (Worthington Biochemical Corporation; Cat#: LK003150) in combination with a modified trehalose-enhanced dissociation protocol³⁴⁰, with minor modifications. P7 pups were decapitated and their brains extracted and placed into Earle's Balanced Salt Solution (EBSS) with trehalose individually in a 6-well plate. Midbrain GFP+ regions were microdissected under a fluorescent microscope and transferred to Worthington papain solution containing DNaseI and an RNase inhibitor. In this solution, the tissue was macerated with a scalpel and placed in a 37°C, 5% CO₂ incubator. After 15 minutes, and every subsequent ten minutes for up to an hour, using a Pasteur pipette, the tissue in papain was triturated, until the tissue has been dissociated, as confirmed under a microscope. The dissociated neurons were pooled on the basis of their common A53T genotypes and transferred to a 15mL conical tube. To each sample, 75µL EBSS containing albumin-ovomucoid inhibitor, DNaseI, RNase inhibitor, and trehalose was added, and were centrifuged for 10 minutes at 100g. The supernatant was removed and the cells were resuspended in 150uL of the same EBSS solution and mechanically dissociated with a P200 pipette. Another 100uL of EBSS solution was added and mechanical dissociation continued with a P1000 pipette. Next, 2.5mL of DMEM/F12 (without phenol red) with trehalose was added and centrifuged as before. The supernatant was discarded and another 200µL of EBSS solution is added and a final mechanical dissociation with a P200 pipette was carried out. Another 2.5mL of DMEM/F12 with trehalose is added and pelleted by centrifugation one final time.

These pellets were resuspended in 750 μ L ice-cold HBSS with calcium and magnesium and passed through a 40 μ m filter. A total of 13,518 GFP+, and therefore likely DA neurons, were sorted by FACS, with each genotype being sorted into a single well of a 96-well plate (wildtype: 5,314; hemizygous: 5,202; homozygous: 3,002), containing 10 μ L of cold PBS. Before submitting the samples to droplet formation by 10X Genomics, the total volume in each well was brought to 35 μ L with PBS.

Single-cell RNA-seq library preparation and sequencing

E15.5 and P7 wildtype mice

Library preparation and amplification of single-cell samples were performed using a modified version of the Smart-Seq2 protocol¹³⁸. Briefly, 96-well plates of single cell lysates were thawed to 4°C, heated to 72°C for 3 min, then immediately placed on ice. Template switching first-strand cDNA synthesis was performed as described above using a 5'-biotinylated TSO oligo. cDNAs were amplified using 20 cycles of KAPA HiFi PCR and 5'-biotinylated ISPCR primer. Amplified cDNA was cleaned with a 1:1 ratio of Ampure XP beads and approximately 200pg was used for a one-quarter standard sized Nextera XT tagmentation reaction. Tagmented fragments were amplified for 14 cycles and dual indexes were added to each well to uniquely label each library. Concentrations were assessed with Quant-iT PicoGreen dsDNA Reagent (Invitrogen) and samples were diluted to ~2nM and pooled. Pooled libraries were sequenced on the Illumina HiSeq 2500 platform to a target mean depth of $\sim 8.0 \times 10^5$ 50-bp paired-end fragments per cell at the Hopkins Genetics Research Core Facility.

Mutant A53T mice

Following FACS, cells were assayed following the Chromium 10X pipeline¹³⁷. Cell capture, cDNA generation, and library preparation were performed with the standard protocol for the Chromium Single Cell 3' V2 reagent kit. Library quality and concentration was assessed with the High Sensitivity DNA assay on the Agilent 2100 Bioanalyzer and the Qubit dsDNA High Sensitivity Assay (Invitrogen).

Single-cell RNA-sequencing libraries were pooled and sequenced on the Illumina HiSeq 2500 and generated ~170,000 reads per cell.

Preliminary single-cell RNA-seq data analysis

E15.5 and P7 wildtype mice

For all libraries, paired-end reads were aligned to the mouse reference genome (mm10) supplemented with the Th-EGFP+ transgene contig, using HISAT2²⁰⁵ with default parameters except: -p 8. Aligned reads from individual samples were quantified against a reference transcriptome (GENCODE vM8)²⁰⁹ supplemented with the addition of the EGFP transcript. Quantification was performed using cuffquant³⁴¹ with default parameters and the following additional arguments: --no-update-check -p 8. Normalized expression estimates across all samples were obtained using cuffnorm³⁴¹ with default parameters.

Gene-level and isoform-level FPKM (fragments per kilobase of transcript per million) values produced by cuffquant³⁴¹ and the normalized FPKM matrix from cuffnorm were used as input for the Monocle 2 single-cell RNA-seq framework³⁴² in R/Bioconductor²⁰⁷. Genes were annotated using the Gencode vM8 release. A

CellDataSet (cds) was then created using Monocle 2 (v2.2.0)³⁴² containing the gene FPKM table, gene annotations, and all available metadata for the sorted cells. All cells labeled as negative controls and empty wells were removed from the data. Relative FPKM values for each cell were converted to estimates of absolute mRNA counts per cell (RPC) using the Monocle 2 Census algorithm³⁴³ using the Monocle function “relative2abs()”. After RPCs were inferred, a new cds was created using the estimated RNA copy numbers with the expression Family set to “negbinomial.size()” and a lower detection limit of 0.1 RPC.

After expression estimates were inferred, the cds containing a total of 473 cells was run through Monocle 2's “detectGenes()” function with the minimum expression level set at 0.1 transcripts. The following filtering criteria were then imposed on the entire dataset: (1) Number of expressed genes: The number of expressed genes detected in each cell in the dataset was plotted and the high and low expressed gene thresholds were set based on observations of each distribution. Only those cells that expressed between 2,000 and 10,000 genes were retained. (2) Cell mass: Cells were then filtered based on the total mass of RNA in the cells calculated by Monocle 2. Again, the total mass of the cell was plotted and mass thresholds were set based on observations from each distribution. Only those cells with a total cell mass between 100,000 and 1,300,000 fragments mapped were retained. (3) Total RNA copies per cell: Cells were then filtered based on the total number of RNA transcripts estimated for each cell. Again, the total RNA copies per cell was plotted and RNA transcript thresholds were set based on observations from each distribution. Only those cells with a total mRNA count between 1,000 and 40,000 RPCs were retained.

A total of 410 individual cells passed these initial filters. Outliers found in subsequent, reiterative analyses described below were analyzed and removed, resulting in a final cell number of 396.

Mutant A53T mice

Sequences were aligned to a modified mm10 genome using the CellRanger v3.0.1 pipeline. The mm10 genome was modified using `cellranger mkref` to include a custom GFP construct to allow for quantification of the dopaminergic fluorescent marker expression. Unique molecular identifier (UMI) counts were quantified per gene per cell (`cellranger count`) and aggregated (`cellranger aggr`) across samples with no normalization. Our data had a high background level of reads not in cells, possibly from the combination of FACS and the microfluidics of the 10X Genomics platform, and as such, our capture efficiency was low, with only 2,430 cells captured (wildtype: 896; hemizygous: 1,059; homozygous: 475).

Using Seurat v3.1.1.9021, cells were examined for the number of genes expressed per cell, number of UMIs, and number of reads to identify empty droplets and doublets/multiplets (***Supplementary Figure 4.2***). Cells with less than 600 or more than 2,700 genes, those with greater than 10,000 UMIs assigned to the cell, and those with more than 40% of reads originating from the mitochondria were removed. A total of 1,357 cells made it through these filters (wildtype: 532; hemizygous: 600; homozygous: 225).

Normalization and cluster analysis

E15.5 and P7 wildtype mice

After initial filtering described above, the entire cds as well as subsets of the cds based on “age” and “region” of cells were created for recursive analysis. Regardless of how the data were subdivided, all data followed a similar downstream analysis workflow.

The genes to be analyzed for each iteration were filtered based on the number of cells that expressed each gene. Genes were retained if they were expressed in >5% of the cells in the dataset being analyzed. These were designated “expressed_genes.” For example, when analyzing all cells collected together ($n = 410$), a gene had to be expressed in 20.5 cells ($410 \times 0.05 = 20.5$) to be included in the analysis. In contrast, when analyzing P7 MB cells ($n = 80$), a gene had to be expressed in just four cells ($80 \times 0.05 = 4$). This was done to include genes that may define rare populations of cells that could be present in any given population.

The data were prepared for Monocle analysis by retaining only the expressed genes that passed the filtering described above. Size factors were estimated using the Monocle 2 “estimateSizeFactors()” function. Dispersions were estimated using the “estimateDispersions()” function.

Genes that have a high biological coefficient of variation (BCV) were identified by first calculating the BCV by dividing the standard deviation of expression for each expressed gene by the mean expression of each expressed gene. A dispersion table was then extracted using the “dispersionTable()” function from Monocle 2. Genes with a

mean expression > 0.5 transcripts and a “dispersion_empirical” $\geq 1.5 \times \text{dispersion_fit}$ or $2.0 \times \text{dispersion_fit}$ were identified as “high variance genes.”

PCA was run using the R “prcomp()” function on the centered and scaled \log_2 expression values of the “high variance genes.” PC1 and PC2 were visualized to scan the data for outliers as well as bias in the PCs for age, region, or plates on which the cells were sequenced. If any visual outliers in the data were observed, those cells were removed from the original subsetted cds and all filtering steps above were repeated. Once there were no visual outliers in PC1 or PC2, a screeplot was used to determine the number of PCs that contributed most significantly to the variation in the data. This was manually determined by inspecting the screeplot and including only those PCs that occur before the leveling-off of the plot.

Once the number of significant PCs was determined, t-SNE¹⁴² was used to embed chosen PC dimensions in a 2D space for visualization. This was done using the “tsne()” function available through the tsne package (v.0.1-3) in R with “whiten = FALSE.” The parameters “perplexity” and “max_iter” were tested with various values and set according to what was deemed to give the cleanest clustering of the data.

After dimensionality reduction via t-SNE, the number of clusters was determined in an unbiased manner by fitting multiple Gaussian distributions over the 2D t-SNE projection coordinates using the R package ADPclust³⁴⁴. t-SNE plots were visualized using a custom R script. The number of genes expressed and the total mRNAs for each cluster were then compared.

Mutant A53T mice

Single-cell RNA-seq data was processed using ``SCTransform()`` to normalize, find variable genes, and scale the data on each genotype separately. Doing so, we regressed out the effects of the proportion of reads mapping to the mitochondria and the replicate. To integrate the data, 3,000 variable features were selected (``SelectIntegrationFeatures()``) and integrated with the ``PrepSCTIntegration()``, ``FindIntegrationAnchors()``, and ``IntegrateData()`` functions.

Principal component (PC) analysis was performed and the distribution of replicates and genotypes were examined to confirm the success of `SCTransform` and integration (***Supplementary Figure 4.3***). ``JackStraw()`` and ``ElbowPlot()`` were used in combination to determine the PC cut-off, PC21, for inclusion into further dimensionality reduction.

t-SNE reduction was used to project the cell relationships onto two dimensions, using ``RunTSNE()`` with perplexity = 50, ``FindNeighbors()``, and ``FindClusters()`` with resolution = 0.2. Seven clusters were identified with 441, 220, 192, 181, 145, 139, and 39 cells. Clusters were examined for whether they contained cells collected from both replicates and all genotypes (***Supplementary Figure 4.4***). Clusters were also examined for bias in percent mitochondrial reads, number of genes expressed, and sequencing depth (***Supplementary Figure 4.5***). Following confirmation of unbiased clustering, the active assay was changed to RNA and expression was normalized with ``NormalizeData()``.

Estimating cell types

E15.5 and P7 wildtype mice

In order to find differentially expressed genes between brain DA populations at each age, the E15.5 and P7 datasets were annotated with regional cluster identity (“subset cluster”). Differential expression analysis was performed using the “differentialGeneTest()” function from Monocle 2 that uses a likelihood ratio test to compare a vector generalized additive model (VGAM) using a negative binomial family function to a reduced model in which one parameter of interest has been removed. In practice, the following model was fit: “~subset.cluster” for E15.5 or P7 dataset. Genes were called as significantly differentially expressed if they had a q value (Benjamini-Hochberg corrected p value) < 0.05.

Mutant A53T mice

Clusters were examined for expression of general marker genes, including pan-neuronal markers, oligodendrocyte/support cell markers (***Supplementary Figure 4.6***), and markers of a variety of different neuronal subtypes (***Supplementary Figure 4.7***). Expression of a variety of known marker genes¹⁵¹ was evaluated and clusters were assigned to anatomical locations (***Supplementary Figure 4.9***). New markers of each region were identified using ‘FindMarkers()’ where only positive markers present in over 25% of cells of that cluster, with an adjusted p-value less than 0.05 under a Wilcoxon rank sum test, were considered (***Table 4.2***). To find markers specific to each cluster, these lists of differential genes were sorted by ‘pct.2’, so as to identify those differentially expressed genes not present in other clusters. These lists of genes in consultation with ISH data from the Allen Mouse Brain Atlas³⁰⁴ were used

to confirm the likely anatomical locations of the clusters, and in the case of contaminating cells (cluster 6), suggest the cells are granule cells (***Supplementary Figure 4.8, Supplementary Figure 4.9***).

Clusters not corresponding to dopaminergic neurons were removed from further analysis, leaving 1,179 dopaminergic neurons for analysis.

Examining the effect of the mutant allele

Mutant transgene expression

Expression of the endogenous mouse *Prnp* was evaluated and visualized. The A53T transgene was randomly inserted into the mouse genome and is bound by the UTRs of the mouse prion protein (*Prnp*)²⁹¹. Since 10X Genomics technologies sequence from the 3' end of the transcript, reads originating from the transgene and the endogenous mouse *Prnp* gene are indistinguishable. Were the transgene included in the genome assembly during alignment, multimapping reads from either of these loci would fail to align and no expression of either the transgene or the endogenous locus would be detected. As such, by not including the transgene in the reference assembly, reads mapping to both the transgene and the endogenous locus are assigned to the endogenous mouse *Prnp* locus. To confirm transgene expression, we simply look for an increase of reads mapping to the *Prnp* gene in the mutant cells over the levels detected in the wildtype cells.

Effect on cell proportions

The percent of cells per genotype per cluster was calculated by dividing the number of cells in a cluster-genotype by the total number of cells in that genotype. Stacked barplots were generated.

Differential gene analysis between genotypes

To identify differentially expressed genes between the genotypes, wildtype cells were compared against homozygous cells, using the `FindMarkers()` function on a per cluster basis and globally and significant genes with an adjusted p-value less than 0.1 under a Wilcoxon rank sum test were identified (**Table 4.3**).

Pathway analysis

All genes up and downregulated overall were submitted to KEGG Search and Color Pathway Mapper³¹³. Querying the mouse genome for associated pathways, all dysregulated genes were submitted and colour coded such that genes that are downregulated in the mutant cells are red, and those that are upregulated in the mutant cells are green.

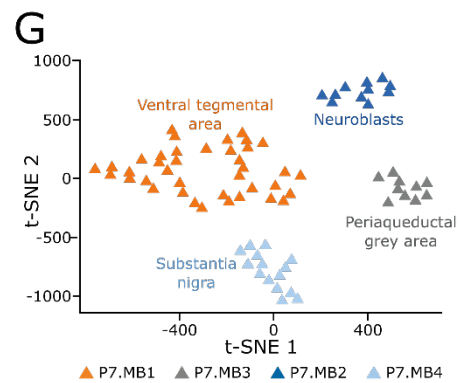
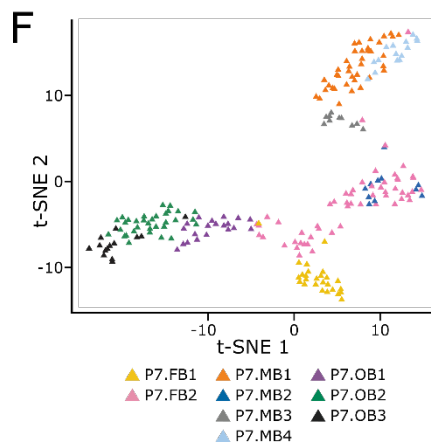
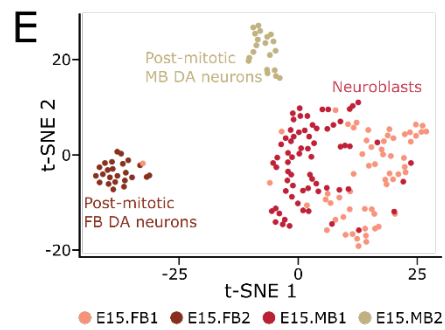
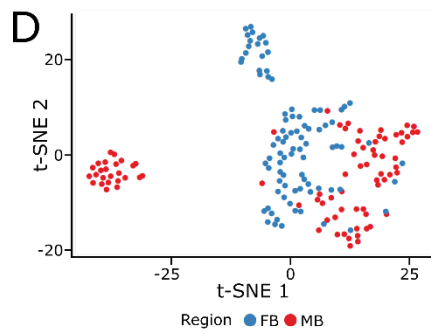
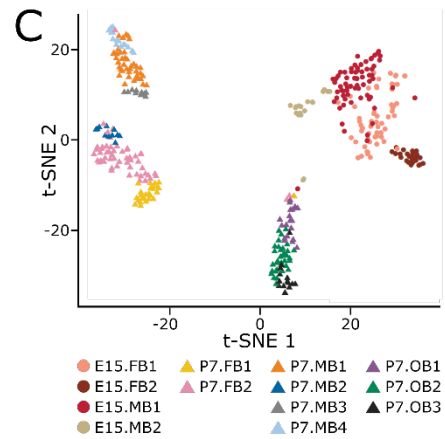
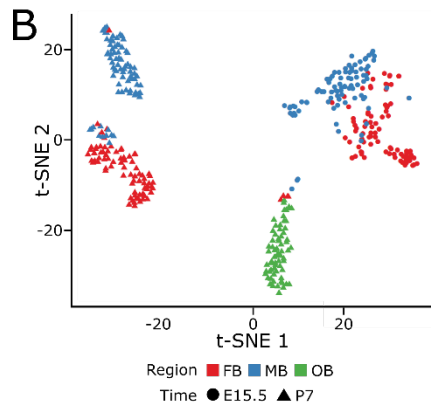
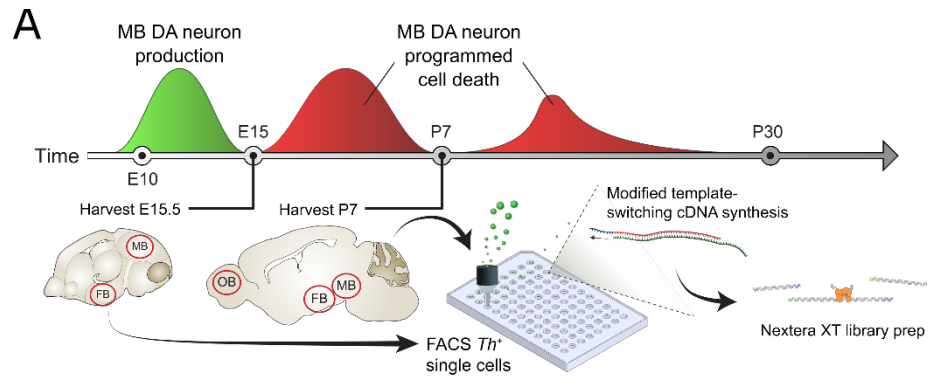
Data availability and sharing

Single-cell RNA-sequencing data will be available at the Gene Expression Omnibus (GEO). All code and related documentation will be made available in a GitHub repository at https://github.com/sarahmcclymont/A53T_scRNAseq. An associated Shiny app is in development.

4.7 Figures and supplementary materials

Figure 4.1: scRNA-seq identifies subpopulations of dopaminergic neurons at E15.5 and P7

(A) A schematic of the scRNA-seq experimental procedures for isolating and sequencing EGFP+ cells from E15.5 and P7 DA neurons of the MB, FB and OB. Timeline adapted from Barallobre *et al.*²⁶³ (B) A t-distributed stochastic neighbour embedding (t-SNE) plot of all collected cells that passed quality control measures colored by regional identity. E15.5 cells cluster together while P7 cells cluster primarily by regional identity. (C) A t-SNE plot of all collected cells coloured by subset cluster identity. Through iterative analysis, time point-regions collected can be separated into 13 total subpopulations. (D) Focusing specifically on all E15.5 cells, coloured by regional identity, we observe an overlap of MB and FB clusters. (E) With marker gene analysis, we can assign functions to these clusters. There are two clusters made each solely of MB and FB neurons, which represent post-mitotic DA neurons however, there also remains a neuroblasts population, composed of both MB and FB cells. (F) Focusing on all P7 neurons collected, coloured by cluster identity, we observe nine subpopulations of cells that mostly cluster by regional identity. (G) Examining specifically the P7 MB neurons, there are four subpopulations of cells: the substantia nigra, the ventral tegmental area, the periaqueductal gray area, and a neuroblast-like population.



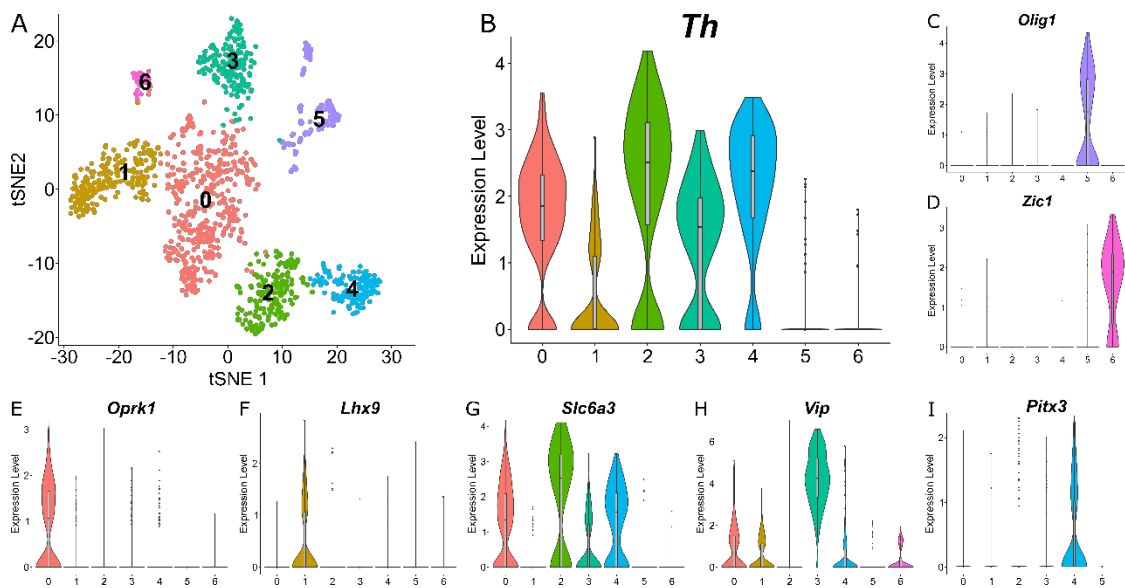


Figure 4.2: Seven transcriptionally distinct clusters of cells are identified

(A) A tSNE plot indicating the seven clusters of neural subpopulations identified post-filtering and quality control measures. (B) Clusters 0 through 4 have detectable levels of *Th* expression, indicating their dopaminergic functioning. Clusters 5 and 6 do not appear to express *Th* and are therefore likely to be contaminating cells. (C-I) The clusters are transcriptionally distinct from each other and marker genes can be identified and used to assign a biological identity to each cluster of cells.

Figure 4.3: All clusters are assigned a biological identity and marker genes of each are identified for functional validation by single-molecule RNA FISH (smFISH)

(A) The seven clusters are assigned to be the ventral tegmental area (VTA), a post-natal neuroblast population, two substantia nigra (SN) clusters, the periaqueductal grey area (PAG), and two contaminating cell types: support cells, like oligodendrocytes and microglia, and granule cells. (B) After removing these contaminating cells from analysis, the remaining clusters are all dopaminergic, with detectable *Th* expression. (C-I) Marker genes for each cluster that will be tested in wildtype P7 mice brain slices by smFISH to confirm the assigned biological identities and anatomical locations.

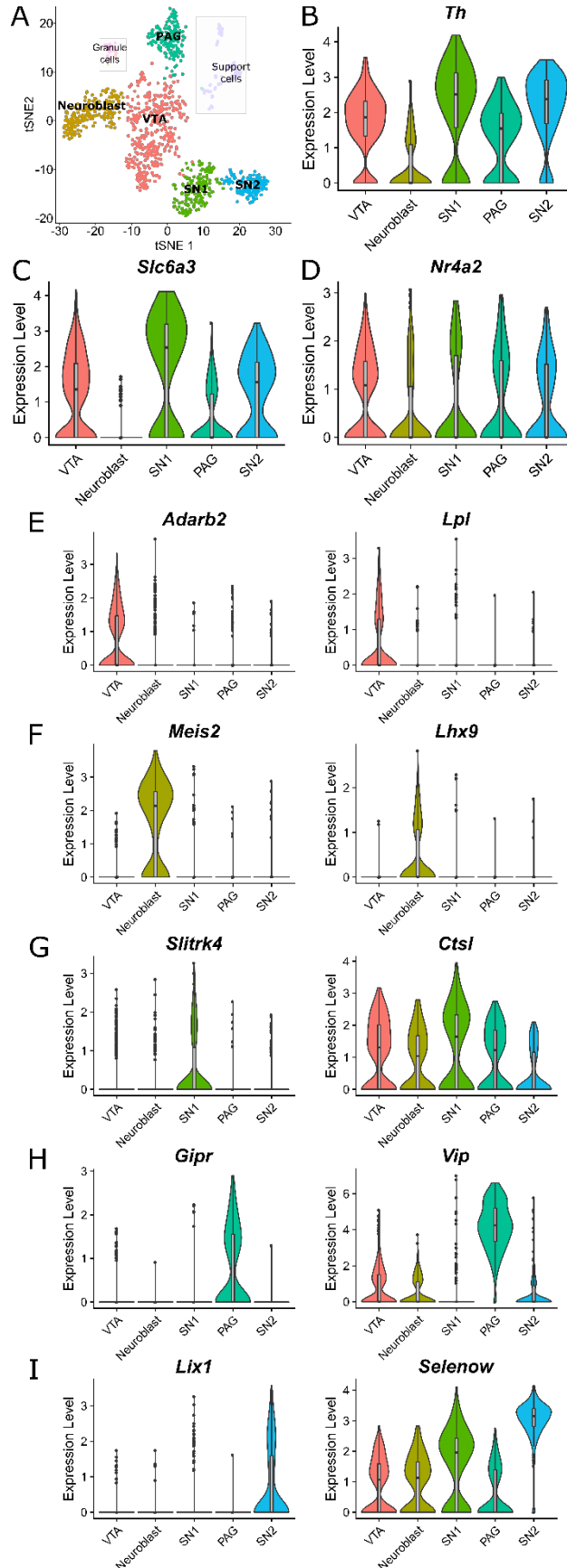
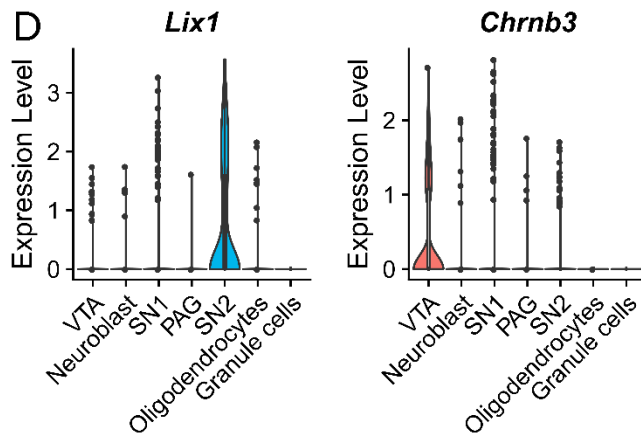
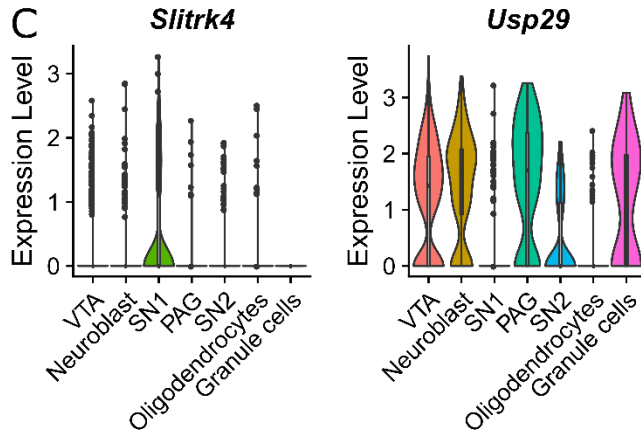
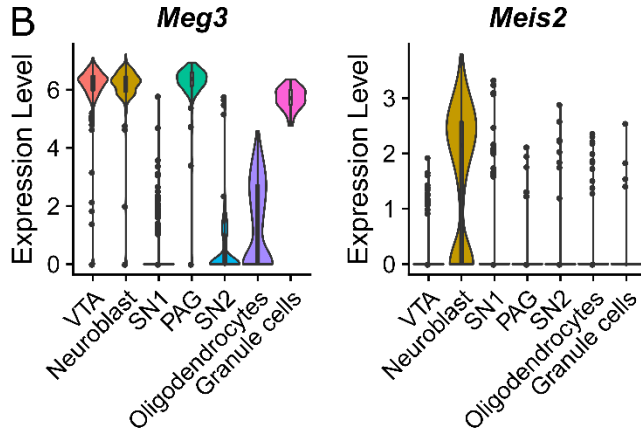
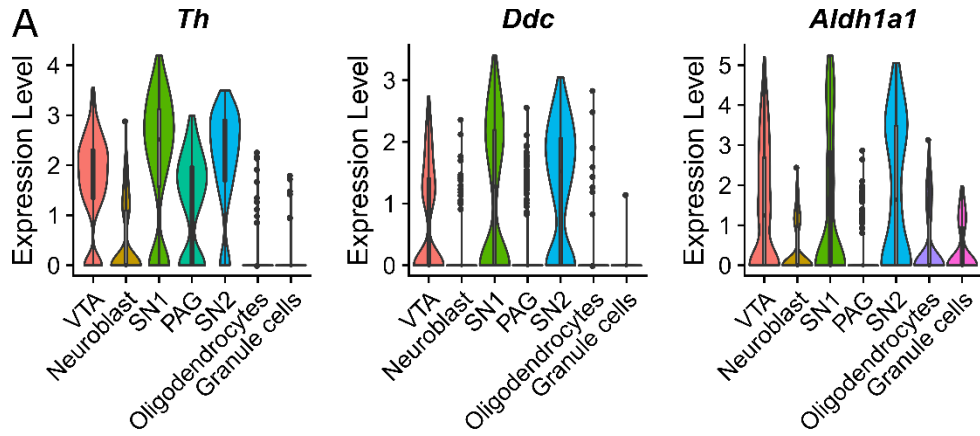


Figure 4.4: The VTA and two SN clusters are transcriptionally related and smFISH will be performed to assess their spatial relationships

(A) All three clusters share elevated expression of *Th*, *Ddc*, and *Aldh1a1*. The expression of these three genes will be probed with smFISH to confirm the location of all three clusters. (B-D) *Aldh1a1* will be used as a probe, along with each pair of genes plotted here, to isolate the locations of each cluster specifically. (B) The VTA cluster is expected to be *Aldh1a1*⁺, *Meg3*⁺, and *Meis2*⁻, to help separate the VTA expression patterns from the nearby neuroblast population that also has high expression of *Meg3*. (C) The SN1 cluster is expected to be *Aldh1a1*⁺, *Slitrk4*⁺, and *Usp29*⁻. *Usp29* was selected as a probe in order to distinguish this cluster from the VTA, neuroblast and SN2 populations. (D) The SN2 cluster is expected to be positive for expression of *Aldh1a1* and *Lix1*, but is not expected to express *Chrnb3*. This combination of probes should allow us to resolve the SN2 population while separating it from the VTA and SN1 clusters.



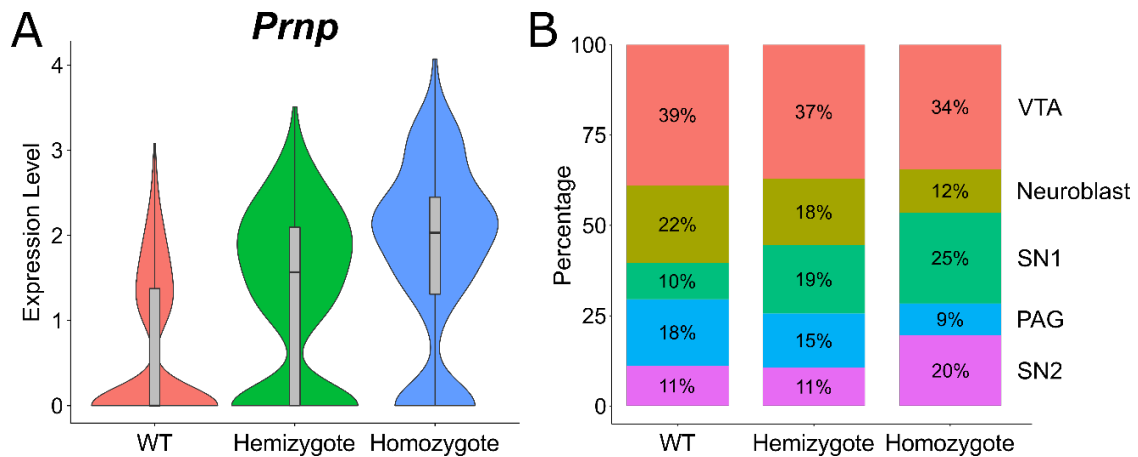


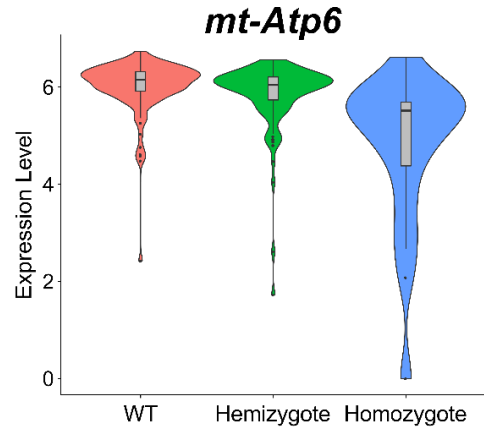
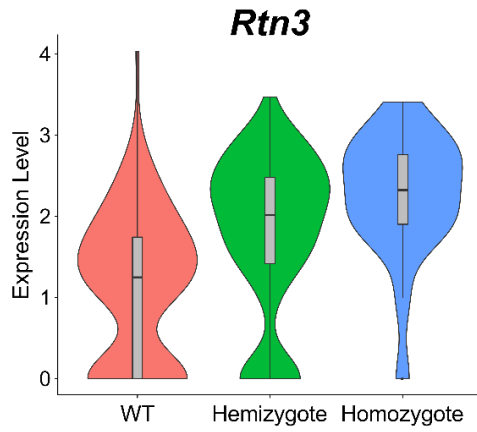
Figure 4.5: The mutant α -synuclein transgene is expressed at post-natal day 7 and alters dopaminergic neuron population proportions

(A) The mutant transgene is expressed in increasing levels commensurate with the number of mutant alleles. (B) The proportions of cells in each cluster are altered in the mutant state. Importantly, the post-natal neuroblast population has half as many cells in the homozygous state as in the wildtype controls, and there is an approximate doubling of the populations of substantia nigra cells in the mutant mice, perhaps indicating a precocious maturation phenotype. The periaqueductal grey area population is also halved in mutant mice, while the VTA remains largely unchanged.

Figure 4.6: There is cluster-specific gene dysregulation of genes previously implicated in neurodegeneration

(A) *Rtn3*, a gene previously implicated in β -amyloid levels in Alzheimer disease, is the only significantly upregulated gene in mutant post-natal neuroblasts. Additionally four genes, like *mt-Atp6*, all involved in oxidative phosphorylation are downregulated in mutant neuroblasts. (B) The VTA has 232 significantly dysregulated genes. *Cck* and *Resp18* are two upregulated genes in the mutant neurons and both are genes previously implicated in PD^{307–309}. Only 27 genes are significantly downregulated, including *Wsb1* and *Gria2*, which are interesting genes given their roles in Lewy body formation and ALS, respectively^{310–312}.

A
Post-natal
neuroblasts



B
Ventral
tegmental area

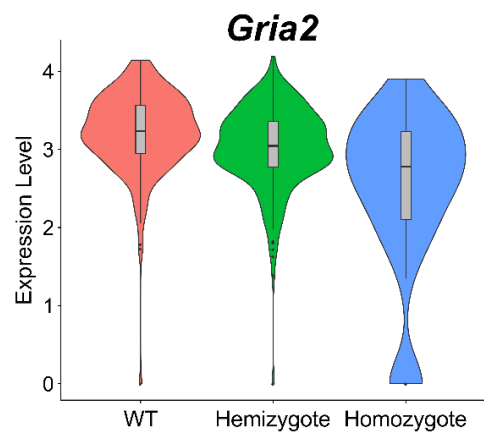
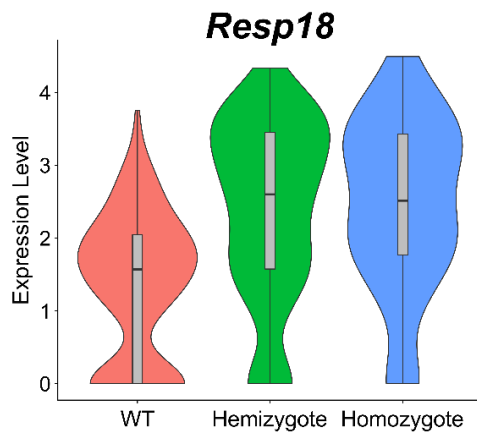
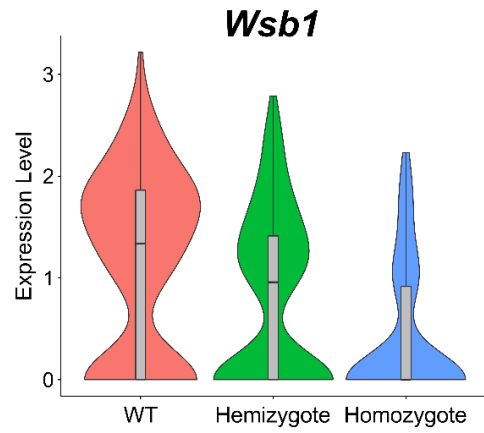
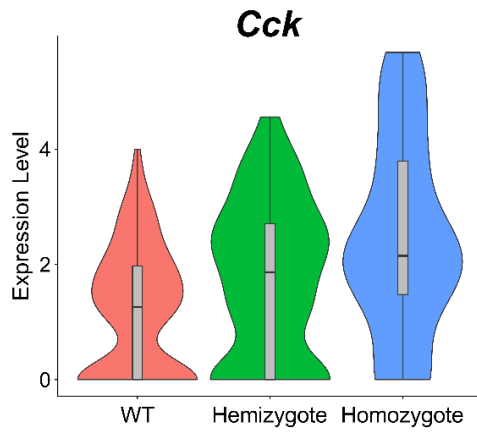
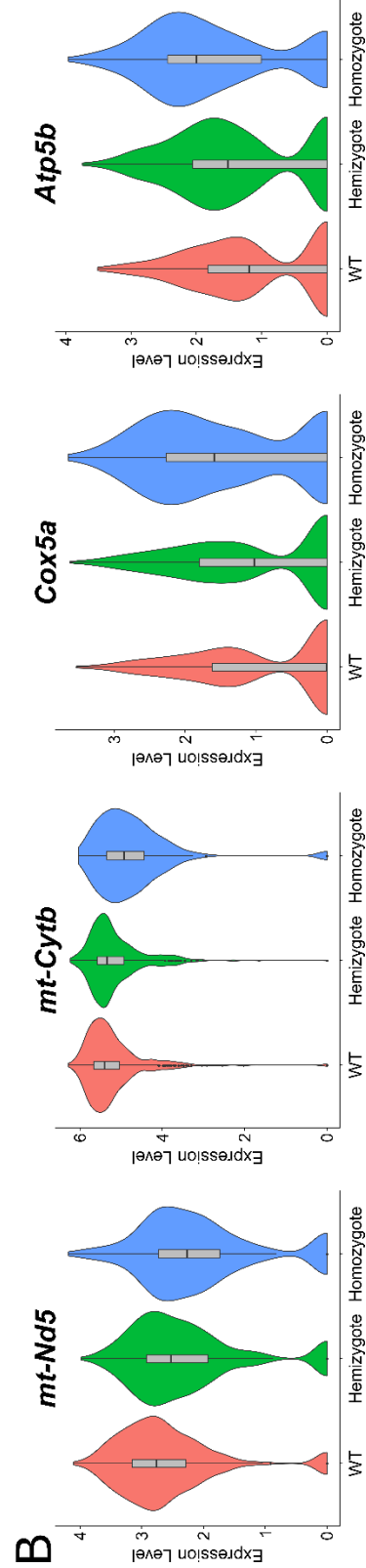
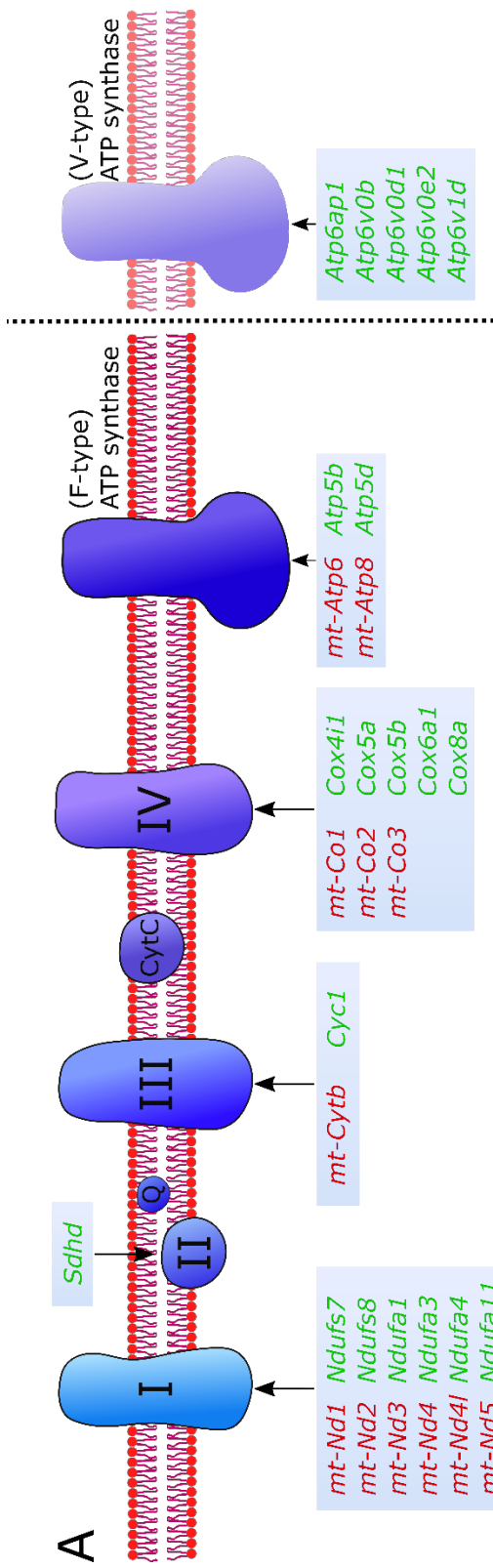


Figure 4.7: There is striking and extensive transcriptional dysregulation of the oxidative phosphorylation pathway in the mutant dopaminergic neurons

(A) A schematic indicating the genes downregulated (red) and upregulated (green) in the mutant state at each complex of the electron transport chain. 37 genes are dysregulated, including 12 downregulated and 25 upregulated genes. Strikingly, the genes that are downregulated are all mitochondrial-encoded while the genes that are upregulated are nuclear-encoded. (B) Examples of oxidative phosphorylation pathway genes that are significantly dysregulated in the mutant condition. *mt-Nd5* and *mt-Cytb*, encoded on the mitochondria, have lower expression in mutants, while *Cox5a* and *Atp5b*, encoded in the nucleus, have increased expression.



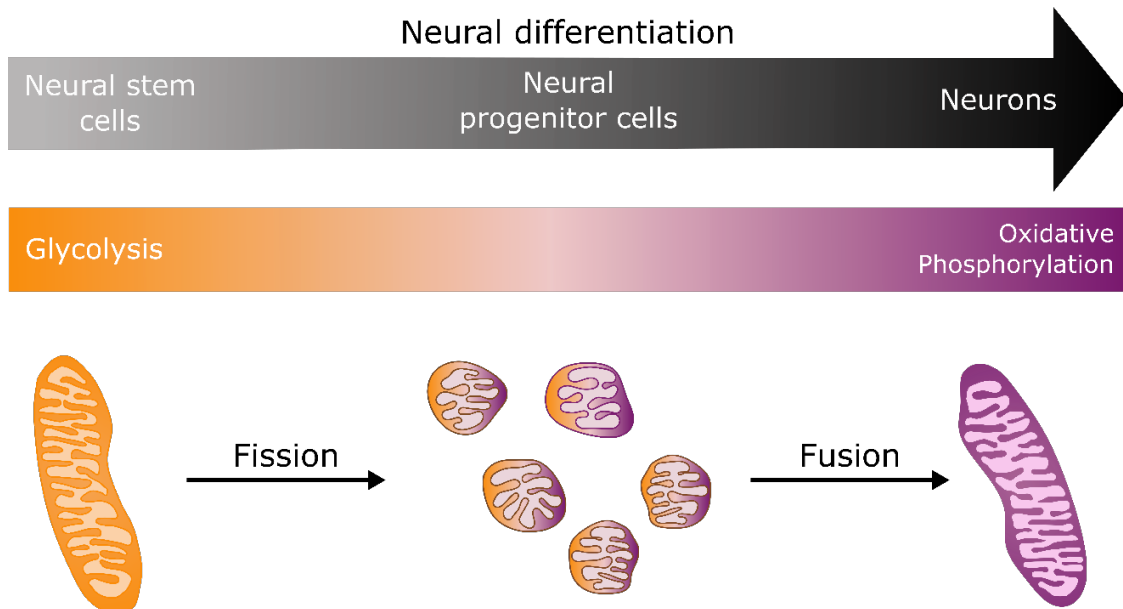


Figure 4.8: A mitochondrial fission or fusion defect could underlie the observed precocious maturation phenotype involving the neuroblasts and substantia nigra populations

In the process of neural differentiation there is a switch in energy production from glycolysis to oxidative phosphorylation. This change is accompanied by changes in mitochondrial conformation – from long, branched mitochondria to small, fragmented mitochondria, and back again^{337–339}. This morphological change is predicated on fission and fusion processes and therefore alterations to these dynamics could explain the maturation phenotype we observe.

Table 4.1: Marker gene expression analysis is used to assign identities to each of the clusters identified at E15.5 and P7

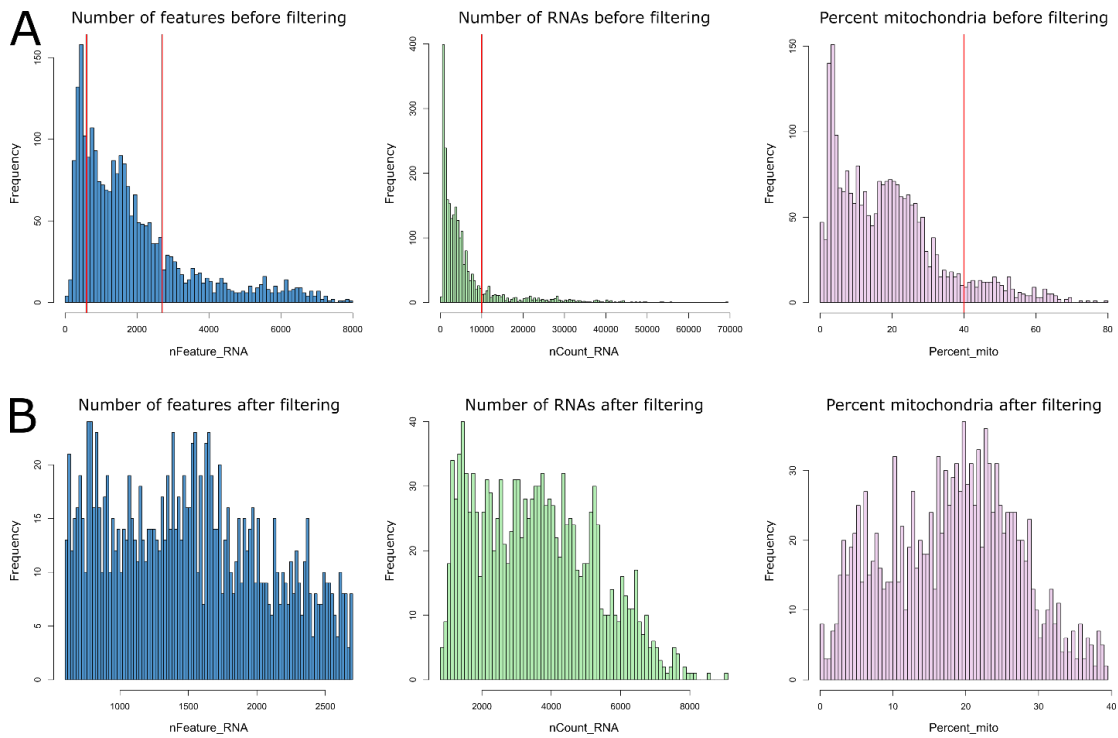
Age	Cluster	Identity	Selected markers
E15.5	FB1	Forebrain neuroblast	<i>Rnd3, Dlx1, Dlx2</i>
	FB2	Post-mitotic forebrain DA neurons	<i>Six3, Six3os1, Sst, Npy</i>
	MB1	Midbrain neuroblast	<i>Lhx9, Ebf1, Pax5, Nrg1</i>
	MB2	Post-mitotic midbrain DA neurons	<i>Foxa1, Lmx1a, Pitx3, Nr4a2</i>
P7	MB1	Ventral tegmental area (VTA)	<i>Otx2, Neurod6</i>
	MB2	Post-natal neuroblast	<i>Fam19a2, Lhx9, Ldb2, Mab21l1, Tmem163, Meis1, Rmst, Crnde, Gm2694</i>
	MB3	Periaqueductal grey area (PAG)	<i>Vip, Pnoc</i>
	MB4	Substantia nigra (SN)	<i>Sox6, Aldh1a7, Ndnf, Serpine2, Rbp4, Fgf20</i>

Table 4.2: Top six marker genes by foldchange difference per cluster

Cluster	Gene	\ln foldchange	Percent of cells in cluster	Percent of cells in all other clusters	Adjusted p- value
Cluster 0 - VTA	<i>Slc18a2</i>	0.982	80.3%	33.3%	5.82E-61
	<i>Epha5</i>	0.950	88.4%	47.2%	1.72E-64
	<i>Cck</i>	0.928	69.4%	41.2%	6.58E-20
	<i>Tenn1</i>	0.921	89.6%	40.1%	2.36E-74
	<i>Cadm1</i>	0.809	95.7%	66.2%	1.09E-46
	<i>Oprk1</i>	0.803	55.1%	12.6%	1.17E-57
Cluster 1 - Neuroblast	<i>Meis2</i>	1.869	67.3%	5.5%	1.94E-121
	<i>Sst</i>	1.311	30.5%	8.3%	1.92E-17
	<i>Rmst</i>	1.149	90.0%	36.4%	8.32E-61
	<i>Rbfox1</i>	1.083	46.8%	5.5%	3.92E-62
	<i>Cntn5</i>	1.028	45.0%	5.3%	1.16E-60
	<i>Mgat4c</i>	0.932	60.0%	20.9%	3.75E-34
Cluster 2 - SN1	<i>Slc6a3</i>	1.420	66.7%	39.2%	2.79E-27
	<i>Tuba1b</i>	1.052	93.8%	76.2%	2.07E-47
	<i>Hsp90aa1</i>	0.979	97.9%	90.9%	2.52E-49
	<i>Tubb2a</i>	0.967	99.5%	91.8%	1.36E-47
	<i>Th</i>	0.964	76.6%	61.3%	5.79E-19
	<i>Atp6v0e2</i>	0.958	71.4%	50.5%	1.53E-19
Cluster 3 - PAG	<i>Vip</i>	2.947	98.3%	33.0%	1.72E-108
	<i>Ebf1</i>	2.029	99.4%	32.1%	2.96E-106
	<i>Fam19a1</i>	1.201	90.6%	30.4%	3.65E-66
	<i>Crhbp</i>	1.130	39.8%	2.8%	1.31E-63
	<i>Gipr</i>	1.128	47.5%	2.0%	3.60E-93
	<i>Dlk1</i>	1.128	93.9%	49.4%	1.49E-44
Cluster 4 - SN2	<i>Tmsb10</i>	1.984	100.0%	91.2%	9.92E-79
	<i>Atp5g1</i>	1.802	97.9%	47.1%	1.16E-78
	<i>Selenow</i>	1.694	97.9%	55.6%	3.50E-73
	<i>Atp5k</i>	1.620	93.8%	34.2%	3.21E-70
	<i>Cox6b1</i>	1.589	97.2%	47.8%	8.80E-72
	<i>Rpl41</i>	1.582	100.0%	80.3%	1.65E-71
Cluster 5 - Oligo	<i>Apoe</i>	4.325	71.9%	7.7%	3.98E-102
	<i>Fabp7</i>	3.091	59.7%	4.6%	7.32E-95
	<i>Bcan</i>	2.873	70.5%	0.7%	3.96E-183
	<i>Atp1a2</i>	2.588	58.3%	2.0%	2.82E-122
	<i>Sparcl1</i>	2.458	35.3%	4.7%	9.59E-36
	<i>Olig1</i>	2.396	53.2%	0.6%	1.14E-132
Cluster 6 - Granule cells	<i>Nfib</i>	2.137	100.0%	11.7%	4.82E-57
	<i>Zic1</i>	1.921	74.4%	1.4%	2.99E-130
	<i>Fgfr1</i>	1.721	79.5%	13.7%	2.14E-31
	<i>Id2</i>	1.641	79.5%	14.7%	2.43E-27
	<i>Arpp21</i>	1.619	79.5%	14.2%	1.48E-28
	<i>Zfp536</i>	1.612	79.5%	4.2%	5.47E-78

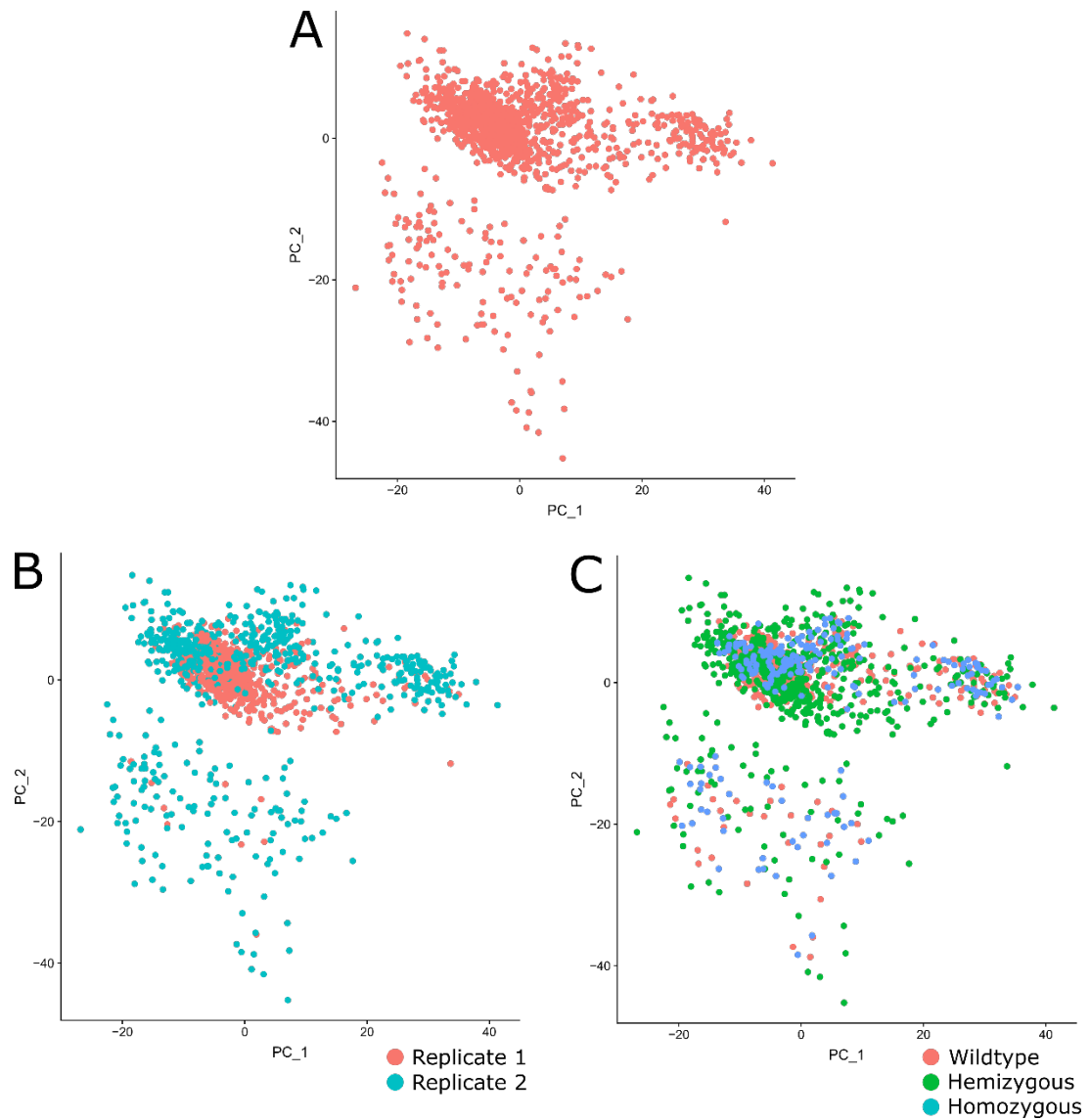
Table 4.3: Genes up- and downregulated globally and in a cluster-specific manner

Cluster	Gene	<i>ln</i> foldchange	Percent of cells in cluster	Percent of cells in all other clusters	Adjusted p-value	
Overall	<i>Prnp</i>	1.191	79.2%	40.2%	9.82E-30	
	<i>Pgrmc1</i>	0.960	88.0%	71.7%	5.54E-18	
	<i>Stmn4</i>	0.910	86.3%	73.8%	8.29E-11	
	<i>Serinc1</i>	0.876	87.4%	60.3%	2.50E-21	
	<i>Stmn3</i>	0.851	96.7%	93.9%	6.68E-23	
	<i>Scg5</i>	0.847	82.0%	76.5%	1.08E-07	
	<i>Atp6v0e2</i>	0.840	77.0%	44.0%	6.85E-17	
	<i>Cst3</i>	0.829	74.9%	62.8%	2.15E-07	
	<i>Fxyd6</i>	0.824	87.4%	71.9%	1.41E-17	
	<i>Bsg</i>	0.816	83.6%	69.8%	3.19E-13	
			...			
		<i>mt-Atp8</i>	-0.683	73.2%	87.1%	6.62E-14
		<i>Arglu1</i>	-0.698	33.9%	64.3%	8.36E-10
		<i>Ogt</i>	-0.755	33.3%	66.2%	6.27E-12
		<i>Rmst</i>	-0.782	28.4%	58.6%	2.27E-08
		<i>Gm42418</i>	-0.784	32.2%	66.4%	1.03E-12
		<i>Malat1</i>	-0.787	69.9%	86.0%	1.20E-13
		<i>Tia1</i>	-0.793	25.7%	62.2%	4.82E-14
		<i>AC149090.1</i>	-0.868	45.4%	78.0%	4.81E-17
	<i>Snhg11</i>	-0.953	47.0%	80.8%	1.21E-22	
	<i>Meg3</i>	-0.974	55.7%	84.8%	6.27E-25	
VTA	<i>Cck</i>	1.931	82.5%	62.0%	3.74E-03	
	<i>Pgrmc1</i>	1.422	93.7%	74.5%	6.04E-11	
	<i>Cst3</i>	1.383	81.0%	62.0%	5.89E-06	
	<i>Stmn4</i>	1.368	87.3%	78.3%	9.63E-08	
	<i>Bsg</i>	1.315	87.3%	69.6%	1.70E-09	
	<i>Stmn3</i>	1.297	95.2%	93.5%	2.58E-06	
	<i>Scg5</i>	1.259	90.5%	76.1%	3.77E-08	
	<i>Fxyd6</i>	1.227	90.5%	70.7%	5.46E-08	
	<i>Ly6h</i>	1.224	96.8%	92.4%	3.20E-04	
	<i>Ndufc2</i>	1.218	71.4%	39.7%	2.41E-06	
			...			
		<i>Tia1</i>	-0.661	42.9%	78.8%	2.83E-03
		<i>Snhg11</i>	-0.667	77.8%	100.0%	4.40E-11
		<i>Rbm5</i>	-0.678	41.3%	74.5%	1.64E-03
		<i>Carmil3</i>	-0.704	30.2%	65.2%	9.82E-03
		<i>Meg3</i>	-0.745	84.1%	100.0%	1.06E-15
		<i>Prpf4b</i>	-0.782	30.2%	61.4%	8.93E-03
		<i>AC149090.1</i>	-0.801	73.0%	95.1%	5.06E-07
		<i>Arglu1</i>	-0.844	38.1%	77.2%	5.60E-05
	<i>Wsb1</i>	-0.857	28.6%	66.8%	1.89E-04	
	<i>Gm42418</i>	-0.896	39.7%	76.6%	4.27E-05	
Neuroblast	<i>Prnp</i>	1.433	90.9%	48.0%	0.0004	
	<i>Rtn3</i>	0.882	95.5%	68.6%	0.016	
	<i>mt-Co3</i>	-0.644	95.5%	100.0%	0.063	
	<i>mt-Nd1</i>	-0.680	90.9%	100.0%	0.034	
	<i>mt-Atp6</i>	-0.734	95.5%	100.0%	0.005	
	<i>mt-Cytb</i>	-0.750	81.8%	99.0%	0.017	
PAG	<i>Nsg1</i>	1.166	100.0%	63.2%	0.092	
	<i>Ctsl</i>	1.086	87.5%	31.0%	0.023	
	<i>Tomm20</i>	1.056	75.0%	21.8%	0.074	
	<i>Dlk1</i>	0.968	100.0%	92.0%	0.005	
	<i>Cd24a</i>	0.869	56.2%	5.7%	0.003	
	<i>AU040320</i>	0.755	31.2%	1.1%	0.091	
	<i>Mtfp1</i>	0.616	25.0%	0.0%	0.074	
	<i>Gm9803</i>	0.590	37.5%	1.1%	0.005	
	<i>Cbr4</i>	0.568	25.0%	0.0%	0.074	
SN 1	NA	NA	NA	NA	NA	
SN2	<i>Prnp</i>	1.552	75.0%	34.0%	0.055	



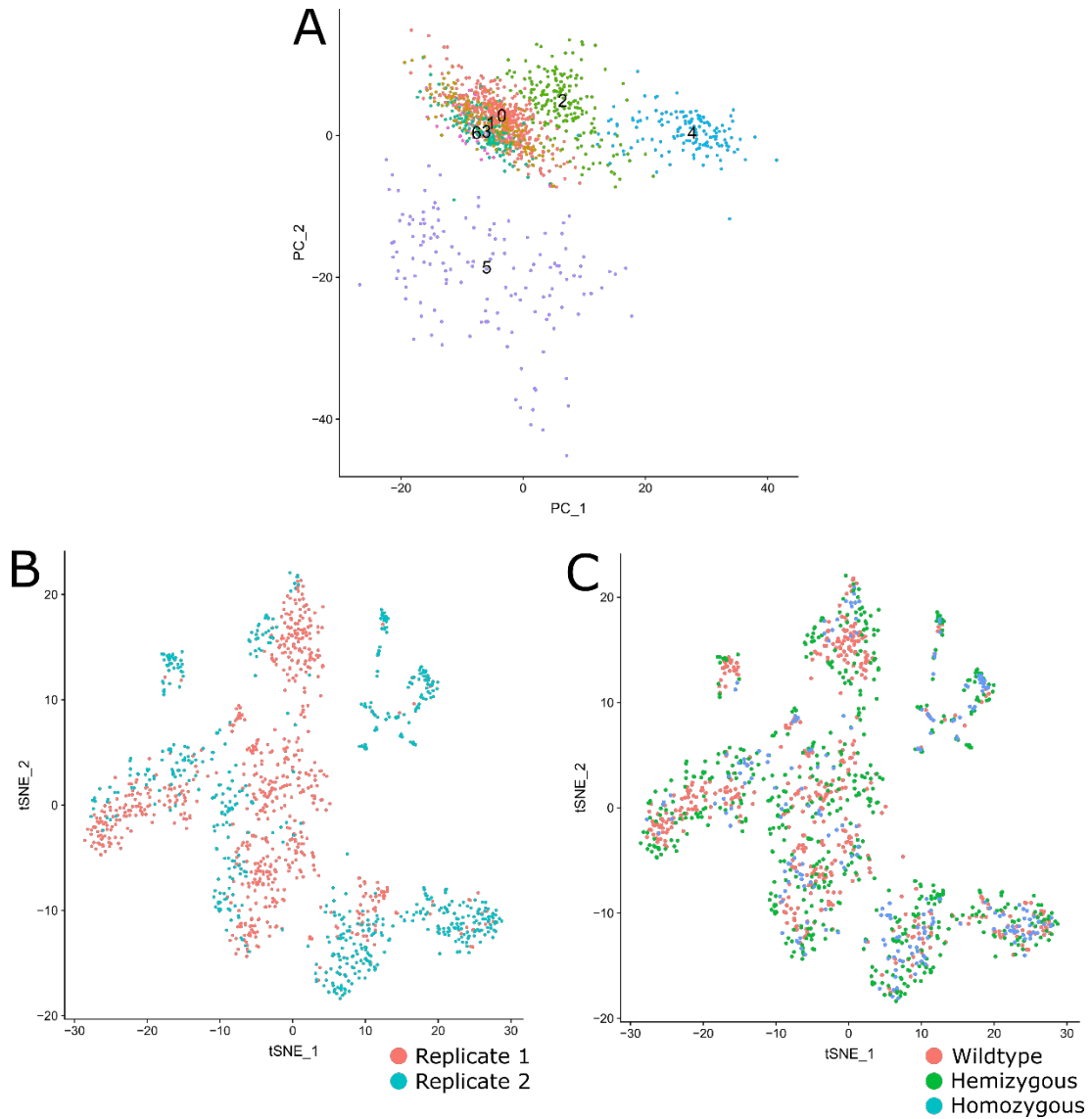
Supplementary Figure 4.2: Quality control metrics used to filter scRNA-seq data

(A) The number of unique genes in each cell, the number of molecules in a cell, and the percent of reads mapping to the mitochondria are plotted and filtering cut-offs are identified to exclude empty droplets, doublets, and low quality cells. (B) Following filtering using these empirically derived cut-offs, the distributions are again examined to confirm the filtering removed low quality cells from the analysis.



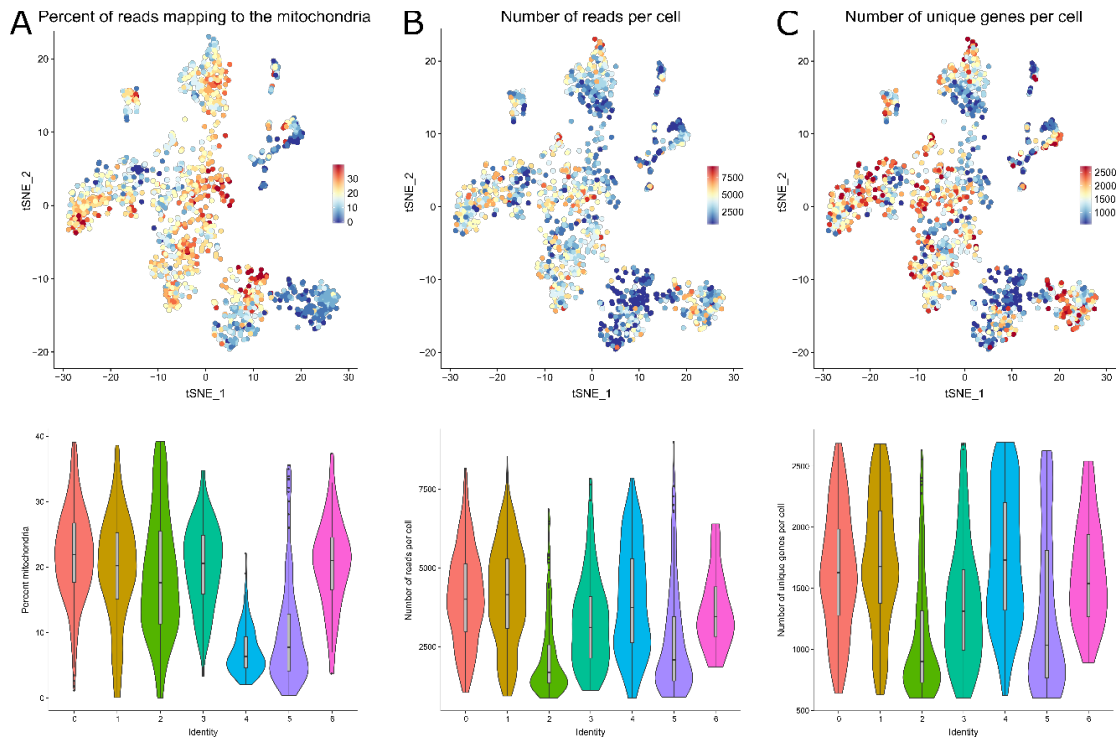
Supplementary Figure 4.3: Replicates and genotypes are well distributed across PC1 and PC2

(A) PC1 and PC2 are visualized. (B) PC1 and PC2 are examined for bias in replicate distribution and (C) genotype distribution.



Supplementary Figure 4.4: Replicates and genotypes are present in each cluster

(A) The clusters as projected onto PC1 and PC2 indicate that cluster 5 represents the group of cells that is separated from the main mass of cells. (B) The replicates are represented in each of the clusters. (C) Each of the three genotypes are present in all clusters.

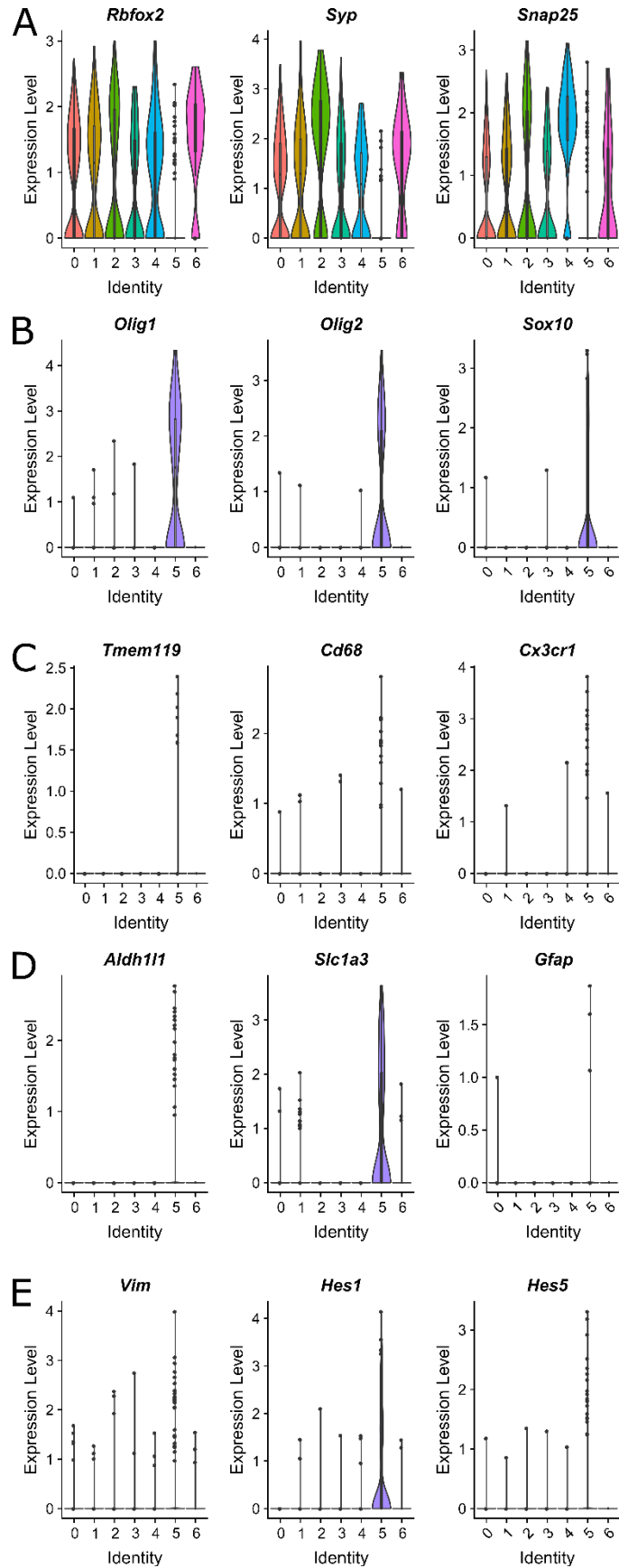


Supplementary Figure 4.5: Assessing the clusters for bias in percent mitochondrial reads, number of reads and number of genes expressed per cell

(A) Cluster 4 and the outlier cluster 5 have few cells with a high proportion of mitochondrial reads. (B) Number of reads and (C) unique genes expressed per cell are well distributed across the clusters, with clusters 2 and 5 having a lower number of each than the other clusters.

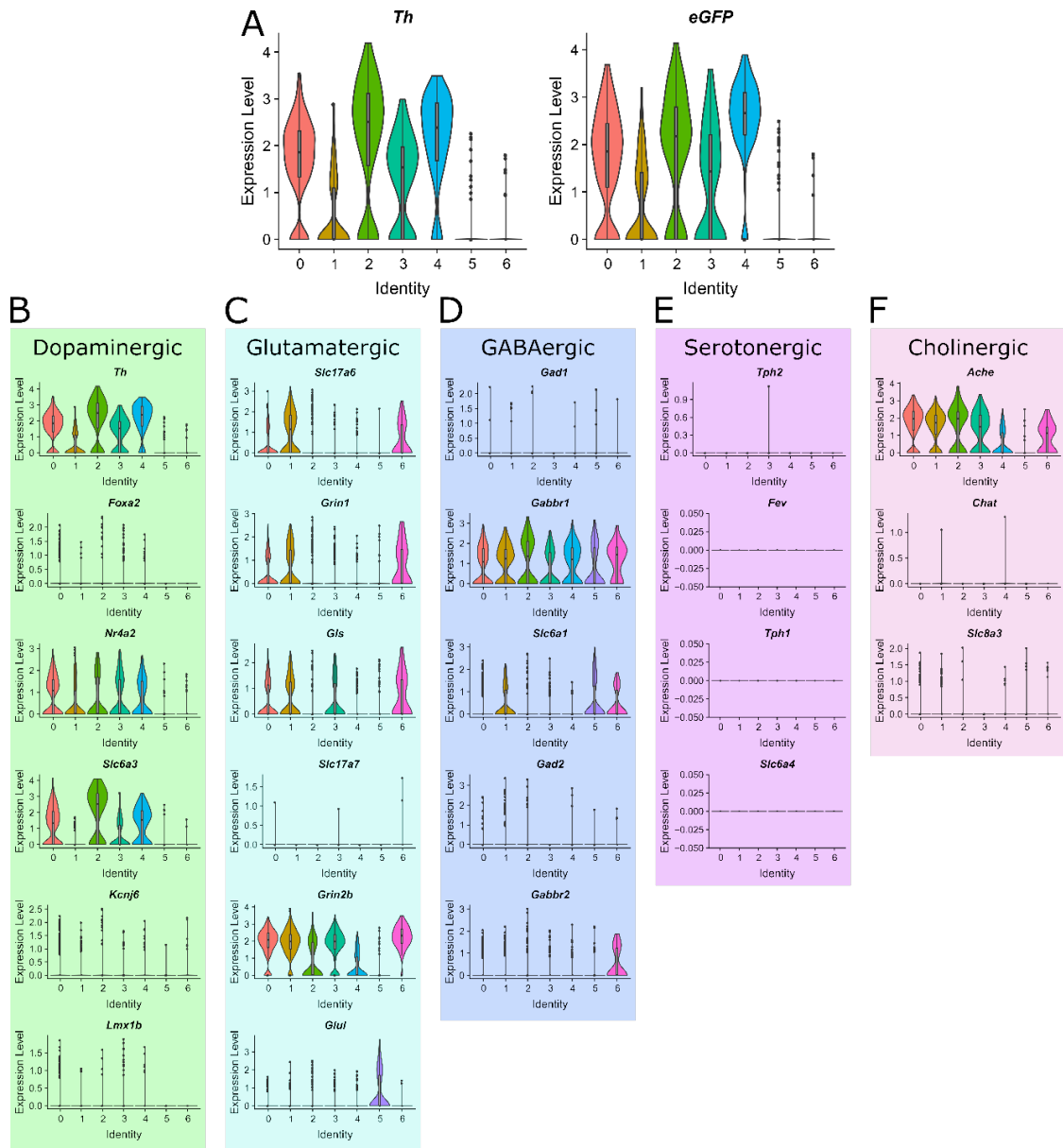
Supplementary Figure 4.6: Cluster 5 displays markers of support cells, like oligodendrocytes and astrocytes

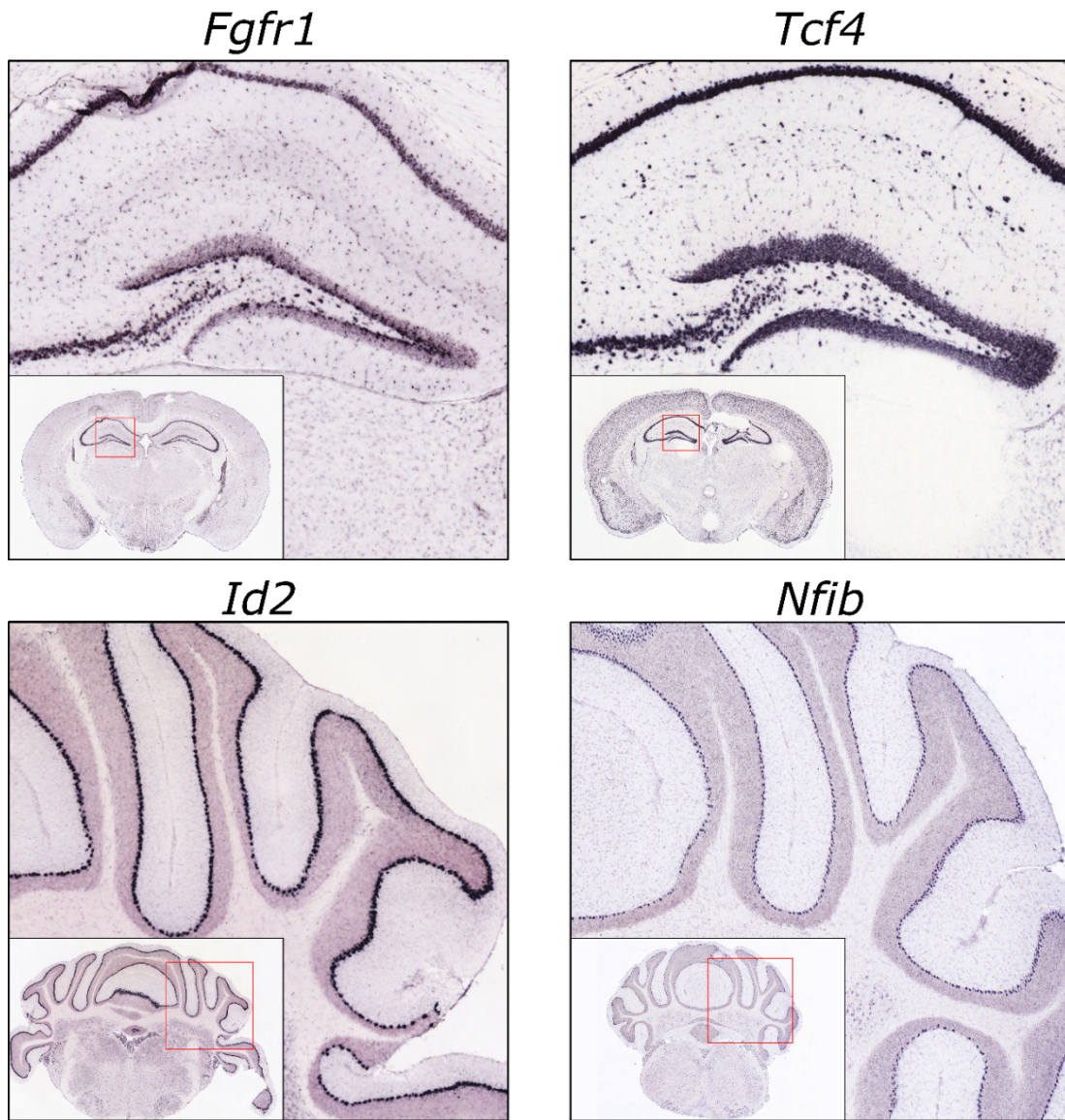
(A) Cluster 0, 1, 2, 3, 4, and 6 express pan-neuronal markers. Cluster 5 does not, instead expressing (B) markers of oligodendrocytes, (C) microglia, (D) astrocytes, and (E) glia.



Supplementary Figure 4.7: Cluster 6 is not dopaminergic and likely represents a contaminating cell type

(A) Cluster 0 through 4 express *Th* and eGFP, which was the marker upon which the cells were sorted by FACS, at robust levels. Cluster 5, oligodendrocytes, and cluster 6 appear to be contaminating cell types, which do not express eGFP. All clusters were evaluated for expression of markers of (B) dopaminergic³⁴⁵⁻³⁴⁸, (C) glutamatergic³⁴⁹⁻³⁵¹, (D) GABAergic³⁵²⁻³⁵⁴, (E) serotonergic³⁵⁵⁻³⁵⁷, and (F) cholinergic neurons³⁵⁸⁻³⁶⁰. Cluster 6 does not appear to be exclusively any one of these subtypes, but does express markers of glutamatergic and GABAergic neurons.



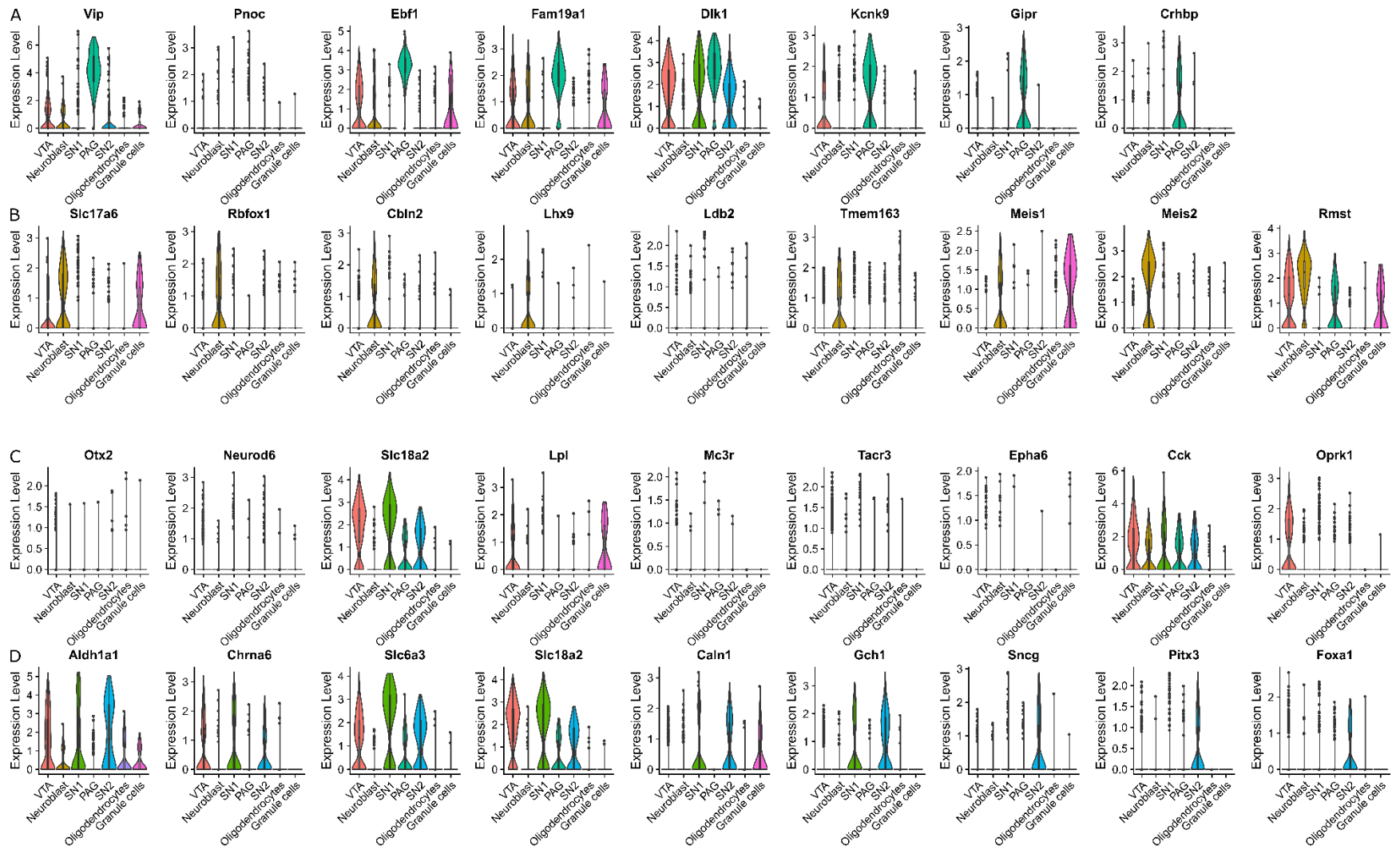


Supplementary Figure 4.8: The contaminating cluster 6 is likely to be granule cells

Consulting the Allen Mouse Brain Atlas for *in situ* hybridization experiments on some of the top marker genes, *Fgfr1*, *Tcf4*, *Id2*, and *Nfib*, all display increased expression in granule cell layers, either in the hippocampus (*Fgfr1*, *Tcf4*) or in the cerebellum (*Id2*, *Nfib*). Experiment and slide identifiers in **Supplementary Table 4.6**.

Supplementary Figure 4.9: Expression of previously established marker genes was used to assign biological identities to each of the remaining clusters

(A, B) Marker gene analysis easily identifies these clusters as (A) the periaqueductal grey area (PAG) and (B) a post-natal neuroblast population. (C, D) Discriminating the ventral tegmental area (VTA) from the remaining two clusters is challenging given some of the marker genes have shared expression patterns, but key genes like (C) *Oprk1* and *Lpl* distinguish this cluster as the VTA. (D) The remaining two clusters are likely substantia nigra clusters (SN1 and SN2), as demonstrated by the high expression of *Aldh1a1*, *Chrna6*, and *Slc6a3*. Specific to the two SN clusters is *Caln1* and *Gch1* expression. Distinguishing SN1 from SN2 is more difficult. However, we can detect differences like genes specific to SN2, like *Pitx3* and *Foxa1*, while SN1 more highly expresses markers of DA neurons, like *Slc6a3* (here), and *Th* and *Ddc* (seen in *Figure 4.4*).



Supplementary Table 4.1: Experiment and image identifiers for the Allen Mouse Brain Atlas *in situ* hybridization slides used in assigning cluster six as granule neurons

Gene	Experiment ID	Image ID	URL
<i>Fgfr3</i>	79588290	79593744	https://mouse.brain-map.org/experiment/siv?id=79588290&imageId=79593744
<i>Id2</i>	71836806	71681421	https://mouse.brain-map.org/experiment/siv?id=71836806&imageId=71681421
<i>Tcf4</i>	79488927	79454916	https://mouse.brain-map.org/experiment/siv?id=79488927&imageId=79454916
<i>Nfib</i>	1555	101362638	https://mouse.brain-map.org/experiment/siv?id=1555&imageId=101362638

Chapter 5: Conclusions

5.1 Summary of significant findings

Comprehensive evaluation of midbrain dopaminergic (DA) neurons is required for an improved understanding of the molecular mechanisms of Parkinson disease (PD) risk and progression. Here, we have examined the chromatin and transcriptional landscapes of midbrain DA neurons to extensively catalogue putative enhancers, identify transcription factors (TFs) important to gene expression regulation, uncover variants that disrupt enhancer functioning and TF binding, and detect significant alterations in gene expression, neuronal maturation, and mitochondrial function in the early disease state.

To identify and prioritize non-coding variants associated with PD, we first queried the chromatin of 50,000 midbrain and forebrain DA neurons. We identified >100,000 open chromatin regions that we assessed for functionality using *in silico* and *in vivo* means. One such region, in an intron of the familial PD gene, *SNCA*, directs reporter expression in catecholaminergic neurons throughout the mouse life course in key regions affected in PD. Not only does this enhancer direct disease-appropriate expression, but in sequencing ~1,000 PD cases and controls, we demonstrate it to contain disease-associated variation. Two tightly linked common variants, rs2737024 and rs2583959, that fall in this enhancer are significantly associated with PD risk (odds ratio ~1.25, p-value = 0.002).

This finding is intriguing, not only because we identify novel disease-associated enhancer variants near to a familial PD gene, but this also represents the possible benefit of a paradigm shift in how common disease variants are

identified. Typically, large genome-wide association studies uncover loci associated with disease but do not indict any one variant. Functional fine-mapping, in which these results are intersected with functional annotations, like those generated by ChIP-seq experiments, can be employed to prioritize variants that likely disrupt functional elements. Our strategy flips this paradigm – starting with a high quality, disease-relevant annotation dataset, we query these annotations and identify disease-associated variants by sequencing a relatively small number of cases and controls.

To uncover how these and other common variants associated with neurodegeneration and neuropsychiatric disorders disrupt gene expression and ultimately modulate disease risk, we employed a machine learning algorithm, gkm-SVM, to learn the regulatory vocabulary of midbrain DA neurons. In combination with TF footprinting and bulk RNA-seq, we identify TFs that are actively engaging the DNA to influence gene expression in MB DA neurons. We identify Rfx3/5, Foxa1/2, Ascl1, and Nr4a2 as important to MB DA neurons. We use this regulatory vocabulary to predict how >7,000 disease-associated variants might impact TF and protein binding. These predictions have been challenging to validate given the cell-type restricted activity of enhancers and the reliance on common validation strategies using *in vitro* cell surrogates. Prevented, as we were, from using standard luciferase assays to validate our predictions, we shifted our focus to protein-binding arrays and have identified four proteins, NOVA1, PEG10, SNRPA, and CHMP5, whose binding may be impacted by the *SNCA* enhancer variants.

The challenge of finding appropriate *in vitro* cellular surrogates for *in vivo*-derived data is not unique to this study. Enhancer usage can be excessively

restricted by cell type, developmental time point, or environmental perturbation. As the field moves forward in querying increasingly more refined populations *in vivo*, the validity of using an *in vitro* cell surrogate not matched for those characteristics becomes untenable. One strategy to combat this limitation is to perform validation experiments *in vivo*, in the same cells as were queried. This strategy would be ideal in terms of the biological validity of the results but is not always feasible given the low throughput nature of many *in vivo* assays and the cell number requirements of many enhancer reporter assays. An alternative strategy is to find more appropriate *in vitro* cell surrogates. Here, we use ATAC-seq and bulk and single-cell RNA-seq (scRNA-seq) to judge the suitability of an immortalized mouse substantia nigra cell line, SN4741. Karyotyping suggests these cells are triploid and highly genetically complex. Preliminary scRNA-seq fails to confirm these cells as necessarily DA neurons, calling into question their suitability. We continue to assess these cells but also continue to search for alternative surrogates and may turn our focus to differentiated iPSCs.

Finally, we investigated the mechanisms underlying the preferential degeneration of DA neurons of the substantia nigra. We performed scRNA-seq in ~1,300 midbrain DA neurons at an early post-natal time point in a mouse model of PD to assess the effects of the mutation on the transcriptomes of developing neurons. We identify five clusters of MB DA neurons, including a novel separation of two substantia nigra populations. The mutant allele has significant effects on the distribution of cells within the clusters, with half as many neuroblasts present in the homozygous mutant mice. This halving appears to be accompanied by a commensurate doubling of the number of mutant cells in the substantia nigra

clusters, likely indicating a precocious maturation defect of DA neurons in relevant and key populations in the pathogenesis of PD.

There is also significant cluster-specific and global gene dysregulation in the mutant state. Genes known to be involved in neurodegenerative diseases, including PD, are among the hundreds of genes identified as differentially expressed, but perhaps most strikingly, there is extensive dysregulation of genes involved in the oxidative phosphorylation pathway. Most interestingly, this dysregulated expression of mitochondria is dependent on the genome on which the implicated gene is encoded on. The mitochondrial-encoded genes of the electron transport chain are all downregulated, while the upregulated genes are all encoded in the nucleus. This is a novel pattern of transcriptional dysregulation in PD and we hypothesize this may indicate a mitochondrial fission or fusion dysfunction that underlies the neuronal maturation defect of neuroblasts into substantia nigra neurons. These observations suggest a developmental origin to the preferential degeneration of substantia nigra DA neurons in PD.

5.2 Future directions

We are eager to continue investigating each of these exciting results. In regards to the enhancer at *SCNA* containing disease-associated variants, we look forward to characterizing the enhancer-deletion mouse for molecular and gross motor phenotypes. To supplement these deletion studies, it would be worthwhile to study the reverse phenomenon and attempt an over-expression assay at the enhancer, perhaps through recruiting activating TFs or chromatin remodellers to the locus with CRISPR-activation experiments. Additionally, the chromatin assay

of midbrain DA neurons yielded a catalogue of over 100,000 open chromatin regions. We have examined six of these. There remains a myriad of information still to be explored in these catalogues, either through further single locus investigations, looking at any one of the other 100 some odd regions that we already know to overlap PD associated variation, or we are moving to massively parallel reporter assays (MPRA) to test many disease associated variants for function simultaneously. We are also exploring using promoter capture HiC (pcHiC) to assign open chromatin regions to their cognate genes.

Both MPRA and pcHiC have high cell number requirements for assay and as such, usually rely on *in vitro* modelling. We continue to search for accurate cell surrogates for our cell types of interest. We have focused our efforts on characterizing SN4741 cells with ATAC-seq, bulk RNA-seq, and scRNA-seq. This characterization strategy can be applied to a variety of cell lines to allow an empirical judgement as to the relatedness of the chromatin, TFs, and transcriptome to the *in vivo* system. While SN4741 cells are perhaps not as proximal to a DA state as we had hoped, we are currently investigating methods to improve the differentiation process, including longer differentiation time or supplementation with differentiation factors like retinoic acid.

Although we have examined the transcriptomes of midbrain DA neurons in a mouse model of PD, we have not yet examined the chromatin landscape of these neurons or how the disease state might alter this. We have presently optimized small-scale ATAC-seq to facilitate analysis of these neurons. This will allow us to assay enhancer usage and develop a TF vocabulary at post-natal time point and will also allow us to assess how changes in the chromatin are related to changes in the

transcriptome. This would be a rich source of data however it would likely be unable to elaborate on the mitochondrial phenotype, given the lack of nucleosomes in the mitochondrial genome. We have begun employing a variety of assays to examine the mitochondria in our mouse model of PD however, these are static evaluations of mitochondria. To investigate mitochondrial dynamics in an *in vivo* model, we can employ mouse models like the MitoMouse lines³⁶¹, wherein the mitochondria are fluorescently-labelled for *in vivo* imaging. Ultimately, we hope an improved understanding of these mechanisms will allow us to modulate the progression or risk of PD.

Assaying the chromatin and transcriptomic landscapes of wildtype and mutant mice has proven effective for increasing our understanding of the biology of DA neurons and PD. We plan to extend these assays to other time points, cell types, and perturbations in order to expand our understanding. These strategies represent powerful paradigms for assaying restricted populations of cells in the wildtype and disease states. We believe that these approaches can be applied to a biological systems in order to provide insight into the molecular mechanisms of health and disease.

Chapter 6: Appendices

6.1 The transcriptional targets of SOX9 in Type II Diabetes⁵

Ultimately, both Type I and Type II diabetes result in a loss of functioning beta-cells. Current treatments, including insulin injections, transplantation of donor beta cells, and differentiation of stem cells are compromised by systemic complications, scarcity of donor tissues, and cost, respectively. An effective treatment for these diseases would ideally involve an increase in beta-cell mass. Thus exploring the induction of beta-cells endogenously from pancreatic progenitors is an alluring treatment target. However, the capacity for beta-cell neogenesis in mammals is controversial. Neogenesis of beta cells after partial ductal ligation has been seen in mice in some studies but not others^{362–367}, and targeted ablation of beta cells in mice is resolved by proliferation and transdifferentiation^{368–372}. Analysis of islets of adult humans with Type I diabetes that died from diabetic ketoacidosis showed that while beta cells continuously apoptosed, there remained beta cells present at death and no markers of proliferation, suggesting some capacity for neogenesis^{373,374}. Conversely, zebrafish have an extraordinary capacity to regenerate beta cells by neogenesis after targeted ablation^{375,376} and we have identified the progenitor cell that contributes to beta-cell neogenesis in the zebrafish: centroacinar cells (CACs)³⁷⁷.

CACs are terminal intercalated duct cells – they exist in acini, have long extensions, are connected by tight junctions so they can successfully line pancreatic ducts, and are Notch responsive^{377,378}. In addition, they are known to express *sox9b*,

⁵ This work was performed jointly by Hannah E. Edelman and Sarah A. McClymont and is in preparation for submission for publication.

a homologue of human *SOX9*, and we have shown that following beta-cell ablation, these *sox9b*-expressing CACs are a source of regenerated beta cells³⁷⁹. Humans also have CACs – terminal intercalated duct cells that express *SOX9*³⁸⁰ and are Notch responsive³⁸¹ – but their role in beta-cell neogenesis is unknown. Given the similarities between these cells in humans and zebrafish, by understanding the mechanisms behind beta-cell neogenesis via CACs in zebrafish, we may be able to exploit these same mechanisms in humans for use in diabetes treatment.

Towards this end, an intriguing target to begin unravelling the mechanisms of beta-cell regeneration is SOX9. The importance of SOX9 in pancreatic identity³⁸², ductal cell identity³⁸³, and pancreatic progenitor identity³⁸⁰ has already been established in mammals. Additionally, we recently found that Sox9b helps maintain the progenitor identity of CACs in zebrafish in that heterozygous loss of *sox9b* results in more efficient differentiation of CACs into beta cells after ablation³⁷⁹. Given its important role in pancreatic progenitor status across species, we are interested in understanding the transcriptional targets of SOX9 (and Sox9b) to elucidate the genetic program behind pancreatic progenitors and beta-cell neogenesis.

We set out to identify effectors of SOX9 transcriptional activity in human pancreatic cells to reveal molecular mechanisms that drive beta-cell neogenesis using high-throughput sequencing methods. To do so, we utilized a human pancreatic adenocarcinoma line, PANC-1, that represents an undifferentiated, ductal pancreas cell population that can be induced to differentiate toward an endocrine fate³⁸⁴. To query the mechanisms of SOX9 regulation in these cells, we integrated RNA-seq and ChIP-seq to identify direct transcriptional targets of SOX9.

We found that SOX9 directly regulates *EPCAM*, which encodes a transmembrane protein expressed in stem and progenitor cells in many epithelial tissues³⁸⁵⁻³⁸⁸ and is an interesting target of future studies about pancreatic progenitor function.

SOX9 modulates the transcription of proliferation and cilia genes in PANC-1 cells

We took a two-step approach to identify direct transcriptional targets of SOX9 in PANC-1 cells; namely, RNA-seq following *SOX9* knockdown and then ChIP-seq to identify SOX9 binding sites. To identify transcripts that are regulated by SOX9 activity, we performed RNA-seq in PANC-1 cells that had been transfected with either a *SOX9* siRNA or a control siRNA. Following *SOX9* siRNA transfection, *SOX9* knockdown was confirmed using both Western blotting (**Appendix Figure 6.1A**) and immunofluorescence of fixed cells (**Appendix Figure 6.1B**). To identify genes regulated by SOX9, we sequenced total RNA extracted from PANC-1 cells transfected with either control or *SOX9* siRNA and identified genes that are differentially expressed between knockdown conditions. We identified 93 differentially expressed genes with 60 genes upregulated and 33 downregulated (**Appendix Figure 6.1C; Appendix Table 6.1**).

To confirm the identity of the top differentially expressed genes, we performed qRT-PCR in siRNA-treated PANC-1 cells for the top five up- and down-regulated genes. In doing so, we confirmed the differential expression of these ten genes, except the downregulated gene *SKIV2L* (**Appendix Figure 6.1D**). While *SKIV2L* demonstrated the largest fold change of the downregulated genes, it was also nearest the acceptable p-value cut-off and thus may be a false positive.

In order to further validate the differential gene expression results, we analyzed publicly available RNA-seq data from 178 pancreatic adenocarcinoma samples in The Cancer Genome Atlas. In these samples, *SOX9* is highly expressed and as such, we expected to observe directions of effect opposite to those observed in our induced knockdown experiments. We clustered our differentially expressed genes based on their expression patterns in the tumor samples and observe that in general, corroborating our knockdown studies, upregulated genes clustered together and were lowly expressed in the tumor samples, and downregulated genes clustered together and with *SOX9* and are highly expressed (**Appendix Figure 6.2A**). Furthermore, at an individual gene level, we correlated *SOX9* expression with each differentially expressed gene to examine how closely the genes are co-regulated in these tumor samples and observed there to be a high degree of correlation between the differentially expressed genes and *SOX9* expression (**Appendix Supplementary Table 6.1**). For example, the upregulated gene with the strongest degree of negative correlation with *SOX9* expression is *CCDC13* ($r = -0.50$), bolstering our observation that *SOX9* negatively regulates this gene (**Appendix Figure 6.2B**). Conversely, expression of the downregulated gene, *ESRP1*, is strongly positively correlated with *SOX9* expression ($r = 0.67$), further suggesting that *SOX9* positively regulates this gene's expression (**Appendix Figure 6.2C**). Overall, these data serve to validate observations from our genome wide RNA-seq analyses and provide additional support for our observed transcriptional targets of *SOX9*.

Finally, to assess the biological consequences of these differentially expressed genes, we explored their individual functions and gene ontology (GO)

terms. Genes that are down-regulated following *SOX9* knockdown include genes with established roles in cancer motility (*ESRP1*³⁸⁹), cell-cell adhesion (*TINAGLI*^{390,391}), obesity and insulin resistance (*RGCC*³⁹²), and cancer stem cell maintenance (*EPCAM*³⁹³). Down-regulated genes were collectively enriched for biological processes associated with Notch signaling and the negative regulation of proliferation (**Appendix Figure 6.1E**), suggesting a role for *SOX9* in negatively regulating proliferation.

As a whole, upregulated genes were enriched for processes associated with cilia development, assembly, and movement (**Appendix Figure 6.1F**), suggesting that *SOX9* typically suppresses these processes. Interestingly, the most highly upregulated genes following *SOX9* knockdown are *LRRC6* and *SPEF1*, which have known roles in regulating primary cilia, important features of ductal epithelial cells^{394–396}. In all, these observations may indicate that *SOX9* serves to restrict the differentiation of pancreas progenitor cells toward an epithelial fate by promoting the expression of genes important for maintaining a progenitor status.

SOX9 binding occurs primarily at transcription start sites and regulates pancreatic functions

We next set out to identify putative *SOX9*-responsive regulatory regions, undertaking anti-*SOX9* ChIP-seq. Following pull-down, sequencing, and alignment, our analysis identified 47,858 *SOX9* binding sites in PANC-1 cells. We then analyzed the sequence underlying the top 1000 most significant *SOX9* binding events to identify transcription factor motifs present at these sites. The top enriched motif was a “tail-to-tail” palindrome with high similarity to the *SOX9* consensus motif

(*Appendix Figure 6.3A*). This result supports previous observations in chondrocytes that SOX9 can function as a homodimer³⁹⁷. The second highest enriched motif matched the binding sequence for FOS::JUN (*Appendix Figure 6.3B*) which has been previously reported to bind in conjunction with SOX9 in chondrocytes³⁹⁸. Similar to findings from other groups^{399,400}, we observe an enrichment of SOX9 binding events at gene promoters (8.8% of SOX9 peaks; *Appendix Figure 6.3C*), with diminishing proportions of SOX9 binding events occurring as the distance to the transcriptional start site increases (*Appendix Figure 6.3D*).

Next, we sought to interrogate the potential functional outcome of SOX9 binding in PANC-1 cells. We performed functional annotation of the genes proximal to SOX9 binding sites and found many GO biological processes that match with SOX9's known roles (*Appendix Figure 6.3E*). These include: 1) endocrine pancreas development, reflecting the known function of SOX9 activity in the pancreas; 2) stem cell maintenance, reflecting the participation of SOX9 in maintaining pancreatic progenitor identity³⁷⁹ and; 3) ossification and osteoblast differentiation, which reflects known functions of SOX9 in bone development^{401,402}. Taken collectively, these data suggest that SOX9 acts in PANC-1 cells to regulate target genes important to pancreas biology, in a manner consistent with previously reported modes of action in chondrocytes.

Direct targets of SOX9 overlap with known pancreatic ductal genes

To narrow down the list of differentially expressed genes to those most biologically relevant, we combined three datasets: 1) SOX9 binding events from our

ChIP-seq experiment to identify direct binding targets of SOX9; 2) genes with enriched expression in ductal cells (versus acinar or endocrine cells) of the adult zebrafish pancreas⁴⁰³ and; 3) genes enriched in zebrafish CACs which can be used as CAC markers³⁷⁷. By combining these three data sets we sought to find genes that were not only direct targets of SOX9 – i.e. those that were differentially expressed following *SOX9* knockdown and had a SOX9 binding site proximal to their promoter – but also relevant to the biology of the zebrafish pancreatic ductal cells, and specifically CACs, not just the molecular underpinnings of human PANC-1 cells.

We observed that the majority of genes upregulated with *SOX9* knockdown are not direct targets of SOX9 (57% of all upregulated genes and 70% of the top ten are not bound by SOX9), while all but one of the top ten genes downregulated with *SOX9* knockdown are direct targets (***Appendix Table 6.2***), and 64% of downregulated genes overall are bound. Several downregulated genes have enriched expression in pancreatic ductal cells while the only upregulated gene with ductal expression is *LRRC6*. Finally, *EPCAM* appears to be a promising candidate gene for further investigation because it is a direct target of SOX9, has enriched expression in zebrafish ductal cells, and is a marker of CACs (***Appendix Table 6.2***).

SOX9 directly regulates expression of EPCAM

Given its promising expression in zebrafish pancreatic ductal cells, specifically CACs, we examined the relationship between SOX9 and *EPCAM*. *EPCAM* is a gene encoding a transmembrane protein that is a well-known marker of epithelium and has roles in the regulation of pancreas progenitor differentiation and cell adhesion^{385,393}. Traces from our RNA-seq and ChIP-seq display the

decreased expression of *EPCAM* with SOX9 knockdown and a large SOX9 binding peak centered on the *EPCAM* transcriptional start site and promoter (**Appendix Figure 6.4A**). Additionally, we performed EPCAM and SOX9 antibody staining in PANC-1 cells transfected with either control or *SOX9* siRNA. In control-transfected cells, SOX9 was expressed in all cell nuclei and EPCAM was expressed at the plasma membrane in all cells, albeit with varying intensity. Following *SOX9* knockdown, SOX9 expression was reduced or absent in nuclei, and EPCAM expression was reduced or absent (**Appendix Figure 6.4B**). These results support the conclusion that SOX9 directly regulates *EPCAM* expression in PANC-1 cells.

Furthermore, as a preliminary validation of the relevance of this relationship *in vivo*, we examined the co-regulation of *SOX9* and *EPCAM* in the TCGA pancreatic adenocarcinoma samples, as above. In doing so, we observe a strong positive correlation of *SOX9* and *EPCAM* expression ($r = 0.58$; **Appendix Figure 6.4C**), suggesting that *in vivo*, as well as *in vitro*, SOX9 is a positive regulator of *EPCAM* expression.

Discussion

We previously demonstrated that Sox9b functions downstream of both Notch signaling and Retinoic Acid signaling in CACs³⁷⁹. All of these pathways are responsible for maintaining the progenitor identity of CACs. Because understanding the balance between progenitor maintenance and endocrine differentiation is central to characterizing beta-cell neogenesis, we sought to further elucidate the differentiation process by finding the downstream targets of Sox9b to better

understand how it functions as a central transcription factor in CAC progenitor identity.

By performing RNA-seq and ChIP-seq on PANC-1 cells we were able to identify both direct targets (change in expression with SOX9 knockdown and SOX9 binding peak near promoter) and indirect targets (change in expression with SOX9 knockdown but no SOX9 binding peak) of SOX9. Because a number of the direct targets – like EpCAM and ESRP1 – are associated with epithelial cells, we wondered if a decrease in these genes results in an epithelial-to-mesenchymal transition (EMT) and this change in cell state alters the overall transcriptional profile of the PANC-1 cells. But simple EMT induction does not explain the changes in transcription we saw - looking at EMT markers by Western, IF, and qPCR showed that not only was there no EMT, but PANC-1 cells are already mesenchymal in nature before SOX9 knockdown (*Appendix Supplementary Figure 6.1*). So a change to a more mesenchymal phenotype is not responsible for the transcriptional differences seen in SOX9 knockdown. Because Sox9b is known to help CACs maintain their progenitor identity³⁷⁹, it is possible that knockdown of SOX9 in PANC-1 cells induces differentiation. This idea fits with our RNA-seq results – knockdown of SOX9 caused an increase in the expression of many ciliary genes which is a known consequence of differentiation⁴⁰⁴.

GO analysis of genes downregulated with SOX9 knockdown showed that SOX9 is important for the positive regulation of Notch receptor targets. This aligns well with previous research, as the role of SOX9 in mediating the expression of downstream Notch signaling targets is well-established^{404–407}. Understanding this

intricate cross-talk between SOX9 and Notch will be important to being able to manipulate progenitor differentiation to endocrine cells^{379,408}.

In comparison to previous work in other cell systems, we were able to easily see the SOX9 binding consensus sequence by pulling out the most enriched sequence at binding peaks – unlike the results of ChIP-seq performed in hair follicle stem cells⁴⁰⁰. They found that SOX9 was promiscuous in the sequences it bound to and never saw the consensus sequence. In addition, we saw that SOX9 preferentially bound to promoters, no matter the target gene, unlike what was seen in developing chondrocytes where SOX9 bound promoters for normal cellular functions and enhancers for chondrocyte-specific transcription⁴⁰⁹. These differences suggest that SOX9 may vary in how it acts as a transcription factor between different cell types.

The direct targets of SOX9 that we found have some interesting connotations for both the ductal nature of CACs and their progenitor role. TINAGL1 is a matricellular protein that is known to bind integrins and laminins and is involved in both post-implantation development in the uterus^{391,410} and vascular smooth muscle adhesion^{390,411,412}. Its homolog TINAG is important in tubulogenesis in the kidney⁴¹³. CACs are known to line the branching ducts of the pancreas, so an extracellular molecule important for cell-adhesion (and the cell-cell signaling that goes along with it) makes biological sense for having an important SOX9-mediated role. ESRP1 is a regulator of splicing and is known for predicting favorable outcomes in cancer – i.e. its expression (and therefore splicing patterns of things like *FGFR2*) is thought to suppress cancer cell motility and therefore metastasis^{389,414–416}. This would make biological sense in the context of CACs and regeneration – expression of Sox9b and therefore ESRP1 could promote the epithelial nature of CACs while

loss of these proteins would promote a more migratory, mesenchymal nature which is necessary for CACs to move to ablated islets and become endocrine cells. All of these genes would be interesting targets to follow up on by examining their expression in the zebrafish pancreas, developing knockouts, and assessing their importance in CAC maintenance and biology.

Finally, we showed that SOX9 regulates the expression of EpCAM, a gene that encodes a single-pass transmembrane glycoprotein that localizes to tight junctions³⁸⁵. In addition to its role in cell adhesion, EpCAM can also undergo proteolytic cleavage to produce intra- and extracellular fragments that participate in signaling and transcriptional activation³⁹³. Although EpCAM is expressed broadly in epithelial tissues, higher expression is found in cells that are actively undergoing proliferation and differentiation events. Thus, EpCAM serves as a marker of stem and progenitor cell populations in the intestine, liver, and salivary gland^{386–388}. It is also expressed in human and murine embryonic stem cells, where it may function upstream of the well-known pluripotency factors OCT4 and SOX2^{417,418}. In the human fetal pancreas, EpCAM expression is enhanced in progenitor cells that are budding from the ductal epithelium to form new endocrine cells³⁸⁵. In the adult pancreas, it is most highly expressed in intercalated ducts³⁸⁵, which closely resemble CACs and have been proposed to serve as an endocrine progenitor population during regeneration³⁷⁸. Intriguingly, transgenic over-expression of EpCAM in the mouse pancreas provokes the development of large endocrine islets⁴¹⁹. Within the developing zebrafish liver EpCAM is expressed in Notch responsive progenitor cells called Biliary Epithelial Cells (BECs) that act as progenitors for hepatocytes during liver damage and development^{420,421}. Many

parallels can be drawn between BECs and the CACs of the pancreas. These cell types both line ducts, and Notch signaling as well as Sox9 play an important role in maintaining the progenitor status of both. Using CRISPR-Cas9, we are developing our own EpCam knockout zebrafish with the same mutation as humans with congenital tufting enteropathy. We are hoping to use this knockout to examine the pancreatic and CAC phenotype when EpCam is lost. Because EpCAM is expressed in the mammalian pancreas and in the CACs of zebrafish, understanding how EpCAM acts downstream of Sox9 in the maintenance and identity of endocrine progenitor cells will help us better define the mechanisms behind endogenous increases in beta-cell mass.

Materials and Methods

RNA-seq library preparation and sequencing

PANC-1 cells were transfected in 24-well plates with either 25 nM of control siRNA (Dharmacon catalog # D-001210-03-05) or 25 nM of *SOX9* siRNA (Dharmacon catalog # M-021507-00-0005) using Lipofectamine 3000 (Thermo). After 48 hours, transfected cells were pooled (4 wells per replicate) and harvested, two replicates from each siRNA condition, and total RNA was isolated using a Qiagen RNeasy kit. RNA-seq libraries were created using the Illumina TruSeq Stranded Total RNA Sample Prep Kit. RNA-seq libraries were pooled and sequenced on the Illumina HiSeq 2000 to a minimum depth of 60 million 2x100 bp reads per library.

RNA-seq alignment, quantification, and analysis

Reads were aligned to hg19 genome with HISAT2 (v2.0.5)²⁰⁵ and visualized with the Integrative Genomics Viewer^{172,173}. Statistical analyses were performed using R²²¹. Gene expression was quantified using the “featureCount” function of the Rsubread package (v1.28.1)²⁰⁶ to count read overlap with RefSeq genes. Genes with greater than one read across all four samples were submitted to DESeq2 (v1.18.1)⁴²² to identify genes differentially expressed across conditions (absolute(log₂(foldchange)) > 1, adjusted p-value < 0.05). To generate the volcano plot, each gene’s log₂(fold change) was plotted against the -log₁₀(adjusted p-value), with genes meeting our criteria for significantly differentially expressed being plotted in red (upregulated) or blue (downregulated). Genes significantly up- and down-regulated were submitted to Enrichr^{423,424}. The GO Biological Process was ranked on the basis of the combined score. Differentially expressed genes were annotated as being directly bound by SOX9 if there was a SOX9 binding event within 1kb (upstream or downstream) of the transcriptional start site (RefSeq, hg19).

ChIP-seq library preparation and sequencing

PANC-1 cells were grown in DMEM supplemented with 10% FBS at 37°C with 5.0% CO₂ and passaged at 70-80% confluency. 2 biological replicates of ChIP-seq were performed essentially as in Lee *et al.*, 2006⁴²⁵. Briefly; approximately 2.0 X 10⁸ cells were crosslinked in 11% formaldehyde and stopped with 2.5 M glycine before being washed in 1X PBS, lysed, and sonicated for 35 minutes in a Bioruptor at 4°C to achieve a fragment size of approximately 200 bp. An input fraction was set

aside and the rest of the lysate was then incubated with 10 μ g anti-SOX9 (AB5535, Millipore) overnight at 4°C. Antibody-bound chromatin was then purified using Protein G Dynabeads (Thermo) and crosslinking was reversed overnight at 65°C. ChIP-seq libraries were created using the Illumina TruSeq DNA Sample Prep Kit and quantified using Quant-iT PicoGreen dsDNA assay (Invitrogen). ChIP-seq libraries were pooled and sequenced on the Illumina HiSeq to a minimum depth of 59 million, 1x50bp reads per library.

ChIP-seq alignment, peak calling, and analysis

Reads were aligned to hg19 with Bowtie2 (v2.2.5)¹⁹⁸ in --local mode following TruSeq adapter removal and quality filtering with fastx toolkit (v0.0.14). Following alignment, reads with mapping score < MAPQ30, reads aligning to the mitochondria, and duplicate reads were removed with SAMtools (v1.3.1)¹⁹⁹. ChIP-seq replicates were combined and peaks were called on this joint file with MACS2 (v2.1.1.20160309)²⁰⁰ using “callpeak”. Peaks with q-value >10⁻³ and those overlapping ENCODE blacklists were removed¹⁹⁵. These peaks were annotated for their genomic location using CEAS (v1.0.0)²⁰¹ of the Cistrome analysis pipeline²⁰² and the distance of each peak to the nearest gene’s transcriptional start site was quantified. The top 1000 most significant SOX9 peaks by q-value were submitted to SeqPos (v1.0.0) under default parameters. The top resulting position weight matrices were matched to motifs in the JASPAR database⁴²⁶. These same 1000 peaks were submitted to GREAT (v3.0.0)²¹⁸ under default settings except that the association rule was expanded such that “proximal” was defined as 5kb both upstream and downstream. The top 20 GO Biological Process terms by binomial rank were chosen for display and manually grouped by function.

Western blot confirmation of SOX9 knockdown

PANC-1 cells were cultured in 6-well plates and transfected with 100 nM control siRNA (catalogue number above) or 100 nM *SOX9* siRNA (catalogue number above). After 48 hours, cells were washed with PBS, isolated in RIPA buffer with complete, EDTA-free protease inhibitor (Roche) and vortexed. Supernatant was collected after centrifugation. Protein concentration was determined using the Pierce BCA Protein Assay Kit (Thermo Scientific) and 10 µg of protein was run on an Any kD Mini-PROTEAN TGX Precast Protein Gel (Bio-Rad). Transferred at 45V for 90 minutes. Membrane blocked for 1 hour and incubated overnight in primary antibody (Rabbit anti-Sox9 Santa Cruz sc-20095 1:500; N-Cadherin D4R1H XP Rabbit mAb Cell Signaling Technology 13116 1:1000; E-Cadherin 24E10 Rabbit mAb Cell Signaling Technology 3195 1:1000), washed 3 times, and incubated in anti-rabbit HRP (Cell Signaling 7074S 1:2500) for 1 hour. Developed using SuperSignal West Dura Extended Duration Substrate (Thermo Scientific) and exposed on a ChemiDoc-It². Stripped membrane using Restore Western Blot Stripping Buffer (Thermo Scientific) and repeated staining and development as above using rabbit anti-beta-tubulin (Cell Signaling 2128 1:1000) primary.

Antibody staining

PANC-1 cells were grown on gelatin-coated coverslips for 48 hours after siRNA transfection and fixed in 4% paraformaldehyde buffered in 1X PBS. Following 4X5 min washes in 1X PBS, coverslips were blocked in PBST + 10% FBS for 1 hour at room temperature and permeablized in 0.5% Triton in PBS for 20 minutes, incubated with primary antibodies (rabbit anti-SOX9 Santa Cruz sc-20095

1:250; mouse anti-EPCAM Santa Cruz sc-66020 1:100; vimentin D21H3 XP Rabbit mAb Cell Signaling Technology 5741 1:100) at 4°C overnight. Coverslips were washed 4X5 min in blocking, then incubated in secondary antibody (Alexa Fluor 488 donkey anti-rabbit, Alexa Fluor 488 donkey anti-mouse, Cy3 donkey anti-rabbit, Alexa Fluor 647 donkey anti-rabbit, all 1:500, Jackson ImmunoResearch 711-546-152, 715-456-150, 711-166-152, 711-606-152 respectively) at 4°C overnight before 4X5 min final PBST washes and a brief DAPI (1:2500 in PBS) stain. Images were collected using a Nikon A1-si Laser Scanning Confocal microscope.

Quantitative PCR confirmation of differentially expressed genes

PANC-1 cells were cultured in 12-well plates and transfected with 100 nM control siRNA (catalogue number above) or 100 nM *SOX9* siRNA (catalogue number above). After 48 hours, RNA was isolated using Qiagen RNeasy Kit (with DNase digestion step) and cDNA was synthesized using Superscript III (Thermo) with random hexamer primers. 3 biological replicates of the quantitative PCR reactions were run in technical triplicate following the default SYBR green cycling conditions on an Applied Biosystems Viia 7 using 2x Power SYBR Green Master Mix (Applied Biosystems). Expression was calculated using the $\Delta\Delta CT$ method normalized to *GAPDH* expression and control siRNA transfected cells. Primer sequences can be found in **Appendix Supplementary Table 6.2**^{389,427,428}.

Correlation with TCGA pancreatic adenocarcinoma expression patterns

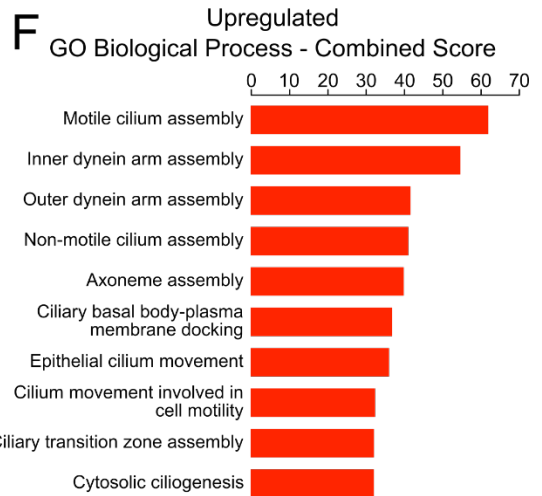
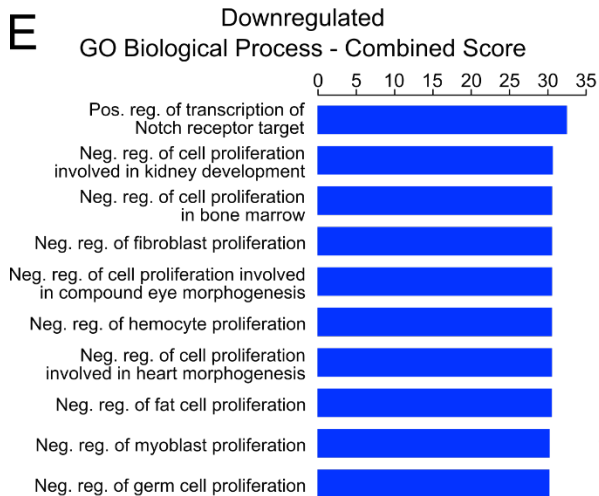
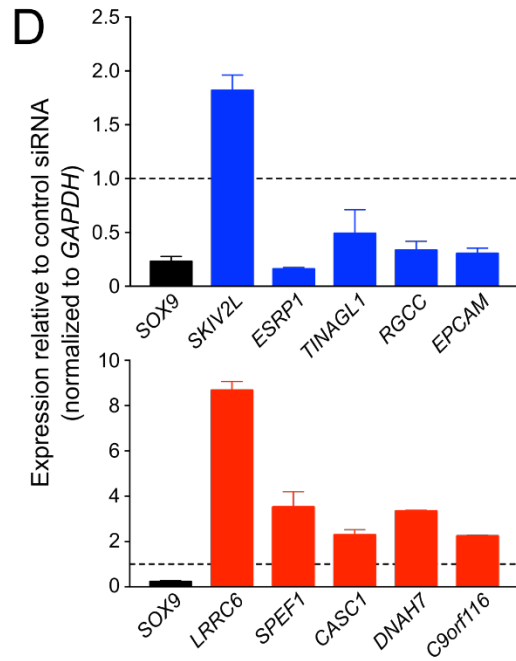
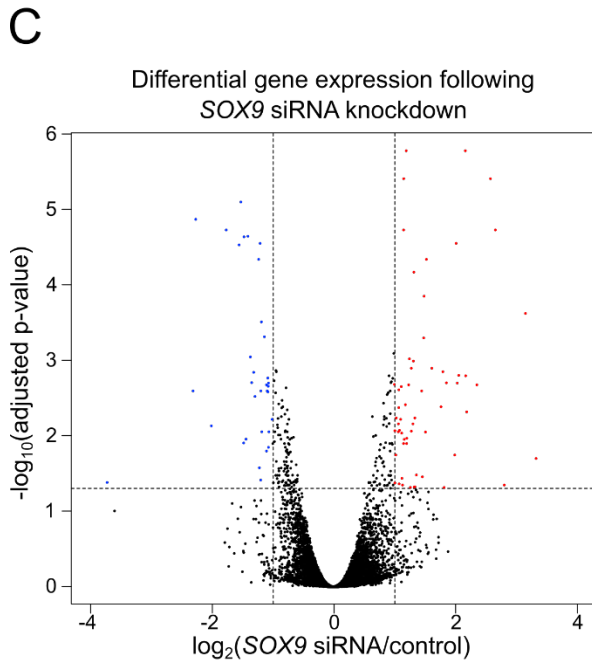
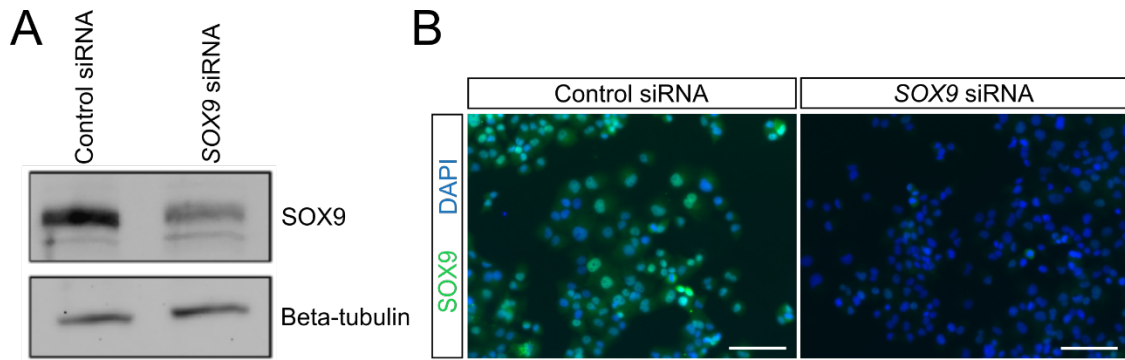
mRNA-sequencing from 178 pancreatic adenocarcinoma samples from the TCGA was accessed and downloaded on March 2, 2018. Read-counts were log₂ normalized after addition of a pseudocount and a heatmap⁴²⁹ was generated for all

significantly upregulated and downregulated genes and SOX9's expression patterns in these samples, using hierarchical clustering to group these selected genes into three clusters, and scaling the expression values by column. Individual gene correlation with SOX9 expression was calculated using the Pearson correlation method and for visualization, normalized expression values were plotted.

Figures and supplementary materials

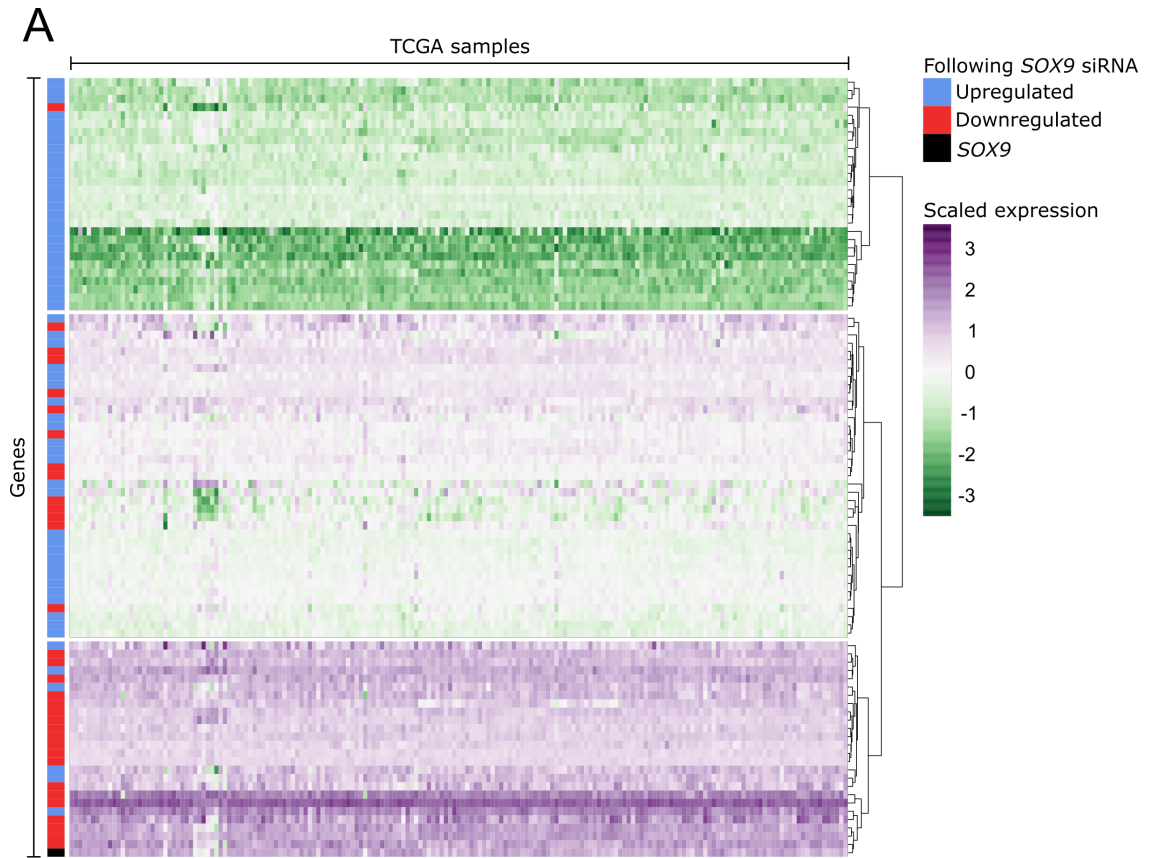
Appendix Figure 6.1: Knockdown of SOX9 results in an increase of ciliary gene expression and a decrease in expression of genes negatively regulating proliferation

(A) Western blot and (B) immunofluorescence confirms knockdown of SOX9 protein following *SOX9* siRNA treatment. Scale bar is 100 um. (C) A volcano plot of adjusted p-value versus fold change upon SOX9 knockdown indicates that 93 genes exhibit significantly altered expression (33 decreased and 60 increased). (D) Quantitative PCR confirms all but one (*SKIV2L*) of the top five upregulated and top five downregulated genes observed with RNA-seq. 3 biological replicates per gene, error bars represent standard deviation (E) GO analysis of downregulated genes reveals a role for SOX9 in Notch signaling as well as in the regulation of proliferation. (F) GO analysis of upregulated genes is enriched for ciliary development and function. Neg, Negative; Pos, Positive; reg, regulation.

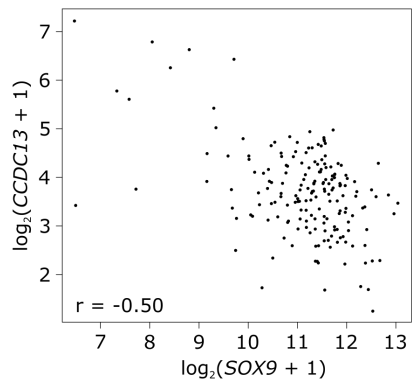


Appendix Figure 6.2: Expression data from pancreatic adenocarcinoma samples corroborate the expression patterns seen in PANC-1 cells

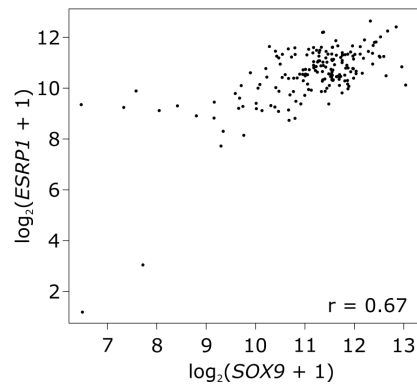
(A) Expression analysis of the TCGA pancreatic adenocarcinoma sample data set indicate that upregulated genes identified by RNA-seq in PANC-1 cells tend to cluster together and are lowly expressed in these samples. Further, downregulated genes identified by RNA-seq in PANC-1 cells tend to cluster together and with *SOX9* and are highly expressed. (B) At an individual gene level, *CCDC13*, a gene that is upregulated following *SOX9* knockdown, is negatively correlated with *SOX9* expression, suggesting that *SOX9* negatively regulates this gene. (C) Conversely, *ESRP1*, a gene that is downregulated following *SOX9* knockdown, is positively correlated with *SOX9* expression, suggesting that *SOX9* positively regulates this gene.



B Correlating *SOX9* expression with upregulated gene, *CCDC13*

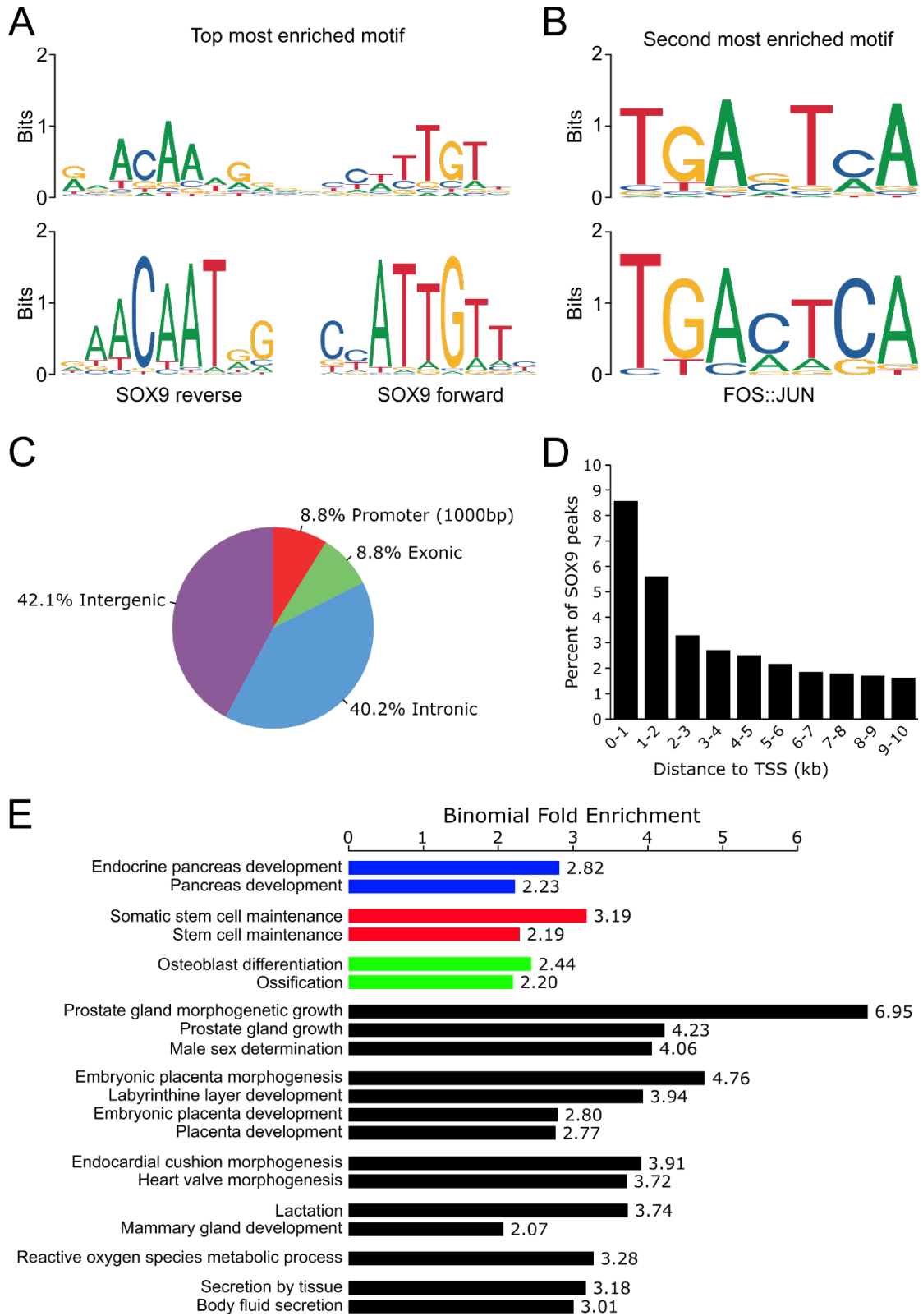


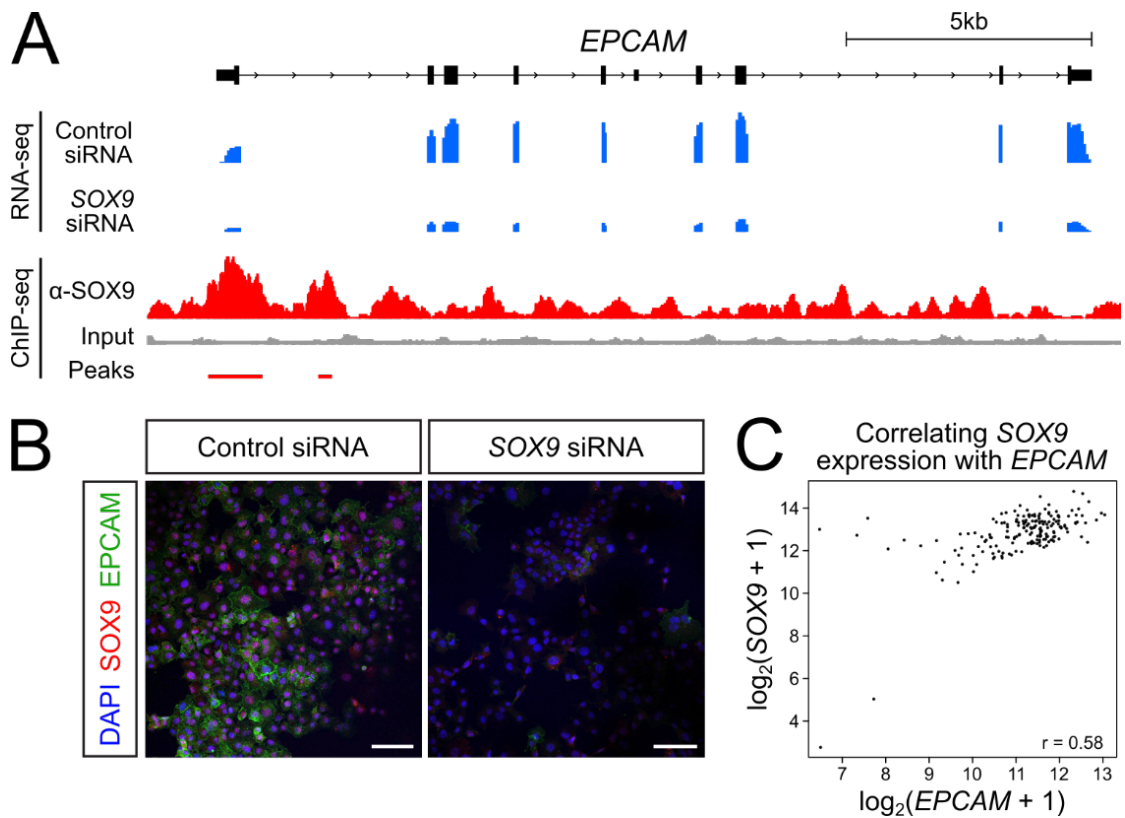
C Correlating *SOX9* expression with downregulated gene, *ESRP1*



Appendix Figure 6.3: Characteristics of SOX9 gene regulation include promoter proximal binding and regulation of genes important to pancreatic biology

(A) The most common motif seen at SOX9 binding sites is a homodimer of the forward and reverse known SOX9 binding sequences (tail to tail). (B) The second most common motif at SOX9 binding sites is recognized by FOS::JUN. (C) SOX9 binding is enriched at promoters (≤ 1000 base-pairs upstream of a transcription start site). (D) As distance from the transcriptional start site increases, the proportion of SOX9 binding events decreases. (E) The nearest genes to SOX9 binding sites are enriched for GO terms related to known SOX9 biology, including endocrine pancreas functions.





Appendix Figure 6.4: RNA-seq and ChIP-seq in combination reveal *EPCAM* to be a direct target of *SOX9*

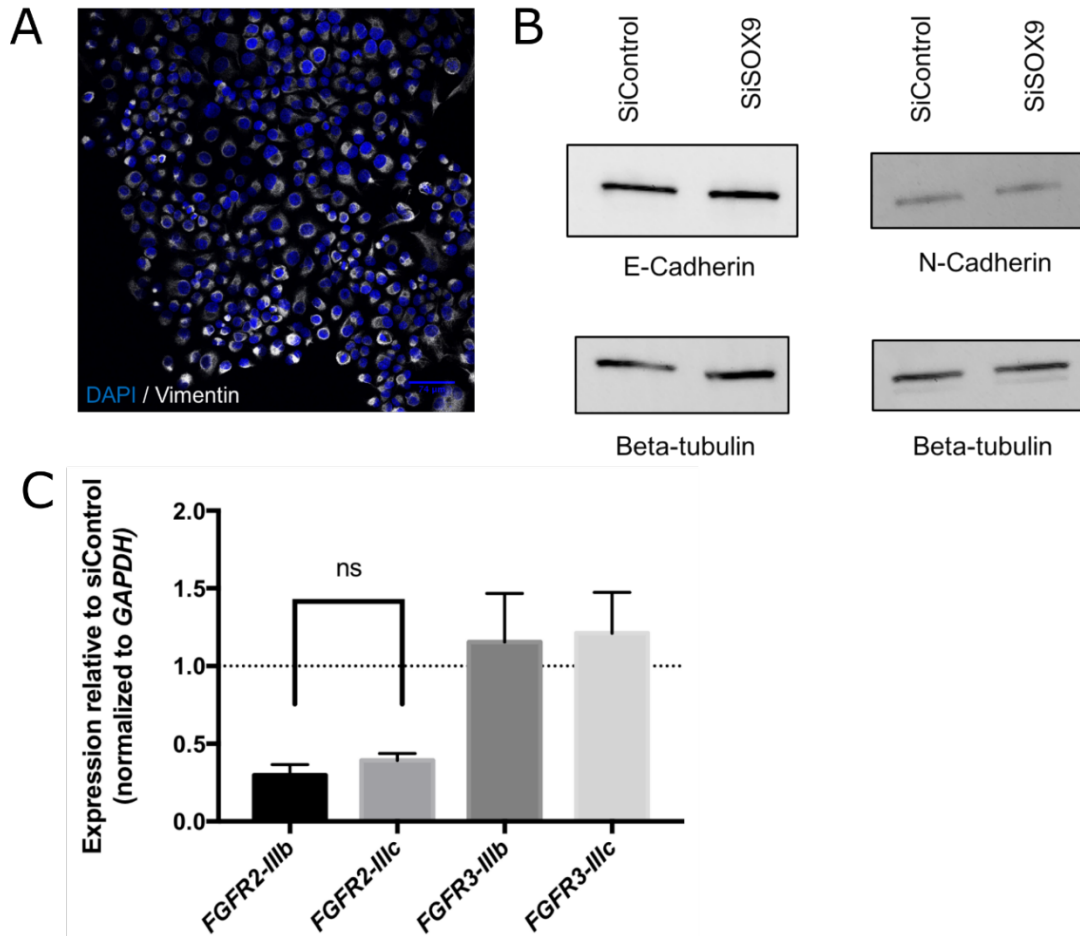
(A) Traces corresponding to RNA-seq and ChIP-seq reads near the *EPCAM* gene confirm *SOX9* knockdown and reveal *SOX9* binding at the *EPCAM* transcription start site. (B) Immunofluorescence staining of PANC-1 cells confirms knockdown of both *SOX9* and *EPCAM* protein by si*SOX9*. Scale bar is 100 μm . (C) Expression of *EPCAM* and *SOX9* in TCGA pancreatic adenocarcinoma samples are positively correlated, suggesting that *EPCAM* expression is upregulated by *SOX9*.

Appendix Table 6.1: Top 10 upregulated and downregulated genes following SOX9 siRNA knockdown

	Gene	log₂ (SOX9 siRNA/control)	p-value (adjusted)
Upregulated	<i>LRRC6</i>	3.332	1.98E-02
	<i>SPEF1</i>	3.158	2.36E-04
	<i>CASC1</i>	2.809	4.45E-02
	<i>DNAH7</i>	2.663	1.85E-05
	<i>C9orf116</i>	2.583	3.86E-06
	<i>CFAP69</i>	2.358	2.09E-03
	<i>CCDC13</i>	2.191	4.77E-03
	<i>AXIN2</i>	2.175	1.59E-03
	<i>C6orf165</i>	2.169	1.64E-06
	<i>ROPN1L</i>	2.061	1.56E-03
Downregulated	<i>TUB</i>	-1.437	1.09E-02
	<i>LIN7C</i>	-1.470	2.28E-05
	<i>SLC45A3</i>	-1.475	1.23E-02
	<i>DDAH1</i>	-1.522	7.86E-06
	<i>CENPM</i>	-1.550	2.91E-05
	<i>EPCAM</i>	-1.762	1.85E-05
	<i>RGCC</i>	-2.006	7.30E-03
	<i>TINAGL1</i>	-2.264	1.33E-05
	<i>ESRP1</i>	-2.307	2.52E-03
	<i>SKIV2L</i>	-3.721	4.11E-02

Appendix Table 6.2: Assessing the top differentially expressed genes for evidence of direct SOX9 regulation, expression in pancreatic ductal cells, and ability to mark CACs

	Gene	Direct target of SOX9?	Enriched ductal expression?	CAC marker?
Upregulated	<i>LRRC6</i>	Yes	Yes	No
	<i>SPEF1</i>	Yes	No	No
	<i>CASC1</i>	No	No	No
	<i>DNAH7</i>	No	No	No
	<i>C9orf116</i>	No	No	No
	<i>CFAP69</i>	Yes	No	No
	<i>CCDC13</i>	No	No	No
	<i>AXIN2</i>	No	No	No
	<i>C6orf165</i>	No	No	No
	<i>ROPN1L</i>	No	No	No
Downregulated	<i>TUB</i>	Yes	No	No
	<i>LIN7C</i>	Yes	No	No
	<i>SLC45A3</i>	Yes	No	No
	<i>DDAH1</i>	Yes	No	No
	<i>CENPM</i>	Yes	No	No
	<i>EPCAM</i>	Yes	Yes	Yes
	<i>RGCC</i>	No	Yes	No
	<i>TINAGL1</i>	Yes	Yes	No
	<i>ESRP1</i>	Yes	Yes	No
	<i>SKIV2L</i>	Yes	No	No




Appendix Supplementary Figure 6.1: Changes in gene expression from SOX9 knockdown do not correspond with changes in cell state

(A) Immunofluorescence for the mesenchymal marker vimentin is present uniformly in PANC-1 cells. (B) There is no change in N-cadherin or E-cadherin expression by Western blot between PANC-1 cells treated with siControl versus siSOX9. (C) Quantitative PCR shows that while there is a general decrease in expression of *FGFR2* with SOX9 knockdown, there is no change in the ratio between its splice isoforms or the splice isoforms of *FGFR3*, indicative of no EMT.^{60,61} Three biological replicates per gene, error bars represent standard deviation.

Appendix Supplementary Table 6.1: Correlation of SOX9 expression with the top ten up- and down-regulated genes

	Gene	Pearson correlation
Upregulated	<i>LRRC6</i>	0.21
	<i>SPEF1</i>	-0.13
	<i>CASC1</i>	-0.14
	<i>DNAH7</i>	-0.02
	<i>C9orf116</i>	0.07
	<i>CFAP69</i>	-0.22
	<i>CCDC13</i>	-0.50
	<i>AXIN2</i>	-0.17
	<i>C6orf165</i>	-0.31
	<i>ROPN1L</i>	-0.35
Downregulated	<i>TUB</i>	-0.56
	<i>LIN7C</i>	0.04
	<i>SLC45A3</i>	0.48
	<i>DDAH1</i>	0.35
	<i>CENPM</i>	0.24
	<i>EPCAM</i>	0.58
	<i>RGCC</i>	-0.30
	<i>TINAGL1</i>	0.57
	<i>ESRP1</i>	0.67
	<i>SKIV2L</i>	0.03



Appendix Supplementary Table 6.2: List of primers used for qPCR

Gene	Forward	Reverse	Source
GAPDH	GCACCGTCAAGGCTGAGAAC	ATGGTGGTGAAGACGCCAGT	Ishii <i>et al.</i> 2014
SOX9	GGGCACCGCCTCTACTCCA	TCCCAGTGCTGGGGCTGT	
LRRC6	CGCCATGGGCTGGATCACAGAA	ATGCAACGAGAGTTCCTCCAGGG	
SPEF1	AGCGATGGAGTCCTTGTTCAGAG	TGGAGAGAGTTGGCGGGGACA	
CASC1	AGGTGTTTTCTGAAGCAGAGA	AGGATCAGGACTCCCATCACA	
DNAH7	TGCCTCTATCGTCCTAGGGG	GCTGGCCGATTTATCCTGCT	
C9orf116	GGAGAGGACCAGCGACTACT	ACAGCCTTCTGGGTCCTGTA	
SKIV2L	CGGGAGCGAATGCAGATACA	GTTCCGAGCACCTCTACTCG	
ESRP1	CAATATTGCCAAGGGAGGTG	GTCCCCATGTGATGTTTGTG	Ishii <i>et al.</i> 2014
TINAGL1	TCCCAAACAGCAGTTGGATGTA	GTTTCTTGGTACACTGCCA	
RGCC	GCACCTGGAGCGCATGAAGC	TGAATCTGCACTCTCCGAGTCGCT	
EPCAM	CGCGTTCGGGCTTCTGCTTG	ATTTGGCAGCCAGCTTTGAGCA	
FGFR2-IIIb	CGTGGAAAAGAACGGCAGTAAATA	GAACTATTTATCCCCGAGTGCTTG	Ranieri <i>et al.</i> 2015
FGFR2-IIIc	TGAGGACGCTGGGAATATACG	TAGTCTGGGGAAGCTGTAATCTCCT	Ranieri <i>et al.</i> 2015
FGFR3-IIIb	TCAAGTCCTGGATCAGTGAGAGT	AGGAAGAAGCCCACCCCG	Tomlinson <i>et al.</i> 2005
FGFR3-IIIc	GAGTTCCTGCAAGGTGTACAGT	GAGAGAACCCTCTAGCTCCTTGTCG	Tomlinson <i>et al.</i> 2005

6.2 References

1. de Lau, L. M. L. & Breteler, M. M. B. Epidemiology of Parkinson's disease. *Lancet. Neurol.* **5**, 525–535 (2006).
2. Fahn, S. Description of Parkinson's disease as a clinical syndrome. *Ann. N. Y. Acad. Sci.* **991**, 1–14 (2003).
3. Postuma, R. B. *et al.* Identifying prodromal Parkinson's disease: pre-motor disorders in Parkinson's disease. *Mov. Disord.* **27**, 617–626 (2012).
4. Fearnley, J. M. & Lees, A. J. Ageing and Parkinson's disease: substantia nigra regional selectivity. *Brain* **114**, 2283–2301 (1991).
5. Corti, O., Lesage, S. & Brice, A. What genetics tells us about the causes and mechanisms of Parkinson's disease. *Physiol. Rev.* **91**, 1161–1218 (2011).
6. Kumar, K. R., Djarmati-Westenberger, A. & Grunewald, A. Genetics of Parkinson's disease. *Semin. Neurol.* **31**, 433–440 (2011).
7. Polymeropoulos, M. H. *et al.* Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* **276**, 2045–2047 (1997).
8. Funayama, M. *et al.* A new locus for Parkinson's disease (PARK8) maps to chromosome 12p11.2-q13.1. *Ann. Neurol.* **51**, 296–301 (2002).
9. Zimprich, A. *et al.* Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. *Neuron* **44**, 601–607 (2004).
10. Zimprich, A. *et al.* A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset Parkinson disease. *Am. J. Hum. Genet.* **89**, 168–175 (2011).
11. Kitada, T. *et al.* Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. *Nature* **392**, 605–608 (1998).
12. Bonifati, V. *et al.* DJ-1(PARK7), a novel gene for autosomal recessive, early onset parkinsonism. *Neurol. Sci. Off. J. Ital. Neurol. Soc. Ital. Soc. Clin. Neurophysiol.* **24**, 159–160 (2003).
13. Valente, E. M. *et al.* Hereditary early-onset Parkinson's disease caused by mutations in PINK1. *Science* **304**, 1158–1160 (2004).
14. Kiely, A. P. *et al.* alpha-Synucleinopathy associated with G51D SNCA mutation: a link between Parkinson's disease and multiple system atrophy? *Acta Neuropathol.* **125**, 753–769 (2013).
15. Kruger, R. *et al.* Ala30Pro mutation in the gene encoding alpha-synuclein in Parkinson's disease. *Nature Genetics* **18**, 106–108 (1998).
16. Zarranz, J. J. *et al.* The new mutation, E46K, of alpha-synuclein causes Parkinson and Lewy body dementia. *Ann. Neurol.* **55**, 164–173 (2004).
17. Pasanen, P. *et al.* Novel alpha-synuclein mutation A53E associated with atypical multiple system atrophy and Parkinson's disease-type pathology. *Neurobiol. Aging* **35**, 2180.e1-5 (2014).
18. Proukakis, C. *et al.* A novel alpha-synuclein missense mutation in Parkinson disease. *Neurology* **80**, 1062–1064 (2013).
19. Chartier-Harlin, M.-C. *et al.* Alpha-synuclein locus duplication as a cause of familial Parkinson's disease. *Lancet (London, England)* **364**, 1167–1169 (2004).

20. Singleton, A. B. *et al.* alpha-Synuclein locus triplication causes Parkinson's disease. *Science* **302**, 841 (2003).
21. Singleton, A. & Gwinn-Hardy, K. Parkinson's disease and dementia with Lewy bodies: a difference in dose? *Lancet (London, England)* **364**, 1105–1107 (2004).
22. Inzelberg, R. *et al.* Onset and progression of disease in familial and sporadic Parkinson's disease. *Am. J. Med. Genet. A* **124A**, 255–258 (2004).
23. Chang, D. *et al.* A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat. Genet.* **49**, 1511–1516 (2017).
24. Nalls, M. A. *et al.* Expanding Parkinson's disease genetics: novel risk loci, genomic context, causal insights and heritable risk. *bioRxiv* 388165 (2019).
25. Nalls, M. A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* **46**, 989–993 (2014).
26. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
27. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
28. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).
29. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
30. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
31. Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116–120 (2012).
32. Bulger, M. & Groudine, M. Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev. Biol.* **339**, 250–257 (2010).
33. Tjian, R. The binding site on SV40 DNA for a T antigen-related protein. *Cell* **13**, 165–179 (1978).
34. Farley, E. K., Olson, K. M., Zhang, W., Rokhsar, D. S. & Levine, M. S. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 6508–6513 (2016).
35. Crocker, J., Noon, E. P.-B. & Stern, D. L. The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution. *Curr. Top. Dev. Biol.* **117**, 455–469 (2016).
36. Grossman, S. R. *et al.* Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E1291–E1300 (2017).
37. Stargell, L. A. & Struhl, K. Mechanisms of transcriptional activation in vivo: two steps forward. *Trends Genet.* **12**, 311–315 (1996).
38. Ptashne, M. & Gann, A. Transcriptional activation by recruitment. *Nature* **386**, 569–577 (1997).
39. Sawado, T., Halow, J., Bender, M. A. & Groudine, M. The beta -globin locus control region (LCR) functions primarily by enhancing the transition from transcription initiation to elongation. *Genes Dev.* **17**, 1009–1018 (2003).

40. Nechaev, S. & Adelman, K. Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation. *Biochim. Biophys. Acta* **1809**, 34–45 (2011).
41. Ptashne, M. Gene regulation by proteins acting nearby and at a distance. *Nature* **322**, 697–701 (1986).
42. de Laat, W. *et al.* Three-dimensional organization of gene expression in erythroid cells. *Curr. Top. Dev. Biol.* **82**, 117–139 (2008).
43. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).
44. Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430–435 (2010).
45. Schmidt, D. *et al.* A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res.* **20**, 578–588 (2010).
46. Phillips-Cremins, J. E. *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281–1295 (2013).
47. Palstra, R.-J. *et al.* The beta-globin nuclear compartment in development and erythroid differentiation. *Nat. Genet.* **35**, 190–194 (2003).
48. Kornberg, R. D. Chromatin structure: a repeating unit of histones and DNA. *Science* **184**, 868–871 (1974).
49. Knezetic, J. A. & Luse, D. S. The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro. *Cell* **45**, 95–104 (1986).
50. Lorch, Y., LaPointe, J. W. & Kornberg, R. D. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell* **49**, 203–210 (1987).
51. Wapinski, O. L. *et al.* Rapid Chromatin Switch in the Direct Reprogramming of Fibroblasts to Neurons. *Cell Rep.* **20**, 3236–3247 (2017).
52. Zaret, K. Developmental competence of the gut endoderm: genetic potentiation by GATA and HNF3/fork head proteins. *Dev. Biol.* **209**, 1–10 (1999).
53. Zaret, K. S. *et al.* Pioneer factors, genetic competence, and inductive signaling: programming liver and pancreas progenitors from the endoderm. *Cold Spring Harb. Symp. Quant. Biol.* **73**, 119–126 (2008).
54. Engelen, E. *et al.* Sox2 cooperates with Chd7 to regulate genes that are mutated in human syndromes. *Nat. Genet.* **43**, 607–611 (2011).
55. Ding, J., Xu, H., Faiola, F., Ma'ayan, A. & Wang, J. Oct4 links multiple epigenetic pathways to the pluripotency network. *Cell Res.* **22**, 155–167 (2012).
56. Takaku, M. *et al.* GATA3-dependent cellular reprogramming requires activation-domain dependent recruitment of a chromatin remodeler. *Genome Biol.* **17**, 36 (2016).
57. Narlikar, G. J., Sundaramoorthy, R. & Owen-Hughes, T. Mechanisms and functions of ATP-dependent chromatin-remodeling enzymes. *Cell* **154**, 490–503 (2013).
58. Kioussis, D., Vanin, E., deLange, T., Flavell, R. A. & Grosveld, F. G. Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. *Nature* **306**, 662–666 (1983).
59. Lettice, L. A. *et al.* Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 7548–7553 (2002).

60. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735 (2003).
61. Weedon, M. N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat. Genet.* **46**, 61–64 (2014).
62. Lauderdale, J. D., Wilensky, J. S., Oliver, E. R., Walton, D. S. & Glaser, T. 3' deletions cause aniridia by preventing PAX6 gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 13755–13759 (2000).
63. Bhatia, S. *et al.* Disruption of autoregulatory feedback by a mutation in a remote, ultraconserved PAX6 enhancer causes aniridia. *Am. J. Hum. Genet.* **93**, 1126–1134 (2013).
64. Smemo, S. *et al.* Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum. Mol. Genet.* **21**, 3255–3263 (2012).
65. Emison, E. S. *et al.* A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature* **434**, 857–863 (2005).
66. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9362–9367 (2009).
67. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.* **373**, 895–907 (2015).
68. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371–375 (2014).
69. Gupta, R. M. *et al.* A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. *Cell* **170**, 522–533.e15 (2017).
70. Kleftogiannis, D., Kalnis, P. & Bajic, V. B. DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res.* **43**, e6 (2015).
71. Lee, D., Karchin, R. & Beer, M. A. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.* **21**, 2167–2180 (2011).
72. Fletez-Brant, C., Lee, D., McCallion, A. S. & Beer, M. A. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic Acids Res.* **41**, W544-56 (2013).
73. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.* **10**, e1003711 (2014).
74. Kim, S. G., Harwani, M., Grama, A. & Chaterji, S. EP-DNN: A Deep Neural Network-Based Global Enhancer Prediction Algorithm. *Sci. Rep.* **6**, 38433 (2016).
75. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**, e7 (2005).
76. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394 (2003).
77. Antonellis, A. *et al.* Identification of neural crest and glial enhancers at the mouse Sox10 locus through transgenesis in zebrafish. *PLoS Genet.* **4**, e1000174 (2008).
78. Pennacchio, L. A. *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
79. Shin, J. T. *et al.* Human-zebrafish non-coding conserved elements act in vivo to

- regulate transcription. *Nucleic Acids Res.* **33**, 5437–5445 (2005).
80. Chen, H., Li, C., Zhou, Z. & Liang, H. Fast-Evolving Human-Specific Neural Enhancers Are Associated with Aging-Related Diseases. *Cell Syst.* **6**, 604–611.e4 (2018).
 81. Moon, J. M., Capra, J. A., Abbot, P. & Rokas, A. Signatures of Recent Positive Selection in Enhancers Across 41 Human Tissues. *G3 (Bethesda)*. **9**, 2761–2774 (2019).
 82. Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K.-D. & Wray, G. A. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat. Genet.* **39**, 1140–1144 (2007).
 83. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
 84. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
 85. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
 86. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21931–21936 (2010).
 87. Ogryzko, V. V, Schiltz, R. L., Russanova, V., Howard, B. H. & Nakatani, Y. The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell* **87**, 953–959 (1996).
 88. May, D. *et al.* Large-scale discovery of enhancers from human heart tissue. *Nat. Genet.* **44**, 89–93 (2011).
 89. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* **38**, 1348–1354 (2006).
 90. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
 91. Schoenfelder, S., Javierre, B.-M., Furlan-Magaril, M., Wingett, S. W. & Fraser, P. Promoter Capture Hi-C: High-resolution, Genome-wide Profiling of Promoter Interactions. *J. Vis. Exp.* (2018).
 92. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
 93. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
 94. Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. & Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* **17**, 877–885 (2007).
 95. 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**, 1213–8 (2013).
 96. Brasier, A. R., Tate, J. E. & Habener, J. F. Optimized use of the firefly luciferase assay as a reporter gene in mammalian cell lines. *Biotechniques* **7**, 1116–1122

- (1989).
97. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
 98. Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C. & Cohen, B. A. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 19498–19503 (2012).
 99. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265–270 (2012).
 100. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
 101. Kothary, R. *et al.* Inducible expression of an hsp68-lacZ hybrid gene in transgenic mice. *Development* **105**, 707–714 (1989).
 102. Rossant, J., Zirngibl, R., Cado, D., Shago, M. & Giguere, V. Expression of a retinoic acid response element-hsplacZ transgene defines specific domains of transcriptional activity during mouse embryogenesis. *Genes Dev.* **5**, 1333–1344 (1991).
 103. Fisher, S. *et al.* Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. *Nat. Protoc.* **1**, 1297–1305 (2006).
 104. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
 105. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
 106. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
 107. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442–451 (2013).
 108. Singh, P. & Schimenti, J. C. The genetics of human infertility by functional interrogation of SNPs in mice. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 10431–10436 (2015).
 109. Soldner, F. *et al.* Parkinson-associated risk variant in distal enhancer of alpha-synuclein modulates target gene expression. *Nature* **533**, 95–99 (2016).
 110. Dickel, D. E. *et al.* Ultraconserved Enhancers Are Required for Normal Development. *Cell* **172**, 491–499.e15 (2018).
 111. Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239–243 (2018).
 112. Towbin, H., Staehelin, T. & Gordon, J. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc. Natl. Acad. Sci. U. S. A.* **76**, 4350–4354 (1979).
 113. Burnette, W. N. ‘Western blotting’: electrophoretic transfer of proteins from sodium dodecyl sulfate--polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. *Anal. Biochem.* **112**, 195–203 (1981).
 114. Engvall, E. & Perlmann, P. Enzyme-linked immunosorbent assay (ELISA). Quantitative assay of immunoglobulin G. *Immunochemistry* **8**, 871–874 (1971).
 115. Bruins, A. P., Covey, T. R. & Henion, J. D. Ion spray interface for combined liquid

- chromatography/atmospheric pressure ionization mass spectrometry. *Anal. Chem.* **59**, 2642–2646 (1987).
116. Bass, J. J. *et al.* An overview of technical considerations for Western blotting applications to physiological research. *Scand. J. Med. Sci. Sports* **27**, 4–25 (2017).
 117. Reeves, J. R. & Bartlett, J. M. Measurement of protein expression a technical overview. *Methods Mol. Med.* **39**, 471–483 (2001).
 118. Greenbaum, D., Colangelo, C., Williams, K. & Gerstein, M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.* **4**, 117 (2003).
 119. Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19**, 1720–1730 (1999).
 120. Alwine, J. C., Kemp, D. J. & Stark, G. R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5350–5354 (1977).
 121. Higuchi, R., Fockler, C., Dollinger, G. & Watson, R. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Biotechnology. (N. Y.)* **11**, 1026–1030 (1993).
 122. Holland, P. M., Abramson, R. D., Watson, R. & Gelfand, D. H. Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 7276–7280 (1991).
 123. Heid, C. A., Stevens, J., Livak, K. J. & Williams, P. M. Real time quantitative PCR. *Genome Res.* **6**, 986–994 (1996).
 124. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242–2246 (2004).
 125. Yamada, K. *et al.* Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* **302**, 842–846 (2003).
 126. Selinger, D. W. *et al.* RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nat. Biotechnol.* **18**, 1262–1268 (2000).
 127. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
 128. Wilhelm, B. T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239–1243 (2008).
 129. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
 130. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523–536 (2008).
 131. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**, 613–619 (2008).
 132. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
 133. Tang, F. *et al.* RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat. Protoc.* **5**, 516–535 (2010).
 134. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by

- highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
135. Ramskold, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
 136. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).
 137. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
 138. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
 139. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
 140. Kharchenko, P. V, Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
 141. Jolliffe, I. Principal component analysis and factor analysis. Aberdeen. (2002).
 142. Maaten, L. van der & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
 143. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv Prepr. arXiv1802.03426* (2018).
 144. Neftel, C. *et al.* An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell* **178**, 835–849.e21 (2019).
 145. Hovestadt, V. *et al.* Resolving medulloblastoma cellular architecture by single-cell genomics. *Nature* **572**, 74–79 (2019).
 146. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
 147. Zeisel, A. *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
 148. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7285–7290 (2015).
 149. La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* **167**, 566–580.e19 (2016).
 150. Marques, S. *et al.* Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science* **352**, 1326–1329 (2016).
 151. Hook, P. W. *et al.* Single-Cell RNA-Seq of Mouse Dopaminergic Neurons Informs Candidate Gene Selection for Sporadic Parkinson Disease. *Am. J. Hum. Genet.* **102**, 427–446 (2018).
 152. Villani, A.-C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, (2017).
 153. Montoro, D. T. *et al.* A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
 154. Plasschaert, L. W. *et al.* A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
 155. Steuerma, Y. *et al.* Dissection of Influenza Infection In Vivo by Single-Cell RNA Sequencing. *Cell Syst.* **6**, 679–691.e4 (2018).
 156. De Baets, S. *et al.* A GFP expressing influenza A virus to report in vivo tropism and

- protection by a matrix protein 2 ectodomain-specific monoclonal antibody. *PLoS One* **10**, e0121491 (2015).
157. McFadden, G., Mohamed, M. R., Rahman, M. M. & Bartee, E. Cytokine determinants of viral tropism. *Nat. Rev. Immunol.* **9**, 645–655 (2009).
 158. See, K. *et al.* Single cardiomyocyte nuclear transcriptomes reveal a lincRNA-regulated de-differentiation and cell cycle stress-response in vivo. *Nat. Commun.* **8**, 225 (2017).
 159. Nomura, S. *et al.* Cardiomyocyte gene programs encoding morphological and functional signatures in cardiac hypertrophy and failure. *Nat. Commun.* **9**, 4435 (2018).
 160. Grubman, A. *et al.* A single cell brain atlas in human Alzheimer’s disease. *bioRxiv* 628347 (2019).
 161. Ma, S. Y., Roytta, M., Rinne, J. O., Collan, Y. & Rinne, U. K. Correlation between neuromorphometry in the substantia nigra and clinical features in Parkinson’s disease using disector counts. *J. Neurol. Sci.* **151**, 83–87 (1997).
 162. Pringsheim, T., Jette, N., Frolkis, A. & Steeves, T. D. L. The prevalence of Parkinson’s disease: a systematic review and meta-analysis. *Mov. Disord.* **29**, 1583–1590 (2014).
 163. Thomas, B. & Beal, M. F. Parkinson’s disease. *Hum. Mol. Genet.* **16 Spec No**, R183–94 (2007).
 164. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22**, 1748–1759 (2012).
 165. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
 166. Forrest, M. P. *et al.* Open Chromatin Profiling in hiPSC-Derived Neurons Prioritizes Functional Noncoding Psychiatric Risk Variants and Highlights Neurodevelopmental Loci. *Cell Stem Cell* **21**, 305–318.e8 (2017).
 167. Fullard, J. F. *et al.* Open chromatin profiling of human postmortem brain infers functional roles for non-coding schizophrenia loci. *Hum. Mol. Genet.* **26**, 1942–1951 (2017).
 168. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
 169. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).
 170. Praetorius, C. *et al.* A polymorphism in IRF4 affects human pigmentation through a tyrosinase-dependent MITF/TFAP2A pathway. *Cell* **155**, 1022–1033 (2013).
 171. Heintz, N. Gene expression nervous system atlas (GENSAT). *Nat. Neurosci.* **7**, 483 (2004).
 172. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29**, 24–26 (2011).
 173. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
 174. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**,

D88-92 (2007).

175. Stott, S. R. W. *et al.* Foxa1 and foxa2 are required for the maintenance of dopaminergic properties in ventral midbrain neurons at late embryonic stages. *J. Neurosci.* **33**, 8022–8034 (2013).
176. Arenas, E. Foxa2: the rise and fall of dopamine neurons. *Cell Stem Cell* **2**, 110–112 (2008).
177. Prakash, N. & Wurst, W. Development of dopaminergic neurons in the mammalian brain. *Cell. Mol. Life Sci.* **63**, 187–206 (2006).
178. Smits, S. M., Ponnio, T., Conneely, O. M., Burbach, J. P. H. & Smidt, M. P. Involvement of Nurr1 in specifying the neurotransmitter identity of ventral midbrain dopaminergic neurons. *Eur. J. Neurosci.* **18**, 1731–1738 (2003).
179. Hook, P. W. *et al.* Single-Cell RNA-Seq of Mouse Dopaminergic Neurons Informs Candidate Gene Selection for Sporadic Parkinson Disease. *Am. J. Hum. Genet.* **102**, 427–446 (2018).
180. Wang, Y. *et al.* The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol.* **19**, 151 (2018).
181. Zarow, C., Lyness, S. A., Mortimer, J. A. & Chui, H. C. Neuronal loss is greater in the locus coeruleus than nucleus basalis and substantia nigra in Alzheimer and Parkinson diseases. *Arch. Neurol.* **60**, 337–341 (2003).
182. Kasthuber, E., Kratochwil, C. F., Ryu, S., Schweitzer, J. & Driever, W. Genetic dissection of dopaminergic and noradrenergic contributions to catecholaminergic tracts in early larval zebrafish. *J. Comp. Neurol.* **518**, 439–458 (2010).
183. Rink, E. & Wullimann, M. F. The teleostean (zebrafish) dopaminergic system ascending to the subpallium (striatum) is located in the basal diencephalon (posterior tuberculum). *Brain Res.* **889**, 316–330 (2001).
184. Seidel, K. *et al.* The brainstem pathologies of Parkinson’s disease and dementia with Lewy bodies. *Brain Pathol.* **25**, 121–135 (2015).
185. Wakabayashi, K., Mori, F., Tanji, K., Orimo, S. & Takahashi, H. Involvement of the peripheral nervous system in synucleinopathies, tauopathies and other neurodegenerative proteinopathies of the brain. *Acta Neuropathol.* **120**, 1–12 (2010).
186. Wakabayashi, K. & Takahashi, H. Neuropathology of autonomic nervous system in Parkinson’s disease. *Eur. Neurol.* **38 Suppl 2**, 2–7 (1997).
187. Braak, H. *et al.* Amygdala pathology in Parkinson’s disease. *Acta Neuropathol.* **88**, 493–500 (1994).
188. Langston, J. W. & Forno, L. S. The hypothalamus in Parkinson disease. *Ann. Neurol.* **3**, 129–133 (1978).
189. Guella, I. *et al.* alpha-synuclein genetic variability: A biomarker for dementia in Parkinson disease. *Ann. Neurol.* **79**, 991–999 (2016).
190. Biedler, J. L., Roffler-Tarlov, S., Schachner, M. & Freedman, L. S. Multiple neurotransmitter synthesis by human neuroblastoma cell lines and clones. *Cancer Res.* **38**, 3751–3757 (1978).
191. Dutta, G., Zhang, P. & Liu, B. The lipopolysaccharide Parkinson’s disease animal model: mechanistic studies and drug discovery. *Fundam. Clin. Pharmacol.* **22**, 453–464 (2008).
192. Perry, V. H. The influence of systemic inflammation on inflammation in the brain:

- implications for chronic neurodegenerative disease. *Brain. Behav. Immun.* **18**, 407–413 (2004).
193. Preissl, S. *et al.* Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.* **21**, 432–439 (2018).
 194. Greffard, S. *et al.* Motor score of the Unified Parkinson Disease Rating Scale as a good predictor of Lewy body-associated neuronal loss in the substantia nigra. *Arch. Neurol.* **63**, 584–588 (2006).
 195. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
 196. Gong, S. *et al.* A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* **425**, 917–925 (2003).
 197. Westerfeld, M. *The Zebrafish Book. A Guide for the Laboratory Use of Zebrafish (Danio rerio)*. (Univ. Oregon Press, 2007).
 198. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
 199. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
 200. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
 201. Shin, H., Liu, T., Manrai, A. K. & Liu, X. S. CEAS: cis-regulatory element annotation system. *Bioinformatics* **25**, 2605–2606 (2009).
 202. Liu, T. *et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.* **12**, R83 (2011).
 203. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493-6 (2004).
 204. Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160-5 (2016).
 205. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
 206. Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41**, e108 (2013).
 207. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).
 208. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
 209. Mudge, J. M. & Harrow, J. Creating reference gene annotation for the mouse C57BL6/J genome assembly. *Mamm. Genome* **26**, 366–378 (2015).
 210. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
 211. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
 212. Ye, J. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, 134 (2012).
 213. Hansen, K. D., Irizarry, R. A. & Wu, Z. Removing technical variability in RNA-seq

- data using conditional quantile normalization. *Biostatistics* **13**, 204–216 (2012).
214. Wei, T. & Simko, V. corrplot: Visualization of a Correlation Matrix. (2016).
 215. Neuwirth, E. RColorBrewer: ColorBrewer Palettes. (2014).
 216. Schwalb, B., Tresch, A., Torkler, P., Duemcke, S. & Demel, C. LSD: Lots of Superior Depictions. (2015).
 217. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
 218. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
 219. Poulin, F. *et al.* In vivo characterization of a vertebrate ultraconserved enhancer. *Genomics* **85**, 774–781 (2005).
 220. Hughes, A. J., Daniel, S. E., Kilford, L. & Lees, A. J. Accuracy of clinical diagnosis of idiopathic Parkinson’s disease: a clinico-pathological study of 100 cases. *J. Neurol. Neurosurg. Psychiatry* **55**, 181–184 (1992).
 221. R Core Team. R: A language and environment for statistical computing. (2017).
 222. Sinnwell, J. P. & Schaid, D. J. haplo.stats: Statistical Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguous. (2016).
 223. Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. & Poland, G. A. Score Tests for Association between Traits and Haplotypes when Linkage Phase Is Ambiguous. *Am. J. Hum. Genet.* **70**, 425–434 (2002).
 224. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).
 225. Labun, K. *et al.* CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res.* **47**, W171–W174 (2019).
 226. Labun, K., Montague, T. G., Gagnon, J. A., Thyme, S. B. & Valen, E. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res.* **44**, W272–W276 (2016).
 227. Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M. & Valen, E. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Res.* **42**, W401–W407 (2014).
 228. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
 229. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
 230. Watkins-Chow, D. E. *et al.* Highly Efficient Cpf1-Mediated Gene Targeting in Mice Following High Concentration Pronuclear Injection. *G3 (Bethesda)*. **7**, 719–722 (2017).
 231. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
 232. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
 233. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248–249 (2010).
 234. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using

- RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
235. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
236. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
237. McClymont, S. A. *et al.* Parkinson-Associated SNCA Enhancer Variants Revealed by Open Chromatin in Mouse Dopamine Neurons. *Am. J. Hum. Genet.* **103**, 874–892 (2018).
238. Sugiaman-Trapman, D. *et al.* Characterization of the human RFX transcription factor family by regulatory and target gene analysis. *BMC Genomics* **19**, 181 (2018).
239. Gajiwala, K. S. *et al.* Structure of the winged-helix protein hRFX1 reveals a new mode of DNA binding. *Nature* **403**, 916–921 (2000).
240. Kittappa, R., Chang, W. W., Awatramani, R. B. & McKay, R. D. G. The *foxa2* gene controls the birth and spontaneous degeneration of dopamine neurons in old age. *PLoS Biol.* **5**, e325 (2007).
241. Caiazzo, M. *et al.* Direct generation of functional dopaminergic neurons from mouse and human fibroblasts. *Nature* **476**, 224–227 (2011).
242. Sung, M.-H., Guertin, M. J., Baek, S. & Hager, G. L. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol. Cell* **56**, 275–285 (2014).
243. Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer’s disease. *Nat. Genet.* **45**, 1452–1458 (2013).
244. Genetic determinants of common epilepsies: a meta-analysis of genome-wide association studies. *Lancet. Neurol.* **13**, 893–903 (2014).
245. Hoglinger, G. U. *et al.* Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. *Nat. Genet.* **43**, 699–705 (2011).
246. Consortium, S. W. G. of the P. G. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
247. Beer, M. A. Predicting enhancer activity and variant impact using gkm-SVM. *Hum. Mutat.* **38**, 1251–1258 (2017).
248. Gorkin, D. U. *et al.* Integration of ChIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes. *Genome Res.* **22**, 2290–2301 (2012).
249. Jeong, J. S. *et al.* Rapid identification of monospecific monoclonal antibodies using a human proteome microarray. *Mol. Cell. Proteomics* **11**, O111.016253 (2012).
250. Son, J. H. *et al.* Neuroprotection and neuronal differentiation studies using substantia nigra dopaminergic cells derived from transgenic mouse embryos. *J. Neurosci.* **19**, 10–20 (1999).
251. Zorita, E., Cusco, P. & Filion, G. J. Starcode: sequence clustering based on all-pairs search. *Bioinformatics* **31**, 1913–1919 (2015).
252. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
253. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
254. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying

- similarity between motifs. *Genome Biol.* **8**, R24 (2007).
255. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21**, 447–455 (2011).
 256. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
 257. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* **11**, R119 (2010).
 258. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
 259. Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847–3849 (2015).
 260. Shannon, P. & Richards, M. MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs. (2017).
 261. Hu, S. *et al.* DNA methylation presents distinct binding sites for human transcription factors. *Elife* **2**, e00726 (2013).
 262. Postuma, R. B. & Berg, D. Advances in markers of prodromal Parkinson disease. *Nat. Rev. Neurol.* **12**, 622–634 (2016).
 263. Barallobre, M. J. *et al.* DYRK1A promotes dopaminergic neuron survival in the developing brain and in a mouse model of Parkinson’s disease. *Cell Death Dis.* **5**, e1289 (2014).
 264. Hegarty, S. V, Sullivan, A. M. & O’Keeffe, G. W. Midbrain dopaminergic neurons: a review of the molecular circuitry that regulates their development. *Dev. Biol.* **379**, 123–138 (2013).
 265. Pacary, E., Azzarelli, R. & Guillemot, F. Rnd3 coordinates early steps of cortical neurogenesis through actin-dependent and -independent mechanisms. *Nat. Commun.* **4**, 1635 (2013).
 266. Peukert, D., Weber, S., Lumsden, A. & Scholpp, S. Lhx2 and Lhx9 determine neuronal differentiation and compartment in the caudal forebrain by regulating Wnt signaling. *PLoS Biol.* **9**, e1001218 (2011).
 267. Pacary, E. *et al.* Proneural transcription factors regulate different steps of cortical neuron migration through Rnd-mediated inhibition of RhoA signaling. *Neuron* **69**, 1069–1084 (2011).
 268. Yin, M. *et al.* Ventral mesencephalon-enriched genes that regulate the development of dopaminergic neurons in vivo. *J. Neurosci.* **29**, 5170–5182 (2009).
 269. Mei, L. & Xiong, W.-C. Neuregulin 1 in neural development, synaptic plasticity and schizophrenia. *Nat. Rev. Neurosci.* **9**, 437–452 (2008).
 270. Petryniak, M. A., Potter, G. B., Rowitch, D. H. & Rubenstein, J. L. R. Dlx1 and Dlx2 control neuronal versus oligodendroglial cell fate acquisition in the developing forebrain. *Neuron* **55**, 417–433 (2007).
 271. Retaux, S., Rogard, M., Bach, I., Failli, V. & Besson, M. J. Lhx9: a novel LIM-homeodomain gene expressed in the developing forebrain. *J. Neurosci.* **19**, 783–793 (1999).
 272. Kramer, R. *et al.* Neuregulins with an Ig-like domain are essential for mouse myocardial and neuronal development. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 4833–4838 (1996).

273. Arenas, E., Denham, M. & Villaescusa, J. C. How to make a midbrain dopaminergic neuron. *Development* **142**, 1918–1936 (2015).
274. Kelsom, C. & Lu, W. Development and specification of GABAergic cortical interneurons. *Cell Biosci.* **3**, 19 (2013).
275. Lavado, A., Lagutin, O. V & Oliver, G. Six3 inactivation causes progressive caudalization and aberrant patterning of the mammalian diencephalon. *Development* **135**, 441–450 (2008).
276. Geng, X., Lavado, A., Lagutin, O. V, Liu, W. & Oliver, G. Expression of Six3 Opposite Strand (Six3OS) during mouse embryonic development. *Gene Expr. Patterns* **7**, 252–257 (2007).
277. Gestri, G. *et al.* Six3 functions in anterior neural plate specification by promoting cell proliferation and inhibiting Bmp4 expression. *Development* **132**, 2401–2413 (2005).
278. Lagutin, O. V *et al.* Six3 repression of Wnt signaling in the anterior neuroectoderm is essential for vertebrate forebrain development. *Genes Dev.* **17**, 368–379 (2003).
279. Panman, L. *et al.* Sox6 and Otx2 control the specification of substantia nigra and ventral tegmental area dopamine neurons. *Cell Rep.* **8**, 1018–1025 (2014).
280. Viereckel, T. *et al.* Midbrain Gene Screening Identifies a New Mesoaccumbal Glutamatergic Pathway and a Marker for Dopamine Cells Neuroprotected in Parkinson's Disease. *Sci. Rep.* **6**, 35203 (2016).
281. Kozicz, T., Vigh, S. & Arimura, A. The source of origin of PACAP- and VIP-immunoreactive fibers in the laterodorsal division of the bed nucleus of the stria terminalis in the rat. *Brain Res.* **810**, 211–219 (1998).
282. Darland, T., Heinricher, M. M. & Grandy, D. K. Orphanin FQ/nociceptin: a role in pain and analgesia, but so much more. *Trends Neurosci.* **21**, 215–221 (1998).
283. Poulin, J.-F. *et al.* Defining midbrain dopaminergic neuron diversity by single-cell gene expression profiling. *Cell Rep.* **9**, 930–943 (2014).
284. Cai, H., Liu, G., Sun, L. & Ding, J. Aldehyde Dehydrogenase 1 making molecular inroads into the differential vulnerability of nigrostriatal dopaminergic neuron subtypes in Parkinson's disease. *Transl. Neurodegener.* **3**, 27 (2014).
285. Itoh, N. & Ohta, H. Roles of FGF20 in dopaminergic neurons and Parkinson's disease. *Front. Mol. Neurosci.* **6**, 15 (2013).
286. Ng, S.-Y., Bogu, G. K., Soh, B. S. & Stanton, L. W. The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis. *Mol. Cell* **51**, 349–359 (2013).
287. Ellis, B. C., Molloy, P. L. & Graham, L. D. CRNDE: A Long Non-Coding RNA Involved in Cancer, Neurobiology, and Development. *Front. Genet.* **3**, 270 (2012).
288. Guttman, M. *et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295–300 (2011).
289. Lin, M. *et al.* RNA-Seq of human neurons derived from iPS cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders. *PLoS One* **6**, e23356 (2011).
290. Uhde, C. W., Vives, J., Jaeger, I. & Li, M. Rmst is a novel marker for the mouse ventral mesencephalic floor plate and the anterior dorsal midline cells. *PLoS One* **5**, e8641 (2010).
291. Giasson, B. I. *et al.* Neuronal alpha-synucleinopathy with severe movement disorder in mice expressing A53T human alpha-synuclein. *Neuron* **34**, 521–533 (2002).

292. Zhou, Q., Wang, S. & Anderson, D. J. Identification of a novel family of oligodendrocyte lineage-specific basic helix-loop-helix transcription factors. *Neuron* **25**, 331–343 (2000).
293. Bennett, M. L. *et al.* New tools for studying microglia in the mouse and human CNS. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E1738-46 (2016).
294. Jung, S. *et al.* Analysis of fractalkine receptor CX(3)CR1 function by targeted deletion and green fluorescent protein reporter gene insertion. *Mol. Cell. Biol.* **20**, 4106–4114 (2000).
295. Cahoy, J. D. *et al.* A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J. Neurosci.* **28**, 264–278 (2008).
296. Yang, Y. *et al.* Molecular comparison of GLT1+ and ALDH1L1+ astrocytes in vivo in astroglial reporter mice. *Glia* **59**, 200–207 (2011).
297. Kageyama, R., Ohtsuka, T. & Kobayashi, T. Roles of Hes genes in neural development. *Dev. Growth Differ.* **50 Suppl 1**, S97-103 (2008).
298. Wang, W. *et al.* Nuclear factor I coordinates multiple phases of cerebellar granule cell development via regulation of cell adhesion molecules. *J. Neurosci.* **27**, 6115–6127 (2007).
299. Aruga, J. *et al.* Mouse *Zic1* is involved in cerebellar development. *J. Neurosci.* **18**, 284–293 (1998).
300. Blank, M. C. *et al.* Multiple developmental programs are altered by loss of *Zic1* and *Zic4* to cause Dandy-Walker malformation cerebellar pathogenesis. *Development* **138**, 1207–1216 (2011).
301. Gleichmann, M. *et al.* Identification of inhibitor-of-differentiation 2 (*Id2*) as a modulator of neuronal apoptosis. *J. Neurochem.* **80**, 755–762 (2002).
302. Jung, M. *et al.* Analysis of the expression pattern of the schizophrenia-risk and intellectual disability gene *TCF4* in the developing and adult brain suggests a role in development and plasticity of cortical and hippocampal neurons. *Mol. Autism* **9**, 20 (2018).
303. Sepp, M., Kannike, K., Eesmaa, A., Urb, M. & Timmusk, T. Functional diversity of human basic helix-loop-helix transcription factor *TCF4* isoforms generated by alternative 5' exon usage and splicing. *PLoS One* **6**, e22138 (2011).
304. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).
305. Chiurchiu, V., Maccarrone, M. & Orlicchio, A. The role of reticulons in neurodegenerative diseases. *Neuromolecular Med.* **16**, 3–15 (2014).
306. Lenka, A., Arumugham, S. S., Christopher, R. & Pal, P. K. Genetic substrates of psychosis in patients with Parkinson's disease: A critical review. *J. Neurol. Sci.* **364**, 33–41 (2016).
307. Wang, J., Si, Y.-M., Liu, Z.-L. & Yu, L. Cholecystokinin, cholecystokinin-A receptor and cholecystokinin-B receptor gene polymorphisms in Parkinson's disease. *Pharmacogenetics* **13**, 365–369 (2003).
308. Su, J. *et al.* *RESP18* deficiency has protective effects in dopaminergic neurons in an MPTP mouse model of Parkinson's disease. *Neurochem. Int.* **118**, 195–204 (2018).
309. Huang, Y. *et al.* *RESP18* is involved in the cytotoxicity of dopaminergic neurotoxins in MN9D cells. *Neurotox. Res.* **24**, 164–175 (2013).

310. Nucifora, F. C. J. *et al.* Ubiquitination via K27 and K29 chains signals aggregation and neuronal protection of LRRK2 by WSB1. *Nat. Commun.* **7**, 11792 (2016).
311. Salpietro, V. *et al.* AMPA receptor GluA2 subunit defects are a cause of neurodevelopmental disorders. *Nat. Commun.* **10**, 3094 (2019).
312. Briand, L. A., Deutschmann, A. U., Ellis, A. S. & Fosnocht, A. Q. Disrupting GluA2 phosphorylation potentiates reinstatement of cocaine seeking. *Neuropharmacology* **111**, 231–241 (2016).
313. Kanehisa, M. & Sato, Y. KEGG Mapper for inferring cellular functions from protein sequences. *Protein Sci.* (2019).
314. Mastroeni, D. *et al.* Nuclear but not mitochondrial-encoded oxidative phosphorylation genes are altered in aging, mild cognitive impairment, and Alzheimer's disease. *Alzheimers. Dement.* **13**, 510–519 (2017).
315. Schapira, A. H. V., Chaudhuri, K. R. & Jenner, P. Non-motor features of Parkinson disease. *Nat. Rev. Neurosci.* **18**, 435–450 (2017).
316. Al-Qassabi, A., Fereshtehnejad, S.-M. & Postuma, R. B. Sleep Disturbances in the Prodromal Stage of Parkinson Disease. *Curr. Treat. Options Neurol.* **19**, 22 (2017).
317. Weingarten, C. P., Sundman, M. H., Hickey, P. & Chen, N. Neuroimaging of Parkinson's disease: Expanding views. *Neurosci. Biobehav. Rev.* **59**, 16–52 (2015).
318. Politis, M. & Loane, C. Serotonergic dysfunction in Parkinson's disease and its relevance to disability. *ScientificWorldJournal.* **11**, 1726–1734 (2011).
319. Ho, D. H. *et al.* LRRK2 Kinase Activity Induces Mitochondrial Fission in Microglia via Drp1 and Modulates Neuroinflammation. *Exp. Neurobiol.* **27**, 171–180 (2018).
320. Ma, B. *et al.* LRRK2 modulates microglial activity through regulation of chemokine (C-X3-C) receptor 1-mediated signalling pathways. *Hum. Mol. Genet.* **25**, 3515–3523 (2016).
321. Moehle, M. S. *et al.* LRRK2 inhibition attenuates microglial inflammatory responses. *J. Neurosci.* **32**, 1602–1611 (2012).
322. Park, J.-S., Davis, R. L. & Sue, C. M. Mitochondrial Dysfunction in Parkinson's Disease: New Mechanistic Insights and Therapeutic Perspectives. *Curr. Neurol. Neurosci. Rep.* **18**, 21 (2018).
323. Pickrell, A. M. & Youle, R. J. The roles of PINK1, parkin, and mitochondrial fidelity in Parkinson's disease. *Neuron* **85**, 257–273 (2015).
324. Mullin, S. & Schapira, A. alpha-Synuclein and mitochondrial dysfunction in Parkinson's disease. *Mol. Neurobiol.* **47**, 587–597 (2013).
325. Hauser, D. N. & Hastings, T. G. Mitochondrial dysfunction and oxidative stress in Parkinson's disease and monogenic parkinsonism. *Neurobiol. Dis.* **51**, 35–42 (2013).
326. Krebiehl, G. *et al.* Reduced basal autophagy and impaired mitochondrial dynamics due to loss of Parkinson's disease-associated protein DJ-1. *PLoS One* **5**, e9367 (2010).
327. Fukui, H. & Moraes, C. T. The mitochondrial impairment, oxidative stress and neurodegeneration connection: reality or just an attractive hypothesis? *Trends Neurosci.* **31**, 251–256 (2008).
328. Richardson, J. R., Quan, Y., Sherer, T. B., Greenamyre, J. T. & Miller, G. W. Paraquat neurotoxicity is distinct from that of MPTP and rotenone. *Toxicol. Sci.* **88**, 193–201 (2005).

329. Nicklas, W. J., Vyas, I. & Heikkila, R. E. Inhibition of NADH-linked oxidation in brain mitochondria by 1-methyl-4-phenyl-pyridine, a metabolite of the neurotoxin, 1-methyl-4-phenyl-1,2,5,6-tetrahydropyridine. *Life Sci.* **36**, 2503–2508 (1985).
330. Marella, M., Seo, B. B., Yagi, T. & Matsuno-Yagi, A. Parkinson's disease and mitochondrial complex I: a perspective on the Ndi1 therapy. *J. Bioenerg. Biomembr.* **41**, 493–497 (2009).
331. Greenamyre, J. T., Sherer, T. B., Betarbet, R. & Panov, A. V. Complex I and Parkinson's disease. *IUBMB Life* **52**, 135–141 (2001).
332. Pozo Devoto, V. M. *et al.* alphaSynuclein control of mitochondrial homeostasis in human-derived neurons is disrupted by mutations associated with Parkinson's disease. *Sci. Rep.* **7**, 5042 (2017).
333. Nakamura, K. *et al.* Direct membrane association drives mitochondrial fission by the Parkinson disease-associated protein alpha-synuclein. *J. Biol. Chem.* **286**, 20710–20726 (2011).
334. Kamp, F. *et al.* Inhibition of mitochondrial fusion by alpha-synuclein is rescued by PINK1, Parkin and DJ-1. *EMBO J.* **29**, 3571–3589 (2010).
335. Pozo Devoto, V. M. & Falzone, T. L. Mitochondrial dynamics in Parkinson's disease: a role for alpha-synuclein? *Dis. Model. Mech.* **10**, 1075–1087 (2017).
336. Xie, W. & Chung, K. K. K. Alpha-synuclein impairs normal dynamics of mitochondria in cell and animal models of Parkinson's disease. *J. Neurochem.* **122**, 404–414 (2012).
337. Son, G. & Han, J. Roles of mitochondria in neuronal development. *BMB Rep.* **51**, 549–556 (2018).
338. Arrazola, M. S. *et al.* Mitochondria in Developmental and Adult Neurogenesis. *Neurotox. Res.* **36**, 257–267 (2019).
339. Khacho, M. *et al.* Mitochondrial Dynamics Impacts Stem Cell Identity and Fate Decisions by Regulating a Nuclear Transcriptional Program. *Cell Stem Cell* **19**, 232–247 (2016).
340. Saxena, A. *et al.* Trehalose-enhanced isolation of neuronal sub-types from adult mouse brain. *Biotechniques* **52**, 381–385 (2012).
341. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
342. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
343. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **14**, 309–315 (2017).
344. Wang, X.-F. & Xu, Y. Fast clustering using adaptive density peak detection. *Stat. Methods Med. Res.* **26**, 2800–2811 (2017).
345. Prakash, N. & Wurst, W. Development of dopaminergic neurons in the mammalian brain. *Cell. Mol. Life Sci.* **63**, 187–206 (2006).
346. Nakatani, T., Kumai, M., Mizuhara, E., Minaki, Y. & Ono, Y. Lmx1a and Lmx1b cooperate with Foxa2 to coordinate the specification of dopaminergic neurons and control of floor plate cell differentiation in the developing mesencephalon. *Dev. Biol.* **339**, 101–113 (2010).
347. Jankovic, J., Chen, S. & Le, W. D. The role of Nurr1 in the development of dopaminergic neurons and Parkinson's disease. *Prog. Neurobiol.* **77**, 128–138 (2005).

348. Chen, N. & Reith, M. E. Structure and function of the dopamine transporter. *Eur. J. Pharmacol.* **405**, 329–339 (2000).
349. Wallen-Mackenzie, A., Wootz, H. & Englund, H. Genetic inactivation of the vesicular glutamate transporter 2 (VGLUT2) in the mouse: what have we learnt about functional glutamatergic neurotransmission? *Ups. J. Med. Sci.* **115**, 11–20 (2010).
350. Rumping, L. *et al.* GLS hyperactivity causes glutamate excess, infantile cataract and profound developmental delay. *Hum. Mol. Genet.* **28**, 96–104 (2019).
351. Zhou, Y. *et al.* Selective deletion of glutamine synthetase in the mouse cerebral cortex induces glial dysfunction and vascular impairment that precede epilepsy and neurodegeneration. *Neurochem. Int.* **123**, 22–33 (2019).
352. Grone, B. P. & Maruska, K. P. Three Distinct Glutamate Decarboxylase Genes in Vertebrates. *Sci. Rep.* **6**, 30507 (2016).
353. Kaupmann, K. *et al.* GABA(B)-receptor subtypes assemble into functional heteromeric complexes. *Nature* **396**, 683–687 (1998).
354. Zhou, Y. & Danbolt, N. C. GABA and Glutamate Transporters in Brain. *Front. Endocrinol. (Lausanne)*. **4**, 165 (2013).
355. Zill, P. *et al.* Analysis of tryptophan hydroxylase I and II mRNA expression in the human brain: a post-mortem study. *J. Psychiatr. Res.* **41**, 168–173 (2007).
356. Krueger, K. C. & Deneris, E. S. Serotonergic transcription of human FEV reveals direct GATA factor interactions and fate of Pet-1-deficient serotonin neuron precursors. *J. Neurosci.* **28**, 12748–12758 (2008).
357. Murphy, D. L. & Lesch, K.-P. Targeting the murine serotonin transporter: insights into human neurobiology. *Nat. Rev. Neurosci.* **9**, 85–96 (2008).
358. Oda, Y. Choline acetyltransferase: the structure, distribution and pathologic changes in the central nervous system. *Pathol. Int.* **49**, 921–937 (1999).
359. Nachmansohn, D. & Machado, A. L. THE FORMATION OF ACETYLCHOLINE. A NEW ENZYME: ‘CHOLINE ACETYLASE’. *J. Neurophysiol.* **6**, 397–403 (1943).
360. Erickson, J. D. *et al.* Functional identification of a vesicular acetylcholine transporter and its expression from a ‘cholinergic’ gene locus. *J. Biol. Chem.* **269**, 21929–21932 (1994).
361. Misgeld, T., Kerschensteiner, M., Bareyre, F. M., Burgess, R. W. & Lichtman, J. W. Imaging axonal transport of mitochondria in vivo. *Nat. Methods* **4**, 559–561 (2007).
362. Kopp, J. L. *et al.* Sox9+ ductal cells are multipotent progenitors throughout development but do not produce new endocrine cells in the normal or injured adult pancreas. *Development* **138**, 653–665 (2011).
363. Xu, X. *et al.* β Cells Can Be Generated from Endogenous Progenitors in Injured Adult Mouse Pancreas. *Cell* **132**, 197–207 (2008).
364. Al-Hasani, K. *et al.* Adult Duct-Lining Cells Can Reprogram into β -like Cells Able to Counter Repeated Cycles of Toxin-Induced Diabetes. *Dev. Cell* **26**, 86–100 (2013).
365. Van de Casteele, M. *et al.* Partial Duct Ligation: β -Cell Proliferation and Beyond. *Diabetes* **63**, 2567–2577 (2014).
366. Xiao, X. *et al.* No evidence for β cell neogenesis in murine adult pancreas. *J. Clin. Invest.* **123**, 2207–2217 (2013).
367. Xiao, X. *et al.* Neurogenin3 Activation Is Not Sufficient to Direct Duct-to-Beta Cell

- Transdifferentiation in the Adult Pancreas. *J. Biol. Chem.* **288**, 25297–25308 (2013).
368. Arnes, L., Hill, J. T., Gross, S., Magnuson, M. A. & Sussel, L. Ghrelin Expression in the Mouse Pancreas Defines a Unique Multipotent Progenitor Population. *PLoS One* **7**, e52026 (2012).
369. Chera, S. *et al.* Diabetes recovery by age-dependent conversion of pancreatic δ -cells into insulin producers. *Nature* **514**, 503–507 (2014).
370. Courtney, M. *et al.* The Inactivation of Arx in Pancreatic α -Cells Triggers Their Neogenesis and Conversion into Functional β -Like Cells. *PLoS Genet.* **9**, e1003934 (2013).
371. Nir, T., Melton, D. A. & Dor, Y. Recovery from diabetes in mice by β cell regeneration. *J. Clin. Invest.* **117**, 2553–2561 (2007).
372. Thorel, F. *et al.* Conversion of adult pancreatic α -cells to β -cells after extreme β -cell loss. *Nature* **464**, 1149–1154 (2010).
373. Meier, J. J., Bhushan, A., Butler, A. E., Rizza, R. A. & Butler, P. C. Sustained beta cell apoptosis in patients with long-standing type 1 diabetes: indirect evidence for islet regeneration? *Diabetologia* **48**, 2221–2228 (2005).
374. Butler, A. E. *et al.* Modestly increased beta cell apoptosis but no increased beta cell replication in recent-onset type 1 diabetic patients who died of diabetic ketoacidosis. *Diabetologia* **50**, 2323–2331 (2007).
375. Moss, J. B. *et al.* Regeneration of the pancreas in adult zebrafish. *Diabetes* **58**, 1844–1851 (2009).
376. Pisharath, H., Rhee, J. M., Swanson, M. A., Leach, S. D. & Parsons, M. J. Targeted ablation of beta cells in the embryonic zebrafish pancreas using *E. coli* nitroreductase. *Mech. Dev.* **124**, 218–229 (2007).
377. Delaspre, F. *et al.* Centroacinar Cells Are Progenitors That Contribute to Endocrine Pancreas Regeneration. *Diabetes* **64**, 3499–3509 (2015).
378. Beer, R. L., Parsons, M. J. & Rovira, M. Centroacinar cells: At the center of pancreas regeneration. *Dev. Biol.* **413**, 8–15 (2016).
379. Huang, W. *et al.* Sox9b is a mediator of retinoic acid signaling restricting endocrine progenitor differentiation. *Dev. Biol.* **418**, 28–39 (2016).
380. Seymour, P. A. *et al.* SOX9 is required for maintenance of the pancreatic progenitor cell pool. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1865–1870 (2007).
381. Miyamoto, Y. *et al.* Notch mediates TGF α -induced changes in epithelial differentiation during pancreatic tumorigenesis. *Cancer Cell* **3**, 565–576 (2003).
382. Seymour, P. A. Sox9: a master regulator of the pancreatic program. *Rev. Diabet. Stud.* **11**, 51–83 (2014).
383. Kopp, J. L. *et al.* Identification of Sox9-dependent acinar-to-ductal reprogramming as the principal mechanism for initiation of pancreatic ductal adenocarcinoma. *Cancer Cell* **22**, 737–750 (2012).
384. Wu, Y. *et al.* c-Kit and stem cell factor regulate PANC-1 cell differentiation into insulin- and glucagon-producing cells. *Lab. Investig.* **90**, 1373–1384 (2010).
385. Cirulli, V. *et al.* KSA antigen Ep-CAM mediates cell-cell adhesion of pancreatic epithelial cells: morphoregulatory roles in pancreatic islet development. *J Cell Biol* **140**, 1519–1534 (1998).
386. Maimets, M. *et al.* Long-Term In Vitro Expansion of Salivary Gland Stem Cells

- Driven by Wnt Signals. *Stem cell reports* **6**, 150–162 (2016).
387. Tanimizu, N., Kobayashi, S., Ichinohe, N. & Mitaka, T. Downregulation of miR122 by grainyhead-like 2 restricts the hepatocytic differentiation potential of adult liver progenitor cells. *Development* **141**, 4448–4456 (2014).
 388. Balzar, M. *et al.* The structural analysis of adhesions mediated by Ep-CAM. *Exp. Cell Res.* **246**, 108–121 (1999).
 389. Ishii, H. *et al.* Epithelial splicing regulatory proteins 1 (ESRP1) and 2 (ESRP2) suppress cancer cell motility via different mechanisms. *J Biol Chem* **289**, 27386–27399 (2014).
 390. Li, D. *et al.* Adrenocortical zonation factor 1 is a novel matricellular protein promoting integrin-mediated adhesion of adrenocortical and vascular smooth muscle cells. *FEBS J.* **274**, 2506–2522 (2007).
 391. Tajiri, Y. *et al.* Tubulointerstitial nephritis antigen-like 1 is expressed in the uterus and binds with integrins in decidualized endometrium during postimplantation in mice. *Biol. Reprod.* **82**, 263–270 (2010).
 392. Cui, X. B., Luan, J. N., Ye, J. & Chen, S. Y. RGC32 deficiency protects against high-fat diet-induced obesity and insulin resistance in mice. *J Endocrinol* **224**, 127–137 (2015).
 393. Schnell, U., Cirulli, V. & Giepmans, B. N. EpCAM: structure and function in health and disease. *Biochim Biophys Acta* **1828**, 1989–2001 (2013).
 394. Chan, S. W., Fowler, K. J., Choo, K. H. & Kalitsis, P. Spef1, a conserved novel testis protein found in mouse sperm flagella. *Gene* **353**, 189–199 (2005).
 395. Xue, J. C. & Goldberg, E. Identification of a novel testis-specific leucine-rich protein in humans and mice. *Biol Reprod* **62**, 1278–1284 (2000).
 396. Zeng, L. *et al.* Identification of a novel human doublecortin-domain-containing gene (DCDC1) expressed mainly in testis. *J Hum Genet* **48**, 393–396 (2003).
 397. Bernard, P. *et al.* Dimerization of SOX9 is required for chondrogenesis, but not for sex determination. *Hum Mol Genet* **12**, 1755–1765 (2003).
 398. He, X., Ohba, S., Hojo, H. & McMahon, A. P. AP-1 family members act with Sox9 to promote chondrocyte hypertrophy. *Development* **143**, 3012–3023 (2016).
 399. Shih, H. P. *et al.* A Gene Regulatory Network Cooperatively Controlled by Pdx1 and Sox9 Governs Lineage Allocation of Foregut Progenitor Cells. *Cell Rep* **13**, 326–336 (2015).
 400. Kadaja, M. *et al.* SOX9: a stem cell transcriptional regulator of secreted niche signaling factors. *Genes Dev* **28**, 328–341 (2014).
 401. Zeng, L., Kempf, H., Murtaugh, L. C., Sato, M. E. & Lassar, A. B. Shh establishes an Nkx3.2/Sox9 autoregulatory loop that is maintained by BMP signals to induce somitic chondrogenesis. *Genes Dev.* **16**, 1990–2005 (2002).
 402. Lefebvre, V. & Dvir-Ginzberg, M. SOX9 and the many facets of its regulation in the chondrocyte lineage. *Connect. Tissue Res.* **58**, 2–14 (2017).
 403. Tarifeño-Saldivia, E. *et al.* Transcriptome analysis of pancreatic cells across distant species highlights novel important regulator genes. *BMC Bol.* **15**, 21 (2017).
 404. Muto, A., Iida, A., Satoh, S. & Watanabe, S. The group E Sox genes Sox8 and Sox9 are regulated by Notch signaling and are required for Müller glial cell development in mouse retina. *Exp. Eye Res.* **89**, 549–558 (2009).

405. Briot, A. *et al.* Repression of Sox9 by Jag1 is continuously required to suppress the default chondrogenic fate of vascular smooth muscle cells. *Dev. Cell* **31**, 707–721 (2014).
406. Capaccione, K. M. *et al.* Sox9 mediates Notch1-induced mesenchymal features in lung adenocarcinoma. *Oncotarget* **5**, 3636–3650 (2014).
407. Haller, R. *et al.* Notch1 signaling regulates chondrogenic lineage determination through Sox9 activation. *Cell Death Differ.* **19**, 461–469 (2012).
408. Huang, W. *et al.* Retinoic acid plays an evolutionarily conserved and biphasic role in pancreas development. *Dev. Biol.* **394**, 83–93 (2014).
409. Ohba, S., He, X., Hojo, H. & McMahon, A. P. Distinct Transcriptional Programs Underlie Sox9 Regulation of the Mammalian Chondrocyte. *Cell Rep.* **12**, 229–243 (2015).
410. Igarashi, T. *et al.* Tubulointerstitial nephritis antigen-like 1 is expressed in extraembryonic tissues and interacts with laminin 1 in the Reichert membrane at postimplantation in the mouse. *Biol. Reprod.* **81**, 948–955 (2009).
411. Wex, T. *et al.* TIN-ag-RP, a novel catalytically inactive cathepsin B-related protein with EGF domains, is predominantly expressed in vascular smooth muscle cells. *Biochemistry* **40**, 1350–1357 (2001).
412. Favre, C. J. *et al.* Expression of genes involved in vascular development and angiogenesis in endothelial cells of adult lung. *Am. J. Physiol. Heart Circ. Physiol.* **285**, H1917-38 (2003).
413. Kalfa, T. A., Thull, J. D., Butkowski, R. J. & Charonis, A. S. Tubulointerstitial nephritis antigen interacts with laminin and type IV collagen and promotes cell adhesion. *J. Biol. Chem.* **269**, 1654–1659 (1994).
414. Ueda, J. *et al.* Epithelial splicing regulatory protein 1 is a favorable prognostic factor in pancreatic cancer that attenuates pancreatic metastases. *Oncogene* **33**, 4485–4495 (2013).
415. Deloria, A. J. *et al.* Epithelial splicing regulatory protein 1 and 2 paralogues correlate with splice signatures and favorable outcome in human colorectal cancer. *Oncotarget* **7**, 73800–73816 (2016).
416. Warzecha, C. C., Sato, T. K., Nabet, B., Hogenesch, J. B. & Carstens, R. P. ESRP1 and ESRP2 Are Epithelial Cell-Type-Specific Regulators of FGFR2 Splicing. *Mol. Cell* **33**, 591–601 (2009).
417. González, B., Denzel, S., Mack, B., Conrad, M. & Gires, O. EpCAM Is Involved in Maintenance of the Murine Embryonic Stem Cell Phenotype. *Stem Cells* **27**, 1782–1791 (2009).
418. Lu, T.-Y. *et al.* Epithelial Cell Adhesion Molecule Regulation Is Associated with the Maintenance of the Undifferentiated Phenotype of Human Embryonic Stem Cells. *J. Biol. Chem.* **285**, 8719–8732 (2010).
419. Vercollone, J. R., Balzar, M., Litvinov, S. V, Yang, W. & Cirulli, V. MMTV/LTR Promoter-Driven Transgenic Expression of EpCAM Leads to the Development of Large Pancreatic Islets. *J. Histochem. Cytochem.* **63**, 613–625 (2015).
420. Rodrigo-Torres, D. *et al.* The biliary epithelium gives rise to liver progenitor cells. *Hepatology* **60**, 1367–1377 (2014).
421. Lorent, K., Moore, J. C., Siekmann, A. F., Lawson, N. & Pack, M. Reiterative use of the notch signal during zebrafish intrahepatic biliary development. *Dev. Dyn.* **239**, 855–864 (2010).

422. 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).
423. Kuleshov, M. V *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90-7 (2016).
424. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
425. Lee, T. I., Johnstone, S. E. & Young, R. A. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nat. Protoc.* **1**, 729–748 (2006).
426. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D260–D266 (2018).
427. Ranieri, D. *et al.* Expression of the FGFR2 mesenchymal splicing variant in epithelial cells drives epithelial-mesenchymal transition. *Oncotarget* **7**, 5440–5460 (2016).
428. Tomlinson, D. C., L'Hôte, C. G., Kennedy, W., Pitt, E. & Knowles, M. A. Alternative splicing of fibroblast growth factor receptor 3 produces a secreted isoform that inhibits fibroblast growth factor-induced proliferation and is repressed in urothelial carcinoma cell lines. *Cancer Res.* **65**, 10441–10449 (2005).
429. Kolde, R. pheatmap: Pretty Heatmaps. (2019). Available at: <https://cran.r-project.org/package=pheatmap>.

6.3 Curriculum Vitae

Sarah McClymont

733 N Broadway
Baltimore, MD 21205
443-301-2909
sarahmcclymont@gmail.com

EDUCATION

- Johns Hopkins University School of Medicine, Baltimore, MD** 2013-2019
PhD candidate in the Predoctoral Training Program in Human Genetics
Thesis: Transcriptional regulation and disruption in Parkinson disease
Advisor: Dr. Andrew McCallion
- University of Guelph, Guelph, ON, Canada** 2008-2012
BSc (Hons) Biological Sciences, Minor in Educational Psychology

RESEARCH EXPERIENCE

- Undergraduate research assistant** – University of Guelph, Plant Agriculture 2011-2012
Identifying a disease resistance QTL for common bacterial blight in the common bean for the University of Guelph Field Bean Breeding Program
Advisors: Dr. Alireza Navabi and Dr. Weilong Xie
- Undergraduate research assistant** – University of Guelph, Plant Agriculture 2008-2011
Designing a novel aeroponics system and characterizing maize root physiology to examine the effects of domestication on fine root architecture and nitrogen use efficiency
Advisors: Dr. Manish Raizada and Dr. Amelie Gaudin

TEACHING EXPERIENCE

- Curriculum and content development** – Getting and Cleaning Data, Coursera 2019
Curriculum and content development – The Data Scientist's Toolbox, Coursera 2018
Teaching assistant – Human and Mammalian Genetics, Jackson Labs 2018
Teaching assistant – Undergraduate genetics, Stevenson University 2017-2018
Content development – Chromebook Data Science Program, Johns Hopkins 2017-2018
Lesson instructor – Science outside the Lines, Art with a Heart 2017
Teaching assistant – Basic Mechanisms of Disease, Johns Hopkins 2015

LEADERSHIP AND SERVICE

- Lead mentor** – Human genetics mentorship program, Johns Hopkins University 2016-present
Poster judge – Fourth-year undergraduate genetics, Stevenson University 2016, 2018
Graduate student committees for the Institute of Genetic Medicine:
Human genetics mentorship program 2016-present
Human genetics seminar speaker selection 2015-present
Barton Childs' seminar speaker selection 2014-present
Recruitment 2014-2015

HONOURS AND AWARDS

- C.W. Cotterman Award from the American Society of Human Genetics 2019
The McKusick Short Course on Human and Mammalian Genetics and Genomics Teaching Scholarship 2018
Natural Sciences and Engineering Research Council of Canada (NSERC) Postgraduate Scholarship 2013-2014
Undergraduate Student Research Award/NSERC Scholarship 2009

PUBLICATIONS

1. **McClymont, S.A.**, Hook, P.W., Soto, A.I., Reed, X., Law, W.D., Kerans, S.J., Waite, E.L., Briceno, N.J., Thole, J.F., Heckman, M.G., Diehl, N.N., Wszolek, Z.K., Moore, C.D., Zhu, H., Akiyama, J.A., Dickel, D.E., Visel, A., Pennacchio, L.A., Ross, O.A., Beer, M.A., McCallion, A.S. (2018). Parkinson-associated *SNCA* enhancer variants revealed by open chromatin in mouse dopamine neurons. *The American Journal of Human Genetics*, 103:874-892.
2. Gould, R.A., Aziz, H., Woods, C.E., ..., **McClymont, S.A.**, ..., *et al.* (2018). *ROBO4* Mutations Predispose Individuals to Bicuspid Aortic Valve and Thoracic Aortic Aneurysm. *Nature Genetics*, 51:42-50.
3. Hook, P.W., **McClymont, S.A.**, Cannon, G.H., Law, W.D., Morton, A.J., Goff, L.A., and McCallion, A.S. (2018). Single-Cell RNA-Seq of Mouse Dopaminergic Neurons Informs Candidate Gene Selection for Sporadic Parkinson Disease. *The American Journal of Human Genetics*, 102: 427–446.
4. Xie, W., Raja, K., **McClymont, S.**, Stonehouse, R., Bett, K., Yu, K., Pauls, K.P., Navabi, A. (2017). Interaction of quantitative trait loci for resistance to common bacterial blight and pathogen isolates in *Phaseolus vulgaris* L. *Molecular Breeding*, 37:55.
5. Turner, T.N., ..., **McClymont, S.A.**, ..., *et al.* (2016). Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. *The American Journal of Human Genetics*, 98: 58-74.
6. Gaudin, A.C.M., **McClymont, S.A.**, Soliman, S.S.M., and Raizada, M.N. (2014). The effect of altered dosage of a mutant allele of *Teosinte branched 1 (tb1-ref)* on the root system of modern maize. *BMC Genetics*, 15:23.
7. Gaudin, A.C.M., **McClymont, S.A.**, Holmes, B.M., Lyons, E. and Raizada, M.N. (2011). Novel temporal, fine-scale and growth variation phenotypes in roots of adult-stage maize (*Zea mays* L.) in response to low nitrogen stress. *Plant, Cell & Environment*, 34: 2122–2137.
8. Gaudin, A.C.M., **McClymont, S.A.** and Raizada, M.N. (2011). The nitrogen adaptation strategy of the wild teosinte ancestor of modern maize, *Zea mays* subsp. *parviglumis*. *Crop Science*, 51: 2780-2795.

ORAL PRESENTATIONS

1. **McClymont, S.A.** Effective undergraduate education for the successful preparation of graduate students and genetics professionals. Invited speaker at the American Society of Human Genetics Meeting; October 2019; Houston, TX.
2. **McClymont, S.A.** Enhancer variants at *SNCA* confer Parkinson disease risk. Invited speaker at Grand Challenges in Parkinson's Disease; August 2019; Grand Rapids, MI.
3. **McClymont, S.A.**, Hook, P.W., Soto, A.I., Briceno, N.J., Thole, J.F., Law, W.D., Kerans, S.J., Heckman, M.G., Diehl, N.N., Waite, E.L., Reed, X., Dickel, D.E., Akiyama, J.A., Visel, A., Pennacchio, L.A., Beer, M.A., Ross, O.A., McCallion, A.S. Exploiting dynamic open chromatin in mouse dopamine neurons reveals Parkinson-associated variation in an *SNCA* enhancer: a paradigm for illuminating functional noncoding variation. Presented at the 32nd International Mammalian Genome Conference; November 2018; San Juan, Puerto Rico.

POSTER PRESENTATIONS

1. **McClymont, S.A.**, McCallion, A.S. Single-cell RNA-seq in a mouse model of Parkinson disease reveals potential disease mechanisms affecting subpopulations of dopaminergic neurons. Presented at the American Society of Human Genetics Meeting; October 2019; Houston, TX.
2. **McClymont, S.A.**, Hook, P.W., Law, W.D., Beer, M.A., Ross, O.A., McCallion, A.S. Dopaminergic neuronal chromatin signatures reveal Parkinson Disease associated variation in a novel aminergic intronic enhancer at *SNCA*. Presented at the 6th Annual Genetics Research Day; April 2019; Baltimore, MD. Third Place.
3. **McClymont, S.A.**, Hook, P.W., Law, W.D., Beer, M.A., Ross, O.A., McCallion, A.S. Dopaminergic neuronal chromatin signatures reveal Parkinson Disease associated variation in a novel aminergic intronic enhancer at *SNCA*. Presented at The Keystone Symposia Conference on Chromatin Architecture and Chromosome Organization; March 2018; Whistler, B.C., Canada.

4. Edelman H.E.* , **McClymont, S.A.***, McCallion, A.S., Parsons, M.J. SOX9 ChIP-seq and RNA-seq in PANC-1 cells reveals interesting target genes for pancreatic progenitor biology. Presented at the 5th Annual Genetics Research Day; February 2018; Baltimore, MD.
5. **McClymont, S.A.**, Hook, P.W., Beer, M.A., Ross, O.A., McCallion, A.S. Chromatin Profiles of Dopaminergic Neurons Refine a Parkinsonian-associated Interval and Prioritize Neurological GWAS-implicated Variants for Functional Validation. Presented at the Lasker to Lasker: Bayview Research Symposium; December 2016; Baltimore, MD.
6. **McClymont, S.A.**, Hook, P.W., Beer, M.A., Ross, O.A., McCallion, A.S. Dopaminergic neuron chromatin signatures refine a novel Parkinsonian-associated interval and establish a pipeline for informing future genetic studies of neurological disease. Presented at the American Society for Human Genetics Meeting; October 2016; Vancouver, B.C., Canada.
7. **McClymont, S.A.**, Hook, P.W., Goff, L.A., McCallion, A.S. Chromatin profiles from *ex vivo* purified dopaminergic neurons establish a promising model to support studies of neurological function and dysfunction. Presented at the Human Genome Meeting; March 2016; Houston, TX.

* co-first authors