UCLA
UCLA Electronic Theses and Dissertations

Title
Alternative splicing in mammalian brain

Permalink
https://escholarship.org/uc/item/6mw6x01d

Author
Zhang, Yida

Publication Date
2019

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Alternative splicing in mammalian brain

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Bioinformatics

by

Yida Zhang

2019

©Copyright by

Yida Zhang

2019

ABSTRACT OF THE DISSERTATION

Alternative splicing in mammalian brain

by

Yida Zhang

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2019

Professor Yi Xing, Chair

Post-transcriptional regulation is a set of important biological functions taking place during the genetic information flow from DNA to RNA. Alternative splicing is one of the post-transcriptional regulation mechanisms, which creates enormous diversity at the RNA level. Compared with other tissues, alternative splicing is highly abundant in brain. Emerging evidence shows that alternative splicing plays an important role in many neurological disorders. In this dissertation, the regulation of alternative splicing in mammalian brain and its relationship with neurological disorders was investigated. Based on our study, we demonstrated the importance of alternative splicing in neurological disorders, which provides insights of disease prevention and treatment.

The first part of the dissertation studies the genetic regulation of alternative splicing in human brain and its implications in neurological disorders. A major question in genetics is how variants of ubiquitously expressed genes produce tissue- and cell type-specific phenotypes. One

mechanism is through affecting alternative splicing that may contribute to disease. We investigated splicing quantitative trait loci (sQTLs) in 1,209 samples from 13 human brain regions, using RNA-seq and genotype data from the GTEx project. Hundreds of sQTLs were identified in each brain region. Some sQTLs were shared across brain regions, whereas others displayed regional specificity. These 'regionally ubiquitous' and 'regionally specific' sQTLs showed distinct positional distributions of single nucleotide polymorphisms (SNPs) within and outside essential splice sites, respectively, suggesting their regulation by divergent molecular mechanisms. Integrating the binding motifs and expression patterns of RNA binding proteins with exon splicing profiles, we uncovered likely causal variants of brain region specific sQTLs. Overall, our study reveals widespread regional variation of sQTLs in human brain, and demonstrates that such variation can be used to fine map causal variants of sQTLs and their associated neurological diseases.

The second part of the dissertation focuses on sexual dimorphism of mammalian brain. Women and men showed difference in terms of neurological disorders. Previous studies investigated this sex difference using neuroanatomy and behavioral approaches. However, little is known about the underlying molecular mechanism. As an important post-transcriptional regulatory mechanism, we studied sexual dimorphic alternative splicing in mammalian brain through the female specific expression of the long noncoding RNA (lncRNA) Xist. This study is still ongoing.

The dissertation of Yida Zhang is approved.

Douglas L. Black

Thomas Graeber

Qing Zhou

Yi Xing, Committee Chair

University of California, Los Angeles

2019

To my mother and father

TABLE OF CONTENTS

Chapter 1 Introduction 1

References 4

Chapter 2 Regional variation of splicing QTLs in human brain 7

2.1 Introduction 7

2.2 Results 8

2.2.1 Gene expression and alternative splicing in human brain 8

2.2.2 Identification of sQTLs in 13 brain regions 10

2.2.3 Relationship among SNP position, significance, and brain region specificity of sQTLs 11

2.2.4 Examples of regionally ubiquitous and regionally specific sQTLs 13

2.2.5 Using regional specificity of sQTL signals to prioritize causal sQTL cis variants and trans regulators 16

2.3 Discussion 19

2.4 Methods 21

2.4.1 Datasets and processing of GTEx brain tissue samples. 21

2.4.2 Covariate and confounding factor analysis for splicing and gene expression data. 22

2.4.3 Correcting PSI values for confounding factors prior to sQTL analysis. 23

2.4.4 Identification of sQTLs. 24

2.4.5 LD calculation. 25

2.4.6 Selection of the most representative sQTL SNP across 13 brain regions. 26

2.4.7 Defining regionally ubiquitous and regionally specific sQTLs. 26

2.4.8 Quantifying the effect of SNPs on RBP-RNA binding using DeepBind. 26

2.4.9 Motif scan of NOVA1. 28
2.5 References 29
2.6 Figures 35
Chapter 3 Sexual dimorphic alternative splicing in mammalian brain 63
3.1 Introduction 63
3.2 Results 65
3.2.1 Development stage and tissue are the main factors contributing to the splicing variation based on the FCG mouse data 65
3.2.2 Pairwise comparison between different genotypes from the FCG model across tissue and development stage reveals distinct splicing pattern 66
3.2.3 Differential splicing events between genotypes show high specificity to development stage and tissue 67
3.2.4 Differential splicing between genotypes within the same development stage and tissue 67
3.2.5 PTBP2 knockdown leads to splicing changes in the mouse cortex across different development stages 67
3.2.6 Xist knockdown leads to splicing changes in the mouse cerebellum and cortex across different development stages 68
3.3. Discussions 69
3.4 Methods 70
3.4.1 FCG mouse data 70
3.4.2 PTBP2 knockdown data 70
3.4.3 Xist knockdown data 71

3.4.4 RNA-seq data processing 71
3.5 References 71
3.6 Figures 76
3.7 Tables 85
Chapter 4 Concluding Remarks 89

LIST OF FIGURES

Figure 2.1 Overview of the study and data 35

Figure 2.2 Identification of sQTLs in each brain region 37

Figure 2.3 Relationship among positional distribution of sQTLs, significance of sQTLs and brain-region specificity 39

Figure 2.4 Two groups of sQTLs with different types of brain-region specificity 41

Figure 2.5 Brain region-specific sQTLs and RBP expression 43

Figure 2.6 Combining deep learning, regional expression and sQTL information to prioritize causal *cis*-acting sQTL variants and their trans-acting regulators 44

Supplementary Figure 2.7 Distribution of GTEx brain tissue samples across age, sex and brain region 46

Supplementary Figure 2.8 Numbers of splicing events for three types of alternative splicing 47

Supplementary Figure 2.9 t-SNE clustering of brain samples 48

Supplementary Figure 2.10 Dependence of splicing and gene expression events on age, sex and brain region factors 49

Supplementary Figure 2.11 Regulation of brain region-specific splicing of *NRXN2* exon 4 by KHDRBS3 and KHDRBS1 50

Supplementary Figure 2.12 Positional distribution and significance of sQTL SNPs. 52

Supplementary Figure 2.13 Relationship between brain region specificity and positional distribution of sQTLs 53

Supplementary Figure 2.14 Positional distribution of regionally specific and ubiquitous sQTLs 54

Supplementary Figure 2.15 Ubiquitous *cis*-acting sQTLs significant in all brain regions 55

Supplementary Figure 2.16 Gene expression level of RBFOX1 and RBFOX3 across 13 brain regions 56
Supplementary Figure 2.17 RBFOX2 regulation of the brain region-specific association between *MAPT* exon 3 and rs17651213 57
Supplementary Figure 2.18 Sashimi plot of *MAPT* exon 3 and rs17651213 in two brain regions: cerebellum and cerebellar hemisphere 58
Supplementary Figure 2.19 Effect of rs6580200 on RBP binding, as predicted by DeepBind 59
Supplementary Figure 2.20 Significance of the sQTL (*CXXC5* exon 2 and rs6580200) agrees with gene expression of HNRNPK 60
Supplementary Figure 2.21 SNP rs4077093 regulates splicing of *POU6F1* exon 4 by affecting NOVA1 binding 61
Supplementary Figure 2.22 Significance of the sQTL (*POU6F1* exon 4 and rs4077093) agrees with gene expression of NOVA1 62
Figure 3.1 PCA result based on RNA-seq data from FCG mouse model 76
Figure 3.2 Heat map of significant differential splicing exons between different pairs of genotypes 77
Figure 3.3 Overlap of significant differential splicing events for each comparison under different development stage and tissue combination 78
Figure 3.4 Overlap of differential splicing events between all pairs of genotypes within the same development stage and tissue 79
Figure 3.5 PTBP2 knockdown leads to splicing changes in the mouse cortex across different development stages 80

Figure 3.6 Differential splicing events between Xist knockdown and control in males and females in the mouse cerebellum across different development stages 81
Figure 3.7 Differential splicing events between Xist knockdown and control in males and females in the mouse cortex at P60 82
Supplementary Figure 3.8 Xist expression in knockdown and control samples 83
Supplementary Figure 3.9 Overlap of the differential splicing events between Xist knockdown vs. control in males and in females in the mouse cortex at P60 84

LIST OF TABLES

Table 3.1 Sample size distribution of FCG mouse data..... 85

Table 3.2 Number of significant differentially splicing events in six pairwise comparisons..... 86

Table 3.3 Sample size distribution of PTBP2 knockdown data..... 87

Table 3.4 Sample size distribution of Xist knockdown data 88

ACKNOWLEDGEMENTS

First, I would express my deepest gratitude to my thesis advisor, Yi Xing, who has been extremely supportive to my research and career development during my PhD. I joined his lab in 2013 with little knowledge of the field and what I want to do in the future. It is him that taught me the basics of how to do research, how to think and work as an independent researcher, how to read and write scientific articles, and how to navigate my career through different paths. I wouldn't come to the current stage without his guidance. All the things I have learned in my PhD career will be extremely valuable for my future career.

I would also like to thank my committee members for their support and insightful advice: Douglas Black, Thomas Graeber and Qing Zhou. I am really grateful to have them in my committee.

I owe a special thanks to my girlfriend, Xuewen Quan. She has been extremely supportive since I knew her in day one. She encouraged me during the low points of my PhD and showed me enormous love, which helped me overcome all the obstacles. I couldn't achieve what I have achieved now without her.

My life would not have been so colorful without my friend and colleagues in the Xing lab: Jinkai Wang, Zhixiang Lu, Juw Won Park, Shihao Shen, Shaofang Li, Shayna Stein, Yu-ting Tseng, Chengyang Wang, Zijun Zhang, Yang Guo, Zhicheng Pan, Yang Pan, Yuanyuan Wang, Yongbo Wang, Samir Adhikari, Eddie Park, Ruijiao Xin, Yan Gao, Amal Katrib, Emad Bahrami-Samani, Xinyuan Chen, Harry Yang, Yang Xu, Levon Demirdjian, Wankun Deng, Yuan Gao, Feng Wang, Yungang Xu, Siwei Luo, and Kate Kadash-Edmondson.

The project for studying the regional variation of splicing QTLs in human brain is under review. I would like to thank all the authors of this project:

Harry Yang for downloading all the GTEx brain data and running rMATs on all the samples.
Kate Kadash-Edmondson for helping me write and edit the manuscript. And also polish all the figures.
Yang Pan for helping me write the code for sQTL calculation.
Zhicheng Pan for helping me with the downstream GWAS analysis.
Beverly L. Davidson and Yi Xing, the corresponding authors of this project, for conceive the project and guidance throughout the whole project.

Besides research, I have been learning violin since the second year of my PhD. Ever since that, my life has been completely changed. Here, I need to express my deepest gratitude to my violin teacher, Mika Krstic. He not only teaches me all the techniques of playing violin, but also teaches me the lessons I need to know to live a life. He is not only great teacher, but also a best friend. The one-hour practice everyday is the best time of my day. I couldn't imagine a life without violin and music.

Different from most PhD students, I don't have a healthy body. There are some health issues that have been with me throughout my PhD. I have experienced miraculous moments that everything seems cured. But I have also experienced moments that everything comes back. During my battle with the disease, there are lots of moments that I have thought about giving up my research career. However, my physician, Dr. Susanne B. Nicholas and Dr. Lawrence Holzman have been encouraging me to do what I really want to do and be the person that I really want to be all the time. In the meantime of setting me free from all the worries and concerns, they have also been trying hard to treat my disease. Although the disease is still with me, it has been more like a treasure to my life instead of an enemy. I owe great thanks to Dr. Susanne B. Nicholas and Dr. Lawrence Holzman. They saved my life.

Last but not the least, I would love to thank my parents. There is no way that I can thank them enough for their life long support and unconditional love. I would not be the person I am right now without them. For all the sufferings I have been through during my PhD, they suffered ten times more because there is nothing they can do. Mom and Dad, I know it is a long way and it is not easy. Thank you and I love you forever.

VITA

EDUCATION

- University of California, Los Angeles 2013 – 2019

  Ph.D. Candidate in Bioinformatics (advanced to doctoral candidacy on Aug. 2015)

  Ph.D. advisor: Dr. Yi Xing

- Tongji University, Shanghai 2009 – 2013

  B.E. in Bioinformatics

HONORS & AWARDS

- Jul. 2015 – Jun. 2016 University Fellowship, UCLA
- Oct. 2017 – Oct. 2018 Dissertation Year Fellowship, UCLA

TEACHING EXPERIENCE

- CHEM C200 - Genomics and Computational Biology (Instructor: Dr. Christopher J. Lee), University of California, Los Angeles (2017 Spring quarter)
  - Teaching assistant.

PUBLICATIONS

[1] **Zhang Y**, Yang H, Kadash-Edmondson KE, Pan Y, Pan Z, Davidson BL, Xing Y. Regional variation of splicing QTLs in human brain. Manuscript under review.

[2] Lin L., Park JW., Ramachandran S., **Zhang Y.**, Tseng YT., Shen S., Waldvogel H., Curtis M., Faull R., Troncoso J., Ross C., Davidson BL.$^{+}$, Xing Y.$^{+}$ Transcriptome sequencing reveals aberrant alternative splicing in Huntington's disease. *Human Molecular Genetics*,

2016; 25:3454-3466. ($^{+}$ joint corresponding authors).

[3] Gao J$^{\#}$, Huang Q$^{\#}$, Wu D, Zhang Q, **Zhang Y**, Chen T, Liu Q, Zhu R, Cao Z, He Y. Study on human GPCR-inhibitor interactions by proteochemometric modeling. *Gene*, 2013; 518 (1): 124-131. ($^{\#}$ equal contribution)

[4] **Zhang Y.,** Baker SS, Baker RD, Zhu R^{+}, Zhu L^{+}. Systematic analysis of the gene expression in the livers of nonalcoholic steatohepatitis: implications on potential biomarkers and molecular pathological mechanism. *PLoS One*, 2012; 7: e51131. ($^{+}$ joint corresponding authors)

[5] Wu D.$^{\#}$, Huang Q.$^{\#}$, **Zhang Y.**, Zhang Q., Liu Q., Gao J., Cao Z., Zhu R. Screening of selective histone deacetylase inhibitors by proteochemometric modeling. *BMC Bioinformatics*, 2012; 13: 212-221 ($^{\#}$ equal contribution)

Chapter 1 Introduction

Based on the central dogma, the genetic information in DNA is transcribed into RNA, and then translated into protein to carry out biological functions[1]. However, different cells in our body, although with identical DNA, can have different molecular features and carry out diverse biological functions. One mechanism enabling this one-to-many relationship to happen is through post-transcriptional regulation[2, 3]. Through post-transcriptional regulation, the same DNA can generate multiple different RNAs, which can be translated into different proteins. Post-transcriptional regulation includes but not limited to capping, alternative splicing, addition of poly(A) tail, RNA editing, transportation, degradation, etc.

Among the many post-transcriptional regulations, alternative splicing is a ubiquitous and important regulation process during which particular exons in a gene may be included or excluded from final mature messenger RNA (mRNA). Due to alternative splicing, a single gene can generate multiple transcripts, which will be translated into different proteins carrying out different biological functions. This leads to enormous diversity and complexity at the RNA level. In addition, more than 95% of human genes undergo this process[4].

Compared to other tissues, alternative splicing is highly abundant in brain[5]. Abnormality during alternative splicing can result in a series of diseases including many neurological disorders[6-9]. For diseases affected by alternative splicing, many of them are caused by genetically regulated alternative splicing events, which are also known as splicing quantitative traits (sQTLs)[10]. It has been reported recently that sQTL is a primary link between genetic variation and disease[11]. There are two studies investigating the role of sQTL in schizophrenia[12] and Alzheimer disease[13]. However, both of them are restricted to only one brain region. Consequently, as efforts to depict a comprehensive picture of sQTLs in human brain across

multiple brain regions and study its role in neurological disorders, Chapter 2 presents a study on sQTLs in human brain across 13 brain regions. In detail, Chapter 2 describes a study of sQTLs using human brain samples from 13 different brain regions. By calculating sQTLs in each brain region separately, hundreds of sQTLs were identified in each brain region. Some sQTLs were shared across brain regions, whereas others displayed regional specificity. These 'regionally ubiquitous' and 'regionally specific' sQTLs showed distinct positional distributions of single nucleotide polymorphisms (SNPs) within and outside essential splice sites, respectively, suggesting their regulation by divergent molecular mechanisms. Integrating the binding motifs and expression patterns of RNA binding proteins with exon splicing profiles, we uncovered likely causal variants of brain region specific sQTLs. Overall, our study reveals widespread regional variation of sQTLs in human brain, and demonstrates that such variation can be used to fine map causal variants of sQTLs and their associated neurological diseases

Although many studies have focused on alternative splicing and other molecular mechanisms in normal brain functions and neurological disorders, one piece that is mostly ignored is the difference between males and females[14-16], which is a phenomenon called sexual dimorphism. Emerging evidence show that there are widespread sex differences in the adult human brain[17-20]. These sex differences manifest not only in normal neurodevelopment[21, 22], but also in neurological disorders[23, 24]. Many studies have investigated this sex difference from neuroanatomical and behavioral perspectives, however, the underlying molecular mechanisms are poorly understood[24-27]. Four core genotype (FCG) system is a system used to study the sex differences on the molecular level due to its ability to separate the effect from sex chromosomes and effect from the animal's gonadal sex[28]. In Chapter 3, we described a study focusing on the sexual dimorphic alternative splicing in mouse brain using the FCG mouse model. In detail, X

chromosome inactivation is a well-studied process for female specific gene expression[29-32]. A long non-coding RNA called Xist is known as the driver of this sexually dimorphic gene expression. Given the key role of alternative splicing in regulating gene expression of most mammalian genes[33], we hypothesized that the sex difference on the alternative splicing level may contribute to the sexual dimorphism in mammalian brain. Recent studies have showed that RNA binding proteins (RBPs), which are the proteins that regulate alternative splicing, interacts with Xist[34-36]. These include the Polypyrimidine Tract Binding Proteins 1 and 2 (PTBP1 and PTBP2). The evidence supported a complex model with interactions between RBPs, alternative splicing events, and Xist in regulating the sexual dimorphism in mammalian brain. In order to decipher the detailed interaction and regulatory mechanism, we generated data using the four core genotype mouse model together with Xist knockdown and PTBP1/2 knockdown data. By identifying the difference on the splicing level between different genotype and knockdown condition, we are able to examine the mechanisms of sexually dimorphic splicing, which will shed light on the underlying mechanisms and regulatory networks of sexual dimorphism in mammalian brain.

References

1. Crick F. Central dogma of molecular biology. *Nature* **227**, 561 (1970).

2. Franks A, Airoldi E, Slavov N. Post-transcriptional regulation across human tissues. *PLoS computational biology* **13**, e1005535 (2017).

3. Keene JD. RNA regulons: coordination of post-transcriptional events. *Nature Reviews Genetics* **8**, 533 (2007).

4. Wang ET, *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470 (2008).

5. Yeo G, Holste D, Kreiman G, Burge CB. Variation in alternative splicing across human tissues. *Genome biology* **5**, R74 (2004).

6. Licatalosi DD, Darnell RB. Splicing regulation in neurologic disease. *Neuron* **52**, 93-101 (2006).

7. Manning KS, Cooper TA. The roles of RNA processing in translating genotype to phenotype. *Nature reviews Molecular cell biology* **18**, 102-114 (2017).

8. Raj B, Blencowe BJ. Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. *Neuron* **87**, 14-27 (2015).

9. Vuong CK, Black DL, Zheng S. The neurogenetics of alternative splicing. *Nature reviews Neuroscience* **17**, 265-281 (2016).

10. Park E, Pan Z, Zhang Z, Lin L, Xing Y. The expanding landscape of alternative splicing variation in human populations. *The American Journal of Human Genetics* **102**, 11-26 (2018).

11. Li YI, *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600-604 (2016).

12. Takata A, Matsumoto N, Kato T. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nature communications* **8**, 14519 (2017).

13. Raj T, *et al.* Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nature genetics* **50**, 1584-1592 (2018).

14. Arnold AP. Conceptual frameworks and mouse models for studying sex differences in physiology and disease: why compensation changes the game. *Experimental neurology* **259**, 2-9 (2014).

15. Cahill L, Aswad D. Sex influences on the brain: an issue whose time has come. *Neuron* **88**, 1084-1085 (2015).

16. Eliot L. The trouble with sex differences. *Neuron* **72**, 895-898 (2011).

17. Grgurevic N, Majdic G. Sex differences in the brain–an interplay of sex steroid hormones and sex chromosomes. *Clinical Science* **130**, 1481-1497 (2016).

18. Sacher J, Neumann J, Okon-Singer H, Gotowiec S, Villringer A. Sexual dimorphism in the human brain: evidence from neuroimaging. *Magnetic resonance imaging* **31**, 366-375 (2013).

19. Trabzuni D*, et al.* Widespread sex differences in gene expression and splicing in the adult human brain. *Nature communications* **4**, 2771 (2013).

20. Cosgrove KP, Mazure CM, Staley JK. Evolving knowledge of sex differences in brain structure, function, and chemistry. *Biological psychiatry* **62**, 847-855 (2007).

21. Chung WCJ, Auger AP. Gender differences in neurodevelopment and epigenetics. *Pflügers Archiv-European Journal of Physiology* **465**, 573-584 (2013).

22. Shi L, Zhang Z, Su B. Sex biased gene expression profiling of human brains at major developmental stages. *Scientific reports* **6**, 21181 (2016).

23. May M. Sex on the brain: unraveling the differences between women and men in neurodegenerative disease. (ed^(eds). Nature Publishing Group (2016).

24. Qureshi IA, Mehler MF. Genetic and epigenetic underpinnings of sex differences in the brain and in neurological and psychiatric disease susceptibility. In: *Progress in brain research* (ed^(eds). Elsevier (2010).

25. Ngun TC, Ghahramani N, Sánchez FJ, Bocklandt S, Vilain E. The genetics of sex differences in brain and behavior. *Frontiers in neuroendocrinology* **32**, 227-246 (2011).

26. Sekido R. The potential role of SRY in epigenetic gene regulation during brain sexual differentiation in mammals. In: *Advances in genetics* (ed^(eds). Elsevier (2014).

27. McCarthy MM, Arnold AP. Reframing sexual differentiation of the brain. *Nature neuroscience* **14**, 677 (2011).

28. Arnold AP, Chen X. What does the "four core genotypes" mouse model tell us about sex differences in the brain and other tissues? *Frontiers in neuroendocrinology* **30**, 1-9 (2009).

29. Gendrel A-V, Heard E. Noncoding RNAs and epigenetic mechanisms during X-chromosome inactivation. *Annual review of cell and developmental biology* **30**, 561-580 (2014).

30. Kamikawa Y, Donohoe ME. The dynamics of X‐chromosome inactivation in mouse development. *Molecular reproduction and development* **81**, 141-147 (2014).

31. Lee JT, Bartolomei MS. X-inactivation, imprinting, and long noncoding RNAs in health and disease. *Cell* **152**, 1308-1323 (2013).

32. Payer B, Lee JT. Coupling of X-chromosome reactivation with the pluripotent stem cell state. *RNA biology* **11**, 798-807 (2014).

33. Braunschweig U, Gueroussov S, Plocik AM, Graveley BR, Blencowe BJ. Dynamic integration of splicing within gene regulatory pathways. *Cell* **152**, 1252-1269 (2013).

34. Chu C*, et al.* Systematic discovery of Xist RNA binding proteins. *Cell* **161**, 404-416 (2015).

35. McHugh CA*, et al.* The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521**, 232 (2015).

36. Minajigi A*, et al.* A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science* **349**, aab2276 (2015).

Chapter 2 Regional variation of splicing QTLs in human brain

2.1 Introduction

Alternative splicing is a crucial post-transcriptional regulatory mechanism that enables enormous RNA-level complexity. Through alternative splicing, specific exons may be included or excluded from the final mature messenger RNA (mRNA), allowing a single gene to generate multiple transcript and protein products with unique biological functions[1]. In humans, more than 95% of multi-exon genes undergo alternative splicing[2], with the nervous system being a prominent site[3, 4] Abnormalities in alternative splicing have been associated with multiple pathologies[5-8], including neurodegenerative disorders like Parkinson disease (PD)[9], Alzheimer disease (AD)[10, 11], amyotrophic lateral sclerosis (ALS)[12], and frontotemporal dementia (FTD)[13], as well as the neuropsychiatric disorders autism spectrum disorder (ASD)[14-17] and schizophrenia[18-20].

Genetic variants can alter exon inclusion or splice site usage, by creating or disrupting splice sites[21] or other cis splicing regulatory elements within precursor mRNAs[22-24]. Such genetically regulated alternative splicing events were recently recognized as a primary link between genetic variation and disease[25]. Splicing quantitative trait loci (sQTL) analysis is a common method for discovering genotype-splicing associations. In an sQTL analysis, the splicing level of an alternative exon or splice site is treated as a quantitative trait and tested for association with genotype across a population. Using this approach, two recent reports described effects of genetic variants on alternative splicing in human brain as well as their associations with schizophrenia[20] and AD[26]. However, both studies were restricted to a single brain region, and therefore the full spectrum and regional variability of sQTLs in human brain were not revealed.

In this study, we systematically discovered sQTLs in 13 regions of human brain using RNA-seq and genotype data from the Genotype-Tissue Expression (GTEx) project (**Figure 2.1A**). By comparing sQTL signals across different brain regions, we found that some sQTLs were shared across brain regions, whereas others displayed regional specificity. These 'regionally ubiquitous' and 'regionally specific' sQTLs were enriched for genetic variants within and outside essential splice sites, respectively, suggesting their regulation by distinct molecular mechanisms. Integrating RNA binding protein (RBP) motifs with RBP expression and exon splicing profiles, we demonstrate that regional variation of sQTLs can be used to fine map causal variants underlying genetically regulated alternative splicing events and their associated neurological diseases.

2.2 Results

2.2.1 Gene expression and alternative splicing in human brain

We downloaded RNA-seq data from the GTEx project website and collected corresponding gene expression, alternative splicing, and genotype information of 1,209 human brain samples comprising 13 brain regions (**Supplementary Figure 2.7**) (**Methods**). We applied our computational tool rMATS[27], developed for quantifying alternative splicing events from large-scale RNA-seq data, to estimate exon splicing levels (percent spliced in, or PSI) from the GTEx RNA-seq data. After applying filters to select high-confidence alternative splicing events with reliable RNA-seq quantitation, we collected 10,665 skipped exons (SE), 1,443 alternative 5' splice site (A5SS) events, and 2,434 alternative 3' splice site (A3SS) events (**Supplementary Figure 2.8**). The t-distributed stochastic neighbor embedding (t-SNE) algorithm was used to inspect relationships among samples. Using this approach, we were able to separate samples

from different brain regions and to recapitulate relationships between brain regions using gene expression and alternative splicing information (**Figure 2.1B** shows results for SE events; similar patterns were obtained for A5SS and A3SS events). Although we tested additional co-variates such as age and gender, as well as various confounding or batch effects, these other factors did not enable clear separation of samples based on gene expression or alternative splicing (**Supplementary Figure 2.9**). Our results indicate that brain region is the dominant contributing factor to variations in gene expression and alternative splicing in human brain.

To quantify contributions of different biological factors to alternative splicing and gene expression, we fit a linear mixed model to each alternative splicing event or gene. We included brain region, age, and gender in the model as main effects and controlled for confounding factors (**Methods**). Using a false-discovery rate (FDR) of <5%, we found that gene expression and alternative splicing showed substantial variation depending on brain region, but much less variation depending on age or gender (**Supplementary Figure 2.10**). These findings were consistent with the t-SNE results.

Given this strong brain-regional variation, we validated our results using previously reported brain region-specific alternative splicing events and genes; *Neurexin* is one prominent example. Through alternative splicing, *neurexins 1, 2,* and *3* generate thousands of mRNA and protein products[28-30]. Ehrmann et al. reported that the tissue-dependent RBP KHDRBS3 (referred to as T-STAR) and its paralog KHDRBS1 (referred to as Sam68) regulate brain region-specific alternative splicing of *neurexin1-3* exon 4[31]. Specifically, exon 4 was included at low levels in brain regions with a high gene expression ratio of *KHDRBS3* versus *KHDRBS1* (e.g., cortex) and at high levels in regions with a low gene expression ratio (e.g., cerebellum). Plotting the exon inclusion level of *neurexin2* exon 4 with the *KHDRBS3*: *KHDRBS1*expression ratio

(**Supplementary Figure 2.11**), we found consistent results with those reported previously[31]. Exon 4 of *neurexin1* and *neurexin3* showed a similar pattern based on our data.

2.2.2 Identification of sQTLs in 13 brain regions

To elucidate genetic regulation of alternative splicing in human brain, we performed sQTL analyses in each of the 13 brain regions separately (**Methods**). After controlling for confounding factors, we identified sQTLs by using linear regression to calculate the association between exon splicing levels (PSI values) and SNP genotypes in each brain region (**Methods**). The top SNP with the smallest *P* value within 200 kb of each alternative splicing event (SE, A5SS, or A3SS) was selected as the sQTL SNP in that brain region. The corresponding alternative splicing event was defined as the sQTL event. 387 to 849 significant sQTL events ($P < 10^{-5}$) were found in each brain region (**Figure 2.2A**), with most being specific to only a few regions (**Figure 2.2B**). We also identified 133 sQTL exons that were significant in all 13 brain regions when summing all three splice types. These data indicate that sQTLs in human brain can be either 'regionally ubiquitous' or 'regionally specific'.

Next, we annotated sQTL exons in terms of disease risk by calculating the linkage disequilibrium (LD) between sQTL SNPs and SNPs identified as disease associated variants by Genome Wide Association Studies (GWAS) (**Methods**). We used disease ontology terms in the NHGRI-EBI GWAS Catalog[32], focusing on terms related to 14 neurological disorders: AD, ALS, PD, FTD, epilepsy, ASD, schizophrenia, bipolar, depression, attention deficit hyperactivity disorder (ADHD), glioma/glioblastoma, multiple sclerosis, narcolepsy, and stroke. Any sQTL event having a significant SNP in high LD ($r^2 > 0.8$) with a neurological disorder-related GWAS

variant was considered to be a disease sQTL event related to that specific disorder. **Figure 2.2C** shows the number of disease sQTL events for each disorder in each brain region.

We identified a predominance of disease sQTL events that were related to neuropsychiatric disorders (e.g., schizophrenia). This result was not surprising, given the greater prevalence of reported neuropsychiatric disease-related GWAS variants in the literature[33, 34]. We also identified sQTL events related to neurodegenerative diseases (e.g., AD and PD). Neurodegenerative disease-related sQTL exons tended to be brain region-specific, whereas neuropsychiatric disorder-related sQTL exons were shared across a larger number of brain regions.

2.2.3 Relationship among SNP position, significance, and brain region specificity of sQTLs

We examined the relationship among SNP position, significance, and brain region specificity of sQTLs. Here, we describe the result for SE events, but similar results were observed for A5SS and A3SS events.

To evaluate the positional distribution of sQTL SNPs, we plotted the –log10 *P* values of significant sQTL SNPs in all 13 brain regions together against the distance to the sQTL exons (**Supplementary Figure 2.12**). As expected and consistent with previous results[20], we found that sQTL SNPs closer to the alternative exons tended to be more significant. Next, we examined SNP positions for all exons included in the sQTL analysis for each brain region. For each exon, we obtained the –log10 *P* values of all SNPs within 200 kb of the exon, together with the distance of the SNPs to the exon. As the *P*-value cutoff for significant sQTLs increased, we observed an increase in the fraction of exons with at least one significant SNP within 300 bp of

exons (**Figure 2.3A**). This result, which was reproducible across all brain regions, indicated enrichment in the local regulation of significant sQTLs by exon-proximal SNPs.

To obtain a fine-grained picture of the relationship between SNP position and significance of sQTLs, we classified all SNPs within 200 kb of all sQTL exons into 5 groups, based on the SNP position relative to the corresponding sQTL exon. In the figures that follow, 5'SS represents the 9 bases (3 exonic and 6 intronic) around the 5' splice site, and 3'SS represents the 23 bases (20 intronic and 3 exonic) around the 3' splice site[35]. We found that SNP position was associated with sQTL significance (**Figure 2.3B**), with 5'SS SNPs having the highest sQTL significance, followed by 3'SS SNPs, exonic SNPs, and proximal intronic SNPs (≤300 bp of exons). SNPs in distal intronic regions (>300 bp from exons) had the largest overall *P* values, indicating their relatively minor impact on splicing. These results are consistent with previous sQTL studies on human B-lymphoblastoid cell lines[36, 37].

Next, we asked for individual sQTLs, whether their numbers of significant brain regions were associated with their SNP positions and levels of sQTL significance. First, for each sQTL exon we plotted the sQTL significance against the number of significant regions (**Figure 2.3C**). An sQTL exon with at least one significantly associated SNP within the 200-kb window in a brain region was considered significant in that region. Overall significance of the sQTL exon was calculated as the median of the smallest *P* value from all regions in which the sQTL was significant. We found that regionally ubiquitous sQTL exons tended to have overall higher sQTL significance than regionally specific sQTL exons. Second, we analyzed the relationship between the SNP position and brain region specificity of sQTLs. For each sQTL exon, we first classified all significant SNPs within 200 kb across all brain regions into 5 categories: splice site dinucleotide (GT and AG for donor and acceptor sites, respectively), splice site (9 nt around

5'SS, 23 nt around 3'SS, excluding the dinucleotide), exon body, proximal intronic region (≤300 bp of the exon), and distal intronic region (>300 bp of the exon). Then we assigned each sQTL exon to a specific category based on whether it had a significant SNP in that category, prioritized sequentially from splice site dinucleotide to distal intronic region. This information was plotted together with the number of significant regions for each sQTL exon (**Figure 2.3D**). Of note, the statistical significance of some sQTLs significant in a limited number of brain regions is on the boundary of the significance cutoff. Those sQTLs are more likely to be false positive. Therefore, we used permutation FDR<10% instead of *P* value < 10^{-5} as the cutoff for significance here to get a high-confident set of sQTLs. We observed that sQTL exons with significant SNPs closer to their splice sites tended to be regionally ubiquitous rather than regionally specific. For example, for sQTLs with significant SNPs located at the splice site dinucleotide, 53% were significant in all 13 brain regions and 93% were significant in the majority (>=7 out of 13) of brain regions. By contrast, for sQTLs with significant SNPs located only in proximal or distal intronic regions, the majority (50% or 90% respectively) were significant in no more than 4 brain regions (**Figure 2.3D**). The result is consistent when we use permutation FDR<5% and permutation FDR<1% as the significance cutoff (**Supplementary Figure 2.13**).

2.2.4 Examples of regionally ubiquitous and regionally specific sQTLs

We used stringent criteria to define high-confident sets of regionally ubiquitous and regionally specific sQTLs in human brain. Specifically, sQTLs reaching transcriptome-wide significance (permutation FDR<10%) in at least 10 brain regions were considered regionally ubiquitous (**Figure 2.4A**), whereas sQTLs reaching transcriptome-wide significance (permutation FDR<10%) in at least one brain region but having no more than 4 brain regions with nominal *P*

value more significant than 10^{-5} were considered regionally specific (**Figure 2.4E**) (**Methods**). 148 and 758 sQTLs were defined as regionally ubiquitous and specific, respectively.

For each sQTL exon we obtained and plotted the coefficient from the linear regression analysis of exon splicing levels against genotypes in each brain region, which represented the effect size of the sQTL signal in that region, using the most representative sQTL SNP across the 13 brain regions (**Methods**) (**Figure 2.4A, 2.4E**). Consistent with **Figure 2.3D**, we found that regionally specific sQTLs were enriched for SNPs in distal intronic regions, while regionally ubiquitous sQTLs were enriched for SNPs in splice sites and exons (**Supplementary Figure 2.14**, χ^2 test *P* value < 2.2×10^{-16}). This result is consistent using permutation FDR<5% and permutation FDR<1% as the significance cutoff (**Supplementary Figure 2.14**, χ^2 test *P* value < 2.2×10^{-16} for permutation FDR<5% and χ^2 test *P* value = 1.8×10^{-7} for permutation FDR<1%).

For regionally ubiquitous sQTLs, we illustrate the example of the sQTL exon in Regulator of Telomere Elongation Helicase 1 (*RTEL1*), which protects telomeres during DNA replication. Several variants in *RTEL1* have been reported to be associated with risk of glioma[38, 39]. We observed significant associations between PSI values of *RTEL1* exon 23 and genotypes of rs6062302 in all brain regions (**Figure 2.4B**). **Figure 2.4C** showed the correlation in the cerebellar hemisphere and the nucleus accumbans basal ganglia (both highly significant). SNP rs6062302 is a glioma-related GWAS variant and it is in high LD with another glioma-related GWAS variant (**Figure 2.4D**).

Beyond the example of *RTEL1* shown above (**Figure 2.4B-D**), our analyses identified a total of 148 regionally ubiquitous sQTLs, including those associated with GWAS signals. For example, we identified significant correlation between rs67573812 and *C8orf59* exon 2 (**Supplementary Figure 2.15A**). This SNP, located at the 5'SS (dinucleotide) of *C8orf59* exon 2,

disrupts a canonical splice site (GT to GA). A second example involved flotillin 1 (*FLOT1*) exon 5 and rs1059612, located in the proximal intronic region of exon 5 (**Supplementary Figure 2.15B**). FLOT1 is a membrane-raft associated protein that is involved in synaptic transmission and synapse formation[40, 41]. Expression QTL and GWAS studies have associated *FLOT1* expression with genetic risks for schizophrenia and major depressive disorder[42, 43]. We found that the SNP was in high LD with GWAS variants related to schizophrenia, bipolar, ADHD, unipolar depression, and ASD. Another example involved *SLC39A13* exon 5 and rs2293576 (**Supplementary Figure 2.15C**), located within the exon body of exon 5 and in high LD with the AD-related GWAS variant rs10838725. This finding is consistent with previous reports that *SLC39A13* is within the LD block of rs10838725, an AD-risk SNP identified by The International Genomics of Alzheimer's Project[44, 45].

For regionally specific sQTLs, we illustrate the correlation between PSI values of *SLC26A10* exon 12 and genotypes of rs1871417, which was significant only in cerebellar regions (**Figure 2.4F**). **Figure 2.4G** shows the correlation in one significant region (cerebellar hemisphere) and one insignificant region (nucleus accumbans basal ganglia). Our analysis showed that the sQTL SNP of *SLC26A10* exon 12 (rs1871417) was in high LD with an immune system disease-related GWAS variant rs10876993 (**Figure 2.4H**). More specifically, rs10876993 was reported to be related to Celiac disease and Rheumatoid arthritis[46].

We also found many other sQTLs restricted to specific brain regions. For example, the association between rs6580200 and *CXXC5* exon 3 was only significant in cerebellar hemisphere (**Figure 2.5A**). Previous pathway-based analysis determined *CXXC5* to be a schizophrenia gene[47]. The SNP that we identified was in high LD with schizophrenia-related GWAS variants. In *TRIM26*, the association between rs971570 and exon 2 was significant only in frontal cortex

(**Figure 2.5B**). Although the function of TRIM26 is unknown, several studies found that *TRIM26* was differentially expressed between schizophrenia patients and controls[48, 49]. Here, we found that this SNP was in high LD with GWAS variants related to schizophrenia, as well as ADHD, ASD, bipolar, and unipolar depression. In *POU6F1*, the association between rs4076884 and exon 4 was significant only in cerebellum and cerebellar hemisphere (**Figure 2.5C**).

2.2.5 Using regional specificity of sQTL signals to prioritize causal sQTL cis variants and trans regulators

To understand why the genetic regulation of alternative splicing could be restricted to specific brain regions, we investigated the underlying regulatory mechanisms of regionally specific sQTLs. One hypothesis is that the brain region-specific expression of RBPs, functioning as trans-acting regulators of alternative splicing, could contribute to the brain region specificity of sQTLs. To test this hypothesis, we analyzed the gene expression patterns of RBPs across brain regions. Many RBPs showed highly brain region-specific expression patterns (**Figure 2.5D**) and, thus, could potentially provide a source of regional regulation.

To prioritize possible cis variants and identify potential trans regulators of brain region-specific sQTLs, we used the deep-learning model DeepBind[50] to predict the impact of a given SNP on RBP-RNA binding. DeepBind predicts RBP sequence specificities based on sequence features from CLIP-seq experiments. We assessed the RBP binding effects of all significant SNPs within 300 bp of a given sQTL exon (**Methods**). We then incorporated RBP expression data to further prioritize candidate cis variants and trans regulators (**Methods**). Intuitively, if a SNP regulates the splicing of an exon through affecting RBP binding, we would expect to see greater sQTL significance and effect size in brain regions where the RBP is highly expressed. To

test this, we classified the brain samples into two groups: 1) samples from brain regions where the sQTL is significant; 2) samples from brain regions where the sQTL is insignificant. For each RBP, a Wilcoxon rank sum test was performed to see if the RBP is differentially expressed between the two groups.

Using this strategy, we fine-mapped possible causal variants for the brain region-specific sQTL for exon 3 of microtubule-associated protein tau (*MAPT*). *MAPT* encodes tau, an abundant protein in the nervous system that promotes microtubule assembly and stability. Aggregation of hyperphosphorylated tau protein is a primary marker of AD[10, 51, 52]. Moreover, numerous studies have implicated tau in the pathogenesis of other neurological disorders, including PD[53]. Human *MAPT* contains 16 exons (**Figure 2.6A**), with exons 2, 3, and 10 being alternatively spliced to generate six isoforms. The sQTL exon that we identified was exon 3. **Figure 2.6B** shows the positional distribution and sQTL *P* value of SNPs within 200 kb of *MAPT* exon 3 in cerebellum. In a local window including 300 bp of upstream and downstream intronic regions, there were six significant SNPs that were in high LD with sQTL SNP rs62055489 (>100 kb from *MAPT* exon 3). Testing the effects of these six SNPs against all RBPs in the DeepBind model, we identified a strong effect of rs17651213 (G>A) on RBFOX binding (**Figure 2.6C**). We found that the SNP created a consensus binding site of splicing factors RBFOX downstream of *MAPT* exon 3 (**Figure 2.6D**). When we checked the expression levels of *RBFOX1*, *RBFOX2*, and *RBFOX3*, we found that *RBFOX2/3* had the highest expression levels in cerebellum and cerebellar hemisphere, two brain regions where the sQTL was also significant (**Figure 2.6E** and **Supplementary Figure 2.16**).

Compared with RBFOX3, the splicing regulation mechanism and RNA binding of RBFOX2 is well studied. Therefore, we picked RBFOX2. Nevertheless, it could be a

combinatorial regulation by both RBFOX2 and RBFOX3. Of note, rs17651213 was located 88 bp downstream of *MAPT* exon 3. When the alternative allele that creates the RBFOX binding site was present, we observed greater inclusion of the exon (**Supplementary Figure 2.17** and **Supplementary Figure 2.18**), consistent with prior reports that RBFOX2 binding downstream of alternative exons promotes exon inclusion[54]. Finally, in accordance with the importance of tau protein in AD and PD pathogenesis, we found that rs17651213 was in high LD with 11 PD GWAS variants, and one AD GWAS variant (**Figure 2.6F**). Overall, by integrating RBP motifs, region-specific RBP gene expression, and sQTL pattern, we identified rs17651213 as the likely causal variant that regulates *MAPT* exon 3 splicing by creating an RBFOX2 binding site downstream of the exon.

We also identified other brain region-specific sQTLs with potential relevant SNPs and RBP regulators. For example, DeepBind prediction indicated that the sQTL SNP rs6580200 regulates splicing of *CXXC5* exon 2 by disrupting an HNRNPK binding site (**Supplementary Figure 2.19**). This sQTL was significant in cerebellar hemisphere, where HNRNPK was highly expressed (**Supplementary Figure 2.20**). A second example was rs4077093, which may regulate splicing of *POU6F1* exon 4 by disrupting NOVA1 binding (**Supplementary Figure 2.21**). The sQTL was significant in cerebellum and cerebellar hemisphere, where NOVA1 was highly expressed (**Supplementary Figure 2.22**). Using the well-defined consensus motif of NOVA1 (YCAY), we performed a motif scan and identified the effect of rs4077093 on NOVA1 binding. Specifically, rs4077093 (A>C) was predicted to disrupt a NOVA1 motif within the exon and found to be associated with increased exon inclusion, consistent with previous reports that NOVA1 binding within alternative exons promotes exon skipping.

2.3 Discussion

Alternative splicing is known to influence biological functions and disease processes, but much remains unknown about the disease causality and underlying regulatory mechanisms of genetically regulated alternative splicing events. A number of studies have performed sQTL analyses to interrogate genotype-splicing associations in human cell lines or tissues, including brain where alternative splicing is highly prevalent[20, 26, 55]. However, previous studies largely conducted sQTL analyses one tissue at a time, and did not compare and contrast sQTL signals across multiple tissues or cell types. In this work, we analyzed sQTLs in 1,209 human brain samples across 13 brain regions. We identified regionally ubiquitous sQTLs with significant signals across a large number of brain regions, as well as regionally specific sQTLs whose significance was restricted to specific brain regions. Many sQTLs were associated with GWAS signals. Together, our study provides a comprehensive view of genetically regulated alternative splicing events in human brain, and reveals their associations with neurological traits and diseases.

One major challenge in genetic association studies of molecular and phenotypic traits is to identify the causal variants and regulatory mechanisms underlying the observed association. In an sQTL analysis, multiple SNPs in LD with each other can be significantly associated with the levels of exon splicing, and it is difficult to pinpoint the specific variant(s) causal for the observed sQTL signal. Our analysis of SNP significance as a function of distance to sQTL exons suggests that the majority of sQTLs, especially those with strong effects on splicing, likely have their causal variants located within the exonic or proximal intronic (≤300 bp of exon) regions. This is consistent with prior findings that sequence information within this window has a high predictive power for alternative splicing patterns[56, 57]. By examining the SNP positions for

regionally ubiquitous and specific sQTLs, we uncover the likely molecular mechanisms responsible for generating these two types of sQTLs. Regionally ubiquitous sQTLs are enriched for SNPs located on the 5' or 3' splice site. As splice sites play an essential role in splicing, it is not surprising that SNPs strengthening or weakening splice site signals have ubiquitous effects across different brain regions or tissue types. By contrast, regionally specific sQTLs tend to have SNPs located outside the 5' or 3' splice site, and are enriched for SNPs located in distal intronic regions. Causal SNPs underlying regionally specific sQTLs likely affect splicing by modulating the interactions of tissue- or cell-type specific splicing factors with the pre-mRNA.

We developed an integrative strategy to fine map the likely causal variants for brain region-specific sQTLs. We assessed SNP effects on RBP-RNA binding, and compared brain region-specific RBP gene expression patterns and sQTL signals, to identify potential causal cis variants and trans regulators for regionally specific sQTLs. Using this strategy, we identified and highlighted the likely molecular mechanisms for several brain region-specific sQTLs (*MAPT*, *CXXC5, POU6F1*), in which SNPs created or disrupted the binding sites of splicing factors with region-specific expression patterns. As the research community continues to accumulate population-scale RNA-seq datasets[58] as well as RNA binding profiles and specificities of RBPs[59], we envision that this integrative strategy can be applicable to fine map causal variants of tissue- or cell-type-specific sQTLs in diverse biological systems.

Our study highlights the molecular mechanism of how genetic polymorphisms in the human genome can have tissue- or cell-type specific effects on splicing. As one example, we found that a G-to-A SNP in *MAPT* (rs17651213) was associated with increased inclusion of *MAPT* exon 3 and created the consensus binding site of the splicing factor RBFOX2. This association was specific to cerebellar tissues, where RBFOX2 was highly expressed. This

example illustrates a general scenario that a SNP alters the putative binding site of a trans splicing regulator with the RNA, but its molecular impact is dependent on the concentration of the regulator such that the SNP only alters splicing in tissues or cell types where the regulator is highly expressed (**Figure 2.6G**).

One limitation of this study is that it used bulk tissues from the GTEx project, and samples from each brain region represented a mixture of multiple cell types. This may dilute the power of detecting tissue- or cell-type specific sQTLs, especially if the cell type affected represents a minor population of cells in the bulk tissues. For example, it is well known that RBFOX2 expression is highest in neuronal cells[60-62] and the cerebellum has the highest density of neurons among brain regions[63, 64]. Therefore, the significant association between rs17651213 and *MAPT* exon 3 in cerebellar tissues could be related to neuron-specific splicing regulation through RBFOX2, and it is possible that a stronger SNP-splicing association can be observed specifically in neurons. Future studies using population-scale RNA-seq data of purified cell types or single cells may further expand the catalog and our mechanistic understanding of genetically regulated alternative splicing events in human brain.

2.4 Methods

2.4.1 Datasets and processing of GTEx brain tissue samples.

RNA-seq data (BAM files; v7, June 2017 release) for human brain tissues from postmortem donor specimens were downloaded from the GTEx Portal website (http://www.gtexportal.org/). RNA-seq data were available for 1,409 samples from 13 brain regions of 201 individuals. An ultra-fast version of rMATS (rMATS-turbo, https://rmats.sourceforge.io/rmats4.0.2/) was run on BAM files to obtain exon inclusion levels (PSI values), with missing PSI values imputed by

using k-nearest neighbors. To ensure reliability of the imputation and downstream analyses, we only included an exon in our analyses if it passed the following filters:

1) Average PSI $\hat{I}$ [0.05 0.95],

2) Average total count (inclusion count + skipping count) ≥10,

3) Missing PSI value < 5%,

4) max(PSI) – min(PSI) > 0.05

Processed gene expression data (in transcripts per kilobase million, TPM) was downloaded from the GTEx Portal. Processed gene expression data was available for 1,258 of the 1,409 samples with RNA-seq data. Median normalization was performed on each sample for between-sample normalization.

Genotype data for GTEx brain samples were downloaded from the database of Genotypes and Phenotypes (dbGaP, https://www.ncbi.nlm.nih.gov/gap/). Variants with a minor allele frequency < 0.05 were removed before analysis. The final dataset for sQTL analysis comprised 1,209 samples for which genotype data, corresponding RNA-seq, and gene expression data were available.

2.4.2 Covariate and confounding factor analysis for splicing and gene expression data.

To determine the extent to which splicing (PSI) results were affected by differences in age, brain region or sex between tissue samples, the following linear mixed model was used:

$$\psi_{ij} = logit^{-1}(\mu_i + \alpha Age_j + \beta BrainRegion_j + \gamma Sex_j + P_{ij} + \sum_{k=1}^{N} \delta_k SV_k + \varepsilon_i)$$

$$P_{ij} \sim N(0, \sigma_{ij}^2)$$

where y_{ij} is the PSI for exon i of sample j; and Age_j, $BrainRegion_j$ and Sex_j are the age, brain region and sex of sample j with regression coefficients a, b and g, respectively. Random effect term P_{ij} was introduced to account for the fact that multiple GTEx samples may come from the same donor. To control for all other known and unknown confounding factors (e.g., ancestry, BMI, batch effects, etc.), the following surrogate variable(s) estimated by Surrogate Variable Analysis (SVA)[65] was used:

$$SV_k (1 \leq k \leq N)$$

where SV_k is the value of surrogate variable (SV) k with regression coefficient d_k; N is the total number of SVs that are not correlated with age, brain region or sex; e_i is the error term; and m_i is the regression intercept for exon i. The model was fitted by using the *lmer* function in the lme4 package in R. For each exon, we used a least-squares approach to estimate regression coefficients and a likelihood ratio test to estimate significance. If a, b or g significantly deviated from 0, then the exon was considered to be dependent on age, brain region or sex. False discovery rate (FDR) < 5% was used as the cutoff for significance.

We performed the same covariate and confounding factor analysis on gene expression data by replacing PSI with normalized TPM, but with one key difference. For gene expression analysis, SVA estimated 1,232 SVs, which generated more parameters than could be analyzed by the linear mixed model. As t-SNE analysis did not reveal any obvious effect of confounding factors on gene expression (**Supplementary Figure 2.9**), we excluded SVs from the analysis.

2.4.3 Correcting PSI values for confounding factors prior to sQTL analysis.

Prior to the sQTL analysis, we corrected the PSI value by removing potential confounding factors. First, we performed principal component analysis in PLINK on GTEx genotype data to

obtain population structure. This analysis was performed on a genome-wide set of LD-pruned variants ($R^2 > 0.2$). The top three principal components captured most of the genotype variance and, therefore, were used as covariates to account for population structure.

Next, we used SVA to remove potential confounding factors from sources other than brain region. As the aim of the sQTL analysis was to identify the sQTL landscape across different brain regions, we included only one factor (brain region) in the sQTL analysis. Any variation that could not be explained by brain region was estimated by SVA and removed.

We fitted the following linear model:

$$\psi_i = logit^{-1}(\mu_i + \sum_{m=1}^{M} SV_m + \sum_{k=1}^{3} PC_k + \varepsilon_i)$$

where M is the total number of SVs uncorrelated with brain region. SV_m is the value of surrogate variable (SV) m. PC_k is the value of principal component k. The top three principal components were used to account for the population structure. We used the residual of the model as the corrected PSI value in sQTL analysis.

2.4.4 Identification of sQTLs.

After the corrected PSI value was obtained, all SNPs within a 200-kb window on each side of an alternative splicing event were fitted with a linear regression. We estimated the FDR using a permutation procedure to obtain the null distribution of *P* values. For each alternative splicing event, we recorded the minimum *P* value over all *cis* SNPs in each of five permutations and used this set of *P* values as the empirical null distribution. For a given FDR value *f*, we defined the *P*-value cutoff *z* as: $P(p_0 < z)/P(p_1 < z) = f$, where $P(p_0 < z)$ is the fraction of expected $P < z$ from the null distribution, and $P(p_1 < z)$ is the fraction of observed $P < z$ from the real data.

We defined the sQTL SNP of an alternative splicing event as the closest SNP with the most significant association, using a two-tier cutoff for significance: 1) FDR < 10%, and 2) an uncorrected $P < 10^{-5}$. The corresponding P value of FDR = 10% is ~10^{-6} for each brain region, thus events meeting this threshold represent a higher confidence set of sQTLs. This calculation was done in all 13 brain regions separately.

2.4.5 LD calculation.

Most GWAS variants have been identified in samples of European ancestry. Therefore, for LD calculations, we used genotype data from the CEU population (Utah residents with ancestry from northern and western Europe). Genotype data were downloaded from the 1000 Genomes Project (HapMap Project Genome Browser version E, using data release #28 (Phase II+III), https://www.internationalgenome.org/) and processed as described previously[66]. To calculate LD between SNPs, we used PLINK (with parameters --no-fid --no-parents --r2 --ld-window-kb 1000 --ld-window 99999 --ld-window-r2 0.8). SNPs with $r^2 > 0.8$ were defined as being in high LD.

We calculated the LD of sQTL SNPs (for SE, A5SS and A3SS events) with SNPs from the 1000 Genomes Project. SNPs that were in high LD with sQTL SNPs were annotated by using GWAS variants from the GWAS Catalog v1.0.1 (ftp://ftp.ebi.ac.uk/pub/databases/gwas/releases/2017/08/01/). We used 14 keywords to classify GWAS traits into disease categories: ‘Alzheimer’, ‘amyotrophic lateral sclerosis’, ‘Parkinson’, ‘frontotemporal dementia’, ‘epilepsy’, ‘autism’, ‘schizophrenia’, ‘bipolar’, ‘depression’, ‘attention deficit hyperactivity disorder’, ‘glio’, ‘multiple sclerosis’, ‘narcolepsy’, and ‘stroke’. After identifying GWAS SNPs that were in high LD with sQTL SNPs, we obtained the GWAS

SNP disease ontology information ('MAPPED_TRAIT' column in GWAS Catalog table). If the disease ontology information contained any of the 14 keywords, then the sQTL SNP and the corresponding alternative splicing event were considered to be related to that specific disease.

2.4.6 Selection of the most representative sQTL SNP across 13 brain regions.

For each sQTL event, we collected the sQTL SNPs for each brain region (see 'Identification of sQTLs'). Among all significant (FDR < 10%) sQTL SNPs, we selected the SNP that was the top (i.e. most significant) SNP in the largest number of brain regions as the most representative sQTL SNP. In the event of a tie, we compared competing SNPs and chose the SNP that was significant (FDR < 10%) in the largest number of brain regions. If sQTL SNPs were still tied, then we chose the SNP closest to the sQTL event as the most representative sQTL SNP across all 13 brain regions.

2.4.7 Defining regionally ubiquitous and regionally specific sQTLs.

We considered an sQTL event to be regionally specific if: 1) the *P* value of the most representative SNP across 13 brain regions was $<10^{-5}$ in no more than 4 regions, and 2) the *P* value exceeded the FDR cutoff (FDR < 10%) in at least one region. We considered an sQTL event to be regionally ubiquitous if the *P* value of the most representative SNP across 13 brain regions exceeded the FDR cutoff (FDR < 10%) in at least 10 brain regions. We applied a two-tier cutoff for sQTL significance (FDR < 10%, uncorrected $P < 10^{-5}$) to ensure a reliable definition of significant and insignificant events.

2.4.8 Quantifying the effect of SNPs on RBP-RNA binding using DeepBind.

We used the DeepBind model[50] to quantify the effect of SNPs on RBP-RNA binding. DeepBind takes RNA sequences as input and output DeepBind scores, which quantify the binding specificity of different RBPs for the input sequences. The DeepBind scores can be used to generate mutation maps, which visually display the impact of sequence variants on RBP-RNA binding.

First, we collected all significant SNPs (and 20-bp flanking sequences on both sides) within 300 bp of the nearest splice site of each sQTL event across all 13 brain regions. Next, we tested the effect of each significant SNP on the binding of 102 human RBPs on which the DeepBind model had been trained. Following the procedure described in Alipanahi et al.[50], we calculated DeepBind scores and generated mutation maps for 41-bp sliding windows using a 20-bp motif detector. We used a modified approach to generate the mutation map for each SNP-RBP pair. Specifically, whereas Alipanahi et al.[50] was primarily concerned with the ability of a sequence variant to decrease the binding score (i.e., base height was unchanged if a sequence variant increased binding), our study was also concerned with the ability of a sequence variant to increase binding. To account for this, we calculated the sum of the score on each base by using the absolute value of the four DeepBind scores on that base, and then scaled the height of each base by using this sum. In this way, the height of each base was proportional to the ability of the base to change (increase or decrease) binding.

To enhance our ability in identifying potential causal variants underlying regionally specific sQTLs, we analyzed RBP expression data across the 13 brain regions. If a SNP regulates exon splicing by altering RBP binding, then we would expect to see higher sQTL significance in brain regions where the RBP is highly expressed. To test this possibility, we grouped samples based on whether they were from brain regions where the sQTL was significant or insignificant.

For each RBP, we applied a Wilcoxon rank sum test to determine whether the RBP was differentially expressed between the two groups, with FDR < 0.05 indicating significant differential expression. FDR is calculated on the *P* values of all the sQTL and RBP combinations.

2.4.9 Motif scan of NOVA1.

NOVA1 is not included in the DeepBind model but has a well-defined consensus motif (YCAY). Therefore, for NOVA1, instead of running the DeepBind model, we carried out a motif scan on the same set of SNPs as were used in the DeepBind analysis. For each SNP, we checked all possible 4-mer sequences overlapping with the SNP to see if the SNP created or disrupted a NOVA1 binding site.

2.5 References

1. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457 (2010).

2. Wang ET*, et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476 (2008).

3. Barbosa-Morais NL*, et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587-1593 (2012).

4. Yeo G, Holste D, Kreiman G, Burge CB. Variation in alternative splicing across human tissues. *Genome biology* **5**, R74 (2004).

5. Licatalosi DD, Darnell RB. Splicing regulation in neurologic disease. *Neuron* **52**, 93-101 (2006).

6. Manning KS, Cooper TA. The roles of RNA processing in translating genotype to phenotype. *Nature reviews Molecular cell biology* **18**, 102-114 (2017).

7. Raj B, Blencowe BJ. Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. *Neuron* **87**, 14-27 (2015).

8. Vuong CK, Black DL, Zheng S. The neurogenetics of alternative splicing. *Nature reviews Neuroscience* **17**, 265-281 (2016).

9. Trabzuni D*, et al.* MAPT expression and splicing is differentially regulated by brain region: relation to genotype and implication for tauopathies. *Human molecular genetics* **21**, 4094-4103 (2012).

10. Buée L, Bussière T, Buée-Scherrer V, Delacourte A, Hof PR. Tau protein isoforms, phosphorylation and role in neurodegenerative disorders. *Brain Research Reviews* **33**, 95-130 (2000).

11. Rockenstein EM, McConlogue L, Tan H, Power M, Masliah E, Mucke L. Levels and alternative splicing of amyloid β protein precursor (APP) transcripts in brains of APP transgenic mice and humans with Alzheimer's disease. *Journal of Biological Chemistry* **270**, 28257-28267 (1995).

12. Da Cruz S, Cleveland DW. Understanding the role of TDP-43 and FUS/TLS in ALS and beyond. *Current opinion in neurobiology* **21**, 904-919 (2011).

13. Ling SC, Polymenidou M, Cleveland DW. Converging mechanisms in ALS and FTD: disrupted RNA and protein homeostasis. *Neuron* **79**, 416-438 (2013).

14. Quesnel-Vallières M, Weatheritt RJ, Cordes SP, Blencowe BJ. Autism spectrum disorder: insights into convergent mechanisms from transcriptomics. *Nature Reviews Genetics*, 1 (2018).

15. De Rubeis S*, et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215 (2014).

16. Iossifov I*, et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-221 (2014).

17. Takata A, Ionita-Laza I, Gogos JA, Xu B, Karayiorgou M. De Novo Synonymous Mutations in Regulatory Elements Contribute to the Genetic Etiology of Autism and Schizophrenia. *Neuron* **89**, 940-947 (2016).

18. Clinton SM, Haroutunian V, Davis KL, Meador-Woodruff JH. Altered transcript expression of NMDA receptor-associated postsynaptic proteins in the thalamus of subjects with schizophrenia. *American Journal of Psychiatry* **160**, 1100-1109 (2003).

19. Xu B*, et al.* De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nature genetics* **44**, 1365-1369 (2012).

20. Takata A, Matsumoto N, Kato T. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nature communications* **8**, 14519 (2017).

21. Eom T*, et al.* NOVA-dependent regulation of cryptic NMD exons controls synaptic protein levels after seizure. *eLife* **2**, e00178 (2013).

22. Faustino NA, Cooper TA. Pre-mRNA splicing and human disease. *Genes & development* **17**, 419-437 (2003).

23. Scotti MM, Swanson MS. RNA mis-splicing in disease. *Nature reviews Genetics* **17**, 19-32 (2016).

24. Sibley CR, Blazquez L, Ule J. Lessons from non-canonical splicing. *Nature reviews Genetics* **17**, 407-421 (2016).

25. Li YI*, et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600-604 (2016).

26. Raj T*, et al.* Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nature genetics* **50**, 1584-1592 (2018).

27. Shen S*, et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E5593-5601 (2014).

28. Craig AM, Kang Y. Neurexin-neuroligin signaling in synapse development. *Current opinion in neurobiology* **17**, 43-52 (2007).

29. Loya CM, Van Vactor D, Fulga TA. Understanding neuronal connectivity through the post-transcriptional toolkit. *Genes & development* **24**, 625-635 (2010).

30. Traunmüller L, Gomez AM, Nguyen T-M, Scheiffele P. Control of neuronal synapse specification by a highly dedicated alternative splicing program. *Science* **352**, 982-986 (2016).

31. Ehrmann I*, et al.* The tissue-specific RNA binding protein T-STAR controls regional splicing patterns of neurexin pre-mRNAs in the brain. *PLoS genetics* **9**, e1003474 (2013).

32. Buniello A*, et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research* **47**, D1005-D1012 (2018).

33. Bray NJ, O'Donovan MC. The genetics of neuropsychiatric disorders. *Brain and neuroscience advances* **2**, (2019).

34. McCarroll SA, Feng G, Hyman SE. Genome-scale neurogenetics: methodology and meaning. *Nature neuroscience* **17**, 756 (2014).

35. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of computational biology* **11**, 377-394 (2004).

36. Pickrell JK*, et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768 (2010).

37. Zhao K, Lu Z-x, Park JW, Zhou Q, Xing Y. GLiMMPS: robust statistical model for regulatory variation of alternative splicing using RNA-seq data. *Genome biology* **14**, R74 (2013).

38. Shete S*, et al.* Genome-wide association study identifies five susceptibility loci for glioma. *Nature genetics* **41**, 899-904 (2009).

39. Wrensch M*, et al.* Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility. *Nature genetics* **41**, 905-908 (2009).

40. Cremona ML*, et al.* Flotillin-1 is essential for PKC-triggered endocytosis and membrane microdomain localization of DAT. *Nature neuroscience* **14**, 469-477 (2011).

41. Pizzo AB*, et al.* The membrane raft protein Flotillin-1 is essential in dopamine neurons for amphetamine-induced behavior in Drosophila. *Molecular psychiatry* **18**, 824-833 (2013).

42. O'Brien HE*, et al.* Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. *Genome biology* **19**, 194 (2018).

43. Zhong J*, et al.* Integration of GWAS and brain eQTL identifies FLOT1 as a risk gene for major depressive disorder. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, (2019).

44. Efthymiou AG, Goate AM. Late onset Alzheimer's disease genetics implicates microglial pathways in disease risk. *Molecular neurodegeneration* **12**, 43 (2017).

45. Karch CM, Ezerskiy LA, Bertelsen S, Alzheimer's Disease Genetics C, Goate AM. Alzheimer's Disease Risk Polymorphisms Regulate Gene Expression in the ZCWPW1 and the CELF1 Loci. *PloS one* **11**, e0148717 (2016).

46. Zhernakova A*, et al.* Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS genetics* **7**, e1002004 (2011).

47. Wu C, Pan W. Integrating eQTL data with GWAS summary statistics in pathway-based analysis with application to schizophrenia. *Genetic epidemiology* **42**, 303-316 (2018).

48. de Jong S*, et al.* Expression QTL analysis of top loci from GWAS meta-analysis highlights additional schizophrenia candidate genes. *European journal of human genetics : EJHG* **20**, 1004-1008 (2012).

49. Watanabe M, Hatakeyama S. TRIM proteins and diseases. *Journal of biochemistry* **161**, 135-144 (2017).

50. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology* **33**, 831 (2015).

51. Ballatore C, Lee VM, Trojanowski JQ. Tau-mediated neurodegeneration in Alzheimer's disease and related disorders. *Nature reviews Neuroscience* **8**, 663-672 (2007).

52. Soto C, Pritzkow S. Protein misfolding, aggregation, and conformational strains in neurodegenerative diseases. *Nature neuroscience* **21**, 1332-1340 (2018).

53. Irwin DJ, Lee VM, Trojanowski JQ. Parkinson's disease dementia: convergence of alpha-synuclein, tau and amyloid-beta pathologies. *Nature reviews Neuroscience* **14**, 626-636 (2013).

54. Yeo GW, Coufal NG, Liang TY, Peng GE, Fu X-D, Gage FH. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature structural & molecular biology* **16**, 130 (2009).

55. Park E, Pan Z, Zhang Z, Lin L, Xing Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *American journal of human genetics* **102**, 11-26 (2018).

56. Xiong HY*, et al.* The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).

57. Zhang Z*, et al.* Deep-learning augmented RNA-seq analysis of transcript splicing. *Nature Methods* **16**, 307 (2019).

58. Park E, Pan Z, Zhang Z, Lin L, Xing Y. The expanding landscape of alternative splicing variation in human populations. *The American Journal of Human Genetics* **102**, 11-26 (2018).

59. Van Nostrand EL*, et al.* A large-scale binding and functional map of human RNA binding proteins. *bioRxiv*, 179648 (2018).

60. Gehman LT*, et al.* The splicing regulator Rbfox2 is required for both cerebellar development and mature motor function. *Genes & development* **26**, 445-460 (2012).

61. Nakahata S, Kawamoto S. Tissue-dependent isoforms of mammalian Fox-1 homologs are associated with tissue-specific splicing activities. *Nucleic acids research* **33**, 2078-2089 (2005).

62. Underwood JG, Boutz PL, Dougherty JD, Stoilov P, Black DL. Homologues of the Caenorhabditis elegans Fox-1 protein are neuronal splicing regulators in mammals. *Molecular and cellular biology* **25**, 10005-10016 (2005).

63. Hariri AR. The Emerging Importance of the Cerebellum in Broad Risk for Psychopathology. *Neuron* **102**, 17-20 (2019).

64. Keller D, Erö C, Markram H. Cell densities in the mouse brain: a systematic review. *Frontiers in neuroanatomy* **12**, (2018).

65. Leek JT*, et al.* sva: Surrogate variable analysis. *R package version* **3**, 10.18129 (2017).

66. Stein S, Lu Z-x, Bahrami-Samani E, Park JW, Xing Y. Discover hidden splicing variations by mapping personal transcriptomes to personal genomes. *Nucleic acids research* **43**, 10612-10622 (2015).

2.6 Figures

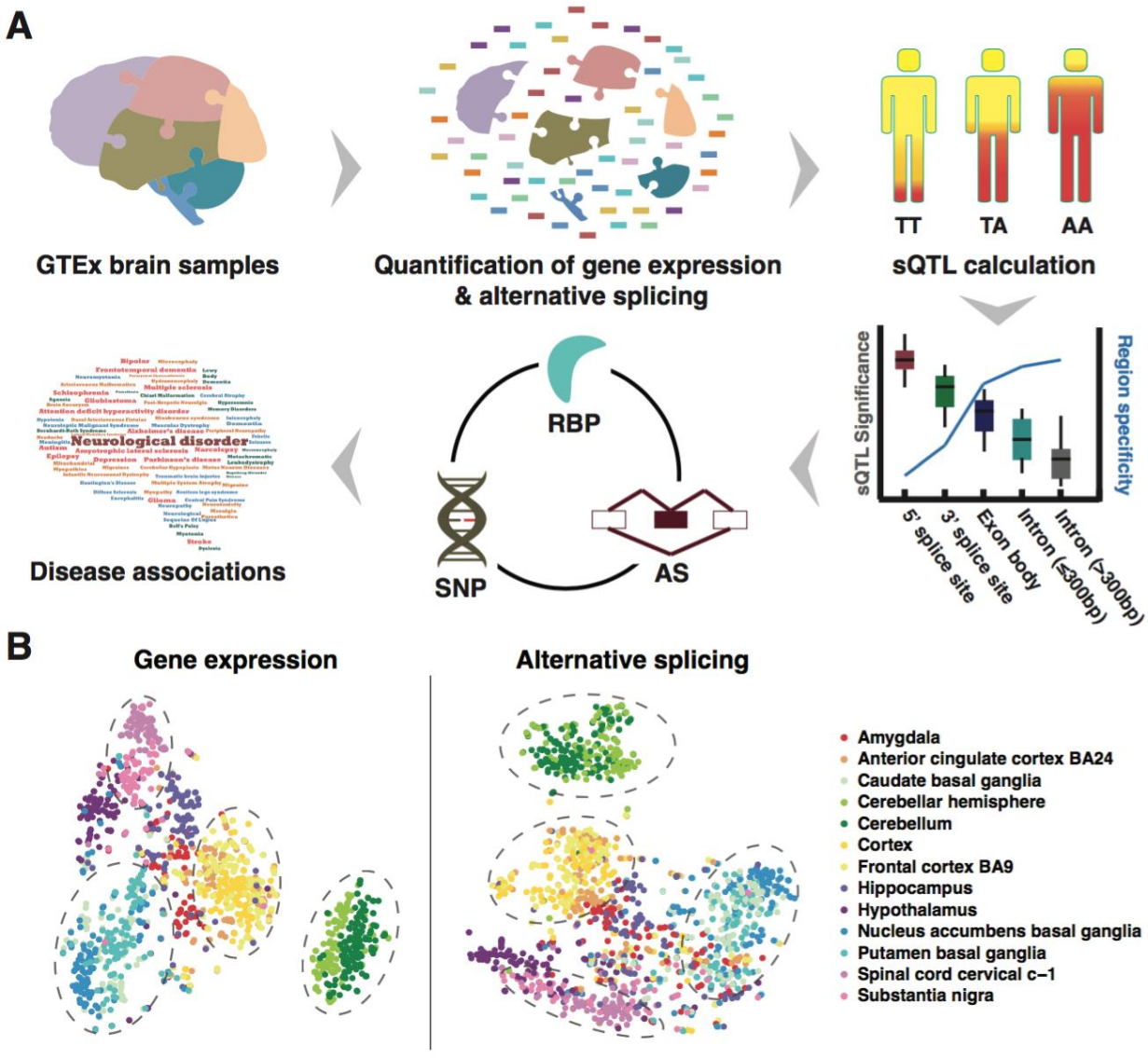


Figure 2.1 Overview of the study and data

(**A**) **1** RNA-seq data derived from human brain tissue samples were downloaded from the GTEx Portal. **2** Data were processed to quantify gene expression (in TPM) and alternative splicing (in PSI) across 13 brain regions. **3** Splicing QTLs (sQTLs) were identified based on the alternative splicing and genotype information. **4** The sQTL data were used to study the relationship among the positional distribution of sQTLs, significance of sQTLs and brain-region specificity. **5** RNA

binding protein (RBP) expression was incorporated with alternative splicing (AS) and single-nucleotide polymorphism (SNP) data to identify potential regulators. **6** Finally, data were used to prioritize causal variants involved in 14 neurological disorders.

(**B**) t-SNE clustering of all brain samples based on gene expression (left) and alternative splicing (right). Samples are color-coded by brain region. Dashed ovals encircle samples from physically proximate brain regions.

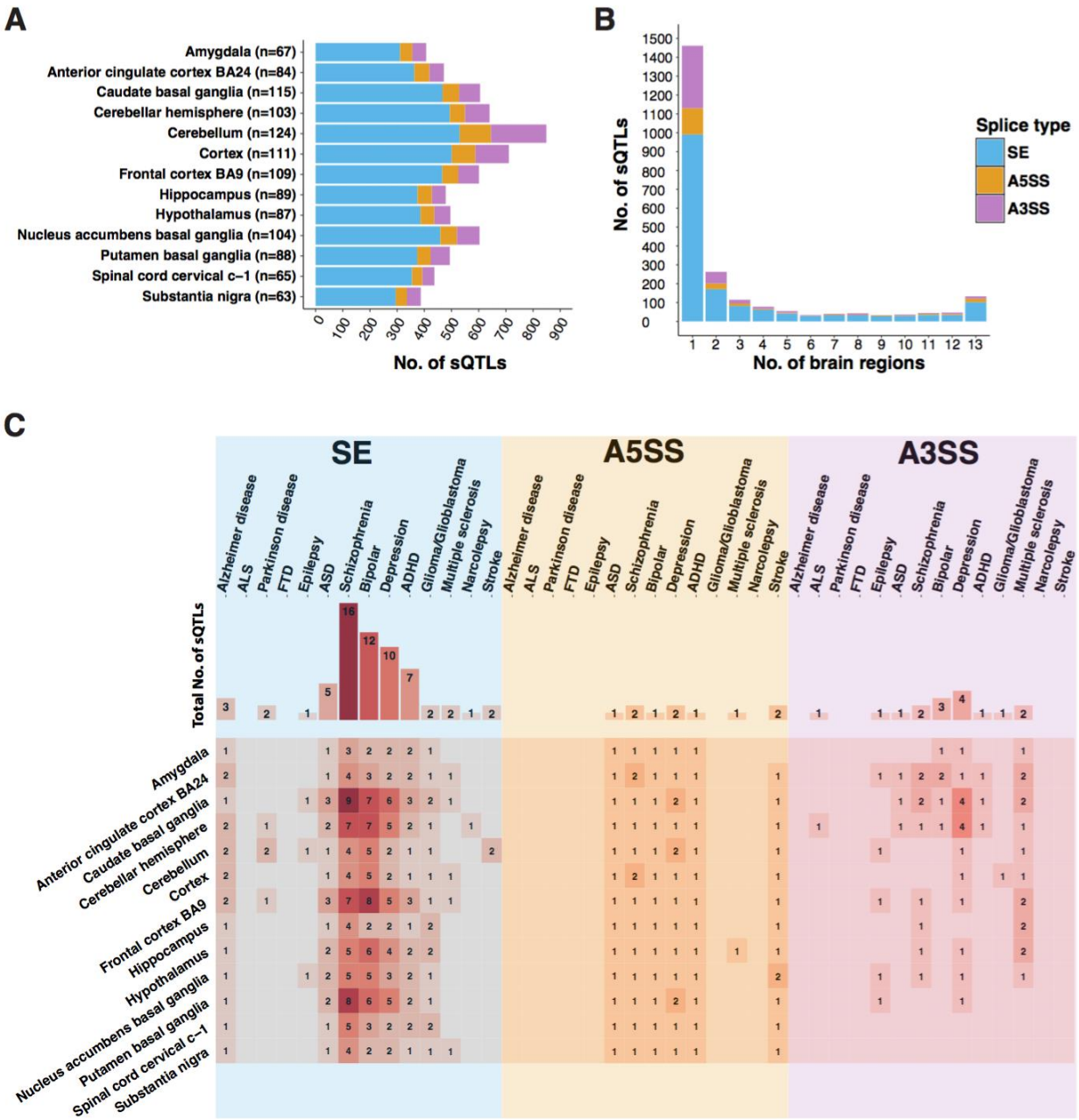


Figure 2.2 Identification of sQTLs in each brain region

(**A**) Stacked bar plot showing the number of sQTLs (including SE, A5SS, and A3SS events) identified in each brain region, with the number in parentheses referring to the number of tissue 747 samples with available genotype information.

(**B**) Stacked bar plot showing the histogram for the number of brain regions where sQTLs (including SE, A5SS, and A3SS events) were determined to be significant.

(**C**) Heat map showing the number of disease sQTL (including SE, A5SS and A3SS events) for each neurological disorder in each brain region. Each row represents one brain region. Each column represents one neurological disorder. Bar plot above heat map shows the total number of unique sQTL exons related to each neurological disorder.

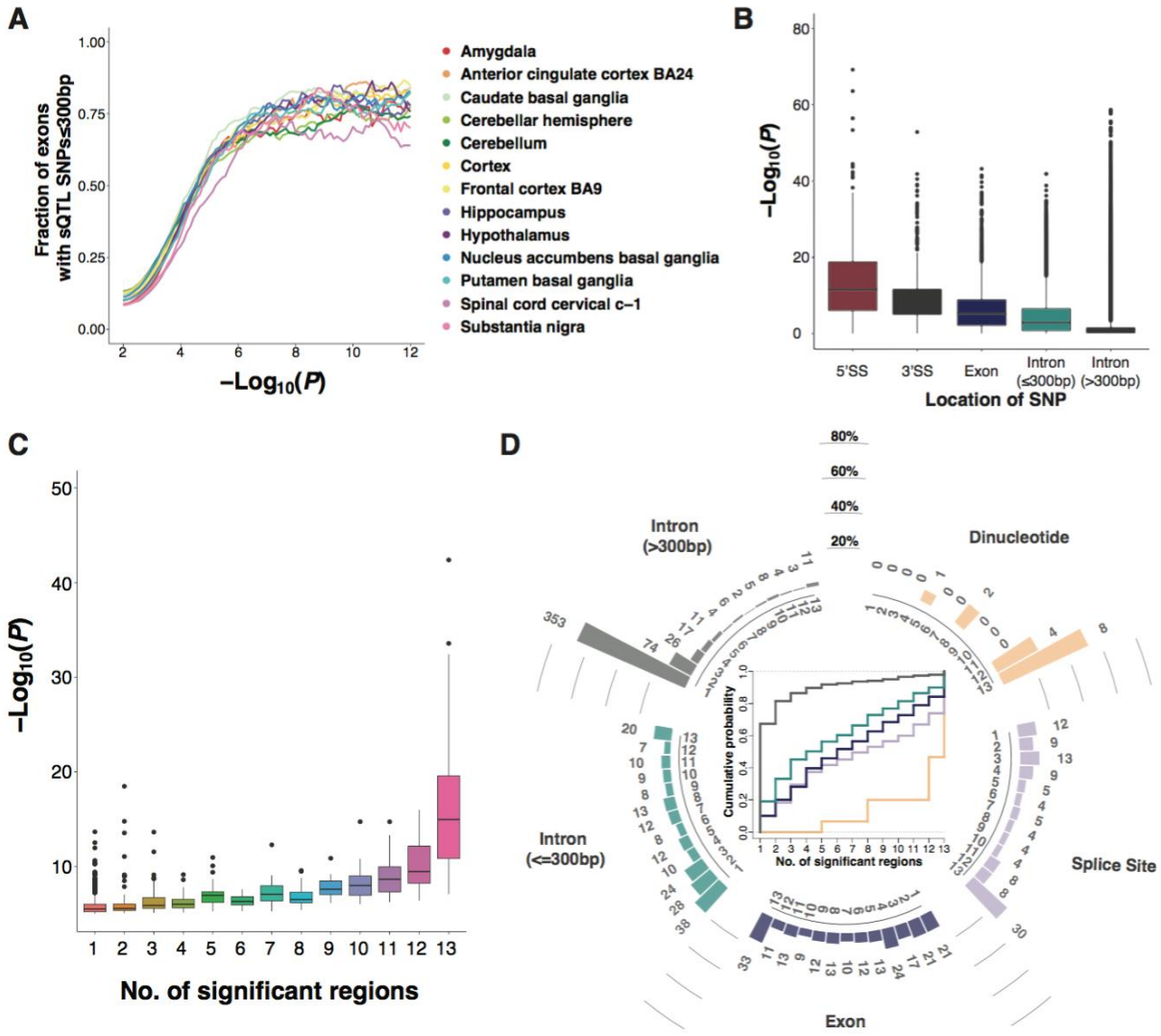


Figure 2.3 Relationship among positional distribution of sQTLs, significance of sQTLs and brain-region specificity

(**A**) Fraction of sQTL exons with any significant SNP within 300 bp of the splice site, as a function of the significance (-$\log_{10}$(*P*-value)) cutoff for significant sQTL exons. Each curve represents the result in one brain region.

(**B**) Box plot showing significance (-$\log_{10}$(*P*-value)) of all SNPs within 200 kb of all significant ($P \leq 10^{-5}$) sQTL exons. Results are grouped based on SNP position relative to the splice site.

(**C**) Box plot showing the relationship between sQTL significance ($-\log_{10}$(*P*-value)) and the number of brain regions where the sQTL is significant. Each dot represents one sQTL exon. An sQTL exon having at least one significantly associated SNP in a brain region is considered significant in that brain region. The sQTL significance is calculated by considering all brain regions where the sQTL is significant, and taking the median of the smallest *P* values.
(**D**) Bar plots (outside) and cumulative distribution function (CDF) (inside) showing the relationship between brain-region specificity and sQTL positional distribution. The sQTL exons are grouped based on the location of significant SNPs. For each group, the bar height is proportional to the percentage of sQTL exons that are significant in a given number of brain regions. Numbers above bars represent absolute numbers of sQTL exons.

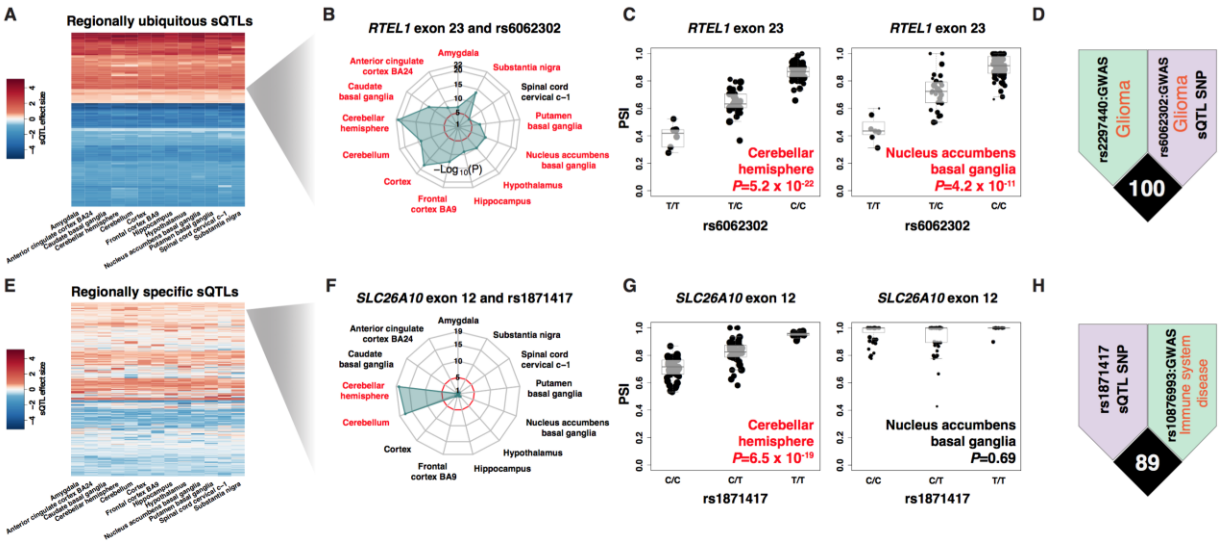


Figure 2.4 Two groups of sQTLs with different types of brain-region specificity

(**A**) Heat map showing correlation coefficients of regionally ubiquitous sQTLs. Each row represents one sQTL. Each column represents one brain region.

(**B**) Radar plot showing significance (-log$_{10}$(*P*-value)) of the sQTL (*RTEL* exon 23 and rs6062302) in each brain region. Red circle indicates the significance cutoff ($P = 10^{-5}$). Significant brain regions are highlighted in red text.

(**C**) Box plot showing the significant association of SNP rs6062302 with PSI value (exon inclusion level) of *RTEL* exon 23 in two brain regions: cerebellar hemisphere and nucleus accumbans basal ganglia.

(**D**) LD plot showing a GWAS variant (rs2297440, glioma; *green*) in high LD with the sQTL SNP (rs6062302; *purple*). SNP rs6062302 itself is a GWAS variant related to glioma.

(**E**) Heat map showing correlation coefficients of regionally specific sQTLs.

(**F**) Radar plot showing significance (-log$_{10}$(*P*-value)) of the sQTL (*SLC26A10* exon 12 and rs1871417) in each brain region.

(**G**) Box plot showing the association of SNP rs1871417 with PSI value (exon inclusion level) of *SLC26A10* exon 12 in two brain regions: cerebellar hemisphere (significant) and nucleus

accumbans basal ganglia (not significant).

(**H**) LD plot showing a GWAS variant (rs10876993, immune system disease; *green*) in high LD with the sQTL SNP (rs1871417; *purple*).

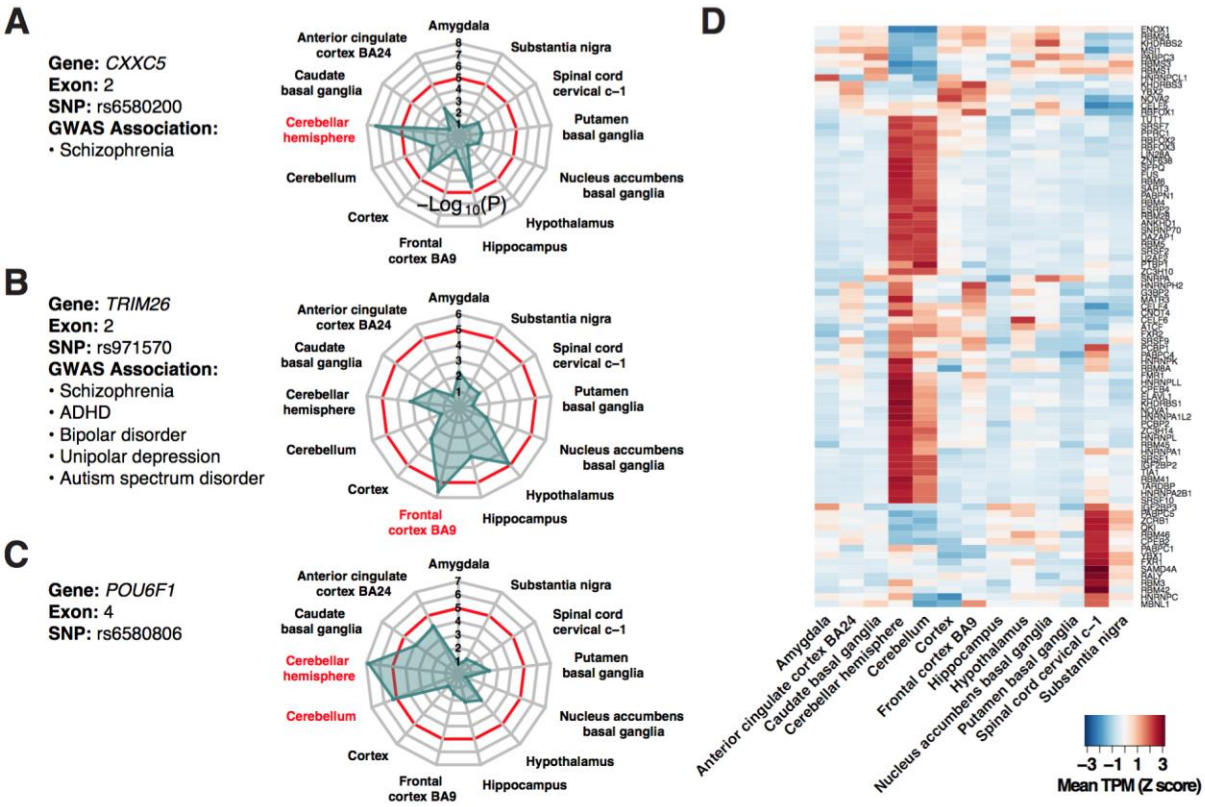


Figure 2.5 Brain region-specific sQTLs and RBP expression

(**A**), (**B**), (**C**) Radar plots showing significance (-$\log_{10}$(*P*-value)) of three brain region-specific sQTLs (**A**: rs6580200 and *CXXC5* exon 3; **B**: rs971570 and *TRIM26* exon 2; **C**: rs6580806 and *POU6F1* exon 4) in each brain region.

(**D**) Heat map showing gene expression (z-score–transformed TPM) of RBPs in the DeepBind model across all 13 brain regions. Each row represents one RBP. Each column represents one brain region. In addition to all 102 *Homo sapiens* RBPs from the DeepBind model, the analysis includes RBFOX2, RBFOX3, NOVA1, and NOVA2.

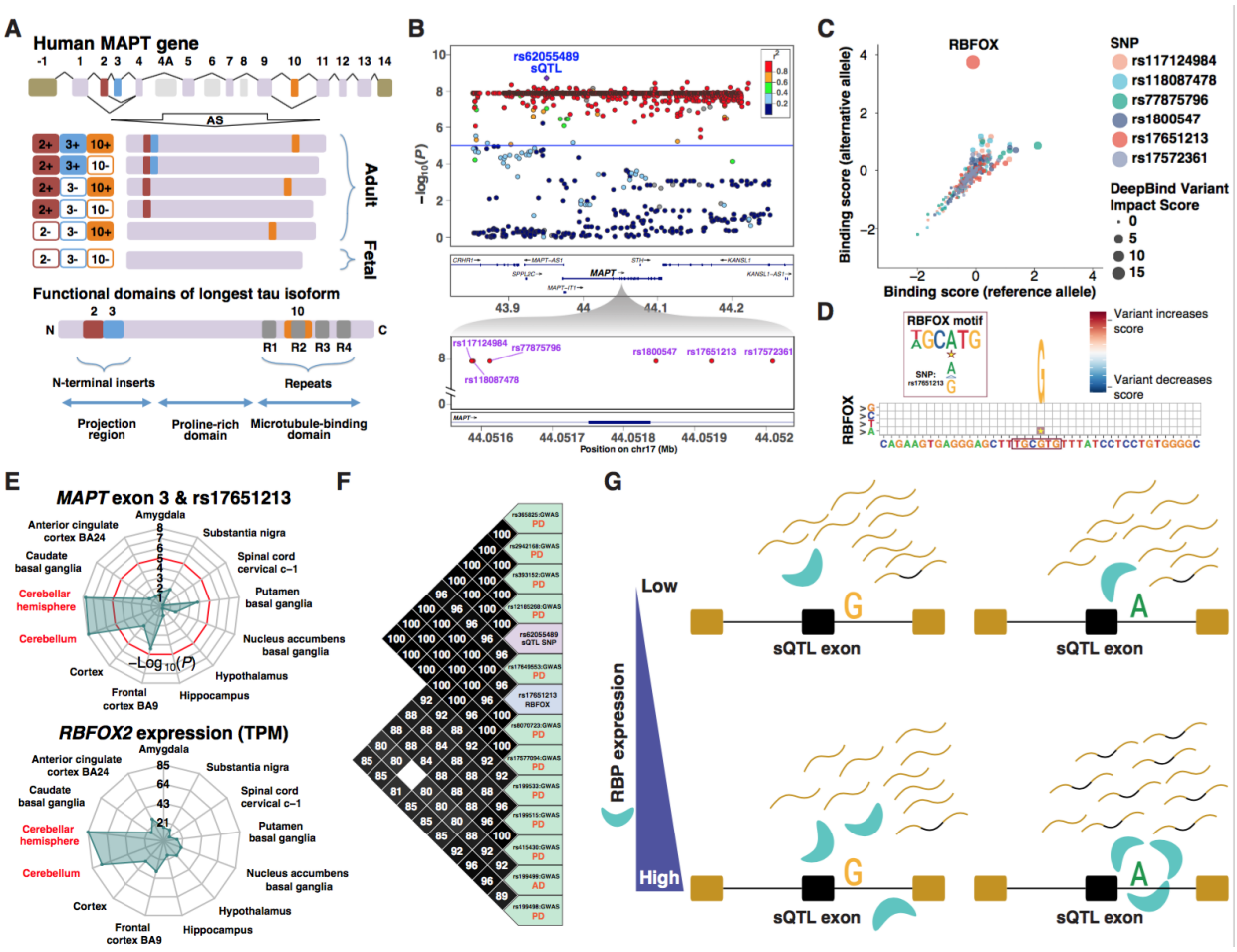


Figure 2.6 Combining deep learning, regional expression and sQTL information to prioritize causal *cis*-acting sQTL variants and their trans-acting regulators

(**A**) (Upper panel) Gene structure and six human tau isoforms of 16-exon human *MAPT*. Exon -1 (in the promoter region) and exon 14 are transcribed but not translated. Primary transcript of human tau contains 13 exons (exons 4A, 6 and 8 are not transcribed). Among the 13 exons, only exons 2, 3, and 10 are alternatively spliced, generating six mRNA isoforms. (Lower panel) Functional domains of longest tau isoform (including exons 2, 3 and 10).

(**B**) Local plot showing distribution of all SNPs within 200 kb (upper) and all significant SNPs within 300 bp (lower) of *MAPT* exon 3 in cerebellum. Window is extended 30 kb (upper) and 30 bp (lower) to include boundary information. Each circle represents one SNP, color-coded

according to its LD with the sQTL SNP. Y-axis shows the significance of association ($-\log_{10}$(*P*-value)) between each SNP and the sQTL exon. Horizontal line represents the significance cutoff ($P = 10^{-5}$). Genes in the UCSC Genome Browser (https://genome.ucsc.edu/) are shown in panels below plots.

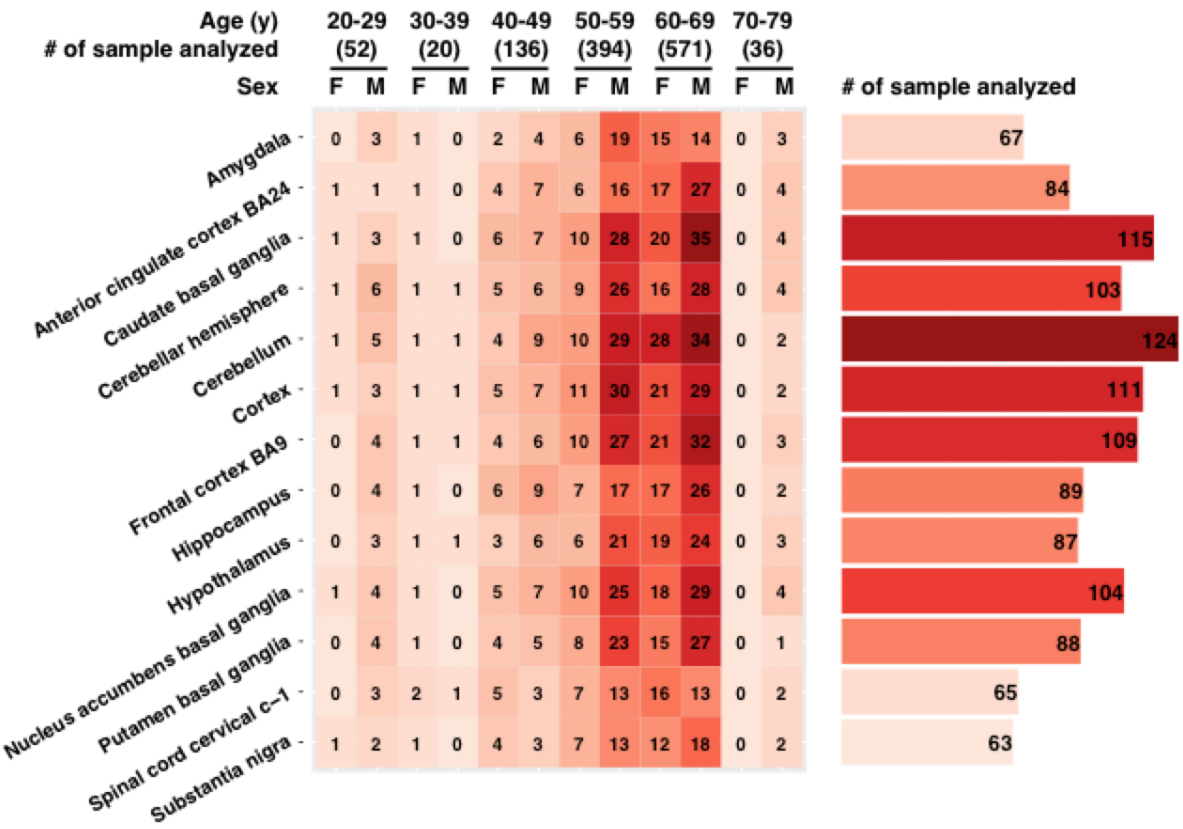
(**C**) Bubble plot showing effects of six SNPs on RBP binding. Axes show RBP binding scores of sequences with reference allele (X-axis) or alternative allele (Y-axis) for each RBP. Bubble size is proportional to the DeepBind score, which represents the SNP-mediated change in RBP binding.

(**D**) DeepBind mutation map for rs17651213 with the RBFOX binding site. Star indicates the mutation position.

(**E**) Radar plot showing significance ($-\log_{10}$(*P*-value)) of the sQTL (*MAPT* exon 3 and rs17651213) and mean gene expression of RBFOX2 in each brain region.

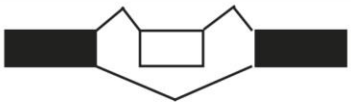
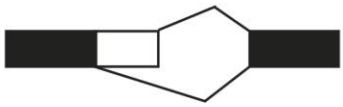
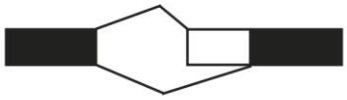
(**F**) LD plot showing SNP rs17651213 (*blue*) in high LD with the sQTL SNP (rs62055489; *purple*), 11 PD GWAS variants (*green*) and one AD GWAS variant (*green*).

(**G**) Diagram showing one proposed mechanism for the observed effects. A SNP may alter the RNA binding site of a *trans*-acting splicing regulator, but the molecular impact depends on the regulator concentration (e.g. a SNP may only alter splicing in tissues or cell types with high regulator expression).



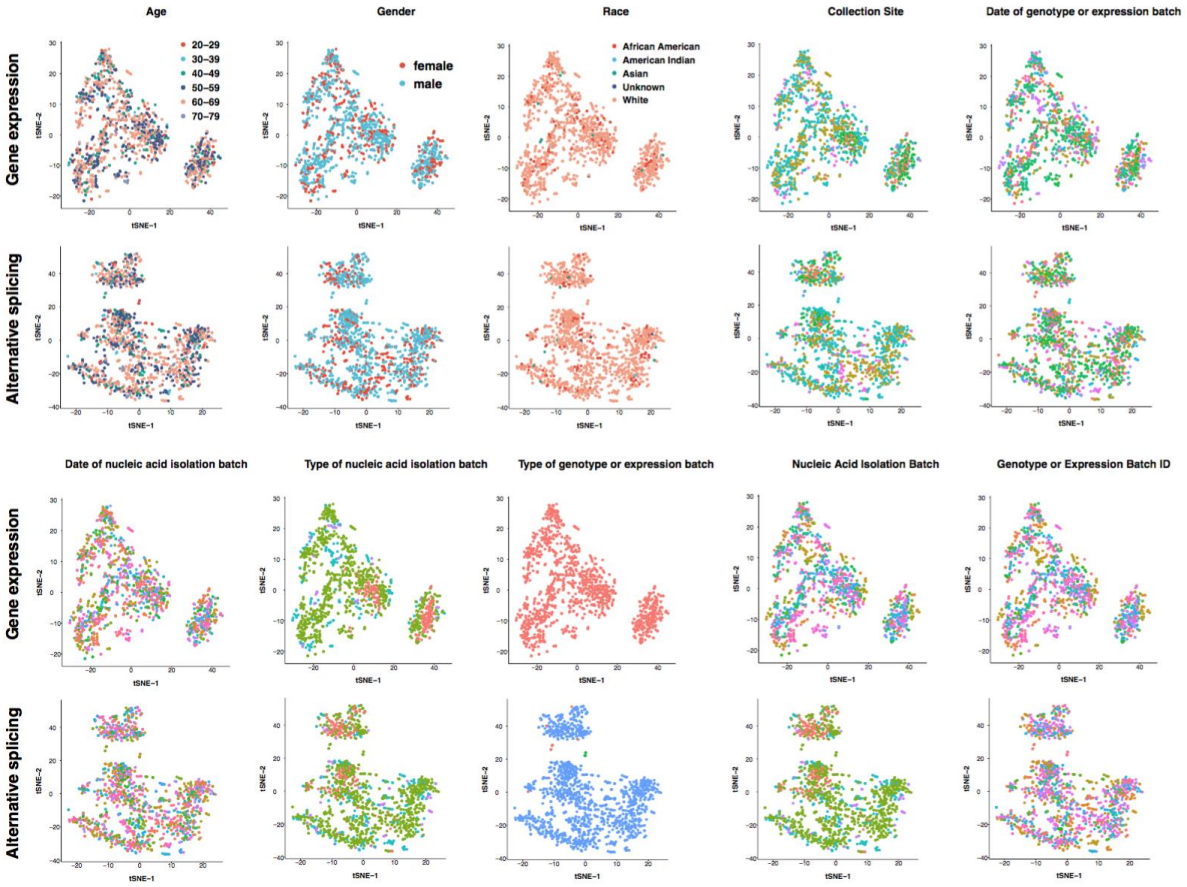
Supplementary Figure 2.7 Distribution of GTEx brain tissue samples across age, sex and brain region

Heat map showing the number of samples in each combination of age, sex and brain region. Bar plots on right show number of samples in each brain region.

Event Types		Events after filtering
Skipped exon		10,665
Alternative 5’ splice site		1,443
Alternative 3’ splice site		2,434

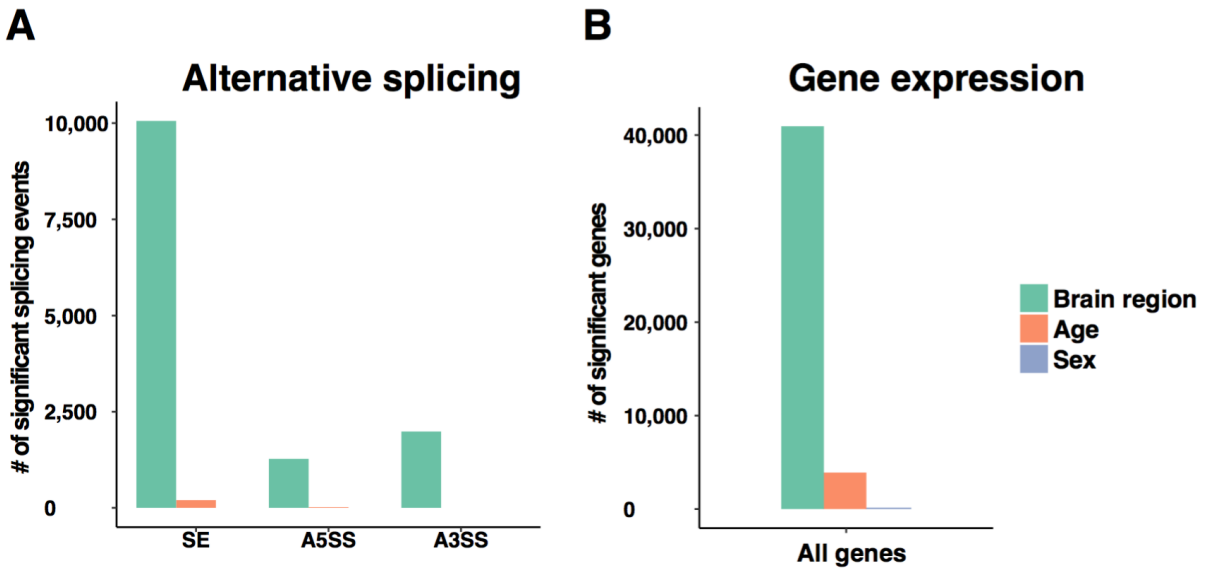
Supplementary Figure 2.8 Numbers of splicing events for three types of alternative splicing

Numbers represent results after filtering. Patterns of the three types of alternative splicing are also shown.



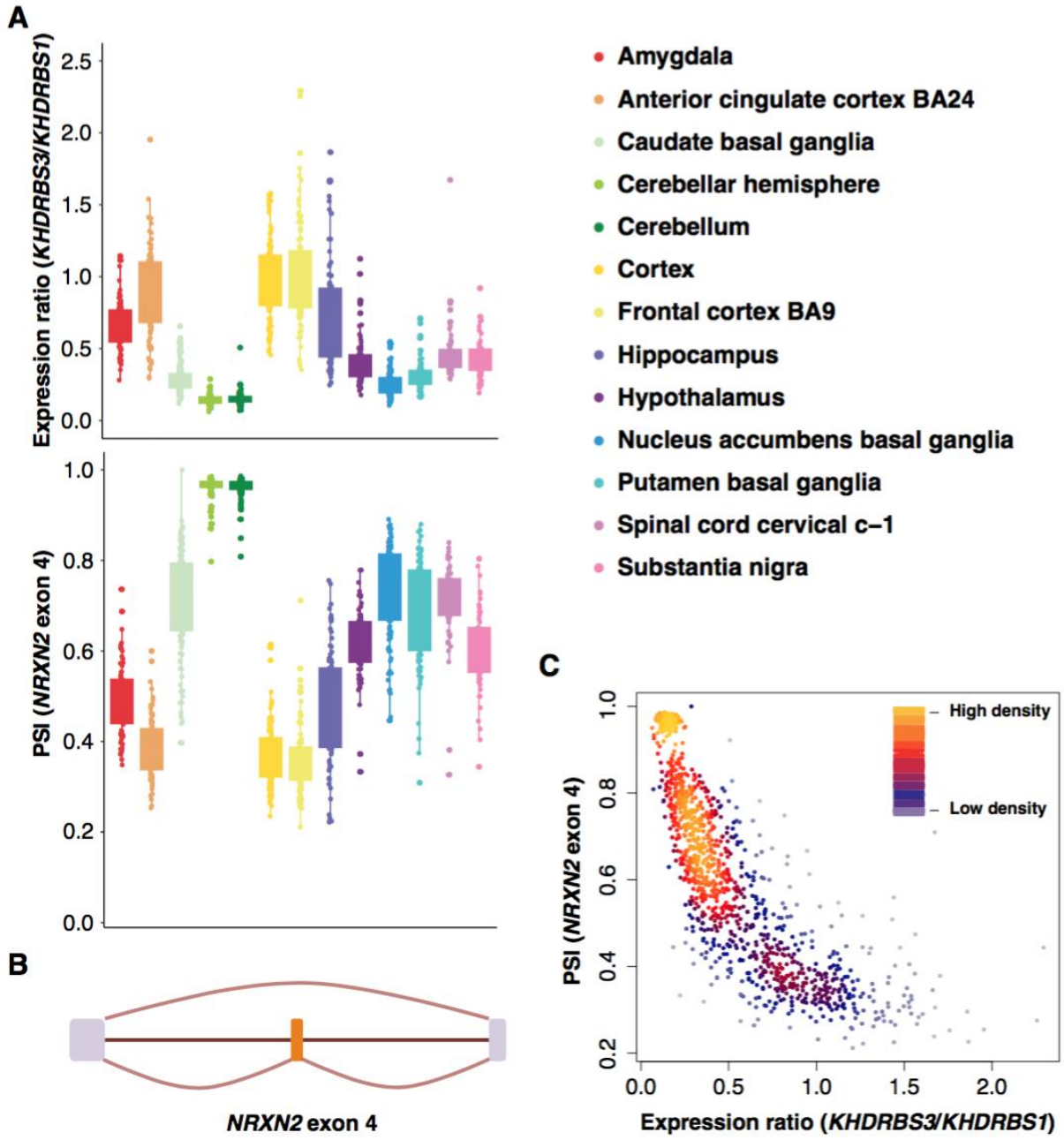
Supplementary Figure 2.9 t-SNE clustering of brain samples

t-SNE clustering of all brain tissue samples based on gene expression (upper) and alternative splicing (lower). Samples are color-coded by age, sex, race, and selected potential batch effects.



Supplementary Figure 2.10 Dependence of splicing and gene expression events on age, sex and brain region factors

Bar plots showing numbers of (**A**) splicing events (including SE, A5SS and A3SS events) and (**B**) expressed genes, with significant differences across age groups, sexes or brain regions.

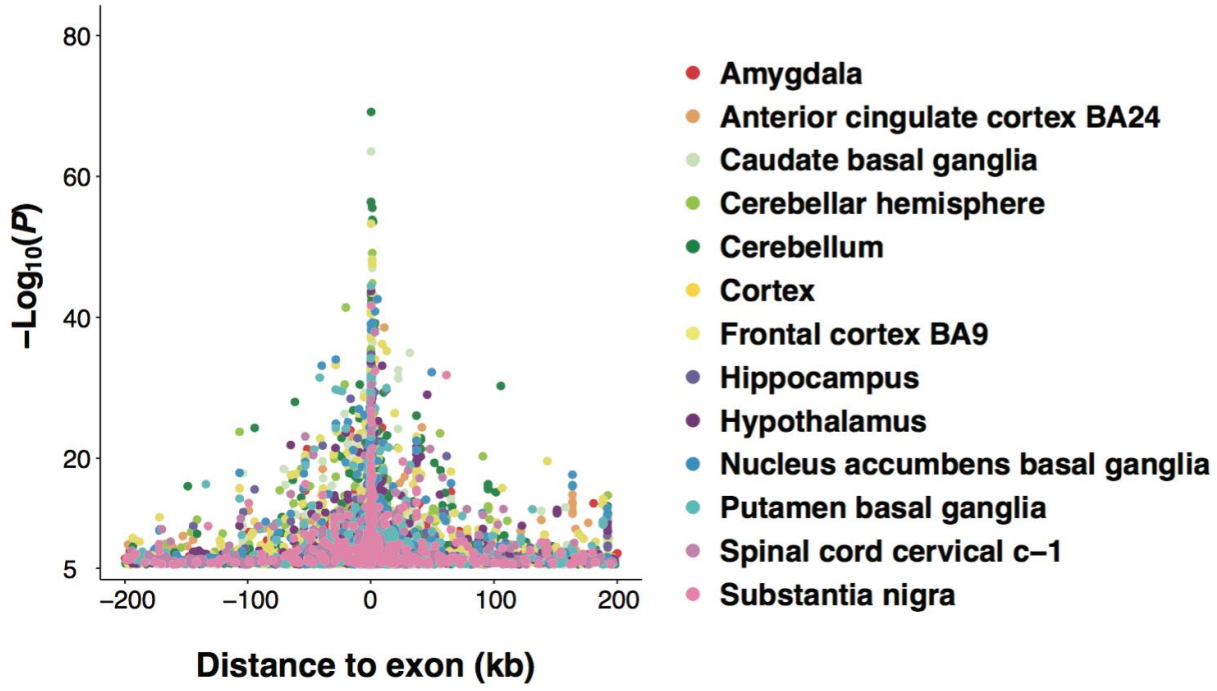


Supplementary Figure 2.11 Regulation of brain region-specific splicing of *NRXN2* exon 4 by KHDRBS3 and KHDRBS1

(**A**) (Top panel) Box plot showing the gene expression ratio between KHDRBS3 and KHDRBS1 in each brain region. (Bottom panel) Box plot showing the PSI value (exon inclusion level) of *NRXN2* exon 4 in each brain region.

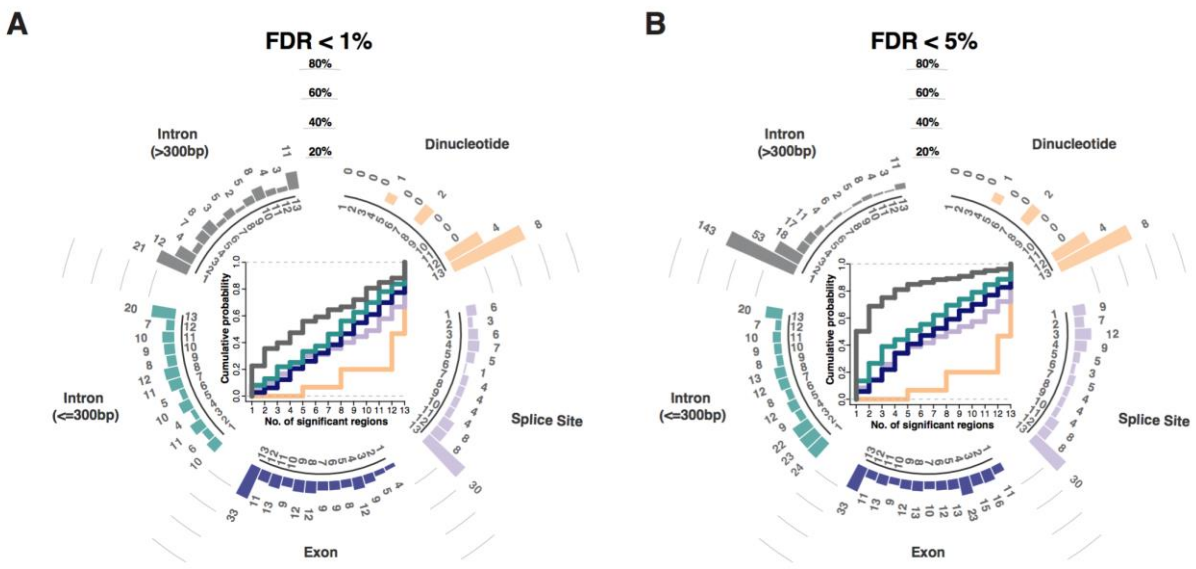
(**B**) Structure of the splicing event.

(**C**) Scatter plot showing the negative correlation between the gene expression ratio (*KHDRBS3*:*KHDRBS1*) and the PSI value (exon inclusion level) of *NRXN2* exon 4.



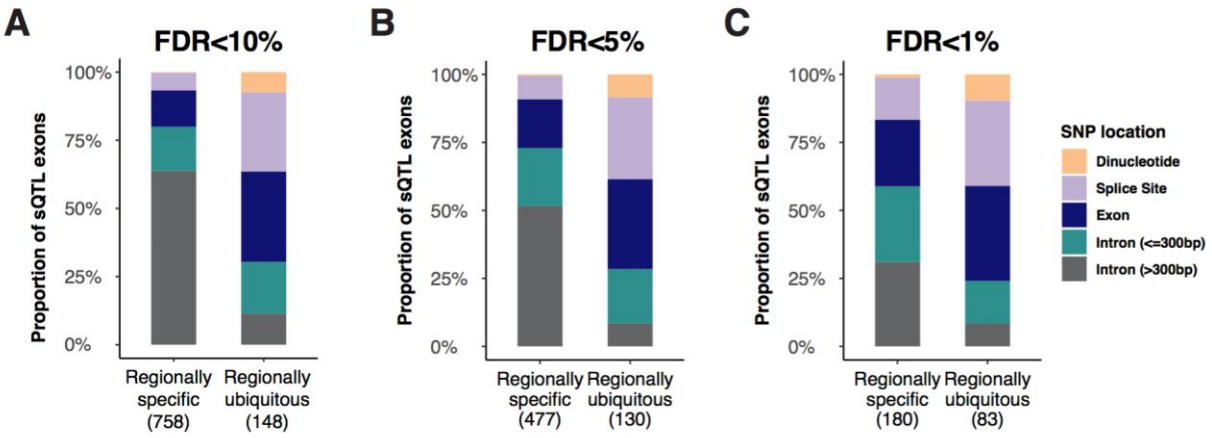
Supplementary Figure 2.12 Positional distribution and significance of sQTL SNPs

Significance (-$\log_{10}$(*P*-value)) of sQTL SNPs in all 13 brain regions as a function of the distance of the sQTL SNP to the nearest alternative splicing event (sQTL exon). Each dot represents one significant sQTL SNP in one brain region. Dot color corresponds to the brain region where the sQTL exon and SNP show a significant association.



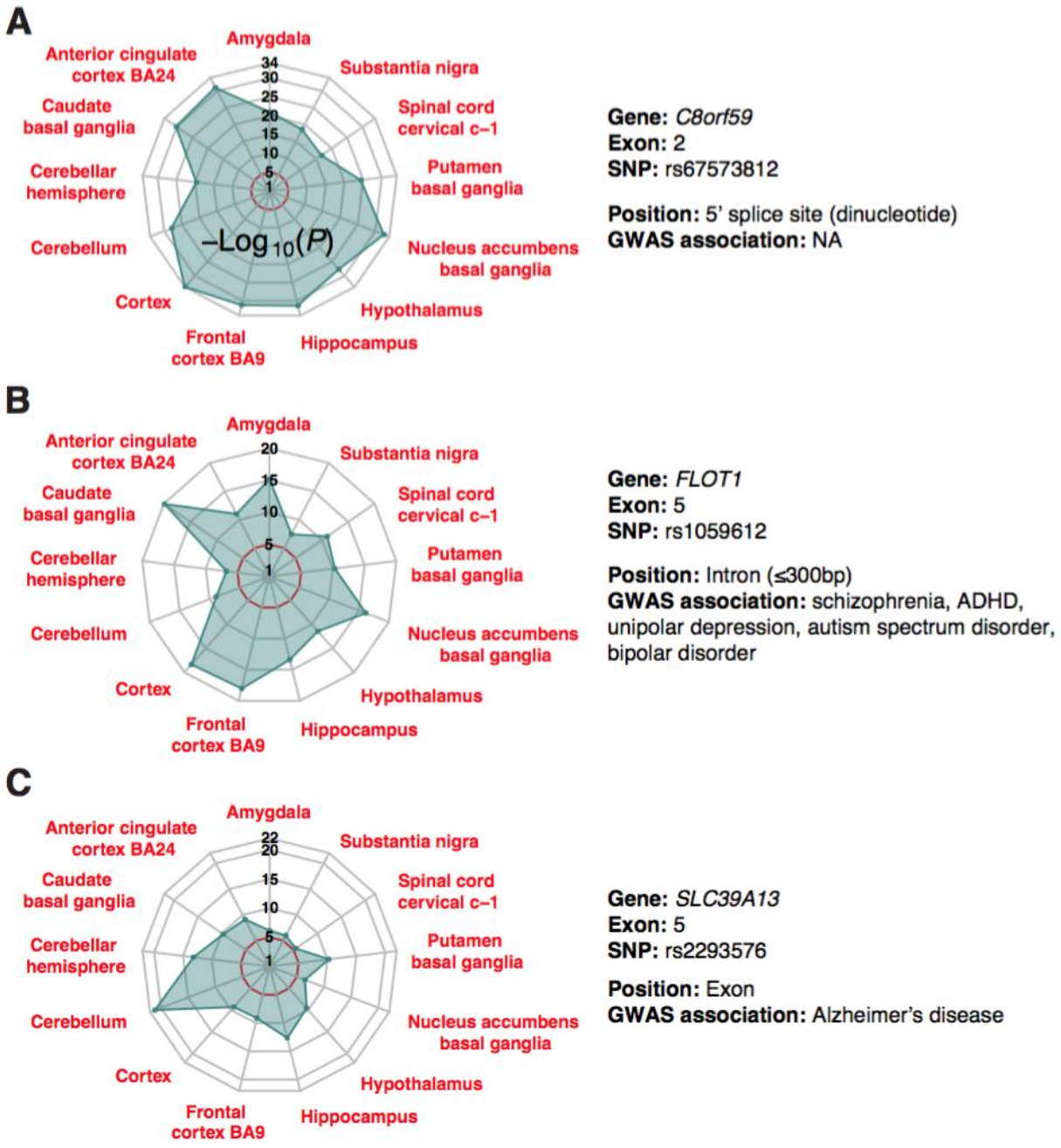
Supplementary Figure 2.13 Relationship between brain region specificity and positional distribution of sQTLs

Bar plots (outside) and cumulative distribution functions (CDFs) (inside) showing the relationship between SNP position and brain region specificity of sQTLs (SE events). The sQTLs are grouped based on the position of their significant SNP (e.g., dinucleotide, splice site, etc.). For each group, the bar plot shows the histogram of the percentage of sQTLs that are significant in a given number of brain regions. Each bar is labeled above with the number of significant sQTLs in the given number of brain regions. Significant sQTLs were defined with different significance cutoffs (left: FDR < 1%; right: FDR < 5%).



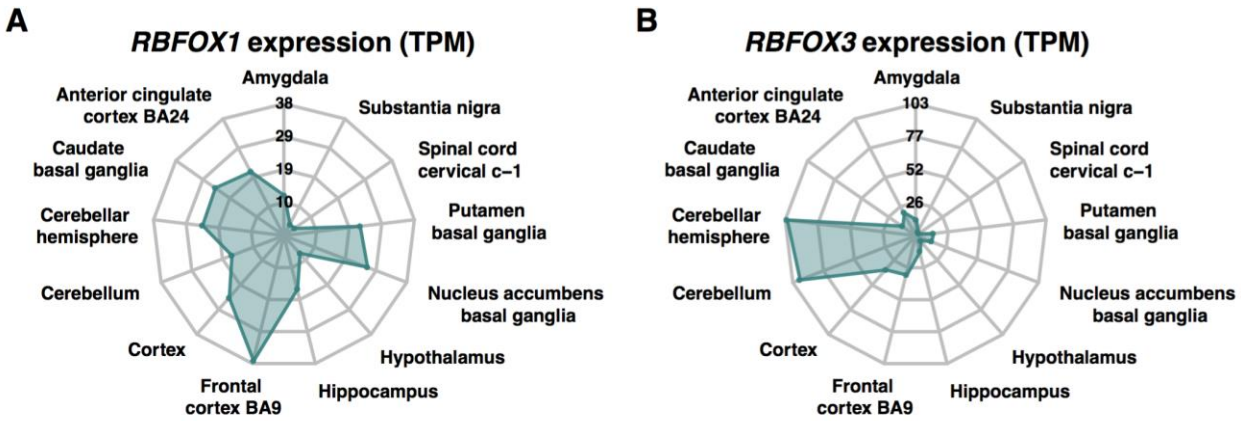
Supplementary Figure 2.14 Positional distribution of regionally specific and ubiquitous sQTLs

Stacked bar plots showing proportions of sQTLs (grouped based on significant SNP positions) for regionally specific and regionally ubiquitous sQTLs. Significant sQTLs are defined with different significance cutoffs (**A**: FDR < 10%; **B**: FDR < 5%; **C**: FDR < 1%).



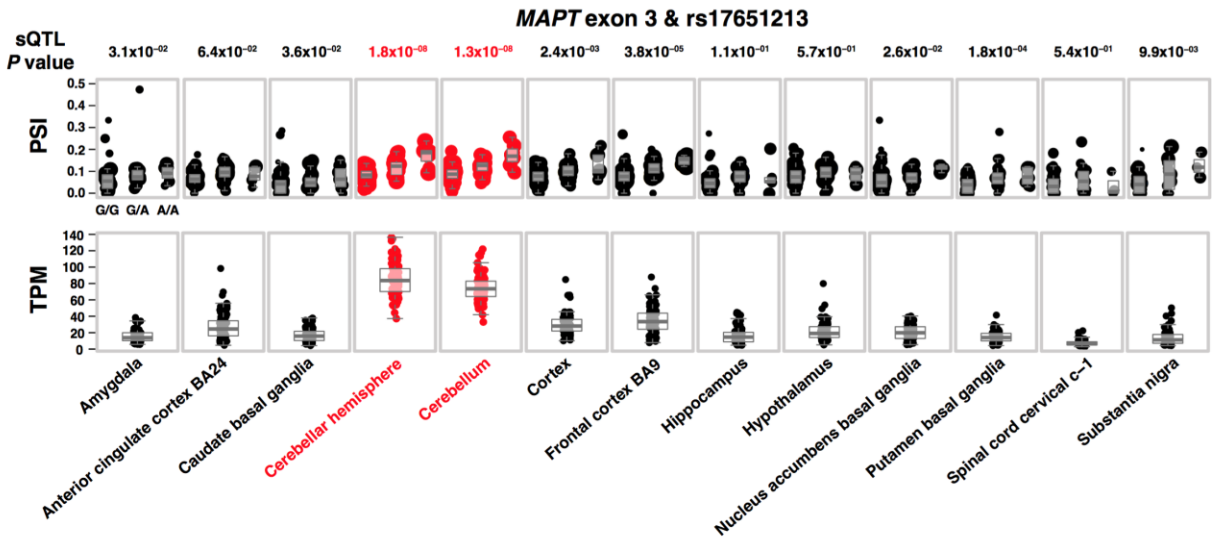
Supplementary Figure 2.15 Ubiquitous *cis*-acting sQTLs significant in all brain regions

(**A**), (**B**), (**C**) Radar plots showing significance (-$\log_{10}$(*P*-value)) of three ubiquitous sQTLs (**A**: rs67573812 and *C8orf59* exon 2; **B**: rs1059612 and *FLOT1* exon 5; **C**: rs2293576 and *SLC39A13* exon 5) in each brain region.



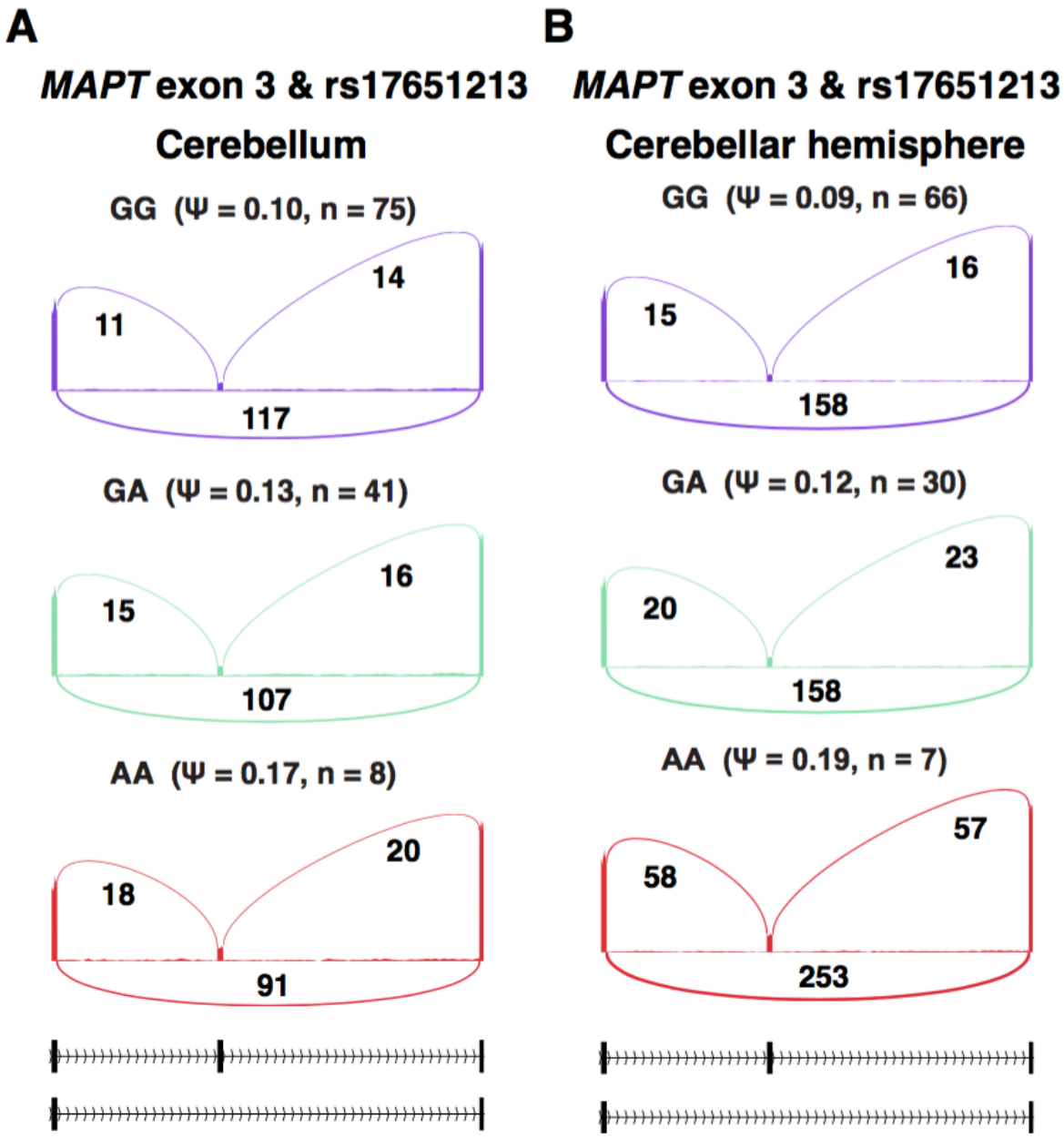
Supplementary Figure 2.16 Gene expression level of RBFOX1 and RBFOX3 across 13 brain regions

Radar plot showing mean gene expression of RBFOX1 (**A**) and RBFOX3 (**B**) in each brain region.



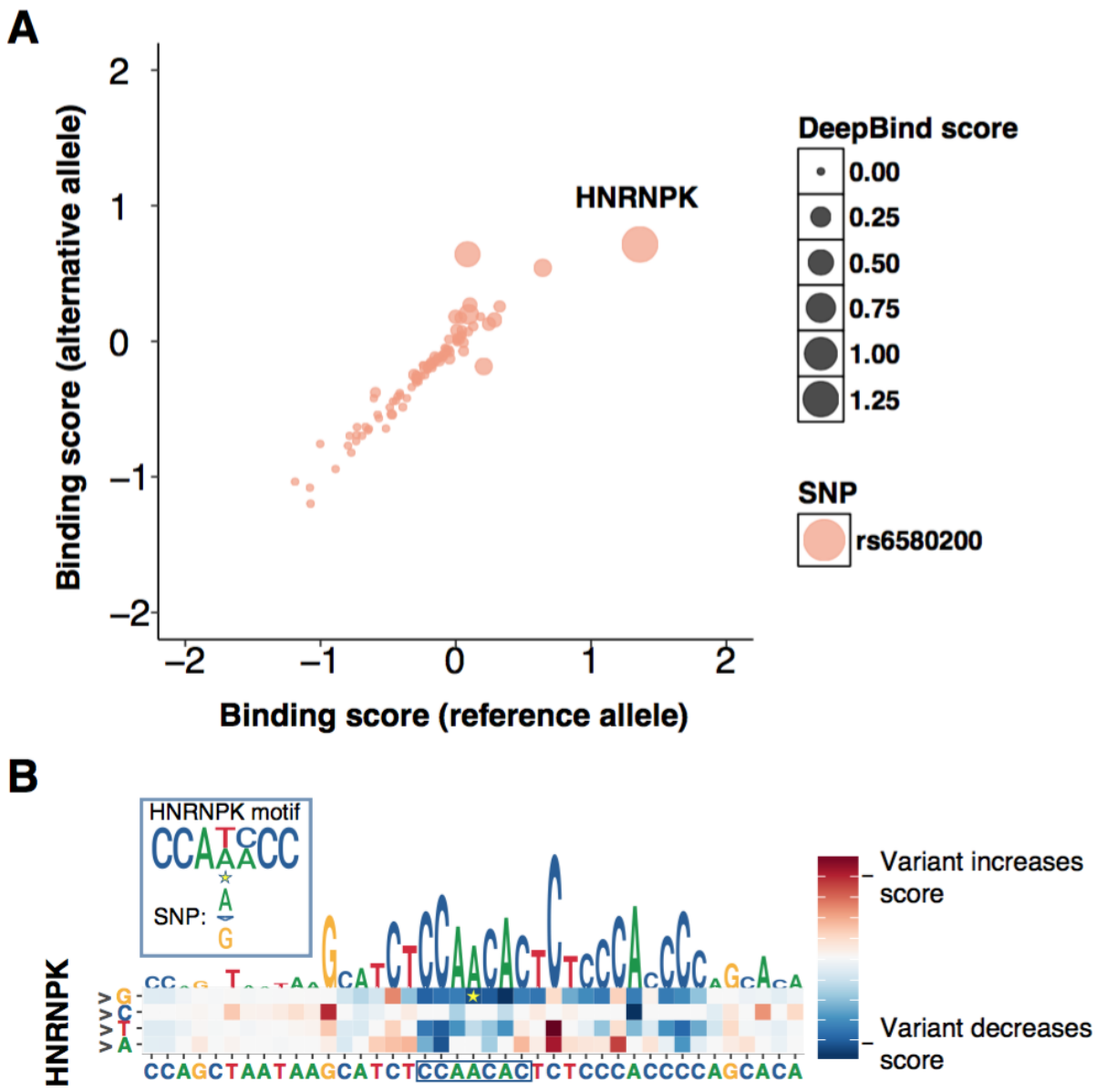
Supplementary Figure 2.17 RBFOX2 regulation of the brain region-specific association between *MAPT* exon 3 and rs17651213

(Top panel) Box plot showing the association between *MAPT* exon 3 and rs17651213 in each brain region. Brain regions where the association is significant are shown in red. Reported PSI value is the original PSI value before confounding factor correction. (Bottom panel) Box plot showing gene expression of *RBFOX2* in each brain region. Each dot represents one tissue sample.



Supplementary Figure 2.18 Sashimi plot of *MAPT* exon 3 and rs17651213 in two brain regions: cerebellum and cerebellar hemisphere

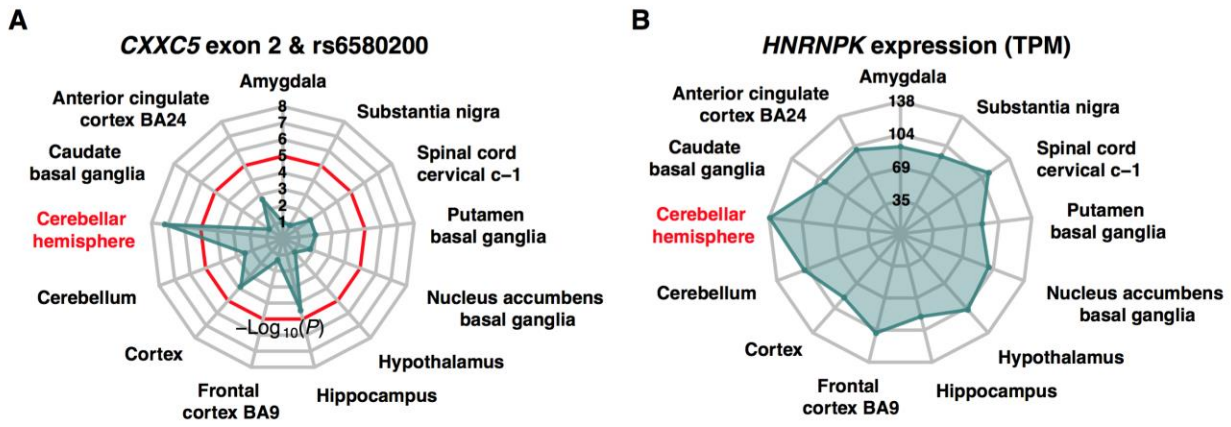
Sashimi plots of the sQTL (*MAPT* exon 3 and rs17651213) in two significant brain regions: (**A**) cerebellum and (**B**) cerebellar hemisphere.



Supplementary Figure 2.19 Effect of rs6580200 on RBP binding, as predicted by DeepBind

(**A**) Bubble plot showing effects of rs6580200 on the binding of different RBPs.

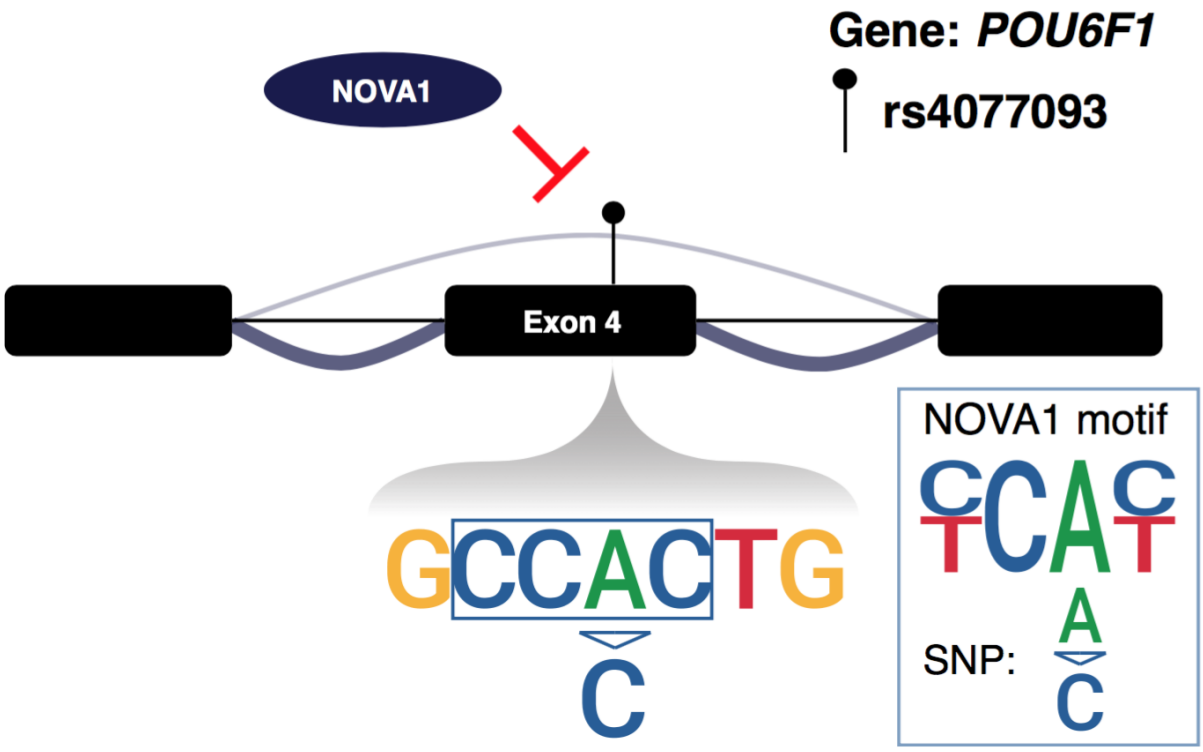
(**B**) DeepBind mutation map showing SNP rs6580200 with the HNRNPK binding site. Star indicates the mutation position.



Supplementary Figure 2.20 Significance of the sQTL (*CXXC5* exon 2 and rs6580200) agrees with gene expression of HNRNPK

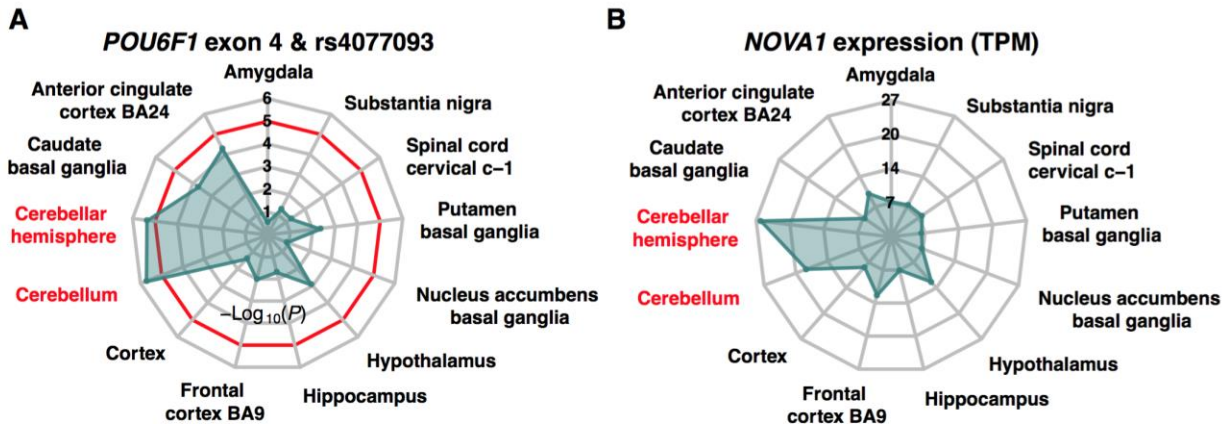
(**A**) Radar plot showing significance (-$\log_{10}$(*P*-value)) of the sQTL (*CXXC5* exon 2 and rs6580200) in each brain region.

(**B**) Radar plot showing the mean gene expression of HNRNPK in each brain region.



Supplementary Figure 2.21 SNP rs4077093 regulates splicing of *POU6F1* exon 4 by affecting NOVA1 binding

SNP rs4077093 regulates splicing of *POU6F1* exon 4 by disrupting a NOVA1 binding site on the exon body, which leads to more inclusion of the *POU6F1* exon 4.



Supplementary Figure 2.22 Significance of the sQTL (*POU6F1* exon 4 and rs4077093) agrees with gene expression of NOVA1

(**A**) Radar plot showing significance (-$\log_{10}$(*P*-value)) of the sQTL (*POU6F1* exon 4 and rs4077093) in each brain region.

(**B**) Radar plot showing the mean gene expression of NOVA1 in each brain region.

Chapter 3 Sexual dimorphic alternative splicing in mammalian brain

3.1 Introduction

Sexual dimorphism is a phenomenon that two sexes of the same species show different characteristics in their non-sexual organs. Emerging evidence show that there are widespread sex differences in the adult human brain[1-4]. These sex differences not only affect normal neurodevelopment[5, 6], but also neurological disorders[7, 8]. Previous studies have studied the mechanism of sexual dimorphism from neuroanatomical and behavioral perspectives, however, the underlying molecular mechanisms are still poorly understood[8-11].

In mammals, the development of male and female is determined based on the sex chromosome in the embryo (XY and XX). The gene *Sry* on the Y chromosome controls the testis development. If *Sry* is absent, it allows the development of ovaries. For mice carrying an autosomal *Sry* transgene in an XX background, they develop testes. On the contrary, for mice carrying an inactivate *Sry* gene in an XY background, they develop ovaries[9, 12]. Although sex specific features are mostly attributed to sex hormone production[13], studies also indicate that the sex chromosome complement on the gene expression level also has effects on functions like development, behavior and disease-susceptibility[10]. However, the mechanism of how gene expression leads to sexual dimorphisms is not well understood. One well-studied process is X chromosome inactivation (XCI)[14-17]. During early embryo development, both X chromosomes in XX females are active. To compensate for the dosage of X-linked genes between XX females and XY males, one X chromosome (Xi) is inactivated. XCI is caused by a long non-coding RNA (lncRNA) called Xist. During XCI, Xist covers the Xi, which makes the X chromosome into silent heterochromatin. After XCI in the early embryo, the Xi stays silent throughout the

development. However, Xist is not essential for the maintenance of the silent X chromosome[18]. Although the role of Xist in XCI is well studied, very few studies have investigated the consequence of Xist loss in mature tissues, especially brain[19].

Alternative splicing is known for regulating gene expression and the output of most mammalian genes[20]. This regulatory mechanism is common in mammalian nervous system. It can affect multiple different aspects of neuronal cell biology, including neuronal differentiation, neuronal connectivity, and synaptic transmission[21]. Previous genetic studies in Drosophila have shown that alternative splicing regulates the somatic sexual development of the fly[22-24]. However, the regulatory mechanism of sex determination is not conserved between insects and vertebrates. In addition, the role of alternative splicing in mammalian sexual dimorphism is not well studied. Although previous studies have identified differential alternative isoform expression in mammalian between males and females, the underlying mechanism causing these differences and its significance are poorly understood[4, 25].

RNA binding proteins (RBPs) are key regulators of alternative splicing. Recent studies have identified several RBPs interacting with Xist[26-28]. These RBPs include PTBP1 and PTBP2. They bind to a repeat sequence that is essential to Xist function. The PTB proteins are a family of proteins that are well known regulators of alternative splicing. They control extensive programs of isoform expression in the brain[29-32]. Both PTBP1 and PTBP2 are splicing repressors for exons carrying their binding site, however, they also show position-dependent effect, which means they can also act as splicing enhancers[31, 32]. They can also bind in 3' UTRs to affect cytoplasmic mRNA utilization[32, 33]. It has been reported by several genome-wide analyses that PTB proteins are involved in diverse cellular functions[29, 31, 32, 34-37]. For example, during neuronal progenitor cell differentiation, PTBP1 expression is turned off and PTBP2 is induced[29, 30, 34]. This

on and off switch of PTBP1 and PTBP2 initiates a transition in isoform expression, i.e., exons that are more sensitive to PTBP1 are no longer repressed and are included in the final mature mRNA. For exons that are affected by both proteins, they will continue to be repressed given the presense of PTBP2 until PTBP2 is reduced later in neuronal maturation, which generates mature neuron signature on the isoform level[31, 32, 37]. In addition, previous studies have shown physical interaction between PTBP1/2 and Xist[38]. Noncoding RNAs are known to affect alternative splicing through multiple molecular mechanisms. One mechanism is through sequestering regulatory proteins away from their targets, and then alters their expression level or modifies their activity[39-45]. Given all the evidence, we hypothesized that Xist may sequester PTBP1 and PTBP2 away from their targets in female cells, which leads to transcriptome-wide sexual dimorphic splicing. In order to test this hypothesis, RNA-seq data with Xist or PTBP1/2 knockdown based on the four core genotype (FCG) mouse models are generated. Comprehensive computational analysis is carried out to elucidate the underlying molecular mechanism and significance of sexual dimorphic alternative splicing.

3.2 Results

3.2.1 Development stage and tissue are the main factors contributing to the splicing variation based on the FCG mouse data

RNA-seq data of 64 samples (**Table 3.1**) from FCG mouse models was generated. We applied our computational tool rMATS[46], developed for quantifying alternative splicing events from large-scale RNA-seq data, to estimate exon splicing levels (percent spliced in, or PSI) from the FCG mouse RNA-seq data. After applying filters to select high-confidence alternative splicing events with reliable RNA-seq quantitation, Principal component analysis (PCA) was used to

inspect relationships among samples. Scree plot showed that the first three PCs captured the most variance on the splicing level (**Figure 3.1A**). By overlaying information from development stage, tissue and genotype to the PCA plot, we found that the samples were well separated based on development stage (PC1) and tissue (PC2) (**Figure 3.1B-D**).

3.2.2 Pairwise comparison between different genotypes from the FCG model across tissue and development stage reveals distinct splicing pattern

There are four genotypes in the FCG mouse model: XXF, XYF, XXM, and XYM. XXF represents wild type female mice. XYF represents mice carrying an inactivate *Sry* gene in an XY background. They develop ovaries. XXM represents mice carrying an autosomal *Sry* transgene in an XX background. They develop testes. XYM represents wild type male mice. Sex chromosome effect can be quantified by comparing XXF vs. XYF and XXM vs. XYM. Sex hormone effect can be quantified by comparing XXF vs. XXM and XYF vs. XYM. In our study, we carried out all six pairwise comparisons to have a comprehensive understanding of the effect from sex chromosome and hormone on the splicing level. Each comparison was performed in a given development stage and tissue. In total, we have 24 comparisons. For each genotype pairwise comparison, exons significant (FDR<5% and the difference between maximal PSI value and minimal PSI value, i.e., delta-PSI, is greater than 10%) in at least one tissue and development stage were plotted (**Figure 3.2**). Hundreds of significant exons were identified for each genotype pairwise comparison (**Table 3.2**), indicating extensive splicing changes induced by sex chromosome and sex hormones.

3.2.3 Differential splicing events between genotypes show high specificity to development stage and tissue

In order to test if the significant differential splicing events are specific to particular development stage or tissue, we calculated the overlap of significant differential splicing events among all the development stages and tissues (**Figure 3.3**). Based on the Venn diagram, the overlap between different development stages and tissues show little overlap, indicating that the differential splicing events are highly specific to development stage and tissue.

3.2.4 Differential splicing between genotypes within the same development stage and tissue

The next question is, for the same development stage and tissue, what is the overlap of significant differential splicing events between different pairwise comparisons? UpSet plot was generated to show the overlap (**Figure 3.4**). Most significant differential splicing events are specific to a particular comparison, indicating that the splicing difference is highly dependent on sex, sex chromosome, and sex hormone. Interestingly, from our result, the overlap between comparison XXM vs. XYM and XXF vs. XYF is very low, although both comparisons reflect effect from sex chromosome. It is the same with XXF vs. XXM and XYF vs. XYM, where both comparisons reflect effect from sex hormone. This result shows that the effect of sex chromosome and sex hormone is different between females and males.

3.2.5 PTBP2 knockdown leads to splicing changes in the mouse cortex across different development stages

In order to study the role of PTBP2 in sexual dimorphic alternative splicing, we generated data with PTBP2 knockdown at two development stages (E18 and P60) in mouse cortex (**Table 3.3**).

Same pipeline was used to quantify alternative splicing and differential splicing events between different conditions. Since full PTBP2 knockout in Ptbp2f/f homozygous is lethal, we used Ptbp2f/+(hetero) genotype to have samples in later stage up to P60. Although PTBP2 protein level in hetero would be not completely gone, there would be effects if PTBP2 is involved in the sexual dimorphic alternative splicing. By comparing the PTBP2 knockdown samples between males and females at E18 and P60, we identified many significant differential splicing events (**Figure 3.5**), indicating a role of PTBP2 in the sexual dimorphic alternative splicing.

3.2.6 Xist knockdown leads to splicing changes in the mouse cerebellum and cortex across different development stages

We hypothesized that Xist is a mediator between PTBP1/2 and sexual dimorphic alternative splicing. To test this hypothesis, we generated data with Xist knockdown in mouse cerebellum and cortex (**Table 3.4**). To make sure the knockdown works, we first checked the expression level of Xist (**Supplementary Figure 3.8**). In P60 Cortex, the expression level of Xist in samples with Xist knockdown is significantly lower than samples with normal Xist expression, which proved the quality of the data.

To test the effect of Xist on alternative splicing, we did pairwise comparison between Xist knockdown group and the control group across males and females. Hundreds of significant differential splicing events were identified in both cerebellum and cortex (**Figure 3.6** and **Figure 3.7**). This shows that Xist has profound impact on alternative splicing. Interestingly, we overlapped the differential splicing result between two comparisons in P60 Cortex, Xist knockdown in male vs. control in male and Xist knockdown in female vs. control in female. The

overlap is small (**Supplementary Figure 3.9**), indicating the effect of Xist on alternative splicing is different in males and females.

3.3. Discussions

This project is still ongoing. More data will be generated in the future. Based on the current result, we can already see transcriptome-wide splicing changes between genotypes, Xist knockdown conditions and PTBP2 knockdown conditions. These changes on the splicing level lend credibility to our study design. This project is innovative in both the biological question to be addressed and the methodology we used to tackle this question. Different from previous studies that focus on the role of sex hormones or the *Sry* gene in the sexual dimorphism of mammalian brain, our study investigate the entire transcriptome for generating the sexually dimorphic gene expression and alternative splicing. Although the role of Xist in XCI is extensively studied, the impact of Xist on the downstream gene expression in mammalian brain is poorly characterized. The fact that PTBP1/2 binds to Xist points to a testable model that Xist sequesters PTBP1/2 away from their targets, and then alters their expression level or modifies their splicing regulatory activity, which finally leads to sexual dimorphic RNA splicing. More importantly, this model can be extended to other proteins that interact with Xist.

This project also develops novel computational and experimental methods to study sexual dimorphic gene expression and alternative splicing. rMATS, a software that identifies splicing across multiple samples, while accounting for sample variance and measurement uncertainty, is used in the project, which provides a new level of rigor. In addition, the genetic models used in the study provide a powerful way to study the sexual dimorphic splicing while ruling out the effect from gonadal hormones. The engineered mouse lines with Xist and PTBP1/2 knockdown

allow us to test the impact of these molecules in controlling the transcriptome-wide gene expression and splicing changes in a more definitive way. Overall, this project will depict a transcriptome-wide picture of sexually dimorphic alternative splicing in the mammalian brain at unprecedented resolution, and shed light on the underlying molecular mechanisms and regulatory networks.

3.4 Methods

3.4.1 FCG mouse data

Male and female mice (C57BL/6J) with Sry gene activated and inactivated were generated. Samples from cerebellum and cortex were obtained for sequencing. In total, we have 64 samples (**Table 3.1**). PolyA selected, 75 base pair-end and stranded RNA-seq data were generated with sequencing depth between 20 million and 30 million per sample.

3.4.2 PTBP2 knockdown data

PTBP2 flox mice that have loxP sites inserted flanking exon 4 of PTBP2 were obtained. We also obtained NestinCre transgenic mice that express Cre recombinase in neuronal progenitor cells. By crossing PTBP2 flox mice with NestinCre transgenic mice, we obtained mice with PTBP2 knockdown expression. Therefore, there are 3 genotypes in females and male (PTBP2f/+, NestinCre/+ and +/+, + in the genotype means wild type allele). We prepared cortex tissues from E18 and P60, n=4 for each genotypes. In total we have 48 samples (**Table 3.3**). RNA sequencing library was prepared with TruSeq Stranded mRNA from total RNA, and run in HiSeq4000, 75 base pair-end.

3.4.3 Xist knockdown data

The same knockdown method was also used to knockdown Xist expression. We prepared cortex tissues from P60 mice and cerebellum tissues from P5 and P60 tissues. For each tissue and development stage, we have 4 samples for each condition (male with/without Xist knockdown, female with/without Xist knockdown). Therefore, we have 16 samples from cortex and 32 samples from cerebellum (**Table 3.4**).

3.4.4 RNA-seq data processing

RNA-seq reads were mapped using STAR 2.5.3a. Gene expression quantification (in transcripts per kilobase million, TPM) was done using Kallisto- v0.43.1. Median normalization was performed on each sample for between-sample normalization. rMATS was run on the BAM files generated by STAR to obtain exon inclusion levels (PSI values), with missing PSI values imputed by using k-nearest neighbors. To ensure reliability of the imputation and downstream analyses, we only included an exon in our analyses if it passed the following filters:

1) Average PSI Î [0.05 0.95],

2) Average total count (inclusion count + skipping count) ≥10,

3) Missing PSI value < 10%,

4) max(PSI) – min(PSI) > 0.05

3.5 References

1. Cosgrove KP, Mazure CM, Staley JK. Evolving knowledge of sex differences in brain structure, function, and chemistry. *Biological psychiatry* **62**, 847-855 (2007).

2. Grgurevic N, Majdic G. Sex differences in the brain–an interplay of sex steroid hormones and sex chromosomes. *Clinical Science* **130**, 1481-1497 (2016).

3. Sacher J, Neumann J, Okon-Singer H, Gotowiec S, Villringer A. Sexual dimorphism in the human brain: evidence from neuroimaging. *Magnetic resonance imaging* **31**, 366-375 (2013).

4. Trabzuni D*, et al.* Widespread sex differences in gene expression and splicing in the adult human brain. *Nature communications* **4**, 2771 (2013).

5. Chung WCJ, Auger AP. Gender differences in neurodevelopment and epigenetics. *Pflügers Archiv-European Journal of Physiology* **465**, 573-584 (2013).

6. Shi L, Zhang Z, Su B. Sex biased gene expression profiling of human brains at major developmental stages. *Scientific reports* **6**, 21181 (2016).

7. May M. Sex on the brain: unraveling the differences between women and men in neurodegenerative disease. (ed^(eds). Nature Publishing Group (2016).

8. Qureshi IA, Mehler MF. Genetic and epigenetic underpinnings of sex differences in the brain and in neurological and psychiatric disease susceptibility. In: *Progress in brain research* (ed^(eds). Elsevier (2010).

9. McCarthy MM, Arnold AP. Reframing sexual differentiation of the brain. *Nature neuroscience* **14**, 677 (2011).

10. Ngun TC, Ghahramani N, Sánchez FJ, Bocklandt S, Vilain E. The genetics of sex differences in brain and behavior. *Frontiers in neuroendocrinology* **32**, 227-246 (2011).

11. Sekido R. The potential role of SRY in epigenetic gene regulation during brain sexual differentiation in mammals. In: *Advances in genetics* (ed^(eds). Elsevier (2014).

12. Arnold AP, Chen X. What does the "four core genotypes" mouse model tell us about sex differences in the brain and other tissues? *Frontiers in neuroendocrinology* **30**, 1-9 (2009).

13. Arnold AP. The end of gonad-centric sex determination in mammals. *Trends in genetics* **28**, 55-61 (2012).

14. Gendrel A-V, Heard E. Noncoding RNAs and epigenetic mechanisms during X-chromosome inactivation. *Annual review of cell and developmental biology* **30**, 561-580 (2014).

15. Kamikawa Y, Donohoe ME. The dynamics of X‐chromosome inactivation in mouse development. *Molecular reproduction and development* **81**, 141-147 (2014).

16. Lee JT, Bartolomei MS. X-inactivation, imprinting, and long noncoding RNAs in health and disease. *Cell* **152**, 1308-1323 (2013).

17. Payer B, Lee JT. Coupling of X-chromosome reactivation with the pluripotent stem cell state. *RNA biology* **11**, 798-807 (2014).

18. Csankovszki G, Nagy A, Jaenisch R. Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. *The Journal of cell biology* **153**, 773-784 (2001).

19. Yildirim E*, et al.* Xist RNA is a potent suppressor of hematologic cancer in mice. *Cell* **152**, 727-742 (2013).

20. Braunschweig U, Gueroussov S, Plocik AM, Graveley BR, Blencowe BJ. Dynamic integration of splicing within gene regulatory pathways. *Cell* **152**, 1252-1269 (2013).

21. Li Q, Lee J-A, Black DL. Neuronal regulation of alternative pre-mRNA splicing. *Nature Reviews Neuroscience* **8**, 819 (2007).

22. Black DL. Mechanisms of alternative pre-messenger RNA splicing. *Annual review of biochemistry* **72**, 291-336 (2003).

23. Manoli DS, Foss M, Villella A, Taylor BJ, Hall JC, Baker BS. Male-specific fruitless specifies the neural substrates of Drosophila courtship behaviour. *Nature* **436**, 395 (2005).

24. Stockinger P, Kvitsiani D, Rotkopf S, Tirián L, Dickson BJ. Neural circuitry that governs Drosophila male courtship behavior. *Cell* **121**, 795-807 (2005).

25. Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y. Sex-specific and lineage-specific alternative splicing in primates. *Genome research* **20**, 180-189 (2010).

26. Chu C*, et al.* Systematic discovery of Xist RNA binding proteins. *Cell* **161**, 404-416 (2015).

27. McHugh CA*, et al.* The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521**, 232 (2015).

28. Minajigi A*, et al.* A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science* **349**, aab2276 (2015).

29. Boutz PL*, et al.* A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes & development* **21**, 1636-1652 (2007).

30. Keppetipola N, Sharma S, Li Q, Black DL. Neuronal regulation of pre-mRNA splicing by polypyrimidine tract binding proteins, PTBP1 and PTBP2. *Critical reviews in biochemistry and molecular biology* **47**, 360-378 (2012).

31. Li Q*, et al.* The splicing regulator PTBP2 controls a program of embryonic splicing required for neuronal maturation. *Elife* **3**, e01201 (2014).

32. Licatalosi DD*, et al.* Ptbp2 represses adult-specific splicing to regulate the generation of neuronal precursors in the embryonic brain. *Genes & development* **26**, 1626-1642 (2012).

33. Xue Y*, et al.* Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Molecular cell* **36**, 996-1006 (2009).

34. Makeyev EV, Zhang J, Carrasco MA, Maniatis T. The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Molecular cell* **27**, 435-448 (2007).

35. Spellman R, Llorian M, Smith CWJ. Crossregulation and functional redundancy between the splicing regulator PTB and its paralogs nPTB and ROD1. *Molecular cell* **27**, 420-434 (2007).

36. Tang ZZ, Sharma S, Zheng S, Chawla G, Nikolic J, Black DL. Regulation of the mutually exclusive exons 8a and 8 in the CaV1. 2 calcium channel transcript by polypyrimidine tract-binding protein. *Journal of Biological Chemistry* **286**, 10007-10016 (2011).

37. Zheng S, Gray EE, Chawla G, Porse BT, O'Dell TJ, Black DL. PSD-95 is post-transcriptionally repressed during early neural development by PTBP1 and PTBP2. *Nature neuroscience* **15**, 381 (2012).

38. König J*, et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology* **17**, 909 (2010).

39. Änkö M-L, Neugebauer KM. Long noncoding RNAs add another layer to pre-mRNA splicing regulation. *Molecular cell* **39**, 833-834 (2010).

40. Bergmann JH, Spector DL. Long non-coding RNAs: modulators of nuclear structure and function. *Current opinion in cell biology* **26**, 10-18 (2014).

41. Chen LL, Carmichael GG. Long noncoding RNAs in mammalian cells: what, where, and why? *Wiley Interdisciplinary Reviews: RNA* **1**, 2-21 (2010).

42. Yin Q-F*, et al.* Long noncoding RNAs with snoRNA ends. *Molecular cell* **48**, 219-230 (2012).

43. Rinn J, Guttman M. RNA and dynamic nuclear organization. *Science* **345**, 1240-1241 (2014).

44. Lee JE, Cooper TA. Pathogenic mechanisms of myotonic dystrophy. (ed^(eds). Portland Press Limited (2009).

45. Tripathi V, *et al.* The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular cell* **39**, 925-938 (2010).

46. Shen S, *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences* **111**, E5593-E5601 (2014).

3.6 Figures

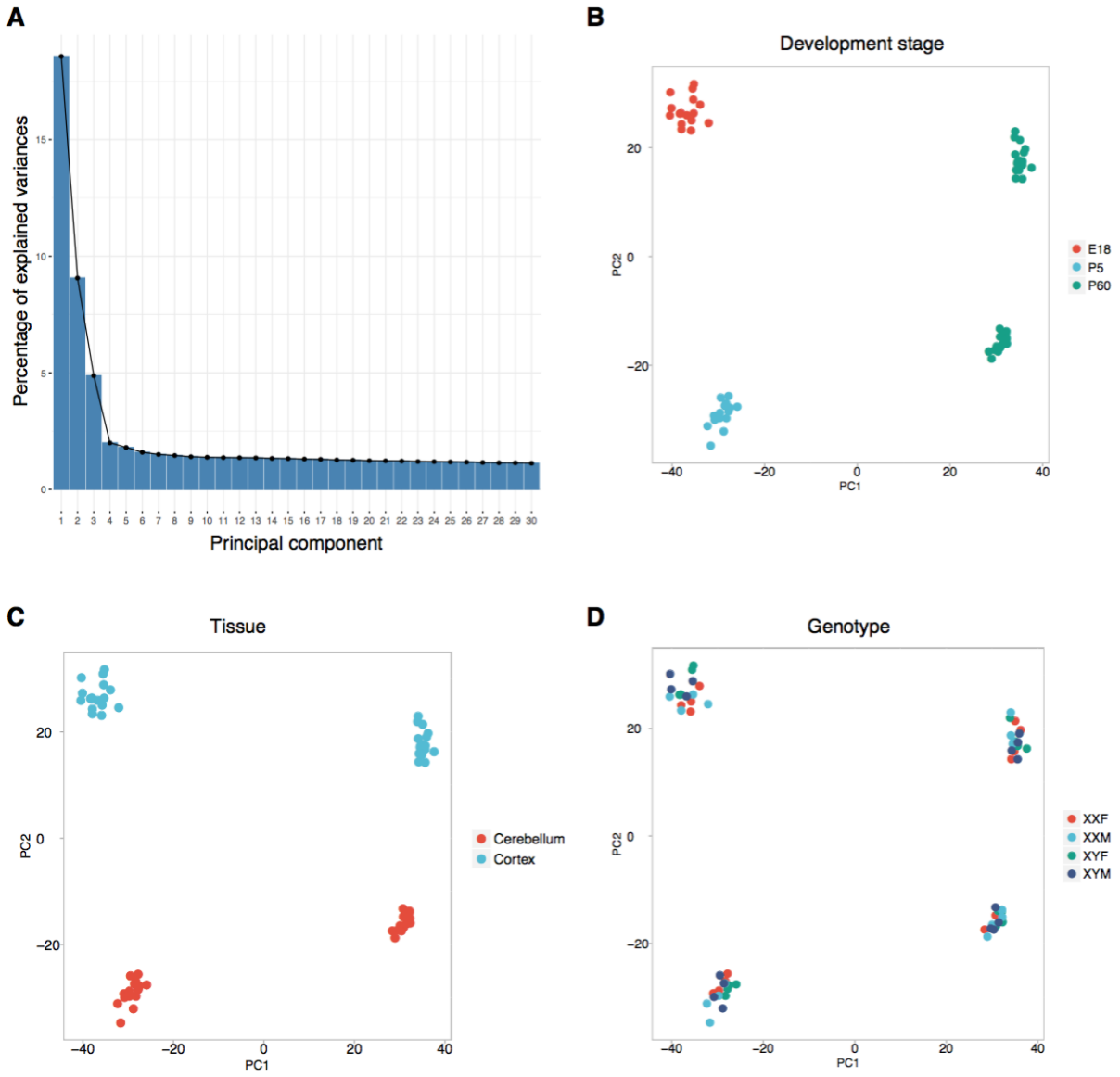


Figure 3.1 PCA result based on RNA-seq data from FCG mouse model

(**A**) Scree plot showing the percentage of explained variance by each principal component

(**B-D**) Plot of PC1 and PC2 overlaid with information of development stage, tissue and genotype

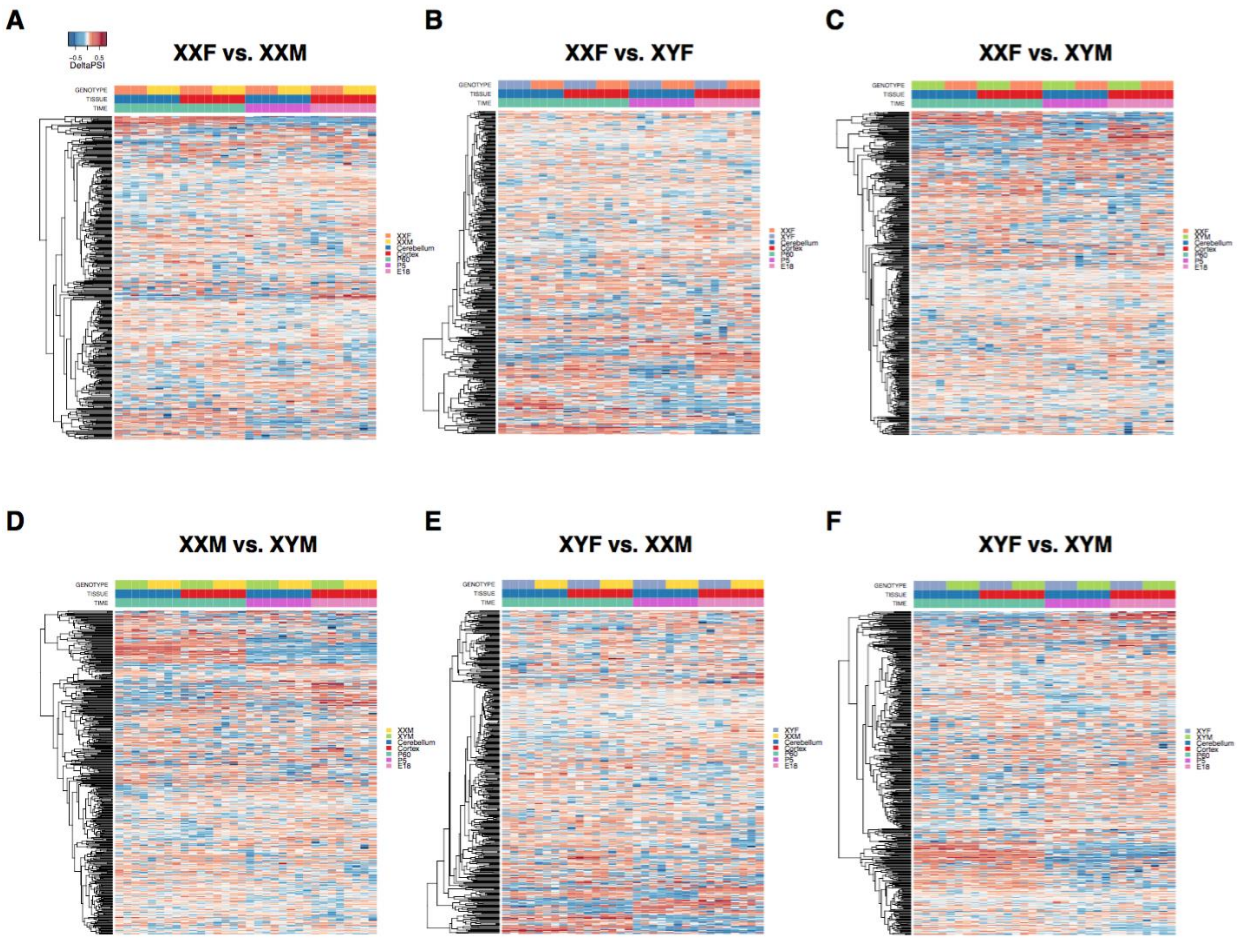


Figure 3.2 Heat map of significant differential splicing exons between different pairs of genotypes

Heat map showing the significant differential splicing events between different pairwise comparisons. Each row represents a splicing event and each column represent a sample. The top annotation bar shows the development stage, tissue and genotype information of each sample.

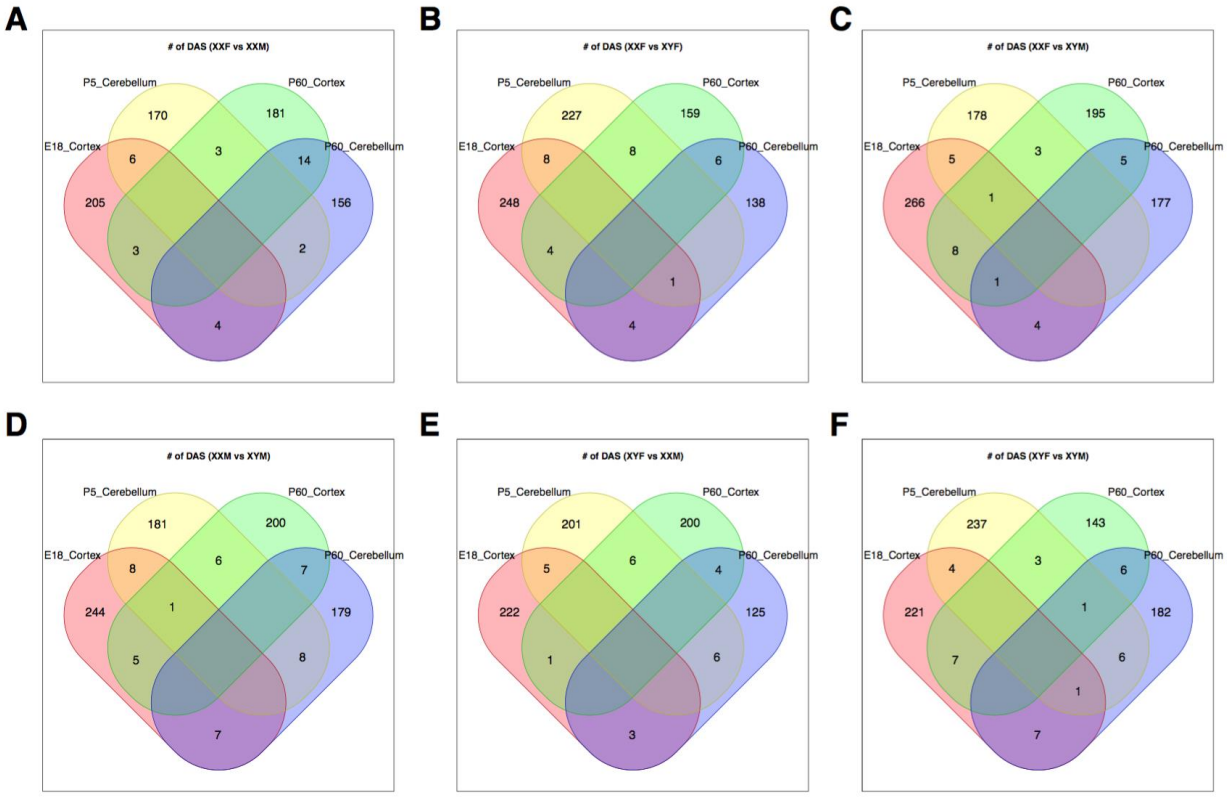


Figure 3.3 Overlap of significant differential splicing events for each comparison under different development stage and tissue combination

Venn diagram showing the overlap of significant differential splicing events for each comparison under different development stage and tissue combination.

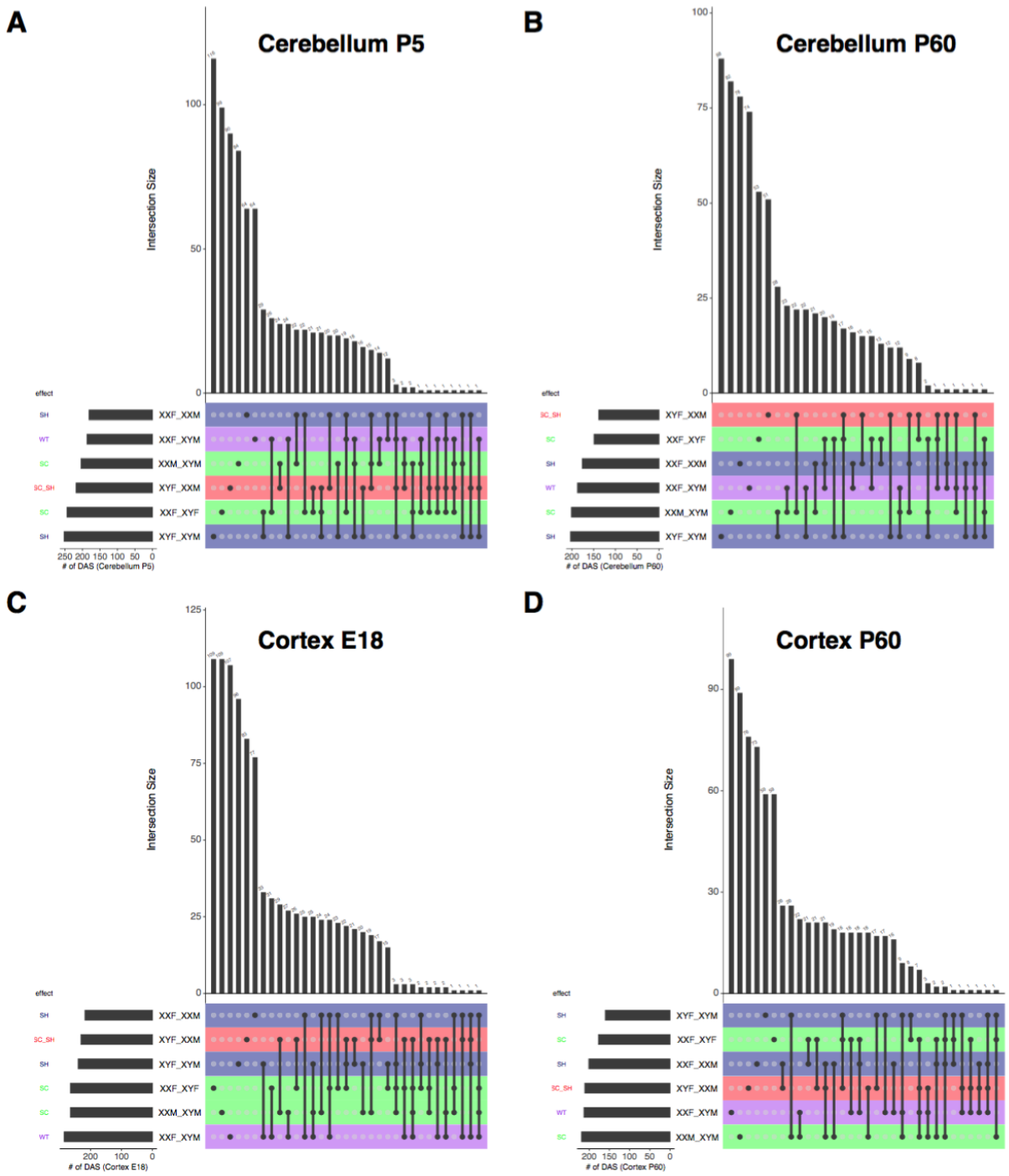


Figure 3.4 Overlap of differential splicing events between all pairs of genotypes within the same development stage and tissue

UpSet plot showing the overlap of differential splicing events between all pairs of genotypes for each development stage and tissue combination

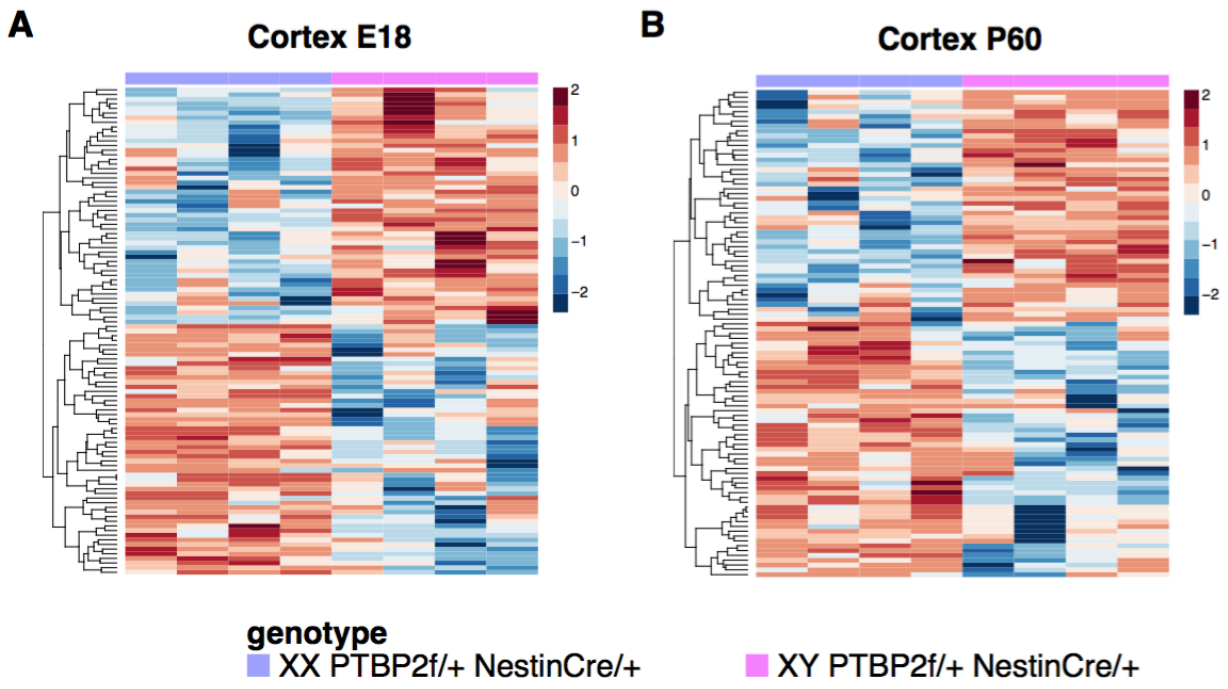


Figure 3.5 PTBP2 knockdown leads to splicing changes in the mouse cortex across different development stages

Heat map of significant differential splicing events between PTBP2 knockdown in males and females in mouse cortex under different development stages. Each row represents a splicing events and each column represents a sample. The top color bar shows the genotype of each sample.

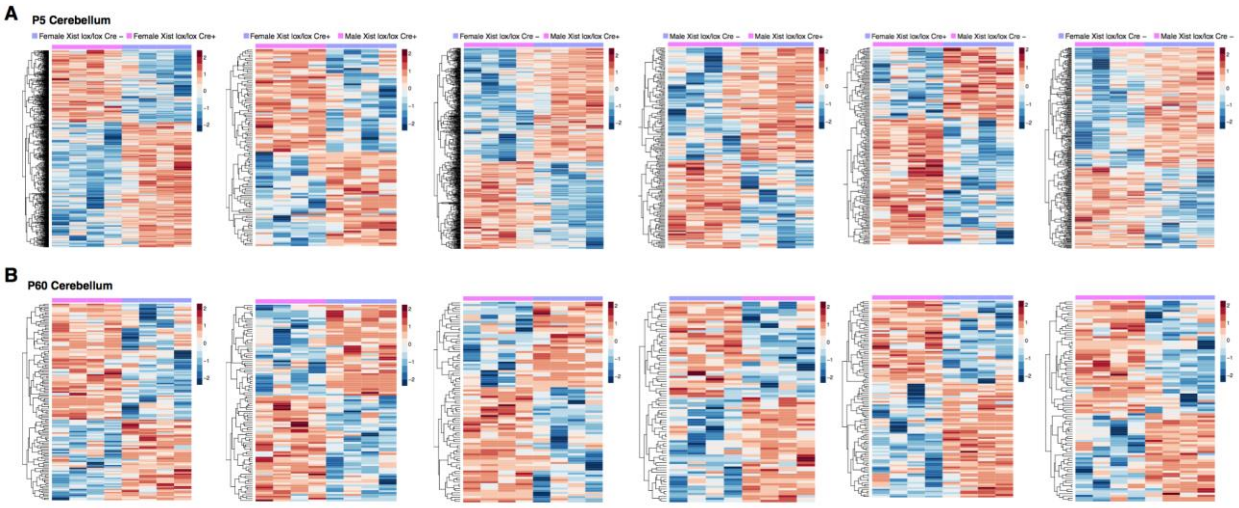


Figure 3.6 Differential splicing events between Xist knockdown and control in males and females in the mouse cerebellum across different development stages

Heat map of significant differential splicing events between Xist knockdown and control in males and females in mouse cerebellum under different development stages. Each row represents a splicing events and each column represents a sample. The top color bar shows the genotype of each sample.

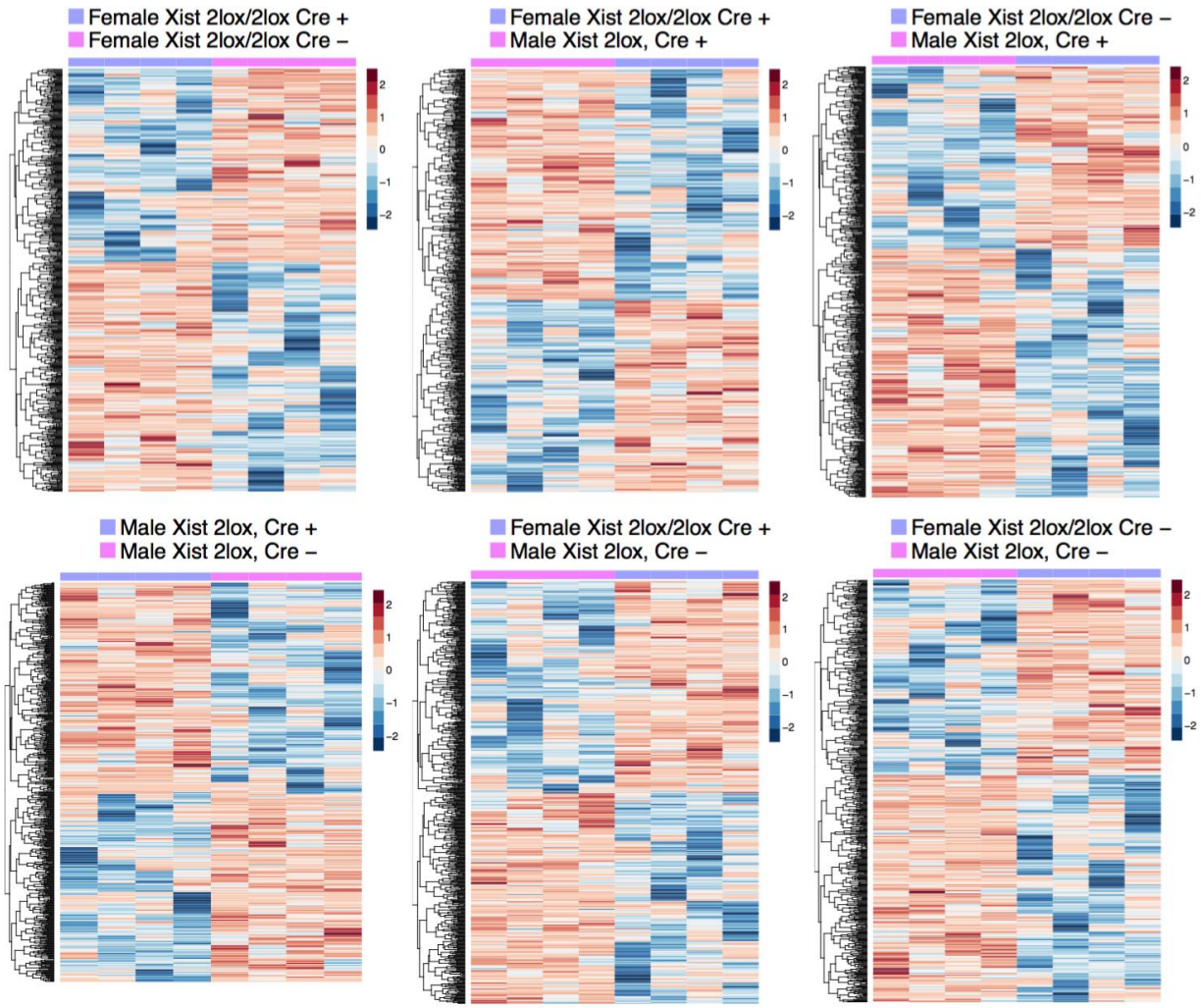
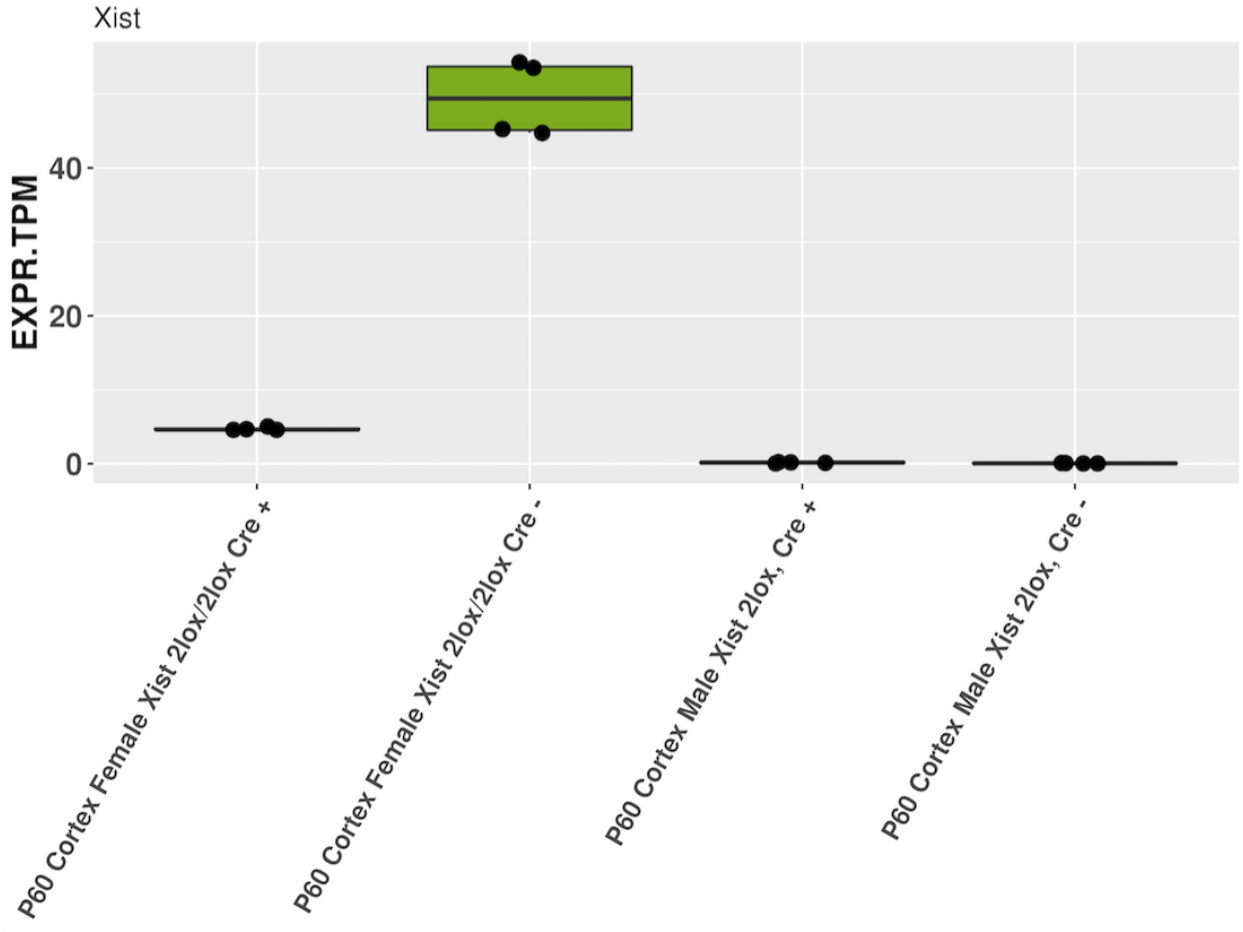


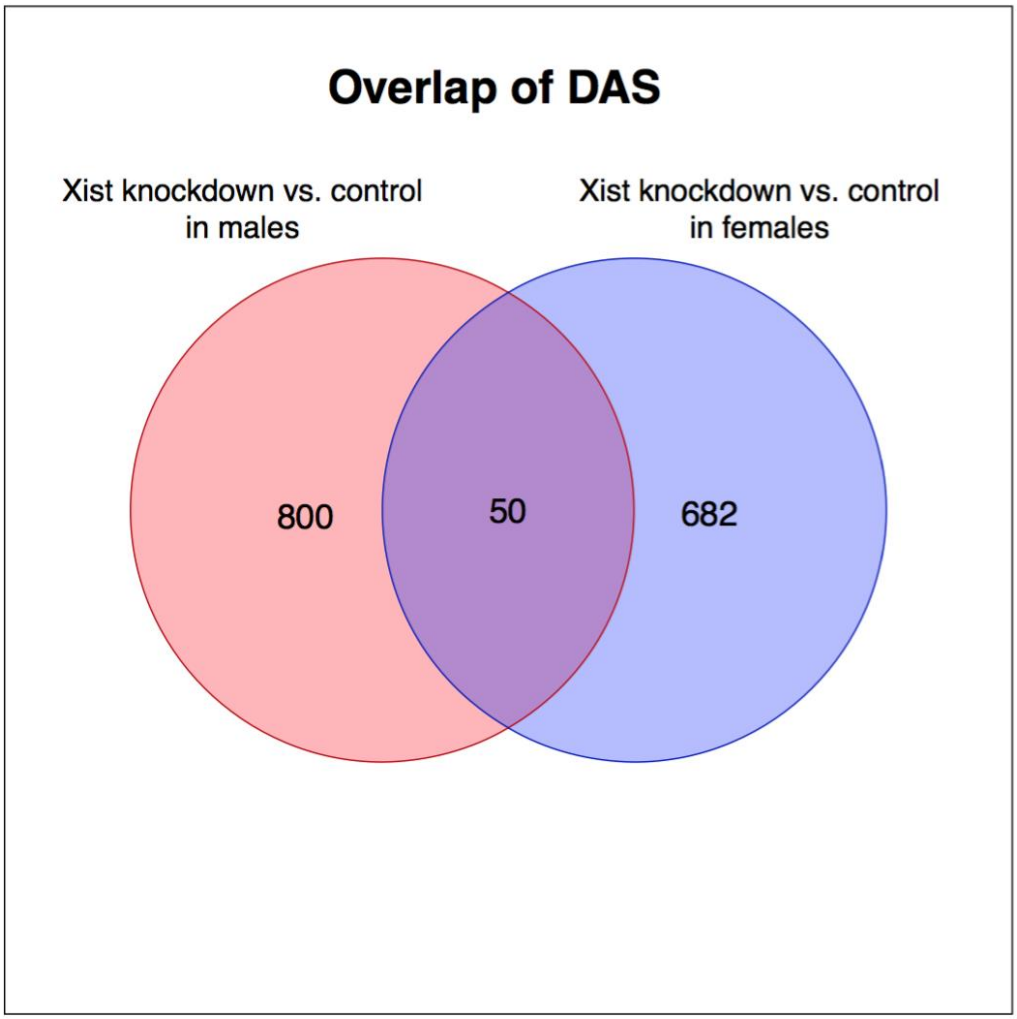
Figure 3.7 Differential splicing events between Xist knockdown and control in males and females in the mouse cortex at P60

Heat map of significant differential splicing events between Xist knockdown and control in males and females in mouse cerebellum at P60. Each row represents a splicing events and each column represents a sample. The top color bar shows the genotype of each sample.



Supplementary Figure 3.8 Xist expression in knockdown and control samples

Box plot showing the expression of Xist (in TPM) in knockdown and control samples.



Supplementary Figure 3.9 Overlap of the differential splicing events between Xist knockdown vs. control in males and in females in the mouse cortex at P60

Venn diagram showing the overlap of the differential splicing events between Xist knockdown and control in males and in females in P60 Cortex

3.7 Tables

Table 3.1 Sample size distribution of FCG mouse data

	Cerebellum		Cortex	
	P5	**P60**	**E18**	**P60**
XXF	4	4	4	4
XYF	4	4	4	4
XXM	4	4	4	4
XYM	4	4	4	4

Table 3.2 Number of significant differentially splicing events in six pairwise comparisons

	Cerebellum		Cortex	
	P5	**P60**	**E18**	**P60**
XXF-XYF	244	149	265	177
XXF-XXM	181	176	218	201
XXF-XYM	187	187	285	213
XYF-XXM	218	138	231	211
XYF-XYM	252	203	240	160
XXM-XYM	204	201	265	219

Table 3.3 Sample size distribution of PTBP2 knockdown data

Sample Genotype			Cortex	
			E18	P60
XX	**PTBP2f/+**	**NestinCre/+**	4	4
XX	**PTBP2f/+**	**+/+**	4	4
XX	**+/+**	**NestinCre/+**	4	4
XY	**PTBP2f/+**	**NestinCre/+**	4	4
XY	**PTBP2f/+**	**+/+**	4	4
XY	**+/+**	**NestinCre/+**	4	4

Table 3.4 Sample size distribution of Xist knockdown data

	Cerebellum		Cortex
	P5	**P60**	**P60**
Male Xist lox/lox Cre +	4	4	4
Male Xist lox/lox Cre -	4	4	4
Female Xist lox/lox Cre+	4	4	4
emale Xist lox/lox Cre -	4	4	4

Chapter 4 Concluding Remarks

We have witnessed a paradigm shift in life science study in the past few decades. Taking advantage of the development of biotechnologies, tons of data is generated, which transforms the previous low throughput small data paradigm into today's high throughput big data paradigm. Besides the large amount of data, we are also faced with increasing complexity of data. Functional assays have characterized and annotated biological processes from DNA, to RNA and protein. Numerous modification and regulation processes are captured. This new ear of big data provides us opportunities to understand life at an unprecedented resolution. However, in the meantime, it also creates challenges of how to manage and interpret the underlying message of the data. Altogether, it is the best of times, it is the worst of times, it is the age of opportunities, it is the age of challenges, it is the epoch of complexity, it is the epoch of clarity, it is the winter of despair, it is the spring of hope.

Among the numerous layers of biological functions, this dissertation has been focused on a post-transcriptional regulation process alternative splicing, which is a major contributor to the complexity of biological regulations, in mammalian brain, the most complex organ. Alternative splicing is a ubiquitous regulatory mechanism that greatly expands the complexity at the RNA level. Compared with research on the genome level, research on the transcriptome level has long been overlooked until recent years when RNA sequencing technology becomes widely used. This gives us the opportunity to take a closer look at the role of alternative splicing in numerous biological conditions.

In Chapter 2, we studied the genetically regulated alternative splicing. Previous studies have identified lots of genetic variants in different diseases. However, it is hard to separate causal effect from secondary effect. In addition, how do the genetic variants affect downstream

biological functions is poorly understood. To fill the gap, we investigated the connection between genotype and phenotype through alternative splicing. By identifying hundreds of splicing QTLs in human brain and their connection with neurological disorders, we not only depicted a comprehensive picture of sQTLs in human brain, but also provided a general framework of how to use variation on the splicing level across different tissues to fine map causal variants underlying genetically regulated alternative splicing events and their associated neurological diseases.

In Chapter 3, we aimed to study the role of alternative splicing in sexual dimorphism in mouse brain. Sex effect has long been overlooked in many experimental design and drug discovery. However, it has been proved that males and females have many differences like development and response to outside stimuli. To better understand the sexual dimorphism in mammalian brain, we generated RNA-esq data using the FCG mouse model to dissect the sexual effect coming from sex chromosomes and sex hormones. Since Xist plays a key role in the X chromosome inactivation, we also generated RNA-seq data with Xist knockdown to study its impact on the sexual dimorphic alternative splicing. In addition, as the regulators of alternative splicing, several RBPs have been reported to interact with Xist, for example, PTBP1 and PTBP2. To test their impact on the sexual dimorphic alternative splicing, we generated RNA-seq data with PTBP1 and PTBP2 knockdown. Altogether, we hypothesized that Xist may sequester PTBP1 and PTBP2 away from their targets in female cells, which leads to transcriptome-wide sexual dimorphic splicing. The data we generated provides us powerful tools to test our hypothesis.

Moving forward, the development of single cell sequencing technology and in situ sequencing technology will provide us with data at even higher resolution. We will be able to

look at the effect from single cells and also the spatial information of a biological process on the molecular level. New computational tools need to be developed to harness the power of data generated by these new technologies. In the meantime, as the cost of collecting data decreases, computational biologists need to put more thoughts on the study design and also the design of computational models to fully utilize the powerful technologies, and provides new insights into biological discoveries and ultimately, our understanding of life.