PhD-FSTM-2010-19
The Faculty of Sciences, Technology and Medicine

DISSERTATION

Defence held on 16/06/2020 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN *BIOLOGIE*

by

Yujuan GUI
Born on 14 January 1990 in city of Guangzhou, China

Genetic regulators of ventral midbrain gene expression and nigrostriatal circuit integrity

Dissertation defence committee

Dr Lasse Sinkkonen, dissertation supervisor
Principal Investigator, University of Luxembourg

Dr Thomas Sauter, Chairman
Professor, Université du Luxembourg

Dr Evan G. Williams
Assistant professor, Université du Luxembourg

Dr Minna Kaikkonen-Määttä
Associate Professor, University of Eastern Finland

Dr Florian Giesert
Team leader, Helmholtz Zentrum München

Affidavit

I hereby confirm that the PhD thesis “Genetic Regulators of ventral midbrain gene expression and nigrostriatal circuit integrity” has been written independently and without any other sources than cited.

Luxembourg, 24/05/2020 Yujuan Gui

Acknowledgements

I had always been thinking about doing a PhD in biology to explore topics that are essential for shaping organisms as who they are. I experienced a great load of joy when I received the congratulation email for acceptance, and this joy has been with me throughout the entire journey. It is a great adventure for me to come this far, and I would never be able to do it myself without the help from so many generous and supportive people I have been meeting along the way.

My big thanks go to my supervisors Dr. Lasse Sinkkonen and Dr. Manuel Buttini. They two gave me tremendous amount of invaluable intellectual input.

Lasse taught me lots of techniques and helped me sort out ideas on how to push the project forward. He also gave me independence to explore around building up my skill set, ranging from molecular biology techniques to bioinformatics. I cannot imagine myself coding thousands of lines and analyzing gigabytes of data four years ago. But with his full support, I am confident that any analysis is doable.

My spark with neuroscience is lighted by discussions with Manuel. He is always very enthusiastic about his work and giving me many new ideas from a different angle. Discussion with him is always fruitful. Manuel also established collaborations for our projects around the world, with the help from whom we achieved a lot.

Many thanks also go to Dr. Melanie Thomas and Dr. Mona Karout for their effort on dissecting mouse brain tissues and neuropathology. The project could not even begin without their work.

I would like to thank Dr. Aurélien Ginolhac and Dr. Anthoula Gaigneaux for their help on statistics and programming. Statistics and bugs are always evil, but these two angels bring lights to chase the dark out.

I would also like to thank Prof. Thomas Sauter and all my dear colleagues in the Systems Biology group. We have uncountable fun times and I cannot be more grateful to work with you.

My family is constantly giving me support and very patient to cope with my plan for future career. I would not even have the chance to go abroad without them having my back.

Special thanks go to Alvaro and his family. They are always enthusiastic to know what I have been doing and have been supportive all the time.

Last but not least, I am thankful to: the Fond National de la Recherche Luxembourg for providing the financial support for my PhD, my collaborators for providing input to help building up our projects, and my friends for supporting me whenever I need.

人生很短，经不起来回的犹豫。

Life is too short to keep hesitating.

- **何炅** *Jiong He*

Table of Contents

List of Figures

Figure 7: Representation of cis and trans effect of regulatory variants (Deplancke et al., 2016). A: In motif dependent manner, genetic variant is directly located in the motif sequence and disrupt TF binding. In motif independent scenario, (B-D) local, proximal, and distal variants can all disrupt individual TF and TF-TF complex binding.

Figure 8: Schematic representation of CC mice breeding with one representative chromosome (Abu Toamih Atamni and Iraqi, 2018). The breeding starts from eight parental strains. Intercross is carried out with independent breeding funnel to maximize recombination events and break linkage disequilibrium blocks.

Figure 9: Schematic representation of sagittal view of rodent and human brains (Cova and Armentero, 2011). The nigrostriatal pathway is shown on both, with black dots connecting SN to Str via mfb. Str: striatum; SVZ: subventricular zone; mfb: median forbrain bundle; SNc: substantia nigra pars compacta; VTA: ventral tegmental area.

Abbreviations

Apoa2	Apolipoprotein A-II
ASCL1	Achaete-Scute family BHLH transcription factor 1
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
Atp1b2	ATPase Na+/K+ Transporting Subunit Beta 2
BMDM	Bone-marrow-derived macrophages
bp	Base pairs
BTB domain	BR-C, ttk and bab
c-Myc	Master regulator of cell cycle entry and proliferative metabolism
C/EBPs	CCAAT-enhancer-binding proteins
C1qa	Complement C1q A Chain
CBP	Cyclic AMP-responsive element-binding protein
CBP	Cyclic AMP-responsive element-binding protein
CC	Collaborative Cross
Ccnd3	G1/S-specific cyclin D3
CDC7	Cell division cycle 7
ChIP-seq	Chromatin immunoprecipitation with massively parallel DNA sequencing
CNV	Copy number variation
Col4a2	Collagen alpha-2 (IV) chain protein
Col4a6	Collagen alpha-6 (IV) chain
CTCF	CCCTC-binding factor
DAT	Dopamine transporter
DBD	DNA binding domain
DHS	DNase I hypersensitive site
DNA	Deoxyribonucleic acid
DNase I	Deoxyribonuclease I
DNase-seq	DNaseI hypersensitivity sites sequencing
ECM	Extracellular matrix
ECM	Extracellular matrix
Eno1b	Enolase 1b
ESC	Embryonic stem cell
ETS	Erythroblast transformation specific
eQTL	Expression QTL
FDR	False discovery rate
Fgf2	Fibroblast growth factor 2
Foxa2	Forkhead box protein A2
Gabra2	Gamma-aminobutyric acid receptor subunit alpha-2
GDNF	Glial cells-derived neutrophic factor
GFP	Green florescent protein
GO	Gene ontology

GWAS	Genome-wide association study
H3.3/H2A.Z	Histone variants H3.3 and H2A.Z
H3K27ac	Histone H3 acetylation at lysine 27
H3K27me3	Histone H3 trimethylation at lysine 27
H3K36me3	Histone H3 trimethylation at lysine 36
H3K4me1/2	Histone H3 mono- or demethylation at lysine 4
H3K4me3	Histone H3 trimethylation at lysine 4
H3K79me3	Histone H3 trimethylation at lysine 79
H3K9me3	Histone H3 trimethylation at lysine 9
hQTL	Histone QTL
HSPC	Hematopoietic steam and progenitor cells
HU	Hydroxyurea
Igh	Immunoglobulin heavy-chain locus
IHEC	International Human Epigenome Consortium
INR	Transcriptional initiator
IsO	Isthmic organizer
Isoc2b	Isochorismatase domain-containing protein 2B
kb	Kilobase
LCR	Locus control region
LD	Linkage disequilibrium
LDTF	Lineage-determining transcription factor
Lef1	Lymphoid enhancer-binding factor 1
Lmx1a/b	Lim-homeobox gene
MAO	Monoamine oxidase
Mcee	Methylmalonyl-CoA Epimerase
mDAn	Midbrain dopaminergic neurons
MED1	Mediator Complex Subunit 1
mQTL	Methylation QTL
MHB	Midbrain-hindbrain boundary
mRNA	Messenger RNA
NPC	Neuronal progenitor cell
NPC	Neuronal precursor cells
NURF	Nucleosome remodeling factor
Nurr1	Nuclear Receptor Related-1 Protein
Otx2	Orthodenticle homolog 2
PD	Parkinson's disease
PHD	Plant homeodomain
PIC	Pre-initiation complex
Pitx3	Pituitary homeobox 3
PMAT	Plasma membrane monoamine transporter
polI	RNA polymerase I

polII	RNA polymerase II
polIII	RNA polymerase III
PRC2	Polycomb repressor complex 2
Pttg1	Pituitary tumor transforming gene 1
PWM	Positioning weight matrix
QTL	Quantitative trait locus
RGC	Radial glial cells
RI	Recombinant inbred
RNA	Ribonucleic acid
RPKM	Reads per kilobase of transcript, per million mapped reads
SAGA	Spt-Ada-Gcn5 acetyltransferase
scATAC-seq	Single cell ATAC-seq
scRNA-seq	Single cell RNA-seq
SET domain	Su(var)3-9, Enhancer-of-zeste and Trithorax
SETDB1	SET domain bifurcated histone lysine methyltransferase 1
SGF29	SAGA Complex Associated Factor 29
Slc1a2	Solute Carrier Family 1 Member 2
Slc6a3	Solute Carrier Family 6 Member 3
SMC1	Structural maintenance of chromosomes protein 1
SN	Substantia nigra
SNP	Single nucleotide polymorphism
Sp1	Specificity protein 1
SSR	Simple sequence repeat
SWI/SNF	SWItch/Sucrose Non-Fermentable
TADs	Topologically associating domains
TATA	5'-TATAAA-3' core DNA sequences
Tcf7l2	Transcription factor 7-like 2 (T-cell specific, HMG-box)
TF	Transcription factor
TFBS	Transcription factor binding site
TFIID	Transcription factor II D
TH	Tyrosine hydroxylase
Tn5	Hyperactive transposase 5
TSS	Transcription start site
VTA	Ventral tegmental area
VZ	Ventricular zone
Zfp68	Zinc finger protein 68

Summary

Complex traits are a fundamental feature of diverse organisms. Understanding the genetic architecture of a complex trait is arduous but paramount because heterogeneity is prevalent in populations and often disease-related. Genome-wide association studies have identified many genetic variants associated with complex human traits, but they can only explain a small portion of the expected heritability. This is partially because human genomes are highly diverse with large inter-personal difference. It has been estimated that every human differs from each other by at least 5 million variants. Moreover, many common variants with small effect can contribute to complex traits, but they cannot survive from stringent statistical cutoff given the currently available sample size. Mice are an ideal substitute. They are maintained in a controlled condition to minimize the variation introduced by environment. Each mouse of an inbred strain is genetically identical, but different strains bear innate genetic heterogeneity between each other, mimicking human diversity. Hence, in this work we used inbred mouse strains to study the genetic variation of complex traits. We focused on ventral midbrain, the brain region controlling motor functions and behaviors such as anxiety and fear learning that differ profoundly between inbred mouse strains. Such phenotypic diversity is directed by differences in gene expression that is controlled by *cis*- and *trans*-acting regulatory variants. Profound understanding on the genetic variation of ventral midbrain and its related phenotypic differences could pave the way to apprehend the whole genetic makeup of its associated disease phenotypes such as Parkinson's disease and schizophrenia. Therefore, we set out to investigate the *cis*- and *trans*-acting variants affecting mouse ventral midbrain by coupling tissue-level and cell type-specific transcriptomic and epigenomic data.

Transcriptomic comparison on ventral midbrains of C57BL/6J, A/J and DBA/2J, three inbred strains segregated by ~ 6 million genetic variants, pinpointed PTTG1 was the only transcription factor significantly altered at transcriptional level between the three strains. *Pttg1* ablation on C57BL/6J background led to midbrain transcriptome to shift closer to A/J and DBA/2J during aging, suggesting *Pttg1* is a novel regulator for ventral midbrain transcriptome.

As ventral midbrain is a mixture of cells, tissue level transcriptome cannot always reveal cell type-specific regulatory variation. Therefore, we set out to generate single nuclei chromatin accessibility profiles on ventral midbrains of C57BL/6J and A/J, providing a rich resource to study the transcriptional control of cellular identity and genetic diversity in this brain region. Data integration with existing single cell transcriptomes predicted the key transcription factors

controlling cell identity. Putative regulatory variants showed differential accessibility across cell types, indicating genetic variation can direct cell type-specific gene expression. Comparing chromatin accessibility between mice revealed potential *trans*-acting variation that can affect strain-specific gene expression in a given cell type.

The diverse transcriptome profiles in ventral midbrain can lead to phenotypic variation. Nigrostriatal circuit, bridging from ventral midbrain to dorsal striatum by dopaminergic neurons, is an important pathway controlling motor activity. To search for phenotypes related to dopaminergic neurons, we measured the dopamine concentration in dorsal striatum of eight inbred mouse strains. Interestingly, dopamine levels were varied among stains, suggesting it is a complex trait linked to genetic variation in ventral midbrain. To understand the genetic variation contributing to dopamine level differences, we conducted quantitative trait locus (QTL) mapping with 32 CC strains and found a QTL significantly associated with the trait on chromosome X. As expression changes are likely to be underlying the phenotypic variation, we leveraged our previous transcriptomic data from C57BL/6J and A/J to search for genes differentially expressed in the QTL locus. *Col4a6* is the most likely QTL gene because of its 9-fold expression difference between C57BL/6J and A/J. Indeed, COL4A6 has been shown to regulate axogenesis during brain development. This coincides with our observation that A/J had less axon branching in dorsal striatum than C57BL/6J, prompting us to propose that *Col4a6* can regulate the axon formation of dopaminergic neurons in embryonic stages.

Our study provides a comprehensive overview on *cis*- and *trans*-regulatory variants affecting expression phenotypes in ventral midbrain, and how they could possibly introduce phenotypic difference associated with this brain region. In addition, our single nuclei chromatin landscapes of ventral midbrain are a rich resource for analysis on gene regulation and cell identity. Our work paves the way to apprehend full genetic makeup on the gene expression control of ventral midbrain, the result of which is important to understand the genetic background of midbrain associated phenotypes.

1. Introduction

Gene expression is fundamental to every organism. Given that there are millions of cells in our body, it is puzzling how cells can achieve heterogenous expression by identical genetic material. The transcriptional control of gene expression is orchestrated by several parties, involving chromatin architecture, transcription factors, and genetic variation.

As gene expression leads to phenotypic outcome, variation in individual genomes affecting gene expression control can ultimately result in phenotypic differences in complex traits, which are often disease-associated. Profound understanding on the genetic background of complex traits require knowledge on gene expression control.

1.1 Transcriptional regulation of gene expression in eukaryotes

Gene expression is prominent in every aspect of life. It is a process using genetic code of deoxyribonucleic acid (DNA) as a template to synthesize functional genetic products like proteins or complementary ribonucleic acid chains (RNAs). Most multicellular organisms start from a single fertilized egg which later give rise to all types of cells in a living organism. This requires exquisite spatiotemporal control of gene expression. How to efficiently drive gene expression occurring in the right cell at the right time with a right amount has puzzled scientists incessantly. Another question regarding gene expression is how different types of cells achieve heterogeneous gene expression with identical genetic materials. Considering there are 25 000 genes being organized in every human cell, it is formidable to think about how precise while error-tolerated gene expression is.

How do cells control spatiotemporal expression in themselves? And how do different cell types achieve heterogeneous expression? To answer these questions, it is necessary to briefly review the process of gene expression. Gene expression is a multi-step process. Genes, the functional units of heredity, are first transcribed into messenger RNAs (mRNA) or non-coding RNAs in a process named transcription. The newly formed mRNAs then serve as blueprints to synthesize proteins which exert a myriad of biological functions to support daily activities. Regulation of gene expression can happen in the mist of all steps, from transcription initiation to post-translational modification.

To initiate transcription, the general transcription factor (TF) TFIID recognizes the TATA box (Figure 1). TATA box is a specific sequence located in promoter region upstream of

transcriptional start sites (TSS) of a gene. Once TFIID binds to TATA box, it recruits RNA polymerase II (polII) along with other general TFs to assemble pre-initiation complex (PIC). PIC is the minimum set of transcriptional machinery. Following the formation of PIC complex, polII begins mRNA synthesis by matching complementary bases to the attached DNA template. The mRNA molecule is elongated and, once the strand is completely synthesized, transcription is terminated (Pierce B.A., 2017). Some promoters do not contain a TATA box. They rely on a transcriptional initiator (INR), a DNA sequence element, to initiate transcription (Javahery et al., 1994).

Sequence-specific regions like the TATA box in promoter are important in gene expression control. They provide docking sites for regulatory factors. Almost all genes have their specific promoters upstream of TSS. Promoters are 100 to 1000 base pairs long sequences containing many binding sites for TFs. Based on vicinity to TSS, promoters can be further divided into proximal promotor and core promotor (Figure 1). Core promoter is particularly essential for binding and correctly positioning PIC (Lenhard et al., 2012). The selectivity of transcriptional initiation is in part due to different sequence architectures of promoters. For instance, high-GC content promoters are found in universally expressed genes like housekeeping genes or developmental genes; while low-GC promoters usually control the expression of tissue-specific genes (Sandelin et al., 2007). Promoter is only one of the many sequence-specific regions that can regulate transcription. Many other sequence elements can participate in transcriptional control by recruiting regulatory factors to interact with transcriptional machinery or modify chromatin structure.

In addition to sequence specific elements, TFs are another group of essential players to control gene expression. Similar to TFIID, other TFs can also recognize specific DNA binding sites to either inhibit or initiate transcription. Their binding events depend on the chromatin states around DNA. If chromatin is highly compact, TFs cannot make their way to physically interact with DNA. On the other hand, if genomic region is exposed, it would be much easier for TFs to locate themselves. In general, multiple parties, including sequence-specific elements, TFs and chromatin states, orchestrate the regulatory control of spatiotemporal and heterogeneous gene expression across cells.

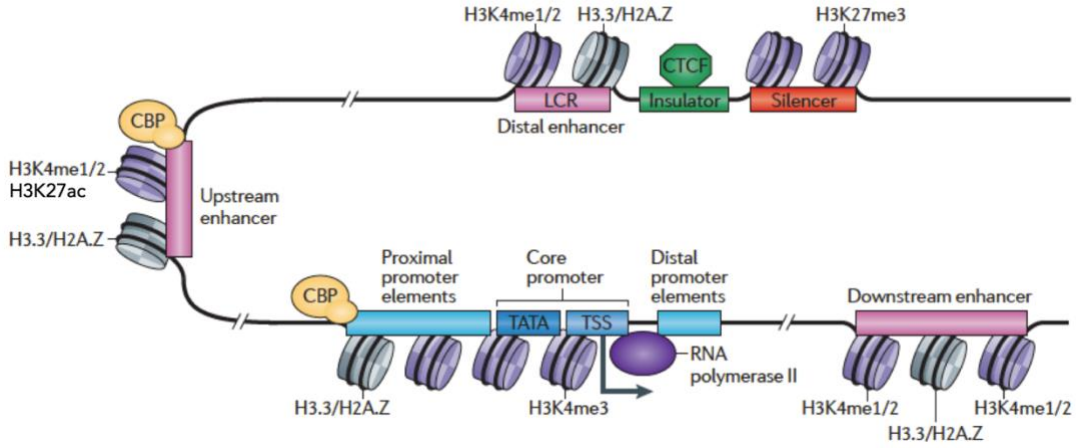


Figure 1: Representation of transcriptional regulatory elements adapted from (Ong and Corces, 2011). Based on the proximity to TSS, promoter can be classified into core promoter and proximal promoter. Core promotor often contains the TATA box for TF binding. Gene transcription can be influenced by upstream or downstream enhancers and distal enhancers. Histone modifications can be found on nucleosomes overlapping with enhancer: H3K4me1/2 for poised enhancer; H3K27ac for gene activation or active enhancers; H3K27me3 for gene repression. Histone variants H3.3/H2A.Z are highly unstable, so they are linked to active gene transcription. CBP is often found coactivating with p300 to interact with TFs to increase the expression of target genes. LCR contains a group of transcriptional regulatory elements. Silencer prevents gene expression by interacting with PRC2. Insulator is a boundary element responsible for domain-bordering and often associated with CTCF. CBP: cyclic AMP-responsive element-binding protein; CTCF: CCCTC-binding factor; H3K4me1/2, histone H3 mono- or dimethylation at lysine 4; H3K4me3, histone H3 trimethylation at lysine 4; H3K27ac, histone H3 acetylation at lysine 27; H3K27me3, histone H3 trimethylation at lysine 27; H3.3/H2A.Z, histone variants H3.3 and H2A.Z; LCR: locus control region; TATA, 5'-TATAAA-3' core DNA sequences; TSS, transcriptional start site.

1.1.1 Chromatin structure and related function

Human genome has more than 6 billion base pairs (bp) directing the development and the biological function of more than 210 cell types (Trapnell, 2015). If each DNA molecule in a human cell is linked head to toe, it will be two meters long. However, a cell nucleus is very tiny so that DNA must be scrupulously packaged to fit in. As a result, DNA endorses a delicate multi-level strategy to suit itself into a confined nucleus. The structure of DNA can be considered at three hierarchical levels: the primary and secondary levels are its nucleotide sequence and the double-stranded helix; and the tertiary level is the higher order that DNA molecules use to fold into the small confinement of nucleus to produce a chromosome (Figure 2) (Pierce B.A., 2017). DNA's higher order structure involves DNA molecule and its associated protein, the two forming a substance called chromatin. Chromatin is not static but constantly undergoing alteration. Dynamic modification of chromatin is one of the major apparatuses for cells to achieve cell type-specific gene expression.

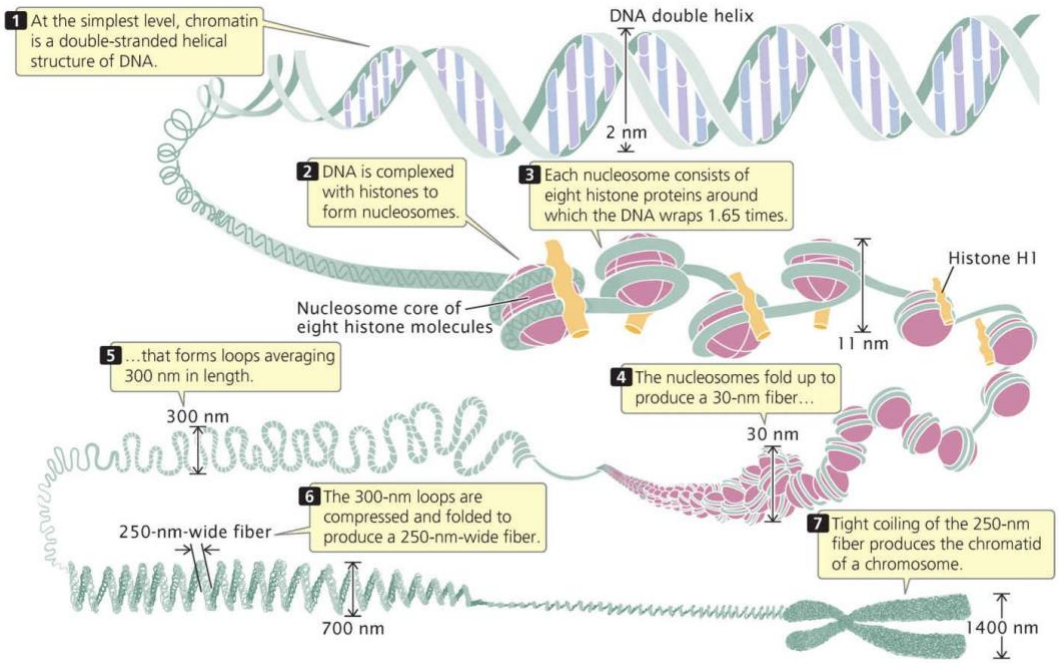


Figure 2: Representation of DNA packaging in eukaryotic cells (Pierce B.A., 2017). Each chromosome contains a single DNA molecule and its associated proteins. The double-helix DNA wraps around histone octamer (with 2 copies of each H2A, H2B, H3, H4) to form nucleosomes. H1 histones guard DNA entering and leaving nucleosomes correctly. Nucleosomes can pack with neighbors to fold into a 30-nm-wide fiber, which further loops around and compress into a 250-nm-wide fiber. Eventually, the fiber tightly coil to produce the chromatid of a chromosome.

1.1.1.1 Higher order of chromatin structure

If we open up the chromosome, we can find DNA wraps around protein cores to create a "beads-on-string" scenario (Figure 2). Each "bead", namely nucleosome, has 145-147 bp of DNA wrapping ~ two times on a protein core. The protein core is an octamer with two copies of each H2A, H2B, H3 and H4. Histone families all have high composition of arginine and lysine whose positive charges attract the negatively charged phosphates of DNA (Erler et al., 2014). Thanks to the electrostatic force, histone proteins can hold onto DNA molecule. Each histone protein has a flexible N-terminal tail, around 11 - 37 amino acids, extruding out of the nucleosome. Histone tails can affect chromatin structure by interacting with nearby nucleosomes to form compact or loose chromatin. "Strings", DNA sequences in between nucleosomes, are linker DNAs bound by non-histone proteins. Following "beads-on-string" structure, nucleosomes fold onto themselves, creating a series of loops and eventually tightly coiling together to produce the chromatid of a chromosome (Pierce B.A., 2017).

The structure of chromatin is highly complex and dynamic. Chromosomal regions that stay constantly compact is called heterochromatin. Heterochromatin is found in centromeres and

telomeres, as well as in most of the Y chromosome. They usually do not have transcription and crossing over happened because its dense nature prevents most transcription events from happening. On the other hand, chromatin undergoing condensation and decondensation during cell cycle is called euchromatin (Klemm et al., 2019). As DNA doubles itself during cell cycle, chromatin undergoes dynamic packaging alongside. During mitosis, euchromatin is dramatically compacted to form mitotic chromatin which can be visualized by microscope (Ginno et al., 2018). Euchromatin is often found on chromosome arms and is enriched for genes. Euchromatin is of important biological relevance, because transcription mainly happens in euchromatin.

As nuclear space is in 3D structure, chromatin has its spatial distribution in nucleus, called the 3D organization of chromatin (Rowley and Corces, 2018). The spatial distribution of chromatin often directs long distance chromatin interaction to affect DNA replication and transcription. Such contact could be obtained from molecular technique Hi-C (Lieberman-Aiden et al., 2009). Based on contact maps, chromatin can be segregated into two types, referred to as A and B compartments. A compartments contain genes that are actively transcribed, and B compartments often have inactive genes. Within each compartment, small scale domains with preferential internal interactions can be found. These are named topologically associating domains (TADs) (Dixon et al., 2012). Chromosome compartmentalization is a key feature of higher-order genome organization and has important function on gene regulation (Szabo et al., 2019).

Because transcription occurs mainly in open chromatin, switching between closed and open chromatin states can achieve selective gene expression. Considering how dramatically chromatin structure can affect gene expression, it is never grandiose for eukaryotes to adopt enough strategies to properly confine DNA in nucleus. Then how do cells control their dynamic chromatin states? The answer lies in dynamic remodeling of chromatin and histone modifications.

1.1.1.2 Chromatin accessibility and histone Modifications

To pack the long stretched DNA molecule into nucleus, DNA wraps around histone proteins to form nucleosomes which subsequently coil into chromatin. As DNA is tightly attached to histone proteins, it is inaccessible to DNA-binding factors such as polII to initiate transcription. Therefore, nucleosome by nature is repressing gene expression. To be able to initiate

transcription, chromatin structure undergoes modification so that the transcriptional signal can reach its destination. There are at least two mechanisms available in eukaryotes to alter chromatin structure, including chromatin remodeling and histone modifications.

Chromatin remodeling relies on chromatin remodelers to alter the structure of chromatin. Chromatin remodelers are enzymes that can alter chromatin structure independent of altering its chemical structure. Though diverse in protein composition, all chromatin remodelers harbor a subunit with a conserved ATPase domain which converts the energy from ATP hydrolysis into mechanical force to alter chromatin structure (Clapier et al., 2017). These chromatin-remodeling complexes can recognize specific chromatin signals such as DNA sequence, histone modification, or histone variants, and expose regulatory elements to the action of transcriptional regulators. The most well-studied chromatin-remodeling complex is the SWI-SNF complex, which deploys the energy from ATP hydrolysis to nucleosome repositioning to free DNA for transcription (Roberts and Orkin, 2004).

On the other hand, unlike chromatin remodeling, histone modifications alter chromatin structure by changing its molecular composition. Histone proteins in nucleosome are composed of a core domain and a tail domain. The tail domain protrudes from the nucleosome and is commonly packed with different chemical modifications such as methylation, acetylation, and phosphorylation. Combinations of chemical modifications on histone tails are referred to as histone code (Jenuwein and Allis, 2001). One of the well-studied histone modifications is histone methylation. Histone methylation is associated with activation or repression of transcription depending on which histone or amino acid is methylated. Histone methyltransferase can methylate lysine or arginine on the histone tail, while histone demethylase can strip the attached methyl groups away. Histone methyltransferase and demethylase often do not interact with DNA directly but rather are recruited by sequence-specific binding proteins or RNA molecules such as Polycomb repressor complex 2 (PRC2) (March and Farrona, 2017). Up to three methyl groups can be transferred on histone lysine in different locations, and the resulting impact on transcription varies dramatically. H3K4me3, indicating tri-methylation of lysine on the 4^{th} residue, associates with open chromatin by recruitment of a plant homeodomain (PHD) finger of nucleosome remodeling factor (NURF) (Wysocka et al., 2006). H3K4me3 is highly enriched in the promoters of active transcribed genes (Figure 1). On the other hand, H3K27me3, tri-methylation of lysine on the 27^{th} residue, is often found in association with gene repression and deposited by PRC2 (Greer and Shi, 2012;

Mierlo et al., 2019), the enzyme with methyltransferase activity playing a key role in gene silencing and establishment of heterochromatin.

Histone acetylation is another prevalent histone mark and often comes together with increasing transcription. Acetyl groups are transferred to lysine residues by histone acetyltransferases. The deposition of acetyl groups reduces positive charges of histone proteins, so DNA are less tightly gripped and exposed to regulatory control (Bannister and Kouzarides, 2011). Similar to histone methylation, acetylation can also be removed by deacetylases. Many TF cofactors have acetyltransferase activity such as p300/CBP which was shown to position at promoter regions via association with TFs (Soutoglou et al., 2001). There are other kinds of modifications that cells employ to control chromatin states such as histone phosphorylation and ubiquitination which are important for the response to DNA damage and gene silencing (Cao and Yan, 2012; Rossetto et al., 2012). Histone marks interact with each other to exert their control on gene expression. Proper deposition and interpretation of histone marks are fully relied on the aforementioned histone modification enzymes, namely writer (setting up histone modifications), eraser (removing histone modifications), and reader (interpreting histone modifications) (Figure3) (Højfeldt et al., 2013; Hyun et al., 2017).

Chromatin state is characterized by a distinct array of histone marks. Histone modifications often associated with active genes are lysine acetylation on H3/H4, H3K4me3, H3K79me3, and H3K36me3. Marks associated with gene repression are H3K27me3 and H3K9me3. The combinatorial complexity of histone marks is vast and has been demonstrated by ChIP-seq (chromatin immunoprecipitation with massively parallel DNA sequencing). Mapping of histone marks on genome found that combinations of modifications either occur together, or are mutually exclusive, suggesting crosstalk is required to have proper transcriptional regulation (Zhang et al., 2015). H3K4me3 can promote downstream H3/H4 acetylation through recruitment of histone acetylation enzyme, indicating a positive feedback loop between the two histone marks. H3K4me3 and H3K9ac are an example of combinatorial interaction of histone marks. SGF29, a reader for H3K4me3, is a component of the SAGA histone acetylation complex. The deletion of SGF29 can lead to the loss of H3K9ac and the absence of SAGA complex at target sites (Bian et al., 2011). Mutual exclusion can also be found on genome. H3K9 methylation is mediated by SETDB1 and often associated with heterochromatin formation (Brower-Toland et al., 2009), while H3K27me3 is mainly deposited by PRC2 to repress gene expression. H3K9 methylation can crosstalk with H3K27me3 and they two are

mutually exclusive present at constitutive heterochromatin. This has been shown in one-cell stage mouse embryo, where H3K27me3 is present in male pericentric heterochromatin but not in female heterochromatin which contains H3K9me3 (Puschendorf et al., 2008). Interestingly, active marks and repressive marks can coexist in the same promoter region, namely bivalent promoter. Bivalent promoters can be switched between active and inactive states to achieve timely control on gene expression. It has been shown that bivalent promoters are prevalent in developmental stages. For example, H3K4me3 can be co-deposited with H3K27me3 to poise expression of developmental genes. During neurogenesis, some bivalent genes become expressed and lose the H3K27me3 mark, while those that stay silent lose H3K4me3 but retain H3K27me3 (Bernstein et al., 2006; Voigt et al., 2013).

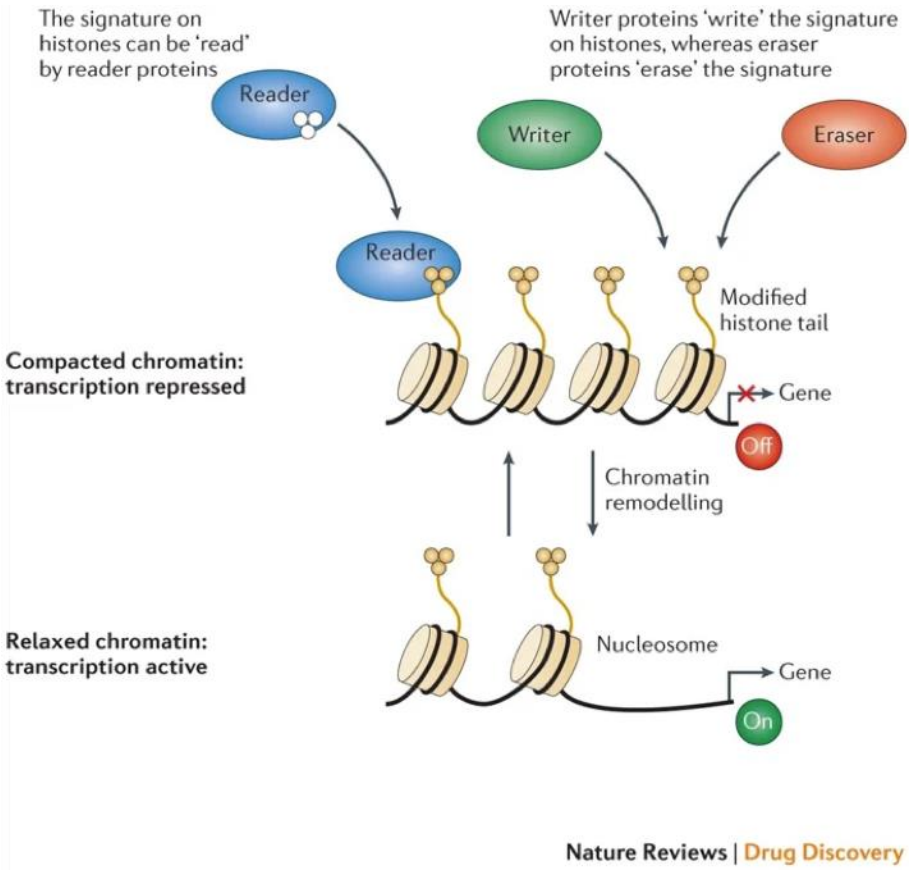


Figure 3: Representation of reader, writer, and eraser enzymes (Højfeldt et al., 2013). Compact chromatin is often transcription repressed, while relaxed/open chromatin is transcription active. Reader can read the histone signatures on nucleosome. Writer can deposit chromatin signatures which can be removed by eraser.

Because the relationship between histone marks and associated cellular activities is quite specific, scientists have been mapping chromatin modifications to help understand their relationship with gene transcriptional control. Such community effort has been collected by

ENCODE (Davis et al., 2018), Roadmap Epigenomics (Kundaje et al., 2015) and IHEC (International Human Epigenome Consortium) (Bujold et al., 2016), which together provide a fruitful collection of histone modification patterns on many mouse and human cell lines and primary tissues. These data were produced by various techniques covering a wide range of histone modifications ready to be analyzed.

ChIP-seq is widely used to detect genome-wide TF binding sites and histone modifications (Barski and Zhao, 2009). It relies on specific antibodies designed to bind against histone modification or TF of interest. After antibody binding to specific target, the complex could be pulled down or precipitated by magnetic beads. ChIP-seq has largely pushed our understanding on TFs binding sites on regulatory elements in the genome. However, it is not easy to have a working antibody. Instead, scientists looked at techniques that could detect open chromatin, chromatin regions that are accessible for regulatory factors. DNase-seq (deoxyribonuclease I [DNase I] hypersensitivity sites sequencing) is a method to identify the location of accessible regulatory regions (Song and Crawford, 2010). Regulatory regions are often free of nucleosomes, so they are sensitive to cleavage of DNase I. Such regions are called DNase I hypersensitivity sites (DHS). DHSs are of particular biological relevance because they are highly correlated with active transcription. Study with twenty diverse human cell lines found cell type-specific gene expression greatly associates with DHSs (Marstrand and Storey, 2014). Similar to DNase-seq, ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is an alternative method to detect open chromatin regions (Buenrostro et al., 2015) but faster and more sensitive, as it does not require enzymatic digestion. ATAC-seq depends on the hyperactive transposase 5 (Tn5) which universally cuts and ligates adapters to open chromatin region. However, the information from DNase-seq or ATAC-seq does not guarantee TF binding. To address this question, many computational methods have been developed in the past years to infer TF binding sites on open chromatin. The idea is based on the fact that when TF occupies genomic region, the same binding site is no longer available for DNase I or Tn5 interaction. Therefore, mapping sequencing reads from DNase-seq or ATAC-seq on genome would create small bumps on top of the pile-up peaks. Such small bumps, namely footprints, are enriched for TF binding sites (Galas and Schmitz, 1978). Tools like HINT-ATAC (Li et al., 2019b) are specifically designed to scan for footprints in ATAC-seq data. As TFs have high binding affinity towards sequence-specific regions, one can decipher which TF likely binds to a footprint by motif enrichment analysis using HOMER (Heinz et al., 2010) or MEME (Bailey et al., 2009).

Chromatin remodeling alongside with histone modifications gives credence to spatiotemporal control of gene expression. However, there are billions of nucleotides on eukaryotic genome, not every free region is given an equal amount of significance. Fortunately, regulatory proteins such as TFs have high affinity to sequence-specific regions like the TATA box in promoter. Enhancers, another group of specific genomic sequences, have a greater implication on fine tuning gene expression.

1.1.1.3 Enhancer

Though enzymes to establish histone modifications are well-characterized, the mechanisms on how to deposit histone marks is still blurry. Enhancers are DNA sequences that contain multiple TF binding sites and participate in histone mark deposition. They can locate in introns of genes or intergenic regions. Enhancers can activate gene transcription independent of their location, sequence orientation and distance to target genes (Ong and Corces, 2011). In some cases, they can even activate transcription on another chromosome (Bateman et al., 2012). One central question in understanding the function of enhancer is how distinct regulatory elements coordinate to act on gene expression through enhancer.

Enhancers can affect transcription in several ways through binding of regulatory factors. If an enhancer is proximal to a gene, the bound TFs can act directly on transcription by either promoting or inhibiting the assembly of transcriptional machinery. On the other hand, if TFs interact with a distal enhancer, long-range enhancer-promoter contacts can be established by TF-mediator complex which loops the chromatin around so that the mediator recruits polII and stabilize PIC formation in the target gene promoter. Some active enhancers can even be transcribed into non-coding RNAs such as enhancer RNAs (eRNAs) (Andersson and Sandelin, 2020; Kaikkonen and Adelman, 2018; Schoenfelder and Fraser, 2019). eRNAs function through interplay with cohesin to stabilize long-range enhancer-promoter interaction (Hsieh et al., 2014; Lai et al., 2015; Pnueli et al., 2015). In general, exerting proper regulatory control with enhancer require dynamic interplay between chromatin structure, regulatory factors, and enhancer itself.

Chromatin alteration such as nucleosome dynamics and histone modifications, can affect enhancer's functionality. It has been shown that nucleosomes are less occupied in TSS, suggesting gene transcription is correlated with nucleosome occupancy (Ong and Corces, 2011). Because recruitment of regulatory proteins requires freeing enhancers from nucleosome,

it is not surprising to find histone variants play a part in nucleosome dynamics as they are relatively unstable by nature. This is supported by observations that H3.3/H2A.Z histone variants are enriched in nucleosome-free regions in HELA cells (Jin et al., 2009) and CD4+ T cells (Barski et al., 2007). In addition, H2A.Z enrichment was shown to be positively correlated with gene expression (Hu et al., 2013). Therefore, histone variants affecting nucleosome dynamics could be an important feature of many enhancers.

Histone modifications is also important for enhancer to control cell-type specific gene expression. This is demonstrated by study of p300 binding sites on multiple mouse tissues (Visel et al., 2009). p300 is a well-characterized co-activator recruited by many TFs to activate gene expression. The binding sites of p300 are correlated with tissue-specific gene expression, indicating histone modifications are heterogeneous across cell types. p300 itself has acetyltransferase activity and can deposit acetyl groups on histone. In fact, enhancers control cell-type specific gene expression by acquiring specific histone marks. These histone marks recruit different TFs to achieve context-dependent gene expression (Smith and Shilatifard, 2010). Such histone marks are acquired even before transcription and they serve as an epigenetic memory for progenitors committed to different cell lineages (Kim et al., 2010). Human embryonic stem cell (ESCs) differentiation involves switching between different histone signatures (Rada-Iglesias et al., 2011). Enhancer elements with H3K4me1 and H3K27ac are found proximally to active transcribed genes in human ESCs, whereas silenced genes are dominant by H3K4me1 and H3K27me3 instead. However, the silenced genes can be turned on by replacing H3K27me3 with H3K27ac in later developmental stages. The switch of histone marks leading to transcriptional changes is also observed in preadipocytes, as poised enhancers (deposited with H3K4m1/2) become activated by acquiring H3K27ac and binding of MED1, SMC1 and p300 (Siersbæk et al., 2017). Taken together, with different combinations of histone marks, enhancers can coordinate with different regulatory proteins to achieve heterogeneous gene expression tailored to different cell types.

Based on p300 binding sites or genome-wide histone marks, a single cell type is estimated to have 10 000 – 150 000 putative enhancers (Pott and Lieb, 2015). How many functional enhancers are actually needed in a given cell type under a give condition? Researchers found out some enhancer regions are far more active than others. Super-enhancers are characterized by groups of enhancers in close genomic proximity with high levels of mediator binding (Pott and Lieb, 2015). Taking mouse ESCs as an example, super-enhancers can be defined as

followed: (1) genomic sites bound by master regulators OCT4, SOX2 and NANOG are defined as enhancers; (2) enhancers within 12.5 kb are stitched together; (3) mediator (MED1) binding affinity is ranked from lowest to highest on every "stitched" enhancer to create MED1 - enhancer affinity curve. Enhancers with slope in MED1 - enhancer affinity curve bigger than one are classified as super-enhancers. Super-enhancers have a lot of unique characteristics comparing to normal enhancers. Because closely located enhancers are stitched together, median size of super-enhancer is an order of magnitude longer than normal enhancers, with 8667 bp comparing to 703 bp in mESCs (Whyte et al., 2013). Therefore, it is not surprising to find super-enhancers are highly enriched with many regulatory elements and constantly active.

Super-enhancers usually establish long-range promoter-enhancer contact. Activate super-enhancers are often marked by their flanking nucleosomes with histone acetylation like H3K27ac (Calo and Wysocka, 2013). They can recruit tissue or developmental-specific TFs and cofactors (Pott and Lieb, 2015). In the mist of bound factors, cohesin and mediator coordinate to form physical contact between enhancer and target gene promoter (Soutourina, 2019). Meanwhile, polII can transcribe eRNAs based on enhancer sequence in both directions (Andersson et al., 2014), suggesting the involvement of eRNAs in mediating super-enhancer function. The transcription of mouse *Igh (immunoglobulin heavy-chain)* is controlled by a downstream super-enhancer dependent on PolII pausing and elongation factor SPT5. Deletion of *Spt5* leads to the loss of super-enhancer-promoter physical contact and *Igh* gene expression, but shows no effect on TF-mediator complex assembly. Restoring transcription on *Igh* supper-enhancers brings back enhancer-promoter interaction (Fitz et al., 2020), indicating eRNAs regulate super-enhancer by stabilizing long-range enhancer-promoter interaction and is independent of TF/cofactor binding on super-enhancer.

Interestingly, multiple enhancers with similar activities can be found near the same gene. The function of these enhancers are redundant, which is important to protect key developmental events from disruption. This is supported by study using individual or combinatorial enhancer deletion on loci required for limb development (Osterwalder et al., 2018). None of the single mutation could lead to obvious defect in limb morphology, but removal of pairs of limb enhancers near the same gene led to obvious phenotypes, suggesting enhancer redundancy gives phenotypic robustness to loss-of-function mutation.

In summary, enhancers are important sequence elements to control cell type-specific gene expression by recruiting regulatory factors to affect a series of events including nucleosome dynamics, histone modifications and eRNA synthesis for long-range contact. All these events cannot happen without recruiting regulatory factors binding to enhancer regions.

1.1.2 Transcription factors in gene expression

TFs are distinctive proteins with at least one DNA binding domain (DBD) and capable of regulating transcription (Lambert et al., 2018). They can recognize specific DNA sequences like enhancers to control chromatin structures and gene expression, forming a complex network to regulate cell differentiation during embryonic development. Context-dependent gene expression is a corollary of the existence of TFs. There are up to 1600 TFs in human genome functioning mostly in collaboration to exert biological function (Lambert et al., 2018). For example, master regulators, a set of TFs controlling gene expression specifying cell types, are important for cells to narrow down their dedicated differentiation lineages (Oestreich and Weinmann, 2012). Some of the master regulators can even induce cells to undertake dedifferentiation or transdifferentiation. Yamanaka factors (*Oct2*, *Sox2*, *Klf4*, *Myc*) can remodel nucleosome dynamics in closed chromatin, promoting terminally differentiated cells retreated to pluripotent stage by inducing expression of developmental genes (Takahashi & Yamanaka, 2006). Moreover, same TFs can regulate different genes in different cell types, as well as recruiting dissimilar cofactors based on context. Because TFs participate in all facets of gene expression control, it is vital to understand how they recognize DNA sequence patterns, and more importantly, how they regulate gene expression upon binding to DNA.

1.1.2.1 TF binding to DNA sequence

The first step for TFs to function on gene expression is recognizing specific binding sequences on DNA. The binding sequences of TFs are represented as motifs, a set of sequence logos underlying by quantitative methods such as the positioning weight matrix (PWM) (Stormo and Zhao, 2010). Extensive effects have been spent to characterize human and mouse TF binding motifs (Jolma et al., 2013, 2015). As proteins which are related in amino acid sequence bind to similar sites, it is possible that one single motif can be bound by multiple TFs. Analysis with PWM models revealed most TFs do not need multiple PWMs to explain the high-affinity to DNA, but multiple binding modes do exist for some factors.

TFs have preferred binding to each base of motifs with up to 1000 times higher affinity comparing to random sequences (Lambert et al., 2018). Because TF binding sites are short (6 to 12 bp) and flexible, genes usually have multiple binding sites for different TFs given a typical human gene is more than 20 kb. Such scenario adds extra complexity for TFs to control gene expression by acting cooperatively. In fact, TFs often do not function alone but rather collaborate to attain required specificity on gene expression control (Long et al., 2016). There are many ways for TFs to achieve cooperative binding. Protein-protein interactions can give additional stability to the established complexes. TFs and cofactors aid each other in DNA binding by forming homodimers, trimers, or other higher-order structures (Lambert et al., 2018). With cooperative binding, occupancy of TF-TF or TF-cofactor complexes on binding sites can temporarily elongate to achieve longer lasting gene expression. In addition, interaction with DNA molecules also facilitates recruitment of regulatory proteins. Sometimes, the binding of the first TFs can remodel DNA structure, which in turn helps the recruitment of subsequent TF. For example, pioneering factors, usually abundantly expressed during development, can bind to heterochromatin and aid the following binding of non-pioneering factors (Cirillo et al., 2002; Mayran et al., 2019).

Up to date, 1639 TFs has been manually curated in human, the information of which is stored on HumanTFs (http://humantfs.ccbr.utoronto.ca/) (Lambert et al., 2018). Most of the curated human TFs contain at least one of the two DBD types: C2H2-ZFs which are less conserved and only one-third of them with a known motif, and homeodomains which are well-conserved and almost all of them have a known motif. Different types of TFs have different specificity in tissues. For example, C2H2-ZFs are more universally present across tissues comparing to other types of TFs (Lambert et al., 2018), probably because they can repress the expression of transposable elements which is required in general across cell types (Ecco et al., 2017).

Because different tissues adopt distinct sets of TFs for gene expression control, cell-type specific information is required to understand how TFs function across tissues, where single-cell RNA-seq can shed some lights on. A study with single-cell transcriptomics on 20 mouse organs identified cell type-enriched TFs, which provides rich information to design reprogramming protocols to study underlying regulatory networks (Schaum et al., 2018). Interestingly, clustering cell types by TFs' expression have clearer separation comparing to other features like cell-surface markers, RNA splicing factors or two combined, further highlighting the importance of cell type-identity TFs on gene expression specificity.

As TFs can affect gene expression which result in different phenotypes in organisms, intuitively, mutations or sequence variants on TFs could lead to phenotypic difference or even disease phenotypes. There are 313 TFs associated with at least one disease phenotype in human (Köhler et al., 2014). Depending on where the mutations or variants are, a range of consequences can be triggered. Mutations on DBD can directly affect TF sequence specificity, and mutations outside DBD can alter protein structure which might alter protein-protein interaction. However, because TFs are in general pivotal for important biological processes like embryonic development, mutations leading to dramatic effects are usually lethal and cannot be observed. Therefore, the wide varieties of phenotypes that we can observe are in fact largely due to genetic variation on TF binding sites instead of on TFs themselves. It has been estimated that up to 85 - 93% of common disease-associated genetic variation can impact gene expression (Hindorff et al., 2009; Maurano et al., 2012).

1.1.2.2 TF effectors

Upon DNA binding, how TFs affect gene transcription is dramatically varied. The simplest way is binding to promoter region to facilitate the assembly of transcriptional machinery. Additionally, TFs can hire regulatory proteins to function on specific phase of transcription through their effectors. The effector domain is the modular subunit that TFs use to bind other regulatory factors (Figure 4) (Frietze and Farnham, 2011). One TF can have multiple effectors to interact with different parties like proteins, ligands, or nearby chromatin.

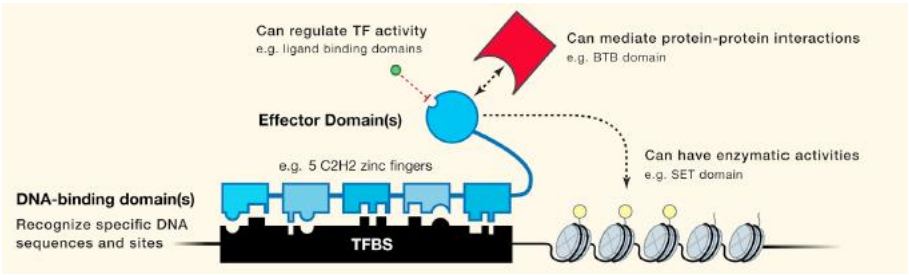


Figure 4: Representation of effector domain of TFs (Lambert et al., 2018). TFs have DNA-binding domains like 5C2H2 zinc fingers to recognize specific DNA sequences, namely TFBS. The effector domain is the subunit of TF that is uses to bind other regulatory factors. For example, ligand binding domains can regulate TF activity, BTB domain can mediate protein-protein interaction, and SET domain can have enzymatic activities on nearby chromatin. TFBS: transcription factor binding site.

Cofactors can be either coactivators or corepressors. They usually have multiple domains acting on a series of biological activities such as chromatin binding, nucleosome remodeling, and histone modification. Though there are many cofactors available for TFs to interact with,

some of them are universally present in cell types. An example of ubiquitously expressed cofactor is the mediator, a complex with more than 30 subunits (Yin and Wang, 2014). Mediator functions as a bridge between TFs and polII and regulate most of the transcribed genes. However, not all TFs have a clear effector. Some of the TFs do not have an effector, either because their effectors are unknown, or these TFs simply occupy the binding sites to occlude the binding of other proteins (Lambert et al., 2018).

TF-cofactor complex interaction is largely dependent on the state and structure of TF, such as whether it is phosphorylated or bound by ligands. For instance, PU.1 and IRF8 are important TFs for B-cell development. The phosphorylation of IRF8 can promote cooperative binding with PU.1 to regulate gene expression (Mohaghegh et al., 2019). Cooperative binding of TF-TF or TF-cofactor complex have specific binding affinity to heterodimeric motifs, where individual TF could have binding affinity but appear to be much weaker comparing to the complex (Jolma et al., 2015).

Taken together, TFs cooperate with each other or other cofactors to control gene expression. Such interaction is high dynamic and context-dependent. Mutations affecting TFs themselves are disease-related but less common due to potential deleterious results (Barrera et al., 2016; Bejerano et al., 2004). The control of gene expression is fundamental to the wide range of phenotypes that we observe in organisms. How do regulatory elements as discussed above cooperate to contribute to phenotypic difference? The answer lies in the genetic variation of genomes.

1.2 Genetic variation

Genetic variation is the difference in DNA among individuals (Pierce B.A., 2017). With the completion of primary sequence of human genomes from 26 populations by the 1000 Genome Project Consortium (Auton et al., 2015), different kinds of genetic variation has been studied in a finer resolution. It has been estimated that there are over 88 million SNPs in human genome (Auton et al., 2015). Small scale variants include single nucleotide polymorphism (SNP) representing a single difference in DNA nucleotide (Figure 4), and various repetitive elements that involve relatively short DNA sequences like micro- and minisatellites. These small scale variants constitute most genetic variation in human genome. Another group of genetic variation is the structural variants often associating with many nucleotides (Ho et al., 2020). Structural variants can be classified into insertion-deletion variant, block substitution, inversion variant

and copy number variant (CNV) (Figure 5) depending on the type of variation (Frazer et al., 2009). Some studies also classified structural variants in submicroscopic and microscopic variants based on the size of affected nucleotides. Submicroscopic structural variants range from ~ 1kb to 3 Mb such as large CNV, and microscopic structural variants with length more than 3 Mb including aneuploidies or reciprocal translocations (Feuk et al., 2006). Though structural variants are far less common, they can lead to phenotypic variation or disease as small scale variants do.

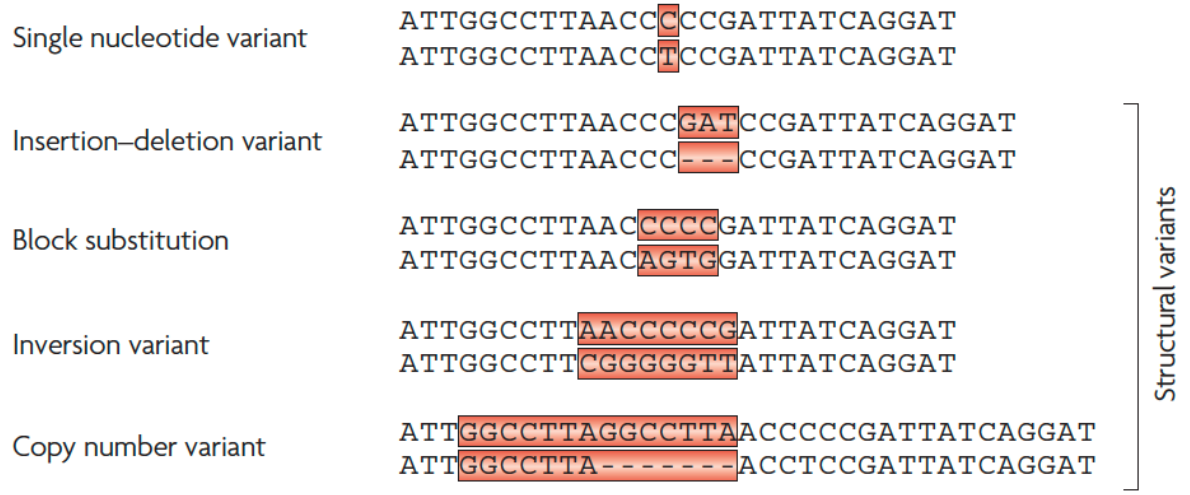


Figure 5: Representation of different types of genetic variation (Frazer et al., 2009). Single nucleotide variant is a single difference on nucleotide. Structural variants usually involve several nucleotides, like insertion-deletion variant, block substitution, inversion variant and copy number variant.

How does genetic variation arise in the genome? The molecular sources for genetic variation are mainly gene mutation and genetic recombination. Gene mutation is a permanent alteration in the DNA sequence. It can appear anywhere is the genome affecting a single or several kilo bases segment of a chromosome. There are two types of mutations, the hereditary mutation in germ cells which appear in every cell of a living organism and can be transferred to the offspring, and the somatic mutation only possessed by a group of cells in the body and generally not inheritable (Pierce B.A., 2017). Germline mutation is important to phenotypes as the outcome can be preserved in a population. Additionally, genetic recombination does not change DNA sequence per se but enables the exchange of genetic materials between paternal and maternal chromosomes which lead to novel combination of variants in the daughter germ cells (Silver LM, 1995). In the aspect of evolution in a population, genetic variation also goes

through natural selection directing by differential reproductive success, genetic drift by random external force, or gene flow moving genes in or out of a population.

The reason why genetic variation attracts so much attention is its causal effect on phenotypes. Genetic variation on protein coding sequence can have a direct effect on protein function. Sickle cell disease is a common inherited blood disorder caused by a SNP on the protein-coding region of hemoglobin, which results in glutamic acid being substituted by valine. The mutated hemoglobin forms a hydrophobic patch under low oxygen condition, giving the sickle shape of blood cells. However, genetic variation on protein-coding region is less common (Auton et al., 2015). In fact, genome-wide association studies (GWAS), a statistical method to relate genetic variation to phenotypic diversity, found most of the variants to locate in non-coding genome (Visscher et al., 2017). Of 500 unique trait/disease-associated SNPs from more than 150 GWAS studies, only 4.9% are located in protein-coding regions, while the rest are found in introns, promoters, and other distal regions away from genes (Auton et al., 2015). Moreover, based on genome-wide characterization of DHSs, 76.6% of GWAS SNPs overlap with DHSs or are in close linkage disequilibrium (LD) with SNPs in DHS (Maurano et al., 2012). As DHSs are enriched for TF-DNA binding sites, trait/disease-associated variants can potentially disrupt TF-DNA interaction to introduce variation in gene expression and ultimately difference in phenotypes.

Since more than 88 million variants are detected on human genome, and on average every human being differs from each other by approximately 5 million (Auton et al., 2015), the tremendous amount of possible permutations establishes the genetic basis of complex traits. Not every genetic variation can contribute to variation of phenotypes, and different phenotypes result from dissimilar sets of variants. All these layers add extra complexity on identifying the causal variants that directly contribute to phenotypic variance.

1.2.1 Regulatory variation

Genetic variants affecting gene expression are regulatory variants or functional variants. Regulatory variants exert impact on gene expression by positioning on regulatory regions of the genome. For example, regulatory variants in gene promoters can alter gene expression by interrupting transcriptional machinery binding. CDC7 is a kinase involved in the regulation of cell cycle at S phase by promoting DNA replication. Functional variant within the promoter region of CDC7 can disrupt GLIS2 binding, hence influence the transcription of *Cdc7* (Yang

et al., 2019). The aforementioned example has the regulatory variant in close proximity to the target gene. Such variants are said to have *cis*-regulatory effect on transcription (Figure 6). *Cis*-regulatory variants can also present on nearby or distal enhancers which are brought closer to target genes by DNA looping. On the contrary, some variants have *trans*-regulatory effect, meaning the regulatory effect from the variants is indirect. *Trans*-regulatory variants can exert their effects on gene expression through numerous different mechanisms. In one prominent scenario, the abundance or the nature of a TF is altered by a variant, leading to changes in gene expression in target genes.

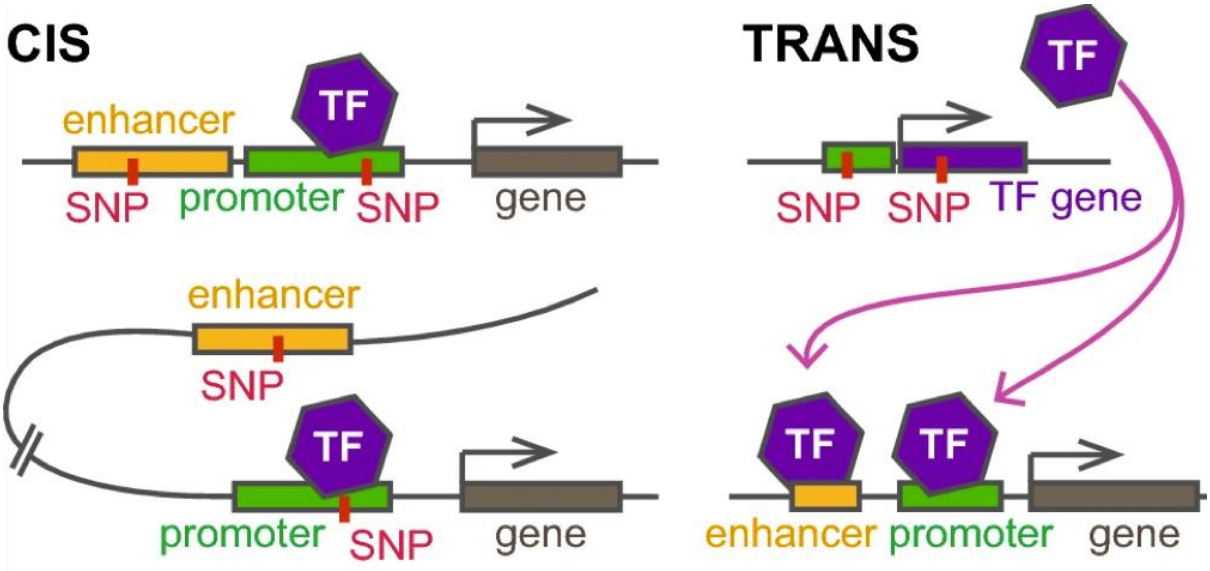


Figure 6: Representation of *cis* and *trans* effect of regulatory variants (Ohnmacht et al., 2020). *Cis*-regulatory variants could be present on enhancer or promoter affecting proximal gene expression, while *trans*-regulatory variants exert their effect on gene expression by altering TF itself or its abundance.

Both *cis*- and *trans*-regulatory variants involve in the interplay between variants themselves and the resulted TF-DNA binding. How trait-associated variants lead to changes in phenotype is rooted in the idea that regulatory variants can disrupt TF-DNA binding; therefore resulting in alteration in gene expression and ultimately introducing variation in phenotype. So intuitively, TF-DNA interaction can be key drivers of phenotypic variation. However, as the majority of TF-DNA binding events are not driven by sequence alterations in the motif, the interplay between regulatory variants and the resulted TF-DNA binding is far more complicated than previously estimated (Deplancke et al., 2016).

The complexity of TF-DNA binding variation lies in various kinds of TF motif interaction (Figure 7), suggesting TF-DNA binding itself is a complex molecular trait. It is straightforward

that local regulatory variants in or right next to a motif can have direct impact on TF-DNA interaction. Such impact can transit to cooperative binding of two or more TFs on genome. This is supported by study showing that proximal variant on motif of C/EBP and AP-1 can lead to differential PU.1 binding in macrophages (Heinz et al., 2013). Such co-binding of TFs on DNA is quite common as regulatory regions tend to harbor multiple binding sites for TFs. In fact, most human TFs function with one or more factors through cooperative binding (motifs of two TFs are next to each other) or collaborative binding (motifs are separated that TFs have no contact). This might be due to the fact that DNA is tightly bound to the histone proteins in nucleosome. One TF is not enough to have enough mechanical force to free DNA, so it evolves to interact with more TFs.

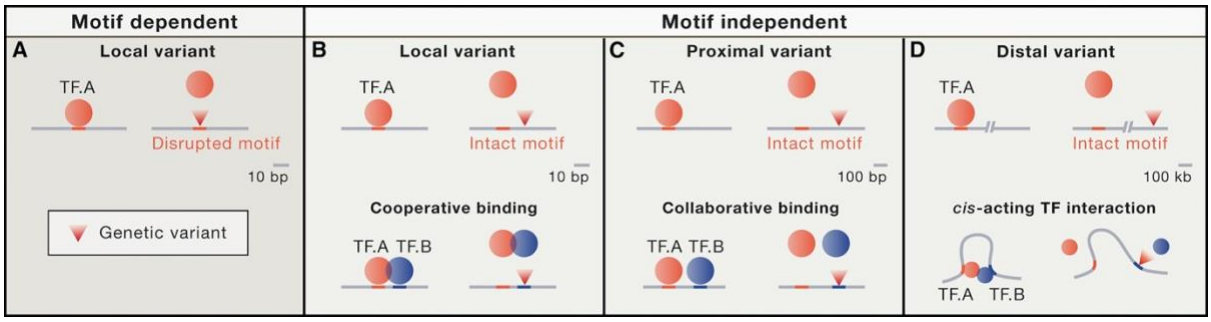


Figure 7: Representation of cis and trans effect of regulatory variants (Deplancke et al., 2016). A: In motif dependent manner, genetic variant is directly located in the motif sequence and disrupt TF binding. In motif independent scenario, (B-D) local, proximal, and distal variants can all disrupt individual TF and TF-TF complex binding.

TF-binding sites are often in enhancer regions where regulatory variants are also found. TF binding affinity on enhancer elements can be altered by genetic variation. A study with two genetically diverse mouse strains aimed to demonstrate the effect of variation on enhancer function (Heinz et al., 2013). PU.1 and C/EBPα are two pioneering factors in macrophages. ChIP-seq of PU.1 and C/EBPα on macrophages found strain-specific binding of both factors are highly correlated with polymorphism frequency, indicating genetic variation in enhancers can potentially affect TF binding.

Another important feature of regulatory variants is that they are capable of affecting gene transcription in a cell-type specific manner. Accessibility variance, the scenario that variants are in the chromatin accessible region of a cell type, is systematically associated with *cis*- and *trans*-regulatory elements (Buenrostro et al., 2015). Cell type-specific accessibility information can be obtained from *in vitro* study, but it is not easy to get such data from tissue. Fortunately,

recent development on single cell open chromatin assay largely facilitates analysis on regulatory control in cell level.

1.2.2 Quantitative trait locus (QTL)

Genetic variants can affect gene expression which ultimately lead to phenotypic difference. The phenotypic difference is specific manifestation of a feature. Based on the how diverse the phenotypic difference is, traits can be qualitative or quantitative. Qualitative traits only have a few distinct phenotypes and are usually controlled by few genes. On the contrary, quantitative traits can have a wide range of phenotypes (Pierce B.A., 2017). Quantitative traits are polygenic, meaning they are influenced by many genetic variants and each of them has a relative small effect on phenotype. It is not feasible to identify regulatory variation or genes for quantitative traits using traditional molecular techniques, as quantitative traits are not simple single-gene case but rather under control of numerous factors from genetics and environment.

Instead of directly searching for regulatory variants or genes, we can look for genomic blocks affecting quantitative traits. Quantitative trait locus (QTL) is a section of DNA that correlates with variation of a quantitative trait. QTL is mapped with genetic markers (SNPs or microsatellites) to identify which genetic markers correlate with trait of interest. The idea behind QTL mapping is if a marker is always correlated with a quantitative phenotype, the QTL with the correlated marker contains genes affecting the specific phenotype. Taking QTL mapping on inbred mice as an example, the analysis needs two types of input: the phenotype measurement on progenies from inbred parental lines, and the genetic markers (Silver LM, 1995). Genetic markers are optimal for genotyping and used to distinguishing between parental animals. There are many mouse panels available for QTL mapping. Usually homologous inbred mice are crossed to get F1 progeny, which later is intercrossed or backcrossed to have F2 generation. The phenotypes of interest and the genetic markers are both measured on F2 cohort. Statistics methods like linear mixed model can be applied to find markers correlating with the observed trait (Silver LM, 1995).

QTL mapping has a wide range of applications on identifying genomic regions correlating with phenotypes in different species, including plants, domestic animals, and humans. Depending on the biological relevance QTL mapping is related to, QTLs can further classified according to the phenotypes studied: expression QTL (eQTL) finding genetic variants explaining difference in gene expression (Nica and Dermitzakis, 2013), methylation QTL (meQTL)

identifying differential methylation (Huan et al., 2019), or histone QTL (hQTL) related to variants causing imbalance in histone modification (Pelikan et al., 2018).

1.3 Mouse strains as a model organism to study human complex traits

Mice and humans are very different in appearance, but they do share many common characteristics. They are both mammals, so a lot of physiological similarities are observed: mice and humans both develop from a single fertilized oocyte which gives rise to a variety of organs: heart, brain, lungs, kidneys, etc. They also have similar circulatory, reproductive, digestive, hormonal and nervous system. At the molecular level, almost all genes in mice have a homolog in human, indicating that mice inherit traits in the same way as humans do (Zhu et al., 2019). Mice and humans are both diploid animals, meaning two copies of chromosomes from parental and maternal parents respectively are present in nuclei. Because of similarities on both physiology and genetics, mice have been used as a model system to study human phenotypes for more than 100 years. With the rich research on mouse population, a lot of specific mouse models have been established, dedicated to studies on a wide range of disease phenotypes or inheritable complex traits.

Except for similarities on physiology and genetics, there are several other advantages to use mice as a model system to study human complex traits. The life span of mice is relatively short, 2 years for laboratory mice and shorter for wild mice. So it is beneficial to study chronic diseases like diabetes and aging on mouse model (Silver LM, 1995). Moreover, mice are small and easy to reproduce. They are economically to maintain in large quantity in animal facility.

1.3.1 Genetic variation of complex traits

Complex traits are prevalent in organisms. Understanding the genetic background of complex traits is intricate but necessary as such information can help to decipher disease phenotypes. Alleles are different copies of genes found in a population. Haplotype is a specific set of linked genetic variants or alleles on a single chromosome or part of a chromosome. Individuals with diverse haplotype composition exhibit a variety of phenotypic difference (Visscher et al., 2017).

One of the main questions in genetics is how genetic variation contributes to phenotypes (Boyle et al., 2017). There was a fierce debate between Mendelians and biometricians in early 1900s. Mendelians were followers of Mendel's theory of heredity and focused on discrete, monogenetic traits, while biometricians were interested in phenotypes with a continuous spread.

Biometricians found out some phenotypes such as human height and weight had a normal distribution in the population and cannot be simply explained by Mendelian genetics. This debate was later resolved by RA Fisher describing the famous infinitesimal model: if there are many genes affecting a trait, sampling in a population with enough replicates can bring a normal distributed phenotype. The more genes are involved, the lower degree of contribution each of them bears. The model has been used to quantify the inheritance of quantitative traits and had a huge impact on plant and animal breeding. It describes inheritance as the sum of genetic factors and non-genetic (environmental) factors.

However, though the genetic variation of complex traits can be quantified by infinitesimal model, how many genes would actually be important for driving complex traits is still unclear. To answer this question, many GWAS studies have been conducted to elucidate associated variants to specific phenotypes. At least two surprising phenomena are observed from the GWAS studies. Many contributed SNPs bear small effect sizes, and even the most significant hits can only account for a modest fraction of the predicted inheritance. This phenomenon is referred as the mystery of missing heritability. One contribution to the missing heritability of complex traits is likely variants which are well below the genome-wide statistical significance. Then how many variants can potentially contribute to the missing heritability? In human height, counting causal SNPs and nearby SNPs in linkage disequilibrium (non-random association between alleles in a haplotype), 62% of all common SNPs can have a non-zero effect (Barton et al., 2017). As the inheritance is proportional to the physical length of chromosome (Shi et al., 2016b), the associated variants tend to be normally distributed across the genome. Though the distribution of causal variants is anticipated to be even, the second surprising observation from GWAS is that phenotype-related variants are more likely to be found in non-coding regions where are enriched for regulatory elements such as promoters and enhancers. This is likely due to the fact that regulatory variants can affect chromatin activity, which further *cis* or *trans*-regulate gene expression via interconnected regulatory networks. This brings us to another question: whether the enrichment of genetic signals is restricted to relevant cell types or broadly. Because only a fraction of GWAS hits in regulatory regions can be explained by eQTL, it is likely that associated SNPs can affect gene expression in a cell-type specific manner.

Intuitively, phenotype-associated variants are expected to be enriched in key genes and regulatory pathways that drive the phenotypes. However, the fact that GWAS hits can only explain a small fraction of phenotypic variance of complex traits suggests an alternative model

taking variants with weak effect into consideration. The cell type-unique disease-associated SNPs and broadly active SNPs do not show difference in contribution to heritability (Barton et al., 2017). This suggests, though genetic variants are enriched in regulatory regions and affect cell type specific gene expression, the inheritance of complex traits does not behave in the similar manner.

Recently, an "omnigenic" model is proposed to explain trait heritability by considering *cis*-effect from core genes and *trans*-regulatory effect from peripheral genes (Barton et al., 2017; Liu et al., 2019). Core genes are genes that directly affect phenotypic outcomes, and peripheral genes have impact on phenotypes by affecting core genes. Variants with modest effect size can affect the expression of core genes, while the ones with minor effect size influence peripheral genes which affect core genes indirectly through inter-connected regulatory networks like gene regulatory networks or protein-protein interaction networks. Because the number of common variants contributing to a complex trait is largely outpaced the number of key variants, the sum of small effects across peripheral genes could be dominant in phenotypic variance.

GWAS studies cannot identify common variants with small effect size as they are not able to survive from the stringent genome-wide significance cutoff considering the sample size available at the moment. Instead, QTL analysis with mouse strains is an alternative approach to identify genomic regions associated with trait of interest.

1.3.2 Mouse strains as a model organism

1.3.2.1 Inbred mouse strains

An inbred strain is a population of animals derived from more than twenty generations of brother-sister mating. All members of the strain result from a single breeding pair of individuals in the twentieth or a later generation. The resulting animals are basically clones of each other in the genetic level with less than 2% genetic variability. On average, at least 98.6% of the loci in each inbred mouse are homozygous (Beck et al., 2000). They have many unique characteristics comparing to wild type mice. Inbred animals are isogenic and homozygous, meaning they are genetically identical. This has direct impact on the resulting phenotypes: animals in some inbred strains are phenotypically uniform, but the ones from different inbred strains are distinct from each other and can be identifiable by their genetic profiles (Silver LM, 1995). These features of inbred strains eliminate genetic variability and thus ensure

reproducibility of research experiments. As inbred strains are almost genetically identical, any outcome observed after treatment, no matter a chemical carcinogen or environmental hazard or drug administration, is a result from the experiment itself. This greatly reduces the number of animals needed to recognize an effect and restricts the confounding influence from unknown factors. If all animals in a study are genetically different, it is difficult to differentiate outcome caused by an experiment from the ones by genetic difference among individuals.

Up to date, there are more than 450 inbred mouse strains available. Many of them have been bred for more than 150 generations. For instance, C57BL/6J reached its 226th generation in 2010 according to the Jackson Laboratory (Casellas, 2011). The first inbred strain, DBA (named after three coat colors: Dilute, D; Brown, B; non-agouti, A), was developed by Little in 1909 (Silver LM, 1995). Subsequently, more inbred strains were generated in the next decades, including C57BL/6, C3H, A/J, and BALB/C. Some phenotypes from the inbred strains are very stable across time. Comparing experiments on ethanol preference and locomotor activity from now and from 30 - 50 years ago showed that strain differences have been highly stable during this time (Wahlsten et al., 2006).

Inbred mouse strains have revolutionized our understanding of gene functions. Scientists have been using inbred mouse strains to create transgenic mice to study genes of interest. Transgenic mice are group of animals with manipulation on genomes by introducing exotic genetic material into genome. In addition, knock-out mouse model with specific gene expression missing can be created by replacing or disrupting target gene sequence with an artificial piece of DNA. By comparing knock-out mouse model with wildtype animals, scientists can learn about the gene function. Observing the characteristics of genetic engineered mice gives us information that can be used to understand how a similar gene many cause or contribute to a disease phenotype in humans (Perlman, 2016).

Embryonic stem cells (ESCs) from early-stage mouse embryos (4 days after fertilization) is the starting material to create transgenic model because it can give rise to all adult cells and subject to long-term storage. There are two methods available to insert artificial DNA into mouse chromosome. The first method takes usage of the homologous recombination (Capecchi, 1989). The arms, upstream and downstream of the foreign DNA, are homologous to the flanking regions of target sequence. Based on homologous recombination, the artificial DNA will swap with the target gene sequence, thus disrupting the gene function. The second method is gene

trapping (Cobellis et al., 2005). Instead of targeting a specific genome sequence, gene trapping endorses random insertion which is designed to target cell's DNA splicing machinery or mRNA degradation. With manipulation on mRNA, some advanced mouse models are created to have a cell-type-specific or even inducible manner. Cre/loxP-mediated recombination system involves a site-specific recombinase Cre and two loxP sites flanking a piece of DNA. Cre efficiently excises DNA sequences located between two loxP sites in the same orientation, leaving one loxP site on the DNA (Walrath et al., 2010). Thus, the system can be used to generate specific genome alterations on genes that are developmentally lethal. For example, glial cells-derived neurotrophic factor (GDNF) promotes survival of dopaminergic neurons and has been a therapeutic target for the Parkinson's disease, but its dosage-dependent effect is not clear due to lack of study system. Mouse models with GDNF hyper-expression was generated by two LoxP sites flanking the 3'UTR region of GDNF, thus disrupting GDNF's mRNA degradation and resulting in over expression of GDNF, leading to the novel discovery of GDNF dosage on brain and kidney development (Li et al., 2019a).

Some inbred mouse strains have phenotypes that are specifically beneficial for producing transgenic animals. The FVB mice have big pronuclei, which makes the microinjection of DNA into the fertilized egg easier. ESCs from 129/Svj have a high successful rate on germline transmission (Beck et al., 2000). Because different mouse models have different characteristics suitable for the research purpose, it is likely that the ESCs with mutation is transferred into a receiver with another genetic background. What if the transgene is in 129/Svj ESCs but inbred mice in C57BL/6J is optimal? To bypass the mixed background problem, the congenic mice can be created by backcrossing the F1 chimera with one of the parental strains for more than 10 generations.

However, in reality, many transgenic mouse models are maintained in two or more backgrounds. This is because some mutations are lethal to animals in single background. For example, *Tgfβ1* germline knock-out is maintained on 129/Svj and CF-1 mixed background to prevent the loss of the autoimmune phenotype. If *Tgfβ1* is aborted on 129/SvJ or C57Bl/6J alone, the animals simply would not be able to survive (Kallapur et al., 1999). Therefore, the phenotypic variability in penetrance is mainly due to knock-out allele being present on a mixed background. Though congenic mice are thought to be identical to the parental strain after more than 10 generations backcrossing, the flanking-gene problem could compromise the estimated "pure" background (Ridgway et al., 2007; Smithies and Maeda, 1995): if the trait-modifying

genes flank the targeted gene, it is hard to know whether the phenotype is due to the targeted candidate gene, the flanked genes from another background, or a complex interaction between the two. Nonetheless, these caveats do not invalidate the usefulness of transgenic mouse models, but rather considerations researchers should be aware of when designing experiments.

Many inbred mouse strains are available and selection of the ideal ones for experiments could be exhausting. However, understanding the phenotypic differences of a trait across mouse strains is paramount but necessary. For example, studies on immune response to a pathogen should not be started with a strain with high innate immune response. Therefore, a lot of studies focused on characterizing phenotypic differences between inbred mouse strains on baseline on topics ranging from neuroscience, metabolic to infectious studies. C57BL/6J and A/J are two commonly used mouse strains to study neurodegenerative diseases. They have been consistently reported to differ in behavioral and physiological processes. Studies have reported that A/J is more prone to behave anxiously and less social comparing to other strains (Moy et al., 2007). A/J is also known to have lower motor activity comparing to C57BL/6J (Thifault et al., 2002).

It is possible to trace back to the genetic culprit underlying the phenotypic difference between inbred strains, as they have limited genetic and environmental variability that is valuable for disentangling gene–phenotype interactions. For instance, the amygdala is an almond-shape cluster of neurons and plays a key role in the processing of emotions. Stress can increase excitatory neurotransmission of amygdala in DBA/2J but not C57BL/6J. Differential gene expression of amygdala between DBA/2J and C57BL/6J under stress found out glutamate receptors NMDA NR1 (*Grin1*) was altered, and the null mutant of *Grin1* was sufficient to produce a DBA/2J-like phenotype, suggesting a causal relationship between *Grin1* and stress (Mozhui et al., 2010).

Another important factor worth noting is that there are many genetic variants between inbred mouse strains, and they can lead to surprising variation on phenotypes. GABA type-A receptors on GABAergic neurons are responsible for fast inhibitory neurotransmission. Receptors with A2 subunit (GABRA2) are abundantly expressed in many brain regions, including frontal cortex, amygdala, dorsal striatum and nucleus accumbens, where are important for emotion behaviors like anxiety and fear. C57BL/6J has a single mutation upstream of *Gabra2* which reduces *Gabra2* expression dramatically comparing to other strains, and correction of the

sequence carried out by CRISPR-*Cas9* in C57BL/6J restored *Gabra2* expression (Mulligan et al., 2019).

Advance in genome sequencing allows researchers to access complete sequence to multiple inbred strains, which are available on the Mouse Genome Project (Keane et al., 2011; Yalcin et al., 2011) and Mouse Genome Database (Bult et al., 2019). With the rich information on inbred mice genomes, studies with murine genomes can reach an unprecedented and finer scale.

1.3.2.2 Using recombinant inbred strains to map QTLs

Scientists have been using inbred mouse strains not only to study the function of individual genes, but also to understand the genetic background of complex traits. Recombinant inbred (RI) strains are a collection of animals that carry random recombination produced from a specific breeding scheme (Silver LM, 1995). The recombination events between homologous chromosomes of a set of RI strains are preserved, and this is the key of RI strains. Initially, the RI strains were developed to map novel loci on mouse genome. With the development of genotyping techniques on genetic markers, the application of RI strains has been expanded from gene mapping to understand the function of a novel locus to QTL mapping searching for genes associated with quantitative traits (Zou et al., 2005).

Breeding of RI strains begins with an outcross between two well-established inbred strains, such as C57BL/6J and DBA/2J, which are considered as progenitor strains. The progenitor strains produce F1 progeny which are hybrid and genetically identical. The F1 animals are bred to each other to produce F2 animals. The F2 animals are no longer identical because of the recombination events and the segregation of progenitor alleles from the heterozygous F1 parents. Each F2 animal has a unique set of loci in which some of them are heterozygous or are identical to the progenitor strains. Some F2 animals are then chosen for brother-sister mating for more than 20 generations to produce new inbred strains. BXD strains, the set of RI strains developed from C57BL/6J and DBA/2J, have 150 inbred strains segregating for over 6 million variants up to date (Ashbrook et al., 2019). Many studies have been using BXD strains to map QTLs of bacterial infection susceptibility (Abdeltawab et al., 2008), auto-immune disease (Alberts et al., 2011), ethanol response , and neuroanatomical traits (Lu et al., 2008).

There are several other RI strains available such as AXB/BXA from C57BL/6J and A/J, CXB from BALB/c x C57BL/6J, or chromosome substitution mice with a full chromosome from

one of the progenitor strains. Using RI strains to map QTLs of quantitative traits lies in the following idea: a nonrandom association between genetic markers and phenotypes of quantitative traits suggest that one or more genes that contribute to the quantitative traits are closely linked to genetic markers (Silver LM, 1995). The genetic markers, such as microsatellites or SNPs, are sets of loci used to differentiate individuals. Because alleles which are physically closely located tend to be segregated together, each genetic marker could be a representative of a genomic sequence with linked alleles. The whole genome can be covered if enough genetic markers are genotyped. Quantitative traits always have genetic sources; therefore, a trait of interest is measured in a group of animals, one would expect a high association between the genetic marker with contributed gene(s) nearby. That brings us to the advantages of RI strains: (1) Because each individual animal in a RI strain is identical, we can expose the animals to various experimental perturbations and interventions, and measure replicates to take use of the mean and standard deviation to have more accurate mapping. (2) Unlike the inbred strains, RI strains are genetically diversified, which results in different levels of penetrance and expressivity of a trait. If inbred strain A has 20% penetrance and inbred strain B has 80% penetrance, the offspring of strain A and strain B would not give information on which allele(s) predisposes the level of penetrance (Silver LM, 1995). (3) RI strains are maintained in a defined environment, which minimizes the influence of environmental factors on phenotypic differences.

RI strains has been widely used in many research fields to search for genes contributing to traits of interest thanks to their unique characteristics to QTL mapping. However, traditional RI mouse panels are not optimal for QTL mapping on complex traits. The level of variability on their genomes is limited because often only two progenitor strains are involved. The breeding history of these mice is complex and sometimes uncertain, leading to unknown confounding factors on their genetics, which prevents inference of causation (The Complex Trait Consortium et al., 2004). Therefore, a new mouse panel was proposed by Complex Trait Consortium specifically dedicated to complex trait analysis, namely the Collaborative Cross (CC) mouse panel (Figure 8).

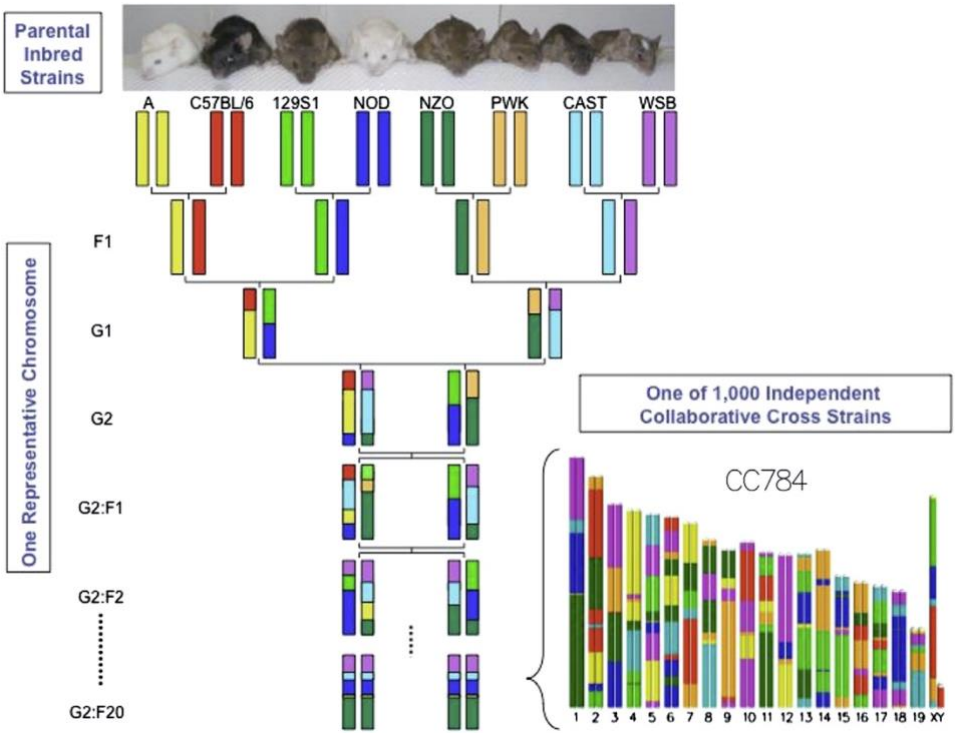


Figure 8: Schematic representation of CC mice breeding with one representative chromosome (Mathes et al., 2011). The breeding starts from eight parental strains. Intercross is carried out with independent breeding funnel to maximize recombination events and break linkage disequilibrium blocks.

CC mice are developed from eight progenitor strains: A/J, C57BL/6J, 129S1/SvImJ, NOD/LtJ, NZO/HiLtJ, CAST/Ei, PWK/PhJ, and WSB/EiJ. CAST/Ei, PWK/PhJ, and WSB/EiJ are wild-derived strains, so they introduce a large level of variation into CC genomes. For example, common inbred strains and C57BL/6J are usually segregated by 4 million SNPs, while PWK/PhJ differs with CAST/Ei by 14 million or with WSB/EiJ by 6 million, respectively. With introduction of three wild-derived strains into the panel, novel QTLs reflecting contrasts with wild-derived strains can be detected (Abu Toamih Atamni and Iraqi, 2018). The breeding scheme of CC panel is well-defined to control randomization to minimize and disperse large LD blocks. Thus, CC mice manage to capture abundant genetic diversity, with the presence of segregating polymorphisms every 100 - 200 bp. This is enough to drive phenotypic diversity in almost any complex trait of interest (Keele et al., 2019). There are 70 CC strains available up to date with readily available genetic markers. It has been estimated that one can achieve a QTL mapping with a resolution of about 1 Mb with 70 lines (Abu Toamih Atamni and Iraqi, 2018). Moreover, candidate genes within the significant QTL interval can be refined by incorporating variation data from CC progenitor strains available from the Sanger Mouse

Genomes Project (Keane et al., 2011). For example, attention can be restricted to variants who behave the same between progenitor strains and the QTL hit.

Deciding how many CC lines and replicates required to conduct an experiment is extremely important for QTL mapping power. With 30 CC strains and each strain with 5 replicates, one can achieve mapping power more than 80% on QTLs with big effect sizes (> 0.444) (Keele et al., 2019). A study with 31 CC lines to perform genome-wide haplotype mapping on trabecular traits yielded six QTLs at 1 % false discovery rate (FDR) and many genes within the significant genomic interval have been shown to associate with bone biology (Levy et al., 2015).

1.4 Dopaminergic circuits and related pathologies

1.4.1 Dopaminergic neurons and related pathways

One of the most intriguing neuroanatomical circuit in mammalian is the nigro-striatal circuit. It is in the central position of neuroscience research because of its relevance to disease. In addition, its inherent complexity makes the nigro-striatal circuit a well suited object to identify novel genetic regulators of dopaminergic neuron's (DAn) function and integrity.

In the nervous system, dopaminergic neurons (DAn) are essential to voluntary movement and behavioral processes like mood, reward, addition and stress. The integrity of DAn and its related pathways is of essential biological relevance and has been in the center of study on neurological diseases.

DAns in the nervous system are mainly located in midbrain (mDAns). There are three dopaminergic circuits in the ventral midbrain based on their locations and related functions. The most well-known dopaminergic system is the nigrostriatal circuit. The cell bodies of DAns are in the zona compacta of substantia nigra (SN), but their fibers innervate the caudate putamen which is also called dorsal striatum. DAns in SN accounting for 3-5 % of total SN neurons are often referred to as A9 group (Chinta and Andersen, 2005). The main function of nigrostriatal circuit is to control the voluntary motor movement. The other two dopaminergic systems overlap with each other in function and location to certain extent. The mesolimbic circuit and the mesocortical circuit both have DAn cell bodies in the ventral tegmental area (VTA) locating close to SN, but their neuronal fibers project to different brain regions. The mesolimbic circuit mainly innervate to nucleus accumbens, amygdala and hippocampus, while

DAns in the mesocortical circuit is projecting to the cortex. Therefore, these two systems are often collectively called mesocorticolimbic circuit with function linking to emotion-based behavior (Wise, 2004). DAns in VTA are often referred to as A10 group.

The importance of DAn lies in the neurotransmitter, dopamine, which is of pivotal biological relevance on a variety of systems. Except for the motor function and the emotion-based behavior as mentioned before in the brain, dopamine participates in the regulation of peripheral systems like the immune system, the kidney and the pancreas (Armando et al., 2011). The synthesis of dopamine begins with an amino acid tyrosine being taken up into brain from liver. Tyrosine hydroxylase (TH) can add a hydroxyl group to tyrosine to produce levodopa (L-DOPA), which is decarboxylated by DOPA decarboxylase to have mature dopamine in the cytoplasm of neurons (Musacchio, 1975).

The way dopamine sends out its signal is by binding to the cell surface receptors in the post-synaptic neurons. The main sensory system on the membrane of post-synaptic neuron is the dopamine receptor. Dopamine receptors are G protein-coupled receptors with 5 types, D1 to D5. They can be further divided into 2 groups: D1-like receptors include D1 and D5, and D2-like receptors include D2-D4 (Chinta and Andersen, 2005). D1-like receptors can induce both excitation and inhibition, depending on the opening of different ion channels like sodium channels or potassium channels. D2-like receptors usually induce inhibition of the target neurons. The most abundant types of dopamine receptors in the nervous system are D1 and D2 (Romanelli et al., 2010). Upon synthesis, dopamine in the cytosol is stored in synaptic vesicles by the vesicular monoamine transporter (VMAT) before ejected to synaptic cleft (Eiden et al., 2004). Once dopamine is released into synaptic cleft, it binds to the dopamine receptors on the membrane of post-synaptic neurons to trigger action potential. The extracellular dopamine can be retaken back by dopamine transporter (DAT) or plasma membrane monoamine transporter (PMAT). When dopamine is transported back to the neuron, it is either broken down by monoamine oxidase (MAO), or repackaged by VMAT for release. Therefore, DAn is a cell group having different anatomical positions and projections with distinct cellular functions.

How do DAns emerge from the embryonic development? During gastrulation, cells migrate from posterior to anterior alongside with the establishment of three germ layers (mesoderm, endoderm and ectoderm). Neuronal tube is established in the rostral end of the embryo, directing the formation of two important signaling centers: the isthmic organizer (IsO) defining

midbrain-hindbrain boundary (MHB), and the floor plate important for the ventral identity of a body. In the mouse model starting from E7.5, progenitors for different brain regions are specified by distinct TFs, with *Otx2* in midbrain and *Gbx2* in hindbrain. The patterning of MHB relies on the expression control of two morphogens from *Otx2* and *Gbx2*: *Wnt1* in the midbrain and *Fgf8* in the hindbrain. Interestingly, it is *Fgf8* required for the anterior-posterior patterning of MHB by establishing a concentration gradient (Basson et al., 2008), with cells in high concentration of FGF8 being committed to hindbrain and the ones receiving low concentration developing to midbrain. On the other hand, the dorsal-ventral patterning in floor plate depends on the morphogen SHH. *Shh* can induce the expression of *Foxa2* at E8 which gives the regional identity of ventral midbrain on floor plate for the specification and proliferation of mDA progenitors, mDA neurogenesis, and differentiation and survival of mDAns.

All mDAns are derived from a common population of neural progenitor cells (NPC) in the ventral midbrain. *Shh*, *Fgf8*, and *Wnt* family members together regulate the expression of several TFs like *Lmx1a/b*, *Otx2*, *Foxa1/2* in mDA NPCs, which later exits cell cycle and migrate to form SN or VTA. Misregulation of DAn specific TFs can cause severe defects. *Pitx3*, strictly expressed in mDAn in embryonic and adult brain, can be induced by *Wnt1* at E11.5. Its expression timing is overlapping with the induction of *Th* (Smidt et al., 1997). Mutation in *Pitx3* in aphakia mouse model leads to DAn selective degeneration (Smidt et al., 2004). In addition, expression of *Foxa1/2* and *Otx2* in mDAn progenitors is required for the induction of *Lmx1a/b*. *Limx1a*, specifically expressing in DAn NPC, is important for DAn specification (Andersson et al., 2006). On the other hand, *Limx1b* is required for neuronal differentiation. It has been shown that though *Limx1b* did not affect the expression of *Th* or *Nurr1*, its ablation failed to induce the expression of *Pitx3*, resulting in the loss of TH-positive neurons during embryonic maturation (Smidt et al., 2000). Moreover, *Nurr1* is another TF generally expressed in mDAn. It is required for the differentiation of post-mitotic mDAn and controlling the expression of genes associated with the synthesis and uptake of dopamine, like *Th*, *Dat*, *Vmat2* (Chinta and Andersen, 2005). Taking together, the development of mDAn is highly complex with sophisticated spatiotemporal control of several TFs.

As there are different groups of mDAns like A9 in SN and A10 in VTA, how does neuronal progenitors commit to anatomically similar cells with distinct innervation? The answer lies in the interplay between *Sox6* and *Otx2* (Panman et al., 2014). The expression pattern of *Sox6* and *Otx2* can distinguish DAn NPC populations, with *Sox6* localized in progenitors migrating to

SN and *Otx2* in cells going to VTA, suggesting the commitment of subpopulation already happens at the stage of NPC.

Intrinsic regulation of TFs expression is undoubtedly important for DAns to maintain its integrity. In addition, many other factors can also exert their effects from different aspects. For example, the extracellular matrix (ECM) is astonishing abundant in the developing and adult nervous system. ECM can provide a microenvironment to modulate cell behavior. Removal of basal lamina, a main component of ECM, leads to detachment of radial glial cells (RGC) fibers and affect its survival (Radakovits et al., 2009). Considering RGCs are the precursors for many brain cells and even DAn precursors show radial glial characteristics (Hebsgaard et al., 2009), intuitively, ECM can potentially modulate neuronal differentiation during development. In fact, a study in Zebra fish model found that type IV collagen controls axogenesis by regulating the integrity of basement membrane (Takeuchi et al., 2015), reinstating the importance of macroenvironment for neurogenesis.

1.4.2 Similarities of dopaminergic circuits between human and mouse

Human and mouse are both mammals, so they share a lot of similarities in terms of genetics and physiology. This similarity is also present in dopaminergic circuits. Human and mouse both have the three dopaminergic circuits. The nigrostriatal circuit, ranging from SN to dorsal striatum, is comparable in both organisms (Figure 9).

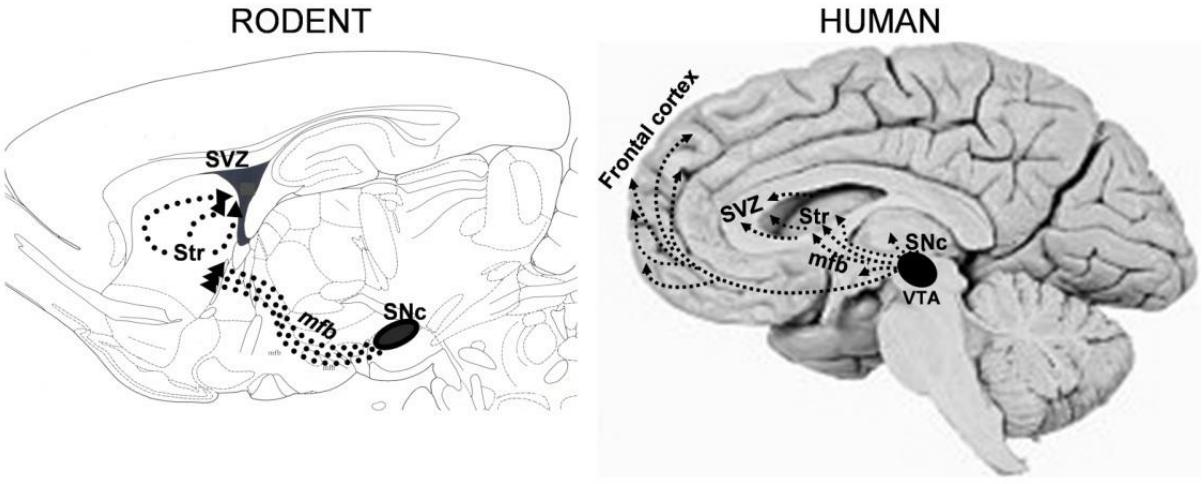


Figure 9: Schematic representation of sagittal view of rodent and human brains (Cova and Armentero, 2011). The nigrostriatal pathway is shown on both, with black dots connecting SN to Str via mfb. Str: striatum; SVZ: subventricular zone; mfb: median forbrain bundle; SNc: substantia nigra pars compacta; VTA: ventral tegmental area.

The number of mDAns are quite diverse in animals probably due to different sizes of brains. It's been estimated that there are 45 000 DAn in rats and 165 000 in macaque monkey (Chinta and Andersen, 2005). The number of mDAn gradually decreases over time in human, with 590 000 at the age of 40 and down to 350 000 at the age of 60 (German and Manaye, 1993). However, such decrease is not observed in rodents probably due to their short life span.

1.4.3 Related pathologies and their genetic backgrounds

DAn is involved in many neuronal pathologies due to their involvement in a wide range of biological functions. One of the most studied neuronal pathology related to DAn is Parkinson's disease (PD). PD is the second-most common neurodegenerative disorder affecting 2-3% of the population more than 65 years of age (Poewe et al., 2017). This disorder was first described clinically by the British physician James Parkinson in 1817 (Arenas et al., 2015). Neurodegeneration of DAns in SN causing striatal dopamine deficiency and intracellular aggregation of α-synuclein are the two hallmarks of PD. Patients with PD experience a series of motor dysfunctions such as bradykinesia (slow movement), rigidity and tremor, and non-motor symptoms like hyposmia (loss of smell), rapid eye movement sleep behavior disorder, depression and constipation (Schapira et al., 2017).

Scientists have dedicated tremendous effort to understand the etiology and progression of PD. PD can be contributed by environmental factors and genetic factors. In terms of genetics, as there are many TFs required to guard the proper neurogenesis of mDAn, misregulation of such factors can cause defects on mDAn and ultimately contribute to the progression of PD. Two mutations in the first exon of *Nurr1* are associated with familial PD (Le et al., 2003). *Pitx3* is essential in DAn differentiation and maintenance. Two polymorphisms located in the first intron of *Pitx3* are shown to associated with the sporadic form of PD (Bergman et al., 2010). Many other mutations can potentially predispose disease risk in human population. Around 5 to 10% PD cases are caused by monogenic mutation which is often found in familial PD. For example, *PRKN* (Parkin), *PINK1* (PTEN induced kinase 1), and *DJ-1* (PARK7) are associated with early onset autosomal recessive PD, whereas autosomal dominant PD cases are often linked to mutations on *SNCA* (α-synuclein), *LRRK2* (leucine rich repeat kinase 2), and *VPS35* (vacuolar protein sorting associated protein 35) (Ohnmacht et al., 2020). However, the majority of PD cases are sporadic with unclear etiology.

PD patients often show a wide range of phenotypes in terms of age of onset (Pagano et al., 2016), disease progression (Parashos et al., 2014), and response to clinical treatment (Poewe et al., 2017). Such diverse phenotypic observation suggests PD is a complex disease with a biological association to the integrity of DAn in nigrostriatal circuit. GWAS studies have identified 90 variants which are estimated to explain 16-36% of the heritable risk of PD (Nalls et al., 2019), suggesting the missing heritability of PD is still up for discovering. Taking all these together, studies to better understand the genetic aspect of the degeneration of DAns is in high demand, especially on disentangling potential contribution from different cell types. For example, one study identified that microglia with *Braf* gene mutation can lead to late-onset neurodegeneration (Mass et al., 2017). In addition, PD patients develop motor symptoms and many other systems that are not related to DAn, suggesting multiple cell types are affected during the disease progress. The recent advance on single cell RNA-seq largely improve the readout resolution comparing to traditional bulk RNA-seq. Single cell RNA-seq on SN and cortical tissue from PD patients uncovered oligodendrocyte gene expression is associating with PD risk (Agarwal et al., 2020), prompting further studies to link more SN cell type expression profiles to specific disease risk.

In addition to PD, there are other neuronal diseases associated with aberrant level of dopamine. Schizophrenia is a severe mental disorder linked to mesocorticolimbic dopamine circuit. It has been characterized that schizophrenia patients have hyperactive dopamine in the mesolimbic area and hypoactive dopamine in the prefrontal cortex (Facchinello et al., 2017), suggesting dopamine level is critically linked to schizophrenia. The responsible factors and the underlying mechanism of schizophrenia have been studied extensively in animal models. In a rat model of schizophrenia, dopamine in nucleus accumbens was found to associate with many receptors like GABA(A) receptors (Rung et al., 2005). Modulation on GABA(A) receptors can decrease cognitive deficits in rats. Recent studies also found Wnt signaling pathway might play a role in schizophrenia. Comparing the expression of Wnt-related genes in blood samples between patients and control, Wnt signaling pathway is aberrantly regulated, suggesting its involvement in disease mechanism (Hoseth et al., 2018). Taken together, the homeostasis of dopamine is essential to maintain at certain level to avoid disease risks .

2. Aim of the study

Inbred mouse strains have substantial phenotypic differences in motor functions and behaviors, the complex traits of which closely link to ventral midbrain. Such phenotypic differences are likely resulted from gene expression changes in the same brain region. Understanding the genetic variation in ventral midbrain can help to apprehend the full genetic makeup about its disease associated phenotypes such as Parkinson's disease or schizophrenia.

Therefore, we aimed at identifying genetic variants behind the gene expression changes in ventral midbrain from the view of *cis* and *trans*-regulatory effects by using genetically diverse inbred mouse strains.

Because TFs can exert *trans*-regulatory effect on thousands of genes, we searched for differential expressed genes coding for TFs by comparing midbrain transcriptomes from C57BL/6J, A/J and DBA/2J. Such TFs could be candidates with *trans* effect to regulate midbrain transcriptome.

Ventral midbrain contains many types of cells. To dissect cell type-specific regulatory variants, we planned to generate single nuclei chromatin accessibility profiles along with tissue-level H3K27ac ChIP-seq assay on ventral midbrains of C57BL/6J and A/J. Cell type-specific *cis* variants could be identified by looking at if they located in the TF binding sites in proximity to differentially expressed genes. Selective *trans* acting candidates could come from comparison with cell type-specific accessible regions between the two strains.

Gene expression changes in ventral midbrain could direct phenotypic difference associated with this region. Nigrostriatal circuit, bridging from SN in ventral midbrain to dorsal striatum, is established by DAns who secret dopamine in dorsal striatum and perish in neurodegenerative diseases. To identify phenotypic difference in nigrostriatal circuit associated with ventral midbrain, we measured if inbred mouse strains carried different levels of dopamine in dorsal striatum, and if so, mouse strain panels could be used to identify QTLs associated with the trait.

Integration of the aforementioned approaches could pinpoint potential cell type specific *cis* or *trans*-regulatory variants on ventral midbrain gene expression, leading the way to link these variation to phenotypic outcome.

3. Materials and methods

Detailed information about the materials and methods are provided together with each manuscript in Results section. Below is a brief summary of experiments and computational analysis I performed.

In "*Pituitary Tumor Transforming Gene 1* orchestrates gene regulatory variation in mouse ventral midbrain during aging" manuscript, I validated the expression of *Pttg1* by RT-PCR, developed and optimized the chromatin immunoprecipitation for ChIP-seq and checked RNA integrity number for RNA-seq. I also performed computational analysis for the manuscript, mainly focusing on RNA-seq, ChIP-seq, PCA, and expression deconvolution.

In "Single nuclei chromatin profiles of midbrain from genetically distinct mouse strains reveal cell identity transcription factors and cell type-specific gene regulatory variation" manuscript, I isolated nuclei from frozen ventral midbrain, and performed snATAC-seq library construction with Dr. Kamil Grzyb. In addition, I performed chromatin immunoprecipitation for ChIP-seq, and optimized library construction for ATAC-seq. I was responsible for computational analysis for the manuscript, mainly focusing on snATAC-seq, ATAC-seq, ChIP-seq, GO enrichment analysis, and motif enrichment analysis.

In "Quantitative trait locus mapping identifies *Col4a6* as a novel regulator of nigrostriatal dopamine level and axonal branching in mice", I performed computational analysis for the manuscript, mainly focusing on QTL mapping with GeneNetwork and permutation to identify genome-wide significance threshold.

4. Results

4.1 Manuscript 1

Pituitary Tumor Transforming Gene 1 orchestrates gene regulatory variation in mouse ventral midbrain during aging

4.1.1 Preface

Inbred mouse strains differ substantially in motor function and behavior which are closely linked to ventral midbrain. Such phenotypic differences are likely affected by gene expression phenotypes in ventral midbrain. To understand the genetic regulators of ventral midbrain transcriptome, we generated midbrain-specific transcriptome profiles from 3 inbred mouse strains, C57BL/6J, A/J and DBA/2J. Pairwise comparisons between gene expression profiles revealed substantial amount of differentially expressed genes, indicating there is phenotypic difference of midbrain transcriptome in these three mice. Searching for potential trans-regulatory effect conducting by TFs found *Pttg1* to be of interest, as it was the only TF gene altered in all comparisons. Interestingly, removing *Pttg1* on C57BL/6J background could cause transcriptome shift towards A/J and DBA/2J along aging. Our work revealed ventral midbrain transcriptome is a complex trait with substantial difference in inbred mouse strains, where *Pttg1* behaves as a *trans*-regulatory variant potentially affecting its expression phenotypes.

The molecular biology experiments and computational analysis were performed by me, except: the animal experiments were performed by Dr. Melanie Thomas; the sequencing was performed by Dr. Rashi Halder.

4.1.2 Manuscript

Pituitary Tumor Transforming Gene 1 orchestrates gene regulatory variation in mouse ventral midbrain during aging

Yujuan Gui[1#], Mélanie H. Thomas[2#], Pierre Garcia[2,3,4], Mona Karout[2], Rashi Halder[2], Alessandro Michelucci[2,5], Heike Kollmus[6], Cuiqi Zhou[7], Shlomo Melmed[7], Klaus Schughart[6,8,9], Rudi Balling[2], Michel Mittelbronn[2,3,4,5,10], Joseph H. Nadeau[11,12], Robert W. Williams[13], Thomas Sauter[1], Manuel Buttini[2*], Lasse Sinkkonen[1*]

[1]Department of Life Sciences and Medicine (DLSM), University of Luxembourg, Belvaux, Luxembourg
[2]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Belvaux, Luxembourg
[3]National Center of Pathology (NCP), Laboratoire National de Santé (LNS), Dudelange, Luxembourg
[4]Luxembourg Centre of Neuropathology (LCNP), Luxembourg.
[5]Department of Oncology (DONC), Luxembourg Institute of Health (LIH), Luxembourg, Luxembourg
[6]Department of Infection Genetics, Helmholtz Centre for Infection Research, Braunschweig, Germany
[7]Cedars Sinai Medical Centre, Los Angeles, California, USA
[8]University of Veterinary Medicine Hannover, Hannover, Germany
[9]Department of Microbiology, Immunology and Biochemistry, University of Tennessee Health Science Center, Memphis, Tennessee, USA.
[10]Pacific Northwest Research Institute, Seattle, Washington, United States
[11]Maine Medical Center Research Institute, Scarborough, Maine USA
[12]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, USA

[*]Corresponding authors: Drs. Lasse Sinkkonen, Manuel Buttini

#These authors contributed equally

Key words: Pttg1 – mouse strains – regulatory variation – midbrain

Pituitary Tumor Transforming Gene 1 orchestrates gene regulatory variation in mouse ventral midbrain during aging

Yujuan Gui[1#], Mélanie H. Thomas[2#], Pierre Garcia[2,3,4], Mona Karout[2], Rashi Halder[2], Alessandro Michelucci[2,5], Heike Kollmus[6], Cuiqi Zhou[7], Shlomo Melmed[7], Klaus Schughart[6,8,9], Rudi Balling[2], Michel Mittelbronn[2,3,4,5], Joseph H. Nadeau[10,11], Robert W. Williams[12], Thomas Sauter[1], Manuel Buttini[2*], Lasse Sinkkonen[1*]

[1]Department of Life Sciences and Medicine (DLSM), University of Luxembourg, Belvaux, Luxembourg

[2]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Belvaux, Luxembourg

[3]National Center of Pathology (NCP), Laboratoire National de Santé (LNS), Dudelange, Luxembourg

[4]Luxembourg Centre of Neuropathology (LCNP), Luxembourg.

[5] Neuro-Immunology Group, Department of Oncology (DONC), Luxembourg Institute of Health (LIH), Luxembourg

[6]Department of Infection Genetics, Helmholtz Centre for Infection Research, Braunschweig, Germany

[7]Cedars Sinai Medical Centre, Los Angeles, California, USA

[8]University of Veterinary Medicine Hannover, Hannover, Germany

[9]Department of Microbiology, Immunology and Biochemistry, University of Tennessee Health Science Center, Memphis, Tennessee, USA.

[10]Pacific Northwest Research Institute, Seattle, Washington, United States

[11]Maine Medical Center Research Institute, Scarborough, Maine USA

[12]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, USA

[*]**Corresponding authors:** Drs. Lasse Sinkkonen (lasse.sinkkonen@uni.lu), Manuel Buttini (manuel.buttini@uni.lu)

#These authors contributed equally

Key words: Pttg1 – mouse strains – regulatory variation – midbrain - aging

Abstract

Dopaminergic neurons in the midbrain are of particular interest due to their role in diseases such as Parkinson's disease and schizophrenia. Genetic variation between individuals can affect the integrity and function of dopaminergic neurons but the DNA variants and molecular cascades modulating dopaminergic neurons and other cells types of ventral midbrain remain poorly defined. Three genetically diverse inbred mouse strains — C57BL/6J, A/J, and DBA/2J — differ significantly in their genomes (~7 million variants), motor and cognitive behavior, and susceptibility to neurotoxins. To further dissect the underlying molecular networks responsible for these variable phenotypes, we generated RNA-seq and ChIP-seq data from ventral midbrains of the 3 mouse strains. We defined 1000–1200 transcripts that are differentially expressed among them. These widespread differences may be due to altered activity or expression of upstream transcription factors. Interestingly, transcription factors were significantly underrepresented among the differentially expressed genes, and only one TF, *Pttg1*, showed significant differences between all three strains. The changes in *Pttg1* expression were accompanied by consistent alterations in histone H3 lysine 4 trimethylation at *Pttg1* transcription start site. The ventral midbrain transcriptome of three-month-old C57BL/6J congenic *Pttg1*$^{-/-}$ mutants was only modestly altered, but shifted towards that of A/J and DBA/2J in nine-month-old mice. Principle component analysis identified the genes underlying the transcriptome shift and deconvolution of these bulk RNA-seq changes using midbrain single cell RNA-seq data suggested that the changes were occurring in several different cell types, including neurons, oligodendrocytes, and astrocytes. Taken together, our results show that *Pttg1* contributes to gene regulatory variation between mouse strains and influences mouse midbrain transcriptome during aging.

Introduction

Two populations of dopaminergic neurons (DAns) in ventral midbrain are of translational interest. One group resides in substantia nigra (SN) controlling motor function, while the other is in ventral tegmental area (VTA) and associated with cognitive function (Vogt Weisenhorn et al., 2016). Many human phenotypes, such as differences in motor learning (Pearson-Fuhrhop et al., 2013) or in disease susceptibility to schizophrenia and Parkinson's disease (PD), are linked to DAns and modulated by genetic variation regulating dopaminergic circuits (Gao and Hong, 2011; Avramopoulos, 2018). Interestingly, recent work has established that most genetic variants associated with human traits and diseases are localized in non-coding genome and significantly enriched in cell type-specific gene regulatory regions (Maurano et al., 2012). Indeed, it has been suggested that most complex traits are explained by cumulative effects of numerous *cis*- and *trans*-regulatory variants that individually contributes to relatively small phenotypic effects (Liu et al., 2019). In particular, peripheral master regulators such as transcription factors (TFs) with tens to hundreds of target genes could be mediating a lot of gene regulatory variation through *trans*-effects while their own expression is altered by local *cis*-variants.

Mouse and human brains share large similarities in dopaminergic circuits and related gene expression, making mouse an ideal model system for neuroscience (Vogt Weisenhorn et al., 2016; Hodge et al., 2019). Three mouse strains, C57BL/6J, A/J, and DBA/2J, are frequently used in biology and show phenotypic differences in their dopaminergic circuits. For example, C57BL/6J has the highest motor activity and sensitivity to addiction (Ingram et al., 1981; Yoshimoto and Komura, 1987; Jong et al., 2010; Eisener-Dorman et al., 2011; Ziółkowska et al., 2012), and its dopamine levels in ventral midbrain are increased compared to the other strains (Cabib et al., 2002). Moreover, the strains respond differently to PD toxins such as methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP), drawing parallels with varied

susceptibility to PD in human population (Hamre et al., 1999). Mouse models are also a fundamental step to study genetic aspects of the brain, with 90% of mouse genes being identical to human genes (Guénet, 2005). Similar to a typical human genome that differs from the reference genome by approximately 5 million variants (Auton et al., 2015), these mouse strains are collectively segregated by around 7 million variants. These characteristics make the mouse an interesting model to study genetic factors and extent of gene regulatory variation in connection to ventral midbrain and dopaminergic circuits.

Here we aimed to elucidate gene regulatory variation underlying the known phenotypic differences within mouse midbrains (Ingram et al., 1981; Yoshimoto and Komura, 1987; Jong et al., 2010; Eisener-Dorman et al., 2011; Ziółkowska et al., 2012) by using a comparative functional genomics approach focusing on transcriptomic and epigenomic analysis of C57BL/6J, A/J, and DBA/2J strains. We identify significant differences between midbrains of the mouse strains with over 1000 genes showing altered expression levels in each comparison. To delineate whether these changes are due to regulatory variation associated with TFs, we looked at which TFs have altered expression. Suprisingly, TFs are significantly under-represented among the altered genes with only *Pttg1 (Pituitary Tumor Transforming Gene 1)* showing significant changes between all three strains. Deletion of *Pttg1* alone is not sufficient to cause major midbrain gene expression changes in young mice, but does lead to substantial transcriptomic shift during aging, resembling the differences distinguishing C57BL/6J from A/J and DBA/2J strains. The changes induced by loss of *Pttg1* are not limited to any specific cell type but instead appear to affect multiple different cell types of the ventral midbrain. Our findings implicate *Pttg1* in the transcriptomic control of the midbrain during aging, and suggest it could contribute to the gene regulatory variation, and possibly also phenotypic variation, between mouse strains.

Materials and Methods

Animals

All experiments were performed in accordance with the European Communities Council Directive 2010/63/EU, approved by appropriate government agencies and respecting the 3 Rs' requirements for Animal Welfare. For the mice bred in the Animal Facility of University of Luxembourg, all experiments in mice were performed according to the national guidelines of the animal welfare law in Luxembourg (*Règlement grand-ducal* adopted on January 11th, 2013). The protocol was reviewed and approved by the Animal Experimentation Ethics Committee (AEEC). For the mice bred in Helmholtz Centre for Infection Research (Braunschweig, Germany), all experiments were performed according to the national guidelines of the animal welfare law in Germany (BGBl. I S. 1206, 1313 and BGBl. I S. 1934). The protocol was reviewed and approved by the 'Niedersächsisches Landesamt für Verbraucherschutz und Lebensmittelsicherheit, Oldenburg, Germany' (Permit Number: 33.9-42502-05-11A193). Mice were housed on a 12 hours-light/dark cycle and provided food and water *ad libitum*. Three mouse strains, C57BL6/6J, A/J and DBA/2J, were used in this study. C57BL/6J and DBA/2J mice were purchased from the provider of Jackson Laboratory in Europe (Charles River). Study cohorts were either directly used after a 2-weeks resting period to allow for acclimatization and control for potential environmental effects, or were bred in house. The A/J breeders were directly purchased from Jackson Laboratory and the study cohorts were bred either at the Helmholtz Centre for Infection Research (Braunschweig, Germany) or in-house at the Animal Facility of University of Luxembourg (Esch-sur-Alzette, Luxembourg). Mice used were within 3-4 generations of breeding cycles. The *Pttg1* knock-out transgenic line was established at Cedars Sinai Medical Center (Wang et al., 2003) and *Pttg1*$^{-/-}$ mice had been backcrossed to C57BL/6J for more than 10 generations. 9- to 13-month old *Pttg1*$^{-/-}$ female mice were bred at the Cedars Sinai Medical Center (Los Angeles, USA. The

Pttg1$^{+/-}$ mice were bred in-house to generate a 3 month-old study cohort (*Pttg1*$^{+/+}$, *Pttg1*$^{+/-}$, and *Pttg1*$^{-/-}$) and to maintain a colony at the local animal facility. Brain samples from the 9 to 13 month-old *Pttg1*$^{-/-}$ mice were collected as described below and 9 month-old strain-matched C57BL/6J were used as a control.

In this study, 12 mice per strain were used at 3 months of age, 4 to 6 mice per strain were used at 9 months of age and 5 to 6 mice per group were used for the *Pttg1* cohorts. For each cohort a comparable number of males and females was used except for aged *Pttg1*$^{-/-}$ cohort where all of the mice were female. At each age group the mice were anesthetized with a ketamine-medetomidine mix (150 and 1 mg/kg, respectively) and intracardially perfused with PBS (phosphate-buffered saline) before extracting the brain. One hemibrain of each mouse was dissected for midbrain. The ventral midbrain was dissected as described in Karunakaran *et al*. (Karunakaran et al., 2007). Briefly, one hemibrain was placed ventral side up on a metal plate over ice, and the region was removed with Dumont forceps caudally of the hypothalamus and thalamus, rostrally of the pons, and ventrally of the Medial Lemniscus, and inferior colliculus. These regions were identified visually on the cut medial surface of the hemibrain. The dissected midbrain was immediately snap-frozen, stored at -80°C, and used for qPCR, RNA-seq, and ChIP-seq analysis as described below.

RT-qPCR

The RNA expression of genes of interest was measured in the midbrains of C57BL/6J, A/J, and DBA/2J. RNA was extracted from the midbrain of each mouse using the RNeasy® Plus Universal Mini Kit (Qiagen, Germany). The reverse transcription was performed using 300 ng of total RNA mixed with 3.8 μM of oligo(dT)20 (Life Technologies) and 0.8 mM of dNTP Mix (Invitrogen). After heating the mixture to 65°C for 5 minutes and an incubation on ice for

1 min, a mix of first-strand buffer, 5 mM of DTT (Invitrogen), RNAse OUT™ (Invitrogen) and 200 units of SuperScript III reverse transcriptase (200 units/μL, Invitrogen) was added to the RNA. The mixture was incubated at 50°C for 60 minutes and then the reaction was inactivated by heating at 70°C for 15 minutes. After adding 80 μL of RNAse free water, the cDNA is stored at -20°C.

RT-qPCR was performed to measure the RNA expression of several genes using the Applied Biosystems 7500 Fast Real-Time PCR System. Each reaction had 5 μL of cDNA, 5 μL of primer mixture (forward and backward primers) (2 μM) and 10 μL of the Absolute Blue qPCR SYBR Green Low ROX Mix (ThermoFisher Scientific, AB4322B). The conditions of the PCR reaction were the following: 95°C for 15 minutes and repeating 40 cycles of 95°C for 15 seconds, 55°C for 15 seconds and 72°C for 30 seconds. The gene expression level was calculated using the $2^{-(\Delta\Delta Ct)}$ method. The $\Delta\Delta Ct$ refers to $\Delta Ct_{(\text{target gene})} - \Delta Ct_{(\text{housekeeping gene})\text{test}} - (\Delta Ct_{(\text{target gene})} - \Delta Ct_{(\text{housekeeping gene})})_{\text{control}}$. *Rpl13a* and *Gapdh* were used as the housekeeping genes and the sequences of the used primers are provided in the Supplementary Table S1.

RNA-seq

The RNA sequencing of 6 C57BL/6J and 6 A/J samples from both 3 month-old and 9-month old mice was done at the sequencing platform of the Genomics Core Facility in EMBL Heidelberg, Germany. The samples were processed by Illumina CBot. The single-end, stranded sequencing was applied by the Illumina NextSeq 500 machine with read length of 80 bp.

The remaining RNA-seq samples were processed at the sequencing platform in the Luxembourg Centre for Systems Biomedicine (LCSB) of the University of Luxembourg. The RNA quality was determined by Agilent 2100 Bioanalyzer and the concentration was quantified by Nanodrop. The TruSeq Stranded mRNA Library Prep kit (Illumina) was used for

library preparation with 1 μg of RNA as input according to the manufacturer's instructions. The libraries were then adjusted to 4 nM. The single-end, stranded sequencing was applied by the Illumina NextSeq 500 machine with read length of 75 bp.

The raw reads quality was assessed by FastQC (v0.11.5) (Andrews, 2010). Using the PALEOMIX pipeline (v1.2.12) (Schubert et al., 2014), AdapterRemoval (v2.1.7) (Lindgreen, 2012) was used to remove adapters, with a minimum length of the remaining reads set to 25 bp. The rRNA reads were removed using SortMeRNA (v2.1) (Kopylova et al., 2012). After removal of adapters and rRNA reads, the quality of the files was re-assessed by FastQC. The mapping was done by STAR (v.2.5.2b) (Dobin et al., 2013). The mouse reference genome, GRCm38.p5 (mm10, patch 5), was downloaded from GENCODE. The suit tool Picard (v2.10.9) (Adams et al., 2000) validated the BAM files. Raw FASTQ files were deposited in ArrayExpress with the accession number E-MTAB-8333.

The reads were counted using *featureCounts* from the R package *Rsubread* (v1.28.1) (Liao et al., 2014). The DEGs were called using R package *DESeq2* (v1.20.0) (Love et al., 2014). RPKM for each gene in each sample was calculated as reads divided by the scale factor and the gene length (kb). The scale factor was calculated as library size divided by 1 million.

Chromatin Immunoprecipitation (ChIP)

ChIP was performed on the dissected mouse midbrain tissue. The fresh tissue was snap frozen for at least a week before crosslinking with formaldehyde (Sigma-Aldrich, F8775-25ML) at a final concentration of 1.5% in PBS (Lonza, BE17-516F) for 10 minutes at room temperature. The formaldehyde was quenched by glycine (Carl Roth, 3908.3) at a final concentration of 125 mM for 5 minutes at room temperature, followed by centrifugation at 1,300 rpm for 5 minutes at 7°C. The fixed tissue was washed twice for 2 minutes with ice-cold PBS plus 1x cOmplete™

mini Proteinase Inhibitor (PI) Cocktail (Roche, 11846145001). The tissue was minced by the Dounce Tissue Grinder (Sigma, D8939-1SET), the lysate of which was centrifuged at 1,300 rpm for 5 minutes at 7°C. The pellet was suspended in ice-cold Lysis Buffer [5 mM 1,4-piperazinediethanesulfonic acid (PIPES) pH8.0 (Carl Roth, 9156.3), 85 mM potassium chloride (KCl) (PanReac AppliChem, A2939), 0.5% 4-Nonylphenyl-polyethylene glycol (NP-40) (Fluka Biochemika, 74385)] with 1xPI, and kept on ice for 30 minutes. The tissue lysate was centrifuged 2,500 rpm for 10 minutes at 7°C. The pellet was suspended with ice-cold Shearing Buffer [50mM Tris Base pH 8.1, 10 mM ethylenediamine tetraacetic acid (EDTA) (Carl Roth, CN06.3), 0.1% sodium odecylsulfate (SDS) (PanReac Applichem, A7249), 0.5% sodium deoxycholate (Fluka Biochemika, 30970)] with 1x PI.

The sonication (Diagenode Bioruptor® Pico Sonication System with minichiller 3000) was used to shear the chromatin with program 30 seconds on, 30 seconds off with 35 cycles at 4°C. After sonication the cell debris was removed by centrifugation at 14,000 rpm for 10 minutes at 7°C. The concentration of the sheared and reverse crosslinked chromatin was measured by Nanodrop 2000c (Thermo Scientific, E597) and shearing was confirmed to produce chromatin fragments of 100 bp to 200 bp.

Each reaction had 10 – 14 μg of chromatin, of which 10% of the aliquot was used as input DNA. The chromatin sample was diluted 1:10 with Modified RIPA buffer [140 mM NaCl (Carl Roth, 3957.2), 10 mM Tris pH 7.5, 1 mM EDTA, 0.5 mM ethylene glycol-bis-N,N,N',N'-tetraacetic acid (EGTA) (Carl Roth, 3054.3), 1% Triton X-100, 0.01% SDS, 0.1% sodium deoxycholate] with 1x PI, followed by addition of 5 μL of H3K4me3 (histone H3 lysine 4 trimethylation) antibody (Millipore, 17-614) and incubation overnight at 4°C with rotation. After incubation, the immunocomplexes were collected with 25 μL of PureProteome™ Protein A Magnetic (PAM) Beads (Millipore, LSKMAGA10) for 2 hours at 4°C with rotation.

The beads were washed twice with 800 μL of Wash Buffer 1 (WB1) [20 mM Tris pH 8.1, 50 mM NaCl, 2mM EDTA, 1% Triton X-100, 0.1% SDS], once with 800 μL of Wash Buffer 2 (WB2) [10 mM Tris pH 8.1, 150 mM NaCl, 1 mM EDTA, 1% NP-40, 1% sodium deoxycholate, 250 mM lithium chloride (LiCl) (Carl Roth, 3739.1)], and twice with 800 μL of Tris-EDTA (TE) Buffer [10 mM Tris PH 8.1, 1 mM EDTA pH 8.0]. The beads were re-suspended in 100 μL of ChIP Elution Buffer [0.1 M sodium bicarbonate ($NaHCO_3$) (Sigma-Aldrich, S5761) and 1% SDS]. After the elution, the chromatin and the 10% input were both reverse-crosslinked at 65°C for 3 hours with 10 μg of RNase A (ThermoFisher, EN0531) and 20 μg of thermoresistant proteinase K (ThermoFisher, EO0491), followed by purification with MiniElute Reaction Cleanup Kit (Qiagen, 28206) according to the manufacture's instruction.

The concentration of the chromatin was measured by Qubit® dsDNA HS Assay Kit (ThermoFisher, Q32851) and Qubit 1.0 fluorometer (Invitrogen, Q32857) according to the manufacturer's instructions and rest of the chromatin was used for high-throughput sequencing.

ChIP-seq

The sequencing of the chromatin samples was done at the sequencing platform in the LCSB of the University of Luxembourg. The single-end, unstranded sequencing was applied by the Illumina NextSeq 500 machine with read length of 75 bp. The raw reads quality was assessed by FastQC (v0.11.5) (Andrews, 2010). The PALEOMIX pipeline (v1.2.12) (Schubert et al., 2014) was used to generate BAM files from the FASTQ files, including steps of adapter removal, mapping and duplicate marking. The mapping was done by BWA (v.0.7.16a) (Li and Durbin, 2009), with backtrack algorithm using the quality offset of Phred score to 33. Duplicate reads were marked but not discarded. The mouse reference genome, GRCm38.p5 (mm10, patch 5), was downloaded from GENCODE (https://www.gencodegenes.org/). The suit tool

Picard (v2.10.9) (Adams et al., 2000) was used to validate the BAM files. Raw FASTQ files were deposited in ArrayExpress with the accession number E-MTAB-8333.

The H3K4me3 ChIP-seq peaks were called by Model-based analysis of ChIP-seq (MACS, v2.1.1) (Zhang et al., 2008). The signal normalization in pairwise comparison was done by THOR (v0.10.2) (Allhoff et al., 2016), with TMM normalization and adjusted p-value cut-off 0.01.

Principle Component Analysis (PCA)

The raw counts were normalized to library size and $\log_2$-transformed using *DESeq2* (v1.20.0). The PCs were calculated with 500 genes which have the most varied expression across samples.

Bulk RNA-seq data deconvolution using single cell RNA-seq data

The bulk RNA-seq deconvolution was done with CIBERSORTx (https://cibersortx.stanford.edu/) (Newman et al., 2019). The signature matrix on SN of single cell RNA-seq was constructed with DropViz (http://dropviz.org/) with default parameters. The expression of 332 genes correlating with PC1 (pvalue < 0.05) from Figure 5A in 5 cell types (neuron, dopaminergic neuron, oligodendrocyte, astrocyte, endothelial cells) were inferred with default parameters.

Statistical Analysis

The p-value of DEGs called from pair-wise comparisons in RNA-seq was adjusted for multiple testing with the Benjamini-Hochberg procedure with cutoff below 0.05. The significance of

peak calling was analysed with MACS2 and the significance in ChIP-seq signal normalization was defined with multiple test correction (Benjamini/Hochberg) for p-values with cutoff below 0.05.

Results

Midbrain transcriptomes are significantly different between common mouse strains

To investigate genetic background driving gene expression differences in ventral midbrain, we performed transcriptomic and epigenomic analyses on isolated ventral midbrains containing SN and VTA from three genetically diverse mouse strains, C57BL/6J, A/J, and DBA/2J (Figure 1A). For transcriptomic profiling, midbrains from 36 individual 3-month old mice were analysed by RNA-seq, corresponding to 12 mice (6 males and 6 females) from each strain. For epigenomic analysis, the enrichment of histone H3 lysine 4 trimethylation (H3K4me3), an established marker of open transcription start sites (TSS) (Santos-Rosa et al., 2002; Barski et al., 2007; Guenther et al., 2007), was analysed by ChIP-seq from dissected ventral midbrain of 6 individual 3-month old mice (2 males from each strain).

A principle component analysis (PCA) of RNA-seq data could clearly separate the samples according to strain of origin (Figure 1B), suggesting significant differences exist at the transcriptomic level between ventral midbrains. Males and females showed only minor differences as indicated in Figure 1B. In total 25 DEGs between males and females could be detected (data not shown), indicating that main driver of gene expression differences was the genetic background of each strain. Indeed, a pair-wise comparison of the individual strains to each other revealed a significant (FDR<0.05) change in expression with a $\log_2$-fold change ($\log_2$FC) higher than 1 for more than 1000 genes (Figure 1C and Supplementary Table S2). Changes could be observed for both high expressed genes as well as lower abundance

transcripts with comparable numbers of up- and down-regulated transcripts in each comparison.

Gene expression levels correlated well with the enrichment of H3K4me3 at the corresponding TSS (Figure 1D), indicating that the ChIP-seq could serve as an indicator of midbrain transcriptional activity.

Pttg1 **is the only transcription factor with altered midbrain expression between all three mouse strains**

Gene expression changes linked to complex traits have been suggested to be explained by both small cumulative effects of *cis*-regulatory variants across numerous genes, and by *cis*-regulatory variants at "peripheral master regulators" such as TFs that can in *trans* influence a number of co-regulated genes directly linked to the trait (Liu et al., 2019). To better understand whether the observed gene expression changes in the mouse midbrain transcriptomes could be due to variants affecting upstream TFs, we further examined TFs included among the differentially expressed genes (DEGs). We first overlapped the DEGs from the pair-wise comparisons of the strains and identified 53 genes to be differentially expressed between all three strains (Figure 2A). Moreover, we identified a total of 1292 genes to be shared between at least two of the pair-wise comparisons of the strains (Supplementary Table S3). These genes are clustered in Figure 2B according to their gene expression profiles across the three strains with comparable numbers of genes showing particularly abundant or low expression levels in one or another strain. Next we used a manually curated list of 950 TFs (Heinäniemi et al., 2013), 841 of which could be detected in the midbrain, and identified 5 genes coding for TFs (*Pttg1*, *Npas1*, *Hes5*, *Scand1*, and *Zfp658*) to be differentially expressed in at least one of the mouse strains (Figure 2B). Interestingly, the number of differentially expressed TFs was much

smaller than the 36 TFs that could be expected among the DEGs just by chance (hypergeometric test, $p = 2.03*10^{-11}$). This lack of variation among TFs indicates a tight control of TF gene expression, which may need to be kept within a narrow range to allow for proper cellular function in the midbrain. Among the five TFs, only *Pttg1* showed a significant difference between all three strains and higher than 2-fold change in each comparison (Figure 2B and Supplementary Table S3). In detail, C57BL/6J midbrain samples showed an average *Pttg1* expression of 14 RPKM (Reads Per Kilobase Million) and DBA/2J samples an average expression of 2.5 RPKM while in A/J midbrains *Pttg1* expression was never higher than 1 RPKM (Figure 2B and Supplementary Table S3). Differential midbrain expression levels of *Pttg1* between different mouse strains was confirmed by RT-qPCR, with A/J showing a particularly low expression level (Supplementary Figure S1A and S1B). Moreover, the H3K4me3 signal from ChIP-seq analysis was clearly reduced at the TSS of *Pttg1* gene in A/J compared to C57BL/6J, while no differences were observed at the TSS of neighbouring genes *Slu7* and *C1qtnf2* (Figure 2C). In addition, the signal in A/J appeared comparable or lower than in DBA/2J, despite the overall enrichment in DBA/2J samples being weaker than in the other two strains. These results suggest that reduced expression of *Pttg1* in the midbrain of A/J is due to decreased transcription at the locus.

Therefore *Pttg1* appears to be a prime candidate for explaining midbrain transcriptomic differences between the mouse strains.

Loss of *Pttg1* leads to changes in the midbrain transcriptome during aging

Given that *Pttg1* encodes the only TF with significantly altered expression levels between all three mouse strains, we investigated the role of midbrain PTTG1 in more detail. To test whether altered expression of *Pttg1* alone can indeed influence the midbrain transcriptome, we

investigated C57BL/6J congenic *Pttg1*$^{-/-}$ mice. In contrast to differences in A/J or DBA/2J, deletion of *Pttg1* in the 3-month old mice leads to minor transcriptomic changes with 3 additional genes differentially expressed compared to the *Pttg1*$^{+/+}$ littermates (Figure 3A). Two of these (*Thg1l* and *Ublcp1*) were previously found to strongly correlate with *Pttg1* expression across different mouse strains, and to be genetically associated with neocortex volume (Gaglani et al., 2009), while the third gene (*Gm12663*) is an anti-sense transcript of *Ublcp1*. Moreover, the expression of these genes is dependent of *Pttg1* expression level when corroborating the analysis with 3-month-old *Pttg1*$^{+/-}$, with *Ublcp1* showing positive, and *Thg1l* and *Gm12663* showing negative correlation with *Pttg1* levels (Figure 3B).

Although the observed midbrain transcriptome changes in *Pttg1*$^{-/-}$ mice were minimal, we were curious to elucidate whether these early changes would lead to additional transcriptomic differences at an older age. We therefore performed further RNA-seq analysis with isolated midbrains of a cohort of six aged mice from each C57BL/6J, A/J, and DBA/2J strains (all 9 months old), and C57BL/6J congenic *Pttg1*$^{-/-}$ mice (9-13 months old) (Figure 4A). Interestingly, comparing samples from 9-month-old wild-type C57BL/6J or A/J to those from younger 3-month-old mice of the respective strains identified almost no genes with strong expression changes of 5-fold or more ($\log_2$FC>2.25) (Figure 4B and Supplementary Table S4). Similarly, comparison of 9-month-old DBA/2J midbrain transcriptome to the younger counterparts revealed only 57 strongly altered genes. Conversely, the midbrain samples of *Pttg1*$^{-/-}$ mice showed over 300 genes that were strongly differentially expressed in the aged mice compared to 3-month-old mice, as shown in the Volcano plot in Figure 4B.

Pttg1 **contributes to gene regulatory variation in the midbrain cell types during aging**

To obtain a broader overview of the extent and the direction of transcriptomic changes across the studied mouse strains and ages, we performed PCA analysis for all 78 midbrain transcriptome profiles. Interestingly, the PCA revealed that over half of the variance between the studied mice was explained by the first and the second principle components (PCs) that separated the mice according to genetic background (Figure 5A). C57BL/6J mice were separated from A/J and DBA/2J along PC1 while A/J and DBA/2J were separated from each other along PC2. Consistent with the small number of DEGs in 3-month-old *Pttg1*$^{-/-}$ mice, they clustered closely together with their heterozygous *Pttg1*$^{+/-}$ littermates, and with wild-type C57BL/6J mice. Also, aged mice clustered largely together with their genetically identical counterparts for each A/J, DBA/2J, and C57BL/6J. However, for the aged cohort of *Pttg1*$^{-/-}$ mice, the transcriptome profiles had significantly shifted along PC1 from C57BL/6J towards A/J and DBA/2J (Figure 5A).

Analysis of genes that contributed most to the differences along PC1 revealed genes that were altered not only in A/J and DBA/2J strains but also in aged *Pttg1*$^{-/-}$ mice when compared to C57BL/6J (Figure 5B). Furthermore, gene changes in *Pttg1*$^{-/-}$, A/J, and DBA/2J mice showed the same directionality, with the *Pttg1*$^{-/-}$ mice clustering together with A/J and DBA/2J rather than C57BL/6J when analysed with hierarchical clustering.

Finally, to see whether the loss of *Pttg1* was specifically affecting only some of the cell types in the midbrain, we performed deconvolution analysis of the DEGs contributing to PC1 using mouse midbrain single cell RNA-seq data (Saunders et al., 2018). Based on the inference, the DEGs included genes preferentially expressed in many different cell types, including different types of neurons such as *Th*$^{+}$ DAns, oligodendrocytes, astrocytes, and endothelial cells (Figure 5C). Additionally, over a third of the genes could not be inferred, suggesting they are expressed broadly across multiple different cell types.

Taken together, the results indicate that loss of *Pttg1* leads to only limited transcriptomic changes in the midbrain of young mice, but can lead to substantial differences during aging, with parts of C57BL/6J transcriptome shifted towards A/J and DBA/2J in aging mice. Thus, our data indicate that PTTG1 contributes to transcriptome differences in multiple cell types of the midbrain between the three genetically diverse mouse strains.

Discussion

We investigated gene expression differences between mouse strains to understand how genetic variation can influence midbrain and its important cell types such as DAns that control motor function and behaviour. Our transcriptomic analysis revealed extensive changes in midbrain gene expression between the 3 mouse strains and highlighted *Pttg1* as an important regulator of midbrain transcriptome during aging.

The observed midbrain transcriptomic differences are comparable to the transcriptome level changes observed between mouse strains in other tissues such as lung (Wilk et al., 2015), striatum (Bottomly et al., 2011), and retina (Wang et al., 2019), or in specific cell types such as macrophages (Link et al., 2018) and other immune cells (Mostafavi et al., 2014). Interestingly, despite the obvious variation in the gene expression between the mouse strains, genes encoding for TFs are under-represented among the DEGs in the mouse midbrain. This finding is consistent with similar results from plants (Lin et al., 2017), where TF coding genes were also found to be under-represented among the genes showing differential expression between genetically diverse strains. Such findings are likely to be due to natural selection against phenotypes arising from major variation in TF expression levels that could be detrimental for the normal functioning of an organism.

In total 5 TFs were found to vary in their expression between the mouse strains. While most changes were weaker than those observed for *Pttg1*, they could nevertheless contribute to the observed gene regulatory variation. Indeed, both *Npas1* and *Hes1* have been previously connected to regulation of neuronal genes and neurogenesis (Ishibashi et al., 1994; Ishibashi et al., 1995; Michaelson et al., 2017) while possible roles for *Scand1* and *Zfp658* in CNS have not yet been described. To investigate the possible contribution of these factors on gene regulatory variation between the mouse strains, we searched the existing data sets for those identifying targets of these TFs in central nervous system. Interestingly, target genes of *Npas1* in hippocampus have been previously described (Michaelson et al., 2017). However, comparison of genes altered upon *Npas1* deletion (FDR<0.05) revealed only 5 genes to be shared with DEGs between A/J, DBA/2J, and C57BL/6J (data not shown). Therefore, *Npas1* is not likely to mediate *trans*-acting variation between the mouse strains.

The only TF showing significant changes in the ventral midbrain between all three mouse strains is *Pttg1*, also known as securin. *Pttg1* was originally described as an oncogene in pituitary tumors (Pei and Melmed, 1997) and found to regulate sister chromatid adhesion in M-phase of cell cycle (Zou et al., 1999). However, the protein has multiple functions and also a role as a DNA-binding transcriptional activator has been described (reviewed in Vlotides *et al.*, 2007 (Vlotides et al., 2007)).

Little is known about the neurological functions of PTTG1. Keeley *et al*., 2014 (Keeley et al., 2014) identified a link between PTTG1 and the central nervous system, showing increased *Pttg1* expression in retinas of C57BL/6J mice compared to A/J due to a *cis* deletion variant at the *Pttg1* promoter, consistent with our findings in the midbrain. Interestingly, differential *Pttg1* expression correlated with mosaic regularity variation across 25 recombinant inbred strains derived from the two parental C57BL/6J and A/J mouse strains, involving PTTG1 in the patterning of a type of retinal neurons, the amacrine cells. Moreover, *Pttg1* expression in

neocortex correlates with neocortical volume and the locus is genetically associated with this trait (Gaglani et al., 2009). Therefore, *Pttg1* appears to play a role in development or maintenance of central nervous system, and our results indicate its possible involvement in genetic control of midbrain cell types. Indeed, previous work using microarrays found >1400 genes to be misregulated across the whole brain of *Pttg1*$^{-/-}$ mice at the age of 3-5 months (Lum et al., 2006). While we identified far fewer DEGs specifically in the midbrain of the 3 month-old *Pttg1*$^{-/-}$ mice at our significance cut-off (FDR<0.05) using RNA-seq, this increased significantly during aging. The differences in the results for younger mice could be due to the use of a specific brain region (rather than the whole brain) and applied methodology with related statistical analysis, but could also be contributed to by unknown differences in the environmental conditions between the studies. Importantly, the overall transcriptomic profile of aged *Pttg1*$^{-/-}$ mice shifted towards the profiles of A/J and DBA/2J (Figure 5), indicating that *Pttg1* might indeed exhibit genetic control over gene expression in the midbrain, although additional genetic factors are likely altered to contribute to these changes already in young mice.

It has been previously reported that *Pttg1* is involved in many biological functions such as regulation of sister chromatid separation, DNA repair or senescence processes (Zou et al., 1999; Bernal et al., 2008; Hsu et al., 2010). Interestingly, a deconvolution analysis of the gene expression changes using single cell RNA-seq analysis indicated that the loss of *Pttg1* influenced gene expression across multiple cell types. However, these changes become observable only during aging. Unlike human brain, mouse brain volume has been shown to increase still during adulthood between 6 and 14 months of age (Maheswaran et al., 2009). Given the abovementioned role of *Pttg1* in regulation of neocortex volume and its effect of gene expression in multiple cell types during aging, it is tempting to speculate that *Pttg1* would

contribute also to control of midbrain volume. Interestingly, a greater brain volume has been reported for C57BL/6J than A/J (Williams, 2000; Wahlsten et al., 2006).

Conclusions

Rather than being entirely explained by the TF expression levels due to *cis*-variation at the *Pttg1* locus, complex traits like midbrain gene expression could be due to cumulative *cis*- and *trans*-regulatory variants across TF binding sites controlling the DEGs. In the future, mapping QTLs associated with the DAn's traits across mouse strains, together with the transcriptomic and epigenomic data generated as part of this work, will enable the identification of further regulatory variants and their impact on midbrain expression phenotype and function of the nigrostriatal circuitry. While linking complex traits such as behaviour and motor function to specific gene expression changes will require further studies, our work highlights the role of *Pttg1* as regulator of mouse midbrain gene expression phenotype and paves way for further identification of additional genetic regulators.

Abbreviations

DAns : dopaminergic neurons ; DEGs : differentially expressed genes ; EDTA : ethylenediamine tetraacetic acid ; H3K4me3 : histone H3 lysine 4 trimethylation ; LCSB : Luxembourg Centre for Systems Biomedicine ; log_2FC : log_2-fold change ; NP-40 : 4-Nonylphenyl-polyethylene glycol ; PC : principle component ; PCA : principle component analysis ; PD : Parkinson's disease ; PI : Proteinase Inhibitor ; *Pttg1 :* Pituitary Tumor Transforming Gene 1 ; RPKM : Reads Per Kilobase Million ; SDS : sodium odecylsulfate ;

SN:substantia nigra ; TFs : transcription factors ; TSS : transcription start sites ; VTA:ventral tegmental area.

Competing interests

The authors declare that they have no competing interests.

Data availability statement

The datasets generated during the current study are available in ArrayExpress with the accession number E-MTAB-8333.

Ethics statement

The protocol for mice bred at University of Luxembourg was reviewed and approved by the Animal Experimentation Ethics Committee (AEEC). For the mice bred in Helmholtz Centre for Infection Research (Braunschweig, Germany), the protocol was reviewed and approved by the 'Niedersächsisches Landesamt für Verbraucherschutz und Lebensmittelsicherheit, Oldenburg, Germany' (Permit Number: 33.9-42502-05-11A193).

Funding

LS and MB would like to thank the Luxembourg National Research Fund (FNR) for the support (FNR CORE C15/BM/10406131 grant). MM would like to thank the Luxembourg National Research Fund (FNR) for the support (FNR PEARL P16/BM/11192868 grant). KS would like

to thank the support by intra-mural grants from the Helmholtz-Association (Program Infection and Immunity).

Author contributions

MB and LS conceived the project with input from AM, KS, RB, JHN, and RWW. YG, MT, MK, MB, and LS designed the experiments and analysis. YG performed all RNA-seq and ChIP-seq experiments and bioinformatic analysis. MT, YG, MK, PG, and MB prepared mouse tissues. YG, MK, and MT performed PCR experiments. HK and KS supported mouse breeding and bioinformatic analysis together with RWW. CZ and SM provided *Pttg1* transgenic mice. RH prepared the libraries and performed the sequencing. YG, MT, MK, MM, TS, MB, and LS analysed the results. YG, MT, MB, and LS wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Drs Aurélien Ginolhac and Anthoula Gaigneaux for their support with bioinformatic analysis and EMBL Gene Core at Heidelberg for support with high-throughput sequencing, and Dr Djalil Coowar (Animal Facility of University of Luxembourg) for help with breeding of experimental mice. KS would like to thank the animal caretakers at the Central Animal Facilities of the HZI for maintaining the mice. The computational analysis presented in this paper were carried out using the HPC facilities of the University of Luxembourg.

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., et al. (2000). The genome sequence of Drosophila melanogaster. *Science* 287, 2185–2195.

Allhoff, M., Seré, K., F Pires, J., Zenke, M., and G Costa, I. (2016). Differential peak calling of ChIP-seq signals with replicates with THOR. *Nucleic Acids Res* 44, e153.

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data.*

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74.

Avramopoulos, D. (2018). Recent Advances in the Genetics of Schizophrenia. *Mol Neuropsychiatry* 4, 35–51.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., et al. (2007). High-resolution profiling of histone methylations in the human genome. *Cell* 129, 823–837.

Bernal, J. A., Roche, M., Méndez-Vidal, C., Espina, A., Tortolero, M., and Pintor-Toro, J. A. (2008). Proliferative potential after DNA damage and non-homologous end joining are affected by loss of securin. *Cell Death Differ* 15, 202–212.

Bottomly, D., Walter, N. A. R., Hunter, J. E., Darakjian, P., Kawane, S., Buck, K. J., et al. (2011). Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLoS ONE* 6, e17820.

Cabib, S., Puglisi-Allegra, S., and Ventura, R. (2002). The contribution of comparative studies in inbred strains of mice to the understanding of the hyperactive phenotype. *Behav Brain Res* 130, 103–109.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21.

Eisener-Dorman, A. F., Grabowski-Boase, L., and Tarantino, L. M. (2011). Cocaine locomotor activation, sensitization and place preference in six inbred strains of mice. *Behav Brain Funct* 7, 29.

Gaglani, S. M., Lu, L., Williams, R. W., and Rosen, G. D. (2009). The genetic control of neocortex volume and covariation with neocortical gene expression in mice. *BMC Neurosci* 10, 44.

Gao, H.-M., and Hong, J.-S. (2011). Gene-environment interactions: key to unraveling the mystery of Parkinson's disease. *Prog Neurobiol* 94, 1–19.

Guénet, J. L. (2005). The mouse genome. *Genome Res* 15, 1729–1740.

Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., and Young, R. A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* 130, 77–88.

Hamre, K., Tharp, R., Poon, K., Xiong, X., and Smeyne, R. J. (1999). Differential strain susceptibility following 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) administration acts in an autosomal dominant fashion: quantitative analysis in seven strains of Mus musculus. *Brain Res* 828, 91–103.

Heinäniemi, M., Nykter, M., Kramer, R., Wienecke-Baldacchino, A., Sinkkonen, L., Zhou, J. X., et al. (2013). Gene-pair expression signatures reveal lineage control. *Nat Methods* 10, 577–583.

Hodge, R. D., Bakken, T. E., Miller, J. A., Smith, K. A., Barkan, E. R., Graybuck, L. T., et al. (2019). Conserved cell types with divergent features in human versus mouse cortex. *Nature* 573, 61–68.

Hsu, Y.-H., Liao, L.-J., Yu, C.-H., Chiang, C.-P., Jhan, J.-R., Chang, L.-C., et al. (2010). Overexpression of the pituitary tumor transforming gene induces p53-dependent senescence through activating DNA damage response pathway in normal human fibroblasts. *J Biol Chem* 285, 22630–22638.

Ingram, D. K., London, E. D., Reynolds, M. A., Waller, S. B., and Goodrick, C. L. (1981). Differential effects of age on motor performance in two mouse strains. *Neurobiol Aging* 2, 221–227.

Ishibashi, M., Ang, S. L., Shiota, K., Nakanishi, S., Kageyama, R., and Guillemot, F. (1995). Targeted disruption of mammalian hairy and Enhancer of split homolog-1 (HES-1) leads to up-regulation of neural helix-loop-helix factors, premature neurogenesis, and severe neural tube defects. *Genes Dev* 9, 3136–3148.

Ishibashi, M., Moriyoshi, K., Sasai, Y., Shiota, K., Nakanishi, S., and Kageyama, R. (1994). Persistent expression of helix-loop-helix factor HES-1 prevents mammalian neural differentiation in the central nervous system. *EMBO J* 13, 1799–1805.

Jong, S. de, Fuller, T. F., Janson, E., Strengman, E., Horvath, S., Kas, M. J. H., et al. (2010). Gene expression profiling in C57BL/6J and A/J mouse inbred strains reveals gene networks specific for brain regions independent of genetic background. *BMC Genomics* 11, 20.

Karunakaran, S., Diwakar, L., Saeed, U., Agarwal, V., Ramakrishnan, S., Iyengar, S., et al. (2007). Activation of apoptosis signal regulating kinase 1 (ASK1) and translocation of death-

associated protein, Daxx, in substantia nigra pars compacta in a mouse model of Parkinson's disease: protection by alpha-lipoic acid. *FASEB J* 21, 2226–2236.

Keeley, P. W., Zhou, C., Lu, L., Williams, R. W., Melmed, S., and Reese, B. E. (2014). Pituitary tumor-transforming gene 1 regulates the patterning of retinal mosaics. *Proc Natl Acad Sci U S A* 111, 9295–9300.

Kopylova, E., Noé, L., and Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 28, 3211–3217.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760.

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930.

Lin, H.-Y., Liu, Q., Li, X., Yang, J., Liu, S., Huang, Y., et al. (2017). Substantial contribution of genetic variation in the expression of transcription factors to phenotypic variation revealed by eRD-GWAS. *Genome Biol* 18, 192.

Lindgreen, S. (2012). AdapterRemoval: Easy Cleaning of Next Generation Sequencing Reads. *BMC Res Notes* 5, 337.

Link, V. M., Duttke, S. H., Chun, H. B., Holtman, I. R., Westin, E., Hoeksema, M. A., et al. (2018). Analysis of Genetically Diverse Macrophages Reveals Local and Domain-wide Mechanisms that Control Transcription Factor Binding and Function. *Cell* 173, 1796-1809.e17.

Liu, X., Li, Y. I., and Pritchard, J. K. (2019). Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* 177, 1022-1034.e6.

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

Lum, P. Y., Chen, Y., Zhu, J., Lamb, J., Melmed, S., Wang, S., et al. (2006). Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. *J Neurochem* 97 Suppl 1, 50–62.

Maheswaran, S., Barjat, H., Rueckert, D., Bate, S. T., Howlett, D. R., Tilling, L., et al. (2009). Longitudinal regional brain volume changes quantified in normal aging and Alzheimer's APP x PS1 mice using MRI. *Brain Res* 1270, 19–32.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190–1195.

Michaelson, J. J., Shin, M.-K., Koh, J.-Y., Brueggeman, L., Zhang, A., Katzman, A., et al. (2017). Neuronal PAS Domain Proteins 1 and 3 Are Master Regulators of Neuropsychiatric Risk Genes. *Biol Psychiatry* 82, 213–223.

Mostafavi, S., Ortiz-Lopez, A., Bogue, M. A., Hattori, K., Pop, C., Koller, D., et al. (2014). Variation and genetic control of gene expression in primary immunocytes across inbred mouse strains. *J Immunol* 193, 4485–4496.

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* 37, 773–782.

Pearson-Fuhrhop, K. M., Minton, B., Acevedo, D., Shahbaba, B., and Cramer, S. C. (2013). Genetic variation in the human brain dopamine system influences motor learning and its modulation by L-Dopa. *PLoS ONE* 8, e61197.

Pei, L., and Melmed, S. (1997). Isolation and characterization of a pituitary tumor-transforming gene (PTTG). *Mol Endocrinol* 11, 433–441.

Santos-Rosa, H., Schneider, R., Bannister, A. J., Sherriff, J., Bernstein, B. E., Emre, N. C. T., et al. (2002). Active genes are tri-methylated at K4 of histone H3. *Nature* 419, 407–411.

Saunders, A., Macosko, E. Z., Wysoker, A., Goldman, M., Krienen, F. M., Rivera, H. de, et al. (2018). Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* 174, 1015-1030.e16.

Schubert, M., Ermini, L., Sarkissian, C. D., Jónsson, H., Ginolhac, A., Schaefer, R., et al. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat Protoc* 9, 1056–1082.

Vlotides, G., Eigler, T., and Melmed, S. (2007). Pituitary tumor-transforming gene: physiology and implications for tumorigenesis. *Endocr Rev* 28, 165–186.

Vogt Weisenhorn, D. M., Giesert, F., and Wurst, W. (2016). Diversity matters - heterogeneity of dopaminergic neurons in the ventral mesencephalon and its relation to Parkinson's Disease. *J Neurochem* 139 Suppl 1, 8–26.

Wahlsten, D., Bachmanov, A., Finn, D. A., and Crabbe, J. C. (2006). Stability of inbred mouse strain differences in behavior and brain size between laboratories and across decades. *Proc Natl Acad Sci U S A* 103, 16364–16369.

Wang, J., Geisert, E. E., and Struebing, F. L. (2019). RNA sequencing profiling of the retina in C57BL/6J and DBA/2J mice: Enhancing the retinal microarray data sets from GeneNetwork. *Mol Vis* 25, 345–358.

Wang, Z., Moro, E., Kovacs, K., Yu, R., and Melmed, S. (2003). Pituitary tumor transforming gene-null male mice exhibit impaired pancreatic beta cell proliferation and diabetes. *Proc Natl Acad Sci U S A* 100, 3428–3432.

Wilk, E., Pandey, A. K., Leist, S. R., Hatesuer, B., Preusse, M., Pommerenke, C., et al. (2015). RNAseq expression analysis of resistant and susceptible mice after influenza A virus infection identifies novel genes associated with virus replication and important for host resistance to infection. *BMC Genomics* 16, 655.

Williams, R. W. (2000). Mapping genes that modulate mouse brain development: a quantitative genetic approach. *Results Probl Cell Differ* 30, 21–49.

Yoshimoto, K., and Komura, S. (1987). Reexamination of the relationship between alcohol preference and brain monoamines in inbred strains of mice including senescence-accelerated mice. *Pharmacol Biochem Behav* 27, 317–322.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9, R137.

Ziółkowska, B., Korostyński, M., Piechota, M., Kubik, J., and Przewłocki, R. (2012). Effects of morphine on immediate-early gene expression in the striatum of C57BL/6J and DBA/2J mice. *Pharmacol Rep* 64, 1091–1104.

Zou, H., McGarry, T. J., Bernal, T., and Kirschner, M. W. (1999). Identification of a vertebrate sister-chromatid separation inhibitor involved in transformation and tumorigenesis. *Science* 285, 418–422. 10.1126/science.285.5426.418

Figure Legends

Figure 1. Functional genomics profiling of isolated midbrains of 3-month-old C57BL/6J, A/J and DBA/2J mice.

A. Schematic representation of the experimental set-up. The ventral midbrains of C57BL/6J, A/J and DBA/2J, dissected using anatomical landmarks directly after mouse euthanasia, were used for RNA-seq and ChIP-seq.
B. Principle component analysis showing transcriptome level differences in the midbrains of the three strains. The individual mice are indicated with black (C57BL/6J), grey (A/J), or brown (DBA/2J). Circles indicate females and triangles males. No bias was observed between females and males.
C. Pairwise comparisons showing DEGs in the midbrains of the three strains. MA plots from left to right: A/J vs. C57BL/6J, DBA/2J vs. C57BL/6J, and DBA/2J vs. A/J. The analysis was done by DEseq2 using ashr shrinkage. The x-axis represents the mean of normalized counts for all replicates and the y-axis represents the $\log_2$-fold change. Each dot represents one gene. Genes with FDR<0.05 and $\log_2$-fold change ($\log_2$FC)>1 are indicated in red and referred to as DEGs.
D. H3K4me3 ChIP-seq signal with corresponding gene expression levels as measured by RNA-seq. The intensity of H3K4me3 ChIP-seq signals are plotted in a window of 3 kb upstream and downstream of the TSS and within-sample normalization was applied. The genes are ranked based on gene expression levels (RPKM) from highest to lowest.

Figure 2. *Pttg1* is the only TF differentially expressed between the midbrains of 3-month-old C57BL/6J, A/J and DBA/2J mice.

A. Venn diagram comparing DEGs of each pair-wise comparison of the mouse strains from Figure 1C. The majority of DEGs are shared by at least two comparisons.
B. Heatmap of the expression of the 1292 DEGs shared between at least two of the comparisons. The read counts were vst-transformed and used for clustering. Expression levels of the five DEGs coding for TFs are shown as dot plots. *=FDR<0.05.
C. The altered expression of *Pttg1* is accompanied by changes in H3K4me3 ChIP-seq signal at the *Pttg1* TSS. The H3K4me3 ChIP-seq was performed on two male replicates. The pair-wise comparisons (C57BL/6J vs. A/J and DBA2J vs. A/J) were performed by THOR with within-sample and between-sample normalizations. Normalized ChIP-seq signals are depicted in black (for C57BL/6J and DBA/2J) or in grey (for A/J). Red rectangle indicates *Pttg1* TSS.

Figure 3. Loss of *Pttg1* leads to minimal changes in the midbrain transcriptome in 3-month old mice.

A. RNA-seq analysis identifies four DEGs in comparison of the congenic C57BL/6J *Pttg1*$^{-/-}$ vs. *Pttg1*$^{+/+}$ mice at the age of 3 months. MA plot was generated as in Figure 1C with labelling of the four DEGs (*Pttg1*, *Thg1l*, *Ublcp1*, *Gm12663*) that are indicated as red dots.
B. The expression of *Ublcp1* is positively correlated with *Pttg1* across genotypes, while *Gm12663* and *Thg1l* show negative correlation with *Pttg1*. The dot plots indicate the expression levels of the DEGs as RPKM in isolated midbrains of *Pttg1*$^{+/+}$, *Pttg1*$^{+/-}$, and *Pttg1*$^{-/-}$ mice.

Figure 4. Loss of *Pttg1* leads to significant transcriptomic changes in the midbrain during aging.

A. Schematic representation of the experimental set-up. The ventral midbrains of 9-month-old C57BL/6J, A/J and DBA/2J mice, and 9-13-month-old congenic C57BL/6J *Pttg1*$^{-/-}$ mice were used for RNA-seq as in Figures 1 and 3.

B. Comparison of midbrain transcriptome of 9-month-old mice to the midbrain transcriptome of the corresponding strains at 3 months of age. *Pttg1* deletion leads to more significant and higher gene expression changes than observed for wild-type mouse strains during aging. Volcano plots from left to right: A/J, C57BL/6J, DBA/2J, and congenic C57BL/6J Pttg1$^{-/-}$. The x-axis represents the mean $\log_2$-fold change for all replicates and the y-axis represents the significance of change as $-\log_{10}$ (p-value). Each dot represents one gene. Genes with FDR<0.05 and $\log_2$-fold change ($\log_2$FC)>2.25 are indicated in red and referred to as DEGs.

Figure 5. C57BL/6J *Pttg1*$^{-/-}$ midbrain transcriptome shift towards A/J during aging.

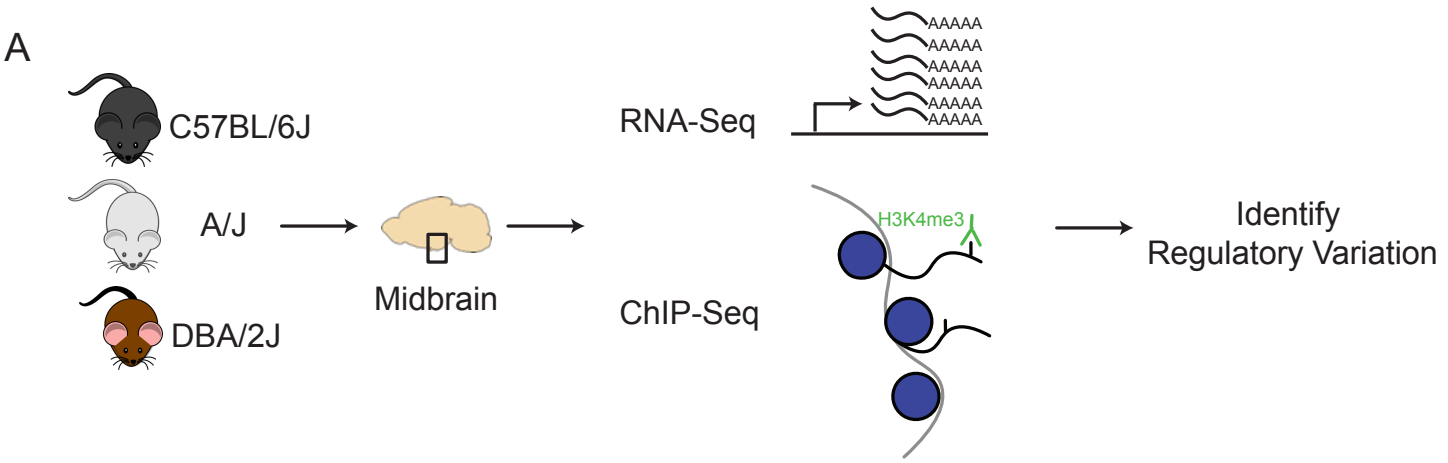
A. Principle component analysis showing transcriptome level differences in the midbrains of C57BL/6J, DNA/2J, and A/J mice at the age of 3 and 9 months, congenic C57BL/6J *Pttg1*$^{+/+}$, *Pttg1*$^{+/-}$, and *Pttg1*$^{-/-}$ at 3 months, and *Pttg1*$^{-/-}$ at 9 to 13 months. Individual mice are indicated with black circles (C57BL/6J 3m), blue circles (C57BL/6J 9m), grey circles (A/J 3m), white circles (A/J 9m), brown circles (DBA/2J 3m), light brown circles (DBA/2J 9m), green circles

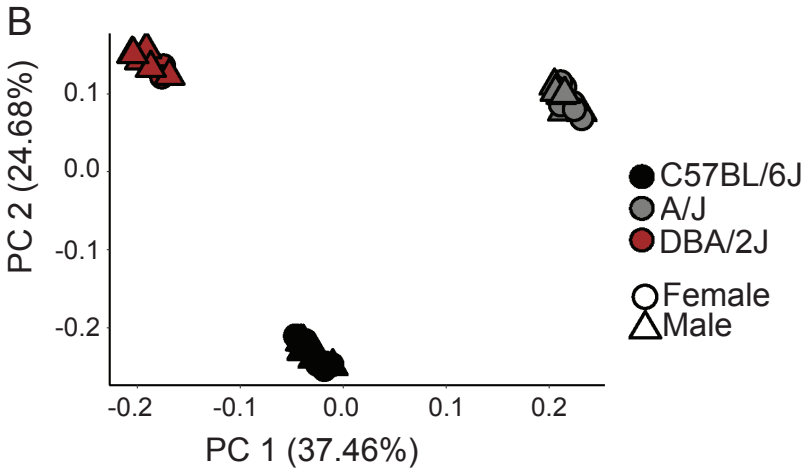
(*Pttg1*$^{+/+}$ 3m), dark green rectangles (*Pttg1*$^{+/-}$ 3m), dark green triangles (*Pttg1*$^{-/-}$ 3m), or light green triangles (*Pttg1*$^{-/-}$ 9-13m). No gender bias was observed.

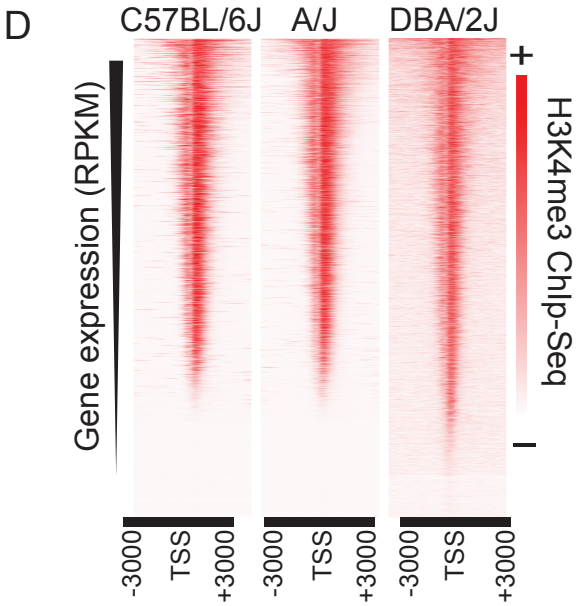
B. Heat map of the differential genes associated with principle component 1 in panel A. Gene expression profile of *Pttg1*$^{-/-}$ mice clusters with A/J and DBA/2J mice instead of C57BL/6J.

C. Deconvolution of differential gene expression using single cell RNA-seq was done for 331 genes contributing the most to PC1 in panel A and detected as expressed in 5 major cell types of the scRNA-seq data. The number of genes and their proportion of all analysed genes are shown for each cell type.

Figure 1


A
C57BL/6J
A/J
DBA/2J
Midbrain
RNA-Seq
ChIP-Seq
H3K4me3
Identify
Regulatory Variation


B
PC 2 (24.68%)
PC 1 (37.46%)
C57BL/6J
A/J
DBA/2J
Female
Male


D
C57BL/6J
A/J
DBA/2J
Gene expression (RPKM)
H3K4me3 ChIp-Seq
-3000
TSS
+3000

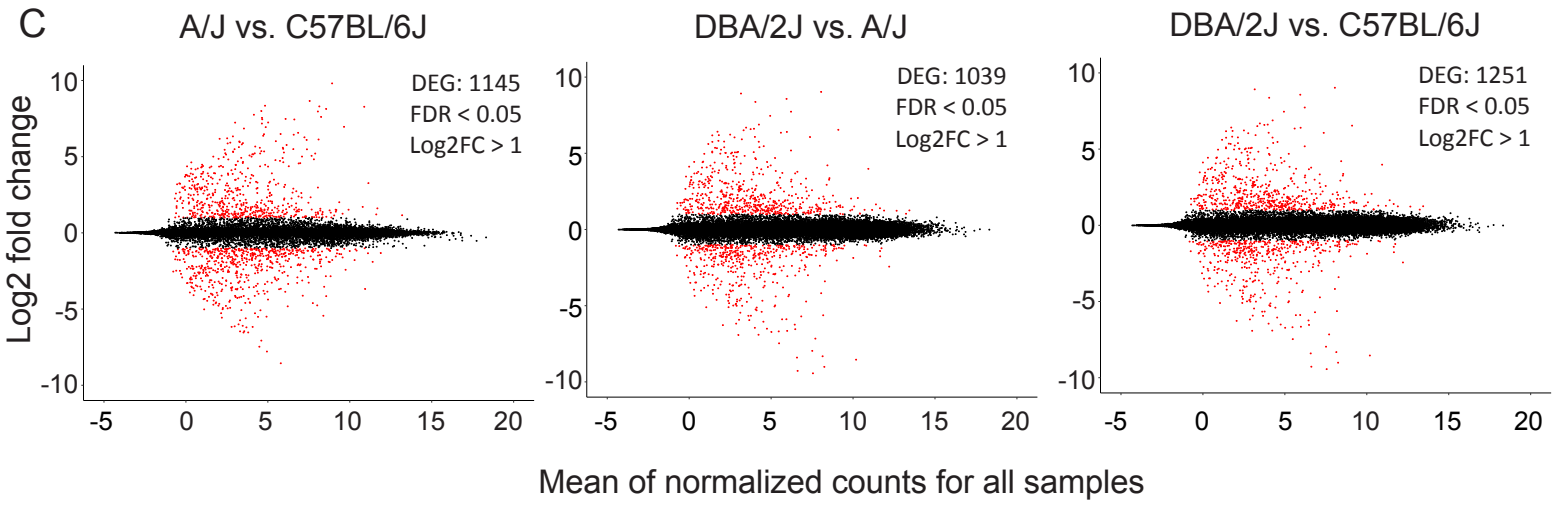

C
A/J vs. C57BL/6J
DBA/2J vs. A/J
DBA/2J vs. C57BL/6J
DEG: 1145
FDR < 0.05
Log2FC > 1
DEG: 1039
FDR < 0.05
Log2FC > 1
DEG: 1251
FDR < 0.05
Log2FC > 1
Log2 fold change
Mean of normalized counts for all samples

Figure 2

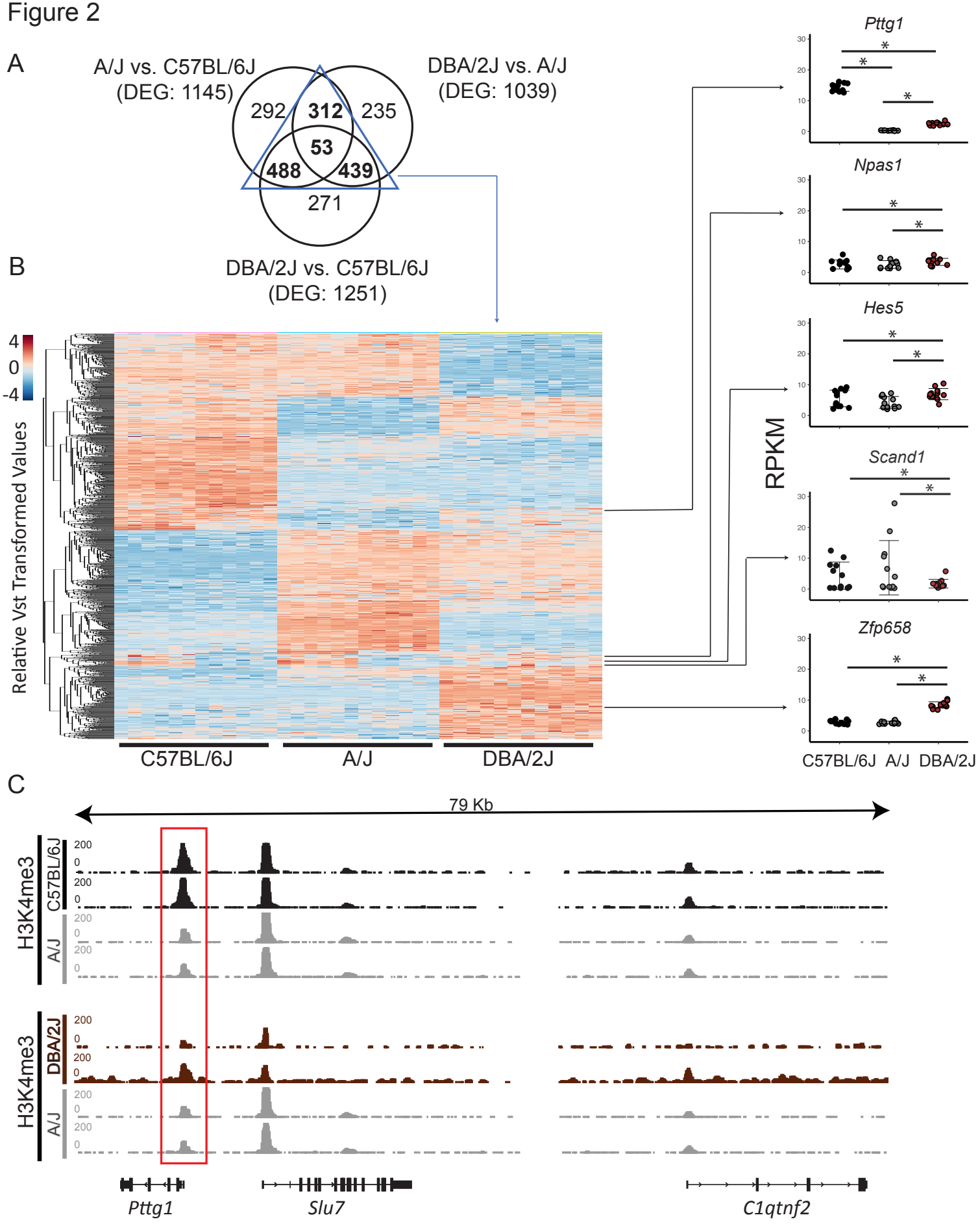

A
A/J vs. C57BL/6J
(DEG: 1145)
DBA/2J vs. A/J
(DEG: 1039)
292
312
235
53
488
439
271
B
DBA/2J vs. C57BL/6J
(DEG: 1251)
4
0
-4
Relative Vst Transformed Values
C57BL/6J
A/J
DBA/2J
RPKM
Pttg1
Npas1
Hes5
Scand1
Zfp658
C57BL/6J A/J DBA/2J
C
79 Kb
H3K4me3
C57BL/6J
A/J
DBA/2J
Pttg1
Slu7
C1qtnf2

Figure 3

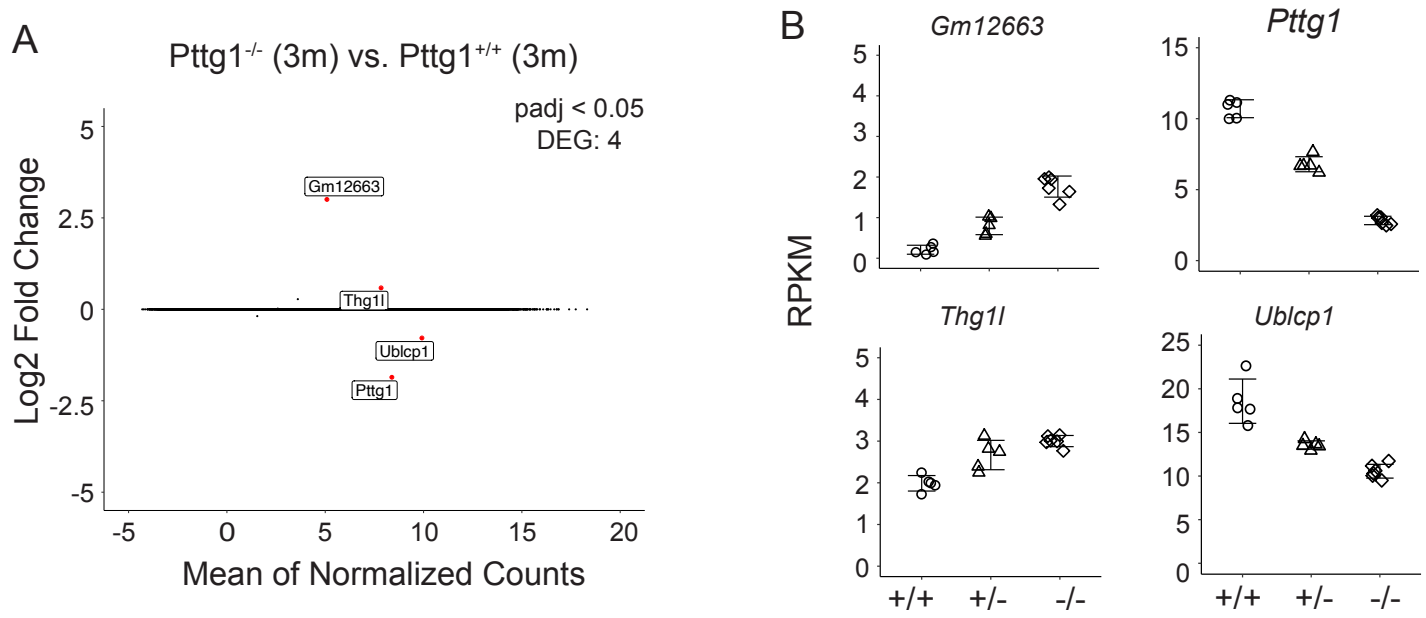
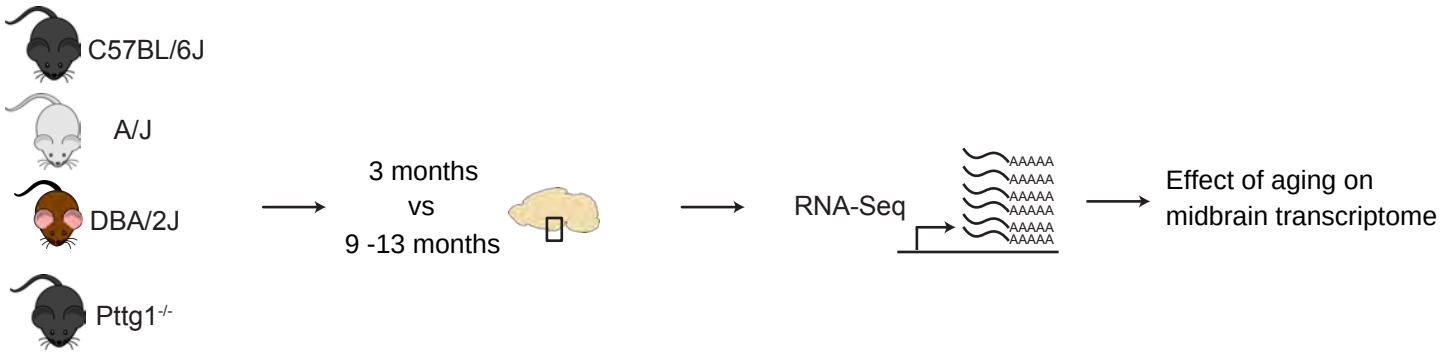

A
Pttg1-/- (3m) vs. Pttg1+/+ (3m)
padj < 0.05
DEG: 4
Gm12663
Thg1l
Ublcp1
Pttg1
Log2 Fold Change
Mean of Normalized Counts
B
Gm12663
Pttg1
Thg1l
Ublcp1
RPKM
+/+
+/-
-/-

Figure 4

A


C57BL/6J
A/J
DBA/2J
Pttg1-/-
3 months
vs
9 -13 months
RNA-Seq
AAAAA
Effect of aging on
midbrain transcriptome

B

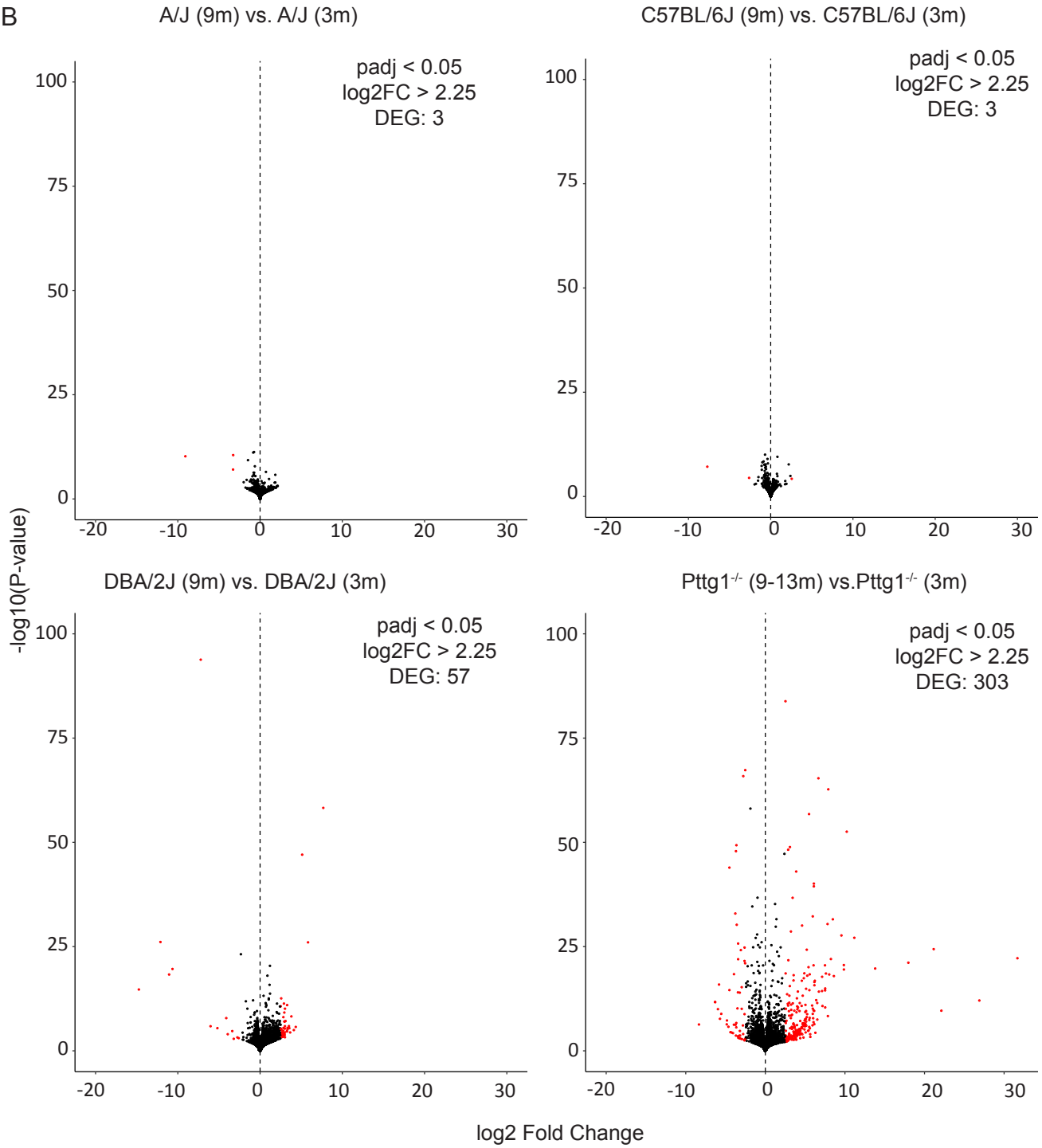
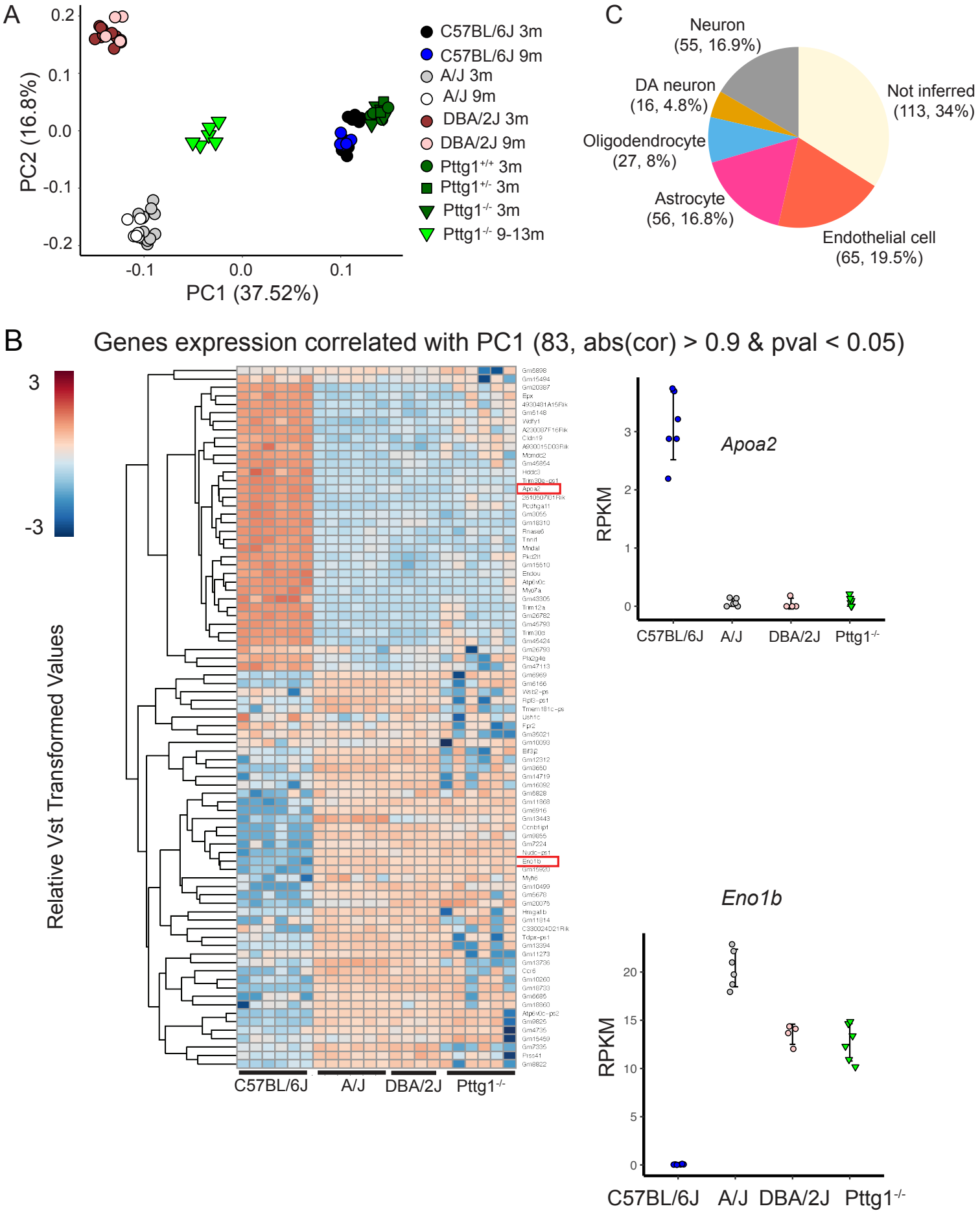

A/J (9m) vs. A/J (3m)
padj < 0.05
log2FC > 2.25
DEG: 3
C57BL/6J (9m) vs. C57BL/6J (3m)
padj < 0.05
log2FC > 2.25
DEG: 3
DBA/2J (9m) vs. DBA/2J (3m)
padj < 0.05
log2FC > 2.25
DEG: 57
Pttg1-/- (9-13m) vs.Pttg1-/- (3m)
padj < 0.05
log2FC > 2.25
DEG: 303
-log10(P-value)
log2 Fold Change

Figure 5


A
PC2 (16.8%)
PC1 (37.52%)
C57BL/6J 3m
C57BL/6J 9m
A/J 3m
A/J 9m
DBA/2J 3m
DBA/2J 9m
Pttg1+/+ 3m
Pttg1+/- 3m
Pttg1-/- 3m
Pttg1-/- 9-13m
C
Neuron (55, 16.9%)
DA neuron (16, 4.8%)
Not inferred (113, 34%)
Oligodendrocyte (27, 8%)
Astrocyte (56, 16.8%)
Endothelial cell (65, 19.5%)
B
Genes expression correlated with PC1 (83, abs(cor) > 0.9 & pval < 0.05)
Relative Vst Transformed Values
C57BL/6J
A/J
DBA/2J
Pttg1-/-
Apoa2
Eno1b
RPKM

Pituitary Tumor Transforming Gene 1 orchestrates gene regulatory variation in mouse ventral midbrain during aging

Yujuan Gui#, Mélanie H. Thomas#, Pierre Garcia, Mona Karout, Rashi Halder, Alessandro Michelucci, Heike Kollmus, Cuiqi Zhou, Shlomo Melmed, Klaus Schughart, Rudi Balling, Michel Mittelbronn, Joseph H. Nadeau, Robert W. Williams, Thomas Sauter, Manuel Buttini*, Lasse Sinkkonen*

Supplementary Information

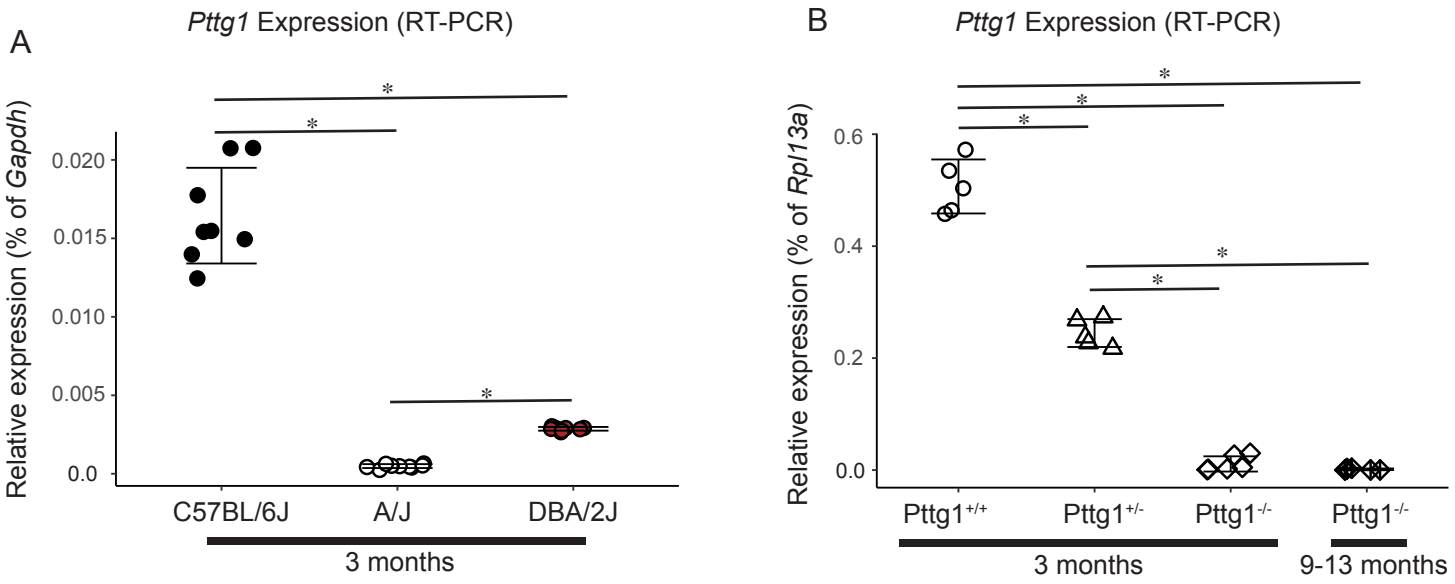
Supplementary Figures

Supplementary Figure S1. RT-PCR measurements of *Pttg1* expression in isolated midbrains are consistent with the RNA-seq results.

A. *Pttg1* expression measured by RT-PCR is consistent with the RNA-seq data across the three strains. Expression levels are presented relative to *Gapdh*. Two-sided Student's t-test was used for statistical testing. *=$p<0.05$.

B. *Pttg1* expression measured by RT-PCR is consistent with the RNA-seq data across the 3 months old $Pttg1^{+/+}$, $Pttg1^{+/-}$, $Pttg1^{-/-}$, and 9-13 months old $Pttg1^{-/-}$ mice. Expression levels are presented relative to *Rpl13a*. Two-sided Student's t-test was used for statistical testing. *=$p<0.05$.

Supplementary Figure S1



Supplementary Tables

Supplementary Table S1. **Primer sequences used in the study.**

Gene	**Forward primer (5‘ – 3‘)**	**Reverse primer (5‘ – 3‘)**
Pttg1	TCAAGGTCGGCTGTTTTGGT	AGTTGCCGAAAAGCCTATGAAG
Rpl13a	TGGTCCCTGCTGCTCTCA	CCCCAGGTAAGCAAACTTTCT
Gapdh	TGCGACTTCAACAGCAACTC	CTGCTCAGTGTCCTTGCTG

Supplementary Table S2. DEGs (FDR < 0.05, log2FC > 1) from three comparisons in Figure 1C. The base mean, log2FC, and FDR are reported for each gene in each comparison: A/J vs. C57BL/6J: 1145 genes; DBA/2J vs. A/J: 1039 genes; DBA/2J vs. C57BL6/J: 1251 genes.

Supplementary Table S3. DEGs (FDR < 0.05, log2FC > 1) shared by at least two comparisons in Figure 2B. The base mean, log2FC, and FDR are reported for each gene in each comparison: A/J vs. C57BL/6J: 853 genes; DBA/2J vs. A/J: 804 genes; DBA/2J vs. C57BL6/J: 980 genes.

Supplementary Table S4. DEGs (FDR < 0.05, log2FC > 2.5) from 3 months old vs. 9 months old mice. The DEGs from C57BL/6J A/J, DBA/2J and *Pttg1*$^{-/-}$ mice are shown on individual worksheets. The base mean, log2FC, and FDR are reported for each gene in each comparison.

4.2 Manuscript 2

Single nuclei chromatin profiles of midbrain from genetically distinct mouse strains reveal cell identity transcription factors and cell type-specific gene regulatory variation

4.2.1 Preface

Ventral midbrain contains many different types of cells. Though tissue level RNA-seq identified *Pttg1* as the potential *trans*-regulatory variant in ventral midbrain transcriptome, *cis*-acting variation along with cell type-specific information are still missing. To identify gene regulatory elements in different cell types, we generated single nuclei accessibility profiles on the ventral midbrains of C57BL/6J and A/J, together with tissue level H3K27ac ChIP-seq to find enhancers. Integration with public available single cell RNA-seq revealed motifs controlling cell identity. Comparing accessibility regions in the two mouse strains across cell types, we found putative regulatory variants enriched in differentially expressed genes, indicating chromatin regions with differed accessibility harbor regulatory elements. Importantly, such differential chromatin regions could revealed *trans*-acting variation potentially contributing to strain-specific gene expression. Our work provides a rich resource to study gene regulatory control in a cell type-specific manner in ventral midbrain.

ChIP-seq and computational analysis were performed by me; snATAC-seq was performed by me and Dr. Kamil Grzyb; the animal work was performed by Dr. Mélanie Thomas; the sequencing was performed by Dr. Rashi Halder.

4.2.2 Manuscript

Single nuclei chromatin profiles of midbrain from genetically distinct mouse strains reveal cell identity transcription factors and cell type-specific gene regulatory variation

Yujuan Gui[1], Kamil Grzyb[2], Mélanie H. Thomas[2], Jochen Ohnmacht[1,2], Pierre Garcia[2], Rashi Halder[2], Manuel Buttini[2], Alexander Skupin[2], Thomas Sauter[1], Lasse Sinkkonen[1]*

[1]Department of Life Sciences and Medicine (DLSM), University of Luxembourg, Belvaux, Luxembourg

[2]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Belvaux, Luxembourg

***Corresponding author:** Dr. Lasse Sinkkonen (lasse.sinkkonen@uni.lu)

Key words: single nuclei ATAC-seq – mouse strains – regulatory variation – midbrain – cell type identity

Single nuclei chromatin profiles of ventral midbrain reveal cell identity transcription factors and cell type-specific gene regulatory variation

Yujuan Gui[1], Kamil Grzyb[2], Mélanie H. Thomas[2], Jochen Ohnmacht[1,2], Pierre Garcia[2], Manuel Buttini[2], Alexander Skupin[2], Thomas Sauter[1], Lasse Sinkkonen[1]*

[1]Department of Life Sciences and Medicine (DLSM), University of Luxembourg, Belvaux, Luxembourg

[2]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Belvaux, Luxembourg

***Corresponding author:** Dr. Lasse Sinkkonen (lasse.sinkkonen@uni.lu)

SUMMARY

Cell types in ventral midbrain are involved in diseases with variable genetic susceptibility such as Parkinson's disease and schizophrenia. Many genetic variants affect regulatory regions and alter gene expression. We report 20 658 single nuclei chromatin accessibility profiles of ventral midbrain from two genetically and phenotypically distinct mouse strains. We distinguish ten cell types based on chromatin profiles and analysis of accessible regions controlling cell identity genes highlights cell type-specific key transcription factors. Regulatory variation segregating the mouse strains manifests more on transcriptome than chromatin level. However, cell type-level data reveals changes not captured at tissue level. To discover the scope and cell-type specificity of *cis*-acting variation in midbrain gene expression, we identify putative regulatory variants and show them to be enriched at differentially expressed loci. Finally, we find TCF7L2 to mediate *trans*-acting variation selectively in midbrain neurons. Our dataset provides an extensive resource to study gene regulation in mesencephalon.

Key words: single nuclei ATAC-seq – chromatin accessibility – mouse strains – genetic variation – *cis*-acting variation – *trans*-acting variation – midbrain – cell type identity – Wnt signaling

INTRODUCTION

The ventral midbrain, or mesencephalon, is one of the most evolutionary conserved brain structures in mammals (Vogt Weisenhorn et al., 2016). It is involved in tasks such as processing of sensory information and eliciting motor and cognitive control through its dopaminergic circuits (Vogt Weisenhorn et al., 2016). It is of particular interest due to its role in human diseases like Parkinson's disease and schizophrenia – diseases whose development and progression are significantly influenced by individual's genetic susceptibility (Klein and Westenberger, 2012; Li et al., 2017; Nalls et al., 2019; Williams et al., 2009).

Like other brain regions, midbrain harbors many different cell types that exhibit both functional and molecular diversity (Gantz et al., 2018; Korotkova et al., 2004; Saunders et al., 2018). A cell type can be distinguished by the profile of its expressed genes. Transcriptomic analysis at single cell level has identified 20 cell types and 58 subtypes in midbrain alone (Saunders et al., 2018). These unique gene expression profiles defining cell state and cellular identity are controlled by epigenetic mechanisms. This is achieved by a dynamic interplay of cell's chromatin structure and expressed transcription factors (TFs). In particular, the combinatorial action of cell type-specific master regulators, TFs that open and specifically bind gene regulatory regions, will result in the correct gene expression program for each cell type (Atlasi and Stunnenberg, 2017). The chromatin landscape and accessibility of TF binding sites can be elucidated using epigenomic analysis such as the assay for transposase-accessible chromatin followed by sequencing (ATAC-seq) (Buenrostro et al., 2015). So far, the ability to isolate pure populations of various brain cell types has been limiting the progress in the field. However, recent developments in single nuclei chromatin assays have now enabled massive parallel analysis of cell type-specific chromatin profiles in their native context (Cusanovich et al., 2018; Lake et al., 2018; Preissl et al., 2018; Sinnamon et al., 2019).

Typical human genomes differ from each other on average by 5 million genetic variants (Auton et al., 2015). Vast majority of these are located in the non-coding genome and those associated with complex traits are enriched at accessible gene regulatory regions in a cell type-specific manner (Maurano et al., 2012). Genetic variation at regulatory regions can influence TF binding and thereby gene expression either in *cis* or in *trans*, hereafter referred to as gene regulatory variation (Deplancke et al., 2016). Identifying genes, regulatory regions and cell types affected by regulatory variants associated with various traits can help to understand the molecular mechanisms underlying the trait in question. C57BL/6J and A/J are two genetically distinct inbred mouse strains often used in neurobiology and to study complex genetic traits. The two strains segregate by ~6 million variants, comparable to the genetic variation between typical human individuals, making them an interesting model system to understand the effects of regulatory genetic variation on the phenotypic expression of complex traits. Indeed, the two strains show genetic differences also in traits associated with midbrain function. For example, A/J is more likely to show anxious and less social behaviour (Moy et al., 2007) and has lower motor activity (Thifault et al., 2002). Moreover, we have recently shown that the two strains exhibit significant differences in their ventral midbrain transcriptomes (Gui et al., 2020). However, the underlying gene regulatory changes and cell type-specific epigenomic profiles of mouse ventral midbrain are not known.

Here we performed chromatin accessibility profiling of mouse ventral midbrain from two mouse strains at single nuclei level (snATAC-seq). We identify >260 000 individual regulatory regions across 20 658 epigenomic profiles. Based on the distinct chromatin profiles, the data can distinguish ten different cell types. For the main cell types we define sets of unique cell identity genes and identify TFs controlling their expression. Comparing gene expression and chromatin accessibility between the mouse strains shows that genetically driven differences are more striking at the transcriptomic than chromatin accessibility level. Nevertheless, regulatory

regions with alterated chromatin accessibility are enriched at differentially expressed genes and can reveal cell type-specific gene regulation. We find *cis*-acting variants to be enriched at differentially expressed genes and pinpoint the extent of cell type-specific gene regulatory variation. Finally, we suggest canonical Wnt signalling to be a mediator of *trans*-acting variation in midbrain neurons.

RESULTS

Single nuclei chromatin profiles of ventral midbrain and identification of major cell types in two mouse strains

To unravel the cell type-specific gene regulation in the midbrain, and how it impacts genetic regulatory variation, we performed ATAC sequencing at single nuclei level (snATAC-seq) for dissected midbrains from two genetically distinct mouse strains, C57BL/6J and A/J (Figure 1, Supplemetary Figure S1). Two perfused ventral midbrain sections from both strains were used for the partitioning and barcoding with a total of 13 640 and 13 259 nuclei from C57BL/6J and A/J, respectively, and afterwards subjected to high throughput sequencing. After filtering the nuclei for multiplets and low coverage, approximately 290 million reads per mouse strain were retained, corresponding to 10 298 (C57BL/6J) and 10 360 (A/J) individual accessibility profiles (Figure 1A). The bulk chromatin accessibility profile aggregated across single nuclei (bulk snATAC-seq) from C57BL/6J showed a total of 231 390 peaks. Notably, 99.7% of regular bulk ATAC-seq peaks obtained from an independent C57BL/6J midbrain section overlapped with bulk snATAC-seq peaks (Figure S1). Moreover, the bulk snATAC-seq profile from A/J with 235 157 peaks was also highly correlated with C57BL/6J profile (Pearson $R > 0.97$). Finally, to be able to distinguish accessible regions at enhancers and promoters actively engaged in transcriptional control, we performed ventral midbrain bulk level ChIP-seq analysis in both

mouse strains for histone H3 lysine 27 acetylation (H3K27ac) (Rada-Iglesias et al., 2011; Siersbæk et al., 2017). Both bulk ATAC-seq and H3K27ac ChIP-seq showed clear correlation with midbrain gene expression levels (Figure S2).

Using an existing single cell genomics toolkit (Butler et al., 2018; Satija et al., 2015), the dimensionality of snATAC-seq was calculated by performing latent semantic indexing (LSI), to allow clustering of the cells with uniform manifold approximation and projection (UMAP) (Figure 1A). A gene activity matrix of snATAC-seq was established by counting reads in the gene body and the promoter region (2 kb upstream of transcription start site (TSS)). To annotate the obtained clusters as individual cell types, we took advantage of existing single cell RNA sequencing (scRNA-seq) of mouse midbrain (Saunders et al., 2018). Through identification of anchor genes shared between the gene activity matrix of snATAC-seq and the highly variable features in scRNA-seq, we could identify 10 different midbrain cell types with distinct chromatin accessibility profiles and sufficient numbers of cells in both strains (Figure 1A, Table S1). The cell types grouped into 6 main clusters. These consisted of glial cell types such as astrocytes (13.5% of cells), microglia (4.2-5.6%), oligodendrocytes (14.-19.5%), and two subtypes of polydendrocytes (*Tnr*+ and *Tnr*+/*Cspg5*+) (3.2-3.9%). They also included two different types of endothelial cells (stalk and tip; 1.1-3.4%). Finally, the largest and most diffuse cluster making up more than half of all cells (57.1-60.8%) consisted of different types of neurons. While scRNA-seq data could distinguish up to 30 neuronal subtypes in the midbrain through combinations of marker gene abundances (Saunders et al., 2018), at chromatin accessibility level these could not be so clearly distinguished. Instead, only three classes of neuronal cell types could be well distinguished: thalamus glutamatergic neurons (referred to as *Slc17a6*+ neurons), dopaminergic neurons (referred to as *Th*+ neurons), and a broader group of neurons consisting to large extent, but not exclusively, from different *Gad2*+ GABAergic neurons (referred to simply as neurons).

Interestingly, an increased proportion of *Th*+ and *Slc17a6*+ neurons and decreased proportions of oligodendrocytes and macrophages could be detected in A/J samples compared to C57BL/6J, while the proportion of astrocytes and *Tnr*+/*Cspg5*+ polydendrocytes remained almost identical (Figure 1A).

Inspection of genomic loci encoding for known cell type-specific marker genes in C57BL/6J samples disclosed highly cell type-selective chromatin accessibility that was well in keeping with gene expression levels from scRNA-seq data of mouse midbrain (Figure 1B). While ubiquitously expressed *Rpl13a* gene had high and consistent levels of accessibility across the cell types, known marker genes for astrocytes (*Aldh1l1*) and microglia (*Tmem119* and *Selplg*) (Bennett et al., 2016; Cahoy et al., 2008) were expressed and most accessible in the respective cell types, especially at their TSS. Similarly, the gene encoding for dopamine transporter (*Slc6a3*) had highest levels of exression and accessibility in the *Th*+ neurons, while in other neurons almost no signal could be detected. At the same time, the adjacent *Clptm1l* gene harboured an accessible promoter in all of the cell types. Finally, the locus encoding for two TFs required for oligodendrocyte generation and maturation, *Olig1* and *Olig2* (Lu et al., 2002; Mei et al., 2013), showed highest accessibility in subtypes of polydendrocytes and oligodendrocytes, as well as astrocytes, again consistent with the gene expression levels.

Importantly, the accessibility profiles between C57BL/6J and A/J were highly comparable also at the level of individual cell types and could equally highlight cell type-specific accessibility consistent with gene expression levels, as shown in Figure S1 for *Aif1* , a known marker gene for microglia.

Taken together, our snATAC-seq profiling produced over 20 000 chromatin profiles of mouse midbrain cell types with comparable quality from two different mouse strains. These data allow

the distinction of 10 different midbrain cell types at epigenomic level that are consistent with known gene expression profiles.

Identification of cell identity genes and associated regulatory regions from single cell data

To leverage the available data for the identification of TFs controlling cellular identity in adult midbrain cell types, we first set out to determine the genes whose expression was selective for each cell type. To obtain these cell identity genes, we used the existing scRNA-seq of the mouse midbrain, and for each gene determined the 85th percentile of its expression across all cell types indicating gene specific "high expression" (Figure 2). To filter out genes being selectively expressed in a specific cell type, a cut-off was applied of at least 60% of the cells of that cell type having the gene under consideration expressed not less than the 85th percentile. Furthermore, to ensure uniqueness, no other cell type was permitted to have the same gene among its top expressed genes in more than 40% of the cells. Through this approach we could define between 47 and 412 identity genes for each cell type (Figure 2; Table S2). On average 170 genes per cell type were determined. To confirm the relevance of the genes for the biology of the cell type in question, GO enrichment analysis for biological processes was performed. In keeping with the genes' role in the molecular and biological identity of the cell types, the top enriched GO terms included: Positive regulation of angiogenesis for endothelial stalk cells; Neutrophil mediated immunity for microglia; Axonogenesis, Neurotransmitter transport, and Regulation of synaptic vesicle exocytosis for the different neurons; Septin ring assembly and Myelination for oligo- and polydendrocytes; and Negative regulation of neuron differentiation for astrocytes. Examples of gene expression profiles of identity genes from selected cell types are shown in Figure 2B. Full list of enriched GO terms are provided in Table S3.

Next, to determine the gene regulatory regions controlling the expression of the cell identity genes, we performed peak calling on the cell type-specific aggregate ATAC-seq signals and associated the peaks to the defined identity genes of the respective cell types using GREAT (basal regulatory region +/-100 kb from TSS or up to nearest gene (McLean et al., 2010)). This resulted in 100 – 1200 accessible regions likely to control cell identity gene expression in each cell type (Figure 2; Table S4).

Cell type-specific chromatin accessibility profiles uncover cell identity regulating transcription factors

Comparison of the chromatin accessibility levels across the cell types confirmed a clear increase in accessibility at the obtained cell identity peaks associated with respective cell identity genes (Figure 3A). The highest increase in signal over background of aggregated midbrain cells was always detected in the corresponding cell types expressing the associated identity genes. At the same time, depletion of signal could be detected in other cell types. Interestingly, the level of accessibility also reflected the developmental relationships of the cell types. The strongest depletion of signal could be detected in the developmentally most distant cell type, the microglia (Ginhoux et al., 2010). And consistently, microglia identity peaks showed the strongest depletion in all other cell types. In contrast, neuron identity peaks showed no major depletion of signal in the related *Th*+ neurons and *vice versa*. Altogether, our approach could accurately identify cell type-specific gene regulatory regions controlling cell type identity.

To identify TFs binding the regulatory regions and controlling the cell type identity genes, we performed TF binding motif analysis in sequences enriched at cell identity peaks (Figure 3B). This analysis was done for eight cell types with highest sequencing coverage. Importantly, the

analysis highlighted motifs for several TFs previously shown to control the differentiation or identity of the respective cell type. These included SOX9 in astrocytes (Stolt et al., 2003), SPI1 in microglia (also known as SFPI1 or PU.1 (Gosselin et al., 2017; Kierdorf et al., 2013)), SOX13 in endothelial stalk (McGary et al., 2010), and SOX10 and SOX8 in oligodendrocytes (Stolt et al., 2002, 2003). The most enriched motif across cell types was the shared binding site for CTCF and CTCFL, the sequence occuring at insulator regions where CTCF mediates chromatin looping events (Atlasi and Stunnenberg, 2017). Thus, indicating the involvement of cell type-selective chromatin loop- and topological domain formation at the cell identity gene loci.

In addition, the enriched motifs included NFI-family motif in astrocytes and polydendrocytes, consistent with the reported role of these factors in the transition from neurogenesis to gliogenesis (Deneen et al., 2006) and the requirement of NFIC for expression of astrocyte marker genes (Wilczynska et al., 2009). Motifs enriched in microglia included TBX20 and MAFF motifs that can also be bound by MAFB, a TF recently shown to be important for maintenance of homeostasis in adult microglia (Matcovitch-Natan et al., 2016). Interestingly, RFX-family motif was highly enriched both in astrocytes and in neurons. Indeed, *Rfx1*, *Rfx3*, *Rfx4*, and *Rfx7* are known to be expressed and to play a role in the brain (Sugiaman-Trapman et al., 2018), with *Rfx4* showing the strongest expression in astrocytes while *Rfx3* and *Rfx7* are abundant in different neurons (Saunders et al., 2018). Thus, our data warrant further investigation of role of individual TFs of the RFX-family in the cellular identity of midbrain neurons and astrocytes.

Together with RFX3, another TF with enriched motif in neurons, ZNF740, also has been shown to localize at gene enhancers active specifically in differentiated human neuronal cell lines, further supporting the relevance of this prediction across species (Pierce et al., 2018). Finally, the motifs enriched uniquely in *Th*+ neurons included binding sites for KLF family TFs, MEF2

TFs, and ZBTB7 TFs (Figure 3B). From these particularly *Klf9*, *Mef2a*, *Mef2d*, and *Zbtb7c* show high expression in *Th*+ neurons (Saunders et al., 2018), with *Mef2d* exhibiting the most selective expression, an observation that could guide more detailed experiments into their role in dopaminergic neuron identity.

Genetically driven chromatin accessibility changes reveal cell type-specific gene expression changes

We have recently shown that the midbrain phenotypic differences between C57BL/6J and A/J mice (and associated behavioural changes) are accompanied by extensive gene regulatory variation (Gui et al., 2020). Based on tissue-level bulk RNA-seq analysis, 1151 genes are significantly differentially expressed (>2-fold, FDR<0.05) in the ventral midbrain between the two strains (Figure 4A). However, the *cis*- and *trans*-acting mechanisms underlying these genetically driven changes, and the affected cell types, are not known.

To address the contribution of chromatin level changes to the gene expression variation, we compared the bulk snATAC-seq signals from the mouse strains and focused on concatenated peak regions with at least two-fold difference in aggreagated read counts (Figure 4A). From this comparison it was clear that both the number of affected genomic regions, and especially the extent of changes at the chromatin accessibility level, were more modest than what could be observed at the transcriptomic level. Nevertheless, 287 of the total of 263 709 called peaks were altered more than two-fold at the bulk level. When associating accessible regions to their target genes we could observe a significant enrichment of these regions at the differentially expressed genes. This indicates, that at least some of the gene expression changes could be linked to changes at the chromatin level. Observing the data at the level of individual cell types allowed the detection of much higher proportion of differentially accessible regions, suggesting

that some of the cell type-specific changes could be masked by the tissue level analysis. While this increase in differential peaks was also associated with a lower number of sequenced cells, and therefore increased noise within the data, the significant enrichment of the differential peaks at differentially expressed genes was true in each cell type (Figure 4A).

For genes like *Isoc2b*, the decreased gene expression in ventral midbrain of A/J was accompanied by reduced accessibility of the promoter across all cell types (Figure 4B). To confirm the lost accessibility was also accompanied by reduced transcriptional activity, we observed H3K27ac levels at the promoter. Consistent with reduced ATAC-seq signal, H3K27ac was also lost at *Isoc2b* locus in A/J.

For ubiquitously expressed genes like *Isoc2b* the altered gene expression could be associated with chromatin level changes even at bulk level analysis. However, for other genes such as *Olfr287* a reduced expression could be observed by RNA-seq although no signal was detectable at bulk chromatin level by any of the methods (bulk ATAC-seq, bulk snATAC-seq, and ChIP-seq). Still, when observing the cell type-specfic snATAC-seq data, an accessible region could be detected at *Olfr287* promoter specifically in astrocytes. And consistently with reduced gene expression, the chromatin was less accessible in A/J (Figure 4B).

In summary, gene regulatory variation in midbrain is associated with chromatin level changes in accessibility, although not at all loci and with lower sensitivity than in transcriptomic analysis. Interestingly, snATAC-seq can reveal cell type-specific regulatory changes not captured in bulk level analysis.

Putative *cis*-acting variants are enriched at midbrain regulatory regions associated with differentially expressed genes

To obtain further insight into the mechanisms underlying the strain-specific gene expression, we next set out to address the extent of *cis*-acting regulatory variation contributing to the observed differences in the midbrain. For this we focused on identification of putative midbrain regulatory variants segregating C57BL/6J and A/J. First we performed TF footprint identification from our midbrain chromatin accessibility profile obtained through the bulk ATAC-seq analysis. Then, these binding sites were overlapped with >6 million variants segregating C57BL/6J and A/J to identify those with the potential to disrupt TF binding. Finally, the binding sites were overlapped with the midbrain H3K27ac profiles from both C57BL/6J and A/J to capture the binding sites engaged in transcriptional activity in either mouse strain, in total yielding 3909 putative regulatory variants of the ventral midbrain (Table S5).

The capacity of the above approach to reduce the number of meaningful variants is illustrated in Figure 5A with the examples of the *Ddhd1*, *Zfp615*, and *4.5S rRNA* loci. Expression of *Ddhd1*, a gene coding for a phospholipase, is modestly but significantly reduced in A/J compared to C57BL/6J and shows accessible chromatin at its TSS and at an upstream enhancer site >20 kb from the TSS. Both regions are marked by H3K27ac signals in both strains. One TF footprint could be identified at both the TSS and the distal enhancer, representing the putative TF binding sites controlling *Ddhd1*. From total of 603 variants at the 61 kb locus, only one coincides with an active TF binding site occupied in the midbrain, corresponding to a putative regulatory variant influencing *Ddhd1*expression in this brain region. Consistently, the affected enhancer shows decreased H3K27ac enrichment in the A/J. This illustrates how majority of genetic variants at any given locus are unlikely to affect gene expression and how, by focusing on those co-localizing within active regulatory regions, those most likely to act as regulatory variants can be identified.

If the midbrain gene regulatory differences between C57BL/6J and A/J indeed depend on the cumulative effect of *cis*-acting variants, such as the variant at the *Ddhd1* locus, the identified regulatory variants would be expected to be enriched in regulatory regions and TF binding sites at the differentially expressed gene loci compared to other expressed genes. To test this directly, we associated all putative regulatory variants to their likely target genes as already outlined in Figure 2, and calculated the number of variants that on average associate with each of the 4794 differentially expressed genes (FDR<0.05, (Gui et al., 2020)). As a control we did the same for all expressed genes found not to change between the strains (FDR>0.05) and for an equal number of randomly selected expressed genes. While unaffected genes and randomly selected genes were associated on average with 0.14 and 0.18 regulatory variants, respectively, this number significantly increased to 0.40 regulatory variants for the differentially expressed genes (Figure 5B). Consequently, genetic variants located in midbrain regulatory regions do not show a random distribution but are instead enriched at the differentially expressed genes, suggesting they play an important role in explaining the observed transcriptomic differences.

Next, we considered whether localization of variants in the TF binding sites of active enhancers could also be associated directly with enhancer activity upstream of gene expression changes. With this aim we used THOR (Allhoff et al., 2016) to identify enhancer regions with significantly altered signal for the H3K27ac enhancer mark between midbrains of C57BL/6J and A/J. Interestingly, 1126 of the 3909 putative regulatory variants localized within an enhancer region with differential H3K27ac enrichment, even when using a stringent cut-off ($p<1x10^{-18}$) for the differential peak calling (Table S6). This indicates that a large proportion of putative regulatory variants associate with enhancers that gain or lose activity between the mouse strains. For example, enhancer harbouring putative regulatory variants in the proximity of *4.5S rRNA* locus exhibits a strong gain of enhancer activity in A/J compared to C57BL/6J (Figure 5A). And at locus like *Zfp619* both gain and loss of enhancer activity can be observed

simultaneously at two separate enhancers associated with multiple putative regulatory variants. Taken together, disruption of TF binding by variants across thousands of enhancer regions is likely to alter enhancer activity, and thereby midbrain gene expression in genetically diverse mouse strains.

Cell type-specificity of *cis*-acting variants in the midbrain

Having identified the putative *cis*-acting regulatory variants contributing to the midbrain gene expression phenotype between C57BL/6J and A/J, we next sought to understand how cell type-selective these variants are. Overlapping the variants with cell type-specific accessibility data suggested that majority of the variant binding sites (57.9%) were accessible, with potential to affect gene expression, in at least 6 out of the 10 cell types (Figure 5C-D). However, just under 14% of the variants were accessible in only 1 or 2 cell types, indicating how non-coding variation can also have very cell type-selective effects on gene expression (Figure 5D). Indeed, also variants with cell type-selective accessibility in only 1-3 cell types were significantly more often occurring at genes with altered expression than at other expressed genes (data not shown).

Nineteen putative regulatory variants were accessible only in *Th*+ neurons and were associated with five differentially expressed genes, representing dopaminergic neuron-specific gene regulatory variation. These included *Zfp68*, a poorly known transcriptional repressor upregulated in C57BL/6J (Figure 5E). Despite the observed differential expression, *Zfp68* promoter appeared comparably accessible in both strains in almost all cell types. However, specifically in *Th*+ neurons of C57BL/6J, an increased accessibility downstream of the TSS can be observed, exactly at a position overlapping a putative regulatory variants within a TF binding site. Providing an example of cell type-specific *cis*-acting variant in dopaminergic

neurons. While some increase could be observed at this position also in *Slc17a6*+ neurons, it could not be detected by applied peak calling parameters.

Taken together, while majority of *cis*-acting variants affect broad array of cell types, a large proportion can also have cell type-specific effects that cannot be dissected without single cell analysis.

TCF7L2 as mediator of *trans*-acting variation in midbrain neurons

A large fraction of the midbrain gene expression variation could be linked to *cis*-acting regulatory variants, even with our strict criteria of presence of the variant in a TF footprint located in an active enhancer (Figure 5A). Still, much of the differential gene expression remained unexplained. This could be due to *cis*-acting variants we have missed, but also due to *trans*-acting variants that can influence a number of target genes by altering a TF's activity, rather than its binding motif. A change in TF activity could be due to change in its expression levels, but could also be due to alternations in other mechanisms controlling TF activity such as post-translational modifations, protein-protein interactions or TF localization.

Genetic differences in non-dopaminergic neurons (such as *Gad2*+ neurons), that make up much of our neuron population (Table S1), have been suggested to contribute to strain specific behavioural differences, including anxiety, reward and motivation traits, such as ethanol consumption (Morales and Margolis, 2017; Ponder et al., 2007; Portugal et al., 2012; Yoneyama et al., 2008). To identify mediators of *trans*-acting variation between C57BL/6J and A/J in neurons, we performed motif enrichment analysis for the 498 regions showing >2-fold change in chromatin accessibility between the mice in neurons (Figure 4A). This revealed the binding motif of LEF1 and TCF7L2, downstream TFs of the canonical Wnt signaling pathway (Nusse and Clevers, 2017), to be the most enriched sequence found at more than third of the

differentially accessible regions (Figure 6A; p = 1e-35). The enrichment was specific for *Gad2*+ and *Slc17a6*+ neurons and could not be found in any of the other cell types (Figure S3) or in the motif enrichment analysis for cell identity genes (Figure 3). LEF1 and TCF7L2 bind the same DNA sequence but have often opposing or cell type-specific functions (Mao and Byers, 2011). Inspecting chromatin accessibility across the cell types for the differential binding sites carrying the LEF1 and TCF7L2 motifs revealed an increased signal specifically in the neurons (Figure 6B-C). To determine whether either factor is expressed in the midbrain neurons and could mediate the observed enrichment and altered accessibility, we visualized their expression using scRNA-seq data. Interestingly, *Lef1* expression was limited to the endothelial cells while *Tcf7l2* had highest expression in the *Gad2*+ and *Slc17a6*+ neurons and polydendrocytes, showing a clear overlap with cells enriched for respective binding motif (Figure 6D). Thus, the transcriptional activity of TCF7L2 is likely to be altered between C57BL/6J and A/J mice in the midbrain neurons.

Taken together, snATAC-seq analysis of tissues from genetically different strains can guide the elucidation of cell type-specific impact of *trans*-acting variants and suggests neuron-specific differences in the canonical Wnt signalling pathway between two commonly used inbred mouse strains.

DISCUSSION

Here we investigated the chromatin accessibility in cell types of mouse ventral midbrain in two different genetic backgrounds and provide a large resource of 20 658 single nuclei chromatin profiles from 10 different cell types. This dataset will benefit future studies on the role of these cell types in various processes involving this brain area such as movement control, cognition and reward mechanisms. A better understanding of midbrain cell types can also profit research

on diseases like PD and schizophrenia. In particular, improved comprehension of genetic variation in gene regulation and how it impacts specific cell types, will pave way for better prediction of genetic susceptibility and affected disease mechanisms.

Our findings on gene accessibility profiles and cell type composition of the midbrain are consistent with existing knowledge from single cell transcriptomics (Saunders et al., 2018). However, neuronal subtypes are more difficult to distinguish at chromatin level than what has been achieved by transcriptomic analysis. Neuron subtypes clustered largely together, often with undefined borders between the subtypes (Figure 1A). This result is expected. While chromatin accessibility is generally known to show positive correlation with gene expression (Liu et al., 2019a; Wu et al., 2016), and this is also true for our data (Figure S2), enhancer accessibility does not necessarily reflect gene regulatory activity (Arnold et al., 2013). Indeed, chromatin accessibility profiles of cell types executing similar functions can be highly similar despite showing different expression patterns and being controlled by different master TFs (Maurano et al., 2012; McKay and Lieb, 2013). Moreover, accessibility can signify priming of a locus for expression without commencing transcription (Lara-Astiaso et al., 2014; Pálfy et al., 2020).

Nevertheless, we could reliable distinguish 10 of the 20 known midbrain cell types also at the chromatin level, with 8 cell types containing enough cells for more detailed analysis (Figure 3). Using this information about gene regulatory regions selectively associated with genes underlying cellular identity, we were able to predict the key regulators of each cell type through motif enrichment analysis. These included many factors previously determined to be necessary for the differentiation or maintenance of the respective cell state (Gosselin et al., 2017; Kierdorf et al., 2013; McGary et al., 2010; Stolt et al., 2002, 2003). Among other insights, the results reveal a cell type-specific role for CTCF binding sites in cellular identity and predict specific roles for MEF2, KLF, and ZBTB7 family TFs in dopaminergic neurons. Moreover, the results

give additional support for a more detailed analysis of RFX family of TFs in regulation of neuro- and gliogenesis balance in the midbrain, something that is consistent with previously identified co-occupancy of RFX factors at SOX2-bound enhancers in neurodevelopment (Lodato et al., 2013).

Genetic variation is known to drive gene expression changes that can trigger phenotypic differences (Fay et al., 2004; Keane et al., 2011; Massouras et al., 2012; Storey et al., 2007). We have recently shown over thousand genes to be differentially expressed in the mouse midbrain due to genetic variation between C57BL/6J and A/J (Gui et al., 2020). Here we leveraged our single cell chromatin accessibility profiles to investigate the mechanisms, cell type-specificity, and extent at which this variation is reflected at the level of chromatin. Interestingly, the transcriptomic differences did not show a linear relationship with midbrain chromatin accessibility (Figure 4A). Both at bulk aggregate levels and in individual cell types the extent of change in accessibility was not comparable to mRNA level change. This is most likely reflecting the observation that TF binding and activity can alter gene expression without change in accessibility, for example if the locus is already open (Cao et al., 2018). Still, when changes in accessibility did occur, this was often associated with local gene expression change. Importantly, some of the chromatin accessibility changes associated with differential gene expression could not be detected at all in tissue level ATAC-seq or ChIP-seq analysis despite high sequencing depth (Figure 4B). Therefore, the improved resolution offered by single cell analysis allows further insights into regulatory interactions that could be missed in tissue level analysis.

Cis-acting variants that influence gene expression have been suggested to explain a significant part of the missing heritability (Ge et al., 2009; Grundberg et al., 2012; Ohnmacht et al., 2020). And increasing number of such variants associated with human traits and diseases have now been experimentally validated (Deplancke et al., 2016). By combining genetic information with

ATAC-seq and ChIP-seq analysis, we found 3909 of the >6 million variants segregating C57BL/6J and A/J to localize in a TF binding site within an active and accessible enhancer in the ventral midbrain (Figure 5). This number is consistent with the previous estimates that as many as one in thousand mouse variants cause *cis*-regulatory effects (Crowley et al., 2015). Importantly, we found the putative *cis*-acting variants to be enriched at differentially expressed genes, indicating they do contribute to the gene expression phenotype.

A large proportion of the putative *cis*-acting variants showed cell type-selective accessibility (Figure 5B). This finding is consistent with the previous work on impact of *cis*-acting variants on disease-associated regulatory variation in the major cell populations of the human brain (Nott et al., 2019). By combining the information on cell type-specific impact of variants on gene regulation with genome-wide association studies and quantitative trait locus mapping for specific traits or diseases can provide insights on the mechanisms how the variants translate into phenotypes. C57BL/6J and A/J differ from each other for numerous phenotypes (Bogue et al., 2020). These include fear-conditioning and reward-releated behaviours like ethanol consumption that are associated with midbrain dopamine signalling (Ponder et al., 2007; Portugal et al., 2012; Yoneyama et al., 2008). Combining our data on cell type selectivity of *cis*-acting variants with genome-wide association studies of such phenotypes could reveal new connections between genes and the underlying mechanisms.

Besides *cis*-acting variation, differential gene expression and altered chromatin accessibility can also be mediated through *trans*-acting variation. *Trans*-acting variants can have many different mechanisms of action and majority of genetically driven gene expression variation has been suggested to be due to *trans*-acting variation, often affecting hundreds or thousands of genes in a cell type-specific manner (Albert et al., 2018; Grundberg et al., 2012; Liu et al., 2019b). Dopaminergic neurons in ventral tegmental area of the midbrain contribute to phenotypes like reward-behaviour and fear (Morales and Margolis, 2017). Moreover, the

associated GABAergic and glutamatergic neurons in midbrain can also control such behaviours, either together with or independent of the dopamine signalling. With this in mind, we asked whether specific signature of *trans*-acting regulatory variation could be found in neurons between our strains that show differential behaviour, based on motif enrichment analysis at regions of differential accessibility (Figure 6). We found an enrichment for the shared binding motif of TCF7L2 and LEF1, corresponding to the Wnt signalling response element recognized by their high mobility group (HMG) DNA-binding domain (Nusse and Clevers, 2017). This suggests that Wnt signalling is altered between C57BL/6J and A/J in neurons. Based on scRNA-seq data TCF7L2, but not LEF1, is abundantly expressed in midbrain *Gad2*$_+$ and *Slc17a6*$_+$ neurons, indicating that alterations in chromatin accessibility and gene expression are due to changes in Wnt signalling and likely to be mediated by TCF7L2 (Figure 6).

Involvement of TCF7L2 as the putative effector of the altered signalling in neurons is particularly intriguing since *TCF7L2* locus has been genetically associated with mental disorders like scizophrenia in humans (Bem et al., 2019). Moreover, transgenic mice have revealed a dose-dependent role for *Tcf7l2* in fear-conditioning and anxiety, traits for which A/J is known to significantly differ from C57BL/6J (Ponder et al., 2007; Portugal et al., 2012). Understanding how Wnt signalling activity is altered between the mouse strains requires further analysis. However, it is interesting to note that expression of *Wnt2b,* upstream ligand of Wnt pathway genetically associated with bipolar disorder (Bem et al., 2019), is significantly decreased in the ventral midbrain of A/J compared to C57BL/6J (Gui et al., 2020). Thus, providing one possible explanation.

Taken together, our single nuclei chromatin analysis provides novel insights into transcriptional control of ventral midbrain cell types and a rich resource for further analysis of cellular identity and gene regulatory variation in this disease-relevant brain region.

ACKNOWLEDGEMENTS

We would like to thank Drs Aurélien Ginolhac and Anthoula Gaigneaux for their support with bioinformatic analysis, Dr Djalil Coowar (Animal Facility of University of Luxembourg) for help with breeding of experimental mice, Dr Rashi Halder at LCSB sequencing facility for high-throughput sequencing, and Sergio Helgueta for help with nuclei isolation. The computational analysis presented in this paper were carried out using the HPC facilities of the University of Luxembourg (Varrette S et al., 2014).

FUNDING

LS and MB would like to thank the Luxembourg National Research Fund (FNR) for the support (FNR CORE C15/BM/10406131 grant). LS and JO would like to thank Fondation du Pélican de Mie et Pierre Hippert-Faber and Luxembourg Rotary Foundation for funding.

AUTHOR CONTRIBUTIONS

YG, MB, and LS conceived the project. YG, KG, JO, AS, TS and LS designed the experiments. YG and KG performed snATAC-seq. YG and JO performed bulk ATAC-seq and YG performed ChIP-seq experiments. MT and PG prepared mouse tissues. YG performed all bioinformatic analysis with help from TS and LS. YG and LS analysed the results. MB, TS, and LS provided funding. YG and LS wrote the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

FIGURE LEGENDS

Figure 1. Midbrain snATAC-seq identifies cell type-specific accessibility.

(A) Clustering of snATAC-seq from C57BL/6J and A/J with corresponding cell type proportions. Major cell types can be identified based on snATAC-seq profiles, with neurons having the biggest proportion on both strains. Cell types in C57BL/6J and A/J have comparable proportions with more than half of neuclei being identified as neurons. (B) Cell type-specific accessibility is observed in marker genes. The genomic tracks are from C57BL/6J midbrain snATAC-seq. The expression profiles measured as transcript per 100 000 in cluster. *Rpl13a* is used as a house keeping gene to normalize the snATAC-seq signal. See also Figure S1 and S2.

Figure 2. Regions controlling cell type identity can be defined by combining snATAC-seq and scRNA-seq.

(A) Schematic workflow to define cell-type specific signatures. Digital gene expression is obtained from DropViz. For each gene, the 85th percentile of its expression across all cell types was calculated. To define a gene as a cell type-identity gene, at least 60% of the cells of a cell type should have expression more than the 85th percentile, while at the same time no other cell type was permitted to have the same gene among its top expressed genes (above the 85th percentile) in more than 40% of the cells. Enrichment analysis with cell type-identity genes found GO terms corresponding to cell type characteristics. The cell type-identity peaks are defined by peaks overlapping with the regulatory regions of cell type-identity genes (basal region +/- 100 kb until nearby genes). Subsequently, the enriched motifs in cell type-identity peaks are detected. (B) Examples of identified cell type identity genes. The identified cell type-identity genes for six major cell types show selective expression in the respective cell types when observing scRNA-seq data of the entire population of midbrain cells.

Figure 3. Identification of cell type specific TFs controlling cellular identity

(A) Heatmap showing the enriched signal of cell type-identity peaks in eight cell types. The analysis was done on C57BL/6J midbrain snATAC-seq. The background is constructed by merging the sampling reads (366278 reads / cell type) from each cell type. The raw signal is normalized to the background and library, following $\log_2$-transformation. The normalized signal is plotted 3 kb up- and downstream of peaks. (B) Motif enrichment analysis on cell type-identity peaks. The PWM logos, names of the associated TFs and p-values are shown for each motif. The motifs are ranked according to p-values.

Figure 4. Association of differentially accessible regions with altered gene expression between C57BL/6J and A/J.

(A) Differential peaks are highly associated with differential genes. The differential peaks (labelled as red) are defined as more than 2 fold change in read count in merged peak area. The read counts in peaks of snATAC-seq bulk are $\log_{10}$-transformed. Peaks with low read count (less than median - 1.5 median absolute deviation) are filtered out. To associate differential peaks to DEGs, peaks are overlapped with the regulatory region of DEGs (basal region +/- 100 kb until nearby genes). As a control, random peaks are selected by bootstrapping with 1000 repetions ($p < 0.0099$). The RPKM from bulk RNA-seq of C57BL/6J and A/J is also $\log_{10}$-transformed, and DEGs are defined as FDR < 0.05 and $\log_2$-fold change > 1 (labelled as red). (B) Cell type-specific differential peaks correlate with gene expression in bulk RNA-seq. The differential peaks are labelled as green.

Figure 5. Putative regulatory variants are associated with differentially expressed genes and show cell type-selective accessibility.

(A) Examples of putative regulatory variants of A/J found in the enhancer region upstream of *Ddhd1*, *Afp619* and *Rn4.5s*. The putative regulatory variants are defined as variants disrupting TF footprints located in active enhancers (defined by H3K27ac). (B) Each DEG (5077, FDR > 0.05) is associated with an average of 0.4 putative regulatory variants, while non-DEGs are associated with only 0.14 variants. Random: 5000 genes are randomly selected from all expressed genes. (C) Putative regulatory variants have differentical accessibility across cell types. Differential accessibility of putative regulatory variants is shown in the heapmap (red: accessible, blue: non-accessible). (D) More than half of the variants are accessible in more than 6 cell types while 7% in only 1 cell type. (E) An example showing how putative regulatory variants with differential accessibility affect cell type-specific gene expression. The variants locating near the TSS of *Zfp68* are associated with a broader signal in *Th*+ neurons of C57BL/6J, potentially resulting in upregulation of *Zfp68* as shown in bulk RNA-seq.

Figure 6. TCF7L2 as a mediator for *trans*-acting variation in neurons.

(A) Motif enrichment analysis for regions of differential accessibility between C57BL/6J and A/J in neurons found TCF7L2 and LEF1 motifs with particularly high enrichment. (B-C) Regions of differential accessibility with motif for LEF1 (B) or (C) TCF7L2 show increased accessibility in neurons compared to other cell types. (D-E) *Lef1* is highly expressed in endothelial stalk, while *Tcf7l2* is abundant in neurons and polydendrocytes. See also Figure S3.

METHODS

MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by Lasse Sinkkonen (lasse.sinkkonen@uni.lu). This study did not generate new unique reagents.

Animals

All experimental procedures in this study were in compliance with the European Communities Council Directive 2010/63/EU, following the 3 Rs' requirements for Animal Welfare. We used two mouse strains in this study, C57BL/6J and A/J, purchased respectively from Charles River and Jackson Laboratory. The study cohorts were bred in-house at the Animal Facility of University of Luxembourg (Esch-sur-Alzette, Luxembourg) and the protocol was approved by the Animal Experimentation Ethics Committee (AEEC) according to the national guidelines of the animal welfare law in Luxembourg (*Règlement grand-ducal* adopted on January 11th, 2013). All mice were housed with a 12-hour light–dark schedule and had free access to food and water.

For each mouse, intracardiac perfusion with PBS was performed after anesthesia with a ketamine-medetomidine mix (150 and 1 mg/kg, respectively). The brain was extracted and both midbrains of each mouse were dissected, immediately snap-frozen, stored at -80°C, and used for single cell partitioning as described below.

Nuclei isolation

For bulk ATAC-seq, frozen midbrains were minced in a Dounce Tissue Grinder (Sigma, D8939-1SET) with A pestle ~10 times following B pestle ~20 times in lysis buffer [5 mM $CaCl_2$ (Merck, A546282), 3 mM $Mg(Ac)_2$ (Roth, P026.1), 10 mM Tris pH 7.8, 0.17 mM β-mercaptoethanol (Gibco, 21985-023), 160 mM sucrose (Sigma, S0389), 0.05 mM EDTA, 0.05% NP40 (Sigma, I3021)]. The lysate was layered on 3 mL sucrose cushion [1.8 M sucrose, 3 mM Mg(Ac)2, 10 mM Tris pH 8.0, 0.167 mM β-mercaptoethanol], following ultra-centrifugation 30,000g for 1 h (Rotor:Beckman Coulter, MLS-50) at 4°C. After centrifugation, the supernatant was discarded and the nuclei pellet was suspended in resuspension buffer (with 0.1% Tween-20, 0.01% digitonin and 0.1% NP40).

For snATAC-seq, the isolation of nuclei was done according to 10X Genomics protocol CG000212 (Rev A) with minor modification. In brief, frozen midbrain sections were minced in a Dounce Tissue Grinder (Sigma, D8939-1SET) with A pestle ~10 times following B pestle ~20 times in 500 uL chilled lysis buffer [10mM Tris-HCL (pH7.4), 10mM NaCl, 3mM $MgCl_2$, 0.1% Tween-20, 0.1% NP40, 0.01% Digitonin, 1% BSA]. The homogenized lysate was incubated on ice for 5 min, following pipette mixing 10x and incubated again on ice for 10 min. Chilled Wash Buffer (500 uL) [10mM Tris-HCl (pH 7.4), 10mM Nacl, 3mM $MgCl_2$, 1% BSA, 0.1% Tween-20] was added to the lysed cells and pipetted mix 5x. The lysate was layered on 3 mL sucrose cushion [1.8 M sucrose, 3 mM Mg(Ac)2, 10 mM Tris pH 8.0, 0.167 mM β-mercaptoethanol], following ultra-centrifugation 30,000g for 1 h (Rotor:Beckman Coulter, MLS-50) at 4°C. After centrifugation, the supernatant was discarded and the nuclei pellet was suspended in Nuclei Buffer (10x PN: 2000153) provided by 10X Genomics Chromium Single Cell ATAC Reagent Kit. The nuclei suspension was passed through a 40 um Flowmi Cell Strainer (Sigma, BAH136800040-50EA).

Bulk ATAC-seq

Tagmentation of mouse midbrain samples was done based on the OMNI-ATAC supplementary protocol 1 (Corces et al., 2017) with minor changes. Briefly, 25,000 nuclei were resuspended in 50 uL resuspension buffer (with 0.1% Tween-20, 0.01% digitonin and 0.1% NP40) and lysed for 3 minutes on ice. After washing in 1 mL resuspension buffer (1% Tween-20) samples were centrifuged for 10 minutes at 500 RCF at 4 °C and supernatant carefully removed. Pellets were resuspended in 25 μL tagmentation mix (Tagment DNA buffer from Illumina, #15027866) containing 2.5 μL Tagment DNA Enzyme (Illumina #15027865) and incubated for 45 min at 37 °C and 1000 rpm in Eppendorf ThermoMixer. Tagmented chromatin was isolated using Zymo Research DNA Clean& Concentrator kit (ZymoResearch ZY-D4013) and eluted in 21 μL elution buffer. Library pre-amplification was done for 5 cycles using primers Ad1 and Ad2.16 (Buenrostro et al., 2015). Five additional cycles of library amplification were done as determined by qPCR (Corces et al., 2017). Library cleanup was done using Zymo Research DNA Clean & Concentrator kit followed by AMPure XP bead (Beckman Coulter #A63880) size selection. A first bead incubation using 0.55x volume of beads removes large fragments. After separation of beads on magnetic stand, supernatant was transferred to a fresh tube and incubated for 5 minutes in 1.5x volumes of beads. After washing with 80% Ethanol, beads were resuspended in 20 μL elution buffer. After separation on magnetic stand eluate was transferred to a fresh tube. Library quality was assessed using Agilent DNA High sensitivity Bioanalyzer chip (Agilent #5067-4626).

Bulk ATAC-seq data analysis

The sequencing of ATAC-seq libraries was done at the sequencing platform in the Luxembourg Centre for Systems Biomedicine (LCSB) of the University of Luxembourg. The paired-end,

unstranded library sequencing was performed using Illumina NextSeq 500/550 75 cycles High Output Kit. Raw FASTQ files and BAM files were processed as described above in the Materials and Methods section of ChIP-seq. After processing of the BAM files, the peaks were called by Genrich (https://github.com/jsh58/Genrich) with parameters "-r -m 30 -j" to remove PCR duplicates and include only reads with mapping quality of at least 30. Footprints were called by HINT-ATAC (Li et al., 2019) with default parameters. Raw FASTQ files were deposited in ArrayExpress with the accession number E-MTAB-8333.

Single nucleus ATAC-seq library preparation and sequencing

The single nucleus ATAC-seq was performed according to Chrominum Single Cell ATAC Reagent Kits User Guide (CG000168 RevB) with Chromium Single Cell E Chip Kit (10X, 100086), Chromium Single Cell ATAC Library & Gel Bead Kit (10X, 1000111), Chromium Single Cell ATAC Gel Bead Kit (10X, 1000085), Chromium Single Cell ATAC Library Kit (10X 100087), Chromium i7 Multiplex Kit N Set (1000084), Dynabeads MyOne Silane (2000048). In brief, nuclei suspension was loaded with a targeted recovery rate of 10 000 nuclei per sample. snATAC-seq libraries quality were assessed using Agilent DNA High sensitivity Bioanalyzer chip (Agilent #5067-4626) and further sequenced on a 150 cycles High Output Kit using Illumina NextSeqTM 500 with targeted sequencing depth of 25 000 read pairs per nucleus.

Single nucleus ATAC-seq analysis pipeline

Cell ranger

The alignment and filtering were done according to 10X running pipelines. In brief, the fastq files was generated from Illumina sequencer's base call files, which were later used as inputs

to align (MAPQ > 30), filter barcode multiplets and generate accessibility counts for each cell in a single library. The technical replicates for each mouse strain were aggregated to create a single peak-barcode matrix. Each unique fragment is associated with a single cell barcode.

Clustering

The snATAC-seq downstream analysis was performed by Signac (version: 0.2.4) in R. The gene activity matrix was calculated with reads in gene body and 2 kb upstream of TSS as a proxy for gene expression. Nuclei with counts less than 5000 were filtered out. The dimensionality was calculated with latent semantic index (LSI) on peaks with at least 100 reads across all cells, which was used as input to generate UMAP graphs.

Cluster annotation

The annotation of snATAC-seq clusters took use of the existing scRNA-seq data on midbrains from adult 3-month-old C57BL/6N mice. The anchors were found between the gene activity matrix of snATAC-seq and the top 5000 variable features of scRNA-seq. The cell labels were transferred from the scRNA-seq to the snATAC-seq with normalization on anchor weights calculated from the LSI dimensional reduction, resulting in 10 347 nuclei of C57BL/6J and 10368 nuclei of A/J annotated.

scRNA-seq data analysis pipeline

Data processing

The DGE (digital gene expression) and cell annotation for midbrains of 3-month-old C57BL/6N mice were downloaded from DropViz. The data analysis was performed by Seurat (version: 3.1.4) in R. Only cells with feature counts between 400 to 7000 and being single or well-curated were used in downstream analysis, resulting in 19 967 cells in total. The DGE

was natural-log transformed and normalized to mitochondrial read counts. The dimentional reduction was done with UMAP. The clusters were annotated with existing annotation from DropViz.

Selection of cell type-identity genes

100 cells were randomly selected from each cell cluster. DGE from each cell type was constructed according to corresponding barcodes. The 85th percentile expression for each gene was calculated on the selected cells. The criteria for cell type-identity genes was defined as: For a particular gene, 60% of cells in a cell type have expression larger than the gene specific 85th percentile; while at most 40% of cells in all other cell types have expression larger than the 85th percentile. This process were repeated for 100 times. Genes that appeared more than 30 times out of 100 were defined as the cell type-identity genes.

Motif enrichment analysis

Generating cell type specific bam files

The cell type specific bamfiles were generated by samtools. The barcodes from each replicate of a strain in bamfile was relabled to avoid barcode collapse. After relabelling, the bamfiles from replicates of a sample were merged. The bamfile for each cell type were subtracted based on corresponding barcodes.

Peak calling

The peak calling was done by MACS2 (2.1.2) (Feng et al., 2012) with custom cutoff on p-values according to cutoff analysis with parameters ‘macs2 callpeak --cutoff-analysis’. The ideal cutoff was chosen based on that the selected p-value would not lead to exponential increase of peak numbers.

Motif enrichment analysis

The motif enrichment analysis was performed by HOMER (4.11.1) (Heinz et al., 2010) with parameters 'findMotifsGenome.pl -size given -mask'.

Chromatin Immunoprecipitation (ChIP)

ChIP was performed on the dissected snap frozen mouse ventral midbrain tissue as previously described (Gui et al., 2020). Each reaction had 10 – 14 μg of chromatin and 10% aliquot was used as input DNA. Immunoprecipitation was performed overnight with 5 μL of H3K27ac antibody (Millipore, 17-614) at 4°C with rotation.

ChIP-seq data analysis

The sequencing of the chromatin samples was done at the sequencing platform in the LCSB of the University of Luxembourg. The single-end, unstranded sequencing with read length of 75 bp was performed with Illumina NextSeq 500 machine. FastQC (v0.11.5) was used for raw reads quality assessment (Andrews S, 2010). Generation of BAM files, including steps of adapter removal, mapping and duplicate marking, were done with PALEOMIX pipeline (v1.2.12) (Schubert et al., 2014), followed by mapping with BWA (v.0.7.16a) (Li et al., 2009). Backtrack algorithm applied the quality offset of Phred score to 33. Duplicate reads were marked and the mouse reference genome, GRCm38.p5 (mm10, patch 5) was downloaded from GENCODE (https://www.gencodegenes.org/). Finally, validation of the Bam files was donw using Picard (v2.10.9) (Adams et al., 2000). Raw FASTQ files were deposited in ArrayExpress with the accession number E-MTAB-8333.

The H3K27ac (H3 lysing 27 acetylation) ChIP-seq peaks, enhancers and super-enhancers, were called by HOMER (Heinz et al., 2010) with default parameters. The signal normalization in pairwise comparison was done by THOR (v0.10.2) (Allhoff et al., 2016), with TMM normalization and adjusted p-value cut-off 0.01.

DATA AND CODE AVAILABILITY

Raw FASTQ files were deposited in ArrayExpress with the accession number E-MTAB-8333 for bulk levels analysis and E-MTAB-9225 for single cell data. The scripts for data analysis can be accessed here: https://github.com/sysbiolux/.

SUPPLEMENTAL INFORMATION

Supplemental Information is available as separate Supplemental Files.

REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. (2000). The Genome Sequence of Drosophila melanogaster. Science 287, 2185–2195.

Albert, F.W., Bloom, J.S., Siegel, J., Day, L., and Kruglyak, L. (2018). Genetics of trans-regulatory variation in gene expression. ELife 7, e35471.

Allhoff, M., Seré, K., F. Pires, J., Zenke, M., and G. Costa, I. (2016). Differential peak calling of ChIP-seq signals with replicates with THOR. Nucleic Acids Res. 44, e153.

Andrews S (2010). FastQC: a quality control tool for high throughput sequence data.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryń, Ł.M., Rath, M., and Stark, A. (2013). Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. Science 339, 1074–1077.

Atlasi, Y., and Stunnenberg, H.G. (2017). The interplay of epigenetic marks during stem cell differentiation and development. Nat. Rev. Genet. 18, 643–658.

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., et al. (2015). A global reference for human genetic variation. Nature 526, 68–74.

Bem, J., Brożko, N., Chakraborty, C., Lipiec, M.A., Koziński, K., Nagalski, A., Szewczyk, Ł.M., and Wiśniewska, M.B. (2019). Wnt/β-catenin signaling in brain development and mental disorders: keeping TCF7L2 in mind. FEBS Lett. 593, 1654–1674.

Bennett, M.L., Bennett, F.C., Liddelow, S.A., Ajami, B., Zamanian, J.L., Fernhoff, N.B., Mulinyawe, S.B., Bohlen, C.J., Adil, A., Tucker, A., et al. (2016). New tools for studying microglia in the mouse and human CNS. Proc. Natl. Acad. Sci. U. S. A. 113, E1738–E1746.

Bogue, M.A., Philip, V.M., Walton, D.O., Grubb, S.C., Dunn, M.H., Kolishovski, G., Emerson, J., Mukherjee, G., Stearns, T., He, H., et al. (2020). Mouse Phenome Database: a data repository and analysis suite for curated primary mouse phenotype data. Nucleic Acids Res. 48, D716–D723.

Buenrostro, J., Wu, B., Chang, H., and Greenleaf, W. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al 109, 21.29.1-21.29.9.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol.

Cahoy, J.D., Emery, B., Kaushal, A., Foo, L.C., Zamanian, J.L., Christopherson, K.S., Xing, Y., Lubischer, J.L., Krieg, P.A., Krupenko, S.A., et al. (2008). A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function. J. Neurosci. 28, 264–278.

Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., Daza, R.M., McFaline-Figueroa, J.L., Packer, J.S., Christiansen, L., et al. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science 361, 1380–1385.

Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B., et al. (2017). An improved

ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat. Methods 14, 959–962.

Crowley, J.J., Zhabotynsky, V., Sun, W., Huang, S., Pakatci, I.K., Kim, Y., Wang, J.R., Morgan, A.P., Calaway, J.D., Aylor, D.L., et al. (2015). Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. Nat. Genet. 47, 353–360.

Cusanovich, D.A., Hill, A.J., Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova, G.N., Huang, X., Christiansen, L., DeWitt, W.S., et al. (2018). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell 174, 1309-1324.e18.

Deneen, B., Ho, R., Lukaszewicz, A., Hochstim, C.J., Gronostajski, R.M., and Anderson, D.J. (2006). The Transcription Factor NFIA Controls the Onset of Gliogenesis in the Developing Spinal Cord. Neuron 52, 953–968.

Deplancke, B., Alpern, D., and Gardeux, V. (2016). The Genetics of Transcription Factor DNA Binding Variation. Cell 166, 538–554.

Fay, J.C., McCullough, H.L., Sniegowski, P.D., and Eisen, M.B. (2004). Population genetic variation in gene expression is associated with phenotypic variation in Saccharomyces cerevisiae. Genome Biol. 5, R26.

Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. Nat. Protoc. 7, 1728–1740.

Gantz, S.C., Ford, C.P., Morikawa, H., and Williams, J.T. (2018). The Evolving Understanding of Dopamine Neurons in the Substantia Nigra and Ventral Tegmental Area. Annu. Rev. Physiol. 80, 219–241.

Ge, D., Fellay, J., Thompson, A.J., Simon, J.S., Shianna, K.V., Urban, T.J., Heinzen, E.L., Qiu, P., Bertelsen, A.H., Muir, A.J., et al. (2009). Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature 461, 399–401.

Ginhoux, F., Greter, M., Leboeuf, M., Nandi, S., See, P., Gokhan, S., Mehler, M.F., Conway, S.J., Ng, L.G., Stanley, E.R., et al. (2010). Fate Mapping Analysis Reveals That Adult Microglia Derive from Primitive Macrophages. Science 330, 841–845.

Gosselin, D., Skola, D., Coufal, N.G., Holtman, I.R., Schlachetzki, J.C.M., Sajti, E., Jaeger, B.N., O'Connor, C., Fitzpatrick, C., Pasillas, M.P., et al. (2017). An environment-dependent transcriptional network specifies human microglia identity. Science.

Grundberg, E., Small, K.S., Hedman, Å.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.-P., Meduri, E., Barrett, A., et al. (2012). Mapping cis - and trans -regulatory effects across multiple tissues in twins. Nat. Genet. 44, 1084–1089.

Gui, Y., Thomas, M.H., Garcia, P., Karout, M., Halder, R., Michelucci, A., Kollmus, H., Zhou, C., Melmed, S., Schughart, K., et al. (2020). Pituitary Tumor Transforming Gene 1 orchestrates gene regulatory variation in mouse ventral midbrain during aging. BioRxiv 2020.05.14.096156.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589.

Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., et al. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477, 289–294.

Kierdorf, K., Erny, D., Goldmann, T., Sander, V., Schulz, C., Perdiguero, E.G., Wieghofer, P., Heinrich, A., Riemke, P., Hölscher, C., et al. (2013). Microglia emerge from erythromyeloid precursors via Pu.1- and Irf8-dependent pathways. Nat. Neurosci. 16, 273–280.

Klein, C., and Westenberger, A. (2012). Genetics of Parkinson's Disease. Cold Spring Harb. Perspect. Med. 2, a008888.

Korotkova, T.M., Ponomarenko, A.A., Brown, R.E., and Haas, H.L. (2004). Functional diversity of ventral midbrain dopamine and GABAergic neurons. Mol. Neurobiol. 29, 243–259.

Lake, B.B., Chen, S., Sos, B.C., Fan, J., Kaeser, G.E., Yung, Y.C., Duong, T.E., Gao, D., Chun, J., Kharchenko, P.V., et al. (2018). Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat. Biotechnol. 36, 70–80.

Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D.A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., et al. (2014). Chromatin state dynamics during blood formation. Science 345, 943–949.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079.

Li, Z., Chen, J., Yu, H., He, L., Xu, Y., Zhang, D., Yi, Q., Li, C., Li, X., Shen, J., et al. (2017). Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. Nat. Genet. 49, 1576–1583.

Li, Z., Schulz, M.H., Look, T., Begemann, M., Zenke, M., and Costa, I.G. (2019). Identification of transcription factor binding sites using ATAC-seq. Genome Biol. 20, 45.

Liu, K., Pan, C., Kuhn, A., Nievergelt, A.P., Fantner, G.E., Milenkovic, O., and Radenovic, A. (2019a). Detecting topological variations of DNA at single-molecule level. Nat. Commun. 10, 3.

Liu, X., Li, Y.I., and Pritchard, J.K. (2019b). Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell 177, 1022-1034.e6.

Lodato, M.A., Ng, C.W., Wamstad, J.A., Cheng, A.W., Thai, K.K., Fraenkel, E., Jaenisch, R., and Boyer, L.A. (2013). SOX2 Co-Occupies Distal Enhancer Elements with Distinct POU Factors in ESCs and NPCs to Specify Cell State. PLOS Genet. 9, e1003288.

Lu, Q.R., Sun, T., Zhu, Z., Ma, N., Garcia, M., Stiles, C.D., and Rowitch, D.H. (2002). Common Developmental Requirement for Olig Function Indicates a Motor Neuron/Oligodendrocyte Connection. Cell 109, 75–86.

Mao, C.D., and Byers, S.W. (2011). Cell-Context Dependent TCF/LEF Expression and Function: Alternative Tales of Repression, De-Repression and Activation Potentials. Crit. Rev. Eukaryot. Gene Expr. 21, 207–236.

Massouras, A., Waszak, S.M., Albarca-Aguilera, M., Hens, K., Holcombe, W., Ayroles, J.F., Dermitzakis, E.T., Stone, E.A., Jensen, J.D., Mackay, T.F.C., et al. (2012). Genomic Variation and Its Impact on Gene Expression in Drosophila melanogaster. PLOS Genet. 8, e1003055.

Matcovitch-Natan, O., Winter, D.R., Giladi, A., Aguilar, S.V., Spinrad, A., Sarrazin, S., Ben-Yehuda, H., David, E., González, F.Z., Perrin, P., et al. (2016). Microglia development follows a stepwise program to regulate brain homeostasis. Science 353.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science 337, 1190–1195.

McGary, K.L., Park, T.J., Woods, J.O., Cha, H.J., Wallingford, J.B., and Marcotte, E.M. (2010). Systematic discovery of nonobvious human disease models through orthologous phenotypes. Proc. Natl. Acad. Sci. 107, 6544–6549.

McKay, D.J., and Lieb, J.D. (2013). A common set of DNA regulatory elements shapes Drosophila appendages. Dev. Cell 27.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. 28, 495–501.

Mei, F., Wang, H., Liu, S., Niu, J., Wang, L., He, Y., Etxeberria, A., Chan, J.R., and Xiao, L. (2013). Stage-Specific Deletion of Olig2 Conveys Opposing Functions on Differentiation and Maturation of Oligodendrocytes. J. Neurosci. 33, 8454–8462.

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

Moy, S.S., Nadler, J.J., Young, N.B., Perez, A., Holloway, L.P., Barbaro, R.P., Barbaro, J.R., Wilson, L.M., Threadgill, D.W., Lauder, J.M., et al. (2007). Mouse behavioral tasks relevant to autism: Phenotypes of 10 inbred strains. Behav. Brain Res. 176, 4–20.

Nalls, M.A., Blauwendraat, C., Vallerga, C.L., Heilbron, K., Bandres-Ciga, S., Chang, D., Tan, M., Kia, D.A., Noyce, A.J., Xue, A., et al. (2019). Identification of novel risk loci, causal

insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol. 18, 1091–1102.

Nott, A., Holtman, I.R., Coufal, N.G., Schlachetzki, J.C.M., Yu, M., Hu, R., Han, C.Z., Pena, M., Xiao, J., Wu, Y., et al. (2019). Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. Science 366, 1134–1139.

Nusse, R., and Clevers, H. (2017). Wnt/β-Catenin Signaling, Disease, and Emerging Therapeutic Modalities. Cell 169, 985–999.

Ohnmacht, J., May, P., Sinkkonen, L., and Krüger, R. (2020). Missing heritability in Parkinson's disease: the emerging role of non-coding genetic variation. J. Neural Transm.

Pálfy, M., Schulze, G., Valen, E., and Vastenhouw, N.L. (2020). Chromatin accessibility established by Pou5f3, Sox19b and Nanog primes genes for activity during zebrafish genome activation. PLOS Genet. 16, e1008546.

Pierce, S.E., Tyson, T., Booms, A., Prahl, J., and Coetzee, G.A. (2018). Parkinson's disease genetic risk in a midbrain neuronal cell line. Neurobiol. Dis. 114, 53–64.

Ponder, C.A., Kliethermes, C.L., Drew, M.R., Muller, J., Das, K., Risbrough, V.B., Crabbe, J.C., Gilliam, T.C., and Palmer, A.A. (2007). Selection for contextual fear conditioning affects anxiety-like behaviors and gene expression. Genes Brain Behav. 6, 736–749.

Portugal, G.S., Wilkinson, D.S., Kenney, J.W., Sullivan, C., and Gould, T.J. (2012). Strain-dependent Effects of Acute, Chronic, and Withdrawal from Chronic Nicotine on Fear Conditioning. Behav. Genet. 42, 133–150.

Preissl, S., Fang, R., Huang, H., Zhao, Y., Raviram, R., Gorkin, D.U., Zhang, Y., Sos, B.C., Afzal, V., Dickel, D.E., et al. (2018). Single-nucleus analysis of accessible chromatin in

developing mouse forebrain reveals cell-type-specific transcriptional regulation. Nat. Neurosci. 21, 432–439.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470, 279–283.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. 33, 495–502.

Saunders, A., Macosko, E.Z., Wysoker, A., Goldman, M., Krienen, F.M., de Rivera, H., Bien, E., Baum, M., Bortolin, L., Wang, S., et al. (2018). Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. Cell 174, 1015-1030.e16.

Schubert, M., Ermini, L., Sarkissian, C.D., Jónsson, H., Ginolhac, A., Schaefer, R., Martin, M.D., Fernández, R., Kircher, M., McCue, M., et al. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat. Protoc. 9, 1056–1082.

Siersbæk, R., Madsen, J.G.S., Javierre, B.M., Nielsen, R., Bagge, E.K., Cairns, J., Wingett, S.W., Traynor, S., Spivakov, M., Fraser, P., et al. (2017). Dynamic Rewiring of Promoter-Anchored Chromatin Loops during Adipocyte Differentiation. Mol. Cell 66, 420-435.e5.

Sinnamon, J.R., Torkenczy, K.A., Linhoff, M.W., Vitak, S.A., Mulqueen, R.M., Pliner, H.A., Trapnell, C., Steemers, F.J., Mandel, G., and Adey, A.C. (2019). The accessible chromatin landscape of the murine hippocampus at single-cell resolution. Genome Res. 29, 857–869.

Stolt, C.C., Rehberg, S., Ader, M., Lommes, P., Riethmacher, D., Schachner, M., Bartsch, U., and Wegner, M. (2002). Terminal differentiation of myelin-forming oligodendrocytes depends on the transcription factor Sox10. Genes Dev. 16, 165–170.

Stolt, C.C., Lommes, P., Sock, E., Chaboissier, M.-C., Schedl, A., and Wegner, M. (2003). The Sox9 transcription factor determines glial fate choice in the developing spinal cord. Genes Dev. 17, 1677–1689.

Storey, J.D., Madeoy, J., Strout, J.L., Wurfel, M., Ronald, J., and Akey, J.M. (2007). Gene-Expression Variation Within and Among Human Populations. Am. J. Hum. Genet. 80, 502–509.

Sugiaman-Trapman, D., Vitezic, M., Jouhilahti, E.-M., Mathelier, A., Lauter, G., Misra, S., Daub, C.O., Kere, J., and Swoboda, P. (2018). Characterization of the human RFX transcription factor family by regulatory and target gene analysis. BMC Genomics 19, 181.

Thifault, S., Lalonde, R., Sanon, N., and Hamet, P. (2002). Comparisons between C57BL/6J and A/J mice in motor activity and coordination, hole-poking, and spatial learning. Brain Res. Bull. 58, 213–218.

Varrette S, Bouvry P, Cartiaux H, and Georgatos F (2014). International Conference on High Performance Computing & Simulation (HPCS). pp. 959–967.

Vogt Weisenhorn, D.M., Giesert, F., and Wurst, W. (2016). Diversity matters – heterogeneity of dopaminergic neurons in the ventral mesencephalon and its relation to Parkinson's Disease. J. Neurochem. 139, 8–26.

Wilczynska, K.M., Singh, S.K., Adams, B., Bryan, L., Rao, R.R., Valerie, K., Wright, S., Griswold-Prenner, I., and Kordula, T. (2009). Nuclear Factor I Isoforms Regulate Gene

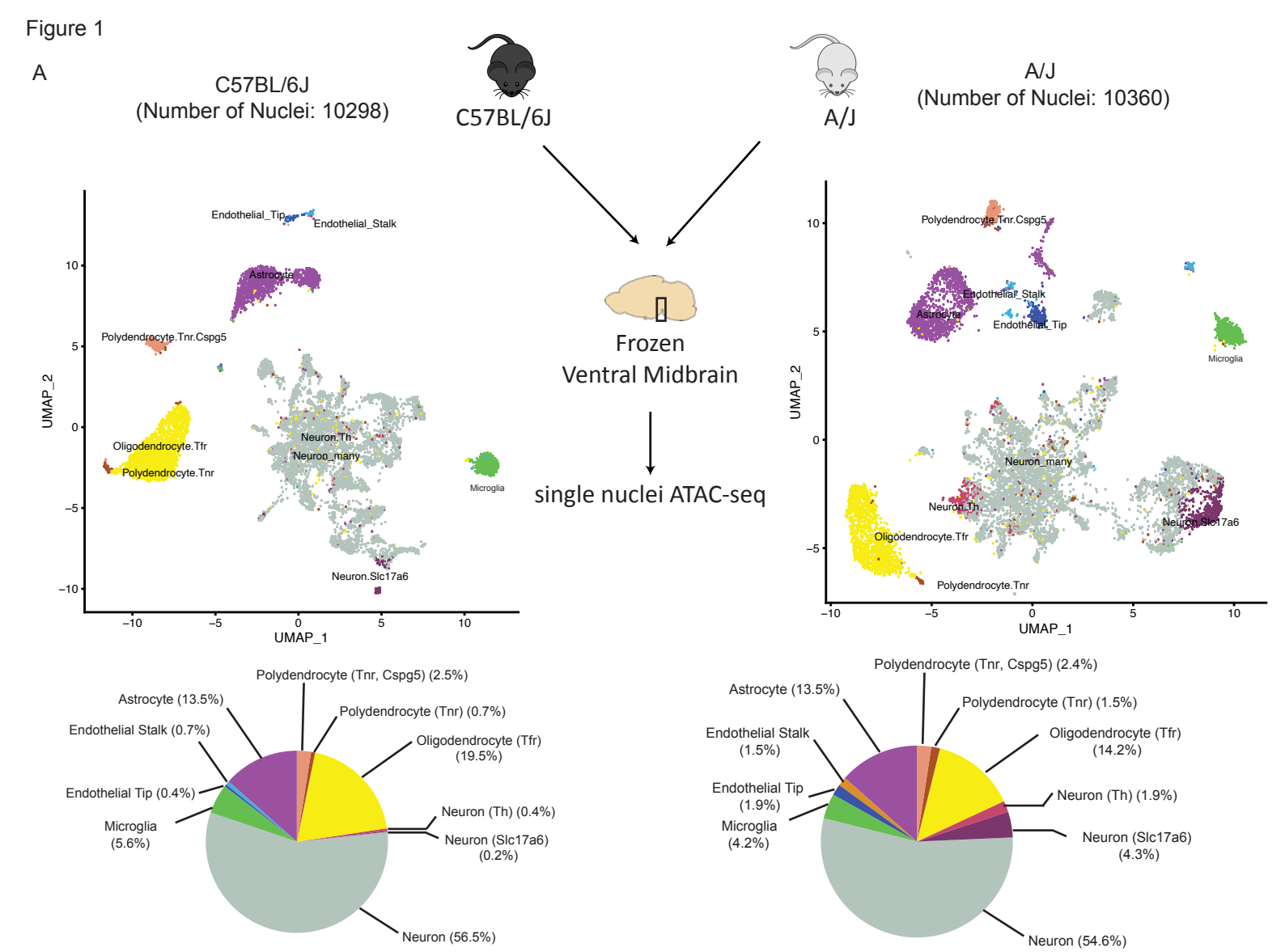
Expression During the Differentiation of Human Neural Progenitors to Astrocytes. STEM CELLS 27, 1173–1181.

Williams, H.J., Owen, M.J., and O'Donovan, M.C. (2009). New findings from genetic association studies of schizophrenia. J. Hum. Genet. 54, 9–14.

Wu, T.P., Wang, T., Seetin, M.G., Lai, Y., Zhu, S., Lin, K., Liu, Y., Byrum, S.D., Mackintosh, S.G., Zhong, M., et al. (2016). DNA methylation on N 6 -adenine in mammalian embryonic stem cells. Nature 532, 329–333.

Yoneyama, N., Crabbe, J.C., Ford, M.M., Murillo, A., and Finn, D.A. (2008). Voluntary ethanol consumption in 22 inbred mouse strains. Alcohol 42, 149–160.

Figure 1


A
C57BL/6J
(Number of Nuclei: 10298)
C57BL/6J
A/J
A/J
(Number of Nuclei: 10360)
Frozen
Ventral Midbrain
single nuclei ATAC-seq
Endothelial_Tip
Endothelial_Stalk
Astrocyte
Polydendrocyte.Tnr.Cspg5
Oligodendrocyte.Tfr
Polydendrocyte.Tnr
Neuron.Th
Neuron_many
Microglia
Neuron.Slc17a6
UMAP_1
UMAP_2
Polydendrocyte (Tnr, Cspg5) (2.5%)
Astrocyte (13.5%)
Polydendrocyte (Tnr) (0.7%)
Endothelial Stalk (0.7%)
Oligodendrocyte (Tfr)
(19.5%)
Endothelial Tip (0.4%)
Neuron (Th) (0.4%)
Microglia
(5.6%)
Neuron (Slc17a6)
(0.2%)
Neuron (56.5%)
Polydendrocyte (Tnr, Cspg5) (2.4%)
Astrocyte (13.5%)
Polydendrocyte (Tnr) (1.5%)
Endothelial Stalk
(1.5%)
Oligodendrocyte (Tfr)
(14.2%)
Endothelial Tip
(1.9%)
Neuron (Th) (1.9%)
Microglia
(4.2%)
Neuron (Slc17a6)
(4.3%)
Neuron (54.6%)

B

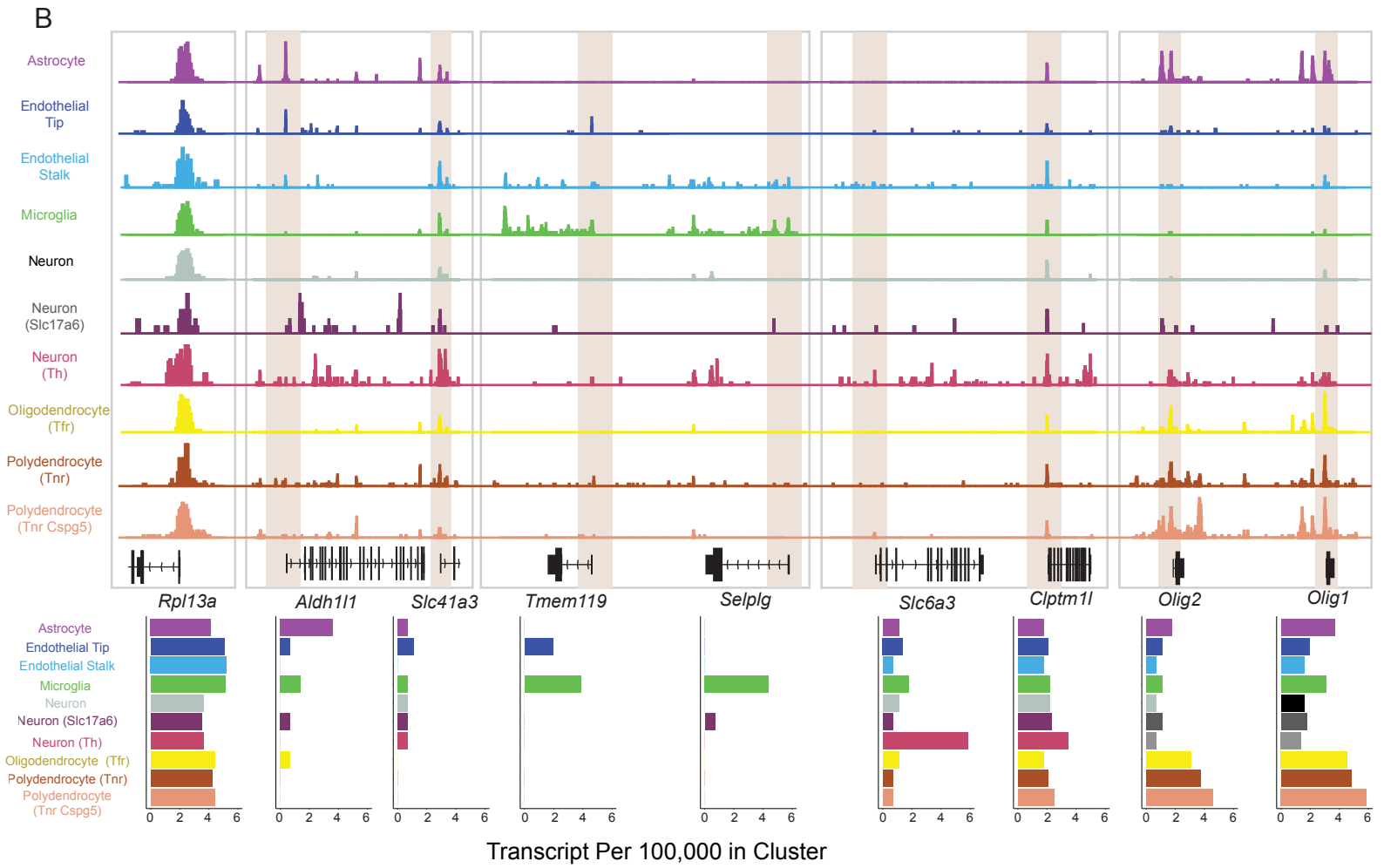
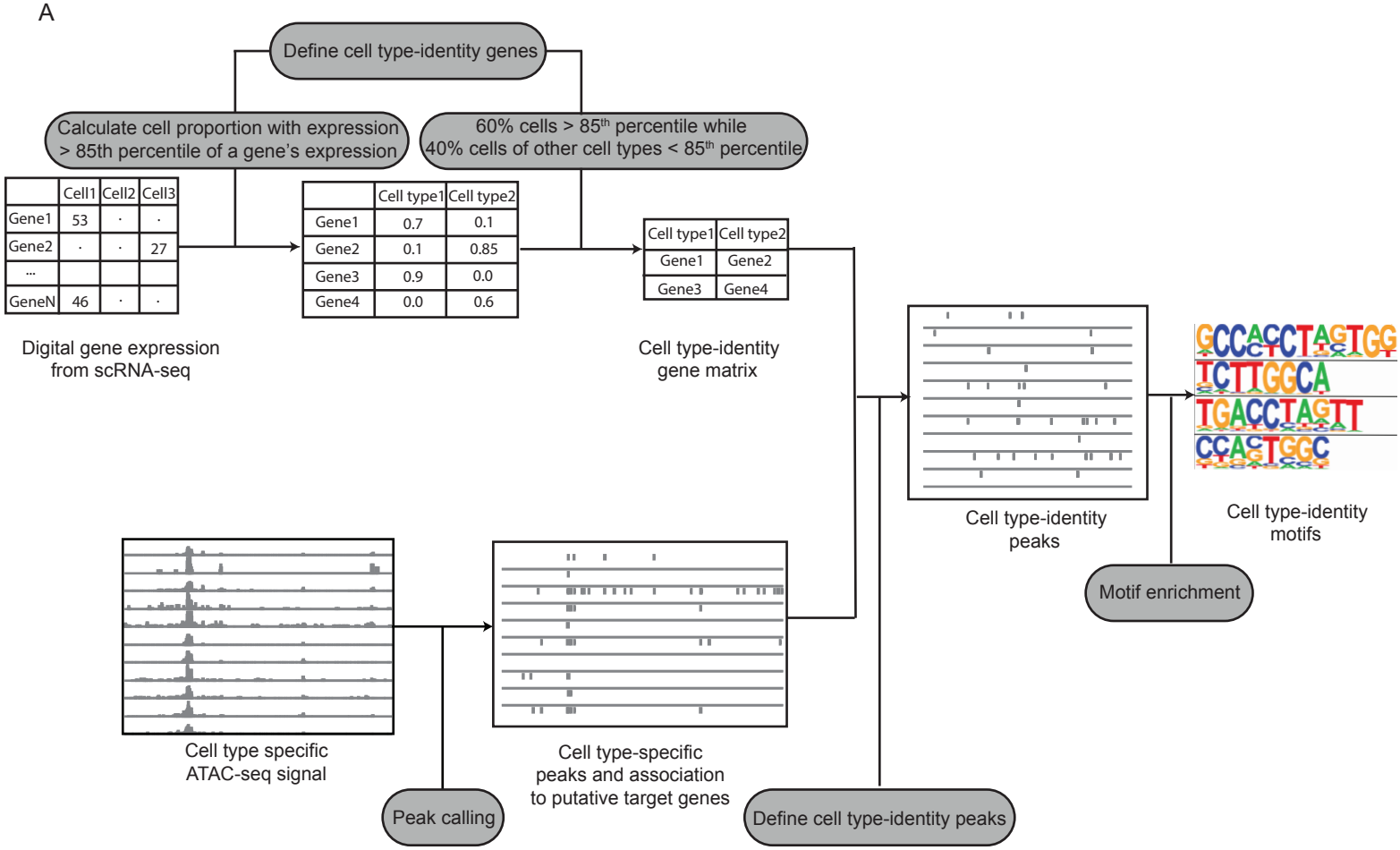

Astrocyte
Endothelial Tip
Endothelial Stalk
Microglia
Neuron
Neuron (Slc17a6)
Neuron (Th)
Oligodendrocyte (Tfr)
Polydendrocyte (Tnr)
Polydendrocyte (Tnr Cspg5)
Rpl13a
Aldh1l1
Slc41a3
Tmem119
Selplg
Slc6a3
Clptm1l
Olig2
Olig1
Transcript Per 100,000 in Cluster

Figure 2

A

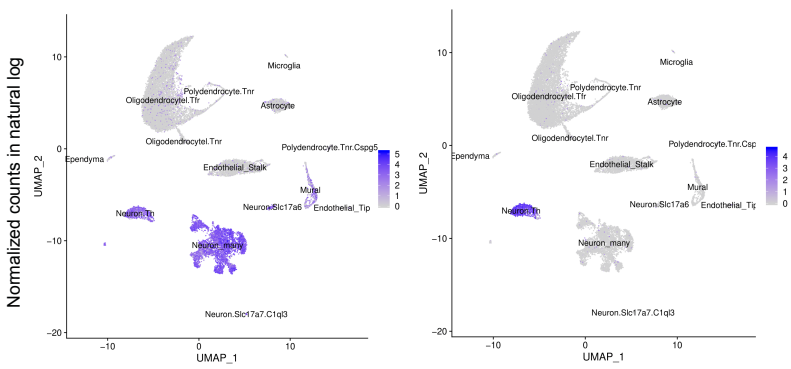


	Cell1	Cell2	Cell3
Gene1	53	·	·
Gene2	·	·	27
...			
GeneN	46	·	·

	Cell type1	Cell type2
Gene1	0.7	0.1
Gene2	0.1	0.85
Gene3	0.9	0.0
Gene4	0.0	0.6

Cell type1	Cell type2
Gene1	Gene2
Gene3	Gene4

B



Cell type	# of identity genes	Top enriched GO term
Astrocyte	100	Negative regulation of neuron differentiation
Endothelial Tip	83	Extracellular matrix organization
Endothelial Stalk	153	Positive regulation of angiogenesis
Microglia	79	Neutrophil mediated immunity
Neuron	194	neurotransmitter transport
Neuron (Slc17a6)	187	Axonogenesis
Neuron (Th)	412	Regulation of synaptic vesicle exocytosis
Oligodendrocyte (Tfr)	168	Septin ring assembly
Polydendrocyte (Tnr)	284	Myelination
Polydendrocyte (Tnr Cspg5)	47	Aminoglycan metabolic process

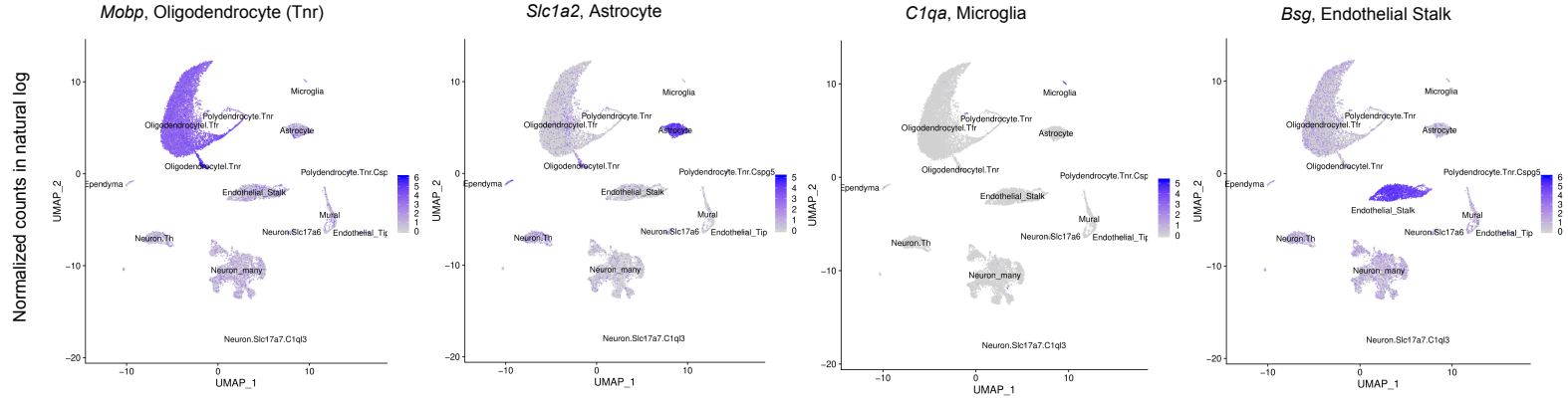
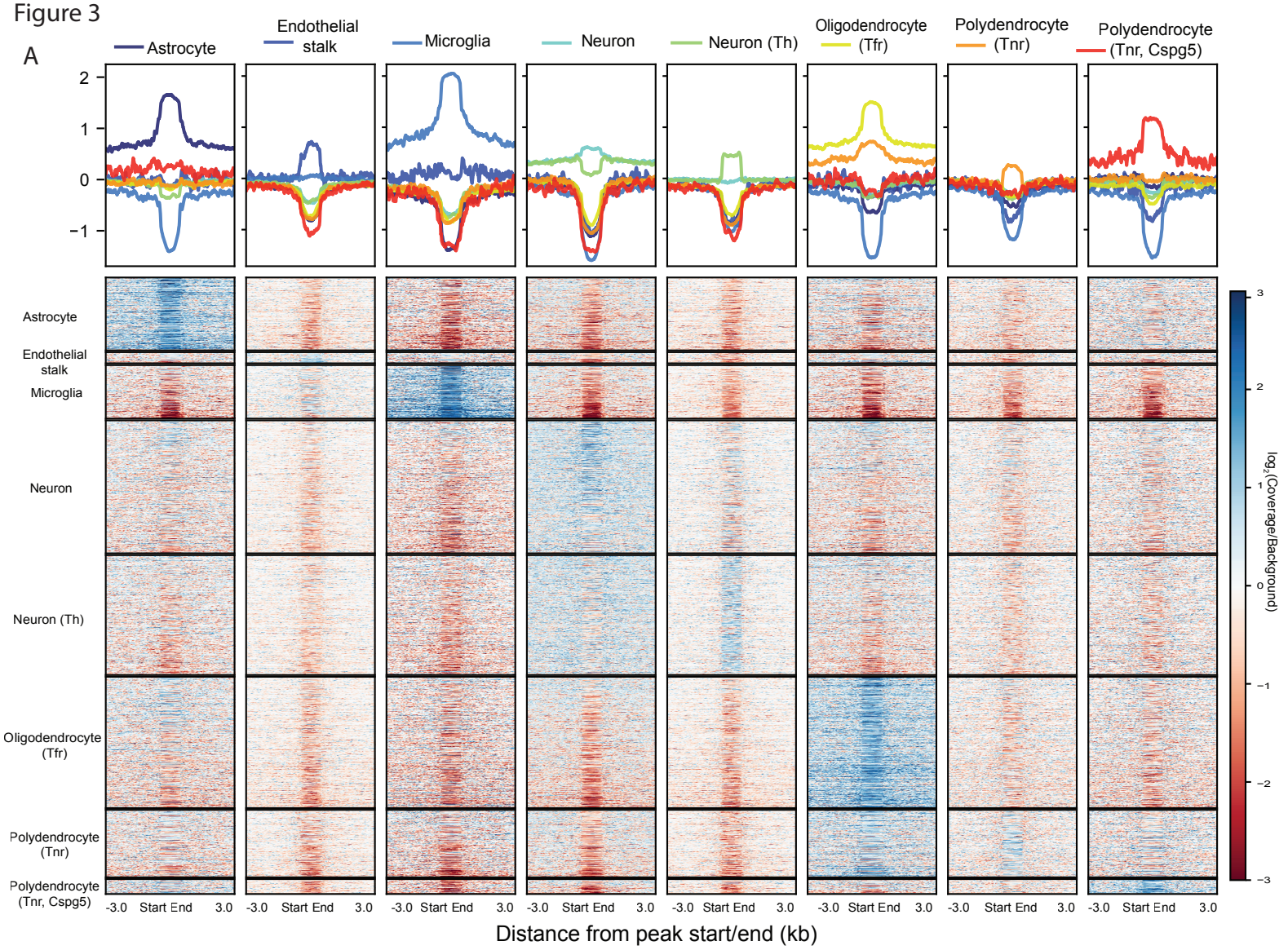


Figure 3


A
Astrocyte
Endothelial stalk
Microglia
Neuron
Neuron (Th)
Oligodendrocyte (Tfr)
Polydendrocyte (Tnr)
Polydendrocyte (Tnr, Cspg5)
2
1
0
−1
Astrocyte
Endothelial stalk
Microglia
Neuron
Neuron (Th)
Oligodendrocyte (Tfr)
Polydendrocyte (Tnr)
Polydendrocyte (Tnr, Cspg5)
-3.0 Start End 3.0
log2(Coverage/Background)
3
2
1
0
−1
−2
−3
Distance from peak start/end (kb)

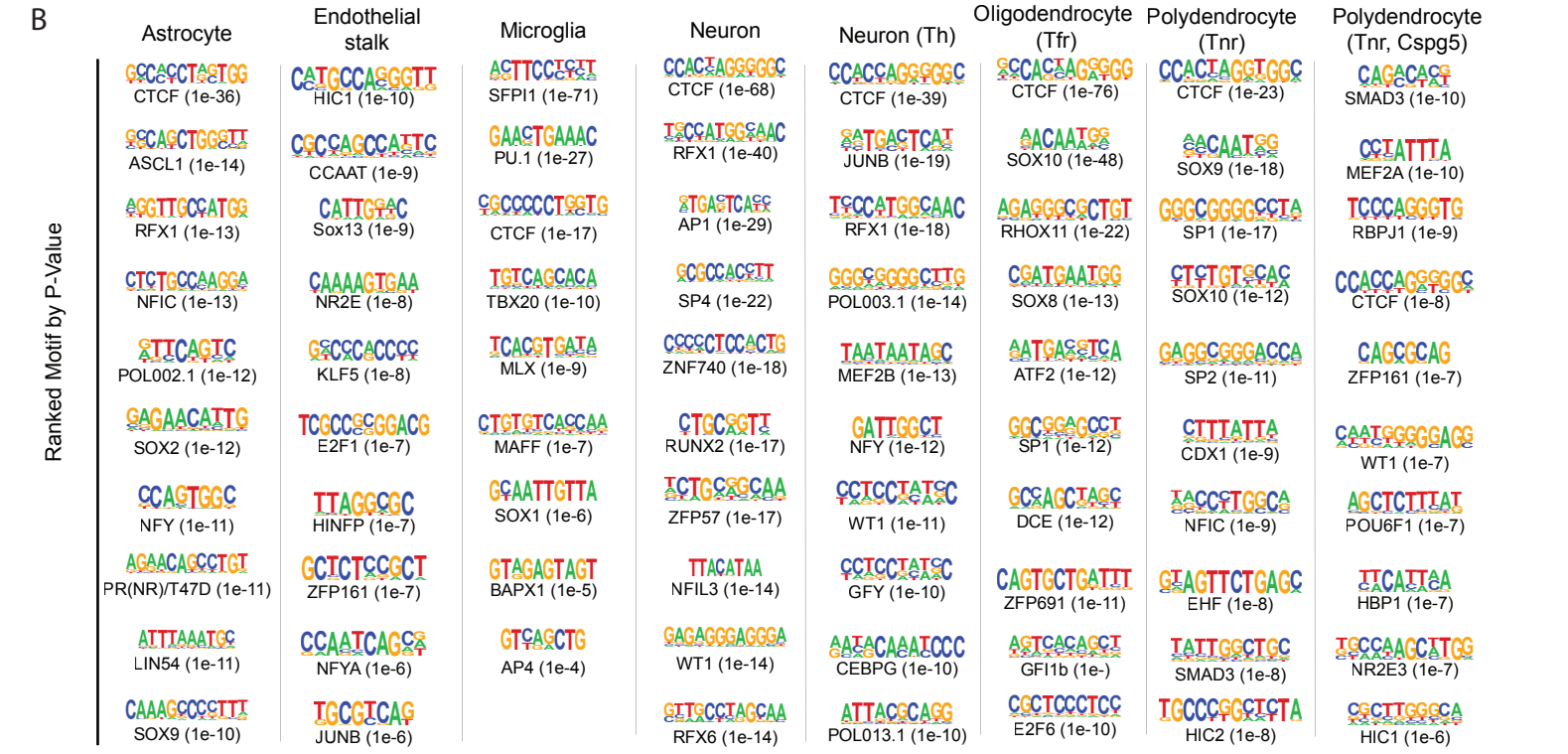
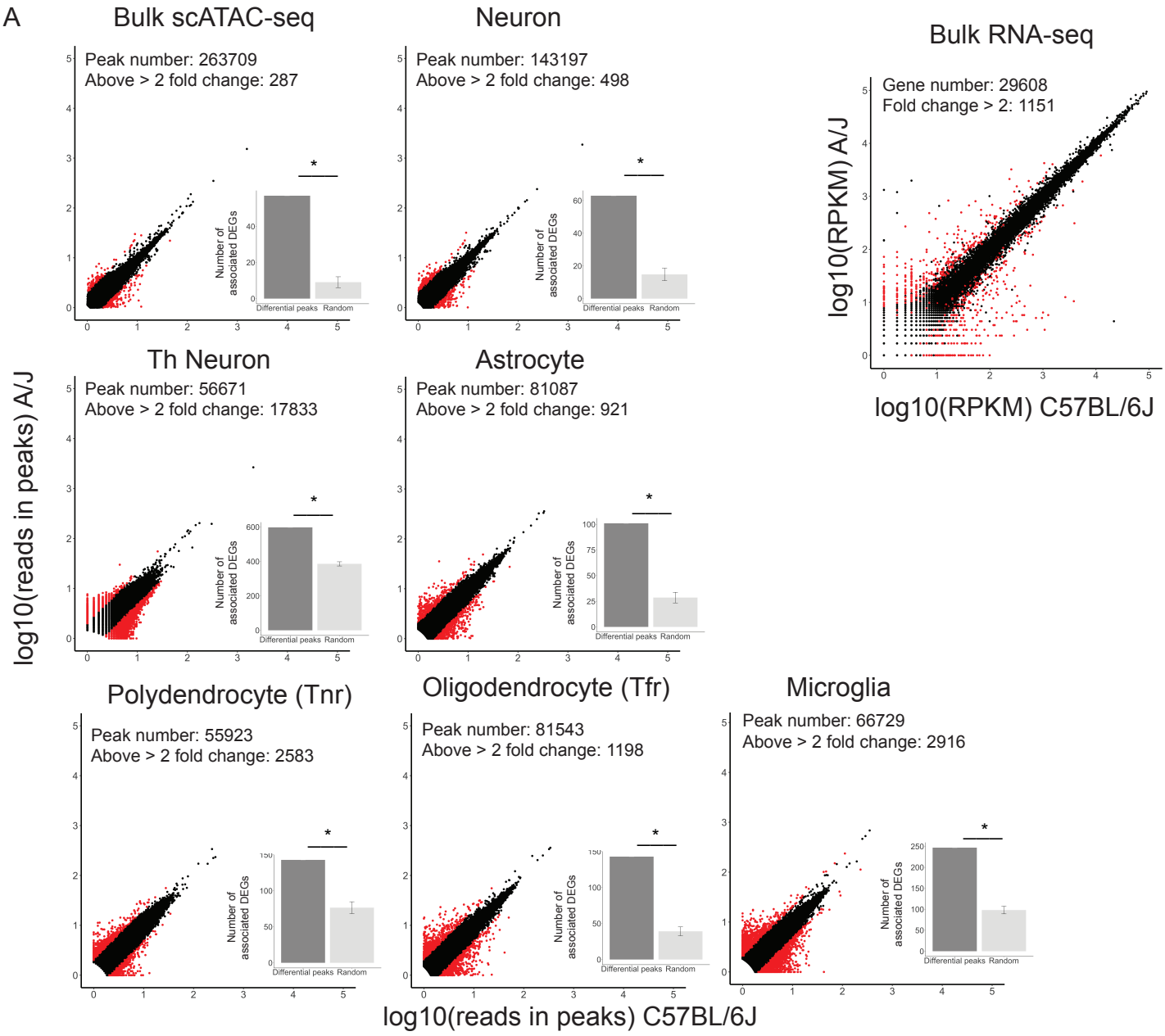

B
Ranked Motif by P-Value
Astrocyte
CTCF (1e-36)
ASCL1 (1e-14)
RFX1 (1e-13)
NFIC (1e-13)
POL002.1 (1e-12)
SOX2 (1e-12)
NFY (1e-11)
PR(NR)/T47D (1e-11)
LIN54 (1e-11)
SOX9 (1e-10)
Endothelial stalk
HIC1 (1e-10)
CCAAT (1e-9)
Sox13 (1e-9)
NR2E (1e-8)
KLF5 (1e-8)
E2F1 (1e-7)
HINFP (1e-7)
ZFP161 (1e-7)
NFYA (1e-6)
JUNB (1e-6)
Microglia
SFPI1 (1e-71)
PU.1 (1e-27)
CTCF (1e-17)
TBX20 (1e-10)
MLX (1e-9)
MAFF (1e-7)
SOX1 (1e-6)
BAPX1 (1e-5)
AP4 (1e-4)
Neuron
CTCF (1e-68)
RFX1 (1e-40)
AP1 (1e-29)
SP4 (1e-22)
ZNF740 (1e-18)
RUNX2 (1e-17)
ZFP57 (1e-17)
NFIL3 (1e-14)
WT1 (1e-14)
RFX6 (1e-14)
Neuron (Th)
CTCF (1e-39)
JUNB (1e-19)
RFX1 (1e-18)
POL003.1 (1e-14)
MEF2B (1e-13)
NFY (1e-12)
WT1 (1e-11)
GFY (1e-10)
CEBPG (1e-10)
POL013.1 (1e-10)
Oligodendrocyte (Tfr)
CTCF (1e-76)
SOX10 (1e-48)
RHOX11 (1e-22)
SOX8 (1e-13)
ATF2 (1e-12)
SP1 (1e-12)
DCE (1e-12)
ZFP691 (1e-11)
GFI1b (1e-)
E2F6 (1e-10)
Polydendrocyte (Tnr)
CTCF (1e-23)
SOX9 (1e-18)
SP1 (1e-17)
SOX10 (1e-12)
SP2 (1e-11)
CDX1 (1e-9)
NFIC (1e-9)
EHF (1e-8)
SMAD3 (1e-8)
HIC2 (1e-8)
Polydendrocyte (Tnr, Cspg5)
SMAD3 (1e-10)
MEF2A (1e-10)
RBPJ1 (1e-9)
CTCF (1e-8)
ZFP161 (1e-7)
WT1 (1e-7)
POU6F1 (1e-7)
HBP1 (1e-7)
NR2E3 (1e-7)
HIC1 (1e-6)

Figure 4


A
Bulk scATAC-seq
Peak number: 263709
Above > 2 fold change: 287
Neuron
Peak number: 143197
Above > 2 fold change: 498
Bulk RNA-seq
Gene number: 29608
Fold change > 2: 1151
log10(RPKM) A/J
log10(RPKM) C57BL/6J
Th Neuron
Peak number: 56671
Above > 2 fold change: 17833
Astrocyte
Peak number: 81087
Above > 2 fold change: 921
log10(reads in peaks) A/J
Polydendrocyte (Tnr)
Peak number: 55923
Above > 2 fold change: 2583
Oligodendrocyte (Tfr)
Peak number: 81543
Above > 2 fold change: 1198
Microglia
Peak number: 66729
Above > 2 fold change: 2916
Number of associated DEGs
Differential peaks
Random
log10(reads in peaks) C57BL/6J

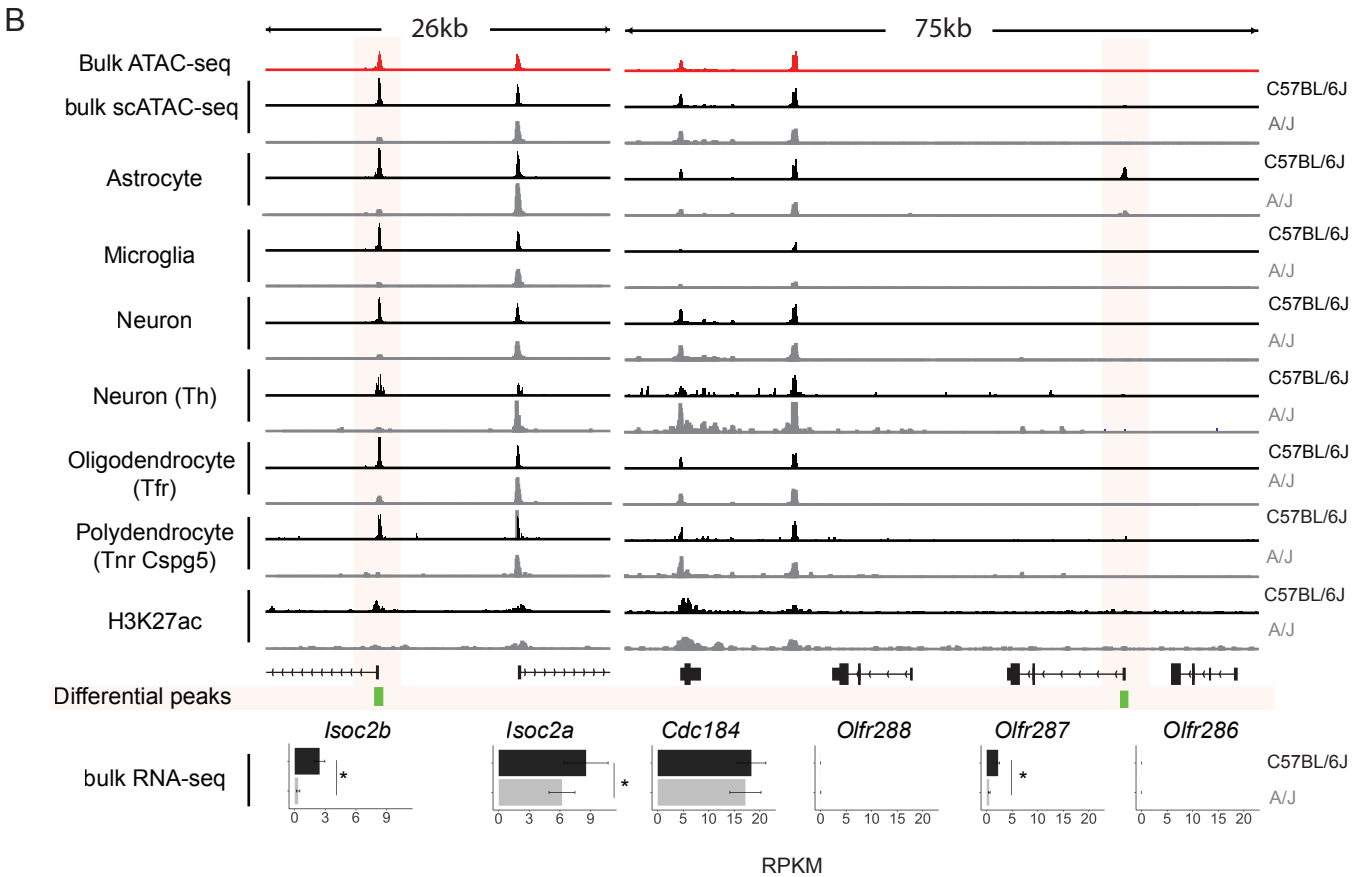

B
26kb
75kb
Bulk ATAC-seq
bulk scATAC-seq
Astrocyte
Microglia
Neuron
Neuron (Th)
Oligodendrocyte (Tfr)
Polydendrocyte (Tnr Cspg5)
H3K27ac
C57BL/6J
A/J
Differential peaks
bulk RNA-seq
Isoc2b
Isoc2a
Cdc184
Olfr288
Olfr287
Olfr286
RPKM

Figure 5

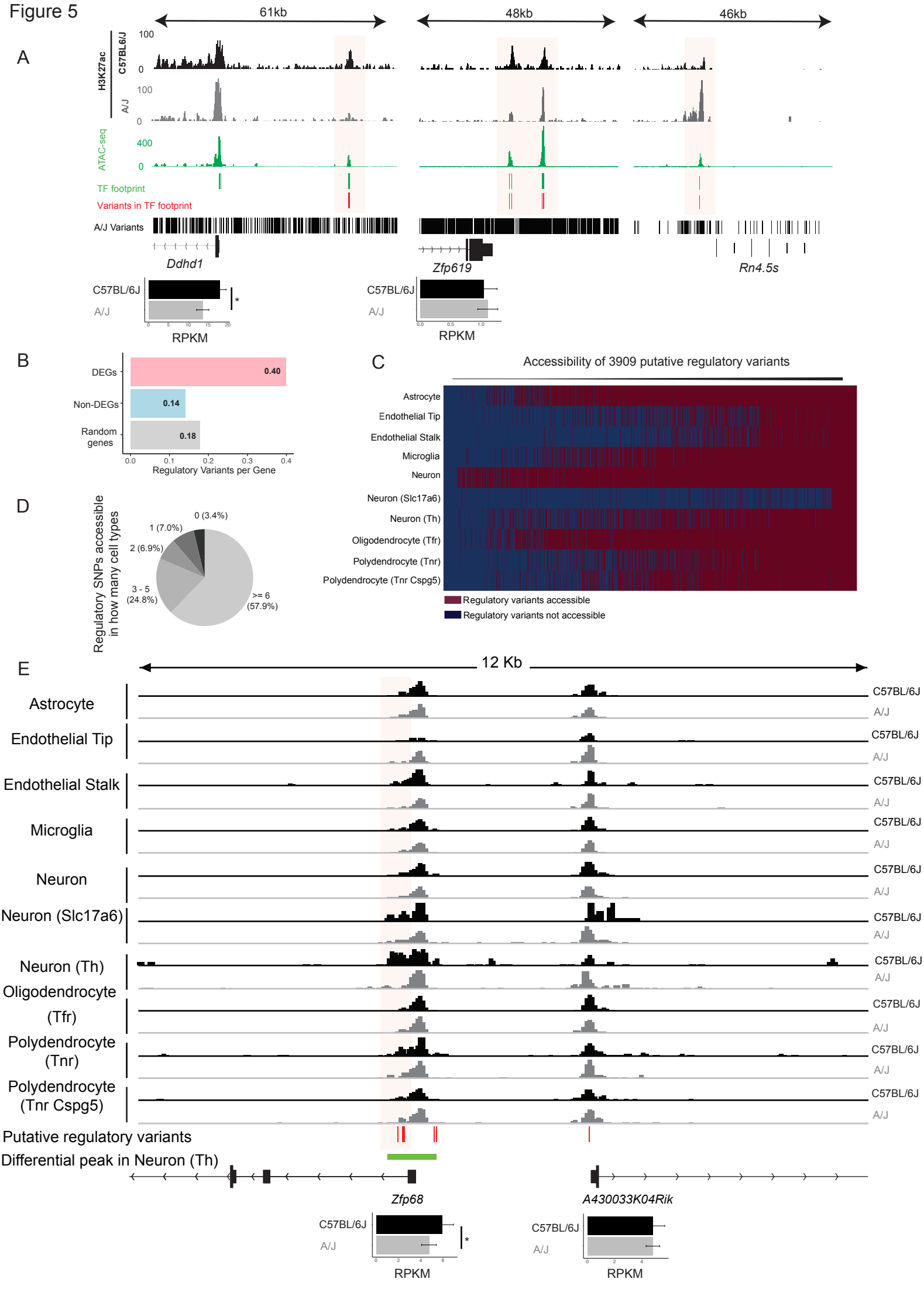
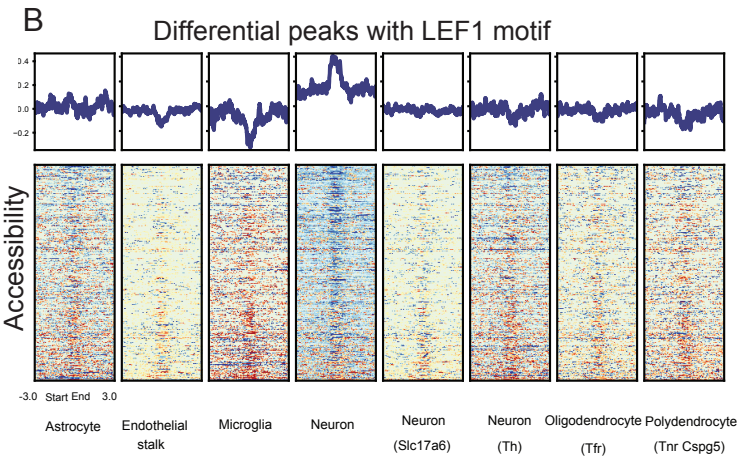

A
61kb
48kb
46kb
H3K27ac
C57BL6J
A/J
ATAC-seq
TF footprint
Variants in TF footprint
A/J Variants
Ddhd1
Zfp619
Rn4.5s
C57BL/6J
A/J
RPKM
B
DEGs
Non-DEGs
Random genes
0.40
0.14
0.18
Regulatory Variants per Gene
C
Accessibility of 3909 putative regulatory variants
Astrocyte
Endothelial Tip
Endothelial Stalk
Microglia
Neuron
Neuron (Slc17a6)
Neuron (Th)
Oligodendrocyte (Tfr)
Polydendrocyte (Tnr)
Polydendrocyte (Tnr Cspg5)
Regulatory variants accessible
Regulatory variants not accessible
D
Regulatory SNPs accessible in how many cell types
0 (3.4%)
1 (7.0%)
2 (6.9%)
3 - 5 (24.8%)
>= 6 (57.9%)
E
12 Kb
Putative regulatory variants
Differential peak in Neuron (Th)
Zfp68
A430033K04Rik

Figure 6

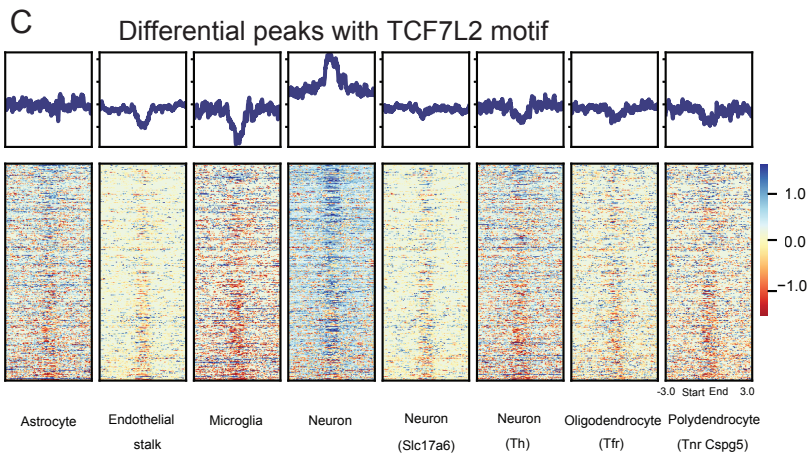
A

	TF	Motif logo	p-value	% of Targets
Ranked by p-value	LEF1		1e-35	34.82%
	FOXO1		1e-18	20.24%
	ZFP691		1e-16	1.65%
	MED1		1e-15	1.41%
	TCF7L2		1e-14	34.12%

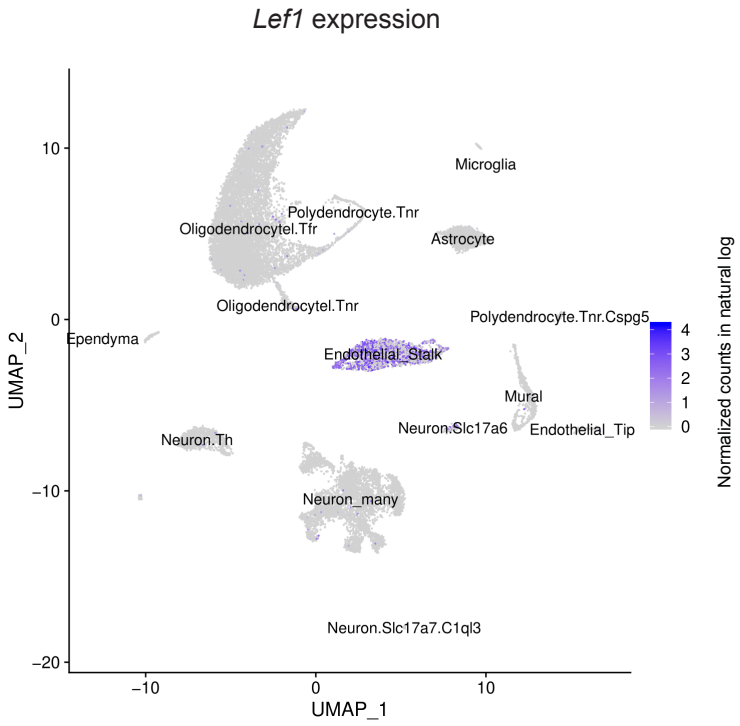
B



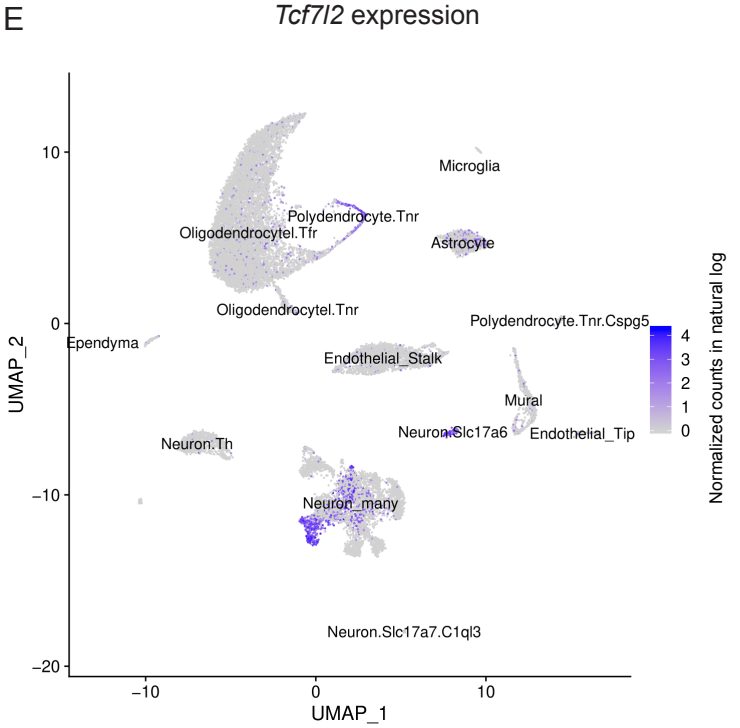
C



D



E



Single nuclei chromatin profiles of midbrain from genetically distinct mouse strains reveal cell identity transcription factors and cell type-specific gene regulatory variation

Yujuan Gui[1], Kamil Grzyb[2], Mélanie H. Thomas[2], Jochen Ohnmacht[1,2], Pierre Garcia[2], Rashi Halder[2], Manuel Buttini[2], Alexander Skupin[2], Thomas Sauter[1], Lasse Sinkkonen[1*]

Supplementary Information

Supplementary Figures

Supplementary Figure S1: snATAC-seq on ventral midbrains of C57BL/6J and A/J revealed cell type-specific chromatin accessibility.

A. The ventral midbrains of the two mouse strains were used as input to snATAC-seq. Ten cell types were identified based on clustering of peak features. The aggregated signal is comparable to bulk ATAC-seq. Different accessibility across cell types can be observed and they can reflect gene expression changes. *Prrrc2a* is globally accessible and its expression can be detected in all cell types. *Aif1* and *Lst1* TSS are selectively accessible in macrophage, and their expression is only abundant in macrophage.

Supplementary Figure S2: H3K27ac ChIP-seq and ATAC-seq correlating with gene expression.

A. H3K27ac ChIP-seq on ventral midbrains of C57BL/6J and A/J. Within-sample normalization is applied to account for gene length. The intensity of H3K27ac ChIP-seq signals are plotted in a window of 2000 bp upstream and downstream of gene body. The genes are ordered based on the highest to the lowest gene expression level.

B. ATAC-seq on ventral midbrain of C57BL/6J. The plotting scheme is the same as Supplementary Figure 2A.

Supplementary Figure S3: Differential peaks can reveal strain-specific TFs.

A. Motif enrichment analysis on differential peaks. The PWM logos, names of the associated TFs and p-values are shown for each motif. The motifs are ranked according to p-values.

Supplementary Tables

Supplementary Table S1: Cell type composition in ventral midbrains of C57BL/6J and A/J in snATAC-seq.

Supplementary table 2: Cell type-identity genes defined from existing scRNA-seq.

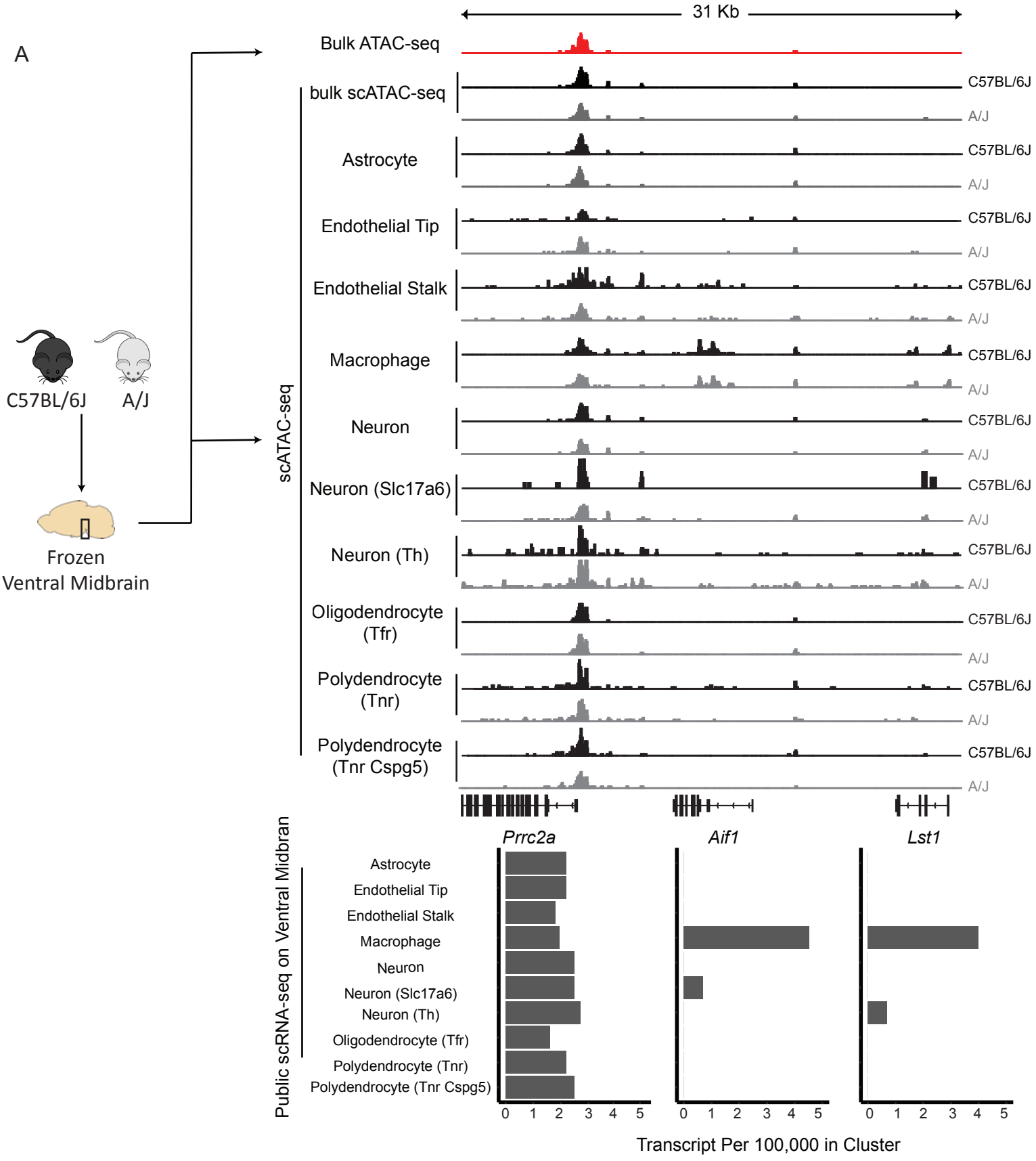
Supplementary table 3: Enrichment analysis on the cell type-identity genes defined from existing scRNA-seq. The GO enrichment analysis was performed by Enrichr.

Supplementary table 4: Cell type-identity peaks defined by associating cell type-specific peaks to the regulatory regions of cell type-identity genes.

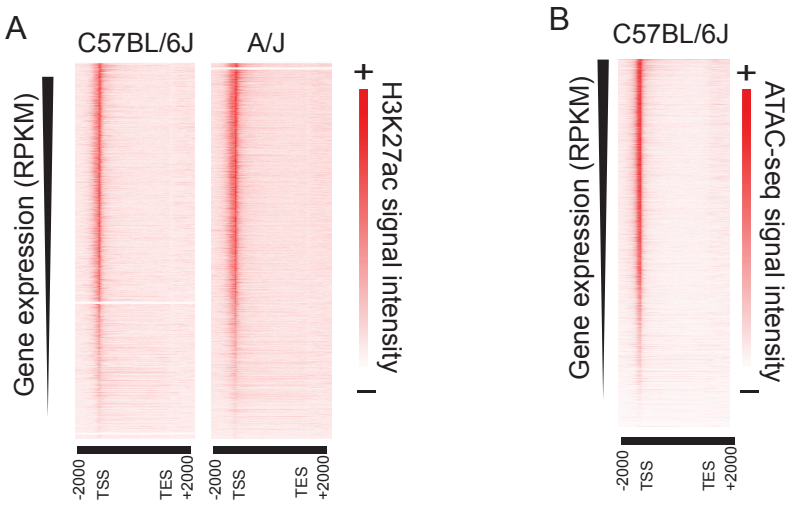
Supplementary table 5: Putative regulatory variants. The location and major / alternative alleles are reported for each variant.

Supplementary table 6: H3K27ac differential peaks between C57BL/6J and A/J.

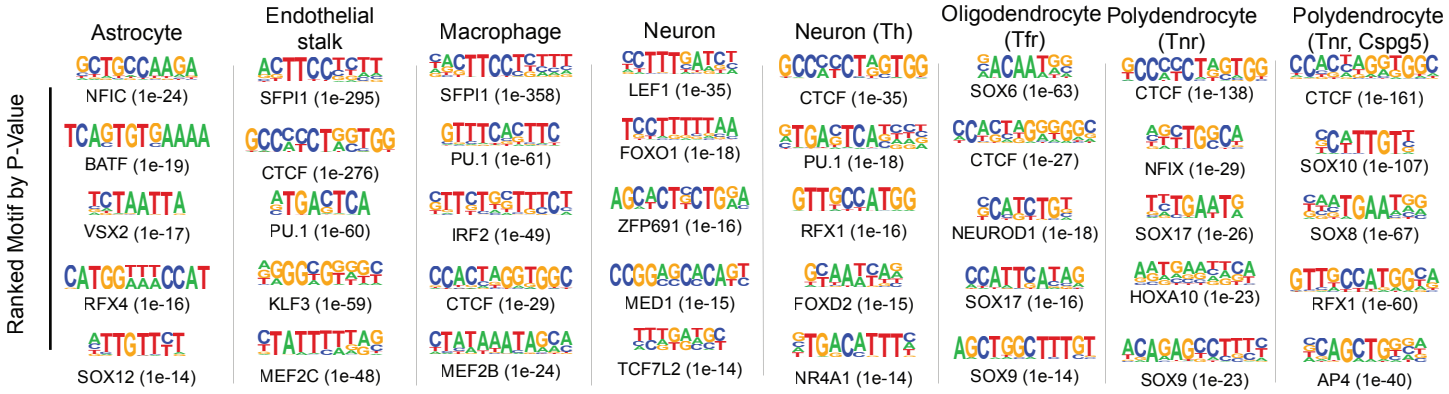
Supplementary figure 1



Supplementary figure 2



Supplementary figure 3



4.3 Manuscript 3

Quantitative trait locus mapping identifies *Col4a6* as a novel regulator of striatal dopamine level and axonal branching in mice

4.3.1 Preface

Substantial difference in ventral midbrain transcriptome was observed in inbreed mouse strains. Such difference in gene expression is likely to affect phenotypes associated with this brain region. DAns, establishing the nigrostriatal circuit by bridging SN in ventral midbrain to dorsal striatum to release dopamine, control motor function and its demise is a hallmark in Parkinson's disease. Dopamine level in dorsal striatum likely reflects the integrity of these neurons. Dopamine measurement on dorsal striation from eight inbred strains found varied dopamine levels across strains, suggesting it is a complex trait likely affected by genetic variation in ventral midbrain. To identify which genetic variants can contribute to dopamine level difference, we did QTL mapping with 32 CC strains and found a significant QTL on chromosome X. Because *Col4a6* was the most differentially expressed gene in tissue level RNA-seq, we proposed it is the QTL gene affecting dopamine level. Interestingly, *Col4a6* was shown to regulate axon branching during development, which could be linked to our observation that A/J has less DAn branching in dorsal striatum. To move from gene expression difference to phenotypic variation in ventral midbrain, we identified *Col4a6* could potentially affect DAn axogenesis during development and ultimately introduce variation in dopamine level in adulthood.

Computational analysis was performed by me; animal dissection was performed by Dr. Manuel Buttini and Dr. Pierre Garcia; dopamine measurement was performed by Dr. Christian Jaeger and Dr. Zdenka Hodak; tissue sectioning, staining, and quantification was performed by Dr. Mélanie Thomas and Dr. Mona Karout; data submission to GeneNetwork was done by Msc. Arthur Centeno.

4.3.2 Manuscript

Quantitative trait locus mapping identifies *Col4a6* as a novel regulator of striatal dopamine level and axonal branching in mice

Mélanie H. Thomas[1#], Yujuan Gui[2#], Pierre Garcia[1,3,4], Mona Karout[1], Christian Jaeger[1], Zdenka Hodak[1], Alessandro Michelucci[1,5], Heike Kollmus[6], Arthur Centeno[7], Klaus Schughart[6,8,9], Rudi Balling[1], Michel Mittelbronn[1,3,4,5], Joseph H. Nadeau[10,11], Robert W. Williams[7], Thomas Sauter[2], Lasse Sinkkonen[2*], Manuel Buttini[1*]

[1]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Belvaux, Luxembourg

[2]Department of Life Sciences and Medicine (DLSM), University of Luxembourg, Belvaux, Luxembourg

[3]National Center of Pathology (NCP), Laboratoire National de Santé (LNS), Dudelange, Luxembourg

[4]Luxembourg Centre of Neuropathology (LCNP), Luxembourg.

[5] Neuro-Immunology Group, Department of Oncology (DONC), Luxembourg Institute of Health (LIH), Luxembourg, Luxembourg

[6]Department of Infection Genetics, Helmholtz Centre for Infection Research, Braunschweig, Germany

[7]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, USA

[8]University of Veterinary Medicine Hannover, Hannover, Germany

[9]Department of Microbiology, Immunology and Biochemistry, University of Tennessee Health Science Center, Memphis, Tennessee, USA

[10]Pacific Northwest Research Institute, Seattle, Washington, United States

[11]Maine Medical Center Research Institute, Scarborough, Maine USA

[*]**Corresponding authors:** Drs. Lasse Sinkkonen (lasse.sinkkonen@uni.lu), Manuel Buttini (manuel.buttini@uni.lu)

#These authors contributed equally

Key words: Col4a6 – mouse strains – dopamine – QTL – regulatory variants – nigrostriatal circuit

Quantitative trait locus mapping identifies *Col4a6* as a novel regulator of striatal dopamine level and axonal branching in mice

Mélanie H. Thomas[1,4#], Yujuan Gui[2#], Pierre Garcia[1,3,4], Mona Karout[1], Christian Jaeger[1], Alessandro Michelucci[1,5], Heike Kollmus[6], Arthur Centeno[7], Klaus Schughart[6,8,9], Rudi Balling[1], Michel Mittelbronn[1,3,4,5], Joseph H. Nadeau[10,11], Robert W. Williams[7], Thomas Sauter[2], Lasse Sinkkonen[2*], Manuel Buttini[1, 4*]

[1]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch/Alzette, Luxembourg; [2]Department of Life Sciences and Medicine (DLSM), University of Luxembourg, Belvaux, Luxembourg; [3]National Center of Pathology (NCP), Laboratoire National de Santé (LNS), Dudelange, Luxembourg; [4]Luxembourg Centre of Neuropathology (LCNP), Luxembourg; [5] Neuro-Immunology Group, Department of Oncology (DONC), Luxembourg Institute of Health (LIH), Luxembourg, Luxembourg; [6]Department of Infection Genetics, Helmholtz Centre for Infection Research, Braunschweig, Germany; [7]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, USA; [8]University of Veterinary Medicine Hannover, Hannover, Germany; [9]Department of Microbiology, Immunology and Biochemistry, University of Tennessee Health Science Center, Memphis, Tennessee, USA; [10]Pacific Northwest Research Institute, Seattle, Washington, United States; [11]Maine Medical Center Research Institute, Scarborough, Maine USA

#These authors contributed equally

*__Correspondance:__ Drs. Lasse Sinkkonen (lasse.sinkkonen@uni.lu), Manuel Buttini (manuel.buttini@uni.lu)

Pages: 21 ; Figures: 5; Tables: 3 (supplemental) ; Words in – Abstract: 240 (*max 250*), Introduction: 631 (*max 650*), Discussion 1055 (*max 1500*)

The authors declare no conflicts of interest.

Acknowledgements: LS and MB thank the Luxembourg National Research Fund (FNR) for funding support (FNR CORE C15/BM/10406131 grant). MM thanks the FNR for funding support (FNR PEARL P16/BM/11192868 grant). KS thanks the support by intra-mural grants from the Helmholtz-Association (Program Infection and Immunity). The authors thank Dr. A. Ginolhac, Dr. A Gaigneaux, Dr. D. Coowar, Dr. R. Halder, Z. Hodak, and the animal caretakes in Luxembourg and Braunschweig, for their contributions.

Abstract

Features of dopaminergic neurons (DAns) in the nigrostriatal circuit are orchestrated by a multitude of yet unknown factors, many of them genetic. Decline of DAns, a characteristic of Parkinson's disease, heralds a decrease in dopamine level, and thus measuring striatal dopamine reflects the integrity of DAns. To identify novel genetic regulators of the integrity of DAns, we used Collaborative Cross (CC) mouse strains to find quantitative trait loci (QTLs) related to dopamine levels in the dorsal striatum. Dopamine levels in this brain region varied greatly in the eight CC founder strains, and the differences were inheritable in 32 derived CC strains. QTL mapping identified a locus associated with dopamine level on chromosome X that contained 393 genes. RNA-seq analysis of the ventral midbrain of two of the founder strains with large striatal dopamine difference (C57BL/6J and A/J) revealed 24 differentially expressed genes within the QTL. The protein-coding gene with the highest expression difference was *Col4a6*, which exhibited a 9-fold reduction in A/J compared to C57BL/6J, consistent with lower dopamine levels in A/J. This was accompanied by reduced striatal axonal branching of DAns in A/J compared to C57BL/6J. Single cell RNA-seq data from developing human midbrain suggests that *Col4a6* is highly expressed in radial glia-like cells and neuronal progenitors, indicating possible involvement in neuronal development. Moreover, *Col4a6* controls axogenesis in non-mammal model organisms. We conclude that dopamine levels and axonal branching in mouse dorsal striatum are modulated by COL4A6 levels during development.

Key words: *Col4a6* – mouse strains – dopamine – QTL – regulatory variants – nigrostriatal circuit

Significance statement

Axonal projections of dopaminergic neurons of the Substantia Nigra are among the longest and highly branched in the mammalian CNS. Their complex arborisation renders them exceedingly susceptible to stressors, and they are one of the first structures to degenerate in Parkinson's disease. Many factors regulating the dopaminergic neuron identity and development are known, but few factors governing axonal branching of these neurons are. Using Quantitative Trait Loci mapping on different inbred mouse strains, phenotypic profiling of dopaminergic neuron features, and mining of public omics databases, we identify a subunit of collagen IV, *Col4a6*, as a putatative key modulator of axonal branching of nigral dopaminergic neurons. Our findings help understand the formation of axonal arborisation of these neurons, and may help design neuroprotective therapies.

Introduction

Dopamine (DA), one of the main neurotransmitters in mammalian brain, is involved in several important activities, including motor and cognitive functions. Two important populations of DAns, with distinct activities, locate in substantia nigra (SN) and ventral tegmental area (VTA) in the ventral midbrain. The DAns in the SN project mainly to dorsal striatum, controlling motor function, while the ones in the VTA project to nucleus accumbens and amygdala, controlling reward and emotion, or to the cortex and hippocampus, modulating cognition and memory (Hassan and Benarroch, 2015; Vogt Weisenhorn et al., 2016). Both DAns are at the centre of research interests because of their involvement in neurological diseases, notably VTA DAn in neuropsychiatric diseases, and SN DAn in Parkinson's disease (PD).

Because of its prevalence and costs to society, PD has received a lot of attention. Environmental and genetic risk factors modulate the variability of PD (Jankovic et al., 1990; Gilgun-Sherki et al., 2004; van Rooden et al., 2011; Del Rey et al., 2018). Age of onset, severity, rate of progression of PD motor symptoms, as well as the response to dopamine replacement therapies vary greatly and are likely due to genetic polymorphism in the nigrostriatal circuit (Kalinderi et al., 2011; Kaplan et al., 2014). It has emerged that genetic factors governing the development of this circuit during ontogenesis and its baseline function in adults are frequently those that are dysregulated in PD (Klafke et al., 2008). Hence, a better understanding of genetic variations associated with these factors could pave the way for the understanding of PD.

As genetic studies with standardized environment are difficult in humans, mouse models are used to study genetic variations. The mouse shares similar brain architecture and 99% of genes with humans, and allows cost-effective and controlled studies (Nadeau and Auwerx, 2019). Genetic variation is associated with phenotypic differences in the dopaminergic circuit and associated behaviours. Differences in the DAn cell number as well as in DA levels and protein trafficking have been shown

between different strains of mice (Baker et al., 1980; Vadász et al., 1987; Vadasz et al., 1998; Zaborszky and Vadasz, 2001; Cabib et al., 2002). Motor behaviour and susceptibility to PD-toxin differ between strains (Ingram et al., 1981; Hamre et al., 1999; Brooks et al., 2004; Jong et al., 2010). Recombinant inbred mouse strains constitute interesting models to identify candidate genes by QTL mapping (Peters et al., 2007). Collaborative Cross (CC) strains are a collection of such strains derived from eight founder strains (Churchill et al., 2004).

In our study, we used CC strains to map QTLs related to the integrity of nigral DAns. We measured dopamine level in the striata of eight CC founders and 32 derived CC strains, and observed that the striatal dopamine level was influenced by the genetic background of the strains. We identified a QTL associated with striatal dopamine level on chromosome X that contained 393 genes. The transcriptomic analysis of C57BL/6J and A/J ventral midbrains, two CC founders having large striatal dopamine level differences, revealed 24 differentially expressed genes within the QTL, with *Col4a6* showing the highest expression difference. Studies using single cell RNA-seq data of developing human midbrain, have revealed a developmental expression profile for *Col4a6* indicating a role in neurogenesis (La Manno et al., 2016), and morphological studies in non-mammalian models (Drosophila, zebrafish) indicate a role in axon guidance and outgrowth (Mirre et al., 1992; Takeuchi et al., 2015). Consistently, measurements of TH-positive axons in projection areas of SN DAns (dorsal striatum) and of VTA DAns (piriform cortex, amygdala) revealed that axonal branching of SN DAns, but not that of VTA DAns, differed between C57BL/6J and A/J. However, the number of TH-positive neurons in the SN did not differ between these 2 strains. These observations indicate that differences in *Col4a6* expression lead to differences in dopaminergic striatal innervation.

Materials and methods

Animals

Eight parental founder strains (A/J, C57BL/6J, 129S1Sv/ImJ, CAST/EiJ, PWK/PhJ, WSB/EiJ, NOD/ShiLtJ, NZO/H1LtJ) and 32 CC strains (Supplementary Table S1), originally obtained from the University of North Carolina, Chapel Hill (UNC), were bred at Chapel Hill or at the Central Animal Facilities of the Helmholtz Centre for Infection Research (Braunschweig, Germany). 10 to 12 mice per group (mixed males and females) were anesthetized with a ketamine-medetomidine mix (150 and 1 mg/kg, respectively). Intracardiac perfusion was performed (phosphate-buffered saline) for each animal before dissecting the striatum and midbrain, immediately snap-frozen. The second hemibrain was fixed in paraformaldehyde (PFA) 4%. The experiments were performed according to the national guidelines of the animal welfare law in Germany (BGBl. I S. 1206, 1313 and BGBl. I S. 1934) and the European Communities Council Directive 2010/63/EU. The protocol was reviewed and approved by the 'Niedersächsisches Landesamt für Verbraucherschutz und Lebensmittelsicherheit, Oldenburg, Germany' (Permit Numbers: 33.9-42502-05-11A193, 33.19-42502-05-19A394), respecting the 3 Rs' requirements for Animal Welfare.

Dopamine measurements by gas chromatography-mass spectrometry (GC-MS)

Striatal DA was measured in 3-month-old CC founders and 32 CC strains. As we used two different methods to extract the metabolites from the tissues, we present the results as percentage of C57BL/6J. The tissue homogenization and metabolite extraction were performed at 4°C or lower to prevent changes in the metabolic profile.

The first method was described by Jaeger *et al*., 2015 (Jaeger et al., 2015). The striatum of each mouse was pulverized in a bead mill with grinding beads (7 mm). The samples were then homogenized in the bead mill with smaller grinding beads (1 mm) and the extraction fluid (methanol/distilled water, 40:8.5 v/v). The metabolites were extracted using a liquid-liquid extraction method first by addition of chloroform to the tissue fluid followed by distilled water. After shaking for 20 minutes at 1300 rpm at 4°C, the mixture was centrifuged for 5 minutes at 5000 x g at 4°C. The

upper phase containing the polar metabolites was transferred to a sample vial for speed vacuum evaporation.

For both methods, the resulting dried samples were derivatized in an established procedure. 20 µL of pyridine (containing 20 mg/mL of methoxyamine hydrochloride) were added to the samples and incubated at 45 °C with continuous shaking for 90 min. Then 20 µL of MSTFA were added to the sample vial and incubated 30 min at 45 °C with continuous shaking.

After derivatization, the GC-MS analysis was performed with an Agilent 7890A GC, or 7890B for the second method, coupled to an Agilent 5975C inert XL mass selective detector (MSD) or 5977A for the second method (Diegem, Belgium). 1 µL of sample was injected into a Split/Splitless inlet operating in split mode (10:1) at 270 °C. Helium was used as a carrier gas with a constant flow rate of 1.2 mL/min. The second method was slightly modified to further reduce the runtime. The GC oven temperature was held 1 min at 80 °C (0.6 min at 90 °C for the second method) and increased at 36 °C/min to 260 °C (at 25 °C/min to 200 °C and held for 6 min for the second method). Then the temperature was increased at 22 °C/min and maintained at a constant temperature of 325 °C for 2 min (4 min post run time at 325 °C for the second method). The transfer line temperature was set constantly to 280 °C and the MSD was operating under electron ionization at 70 eV. As described by Jäger *et al*., 2016 (Jäger et al., 2016), a multi-analyte detection using a quadrupole analyzer in selected ion monitoring mode was used for a sensitive and precise quantification of DA and the internal standard DA-*d4*.

Statistical analysis was performed using the GraphPad Prism 8 software. After applying the Shapiro-Wilk test to assess the normality of our data, a one-way ANOVA was applied to analyse the striatal dopamine levels.

Immunofluorescence

TH protein was measured by immunofluorescence in the dorsal striatum, amygdala, piriform cortex, and SN of 3-month-old C57BL/6J and A/J. From 6 to 14 hemibrains per group were fixed (PFA 4%) for 48h and stored in PBS with 0.2% of sodium azide. Parasagittal free floating sections (50 µm) were generated using a vibratome (Leica; VT 1000S) collected every 4_{th} sections in a tube containing a cryoprotective medium (polyvinyl pyrrolidone 1% w/v in PBS/ethylene glycol 1:1) and stored at -20°C. The lateral sections were collected for the striatum, amygdala and piriform cortex measurements, and the medial sections were collected for the SN measurements.

The sections were washed in PBS with 0.1% Triton X-100 (PBST) and permeabilized in PBS with 3% H_20_2 and1.5% Triton X-100. The sections were then blocked for 1h in PBST with 5% of Bovine Serum Albumin (BSA) and incubated overnight with rabbit anti-TH antibody (1:1000, Millipore, AB152) diluted in PBST with 2% of BSA. After several washing, the sections were incubated for 2 hours with the secondary antibody (Alexa fluor™ 488 goat anti-rabbit 1:1000, Invitrogen), mounted on slides and embedded in fluoromount.

Imaging was performed using a Zeiss AxioImager Z1 upright microscope, coupled to a "Colibri" LED system, and an Mrm3 digital camera for image capture using the software Zeiss Zen 2 Blue. For each striatum, amygdala and piriform cortex section, three images of each brain area were taken at 40x magnification using the apotome system. After thresholding, the area occupied by TH stainings in each picture was determined using the FIJI imaging software (Schindelin et al., 2012; Masliah et al., 2000). For the SN sections, the pictures were taken at 10x magnification. The area occupied by TH positive neurons was measured in the region of interest corresponding to the SN using ImageJ FIJI software. We can distinguish four different areas of the SN. Each area was quantified and averaged separately and summed as a cumulated surface (mm_2) (Masliah et al., 2000; Ashrafi et al., 2017).

GraphPad Prism 8 software was used for the statistical analysis. After applying the Shapiro-Wilk test to assess the normality of our data, an unpaired t-test was applied to analyse the TH measurement in different areas.

Quantitative trait locus mapping

The QTL mapping was done with http://gn2.genenetwork.org/. The dataset containing dopamine measurements of dorsal striata of 32 CC strains were located with search terms (Species: Mouse (mm10); Group: CC Family; Type: Phenotypes; Dataset: CC Phenotypes) and navigated to Record CCF_10001 and CCF_10002. The QTL mapping was done with GEMMA on all chromosomes, MAF >= 0.05 with LOCO method. The genome wide significance of QTL mapping on male and female are set by 500 permutation simulations with FDR under 5% for each scan.

Estimation of ventral dopamine level heritability in CC strains

The broad sense heritability is estimated based on (Belknap, 1998). Briefly, the total phenotypic variance (Vp) is calculated on all CC strains. The genetic variance (Va) is estimated by the mean of within-strain variance. The heritability (H^2) is calculated as Va/Vp.

Results

Differences in striatal dopamine levels across Collaborative Cross mice are under genetic control

To determine whether the reported phenotypic differences between CC mouse strains (Schoenrock et al., 2018; Schoenrock et al., 2020) are accompanied by differences in striatal DA levels, we measured DA levels in isolated dorsal striata from the eight inbred founder CC strains (Supplementary Table S1). In total, dopamine from 102 mice at the age of 3 months was measured by GC-MS. DA levels

varied significantly across the founders (one-way ANOVA, p=0.0004, F=4.507), indicating strain-specific differences in DA levels in the nigrostriatal dopaminergic circuit (Figure 1). PWK/PhJ, A/J, and NOD/LtJ strains showed the lowest levels of DA, while the highest levels were detected in NZO/HILtJ, CAST/EiJ, and C57BL/6J mice. Thus, striatal DA levels appear to be under genetic control.

To investigate if variation in DA level is indeed inheritable, we measured the striatal DA level across 32 strains of CC mice. In total, we analysed 327 CC mice with similar number of mice from both sexes. The CC strains showed considerable variation in DA levels with a range of around 10 pmol/mg in both sexes (Figure 2). From these values, the estimated broad-sense heritability (H^2) was calculated to be 0.52, indicating the DA level differences are inheritable, and associated genetic variation could be detected by QTL mapping. (Hegmann and Possidente, 1981) (see Methods for details).

QTL mapping associates a genomic locus on chromosome X with striatal dopamine levels

Identifying novel genetic regulators associated with striatal DA levels could help better understand the development of dopaminergic circuits and susceptibility to diseases, like PD. Therefore, to leverage the power of CC strains to identify trait-associated genetic loci at a good resolution, we performed QTL mapping based on the measured DA levels across the 32 CC strains. The mapping was performed separately for males and females, and the genome-wide significance of results for all genetic markers are presented in Figure 3 and Supplementary Table S2. QTL mapping using the female data identified a genetic marker located on chromosome X at position 144.300241 Mb to be associated with DA levels, when applying the 95^{th} percentile threshold for genome-wide significance ($-\log_{10}$ p-value = 5.23) (Figure 3A). Moreover, an adjacent upstream marker at position 136.176403 Mb showed strong association in males, with a p-value just below the applied genome-wide significance cut-off (-log10 p-value = 4.91) (Figure 3B). This indicated that the region from 136.176403 Mb to 144.300241 Mb harbors a QTL associated with DA levels in mice. In addition,

markers at downstream positions of 157.823410 Mb and 158.259643 Mb were also highly associated in females (-$\log_{10}$ p-value = 4.96 for both markers). Taken together, our QTL analysis identified a combined region spanning over 32 Mb with high association to striatal DA levels on chromosome X, with 8 Mb region from 136.176403 Mb to 144.300241 Mb showing highest significance in both males and females.

Col4a6 is a developmental gene with altered expression between mouse strains

The identified 32 Mb locus from position 131 Mb to 163 Mb on chromosome X includes 393 genes that could potentially be underlying the association with striatal DA levels, with a region of 8 MB containing 163 genes with most significant association. However, the vast majority (>95%) of trait-associated genetic variants are located outside of protein-coding genomic regions (Maurano et al., 2012). Recent advances in functional genomics analysis have revealed these non-coding variants to be highly enriched in gene regulatory regions such as enhancers where they can disrupt transcription factor binding and alter the target gene expression. Therefore, we asked whether such *cis*-acting regulatory variants could be affecting gene expression at our locus of interest in the SN of the midbrain, from where the DA neurons project to the dorsal striatum. To this aim, we took advantage of our recent transcriptomic profiling of ventral midbrains from C57BL/6J and A/J mice (Gui et al., 2020), two CC founder strains with significantly different levels of striatal DA (Figure 1, p=0.026, unpaired t-test). These two strains are also widely used laboratory mouse strains and have one of the largest differences in striatal dopamine (see above). Using the RNA-seq data we plotted the absolute $\log_2$-fold change for all 393 genes between C57BL/6J and A/J to identify those with altered gene expression (Figure 4A). Interestingly, only 24 genes showed significant differential gene expression (Supplementary Table S3). By far the largest fold change of all protein-coding genes was found for *Collagen 4a6* (*Col4a6*) gene, which showed over 9-fold difference between the 2 strains. Moreover, *Col4a6* transcription start site is located at position 141.474076 Mb, within the 8 Mb region flanked

by the two genetic markers most significantly associated with DA levels in males and females. The transcriptomic analysis was based on a total 24 mice, 12 from each strain and sex, with A/J displaying a significant reduction for *Col4a6* in both females and males compared to age-matched C57BL/6J (Figure 4B), consistent with a lower level of DA in the striatum of A/J (Figure 1).

While *Col4a6* expression was significantly lower in A/J, the overall abundance of expression in the adult C57BL/6J midbrain was also very low (<0.3 RPKM), indicating that, in adult mice, its expression is limited to only one or a few cell types, most likely endothelial cells, which have been reported to produce collagens (Gelse et al., 2003; Ricard-Blum, 2011). Collagen IV is an essential and abundant component of the basement membrane (Mao et al., 2015). In the nervous system, its function has been associated with axon guidance and neurite outgrowth in non-mammal model systems (Mirre et al., 1992; Takeuchi et al., 2015), and in cultured sympathetic neurons (Firla, 1990). The two angles of collagen IV function in the nervous system (regulation of neurogenesis and of neurite outgrowth and guidance) are probably intertwined.

To get a better idea about the cellular source of *COL4A6* during development, we observed its expression in published single cell RNA-seq (scRNA-seq) data corresponding to 26 cell types of the developing human midbrain (La Manno et al., 2016). *COL4A6* expression was highest in floor plate progenitors and selected subtypes radial glia-like cells, with very low expression detected in other cell types (Figure 4C). The expression profile of *COL4A6* closely followed the expression of *SOX2*, a key regulator of neurogenesis (Ferri et al., 2004). Consistently, previous screens for primary SOX2 target genes have found *COL4A6* expression to depend on SOX2 (Fang et al., 2011; Berezovsky et al., 2014).

Taken together, the expression profile of *COL4A6* implicates it as a developmental gene and its dependence on SOX2 indicates a possible role in neurogenesis. Indeed, previous work has shown that the zebrafish orthologs of type IV collagens, *col4a5* and *col4a6*, can control proper axonal guidance during zebrafish development (Takeuchi et al., 2015).

Differences in striatal axonal branching between C57BL/6J and A/J mice

Based on the observed differences in striatal DA levels, the localization of *Col4a6* gene in the associated QTL, the distinct neurodevelopmental expression profile of *COL4A6*, and the previously described role of collagen IV in neurite outgrowth and guidance (see above), we hypothesized that DAn axonal fiber density could be different in the striatum of mouse strains. To test this, we performed TH immunostaining on brain sections from the two founder strains C57BL/6J and A/J. The percentage of area occupied by TH in the DAn projection areas (dorsal striatum for SN and piriform cortex and amyglada for VTA) was used to compare axonal fiber density between the strains. Interestingly, 3-month-old A/J mice showed 29% lower TH fiber density in dorsal striatum compared to C57BL/6J (unpaired t-test, p=0.0059), consistent with lower DA levels in A/J (Figure 5). No such differences could be observed in amygdala or in piriform cortex, suggesting that the differences observed between the two mouse strains appear to be specific to the dorsal striatum.

To determine if the number of nigral DAn differed between C57BL/6J and A/J mice, we estimated the amount of these neurons in the 2 strains as described in Materials and Methods. We observed no difference in the number of TH-positive neurons between the two strains, implying that differences observed in striatal TH positive axons reflect a difference in branching of DAn, rather than their number of the SN (Figure 5). Hence, our data points to a role of COL4A6 in modulation of the branching of DAn of the SN, but not that of DAn of the VTA.

Taken together, our QTL analysis allowed us to identify *Col4a6* as a new gene involved in nigral DAns development and their axonal branching in the dorsal striatum.

Discussion

While the DAns residing in the SN are of central interest to translational neuroscience research because of their unique properties that renders them susceptible in PD (Surmeier, 2018), a lot of the mechanisms surrounding their basic properties remain unknown. Uncovering factors that govern their development, structure, and function can help understand how they interrelate with other nervous system components in assuring the proper operations of the brain, or how they can contribute to disease (Klafke et al., 2008).

In this study, we used CC strains to identify QTL and new candidate genes regulating the integrity of the nigrostriatal circuit. The variations of striatal dopamine levels between CC strains demonstrate an inheritable part of this trait. Together with previous transcriptomic data in the ventral midbrain of two founder strains, our results point to *Col4a6* playing an important role modulating axonal branching of striatal DAns, the groundwork of which may be set during early phases of development.

Despite the success of human GWAS methodologies to decipher phenotype-genotype associations, human tools lack proper standardized and controlled conditions. To overcome these limitations, CC mouse strains were generated to provide a model for heterogeneous human population (Churchill et al., 2004). Genetic diversity of CC mouse model provides more precise QTL mapping results than conventional mapping populations. The wide phenotypic range of around 10 pmol/mg DA enabled us to map a significant QTL of about 32 Mb with a reasonable number of 32 CC strains and eight founder strains. Our previous transcriptomic analysis coming from two CC founders with significantly different levels of striatal DA (Gui et al., 2020) provide useful data to narrow the QTL on chromosome X to 24 DEG, *Col4a6* showing a 9-fold difference in C57BL/6J compared to A/J, two of the CC founder strains that had one of the largest striatal DA and DAn's axonal branching differences. To find out more about the role of *Col4a6* in this brain region, we started to look at data available from the developing human midbrain (La Manno et al., 2016), which suggests an important role of *Col4a6* in the DAns neurogenesis. Collagen IV alpha-6 chain is one of the six subunits of type IV collagen, a major component of basement membranes. Collagen IV is a member of the

collagen family of glycoproteins, which themselves are constituents of the extracellular matrix (ECM), and among the most abundant proteins in the animal kingdom (Vecino and Kwok, 2016). ECM proteins, in particular collagens, in the nervous system play key roles in development, in cellular maintenance and repair, and in tissue responses to diseases involving injury or neoplasia (Rutka et al., 1988). Collagens in the PNS provide a scaffold for Schwann cells and support neurite outgrowth (Lein et al., 1991; Chen et al., 2015). As stated above, a central role for collagen IV in axon guidance and neurite outgrowth is also supported by studies in simple model organisms (Mirre et al., 1992; Takeuchi et al., 2015). Collagens, including type IV, in the nervous system are produced primarily by cells of mesodermal origin, the endothelial cells, and are found, together with other forms ECM proteins, in the basement membrane of cerebral blood vessels and at the glia limitans (Rutka et al., 1988). Upon injury, collagen expression and secretion together with that of other ECM proteins, by glial cells appears (Liesi and Kauppila, 2002). While evidence suggest that, in some scenarios, this process is supportive of neurite outgrowth (see above), in rats, by contributing to the formation of the glial scar, it is thought to inhibit axonal regeneration (Liesi and Kauppila, 2002). Interestingly though, engineered biopolymer scaffolds containing collagen are being explored as therapeutic support for nerve repair after injury (Li and Dai, 2018). Thus, the role of collagens in neuronal maintenance and repair may depend on a delicate balance of opposing actions.

Our study shows for the first time that one subunit of collagen IV, subunit 6, is a candidate for regulating axonal branching in the CNS of a mammal. Because we had already observed large striatal DA differences and large midbrain *Col4a6* expression differences in C56BL/6 versus A/J mice, we then tested if these two mouse strains also showed differences in striatal axonal branching. We therefore measured TH-positive axons in the dorsal striatum C57BL/6J and A/J by quantitative immunofluorescence. We observed less TH-positive axons in the dorsal striatum of A/J compared to C57BL/6J mice. To determine if this was due to a difference in the number of TH positive neurons in the SN between these two strains, we also estimated those numbers, but did not find a difference.

Thus, the lower TH-positive axons we observe in the striata of A/J mice most likely reflect a lesser axonal branching of the DAns. To ensure that our strain differences affect the nigrostriatal circuitry and not all projecting DAns of the ventral midbrain, we assessed brain areas that receive projections from the VTA. Based on our results from the amygdala and piriform cortex, phenotypic differences observed in the striata of C57BL/6J and A/J are not present in the other midbrain dopaminergic circuits. Other studies indicate that more factors regulating nigrostriatal integrity remain to be found. Studying mouse strains other than those used in our study, Baker *et al.*, 1980 (Baker et al., 1980) and Zaborsky *et al.*, 2001 (Zaborszky and Vadasz, 2001) reported strain-dependent differences in the number of midbrain populations of DAns. On a functional level, other studies found strain-specific differences in motor behaviour, such as lower motor activity, balance and exploratory skills displayed by A/J compared to C57BL/6J (Jong et al., 2010). Finally, recent studies showed significant effect of the genetic background on locomotor behaviour (Schoenrock et al., 2020) and response to cocaine (Schoenrock et al., 2018) using recombinant inbred intercrosses generated from CC strains, illustrating how the CC strains can serve as useful model for identifying further QTLs and genetic variants that govern structure and function of DAns.

In this study, using biochemical and neuropathological analyses in combination with QTL mapping in the CC mouse population, we highlight *Col4a6* as a new gene candidate regulating the axonal branching in the nigrostriatal dopaminergic system of mammals. Because these are the structures that are affected early in PD (Kordower et al., 2013), we propose that a better understanding of the actions of collagen IV on these neurons may open the way for novel neuroprotection therapies.

References

Ashrafi, A., Garcia, P., Kollmus, H., Schughart, K., Del Sol, A., Buttini, M., and Glaab, E. (2017). Absence of regulator of G-protein signaling 4 does not protect against dopamine neuron dysfunction and injury in the mouse 6-hydroxydopamine lesion model of Parkinson's disease. Neurobiology of aging *58*, 30-33.

Baker, H., Joh, T.H., and Reis, D.J. (1980). Genetic control of number of midbrain dopaminergic neurons in inbred strains of mice: relationship to size and neuronal density of the striatum. Proceedings of the National Academy of Sciences of the United States of America *77*, 4369-4373.

Belknap, J.K. (1998). Effect of within-strain sample size on QTL detection and mapping using recombinant inbred mouse strains. Behavior genetics *28*, 29-38.

Berezovsky, A.D., Poisson, L.M., Cherba, D., Webb, C.P., Transou, A.D., Lemke, N.W., Hong, X., Hasselbach, L.A., Irtenkauf, S.M., and Mikkelsen, T., et al. (2014). Sox2 promotes malignancy in glioblastoma by regulating plasticity and astrocytic differentiation. Neoplasia (New York, N.Y.) *16*, 193-206, 206.e19-25.

Brooks, S.P., Pask, T., Jones, L., and Dunnett, S.B. (2004). Behavioural profiles of inbred mouse strains used as transgenic backgrounds. I: motor tests. Genes, brain, and behavior *3*, 206-215.

Cabib, S., Puglisi-Allegra, S., and Ventura, R. (2002). The contribution of comparative studies in inbred strains of mice to the understanding of the hyperactive phenotype. Behavioural brain research *130*, 103-109.

Chen, P., Cescon, M., and Bonaldo, P. (2015). The Role of Collagens in Peripheral Nerve Myelination and Function. Molecular neurobiology *52*, 216-225.

Churchill, G.A., Airey, D.C., Allayee, H., Angel, J.M., Attie, A.D., Beatty, J., Beavis, W.D., Belknap, J.K., Bennett, B., and Berrettini, W., et al. (2004). The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nature genetics *36*, 1133-1137.

Del Rey, N.L.-G., Quiroga-Varela, A., Garbayo, E., Carballo-Carbajal, I., Fernández-Santiago, R., Monje, M.H.G., Trigo-Damas, I., Blanco-Prieto, M.J., and Blesa, J. (2018). Advances in Parkinson's Disease: 200 Years Later. Frontiers in neuroanatomy *12*, 113.

Fang, X., Yoon, J.-G., Li, L., Yu, W., Shao, J., Hua, D., Zheng, S., Hood, L., Goodlett, D.R., and Foltz, G., et al. (2011). The SOX2 response program in glioblastoma multiforme: an integrated ChIP-seq, expression microarray, and microRNA analysis. BMC genomics *12*, 11.

Ferri, A.L.M., Cavallaro, M., Braida, D., Di Cristofano, A., Canta, A., Vezzani, A., Ottolenghi, S., Pandolfi, P.P., Sala, M., and DeBiasi, S., et al. (2004). Sox2 deficiency causes neurodegeneration and impaired neurogenesis in the adult mouse brain. Development (Cambridge, England) *131*, 3805-3819.

Firla, M.T. (1990). Asthetische Aspekte zur biomimetischen Schichttechnik. Zahnarztliche Mitteilungen *80*, 1957-1962.

Gelse, K., Pöschl, E., and Aigner, T. (2003). Collagens--structure, function, and biosynthesis. Advanced drug delivery reviews *55*, 1531-1546.

Gilgun-Sherki, Y., Djaldetti, R., Melamed, E., and Offen, D. (2004). Polymorphism in candidate genes: implications for the risk and treatment of idiopathic Parkinson's disease. The pharmacogenomics journal *4*, 291-306.

Gui, Y., Thomas, M.H., Garcia, P., Karout, M., Halder, R., Michelucci, A., Kollmus, H., Zhou, C., Melmed, S., and Schughart, K., et al. (2020). Pituitary Tumor Transforming Gene 1 orchestrates gene regulatory variation in mouse ventral midbrain during aging.

Hamre, K., Tharp, R., Poon, K., Xiong, X., and Smeyne, R.J. (1999). Differential strain susceptibility following 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) administration acts in an autosomal dominant fashion: quantitative analysis in seven strains of Mus musculus. Brain research *828*, 91-103.

Hassan, A., and Benarroch, E.E. (2015). Heterogeneity of the midbrain dopamine system: Implications for Parkinson disease. Neurology *85*, 1795-1805.

Hegmann, J.P., and Possidente, B. (1981). Estimating genetic correlations from inbred strains. Behavior genetics *11*, 103-114.

Ingram, D.K., London, E.D., Reynolds, M.A., Waller, S.B., and Goodrick, C.L. (1981). Differential effects of age on motor performance in two mouse strains. Neurobiology of aging *2*, 221-227.

Jaeger, C., Glaab, E., Michelucci, A., Binz, T.M., Koeglsberger, S., Garcia, P., Trezzi, J.-P., Ghelfi, J., Balling, R., and Buttini, M. (2015). The mouse brain metabolome: region-specific signatures and response to excitotoxic neuronal injury. The American journal of pathology *185*, 1699-1712.

Jäger, C., Hiller, K., and Buttini, M. (2016). Metabolic Profiling and Quantification of Neurotransmitters in Mouse Brain by Gas Chromatography-Mass Spectrometry. Current protocols in mouse biology *6*, 333-342.

Jankovic, J., McDermott, M., Carter, J., Gauthier, S., Goetz, C., Golbe, L., Huber, S., Koller, W., Olanow, C., and Shoulson, I. (1990). Variable expression of Parkinson's disease: a base-line analysis of the DATATOP cohort. The Parkinson Study Group. Neurology *40*, 1529-1534.

Jong, S. de, Fuller, T.F., Janson, E., Strengman, E., Horvath, S., Kas, M.J.H., and Ophoff, R.A. (2010). Gene expression profiling in C57BL/6J and A/J mouse inbred strains reveals gene networks specific for brain regions independent of genetic background. BMC genomics *11*, 20.

Kalinderi, K., Fidani, L., Katsarou, Z., and Bostantjopoulou, S. (2011). Pharmacological treatment and the prospect of pharmacogenetics in Parkinson's disease. International journal of clinical practice *65*, 1289-1294.

Kaplan, N., Vituri, A., Korczyn, A.D., Cohen, O.S., Inzelberg, R., Yahalom, G., Kozlova, E., Milgrom, R., Laitman, Y., and Friedman, E., et al. (2014). Sequence variants in SLC6A3, DRD2, and BDNF genes and time to levodopa-induced dyskinesias in Parkinson's disease. Journal of molecular neuroscience : MN *53*, 183-188.

Klafke, R., Wurst, W., and Prakash, N. (2008). Genetic control of rodent midbrain dopaminergic neuron development in the light of human disease. Pharmacopsychiatry *41 Suppl 1*, S44-50.

Kordower, J.H., Olanow, C.W., Dodiya, H.B., Chu, Y., Beach, T.G., Adler, C.H., Halliday, G.M., and Bartus, R.T. (2013). Disease duration and the integrity of the nigrostriatal system in Parkinson's disease. Brain : a journal of neurology *136*, 2419-2431.

La Manno, G., Gyllborg, D., Codeluppi, S., Nishimura, K., Salto, C., Zeisel, A., Borm, L.E., Stott, S.R.W., Toledo, E.M., and Villaescusa, J.C., et al. (2016). Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. Cell *167*, 566-580.e19.

Lein, P.J., Higgins, D., Turner, D.C., Flier, L.A., and Terranova, V.P. (1991). The NC1 domain of type IV collagen promotes axonal growth in sympathetic neurons through interaction with the alpha 1 beta 1 integrin. The Journal of cell biology *113*, 417-428.

Li, X., and Dai, J. (2018). Bridging the gap with functional collagen scaffolds: tuning endogenous neural stem cells for severe spinal cord injury repair. Biomaterials science *6*, 265-271.

Liesi, P., and Kauppila, T. (2002). Induction of type IV collagen and other basement-membrane-associated proteins after spinal cord injury of the adult rat may participate in formation of the glial scar. Experimental neurology *173*, 31-45.

Mao, M., Alavi, M.V., Labelle-Dumais, C., and Gould, D.B. (2015). Type IV Collagens and Basement Membrane Diseases: Cell Biology and Pathogenic Mechanisms. Current topics in membranes *76*, 61-116.

Masliah, E., Rockenstein, E., Veinbergs, I., Mallory, M., Hashimoto, M., Takeda, A., Sagara, Y., Sisk, A., and Mucke, L. (2000). Dopaminergic loss and inclusion body formation in alpha-synuclein mice: implications for neurodegenerative disorders. Science (New York, N.Y.) *287*, 1265-1269.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., and Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science (New York, N.Y.) *337*, 1190-1195.

Mirre, C., Le Parco, Y., and Knibiehler, B. (1992). Collagen IV is present in the developing CNS during Drosophila neurogenesis. Journal of neuroscience research *31*, 146-155.

Nadeau, J.H., and Auwerx, J. (2019). The virtuous cycle of human genetics and mouse models in drug discovery. Nature reviews. Drug discovery *18*, 255-272.

Peters, L.L., Robledo, R.F., Bult, C.J., Churchill, G.A., Paigen, B.J., and Svenson, K.L. (2007). The mouse as a model for human biology: a resource guide for complex trait analysis. Nature reviews. Genetics *8*, 58-69.

Ricard-Blum, S. (2011). The collagen family. Cold Spring Harbor perspectives in biology *3*, a004978.

Rutka, J.T., Apodaca, G., Stern, R., and Rosenblum, M. (1988). The extracellular matrix of the central and peripheral nervous systems: structure and function. Journal of neurosurgery *69*, 155-170.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., and Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nature methods *9*, 676-682.

Schoenrock, S.A., Kumar, P., Gómez-A, A., Dickson, P.E., Kim, S.-M., Bailey, L., Neira, S., Riker, K.D., Farrington, J., and Gaines, C.H., et al. (2020). Characterization of genetically complex Collaborative Cross mouse strains that model divergent locomotor activating and reinforcing properties of cocaine. Psychopharmacology *237*, 979-996.

Schoenrock, S.A., Oreper, D., Farrington, J., McMullan, R.C., Ervin, R., Miller, D.R., Pardo-Manuel de Villena, F., Valdar, W., and Tarantino, L.M. (2018). Perinatal nutrition interacts with genetic background to alter behavior in a parent-of-origin-dependent manner in adult Collaborative Cross mice. Genes, brain, and behavior *17*, e12438.

Surmeier, D.J. (2018). Determinants of dopaminergic neuron loss in Parkinson's disease. The FEBS journal *285*, 3657-3668.

Takeuchi, M., Yamaguchi, S., Yonemura, S., Kakiguchi, K., Sato, Y., Higashiyama, T., Shimizu, T., and Hibi, M. (2015). Type IV Collagen Controls the Axogenesis of Cerebellar Granule Cells by Regulating Basement Membrane Integrity in Zebrafish. PLoS genetics *11*, e1005587.

Vadasz, C., Sziraki, I., Sasvari, M., Kabai, P., Murthy, L.R., Saito, M., and Laszlovszky, I. (1998). Analysis of the mesotelencephalic dopamine system by quantitative-trait locus introgression. Neurochemical research *23*, 1337-1354.

Vadász, C., Sziráki, I., Murthy, L.R., Vadász, I., Badalamenti, A.F., Kóbor, G., and Lajtha, A. (1987). Genetic determination of mesencephalic tyrosine hydroxylase activity in the mouse. Journal of neurogenetics *4*, 241-252.

van Rooden, S.M., Colas, F., Martínez-Martín, P., Visser, M., Verbaan, D., Marinus, J., Chaudhuri, R.K., Kok, J.N., and van Hilten, J.J. (2011). Clinical subtypes of Parkinson's disease. Movement disorders : official journal of the Movement Disorder Society *26*, 51-58.

Vecino, E., and Kwok, J.C.F. (2016). The Extracellular Matrix in the Nervous System: The Good and the Bad Aspects. In Composition and Function of the Extracellular Matrix in the Human Body, F. Travascio, ed. (InTech).

Vogt Weisenhorn, D.M., Giesert, F., and Wurst, W. (2016). Diversity matters - heterogeneity of dopaminergic neurons in the ventral mesencephalon and its relation to Parkinson's Disease. Journal of neurochemistry *139 Suppl 1*, 8-26.

Zaborszky, L., and Vadasz, C. (2001). The midbrain dopaminergic system: anatomy and genetic variation in dopamine neuron number of inbred mouse strains. Behavior genetics *31*, 47-59.

Figure Legends

Figure 1. Striatal dopamine levels measured in the different Collaborative Cross (CC) founders. (A) Schematic representation of the experimental set-up. (B) Level of dopamine measured by GC-MS in the striatum of the different CC founders, expressed relative to the dopamine level of the common C57BL/6J strain. Data are expressed in mean ± standard deviation and the significance of differences was tested with one-way ANOVA ($p=0.0004$, $F=4.507$).

Figure 2. Striatal dopamine levels measured in the different collaborative cross (CC) strains. Levels of dopamine measured by GC-MS in the striatum of the different CC strains, expressed in pg/mg of tissue (mean ± standard deviation) and the significance of differences was tested with one-way ANOVA. (A) Level of dopamine in the striatum of CC males ($p<0.0001$, $F=3.964$). (B) Level of dopamine in the striatum of CC females ($p<0.0001$, $F=7.435$).

Figure 3. QTL mapping for dorsal striatum dopamine levels in CC strains. Plots show $-\log_{10}$ p-values (y-axis) of genetic markers across chromosome locations (x-axis). Horizontal dashed lines represent the 95^{th} percentile thresholds for genome-wide significance. (A) QTL mapping with female CC strains yielded the most significant QTL at chromosome X 144.3 Mb with -log(P) 5.23. (B) QTL mapping with male CC strains yielded the most significant QTL at chromosome X 136.2 Mb with -log(P) 4.91.

Figure 4. *Col4a6* is the gene with highest differential expression between C57BL/6J and A/J at the identified chromosome X QTL. (A) Log2-fold changes of genes in chromosome X locus between 131Mb and 163 MB. Genes with adjusted p-value below 0.05 are labelled in red. (B) The expression (RPKM) of *Col4a6* in the ventral midbrain of C57BL/6J and A/J. (C) The expression of *Col4a6* and *Sox2* during human midbrain development based on scRNA-seq data from La Manno et al (La Manno et al., 2016). The two genes show similar expression profiles. Cell types are named with "h" to indicate human: Endo, endothelial cells ; Peric, pericytes; Mgl, microglia; OPC, oligodendrocyte precursor cells; Rgl1-3, radial glia-like cells; NProg, neuronal progenitor; Prog, progenitor medial floorplate (FPM), lateral floorplate (FPL), midline (M), basal plate (BP); NbM, medial neuroblast; NbML1&5, mediolateral neuroblasts; RN, red nucleus; DA0-2, dopaminergic neurons; Gaba, GABAergic neurons; Sert, serotonergic; OMTN, oculomotor and trochlear nucleus.

Figure 5. Measure of tyrosine hydroxylase (TH) by immunofluorescence in the brain of 3-month-old C57BL/6J and A/J. (A, B) TH in striatum. (C, D) TH in amygdala. (E, F) TH in piriform cortex. (G, H) TH in substantia nigra. (A, C, E) Images of TH stainings in striatum, amygdala and piriform cortex, magnification 40X. Scales bar: 50 µm (G) Images of TH stainings in substantia nigra, magnification 10X. Scale bar: 200 µm (B, D, F) Quantification of the percentage area occupied by TH from stainings in striatum, amygdala and piriform cortex. (H) Quantification of the TH-positive area in substantia nigra. Data are expressed in mean ± standard deviation and were analysed with unpaired t-test ($p=0.0059$ in the striatum (**), $p=0.47$ in amygdala, $p=0.43$ in piriform cortex and $p=0.78$ in SN).

Figure 1

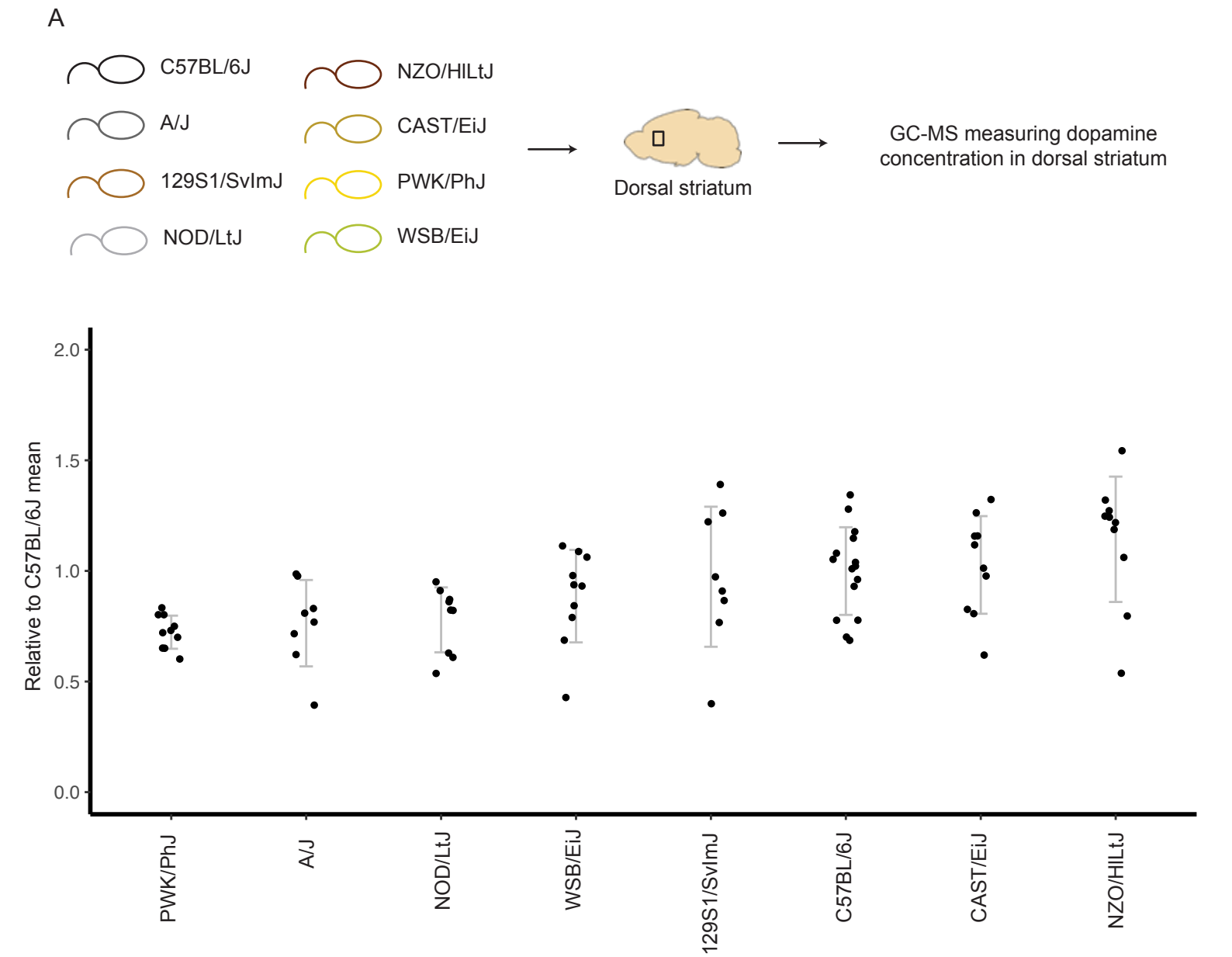
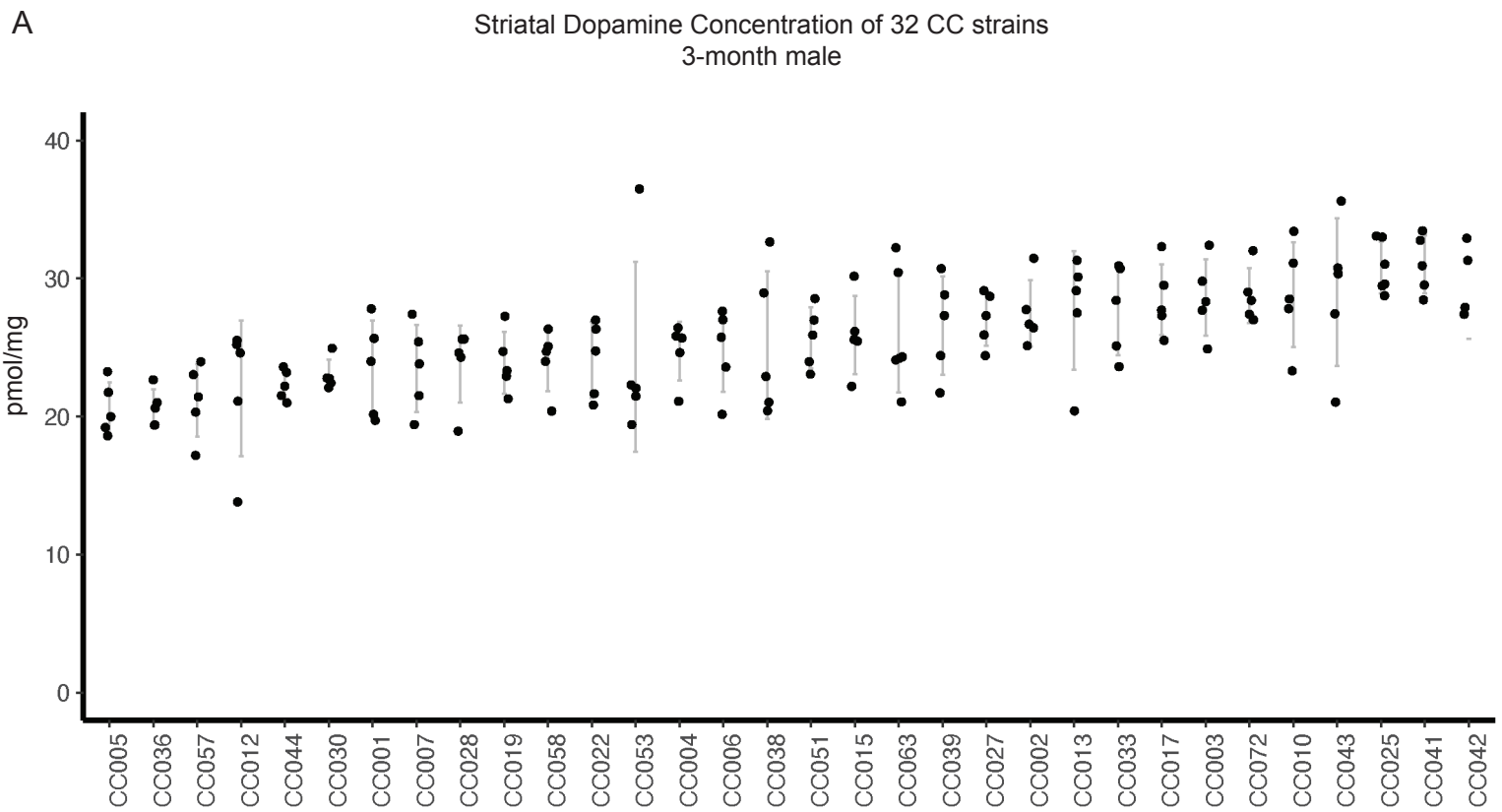

A
C57BL/6J
A/J
129S1/SvImJ
NOD/LtJ
NZO/HlLtJ
CAST/EiJ
PWK/PhJ
WSB/EiJ
Dorsal striatum
GC-MS measuring dopamine concentration in dorsal striatum
Relative to C57BL/6J mean
2.0
1.5
1.0
0.5
0.0
PWK/PhJ
A/J
NOD/LtJ
WSB/EiJ
129S1/SvImJ
C57BL/6J
CAST/EiJ
NZO/HlLtJ

Figure 2


A
Striatal Dopamine Concentration of 32 CC strains
3-month male
pmol/mg
40
30
20
10
0
CC005
CC036
CC057
CC012
CC044
CC030
CC001
CC007
CC028
CC019
CC058
CC022
CC053
CC004
CC006
CC038
CC051
CC015
CC063
CC039
CC027
CC002
CC013
CC033
CC017
CC003
CC072
CC010
CC043
CC025
CC041
CC042

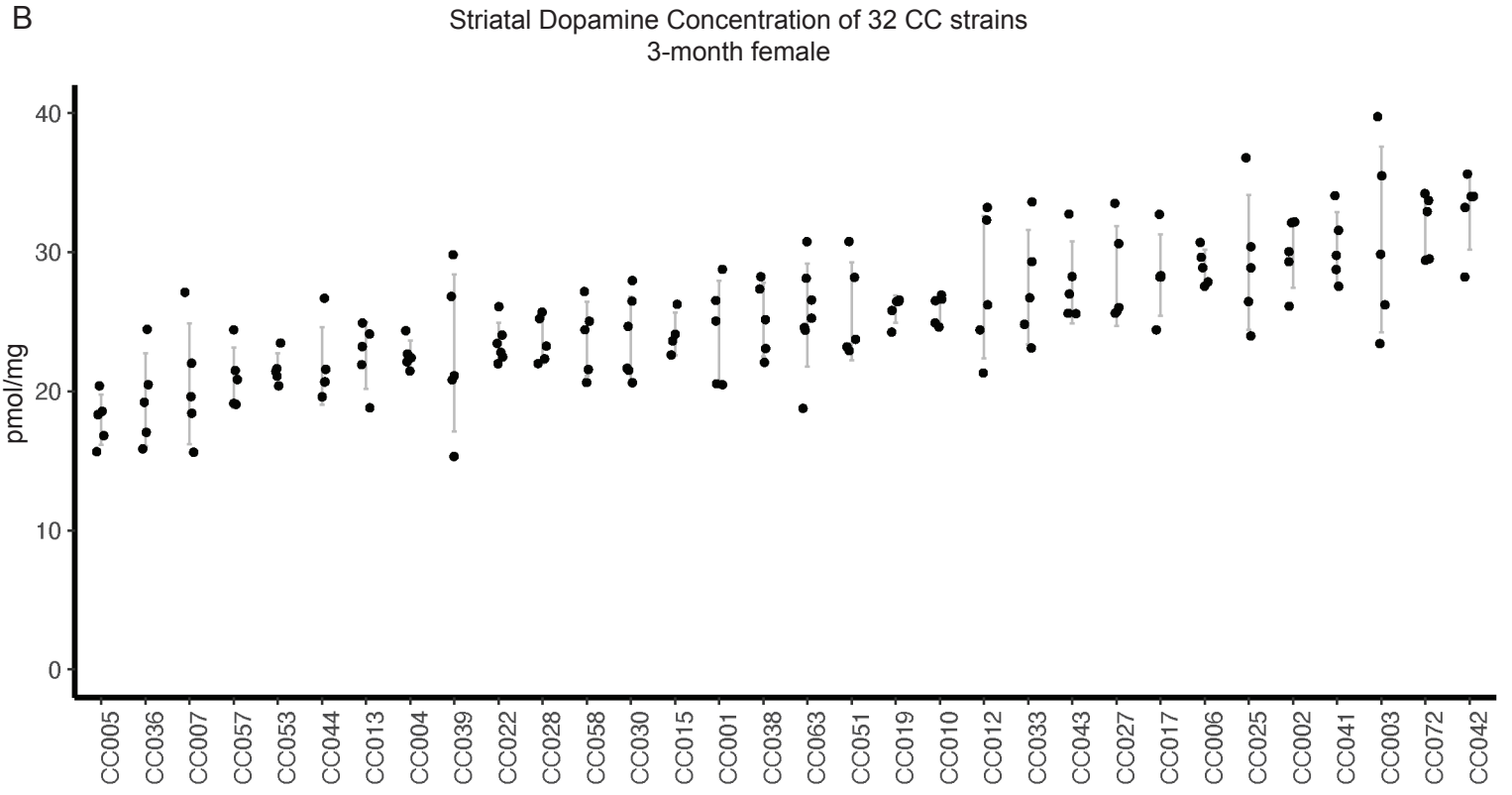
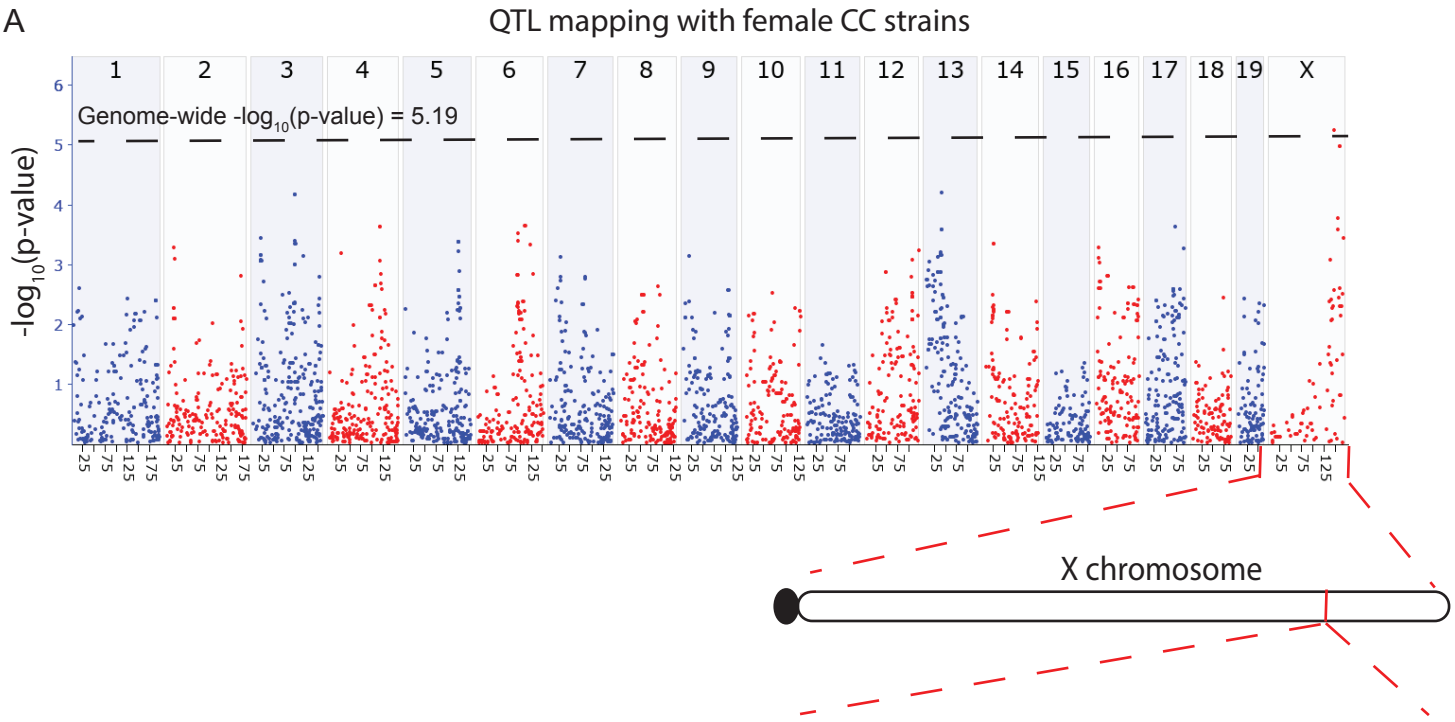
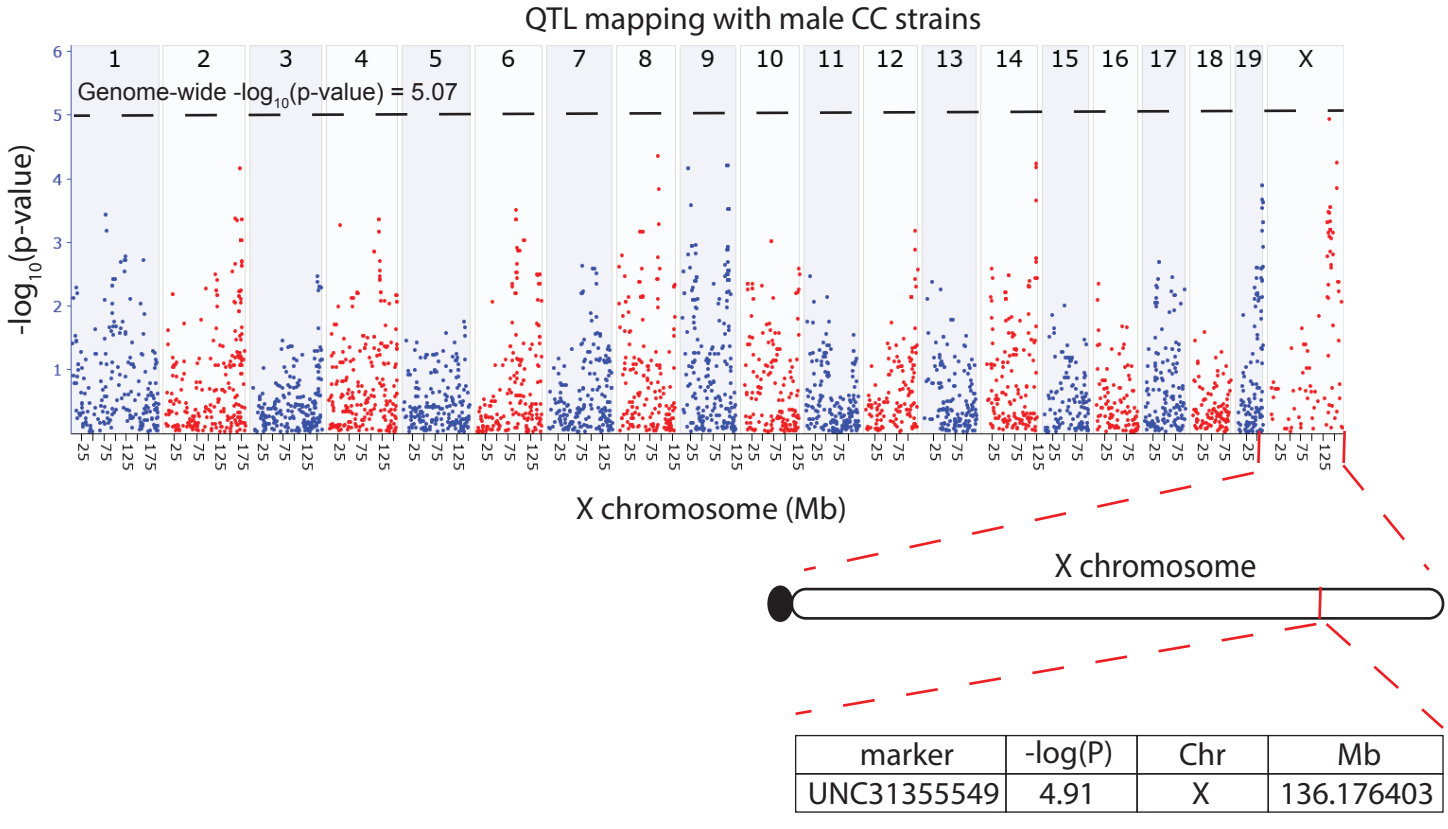

B
Striatal Dopamine Concentration of 32 CC strains
3-month female
pmol/mg
40
30
20
10
0
CC005
CC036
CC007
CC057
CC053
CC044
CC013
CC004
CC039
CC022
CC028
CC058
CC030
CC015
CC001
CC038
CC063
CC051
CC019
CC010
CC012
CC033
CC043
CC027
CC017
CC006
CC025
CC002
CC041
CC003
CC072
CC042

Figure 3

A

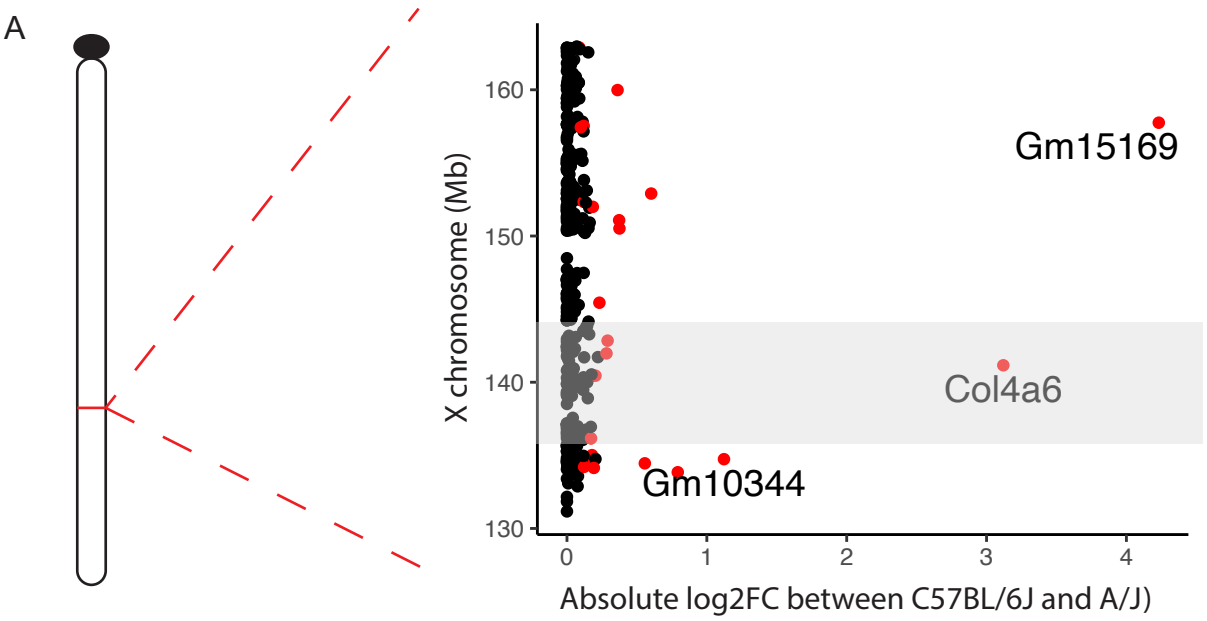


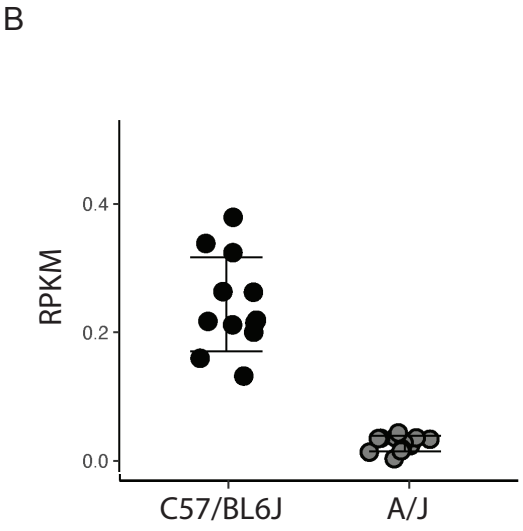
marker	-log(P)	Chr	Mb
UNC31420222	5.23	X	144.300241
UNC31503677	4.96	X	157.823410
JAX00721922	4.96	X	158.259643



marker	-log(P)	Chr	Mb
UNC31355549	4.91	X	136.176403

Figure 4


A
X chromosome (Mb)
160
150
140
130
0
1
2
3
4
Absolute log2FC between C57BL/6J and A/J)
Gm15169
Col4a6
Gm10344


B
RPKM
0.4
0.2
0.0
C57/BL6J
A/J

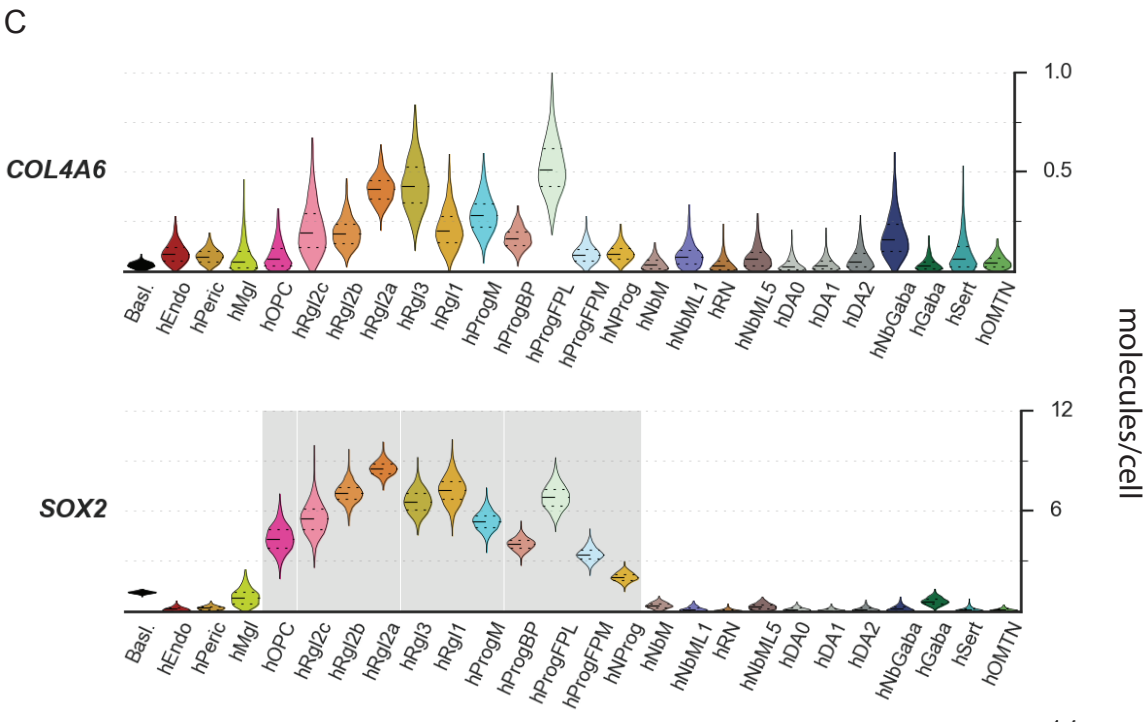
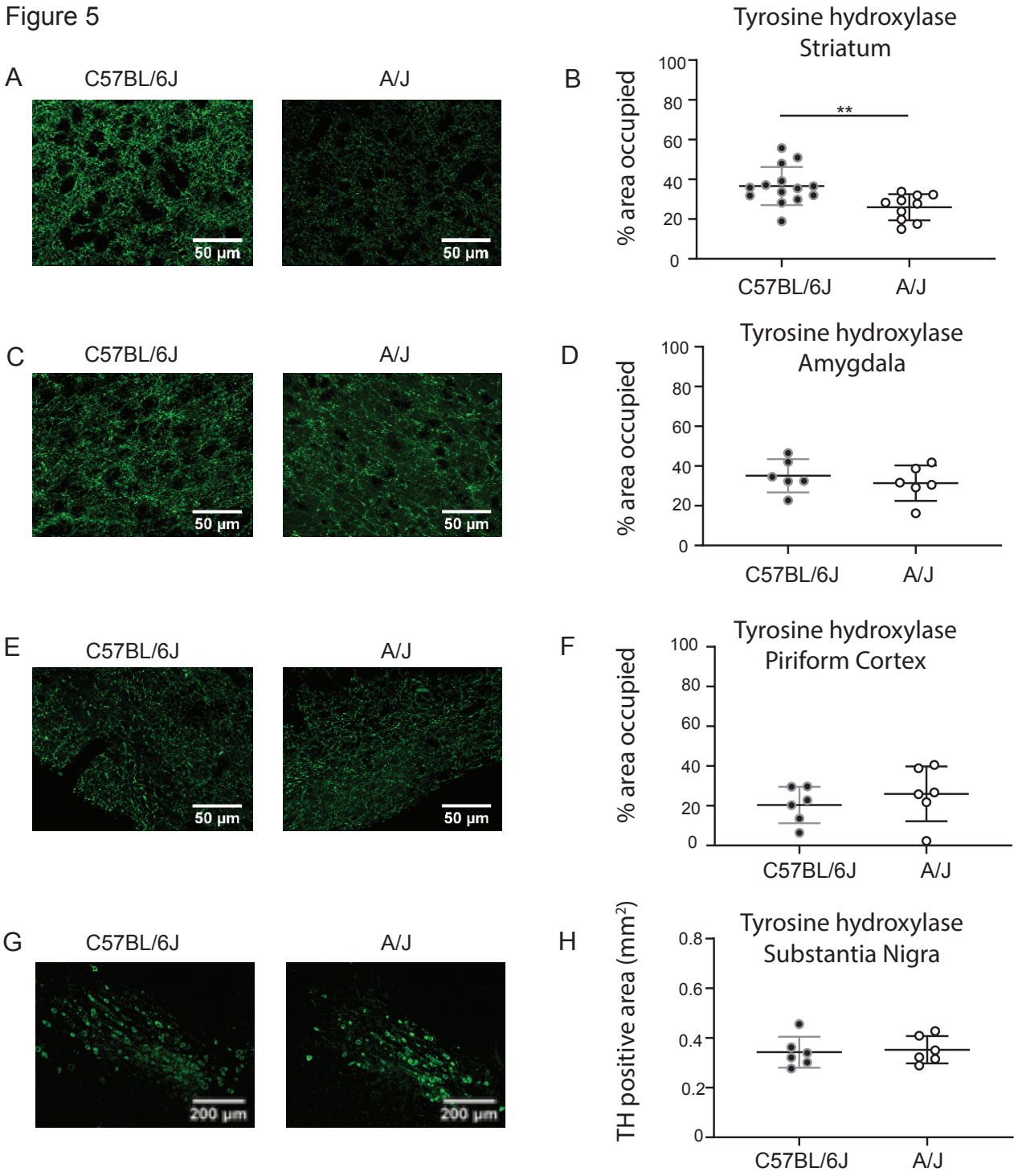

C
COL4A6
SOX2
1.0
0.5
12
6
molecules/cell
Basl.
hEndo
hPeric
hMgl
hOPC
hRgl2c
hRgl2b
hRgl2a
hRgl3
hRgl1
hProgM
hProgBP
hProgFPL
hProgFPM
hNProg
hNbM
hNbML1
hRN
hNbML5
hDA0
hDA1
hDA2
hNbGaba
hGaba
hSert
hOMTN

Figure 5


A
C57BL/6J
A/J
50 µm
B
Tyrosine hydroxylase
Striatum
% area occupied
**
C57BL/6J
A/J
C
D
Tyrosine hydroxylase
Amygdala
E
Tyrosine hydroxylase
Piriform Cortex
F
G
200 µm
H
Tyrosine hydroxylase
Substantia Nigra
TH positive area (mm2)

Quantitative trait locus mapping identifies *Col4a6* as a novel regulator of nigrostriatal dopamine level and axonal branching in mice

Mélanie H. Thomas[1#], Yujuan Gui[2#], Pierre Garcia[1,3,4], Mona Karout[1], Christian Jaeger[1], Zdenka Hodak[1], Alessandro Michelucci[1,5], Heike Kollmus[6], Arthur Centeno[7], Klaus Schughart[6,8,9], Rudi Balling[1], Michel Mittelbronn[1,3,4,5], Joseph H. Nadeau[10,11], Robert W. Williams[7], Thomas Sauter[2], Lasse Sinkkonen[2*], Manuel Buttini[1*]

Supplementation information

Supplementary Tables

Supplementary Table S1: List of mouse strains used in the study. Table S1.a. List of collaborative cross founders. Table S1.b. List of collaborative cross strains.

Supplementary Table S2: List of genetic markers from QTL mapping. The QTL mapping results from males and females are reported.

Supplementary Table S3: List of differentially expressed genes in chromosome X 131 Mb to 163 Mb. The differentially expressed genes are from the comparison in ventral midbrain transcriptomes between C57BL/6J and A/J.

5. Discussion

Complex traits are common in organisms and often disease-associated. Dissecting genetic variation affecting complex traits can help to pinpoint risk variants. Here we focus on ventral midbrain. The ventral midbrain directs a lot of biological functions, such as motor activities and behaviors. Inbred mouse strains have long been shown to differ in motor capabilities (McFadyen et al., 2003; Thifault et al., 2002) and behavioral activities (Laarakker et al., 2011; Yoneyama et al., 2008), indicating these traits are complex and linked to ventral midbrain. Phenotypic difference is largely due to gene expression changes, and such changes are introduced by many *cis* and *trans*-acting regulatory variants. Therefore, to obtain the full genetic makeup of complex traits and disease-related phenotypes, understanding genetic variation underlying ventral midbrain gene expression is indispensable.

PTTG1 as a regulator of ventral midbrain gene expression

To identify genetic variation affecting the expression in ventral midbrain, we first compared the midbrain transcriptomes of three commonly used mouse strains, C57BL/6J, A/J and DBA/2J, which are segregated by ~6 million variants. More than 1000 differentially expressed genes were detected in pairwise comparison, suggesting substantial gene expression differences in ventral midbrains of these mice. Because TFs could have *trans*-regulatory effect affecting the expression of multiple genes, we set out to see if any genes coding for TFs differentially expressed. There were five TF coding genes with altered expression: *Pttg1*, *Npas1*, *Hes5*, *Scand1*, and *Zfp658*, and *Pttg1* was the sole differentially expressed in all three strains. It had high expression in C57BL/6J, moderate in DBA/2J, and almost none in A/J. Its expression changes accompanied with alteration in histone modification, as H3K4me3 signal was less abundant in A/J and DBA/2J comparing to C57BL/6J. Next, we moved on to examine the effect of *Pttg1* on ventral midbrain transcriptome by acquiring *Pttg1*$^{-/-}$ mouse line congenic to C57BL/6J. Interestingly, only four genes were differentially expressed at the age of 3 months between knockout and wildtype: *Gm12663*, *Pttg1*, *Thg1l* and *Ublcp1*. These genes are in close LD, meaning they are likely under the same transcriptional regulation. At the age of 9 months, significant transcriptomic differences was observed with almost 1000 differentially expressed genes found. *Pttg1*$^{-/-}$ transcriptome even shifted towards A/J and DBA/2J, where genes such as *Apoa2* and *Eno1b* contributing the most to transcriptome variation acquired similar expression profiles.

Many differentially expressed genes were found but only 5 genes coding for TFs, suggesting genes encoding for TFs are under-represented. This is in line with results found in plants (Lin et al., 2017). Besides, TFs play an essential role in gene regulation. Considerable defect in TFs is deleterious to organisms. Considering these mice have normal life spans and free of external stimuli, it is unlikely to detect many differences in TF expression.

We found *Pttg1* is the only TF-coding gene altered significantly. *Pttg1* is originally discovered in rat pituitary tumors (Pei & Melmed, 1997). It encodes a mammalian protein, securin. Securin is an inhibitor of separase, a protease required for the separation of sister chromatid in mitosis (Zou et al., 1999). The activation of separase needs the degradation of securin at the onset of metaphase to anaphase transition. In fact, its expression is topped in G1 and M phases (Santos et al., 2015). The biological function of *Pttg1* indicates it can potentially regulate cell cycle. It has been shown in budding yeast that removal of *Pttg1* can lead to early separation of sister chromatids and chromosome instability (Yamamoto et al., 1996). Similar situation was not observed in knockout mouse model, probably due to the fact that separase in vertebrate can be inhibited alternatively by phosphorylation (Mei et al., 2001).

Nonetheless, many studies focused on *Pttg1*'s relationship with tumorigenesis where very often cell cycle is misregulated. In fact, overexpression of *Pttg1* has been found in many human malignancies, including those of the breast (Puri et al., 2001), colon (Heaney et al., 2000), lung (Li et al., 2013) and thyroid (Boelaert et al., 2003). *Pttg1* can also have inhibitory effect on tumorigenesis at least in mammary gland. Ablation of *Pttg1* found precocious branching morphogenesis and the mutant female mice can spontaneously develop mammary gland tumors, corresponding to the observation that PTTG1 protein level was down-regulated in human breast tumors (Hatcher et al., 2014).

The reason why *Pttg1* is involved in cancer progression is partly due to its nature as a TF associated with genes directing the cell cycle. PTTG1 is broadly activated and functioning through direct DNA binding or interaction with other regulatory factors (Tong and Eigler, 2009). One of the downstream targets of PTTG1 is *c-Myc* which can regulate growth and cell cycle entry (Hamid and Kakar, 2004). *Fgf2* is a mitogen that induces proliferation, differentiation and migration of cells. It coincides with *Pttg1* overexpression in tumor angiogenesis. In addition, PTTG1 can interact with SP1, forming a complex to co-localize the

promoter of *Ccnd3* to activate transcription of genes responsible for G1/S phase transition (Tong et al., 2007).

Pttg1 expression was highly downregulated in A/J, with expression dropping down to almost none comparing to 15 RPKM on average in C57BL/6J and 3 RPKM in DBA/2J. An interesting question is if the varied levels of *Pttg1* correlates with tumor susceptibility considering *Pttg1* expression level has been shown to link with several cancers. So far, there is no study focusing on the contribution of *Pttg1* on tumor susceptibility in different mouse strains, despite the fact that inbred mouse strains do show variability when exposed to carcinogens. For example, A/J is highly susceptible to lung cancers, whereas C57BL/6J and C3H are relatively less sensitive (Malkinson, 1989). Considering these three mouse strains have normal life spans and are free from extra treatment, the intrinsic level of *Pttg1* is unlikely to cause natural tumorigenesis.

It is curious why *Pttg1* expression is so low in A/J. A study using a panel of AXB/BXA strains (developed by crossing C57BL/6J with A/J) to identity prospective genes participating in the patterning of retinal amacrine cells shed some lights on this topic (Keeley et al., 2014). Variation in cell mosaic patterning was observed across AXB/BXA strains, and the identified QTL on chromosome 11 yielded *Pttg1* as the most promising candidate. The high expression of *Pttg1* in C57BL/6J is due to a 7-nucleotide deletion in immediate upstream of its TSS. The deletion creates a binding site for AP-1 transcription factor, which usually associates with gene activation. The gene expression changes could also be observed in our ChIP-seq data, with H3K4me3 signal being highly depleted in the TSS of *Pttg1* in A/J. Interestingly, *Pttg1* knockout mice had a reduction in cell mosaic patterning in retina. Similar phenotype could be also found in A/J.

At the age of 3 months, only 4 genes differentially expressed including *Pttg1*, *Thg1l*, *Ublcp1*, and *Gm12663*. However, at the age of 9-13 months, almost 1000 genes showed expression changes. Importantly, *Pttg1*$^{-/-}$ midbrain transcriptome shifted to A/J and DBA/2J. The observation suggested aging might participate in the effect of *Pttg1* ablation. Lifespan differences have been recorded for these mouse strains. C57BL/6J is long-lived with 50% mortality in captivity by 914 days, while DBA/2J and A/J both have a shorter life span (Goodrick, 1975). Much effort has been invested to understand the genetic underpinning of lifespan differences between the long-lived C57BL/6J and the short-lived DBA/2J. Previous study demonstrated that the percentage loss of murine hematopoietic stem and progenitor cells

(HSPC) upon genotoxic agent hydroxyurea (HU) treatment is inversely correlated with the mean lifespan of inbred mice (Schütz et al., 2017). C57BL/6J has low frequency of HSPC dysfunction in response to HU comparing to DBA/2J. A QTL in chromosome 11 is linked to the percentage of dysfunctional HSPC and the regulation of life span in BXD mice (mice produced by crossing C57BL/6J and DBA/2J) (Brown et al., 2020). Further analysis with mice that are reciprocally congenic with either C57BL/6J or DBA/2J background to this locus pointed out that *Pttg1* is the most likely QTL gene. The underlying mechanism is still unknown, but it might linked to CpG methylation as overexpression of *Pttg1* in C57BL/6J HSBC acquired a faster age-associated DNA methylation. What contradicts to our result is that the expression of *Pttg1* in HSPC is higher in DBA/2J comparing to C57BL/6J, but we saw the opposite in our data, suggesting *Pttg1* has a cell type-specific expression pattern. In addition, striatum and retina also have higher expression of *Pttg1* in DBA/2J but not C57BL/6J (Bottomly et al., 2011; Geisert et al., 2009), further prompting us to suspect the effect of *Pttg1* might not be the same in ventral midbrain as in other tissues and there might be an upstream regulator on *Pttg1* expression specific to ventral midbrain. However, the complexity of these topics is beyond the scope of this thesis.

The ablation of *Pttg1* in C57BL/6J leads to midbrain transcriptome shifting towards A/J and DBA/2J along aging, but which type of cells being affected is unclear to us. Based on single cell RNA-seq data on mouse ventral midbrain (Saunders et al., 2018), *Pttg1* expression can be found in various cell types including ependyma, mural cells, endothelial cells, oligodendrocytes, and microglia. This observation is in line with our deconvolution results. Taking together, *Pttg1* is possibly involved in the transcriptional changes of multiple cell types. However, not many studies so far focus on the function of *Pttg1* in the brain. Except for its involvement in retinal mosaic patterning, *Pttg1* is downregulated in a hearing-impaired mouse model proposing contribution to presbycusis (age-related hearing loss) (Lubka-Pathak et al., 2011). *Pttg1* knockout mice showed impairment in spatial learning (Manyes et al., 2018), suggesting a defect in hippocampal system. Another study also found *Pttg1* expression is associated with neocortex volume (Gaglani et al., 2009), proposing its potential involvement in brain development.

An interesting study using F2 intercross mice from C57BL/6J and C3H to map 23 000 gene expression traits in whole brain identified *Pttg1* as the strongest brain cis eQTL (Lum et al., 2006). Several genes have similar expression patterns with *Pttg1*, indicating they might be the

possible downstream targets. To validate *Pttg1* is casual to the expression changes in these genes, they compared transcriptomic profiles between *Pttg1*$^{-/-}$ and wildtype animals and found five genes differentially expressed. We can perform similar analysis to find genes that are changed as a result of *Pttg1* downregulation, by coupling information from knockout mouse line and BXD strains. Our PCA showed *Pttg1*$^{-/-}$ midbrain transcriptome moves towards A/J and DBA/2J. So in theory, mapping with BXD strains can uncover QTLs explaining the expression phenotypes on midbrain transcriptomes between the two strains. QTL mapping with 37 BXD ventral midbrain transcriptome profiles using GeneNetwork (Druka et al., 2008) found a strong QTL on chromosome 11 with *Pttg1* locus in, corresponding to previous study on mouse whole brain (Lum et al., 2006). However, few expression changes can be detected at 3-month, whereas Lum et al., using whole brain from 3-5 months found 1483 differential genes. This is likely because transcriptome changes as a result of *Pttg1* removal might happen in other brain regions prior to ventral midbrain. Instead of looking at 3 months old knockout mice, we switched to animals upon 9-month old where a lot of gene expression changes can be found. Eleven differentially expressed genes, such as *Gabra2* and *Mcee,* were in high expression correlation with *Pttg1* in BXD strains. *Gabra2* is likely under *cis*-regulation as it is closed to *Pttg1*. *Mcee* is possibly under *trans*-regulation and it was shown to have negative genetic interaction with *Pttg1* (Vizeacoumar et al., 2013). However, such results should be taken with caution, as the QTL mapping was performed on BXD mice at 3-month old, but the validation was done with older animals where age might introduce confounding noise.

When looking at genes that highly contributed to the midbrain transcriptome shifting, we found their expression patterns were similar to A/J and DBA/2J. *Apoa2* has high expression in C57BL/6J, but after removing *Pttg1*, its expression was downregulated to a similar level in A/J and DBA/2J. A study showed that *Apoa2* is deterministic for apnea in C57BL/6J, by showing *Apoa2*$^{-/-}$ *and Apoa2*$^{+/-}$ mice do not exhibit symptoms anymore (Gillombardo et al., 2017). Such mechanism could also exist in *Pttg1*$^{-/-}$ mice. On the other hand, *Eno1b*, an isoform of the glycolytic enzyme enolase, overexpressed in α-crystallin mutant mice with defect in eye lens (Andley et al., 2018), which could also link to the retinal mosaic patterning defects of *Pttg1*$^{-/-}$ mice. Taking together, *Pttg1* ablation can globally affect ventral midbrain transcriptome, with affected genes showing biological relevance being in line with previous publications.

The analysis was fully relied on RNA-seq on ventral midbrains from different mouse strains, where many factors could potentially bring bias. Though minor difference in the dissection of

ventral midbrain is unavoidable, handling of animals was performed by professionals to minimize animal-to-animal variation. Other factors that could bring bias are time of experiments and different facilities for sequencing. Two procedures were included in the analysis pipeline to account for sample variation in general. Firstly, all samples were upper-quartile normalized. The normalization managed to bring the mean to zero and the variance comparable across samples. Secondly, wald test was performed on individual potential covariates, including sex and difference in sequencing facilities. Facility difference could bring significant bias in downstream analysis, while few genes (25 genes) were differentially expressed between males and females, including ones that are known to differ in sex like *Xist*. As a result, sequencing facility was included in statistical model to account for its bias. Nonetheless, variation within a strain could be detected. One of DBA/2J 9-month samples had high expression of genes specific to pituitary gland, suggesting potential tissue contamination coming from dissection. Considering this sample was grouped with other DBA/2J 9-month animals on PCA and our analysis mainly focused on 3-month old mice, we did not exclude this sample from downstream analysis. Besides, expression profiles from *Pttg1*$^{-/-}$ 9-13 months old were considerably varied comparing to other strains, which is likely caused by variable expressivity from the effect of *Pttg1*$^{-/-}$ knockout.

Ventral midbrain transcriptome is a complex trait where many *cis* and *trans* regulatory factors can together contribute. Our data found *Pttg1* is a potential *trans*-regulator affecting gene expression phenotypes in this brain region by showing its ablation leads to transcriptome shifting along aging.

Single nuclei chromatin profiles reveling cell identity TFs and cell type-specific gene regulatory variation

Ventral midbrain is a mixture of many different types of cells. The development and function of each cell requires the appropriate set of genes expressed. Cell type-specific gene expression relies on the orchestration conducting by *cis* and *trans*-regulatory elements. Our tissue level transcriptome revealed *Pttg1* as the potential *trans*-regulator affecting the expression phenotypes in ventral midbrain, but information about cell type-specific variants between strains are still missing.

Therefore, we generated single nuclei chromatin profiles on the ventral midbrains of C57BL/6J and A/J. Such datasets have not yet been published and provide a rich resource for the community to study regulatory elements such as promoters and enhancers in a cell type-specific manner. The reason that we would like to look at chromatin profiles in cell level is many studies have shown strain-specific genetic variation can affect gene expression in a given cell type by influencing TF binding or *cis* regulation (Alasoo et al., 2018). Macrophage has specific lineage-determining TFs (LDTFs) including PU.1 and C/EBPs (Heinz et al., 2010). Characterization on collaborative binding of PU.1 and C/EBPs in macrophages derived from C57BL/6J and BALB/cJ mice found hundreds of strain-specific binding sites (Heinz et al., 2013). Strain-specific variants in PU.1 motif lead to loss of PU.1 binding as well as C/EBPs binding, though C/EBPs motif is intact. A subsequent study on bone-marrow-derived macrophages (BMDM) from five inbred strains restated strain-specific genetic variation has regulatory effect (Link et al., 2018). Interestingly, their study also showed genetic variation primarily alters transcription through distal *cis*-regulatory elements. Taking together, with the help of single nuclei chromatin profiles, we could identify *cis* and *trans* genetic variants on cell types of interest in ventral midbrain.

The method we used to generate chromatin profiles in cell level is the single nuclei ATAC-seq (snATAC-seq) developed by 10X Genomics. snATAC-seq enables the profiling of epigenetic landscapes of thousands of individual cells. Based on accessibility on genome, nuclei could be grouped into several clusters. However, it is unfeasible to label clusters based on their epigenomic signatures. There are several reasons for this. Because different cells could adopt different sets of TFs to control gene expression, software like cellranger looks at which motifs are enriched at each cluster. But some TFs like CTCF are shared by multiple cell types, so it is not easy to distinguish clusters based on motif enrichment. Besides, open chromatin region does not guarantee actual gene expression, as many additional factors, like TF binding and histone modifications, can also regulate gene expression (de la Torre-Ubieta et al., 2018).

To efficiently label nuclei clusters, we took use of the existing scRNA-seq data generated on mouse ventral midbrain in C57BL/6 background from the matching age (Saunders et al., 2018). Integration of scRNA-seq and snATAC-seq was done by Signac (Stuart et al., 2019). The idea is based on the positive correlation between gene expression and accessibility at a gene's promoter (Lara-Astiaso et al., 2014). We can first identify signature genes, genes that vary across cell types, in scRNA-seq, then look at the accessibility of these signature genes in nuclei

clusters. With this approach, we managed to identify ten major cell types in ventral midbrains from both mouse strains: endothelial tip, endothelial stalk, astrocyte, microglia, polydendrocyte (Tnr, Cspg5), polydendrocyte (Tnr), Oligodendrocyte (Tfr), neuron, neuron (Th), and neuron (Slc17a6). The proportions of cells are similar between C57BL/6J and A/J, with more than half of the nuclei being assigned to neuronal population. As expected, the accessibility of marker genes is different in cell types. *Tmem119* is a microglia marker gene. Its accessibility in microglia is uniquely high with signal ranging from TSS through gene body and expand to distal region. This high accessibility corresponds to its high expression only in microglia. Noticeably, genes like housekeeping gene *Rpl13a* that do not have a cell type-specific expression, would have their TSS accessible for most of the cells.

When looking at nuclei clusters between the two mouse strains, Th+ neurons in C57BL/6J did not cluster as a whole as it did in A/J. This is likely due to very low number of Th+ neurons found in C57BL/6J, with only 45 nuclei comparing to 196 in A/J. Similar difference in nuclei number can be found in Slc17a6+ neuron, where A/J also got more nuclei. Considering there was only one replicate from each strain, it is impossible to tell if the difference in nuclei number between the two strains is due to biology or technical bias. In addition, The neuronal cluster looks quite spread (even A/J has one part of the neuronal cluster separated from the main), suggesting it is possible to distinguish neuronal subtypes based on epigenomic profiles.

As regulatory elements in open chromatin can affect gene expression, selective chromatin profiles are important for transcriptional control in a given cell type. To identify such chromatin profiles in cell level, we again leveraged the public available scRNA-seq data on ventral midbrain (Saunders et al., 2018). By comparing gene expression across cell types, we could select genes that are highly expressed in a given cell type. These genes are defined as cell type-identity genes which are likely to shape cell specification. Identity genes overlap with marker genes to some degree but might differ in expression pattern, as they are strictly defined to express in one single cell type. For example, *Slc6a3* is one of the top enriched cell type-identity genes for Th+ neurons with expression highly specified. Similar cases could be found in microglia with *C1qa* and in astrocyte with *Slc1a2*. We did notice that cell types from close lineages tend to share gene expression. Though our expression cutoff specified *Atp1b2* to be highly abundant in neuron comparing to other types, Th+ neuron still had certain level of expression. The number of cell type-identity genes was ranging from 47 in polydendrocyte (Tnr, Cspg5) to 412 in Th+ neuron. The large discrepancy in identity gene numbers across cell

types could be due to both the technical bias and their biological nature. Our approach to define cell type-identity genes was a heuristic approach. To setup the expression cutoff, 100 cells were selected from every cell group to have information equally coming from each cell type. Given the fact that many more cells were available in some groups like neuron or oligodendrocyte, it is possible that certain bias was introduced into the pipeline. To overcome the technical bias generated from cell selection, we repeated our pipeline for 100 times and defined genes appearing more than 30 times out of the 100 as cell type-identity genes. On the other hand, cell types like Th+ neurons tend to have distinct expression pattern considering their unique biological functions like dopamine secretion, so intuitively they would adopt more identity genes. As identity genes were highly expressed only in the corresponding cell type, next we moved on to see if they related with cell function by performing gene ontology (GO) enrichment analysis (Chen et al., 2013). Cell type-specific GO terms are enriched in corresponsive identity genes. Therefore, the defined identity genes do bear biological relevance despite technical bias in the pipeline. With the set of cell-type identity genes outlined from scRNA-seq, we could associate open chromatin to these genes. Peaks that fall into the regulatory regions of cell-type identity genes were defined as cell type-identity peaks. Cell type-identity peaks are selective chromatin profiles with *cis*-regulatory elements affecting the expression of identity genes.

The defined identity peaks were enriched in the matching cell types. Interestingly, cells from different lineages tend to have stronger less enrichment on peaks that are not specific for them. For example, microglia arise at very early stages of embryonic development from progenitors in embryonic yolk sac (Ginhoux et al., 2010; Han et al., 2018). Thus, microglia come from a different origin compared to other brain cells. The accessibility in microglia was strongly depleted in identity peaks of other cell types. As TFs could have profound influence on gene expression, we moved on to search for motifs that were enriched in cell type-identity peaks. Many of the found motifs were closely related to cell identity. For instance, the motif of ASCL1 was one of the top-ranked binding sites in the cell type-identity peaks of astrocyte. ASCL1 was shown to regulate astrocyte differentiation by activating notch-signaling pathway (Henke et al., 2009). The notch signaling also upregulates the expression of *Sox9* to induce astrogenesis, the motif of which was also found in astrocyte (Martini et al., 2013). SPI1 (or PU.1), binding to the top enriched motif in microglia, is a master regulator of myeloid cells and controls microglial development and function (Smith et al., 2013). Some motifs are shared by more than one cell types, such as RFX1 in astrocyte, neuron and Th+ neuron. The expression of *Rfx1* was

mainly found in the nuclei of neurons and microglial cells but not in astrocytes. However, another study in monkey discovered RFX motif enrichment in NPCs (neuronal precursor cells) and astrocytes (Goodnight et al., 2019), which corresponded to our results. In addition, CTCF enrichment was present commonly, probably because its binding on insulators are required broadly to maintain heterochromatin boundaries and itself has critical roles in transcriptional regulation (Chen et al., 2012). Some well-defined TFs were not found based on our cell type-identity peaks, such as NURR1 and EN1/2 for Th+ neuron. This is probably due to the analysis used a subset of peaks associated with genes having high expression in one cell type. Taking together, integration of scRNA-seq and snATAC-seq from ventral midbrain revealed cell identity TFs defining cell identity.

To discover the regulatory effect from strain-specific variation, we compared the open chromatin profiles between C57BL/6J and A/J for each cell type to find peaks with differential accessibility. Substantial amount of differential peaks with reads more than 2 fold could be identified from each cell type, ranging from 498 in neuron to 17 833 in Th+ neuron. The number of differential peaks was biased inversely to nuclei number, as Th+ neuron had the least cells but the most differential peaks. To test their biological relevance, differential peaks were assigned to the regulatory regions of differential genes from our previous bulk RNA-seq data. Differential peaks were highly enriched in the regulatory regions of genes with varied expression, suggesting they harbored cis-regulatory elements despite certain level of technical bias. Indeed, peaks with certain fold difference could reflect gene expression change. For example, the expression of *Isoc2b* was detected in C57BL/6J but not in A/J, corresponding with accessible signal only found in its TSS in C57BL/6J. Additionally, the differential peaks were also associated with cell type-specific expression. *Olfr287* had accessibility in its TSS and gene expression specifically in astrocyte of C57BL/6J. Such information could not be dissected out on tissue-level data. In summary, comparing accessibility of each cell type between strains can identify specific differential peaks, and they could reflect gene expression changes found in tissue-level transcriptomes.

Some methods with stricter statistical definition could also be used to identify differential peaks. One paper with snATAC-seq on mouse cortex used Cicero package to create many clusters of cells based on their low-dimensional t-SNE coordinates. Then the accessibility profiles in each cluster could be aggregated as one replicate (Sinnamon et al., 2019). With this approach, sufficient replicates are created as input to DESeq2 (Love et al., 2014), which applies Wald

Test to find sites differentially accessible between cell types or against all other cell types. Such computational framework should be also explored in our dataset. But given the time when the manuscript was written, it has not been performed yet and will be implemented later.

Gene expression phenotype is greatly contributed by *cis*-regulatory variants (Link et al., 2018). To identify putative regulatory variants with *cis*-acting effect, we generated tissue-level H3K27ac ChIP-seq to find enhancers and promoters, as well as ATAC-seq to search for DNA-binding sites. Putative *cis*-acting regulatory variants were defined by locating in enhancer regions and DNA binding sites simultaneously. They were highly enriched in the regulatory region of differentially expressed genes, prompting us to leverage our single cell accessibility profiles to identify cell type-specific *cis*-regulatory variants. Indeed, most of the regulatory variants showed differential accessibility across cell types, with almost 40% accessible in less than 6 cell types. Some of these regulatory variants could potentially affect gene expression in a cell type-specific manner. *Zfp68*, coding a zinc-finger protein that could repress gene expression (Agata et al., 1999), was down regulated in A/J. Putative regulatory variants were differentially accessible in the TSS of *Zfp68* in Th+ neuron , resulting in the depletion of ATAC signal in A/J and ultimately downregulating *Zfp68* expression. In short, putative regulatory variants have differential accessibility across cell types, and they can have *cis*-effect on gene expression in a strain-specific manner.

Except for *cis*-regulatory variation, potential *trans*-acting regulators could introduce expression difference in hundreds of genes in a given cell type (Liu et al., 2019). Differentially accessible peaks between C57BL/6J and A/J were enriched in the regulatory region of differentially expressed genes, indicating they harbored regulatory elements that are bound by *trans*-regulators like TFs. To identity such *trans*-regulators, we performed motif enrichment analysis on cell type-specific differential peaks. In neurons, we found an enrichment of the shared binding motif for TCF7L2 and LEF1. LEF1 and TCF7L2 are members of the TCF family and they closely associated with the Wnt signaling pathway (Jin and Liu, 2008). When Wnt signaling pathway is activated, β-catenin in cytosol would not be degraded by proteasome-mediated degradation process but enters into nucleus to form the β-catenin/LEF1 complex (also referred as the canonical Wnt signaling pathway (Bem et al., 2019)) or β-catenin/TCF7L2 complex. The complex can activate the downstream genes of Wnt signaling pathway (He et al., 1998). Polymorphisms on *Tcf7l2* is associated with the disease risk of type 2 diabetes by affecting glucose metabolism in pancreas and liver (Facchinello et al., 2017). In brain, *Tcf7l2*

expression could be found in thalamus and midbrain (LEE et al., 2009), suggesting its potential involvement in brain function. Studies found that *Tcf7l2* is essential to establish neuronal identity and circuits of the caudal forebrain by affecting post-mitotic neuronal identity switch between thalamic and habenular neurons (Lee et al., 2017). In addition, the Wnt signaling pathway regulated by *Lmx1b-miR125a2* regulatory loop can modulate the size of the midbrain dopaminergic progenitor pool (Anderegg et al., 2013). scRNA-seq on ventral midbrain (Saunders et al., 2018) found *Tcf7l2* was highly expressed in neurons and moderately in polydendroctye. However, the expression of *Lef1* was only detectable in endothelial stalk. Because the binding sites for TCF7L2 and LEF1 are almost identical, it is impossible to distinguish their cellular activity based on open chromatin regions. By combining both scRNA-seq and sn-ATACseq, we could decipher that TCF7L2 but not LEF1 is the one with altered chromatin accessibility in midbrain neurons between C57BL/6J and A/J.

The neuronal population in Saunders et al. was mainly composed of *Gad2*+ and *Slc17a6*+ neurons, which were likely to be GABAergic neurons and glutamatergic neurons, respectively (Mickelsen et al., 2019). Both GABAergic neurons and glutamatergic neurons are associated with the ventral tegmental area (VTA) in ventral midbrain. VTA involves in one of the dopaminergic circuits, the mesocorticalimbic circuit which controls behavioral traits like reward and fear. Alterations in the activity of TCF7L2 were linked to neuropsychiatric disorders in human such as schizophrenia (Bem et al., 2019) and bipolar disorder (Cuellar-Barboza et al., 2016). Mice model with *Tcf7l2* expression altered showed behavioral changes in anxiety and fear learning (Savic et al., 2011). Interestingly, C57BL/6J and A/J are also known to differ in anxiety with A/J behaving more anxious (Laarakker et al., 2011). Based on the distinct expression pattern of *Tcf7l2* in neurons and its altered accessibility between C57BL/6J and A/J, it is likely that TCF7L2 can potentially contribute to phenotypic difference in behaviors in the two mouse strains.

How genetic variants shape the phenotypes of complex traits has long been discussed. GWAS studies have been used to study the genetic basis of many complex traits, but the genome-wide significant hits in combination can only explain a small fraction of the expected heritability (Manolio et al., 2009). This is known to be the missing heritability problem, which is due to large numbers of small-effect common variants cannot pass stringent significance test at the current sample sizes (Loh et al., 2015; Shi et al., 2016a; Yang et al., 2010). A recent model suggested that phenotypic variance is mediated through many genes, including genes that are

not directly involved in the trait in question (Boyle et al., 2017). Based on this concept, they proposed to divide genes into core genes and peripheral genes based on if they can affect a given trait directly or not (Liu et al., 2019). A *cis*-variant for a peripheral gene could be a *trans*-QTL for core genes. To globally explore *cis* and *trans*-regulatory variants affecting the expression phenotype in ventral midbrain, we integrated our snATAC-seq with existing scRNA-seq, as well as tissue level epigenomic profiles. Such integration revealed cell type-specific *cis* and *trans*-regulatory variants that could direct strain-specific gene expression changes, the information of which helps to fill up the missing heritability in gene expression phenotype of ventral midbrain. In addition, we discovered many potential regulatory variants that could direct or indirect contribute to our trait of interest.

In summary, to understand *cis* and *trans*-regulatory variation in cell type level, we generated single cell chromatin accessibility profiles in the ventral midbrains of C57BL/6J and A/J. Integration with public available scRNA-seq found enrichment of cell type-identity motifs. Putative *cis*-regulatory variants had differential accessibility across cell types and could selectively affect gene expression. Analysis at regions with differential accessibility revealed *trans*-acting variation that could potentially affect strain-specific gene expression. Our data provide a rich resource to study cell type regulatory variation in ventral midbrain and opens new venue to identify strain-specific variation.

Bridging ventral midbrain transcriptome variance to associated phenotypes

The transcriptome of ventral midbrains between C57BL/6J and A/J are different, and such expression difference are under control by many regulatory variants in cell type-specific manner. Here we would like to focus on one cell type, the dopaminergic neurons in nigrostriatal circuit (mDAn), to see how gene expression difference contributes to phenotypic variation.

It has long been suspected that the original dopamine level, either caused by different cell numbers or varied intrinsic dopamine release, can affect individual risk in neurodegenerative disease. This is inspired by many studies that found inbred mouse strains have considerable differences related to TH activity (Ciaranello et al., 1972), DAn numbers (Muthane et al., 1994; Ross et al., 1976) and striatum size (Rosen and Williams, 2001). CD-1 mice have a higher number of mDAns comparing to C57/BL, while in BALB/cJ and CBA/J, the number of mDAns did not vary but TH activity is much higher in BALB/cJ, suggesting dopamine level in dorsal striatum is a quantitative trait which can be contributed by cell number and enzyme activity. In

fact, the degenerative process of mDAns is triggered several years before there is sufficient motor symptoms for clinical diagnosis. If a person has less dopamine as a starting point, it is likely that the "tipping point" is reached earlier to cross the critical threshold for motor symptoms (von Linstow et al., 2020).

Therefore, we set out to look at the dopamine concentration on dorsal striatum as phenotypic readout because of its biological relevance to mDAn. mDAn have cell bodies in the SN of ventral midbrain, and their fibers can long project into dorsal striatum to release neurotransmitter dopamine which directs the motor function of organisms. mDAn degeneration causing dopamine deficiency in dorsal striatum is one of the hallmarks of Parkinson's disease, the notorious neurodegenerative disease affecting 2-3% of the population beyond 65 years of age (Poewe et al., 2017). If mDAn contributes to the overall gene expression difference we observed in C57BL/6J and A/J, it is likely that the transcriptome phenotypes in ventral midbrain can lead to detectable trait difference, such as varying levels of dopamine in dorsal striatum.

To understand the genetic background of dopamine levels in inbred mouse strains, we measured the dopamine concentration on dorsal striatum of eight mouse lines: PWK/PhJ, A/J, NOD/LtJ, WSB/EiJ, 129S1/SvlmJ, C57BL/6J, CAST/EjJ, and NZO/HiLtJ. They differed considerably in dopamine concentration, with around 1.5-fold difference between the lowest level in PWK/PhJ and the highest in NZO/HiLTJ, suggesting dopamine concentration in dorsal striatum is a quantitative trait.

Next, we performed QTL mapping with CC strains on dorsal striatum dopamine level to identify genomic locus that could be associated to our trait of interest. The CC strains are designed specifically for the study of complex traits by incorporate genetic variation from the above eight founder strains to maximize genetic diversity (The Complex Trait Consortium et al., 2004). With more than 30 CC strains and 5 replicates each, one can achieve mapping power more than 0.8 on QTLs with relatively large effect size (Keane et al., 2011). Thus, 32 CC strains with 10 replicates each (5 for male and 5 for female) were used in our study. The dopamine measurement on the dorsal stratum across 32 CC strains also had considerable difference, indicating dopamine level is an inheritable trait. Only two strains (CC006 and CC072) showed significant sex difference, both with females having slightly higher dopamine levels. Hence, the QTL mapping was performed separately on each sex.

Interestingly, the top hits of QTL mapping in both sexes were located on chromosome X within 8 Mb away from each other. Female QTL mapping yielded one locus survived from genome-wide significance penalty: UNC31420222 at 144 Mb of chromosome X. To select the genes of interest in proximity with the QTL hit, we looked at the differentially expressed genes between C57BL/6J and A/J, two out of the eight founder strains of CC lines. The hypothesis is that, if a QTL is shown to associate with a trait of interest, genes responsible for phenotypic difference in the QTL would have varied expression profiles. *Col4a6* was the only gene with 9-fold expression change between C57BL6/J and A/J after strict cutoff. Its expression pattern also corresponded to the dopamine measurement in the two founder strains with AJ having 15% less dopamine comparing to C57BL/6J.

Col4a6 encodes for collagen alpha-6 (IV) chain protein which is one of the six subunits of type IV collagen. Type IV collagen is the major structural component of basement membrane. Basement membrane is a thin layer of extracellular matrix (ECM) between epithelial and stromal cells (Ingber, 2003). It provides structural support for cells with unique components to conduct important biological functions like differentiation, proliferation and migration during development. Given collagen alpha-6 (IV) chain is one of the components required for proper ECM formation, any defect on *Col4a6* could potentially cause detrimental effect. This is supported by a study using zebra fish model to show *Col4a6* can affect the integrity of basement membranes which gives support for axogenesis (Takeuchi et al., 2015). Granule cells are the major type of glutamatergic neurons in cerebellum. Functional cerebellar circuits fully rely on the axon formation of granule cells. *Col4a6* reduction on zebrafish resulted in abnormal axogenesis coupling with disorganized basement membrane surroundings, suggesting that type IV collagen controls neuron axogenesis by regulating the integrity of the basement membrane during development. Several other studies also linked collagen coding genes to neuron formation. One study found α-synuclein associated with *Col4a2*, the gene coding subunit collagen alpha-2 (IV) chain protein (Paiva et al., 2018). *Col4a2* was upregulated both in transgenic mice and DAn with α-synuclein overexpression, suggesting collagen related genes might play a role in α-synuclein-induced toxicity. In addition, *Col4a6* was down regulated together with *Sox2*, a gene indispensable for neurogenesis (Ferri et al., 2004). This is likely due to SOX2 can bind to the promoter of *Col4a6* and regulate its expression (Berezovsky et al., 2014). Additionally, the expression patterns of *Col4a6* and *Sox2* during human ventral midbrain development were similar (Manno et al., 2016), suggesting they two could cooperate during neurogenesis.

Interestingly, our data found A/J has less DAn fibers in dorsal striatum comparing to C57BL/6J, which could be due to A/J bearing either smaller mDAn number, or less fiber branching. Analysis on amygdala, piriform cortex, and SN did not find difference in mDAn number between C57BL/6J and A/J; thus, the observation of less DAn fibers of A/J was likely due to less branching. Less branching in A/J could lead to lower dopamine concentration. Given the fact that *Col4a6* can affect axon formation, we suspect similar scenario could happen on the axogenesis of mDAn during development. By looking at the single cell RNA-seq data on human ventral midbrain at developmental stages (Saunders et al., 2018), we found the expression of *Col4a6* is mainly in progenitor medial floorplate cells and radial glia-like cells, both of which are present at developmental stages. In fact, though *Col4a6* expression difference was substantial between C57BL/6J and A/J at the age of 3 months, the expression itself was not abundant, which is likely because the main expression event happens in embryonic stages but the expression discrepancy carries on to adulthood.

There are certain questions not being addressed in our study. First of all, CC strains bear the genetic variation from eight founder strains. Therefore, the significant QTL hit could include potential causal genes or variants that come from any of the eight founders, not necessarily reflecting the true difference between C57BL/6J and A/J. Additional experiments to further validate *Col4a6* downregulation effect on axongenesis is in need. For example, we can reduce *Col4a6* expression by morpholino on DAn reporter line of zebrafish to see if there is any defect regarding axon formation. Secondly, dopamine concentration in ventral midbrain is a complex trait and dopamine synthesis is a multi-step process. As discussed before, mouse strains could have different TH activity, so the variation in dopamine concentration could come from enzymatic activity. Here our study specifically focused on C57BL/6J and A/J, the two strains who do not differ in DAn cell number but indeed have difference in DAn branching, fixing our attention on factors affecting axongenesis.

In summary, QTL mapping on dopamine level variation in ventral midbrain with 32 CC strains found a top QTL hit on chromosome X. Further analysis with our previous generated RNA-seq on midbrain identified *Col4a6* is the most likely gene candidate. Given the fact that *Col4a6* controls neuron axon formation by regulating the integrity of basement membrane during development, we propose similar role of this gene could exist for mDAn.

6. Summary and Perspectives

Ventral midbrain is an essential brain region controlling motor function and behavior. Understanding how genetic variation can contribute to ventral midbrain gene expression difference can help to better decipher the genetic background of its associated disease phenotypes.

Our study looked at *cis* and *trans*-acting variation by coupling tissue-level transcriptomic and epigenomic data with single cell level open chromatin assays. Though many studies focus on individual cell types in ventral midbrain, exploring global expression changes and their associated genetic variation in this brain region have not yet been systematically demonstrated. We first confirmed substantial gene expression changes could be found in inbred mouse strains by conducting RNA-seq on ventral midbrains of C57BL/6J, A/J and DBA/2J, and identified *Pttg1* as a *trans*-acting regulator directing ventral midbrain transcriptome difference.

Ventral midbrain is a mixture of many cell types. To decipher cell type-specific regulatory variants, we generated single cell chromatin profiles in this brain region from C57BL/6J and A/J. Integration of our snATC-seq and existing scRNA-seq found cell type-identity motifs which are selectively enriched. Putative regulatory variants have differential accessibility across cell types, suggesting cell type-specific gene expression is under control of these variants. By comparing chromatin accessibility in neurons between strains, we propose TCF7L2 is a mediator of *trans*-effect in midbrain neurons.

Next, we explored how genetic variation could link to phenotypic difference by focusing on mDAn. Dopamine concentration in dorsal striatum is indicative to the integrity of mDAn in nigrostriatal circuit. Here we measured the dopamine concentration in inbred strains and found it is a complex trait likely resulted from the genetic difference in ventral midbrains. Then we used CC mouse panel optimized for studying complex traits to identify QTLs associated with the varied level of dopamine in dorsal striatum. A significant QTL on chromosome X led us to identify an interesting gene *Col4a6*, with potential regulatory effect on axogenesis of mDAn during development.

Many challenges are still ahead. We identified *Pttg1* as a *trans*-regulator in midbrain transcriptome, but its biological outcome is not clear. Phenotype screening like behavioral tests on *Pttg1*$^{-/-}$ mouse line is necessary to link the genetic variation to phenotypic difference. Our single cell data integration revealed many putative regulatory variants that might have direct or indirect effect on cell type-specific expression changes. Validation of such results could

expand to future projects. For example, we suggest TCF7L2 as a mediator of *trans*-acting regulation directing strain-specific gene expression phenotypes in neurons. *In vivo* or *in vitro* testing with system congenic to C57BL/6J or A/J, or QTL mapping could provide valuable insight on the underlying mechanisms. Our QTL mapping on varied level of dopamine concentration in dorsal striatum found *Col4a6* is the likely QTL gene. Validation by knockdown experiment in zebra fish model is on-going and could show its potential involvement in axon formation of mDAn.

Taken together, our study explored the genetic regulators of ventral midbrain gene expression and nigrostriatal circuit integrity by coupling tissue level and single cell level data from the view of *cis*- and *trans*-acting regulatory effects. Such datasets provide a rich resource for future study about the regulatory elements in this brain region.

7. References

Abdeltawab, N.F., Aziz, R.K., Kansal, R., Rowe, S.L., Su, Y., Gardner, L., Brannen, C., Nooh, M.M., Attia, R.R., Abdelsamed, H.A., et al. (2008). An Unbiased Systems Genetics Approach to Mapping Genetic Loci Modulating Susceptibility to Severe Streptococcal Sepsis. PLOS Pathog. *4*, e1000042.

Abu Toamih Atamni, H.J., and Iraqi, F.A. (2018). Chapter 9 - Collaborative Cross as the Next-Generation Mouse Genetic Reference Population Designed for Dissecting Complex Traits. In Molecular-Genetic and Statistical Techniques for Behavioral and Neural Research, R.T. Gerlai, ed. (San Diego: Academic Press), pp. 191–224.

Agarwal, D., Sandor, C., Volpato, V., Caffrey, T., Monzon-Sandoval, J., Bowden, R., Alegre-Abarrategui, J., Wade-Martins, R., and Webber, C. (2020). A human single-cell atlas of the Substantia nigra reveals novel cell-specific pathways associated with the genetic risk of Parkinson's disease and neuropsychiatric disorders. BioRxiv 2020.04.29.067587.

Agata, Y., Matsuda, E., and Shimizu, A. (1999). Two Novel Krüppel-associated Box-containing Zinc-finger Proteins, KRAZ1 and KRAZ2, Repress Transcription through Functional Interaction with the Corepressor KAP-1 (TIF1β/KRIP-1). J. Biol. Chem. *274*, 16412–16422.

Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A.J., Mann, A.L., Kundu, K., Hale, C., Dougan, G., and Gaffney, D.J. (2018). Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat. Genet. *50*, 424–431.

Alberts, R., Chen, H., Pommerenke, C., Smit, A.B., Spijker, S., Williams, R.W., Geffers, R., Bruder, D., and Schughart, K. (2011). Expression QTL mapping in regulatory and helper T

cells from the BXD family of strains reveals novel cell-specific genes, gene-gene interactions and candidate genes for auto-immune disease. BMC Genomics *12*, 610.

Anderegg, A., Lin, H.-P., Chen, J.-A., Caronia-Brown, G., Cherepanova, N., Yun, B., Joksimovic, M., Rock, J., Harfe, B.D., Johnson, R., et al. (2013). An Lmx1b-miR135a2 Regulatory Circuit Modulates Wnt1/Wnt Signaling and Determines the Size of the Midbrain Dopaminergic Progenitor Pool. PLoS Genet. *9*.

Andersson, R., and Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. Nat. Rev. Genet. *21*, 71–87.

Andersson, E., Tryggvason, U., Deng, Q., Friling, S., Alekseenko, Z., Robert, B., Perlmann, T., and Ericson, J. (2006). Identification of Intrinsic Determinants of Midbrain Dopamine Neurons. Cell *124*, 393–405.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Andley, U.P., Tycksen, E., McGlasson-Naumann, B.N., and Hamilton, P.D. (2018). Probing the changes in gene expression due to α-crystallin mutations in mouse models of hereditary human cataract. PLOS ONE *13*, e0190817.

Arenas, E., Denham, M., and Villaescusa, J.C. (2015). How to make a midbrain dopaminergic neuron. Development *142*, 1918–1936.

Armando, I., Villar, V.A.M., and Jose, P.A. (2011). Dopamine and Renal Function and Blood Pressure Regulation. Compr. Physiol. *1*, 1075–1117.

Ashbrook, D.G., Arends, D., Prins, P., Mulligan, M.K., Roy, S., Williams, E.G., Lutz, C.M., Valenzuela, A., Bohl, C.J., Ingels, J.F., et al. (2019). The expanded BXD family of mice: A cohort for experimental systems genetics and precision medicine. BioRxiv 672097.

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202–W208.

Bannister, A.J., and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Res. *21*, 381–395.

Barrera, L.A., Vedenko, A., Kurland, J.V., Rogers, J.M., Gisselbrecht, S.S., Rossin, E.J., Woodard, J., Mariani, L., Kock, K.H., Inukai, S., et al. (2016). Survey of variation in human transcription factors reveals prevalent DNA binding changes. Science *351*, 1450–1454.

Barski, A., and Zhao, K. (2009). Genomic location analysis by ChIP-Seq. J. Cell. Biochem. *107*.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. Cell *129*, 823–837.

Barton, N.H., Etheridge, A.M., and Véber, A. (2017). The infinitesimal model: Definition, derivation, and implications. Theor. Popul. Biol. *118*, 50–73.

Bateman, J.R., Johnson, J.E., and Locke, M.N. (2012). Comparing Enhancer Action in Cis and in Trans. Genetics *191*, 1143–1155.

Beck, J.A., Lloyd, S., Hafezparast, M., Lennon-Pierce, M., Eppig, J.T., Festing, M.F.W., and Fisher, E.M.C. (2000). Genealogies of mouse inbred strains. Nat. Genet. *24*, 23–25.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved Elements in the Human Genome. Science *304*, 1321–1325.

Bem, J., Brożko, N., Chakraborty, C., Lipiec, M.A., Koziński, K., Nagalski, A., Szewczyk, Ł.M., and Wiśniewska, M.B. (2019). Wnt/β-catenin signaling in brain development and mental disorders: keeping TCF7L2 in mind. FEBS Lett. *593*, 1654–1674.

Berezovsky, A.D., Poisson, L.M., Cherba, D., Webb, C.P., Transou, A.D., Lemke, N.W., Hong, X., Hasselbach, L.A., Irtenkauf, S.M., Mikkelsen, T., et al. (2014). Sox2 Promotes Malignancy in Glioblastoma by Regulating Plasticity and Astrocytic Differentiation. Neoplasia N. Y. N *16*, 193-206.e25.

Bergman, O., Håkansson, A., Westberg, L., Nordenström, K., Carmine Belin, A., Sydow, O., Olson, L., Holmberg, B., Eriksson, E., and Nissbrandt, H. (2010). PITX3 polymorphism is associated with early onset Parkinson's disease. Neurobiol. Aging *31*, 114–117.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell *125*, 315–326.

Bian, C., Xu, C., Ruan, J., Lee, K.K., Burke, T.L., Tempel, W., Barsyte, D., Li, J., Wu, M., Zhou, B.O., et al. (2011). Sgf29 binds histone H3K4me2/3 and is required for SAGA complex recruitment and histone H3 acetylation. EMBO J. *30*, 2829–2842.

Boelaert, K., McCabe, C.J., Tannahill, L.A., Gittoes, N.J.L., Holder, R.L., Watkinson, J.C., Bradwell, A.R., Sheppard, M.C., and Franklyn, J.A. (2003). Pituitary Tumor Transforming Gene and Fibroblast Growth Factor-2 Expression: Potential Prognostic Indicators in Differentiated Thyroid Cancer. J. Clin. Endocrinol. Metab. *88*, 2341–2347.

Bottomly, D., Walter, N.A.R., Hunter, J.E., Darakjian, P., Kawane, S., Buck, K.J., Searles, R.P., Mooney, M., McWeeney, S.K., and Hitzemann, R. (2011). Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. PloS One *6*, e17820.

Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An expanded view of complex traits: from polygenic to omnigenic. Cell *169*, 1177–1186.

Brower-Toland, B., Riddle, N.C., Jiang, H., Huisinga, K.L., and Elgin, S.C.R. (2009). Multiple SET Methyltransferases Are Required to Maintain Normal Heterochromatin Domains in the Genome of Drosophila melanogaster. Genetics *181*, 1303–1319.

Brown, A., Schuetz, D., Han, Y., Daria, D., Nattamai, K.J., Eiwen, K., Sakk, V., Pospiech, J., Saller, T., van Zant, G., et al. (2020). The lifespan quantitative trait locus gene Securin controls hematopoietic progenitor cell function. Haematologica *105*, 317–324.

Buenrostro, J., Wu, B., Chang, H., and Greenleaf, W. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al *109*, 21.29.1-21.29.9.

Bujold, D., Morais, D.A. de L., Gauthier, C., Côté, C., Caron, M., Kwan, T., Chen, K.C., Laperle, J., Markovits, A.N., Pastinen, T., et al. (2016). The International Human Epigenome Consortium Data Portal. Cell Syst. *3*, 496-499.e2.

Bult, C.J., Blake, J.A., Smith, C.L., Kadin, J.A., Richardson, J.E., Anagnostopoulos, A., Asabor, R., Baldarelli, R.M., Beal, J.S., Bello, S.M., et al. (2019). Mouse Genome Database (MGD) 2019. Nucleic Acids Res. *47*, D801–D806.

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how and why? Mol. Cell *49*.

Cao, J., and Yan, Q. (2012). Histone Ubiquitination and Deubiquitination in Transcription, DNA Damage Response, and Cancer. Front. Oncol. *2*.

Capecchi, M.R. (1989). Altering the genome by homologous recombination. Science *244*, 1288–1292.

Casellas, J. (2011). Inbred mouse strains and genetic stability: a review. Animal *5*, 1–7.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics *14*, 128.

Chen, H., Tian, Y., Shu, W., Bo, X., and Wang, S. (2012). Comprehensive Identification and Annotation of Cell Type-Specific and Ubiquitous CTCF-Binding Sites in the Human Genome. PLOS ONE *7*, e41374.

Chinta, S.J., and Andersen, J.K. (2005). Dopaminergic neurons. Int. J. Biochem. Cell Biol. *37*, 942–946.

Ciaranello, R.D., Barchas, R., Kessler, S., and Barchas, J.D. (1972). Catecholamines: strain differences in biosynthetic enzyme activity in mice. Life Sci. Pt 1 Physiol. Pharmacol. *11*, 565–572.

Cirillo, L.A., Lin, F.R., Cuesta, I., Friedman, D., Jarnik, M., and Zaret, K.S. (2002). Opening of Compacted Chromatin by Early Developmental Transcription Factors HNF3 (FoxA) and GATA-4. Mol. Cell *9*, 279–289.

Clapier, C.R., Iwasa, J., Cairns, B.R., and Peterson, C.L. (2017). Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. Nat. Rev. Mol. Cell Biol. *18*, 407–422.

Cobellis, G., Nicolaus, G., Iovino, M., Romito, A., Marra, E., Barbarisi, M., Sardiello, M., Di Giorgio, F.P., Iovino, N., Zollo, M., et al. (2005). Tagging genes with cassette-exchange sites. Nucleic Acids Res. *33*, e44.

Cova, L., and Armentero, M.-T. (2011). 1980-2011: Parkinson's Disease and Advance in Stem Cell Research. New Ther. Park. Dis.

Cuellar-Barboza, A.B., Winham, S.J., McElroy, S.L., Geske, J.R., Jenkins, G.D., Colby, C.L., Prieto, M.L., Ryu, E., Cunningham, J.M., Frye, M.A., et al. (2016). Accumulating evidence for a role of TCF7L2 variants in bipolar disorder with elevated body mass index. Bipolar Disord. *18*, 124–135.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46*, D794–D801.

Deplancke, B., Alpern, D., and Gardeux, V. (2016). The Genetics of Transcription Factor DNA Binding Variation. Cell *166*, 538–554.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Druka, A., Druka, I., Centeno, A.G., Li, H., Sun, Z., Thomas, W.T., Bonar, N., Steffenson, B.J., Ullrich, S.E., Kleinhofs, A., et al. (2008). Towards systems genetic analyses in barley: Integration of phenotypic, expression and genotype data into GeneNetwork. BMC Genet. *9*, 73.

Ecco, G., Imbeault, M., and Trono, D. (2017). KRAB zinc finger proteins. Development *144*, 2719–2729.

Eiden, L.E., Schäfer, M.K.-H., Weihe, E., and Schütz, B. (2004). The vesicular amine transporter family (SLC18): amine/proton antiporters required for vesicular accumulation and regulated exocytotic secretion of monoamines and acetylcholine. Pflugers Arch. *447*, 636–640.

Erler, J., Zhang, R., Petridis, L., Cheng, X., Smith, J.C., and Langowski, J. (2014). The Role of Histone Tails in the Nucleosome: A Computational Study. Biophys. J. *107*, 2911–2922.

Facchinello, N., Tarifeño-Saldivia, E., Grisan, E., Schiavone, M., Peron, M., Mongera, A., Ek, O., Schmitner, N., Meyer, D., Peers, B., et al. (2017). Tcf7l2 plays pleiotropic roles in the control of glucose homeostasis, pancreas morphology, vascularization and regeneration. Sci. Rep. *7*.

Ferri, A.L.M., Cavallaro, M., Braida, D., Cristofano, A.D., Canta, A., Vezzani, A., Ottolenghi, S., Pandolfi, P.P., Sala, M., DeBiasi, S., et al. (2004). Sox2 deficiency causes neurodegeneration and impaired neurogenesis in the adult mouse brain. Development *131*, 3805–3819.

Feuk, L., Carson, A.R., and Scherer, S.W. (2006). Structural variation in the human genome. Nat. Rev. Genet. *7*, 85–97.

Fitz, J., Neumann, T., Steininger, M., Wiedemann, E.-M., Garcia, A.C., Athanasiadis, A., Schoeberl, U.E., and Pavri, R. (2020). Spt5-mediated enhancer transcription directly couples enhancer activation with physical promoter interaction. Nat. Genet. *52*, 505–515.

Frazer, K.A., Murray, S.S., Schork, N.J., and Topol, E.J. (2009). Human genetic variation and its contribution to complex traits. Nat. Rev. Genet. *10*, 241–251.

Frietze, S., and Farnham, P.J. (2011). Transcription Factor Effector Domains. Subcell. Biochem. *52*, 261–277.

Gaglani, S.M., Lu, L., Williams, R.W., and Rosen, G.D. (2009). The genetic control of neocortex volume and covariation with neocortical gene expression in mice. BMC Neurosci. *10*, 44.

Galas, D.J., and Schmitz, A. (1978). DNAase footprinting a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res. *5*, 3157–3170.

Geisert, E.E., Lu, L., Freeman-Anderson, N.E., Templeton, J.P., Nassr, M., Wang, X., Gu, W., Jiao, Y., and Williams, R.W. (2009). Gene expression in the mouse eye: an online resource for genetics using 103 strains of mice. Mol. Vis. *15*, 1730–1763.

German, D.C., and Manaye, K.F. (1993). Midbrain dopaminergic neurons (nuclei A8, A9, and A10): Three-dimensional reconstruction in the rat. J. Comp. Neurol. *331*, 297–309.

Gillombardo, C.B., Darrah, R., Dick, T.E., Moore, M., Kong, N., Decker, M.J., Han, F., Yamauchi, M., Dutschmann, M., Azzam, S., et al. (2017). C57BL/6J Mouse Apolipoprotein A2 Gene is Deterministic for Apnea. Respir. Physiol. Neurobiol. *235*, 88–94.

Ginhoux, F., Greter, M., Leboeuf, M., Nandi, S., See, P., Gokhan, S., Mehler, M.F., Conway, S.J., Ng, L.G., Stanley, E.R., et al. (2010). Fate Mapping Analysis Reveals That Adult Microglia Derive from Primitive Macrophages. Science *330*, 841–845.

Ginno, P.A., Burger, L., Seebacher, J., Iesmantavicius, V., and Schübeler, D. (2018). Cell cycle-resolved chromatin proteomics reveals the extent of mitotic preservation of the genomic regulatory landscape. Nat. Commun. *9*, 4048.

Goodnight, A.V., Kremsky, I., Khampang, S., Jung, Y.H., Billingsley, J.M., Bosinger, S.E., Corces, V.G., and Chan, A.W.S. (2019). Chromatin accessibility and transcription dynamics during in vitro astrocyte differentiation of Huntington's Disease Monkey pluripotent stem cells. Epigenetics Chromatin *12*, 67.

Goodrick, C.L. (1975). Life-Span and the Inheritance of Longevity of Inbred Mice. J. Gerontol. *30*, 257–263.

Greer, E.L., and Shi, Y. (2012). Histone methylation: a dynamic mark in health, disease and inheritance. Nat. Rev. Genet. *13*, 343–357.

Hamid, T., and Kakar, S.S. (2004). PTTG/securin activates expression of p53 and modulates its function. Mol. Cancer *3*, 18.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., et al. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. Cell *172*, 1091-1107.e17.

Hatcher, R.J., Dong, J., Liu, S., Bian, G., Contreras, A., Wang, T., Hilsenbeck, S.G., Li, Y., and Zhang, P. (2014). Pttg1/securin is required for the branching morphogenesis of the mammary gland and suppresses mammary tumorigenesis. Proc. Natl. Acad. Sci. *111*, 1008–1013.

He, T.-C., Sparks, A.B., Rago, C., Hermeking, H., Zawel, L., Costa, L.T. da, Morin, P.J., Vogelstein, B., and Kinzler, K.W. (1998). Identification of c-MYC as a Target of the APC Pathway. Science *281*, 1509–1512.

Heaney, A.P., Singson, R., McCabe, C.J., Nelson, V., Nakashima, M., and Melmed, S. (2000). Expression of pituitary-tumour transforming gene in colorectal tumours. The Lancet *355*, 716–719.

Hebsgaard, J.B., Nelander, J., Sabelström, H., Jönsson, M.E., Stott, S., and Parmar, M. (2009). Dopamine neuron precursors within the developing human mesencephalon show radial glial characteristics. Glia *57*, 1648–1659.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Heinz, S., Romanoski, C.E., Benner, C., Allison, K.A., Kaikkonen, M.U., Orozco, L.D., and Glass, C.K. (2013). Effect of natural genetic variation on enhancer selection and function. Nature *503*, 487–492.

Henke, R.M., Meredith, D.M., Borromeo, M.D., Savage, T.K., and Johnson, J.E. (2009). Ascl1 and Neurog2 form novel complexes and regulate Delta-like3 (Dll3) expression in the neural tube. Dev. Biol. *328*, 529–540.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. *106*, 9362–9367.

Ho, S.S., Urban, A.E., and Mills, R.E. (2020). Structural variation in the sequencing era. Nat. Rev. Genet. *21*, 171–189.

Højfeldt, J.W., Agger, K., and Helin, K. (2013). Histone lysine demethylases as targets for anticancer therapy. Nat. Rev. Drug Discov. *12*, 917–930.

Hoseth, E.Z., Krull, F., Dieset, I., Mørch, R.H., Hope, S., Gardsjord, E.S., Steen, N.E., Melle, I., Brattbakk, H.-R., Steen, V.M., et al. (2018). Exploring the Wnt signaling pathway in schizophrenia and bipolar disorder. Transl. Psychiatry *8*, 1–10.

Hsieh, C.-L., Fei, T., Chen, Y., Li, T., Gao, Y., Wang, X., Sun, T., Sweeney, C.J., Lee, G.-S.M., Chen, S., et al. (2014). Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. Proc. Natl. Acad. Sci. U. S. A. *111*, 7319–7324.

Hu, G., Cui, K., Northrup, D., Liu, C., Wang, C., Tang, Q., Ge, K., Levens, D., Crane-Robinson, C., and Zhao, K. (2013). H2A.Z Facilitates Access of Active and Repressive Complexes to Chromatin in Embryonic Stem Cell Self-Renewal and Differentiation. Cell Stem Cell *12*, 180–192.

Huan, T., Joehanes, R., Song, C., Peng, F., Guo, Y., Mendelson, M., Yao, C., Liu, C., Ma, J., Richard, M., et al. (2019). Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. Nat. Commun. *10*, 4267.

Hyun, K., Jeon, J., Park, K., and Kim, J. (2017). Writing, erasing and reading histone lysine methylations. Exp. Mol. Med. *49*, e324–e324.

Ingber, D.E. (2003). Mechanical control of tissue morphogenesis during embryological development. Int. J. Dev. Biol. *50*, 255–266.

Javahery, R., Khachi, A., Lo, K., Zenzie-Gregory, B., and Smale, S.T. (1994). DNA sequence requirements for transcriptional initiator activity in mammalian cells. Mol. Cell. Biol. *14*, 116–127.

Jenuwein, T., and Allis, C.D. (2001). Translating the Histone Code. Science *293*, 1074–1080.

Jin, T., and Liu, L. (2008). Minireview: The Wnt Signaling Pathway Effector TCF7L2 and Type 2 Diabetes Mellitus. Mol. Endocrinol. *22*, 2383–2392.

Jin, C., Zang, C., Wei, G., Cui, K., Peng, W., Zhao, K., and Felsenfeld, G. (2009). H3.3/H2A.Z double variant–containing nucleosomes mark "nucleosome-free regions" of active promoters and other regulatory regions. Nat. Genet. *41*, 941–945.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-Binding Specificities of Human Transcription Factors. Cell *152*, 327–339.

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *527*, 384–388.

Kaikkonen, M.U., and Adelman, K. (2018). Emerging Roles of Non-Coding RNA Transcription. Trends Biochem. Sci. *43*, 654–667.

Kallapur, S., Ormsby, I., and Doetschman, T. (1999). Strain dependency of TGFβ1 function during embryogenesis. Mol. Reprod. Dev. *52*, 341–349.

Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., et al. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature *477*, 289–294.

Keele, G.R., Crouse, W.L., Kelada, S.N.P., and Valdar, W. (2019). Determinants of QTL Mapping Power in the Realized Collaborative Cross. G3 Genes Genomes Genet. *9*, 1707–1727.

Keeley, P.W., Zhou, C., Lu, L., Williams, R.W., Melmed, S., and Reese, B.E. (2014). Pituitary tumor-transforming gene 1 regulates the patterning of retinal mosaics. Proc. Natl. Acad. Sci. U. S. A. *111*, 9295–9300.

Kim, K., Doi, A., Wen, B., Ng, K., Zhao, R., Cahan, P., Kim, J., Aryee, M., Ji, H., Ehrlich, L., et al. (2010). Epigenetic memory in induced pluripotent stem cells. Nature *467*, 285–290.

Klemm, S.L., Shipony, Z., and Greenleaf, W.J. (2019). Chromatin accessibility and the regulatory epigenome. Nat. Rev. Genet. *20*, 207–220.

Köhler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C.M., Brown, D.L., Brudno, M., Campbell, J., et al. (2014). The Human Phenotype Ontology

project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. *42*, D966–D974.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Laarakker, M.C., van Lith, H.A., and Ohl, F. (2011). Behavioral characterization of A/J and C57BL/6J mice using a multidimensional test: Association between blood plasma and brain magnesium-ion concentration with anxiety. Physiol. Behav. *102*, 205–219.

Lai, F., Gardini, A., Zhang, A., and Shiekhattar, R. (2015). Integrator mediates the biogenesis of enhancer RNAs. Nature *525*, 399–403.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell *172*, 650–665.

Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D.A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., et al. (2014). Chromatin state dynamics during blood formation. Science *345*, 943–949.

Le, W., Xu, P., Jankovic, J., Jiang, H., Appel, S.H., Smith, R.G., and Vassilatis, D.K. (2003). Mutations in NR4A2 associated with familial Parkinson disease. Nat. Genet. *33*, 85–89.

Lee, M., Yoon, J., Song, H., Lee, B., Lam, D.T., Yoon, J., Baek, K., Clevers, H., and Jeong, Y. (2017). Tcf7l2 plays crucial roles in forebrain development through regulation of thalamic and habenular neuron identity and connectivity. Dev. Biol. *424*, 62–76.

LEE, S., LEE, C.E., ELIAS, C.F., and ELMQUIST, J.K. (2009). Expression of the Diabetes-Associated Gene TCF7L2 in Adult Mouse Brain. J. Comp. Neurol. *517*, 925–939.

Lenhard, B., Sandelin, A., and Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat. Rev. Genet. *13*, 233–245.

Levy, R., Mott, R.F., Iraqi, F.A., and Gabet, Y. (2015). Collaborative cross mice in a genetic association study reveal new candidate genes for bone microarchitecture. BMC Genomics *16*.

Li, H., Yin, C., Zhang, B., Sun, Y., Shi, L., Liu, N., Liang, S., Lu, S., Liu, Y., Zhang, J., et al. (2013). PTTG1 promotes migration and invasion of human non-small cell lung cancer cells and is modulated by miR-186. Carcinogenesis *34*, 2145–2155.

Li, H., Jakobson, M., Ola, R., Gui, Y., Kumar, A., Sipilä, P., Sariola, H., Kuure, S., and Andressoo, J.-O. (2019a). Development of the urogenital system is regulated via the 3′UTR of GDNF. Sci. Rep. *9*, 5302.

Li, Z., Schulz, M.H., Look, T., Begemann, M., Zenke, M., and Costa, I.G. (2019b). Identification of transcription factor binding sites using ATAC-seq. Genome Biol. *20*, 45.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Lin, H., Liu, Q., Li, X., Yang, J., Liu, S., Huang, Y., Scanlon, M.J., Nettleton, D., and Schnable, P.S. (2017). Substantial contribution of genetic variation in the expression of transcription factors to phenotypic variation revealed by eRD-GWAS. Genome Biol. *18*, 192.

Link, V.M., Duttke, S.H., Chun, H.B., Holtman, I.R., Westin, E., Hoeksema, M.A., Abe, Y., Skola, D., Romanoski, C.E., Tao, J., et al. (2018). Analysis of Genetically Diverse Macrophages Reveals Local and Domain-wide Mechanisms that Control Transcription Factor Binding and Function. Cell *173*, 1796-1809.e17.

von Linstow, C.U., DeLano-Taylor, M., Kordower, J.H., and Brundin, P. (2020). Does Developmental Variability in the Number of Midbrain Dopamine Neurons Affect Individual Risk for Sporadic Parkinson's Disease? J. Park. Dis. *10*, 405–411.

Liu, X., Li, Y.I., and Pritchard, J.K. (2019). Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell *177*, 1022-1034.e6.

Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., de Candia, T.R., Lee, S.H., Wray, N.R., Kendler, K.S., et al. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat. Genet. *47*, 1385–1392.

Long, H.K., Prescott, S.L., and Wysocka, J. (2016). Ever-changing landscapes: transcriptional enhancers in development and evolution. Cell *167*, 1170–1187.

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

Lu, L., Wei, L., Peirce, J.L., Wang, X., Zhou, J., Homayouni, R., Williams, R.W., and Airey, D.C. (2008). Using gene expression databases for classical trait QTL candidate gene discovery in the BXD recombinant inbred genetic reference population: Mouse forebrain weight. BMC Genomics *9*, 444.

Lubka-Pathak, M., Shah, A.A., Gallozzi, M., Müller, M., Zimmermann, U., Löwenheim, H., Pfister, M., Knipper, M., Blin, N., and Schimmang, T. (2011). Altered expression of securin (Pttg1) and serpina3n in the auditory system of hearing-impaired Tff3-deficient mice. Cell. Mol. Life Sci. *68*, 2739–2749.

Lum, P.Y., Chen, Y., Zhu, J., Lamb, J., Melmed, S., Wang, S., Drake, T.A., Lusis, A.J., and Schadt, E.E. (2006). Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. J. Neurochem. *97*, 50–62.

Malkinson, A.M. (1989). The genetic basis of susceptibility to lung tumors in mice. Toxicology *54*, 241–271.

Manno, G.L., Gyllborg, D., Codeluppi, S., Nishimura, K., Salto, C., Zeisel, A., Borm, L.E., Stott, S.R.W., Toledo, E.M., Villaescusa, J.C., et al. (2016). Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. Cell *167*, 566-580.e19.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Manyes, L., Holst, S., Lozano, M., Santos, E., and Fernandez-Medarde, A. (2018). Spatial learning and long-term memory impairments in RasGrf1 KO, Pttg1 KO, and double KO mice. Brain Behav. *8*, e01089.

March, E., and Farrona, S. (2017). Polycomb silencing mediated by specific DNA-binding recruiters. Nat. Genet. *49*, 1416–1417.

Marstrand, T.T., and Storey, J.D. (2014). Identifying and mapping cell-type-specific chromatin programming of gene expression. Proc. Natl. Acad. Sci. *111*, E645–E654.

Martini, S., Bernoth, K., Main, H., Ortega, G.D.C., Lendahl, U., Just, U., and Schwanbeck, R. (2013). A Critical Role for Sox9 in Notch-Induced Astrogliogenesis and Stem Cell Maintenance. STEM CELLS *31*, 741–751.

Mass, E., Jacome-Galarza, C.E., Blank, T., Lazarov, T., Durham, B.H., Ozkaya, N., Pastore, A., Schwabenland, M., Chung, Y.R., Rosenblum, M.K., et al. (2017). A somatic mutation in erythro-myeloid progenitors causes neurodegenerative disease. Nature *549*, 389–393.

Mathes, W.F., Kelly, S.A., and Pomp, D. (2011). Advances in comparative genetics: influence of genetics on obesity. Br. J. Nutr. *106*, S1–S10.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science *337*, 1190–1195.

Mayran, A., Sochodolsky, K., Khetchoumian, K., Harris, J., Gauthier, Y., Bemmo, A., Balsalobre, A., and Drouin, J. (2019). Pioneer and nonpioneer factor cooperation drives lineage specific chromatin opening. Nat. Commun. *10*, 3807.

McFadyen, M.P., Kusek, G., Bolivar, V.J., and Flaherty, L. (2003). Differences among eight inbred strains of mice in motor ability and motor learning on a rotorod. Genes Brain Behav. *2*, 214–219.

Mei, J., Huang, X., and Zhang, P. (2001). Securin is not required for cellular viability, but is required for normal growth of mouse embryonic fibroblasts. Curr. Biol. *11*, 1197–1201.

Mickelsen, L.E., Bolisetty, M., Chimileski, B.R., Fujita, A., Beltrami, E.J., Costanzo, J.T., Naparstek, J.R., Robson, P., and Jackson, A.C. (2019). Single-cell transcriptomic analysis of the lateral hypothalamic area reveals molecularly distinct populations of inhibitory and excitatory neurons. Nat. Neurosci. *22*, 642–656.

Mierlo, G. van, Veenstra, G.J.C., Vermeulen, M., and Marks, H. (2019). The Complexity of PRC2 Subcomplexes. Trends Cell Biol. *29*, 660–671.

Mohaghegh, N., Bray, D., Keenan, J., Penvose, A., Andrilenas, K.K., Ramlall, V., and Siggers, T. (2019). NextPBM: a platform to study cell-specific transcription factor binding and cooperativity. Nucleic Acids Res. *47*, e31.

Moy, S.S., Nadler, J.J., Young, N.B., Perez, A., Holloway, L.P., Barbaro, R.P., Barbaro, J.R., Wilson, L.M., Threadgill, D.W., Lauder, J.M., et al. (2007). Mouse behavioral tasks relevant to autism: Phenotypes of 10 inbred strains. Behav. Brain Res. *176*, 4–20.

Mozhui, K., Karlsson, R.-M., Kash, T.L., Ihne, J., Norcross, M., Patel, S., Farrell, M.R., Hill, E.E., Graybeal, C., Martin, K.P., et al. (2010). Strain Differences in Stress Responsivity Are Associated with Divergent Amygdala Gene Expression and Glutamate-Mediated Neuronal Excitability. J. Neurosci. *30*, 5357–5367.

Mulligan, M.K., Abreo, T., Neuner, S.M., Parks, C., Watkins, C.E., Houseal, M.T., Shapaker, T.M., Hook, M., Tan, H., Wang, X., et al. (2019). Identification of a Functional Non-coding Variant in the GABAA Receptor α2 Subunit of the C57BL/6J Mouse Reference Genome: Major Implications for Neuroscience Research. Front. Genet. *10*.

Musacchio, J.M. (1975). Enzymes Involved in the Biosynthesis and Degradation of Catecholamines. In Biochemistry of Biogenic Amines, L.L. Iversen, S.D. Iversen, and S.H. Snyder, eds. (Boston, MA: Springer US), pp. 1–35.

Muthane, U., Ramsay, K.A., Jiang, H., Jackson-Lewis, V., Donaldson, D., Fernando, S., Ferreira, M., and Przedborski, S. (1994). Differences in Nigral Neuron Number and Sensitivity to 1-Methyl-4-phenyl-1,2,3,6-tetrahydropyridine in C57/bl and CD-1 Mice. Exp. Neurol. *126*, 195–204.

Nalls, M.A., Blauwendraat, C., Vallerga, C.L., Heilbron, K., Bandres-Ciga, S., Chang, D., Tan, M., Kia, D.A., Noyce, A.J., Xue, A., et al. (2019). Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol. *18*, 1091–1102.

Nica, A.C., and Dermitzakis, E.T. (2013). Expression quantitative trait loci: present and future. Philos. Trans. R. Soc. B Biol. Sci. *368*, 20120362.

Oestreich, K.J., and Weinmann, A.S. (2012). Master regulators or lineage-specifying? Changing views on CD4+ T cell transcription factors. Nat. Rev. Immunol. *12*, 799–804.

Ohnmacht, J., May, P., Sinkkonen, L., and Krüger, R. (2020). Missing heritability in Parkinson's disease: the emerging role of non-coding genetic variation. J. Neural Transm.

Ong, C.-T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat. Rev. Genet. *12*, 283–293.

Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B.J., Afzal, S.Y., Lee, E.A., Zhu, Y., Plajzer-Frick, I., Pickle, C.S., et al. (2018). Enhancer Redundancy Allows for Phenotypic Robustness in Mammalian Development. Nature *554*, 239–243.

Pagano, G., Ferrara, N., Brooks, D.J., and Pavese, N. (2016). Age at onset and Parkinson disease phenotype. Neurology *86*, 1400–1407.

Paiva, I., Jain, G., Lázaro, D.F., Jerčić, K.G., Hentrich, T., Kerimoglu, C., Pinho, R., Szegő, È.M., Burkhardt, S., Capece, V., et al. (2018). Alpha-synuclein deregulates the expression of COL4A2 and impairs ER-Golgi function. Neurobiol. Dis. *119*, 121–135.

Panman, L., Papathanou, M., Laguna, A., Oosterveen, T., Volakakis, N., Acampora, D., Kurtsdotter, I., Yoshitake, T., Kehr, J., Joodmardi, E., et al. (2014). Sox6 and Otx2 Control the Specification of Substantia Nigra and Ventral Tegmental Area Dopamine Neurons. Cell Rep. *8*, 1018–1025.

Parashos, S.A., Luo, S., Biglan, K.M., -Wollner, I.B., He, B., Liang, G.S., Ross, G.W., Tilley, B.C., and Shulman, L.M. (2014). Measuring Disease Progression in Early Parkinson Disease: the National Institutes of Health Exploratory Trials in Parkinson Disease (NET-PD) Experience. JAMA Neurol. *71*, 710–716.

Pei, L., and Melmed*, S. (1997). Isolation and Characterization of a Pituitary Tumor-Transforming Gene (PTTG). Mol. Endocrinol. *11*, 433–441.

Pelikan, R.C., Kelly, J.A., Fu, Y., Lareau, C.A., Tessneer, K.L., Wiley, G.B., Wiley, M.M., Glenn, S.B., Harley, J.B., Guthridge, J.M., et al. (2018). Enhancer histone-QTLs are enriched on autoimmune risk haplotypes and influence gene expression within chromatin networks. Nat. Commun. *9*, 2905.

Perlman, R.L. (2016). Mouse models of human disease. Evol. Med. Public Health *2016*, 170–176.

Pierce B.A. (2017). Genetics: A conceptual Approach (6th eddition) (United States of America: W. H. Freeman).

Pnueli, L., Rudnizky, S., Yosefzon, Y., and Melamed, P. (2015). RNA transcribed from a distal enhancer is required for activating the chromatin at the promoter of the gonadotropin α-subunit gene. Proc. Natl. Acad. Sci. *112*, 4369–4374.

Poewe, W., Seppi, K., Tanner, C.M., Halliday, G.M., Brundin, P., Volkmann, J., Schrag, A.-E., and Lang, A.E. (2017). Parkinson disease. Nat. Rev. Dis. Primer *3*, 1–21.

Pott, S., and Lieb, J.D. (2015). What are super-enhancers? Nat. Genet. *47*, 8–12.

Puri, R., Tousson, A., Chen, L., and Kakar, S.S. (2001). Molecular cloning of pituitary tumor transforming gene 1 from ovarian tumors and its expression in tumors. Cancer Lett. *163*, 131–139.

Puschendorf, M., Terranova, R., Boutsma, E., Mao, X., Isono, K., Brykczynska, U., Kolb, C., Otte, A.P., Koseki, H., Orkin, S.H., et al. (2008). PRC1 and Suv39h specify parental asymmetry at constitutive heterochromatin in early mouse embryos. Nat. Genet. *40*, 411–420.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Radakovits, R., Barros, C.S., Belvindrah, R., Patton, B., and Müller, U. (2009). Regulation of radial glial survival by signals from the meninges. J. Neurosci. Off. J. Soc. Neurosci. *29*, 7694–7705.

Ridgway, W.M., Healy, B., Smink, L.J., Rainbow, D., and Wicker, L.S. (2007). New tools for defining the “genetic background” of inbred mouse strains. Nat. Immunol. *8*, 669–673.

Roberts, C.W.M., and Orkin, S.H. (2004). The SWI/SNF complex — chromatin and cancer. Nat. Rev. Cancer *4*, 133–142.

Romanelli, R.J., Williams, J.T., and Neve, K.A. (2010). Dopamine Receptor Signaling: Intracellular Pathways to Behavior. In The Dopamine Receptors, K.A. Neve, ed. (Totowa, NJ: Humana Press), pp. 137–173.

Rosen, G.D., and Williams, R.W. (2001). Complex trait analysis of the mouse striatum: independent QTLs modulate volume and neuron number. BMC Neurosci. *2*, 5.

Ross, R.A., Judd, A.B., Pickel, V.M., Joh, T.H., and Reis, D.J. (1976). Strain-dependent variations in number of midbrain dopaminergic neurones. Nature *264*, 654–656.

Rossetto, D., Avvakumov, N., and Côté, J. (2012). Histone phosphorylation. Epigenetics *7*, 1098–1108.

Rowley, M.J., and Corces, V.G. (2018). Organizational principles of 3D genome architecture. Nat. Rev. Genet. *19*, 789–800.

Rung, J.P., Carlsson, A., Markinhuhta, K.R., and Carlsson, M.L. (2005). The dopaminergic stabilizers (−)-OSU6162 and ACR16 reverse (+)-MK-801-induced social withdrawal in rats. Prog. Neuropsychopharmacol. Biol. Psychiatry *29*, 833–839.

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., and Hume, D.A. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat. Rev. Genet. *8*, 424–436.

Santos, A., Wernersson, R., and Jensen, L.J. (2015). Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. Nucleic Acids Res. *43*, D1140–D1144.

Saunders, A., Macosko, E.Z., Wysoker, A., Goldman, M., Krienen, F.M., de Rivera, H., Bien, E., Baum, M., Bortolin, L., Wang, S., et al. (2018). Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. Cell *174*, 1015-1030.e16.

Savic, D., Distler, M.G., Sokoloff, G., Shanahan, N.A., Dulawa, S.C., Palmer, A.A., and Nobrega, M.A. (2011). Modulation of Tcf7l2 Expression Alters Behavior in Mice. PLOS ONE *6*, e26897.

Schapira, A.H.V., Chaudhuri, K.R., and Jenner, P. (2017). Non-motor features of Parkinson disease. Nat. Rev. Neurosci. *18*, 435–450.

Schaum, N., Karkanias, J., Neff, N.F., May, A.P., Quake, S.R., Wyss-Coray, T., Darmanis, S., Batson, J., Botvinnik, O., Chen, M.B., et al. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature *562*, 367–372.

Schoenfelder, S., and Fraser, P. (2019). Long-range enhancer–promoter contacts in gene expression control. Nat. Rev. Genet. *20*, 437–455.

Schütz, D., Brown, A., Geiger, H., and Nattamai, K. (2017). PTTG1/securin as a quantitative trait locus candidate gene controlling progenitor cell survival and organismal lifespan. Exp. Hematol. *53*, S68.

Shi, H., Kichaev, G., and Pasaniuc, B. (2016a). Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. Am. J. Hum. Genet. *99*, 139–153.

Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., Qin, X.-C., Li, J., Cao, J.-P., Eden, J.-S., et al. (2016b). Redefining the invertebrate RNA virosphere. Nature *540*, 539–543.

Siersbæk, R., Madsen, J.G.S., Javierre, B.M., Nielsen, R., Bagge, E.K., Cairns, J., Wingett, S.W., Traynor, S., Spivakov, M., Fraser, P., et al. (2017). Dynamic Rewiring of Promoter-Anchored Chromatin Loops during Adipocyte Differentiation. Mol. Cell *66*, 420-435.e5.

Silver LM (1995). Mouse genetics (Oxford University Press).

Sinnamon, J.R., Torkenczy, K.A., Linhoff, M.W., Vitak, S.A., Mulqueen, R.M., Pliner, H.A., Trapnell, C., Steemers, F.J., Mandel, G., and Adey, A.C. (2019). The accessible chromatin landscape of the murine hippocampus at single-cell resolution. Genome Res. *29*, 857–869.

Smidt, M.P., van Schaick, H.S.A., Lanctôt, C., Tremblay, J.J., Cox, J.J., van der Kleij, A.A.M., Wolterink, G., Drouin, J., and Burbach, J.P.H. (1997). A homeodomain gene Ptx3 has highly restricted brain expression in mesencephalic dopaminergic neurons. Proc. Natl. Acad. Sci. U. S. A. *94*, 13305–13310.

Smidt, M.P., Asbreuk, C.H.J., Cox, J.J., Chen, H., Johnson, R.L., and Burbach, J.P.H. (2000). A second independent pathway for development of mesencephalic dopaminergic neurons requires Lmx1b. Nat. Neurosci. *3*, 337–341.

Smidt, M.P., Smits, S.M., and Burbach, J.P.H. (2004). Homeobox gene Pitx3 and its role in the development of dopamine neurons of the substantia nigra. Cell Tissue Res. *318*, 35–43.

Smith, E., and Shilatifard, A. (2010). The Chromatin Signaling Pathway: Diverse Mechanisms of Recruitment of Histone-Modifying Enzymes and Varied Biological Outcomes. Mol. Cell *40*, 689–701.

Smith, A.M., Gibbons, H.M., Oldfield, R.L., Bergin, P.M., Mee, E.W., Faull, R.L.M., and Dragunow, M. (2013). The transcription factor PU.1 is critical for viability and function of human brain microglia. Glia *61*, 929–942.

Smithies, O., and Maeda, N. (1995). Gene targeting approaches to complex genetic diseases: atherosclerosis and essential hypertension. Proc. Natl. Acad. Sci. U. S. A. *92*, 5266–5272.

Song, L., and Crawford, G.E. (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb. Protoc. *2010*, pdb.prot5384.

Soutoglou, E., Viollet, B., Vaxillaire, M., Yaniv, M., Pontoglio, M., and Talianidis, I. (2001). Transcription factor-dependent regulation of CBP and P/CAF histone acetyltransferase activity. EMBO J. *20*, 1984–1992.

Soutourina, J. (2019). Mammalian Mediator as a Functional Link between Enhancers and Promoters. Cell *178*, 1036–1038.

Stormo, G.D., and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. Nat. Rev. Genet. *11*, 751–760.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888-1902.e21.

Szabo, Q., Bantignies, F., and Cavalli, G. (2019). Principles of genome folding into topologically associating domains. Sci. Adv. *5*, eaaw1668.

Takeuchi, M., Yamaguchi, S., Yonemura, S., Kakiguchi, K., Sato, Y., Higashiyama, T., Shimizu, T., and Hibi, M. (2015). Type IV Collagen Controls the Axogenesis of Cerebellar Granule Cells by Regulating Basement Membrane Integrity in Zebrafish. PLOS Genet. *11*, e1005587.

The Complex Trait Consortium, Churchill, G.A., Airey, D.C., Allayee, H., Angel, J.M., Attie, A.D., Beatty, J., Beavis, W.D., Belknap, J.K., Bennett, B., et al. (2004). The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat. Genet. *36*, 1133–1137.

Thifault, S., Lalonde, R., Sanon, N., and Hamet, P. (2002). Comparisons between C57BL/6J and A/J mice in motor activity and coordination, hole-poking, and spatial learning. Brain Res. Bull. *58*, 213–218.

Tong, Y., and Eigler, T. (2009). Transcriptional targets for pituitary tumor-transforming gene-1. J. Mol. Endocrinol. *43*, 179–185.

Tong, Y., Tan, Y., Zhou, C., and Melmed, S. (2007). Pituitary tumor transforming gene interacts with Sp1 to modulate G1/S cell phase transition. Oncogene *26*, 5596–5605.

de la Torre-Ubieta, L., Stein, J.L., Won, H., Opland, C.K., Liang, D., Lu, D., and Geschwind, D.H. (2018). The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. Cell *172*, 289-304.e18.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Res. *25*, 1491–1498.

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854–858.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22.

Vizeacoumar, F.J., Arnold, R., Vizeacoumar, F.S., Chandrashekhar, M., Buzina, A., Young, J.T.F., Kwan, J.H.M., Sayad, A., Mero, P., Lawo, S., et al. (2013). A negative genetic interaction map in isogenic cancer cell lines reveals cancer cell vulnerabilities. Mol. Syst. Biol. *9*, 696.

Voigt, P., Tee, W.-W., and Reinberg, D. (2013). A double take on bivalent promoters. Genes Dev. *27*, 1318–1338.

Wahlsten, D., Bachmanov, A., Finn, D.A., and Crabbe, J.C. (2006). Stability of inbred mouse strain differences in behavior and brain size between laboratories and across decades. Proc. Natl. Acad. Sci. *103*, 16364–16369.

Walrath, J.C., Hawes, J.J., Van Dyke, T., and Reilly, K.M. (2010). Genetically Engineered Mouse Models in Cancer Research. Adv. Cancer Res. *106*, 113–164.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell *153*, 307–319.

Wise, R.A. (2004). Dopamine, learning and motivation. Nat. Rev. Neurosci. *5*, 483–494.

Wysocka, J., Swigut, T., Xiao, H., Milne, T.A., Kwon, S.Y., Landry, J., Kauer, M., Tackett, A.J., Chait, B.T., Badenhorst, P., et al. (2006). A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. Nature *442*, 86–90.

Yalcin, B., Wong, K., Agam, A., Goodson, M., Keane, T.M., Gan, X., Nellåker, C., Goodstadt, L., Nicod, J., Bhomra, A., et al. (2011). Sequence-based characterization of structural variation in the mouse genome. Nature *477*, 326–329.

Yamamoto, A., Guacci, V., and Koshland, D. (1996). Pds1p is required for faithful execution of anaphase in the yeast, Saccharomyces cerevisiae. J. Cell Biol. *133*, 85–97.

Yang, C.-C., Kato, H., Shindo, M., and Masai, H. (2019). Cdc7 activates replication checkpoint by phosphorylating the Chk1-binding domain of Claspin in human cells. ELife *8*, e50796.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.

Yin, J., and Wang, G. (2014). The Mediator complex: a master coordinator of transcription and cell lineage development. Development *141*, 977–987.

Yoneyama, N., Crabbe, J.C., Ford, M.M., Murillo, A., and Finn, D.A. (2008). Voluntary ethanol consumption in 22 inbred mouse strains. Alcohol *42*, 149–160.

Zhang, T., Cooper, S., and Brockdorff, N. (2015). The interplay of histone modifications – writers that read. EMBO Rep. *16*, 1467–1481.

Zhu, F., Nair, R.R., Fisher, E.M.C., and Cunningham, T.J. (2019). Humanising the mouse genome piece by piece. Nat. Commun. *10*, 1845.

Zou, F., Gelfond, J.A.L., Airey, D.C., Lu, L., Manly, K.F., Williams, R.W., and Threadgill, D.W. (2005). Quantitative Trait Locus Analysis Using Recombinant Inbred Intercrosses. Genetics *170*, 1299–1311.

Zou, H., McGarry, T.J., Bernal, T., and Kirschner, M.W. (1999). Identification of a Vertebrate Sister-Chromatid Separation Inhibitor Involved in Transformation and Tumorigenesis. Science *285*, 418–422.