Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Hye Rim Kim Date

Identification of Genetic and Epigenetic Factors regulating cell death and proliferation in cerebellum-related brain disorders

By

Hye Rim Kim
Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences
Cancer Biology

Peng Jin, Ph.D.
Advisor

Paula M. Vertino, Ph.D.
Co-Advisor

Jing Chen, Ph.D.
Committee Member

Renee D. Read, PhD.
Committee Member

Karen N. Conneely, Ph.D
Committee Member

Thomas S. Wingo, MD
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Identification of Genetic and Epigenetic Factors regulating cell death and proliferation in cerebellum-related brain disorders

By

Hye Rim Kim

B.S. College of Pharmacy, Seoul National University, 2009

M.S. College of Pharmacy, Seoul National University, 2011

Advisor: Peng Jin, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Graduate Division of Biological and Biomedical Sciences
Cancer Biology
2019

ABSTRACT

Identification of Genetic and Epigenetic Factors regulating cell death and proliferation in cerebellum-related brain disorders
By Hye Rim Kim

Molecular characterization is the key to understanding disease pathophysiology and developing effective therapeutic agents. However, incomplete (or reduced) penetrance and numerous genetic and epigenetic alterations irrelevant to disease progression mask the identification of bona fide disease-associated factors. Furthermore, genome-wide association studies are limited for the discovery of common variants associated with complex and common disorders, and therefore, novel approaches are needed to determine true pathological variants in rare and complex disorders. In addition, abnormal changes in the epigenome are deemed as key determinants in many diseases, but their role in pathogenesis remains to be understood.

We recently utilized a three-step gene discovery strategy to facilitate the identification of novel genetic factors implicated in amyotrophic lateral sclerosis (ALS), a progressive neurodegenerative disease. Using whole-genome sequencing (WGS), we first identified genetic variants in 135 candidate genes associated with age-of-onset in patients with the G_4C_2 repeat expansion in the *C9orf72* gene (step 1). We then performed an unbiased genetic screen using a *Drosophila* model expressing 30 repeats of G_4C_2, identifying 18 genetic factors modifying G_4C_2 repeat-associated toxicity (step 2). To further test the association of the 18 genes with sporadic ALS risk, gene-based statistical analyses of targeted resequencing and WGS identified rare variants in *MYH15* as a modifying factor of ALS risk. We further demonstrated that *MYH15* modulates the toxicity caused by poly-dipeptides produced from the expanded G_4C_2 repeat.

The cerebellum is critical for motor movements, and thus, neurogenesis in the cerebellum must be sophisticatedly orchestrated for normal neuronal activity. Epigenetic modifications play a critical role in postnatal and adult neurogenesis, but the role of 5-hydroxymethylcytosine (5hmC), an abundant epigenetic factor, in this process remains to be elucidated. We performed genome-wide 5hmC profiling to characterize the genomic loci enriched with 5hmC throughout the processes of neurodevelopment and aging. We further investigated the role of 5hmC alterations in Medulloblastoma (MB), a tumor of the cerebellum. Collectively, these studies highlight the effectiveness of our novel approach to facilitate the identification of genetic modifiers in rare and complex disorders and expand our understanding of epigenetic dynamics in the context of both normal development/aging and diseases.

Identification of Genetic and Epigenetic Factors regulating cell death and proliferation in cerebellum-related brain disorders

By

Hye Rim Kim
B.S. College of Pharmacy, Seoul National University, 2009
M.S. College of Pharmacy, Seoul National University, 2011

Advisor: Peng Jin, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Graduate Division of Biological and Biomedical Sciences
Cancer Biology
2019

Acknowledgement

I deeply appreciate many individuals who have provided me with intellectual and personal support throughout my time at Emory. First, thank to my mentor, Peng. This is a first time for me to think of the meaning of what is a true mentor. I've been through many P.I.s till now and of course, they were great but I was under pressure since the feeling that I was judged by my behaviors and words. So maybe during a couple of years, I assumed you might be similar so did not have a lot of chances to talk about myself and what I truly want to do. You (and all my other committee members) may still doubt about what was is good for me but at least, you appreciated my decision and helped me to get through this. Without your support, I cannot write this acknowledgement. Thank you for me to write this on "April 25th, 2019". And my co-advisor, Paula. You always encouraged me to move forward my MB project and without you and your lab members (particularly Ben) who were always interested in this project, I cannot do this much. I am really sad that I did not take a picture with you before you left. I cannot forget your support in Emory. And my other committee members, Jing, Renee, Karen, and Thomas. When I recall my previous committee meetings now (5 in total), my presentations were not intended to communicate with you my project but rather to report you my progress since the meeting is mandatory. And I did not have a much time with you all to talk in person. That time, I thought it would be better to show you how much I did but now, I think you may want me to show how much I am improved as a scientist under your guidance (it maybe be still not true, though). But at least during the last couples of month, I am happy to have a chance to talk about me with you in person and a little bit open myself to you. If I could go back again, I would visit you more often to ask your help… Even though I cannot go back, I am really glad to have such great people in my committee. Since I am a kind of person who opens everything once it starts, I really want to keep in touch even after leaving Emory (at least, I will!).

Next, I need to thank all my lab members, Junghwa, Yujing, Yunhee, Ying, Feiran, Yaran, Ronghua, Keqin, Ha Eun, Zhiqin, Lin, Mengli, Ruibing, Yulin, Zhen, Kailin, Lei, Cuida, Hwanwook, Matt, and Andrew (OMG, so many!!). Some of you may know, without I list all your names, how much I have bugged you especially when I needed. Sorry if you feel uncomfortable. You definitely deserve beer or coffee with sweets (or you can choose my big hug ☺). And to some of you I did not have a chance to get closer yet. I am a little shy at first (and pretty bad at talking with whom I am not family with), but once I start to show myself, I am a pretty easy person to get along. So I want to have a more time with you while I am here! I will ask you once I am done with all paper works. Please do not say no! ☺

To CB buddies. Without you all, I cannot survive here. I still remember that we studied together for written and oral quals and had a lunch together before going to a seminar class. Of course anyone can think that was just a routine but I missed the time so much. Of course, we cannot do that anymore (and no one wants to go back before quals!), but I also remember "that time" and "you all" wherever I go. You guys are really awesome (younger) brothers and sisters to me!!! Please keep in touch and DO NOT IGNORE ME when you see later! I will kick you if you do! ☺

To my special friends, Ju Hyun and Se Yeon. Without you, I cannot imagine life in Atlanta. You are such good friends and sisters (even though I already have real two in Korea). I was not good for you like you did to me.. Forgive me if I did not treat you well during the time. And please keep in touch and you should invite me your defense! I virtually attend or send my avatar ☺ I will miss you so much!!

To my boyfriend, Jaehwa. First, I thank you for listening my public defense presentation SIX times (maybe if you count about SKAT part, more than 20 times..) and helping me to make that better. And like I said, you is a man version of "Hyerim Kim" so that is why we could get through so long although we live so apart and have lots of issues during the time. Again, thank you for sharing all times with me love you!!

Finally, to my family and especially Mom and Dad: no one can dedicate their time to their children equally. I know you (especially mom) are really sad when I said to stay abroad but you constantly believe me what I have done. I deeply admire your support, encouragement, and love to me. I am so lucky to have you as my parents. Love you so much and I am so excited to see you on May! And my lovely sisters. Thank you for being awesome sisters and always saying to me "we are always proud of you whatever you do". Whenever I was in tough time, that word encouraged me a lot. Cannot wait seeing you again this summer!

Sorry for not mentioning all who have supported me so far in details. Due to your support, I can finally complete my journey as a student. I will keep trying to be a better person, and I believe that is one of the ways to show my gratitude to others. I will share my experience to others to make a better world ☺ Thank you so much all!!

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION..... 1

1.1. Author's Contribution and Acknowledgement of Reproduction..... 1

1.2. Challenges in rare variant association studies of complex neurological disorders.. 2

1.2.1. Family-based sequencing studies 3

1.2.2. Sampling of affected patients with extreme phenotypes 3

1.2.3. Whole-Genome Sequencing 4

1.2.4. Targeted sequencing of prioritized candidate genes..... 5

1.2.5. Rare-variant association testing for sequencing data with the Sequence Kernel Association Test (SKAT)..... 6

1.2.6. Functional annotation of rare variants 7

1.2.7. Functional genomics using a *Drosophila* model 8

1.3. The significance of epigenetics in cellular functions and diseases..... 10

1.3.1. DNA modifications in genome: cytosine modifications and beyond..... 11

1.3.2. Technologies for genome-wide DNA modifications..... 14

1.3.3. Distinct genomic localization of TET proteins: intrinsic structural difference and the interaction with extrinsic factors 16

1.3.4. Significance of epigenetic alterations in pediatric cancer 17

1.4. Summary of background information and dissertation goals 18

CHAPTER 2: Rare Variants in MYH15 Modify Amyotrophic Lateral Sclerosis Risk..... 20

2.1. Author's Contribution and Acknowledgement of Reproduction..... 20

2.2. Introduction..... 21

2.3. Materials and Methods..... 23

2.4. Results..... 29

2.5. Discussion..... 39

2.6. Acknowledgements..... 42

CHAPTER 3: Aging-related epigenetic dynamics in cerebellum..... 43

3.1. Introduction..... 43

3.2. Materials and Methods..... 45

3.3. Results..... 48

3.4. Discussion..... 60

CHAPTER 4: TET1-mediated 5-hydroxymethylcytosine Alteration in the pathogenesis of Medulloblastoma 62
4.1. Introduction 62
4.2. Materials and Methods 64
4.3. Results 69
4.4. Discussion 86
CHAPTER 5: Summary 89
5.1. Summary of key findings 89
5.2. Clinical implications 92
REFERENCES 94
SUPPLEMENTAL TABLES 133

TABLE OF FIGURES

Figure 1.1. The cycle of cytosine modifications. 13
Figure 2.1. 3-step strategy to identify genetic factors associated with ALS risk using a hypothesis-driven and targeted genetic association study (step 1 and step 3) and fly genetics (step 2). 30
Figure 2.2. Functional screen identifies multiple genetic modifiers of $(G_4C_2)_{30}$ toxicity 32
Figure 2.3. Coding variants of *MYH15* identified in either ALS cases or controls during the targeted resequencing (A) and validation dataset (B). 36
Figure 2.4. MYH15 is a potential genetic modifier of dipeptide-mediated toxicity 38
Figure 3.1. Distinct age-dependent 5hmC patterns 49
Figure 3.2. Age-dependent DhMRs are enriched at different genomic loci. 52
Figure 3.3. The functional relevance of age-dependent DhMRs in biological pathways. 55
Figure 3.4. Correlation between DNA hydroxymethylation and gene expression in each age group. 58
Figure 4.1. Loss of 5-hydroxymethylation is a hallmark of MBs. 70
Figure 4.2. Deregulated expression of TET proteins in MBs. 72
Figure 4.3. 5hmC gain in MBs is implicated in stem-like properties. 77
Figure 4.4. 5hmC signature of SmoA1-MBs recapitulates 5hmC signature of human MBs. 80
Figure 4.5. Elevated Tet1 is essential for MB progression. 83
Figure 4.6. TET1 inhibition confers cytotoxic effect on both murine and human MBs 85

TABLE OF TABLES

Table 1.1. Experimental methods for genome wide profiling of DNA modifications 15
Table 2.1. The 18 candidate genes in the table either suppress or enhance the neuronal toxicity, from the repeat expansion. 33
Table 2.2. Gene-based analysis of rare variants for targeted resequencing dataset, replication (WGS) dataset, and meta-analysis which combines two datasets. 36

CHAPTER 1: INTRODUCTION

1.1. Author's Contribution and Acknowledgement of Reproduction

This chapter of this dissertation includes contents from the previously published article: Kim, H., Wang X., and Jin P., Developing DNA methylation-based diagnostic biomarkers. *Journal of Genetics and Genomics*, 2018 (DOI: 10.1016/j.jgg.2018.02.003).

1.2. Challenges in rare variant association studies of complex neurological disorders

Neurological and psychiatric disorders are caused by damage to the central and peripheral nervous systems, leading to a worldwide increase in morbidity, chronic disability and mortality (Group, 2017). Despite millions affected globally, therapeutic options are highly limited mainly due to lack of our understanding of underlying genetic mechanism of pathogenesis. A number of genome-wide association (GWA) studies have contributed to the identification of risk loci and genetic variations associated with common neurological and psychiatric diseases including autism spectrum disorder (ASD), migraine, schizophrenia (SZ), epilepsy, Alzheimer's disease (AD) and Parkinson's disease (PD) (Autism Spectrum Disorders Working Group of The Psychiatric Genomics, 2017; Van Cauwenberghe et al., 2016; Chang et al., 2017; Gormley et al., 2016; Poduri, 2015; Schizophrenia Working Group of the Psychiatric Genomics, 2014). For instance, ASD is a complex developmental disorder characterized by the disability of social communication and interaction with a high incidence rate in the United States (1 out of 59 eight-year-old children, (Hall-Lande et al., 2018)). Multiple GWA studies identified the significant association of variants of the oxytocin receptor gene (OXTR) with ASD, and following functional studies demonstrated that OXTR and oxytocin (OXT; a substrate of OXTR) play a critical role in social perceptual process and the regulation of affiliative behavior (LoParo and Waldman, 2015; Ylisaukko-oja et al., 2006). In addition to the APOE locus (encoding apolipoprotein E) known for AD genetic risk, 19 novel loci associated with AD, a deleterious neuro-degenerative disorder in the elderly, were identified through meta-analysis using 4 independent GWAS data sets (Lambert et al., 2013). Even though GWA studies expand our understanding of common neurological and psychiatric disorders, this strategy was not successful in discovering rare genetic factors (<1%) associated with complex traits such as Amyotrophic Lateral Sclerosis (ALS) and Fragile X-associated tremor/ataxia

syndrome (FXTAS) since most of disease-associated rare variants exhibit modest-to-small effect size and a large sample size is mandatory for statistically reliable detection of rare genetic variants (Auer and Lettre, 2015). Thus, different types of study designs, methodologies and statistical tests are needed for rare variant association studies.

1.2.1. Family-based sequencing studies

Deep sequencing technologies open an era to identify disease-associated rare variants in complex disorders. Family-based sequencing studies, frequently referred to as 'family studies', are often used to investigate shared genetic variants of families with multiple affected members, which likely co-segregate with the disease phenotype (Auer and Lettre, 2015). This family-based study design is reminiscent of traditional linkage-based and genetic association methodologies. In the case of high frequency variants within affected families, studies based on affected relatives highly enhance the detection power compared to studies with unrelated affected individuals, and therefore, contribute to identifying complex-disease-associated rare variants of high penetrance and moderate to large effect (Genotype Relative risk (GRR) = 5~10) (Ionita-Laza and Ottman, 2011). In addition, genotype data from trios (an affected offspring and his or her parents) are often accompanied with a family-based design (Spielman et al., 1993). However, the performance of family studies is not sufficient for rare variant identification with low-to-moderate effect size, which is frequently observed in rare complex disorders (Cirulli and Goldstein, 2010).

1.2.2. Sampling of affected patients with extreme phenotypes

Even in affected individuals carrying the same known genetic mutations, disease progression can

be highly variable. Such a phenotypic variability may be due to the presence of genetic modifiers regulating expressivity and penetrance of causal genes (Cooper et al., 2013). If the disease phenotype is quantitative, it has been shown that extreme sampling strategies in affected populations can boost the statistical power to detect disease-associated rare variants (Kryukov et al., 2009; Li et al., 2011a). To do so, based on the assumption that quantitative (continuous) traits follow a normal distribution, the largest and smallest nth percentile of the distribution, typically less than the 5th percentile, are selected for the association study. However, sampling bias, particularly occurring in small-sample size studies, needs to be removed through mature statistical testing (Barnett et al., 2013); therefore, tens of thousands of samples may still be necessary for the identification of causal rare variants with low-to-modest effect size (Kryukov et al., 2009).

1.2.3. Whole-Genome Sequencing

An estimated total of 20,000-25,000 protein-coding genes in the human genome are involved in important cellular functions (International Human Genome Sequencing Consortium, 2004); thus, mutations and copy number alterations identified in the coding regions engender a deleterious consequence by disrupting protein function or dosage. For example, the most recurrent mutated gene in many types of cancer is the *TP53*, the tumor suppressor, and about 86% of mutation regions are identified between codons 125 and 300 where a DNA binding domain is located (Olivier et al., 2010). In addition, patients affected by monogenic disorders have genetic mutations of key genes proteins; for instance, mutations in Cystic fibrosis conductance transmembrane regulator (CFTR) found in Cystic fibrosis lead to a loss of the amino acid phenylalanine (F), which is critical for channel processing and gating (Choi et al., 2001). Accordingly, whole-exome sequencing (WES), an unbiased screening of de novo genetic variants at coding regions, has enriched our

understanding of causal genetic drivers implicated in numerous diseases and has advanced our treatment and management of patients (Bamshad et al., 2011; Rabbani et al., 2014). Indeed, the first study using four unrelated affected individuals showed the promise of candidate gene identification using WES (Ng et al., 2009). Since then, WES has been used as an effective tool for the investigation of genetic causality in many Mendelian disorders and complex diseases (Chong et al., 2015; Cirulli et al., 2015; Yang et al., 2014). However, the coding regions constitute about 1% of the human genome; the rest of the human genome (~99%) is not translated as proteins (e.g. non-coding regions including introns, 5' and 3' untranslated regions and intergenic regions) (Maston et al., 2006). This large portion of the genome contains functional domains regulating transcription by positioning at cis-regulatory elements (promoters, enhancers, insulators/boundary elements, and silencers) or trans-regulatory elements (distal enhancers and micro-RNA) (Davis et al., 2018; Maston et al., 2006; Plank and Dean, 2014). To do so, whole-genome sequencing (WGS) is the best option for comprehensive genetic screening across the genome (Vincent et al., 2015). Multicenter driven WGS projects such as Project MinE play a central role in understanding of disease-associated rare variants (Kenna et al., 2016; Van Rheenen et al., 2018). However, this method has had limited application in both research and clinic since it is still too expensive to cover large cohorts and necessitates sophisticated statistical tests to identify bona fide associations in a study with a large sample size at lower depth compared with a study with a small sample size at high coverage (Le and Durbin, 2011; Li et al., 2011b).

1.2.4. Targeted sequencing of prioritized candidate genes

As discussed above, WGS is not suitable for large-cohorts due to cost and difficult interpretation of data although the unbiased screening provides an opportunity to identify novel genetic factors.

Recently, multiplex targeted sequencing has offered a more accessible approach to investigate two or more candidate genomic regions in very large cohorts with high specificity and sensitivity (O'Roak et al., 2012; Wingo et al., 2017). This approach not only significantly increases detection power but allows the detection of rare and subclonal variants in heterogeneous population of cancer, which are often implicated in cancer stem cells or drug-resistant clones (Goodhead et al., 2008; Salk et al., 2018). To do so, the selection of prioritization methods is critical. There are many prioritization tools which depend on prior biological knowledge or previous reports of association with different disorders that share phenotypes with a disease of interest (Hoischen et al., 2014; Moreau and Tranchevent, 2012). For example, network analysis based on initial small-scale exome sequencing and GWA studies and data mining of recurrently mutated genes in different neurological disorders are commonly used for gene prioritization of neurological disease (Bromberg, 2013; Hoischen et al., 2014; Lee et al., 2011). In addition, web-based prioritization tools such as Suspects (Adie et al., 2006), GeneWanderer (Köhler et al., 2008), and Posmed (Yoshida et al., 2009) are freely available and user-friendly so that even biologists who have little bioinformatic knowledge can easily utilize these resources (Moreau and Tranchevent, 2012). This targeted approach also enhances the yield of downstream screens compared to the unbiased screening using WGS and WES, but causal relationships between genotype and phenotype are still determined through functional validation (Hoischen et al., 2014; Moreau and Tranchevent, 2012).

1.2.5. Rare-variant association testing for sequencing data with the Sequence Kernel Association Test (SKAT)

Numerous common genetic variants associated with disease have been identified by GWA studies, but rare variants, although they are significant in pathogenesis, cannot be detectable under GWA

setting due to rare allele frequency among affected individuals. Careful sample collection and selection of proper arrays or sequencing platforms are critical to eliminate ascertainment bias and to ensure enough coverage and accuracy for very rare variants, respectively, but sophisticated statistical testing is also essential to increase detection power of rare disease associated variants found in patients carrying heterogenous genetic background (Lee et al., 2014). Unlike common variants, collective rare variants which contribute to disruption of gene functions in various ways are more enriched in cases versus control (Bansal et al., 2010), suggesting that gene- or region-based tests enable the identification of disease-associated rare genetic components. The frequently used statistical method is a burden test which combines all identified rare variants into genetic scores (Auer and Lettre, 2015; Lee et al., 2014; Li and Leal, 2008). This method, however, is limited only for rare variants modulating phenotype in the same direction (Lee et al., 2014). A variance-component test can be used for a comprehensive rare-variant association test considering both trait-increasing and trait-decreasing rare variants (Auer and Lettre, 2015; Lee et al., 2014). In particular, the Sequence Kernel Association Test (SKAT) allows covariate-adjusted association test for both common and rare variants in a region (Wu et al., 2011c). It also computes estimated sample-size and average power which are useful for initial study design (Wu et al., 2011c). The recently introduced combined test, SKAT-O, leverages both burden and variance-component modules, which is more robust to identify variants showing their effects in both the same and different directions (Lee et al., 2012).

1.2.6. Functional annotation of rare variants

Regardless of the research platform, functional annotation of variants after the discovery phase gives a clue whether the variants are associated with pathogenesis. Public databases of functional

annotation including Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) are available to understand previously investigated biological functions of genes (Ashburner et al., 2000; Ogata et al., 1999; The Gene Ontology Consortium, 2019). In addition, phylogeny-based protein function prediction methods such as statistical inference of function through evolutionary relationships (SIFTER) and annotation of clinical significance using ClinVar are widely used to assess the disease relationship of variants (Engelhardt et al., 2011; Landrum et al., 2016; Sahraeian et al., 2015). However, it is challenging to obtain integrated insight from dispersed information stored in different databases. A recently introduced bioinformatic tool, Combined Annotation Dependent Depletion (CADD), quantitates potential deleterious effects of variants, which shows a better ability to predict pathogenicity of novel variants because it uses a strong correlation with allelic diversity, experimentally validated regulatory effects of noncoding regions, and known disease associated variants (Kircher et al., 2014a; Rentzsch et al., 2019). In addition, instead of manual individual annotation via web browsers, a rapid and integrated online annotation method, Bystro (https://bystro.io/), is more useful to investigate a large number of variants at once and filter out unnecessary variants based on customized criteria with natural language (Kotlar et al., 2018).

1.2.7. Functional genomics using a *Drosophila* model

Sequencing and bioinformatic tools are major resources to identify disease-associated genes, but the tools are not a definite indicator of causality. To uncover biological significance of genes in disease or certain biological conditions, genome-wide functional screening known as functional genomics is necessary. High-throughput screening is widely used, which is based on altering gene expression using siRNA, shRNA, microRNA, and CRISPR-Cas9 gRNA libraries in cultured cell

lines, and then assesses the functional importance of genes, for example, by measuring cell viability and proliferation upon knockdown of genes (Echeverri and Perrimon, 2006; Lemons et al., 2013; Shalem et al., 2015; Sims et al., 2011). Mammalian cell lines are easy to maintain and recapitulate biological process in the organisms; however, the assessment of phenotypes expressed via complex biological network is difficult in cell-based assays although there are attempts to overcome such limitations by co-culturing different type of cells (Goers et al., 2014). In this sense, *Drosophila melanogaster*, a fruit fly, is a powerful model organism for genome-wide genetic interactions *in vivo* (Pandey and Nichols, 2011). Compared to rodent models, with *Drosophila* it is relatively easy to expand a large number of progeny within a short period of time that requires a lower infrastructure cost (Pandey and Nichols, 2011). In addition, the genome of *Drosophila* was fully sequenced in 2000 (Adams et al., 2000), and with the completion of the Human Genome Project in 2003 and functional annotation of the human genome, substantial homologies between the two genomes were identified (Pandey and Nichols, 2011). Furthermore, online resources enable researchers to search for ortholog of human genes (DIOPT), and Transgenic RNAi (TRiP) lines facilitate rapid and high throughput functional genomic screening using *Drosophila* (Ni et al., 2011). Although *Drosophila* is limited for a comprehensive understanding of human-specific genes, it is an important model system for large-scale screening combined with functional validation using cell lines and mouse models.

1.3. The significance of epigenetics in cellular functions and diseases

Inter-individual phenotypic diversity is not sufficiently explained by the approximately 0.1 % of genetic variation between individuals (Altshuler et al., 2015). In addition, obtaining the sequence of the human genome could not account for how cells in multicellular organisms, sharing the same genetic code, exhibit unique gene expression for their cellular functions within different tissues (Waddington, 2012). Epigenetics, first introduced by C. H. Waddington in 1939, was proposed as an additional layer of gene regulation in the limited context of primary DNA sequence differences (Bernstein et al., 2010; Jenuwein and Allis, 2001; Waddington, 2012). Epigenetic modifications including DNA methylation, histone modifications, and chromatin remodeling, can contribute to cell-type specific gene expression signatures necessary for cellular function (Doi et al., 2009; Mack et al., 2016; Nestor et al., 2012). These epigenetic modifications are heritable but reversible and dynamic, thereby not only establishing specific cellular states but also being able to respond to changes in the microenvironment, which confers cellular plasticity (Mack et al., 2016). In addition, epigenetic dysregulation could contribute to the development and progression of many diseases (Esteller, 2008; Hwang et al., 2017). For example, hypermethylation at CpG island promoters of non-mutated tumor suppressor genes is recurrently identified in pediatric and adult brain tumors, conferring proliferative advantages and aggressive phenotypes during tumorigenesis (Mack et al., 2016; Suva et al., 2013). Abnormal epigenetic programs are also strongly associated with neurodegeneration by modifying disease risk, age of manifestation, and progression (Hwang et al., 2017; Qureshi and Mehler, 2013). In addition to their biological role as pathogenic factors, epigenetic marks associated with specific diseases are considered as emerging biomarkers for diagnosis and predictors of treatment response and prognosis in many diseases (Paluszczak and Baer-Dubowska, 2006a; Qureshi and Mehler, 2013). Intriguingly, it has been shown that the

epigenetic changes can also be detected in different biologic fluids, such as blood, urine, and fecal samples (Diaz-Lagares et al., 2016; Haggarty, 2015; Jakubowski and Labrie, 2017; Van Neste et al., 2012). Among epigenetic modification assays, the DNA methylation-based assay was the first FDA approved screening test for colorectal cancer (Song et al., 2017b), suggesting that methylation analysis can be implemented in clinics as a screening test of different disease types. However, an understanding of dynamic epigenetics in normal brain development and diseases largely remains to be elucidated.

1.3.1. DNA modifications in genome: cytosine modifications and beyond

Modified DNA bases are essential for epigenetic gene regulation. The most abundant DNA modification is the addition of a methyl group to the 5′ position of the cytosine pyrimidine ring (5-methylcytosin, 5mC). This direct chemical modification to the DNA is conserved throughout evolution and plays a critical role in various cellular processes. Hyper-methylation at promoters, for instance, suppresses gene expression by either inhibiting the binding of transcription factors or by recruiting complex proteins known as methyl-CpG-binding domain proteins (MBDs) (Robertson, 2005a; Schubeler, 2015; Yao et al., 2016). 5mC is also involved in the repression of transposable elements, contributing to genome integrity (Slotkin and Martienssen, 2007). In addition to the identification of biological functions of 5mC, the discovery of DNA methyltransferases (DNMTs) including DNMT1, DNMT3A, and DNMT3B provides the mechanism for maintaining or generating 5mC in the genome (Okano et al., 1999; Schubeler, 2015; Yoder and Bestor, 1998). Initially, methylation was assumed to be a permanent modification due to the chemical stability of the methyl group and a lack of detection of demethylase and other modifications. Two independent studies, however, dramatically changed the understanding of the

dynamic regulation of 5mC by identifying ten-eleven translocation (TET) proteins that can oxidize 5mC to 5-hydroxymethylcytosine (5hmC) and generate further oxidative derivatives 5-formyl-cytosine (5fC) and 5-carboxylcytosine (5caC), which become converted into unmodified cytosine by thymine DNA glycosylase (TDG) mediated base excision DNA repair (BER) (Ito et al., 2011; Plongthongkum et al., 2014; Schubeler, 2015; Yao et al., 2016) (Figure 1.1). The oxidative modifications (5hmC, 5fC and 5caC) cannot be maintained by DNMT1 and are demethylated by either passive dilution or active demethylation pathways (He et al., 2011; Maiti and Drohat, 2011). In addition, the global abundance of the derivatives is much less than 5mC so that the 5mC oxidized derivatives were considered as intermediates generated during the demethylation process. However, further investigation demonstrated the independent roles in transcriptional regulation during embryogenesis and neurodevelopment beyond the demethylation process (Song and He, 2013). Interestingly, the presence of another DNA methylation, adenine methylation (N6-methyladenine, 6mA), was recently identified in mammals, though the abundance is lower than that observed in prokaryotes (Heyn and Esteller, 2015). While host defense is the main role of 6mA in prokaryotic systems, 6mA is deemed as a suppressive mark in eukaryotes based on its significant enrichment at transposable elements (Heyn and Esteller, 2015). Recent studies support the role of 6mA in regulating neuronal gene expression (Yao et al., 2017); however, detailed functional investigation of 6mA's role in transcriptional regulation is needed. Identification of methylation and demethylation enzymes involved in adenine methylation will be critical.

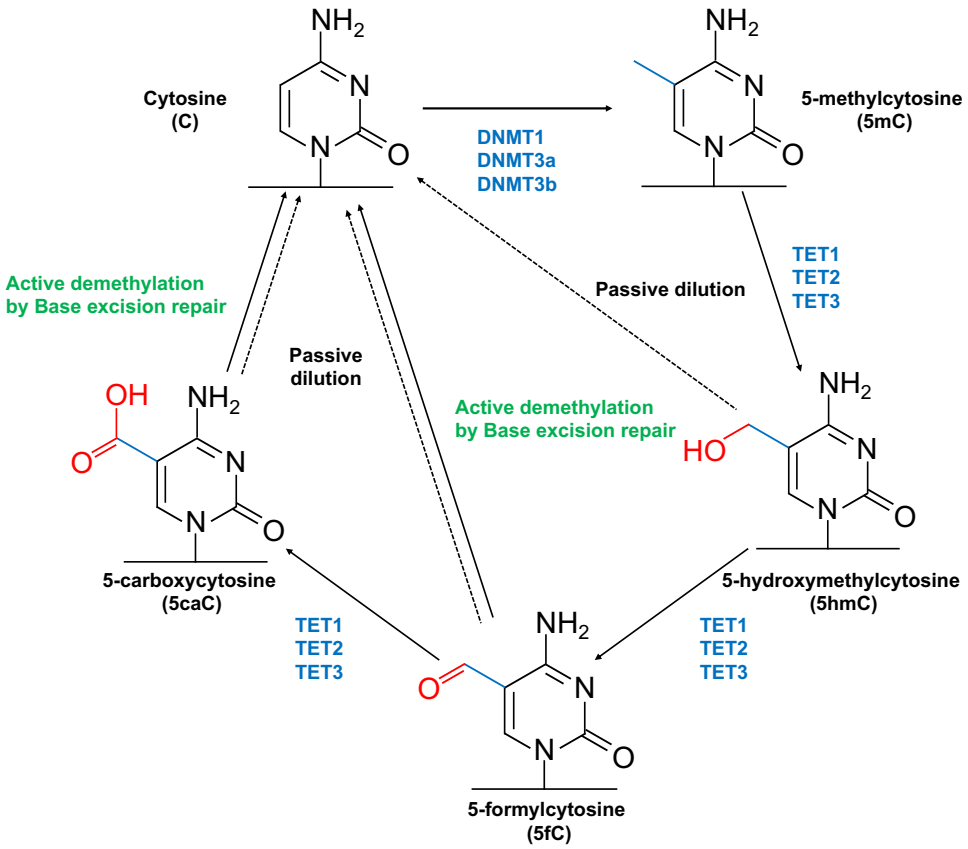


Figure 1.1. The cycle of cytosine modifications. The fifth position of cytosine can be methylated by DNA methyltransferases (DNMT1, DNMT3A, and DNMT3B) to generate 5-methylcytosine (5mC). The methyl group of 5mC can be oxidized by ten-eleven translocation (TET) family enzymes (TET1, TET2, and TET3), generating 5-hydroxymethylcytosine (5hmC), 5-formlycytosine (5fC), and 5-carboxycytosine (5caC). While 5mC is maintained by the interaction between the replication machinery and DNMT1, no maintenance mechanisms exist for the oxidative derivatives; therefore, the levels of 5hmC, 5fC and 5caC are diminished over replication (passive dilution). In addition, 5fC and 5caC can be excised by thymine DNA glycosylase (TDG), and eventually replaced with cytosine (active demethylation by base excision repair (BER)). Both passive and active demethylation mechanisms contribute to dynamics of cytosine modification.

1.3.2. Technologies for genome-wide DNA modifications

The advancement of sequencing technology (Table 1.1) not only enables us to increase our knowledge of epigenomes and identify many disease-associated DNA modifications but also benefits healthcare in terms of disease diagnosis, precise classification, and prognosis (Fernandez et al., 2012; Hwang et al., 2017; Moran et al., 2016; Paluszczak and Baer-Dubowska, 2006b). Indeed, multicenter consortiums including the Encyclopedia of DNA Elements (ENCODE) project and the Roadmap Epigenomics project have uncovered regulatory functions of DNA methylation, histone modifications and chromatin remodelers in various types of cells and *ex vivo* tissues (Bernstein et al., 2010; Dunham et al., 2012). Moreover, sequencing methods to map 5mC oxidized derivatives (5hmC, 5fC, and 5caC) have been developed, allowing a comprehensive understanding of interaction and dynamics of DNA modifications (Booth et al., 2013; Song et al., 2011b; Wu et al., 2016; Yu et al., 2012).

Formalin-fixed and paraffin-embedding is a standard clinical method for long-term storage, leading to the fragmentation of DNA and cross-linking to other macromolecules (e.g. proteins) which significantly affect the yield of successful mapping of epigenetic modifications. Many clinical studies have exhibited that methylation patterns are reliably reproduced in both fresh tissues and formalin-fixed, paraffin-embedded (FFPE) human tissues, demonstrating a powerful tool of DNA methylation as clinical application (Bock et al., 2016). In addition, tissue-specific DNA methylation patterns enable the quantitative analysis of tissue-specific cell-free circulating DNA (cfDNA) in plasma, suggesting its potential use as a disease biomarker for diagnosis and prognosis (Kim et al., 2018).

Method	Description
Whole-genome bisulfite sequencing (WGBS)	WGBS (aslo known as MethylC-seq) is a sequencing-based approach to profile DNA methylomes in an unbiased manner (Frommer et al., 1992). This approach utilizes the selective chemical conversion of unmethylated cytosines by treating with bisulfite, allowing the detection of genome-wide differential DNA methylation regions (DMRs). WGBS is widely used to identify novel DMRs under disease status (Heyn and Esteller, 2012).
Methylation DNA immunoprecipitation sequencing (MeDIP-seq)	MeDIP-seq utilizes the 5mC-specific antibody to enrich genome-wide DNA methylation regions (Heyn and Esteller, 2012; Jacinto et al., 2008). Despite containing less information than WGBS, MeDIP-seq can cover large genomic regions with a substantial resolution and low-cost. In discovery phase, MeDIP-seq is a suitable alternative to detect disease-specific hypermethylated regions in a quantitative manner (Heyn and Esteller, 2012).
5-hmC selective chemical labeling (hMe-Seal)	Selective chemical labeling of 5hmC (hMe-seal) is based on selective enzymatic reaction to add a modified glucose moiety to the hydroxyl-group of 5hmC (Song et al., 2011c). The modified glucose is further labeled with biotin via click reaction, which is an efficient and reliable approach with high yield (Hein et al., 2008).
Tet-assisted bisulfite sequencing (TAB-seq) and Oxidative bisulfite sequencing (OxBS-seq)	TAB-seq and OxBS-seq are sequencing tools to map genome-wide 5hmC distribution at base resolution by coupling to bisulfite (Booth et al., 2013; Yu et al., 2012). WGBS approach cannot distinguish 5mC from 5hmC since both 5mC and 5hmC are resistant to bisulfite treatment. Therefore, either a chemical protection approach by selective glycosylation of 5hmC (TAB-seq) or specific oxidation of 5hmC to 5fC (OxBS-seq) allow precise differentiation of 5hmC from 5mC.
Methylase-assisted bisulfite sequencing (MAB-seq) and 5caC methylase-assisted bisulfite sequencing (caMAB-seq)	MAB-seq and caMAB-seq use chemical reaction of bisulfite treatment and sodium borohydride ($NaBH_4$) treatment, which induces reduction of 5fC to 5hmC. Combination of chemical treatment with the treatment of the bacterial DNA CpG methyltransferase M.SssI, an enzyme to methylate cytosines within CpG dinucleotides, allows base-resolution mapping of 5fC and 5caC (Wu et al., 2016).

Table 1.1. Experimental methods for genome wide profiling of DNA modifications

1.3.3. Distinct genomic localization of TET proteins: intrinsic structural difference and the interaction with extrinsic factors

The conversion of 5mC to 5hmC is mediated by TET family proteins in a Fe(II)/α-ketoglutarate (α-KG)-dependent manner (Tahiliani et al., 2009). Before the demonstration of TET1-mediated enzymatic oxidation in 2009, the initial discovery of *TET1* gene was the study to identify a fusion partner of *MLL* in acute myeloid leukemia (AML) (Lorsback et al., 2003). Subsequent studies of the other two family proteins, TET2 and TET3, showed that all three TET family proteins share the core catalytic domain at the C-terminus to convert 5mC to 5hmC, which includes a double-strand β-helix (DSBH) domain and a cysteine-rich domain (Ito et al., 2010). TET1 and TET3 have a CXXC domain at their N-terminus, which provides a preferential binding of those proteins to non-methylated CpG-rich regions while TET2 lacks a CXXC interaction motif (Deaton and Bird, 2011). The enrichment of Tet1 at CpG islands, active promoters, and bivalent promoters (marked by both H3 lysine 4 tri-methylation (H3K4me3) and H3K27me3) was identified in mouse embryonic stem cells (ESCs) (Williams et al., 2011; Wu et al., 2011a; Xu et al., 2011b); on the other hand, the loss of TET2 activity yields abnormal methylation at enhancer regions in hematopoietic cells (Rasmussen et al., 2015), suggesting that the different genomic occupancy and the regulation of 5hmC levels at different genomic regions is partially explained by the presence of the CXXC domain (Wu and Zhang, 2017).

However, the genomic localization of TET proteins to the corresponding genomic regions is not fully determined by their intrinsic structural properties, but other factors as well as the local chromatin environment modulate the interaction of TET proteins and specific genomic regions (Wu and Zhang, 2017). For example, the protein-protein interaction between stem cell transcription factor Nanog and either Tet1 or Tet2 facilitates the recruitment of Tet1 and Tet2 to Nanog target regions (Costa et al., 2013). In addition, Tet1 exerts dual regulatory functions in mouse ESCs by mediating demethylation at

active promoters and bivalent promoters (Wu et al., 2011a) and different factors are responsible for the recruitment of Tet 1 at different promoters. Lin28A, a well-known RNA-binding protein, directly binds to Tet1, leading to active gene expression through dynamics of DNA modifications (Zeng et al., 2016). On the other hand, Polycomb repressive complex 2 (PRC2) is responsible for the recruitment of Tet1 at H3K27me3 positive regions of the genome (Neri et al., 2013, 2015). Likewise, genomic binding of TET2 is regulated by transcription factors such as Wilms tumor 1 (WT1), a key transcription factor regulating hematopoiesis, and peroxisome proliferator-activated receptor-γ (PPARγ), a nuclear receptor controlling fatty acid storage and glucose metabolism mediates (Wang et al., 2015; Yoo et al., 2017), and RE1-silencing transcription factor (REST) mediates the recruitment of TET3 at its target regions (Perera et al., 2015). These context-dependent recruitments of TET proteins mediated by distinct transcription factors orchestrate dynamics of cytosine modifications at proper target regions.

1.3.4. Significance of epigenetic alterations in pediatric cancer

Medulloblastoma (MB) is the most common pediatric brain tumor arising in the cerebellum, a brain region responsible for the maintenance of balance and posture as well as cognitive function (Massimino et al., 2011; Northcott et al., 2012). More than 80 percent of MBs are diagnosed before age 15, and the incidence among adults (patients >16 years of age) is much rarer (Massimino et al., 2011). Standard treatment based on surgery followed by radiation and adjuvant chemotherapy improves the 5-year survival of many patients, but the treatment-induced toxicity frequently leads to adverse effects such as hormone imbalance and deficits in learning and memory (Packer et al., 2013). Such permanent brain damage severely affects the quality of life for survivors; hence it is necessary to develop targeted therapeutic agents with few side effects. Lack of definitive disease-risk stratification in clinical diagnosis is another clinical challenge. In most cases, prognosis is assessed by clinical/pathological variables

including age at diagnosis, the amount of remaining tumor after surgery and metastases, where patients diagnosed at age of less than 3 years and residual tumor larger than 1.5 cc after surgery are classified into a high-risk group (Thompson et al., 2016). However, the 5-year survival rate for the high-risk group, albeit worse compared to an average-risk group, varies from 30% to 65%. Misclassified patients may lead to overtreatment causing adverse health and economic outcomes (Thompson et al., 2016). As such, the need for targeted agents selective for tumor cells and molecular prognostic/predictive biomarkers drives researchers to conduct genome-wide studies.

Overall mutation rate of MBs is lower than that of adult tumors, consistent with other pediatric malignancies (Greenman et al., 2007), but mutations in epigenetic regulators such as *SMARCA4*, *MLL2*, *BCOR*, and *KDM6A* are frequently observed (Pugh et al., 2012a), suggesting that epigenetic alterations play a substantial role in MB progression. Genome-scale analysis of changes in DNA methylation (Hovestadt et al., 2014a) and the identification of active- and super- enhancers by H3K27ac ChIP-seq (Lin et al., 2016) and their regulatory role in gene expression signatures of MB subgroups have demonstrated the influence of aberrant epigenome on differential transcription in MB subgroups. However, the function of another abundant cytosine modification in the cerebellum, 5hmC has not been elucidated in MB despite the key role of 5hmC and TET proteins in cerebellar development.

1.4. Summary of background information and dissertation goals

Exploring disease etiology starts from the identification of genetic and epigenetic alterations. Technical advances mainly based on next-generation sequencing (NGS) enables a discovery of disease-associated genetic/epigenetic factors by extensive screening at a genome-wide scale. Along with the completion of the Human Genome Project, NGS led to a huge surge of growth rate in the dbSNP catalog. However,

variants associated with orphan diseases and dynamic epigenetic modifications during aging and disease progression still remain to be elucidated. The primary objective of the studies in this thesis is 1) to evaluate the validity of our novel approach to identify genetic factors involved in rare neurological disorder. 2) to understand abnormal epigenetic programs in pediatric brain tumor based on knowledge of age-related epigenetic progression.

This thesis consists of five chapters. In Chapter 2, I present data of a step-wise approach to identify a novel genetic factor, *MYH15* that increases Amyotrophic lateral sclerosis (ALS) risk. The aim was to introduce a methodology for increasing detection power of genetic etiology of ALS with limited sample sizes by using both sequencing approaches (WGS and targeted resequencing) and *Drosophila* based functional genomic analysis screening. In Chapter 3, I present data of genome-wide 5hmC profiles and transcription profiles in human cerebellar tissues categorized into two different groups depending on ages: Young-age cerebellum (YCB) and Old-age cerebellum (OCB). The aim was to better understand epigenetic dynamics during aging in human and its relationship with gene expression. In Chapter 4, I present data of MB-specific 5hmC signature and its implication in tumorigenesis. In addition to 5hmC profiles, I present data to demonstrate TET1 as a putative tumor promoter and a therapeutic target in MBs. The aim was to identify the functional roles of 5hmC and TET enzymes, responsible for 5hmC generation in MBs. Collectively, these studies broaden our perspective of an effective approach to identify genetic and epigenetic alterations by coupling advanced technologies with biological rationale.

CHAPTER 2: Rare Variants in MYH15 Modify Amyotrophic Lateral Sclerosis Risk

2.1. Author's Contribution and Acknowledgement of Reproduction

This chapter is reproduced with minor edits from the previously published article: Kim, H., et al., Rare Variants in MYH15 Modify Amyotrophic Lateral Sclerosis Risk. *Human Molecular Genetics*, 2019 (DOI: 10.1093/hmg/ddz063). T.S.W. and P.J. conceived the study. H.K., T.S.W. and P.J. wrote the manuscript. All of the authors read and commented on the manuscript. S.M.C., Y.L. extracted DNA from samples used in this study. T.S.W. performed WGS data analysis and H.K. selected genes based on variants identified in the WGS. J.L. crossed G_4C_2 repeat stable line and RNAi lines and J.L., HK., H.B. and B.J. observed eye phenotypes for genetic screening. J.L. performed thin-section analysis of adult *Drosophila* eye and SEM imaging. K.H.M. provided expert interpretation of fly imaging data. H.K. performed targeted resequencing and statistical testing of targeted resequencing and additional whole-genome sequencing dataset using SKAT/MetaSKAT package. M.P.E. provided expert statistical advice. J.E.L., C.F, and J.D.G. provided expert clinical interpretation and details of the phenotype for affected individuals. J.J. and J.P. provided dipeptide-repeat (DPR) constructs and H.K. and K.X. performed toxicity assay using the constructs.

2.2. Introduction

Amyotrophic Lateral Sclerosis (ALS) is a complex neurodegenerative disease that can develop at any age, but most commonly occurs between the ages of 40 and 70 years (at a mean age of 55 years) (Taylor et al., 2016). This rare neurological disorder is characterized by progressive degeneration of the upper and lower motor neurons and leads to weakness and death an average of 2 to 5 years after initial clinical symptoms develop (Al-Chalabi et al., 2017; Robberecht and Philips, 2013). Approximately 5-20% of ALS patients exhibit a discernible family history defined as familial ALS (fALS) (Cirulli et al., 2015). Genetic factors are considered as obvious drivers for the pathogenesis in fALS cases (Byrne et al., 2013), but a number of twin and other large-scale genomic studies have also shown a substantial genetic contribution in sporadic ALS (sALS), estimating approximately 60% of the heritability of sALS (Al-Chalabi et al., 2010; McLaughlin et al., 2015; Wingo et al., 2011). Given this, many genetic studies have been conducted to understand the genetic etiology of ALS and have identified rare genetic variants in multiple genes such as SOD1, FUS, and TARDBP (Al-Chalabi et al., 2017; Geevasinga et al., 2016).

Among the known pathogenic mutations, the recently identified hexanucleotide (G_4C_2) repeat expansion in the C9orf72 gene is the most common genetic cause of ALS (C9ALS), although this mutation has an intermediate effect on ALS risk compared to traditional pathologic mutations (Al-Chalabi et al., 2017; Haeusler et al., 2016). C9ALS has a wide range of phenotypic variability in terms of age-at-onset, duration and regions of motor neuron involvement (Renton et al., 2014; Umoh et al., 2016), suggesting the burden of genetic variants in multiple genes may contribute to modulating ALS risk even in the patients sharing the same genetic alteration (Al-Chalabi et al., 2017; Chi et al., 2016; Pang et al., 2017). However, the successful identification of novel genetic components involved in ALS pathogenesis is limited by only using genome-wide association study (GWAS) or whole-genome sequencing (WGS) if

pathogenic mutations have a moderate or small effect on ALS risk (Al-Chalabi et al., 2017).

In this study, we hypothesized that a gene differentially identified among C9ALS groups who have extremely distinct age-at-onset can be a novel genetic factor implicated in ALS. Using functional screening, we were able to prioritize candidate genes more biologically relevant to ALS causality. Here we performed a hypothesis-driven genetic association study using WGS to identify novel genetic candidates associated with ALS risk (step 1), followed by a genetic screen using a *Drosophila* model stably expressing the G_4C_2 repeat expansion (step 2). Prioritized candidate genes were further assessed by a candidate gene association study using sALS cases and non-ALS controls (step 3), consequently leading to the identification of rare variants in *MYH15* as a novel genetic factor of ALS. Furthermore, we show that *MYH15* could modulate the toxicity of dipeptides produced from the expanded G_4C_2 repeat. Our data together demonstrate the utility of combining WGS with fly genetics to facilitate the discovery of fundamental genetic components of complex traits with a limited number of samples.

2.3. Materials and Methods

Study Subjects

In the discovery phase, we performed WGS on DNA samples from four unrelated patients carrying the G4C2 expansion mutation. We included two patients with an early age of onset (31.3 and 41.7 years old; young ALS [YALS]) and two with late age of onset (72.4 and 72.9 years old; old ALS [OALS]). All patients in this phase are unrelated to each other (Table S1). In the replication phase, 576 samples, including 310 sALS patients and 266 unaffected individuals, were used for targeted resequencing (Table S5). Validation of candidate genes using an independent WGS dataset was done with 170 sALS patients and 42 non-ALS controls. The protocols and consent forms for enrollment were approved by the Institutional Review Board at Emory University. Written and informed consent were obtained for all participants.

Genotyping G_4C_2 Repeat Size of study subjects

Genomic DNA from human white blood cells was extracted with the Gentra Puregene kit (Qiagen) according to the manufacturer's protocols. The C9orf72 hexanucleotide repeat for study subjects was determined using the repeat-primed protocol, as described previously (Umoh et al., 2016). Briefly, 4 primers (two forward primers, one reverse primer, and a fluorescently labeled primer) were used for PCR amplification of DNA. Amplified products incorporating a fluorescently labeled primer were separated using a capillary electrophoresis DNA system (ABI3730; Thermo Fisher, Waltham, MA). Based on a cutoff of 30 repeats as a positive indicator and DNA from a C9Pos control from Coriell Institute for Medical Research (6769B1), the status of C9Pos for each sample was determined using amplified fragment length polymorphism analysis in GeneMarker software (Softgenetics, State College, PA).

Whole genome sequencing (WGS) and candidate gene identification

The Hudson Alpha Institute for Biotechnology provided sequencing services (Huntsville, AL, USA). Raw sequencing data were aligned to the hg38 build of the human genome using PEMapper, and variants called using PECaller with default settings (Johnston et al., 2017). Variant annotation and summary sequencing statistics were performed using Bystro (Shetty et al., 2010). To identify candidate ALS phenotypic modifying genes, rare genetic variants (MAF < 0.01) commonly found in either YALS or OALS were considered along with a Combined Annotation Dependent Depletion (CADD) phred-scaled score above 10 (Rentzsch et al., 2019). We selected 89 candidate modifiers (67 genes from YALS and 22 genes from OALS), which have *Drosophila* orthologues searched by *Drosophila* RNAi Screening Center (DRSC) Integrative Ortholog Prediction Tool (DIOPT) (Hu et al., 2011), to determine the genetic interaction between G_4C_2 repeat-associated ALS and genes with rare, potentially damaging variants in both groups of patients (Table S3).

Genetic screen using fly model

The G_4C_2 repeat stable line was established by crossing a *GMR-Gal4* driver with a UAS-$(G_4C_2)_{30}$ repeat transgene (Xu et al., 2013). The RNAi lines were obtained from either the Bloomington Stock Center or the Vienna *Drosophila* RNAi Center (Table S4). The knockdown efficiency of RNAi lines crossed with the *ELAV-Gal4* driver was measured by quantitative RT–PCR (qPCR) (Figure 4A, Right). To determine the genetic interaction between G_4C_2 repeat and candidate genes, eye phenotypes of RNAi lines mated with the G_4C_2 repeat stable line were compared with the eye phenotype of the G_4C_2 repeat stable line, and images were obtained by light microscopy. Eye phenotypes in the figure are representative images of functional screening. All crosses were conducted at 25ºC, replicated three times to validate the specific phenotype, and a minimum of 10 flies were used to determine phenotypic change. Scanning electron

microscopy (SEM) images of whole flies were obtained after dehydrating them in an ethanol gradient (25%, 50%, 75%, and 100%) followed by incubation with hexamethyldisilazane for 1 hour (Electron Microscopy Sciences). After the removal of all chemicals by drying overnight in the fume hood, the flies were coated with argon gas under an electric field and analyzed with a Topcon DS-130F and DS-150F Field Emission Scanning Electron Microscope. For further morphological analysis to confirm the recovery of organized ommatidia in the case of RNAi showing suppressed toxicity when crossing with the G_4C_2 repeat stable line, thin-section analysis of adult *Drosophila* eyes was conducted according to standard protocols (Moberg et al., 2001). In brief, fly heads were exposed to 2% glutaraldehyde in 0.1M PO_4 on ice followed by 2% OsO_4 in 0.1M PO_4 on ice. After dehydration in an ethanol gradient (30%, 50%, 70%, 80%, 90%, 100%), 100% ethanol was replaced with propylene oxide, and an equal volume of resin was added. The fixed heads were transferred to a silicone rubber flat mold for embedding with resin. One μm sections were mounted on glass slides and stained with toluidine blue.

Targeted resequencing

Genomic DNA was extracted from white blood cells of 310 ALS patients and 266 non-ALS subjects using the Gentra Puregene kit (Qiagen) according to the manufacturer's protocols. For targeted resequencing, two sets of primers were designed by using the MPD (multiplex primer design) software with >90% coverage for each gene (Wingo et al., 2017) and optimal multiplex design for the Access Array System (Fluidigm). The first set was designed to capture 14 candidate genes including *DLG2*, *MYH15*, *KIF27*, and *ABCC2*, and 5 known ALS genes (*GRN*, *SOD1*, *FUS*, *TARDBP*, and *TBK1*) (Table 1). The second set covered 400 ancestrally informative and 25 common X chromosome markers. The samples were randomly plated concerning affectation, sex, and age to minimize batch effects. Sequence capture was performed using the Access Array with 48 samples per batch according to the manufacturer's protocol. All samples were barcoded according to the manufacturer's protocol and 250bp

paired-ended sequencing was performed on an Illumina MiSeq.

Base calling and quality control

Mapping of raw targeted resequencing reads to the hg38 of the human genome was performed with PEMapper followed by variant calling using PECaller with default values (Johnston et al., 2017). Variants were annotated and summarized using Bystro (Shetty et al., 2010). As a quality control, samples with apparently different ethnicity according to demographic information were removed (n=18) (Figure S1A). Using unlinked ancestrally informative markers for principal-component analysis (PCA) with EIGENSOFT, we excluded samples whose eigenvectors were >6 SD away from the mean (n=25) (Figures S1A and S1B) (44). Samples within batches having amplicons with > 3 standard deviations (SD) missing sites and batches with > 3SD sample failure were eliminated from further analysis. Moreover, samples with > 3SD excess heterozygosity or genotype rate less than 95% were further dropped (n=44) (Figure S1A). Two samples with known ALS associated mutations in *TARDBP* and *SOD1* were excluded from further analysis as well. In total, 270 ALS and 217 non-ALS samples were used for further statistical analysis. Variants that failed Hardy-Weinberg filtering at 10^{-7} and > 1% of minor allele frequency (MAF) were excluded.

Genotype identification of target genes from replication dataset

The WGS replication dataset is based on whole genome sequencing with approximately 100 maxdepth of coverage for each chromosome and mapped to hg19 of the human genome. Individual 212 vcf files were obtained and combined using bcftools with -0 flag. Variants of merged vcf files were intersected using intersectBed, with genomic regions of interest converted from the hg38 to the hg19 using LiftOver to obtain variants of targeted genes and PCA markers (Kent et al., 2002). A total of 571 variants for PCA markers and 1294 variants for targeted genes were identified resulting in 208 samples for analysis.

Through PCA, samples of outliers (n=2) with known ALS associated mutations in *SOD1* were excluded from the further analysis (Figures S1C and S1D).

Statistical Analysis

We performed gene-based testing of rare variants in the targeted resequencing and the whole genome sequencing datasets using sequence kernel association test (SKAT) and optimized SKAT (SKAT-O) implemented in the R package SKAT v1.2.1. We adjusted for population stratification by incorporating the top 2 eigenvectors from PCA as covariates within the analysis. For multi-allelic sites, the two minor alleles were combined to convert the site to a bi-allelic site using a custom R script (n=9) prior to analysis. Since our genetic interaction screening with a *Drosophila* model is based on the interaction of genes, not regulatory regions, we focused our analysis on those that alter coding sequence including missense or nonsense changes. In addition, we employed a Combined Annotation Dependent Depletion (CADD) phred-scaled score above 20 (Rentzsch et al., 2019), which is a more rigorous way to identify genetic variants leading to protein changes. We derived p-values for SKAT/SKAT-O adjusted for a sample size of less than 2000 and binary traits with asymptotic and efficient resampling methods (Lee et al., 2012; Wu et al., 2011c). In the targeted resequencing project, we used an unadjusted Type-I error rate of 0.05 to identify genes with suggestive evidence of association with ALS risk. For those genes passing this suggestive threshold, we interrogated replication using Emory ALS WGS dataset and identified those genes significantly associated with ALS risk using a Type-I error rate adjusted for multiple testing based on a Bonferroni correction.

For Meta-analysis for the gene-based association test, we used MetaSKAT (Lee et al., 2013), v0.6.0, with individual level genotype data of targeted resequencing and Emory ALS WGS dataset. We adjusted for the top 2 eigenvectors of PCA within each dataset. The genomic coordinates of the Emory ALS WGS

dataset was converted from hg19 to hg38 using LiftOver to unify genotype assembly (Kent et al., 2002).

Toxicity assay using poly-dipeptides constructs

To determine the role of *Myh15* in mammalian system, we generated the constructs expressing 50 repeats of either PR or GA peptides (Zhang et al., 2016). CellTiter-Blue Cell viability Assay (Promega) was used to assess cell viability on 3 days after transfection with poly-dipeptide constructs (150ng) and siRNA (50nM). Briefly, 20 µl of solution was added to each well directly 1 hr before measurement. The fluorescence was measured using FLUOstar Omega (BMG Labtech) microplate reader. All measurements were taken in triplicate and each experiment was replicated at least three times.

Data availability

Targeted resequencing data from this study have been deposited in the NCBI Sequence Read Archive (SRA) under accession number SRP136672. Whole genome sequencing data and supporting data are available on request from the corresponding author.

2.4. Results

Candidate gene discovery through whole-genome sequencing

To facilitate the identification of novel genetic factors of ALS risk with a limited sample size, we employed a stepwise approach of candidate gene selection based on the assumption that genetic risk factors can be identified in even a small number of ALS patients who have the same G_4C_2 repeat expansion but develop clinical symptoms at different ages (Figure 2.1). Therefore, in the discovery phase, we performed WGS on two distinct age-at-onset groups of four unrelated G_4C_2 repeat expansion carriers. Two of the individuals developed ALS at 31 and 41 years old (and referred to here as young ALS [YALS]), and the other two individuals developed ALS at 72 years old (and referred to as old ALS [OALS]) (Table S1). To identify disease-relevant variants, we selected rare and deleterious sites that were unique to either the YALS or OALS groups on the basis of the following criteria: variants that had a minor allele frequency (MAF) <1% in the Genome Aggregation Database (gnomAD) (Lek et al., 2016) and variants that had a CADD (Rentzsch et al., 2019) score higher than 10 (Johnston et al., 2017; Kircher et al., 2014b). In total, we identified 190 variants (159 variants from YALS, 31 variants from OALS) and 135 unique genes (105 genes from YALS and 30 genes from OALS) (Table S2).

<table>
<tr><td rowspan="2">Step 1</td><td rowspan="2">Whole Genome Sequencing</td><td>Samples
• c9ALS patients with Early onset (< 45 years old)
• c9ALS patients with Late onset (> 70 years old)</td></tr>
<tr><td>Criteria used in analysis
• Rare variants (MAF < 0.01)
• Non-synonymous or variants near genes
• High Cadd score (Cadd > 10)
• Drosophila Ortholog Prediction</td></tr>
<tr><td colspan="3">49 age of onset group-specific genes
(42 YALS genes, 7 OALS genes)</td></tr>
<tr><td rowspan="3">Step 2</td><td rowspan="2">Functional screening</td><td>Models
• $(G_4C_2)_{30}$ transgenic line
• 90 RNAi lines corresponding to 49 fly genes</td></tr>
<tr><td>Criteria used in analysis
• Eye morphology
• Degrees of cell death
• Ommatidial disruption
• Known for ALS association and neurological disorder</td></tr>
<tr><td colspan="3">14 G_4C_2 toxicity-modifying genes
(7 Suppressed toxicity genes, 7 Enhanced toxicity genes)</td></tr>
<tr><td rowspan="5">Step 3</td><td rowspan="4">Statistical testing of prioritized genes</td><td rowspan="2">Targeted resequencing</td><td>Samples
• 310 sALS cases
• 266 non-sALS controls</td></tr>
<tr><td>Criteria used in analysis
• Rare variants (MAF < 0.01)
• Non-synonymous or variants near genes
• SKAT analysis</td></tr>
<tr><td rowspan="2">Validation of candidate genes using an independent sequencing dataset</td><td>Samples
• 170 sALS cases
• 42 non-sALS controls</td></tr>
<tr><td>Criteria used in analysis
• Rare variants (MAF < 0.01)
• Non-synonymous or variants near genes
• SKAT and meta analysis (adjustment for multiple testing using a Bonferroni correction)</td></tr>
<tr><td colspan="3">1 novel gene (MYH15) associated with ALS risk</td></tr>
</table>

Figure 2.1. 3-step strategy to identify genetic factors associated with ALS risk using a hypothesis-driven and targeted genetic association study (step 1 and step 3) and fly genetics (step 2).

***Drosophila* genetic screen**

Given transgenic fly lines expressing G_4C_2 repeats display progressive neurodegeneration in eye and motor neurons similar to ALS patients (Freibaum et al., 2015; Xu et al., 2013), we performed a genetic screen using a transgenic fly expressing 30 repeats of G_4C_2 under a *GMR-Gal4* driver (eye-specific) as reported previously (Xu et al., 2013) and tested whether the 135 selected genes could modulate G_4C_2 repeat-associated toxicity (Figure 2.1). Of the 135 genes, 89 genes (65.9%) had a functional homolog in the fly genome with at least a moderate rank score according to the DRSC Integrative Ortholog Prediction Tool (DIOPT, Version 6.0.2 [June 2017]) (Hu et al., 2011) (Figure 2.1, Table S3). A total of 90 RNAi lines corresponding to 49 fly genes were crossed with the $(G_4C_2)_{30}$ repeat transgenic line to determine the genetic interaction between the G_4C_2 repeat and candidate genes (Freibaum et al., 2015) (Figure 2.1, Table S4). All RNAi lines crossed with flies carrying the *Gmr-GAL4* driver alone showed no pathological eye findings (data not shown). However, 11 RNAi/G_4C_2 lines suppressed the $(G_4C_2)_{30}$-related toxicity, and 7 lines showed an evident enhancement of the disrupted eye morphology accompanied by severe necrosis (Figure 2.2A, Table 2.1). Thin-section analysis of $(G_4C_2)_{30}$ flies crossed with suppressors verified the recovery of photoreceptor cells and fewer vacuolated materials compared to the $(G_4C_2)_{30}$ flies itself (Figure 2B). The genes identified in this screening are involved in various cellular functions including cell adhesion, DNA or RNA binding, and regulation of oxidative stress (Table 2.1). Interestingly, the WGS and *Drosophila* screen identified *HIPK2* as a candidate gene of interest (Table 2.1, Table S2). This gene was recently implicated in ALS neurodegeneration, lending validity to our approach (Lee et al., 2016b).

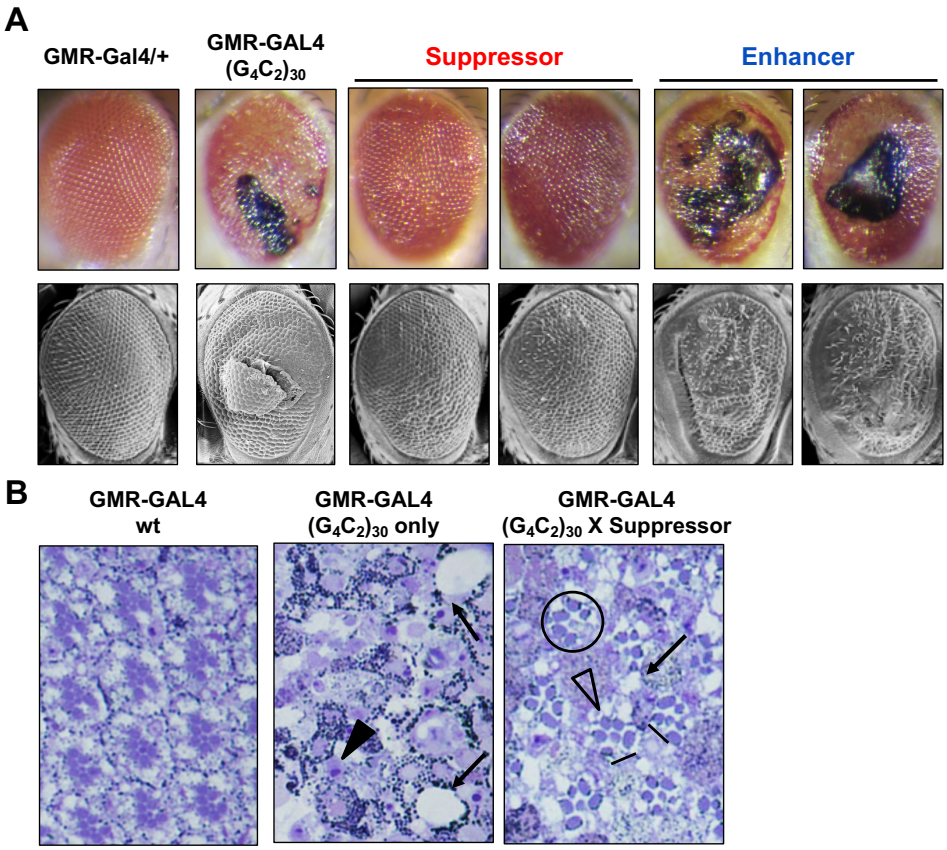


Figure 2.2. Functional screen identifies multiple genetic modifiers of $(G_4C_2)_{30}$ toxicity. (A) The expression of $(G_4C_2)_{30}$ driven by the *GMR-Gal4* driver causes rough eye phenotypes as shown in light microscope image and ommatidial disruption as shown in scanning electron microscope image. In screening the flies, we selected genes with rescued phenotypes as a suppressor and genes with aggravated phenotypes as an enhancer. (B) A representative thin section for GMR-Gal4 flies with either $(G_4C_2)_{30}$ alone or both the $(G_4C_2)_{30}$ and RNAi of suppressor genes. The $(G_4C_2)_{30}$ flies showed a loss of photoreceptor cells (arrowheads) and vacuolated material (arrows). However, $(G_4C_2)_{30}$ flies crossed with RNAi of suppressor exhibited rescued phenotypes regarding recovered photoreceptor cells (circle) and smaller size of vacuolated material (arrows) although there are an abnormal number of photoreceptor cells (open arrowheads) and polarity defects (bars).

Patient type	Gene symbol	Fly ortholog	Fly screening Result	Targeted resequencing	Biological function
YALS	ABCC2	MRP	**Enhancer**	Yes	Protein transporter and regulation of oxidative stress
YALS	MYH15	Mhc	**Enhancer**	Yes	Tight junction pathway
YALS	PLEKHG2	GEFmeso	**Enhancer**	Yes	Postsynaptic signaling pathway
YALS	PPARD	Eip75B	**Enhancer**	Yes	Peroxisome
YALS	SVEP1	uif	**Enhancer**	Yes	Cell adhesion process
YALS	UTP20	CG4554	**Enhancer**	Yes	18s rRNA processing
OALS	CDK11A	Pitslre	**Enhancer**	Yes	Cell cycle and apoptosis
YALS	CELF5	bru-3	**Suppressor**	Yes	mRNA editing and translation
YALS	DBF4	chif	**Suppressor**	No	Cell Cycle Checkpoints in DNA replication
YALS	DLG2	dlg1	**Suppressor**	Yes	Postsynaptic signaling pathway
YALS	EGR3	sr	**Suppressor**	No	Transcriptional regulator in mitogenic stimulus
YALS	FAM98B	CG5913	**Suppressor**	Yes	tRNA processing and gene expression
YALS	FXR2	Fmr1	**Suppressor**	No	RNA binding
YALS	HIPK2	Hipk	**Suppressor**	No	Endoplasmic reticulum (ER) stress
YALS	HK3	Hex-A	**Suppressor**	Yes	Metabolism
YALS	PDK3	Pdk	**Suppressor**	Yes	Metabolism
OALS	KDM2A	Kdm2	**Suppressor**	Yes	Epigenetic modification
OALS	KIF27	cos	**Suppressor**	Yes	Cytokinesis

Table 2.1. The 18 candidate genes in the table either suppress or enhance the neuronal toxicity, from the repeat expansion. Among the 18 genes, 14 genes which are previously unknown for ALS and other neurological association were validated by targeted resequencing.

Targeted resequencing of prioritized gene lists

To further understand the contribution of the genes resulting from WGS analysis and functional screening to ALS without the G_4C_2 repeats and known ALS associated genes, we tested whether individuals with sporadic ALS (sALS) were enriched for rare, likely deleterious variants at selected genes compared to non-ALS controls. For a novel gene finding, we excluded 4 candidate genes which were already known for ALS association in the literature (*DBF4*, *EGR3,* and *HIPK2*) and other neurological disorders (*FXR2*), and then performed targeted resequencing that focused on exonic regions of 14 candidate genes and 5 known ALS-associated genes (i.e., *FUS*, *GRN*, *SOD1*, *TARDBP,* and *TBK1*) in a collection of ALS subjects who were negative for the G_4C_2 expanded repeat (Umoh et al., 2016) (Figure 2.1, Table 2.1, Table S5). After filtering outliers from principal-component analysis (PCA) and samples with low sequencing quality, 489 samples (272 sALS and 217 non-ALS) were included in the analysis (Figure S1A and S1B). We filtered the 1447 variants by those predicted to cause coding changes (e.g., missense, nonsense, and frameshift mutations) and having MAF <1% among controls (Figure S1A). Two known pathogenic mutations in *SOD1* (Ile114Thr) and *TARDBP* (Gly287Ser mutation) were identified in ALS cases (*SOD1* carrier: female, the age of onset: 36 years; *TARDBP* carrier: female, the age of onset: 76.3 years) (Kabashi et al., 2008); these cases were excluded from further analysis. We performed sequence kernel association test (SKAT) analysis (Lee et al., 2012; Wu et al., 2011c) using all variants from each gene, controlling for population structure using eigenvectors from PCA. Using bootstrap to estimate empirical p-values, we found two genes, *DLG2* (10 variants; *p-value = 0.04180*) and *MYH15* (16 variants; *p-value = 0.01950*), which showed suggestive evidence of association with ALS (Figure 2.3A, Table 2.2, Table S6, Table S8). *MYH15* also showed suggestive association with ALS in the unified rare variant association test, SKAT-O (*p-value* = 0.03697, Table 2.2).

To examine whether the *DLG2* and *MYH15* genes that showed suggestive association in the targeted resequencing dataset replicated in a validation dataset, we obtained WGS data of 212 people recruited

at Emory University School of Medicine (Figure 2.1). We followed the same quality control procedures as were used for the targeted resequencing (Figure S1C and S1D), which resulted in the removal of 4 individuals (2 were outliers for ancestry, and 2 were carriers of known pathogenic mutations in *SOD1*). The same selection criteria for variants were applied, which were tested for the resequencing experiment. We performed gene-based analysis using SKAT and SKAT-O and found significant association between ALS and *MYH15* (6 variants; *p-value* = 0.01233 and 0.01708, respectively) after adjustment for multiple testing using a Bonferroni correction (Figure 2.3B, Table 2.2, Table S7). We observed no association between ALS and *DLG2* (2 variants; *p-value* = 0.30207, Table 2.2, Table S7, Table S8).

Meta-analysis

To improve statistical power, we performed a SKAT-based meta-analysis of the two independent datasets using the MetaSKAT (Lee et al., 2013) package. Since all samples used in this meta-analysis had the same ethnicity, we expected homogeneous genetic effects across the samples. The genomic coordinates from the WGS dataset were converted from hg19 to hg38 using LiftOver to pool individual-level genotype data from the two datasets mapped with different assemblies (Kent et al., 2002). Consistent with SKAT results for each dataset, *MYH15* showed borderline significant association with ALS in the SKAT test (20 variants; adjusted *p-value* = 0.02511, Table 2.2). Two variants in *MYH15* were shared in *MYH15* by both datasets, one of which (rs61744539; R1141*) is a nonsense mutation potentially leading to downregulation of *MYH15* gene expression (Figure 2.3, Table S8).

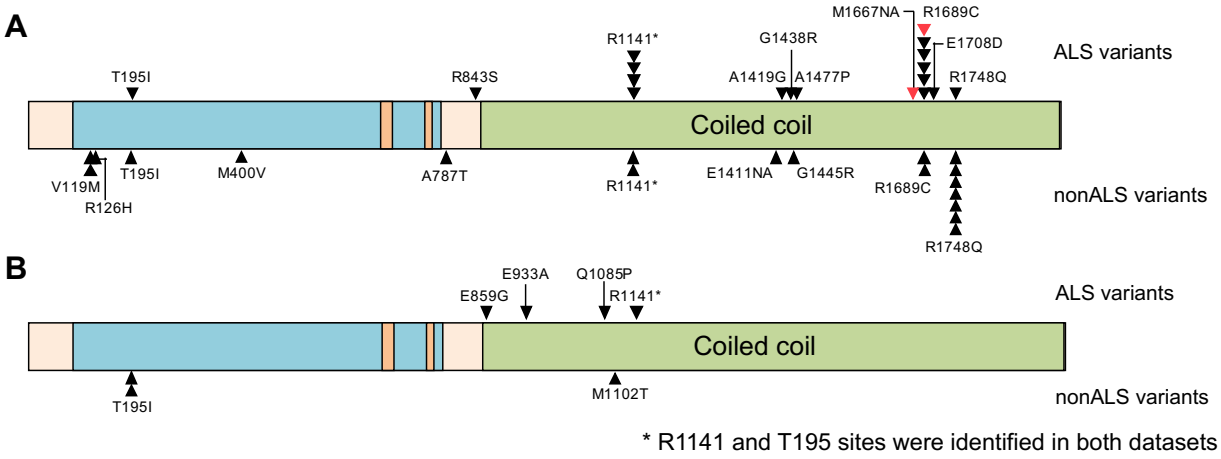


Figure 2.3. Coding variants of *MYH15* identified in either ALS cases or controls during the targeted resequencing (A) and validation dataset (B). Black arrow indicates heterozygous variant while red arrow indicates homozygous variant.

Group	Genes	Targeted Resequencing			Replication (WGS)			Meta-Analysis		
		variant number	SKAT p-value	SKAT-O p-value	variant number	SKAT p-value	SKAT-O p-value	variant number	SKAT p-value	SKAT-O p-value
Suppressors	DLG2	10	0.0418	0.09971	2	0.30207	0.30207	12	0.23962	0.37744
Enhancers	MYH15	16	0.0195	0.03697	6	0.01233	0.01708	20	0.02511	0.0472

Table 2.2. Gene-based analysis of rare variants for targeted resequencing dataset, replication (WGS) dataset, and meta-analysis which combines two datasets. Empirical p-values based on resampling techniques are provided.

MYH15 modulates dipeptide-mediated toxicity associated with G4C2 repeat expansion

The G_4C_2 expansions exert neuronal toxicity through direct RNA transcripts (Kumar et al., 2017; Xu et al., 2013) and repeat-associated non-AUG (RAN) translated dipeptide repeat proteins (DRPs) (Lopez-Gonzalez et al., 2016; Mori et al., 2013; Wen et al., 2014). Of the dipeptides, the arginine-rich proteins, proline-arginine (PR) and glycine-arginine (GR), lead to a significant decrease in survival and aggregate eye phenotypes in the *Drosophila* model (Mizielinska et al., 2014). To determine whether *MYH15* could modulate DPR-mediated toxicity, we performed a functional assay using a transgenic fly expressing 36 repeats of either PR or GR under *GMR-Gal4* driver (Mizielinska et al., 2014). The knockdown of the *MYH15 Drosophila* ortholog, *Mhc*, resulted in enhanced retinal toxicity when crossed with both the G_4C_2-repeat line and the PR repeat line (Figure 2.4A). We also observed substantial lethality in the GR repeat line when crossed with *Mhc*-KD line (data not shown). In addition to a fly model, the downregulation of *Myh15* in Neuro2A cell lines could enhance poly(PR)- and poly(GR)-mediated cell toxicity (Zhang et al., 2016) (Figure 2.4B). A recent report identified the moderate interaction between PR50 and *MYH9* as well as *MYH10* (Lee et al., 2016a), consistent with our findings. Given that myosin heavy chain genes are involved in vesicle transport (Hirokawa et al., 2010), *MYH15* can potentially modulate PR aggregate-mediated toxicity via the impairment of vesicle trafficking.

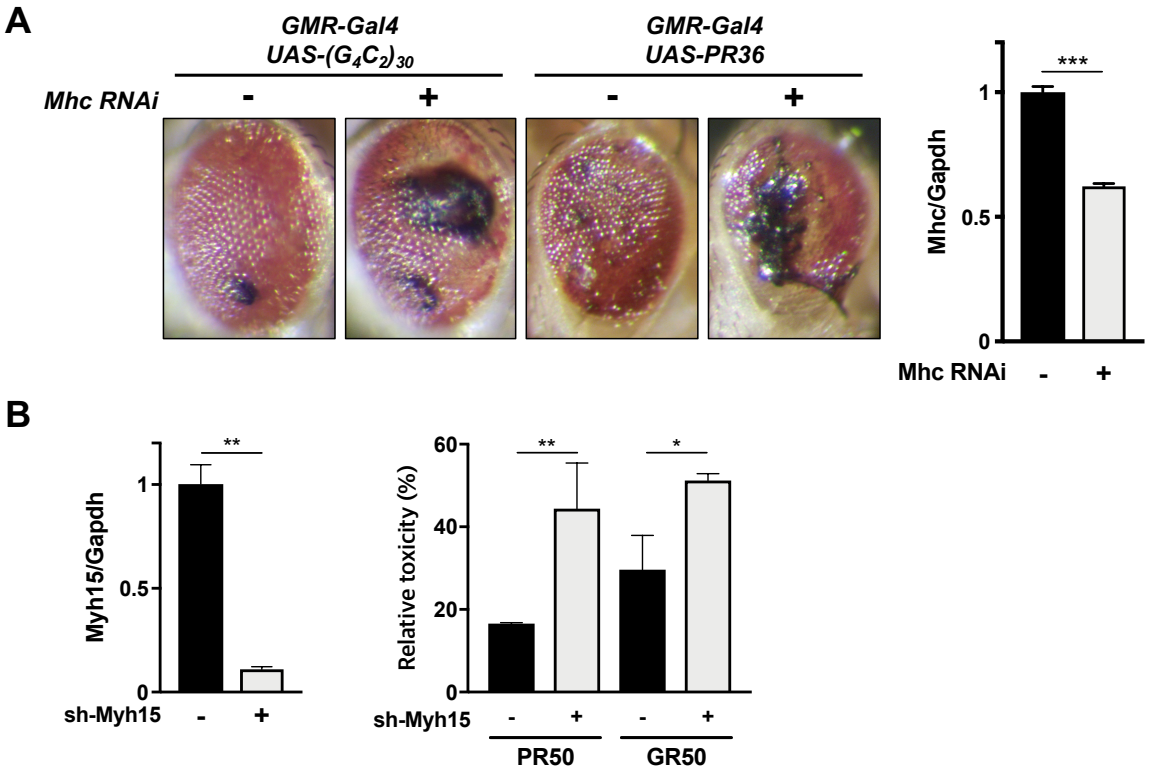


Figure 2.4. MYH15 is a potential genetic modifier of dipeptide-mediated toxicity. (A) Left: Transgenic lines expressing either $(G_4C_2)_{30}$ or $(PR)_{36}$ under *GMR-Gal4* driver cause progressive neurodegeneration in eye. Both transgenic lines displayed aggravated phenotypes when crossing with a RNAi line of *Mhc*, a *Drosophila* ortholog of *MYH15*, implying *MYH15* can modify RNA- and dipeptide-mediated toxicity. Right: The knockdown efficiency of *Mhc* RNAi lines crossed with *ELAV-Gal4* driver was confirmed by quantitative RT–PCR (qPCR). (B) Left: Relative *Myh15* expression after siRNA treatment (50nM) in Neuro2A cell lines. Right: Relative cell viability measured on 3 days after plasmid and siRNA co-transfection. Control: GFP, PR50: GFP-$(PR)_{50}$, GR50: GFP-$(GR)_{50}$

2.5. Discussion

Here we present a 3-step strategy to identify ALS risk-associated genes by integrating fly genetics with whole-genome sequencing (Figure 2.1). Our hypothesis is that genetic factors modulating phenotypic variability of G_4C_2 expansion carriers are associated with ALS risk. As such, initial candidate genes were selected by whole-genome sequencing (WGS) on four unrelated G_4C_2 expansion carriers who developed ALS approximately 30 years apart, identifying 135 potential risk genes (step 1). To prioritize candidate genes from WGS of a small number of C9ALS individuals, we used a *Drosophila* genetic screen to test for genetic interactions between candidate genes and the G_4C_2 model of neurodegeneration (step 2). Through this unbiased screen, we identified novel genetic interactions as well as a known interaction with G_4C_2 toxicity (HIPK2), which supports that our approach is suitable for novel gene identification. Finally, rather than sequencing all genes, most of which are irrelevant to ALS risk, only targeted candidate genes were analyzed to investigate their association with ALS without any known pathogenic mutations in *C9orf72*, *FUS*, *GRN*, *SOD1*, *TARDBP*, and *TBK1* (step 3). Gene-based statistical testing of targeted resequencing and WGS on sporadic ALS (sALS) cases and controls suggests rare variants in *MYH15* represent a likely genetic risk factor for ALS. A further functional assay revealed that *MYH15* can a genetic modifier of dipeptide-mediated toxicity of C9ALS (Figure 2.4).

MYH15, myosin heavy chain 15, was a recently characterized as a slow-type myosin involved in muscle contraction and cytoskeleton remodeling (Barany, 1967; Desjardins et al., 2002). Well-known class-II *MHC* genes are considered to be divergent products from an ancestral gene through the series of gene duplications due to structural similarity of myofilaments within the same class genes (Cope et al., 1996; Desjardins et al., 2002). However, *MYH15*, along with *MYH14* and *MYH16*, displayed unrelated structural features from classical *MYH* genes. In particular, loop domains of *MYH15* are highly divergent;

for instance, the N-terminal positive charge cluster in loop 1 is lost and there is no matched sequence in loop 2 (Desjardins et al., 2002). In addition, the unexpected large size of *MYH15* (>142,000 bp) provides greater possibility of having genetic variants in both exons and introns (Desjardins et al., 2002). Indeed, recent genetic studies identified a common male-specific association of single-nucleotide polymorphisms (SNPs) in *MYH15* with increased coronary microvascular dysfunction risk: there was a nominal association of variants in *MYH15* with increased risk of stroke and coronary heart disease (CHD) (Luke et al., 2009; Yoshino et al., 2014). One of the notable variants in *MYH15* identified in aforementioned studies is Thr1125Ala (rs3900940), which is located in the coiled-coil tail domain of *MYH15* (Luke et al., 2009). In our study, aside from R1748Q (rs56118396), variants identified in the ALS population of our study are distributed within the rod-like tail sequence while variants at N-terminus skew toward the non-ALS population (Figure 2.3). Given that the combination of van der Walls forces and electrostatic interactions between proper amino acids is critical for home-dimerization of the tail domain, the disruption of the coiled structure resulting from nonsynonymous variants is likely associated with ALS progression.

In addition to genetic association studies in cardiovascular diseases, the association of variants in *MYH15* was investigated in a study of common mental disorders, schizophrenia and bipolar disorder (O'Dushlaine et al., 2011). The genome-wide association study (GWAS) was performed and a statistical association was seen using two independent data sets: The International Schizophrenia Consortium (ISC) data set including 3322 schizophrenia cases and 3587 controls from the same ethnic population and the Genetic Association Information Network (Ha Thi et al., 2014) data set, including 1351 cases and 1378 controls. Among associated genes involved in a tight junction pathway, a SNP in *MYH15* (rs16854665, MAF = 0.1287, gnomAD) displayed statistical significance in both studies (O'Dushlaine et al., 2011), suggesting that genetic variants in *MYH15* can be associated with other brain disorders. In addition,

MYH15 is highly expressed in brain-spinal cord tissues compared to other organs (Figure S2) (Gamazon et al., 2015). However, there is no previous evidence about the implication of *MYH15* variants in ALS pathogenesis. This is a first report that links *MYH15* to ALS.

In summary, we have identified *MYH15* as a potential genetic factor associated with ALS risk. Our analyses demonstrate that the combination of WGS with fly genetics facilitates the discovery of fundamental genetic components of complex traits with a limited number of samples.

2.6. Acknowledgements

We thank Hong Yi and Jeannette V. Taylor (The Robert P. Apkarian Integrated Electron Microscopy Core (IEMC), Emory University) for helping Scanning Electron Microscope (SEM) imaging of *Drosophila* eyes. We also appreciate core members of the Emory Integrated Genomic core (EIGC)) for running MiSeq and helpful discussion about sample preparation. This work was supported in part by the National Institutes of Health [NS051630, NS091859, and NS097206 to P.J. and NS073873 to J.E.L. AG056533 to T.S.W. and AG025688] and the Veterans Health Administration [BX001820 to T.S.W.], and the American ALS Association [to J.E.L. and J.D.G]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Veterans Health Administration. Support for patient recruitment and genetic analysis is provided to JDG and JEL by the ALS Association and the Muscular Dystrophy Association.

CHAPTER 3: Aging-related epigenetic dynamics in cerebellum

3.1. Introduction

The cerebellum, located in the posterior cranial fossa, is a central brain part attributed to motor control as well as several sensory and perceptual processes (Buckner, 2013). Although its volume constitutes only 10 percent of the total brain volume and it has relatively simple cellular organization, it consists of more neurons (estimated 101 billion neurons) than any other brain part, and functional abnormalities of the cerebellum are substantially associated with various neurological disorders (Herculano-Houzel, 2009; Schmahmann, 2004). Therefore, it is critical to understand how dynamic changes of gene expression are orchestrated during normal neurodevelopment and aging. A number of studies have demonstrated the key role of epigenetic mechanisms in gene expression regulation during embryonic and adult neurogenesis (Hsieh and Zhao, 2016; Munoz et al., 2012; Yao et al., 2016). Indeed, dynamic changes in DNA methylation during development and aging contribute to shaping the age-dependent transcriptional landscape by regulating transcription factor binding in a timely and spatially distinct manner (Curradi et al., 2002; Field et al., 2018). Another DNA modification, 5-hydroxymethylation (5hmC), is highly abundant in mature cerebellar cells such as Purkinje cells (GABAergic neuron) and granule cells (Glutamatergic neurons) and undergoes dynamic changes in the mouse (Szulwach et al., 2011). In addition, 5hmC alterations are strongly associated with cerebellar disorders such as Huntington's disease (HD), Fragile X-associated tremor/ataxia syndrome (FXTAS), and autism spectrum disorder (ASD) (Cheng et al., 2018; Wang et al., 2013; Yao et al., 2014), but age-dependent 5hmC status in healthy human cerebella remained poorly understood.

In this study, we used 12 healthy cerebellar tissues of two distinct age groups (YCB: 5-19 years old,

OCB: 70-89 years old) to map genome-wide distribution of 5hmC. We found age-dependent differentially hydroxymethylated regions (DhMRs), although the global abundance of both 5mC and 5hmC was comparable regardless of ages. Age-dependent DhMRs were enriched at different genomic loci, suggesting the dynamic changes of 5hmC across the genome could play essential roles in postnatal neurodevelopment and aging. Further motif and pathway analyses revealed that YCB-associated DhMRs were identified around genes involved in proliferation and neuro-transmission while the genes near OCB-associated DhMRs were implicated in immunity and protective pathway against brain aging. We also performed transcriptome analysis and identified that age-dependent differentially expressed genes exerted age-related biological activities such as the regulation of cell-cell communication in young children (5-11 years old), neuronal maturation in young adult (19 years old), and age-defense response in old adult (70-89 years old). In addition, we found that genes differentially expressed in young children showed positive correlation with gene-body 5hmC levels. These results together suggest that age-dependent 5hmC dynamics play a pivotal role in regulating genes involved in neurodevelopment and aging.

3.2. Materials and Methods

Human Tissues

Twelve normal cerebellum samples with no neurological disorders (Table S9) were collected from the NIH NeuroBioBank tissue repositories, which were used for genome-wide 5hmC profiling and transcriptome analysis. Twelve samples are classified into two different age groups: YCB group with 5 to 19 years old and OCB group with 70 to 89 years old.

Quantification of 5mC and 5hmC using High Performance Liquid Chromatography (HPLC)

Genomic DNA (gDNA) preparation was performed as described previously (Song et al., 2011c). In brief, after homogenization of different brain tissues in 600 µl of digestion buffer (100 mM Tris-HCl, pH 8.5, 5 mM EDTA, 0.2% SDS, 200 mM NaCl), samples treated with Proteinase K (Thermo) were incubated at 55°C overnight. On the next day, the same volume (600 µl) of phenol:chloroform:isoamyl alcohol (25:24:1 saturated with 10 mM Tris, pH 8.0, 1 mM EDTA) (P-3803, Sigma) was added to samples, mixed thoroughly by shaking, and then centrifuged for 10 min at 12,000 rpm. After careful transfer of the upper layer (aqueous layer) into a new Eppendorf tube, 600 µl of isopropanol was added to samples and then, precipitated gDNA was reconstituted into distilled water. For the measurement of 5mC and 5hmC using HPLC, extracted gDNA was hydrolyzed to nucleosides and run on a Zorbax XDB-C182.1 3 50 mm column (1.8 mm particle size) attached to an Agilent 1200 Series HPLC system coupled to an Agilent 6410 Triple Quad MassSpectrometer.

Genome-wide 5hmC profiling (hMe-seal sequencing)

The enrichment of 5hmC containing genomic regions was performed for genome-wide 5hmC mapping as previously described (Song et al., 2011c). Briefly, 1 µg of sonicated gDNA with major peak at 200bp

was incubated for 2 hours at 37°C in a 30 µl reaction volume containing 100 µM UDP-6-N_3-Glu, β-glucosyltransferase (β-GT) and NEB buffer 4. DNA purification was performed using AMPure XP beads according to manufacturer recommendation and reconstituted in 30 µl of H_2O. For biotinylation, modified DNA samples were incubated for 2 h at 37°C with addition of 150µM dibenzocyclooctyne modified biotin (click chemistry), and then biotinylated gDNA was captured using Streptavidin. After purification using AMPure XP beads and quantification using Qubit, DNA libraries were generated using NEBNext® Ultra™ II DNA Library Prep Kit, which was ready for sequencing.

Identification of age-specific differential hydroxymethylation regions (DhMRs)

Raw sequencing fastq files were mapped to human genome, hg19 using bowtie2 (Langmead and Salzberg, 2012). After filtering low quality reads and sorting with samtools (Li et al., 2009), PCR duplicates were removed using Picards (Broad Institute, 2016). To understand correlation between age and genome-wide 5hmC distribution, binned matrix (binsize: 2kb) generated using final bam files was used to perform a principal component analysis (PCA) with the built-in R function, 'prcomp()' and Pearson correlation with the built-in R function, 'cor()'. 5hmC peak identification was conducted using with MACS2 with default parameters (Zhang et al., 2008), and then initial differential hydroxyl-methylation regions (DhMRs) were determined using DESeq2 with default parameter (Love et al., 2014). Among the DhMRs identified by DESeq2, age group-specific genomic regions which were not called as peaks within 80 % of samples in the corresponding age group were filtered out, and final group-specific DhMRs were identified after merging adjacent genomic regions using bedtools. All identified genomic regions were annotated by HOMER (Heinz et al., 2010). To understand cis-regulatory function of DhMRs, GREAT was used with default "Basal plus extension" settings (McLean et al., 2010). Significant terms were selected based on less than FDR threshold 0.05 from region-based binomial test and greater than 1.5 region-based fold-change. Motif scanning on DhMRs was performed using Homer

software.

Analysis of RNA-sequencing data

Cerebellum total RNA for sequencing was prepared using Trizol Reagent. Raw sequencing data were mapped to hg19 using HiSAT2 and annotated using StringTie (Pertea et al., 2016). Differentially expressed (DE) genes were identified using DESeq2 (Love et al., 2014) with default setting, and heatmap of DE genes were drawn using Bioconductor package, 'pheatmap'. To understand biological pathways of DE genes for each group, we performed GO enrichment analysis using DAVID (Huang et al., 2009a, 2009b).

3.3. Results

Genome-wide 5hmC distribution is distinct depending on biological ages

In normal mouse development and aging, the global abundance of 5mC and 5hmC changes in an age-dependent manner: while 5mC gradually decreases, 5hmC levels significantly increase during postnatal development and slightly increase with aging (Szulwach et al., 2011; Wilson et al., 1987). To assess if the features identified in mice are conserved in human, we measured 5mC and 5hmC levels of postmortem human cerebellar tissues with different ages and without definite neurological disorders using high performance liquid chromatography (HPLC) (Figure 3.1A and Table S9) (Kriaucionis and Heintz, 2009a). Unlike mice, there was no significant difference in terms of global levels of both 5mC and 5hmC in human cerebella while there was individual variability (Figure 3.1B). However, genomic mapping of 5hmC using the same cerebellar tissues (Song et al., 2011d) exhibited distinct 5hmC distribution depending on the biological ages (Figure 3.1C). In addition, unsupervised hierarchical clustering successfully distinguished the two age groups (Figure 3.1D), suggesting distinct age-related 5hmC genomic loci regardless of its similar abundance.

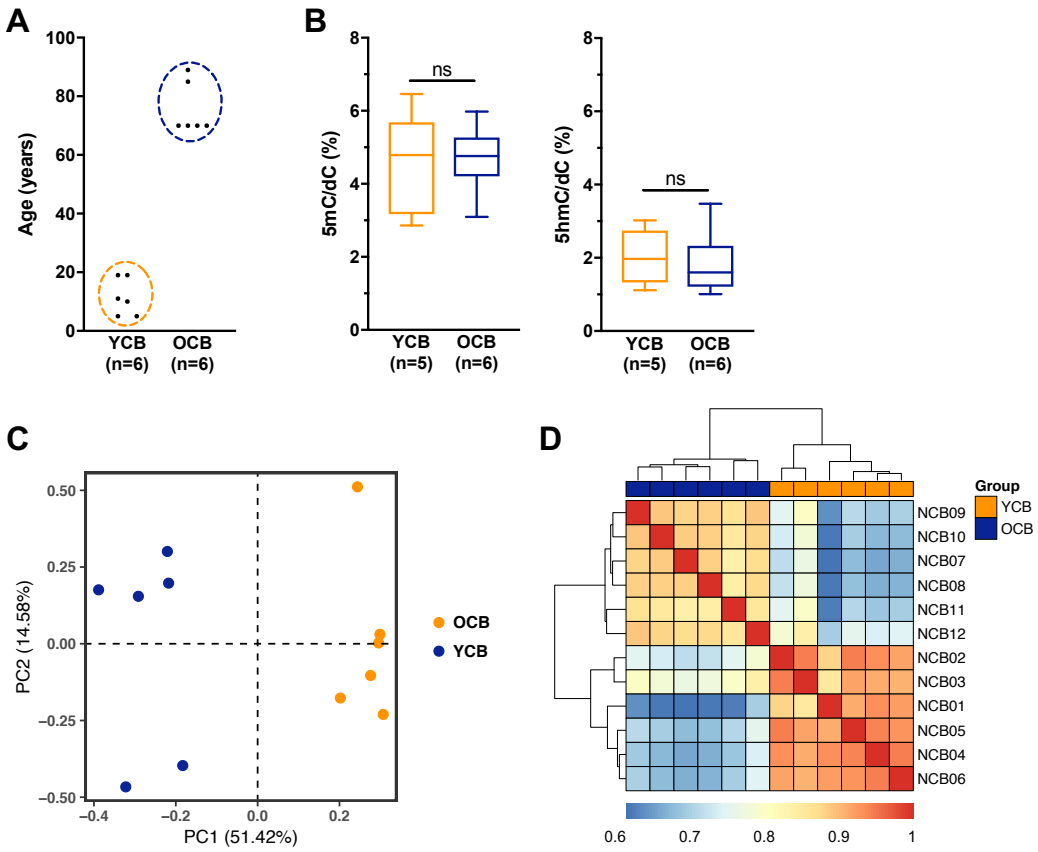


Figure 3.1. Distinct age-dependent 5hmC patterns (A) The summary of age distribution of each group (YCB: 5-19 years old, OCB: 70-89 years old). (B) Global levels (%) of 5mC and 5hmC over total cytosine (C) Principal Component Analysis (PCA) plot showing the first two eigenvectors which explain variance of the variables. Colored symbols correspond to each sample data point in this study. (D) Heatmap showing correlation of 5hmC patterns of each sample sorted by unsupervised hierarchical clustering (Pearson correlation, $p < 0.001$).

Identification of age-dependent DhMRs in healthy cerebellum

To detect age-dependent differential hydroxymethylation regions (DhMRs), we divided samples into a young (age < 20; YCB) and old age group (age ≥ 70; OCB). The statistical analysis was performed using DESeq2 with the counts of mapped reads in each 2kb bin of human genome (hg19) to identify initial DhMRs, which yielded 98,791 DhMRs showing decreased hydroxymethylation in OCB, and 88,073 DhMRs showing increased hydroxymethylation in OCB (Figure 3.2A). Through the steps to filter out DhMRs which were not detected as peaks in MACS2 and to merge adjacent regions, we finally identified 62,032 DhMRs associated with YCB and 60,041 DhMRs associated with OCB. Genomic annotation of these DhMRs revealed that YCB-associated DhMRs were enriched at cis-regulatory regions such as promoters and CpG-islands ($p < 0.05$ and $p < 0.001$, respectively) whereas OCB-associated DhMRs were enriched at coding regions including exon and 3'UTR (both $p < 0.001$) (Figure 3.2B). However, CpG-island containing 5hmC marks were only located within introns (62.72%) and intergenic regions (37.28%) (Figure 3.2C), consistent with the general notion that CG rich-promoters are usually DNA modification-free whereas CG low promoters tend to be methylated (Weber et al., 2007), as well as a prior finding that 5hmC is depleted at CGI promoters in mice (Szulwach et al., 2011).

For further understanding of regulatory functions of YCB-associated DhMRs, we performed motif analysis using HOMER. Notably, we found the significant enrichment of YCB-associated DhMRs at development-related transcription factors such as PITX1, THRb, and PTF1A (Figure 3.2D). PITX1 is a homeodomain transcription factor (TF) critical for neurodevelopment (Szeto et al., 1999) and the enrichment of hydroxymethylation at the binding motif of PITX1 was previously identified (Madrid et al., 2018). In addition, PTF1A is one of the basic helix–loop–helix (bHLH) TFs involved in the differentiation of neural precursors into GABAergic neurons in the cerebellum (Hoshino et al., 2005). PTF1A also plays a pivotal role in pancreatic differentiation of human embryonic stem cell (hESC), and

5hmC enriched genomic loci near *PTF1A* gene showed enhanced chromatin accessibility (Li et al., 2018). Because the cerebellum undergoes an increase in volume and circuit maturation in early life (Tiemeier et al., 2010; Wang et al., 2014), the 5hmC signature found in children may contribute to activating development-associated neurogenesis by recruiting appropriate TFs.

Age-dependent dynamics of 5hmC enrichment on repeat elements

More than 50% of the human genome consists of repetitive DNA elements, mainly categorized into two types: tandem repeats and interspersed repeats. Since the elements are not translated to form proteins, they were previously considered to be useless or 'junk' DNA. However, the discovery of jumping genes (transposons) and their role in regulating gene expression by Barbara McClintock drew attention to the biological roles of repeated sequences (Shapiro and Von Sternberg, 2005). Indeed, they serve as a core domain for hetero-chromatin formation and mitotic chromosome folding during cell division (Shapiro and Von Sternberg, 2005). They are also critical for organizing 3D genome structure by building the boundary of topologically associated domain (TAD) through CTCF recruitment (Harmston et al., 2017; Winter et al., 2018); hence, the proper regulation of repetitive regions is important. Previous data have revealed the significance of DNA methylation in genomic stability and differential methylation patterns at repetitive elements during development (Papin et al., 2017; Putiri and Robertson, 2011). 5hmC patterns at repeat elements are also dynamic during mouse development and aging (Szulwach et al., 2011). In human cerebella, we found significant enrichment of 5hmC on short interspersed nuclear element (SINE) (59.56%) and long tandem repeat (LTR) (9.77%) in YCB group whereas OCB-associated DhMRs were also enriched on simple repeats (2.74%) and DNA transposons (11.51%) (Figure 3.2E). Given that the alteration of transposon activity can be mediated by 5hmC levels (Sun et al., 2016), our data imply the association of epigenetic dynamics at divergent repeat elements with aging.

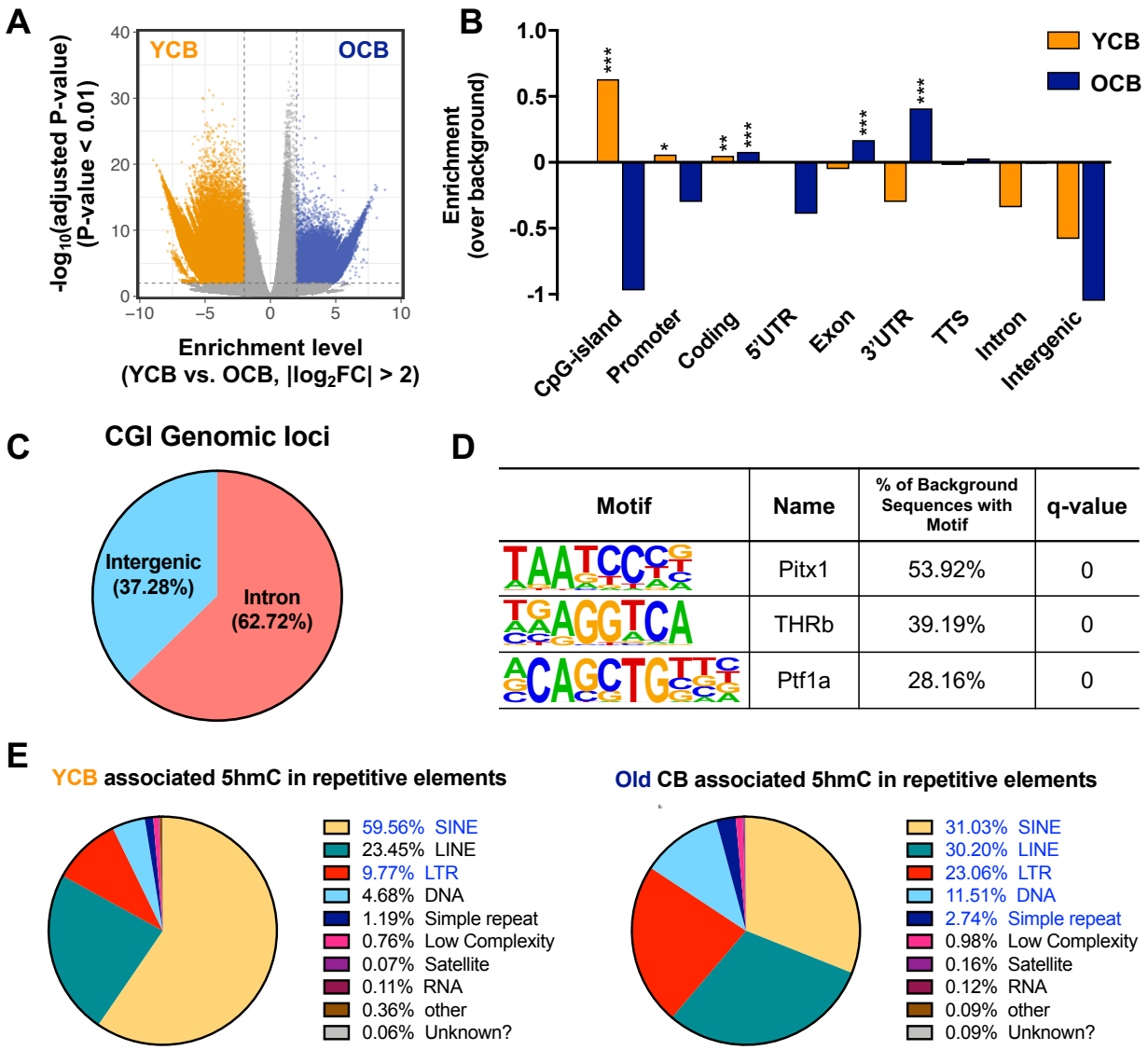


Figure 3.2. Age-dependent DhMRs are enriched at different genomic loci. (A) Volcano plot showing the hMe-seal data of 6 YCB and 6 OCB samples. Orange dots indicate 5hmC-gain regions in YCB, and blue dots indicate 5hmC-gain regions in OCB. Two-thousand-base-pair binning was performed, and the criteria were set as an absolute value of the log 2 fold change (OCB/YCB) > 2 and $p < .01$. (B) Genomic annotation of identified age-dependent DhMRs using HOMER. Statistical significance was marked above each corresponding bar graph. (C) Pie charts illustrating genomic regions of CpG island of YCB-specific DhMRs. (D) Sequence logos shown for the highly enriched sequence motifs in YCB-specific

DhMRs. (E) Pie charts illustrating annotation summary of repeat elements found in each group specific DhMRs. Annotations with absolute fold change > 2 and adjusted p-value < 0.05 than background are marked as blue.

Predicted cis-regulatory functions of age-dependent DhMRs

To further understand the cis-regulatory functions of age-associated DhMRs, we used GREAT (Genomic Regions Enrichment of Annotations Tool, version 3.0), which performs statistical tests using the annotation of genes nearby the input regions. By filtering MSigDB pathway terms with p-value less than 0.05 and 1.5-fold enrichment in binomial test, we identified 12 and 18 pathway terms associated with YCB-specific DhMRs and OCB-specific DhMRs, respectively (Figure 3.3A and 3.3B). Notably, proliferation-related pathways (e.g., the G2 and EIF pathways) and pathways related to the maturation of neural circuitry (Neurotransmitters pathway) were highly enriched using YCB-specific DhMRs (Figure 3.3A). Indeed, high levels of 5hmC were identified nearby *EN2*, the Engrailed homeobox gene. EN2, a key regulator of cerebellar pattern formation, is highly expressed in the cerebellum and its expression is affected by dynamic epigenetic programs during the development (James et al., 2013). In mouse cerebellum, high levels of 5hmC upstream of *En2* are observed across all ages in the cerebellum, but we found that 5hmC peaks were much higher in YCB (Figure 3.3C, left), suggesting EN2 plays a more critical role in developmental neurogenesis in human. Interestingly, immunity-related pathways (e.g. CCR5 and NOS1 pathways) were enriched at OCB-specific DhMRs (Figure 3.3B), and strong 5hmC signals were found in the coding regions of genes implicated in immunity (CCR5; Figure 3.3C, right). Another OCB-enriched pathway was the calcineurin (Ca2+-dependent protein phosphatase) pathway, previously identified because it is associated with brain aging and abnormal calcineurin activity leads to memory deficits (Foster et al., 2001). Collectively, these data suggest that 5hmC signature is strongly linked to age-related pathways.

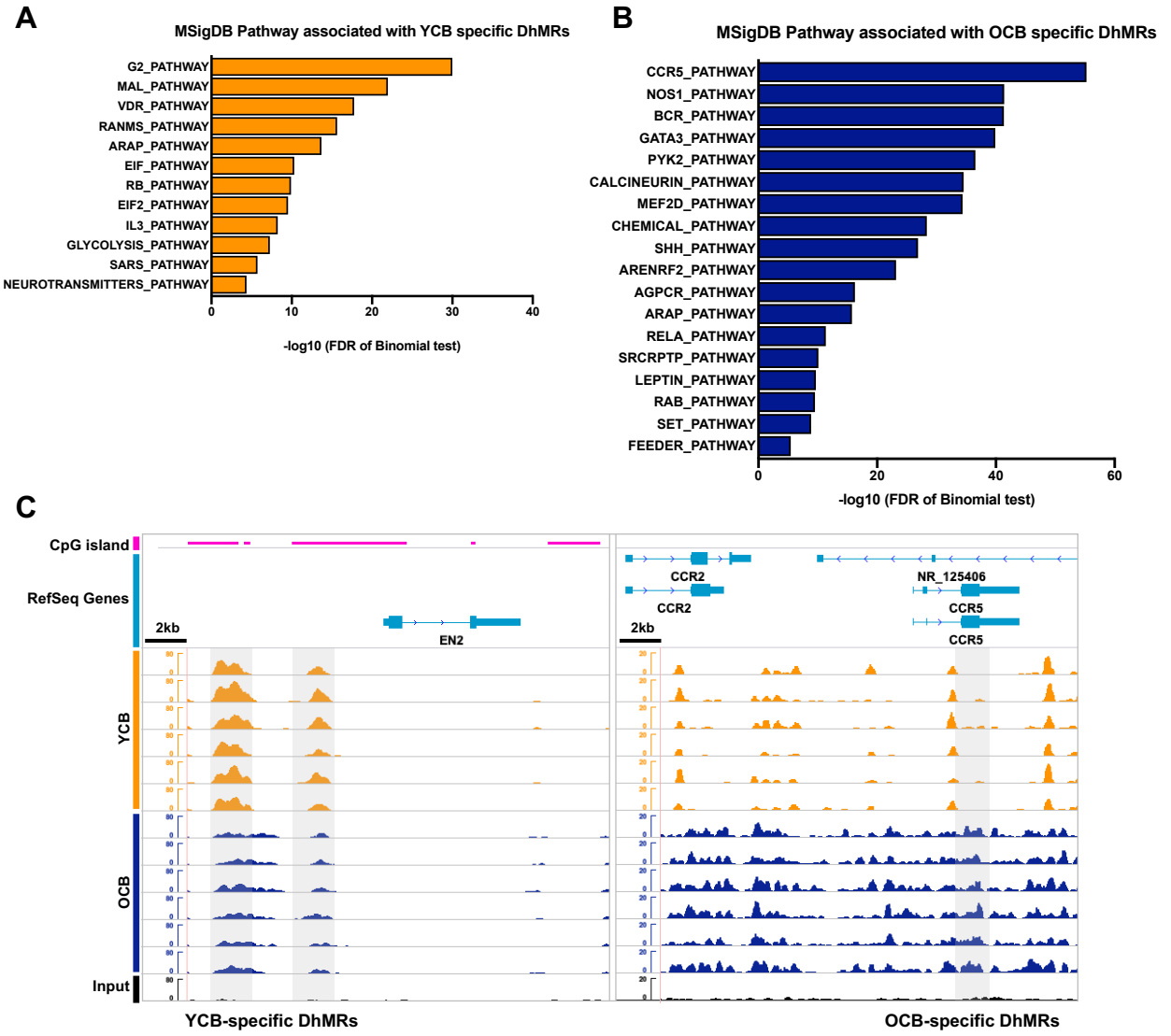


Figure 3.3. The functional relevance of age-dependent DhMRs in biological pathways. (A, B) The summary result of Molecular Signatures Database (MSigDB) enrichment analysis with GREAT. Each term shows statistical significance in binomial test ($q < 0.01$) and greater than 1.5 region-fold enrichment. (C) Representative IGV snapshot of 5hmC signals around *EN2* and *CCR5* genes. YCB- (left) and OCB- (right) specific DhMRs were highlighted as grey. Top panel shows CpG island.

Correlation between age-associated DhMRs and gene expression

In mouse ESCs, gene-body 5hmC levels show a significant positive correlation with gene expression whereas the levels at the transcription start site (TSS) are depleted in highly expressed genes (Tan et al., 2013); however, such a correlation is not replicated in mouse ESC-derived neural progenitor cells (NPCs) (Tan et al., 2013), suggesting that relationship between 5hmC enrichment at genic regions and gene expression is prominent in the undifferentiated cells, but not in differentiated cells in mouse. To explore the correlation between gene-body 5hmC levels and age-dependent differential gene expression in human, we performed total RNA sequencing. Although we clearly identified distinct 5hmC patterns in the two groups, we observed three distinct expression clusters (Figure 3.4A). Interestingly, one of the clusters (C1) is enriched in young adult (n=2, 19 years old), which showed an intermediate expression pattern to the other gene expression clusters (C2 and C3) from young children (n=3, 5-11 years old) and aged adult (n=7, 70-89 years old), respectively (Figure 3.4A). Interestingly, gene ontology (GO) analysis ($p < 0.05$) showed differential genes corresponding to each cluster, revealing that C1-associated genes were enriched at the pathways involved in cell proliferation and neural maturation (myelination, neuron migration, and neuron projection development) (Figure 3.4B). In addition, highly expressed genes in young children (categorized into C2) were involved in cell junction and embryo-development pathways, whereas genes more highly expressed in aged cerebellum were enriched at pathways of cell death and immunity, consistent with enrichment patterns of OCB-specific DhMRs (Figure 3.4B).

We further investigated gene-body 5hmC levels in each cluster. To elucidate whether 5hmC signals within YCB group are separated depending on different expression cluster, we divided YCB into two groups: YCB1 (n=4, young children, 5-11 years old) and YCB2 (n=2, young adult, 19 years old). Consistent with the correlations found in mouse ESCs, we found positive correlation between 5hmC signals and gene expression of genes in the C2 cluster, but there was no significant difference of 5hmC

patterns between YCB1 and YCB2 (Figure 3.4C). In the C2 cluster, the 5hmC levels of OCB group were high around TSS and significantly dropped after transcription end site (TES) (Figure 3.4C, left). In the C1 cluster, there was no substantial enrichment across genic regions, but we found substantial 5hmC levels were detected in both upstream and downstream of TSS in all YCB groups whereas 5hmC levels of OCB group were slightly higher within genic regions (Figure 3.4C, middle). Compared to other clusters, there was no distinct 5hmC signaling on highly expressed genes in OCB (C3 cluster) although general 5hmC levels of OCB group were lower (Figure 3.4C, right). Altogether, these data indicate that the correlation between gene-body 5hmC patterns and gene expression is only valid in genes involved in development, and therefore, another mechanism how 5hmC regulates age-dependent expression remains to be understood.

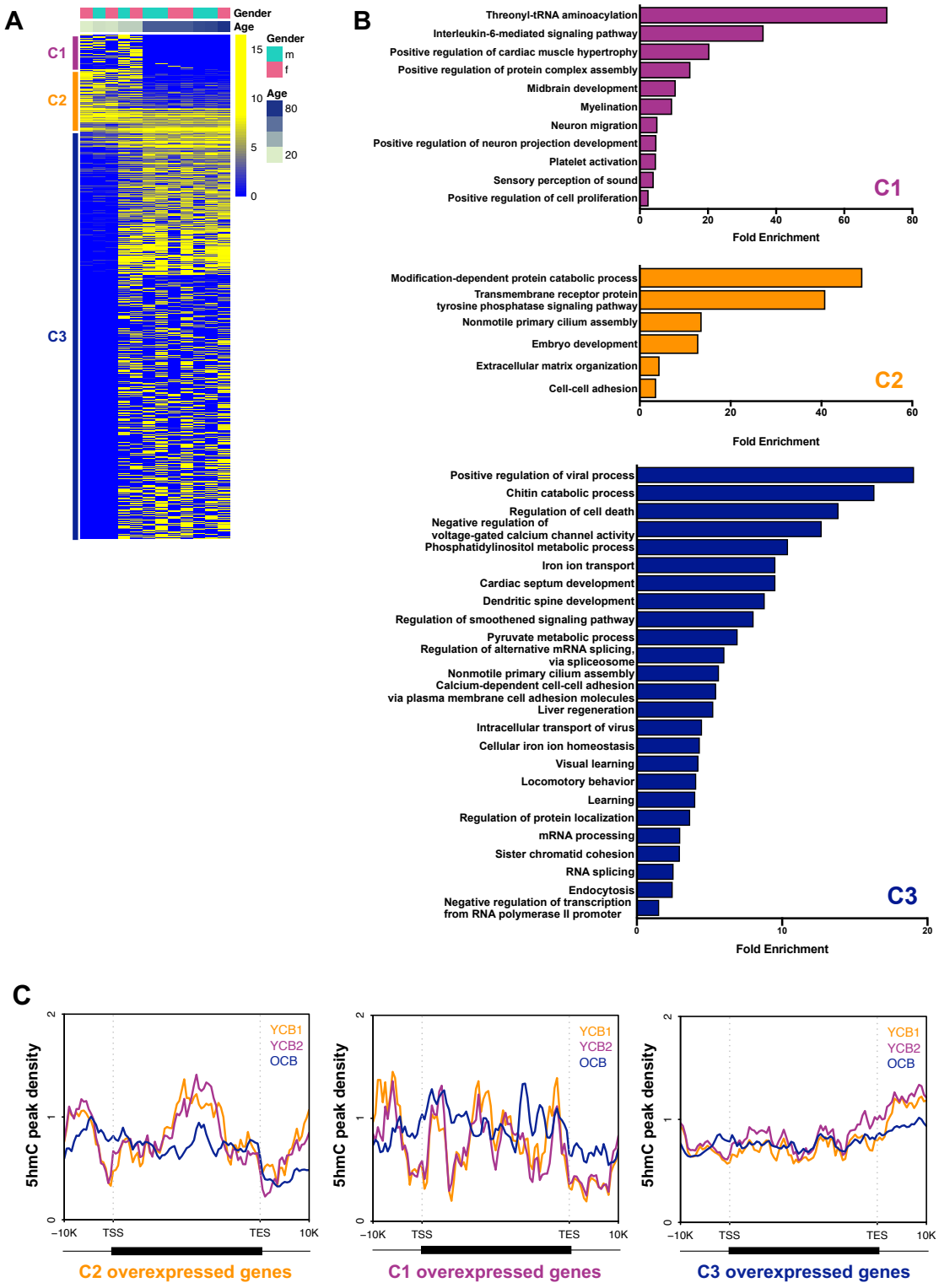


Figure 3.4. Correlation between DNA hydroxymethylation and gene expression in each age group.

Heatmap showing differentially expressed genes identified using DESeq2 (adjusted *p-value* < 0.05).

Three clusters are based on differential expression pattern. (B) Gene Ontology (GO) analysis of each cluster gene using DAVID (*p-value* < 0.05). (C) The average 5hmC densities of each age group were plotted across the gene body regions of each cluster gene. YCB1 (n=4, 5-11 years old), YCB2 (n=2, 19 years old), and OCB (n=6, 70-89 years old).

3.4. Discussion

The most prevalent form of DNA modifications is DNA methylation, the addition of a methyl group at the 5' position of cytosine. Such simple chemical modification not only can alter the binding affinity of transcription factors but can serve as a docking site of various proteins such as methyl-CpG binding domain (MBD) proteins, leading to the formation of repressive complex (Bogdanović and Veenstra, 2009). Through the mechanisms, numerous studies have shown that DNA methylation plays a critical role in fine-tuning gene expression during neurogenesis, the process of new neuron formation (Yao et al., 2016). While most of development and maturation takes place in the embryonic stages, neurogenesis and maturation occurs mainly during early life after birth and some areas of the brain are continuously producing new neurons, even in adult (Kempermann et al., 2018), suggesting differential DNA methylation state at each stage. However, underlying mechanisms for how chemically stable methylation of DNA is removed and reconfigured at the proper genomic regions has been poorly understood until the discovery of the ten-eleven translocation (TET) protein family and the oxidative derivatives of 5-methylcytosine (5mC) (Esteller, 2008; Heyn and Esteller, 2012; Tahiliani et al., 2009). Of the derivatives, 5-hydroxy-methylcytosine (5hmC), the first oxidative derivative generated by TET enzymes, is highly abundant in brain; therefore, dynamic changes of 5hmC have been intensively investigated in neurogenesis and neurological disorders. Genome-wide 5hmC profiling in mouse brain has identified distinct 5hmC enrichment at tissue-specific enhancers (Szulwach et al., 2011; Tahiliani et al., 2009; Wu et al., 2011b). In addition, compared to DNA methylation, the formation of 5hmC is sensitive to environmental cues because TET enzymes are dependent on the level of α-Ketoglutarate and oxygen, suggesting the central role of 5hmC as an indicator of dynamic cellular state (Laird et al., 2013). Nevertheless, a comprehensive understanding of 5hmC dynamics during developmental aging in human cerebellum has not been elucidated.

Here we profiled genome-wide 5hmC distribution using postmortem human cerebellum tissues of different ages without any distinguished neurological symptoms and identified the number of age-dependent differential hydroxymethylation regions (DhMRs) (Figure 3.2A). Age-dependent DhMRs were enriched at distinct genomic loci (Figure 3.2B and Figure 3.2C); DhMRs associated with cerebella of young age group (YCB) were enriched at promoters and CpG island while DhMRs associated with cerebella of the older age group (OCB) were detected within genic regions more. Interestingly, significant enrichment of YCB-specific 5hmC at neurodevelopment-associated transcription factors such as PITX1, THRb, and PTF1A, were identified, suggesting the regulatory role of 5hmC in early life. Indeed, the prediction of cis-regulatory functions of age-dependent DhMRs showed strong enrichment for cell proliferation and neural maturation (Figure 3.3A). Notably, genes near OCB-specific DhMRs were involved in immunity and oxidative stress response (Figure 3.3B). Given neurons are continuously exposed to oxidative stress during aging, 5hmC configuration can be altered accordingly, leading to the activation of the immune system as a protective response. Thus, age-specific 5hmC marks can be a determinant of epigenetic age similar to methylation status (Field et al., 2018). To further understand the correlation between gene-body 5hmC levels and gene expression, we performed transcriptome analysis. While large number of genes were differentially expressed in the OCB group, we found no significant correlation between gene-body 5hmC signals and expression, implying gene-body 5hmC mark is not a critical factor to regulate gene expression in mature neurons. Considering the strong enrichment of 5hmC at active enhancers, the investigation of age-dependent enhancers will provide a clue as to how age-dependent 5hmC marks are involved in age-associated gene expression.

In summary, genome-wide 5hmC profiling in this study revealed age-dependent dynamic changes of 5hmC and its potential role in gene expression in the cerebellum. Therefore, in the future, these data can contribute to a better understanding of age-related neurological disorders in the cerebellum.

CHAPTER 4: TET1-mediated 5-hydroxymethylcytosine Alteration in the pathogenesis of Medulloblastoma

4.1. Introduction

Medulloblastoma (MB) is the most common malignant pediatric brain tumor originating from the cerebellum and brainstem. While this embryonal tumor has a lower mutation rate than adult solid tumors, frequent somatic mutations and altered expression of epigenetic regulators, including chromatin remodeling genes and histone modifiers, highlight the substantial role of epigenetic alterations in MB (Pugh et al., 2012b; Wang et al., 2018). Indeed, methylation signature is highly correlated with transcriptional signature (Hovestadt et al., 2014b), and therefore, both are the gold standard for MB molecular stratification into four subgroups: Wingless (WNT), Sonic Hedgehog (SHH), Group 3 and Group 4 (Cavalli et al., 2017; Ellison et al., 2011; Taylor et al., 2012; Wang et al., 2018). Methylation profiles further refine the substantial intertumoral heterogeneity frequently found in Group 3 and Group 4 patients (Cavalli et al., 2017; Northcott et al., 2017). In addition to stratification for clinically relevant patients, the combination of the DNMT inhibitor 5-aza-2'deoxycytidine (5-aza-dC) and HDAC inhibitor valproic acid (VPA) effectively inhibits tumor formation in *Ptch*-associated tumors (Ecke et al., 2009). Despite the clear significance of the epigenome in MB, there have been limited studies to investigate the dynamic nature of cytosine modifications and enzymes involved in the process.

The cerebellar cortex continuously undergoes neuronal maturation and circuit formation during the postnatal period (Sidman and Rakic, 1973); thus, precise timing of gene expression is critical for neurogenesis. In mouse, while a small increase in global levels of 5-methylcytosine (5mC) is detected during cerebellar maturation, the abundance of its oxidative derivative, 5-hydroxymethylcytosine

(5hmC), dramatically increases up to approximately 0.4% to 0.9% of total cytosines, exclusively in differentiated cells located at the Purkinje layer (PL) and the internal granular layer (IGL), but not in proliferating cerebellar cells at the external granular layer (EGL) (Szulwach et al., 2011; Zhu et al., 2016). In addition, 5hmC is enriched at cerebellar specific enhancers and the exon start site of highly expressed genes involved in axon guidance and ion channels (Wang et al., 2012; Zhu et al., 2016). This evidence suggests the role of 5hmC in establishing and maintaining cell identity during the period of circuit formation.

On the contrary, 5hmC abundance is significantly reduced in many types of human malignancies, such as melanoma (Lian et al., 2012a), prostate, breast, liver, colon cancers (Uribe-Lewis et al., 2015; Yang et al., 2012), and brain tumors (Jin et al., 2011a; Kraus et al., 2012; Yang et al., 2012). Even within the same tumor, 5hmC levels vary with the different stages: high levels of 5hmC are usually observed in low grade gliomas whereas malignant gliomas tend to show low levels. In addition, the decrease in 5hmC levels correlates with shorter postoperative survival (Orr et al., 2012). Moreover, enzymatic impairment caused by somatic mutations and copy number alterations or the deregulated expression of TET enzymes are frequently identified in many cancers and often associated with unfavorable prognosis (Chou et al., 2011; Good et al., 2017; Kudo et al., 2012; Müller et al., 2012). However, the alterations of 5hmC and TET enzymes in MB and their significance to cancer formation have not been demonstrated. Here, we explore the association of 5hmC signature with cancer formation and determine whether TET enzymes could be targeted for therapeutic advantage.

4.2. Materials and Methods

Human Tissues

Medulloblastoma (MB) samples were obtained from three different sources: Aflac cancer center (n = 5), the Xiangya Hospital Department of Neurosurgery (n = 24), and Dr. Erwin G. Van Meir (n = 8). The patient data were anonymized prior to use in this study. For MB tissue samples from Aflac cancer center, molecular subgroup affiliation was determined by NanoString nCounter system using 22 MB subgroup specific gene expression profiles. The protocols were approved by the Institutional Review Board at Emory University. Twenty-four MB tissue samples ranging from 3 to 18 years old with average-risk (children older than 3 years of age with no evidence of metastatic disease and less than 1.5 cm^3 of residual disease) were collected at the Department of Neurosurgery of Xiangya Hospital. Five normal cerebellum samples were collected from patients from 3 to 18 years old with cerebral injury who underwent internal decompression as normal control for dot blot assay. Six normal cerebellum samples from 5 to 19 years old with no neurological disorders were collected from the NIH NeuroBioBank tissue repositories, which were used as normal control for dot blot assay genome-wide 5hmC profiling.

Mice

All protocols for mouse experiments were approved by the Institutional Animal Care and Use Committee (IACUC) at Emory University. $TET1^{+/-}$ mice (Dawlaty et al., 2011) were initially on a mixed C57BL/6x129S4/Sv background and were backcrossed with WT C57BL/6 mice for more than 10 generations before any experiments in this paper. SmoA1 homozygous mice (Hatton et al., 2008) were crossed with $TET1^{+/-}$ mice to generate cohorts in this study. Mice were aged and humanely euthanized upon the signs of disease-related symptoms.

Primary MB Culture

Mice were euthanized by isoflurane inhalation when they showed disease-associated symptoms including hunched posture, tilted head, and lethargy. Isolated tumor tissues were minced in sterile HBSS to obtain a single-cell suspension. To remove cell aggregates and extraneous tissue, the suspension was passed through two different size cell strainers (100 μm and 40 μm) and spun down to collect a cell pellet. The cell pellet was resuspended in Neurobasal medium supplemented with B-27 supplements, L-glutamine, sodium pyruvate, and Pen/Strep and plated at 1.5×10^6 cells per well in a 24-well plate on Matrigel-coated wells. For shRNA-TET1 treatment, wells were infected with lentivirus at a multiplicity of infection (MOI) of 6 and incubated for 5 days. For TET1 inhibitor (UC-51432) treatment (Jiang et al., 2017), wells were incubated with appropriate concentration of the chemicals for 48 hrs.

Human MB Cell Line Culture

Human MB cell lines (ONS-76, Daoy, D556, and D425) were cultured with DMEM with 10% fetal bovine serum, 100 U/ml penicillin, and 100 mg/ml streptomycin at 37C in an atmosphere of 5% CO_2.

Analysis of Gene Expression Array Datasets

Gene expression data of 273 human medulloblastoma samples and 31 human cerebellar tissue control samples were obtained from GEO Series accession numbers GSE49243, GSE12992, GSE10327, GSE37418, GSE50161, GSE44971, GSE7307, and GSE3526. Data analyses were performed using Bioconductor package, ‘simpleaffy’. Briefly, data were normalized using the gcrma algorithm and then, molecular subgroups of tumor samples unclassified in the previous studies were determined by unsupervised hierarchical clustering based on 1-Pearson correlation. Differential gene expression analysis was performed using Bioconductor package, ‘limma’, and then, vocanoplot and boxplot were generated using Bioconductor package, ‘ggplot2’.

Genomic DNA preparation and dot blot assay

Genomic DNA preparation and dot blot assay of 5hmC was performed as described previously. (Szulwach et al., 2011) DNA purification was performed by phenol-chloroform precipitation and reconstituted in DNase -free water. Image-J was used to quantify the level of 5-hmC in dot blot and then data was analyzed using GraphPad Prism 8.0 (Graphpad, Inc.). Beyond standard descriptive and graphical analyses, the association of quantitative variables was evaluated by means of t-test.

Survival analysis using human samples and mouse model

Survival of human patients was measured from the time of initial diagnosis until the date of death due to progressive disease. Disease-associated symptom-free survival of SmoA1 mice was measured from the birth date until the first date showing MB-associated symptoms such as hunched posture and tiled head. The survival distribution was estimated using Kaplan–Meier curves. Survival curves were compared by means of the log-rank test. Results were considered statistically significant when the p-value of the Log-rank (Mantel-Cox) test was below 0.01.

Genome-wide 5hmC profiling (hMe-Seal sequencing)

Tumor tissues and matched normal cerebellar tissues were used for hMe-Seal sequencing (Song et al., 2011a, 2011e) to identify differential hydroxymethylation regions (DhMRs). For labeling of 5hmC-containing genomic regions, 1 μg sonicated genomic DNA (100-300 bp) was incubated for 2 hrs at 37ºC in a 30 μl solution containing 100 μM UDP-6-N_3-Glu, β-glucosyltransferase (β-GT) and NEB buffer 4. After purification using AMPure XP beads, N_3-glucose labeled DNA was incubated for 2 hrs at 37ºC with the addition of 150μM dibenzocyclooctyne-modified biotin (click chemistry), which is enriched by streptavidin. DNA libraries were generated using NEBNext® Ultra™ II DNA Library Prep Kit, which were then ready for sequencing.

Analysis of hMe-seal sequencing data to identify DhMRs

Sequencing data were mapped to either human genome, hg19, for human MBs and age-matched normal cerebella or mouse genome, mm10, for SmoA1 MBs using bowtie2 (Langmead and Salzberg, 2012). Mapped reads were filtered and sorted using samtools (Li et al., 2009) and then PCR duplicates were removed using Picards (Broad Institute, 2016). Binned matrices (binsize: 2kb) were generated using final bam files and were used for peak identification using MACS2 with default parameters (Zhang et al., 2008), and differential hydroxymethylation regions (DhMRs) were determined using DESeq2 with default settings (Love et al., 2014). Among the DhMRs identified by DESeq2, MB or normal samples-specific genomic regions which were not called as peaks within 80 % of samples in the corresponding samples were filtered out, and final either MB or normal-specific DhMRs were identified after merging adjacent genomic regions using bedtools. Identified DhMRs were annotated using Homer (Heinz et al., 2010) and CEAS (Shin et al., 2009). To understand the biological meaning of DhMRs, we used GREAT using default "Basal plus extension" settings (McLean et al., 2010). Enrichment terms less than FDR threshold 0.01 (both region-based binomial and hypergenometric tests) and greater than 1.5 region-based fold-change were regarded as statistically significant. For motif scanning on DhMRs and TET1 binding sites, we used Homer software.

Immunoblotting

Tumor and matched normal tissues were collected from euthanized SmoA1 mice with MB-associated symptoms. Tissues were rinsed with ice-cold PBS, homogenized after the addition of radioimmuno-precipitation assay (RIPA) buffer supplemented with protease inhibitor cocktail. After incubation on ice for 20 min, lysates were then centrifuged in a microfuge at 13,000 rpm for 15 min, and the supernatants were quantified using Bicinchoninic acid (BCA) assay and 50 μg of each sample was loaded in 10% and 6% of acrylamide gels for GAPDH and TET1 detection, respectively. All immunoblotting was repeated

at least three times. For quantitative analysis, autoradiographic films were scanned with an Epson 1680 scanner, and the captured image was analyzed with NIH ImageJ software.

Brain transcardiac perfusion and Histology

For histological staining, 12 week old mice from the different genotype backgrounds with the presence or absence of tumor were transcardially perfused with 4% paraformaldehyde in PBS. Brains collected were post-fixed with 2% paraformaldehyde in PBS overnight, cryoprotected in 20%, then 30% (w/v) sucrose in PBS at 4ºC, and rapidly frozen. Cryostat sections (10 μm) were stained with Hematoxylin and Eosin according to the previous Cold Spring Harbor protocol (Fischer et al., 2008).

RNA Extraction and RT-PCR

RNA was extracted from pellets using Trizol reagent (Thermo Fisher Scientific) according to the manufacture's procedure. After Nanodrop quantification of RNA, 1 μg of RNA was used to generate cDNA with SuperScript III First-Strand Synthesis System for RT-PCR. Quantitative PCR for mRNA of *TET1* and proper internal control (*GAPDH* for Mouse and Actin for Human) detection was carried out using SYBR green (Thermo Fisher Scientific) and a 7500 Fast Real-Time PCR machine (Applied Biosystems) with an initial denaturing step at 95ºC for 10 min, then 40 cycles of PCR (95ºC for 15 s, 60ºC for 1 min) and a further extension at 60ºC for 10 min.

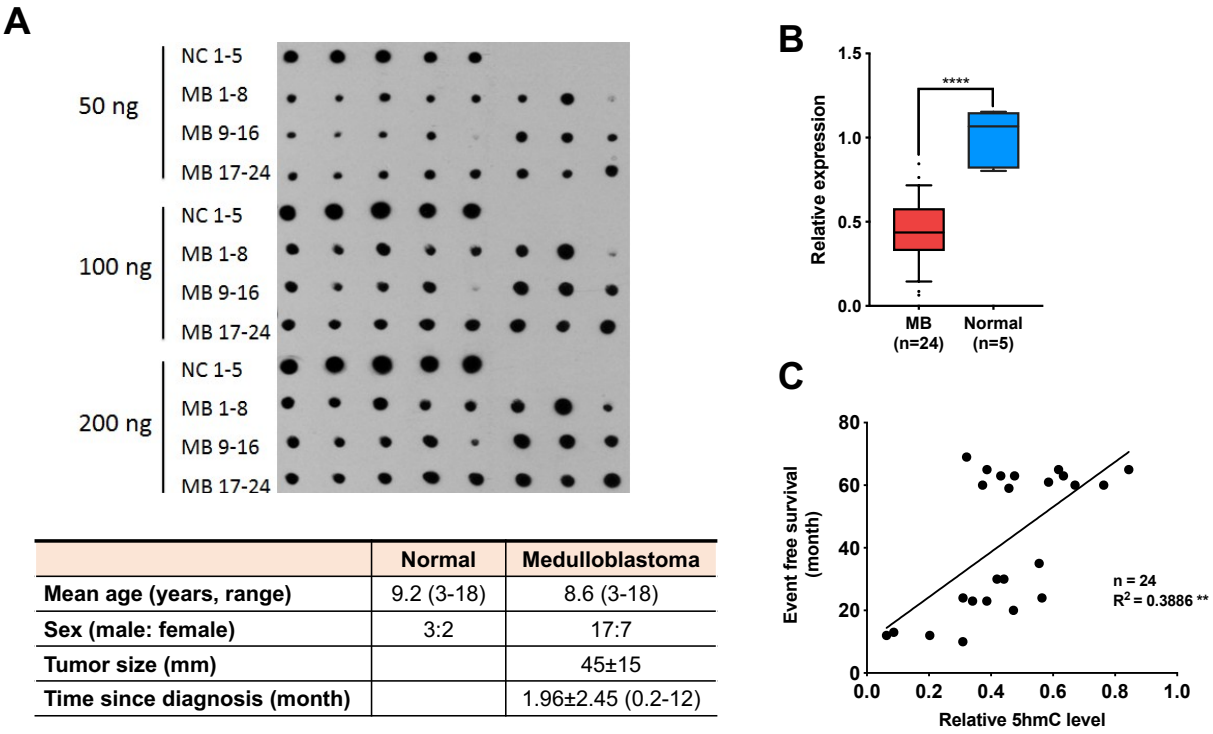
Cell viability Assay

To assess cell viability after treatment of shRNA and TET1 inhibitor, CellTiter-Blue Cell viability Assay (Promega) was used. Briefly, 20 μl of solution was added to each well directly 1 hr before measurement. The fluorescence was measured using FLUOstar Omega (BMG Labtech) microplate reader. All measurements were taken in triplicate and each experiment was replicated at least three times.

4.3. Results

Decrease in 5hmC level is associated with MB prognosis

Previous studies have found significant depletion of 5hmC abundance in many types of human cancer compared to corresponding normal tissues (Ficz and Gribben, 2014; Jin et al., 2011b; Lian et al., 2012b). To further explore this epigenetic characteristic in MB, we first examined 5hmC levels using a dot blot assay with the postmortem cerebellar tissues without neurological disorders cerebellar tissues (n=5) and primary MB tissues (n=24). In the normal tissues, 5hmC levels were confined to a narrow range (0.81 to 1.15-fold compared to the mean value), although the age of the patients in the control group varied from 3 to 18 years old (Figure 4.1A) (Wang et al., 2012). In contrast to the high 5hmC levels found in the normal tissues, we identified a substantial decrease of global 5hmC levels of MB tissues ($p < 0.001$; Figures 4.1A and 4.1B) with an average of 0.45-fold 5-hmC relative to normal. In the independent cohort, we consistently observed a significant reduction of 5hmC levels in MB (n=5) compared to age-matched normal without marked neurological disorders (n=6) using both dot blot assay and HPLC (Figures S3A and S3B), suggesting that the reduction of 5hmC is a hallmark for MB. However, unlike the relatively stable 5hmC levels in normal, MB exhibited significant intertumoral variations of 5hmC levels (0.04 to 0.96-fold difference, Figure 4.1C); in some cases, 5hmC levels were almost undetected (e.g. patient NO.13), whereas 5hmC levels were almost at the same levels as in normal among several patient cases (e.g. patient NO.7) (Figure 4.1A). Consistent with the notion that low 5hmC levels are associated with poor prognosis in glioblastoma, the most common brain tumor in adults (Orr et al., 2012), we identified a notable linear correlation between 5hmC levels and prognosis ($n = 24$, $R^2=0.3886$, $p < 0.001$; Figure 4.1C), which is a better indicator of prognosis than age at diagnosis (Figure S3C).



	Normal	Medulloblastoma
Mean age (years, range)	9.2 (3-18)	8.6 (3-18)
Sex (male: female)	3:2	17:7
Tumor size (mm)		45±15
Time since diagnosis (month)		1.96±2.45 (0.2-12)

Figure 4.1. Loss of 5-hydroxymethylation is a hallmark of MBs. (A, B) Top, 5hmC dot blot analysis shows a significant decrease of total 5hmC levels in MBs (n=24) compared with age-matched normal cerebella (n=5) ($p < 0.001$). Bottom, Table illustrating sample information (age, sex, tumor size, and time since diagnosis) of MBs and normal cerebella. (C) Linear correlation between 5hmC abundance in MBs and prognosis. Low 5hmC level is associated with worse prognosis ($p < 0.01$).

Deregulated expression of TET proteins in MBs

DNA methylation is a prerequisite for 5hmC generation *in vivo* (Figure 4.2A); therefore, genomic hypomethylation can lead to loss of 5hmC in tumors. Another plausible mechanism to explain 5hmC depletion in tumors is either inactivating mutations of TET proteins (TET1, TET2, TET3) which are responsible for converting 5mC to 5hmC or indirect inhibition of enzymatic activity caused by *IDH1/2* mutations (Massé et al., 2009; Xu et al., 2011a). Only three patients among 300 cases have putative driving mutations in either TET proteins (truncating mutations of TET1 or TET2) or IDH1 (missense mutation), which cannot account for loss of 5hmC in MBs in general (Figure S3D). To explore the deregulated expression pattern of enzymes involved in 5hmC generation, we performed meta-analysis using 8 different publicly available gene expression datasets containing 273 human MB samples and 31 human cerebellar tissue control samples (Table S10). Notably, we identified significant up-regulation of *DNMT3A*, *TET1* and *TET2* in MBs compared to adult cerebellar tissues (adjusted p-value < 0.05 and log2 fold changes > 2, Figure 4.2B). In addition, the expression of *DNMT3A*, *TET1* and *TET2* in MBs was comparable with the expression in fetal tissues (Figures 4.2C and S3E). While there is intertumoral variation of expression levels in *TET1* and *TET2*, there is no significant difference of expression levels across MB molecular subgroups (Figure S3F). Considering that high levels of *TET1* and *TET2* mediate epigenetic re-programming by passive and active DNA demethylation process in primordial germ cells (PGCs) (Hackett et al., 2013; Hashimoto et al., 2013; Vincent et al., 2013), high levels of *TET1* and *TET2* in MBs can result in global hypomethylation found in MBs (Hovestadt et al., 2014b).

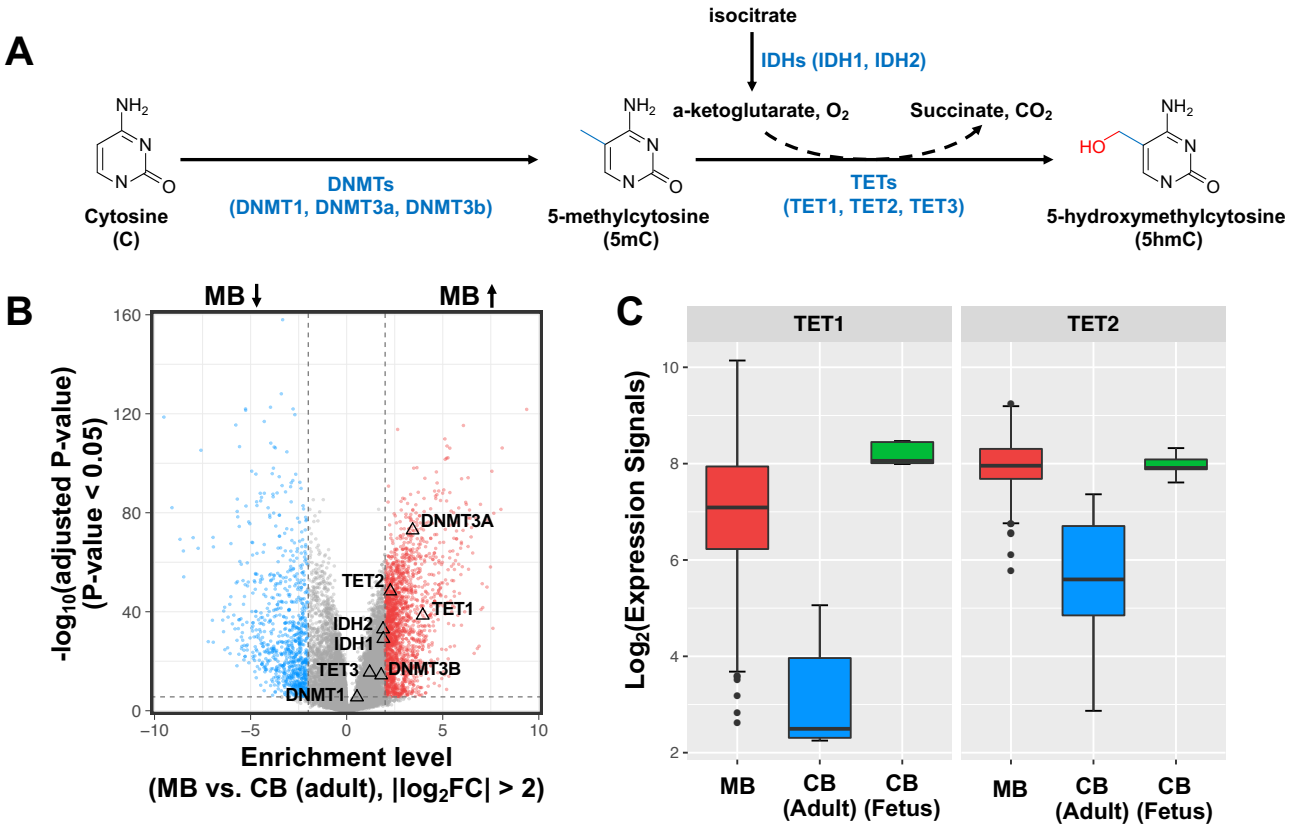


Figure 4.2. Deregulated expression of TET proteins in MBs. (A) Diagram of cytosine modifications. Cytosine can be methylated by DNA methyl-transferases (DNMT1, DNMT3A, and DNMT3B), and further oxidized by Fe (II)/a-ketoglutarate dependent TET proteins (TET1, TET2, and TET3), generating 5-hydroxymethylcytosine (5hmC). IDH1 and IDH2 are also indirectly involved in 5hmC production by regulating the level of a-ketoglutarate. (B) Volcano plot showing distribution of differential expression in MBs, with log2 fold change of Tumor/normal on x-axis and P values on y-axis. Genes with absolute log2 fold change ≥ 2 and adjusted p-value < 0.05 are indicated in either red (upregulated in MBs) or blue (downregulated in MBs). (C) Expression level of *TET1* and *TET2* in MBs (n=273), normal adult cerebella (CB-adult, n=26), and fetal cerebella (CB-fetus, n=5).

5hmC signature is distinct in MBs

Global 5hmC levels are highly variable depending on tissues of origin and developmental stage, but the genomic distribution of 5hmC is highly tissue-specific (Li and Liu, 2011; Nestor et al., 2012), inspiring us to investigate differentially hydroxymethylated genomic regions (DhMRs) in MBs by employing a previously established chemical labeling and affinity purification method coupled with high-throughput sequencing (Song et al., 2011a). Principal component analysis showed substantial similarity of 5hmC patterns in normal cerebella (n=6) regardless of ages, but tumors (n=16) showed divergent 5hmC patterns (Figures S4A and S4B and Table S11). We first identified 87,830 DhMRs showing increased hydroxy-methylation in MBs (5hmC gain in MBs) and 2,222 DhMRs showing decreased hydroxymethylation in normal samples (5hmC loss in MBs) using DESeq2. Due to inter-tumor heterogeneity, many 5hmC gain identified in initial steps were found only in 1 or 2 samples out of 16. To identify more common 5hmC gain, we further filtered out any DhMRs not detected as peaks within at least 80% of corresponding either MBs or normal samples, and then finally identified 9,766 5hmC gain in MBs and 1,965 5hmC loss in MBs (Figure 4.3A). 5hmC gain showed greater than 2-fold enrichment at chromosome 17, 20, and 22, whereas 5hmC loss was mainly enriched at chromosome 10 and 19 (Figure S4B). Given the significant association between elevated C to G transversions with asymmetrically hydroxymethylated sites in cancer genomes (Supek et al., 2014), further investigation is needed to determine the direct correlation between 5hmC alterations and mutagenic events in MBs.

5hmC gain is enriched at regulatory regions of genes involved in stem-cell like properties

Recent studies have revealed significant enrichment of 5hmC at gene bodies of actively transcribed genes in embryonic stem cells (ESCs) and cis-regulatory regions such as promoter and enhancer regions (Hahn et al., 2013; Sardina et al., 2018; Wu et al., 2011b; Yao et al., 2014). Interestingly, 5hmC gain in MBs was significantly located at promoter regions (log2 enrichment=1.721) and transcription start sites (log2

enrichment=1.752), whereas 5hmC loss in MBs was enriched at genic regions including exon (log2 enrichment=1.399) and pseudo genes (log2 enrichment=1.666), indicating that 5hmC gain can play a more active role in gene transcriptional regulation (Figure 4.3B and Table S12). Notably, 5hmC signals of MBs were highly enriched at previously reported MB enhancer regions (Lin et al., 2016) while 5hmC signals of normal samples were even across the background genome (Figures 4.3C and 4.3D). To explore the motif enrichment of 5hmC gain in MBs, we performed motif analysis with HOMER (Heinz et al., 2010) and identified that 5hmC gain was enriched at over 40% of background sequences with SCL and NANOG binding motifs (q-value < 0.05, Figure 4.3E and Table S12). SCL is a basic helix-loop-helix (bHLH) transcription factor which was initially identified as playing a critical role in hematopoiesis (Elefanty et al., 1998), but recent studies have identified its elevated gene expression in post-neurogenic periods and its key role in neuronal growth and brain morphological development (Bradley et al., 2006; Herberth et al., 2005). 5hmC gain was also highly enriched at other bHLH transcription factors including PTF1A (31.04%), HEB (23.79%), and OLIG2 (18.81%) (Table S13), which are involved in neurogenesis as well as maintenance of stem-cell like properties (Hoshino et al., 2005; Li et al., 2017; Schüller et al., 2008). NANOG is a homeobox transcription factor which is an essential factor to maintain self-renewal and cell growth of human ESCs (Pan and Thomson, 2007). NANOG is overexpressed in MB stem cells, which potentially prevents neuronal differentiation and maintain stemness of MBs (Po et al., 2010). 5hmC gain is also found in about 15% of binding sites of LIM homeobox gene families (LHX1, LHX2, and LHX3), suggesting that 5hmC gain has the potential to regulate super-enhancer regions in MBs (Lin et al., 2016). Functional prediction of cis-regulatory regions using GREAT identified that 5hmC gain was enriched at genes involved in the Notch signaling pathway (Figure 4.3F and Table S14). The activation of Notch signaling not only induces stem-like markers and cell growth in tumors but also results in drug resistance through the up-regulation of multidrug resistance ABC transporters (Barnes et al., 2006; Ee et al., 2011). To further investigate if DhMRs in MBs are indeed involved in controlling

stem-like properties, we compared genome-scale patterns of 5hmC gain and loss with normal fetal or adult DhMRs that were published previously (Szulwach et al., 2011). As expected, 5hmC containing genomic regions from fetus were prone to mapping at 5hmC gains (R^2=0.4319, adjusted p-value < 0.001, Figure 4.3G, Top), whereas 5hmC containing genomic regions from adult substantially mapped at 5hmC loss (R^2=0.8035, adjusted p-value < 0.001, Figure 4.3G, Bottom). All these things together suggest that distinct 5hmC patterns in MBs are reminiscent of fetal 5hmC, which can play a key determinant in maintaining stem-like properties in tumorigenesis.

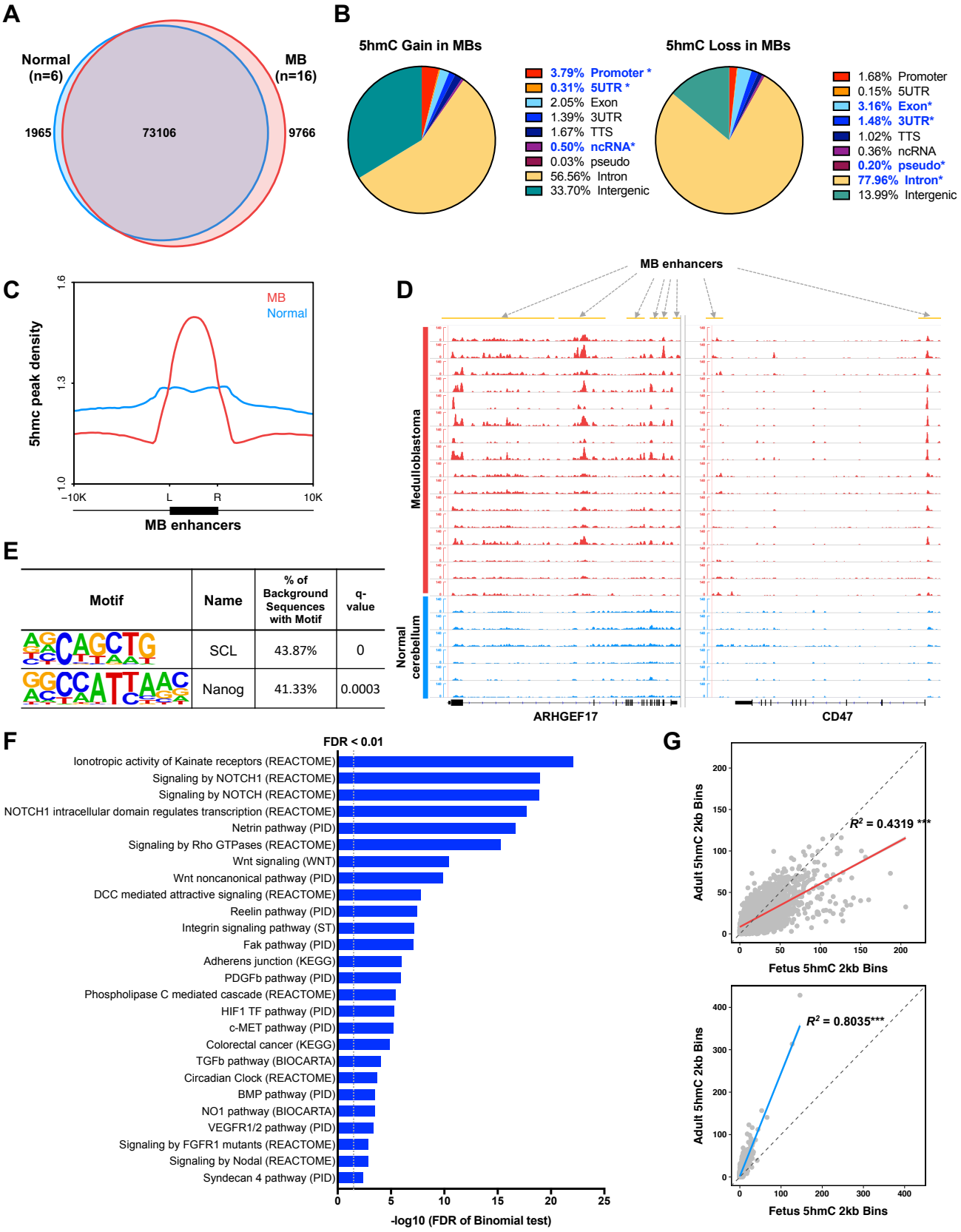
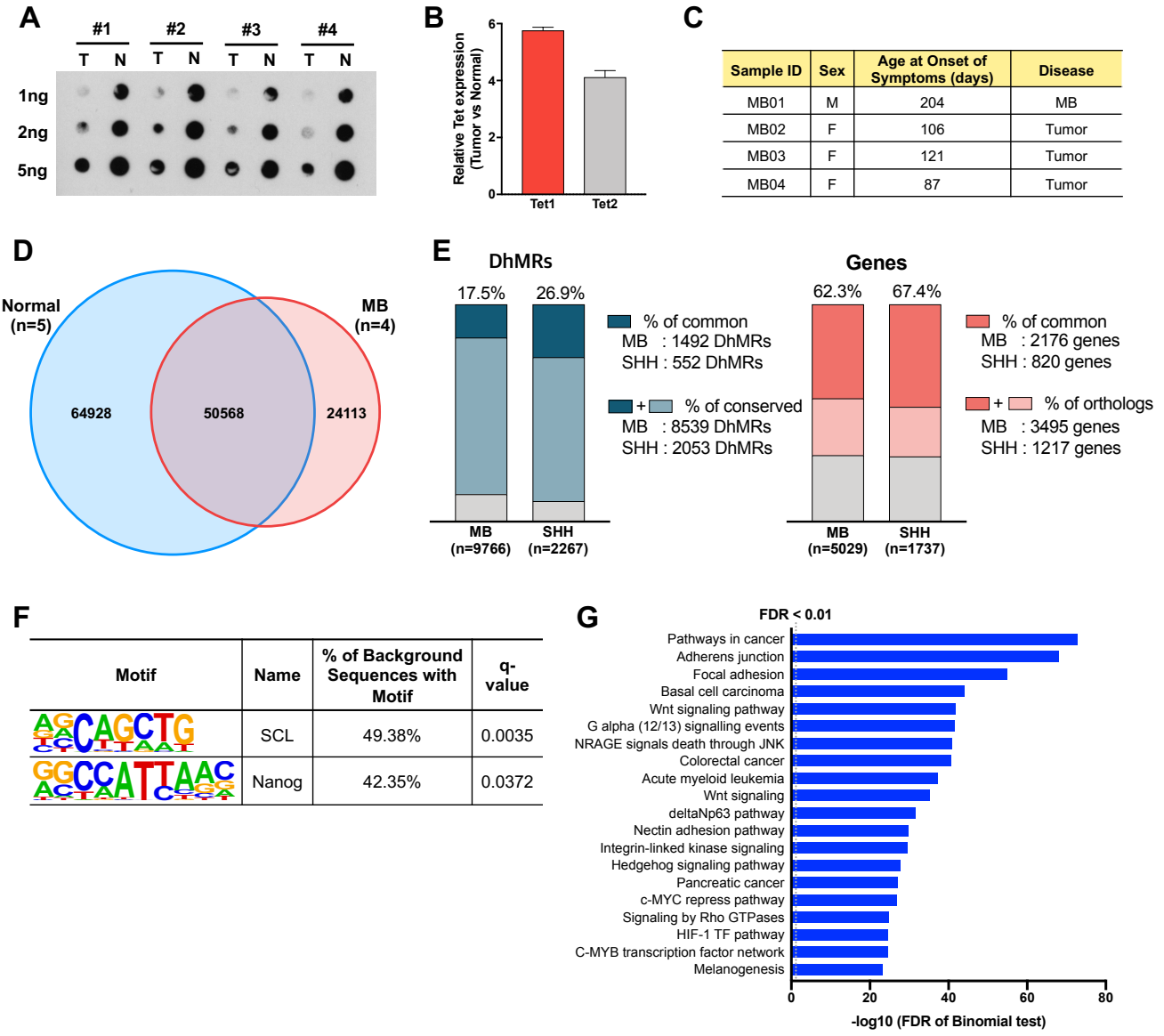

A
Normal (n=6)
MB (n=16)
1965
73106
9766
B
5hmC Gain in MBs
3.79% Promoter *
0.31% 5UTR *
2.05% Exon
1.39% 3UTR
1.67% TTS
0.50% ncRNA*
0.03% pseudo
56.56% Intron
33.70% Intergenic
5hmC Loss in MBs
1.68% Promoter
0.15% 5UTR
3.16% Exon*
1.48% 3UTR*
1.02% TTS
0.36% ncRNA
0.20% pseudo*
77.96% Intron*
13.99% Intergenic
C
MB
Normal
5hmc peak density
1.6
1.3
1.0
−10K
L
R
10K
MB enhancers
D
MB enhancers
Medulloblastoma
Normal cerebellum
ARHGEF17
CD47
E
Motif
Name
% of Background Sequences with Motif
q-value
SCL
43.87%
0
Nanog
41.33%
0.0003
F
FDR < 0.01
Ionotropic activity of Kainate receptors (REACTOME)
Signaling by NOTCH1 (REACTOME)
Signaling by NOTCH (REACTOME)
NOTCH1 intracellular domain regulates transcription (REACTOME)
Netrin pathway (PID)
Signaling by Rho GTPases (REACTOME)
Wnt signaling (WNT)
Wnt noncanonical pathway (PID)
DCC mediated attractive signaling (REACTOME)
Reelin pathway (PID)
Integrin signaling pathway (ST)
Fak pathway (PID)
Adherens junction (KEGG)
PDGFb pathway (PID)
Phospholipase C mediated cascade (REACTOME)
HIF1 TF pathway (PID)
c-MET pathway (PID)
Colorectal cancer (KEGG)
TGFb pathway (BIOCARTA)
Circadian Clock (REACTOME)
BMP pathway (PID)
NO1 pathway (BIOCARTA)
VEGFR1/2 pathway (PID)
Signaling by FGFR1 mutants (REACTOME)
Signaling by Nodal (REACTOME)
Syndecan 4 pathway (PID)
0
5
10
15
20
25
-log10 (FDR of Binomial test)
G
Adult 5hmC 2kb Bins
Fetus 5hmC 2kb Bins
R² = 0.4319 ***
R² = 0.8035***

Figure 4.3. 5hmC gain in MBs is implicated in stem-like properties. (A) Venn diagram of DhMRs identified by hMe-seal sequencing using 16 MB samples and 6 age-matched normal samples. (B) Pie charts illustrating summary of DhMRs annotation using HOMER. Annotations with absolute fold change > 2 and adjusted p-value < 0.05 from background are indicated in blue. (C) Average plots of 5hmC containing 2kb bins around previously reported MB active enhancers. (D) Representative IGV snapshot at promoter regions and genic regions of *ARHGEF17* and *CD47* showing signals for hMe-seal seq in MBs and corresponding normal samples. Top panel shows MB active enhancers. (E) Sequence logos are shown for the highly enriched sequence motifs in 5hmC acquiring genomic regions in MBs. (F) Molecular Signatures Database (MSigDB) enrichment analysis with GREAT. Each term shows statistical significance in both binomial test ($q < 0.01$) and hypergeometric test ($q < 0.05$) as well as greater than 1.5 region-fold enrichment. (G) Plots using mapped 5hmC containing bins of fetus and adult samples at either 5hmC gained peaks in MBs (n=9,766, top) or 5hmC lost peaks in MBs (n=1,965, bottom). Sequencing reads of fetus and adult are used to generate binned matrix (binsize: 2kb). Linear regression analysis determines statistical significance (adjusted p-value < 0.001, R^2=0.4319 and R^2=0.8035).

5hmC signature of SmoA1-MBs recapitulate 5hmC signature of human MBs

Human cell line models are a good *in vitro* tool to understand oncogenic processes by manipulating the expression of target genes using siRNA/shRNA or ectopic overexpression by transfection. However, 5hmC signatures are not maintained during the transition from *in vivo* tissue to *in vitro* culture, mainly due to the strong reduction of TET expression (Ficz and Gribben, 2014; Nestor et al., 2012). Therefore, to further explore the role of 5hmC and Tet proteins in MB progression, we utilized the SmoA1 mouse model, which expresses constitutively active Smo specifically in granule neuron precursors (GNPs) (Hallahan et al., 2004; Hatton et al., 2008). Consistent with human MBs, global 5hmC levels were significantly depleted in murine tumors compared to surrounding normal tissues, and *Tet1* and *Tet2* were overexpressed compared to adjacent normal (Figures 4.4A and 4.4B). With this model, we profiled the genome-wide 5hmC distribution using 4 MBs from SmoA1 mice (SmoA1-MBs) and 5 normal samples from adjacent normal tissues as well as tissues from age-matched C57BL/6J mice (Figure 4.4C). Similar to human MBs, 5hmC profiles were distinct in SmoA1-MBs (Figures S5A and S5B), and we identified 24,113 5hmC gain and 64,928 5hmC loss in SmoA1-MBs (Figure 4.4D). To determine whether 5hmC signature of SmoA1-MBs recapitulate what we found in human MB, we investigated the similarities of 5hmC gain in SmoA1-MBs with 5hmC gain in human MBs. We first compared 5hmC genomic regions from different species directly. Since SmoA1 mouse model was developed to produce mice with a high incidence of Hedgehog (Hh) signaling associated MBs (SHH-MB patients), we additionally identified 2,267 5hmC gain exclusively found in 4 SHH-MB samples (Figure 4.4E), and then, identified the conserved regions of 5hmC gain from either SHH-MBs or all MBs including different subgroups in the mouse genome (conservation rates from the human genome to the mouse genome were 90.6% and 87.4%, respectively) (Figure 4.4E and Tables S11). 26.9% of 2,053 conserved 5hmC regions associated with SHH-MBs were commonly identified in SmoA1-MBs (Figure 4.4E). Intriguingly, a substantial number of conserved 5hmC regions associated with all MBs (17.5% of 8,539 conserved 5hmC regions)

were detected in 5hmC gain of SmoA1-MBs (Figure 4.4E). Given that 5hmC plays an important role in regulating gene expression and context-dependent 5hmC signature may not be well conserved between species, we also explored how much genes nearby human 5hmC gains are overlapped with genes nearby SmoA1 5hmC gain (Figure 4.4E). Approximately 65% of mouse orthologs of genes nearby 5hmC gain in all human MBs (mouse orthologs=3,495, total=5,029) as well genes nearby 5hmC gain in SHH-MBs (mouse orthologs=1,217, total=1,737) were overlapped with genes nearby 5hmC gain in SmoA1-MBs (Figure 4.4E). In addition, 5hmC gain in SmoA1-MBs was substantially located at promoter regions (log2 enrichment=1.28) and transcriptional start sites (log2 enrichment=1.532) consistent with human annotation features (Figure S5D and Table S15). Motif analysis and functional prediction of cis-regulatory regions using GREAT analysis exhibited high concordance with the 5hmC signature identified in human MBs (Figures 4.4F and 4.4G, Tables S16 and S17). Altogether, these data indicate a strong epigenetic similarity between SmoA1- MBs and human MBs regardless of molecular subgroups.



Sample ID	Sex	Age at Onset of Symptoms (days)	Disease
MB01	M	204	MB
MB02	F	106	Tumor
MB03	F	121	Tumor
MB04	F	87	Tumor

Motif	Name	% of Background Sequences with Motif	q-value
	SCL	49.38%	0.0035
	Nanog	42.35%	0.0372

Figure 4.4. 5hmC signature of SmoA1-MBs recapitulates 5hmC signature of human MBs. (A) 5hmC dot blot analysis using SmoA1-MBs (n=4) and surrounding normal tissues (n=4). Left column indicates the total amount of DNA used in this study. (B) Relative mRNA expression level of *Tet1* and *Tet2* in SmoA1-MBs compared to normal. Each region is normalized using *Gaphd* signal. (C) Summary of 4 SmoA1-MBs used for 5hmC profiling. (D) Venn diagram of DhMRs identified by hMe-seal sequencing using 4 SmoA1-MBs and 5 normal tissues. (E) Bar graphs displaying commonly identified DhMRs and genes in both human MBs (either 16 all human MBs or 4 SHH-MBs) and SmoA1-MBs.

For common DhMR identification, human 5hmC gained DhMRs (hg19) are converted to mouse DhMRs (mm10) using batch coordinate conversion (liftOver), and then common DhMRs are identified using intersectBed (bedtools). For common gene identification, mouse orthologs corresponding to human genes nearby human tumor-associated peaks are identified using BioMart and then, compared with genes nearby murine tumor-associated peaks. The numbers next to bar graph indicate the number of peaks or number of genes for that respective category. (F) Sequence logos are shown for the highly enriched sequence motifs in 5hmC acquiring genomic regions in murine MBs. (F) Molecular Signatures Database (MSigDB) enrichment analysis with GREAT using 5hmC gained peaks in murine MBs. All statistical tests were performed using the same parameters used in human MB data analysis.

Tet1 is a key enzyme to modulate MB progression

As *TET1* and *TET2* were consistently overexpressed in both human and murine MBs (Figure 4.2C and Figure 4.4B), we examined whether loss of Tet1 or Tet2 expression may alter progression of murine MBs. Interestingly, SmoA1$^{+/+}$ mice crossing with *Tet1*$^{+/-}$ displayed dramatic delay of the age-of-onset and a decrease in incidence of MB ($p < 0.0001$; log rank test, Figure 4.5A left), but no significant change in the age-of-onset was observed in SmoA1$^{+/+}$ mice crossing with *Tet2*$^{+/-}$ (p=0.5830; log rank test, Figure 4.5A right). The same phenomenon was observed in SmoA1$^{+/-}$ mice crossing with either *Tet1*$^{+/-}$ or *Tet2*$^{+/-}$ mice (Figure S6). MBs derived from SmoA1$^{+/-}$;*Tet1*$^{+/-}$ mice showed higher global 5hmC levels than tumors from SmoA1$^{+/-}$ (Figure 4.5B), indicating that Tet1 is not a major enzyme for increasing global 5hmC levels and 5hmC levels. We then determined the correlation between Tet1 expression and age-of-onset. Consistent with mRNA expression level, Tet1 protein levels were significantly elevated in tumor tissues compared to corresponding normal tissues ($p < 0.05$) and exhibited significant inverse correlation with age-of-onset (Pearson R^2=0.5059, p=0.0366; Figure 4.5C). Upon general brain size examination using 12 weeks old mice, SmoA1$^{+/-}$ mice showed significantly larger cerebellar size than SmoA1$^{+/-}$;*Tet1*$^{+/-}$ mice ($p < 0.001$), but there was no significant size difference in cerebral cortex (Figure 4.5D). In addition, histological examination revealed a high incidence of hyperplasia and invasive tumors as well as an increase in abnormal foliation rate in 12 weeks old SmoA1$^{+/-}$ mice, whereas we observed normal cerebellar morphology from the most of SmoA1$^{+/-}$;*Tet1*$^{+/-}$ mice (Figure 4.5E). Taken together, these results indicate that Tet1 plays an essential factor for abnormal proliferation in MBs.

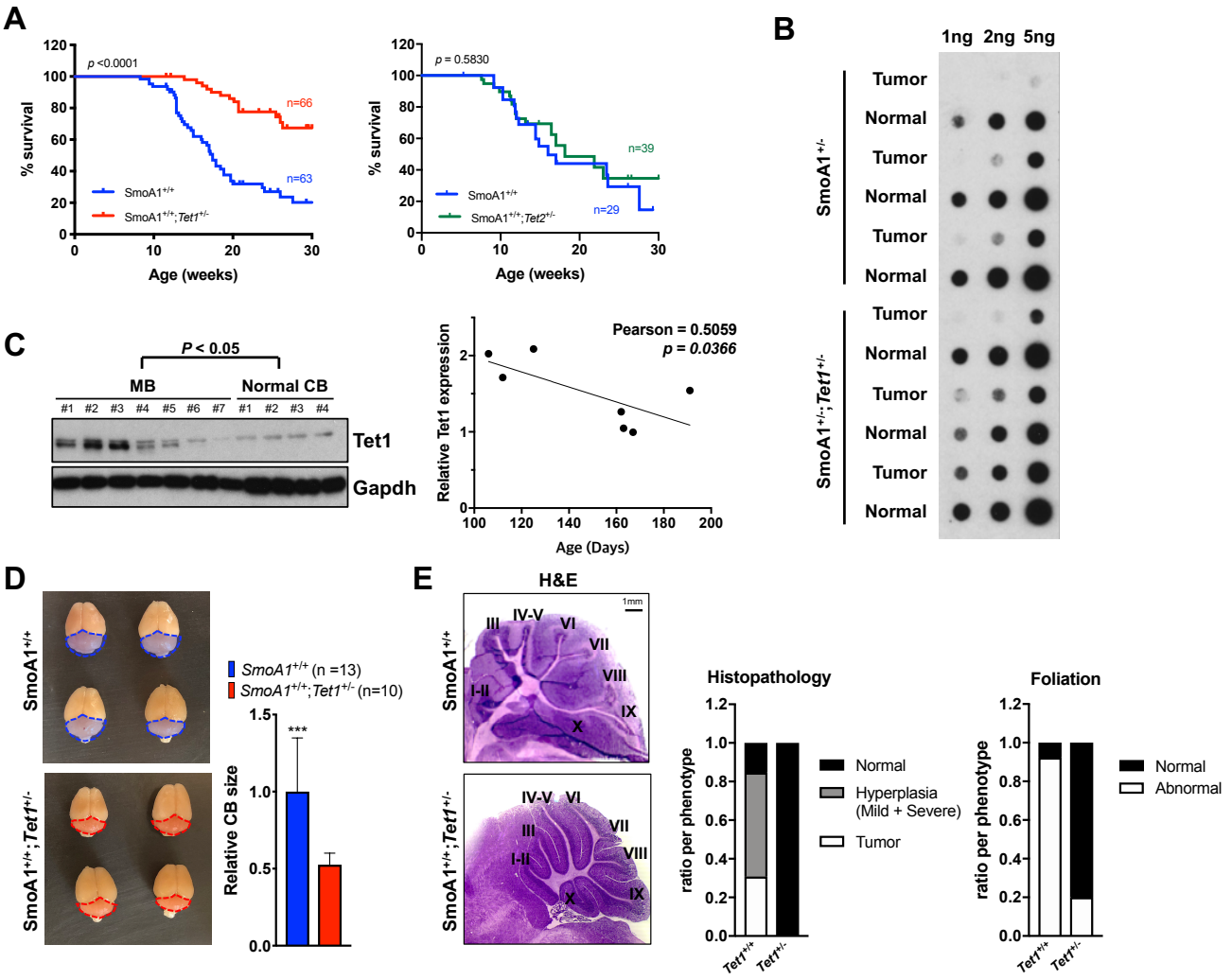


Figure 4.5. Elevated Tet1 is essential for MB progression. (A) Kaplan-Meier curves show the significant survival difference of SmoA1$^{+/+}$ mice when crossed with the hemizygous deletion of *Tet1* (Left: p < 0.0001; log rank test), but not when crossed with the hemizygous deletion of *Tet2* ($p = 0.5830$; log rank test). (B) 5hmC dot blot analysis shows elevated 5hmC levels in SmoA1$^{+/+}$;*Tet1*$^{+/-}$ mice compared to SmoA1$^{+/+}$. (C) Left: Tet1 protein expression in SmoA1-derived MBs (n=7) and corresponding normal (n=4). Right: Pearson correlation between Tet1 expression and age-of-onset (Pearson $R^2 = 0.5059$, $p = 0.0366$). (D) Brain size examination of 12-weeks-old SmoA1$^{+/+}$ mice in the presence of either wild-type or hemizygous deletion of *Tet1* (p < 0.0001; Welch's t-test). (E) H&E staining of 12-weeks-old SmoA1$^{+/+}$ and SmoA1$^{+/+}$;*Tet1*$^{+/-}$ mice cerebella. Scale bar, 1 mm.

Inhibition of Tet1 expression in MB cells leads to tumor growth *in vitro*

To determine whether abrogation of Tet1 expression attenuates cell growth, we treated with small hairpin RNAs (shRNA) targeting *Tet1* in primary SmoA1-derived MB cells. shRNA-mediated *Tet1* inhibition resulted in a dramatic decrease in cell viability (Figure 4.6A). To further investigate whether pharmacological inhibition of *Tet1* shows the same therapeutic effect on MB cells, we used UC-514321, a small molecule that suppresses the expression of *Tet1* by inhibiting binding of STAT transcription factors (STAT3 and STAT5) at promoter regions of *Tet1* in TET1-overexpressed acute myeloid leukemia (AML) (Jiang et al., 2017). Consistent with shRNA-mediated *Tet1* inhibition, we observed a dose-dependent cytotoxic effect of the inhibitor in primary SmoA1-derived MB cells, but not in normal neuronal stem cells (NSCs) (Figures 4.6C and 4.6D), indicating that this inhibitor can selectively induce cell death of abnormally proliferating cells. We then examined whether the pharmacological benefit observed in SmoA1-MBs can be achieved in human MBs by using human MB cell lines (Figure S7A). Notably, only TET1 expressing lines (Daoy, ONS-76, and D556, TET1-positive MBs) were responsive to the inhibitor, but there was no effect on non-TET1 expressing cell line (D425, TET1-negative MB) (Figures 4.6E, 4.6F, and Figure S7A). In summary, these data indicate that UC-514321 targeting overexpressed TET1 selectively suppresses the growth of TET1-positive MBs without adverse effect on normal NSCs, and thereby, TET1 is a promising therapeutic target to treat TET1-positive MBs.

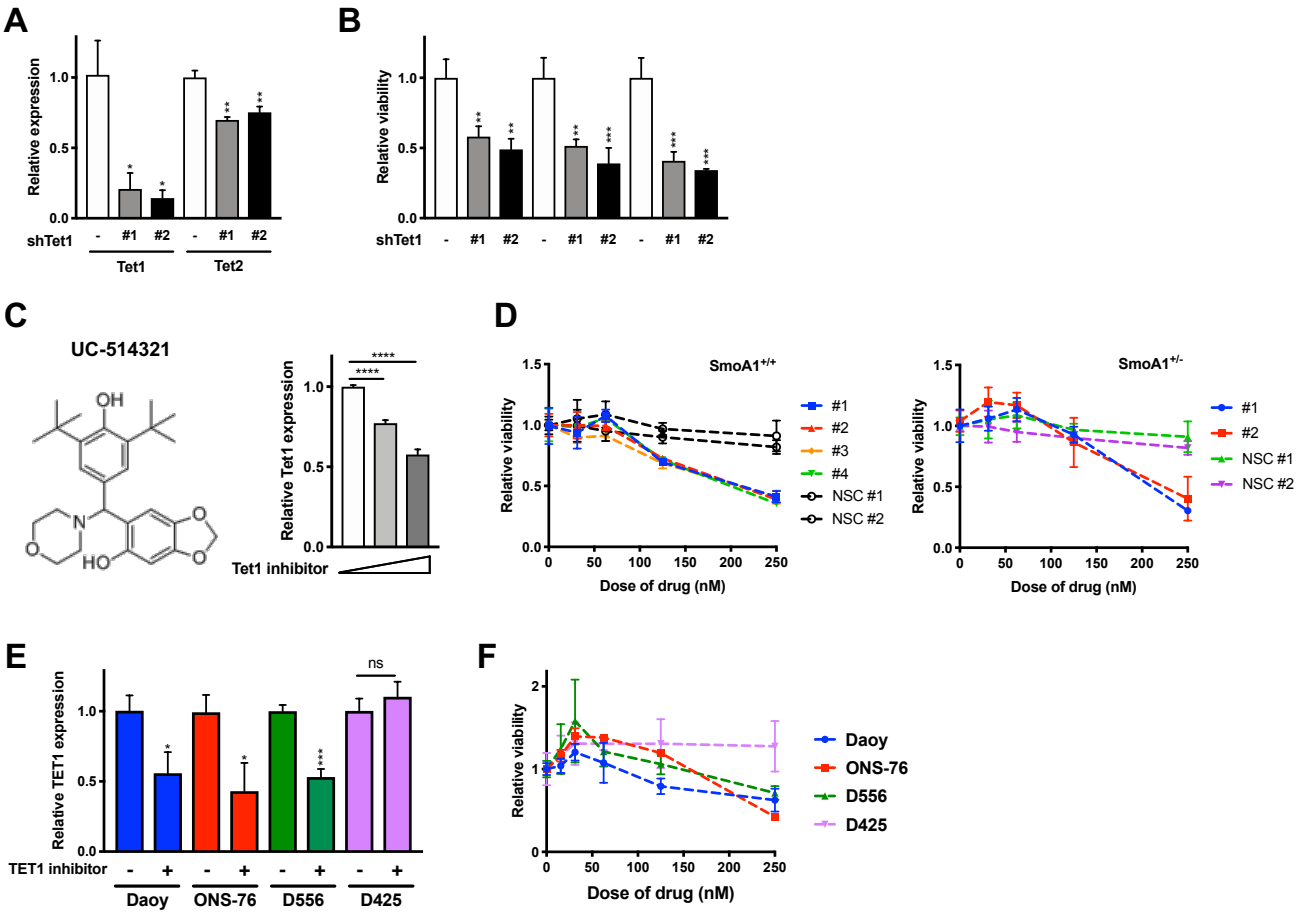


Figure 4.6. TET1 inhibition confers cytotoxic effect on both murine and human MBs. (A) *Tet1* mRNA expression upon shRNA treatment targeting *Tet1* in primary SmoA1-derived MBs. Expression was normalized with *Gapdh* expression. (B) Relative cell viability at 5 days after shRNA treatment. (C) Left: structure of TET1 inhibitor UC-514321. Right: dose-dependent expression of *Tet1* 2 days after chemical treatment (0nM, 100nM, and 200nM, respectively) (D) Relative cell viability depending on dose of drug (nM) in SmoA1$^{+/+}$ and SmoA1$^{+/-}$. NSC: Neuronal stem cell. (E) *TET1* mRNA expression upon 200nM of UC-514321 treatment in MB cell lines. (F) Relative cell viability depending on dose of drug (nM) in MB cell lines.

4.4. Discussion

As one of the most aggressive brain tumors, MB is no longer considered to be a single disease, but transcriptome and methylation profiling divide MBs into four distinct molecular subgroups, WNT, SHH, Group 3, and Group 4 (Northcott et al., 2011). The molecular subgrouping shifted standard clinical risk stratification criteria from established clinical markers (e.g. age of diagnosis, the presence of metastasis at diagnosis, the size of residual disease, and histological variants) to molecular markers such as *CTNBB1* mutation and *MYC* amplification (Pietsch et al., 2014; Taylor et al., 2012; Thompson et al., 2016). Indeed, the prognosis of *CTNBB1* mutated subgroup (WNT) is better than those in other groups, while Group 3, often with *MYC* amplification, have the worst prognosis (Pietsch et al., 2014; Thompson et al., 2016).

Recent studies, however, demonstrate that current molecular subgrouping is insufficient to predict patients' prognosis. For instance, *TP53* mutation is a critical risk factor in both WNT and SHH-MB, (Zhukova et al., 2013). In addition, age-dependent SHH subgroup corresponding to infant ($MB_{SHH\text{-}Infant}$) has a worse prognosis than SHH subgroup diagnosed after 5 years old ($MB_{SHH\text{-}Children}$) although mutation events in *TP53* are rare in $MB_{SHH\text{-}Infant}$ (Schwalbe et al., 2017). In $MB_{SHH\text{-}Infant}$, aberrant DNA methylation at the genes involved in embryonic morphogenesis is considered as a potential driver of the oncogenic process (Schwalbe et al., 2017). Loss of 5hmC, the first oxidative derivative of 5mC, has served as an unfavorable indicator for several malignant tumors, including high-grade glioma (GBM) and leukemia. Consistent with prior findings in different types of tumors, we have identified a significant depletion of 5hmC in MBs and strong inverse correlation between 5hmC levels and prognosis.

Importantly, we identified the enrichment of MB specific 5hmC signature at SCL and NANOG, key

transcription factors involved in self-renewal and pluripotency of ESCs. Moreover, compared to normal tissues, 5hmC signals of MBs were higher at MB active enhancers cis-acting transcription activating elements, suggesting that 5hmC plays a critical role in regulating MB-associated gene expression. We also identified that 5hmC gain was involved in the activation of embryonic development signaling pathways such as the Notch signaling pathway. Indeed, the resemblance of 5hmC gain to fetal 5hmC patterns indicates that 5hmC gain plays a critical role in maintaining stem cell like properties in MBs.

Ten-eleven Translocation (TET) enzymes (TET1, TET2, and TET3) are α-ketoglutarate dependent dioxygenases that convert 5mC to 5hmC and mediate further oxidation processes (Kriaucionis and Heintz, 2009b; Tahiliani et al., 2009). All TET proteins have a core catalytic domain consisting of a double-stranded β-helix (DSBH) domain, a cysteine-rich domain, and binding site for the cofactors Fe(II) and α-ketoglutarate. Crystal structure analysis revealed that the core domain confers the preferential binding of TET proteins to genomic regions containing cytosines in a CpG context (Hu et al., 2013, 2015). In addition, N-terminal CXXC domain found in TET1 and TET3 provides additional binding affinity to target genomic loci (Xu et al., 2012). Regardless of binding sites, TET proteins are responsible for 5hmC generation, so that high expression level of TET proteins tend to increase 5hmC levels. Indeed, the overexpression of TET2 in melanoma cells suppresses tumor initiation and progression by increasing 5hmC level (Bonvin et al., 2019), and elevated 5hmC by overexpressed TET1 promotes glioblastoma-genesis by recruiting the CHTOP-methylosome complex (Takai et al., 2014). Unexpectedly, we observed loss of 5hmC and elevated *TET1* and *TET2* in clinical samples. In addition, when crossing *Tet1*$^{+/-}$ mice with SmoA1 mice, which have a high incidence of spontaneous MB development, we found a dramatic decrease in tumor incidence and tumor onset while the abolishment of *Tet2* did not change tumor incidence and age-of-onset. Further shRNA- and chemical-mediated downregulation of *Tet1* promoted cell-death in both murine and human tumors. Although additional investigation is needed to

determine whether TET1 is a bona fide oncoprotein in MB tumorigenesis, we can conclude that TET1 is an indispensable factor to promote MB development. In addition, further investigation to demonstrate the relationship of overexpressed TET1 and tumor-specific 5hmC signature is needed.

In summary, we present the first comprehensive genome-wide profiling of 5hmC and its potential role in maintaining stemness in MB tumorigenesis. We also identify an unknown MB promoter *TET1* and its role in tumor progression. Our data further show that small molecule-mediated suppression of TET1 can be a therapeutic option for MB subgroups having positive TET1 expression. These findings provide insight into an epigenetic driver, “epi-driver”, in pediatric brain tumors and the biological importance of the drivers in tumor associated signaling pathway.

CHAPTER 5: Summary

5.1. Summary of key findings

There is little commonality between the phenotypic characteristics of neurodegeneration and cancer. However, both are the consequence of disrupted cellular balance between proliferation and death: accelerated neuronal cell death attributes to neurodegeneration while cancer is caused by abnormal resistance to cell death (Plun-Favreau et al., 2010). Indeed, ectopic expression of *MYC* and the dysfunction of *TP53*, a well-known proto-oncogene and tumor suppressor gene, respectively, are also strongly associated with neurodegenerative pathophysiology (Chang et al., 2012; Lee et al., 2009). As such, cross-understanding of two extreme disorders enables an enhanced understanding of underlying biological mechanisms of both diseases. In this thesis, I investigated etiology of two cerebellum-related brain diseases: amyotrophic lateral sclerosis (ALS), a rare, complex neurodegenerative disorder and medulloblastoma, a tumor of the cerebellum. In addition, I also explored the dynamics of 5-hydroxy-methylcytosine (5hmC) during development and aging in the human cerebellum. The overarching goal of this thesis is a better understanding of genetic and epigenetic drivers associated with cerebellum-related brain diseases.

Genetic mutation can directly lead to the production of malfunctional proteins, resulting in disease-causing cellular abnormalities. Technical innovation of high-through sequencing has tremendously accelerated the decoding of disease-associated variants in the last decades, but causal rare genetic variants are often filtered out during statistical testing of sequencing data (MacArthur et al., 2014); therefore, special study designs are necessary to identify causal rare genetic variants. In Chapter 2, I introduced a three-step gene discovery strategy to facilitate the discovery of genetic factors modifying

the risk of ALS, a fatal neurodegenerative disorder. Overall, the hypothesis was that genetic modifiers involved in phenotypic variability of ALS patients carrying G_4C_2 repeat expansion in the *C9orf72* gene, the most prevalent ALS genetic risk (C9ALS), can be unidentified genetic risk factors of ALS. Based on the hypothesis, I first used whole-genome sequencing (WGS) of two pairs of extreme C9ALS cases diagnosed approximately 30 years apart and identified 135 candidate genetic modifiers of C9ALS (step 1). I then performed an unbiased genetic screen using a *Drosophila* model of the G_4C_2 repeat expansion with the genes identified from WGS analysis (step 2). This genetic screen identified the novel genetic interaction between G_4C_2 repeat-associated toxicity and 18 genetic factors, suggesting their potential association with C9ALS risk. I went on to test if 14 out of the 18 genes, those which were not known to be risk factors for ALS previously, are also associated with ALS risk in the sALS cases. Gene-based-statistical analyses of targeted resequencing and WGS were performed (step 3). These analyses together revealed that rare variants in *MYH15* represent a likely genetic risk factor for ALS. In addition, I found that *MYH15* could modulate the toxicity of dipeptides produced from the expanded G_4C_2 repeat. These data demonstrate the power of combining WGS with fly genetics to facilitate the discovery of fundamental genetic components of complex traits with a limited number of samples.

DNA methylation was first discovered in 1948 just after DNA was appreciated as the genetic material (Avery and Macleod, 1944; HOTCHKISS, 1948; McCARTY and AVERY, 1946). Since then, methylated cytosine (5mC) has been considered as a key epigenetic mark involved in various cellular processes such as gene expression regulation, imprinting, and maintenance of genomic integrity (Bird, 2002). In addition, many studies have demonstrated the association of abnormal methylation patterns with diseases such as cancer and neurological disorders (Robertson, 2005a). However, the role of a recently identified DNA modification, 5-hydroxymethylcytosine (5hmC), is less understood. After its first discovery in 2009, 5hmC is now recognized as a highly tissue-specific epigenetic mark due to its

strong enrichment at tissue-specific enhancers and dynamic changes of 5hmC in brain along with 5mC enable plasticity of neuronal circuitry (Guo et al., 2011; Wu et al., 2011b). From a clinical perspective, unique 5hmC signature in cell-free DNA (cfDNA) depending on tissue-of-origin can be used as a biomarker for disease diagnosis and progression (Song et al., 2017a). Therefore, a comprehensive understanding of the cellular role of 5hmC will shed light on basic research as well as have clinical implication. However, there has been limited studies to investigate the dynamics of 5hmC in the human cerebellum. In Chapter 3, I characterized age-dependent dynamics of 5hmC and its correlation with differential gene expression related to development and aging in healthy human cerebella. The cerebellum is the critical brain part to control movement and some cognitive functions, and for proper cerebellar functions, coordinated neurogenesis is necessary. Neurogenesis takes place in embryonic states when new neurons are generated as well as in adult when mature neural circuits are formed. Recovery after traumatic brain injury (TBI) and aging-related oxidative stress also requires active neurogenesis throughout life; thus, organized regulation of gene expression in the precise sites and times is important. Epigenetic mechanisms play a critical role in this process. Epigenetics control active or repressed expression state in the cell by the chemical modifications of DNA and histone proteins, as well as the regulation of non-coding RNAs. Given that abnormal epigenetic programs contribute to development and degenerative disorders, numerous studies have investigated the role of epigenetic mechanisms in brain using animal models (e.g. *Drosophila* and mice). However, dynamics of DNA modifications during development and aging remain poorly understood in humans; therefore, I performed genome-wide mapping of 5hmC in the cerebellum using healthy human cerebellar tissues with different ages and identified age-related distinct 5hmC signature significantly linked to age-associated biological pathways. In addition, development-associated 5hmC marks were significantly enriched at genic regions highly expressed in children. In summary, these results suggest the essential role of 5hmC during development and aging in age-related gene expression.

To understand the role of 5hmC in disease pathogenesis, in Chapter 4, I explored the role of 5hmC alterations in Medulloblastoma (MB), a tumor originating from the cerebellum. 5hmC is highly abundant in Purkinje cells and granular cells, mature neurons of the cerebellum, and regulated by TET (ten-eleven translocation) family proteins. I identified that MB had significant depletion of 5hmC, which is associated with poor prognosis. Despite the global depletion, tumor-specific 5hmC marks in both human and mice MBs were enriched at regulatory regions of genes involved in stemness-related signaling pathways. While TET1 and TET2 expression levels remained high in human MBs like in fetus, only knockout of TET1 in mice attenuated uncontrolled cell growth and prevented abnormal foliation, leading to favorable prognosis in the SmoA1 transgenic MB mouse model. The inhibition of TET1 expression through both shRNA and chemical treatment reduced cell viability in both primary MB cells and human MB cell lines. These results together suggest a potentially key role of 5hmC in MB tumorigenesis and indicate an oncogenic nature of TET1 in this process.

5.2. Clinical implications

It is inevitable that the identification of genetic and epigenetic predisposition for disease cannot prevent diagnosis and disease progression. However, a better understanding of fundamental pathogenetic biology related to genetic and epigenetic programs fuel the development of new treatment for complete cure or slow progression. Unfortunately, there is no definite therapeutic option for complete remission of ALS yet; only supportive cares to alleviate symptoms and prevent unwanted complications are available, which is significantly linked to quality of life and survival for patients (Hobson and McDermott, 2016; Paez-Colasante et al., 2015). Early diagnosis, therefore, helps to provide proper care at the right time and to prolong life expectation (Hobson and McDermott, 2016). Due to phenotypic variability, making

an accurate diagnosis in ALS is complicated. In this sense, a genetic testing of variants in *MYH15*, novel genetic risk factor identified in this thesis (Chapter 2) could contribute to monitoring of the mutation carriers earlier than clinical manifestation of symptoms. In addition to genetic variants, abnormalities of modified cytosines in the genome enable initiation and accelerate disease progression (Esteller, 2007; Maier and Olek, 2002; Robertson, 2005b). Many of the previous studies have utilized animal models to investigate the role of DNA modifications in normal neurogenesis and aging in the brain as well as brain-related disorders found in the human. Even though substantial sequence conservation between species enables us to predict the role of DNA modifications in the human, complex and species-specific features are not fully recapitulated; therefore, my thesis work of genome-wide 5hmC mapping using the human cerebellum tissues provides valuable insights into natural epigenetic dynamics during neuron maturation and aging of the human brain and helps to understand cerebellum-originated disease by using controls (Chapter 3). Indeed, I characterized 5hmC profiles specific to MB, a tumor of the cerebellum through comparison with 5hmC profiles of age-matched normal cerebella (Chapter 4). The levels of 5hmC showed strong negative correlation with prognosis, which can serve as a biomarker of risk stratification. I also identified that the abolishment of TET1, not TET2, leads to cell death, implying that TET1 plays an indispensable role in tumor-specific 5hmC landscape. Therefore, small molecules targeting TET1 may be used as a therapeutic option to cure a subset of MB patients who have high expression of TET1.

REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. (2000). The genome sequence of Drosophila melanogaster. Science (80-.).

Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J., and Pickard, B.S. (2006). SUSPECTS: Enabling fast and effective prioritization of positional candidates. Bioinformatics.

Al-Chalabi, A., Fang, F., Hanby, M.F., Leigh, P.N., Shaw, C.E., Ye, W., and Rijsdijk, F. (2010). An estimate of amyotrophic lateral sclerosis heritability using twin data. J. Neurol. Neurosurg. Psychiatry *81*, 1324–1326.

Al-Chalabi, A., van den Berg, L.H., and Veldink, J. (2017). Gene discovery in amyotrophic lateral sclerosis: implications for clinical management. Nat. Rev. Neurosci. *13*, 96–104.

Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B., et al. (2015). A global reference for human genetic variation. Nature *526*, 68-+.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: Tool for the unification of biology. Nat. Genet.

Auer, P.L., and Lettre, G. (2015). Rare variant association studies: Considerations, challenges and opportunities. Genome Med.

Autism Spectrum Disorders Working Group of The Psychiatric Genomics, C. (2017). Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. Mol Autism *8*, 21.

Avery, O.T., and Macleod, C.M. (1944). Studies on the Chemical Inducing Nature Types of the Substance Transformation. J. Exp. Med.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet.

Bansal, V., Libiger, O., Torkamani, A., and Schork, N.J. (2010). Statistical analysis strategies for association studies involving rare variants. Nat. Rev. Genet.

Barany, M. (1967). ATPase activity of myosin correlated with speed of muscle shortening. J. Gen. Physiol. *50*, 197–218.

Barnes, K.C., Sigaux, F., Margolin, A., Young, R.A., O'Neil, J., Aster, J.C., Soulier, J., Weng, A.P., Look, A.T., Odom, D.T., et al. (2006). NOTCH1 directly regulates c-MYC and activates a feed-forward-loop transcriptional network promoting leukemic cell growth. Proc. Natl. Acad. Sci.

Barnett, I.J., Lee, S., and Lin, X. (2013). Detecting Rare Variant Effects Using Extreme Phenotype Sampling in Sequencing Association Studies. Genet. Epidemiol.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol *28*, 1045–1048.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev.

Bock, C., Halbritter, F., Carmona, F.J., Tierling, S., Datlinger, P., Assenov, Y., Berdasco, M., Bergmann, A.K., Booher, K., Busato, F., et al. (2016). Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. Nat. Biotechnol.

Bogdanović, O., and Veenstra, G.J.C. (2009). DNA methylation and methyl-CpG binding proteins: Developmental requirements and function. Chromosoma.

Bonvin, E., Radaelli, E., Bizet, M., Luciani, F., Calonne, E., Putmans, P., Nittner, D., Singh, N.K., Santagostino, S.F., Petit, V., et al. (2019). TET2-dependent hydroxymethylome plasticity reduces melanoma initiation and progression. Cancer Res.

Booth, M.J., Ost, T.W.B., Beraldi, D., Bell, N.M., Branco, M.R., Reik, W., and Balasubramanian, S.

(2013). Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. Nat. Protoc. *8*, 1841.

Bradley, C.K., Takano, E.A., Hall, M.A., Göthert, J.R., Harvey, A.R., Begley, C.G., and Van Eekelen, J.A.M. (2006). The essential haematopoietic transcription factor Scl is also critical for neuronal development. Eur. J. Neurosci.

Broad Institute (2016). Picard:A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF. Http://Broadinstitute.Github.Io/Picard/.

Bromberg, Y. (2013). Chapter 15: disease gene prioritization. PLoS Comput. Biol. *9*, e1002902–e1002902.

Buckner, R.L. (2013). The cerebellum and cognitive function: 25 years of insight from anatomy and neuroimaging. Neuron.

Byrne, S., Heverin, M., Elamin, M., Bede, P., Lynch, C., Kenna, K., MacLaughlin, R., Walsh, C., Al Chalabi, A., and Hardiman, O. (2013). Aggregation of neurologic and neuropsychiatric disease in amyotrophic lateral sclerosis kindreds: a population-based case-control cohort study of familial and sporadic amyotrophic lateral sclerosis. Ann. Neurol. *74*, 699–708.

Van Cauwenberghe, C., Van Broeckhoven, C., and Sleegers, K. (2016). The genetic landscape of Alzheimer disease: clinical implications and perspectives. Genet Med *18*, 421–430.

Cavalli, F.M.G., Remke, M., Rampasek, L., Peacock, J., Shih, D.J.H., Luu, B., Garzia, L., Torchia, J., Nor, C., Morrissy, A.S., et al. (2017). Intertumoral Heterogeneity within Medulloblastoma Subgroups. Cancer Cell.

Chang, D., Nalls, M.A., Hallgrimsdottir, I.B., Hunkapiller, J., van der Brug, M., Cai, F., International Parkinson's Disease Genomics, C., andMe Research, T., Kerchner, G.A., Ayalon, G., et al. (2017). A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. Nat

Genet *49*, 1511–1516.

Chang, J.R., Ghafouri, M., Mukerjee, R., Bagashev, A., Chabrashvili, T., and Sawaya, B.E. (2012). Role of p53 in neurodegenerative diseases. Neurodegener. Dis.

Cheng, Y., Li, Z., Manupipatpong, S., Lin, L., Li, X., Xu, T., Jiang, Y.H., Shu, Q., Wu, H., and Jin, P. (2018). 5-Hydroxymethylcytosine alterations in the human postmortem brains of autism spectrum disorder. Hum. Mol. Genet.

Chi, S., Jiang, T., Tan, L., and Yu, J.T. (2016). Distinct neurological disorders with C9orf72 mutations: genetics, pathogenesis, and therapy. Neurosci. Biobehav. Rev. *66*, 127–142.

Choi, J.Y., Muallem, D., Kiselyov, K., Lee, M.G., Thomas, P.J., and Muallem, S. (2001). Aberrant CFTR-dependent HCO3- transport in mutations associated with cystic fibrosis. Nature.

Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith, J.D., Harrell, T.M., McMillin, M.J., Wiszniewski, W., Gambin, T., et al. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am. J. Hum. Genet.

Chou, W.C., Chou, S.C., Liu, C.Y., Chen, C.Y., Hou, H.A., Kuo, Y.Y., Lee, M.C., Ko, B.S., Tang, J.L., Yao, M., et al. (2011). TET2 mutation is an unfavorable prognostic factor in acute myeloid leukemia patients with intermediate-risk cytogenetics. Blood.

Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet.

Cirulli, E.T., Lasseigne, B.N., Petrovski, S., Sapp, P.C., Dion, P.A., Leblond, C.S., Couthouis, J., Lu, Y.F., Wang, Q., Krueger, B.J., et al. (2015). Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. Science. *347*, 1436–1441.

Cooper, D.N., Krawczak, M., Polychronakos, C., Tyler-Smith, C., and Kehrer-Sawatzki, H. (2013). Where genotype is not predictive of phenotype: Towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum. Genet.

Cope, M.J., Whisstock, J., Rayment, I., and Kendrick-Jones, J. (1996). Conservation within the myosin motor domain: implications for structure and function. Structure *4*, 969–987.

Costa, Y., Ding, J., Theunissen, T.W., Faiola, F., Hore, T.A., Shliaha, P. V, Fidalgo, M., Saunders, A., Lawrence, M., Dietmann, S., et al. (2013). NANOG-dependent function of TET1 and TET2 in establishment of pluripotency. Nature *495*, 370–374.

Curradi, M., Izzo, A., Badaracco, G., and Landsberger, N. (2002). Molecular Mechanisms of Gene Silencing Mediated by DNA Methylation. Mol. Cell. Biol.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res *46*, D794–D801.

Dawlaty, M.M., Ganz, K., Powell, B.E., Hu, Y.C., Markoulaki, S., Cheng, A.W., Gao, Q., Kim, J., Choi, S.W., Page, D.C., et al. (2011). Tet1 is dispensable for maintaining pluripotency and its loss is compatible with embryonic and postnatal development. Cell Stem Cell.

Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev.

Desjardins, P.R., Burkman, J.M., Shrager, J.B., Allmond, L.A., and Stedman, H.H. (2002). Evolutionary implications of three novel members of the human sarcomeric myosin heavy chain gene family. Mol. Biol. Evol. *19*, 375–393.

Diaz-Lagares, A., Mendez-Gonzalez, J., Hervas, D., Saigi, M., Pajares, M.J., Garcia, D., Crujerias, A.B., Pio, R., Montuenga, L.M., Zulueta, J., et al. (2016). A Novel Epigenetic Signature for Early Diagnosis in Lung Cancer. Clin Cancer Res *22*, 3361–3371.

Doi, A., Park, I.H., Wen, B., Murakami, P., Aryee, M.J., Irizarry, R., Herb, B., Ladd-Acosta, C., Rho, J.S., Loewer, S., et al. (2009). Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nat. Genet. *41*, 1350-U123.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature.

Echeverri, C.J., and Perrimon, N. (2006). High-throughput RNAi screening in cultured cells: A user's guide. Nat. Rev. Genet.

Ecke, I., Petry, F., Rosenberger, A., Tauber, S., Mönkemeyer, S., Hess, I., Dullin, C., Kimmina, S., Pirngruber, J., Johnsen, S.A., et al. (2009). Antitumor effects of a combined 5-aza-2′deoxycytidine and valproic acid treatment on rhabdomyosarcoma and medulloblastoma in Ptch mutant mice. Cancer Res.

Ee, P.-L.R., Bhat, U., Cho, S., He, X., Schneider, E., Lu, M., Miele, L., and Beck, W.T. (2011). Notch1 regulates the expression of the multidrug resistance gene ABCC1/MRP1 in cultured cancer cells. Proc. Natl. Acad. Sci.

Elefanty, a G., Begley, C.G., Metcalf, D., Barnett, L., Köntgen, F., and Robb, L. (1998). Characterization of hematopoietic progenitor cells that express the transcription factor SCL, using a lacZ "knock-in" strategy. Proc. Natl. Acad. Sci. U. S. A.

Ellison, D.W., Dalton, J., Kocak, M., Nicholson, S.L., Fraga, C., Neale, G., Kenney, A.M., Brat, D.J., Perry, A., Yong, W.H., et al. (2011). Medulloblastoma: clinicopathological correlates of SHH, WNT, and non-SHH/WNT molecular subgroups. Acta Neuropathol *121*, 381–396.

Engelhardt, B.E., Jordan, M.I., Srouji, J.R., and Brenner, S.E. (2011). Genome-scale phylogenetic function annotation of large and diverse protein families. Genome Res.

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. Nat. Rev. Genet. *8*, 286–298.

Esteller, M. (2008). Epigenetics in cancer. N Engl J Med *358*, 1148–1159.

Fernandez, A.F., Assenov, Y., Martin-Subero, J.I., Balint, B., Siebert, R., Taniguchi, H., Yamamoto, H., Hidalgo, M., Tan, A.C., Galm, O., et al. (2012). A DNA methylation fingerprint of 1628 human samples.

Genome Res *22*, 407–419.

Ficz, G., and Gribben, J.G. (2014). Loss of 5-hydroxymethylcytosine in cancer: Cause or consequence? Genomics *104*, 352–357.

Field, A.E., Robertson, N.A., Wang, T., Havas, A., Ideker, T., and Adams, P.D. (2018). DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. Mol. Cell.

Fischer, A.H., Jacobson, K.A., Rose, J., and Zeller, R. (2008). Hematoxylin and eosin staining of tissueand cell sections. Cold Spring Harb. Protoc.

Foster, T.C., Sharrow, K.M., Masse, J.R., Norris, C.M., and Kumar, A. (2001). Calcineurin links Ca2+ dysregulation with brain aging. J. Neurosci.

Freibaum, B.D., Lu, Y., Lopez-Gonzalez, R., Kim, N.C., Almeida, S., Lee, K.H., Badders, N., Valentine, M., Miller, B.L., Wong, P.C., et al. (2015). GGGGCC repeat expansion in C9orf72 compromises nucleocytoplasmic transport. Nature *525*, 129–133.

Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L., and Paul, C.L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A *89*, 1827–1831.

Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S. V, Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Consortium, G.Te., Nicolae, D.L., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.

Geevasinga, N., Menon, P., Ozdinler, P.H., Kiernan, M.C., and Vucic, S. (2016). Pathophysiological and diagnostic implications of cortical dysfunction in ALS. Nat. Rev. Neurol. *12*, 651–661.

Goers, L., Freemont, P., and Polizzi, K.M. (2014). Co-culture systems and technologies: Taking synthetic biology to the next level. J. R. Soc. Interface.

Good, C.R., Madzo, J., Patel, B., Maegawa, S., Engel, N., Jelinek, J., and Issa, J.P.J. (2017). A novel isoform of TET1 that lacks a CXXC domain is overexpressed in cancer. Nucleic Acids Res.

Goodhead, I., Campbell, P.J., Dicks, E., Futreal, P.A., Green, A.R., Follows, G.A., Stratton, M.R., Pleasance, E.D., Stephens, P.J., and Rance, R. (2008). Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc. Natl. Acad. Sci.

Gormley, P., Anttila, V., Winsvold, B.S., Palta, P., Esko, T., Pers, T.H., Farh, K.H., Cuenca-Leon, E., Muona, M., Furlotte, N.A., et al. (2016). Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. Nat Genet *48*, 856–866.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. Nature.

Group, G.B.D.N.D.C. (2017). Global, regional, and national burden of neurological disorders during 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet Neurol *16*, 877–897.

Guo, J.U., Su, Y., Zhong, C., Ming, G., and Song, H. (2011). Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. Cell *145*, 423–434.

Ha Thi, B.M., Campolmi, N., He, Z., Pipparelli, A., Manissolle, C., Thuret, J.Y., Piselli, S., Forest, F., Peoc'h, M., Garraud, O., et al. (2014). Microarray analysis of cell cycle gene expression in adult human corneal endothelial cells. PLoS One *9*, e94349.

Hackett, J.A., Sengupta, R., Zylicz, J.J., Murakami, K., Lee, C., Down, T.A., and Surani, M.A. (2013). Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. Science (80-.).

Haeusler, A.R., Donnelly, C.J., and Rothstein, J.D. (2016). The expanding biology of the C9orf72 nucleotide repeat expansion in neurodegenerative disease. Nat. Rev. Neurosci. *17*, 383–395.

Haggarty, S.J. (2015). Epigenetic diagnostics for neuropsychiatric disorders: Above the genome. Neurology *84*, 1618–1619.

Hahn, M.A., Qiu, R., Wu, X., Li, A.X., Zhang, H., Wang, J., Jui, J., Jin, S.-G., Jiang, Y., Pfeifer, G.P.,

et al. (2013). Dynamics of 5-hydroxymethylcytosine and chromatin marks in Mammalian neurogenesis. Cell Rep. *3*, 291–300.

Hall-Lande, J., White, T., Kurzius-Spencer, M., Warren, Z., Wiggins, L., Rosenberg, Pettygrove, S., Van, K., Imm, P., Lee, L.-C., et al. (2018). Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014. MMWR. Surveill. Summ.

Hallahan, A.R., Pritchard, J.I., Hansen, S., Benson, M., Stoeck, J., Hatton, B.A., Russell, T.L., Ellenbogen, R.G., Bernstein, I.D., Beachy, P.A., et al. (2004). The SmoA1 mouse model reveals that notch signaling is critical for the growth and survival of sonic hedgehog-induced medulloblastomas. Cancer Res *64*, 7794–7800.

Harmston, N., Ing-Simmons, E., Tan, G., Perry, M., Merkenschlager, M., and Lenhard, B. (2017). Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. Nat. Commun.

Hashimoto, S., Kojima, N., Okamoto, Y., Seki, Y., Sugasawa, K., Okashita, N., Takada, T., Ebi, K., Kumaki, Y., Nakamura, T., et al. (2013). PRDM14 promotes active DNA demethylation through the Ten-eleven translocation (TET)-mediated base excision repair pathway in embryonic stem cells. Development.

Hatton, B.A., Villavicencio, E.H., Tsuchiya, K.D., Pritchard, J.I., Ditzler, S., Pullar, B., Hansen, S., Knoblaugh, S.E., Lee, D., Eberhart, C.G., et al. (2008). The Smo/Smo model: Hedgehog-induced medulloblastoma with 90% incidence and leptomeningeal spread. Cancer Res.

He, Y.F., Li, B.Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., et al. (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science (80-.). *333*, 1303–1307.

Hein, C.D., Liu, X.M., and Wang, D. (2008). Click chemistry, a powerful tool for pharmaceutical

sciences. Pharm. Res.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol. Cell.

Herberth, B., Minkó, K., Csillag, A., Jaffredo, T., and Madarász, E. (2005). SCL, GATA-2 and Lmo2 expression in neurogenesis. Int. J. Dev. Neurosci.

Herculano-Houzel, S. (2009). The human brain in numbers: a linearly scaled-up primate brain. Front. Hum. Neurosci.

Heyn, H., and Esteller, M. (2012). DNA methylation profiling in the clinic: applications and challenges. Nat Rev Genet *13*, 679–692.

Heyn, H., and Esteller, M. (2015). An Adenine Code for DNA: A Second Life for N6-Methyladenine. Cell *161*, 710–713.

Hirokawa, N., Niwa, S., and Tanaka, Y. (2010). Molecular motors in neurons: transport mechanisms and roles in brain function, development, and disease. Neuron *68*, 610–638.

Hobson, E. V, and McDermott, C.J. (2016). Supportive and symptomatic management of amyotrophic lateral sclerosis. Nat Rev Neurol *12*, 526–538.

Hoischen, A., Krumm, N., and Eichler, E.E. (2014). Prioritization of neurodevelopmental disease genes by discovery of new mutations. Nat. Neurosci.

Hoshino, M., Nakamura, S., Mori, K., Kawauchi, T., Terao, M., Nishimura, Y. V., Fukuda, A., Fuse, T., Matsuo, N., Sone, M., et al. (2005). Ptf1a, a bHLH transcriptional gene, defines GABAergic neuronal fates in cerebellum. Neuron.

HOTCHKISS, R.D. (1948). The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. J. Biol. Chem.

Hovestadt, V., Jones, D.T.W., Picelli, S., Wang, W., Kool, M., Northcott, P.A., Sultan, M., Stachurski,

K., Ryzhova, M., Warnatz, H.J., et al. (2014a). Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. Nature *510*, 537-+.

Hovestadt, V., Jones, D.T.W., Picelli, S., Wang, W., Kool, M., Northcott, P. a, Sultan, M., Stachurski, K., Ryzhova, M., Warnatz, H.-J., et al. (2014b). Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. Nature *510*, 537–541.

Hsieh, J., and Zhao, X. (2016). Genetics and epigenetics in adult neurogenesis. Cold Spring Harb. Perspect. Biol.

Hu, L., Li, Z., Cheng, J., Rao, Q., Gong, W., Liu, M., Shi, Y.G., Zhu, J., Wang, P., and Xu, Y. (2013). Crystal Structure of TET2-DNA Complex: Insight into TET-Mediated 5mC Oxidation. Cell.

Hu, L., Lu, J., Cheng, J., Rao, Q., Li, Z., Hou, H., Lou, Z., Zhang, L., Li, W., Gong, W., et al. (2015). Structural insight into substrate preference for TET-mediated oxidation. Nature.

Hu, Y., Flockhart, I., Vinayagam, A., Bergwitz, C., Berger, B., Perrimon, N., and Mohr, S.E. (2011). An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics *12*, 357.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009a). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc.

Hwang, J.Y., Aromolaran, K.A., and Zukin, R.S. (2017). The emerging field of epigenetics in neurodegeneration and neuroprotection. Nat Rev Neurosci *18*, 347–361.

International Human Genome Sequencing Consortium (2004). International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature.

Ionita-Laza, I., and Ottman, R. (2011). Study designs for identification of rare disease variants in complex diseases: The utility of family-based designs. Genetics.

Ito, S., D'Alessio, A.C., Taranova, O. V, Hong, K., Sowers, L.C., and Zhang, Y. (2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature *466*, 1129–1133.

Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C., and Zhang, Y. (2011). Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. Science (80-.). *333*, 1300–1303.

Jacinto, F. V, Ballestar, E., and Esteller, M. (2008). Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. Biotechniques *44*, 35, 37, 39 passim.

Jakubowski, J.L., and Labrie, V. (2017). Epigenetic Biomarkers for Parkinson's Disease: From Diagnostics to Therapeutics. J Park. Dis *7*, 1–12.

James, S.J., Shpyleva, S., Melnyk, S., Pavliv, O., and Pogribny, I.P. (2013). Complex epigenetic regulation of Engrailed-2 (EN-2) homeobox gene in the autism cerebellum. Transl. Psychiatry.

Jenuwein, T., and Allis, C.D. (2001). Translating the histone code. Science (80-.). *293*, 1074–1080.

Jiang, X., Hu, C., Ferchen, K., Nie, J., Cui, X., Chen, C.H., Cheng, L., Zuo, Z., Seibel, W., He, C., et al. (2017). Targeted inhibition of STAT/TET1 axis as a therapeutic strategy for acute myeloid leukemia. Nat. Commun.

Jin, S.G., Wu, X., Li, A.X., and Pfeifer, G.P. (2011a). Genomic mapping of 5-hydroxymethylcytosine in the human brain. Nucleic Acids Res *39*, 5015–5024.

Jin, S.G., Jiang, Y., Qiu, R., Rauch, T. a., Wang, Y., Schackert, G., Krex, D., Lu, Q., and Pfeifer, G.P. (2011b). 5-hydroxymethylcytosine is strongly depleted in human cancers but its levels do not correlate with IDH1 mutations. Cancer Res. *71*, 7360–7365.

Johnston, H.R., Chopra, P., Wingo, T.S., Patel, V., International Consortium on, B., Behavior in 22q11.2 Deletion, S., Epstein, M.P., Mulle, J.G., Warren, S.T., Zwick, M.E., et al. (2017). PEMapper and PECaller provide a simplified approach to whole-genome sequencing. Proc. Natl. Acad. Sci. U. S. A.

114, E1923–E1932.

Kabashi, E., Valdmanis, P.N., Dion, P., Spiegelman, D., McConkey, B.J., Vande Velde, C., Bouchard, J.P., Lacomblez, L., Pochigaeva, K., Salachas, F., et al. (2008). TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis. Nat. Genet. *40*, 572–574.

Kempermann, G., Gage, F.H., Aigner, L., Song, H., Curtis, M.A., Thuret, S., Kuhn, H.G., Jessberger, S., Frankland, P.W., Cameron, H.A., et al. (2018). Human Adult Neurogenesis: Evidence and Remaining Questions. Cell Stem Cell.

Kenna, K.P., Van Doormaal, P.T.C., Dekker, A.M., Ticozzi, N., Kenna, B.J., Diekstra, F.P., Van Rheenen, W., Van Eijk, K.R., Jones, A.R., Keagle, P., et al. (2016). NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. Nat. Genet.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

Kim, H., Wang, X., and Jin, P. (2018). Developing DNA methylation-based diagnostic biomarkers. J. Genet. Genomics.

Kircher, M., Witten, D.M., Jain, P., O'roak, B.J., Cooper, G.M., and Shendure, J. (2014a). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014b). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.

Köhler, S., Bauer, S., Horn, D., and Robinson, P.N. (2008). Walking the Interactome for Prioritization of Candidate Disease Genes. Am. J. Hum. Genet.

Kotlar, A. V., Trevino, C.E., Zwick, M.E., Cutler, D.J., and Wingo, T.S. (2018). Bystro: Rapid online variant annotation and natural-language filtering at whole-genome scale. Genome Biol.

Kraus, T.F.J., Globisch, D., Wagner, M., Eigenbrod, S., Widmann, D., Münzel, M., Müller, M., Pfaffeneder, T., Hackner, B., Feiden, W., et al. (2012). Low values of 5-hydroxymethylcytosine (5hmC),

the “sixth base,” are associated with anaplasia in human brain tumors. Int. J. Cancer *131*, 1577–1590.

Kriaucionis, S., and Heintz, N. (2009a). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science *324*, 929–930.

Kriaucionis, S., and Heintz, N. (2009b). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science (80-.). *324*, 929–930.

Kryukov, G. V., Shpunt, A., Stamatoyannopoulos, J.A., and Sunyaev, S.R. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. Proc. Natl. Acad. Sci.

Kudo, Y., Tateishi, K., Yamamoto, K., Yamamoto, S., Asaoka, Y., Ijichi, H., Nagae, G., Yoshida, H., Aburatani, H., and Koike, K. (2012). Loss of 5-hydroxymethylcytosine is accompanied with malignant cellular transformation. Cancer Sci. *103*, 670–676.

Kumar, V., Hasan, G.M., and Hassan, M.I. (2017). Unraveling the Role of RNA Mediated Toxicity of C9orf72 Repeats in C9-FTD/ALS. Front. Neurosci. *11*, 711.

Laird, A., Thomson, J.P., Harrison, D.J., and Meehan, R.R. (2013). 5-hydroxymethylcytosine profiling as an indicator of cellular state. Epigenomics.

Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., Jun, G., DeStefano, A.L., Bis, J.C., Beecham, G.W., et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer’s disease. Nat. Genet.

Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res *44*, D862-8.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods.

Le, S.Q., and Durbin, R. (2011). SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome Res.

Lee, H.G., Casadesus, G., Nunomura, A., Zhu, X., Castellani, R.J., Richardson, S.L., Perry, G., Felsher,

D.W., Petersen, R.B., and Smith, M.A. (2009). The neuronal expression of MYC causes a neurodegenerative phenotype in a novel transgenic mouse. Am. J. Pathol.

Lee, I., Blom, U.M., Wang, P.I., Shim, J.E., and Marcotte, E.M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res.

Lee, K.H., Zhang, P., Kim, H.J., Mitrea, D.M., Sarkar, M., Freibaum, B.D., Cika, J., Coughlin, M., Messing, J., Molliex, A., et al. (2016a). C9orf72 Dipeptide Repeats Impair the Assembly, Dynamics, and Function of Membrane-Less Organelles. Cell *167*, 774-788 e17.

Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Christiani, D.C., Wurfel, M.M., and Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. *91*, 224–237.

Lee, S., Teslovich, T.M., Boehnke, M., and Lin, X. (2013). General framework for meta-analysis of rare variants in sequencing association studies. Am. J. Hum. Genet. *93*, 42–53.

Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. Am. J. Hum. Genet.

Lee, S., Shang, Y., Redmond, S.A., Urisman, A., Tang, A.A., Li, K.H., Burlingame, A.L., Pak, R.A., Jovicic, A., Gitler, A.D., et al. (2016b). Activation of HIPK2 Promotes ER Stress-Mediated Neurodegeneration in Amyotrophic Lateral Sclerosis. Neuron *91*, 41–55.

Lek, M., Karczewski, K.J., Minikel, E. V, Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Lemons, D., Maurya, M.R., Subramaniam, S., and Mercola, M. (2013). Developing microRNA screening as a functional genomics tool for disease research. Front. Physiol.

Li, B., and Leal, S.M. (2008). Methods for Detecting Associations with Rare Variants for Common

Diseases: Application to Analysis of Sequence Data. Am. J. Hum. Genet.

Li, W., and Liu, M. (2011). Distribution of 5-Hydroxymethylcytosine in Different Human Tissues. J. Nucleic Acids.

Li, D., Lewinger, J.P., Gauderman, W.J., Murcray, C.E., and Conti, D. (2011a). Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. Genet. Epidemiol.

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

Li, J., Wu, X., Zhou, Y., Lee, M., Guo, L., Han, W., Mo, W., Cao, W.M., Sun, D., Xie, R., et al. (2018). Decoding the dynamic DNA methylation and hydroxymethylation landscapes in endodermal lineage intermediates during pancreatic differentiation of hESC. Nucleic Acids Res.

Li, Y., Sidore, C., Kang, H.M., Boehnke, M., and Abecasis, G.R. (2011b). Low-coverage sequencing: Implications for design of complex trait association studies. Genome Res.

Li, Y., Brauer, P.M., Singh, J., Xhiku, S., Yoganathan, K., Zúñiga-Pflücker, J.C., and Anderson, M.K. (2017). Targeted Disruption of TCF12 Reveals HEB as Essential in Human Mesodermal Specification and Hematopoiesis. Stem Cell Reports.

Lian, C.G., Xu, Y., Ceol, C., Wu, F., Larson, A., Dresser, K., Xu, W., Tan, L., Hu, Y., Zhan, Q., et al. (2012a). Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of melanoma. Cell *150*, 1135–1146.

Lian, C.G., Xu, Y., Ceol, C., Wu, F., Larson, A., Dresser, K., Xu, W., Tan, L., Hu, Y., Zhan, Q., et al. (2012b). Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of Melanoma. Cell.

Lin, C.Y., Erkek, S., Tong, Y., Yin, L., Federation, A.J., Zapatka, M., Haldipur, P., Kawauchi, D., Risch, T., Warnatz, H.J., et al. (2016). Active medulloblastoma enhancers reveal subgroup-specific cellular origins. Nature.

LoParo, D., and Waldman, I.D. (2015). The oxytocin receptor gene (OXTR) is associated with autism

spectrum disorder: a meta-analysis. Mol Psychiatry *20*, 640–646.

Lopez-Gonzalez, R., Lu, Y., Gendron, T.F., Karydas, A., Tran, H., Yang, D., Petrucelli, L., Miller, B.L., Almeida, S., and Gao, F.B. (2016). Poly(GR) in C9ORF72-Related ALS/FTD Compromises Mitochondrial Function and Increases Oxidative Stress and DNA Damage in iPSC-Derived Motor Neurons. Neuron *92*, 383–391.

Lorsback, R.B., Moore, J., Mathew, S., Raimondi, S.C., Mukatira, S.T., and Downing, J.R. (2003). TET1, a member of a novel protein family, is fused to MLL in acute myeloid leukemia containing the t(10;11)(q22;23) [3]. Leukemia.

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

Luke, M.M., O'Meara, E.S., Rowland, C.M., Shiffman, D., Bare, L.A., Arellano, A.R., Longstreth Jr., W.T., Lumley, T., Rice, K., Tracy, R.P., et al. (2009). Gene variants associated with ischemic stroke: the cardiovascular health study. Stroke *40*, 363–368.

MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. Nature.

Mack, S.C., Hubert, C.G., Miller, T.E., Taylor, M.D., and Rich, J.N. (2016). An epigenetic gateway to brain tumor cell identity. Nat Neurosci *19*, 10–19.

Madrid, A., Chopra, P., and Alisch, R.S. (2018). Species-Specific 5 mC and 5 hmC Genomic Landscapes Indicate Epigenetic Contribution to Human Brain Evolution. Front. Mol. Neurosci.

Maier, S., and Olek, A. (2002). Diabetes: a candidate disease for efficient DNA methylation profiling. J Nutr *132*, 2440S-2443S.

Maiti, A., and Drohat, A.C. (2011). Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: Potential implications for active demethylation of CpG sites. J. Biol. Chem.

Massé, A., Vainchenker, W., Dupont, S., Alberdi, A., Delhommeau, F., Fontenay, M., Robert, F., Lécluse, Y., Plo, I., Viguié, F., et al. (2009). Mutation in TET2 in Myeloid Cancers . N. Engl. J. Med.

Massimino, M., Giangaspero, F., Garre, M.L., Gandola, L., Poggi, G., Biassoni, V., Gatta, G., and Rutkowski, S. (2011). Childhood medulloblastoma. Crit Rev Oncol Hematol *79*, 65–83.

Maston, G. a, Evans, S.K., and Green, M.R. (2006). Transcriptional regulatory elements in the human genome. Annu. Rev. Genomics Hum. Genet. *7*, 29–59.

McCARTY, M., and AVERY, O.T. (1946). Studies on the chemical nature of the substance inducing transformation of pneumococcal types; effect of desoxyribonuclease on the biological activity of the transforming substance. J. Exp. Med.

McLaughlin, R.L., Vajda, A., and Hardiman, O. (2015). Heritability of Amyotrophic Lateral Sclerosis: Insights From Disparate Numbers. JAMA Neurol. *72*, 857–858.

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

Mizielinska, S., Gronke, S., Niccoli, T., Ridler, C.E., Clayton, E.L., Devoy, A., Moens, T., Norona, F.E., Woollacott, I.O.C., Pietrzyk, J., et al. (2014). C9orf72 repeat expansions cause neurodegeneration in Drosophila through arginine-rich proteins. Science. *345*, 1192–1194.

Moberg, K.H., Bell, D.W., Wahrer, D.C., Haber, D.A., and Hariharan, I.K. (2001). Archipelago regulates Cyclin E levels in Drosophila and is mutated in human cancer cell lines. Nature *413*, 311–316.

Moran, S., Arribas, C., and Esteller, M. (2016). Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics *8*, 389–399.

Moreau, Y., and Tranchevent, L.C. (2012). Computational tools for prioritizing candidate genes: Boosting disease gene discovery. Nat. Rev. Genet.

Mori, K., Weng, S.M., Arzberger, T., May, S., Rentzsch, K., Kremmer, E., Schmid, B., Kretzschmar,

H.A., Cruts, M., Van Broeckhoven, C., et al. (2013). The C9orf72 GGGGCC repeat is translated into aggregating dipeptide-repeat proteins in FTLD/ALS. Science. *339*, 1335–1338.

Müller, T., Gessi, M., Waha, A., Isselstein, L.J., Luxen, D., Freihoff, D., Freihoff, J., Becker, A., Simon, M., Hammes, J., et al. (2012). Nuclear Exclusion of TET1 Is Associated with Loss of 5-Hydroxymethylcytosine in IDH1 Wild-Type Gliomas. Am. J. Pathol. *181*, 675–683.

Munoz, P., Iliou, M.S., and Esteller, M. (2012). Epigenetic alterations involved in cancer stem cell reprogramming. Mol Oncol *6*, 620–636.

Neri, F., Incarnato, D., Krepelova, A., Rapelli, S., Pagnani, A., Zecchina, R., Parlato, C., and Oliviero, S. (2013). Genome-wide analysis identifies a functional association of Tet1 and Polycomb repressive complex 2 in mouse embryonic stem cells. Genome Biol. *14*, R91.

Neri, F., Incarnato, D., Krepelova, A., Dettori, D., Rapelli, S., Maldotti, M., Parlato, C., Anselmi, F., Galvagni, F., and Oliviero, S. (2015). TET1 is controlled by pluripotency-associated factors in ESCs and downmodulated by PRC2 in differentiated cells and tissues. Nucleic Acids Res *43*, 6814–6826.

Van Neste, L., Herman, J.G., Otto, G., Bigley, J.W., Epstein, J.I., and Van Criekinge, W. (2012). The epigenetic promise for prostate cancer diagnosis. Prostate *72*, 1248–1261.

Nestor, C.E., Ottaviano, R., Reddington, J., Sproul, D., Reinhardt, D., Dunican, D., Katz, E., Dixon, J.M., Harrison, D.J., and Meehan, R.R. (2012). Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes. Genome Res.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature.

Ni, J.Q., Zhou, R., Czech, B., Liu, L.P., Holderbaum, L., Yang-Zhou, D., Shim, H.S., Tao, R., Handler, D., Karpowicz, P., et al. (2011). A genome-scale shRNA resource for transgenic RNAi in Drosophila. Nat. Methods.

Northcott, P.A., Korshunov, A., Witt, H., Hielscher, T., Eberhart, C.G., Mack, S., Bouffet, E., Clifford, S.C., Hawkins, C.E., French, P., et al. (2011). Medulloblastoma comprises four distinct molecular variants. J. Clin. Oncol. *29*, 1408–1414.

Northcott, P.A., Jones, D.T., Kool, M., Robinson, G.W., Gilbertson, R.J., Cho, Y.J., Pomeroy, S.L., Korshunov, A., Lichter, P., Taylor, M.D., et al. (2012). Medulloblastomics: the end of the beginning. Nat. Rev. Cancer *12*, 818–834.

Northcott, P.A., Buchhalter, I., Morrissy, A.S., Hovestadt, V., Weischenfeldt, J., Ehrenberger, T., Gröbner, S., Segura-Wang, M., Zichner, T., Rudneva, V.A., et al. (2017). The whole-genome landscape of medulloblastoma subtypes. Nature.

O'Dushlaine, C., Kenny, E., Heron, E., Donohoe, G., Gill, M., Morris, D., International Schizophrenia, C., and Corvin, A. (2011). Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. Mol. Psychiatry *16*, 286–292.

O'Roak, B.J., Vives, L., Fu, W., Egertson, J.D., Stanaway, I.B., Phelps, I.G., Carvill, G., Kumar, A., Lee, C., Ankenman, K., et al. (2012). Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science (80-.).

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247–257.

Olivier, M., Hollstein, M., and Hainaut, P. (2010). TP53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harb Perspect Biol *2*, a001008.

Orr, B.A., Haffner, M.C., Nelson, W.G., Yegnasubramanian, S., and Eberhart, C.G. (2012). Decreased 5-hydroxymethylcytosine is associated with neural progenitor phenotype in normal brain and shorter survival in malignant glioma. PLoS One *7*, e41036.

Packer, R.J., Zhou, T., Holmes, E., Vezina, G., and Gajjar, A. (2013). Survival and secondary tumors in children with medulloblastoma receiving radiotherapy and adjuvant chemotherapy: results of Children's Oncology Group trial A9961. Neuro. Oncol. *15*, 97–103.

Paez-Colasante, X., Figueroa-Romero, C., Sakowski, S.A., Goutman, S.A., and Feldman, E.L. (2015). Amyotrophic lateral sclerosis: mechanisms and therapeutics in the epigenomic era. Nat Rev Neurol *11*, 266–279.

Paluszczak, J., and Baer-Dubowska, W. (2006a). Epigenetic diagnostics of cancer--the application of DNA methylation markers. J Appl Genet *47*, 365–375.

Paluszczak, J., and Baer-Dubowska, W. (2006b). Epigenetic diagnostics of cancer - The application of DNA methylation markers. J. Appl. Genet.

Pan, G., and Thomson, J.A. (2007). Nanog and transcriptional networks in embryonic stem cell pluripotency. Cell Res.

Pandey, U.B., and Nichols, C.D. (2011). Human Disease Models in Drosophila melanogaster and the Role of the Fly in Therapeutic Drug Discovery. Pharmacol. Rev.

Pang, S.Y., Hsu, J.S., Teo, K.C., Li, Y., Kung, M.H.W., Cheah, K.S.E., Chan, D., Cheung, K.M.C., Li, M., Sham, P.C., et al. (2017). Burden of rare variants in ALS genes influences survival in familial and sporadic ALS. Neurobiol. Aging *58*, 238 e9-238 e15.

Papin, C., Ibrahim, A., Le Gras, S., Velt, A., Stoll, I., Jost, B., Menoni, H., Bronner, C., Dimitrov, S., and Hamiche, A. (2017). Combinatorial DNA methylation codes at repetitive elements. Genome Res.

Perera, A., Eisen, D., Wagner, M., Laube, S.K., Künzel, A.F., Koch, S., Steinbacher, J., Schulze, E., Splith, V., Mittermeier, N., et al. (2015). TET3 is recruited by REST for context-specific hydroxymethylation and induction of gene expression. Cell Rep.

Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., and Salzberg, S.L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protoc.

Pietsch, T., Schmidt, R., Remke, M., Korshunov, A., Hovestadt, V., Jones, D.T.W., Felsberg, J., Kaulich, K., Goschzik, T., Kool, M., et al. (2014). Prognostic significance of clinical, histopathological, and molecular characteristics of medulloblastomas in the prospective HIT2000 multicenter clinical trial cohort. Acta Neuropathol.

Plank, J.L., and Dean, A. (2014). Enhancer function: Mechanistic and genome-wide insights come together. Mol. Cell.

Plongthongkum, N., Diep, D.H., and Zhang, K. (2014). Advances in the profiling of DNA modifications: cytosine methylation and beyond. Nat Rev Genet *15*, 647–661.

Plun-Favreau, H., Lewis, P.A., Hardy, J., Martins, L.M., and Wood, N.W. (2010). Cancer and neurodegeneration: Between the devil and the deep blue sea. PLoS Genet.

Po, A., Ferretti, E., Miele, E., De Smaele, E., Paganelli, A., Canettieri, G., Coni, S., Di Marcotullio, L., Biffoni, M., Massimi, L., et al. (2010). Hedgehog controls neural stem cells through p53-independent regulation of Nanog. EMBO J.

Poduri, A. (2015). Meta-Analysis Revives Genome-Wide Association Studies in Epilepsy. Epilepsy Curr *15*, 122–123.

Pugh, T.J., Weeraratne, S.D., Archer, T.C., Pomeranz Krummel, D.A., Auclair, D., Bochicchio, J., Carneiro, M.O., Carter, S.L., Cibulskis, K., Erlich, R.L., et al. (2012a). Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. Nature *488*, 106–110.

Pugh, T.J., Weeraratne, S.D., Archer, T.C., Pomeranz Krummel, D. a., Auclair, D., Bochicchio, J., Carneiro, M.O., Carter, S.L., Cibulskis, K., Erlich, R.L., et al. (2012b). Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. Nature *488*, 106–110.

Putiri, E.L., and Robertson, K.D. (2011). Epigenetic mechanisms and genome stability. Clin. Epigenetics.

Qureshi, I.A., and Mehler, M.F. (2013). Developing epigenetic diagnostics and therapeutics for brain

disorders. Trends Mol Med *19*, 732–741.

Rabbani, B., Tekin, M., and Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. J. Hum. Genet.

Rasmussen, K.D., Jia, G., Johansen, J. V., Pedersen, M.T., Rapin, N., Bagger, F.O., Porse, B.T., Bernard, O.A., Christensen, J., and Helin, K. (2015). Loss of TET2 in hematopoietic cells leads to DNA hypermethylation of active enhancers and induction of leukemogenesis. Genes Dev.

Renton, A.E., Chio, A., and Traynor, B.J. (2014). State of play in amyotrophic lateral sclerosis genetics. Nat. Neurosci. *17*, 17–23.

Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. *47*, D886–D894.

Van Rheenen, W., Pulit, S.L., Dekker, A.M., Al Khleifat, A., Brands, W.J., Iacoangeli, A., Kenna, K.P., Kavak, E., Kooyman, M., McLaughlin, R.L., et al. (2018). Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. Eur. J. Hum. Genet.

Robberecht, W., and Philips, T. (2013). The changing scene of amyotrophic lateral sclerosis. Nat. Rev. Neurosci. *14*, 248–264.

Robertson, K.D. (2005a). DNA methylation and human disease. Nat Rev Genet *6*, 597–610.

Robertson, K.D. (2005b). DNA methylation and human disease. Nat. Rev. Genet.

Sahraeian, S.M., Luo, K.R., and Brenner, S.E. (2015). SIFTER search: A web server for accurate phylogeny-based protein function prediction. Nucleic Acids Res.

Salk, J.J., Schmitt, M.W., and Loeb, L.A. (2018). Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nat. Rev. Genet.

Sardina, J.L., Collombet, S., Tian, T. V., Gómez, A., Di Stefano, B., Berenguer, C., Brumbaugh, J., Stadhouders, R., Segura-Morales, C., Gut, M., et al. (2018). Transcription Factors Drive Tet2-Mediated Enhancer Demethylation to Reprogram Cell Fate. Cell Stem Cell.

Schizophrenia Working Group of the Psychiatric Genomics, C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.

Schmahmann, J.D. (2004). Disorders of the Cerebellum: Ataxia, Dysmetria of Thought, and the Cerebellar Cognitive Affective Syndrome. J. Neuropsychiatry Clin. Neurosci. *16*, 367–378.

Schubeler, D. (2015). Function and information content of DNA methylation. Nature *517*, 321–326.

Schüller, U., Heine, V.M., Mao, J., Kho, A.T., Dillon, A.K., Han, Y.G., Huillard, E., Sun, T., Ligon, A.H., Qian, Y., et al. (2008). Acquisition of Granule Neuron Precursor Identity Is a Critical Determinant of Progenitor Cell Competence to Form Shh-Induced Medulloblastoma. Cancer Cell.

Schwalbe, E.C., Lindsey, J.C., Nakjang, S., Crosier, S., Smith, A.J., Hicks, D., Rafiee, G., Hill, R.M., Iliasova, A., Stone, T., et al. (2017). Novel molecular subgroups for clinical classification and outcome prediction in childhood medulloblastoma: a cohort study. Lancet Oncol.

Shalem, O., Sanjana, N.E., and Zhang, F. (2015). High-throughput functional genomics using CRISPR–Cas9. Nat. Rev. Genet. *16*, 299–311.

Shapiro, J.A., and Von Sternberg, R. (2005). Why repetitive DNA is essential to genome function. Biol. Rev. Camb. Philos. Soc.

Shetty, A.C., Athri, P., Mondal, K., Horner, V.L., Steinberg, K.M., Patel, V., Caspary, T., Cutler, D.J., and Zwick, M.E. (2010). SeqAnt: a web service to rapidly identify and annotate DNA sequence variations. BMC Bioinformatics *11*, 471.

Shin, H., Liu, T., Manrai, A.K., and Liu, S.X. (2009). CEAS: Cis-regulatory element annotation system. Bioinformatics.

Sidman, R.L., and Rakic, P. (1973). Neuronal migration, with special reference to developing human brain: a review. Brain Res.

Sims, D., Mendes-Pereira, A.M., Frankum, J., Burgess, D., Cerone, M.A., Lombardelli, C., Mitsopoulos, C., Hakas, J., Murugaesu, N., Isacke, C.M., et al. (2011). High-throughput RNA interference screening

using pooled shRNA libraries and next generation sequencing. Genome Biol.

Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet *8*, 272–285.

Song, C.X., and He, C. (2013). Potential functional roles of DNA demethylation intermediates. Trends Biochem Sci *38*, 480–484.

Song, C.-X., Szulwach, K.E., Fu, Y., Dai, Q., Yi, C., Li, X., Li, Y., Chen, C.-H., Zhang, W., Jian, X., et al. (2011a). Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. Nat. Biotechnol. *29*, 68–72.

Song, C.-X., Yin, S., Ma, L., Wheeler, A., Chen, Y., Zhang, Y., Liu, B., Xiong, J., Zhang, W., Hu, J., et al. (2017a). 5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. Cell Res *27*, 1231–1242.

Song, C.X., Szulwach, K.E., Fu, Y., Dai, Q., Yi, C., Li, X., Li, Y., Chen, C.H., Zhang, W., Jian, X., et al. (2011b). Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. Nat Biotechnol *29*, 68–72.

Song, C.X., Szulwach, K.E., Fu, Y., Dai, Q., Yi, C., Li, X., Li, Y., Chen, C.H., Zhang, W., Jian, X., et al. (2011c). Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. Nat Biotechnol *29*, 68–72.

Song, C.X., Szulwach, K.E., Fu, Y., Dai, Q., Yi, C., Li, X., Li, Y., Chen, C.H., Zhang, W., Jian, X., et al. (2011d). Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. Nat Biotechnol *29*, 68–72.

Song, C.X., Clark, T.A., Lu, X.Y., Kislyuk, A., Dai, Q., Turner, S.W., He, C., and Korlach, J. (2011e). Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. Nat Methods *9*, 75–77.

Song, L., Jia, J., Peng, X., Xiao, W., and Li, Y. (2017b). The performance of the SEPT9 gene methylation assay and a comparison with other CRC screening tests: A meta-analysis. Sci Rep *7*, 3032.

Spielman, R.S., McGinnis, R.E., and Ewens, W.J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet.

Sun, M., Song, M.M., Wei, B., Gao, Q., Li, L., Yao, B., Chen, L., Lin, L., Dai, Q., Zhou, X., et al. (2016). 5-Hydroxymethylcytosine-mediated alteration of transposon activity associated with the exposure to adverse in utero environments in human. Hum. Mol. Genet.

Supek, F., Lehner, B., Hajkova, P., and Warnecke, T. (2014). Hydroxymethylated Cytosines Are Associated with Elevated C to G Transversion Rates. PLoS Genet.

Suva, M.L., Riggi, N., and Bernstein, B.E. (2013). Epigenetic reprogramming in cancer. Science (80-.). *339*, 1567–1570.

Szeto, D.P., Rodriguez-Esteban, C., Ryan, A.K., O'Connell, S.M., Liu, F., Kioussi, C., Gleiberman, A.S., Izpisúa-Belmonte, J.C., and Rosenfeld, M.G. (1999). Role of the Bicoid-related homeodomain factor Pitx1 in specifying hindlimb morphogenesis and pituitary development. Genes Dev.

Szulwach, K.E., Li, X., Li, Y., Song, C.X., Wu, H., Dai, Q., Irier, H., Upadhyay, A.K., Gearing, M., Levey, A.I., et al. (2011). 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. Nat. Neurosci.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W. a, Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., et al. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science *324*, 930–935.

Takai, H., Masuda, K., Sato, T., Sakaguchi, Y., Suzuki, T., Suzuki, T., Koyama-Nasu, R., Nasu-Nishimura, Y., Katou, Y., Ogawa, H., et al. (2014). 5-Hydroxymethylcytosine Plays a Critical Role in Glioblastomagenesis by Recruiting the CHTOP-Methylosome Complex. Cell Rep. *9*, 48–60.

Tan, L., Xiong, L., Xu, W., Wu, F., Huang, N., Xu, Y., Kong, L., Zheng, L., Schwartz, L., Shi, Y., et al. (2013). Genome-wide comparison of DNA hydroxymethylation in mouse embryonic stem cells and neural progenitor cells by a new comparative hMeDIP-seq method. Nucleic Acids Res.

Taylor, J.P., Brown Jr., R.H., and Cleveland, D.W. (2016). Decoding ALS: from genes to mechanism. Nature *539*, 197–206.

Taylor, M.D., Northcott, P. a., Korshunov, A., Remke, M., Cho, Y.J., Clifford, S.C., Eberhart, C.G., Parsons, D.W., Rutkowski, S., Gajjar, A., et al. (2012). Molecular subgroups of medulloblastoma: The current consensus. Acta Neuropathol. *123*, 465–472.

The Gene Ontology Consortium (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. *47*, D330–D338.

Thompson, E.M., Hielscher, T., Bouffet, E., Remke, M., Luu, B., Gururangan, S., McLendon, R.E., Bigner, D.D., Lipp, E.S., Perreault, S., et al. (2016). Prognostic value of medulloblastoma extent of resection after accounting for molecular subgroup: a retrospective integrated clinical and molecular analysis. Lancet Oncol.

Tiemeier, H., Lenroot, R.K., Greenstein, D.K., Tran, L., Pierson, R., and Giedd, J.N. (2010). Cerebellum development during childhood and adolescence: A longitudinal morphometric MRI study. Neuroimage.

Umoh, M.E., Fournier, C., Li, Y., Polak, M., Shaw, L., Landers, J.E., Hu, W., Gearing, M., and Glass, J.D. (2016). Comparative analysis of C9orf72 and sporadic disease in an ALS clinic population. Neurology *87*, 1024–1030.

Uribe-Lewis, S., Stark, R., Carroll, T., Dunning, M.J., Bachman, M., Ito, Y., Stojic, L., Halim, S., Vowler, S.L., Lynch, A.G., et al. (2015). 5-hydroxymethylcytosine marks promoters in colon that resist DNA hypermethylation in cancer. Genome Biol. *16*, 30–35.

Vincent, J.J., Huang, Y., Chen, P.Y., Feng, S., Calvopiña, J.H., Nee, K., Lee, S.A., Le, T., Yoon, A.J., Faull, K., et al. (2013). Stage-specific roles for Tet1 and Tet2 in DNA demethylation in primordial germ cells. Cell Stem Cell.

Vincent, Q.B., Shang, L., Casanova, J.-L., Belkadi, A., Antipenko, A., Abel, L., Cobat, A., Bolze, A., Boisson, B., and Itan, Y. (2015). Whole-genome sequencing is more powerful than whole-exome

sequencing for detecting exome variants. Proc. Natl. Acad. Sci.

Waddington, C.H. (2012). The epigenotype. 1942. Int J Epidemiol *41*, 10–13.

Wang, F., Yang, Y., Lin, X., Wang, J.Q., Wu, Y.S., Xie, W., Wang, D., Zhu, S., Liao, Y.Q., Sun, Q., et al. (2013). Genome-wide loss of 5-hmC is a novel epigenetic feature of huntington's disease. Hum. Mol. Genet.

Wang, J., Garancher, A., Ramaswamy, V., and Wechsler-Reya, R.J. (2018). Medulloblastoma: From Molecular Subgroups to Molecular Targeted Therapies. Annu Rev Neurosci *41*, 207–232.

Wang, S.S.H., Kloth, A.D., and Badura, A. (2014). The Cerebellum, Sensitive Periods, and Autism. Neuron.

Wang, T., Pan, Q., Lin, L., Szulwach, K.E., Song, C.X., He, C., Wu, H., Warren, S.T., Jin, P., Duan, R., et al. (2012). Genome-wide DNA hydroxymethylation changes are associated with neurodevelopmental genes in the developing human cerebellum. Hum. Mol. Genet. *21*, 5500–5510.

Wang, Y., Xiao, M., Chen, X., Chen, L., Xu, Y., Lv, L., Wang, P., Yang, H., Ma, S., Lin, H., et al. (2015). WT1 recruits TET2 to regulate its target gene expression and suppress leukemia cell proliferation. Mol. Cell.

Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Pääbo, S., Rebhan, M., and Schübeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat. Genet.

Wen, X., Tan, W., Westergard, T., Krishnamurthy, K., Markandaiah, S.S., Shi, Y., Lin, S., Shneider, N.A., Monaghan, J., Pandey, U.B., et al. (2014). Antisense proline-arginine RAN dipeptides linked to C9ORF72-ALS/FTD form toxic nuclear aggregates that initiate in vitro and in vivo neuronal death. Neuron *84*, 1213–1225.

Williams, K., Christensen, J., Pedersen, M.T., Johansen, J. V, Cloos, P.A., Rappsilber, J., and Helin, K. (2011). TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature *473*,

343–348.

Wilson, V.L., Smith, R.A., Ma, S., and Cutler, R.G. (1987). Genomic 5-methyldeoxycytidine decreases with age. J. Biol. Chem.

Wingo, T.S., Cutler, D.J., Yarab, N., Kelly, C.M., and Glass, J.D. (2011). The heritability of amyotrophic lateral sclerosis in a clinically ascertained United States research registry. PLoS One *6*, e27985.

Wingo, T.S., Kotlar, A., and Cutler, D.J. (2017). MPD: multiplex primer design for next-generation targeted sequencing. BMC Bioinformatics *18*, 14.

Winter, D.J., Ganley, A.R.D., Young, C.A., Liachko, I., Schardl, C.L., Dupont, P.Y., Berry, D., Ram, A., Scott, B., and Cox, M.P. (2018). Repeat elements organise 3D genome structure and mediate transcription in the filamentous fungus Epichloë festucae. PLoS Genet.

Wu, X., and Zhang, Y. (2017). TET-mediated active DNA demethylation: mechanism, function and beyond. Nat Rev Genet *18*, 517–534.

Wu, H., D'Alessio, A.C., Ito, S., Xia, K., Wang, Z., Cui, K., Zhao, K., Sun, Y.E., and Zhang, Y. (2011a). Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. Nature *473*, 389–393.

Wu, H., D'Alessio, A.C., Ito, S., Wang, Z., Cui, K., Zhao, K., Sun, Y.E., and Zhang, Y. (2011b). Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. Genes Dev.

Wu, H., Wu, X., and Zhang, Y. (2016). Base-resolution profiling of active DNA demethylation using MAB-seq and caMAB-seq. Nat. Protoc.

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011c). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.

Xu, W., Yang, H., Liu, Y., Yang, Y., Wang, P., Kim, S.H., Ito, S., Yang, C., Wang, P., Xiao, M.T., et al. (2011a). Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α-ketoglutarate-dependent

dioxygenases. Cancer Cell.

Xu, Y., Wu, F., Tan, L., Kong, L., Xiong, L., Deng, J., Barbera, A.J., Zheng, L., Zhang, H., Huang, S., et al. (2011b). Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. Mol Cell *42*, 451–464.

Xu, Y., Xu, C., Kato, A., Tempel, W., Abreu, J.G., Bian, C., Hu, Y., Hu, D., Zhao, B., Cerovina, T., et al. (2012). Tet3 CXXC domain and dioxygenase activity cooperatively regulate key genes for xenopus eye and neural development. Cell.

Xu, Z., Poidevin, M., Li, X., Li, Y., Shu, L., Nelson, D.L., Li, H., Hales, C.M., Gearing, M., Wingo, T.S., et al. (2013). Expanded GGGGCC repeat RNA associated with amyotrophic lateral sclerosis and frontotemporal dementia causes neurodegeneration. Proc. Natl. Acad. Sci. U. S. A. *110*, 7778–7783.

Yang, H., Liu, Y., Bai, F., Zhang, J.-Y., Ma, S.-H., Liu, J., Xu, Z.-D., Zhu, H.-G., Ling, Z.-Q., Ye, D., et al. (2012). Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. Oncogene 663–669.

Yang, Y., Muzny, D.M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang, M., Buhay, C., et al. (2014). Molecular findings among patients referred for clinical whole-exome sequencing. JAMA - J. Am. Med. Assoc.

Yao, B., Lin, L., Street, R.C., Zalewski, Z.A., Galloway, J.N., Wu, H., Nelson, D.L., and Jin, P. (2014). Genome-wide alteration of 5-hydroxymethylcytosine in a mouse model of fragile X-associated tremor/ataxia syndrome. Hum. Mol. Genet. *23*, 1095–1107.

Yao, B., Christian, K.M., He, C., Jin, P., Ming, G.L., and Song, H. (2016). Epigenetic mechanisms in neurogenesis. Nat Rev Neurosci *17*, 537–549.

Yao, B., Cheng, Y., Wang, Z., Li, Y., Chen, L., Huang, L., Zhang, W., Chen, D., Wu, H., Tang, B., et al. (2017). DNA N6-methyladenine is dynamically regulated in the mouse brain following environmental stress. Nat Commun *8*, 1122.

Ylisaukko-oja, T., Alarcon, M., Cantor, R.M., Auranen, M., Vanhala, R., Kempas, E., von Wendt, L., Jarvela, I., Geschwind, D.H., and Peltonen, L. (2006). Search for autism loci by combined analysis of Autism Genetic Resource Exchange and Finnish families. Ann Neurol *59*, 145–155.

Yoder, J.A., and Bestor, T.H. (1998). A candidate mammalian DNA methyltransferase related to pmt1p of fission yeast. Hum Mol Genet *7*, 279–284.

Yoo, Y., Park, J.H., Weigel, C., Liesenfeld, D.B., Weichenhan, D., Plass, C., Seo, D.G., Lindroth, A.M., and Park, Y.J. (2017). TET-mediated hydroxymethylcytosine at the Pparγ locus is required for initiation of adipogenic differentiation. Int. J. Obes.

Yoshida, Y., Makita, Y., Heida, N., Asano, S., Matsushima, A., Ishii, M., Mochizuki, Y., Masuya, H., Wakana, S., Kobayashi, N., et al. (2009). PosMed (Positional Medline): Prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. Nucleic Acids Res.

Yoshino, S., Cilluffo, R., Best, P.J., Atkinson, E.J., Aoki, T., Cunningham, J.M., de Andrade, M., Choi, B.J., Lerman, L.O., and Lerman, A. (2014). Single nucleotide polymorphisms associated with abnormal coronary microvascular function. Coron. Artery Dis. *25*, 281–289.

Yu, M., Hon, G.C., Szulwach, K.E., Song, C.X., Zhang, L., Kim, A., Li, X., Dai, Q., Shen, Y., Park, B., et al. (2012). Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell *149*, 1368–1380.

Zeng, Y., Yao, B., Shin, J., Lin, L., Kim, N., Song, Q., Liu, S., Su, Y., Guo, J.U., Huang, L., et al. (2016). Lin28A Binds Active Promoters and Recruits Tet1 to Regulate Gene Expression. Mol. Cell.

Zhang, Y., Liu, T., Meyer, C. a, Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

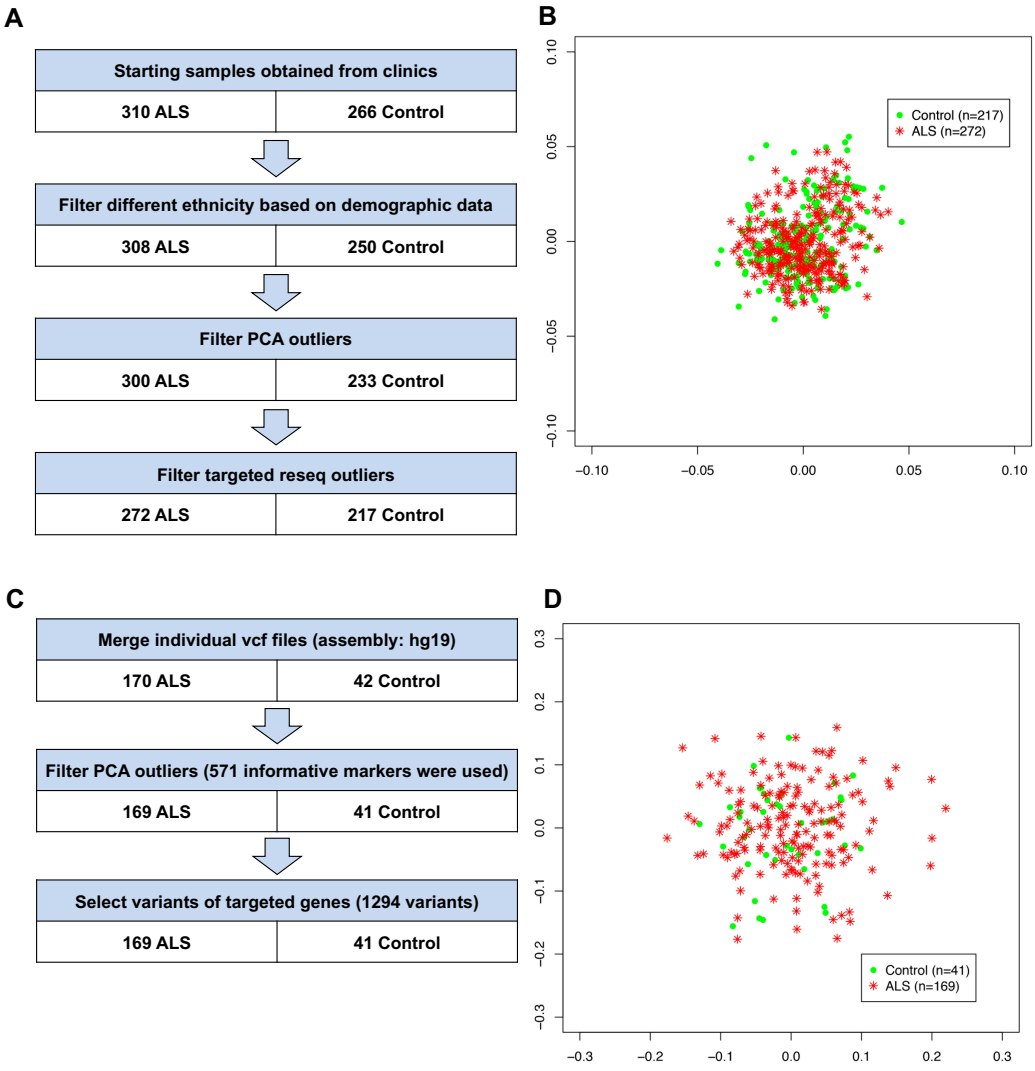
Zhang, Y.J., Gendron, T.F., Grima, J.C., Sasaguri, H., Jansen-West, K., Xu, Y.F., Katzman, R.B., Gass,

J., Murray, M.E., Shinohara, M., et al. (2016). C9ORF72 poly(GA) aggregates sequester and impair HR23 and nucleocytoplasmic transport proteins. Nat. Neurosci. *19*, 668–677.

Zhu, X., Girardo, D., Govek, E.E., John, K., Mellén, M., Tamayo, P., Mesirov, J.P., and Hatten, M.E. (2016). Role of Tet1/3 Genes and Chromatin Remodeling Genes in Cerebellar Circuit Formation. Neuron.

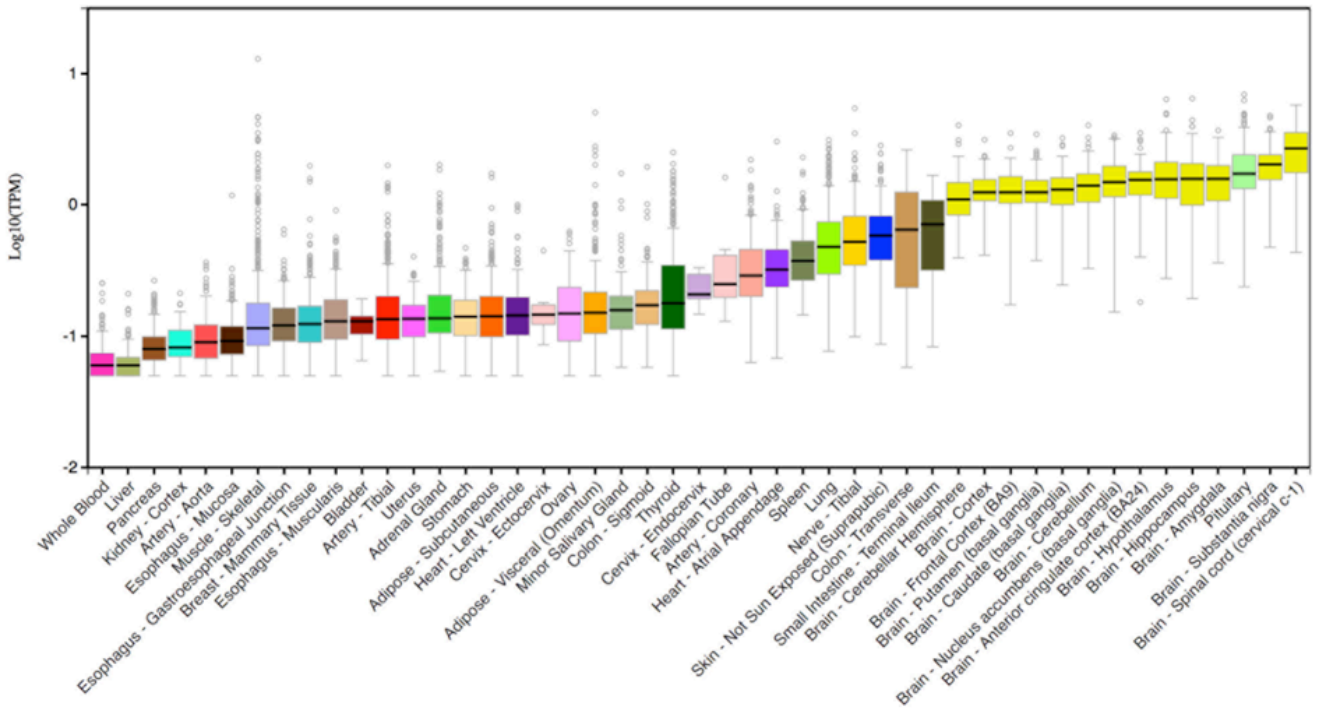
Zhukova, N., Ramaswamy, V., Remke, M., Pfaff, E., Shih, D.J.H., Martin, D.C., Castelo-Branco, P., Baskin, B., Ray, P.N., Bouffet, E., et al. (2013). Subgroup-specific prognostic implications of TP53 mutation in medulloblastoma. J. Clin. Oncol.

SUPPLEMENTAL FIGURES

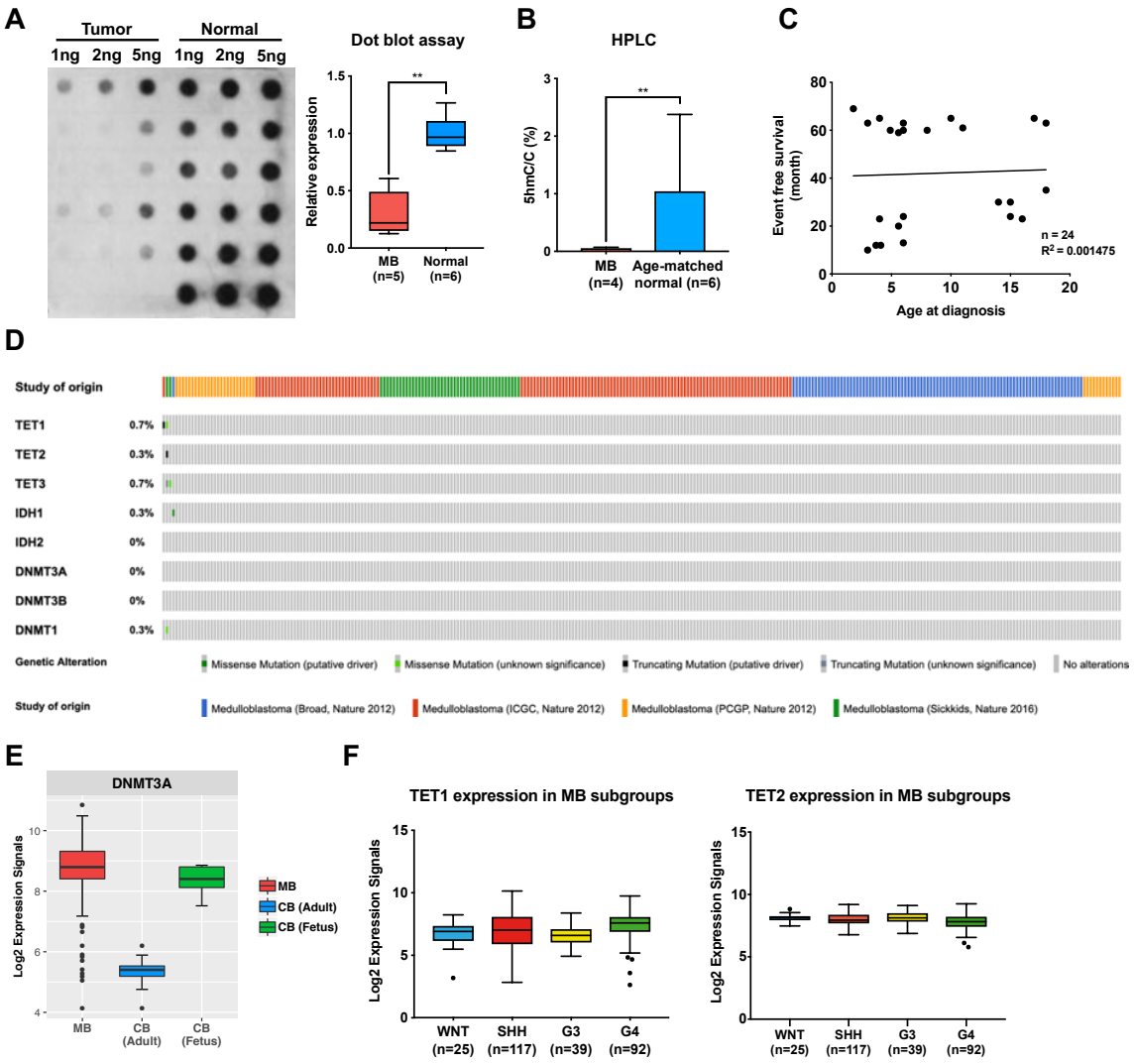


Supplemental Figure 1 Quality control of samples used for statistical testing using targeted resequencing and a validation Project MinE dataset. (A) 38 ALS cases and 49 controls were filtered out during each quality control step, and final 272 ALS cases and 217 controls passed quality control. Additional 2 ALS cases that have known ALS genes, SOD1 and TARDBP, were excluded for SKAT analysis. (B) Principal component analysis (PCA) plot of samples that passed quality control. (C) 169 ALS cases and 41 controls passed after PCA. Additional 2 ALS cases that have mutations in SOD1 were

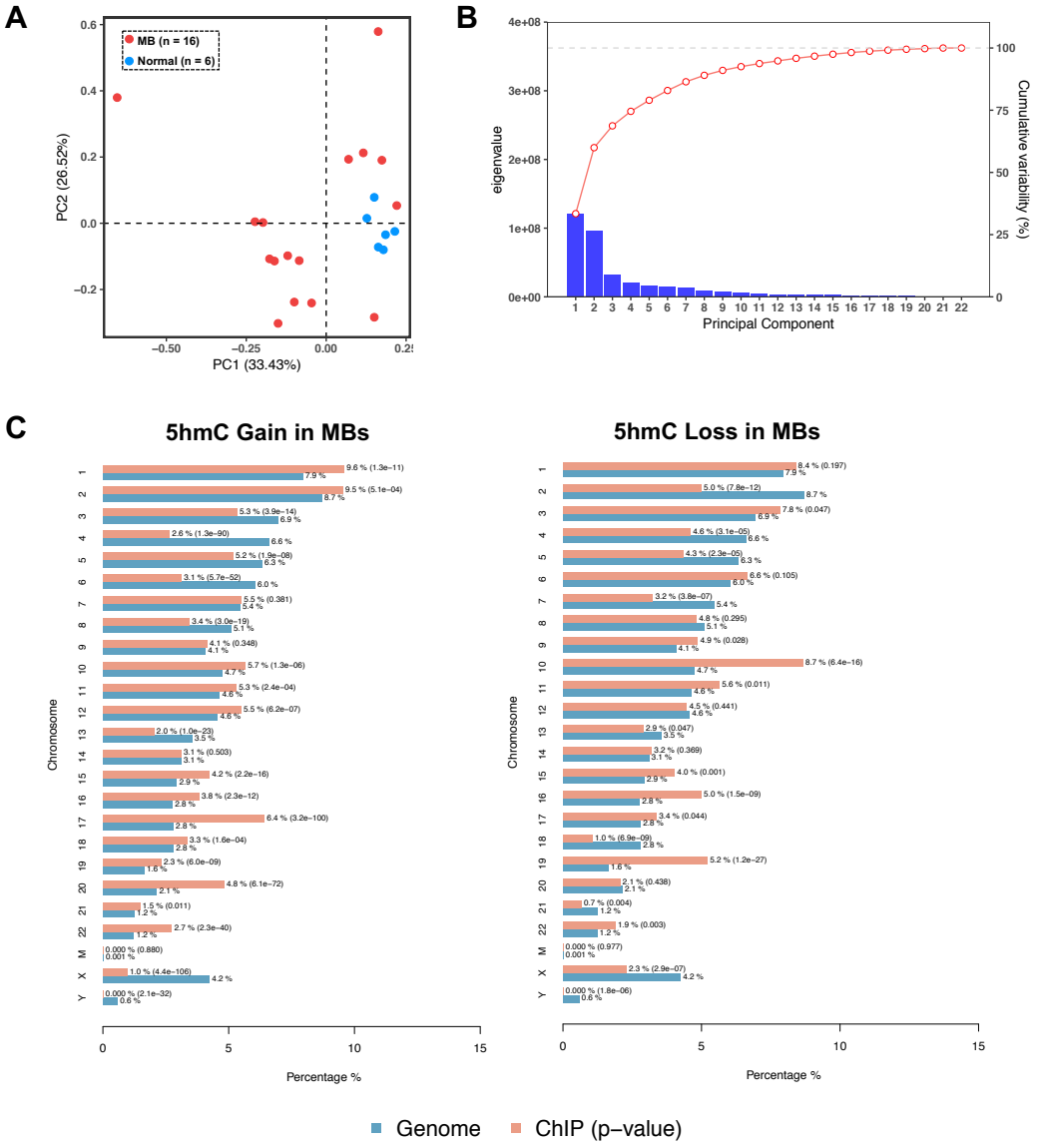
excluded for further analysis. (D) Principal component analysis (PCA) plot of ALS cases and controls passed quality control.



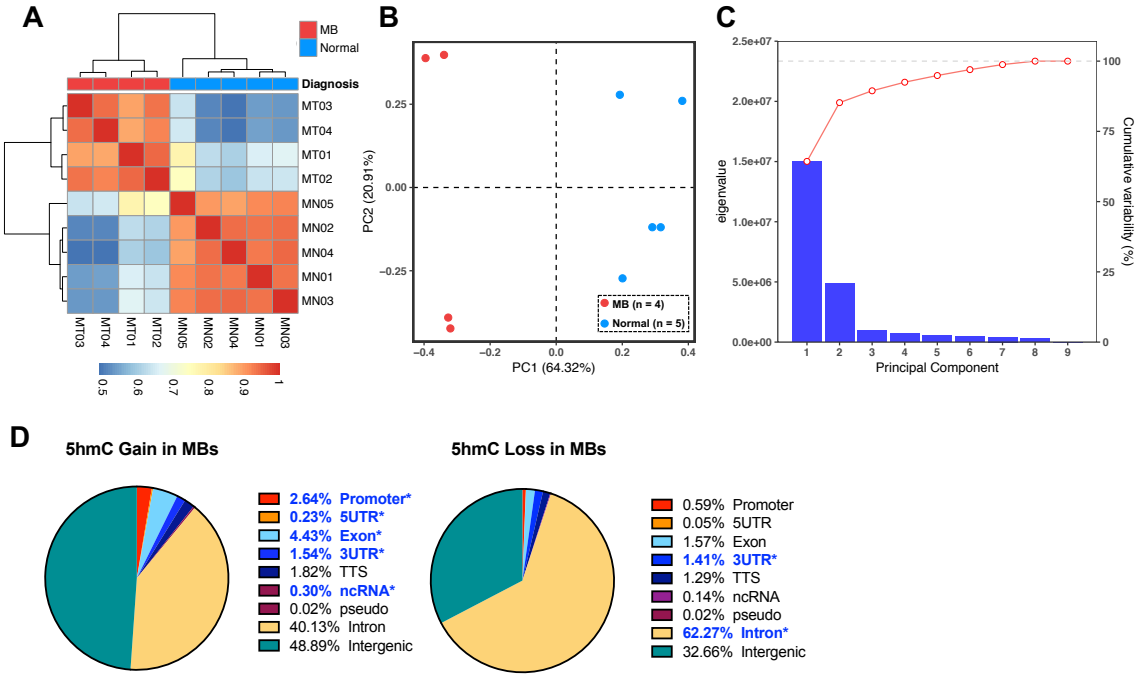
Supplemental Figure 2 Gene expression for MYH15 in 47 tissues obtained from GTEx Analysis Release V7 (dbGaP Accession phs000424.v7.p2). Expression values are shown in TPM (Transcripts Per Million)



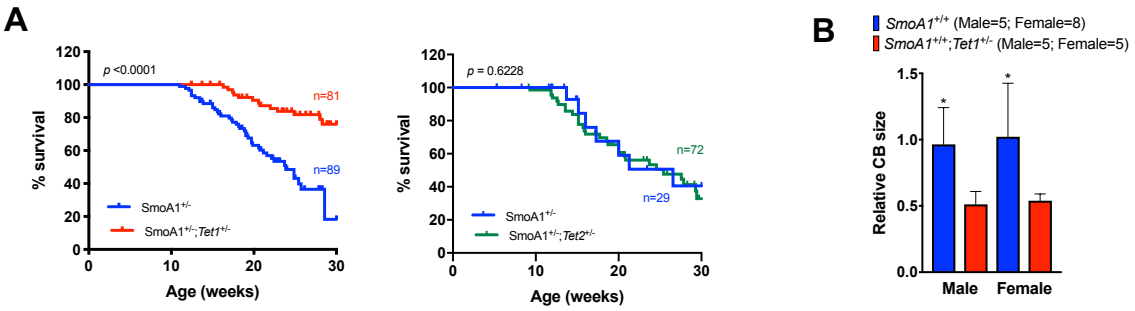
Supplemental Figure 3 (A) Dot blot assay of additional MB and normal cerebellum cohort. (B) 5hmC was significantly depleted in HPLC analysis. (C) Pearson correlation analysis between age at diagnosis and event-free survival (month). (n=24, $R^2 = 0.001476$, $p > 0.05$ (D) Mutation rates of enzymes involved in cytosine modifications. (D) Expression levels of DNMT3A in MB, CB (adult), and CB (fetus). (F) Expression levels of TET1 and TET2 in MB subgroups.



Supplemental Figure 4 (A, B) PCA analysis using human 5hmC profiles (left: PCA plot, right: scree plot). (C) CEAS annotation of human 5hmC gain (left) and loss (right) in MB



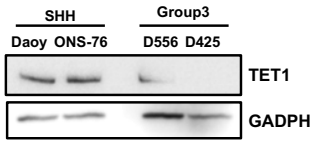
Supplemental Figure 5 (A) Heatmap using the result of Pearson correlation using 5hmC profiles in the murine MB. (B, C) PCA analysis using mouse 5hmC profiles (left: PCA plot, right: scree plot). (D) Genomic enrichment analysis of 5hmC gain and loss in murine MB.

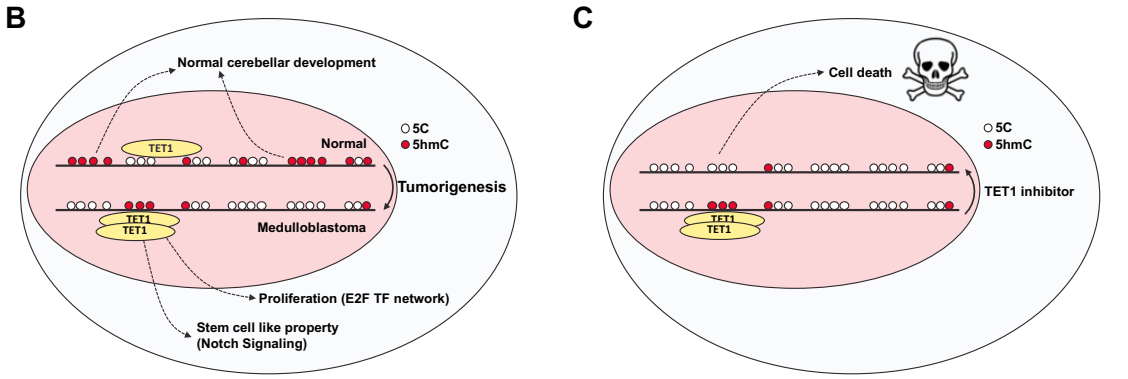


Supplemental Figure 6 (A) Kaplan-Meier curves show the significant survival difference of SmoA1$^{+/-}$ mice crossed with hemizygous deletion of *Tet1* (Left: $p < 0.0001$; log rank test), but not crossed with hemizygous deletion of *Tet2* ($p = 0.6228$; log rank test). (B) Cerebellum size is significantly different depending on genotype, but not sex.

A

Human: medulloblastoma cell line	Gender	Subtypes	p53	MYC
Daoy	male	SHH	mut	unknown
ONS-76	female	SHH	WT	unknown
D556	female	Grp3	unknown	amp
D425	male	Grp3	unknown	amp





Supplemental Figure 7 (A) Table showing characteristics of human MB cell lines used in this study (left) and TET1 expression in each cell line (right). (B) Model of oncogenic functions of 5hmC and TET1 in MB progression. (C) Model of TET1 inhibitor to lead to cell death.

SUPPLEMENTAL TABLES

Supplementary Table 1: Clinical information of four C9ALS patients included in whole-genome sequencing

Sample Group	Sample ID	Sex	Age of onset
YALS	SL119751	F	34.33
	SL119752	F	41.65
OALS	SL119753	F	72.37
	SL119754	M	72.86

Supplementary Table 2: List of genic variants exclusively found in either YALS or OALS in whole-genome sequencing

Patient type	chr	pos	snp id	ref	alt	maf	gene symbol	annotation type	aa change	str	Cadd	Phast Cons	phyloP	Heterozygotes ids	homozygote ids
YALS	chr1	15445709	rs515726210	T	A	NA	CTRC	Replacement	V251D	+	26.772	0.543	0.472	SL119751; SL119752	NA
YALS	chr1	15445753	NA	G	A	NA	CTRC	Non-Coding		+	12.382	0.74	0.709	SL119751; SL119752	NA
YALS	chr1	47794752	NA	G	C	NA	TRABD2B	Replacement	N274K	-	23.425	0.996	0.945	SL119751; SL119752	NA
YALS	chr1	67412777	NA	A	C	NA	SERBP1	Non-Coding		-	13.386	0.949	0.945	SL119751; SL119752	NA
YALS	chr1	67412839	NA	A	T	NA	SERBP1	Non-Coding		-	14.39	1	0.945	SL119751; SL119752	NA
YALS	chr1	67412861	NA	G	A	NA	SERBP1	Non-Coding		-	14.724	1	0.945	SL119751; SL119752	NA
YALS	chr1	92483381	rs34631763	C	T	0.009684505	GFI1	Replacement	S36N	-	18.406	0.201	0	SL119751; SL119752	NA
YALS	chr1	112600798	rs116580400	T	C	0.009784345	ST7L	Replacement	R168G	-	22.087	0.996	0.945	SL119751; SL119752	NA
YALS	chr1	115925198	rs10923803	G	T	0.002196486	RP11-12L8.1	Non-Coding		+	13.051	0.984	0.945	SL119751; SL119752	NA
YALS	chr1	155795096	NA	T	G	NA	GON4L	Replacement	E567D	-	17.067	1	0.945	SL119751; SL119752	NA
YALS	chr1	155795125	NA	C	A	NA	GON4L	Replacement	E558*	-	21.083	1	0.945	SL119752	NA
YALS	chr1	180114558	rs142500573	C	T	0.00778754	CEP350	Non-Coding		+	14.055	0.996	0.945	SL119751; SL119752	NA
YALS	chr1	247155424	NA	A	T	NA	ZNF124	Non-Coding		-	12.717	0.917	0.709	SL119751; SL119752	NA
YALS	chr1	247155432	NA	G	T	NA	ZNF124	Non-Coding		-	14.055	0.913	0.709	SL119751; SL119752	NA
YALS	chr1	247155436	NA	A	T	NA	ZNF124	Non-Coding		-	13.386	0.858	0.709	SL119751; SL119752	NA
YALS	chr1	247155444	NA	C	T	NA	ZNF124	Non-Coding		-	15.059	0.713	0.709	SL119751; SL119752	NA
YALS	chr1	247155445	NA	A	T	NA	ZNF124	Non-Coding		-	12.382	0.709	0.709	SL119751; SL119752	NA
YALS	chr1	248488874	rs199519459	G	A	NA	OR2T5	Replacement	V96I	+	17.736	0.024	0	SL119751; SL119752	NA
YALS	chr1	248489049	rs200959275	T	C	NA	OR2T5	Replacement	F154S	+	13.051	0	0	SL119751; SL119752	NA
YALS	chr1	248558785	NA	T	G	NA	OR2T29	Replacement	E236A	-	10.039	0.063	0	SL119751; SL119752	NA
YALS	chr2	96848428	NA	T	G	NA	ANKRD39	Replacement	H142P	-	18.74	0.937	0.709	SL119751; SL119752	NA

YALS	chr2	107892532	NA	A	G	NA	RGPD4	Non-Coding		+	13.72	0.988	0.945	SL119751; SL119752	NA
YALS	chr2	189873653	NA	A	C	NA	PMS1	Replacement	L715F	+	24.094	0.996	0.945	SL119751; SL119752	NA
YALS	chr2	203296493	NA	G	A	NA	CYP20A1	Replacement	D390N	+	22.756	0.965	0.945	SL119751; SL119752	NA
YALS	chr2	218459828	rs563821163	T	C	0.000199681	USP37	Replacement	T869A	-	21.417	1	0.945	SL119751; SL119752	NA
YALS	chr2	218459832	NA	C	A	NA	USP37	Replacement	E867D	-	20.748	0.996	0.945	SL119751; SL119752	NA
YALS	chr2	218459858	NA	C	T	NA	USP37	Replacement	G859R	-	32.795	0.996	0.945	SL119751; SL119752	NA
YALS	chr2	218459908	NA	G	A	NA	USP37	Non-Coding		-	14.724	0.98	0.945	SL119751; SL119752	NA
YALS	chr3	50256215	NA	T	C	NA	GNAI2	Replacement	I111T	+	22.087	0.965	0.945	SL119751; SL119752	NA
YALS	chr3	108493131	NA	G	T	NA	MYH15	Replacement	S273Y	-	22.087	0.039	0	SL119751; SL119752	NA
YALS	chr3	170038072	rs187596124	C	G	NA	GPR160	Non-Coding		+	13.051	0.15	0	SL119751; SL119752	NA
YALS	chr3	195292453	NA	C	T	NA	ACAP2	Non-Coding		-	26.102	0.996	0.945	SL119751; SL119752	NA
YALS	chr3	195297212	NA	T	A	NA	ACAP2	Replacement	K489*	-	22.421	1	0.945	SL119752	NA
YALS	chr4	86453	NA	T	G	NA	ZNF595	Replacement	C134G	+	15.394	0.906	0.709	SL119751; SL119752	NA
YALS	chr4	39823037	rs62307873	C	T	0.001797125	PDS5A	Non-Coding		-	15.059	1	0.945	SL119751; SL119752	NA
YALS	chr4	70824318	NA	C	A	NA	GRSF1	Non-Coding		-	13.72	0.437	0.236	SL119751; SL119752	NA
YALS	chr4	78499716	rs386676424; rs7695038	C	G	0.001397764	FRAS1	Non-Coding		+	10.374	0.98	0.945	NA	SL119751; SL119752
YALS	chr4	145138188	NA	A	G	NA	OTUD4	Replacement	Y798H	-	27.106	1	0.945	SL119751; SL119752	NA
YALS	chr4	145152617	NA	C	A	NA	OTUD4	Replacement	G233*	-	22.421	0.996	0.945	SL119751	NA
YALS	chr4	176151924	NA	G	T	NA	WDR17	Replacement	G830V	+	28.78	1	0.945	SL119751; SL119752	NA
YALS	chr5	132604931	NA	C	A	NA	RAD50	Replacement	R884S	+	25.433	0.996	0.945	SL119751; SL119752	NA
YALS	chr5	132604946	NA	G	A	NA	RAD50	Replacement	E889K	+	23.091	1	0.945	SL119751; SL119752	NA
YALS	chr5	132604949	NA	C	A	NA	RAD50	Replacement	Q890K	+	22.756	1	0.945	SL119751; SL119752	NA
YALS	chr5	132604956	NA	T	A	NA	RAD50	Replacement	V892E	+	22.756	0.996	0.945	SL119751; SL119752	NA

YALS	chr5	132604957	NA	G	A	NA	RAD50	Silent	V892V	+	10.709	0.996	0.945	SL119751; SL119752	NA
YALS	chr5	132604987	NA	C	A	NA	RAD50	Replacement	Y902*	+	35.807	0.862	0.709	SL119751; SL119752	NA
YALS	chr5	134360096	NA	T	G	NA	CDKL3	Non-Coding		-	18.74	0.996	0.945	SL119751; SL119752	NA
YALS	chr5	176881302	rs145827614	C	T	0.00778754	HK3	Replacement	R876H	-	16.063	0.925	0.709	SL119751; SL119752	NA
YALS	chr6	34876038	rs17703221	T	C	0.004193291	UHRF1BP1	Non-Coding		+	10.374	0.76	0.709	SL119751; SL119752	NA
YALS	chr6	35342634	rs11571502	G	C	0.00399361	PPARD	Non-Coding		+	15.059	0.043	0	SL119751; SL119752	NA
YALS	chr6	35411220	rs9658135	G	A	0.006988818	PPARD	Non-Coding		+	16.398	0.878	0.709	SL119751; SL119752	NA
YALS	chr6	35426727	rs9658170	G	A	0.006988818	PPARD	Non-Coding		+	10.709	0	0	SL119751; SL119752	NA
YALS	chr6	38906260	NA	C	G	NA	DNAH8	Replacement	Y3067*	+	43.839	0.988	0.945	SL119752	NA
YALS	chr6	38906299	NA	C	A	NA	DNAH8	Replacement	Y3080*	+	43.839	0.98	0.945	SL119752	NA
YALS	chr6	38906325	NA	G	A	NA	DNAH8	Replacement	G3089D	+	31.791	0.996	0.945	SL119751; SL119752	NA
YALS	chr6	75891294	NA	A	C	NA	MYO6	Replacement	E978D	+	15.059	0.693	0.472	SL119751; SL119752	NA
YALS	chr6	116431127	rs34994230	A	G	0.006888978	DSE	Replacement	I282V	+	23.425	0.996	0.945	SL119751; SL119752	NA
YALS	chr6	121449716	NA	A	G	NA	GJA1	Non-Coding		+	15.059	0.996	0.945	SL119751; SL119752	NA
YALS	chr6	138339851	NA	C	A	NA	KIAA1244	Non-Coding		+	11.043	0.039	0	SL119751; SL119752	NA
YALS	chr6	157208928	NA	C	T	NA	ARID1B	Non-Coding		+	14.055	0.996	0.945	SL119751; SL119752	NA
YALS	chr6	157208934	NA	C	T	NA	ARID1B	Non-Coding		+	12.717	0.992	0.945	SL119751; SL119752	NA
YALS	chr6	157208941	NA	G	T	NA	ARID1B	Non-Coding		+	13.051	0.996	0.945	SL119751; SL119752	NA
YALS	chr7	5943927	NA	A	C	NA	RSPH10B2	Replacement	N531K	-	23.091	0.024	0	SL119751; SL119752	NA
YALS	chr7	5943938	NA	T	G	NA	RSPH10B2	Replacement	I528L	-	24.429	0.449	0.236	SL119751; SL119752	NA
YALS	chr7	87904377	NA	A	C	NA	DBF4	Replacement	D337A	+	26.102	0.996	0.945	SL119751; SL119752	NA
YALS	chr7	87904378	NA	C	T	NA	DBF4	Silent	D337D	+	13.051	0.996	0.945	SL119751; SL119752	NA
YALS	chr7	87904409	NA	A	T	NA	DBF4	Replacement	K348*	+	40.827	0.972	0.945	SL119751	NA
YALS	chr7	92519028	NA	C	A	NA	PEX1	Silent	V108V	-	15.059	0.996	0.945	SL119751; SL119752	NA

YALS	chr7	107657564	NA	A	G	NA	SLC26A4-AS1	Non-Coding		-	10.039	0.008	0	SL119751; SL119752	NA
YALS	chr7	139561645	NA	A	T	NA	HIPK2	Non-Coding		-	15.394	1	0.945	SL119751; SL119752	NA
YALS	chr8	22693189	NA	T	C	NA	EGR3	Non-Coding		-	13.051	0.815	0.709	SL119751; SL119752	NA
YALS	chr8	41931283	rs117665105	T	C	0.006988818	KAT6A	Non-Coding		-	10.039	0.902	0.709	SL119751; SL119752	NA
YALS	chr8	93804878	rs201791586	G	A	0.000199681	TMEM67	Silent	A732A	+	14.39	0.992	0.945	SL119751; SL119752	NA
YALS	chr8	93804884	NA	A	G	NA	TMEM67	Non-Coding		+	10.374	0.791	0.709	SL119751; SL119752	NA
YALS	chr9	66918360	rs531820325	C	A	0.0000998	ZNF658	Replacement	T265K	+	14.39	0.004	0	SL119751; SL119752	NA
YALS	chr9	85677459	NA	A	T	NA	AGTPBP1	Replacement	I138N	-	26.772	0.984	0.945	SL119751; SL119752	NA
YALS	chr9	110429999	rs142508835	A	T	0.009185304	SVEP1	Replacement	S1846T	-	22.421	0.748	0.709	SL119751; SL119752	NA
YALS	chr9	137171651	rs72763276	G	A	0.003194888	TMEM210	Replacement	R72W	-	10.374	0.551	0.472	SL119751; SL119752	NA
YALS	chr9	137452391	NA	T	G	NA	NSMF	Replacement	I374L	-	23.76	1	0.945	SL119751; SL119752	NA
YALS	chr9	137452394	NA	T	G	NA	NSMF	Replacement	K373Q	-	23.425	0.996	0.945	SL119751; SL119752	NA
YALS	chr10	27195886	NA	T	G	NA	ACBD5	Non-Coding		-	10.374	0.094	0	SL119751; SL119752	NA
YALS	chr10	47001445	NA	C	T	NA	PTPN20A	Non-Coding		+	10.039	0.028	0	SL119752	SL119751
YALS	chr10	47523418	NA	C	T	NA	AGAP10	Replacement	E37K	-	12.717	0.008	0	SL119751; SL119752	NA
YALS	chr10	99794460	NA	G	A	NA	ABCC2	Replacement	W208*	+	37.815	0.996	0.945	SL119751; SL119752	NA
YALS	chr10	133289836	NA	T	C	NA	TUBGCP2	Replacement	I320V	-	17.067	0.996	0.945	SL119751; SL119752	NA
YALS	chr11	2138714	NA	G	C	NA	IGF2	Non-Coding		-	15.728	0.972	0.945	SL119751; SL119752	NA
YALS	chr11	47261271	NA	A	G	NA	NR1H3	Replacement	K177R	+	27.106	0.996	0.945	SL119751; SL119752	NA
YALS	chr11	62702462	NA	A	C	NA	BSCL2	Non-Coding		-	15.394	0.976	0.945	SL119751; SL119752	NA
YALS	chr11	84923413	rs146213168	C	T	0.009784345	DLG2	Non-Coding		-	19.075	0.988	0.945	SL119751; SL119752	NA
YALS	chr12	56586866	NA	T	A	NA	RBMS2	Silent	S297S	+	21.083	0.996	0.945	SL119751; SL119752	NA
YALS	chr12	56586893	NA	G	C	NA	RBMS2	Replacement	W306C	+	22.087	0.992	0.945	SL119751; SL119752	NA

YALS	chr12	68841234	NA	T	G	NA	MDM2	Non-Coding		+	11.043	0.433	0.236	SL119751; SL119752	NA
YALS	chr12	69820897	NA	A	T	NA	RAB3IP	Non-Coding		+	13.386	0.142	0	SL119751; SL119752	NA
YALS	chr12	101352119	NA	C	A	NA	UTP20	Replacement	S1650*	+	39.823	0.988	0.945	SL119752	NA
YALS	chr12	101352135	NA	C	A	NA	UTP20	Replacement	Y1655*	+	37.815	1	0.945	SL119752	NA
YALS	chr12	101352149	NA	T	A	NA	UTP20	Replacement	I1660N	+	31.791	1	0.945	SL119751; SL119752	NA
YALS	chr12	109188118	NA	T	G	NA	ACACB	Replacement	F700L	+	22.756	0.996	0.945	SL119751; SL119752	NA
YALS	chr13	101722800	rs149661933	C	T	0.004992013	FGF14	Non-Coding		-	11.043	0.555	0.472	SL119751; SL119752	NA
YALS	chr14	49895956	rs1064615	T	G	0.008985623	ARF6	Non-Coding		+	14.39	0.075	0	SL119751; SL119752	NA
YALS	chr14	52756579	NA	C	A	NA	STYX	Replacement	P91T	+	20.748	1	0.945	SL119751; SL119752	NA
YALS	chr14	52756605	NA	C	A	NA	STYX	Replacement	F99L	+	17.402	0.996	0.945	SL119751; SL119752	NA
YALS	chr14	55067260	NA	C	A	NA	MAPK1IP1L	Non-Coding		+	11.043	0.201	0	SL119751; SL119752	NA
YALS	chr15	38484265	NA	G	C	NA	FAM98B	Replacement	G303A	+	20.748	0.996	0.945	SL119751; SL119752	NA
YALS	chr15	43154830	NA	A	T	NA	TMEM62	Replacement	Q394L	+	18.406	0.992	0.945	SL119751; SL119752	NA
YALS	chr15	89195615	rs146738558	G	A	0.002995208	ABHD2	Non-Coding		+	15.059	1	0.945	SL119751; SL119752	NA
YALS	chr16	10747281	NA	G	A	NA	NUBP1	Non-Coding		+	10.374	0.276	0.236	SL119751; SL119752	NA
YALS	chr16	23562624	NA	G	A	NA	UBFD1	Non-Coding		+	21.752	0.925	0.709	SL119751; SL119752	NA
YALS	chr16	77284045	NA	T	A	NA	ADAMTS18	Replacement	N1193Y	-	16.732	0.996	0.945	SL119751; SL119752	NA
YALS	chr16	87419785	NA	G	A	NA	ZCCHC14	Replacement	P211L	-	22.421	1	0.945	SL119751; SL119752	NA
YALS	chr16	87419794	NA	A	G	NA	ZCCHC14	Replacement	L208P	-	25.433	1	0.945	SL119751; SL119752	NA
YALS	chr16	87419831	NA	T	A	NA	ZCCHC14	Replacement	R196W	-	26.102	0.945	0.945	SL119751; SL119752	NA
YALS	chr16	89919683	rs11547464	G	A	0.002795527	MC1R	Replacement	R142H	+	22.087	1	0.945	SL119751; SL119752	NA
YALS	chr17	6695727	NA	T	A	NA	SLC13A5	Replacement	K352*	-	33.799	0.012	0	SL119751; SL119752	NA
YALS	chr17	7602906	NA	C	T	NA	FXR2	Non-Coding		-	16.063	0.866	0.709	SL119751; SL119752	NA

YALS	chr17	18805539	rs2589696	A	G	NA	TVP23B	Non-Coding		+	11.378	0.839	0.709	SL119751; SL119752	NA
YALS	chr17	39066886	NA	G	C	NA	PLXDC1	Non-Coding		-	11.713	0.039	0	SL119751; SL119752	NA
YALS	chr17	40701933	NA	T	A	NA	KRT24	Replacement	I206F	-	29.783	0.5	0.472	SL119751; SL119752	NA
YALS	chr17	50135246	NA	A	G	NA	PPP1R9B	Non-Coding		-	13.72	0.992	0.945	SL119751; SL119752	NA
YALS	chr17	65225096	rs34797451	G	A	0.00399361	RGS9	Replacement	R498H	+	22.756	0.555	0.472	SL119751; SL119752	NA
YALS	chr18	11854252	NA	T	C	NA	CHMP1B	Non-Coding		+	19.744	0.409	0.236	SL119751; SL119752	NA
YALS	chr18	11854300	NA	C	T	NA	CHMP1B	Non-Coding		+	12.717	0.067	0	SL119751; SL119752	NA
YALS	chr18	79342309	NA	C	A	NA	ATP9B	Replacement	C775*	+	12.047	0.992	0.945	SL119751	NA
YALS	chr19	111016	rs200336441	T	G	NA	OR4F17	Replacement	F113C	+	10.709	0.913	0.709	SL119751; SL119752	NA
YALS	chr19	3275972	NA	G	C	NA	CELF5	Replacement	G171R	+	18.74	0.953	0.945	SL119751; SL119752	NA
YALS	chr19	3275973	NA	G	C	NA	CELF5	Replacement	G171A	+	18.74	0.961	0.945	SL119751; SL119752	NA
YALS	chr19	3275975	NA	A	C	NA	CELF5	Replacement	S172R	+	17.402	0.972	0.945	SL119751; SL119752	NA
YALS	chr19	3275976	NA	G	C	NA	CELF5	Replacement	S172T	+	17.067	0.972	0.945	SL119751; SL119752	NA
YALS	chr19	5744446	NA	T	A	NA	CATSPERD	Replacement	L198*	+	33.799	0.008	0	SL119751; SL119752	NA
YALS	chr19	5744449	NA	G	A	NA	CATSPERD	Replacement	G199D	+	15.728	0.012	0	SL119751; SL119752	NA
YALS	chr19	5768223	NA	T	A	NA	CATSPERD	Replacement	Y539N	+	23.425	0.028	0	SL119751; SL119752	NA
YALS	chr19	6667118	NA	A	T	NA	TNFSF14	Replacement	L62H	-	12.047	0.992	0.945	SL119751; SL119752	NA
YALS	chr19	6667130	NA	G	T	NA	TNFSF14	Replacement	P58Q	-	10.039	0.921	0.709	SL119751; SL119752	NA
YALS	chr19	10173904	NA	T	A	NA	DNMT1	Replacement	D217V	-	18.071	0.819	0.709	SL119751; SL119752	NA
YALS	chr19	16552144	NA	A	G	NA	SLC35E1	Non-Coding		-	10.374	0.165	0	SL119751; SL119752	NA
YALS	chr19	19933957	NA	T	A	NA	ZNF93	Silent	I334I	+	16.732	0.858	0.709	SL119751; SL119752	NA
YALS	chr19	19933962	NA	C	T	NA	ZNF93	Replacement	T336I	+	16.398	0.78	0.709	SL119751; SL119752	NA
YALS	chr19	19933969	NA	G	T	NA	ZNF93	Replacement	E338D	+	15.059	0.602	0.472	SL119751; SL119752	NA

YALS	chr19	19933971	NA	A	T	NA	ZNF93	Replacement	K339I	+	15.728	0.591	0.472	SL119751; SL119752	NA
YALS	chr19	21808503	NA	A	G	NA	ZNF43	Replacement	C512R	-	24.094	0.933	0.709	SL119751; SL119752	NA
YALS	chr19	39422251	NA	T	A	NA	PLEKHG2	Replacement	L547Q	+	17.402	0.969	0.945	SL119751; SL119752	NA
YALS	chr19	47874750	NA	T	C	NA	SULT2A1	Replacement	S218G	-	19.409	0.748	0.709	SL119751; SL119752	NA
YALS	chr19	47874755	NA	T	C	NA	SULT2A1	Replacement	K216R	-	22.087	0.795	0.709	SL119751; SL119752	NA
YALS	chr19	49625977	NA	T	G	NA	PRR12	Non-Coding		+	13.051	0.098	0	SL119751; SL119752	NA
YALS	chr19	49959773	rs201740168	C	A	NA	SIGLEC11	Replacement	A265S	-	12.717	0.354	0.236	SL119751; SL119752	NA
YALS	chr19	50422933	NA	G	C	NA	SPIB	Replacement	E79Q	+	12.717	0.016	0	SL119751; SL119752	NA
YALS	chr19	50422935	NA	A	C	NA	SPIB	Replacement	E79D	+	11.713	0.016	0	SL119751; SL119752	NA
YALS	chr19	53914590	NA	A	C	NA	CACNG7	Non-Coding		+	14.724	0.988	0.945	SL119751; SL119752	NA
YALS	chr19	54927666	NA	T	G	NA	NLRP7	Replacement	S946R	-	22.087	0.055	0	SL119751; SL119752	NA
YALS	chr20	45078818	NA	C	T	NA	STK4	Non-Coding		+	10.374	0.382	0.236	SL119751; SL119752	NA
YALS	chr21	26841068	NA	G	T	NA	ADAMTS1	Replacement	N436K	-	12.047	0.917	0.709	SL119751; SL119752	NA
YALS	chr21	26841073	NA	A	C	NA	ADAMTS1	Replacement	S435A	-	16.063	0.992	0.945	SL119751; SL119752	NA
YALS	chr21	26841076	NA	G	C	NA	ADAMTS1	Replacement	L434V	-	17.067	0.996	0.945	SL119751; SL119752	NA
YALS	chr21	26841089	NA	C	A	NA	ADAMTS1	Replacement	M429I	-	21.752	0.984	0.945	SL119751; SL119752	NA
YALS	chr21	26841093	NA	T	A	NA	ADAMTS1	Replacement	H428L	-	20.079	0.972	0.945	SL119751; SL119752	NA
YALS	chr21	26841094	NA	G	C	NA	ADAMTS1	Replacement	H428D	-	19.075	0.969	0.945	SL119751; SL119752	NA
YALS	chr21	26841097	NA	A	G	NA	ADAMTS1	Replacement	S427P	-	17.736	0.925	0.709	SL119751; SL119752	NA
YALS	chrX	13816977	rs3747420	T	C	0.000455789	GPM6B	Non-Coding		-	10.039	0.992	0.945	SL119752	SL119751
YALS	chrX	24550436	rs1055186	A	C	0.01	PDK3	Non-Coding		+	12.047	0.063	0	SL119751	SL119752
YALS	chrX	40680755	NA	T	C	NA	MED14	Non-Coding		-	15.394	0.984	0.945	SL119751; SL119752	NA
OALS	chr1	1722769	NA	A	G	NA	CDK11A	Replacement	L17P	-	21.752	1	0.945	SL119753; SL119754	NA
OALS	chr1	166939574	rs41269698	G	A	0.007587859	ILDR2	Non-Coding		-	15.394	0.996	0.945	SL119753; SL119754	NA

OALS	chr1	167126528	rs6668826	G	A	0.000798722	DUSP27	Replacement	R466H	+	12.047	0.083	0	SL119753	SL119754
OALS	chr2	110543432	NA	T	C	NA	RGPD8	Replacement	E798G	-	18.071	0.976	0.945	SL119753; SL119754	NA
OALS	chr3	10505548	rs111408739	T	C	NA	ATP2B2	Non-Coding		-	16.732	0.909	0.709	SL119753; SL119754	NA
OALS	chr3	113396696	rs77501585	C	G	0.009984026	CFAP44	Replacement	G534A	-	29.783	0.909	0.709	SL119753; SL119754	NA
OALS	chr5	1240642	rs7447815	C	G	0.000599042	SLC6A18	Replacement	Y319*	+	26.102	0.063	0	SL119753; SL119754	NA
OALS	chr5	140552044	rs202193903; rs368142622	C	G	NA	SRA1	Non-Coding		-	19.744	0.382	0.236	SL119753	SL119754
OALS	chr5	160259745	rs139134014	C	T	0.004592652	CCNJL	Replacement	V151I	-	13.72	0.417	0.236	SL119753; SL119754	NA
OALS	chr6	149572573	NA	G	A	NA	GINM1	Replacement	V83I	+	22.087	1	0.945	SL119753; SL119754	NA
OALS	chr6	160707780	rs143079629	G	A	0.001797125	PLG	Replacement	R89K	+	15.059	0.823	0.709	SL119753; SL119754	NA
OALS	chr7	99434936	rs150504114	G	A	0.007188498	PTCD1	Replacement	R103C	-	15.728	0.228	0	SL119753; SL119754	NA
OALS	chr7	100107269	rs146348021	C	T	0.001397764	TAF6	Replacement	G671S	-	14.39	0.161	0	SL119753; SL119754	NA
OALS	chr7	100127910	rs201045178	A	C	0.000399361	MBLAC1	Replacement	Q172P	+	21.752	0.413	0.236	SL119753; SL119754	NA
OALS	chr8	38393827	NA	A	T	NA	LETM2	Non-Coding		+	14.055	0.508	0.472	SL119753; SL119754	NA
OALS	chr9	83880361	rs200952027	C	T	0.001500001	KIF27	Replacement	R860Q	-	22.756	0.953	0.945	SL119754	SL119753
OALS	chr10	45773308	NA	G	C	NA	FAM21C	Replacement	G674R	+	23.091	0.98	0.945	SL119753; SL119754	NA
OALS	chr10	45826107	NA	G	C	NA	AGAP6	Replacement	D515E	-	16.398	0.807	0.709	SL119753; SL119754	NA
OALS	chr10	50129923	NA	C	A	NA	FAM21A	Replacement	P1177T	+	20.748	0.921	0.709	SL119753; SL119754	NA
OALS	chr10	73763876	rs35528438	A	G	0.00798722	SEC24C	Replacement	I374V	+	21.752	0.996	0.945	SL119753; SL119754	NA
OALS	chr11	67253595	rs34925153	A	G	0.006389776	KDM2A	Silent	P1025P	+	14.39	0.988	0.945	SL119753; SL119754	NA
OALS	chr12	133080312	rs116668890	G	A	NA	ZNF140	Non-Coding		+	11.043	0	0	SL119753; SL119754	NA
OALS	chr19	7613156	NA	T	C	NA	CAMSAP3	Replacement	F915S	+	23.091	0.98	0.945	SL119753; SL119754	NA
OALS	chr19	7613161	NA	A	C	NA	CAMSAP3	Replacement	K917Q	+	23.425	0.992	0.945	SL119753; SL119754	NA
OALS	chr19	32973495	rs531017428	T	C	NA	C19orf40	Replacement	L59P	+	21.417	0.98	0.945	SL119753; SL119754	NA
OALS	chr19	49782063	NA	T	C	NA	AP2A1	Replacement	S85P	+	25.768	0.996	0.945	SL119753; SL119754	NA

OALS	chr21	36387733	NA	G	A	NA	CHAF1B	Non-Coding		+	18.74	0.988	0.945	SL119753; SL119754	NA
OALS	chr22	15528700	NA	T	A	NA	OR11H1	Replacement	M181K	+	23.76	0.917	0.709	SL119753; SL119754	NA
OALS	chr22	32194461	rs3986037	C	T	0.003194888	RFPL2	Replacement	R50H	-	12.382	0.087	0	SL119753; SL119754	NA
OALS	chr22	37655664	NA	C	G	NA	SH3BP1	Replacement	L696V	+	15.394	0.984	0.945	SL119753; SL119754	NA
OALS	chrX	88753806	rs5984611	G	A	0.00375	CPXCR1	Replacement	R131H	+	15.059	0	0	SL119753	SL119754

Supplementary Table 3: fly orthologs of human genes

patient type	human symbol	fly symbol	Fly Gene ID	FlyBaseID	DIOPT Score	Weighted Score	Rank	Best Score	Best Score Reverse	Prediction Derived From
YALS	ST7L	CG3634	40301	FBgn0037026	8	7.706	moderate	Yes	No	Compara, eggNOG, Inparanoid, OrthoDB, orthoMCL, Panther, Phylome, TreeFam
YALS	CTRC	CG32260	317943	FBgn0052260	2	1.933	high	Yes	Yes	Compara, RoundUp
YALS	GON4L	mute	2768848	FBgn0085444	6	5.908	high	Yes	Yes	Compara, Inparanoid, Panther, Phylome, RoundUp, TreeFam
YALS	SERBP1	vig2	43016	FBgn0046214	8	7.708	high	Yes	Yes	Compara, eggNOG, Inparanoid, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	SERBP1	vig	34885	FBgn0024183	8	7.708	high	Yes	Yes	Compara, eggNOG, Inparanoid, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	GFI1	sens-2	33957	FBgn0051632	6	5.803	moderate	Yes	No	Compara, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	ANKRD39	CG44001	326173	FBgn0264743	8	7.813	high	Yes	Yes	Compara, Homologene, OMA, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	PMS1	Pms2	36705	FBgn0011660	2	1.903	moderate	Yes	No	eggNOG, RoundUp
YALS	RGPD4	Nup358	43041	FBgn0039302	8	7.706	moderate	Yes	No	Compara, eggNOG, Inparanoid, OrthoDB, orthoMCL, Panther, Phylome, TreeFam
YALS	GNAI2	Galphai	38765	FBgn0001104	10	9.719	moderate	Yes	No	Compara, eggNOG, Inparanoid, OMA, OrthoDB, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	ACAP2	CenB1A	42735	FBgn0039056	11	10.719	high	Yes	Yes	Compara, eggNOG, Homologene, Inparanoid, OMA, OrthoDB, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	MYH15	Mhc	35007	FBgn0264695	8	7.706	moderate	Yes	No	Compara, eggNOG, Inparanoid, OrthoDB, orthoMCL, Panther, Phylome, TreeFam
YALS	FRAS1	kon	35104	FBgn0032683	3	2.74	moderate	Yes	No	Compara, eggNOG, Phylome
YALS	PDS5A	pds5	36286	FBgn0260012	9	8.716	moderate	Yes	No	Compara, eggNOG, Inparanoid, OMA, OrthoDB, orthoMCL, Panther, Phylome, TreeFam
YALS	GRSF1	glo	41431	FBgn0259139	4	3.941	moderate	Yes	No	Compara, OrthoDB, Panther, Phylome
YALS	OTUD4	otu	31789	FBgn0003023	4	3.943	moderate	Yes	No	Compara, Panther, Phylome, RoundUp
YALS	RAD50	rad50	37564	FBgn0034728	9	8.818	high	Yes	Yes	Compara, eggNOG, Homologene, Inparanoid, OMA, Panther, Phylome, RoundUp, TreeFam
YALS	HK3	Hex-A	45875	FBgn0001186	7	6.744	moderate	Yes	No	Compara, eggNOG, OrthoDB, orthoMCL, Panther, Phylome, RoundUp
YALS	CDKL3	CG7236	33798	FBgn0031730	2	1.901	moderate	Yes	No	eggNOG, OrthoDB
YALS	PPARD	Hr96	42993	FBgn0015240	3	2.96	moderate	Yes	No	eggNOG, Panther, TreeFam
YALS	PPARD	Eip75B	39999	FBgn0000568	3	2.95	moderate	Yes	No	eggNOG, Isobase, Panther
YALS	KIAA1244	CG5937	31537	FBgn0029834	10	9.568	high	Yes	Yes	Compara, eggNOG, Homologene, Inparanoid, Isobase, OMA, orthoMCL, Phylome, RoundUp, TreeFam
YALS	ARID1B	osa	42130	FBgn0261885	8	7.749	high	Yes	Yes	Compara, eggNOG, Inparanoid, OrthoDB, orthoMCL, Panther, Phylome, RoundUp
YALS	MYO6	jar	42889	FBgn0011225	10	9.768	high	Yes	Yes	Compara, eggNOG, Homologene, Inparanoid, Isobase, OMA, Panther, Phylome, RoundUp, TreeFam

YALS	UHRF1BP1	CG34126	318872	FBgn0083962	9	8.709	high	Yes	Yes	Compara, eggNOG, Inparanoid, OrthoDB, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	DNAH8	CG9492	41171	FBgn0037726	7	6.705	moderate	Yes	No	Compara, eggNOG, Inparanoid, orthoMCL, Panther, Phylome, TreeFam
YALS	DBF4	chif	34974	FBgn0000307	4	3.773	moderate	Yes	No	eggNOG, Phylome, RoundUp, TreeFam
YALS	HIPK2	Hipk	38070	FBgn0035142	9	8.729	high	Yes	Yes	eggNOG, Inparanoid, Isobase, OrthoDB, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	PEX1	Pex1	45460	FBgn0013563	8	7.708	high	Yes	Yes	Compara, eggNOG, Inparanoid, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	TMEM67	CG15923	42443	FBgn0038814	10	9.718	high	Yes	Yes	Compara, eggNOG, Homologene, Inparanoid, OMA, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	KAT6A	enok	37859	FBgn0034975	7	6.608	high	Yes	Yes	Compara, eggNOG, Inparanoid, orthoMCL, Phylome, RoundUp, TreeFam
YALS	EGR3	sr	42162	FBgn0003499	3	2.893	moderate	Yes	No	Compara, RoundUp, TreeFam
YALS	SVEP1	uif	33983	FBgn0031879	2	1.905	high	Yes	Yes	eggNOG, Inparanoid
YALS	AGTPBP1	CG31019	318558	FBgn0051019	2	1.903	moderate	Yes	No	eggNOG, RoundUp
YALS	AGTPBP1	NnaD	32329	FBgn0265726	2	1.8	moderate	Yes	No	eggNOG, orthoMCL
YALS	AGAP10	CenG1A	34803	FBgn0028509	2	1.81	moderate	Yes	No	orthoMCL, Phylome
YALS	ABCC2	MRP	34686	FBgn0032456	6	5.601	moderate	Yes	No	Compara, eggNOG, OrthoDB, orthoMCL, Phylome, TreeFam
YALS	ACBD5	CG8814	33492	FBgn0031478	8	7.706	high	Yes	Yes	Compara, eggNOG, Inparanoid, OrthoDB, orthoMCL, Panther, Phylome, TreeFam
YALS	TUBGCP2	Grip84	32946	FBgn0026430	12	11.669	high	Yes	Yes	Compara, eggNOG, Homologene, Inparanoid, Isobase, OMA, OrthoDB, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	NR1H3	EcR	35540	FBgn0000546	6	5.783	high	Yes	Yes	Compara, eggNOG, Isobase, orthoMCL, Panther, RoundUp
YALS	DLG2	dlg1	32083	FBgn0001624	8	7.609	moderate	Yes	No	Compara, eggNOG, Inparanoid, OrthoDB, orthoMCL, Phylome, RoundUp, TreeFam
YALS	BSCL2	Seipin	31245	FBgn0040336	8	7.708	high	Yes	Yes	Compara, eggNOG, Inparanoid, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	RBMS2	shep	38605	FBgn0052423	7	6.709	high	Yes	Yes	Compara, eggNOG, Inparanoid, OrthoDB, Phylome, RoundUp, TreeFam
YALS	ACACB	ACC	35761	FBgn0033246	8	7.708	moderate	Yes	No	Compara, eggNOG, Inparanoid, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	UTP20	CG4554	37570	FBgn0034734	10	9.718	high	Yes	Yes	Compara, eggNOG, Homologene, Inparanoid, OMA, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	FGF14	bnl	42356	FBgn0014135	3	2.96	moderate	Yes	No	eggNOG, Panther, TreeFam
YALS	STYX	Mkp	4379907	FBgn0083992	4	3.804	moderate	Yes	No	eggNOG, OrthoDB, orthoMCL, RoundUp
YALS	ARF6	Arf51F	36699	FBgn0013750	12	11.669	high	Yes	Yes	Compara, eggNOG, Homologene, Inparanoid, Isobase, OMA, OrthoDB, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	ABHD2	Hydr2	33532	FBgn0014906	11	10.719	high	Yes	Yes	Compara, eggNOG, Homologene, Inparanoid, OMA, OrthoDB, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	FAM98B	CG5913	43132	FBgn0039385	8	7.909	moderate	Yes	No	Compara, eggNOG, Inparanoid, OMA, OrthoDB, Panther, RoundUp, TreeFam
YALS	NUBP1	CG17904	35000	FBgn0032597	10	9.768	high	Yes	Yes	Compara, Homologene, Inparanoid, Isobase, OMA, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	ADAMTS18	CG4096	31490	FBgn0029791	4	3.931	moderate	Yes	No	Compara, eggNOG, OrthoDB, Panther

YALS	TVP23B	CG5021	39025	FBgn0035944	10	9.719	high	Yes	Yes	Compara, eggNOG, Inparanoid, OMA, OrthoDB, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	PPP1R9B	Spn	46194	FBgn0010905	5	4.948	moderate	Yes	No	Compara, Inparanoid, Panther, Phylome, RoundUp
YALS	FXR2	Fmr1	37528	FBgn0028734	8	7.706	moderate	Yes	No	Compara, eggNOG, Inparanoid, OrthoDB, orthoMCL, Panther, Phylome, TreeFam
YALS	SLC13A5	Indy	40049	FBgn0036816	9	8.759	high	Yes	Yes	Compara, eggNOG, Inparanoid, OMA, OrthoDB, orthoMCL, Panther, Phylome, RoundUp
YALS	RGS9	CG42450	32874	FBgn0259927	5	4.77	high	Yes	Yes	eggNOG, orthoMCL, Panther, Phylome, TreeFam
YALS	PLXDC1	l(1)G0289	31964	FBgn0028331	7	6.808	moderate	Yes	No	Compara, eggNOG, Inparanoid, Panther, Phylome, RoundUp, TreeFam
YALS	CHMP1B	Chmp1	40036	FBgn0036805	9	8.819	high	Yes	Yes	Compara, eggNOG, Inparanoid, OMA, OrthoDB, Panther, Phylome, RoundUp, TreeFam
YALS	ATP9B	CG31729	34736	FBgn0051729	10	9.665	high	Yes	Yes	Compara, eggNOG, Homologene, Inparanoid, Isobase, OMA, orthoMCL, Panther, Phylome, TreeFam
YALS	CACNG7	stg1	318064	FBgn0064123	4	3.86	high	Yes	Yes	eggNOG, orthoMCL, Panther, TreeFam
YALS	CELF5	bru-3	39527	FBgn0264001	8	7.606	high	Yes	Yes	Compara, eggNOG, Homologene, Inparanoid, OrthoDB, orthoMCL, Phylome, TreeFam
YALS	SLC35E1	CG14621	33128	FBgn0031183	10	9.656	high	Yes	Yes	Compara, eggNOG, Homologene, Inparanoid, Isobase, OrthoDB, orthoMCL, Panther, Phylome, TreeFam
YALS	PLEKHG2	GEFmeso	37134	FBgn0050115	7	6.844	high	Yes	Yes	Compara, eggNOG, Isobase, OrthoDB, Panther, RoundUp, TreeFam
YALS	SULT2A1	St2	41098	FBgn0037665	6	5.749	moderate	Yes	No	Compara, eggNOG, Inparanoid, OrthoDB, Phylome, RoundUp
YALS	STK4	hpo	37247	FBgn0261456	7	6.808	moderate	Yes	No	Compara, Inparanoid, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	ADAMTS1	AdamTS-A	41887	FBgn0038341	7	6.744	moderate	Yes	No	Compara, eggNOG, OrthoDB, orthoMCL, Panther, Phylome, RoundUp
YALS	GPM6B	M6	40383	FBgn0037092	5	4.875	moderate	Yes	No	eggNOG, Inparanoid, Panther, Phylome, TreeFam
YALS	PDK3	Pdk	35970	FBgn0017558	12	11.669	high	Yes	Yes	Compara, eggNOG, Homologene, Inparanoid, Isobase, OMA, OrthoDB, orthoMCL, Panther, Phylome, RoundUp, TreeFam
YALS	MED14	MED14	38073	FBgn0035145	11	10.719	high	Yes	Yes	Compara, eggNOG, Homologene, Inparanoid, OMA, OrthoDB, orthoMCL, Panther, Phylome, RoundUp, TreeFam
OALS	DUSP27	CG7378	32888	FBgn0030976	3	2.831	moderate	Yes	No	Compara, eggNOG, OrthoDB
OALS	CDK11A	Pitslre	40292	FBgn0016696	7	6.919	moderate	Yes	No	eggNOG, Inparanoid, Isobase, OrthoDB, Panther, RoundUp, TreeFam
OALS	RGPD8	Nup358	43041	FBgn0039302	8	7.706	moderate	Yes	No	Compara, eggNOG, Inparanoid, OrthoDB, orthoMCL, Panther, Phylome, TreeFam
OALS	CFAP44	CG34124	4379887	FBgn0083960	9	8.713	high	Yes	Yes	Compara, eggNOG, Homologene, OMA, orthoMCL, Panther, Phylome, RoundUp, TreeFam
OALS	ATP2B2	PMCA	43787	FBgn0259214	8	7.708	moderate	Yes	No	Compara, eggNOG, Inparanoid, orthoMCL, Panther, Phylome, RoundUp, TreeFam
OALS	SLC6A18	CG43066	37129	FBgn0262476	7	6.774	high	Yes	Yes	eggNOG, OrthoDB, orthoMCL, Panther, Phylome, RoundUp, TreeFam
OALS	CCNJL	CycJ	38428	FBgn0010317	4	3.795	moderate	Yes	No	Compara, eggNOG, Inparanoid, TreeFam
OALS	MBLAC1	CG9117	33846	FBgn0031766	10	9.709	high	Yes	Yes	Compara, eggNOG, Homologene, Inparanoid, OrthoDB, orthoMCL, Panther, Phylome, RoundUp, TreeFam
OALS	PTCD1	CG4611	38601	FBgn0035591	7	6.679	high	Yes	Yes	eggNOG, Inparanoid, OrthoDB, orthoMCL, Phylome, RoundUp, TreeFam
OALS	TAF6	Taf6	40134	FBgn0010417	9	8.708	high	Yes	Yes	Compara, eggNOG, Homologene, Inparanoid, orthoMCL, Panther, Phylome, RoundUp, TreeFam
OALS	LETM2	Letm1	37912	FBgn0019886	6	5.803	moderate	Yes	No	Compara, eggNOG, Panther, Phylome, RoundUp, TreeFam

OALS	KIF27	cos	35653	FBgn0000352	3	3.02	moderate	Yes	No	OMA, Panther, Phylome
OALS	FAM21C	FAM21	37331	FBgn0034529	6	5.728	high	Yes	Yes	eggNOG, Inparanoid, OMA, orthoMCL, Phylome, RoundUp
OALS	SEC24C	Sec24CD	33409	FBgn0262126	9	8.758	high	Yes	Yes	Compara, Homologene, Inparanoid, Isobase, orthoMCL, Panther, Phylome, RoundUp, TreeFam
OALS	FAM21A	FAM21	37331	FBgn0034529	6	5.728	high	Yes	Yes	eggNOG, Inparanoid, OMA, orthoMCL, Phylome, RoundUp
OALS	AGAP6	CenG1A	34803	FBgn0028509	4	3.67	moderate	Yes	No	eggNOG, orthoMCL, Phylome, TreeFam
OALS	KDM2A	Kdm2	41090	FBgn0037659	10	9.768	high	Yes	Yes	Compara, eggNOG, Homologene, Inparanoid, Isobase, OMA, Panther, Phylome, RoundUp, TreeFam
OALS	C19orf40	Ercc1	36654	FBgn0028434	2	1.81	moderate	Yes	No	eggNOG, Phylome
OALS	AP2A1	AP-2alpha	33211	FBgn0264855	10	9.719	moderate	Yes	No	Compara, eggNOG, Inparanoid, OMA, OrthoDB, orthoMCL, Panther, Phylome, RoundUp, TreeFam
OALS	CAMSAP3	Patronin	36978	FBgn0263197	7	6.703	moderate	Yes	No	Compara, eggNOG, orthoMCL, Panther, Phylome, RoundUp, TreeFam
OALS	CHAF1B	Caf1-105	36107	FBgn0033526	9	8.708	high	Yes	Yes	Compara, eggNOG, Homologene, Inparanoid, orthoMCL, Panther, Phylome, RoundUp, TreeFam
OALS	SH3BP1	RhoGAP92B	42371	FBgn0038747	6	5.885	moderate	Yes	No	eggNOG, Inparanoid, OMA, Panther, Phylome, TreeFam

Supplementary Table 4: RNAi lines used in this study

patient type	gene symbol	fly ortholog	stock source	Reagent ID
YALS	ABCC2	MRP	TRiP Short Hairpin	38316
YALS	ABCC2	MRP	VDRC KK	105419
YALS	ACACB	ACC	TRiP Short Hairpin	34885
YALS	ADAMTS18	CG4096	TRiP Short Hairpin	44522
YALS	AGTPBP1	NnaD	TRiP Short Hairpin	33549
YALS	AGTPBP1	NnaD	TRiP Short Hairpin	44036
YALS	ARF6	Arf51F	TRiP Short Hairpin	51417
YALS	ARF6	Arf51F	TRiP Short Hairpin	27261
YALS	ARID1B	osa	TRiP Short Hairpin	35447
YALS	ARID1B	osa	TRiP Short Hairpin	31266
YALS	ATP9B	CG31729	TRiP Short Hairpin	51819
YALS	BSCL2	Seipin	TRiP Short Hairpin	37501
YALS	CELF5	bru-3	TRiP Short Hairpin	50734
YALS	CELF5	bru-3	TRiP Short Hairpin	43318
YALS	CELF5	bru-3	VDRC KK	109946
YALS	CHMP1B	Chmp1	TRiP Short Hairpin	33928
YALS	CHMP1B	Chmp1	TRiP Short Hairpin	28906
YALS	DBF4	chif	TRiP Short Hairpin	33365
YALS	DLG2	dlg1	TRiP Short Hairpin	31521
YALS	DLG2	dlg1	TRiP Short Hairpin	35286
YALS	DLG2	dlg1	VDRC KK	109274
YALS	DNAH8	CG9492	TRiP Short Hairpin	51725
YALS	EGR3	sr	TRiP Short Hairpin	27701
YALS	FAM98B	CG5913	TRiP Short Hairpin	53965
YALS	FAM98B	CG5913	VDRC GD	40336
YALS	FGF14	bnl	TRiP Short Hairpin	34572
YALS	FXR2	Fmr1	TRiP Short Hairpin	35200
YALS	FXR2	Fmr1	TRiP Short Hairpin	27484
YALS	GFI1	sens-2	TRiP Short Hairpin	34984
YALS	GPM6B	M6	TRiP Short Hairpin	37503
YALS	GPM6B	M6	TRiP Short Hairpin	54032
YALS	GRSF1	glo	TRiP Short Hairpin	36066
YALS	GRSF1	glo	TRiP Short Hairpin	33668
YALS	HIPK2	Hipk	TRiP Short Hairpin	35363
YALS	HIPK2	Hipk	TRiP Short Hairpin	56458
YALS	HIPK2	Hipk	TRiP Short Hairpin	20760
YALS	HIPK2	Hipk	VDRC KK	108254
YALS	HK3	Hex-A	TRiP Short Hairpin	35155
YALS	HK3	Hex-A	VDRC KK	103536
YALS	KAT6A	enok	TRiP Short Hairpin	42941
YALS	KAT6A	enok	TRiP Short Hairpin	29518
YALS	MED14	MED14	TRiP Short Hairpin	34575
YALS	MYH15	Mhc	TRiP Short Hairpin	35729
YALS	MYH15	Mhc	TRiP Short Hairpin	26299
YALS	MYH15	Mhc	VDRC KK	105355

YALS	NR1H3	EcR	TRiP Short Hairpin	58286
YALS	NR1H3	EcR	TRiP Short Hairpin	29374
YALS	OTUD4	otu	TRiP Short Hairpin	34065
YALS	PDK3	Pdk	TRiP Short Hairpin	28635
YALS	PDK3	Pdk	TRiP Short Hairpin	35142
YALS	PDK3	Pdk	VDRC KK	106641
YALS	PDS5A	pds5	TRiP Short Hairpin	35632
YALS	PEX1	Pex1	TRiP Short Hairpin	51497
YALS	PEX1	Pex1	TRiP Short Hairpin	28979
YALS	PLEKHG2	GEFmeso	TRiP Short Hairpin	42545
YALS	PLEKHG2	GEFmeso	VDRC GD	39952
YALS	PLXDC1	l(1)G0289	TRiP Short Hairpin	33690
YALS	PLXDC1	l(1)G0289	TRiP Short Hairpin	32910
YALS	PMS1	Pms2	TRiP Short Hairpin	55614
YALS	PPARD	Eip75B	TRiP Short Hairpin	43231
YALS	PPARD	Eip75B	TRiP Short Hairpin	26717
YALS	PPARD	Eip75B	VDRC KK	108399
YALS	RBMS2	shep	TRiP Short Hairpin	43545
YALS	RBMS2	shep	TRiP Short Hairpin	38218
YALS	RGPD4	Nup358	TRiP Short Hairpin	34967
YALS	RGPD4	Nup358	TRiP Short Hairpin	33003
YALS	SERBP1	vig	TRiP Short Hairpin	35183
YALS	SERBP1	vig	TRiP Short Hairpin	35184
YALS	STK4	hpo	TRiP Short Hairpin	35176
YALS	STK4	hpo	TRiP Short Hairpin	27661
YALS	SVEP1	uif	TRiP Short Hairpin	38354
YALS	SVEP1	uif	TRiP Short Hairpin	38365
YALS	SVEP1	uif	VDRC KK	101153
YALS	TMEM67	CG15923	TRiP Short Hairpin	53289
YALS	TUBGCP2	Grip84	TRiP Short Hairpin	33548
YALS	UTP20	CG4554	TRiP Short Hairpin	53270
YALS	UTP20	CG4554	VDRC GD	103706
YALS	UTP20	CG4554	VDRC KK	103706
YALS	UTP20	CG4554	VDRC GD	21620
OALS	CAMSAP3	Patronin	TRiP Short Hairpin	36659
OALS	CCNJL	CycJ	TRiP Short Hairpin	37521
OALS	CDK11A	Pitslre	TRiP Short Hairpin	35157
OALS	CDK11A	Pitslre	VDRC KK	107303
OALS	KDM2A	Kdm2	TRiP Short Hairpin	33699
OALS	KDM2A	Kdm2	TRiP Short Hairpin	31360
OALS	KDM2A	Kdm2	VDRC KK	109295
OALS	KIF27	cos	TRiP Short Hairpin	44472
OALS	KIF27	cos	VDRC KK	108914
OALS	LETM2	Letm1	TRiP Short Hairpin	37502
OALS	SH3BP1	RhoGAP92B	TRiP Short Hairpin	33391

Supplementary Table 5: Summary of clinical information on ALS cases and controls included in targeted resequencing

		ALS	Unaffected control
Gender	**Female**	120 (38.7%)	141 (53.0%)
	Male	190 (61.3%)	125 (47.0%)
Average of age-at-onset		56.8	N/A
Average of age-at-recruitment		N/A	75.4

Supplementary Table 6: Gene-based analysis of rare variants for targeted resequencing dataset

Group	Genes	Targeted Resequencing		
		variant number	SKAT p-value	SKAT-O p-value
Suppressors	CELF5	2	0.8699	0.8699
	DLG2	10	**0.0418**	0.09971
	FAM98B	3	0.48711	0.62493
	HK3	19	0.885	0.97769
	PDK3	1	0.86912	0.86912
	KDM2A	5	0.69728	0.87765
	KIF27	19	0.64552	0.88058
Enhancers	ABCC2	12	0.65511	0.60758
	MYH15	16	**0.0195**	**0.03697**
	PLEKHG2	12	0.7032	0.87329
	PPARD	3	0.3854	0.62406
	SVEP1	39	0.82087	0.89962
	UTP20	27	0.48725	0.7171
	CDK11A	15	0.8892	0.41197

Supplementary Table 7: Gene-based analysis of rare variants for WGS dataset

Group	Genes	WGS		
		variant number	SKAT p-value	SKAT-O p-value
Suppressors	CELF5	0	-	-
	DLG2	2	0.30207097	0.30207097
	FAM98B	6	0.87353369	0.31996341
	HK3	19	0.43986132	0.67842778
	PDK3	1	0.60350993	0.60350993
	KDM2A	1	0.59609034	0.59609034
	KIF27	7	0.33940203	0.42108945
Enhancers	ABCC2	16	0.96040603	0.34382896
	MYH15	6	0.01232624	0.01707692
	PLEKHG2	12	0.8085755	0.36646429
	PPARD	0	-	-
	SVEP1	23	0.71040196	0.74890023
	UTP20	18	0.72497056	0.96158609
	CDK11A	3	0.23908214	0.23908214

Supplementary Table 8: List of variants with a statistical significance found in targeted resequencing and WGS

Gene	Chr	Position	ID	ref	alt	aa change	Function	Cadd	Targeted resequencing				Replication (WGS)			
									Hetero-zygotes in ALS	Homo-zygotes in ALS	Hetero-zygotes in Control	Homo-zygotes in Control	Hetero-zygotes in ALS	Homo-zygotes in ALS	Hetero-zygotes in Control	Homo-zygotes in Control
MYH15	chr3	108394107	rs56118396	C	T	R1748Q	nonSynonymous	2	1		7					
MYH15	chr3	108398706		G	G	E1708D	nonSynonymous	3.5	1							
MYH15	chr3	108398765	rs76478083	G	A	R1689C	nonSynonymous	24	5	1	2					
MYH15	chr3	108398829		G	+A	M1667NA	nonSynonymous	NA		1						
MYH15	chr3	108410709	rs377336538	C	G	A1477P	nonSynonymous	22.5	1							
MYH15	chr3	108410805	rs368421301	C	T	G1445R	nonSynonymous	23.5			1					
MYH15	chr3	108410826	rs368538771	C	T	G1438R	nonSynonymous	26	1							
MYH15	chr3	108410882		C	C	A1419G	nonSynonymous	21	1							
MYH15	chr3	108410906		A	-A	E1411NA	nonSynonymous	NA			1					
MYH15	chr3	108428833	rs61744539	G	A	R1141*	nonSynonymous	37	4		2		1			
MYH15	chr3	108430899	rs534599773	A	G	M1102T	nonSynonymous	13.2							1	
MYH15	chr3	108437581	rs148843085	T	G	Q1085P	nonSynonymous	23.4					1			
MYH15	chr3	108441178	rs368092347	T	G	E933A	nonSynonymous	14.6					1			
MYH15	chr3	108444779		T	C	E859G	nonSynonymous	15.1					1			
MYH15	chr3	108444826		G	A	R843S	nonSynonymous	21.5	1							
MYH15	chr3	108454106	rs199678295	C	T	A787T	nonSynonymous	23.5			1					
MYH15	chr3	108476492		A	C	M400V	nonSynonymous	11.5			1					
MYH15	chr3	108499455	rs202126707	G	A	T195I	nonSynonymous	24.5	1		1				2	
MYH15	chr3	108501734		G	T	R126H	nonSynonymous	25.5			1					
MYH15	chr3	108501756	rs200749942	C	T	V119M	nonSynonymous	23.5			2					
DLG2	chr11	83462065		C	A	Q815*	StopGain	40			1					
DLG2	chr11	83786749		C	A	A484V	nonSynonymous	32			1					
DLG2	chr11	83930478	rs185568966	G	A	P344L	nonSynonymous	34	1		1					
DLG2	chr11	83965464	rs373325643	T	C	N249S	nonSynonymous	23	1							
DLG2	chr11	84163487		G	T	E95K	nonSynonymous	26.5	1							
DLG2	chr11	84316847	rs373138134	T	G	H100P	nonSynonymous	16.5			1					
DLG2	chr11	84316958	rs557645711	C	G	R63T	nonSynonymous	10.5			1					
DLG2	chr11	84317123		A	C	Q8R	nonSynonymous	11	1							
DLG2	chr11	85111692		C	T	A109D	nonSynonymous	18			1					
DLG2	chr11	85285273	rs373336609	C	T	E45K	nonSynonymous	25			1					

Supplementary Table 9: Sample information used in this study

SampleID	Age	Sex	PMI	Race	Source	Neuropathology Report
NCB01	11	M	19	Black or African-American	Maryland	Individual went into the river and according to witnesses, lost his footing and got swept away by the rapids and had trouble with the undertow. 911 was called. When rescue team responded, individual was found lodged under a rock. Estimated that he was submerged in the water for 25-30 minutes. He could not swim. Transported to hospital, where he was unable to be resuscitated.
NCB02	5	F	24	Black or African-American	Maryland	Donor was a 5-year black girl with history chronic kidney disease due to congenital nephrotic syndrome, renal osteodystophy, secondary hyperparathyroidism with partial parathyroidectomy, anemia, liver failure, hepatosplenomegaly, patent ductus arteriosys with transcatheter closure and chronic malnutrition. NEUROPATHOLOGY FINAL DIAGNOSIS Subdural hematoma, organizing. Histiocytic infiltrates in dura and choroid plexus, similar to that in the systemic organs.
NCB03	19	M	5	White	Maryland	Donor was a 19-year-old male who drowned after a car accident. He had no positive medical history and was not currently taking any medications.
NCB04	10	F	10	White	Maryland	Donor was a 9-year-old girl with a history of asthma and GERD. Donor routinely had a treatment at 2200 hours every day. Donor complained about not feeling well during her treatment and 15 mins. after the treatment donor stated her heart hurt.
NCB05	5	M	19	White	Maryland	Donor has a medical history of H1N1, seasonal allergies, and family history of arrhythmias. Pathologic Diagnoses concludes that due to the Anomalous Left Coronary Artery with Complications the subject had an anomalous left main coronary artery arising in right sinus with acute angle take-off and proximal course between aorta and pulmonary artery. Subject also had circumferential subendocardial necrosis with contraction bands and interstitial hemorrhage, left ventricle, consistent with reperfusion injury, diffuse pulmonary hemorrhage with focal acute bronchopneumonia, early hypoxic-ischemic encephalopathy, focal ischemic acute renal tubular necrosis, focal ischemic change of small intestines and status post extracorporeal membrane oxygenation. Medical Examiners Opinion: This 5-year-old white male died of Anomalous Left Coronary Artery with Complications. The manner of death is NATURAL. Toxicological analyses were negative for alcohol and drugs.
NCB06	19	F	5	White	Maryland	PATHOLOGIC DIAGNOSES: Evidence of compressional asphyxia is shown with petechial hemorrhages of the eyelids, conjunctivae, face, chin, right side of shoulder and upper chest. A small amount of coal dust in the external nares and oral cavity was noted as well. Additional injuries include a laceration with subgaleal hemorrhage of right frontal scalp, abrasions of forehead and philtrum, as well as abraded contusion of right leg. MEDICAL EXAMINERS OPINION: This 19-year-old white female, died of Compressional

						Asphyxia. Police investigation revealed that she was buried in the coal when a train with multiple cars carrying coal derailed and overturned. She was found unresponsive buried in the coal, with only feet and small portions of legs visible by the first responders. The manner of death is ACCIDENT. This case is associated with OCME Case: The deceased had been consuming alcoholic beverages prior to death.
NCB07	70	M	28	Black or African-American	Maryland	Clinical History: 2 years prior to death, Patient is a black who at age 68-year-old man was admitted with a sudden, severe headache that was very unusual for him. MRI scan revealed a 1.8 cm pituitary tumor (lesion) with mild elevation of the chiasm. There is some heterogeneity in the superior aspect of the mass. One year prior to death - Status post resection of pituitary tumor; CT Scan showed no evidence of any remaining tumor. FINAL NEUROPATHOLOGIC DIAGNOSIS: Brain with no significant pathology.
NCB08	70	M	12	Black or African-American	Maryland	Found unresponsive and cold. No history of smoking, drug use, falls or fractures. Drinks beer daily. Autopsy Findings: Subject died of hypertensive atherosclerotic cardiovascular disease. Diabetes mellitus and chronic renal disease are contributory factors in their death. The manner of death is NATURAL. Final NP Findings: moderately advanced cerebrovascular atherosclerosis, otherwise unremarkable adult brain.
NCB09	89	F	18.5	Unknown	Harvard	1. Cerebrovascular disease, with atherosclerosis, arteriosclerosis, arteriolosclerosis, myelin pallor in deep white matter of frontal lobe, semiacute infarct in the pons and remote infarct in cerebellar posterior lobe. 2. Neurofibrillary tangles, Braak stage I
NCB10	70	F	17.18	Unknown	Harvard	1.Neurofibrillary degeneration, Braak stage I, with non-neuritic neocortical amyloid plaques. 2. Atherosclerosis 3. Autolysis, mild
NCB11	70	F	21.33	Unknown	Harvard	1. Slight loss of neurons in the substantia nigra 2. Neurofibrillary degeneration, Braak stage I 3. Arteriosclerosis, with widespread mineralization of vessel walls in the globus pallidus
NCB12	85	M	29.05	Unknown	Harvard	1.Neurofibrillary degeneration, Braak stage I, with mild amyloid angiopathy 2. Atherosclerosis and arteriosclerosis 3. Autolysis, mild

Supplementary Table 10: Summary of GEO datasets related to Figures 4.2B and C

Sample	Disease	Gender	Age	Subgroup	Source
GSM1094863_DKFZ0888.CEL	Normal	F	Fetus	Normal	GSE44971
GSM1094864_DKFZ0889.CEL	Normal	F	Fetus	Normal	GSE44971
GSM1094865_DKFZ0890.CEL	Normal	M	Fetus	Normal	GSE44971
GSM1094866_DKFZ0891.CEL	Normal	M	Fetus	Normal	GSE44971
GSM1094867_DKFZ0892.CEL	Normal	F	Fetus	Normal	GSE44971
GSM1094868_DKFZ0894.CEL	Normal	M	24	Normal	GSE44971
GSM1094869_DKFZ0895.CEL	Normal	M	Adult	Normal	GSE44971
GSM1094870_DKFZ0896.CEL	Normal	M	Adult	Normal	GSE44971
GSM1094871_DKFZ0897.CEL	Normal	M	26	Normal	GSE44971
GSM1214936_NORMAL_E514_Cerebellum.CEL	Normal	Unknown	Children	Normal	GSE50161
GSM1214944_NORMAL_E605_Cerebellum.CEL	Normal	Unknown	Children	Normal	GSE50161
GSM175852.CEL	Normal	M	Unknown	Normal	GSE7307
GSM175853.CEL	Normal	M	Unknown	Normal	GSE7307
GSM175854.CEL	Normal	M	Unknown	Normal	GSE7307
GSM175907.CEL	Normal	Unknown	Unknown	Normal	GSE7307
GSM176030.CEL	Normal	M	Unknown	Normal	GSE7307
GSM176031.CEL	Normal	M	Unknown	Normal	GSE7307
GSM176048.CEL	Normal	M	Unknown	Normal	GSE7307
GSM176157.CEL	Normal	F	Unknown	Normal	GSE7307
GSM176158.CEL	Normal	F	Unknown	Normal	GSE7307
GSM176159.CEL	Normal	F	Unknown	Normal	GSE7307
GSM176160.CEL	Normal	F	Unknown	Normal	GSE7307
GSM80616.CEL	Normal	M	25	Normal	GSE3526
GSM80617.CEL	Normal	M	38	Normal	GSE3526
GSM80618.CEL	Normal	F	39	Normal	GSE3526
GSM80619.CEL	Normal	M	30	Normal	GSE3526
GSM80626.CEL	Normal	M	35	Normal	GSE3526
GSM80636.CEL	Normal	F	50	Normal	GSE3526
GSM80637.CEL	Normal	F	48	Normal	GSE3526
GSM80638.CEL	Normal	F	53	Normal	GSE3526
GSM80639.CEL	Normal	F	23	Normal	GSE3526
GSM1195778_DKFZ0003.CEL	Medulloblastoma	M	26	SHH	GSE49243
GSM1195779_DKFZ0019.CEL	Medulloblastoma	M	38	SHH	GSE49243
GSM1195780_DKFZ0023.CEL	Medulloblastoma	M	17	SHH	GSE49243
GSM1195781_DKFZ0025.CEL	Medulloblastoma	M	19	SHH	GSE49243
GSM1195782_DKFZ0029.CEL	Medulloblastoma	F	9	SHH	GSE49243
GSM1195783_DKFZ0030.CEL	Medulloblastoma	M	35	SHH	GSE49243
GSM1195784_DKFZ0031.CEL	Medulloblastoma	M	3	SHH	GSE49243
GSM1195785_DKFZ0032.CEL	Medulloblastoma	F	7	SHH	GSE49243
GSM1195786_DKFZ0036.CEL	Medulloblastoma	F	32	SHH	GSE49243
GSM1195787_DKFZ0037.CEL	Medulloblastoma	M	5	SHH	GSE49243
GSM1195788_DKFZ0048.CEL	Medulloblastoma	F	44	SHH	GSE49243
GSM1195789_DKFZ0051.CEL	Medulloblastoma	F	24	SHH	GSE49243
GSM1195790_DKFZ0052.CEL	Medulloblastoma	M	18	SHH	GSE49243
GSM1195791_DKFZ0053.CEL	Medulloblastoma	M	46	SHH	GSE49243
GSM1195792_DKFZ0057.CEL	Medulloblastoma	M	18	SHH	GSE49243
GSM1195793_DKFZ0059.CEL	Medulloblastoma	F	34	SHH	GSE49243
GSM1195794_DKFZ0060.CEL	Medulloblastoma	M	48	SHH	GSE49243
GSM1195795_DKFZ0201.CEL	Medulloblastoma	F	16	SHH	GSE49243
GSM1195796_DKFZ0216.CEL	Medulloblastoma	F	4	SHH	GSE49243
GSM1195797_DKFZ0217.CEL	Medulloblastoma	M	1	SHH	GSE49243

GSM1195798_DKFZ0218.CEL	Medulloblastoma	F	10	SHH	GSE49243
GSM1195799_DKFZ0222.CEL	Medulloblastoma	M	1	SHH	GSE49243
GSM1195800_DKFZ0223.CEL	Medulloblastoma	M	8	SHH	GSE49243
GSM1195801_DKFZ0226.CEL	Medulloblastoma	F	13	SHH	GSE49243
GSM1195802_DKFZ0227.CEL	Medulloblastoma	F	4	SHH	GSE49243
GSM1195803_DKFZ0229.CEL	Medulloblastoma	F	1	SHH	GSE49243
GSM1195804_DKFZ0241.CEL	Medulloblastoma	M	20	SHH	GSE49243
GSM1195805_DKFZ0244.CEL	Medulloblastoma	F	42	SHH	GSE49243
GSM1195806_DKFZ0248.CEL	Medulloblastoma	M	23	SHH	GSE49243
GSM1195807_DKFZ0249.CEL	Medulloblastoma	M	28	SHH	GSE49243
GSM1195808_DKFZ0250.CEL	Medulloblastoma	M	50	SHH	GSE49243
GSM1195809_DKFZ0251.CEL	Medulloblastoma	M	27	SHH	GSE49243
GSM1195810_DKFZ0345.CEL	Medulloblastoma	M	35	SHH	GSE49243
GSM1195811_DKFZ0347.CEL	Medulloblastoma	M	23	SHH	GSE49243
GSM1195812_DKFZ0349.CEL	Medulloblastoma	F	2	SHH	GSE49243
GSM1195813_DKFZ0350.CEL	Medulloblastoma	F	1.5	SHH	GSE49243
GSM1195814_DKFZ0361.CEL	Medulloblastoma	F	Unknown	SHH	GSE49243
GSM1195815_DKFZ0362.CEL	Medulloblastoma	M	31	SHH	GSE49243
GSM1195816_DKFZ0363.CEL	Medulloblastoma	M	36	SHH	GSE49243
GSM1195817_DKFZ0364.CEL	Medulloblastoma	F	39	SHH	GSE49243
GSM1195818_DKFZ0365.CEL	Medulloblastoma	M	26	SHH	GSE49243
GSM1195819_DKFZ0366.CEL	Medulloblastoma	F	49	SHH	GSE49243
GSM1195820_DKFZ0368.CEL	Medulloblastoma	M	32	SHH	GSE49243
GSM1195821_DKFZ0373.CEL	Medulloblastoma	F	0	SHH	GSE49243
GSM1195822_DKFZ0407.CEL	Medulloblastoma	M	2.7	SHH	GSE49243
GSM1195823_DKFZ0410.CEL	Medulloblastoma	M	25.6	SHH	GSE49243
GSM1195824_DKFZ0412.CEL	Medulloblastoma	F	1.6	SHH	GSE49243
GSM1195825_DKFZ0416.CEL	Medulloblastoma	F	1.6	SHH	GSE49243
GSM1195826_DKFZ0463.CEL	Medulloblastoma	F	30.7	SHH	GSE49243
GSM1195827_DKFZ0464.CEL	Medulloblastoma	F	28.6	SHH	GSE49243
GSM1195828_DKFZ0472.CEL	Medulloblastoma	M	1.5	SHH	GSE49243
GSM1195829_DKFZ0474.CEL	Medulloblastoma	F	27	SHH	GSE49243
GSM1195830_DKFZ0548.CEL	Medulloblastoma	M	13	SHH	GSE49243
GSM1195831_DKFZ0552.CEL	Medulloblastoma	M	Unknown	SHH	GSE49243
GSM1195832_DKFZ0557.CEL	Medulloblastoma	F	6	SHH	GSE49243
GSM1195833_DKFZ0558.CEL	Medulloblastoma	F	9	SHH	GSE49243
GSM1195834_DKFZ0561.CEL	Medulloblastoma	F	22	SHH	GSE49243
GSM1195835_DKFZ0579.CEL	Medulloblastoma	M	38	SHH	GSE49243
GSM1195836_DKFZ0581.CEL	Medulloblastoma	M	23	SHH	GSE49243
GSM1195837_DKFZ0582.CEL	Medulloblastoma	F	11	SHH	GSE49243
GSM1195838_DKFZ0583.CEL	Medulloblastoma	M	17	SHH	GSE49243
GSM1195839_DKFZ0584.CEL	Medulloblastoma	M	37	SHH	GSE49243
GSM1195840_DKFZ0596.CEL	Medulloblastoma	F	19	SHH	GSE49243
GSM1195841_DKFZ0597.CEL	Medulloblastoma	F	31	SHH	GSE49243
GSM1195842_DKFZ0599.CEL	Medulloblastoma	F	25	SHH	GSE49243
GSM1195843_DKFZ0614.CEL	Medulloblastoma	F	22	SHH	GSE49243
GSM1195844_DKFZ0661.CEL	Medulloblastoma	F	32	SHH	GSE49243
GSM1195845_DKFZ0761.CEL	Medulloblastoma	F	39	SHH	GSE49243
GSM1195846_DKFZ0762.CEL	Medulloblastoma	F	23	SHH	GSE49243
GSM1195847_DKFZ0764.CEL	Medulloblastoma	M	35	SHH	GSE49243
GSM1195848_DKFZ0782.CEL	Medulloblastoma	M	1.4	SHH	GSE49243
GSM1195849_DKFZ0787.CEL	Medulloblastoma	M	9	SHH	GSE49243
GSM1195850_DKFZ0788.CEL	Medulloblastoma	M	17	SHH	GSE49243
GSM1214914_MED_945.CEL	Medulloblastoma	Unknown	Unknown	G3	GSE50161

GSM1214915_MED_186.CEL	Medulloblastoma	Unknown	Unknown	G4	GSE50161
GSM1214916_MED_254.CEL	Medulloblastoma	Unknown	Unknown	G4	GSE50161
GSM1214917_MED_258.CEL	Medulloblastoma	Unknown	Unknown	SHH	GSE50161
GSM1214918_MED_262.CEL	Medulloblastoma	Unknown	Unknown	G4	GSE50161
GSM1214919_MED_277.CEL	Medulloblastoma	Unknown	Unknown	G3	GSE50161
GSM1214920_MED_288.CEL	Medulloblastoma	Unknown	Unknown	SHH	GSE50161
GSM1214921_MED_330.CEL	Medulloblastoma	Unknown	Unknown	G4	GSE50161
GSM1214922_MED_437.CEL	Medulloblastoma	Unknown	Unknown	G4	GSE50161
GSM1214923_MED_529.CEL	Medulloblastoma	Unknown	Unknown	SHH	GSE50161
GSM1214924_MED_565.CEL	Medulloblastoma	Unknown	Unknown	SHH	GSE50161
GSM1214925_MED_613.CEL	Medulloblastoma	Unknown	Unknown	WNT	GSE50161
GSM1214926_MED_676.CEL	Medulloblastoma	Unknown	Unknown	G3	GSE50161
GSM1214927_MED_719.CEL	Medulloblastoma	Unknown	Unknown	SHH	GSE50161
GSM1214928_MED_791.CEL	Medulloblastoma	Unknown	Unknown	G4	GSE50161
GSM1214929_MED_797.CEL	Medulloblastoma	Unknown	Unknown	G4	GSE50161
GSM1214930_MED_801.CEL	Medulloblastoma	Unknown	Unknown	SHH	GSE50161
GSM1214931_MED_801b.CEL	Medulloblastoma	Unknown	Unknown	SHH	GSE50161
GSM1214932_MED_877.CEL	Medulloblastoma	Unknown	Unknown	SHH	GSE50161
GSM1214933_MED_898.CEL	Medulloblastoma	Unknown	Unknown	SHH	GSE50161
GSM1214934_MED_925.CEL	Medulloblastoma	Unknown	Unknown	G3	GSE50161
GSM1214935_MED_B40.CEL	Medulloblastoma	Unknown	Unknown	SHH	GSE50161
GSM260959.CEL	Medulloblastoma	M	19	WNT	GSE10327
GSM260960.CEL	Medulloblastoma	M	7	G4	GSE10327
GSM260961.CEL	Medulloblastoma	M	3	SHH	GSE10327
GSM260962.CEL	Medulloblastoma	M	15	G3	GSE10327
GSM260963.CEL	Medulloblastoma	M	8	WNT	GSE10327
GSM260964.CEL	Medulloblastoma	M	4	G3	GSE10327
GSM260965.CEL	Medulloblastoma	M	14	G4	GSE10327
GSM260966.CEL	Medulloblastoma	F	7	G4	GSE10327
GSM260967.CEL	Medulloblastoma	M	3	SHH	GSE10327
GSM260968.CEL	Medulloblastoma	F	10	G4	GSE10327
GSM260969.CEL	Medulloblastoma	F	5	G4	GSE10327
GSM260970.CEL	Medulloblastoma	M	2	G3	GSE10327
GSM260971.CEL	Medulloblastoma	F	20	WNT	GSE10327
GSM260972.CEL	Medulloblastoma	M	2	SHH	GSE10327
GSM260973.CEL	Medulloblastoma	M	3	G4	GSE10327
GSM260974.CEL	Medulloblastoma	M	4	G4	GSE10327
GSM260975.CEL	Medulloblastoma	F	11	G4	GSE10327
GSM260976.CEL	Medulloblastoma	M	13.5	G4	GSE10327
GSM260977.CEL	Medulloblastoma	M	7.3	G4	GSE10327
GSM260978.CEL	Medulloblastoma	M	8	WNT	GSE10327
GSM260979.CEL	Medulloblastoma	M	6.4	G3	GSE10327
GSM260980.CEL	Medulloblastoma	M	1.8	SHH	GSE10327
GSM260981.CEL	Medulloblastoma	M	3.3	G3	GSE10327
GSM260982.CEL	Medulloblastoma	M	2.8	G3	GSE10327
GSM260983.CEL	Medulloblastoma	M	2.2	SHH	GSE10327
GSM260984.CEL	Medulloblastoma	M	3.1	G4	GSE10327
GSM260985.CEL	Medulloblastoma	M	5.9	G4	GSE10327
GSM260986.CEL	Medulloblastoma	F	4.8	G4	GSE10327
GSM260987.CEL	Medulloblastoma	M	27.1	SHH	GSE10327
GSM260988.CEL	Medulloblastoma	F	35.3	SHH	GSE10327
GSM260989.CEL	Medulloblastoma	M	10.3	G4	GSE10327
GSM260990.CEL	Medulloblastoma	M	16.6	G4	GSE10327
GSM260991.CEL	Medulloblastoma	F	5	G4	GSE10327

GSM260992.CEL	Medulloblastoma	M	5.3	G4	GSE10327
GSM260993.CEL	Medulloblastoma	F	7.8	WNT	GSE10327
GSM260994.CEL	Medulloblastoma	M	12.2	G4	GSE10327
GSM260995.CEL	Medulloblastoma	F	1.5	SHH	GSE10327
GSM260996.CEL	Medulloblastoma	F	25.6	G3	GSE10327
GSM260997.CEL	Medulloblastoma	M	10	G4	GSE10327
GSM260998.CEL	Medulloblastoma	F	6	G4	GSE10327
GSM260999.CEL	Medulloblastoma	M	7.4	G4	GSE10327
GSM261000.CEL	Medulloblastoma	M	10.4	WNT	GSE10327
GSM261001.CEL	Medulloblastoma	M	4.9	G4	GSE10327
GSM261002.CEL	Medulloblastoma	F	11.2	WNT	GSE10327
GSM261003.CEL	Medulloblastoma	M	12.7	WNT	GSE10327
GSM261004.CEL	Medulloblastoma	M	2.8	SHH	GSE10327
GSM261005.CEL	Medulloblastoma	F	Unknown	G3	GSE10327
GSM261006.CEL	Medulloblastoma	M	5.4	G4	GSE10327
GSM261007.CEL	Medulloblastoma	M	10	WNT	GSE10327
GSM261008.CEL	Medulloblastoma	M	6	G3	GSE10327
GSM261009.CEL	Medulloblastoma	M	5	G3	GSE10327
GSM261010.CEL	Medulloblastoma	F	6	WNT	GSE10327
GSM261011.CEL	Medulloblastoma	M	3	G4	GSE10327
GSM261012.CEL	Medulloblastoma	M	7	SHH	GSE10327
GSM261013.CEL	Medulloblastoma	F	12	G4	GSE10327
GSM261014.CEL	Medulloblastoma	F	3	SHH	GSE10327
GSM261015.CEL	Medulloblastoma	F	3.7	G4	GSE10327
GSM261016.CEL	Medulloblastoma	F	6.3	SHH	GSE10327
GSM261017.CEL	Medulloblastoma	F	31	SHH	GSE10327
GSM261018.CEL	Medulloblastoma	M	2.5	SHH	GSE10327
GSM261019.CEL	Medulloblastoma	F	2.4	G3	GSE10327
GSM261020.CEL	Medulloblastoma	M	3.8	G3	GSE10327
GSM324062.CEL	Medulloblastoma	Unknown	3	SHH	GSE12992
GSM324063.CEL	Medulloblastoma	Unknown	8	G4	GSE12992
GSM324064.CEL	Medulloblastoma	Unknown	9.1	SHH	GSE12992
GSM324065.CEL	Medulloblastoma	Unknown	3.4	SHH	GSE12992
GSM324066.CEL	Medulloblastoma	Unknown	7.3	G4	GSE12992
GSM324067.CEL	Medulloblastoma	Unknown	9.5	G4	GSE12992
GSM324068.CEL	Medulloblastoma	Unknown	11.2	G4	GSE12992
GSM324069.CEL	Medulloblastoma	Unknown	4.9	G3	GSE12992
GSM324082.CEL	Medulloblastoma	Unknown	8.9	G4	GSE12992
GSM324083.CEL	Medulloblastoma	Unknown	5.8	SHH	GSE12992
GSM324084.CEL	Medulloblastoma	Unknown	3.4	G3	GSE12992
GSM324085.CEL	Medulloblastoma	Unknown	3.7	G4	GSE12992
GSM324090.CEL	Medulloblastoma	Unknown	8.2	G4	GSE12992
GSM324091.CEL	Medulloblastoma	Unknown	7.6	G4	GSE12992
GSM324092.CEL	Medulloblastoma	Unknown	5.6	G4	GSE12992
GSM324093.CEL	Medulloblastoma	Unknown	9	G4	GSE12992
GSM324104.CEL	Medulloblastoma	Unknown	1	SHH	GSE12992
GSM324111.CEL	Medulloblastoma	Unknown	4.2	G3	GSE12992
GSM324112.CEL	Medulloblastoma	Unknown	10.5	SHH	GSE12992
GSM324113.CEL	Medulloblastoma	Unknown	8	G4	GSE12992
GSM324115.CEL	Medulloblastoma	Unknown	5.3	G4	GSE12992
GSM324119.CEL	Medulloblastoma	Unknown	11.5	G4	GSE12992
GSM324137.CEL	Medulloblastoma	Unknown	12.3	WNT	GSE12992
GSM324138.CEL	Medulloblastoma	Unknown	0.3	SHH	GSE12992
GSM324139.CEL	Medulloblastoma	Unknown	9.6	G4	GSE12992

GSM324140.CEL	Medulloblastoma	Unknown	7.8	G3	GSE12992
GSM324141.CEL	Medulloblastoma	Unknown	9	WNT	GSE12992
GSM324508.CEL	Medulloblastoma	Unknown	5.8	G4	GSE12992
GSM324512.CEL	Medulloblastoma	Unknown	7.1	G3	GSE12992
GSM324513.CEL	Medulloblastoma	Unknown	3.1	G4	GSE12992
GSM324514.CEL	Medulloblastoma	Unknown	10.2	G4	GSE12992
GSM324515.CEL	Medulloblastoma	Unknown	3.5	G3	GSE12992
GSM324516.CEL	Medulloblastoma	Unknown	5.6	WNT	GSE12992
GSM324517.CEL	Medulloblastoma	Unknown	6.5	SHH	GSE12992
GSM324526.CEL	Medulloblastoma	Unknown	2	SHH	GSE12992
GSM325233.CEL	Medulloblastoma	Unknown	10.9	G4	GSE12992
GSM325278.CEL	Medulloblastoma	Unknown	13.2	WNT	GSE12992
GSM325280.CEL	Medulloblastoma	Unknown	11	G4	GSE12992
GSM325281.CEL	Medulloblastoma	Unknown	2.8	G4	GSE12992
GSM325282.CEL	Medulloblastoma	Unknown	5	G3	GSE12992
GSM918578_mbt003-u133v2.CEL	Medulloblastoma	M	8.4	G4	GSE37418
GSM918579_mbt004-u133v2.CEL	Medulloblastoma	M	8.3	G4	GSE37418
GSM918580_mbt006-u133v2.CEL	Medulloblastoma	F	6.6	WNT	GSE37418
GSM918581_mbt008-u133v2.CEL	Medulloblastoma	M	9.2	G4	GSE37418
GSM918582_mbt009-u133v2.CEL	Medulloblastoma	M	8.6	SHH	GSE37418
GSM918583_mbt010-u133v2.CEL	Medulloblastoma	F	8.1	G4	GSE37418
GSM918584_mbt011-u133v2.CEL	Medulloblastoma	M	8.8	G4	GSE37418
GSM918585_mbt013-u133v2.CEL	Medulloblastoma	F	9	G4	GSE37418
GSM918586_mbt016-u133v2.CEL	Medulloblastoma	M	10.1	G3	GSE37418
GSM918587_mbt020-u133v2.CEL	Medulloblastoma	M	9.8	G3	GSE37418
GSM918588_mbt024-u133v2.CEL	Medulloblastoma	M	4.11	G3	GSE37418
GSM918589_mbt031-u133v2.CEL	Medulloblastoma	M	5.2	G3	GSE37418
GSM918590_mbt032-u133v2.CEL	Medulloblastoma	M	8.9	G4	GSE37418
GSM918591_mbt033-u133v2.CEL	Medulloblastoma	M	9.4	G4	GSE37418
GSM918592_mbt034-u133v2.CEL	Medulloblastoma	M	14.8	G4	GSE37418
GSM918593_mbt035-u133v2.CEL	Medulloblastoma	F	9.1	WNT	GSE37418
GSM918594_mbt037-u133v2.CEL	Medulloblastoma	M	4.8	G3	GSE37418
GSM918595_mbt045-u133v2.CEL	Medulloblastoma	F	7.9	G4	GSE37418
GSM918596_mbt046-u133v2.CEL	Medulloblastoma	F	3.4	G4	GSE37418
GSM918597_mbt048-u133v2.CEL	Medulloblastoma	M	5.5	G4	GSE37418
GSM918598_mbt050-u133v2.CEL	Medulloblastoma	M	11.9	G4	GSE37418
GSM918599_mbt051-u133v2.CEL	Medulloblastoma	M	6.9	G4	GSE37418
GSM918600_mbt053-u133v2.CEL	Medulloblastoma	M	12.11	G3	GSE37418
GSM918601_mbt058-u133v2.CEL	Medulloblastoma	M	3.4	G3	GSE37418
GSM918602_mbt062-u133v2.CEL	Medulloblastoma	F	13.5	G4	GSE37418
GSM918603_mbt063-u133v2.CEL	Medulloblastoma	M	11.7	WNT	GSE37418
GSM918604_mbt068-u133v2.CEL	Medulloblastoma	M	3.8	G4	GSE37418
GSM918605_mbt069-u133v2.CEL	Medulloblastoma	M	5	G4	GSE37418
GSM918606_mbt075-u133v2.CEL	Medulloblastoma	F	16.1	SHH	GSE37418
GSM918607_mbt078-u133v2.CEL	Medulloblastoma	F	10	SHH	GSE37418
GSM918608_mbt079-u133v2.CEL	Medulloblastoma	F	3.2	SHH	GSE37418
GSM918609_mbt081-u133v2.CEL	Medulloblastoma	F	8.7	SHH	GSE37418
GSM918610_mbt083-u133v2.CEL	Medulloblastoma	M	11.9	G4	GSE37418
GSM918611_mbt085-u133v2.CEL	Medulloblastoma	M	8.7	G3	GSE37418
GSM918612_mbt087-u133v2.CEL	Medulloblastoma	M	3.2	G3	GSE37418
GSM918613_mbt089-u133v2.CEL	Medulloblastoma	M	4.3	G4	GSE37418
GSM918614_mbt093-u133v2.CEL	Medulloblastoma	M	6.11	G3	GSE37418
GSM918615_mbt095-u133v2.CEL	Medulloblastoma	M	6.2	G4	GSE37418
GSM918616_mbt098-u133v2.CEL	Medulloblastoma	M	11.7	G4	GSE37418

GSM918617_mbt099-u133v2.CEL	Medulloblastoma	F	12.11	G4	GSE37418
GSM918618_mbt102-u133v2.CEL	Medulloblastoma	M	10.1	WNT	GSE37418
GSM918619_mbt103-u133v2.CEL	Medulloblastoma	M	16.8	SHH	GSE37418
GSM918620_mbt105-u133v2.CEL	Medulloblastoma	F	7	SHH	GSE37418
GSM918621_mbt106-u133v2.CEL	Medulloblastoma	M	8.9	SHH	GSE37418
GSM918622_mbt109-u133v2.CEL	Medulloblastoma	F	6.9	G4	GSE37418
GSM918623_mbt110-u133v2.CEL	Medulloblastoma	M	8.8	G4	GSE37418
GSM918624_mbt124-u133v2.CEL	Medulloblastoma	M	5.2	G3	GSE37418
GSM918625_mbt126-u133v2.CEL	Medulloblastoma	F	9.2	WNT	GSE37418
GSM918626_mbt127-u133v2.CEL	Medulloblastoma	M	6.4	G4	GSE37418
GSM918627_mbt135-u133v2.CEL	Medulloblastoma	M	6.11	G4	GSE37418
GSM918628_mbt136-u133v2.CEL	Medulloblastoma	M	8.4	SHH_OUTLIER	GSE37418
GSM918629_mbt140-u133v2.CEL	Medulloblastoma	M	13.7	G3	GSE37418
GSM918630_mbt141-u133v2.CEL	Medulloblastoma	F	6.4	G4	GSE37418
GSM918631_mbt144-u133v2.CEL	Medulloblastoma	F	5.9	G4	GSE37418
GSM918632_mbt145-u133v2.CEL	Medulloblastoma	M	9.2	G4	GSE37418
GSM918633_mbt146-u133v2.CEL	Medulloblastoma	M	8.8	G4	GSE37418
GSM918634_mbt147-u133v2.CEL	Medulloblastoma	M	8.2	G4	GSE37418
GSM918635_mbt148-u133v2.CEL	Medulloblastoma	M	5.1	G4	GSE37418
GSM918636_mbt149-u133v2.CEL	Medulloblastoma	F	7.1	G3	GSE37418
GSM918637_mbt150-u133v2.CEL	Medulloblastoma	M	3.11	G3	GSE37418
GSM918638_mbt151-u133v2.CEL	Medulloblastoma	F	8.2	WNT	GSE37418
GSM918639_mbt156-u133v2.CEL	Medulloblastoma	M	5	G3	GSE37418
GSM918640_mbt158-u133v2.CEL	Medulloblastoma	M	10.3	G3	GSE37418
GSM918641_mbt161-u133v2.CEL	Medulloblastoma	M	8.5	WNT	GSE37418
GSM918642_mbt166-u133v2.CEL	Medulloblastoma	F	8.5	WNT	GSE37418
GSM918643_mbt167-u133v2.CEL	Medulloblastoma	F	8.4	WNT	GSE37418
GSM918644_mbt168-u133v2.CEL	Medulloblastoma	F	9.1	WNT	GSE37418
GSM918645_tbm055-u133v2.CEL	Medulloblastoma	M	6.11	G4	GSE37418
GSM918646_tbm061-u133v2.CEL	Medulloblastoma	M	9.1	G4	GSE37418
GSM918647_tbm082-u133v2.CEL	Medulloblastoma	M	11.8	G4	GSE37418
GSM918648_tbm084-u133v2.CEL	Medulloblastoma	M	12.1	G4	GSE37418
GSM918649_tbm091-u133v2.CEL	Medulloblastoma	M	4.1	SHH	GSE37418
GSM918650_tbm092-u133v2.CEL	Medulloblastoma	M	12.2	G4	GSE37418
GSM918651_tbm107-u133v2.CEL	Medulloblastoma	M	5.1	SHH	GSE37418
GSM918652_tbm111-u133v2.CEL	Medulloblastoma	M	10.1	G4	GSE37418
GSM918653_tbm143-u133v2.CEL	Medulloblastoma	M	5.9	G4	GSE37418

Supplementary Table 11: MB sample summary, related to Figure 4.3

ID #	Diagnosis	Moleclar Subgroup	Age (years)	Dot blot assay	hME-seal Seq	Institution
MB01	Medulloblastoma	Unknown	7	Yes	Yes	Xiangya Hospital
MB02	Medulloblastoma	Unknown	6	Yes	Yes	Xiangya Hospital
MB03	Medulloblastoma	Unknown	7	Yes	Yes	Xiangya Hospital
MB04	Medulloblastoma	SHH	7	Yes	Yes	Aflac cancer center
MB05	Medulloblastoma	Grp3/4	5	Yes	Yes	Aflac cancer center
MB06	Medulloblastoma	SHH	19	Yes	Yes	Aflac cancer center
MB07	Medulloblastoma	Grp3/4	3	Yes	Yes	Aflac cancer center
MB08	Medulloblastoma	Grp3/4	11	Yes	Yes	Aflac cancer center
MB09	Medulloblastoma	WNT	7	No	Yes	Dr. Erwin G. Van Meir
MB10	Medulloblastoma	WNT	11	No	Yes	Dr. Erwin G. Van Meir
MB11	Medulloblastoma	SHH	34	No	Yes	Dr. Erwin G. Van Meir
MB12	Medulloblastoma	SHH	1.5	No	Yes	Dr. Erwin G. Van Meir
MB13	Medulloblastoma	Group3	2	No	Yes	Dr. Erwin G. Van Meir
MB14	Medulloblastoma	Group3	7	No	Yes	Dr. Erwin G. Van Meir
MB15	Medulloblastoma	Group4	11	No	Yes	Dr. Erwin G. Van Meir
MB16	Medulloblastoma	Group4	12	No	Yes	Dr. Erwin G. Van Meir
CB01	No brain disorder	-	11	Yes	Yes	University of Maryland
CB02	No brain disorder	-	5	Yes	Yes	University of Maryland
CB03	No brain disorder	-	19	Yes	Yes	University of Maryland
CB04	No brain disorder	-	10	Yes	Yes	University of Maryland
CB05	No brain disorder	-	5	Yes	Yes	University of Maryland
CB06	No brain disorder	-	19	Yes	Yes	University of Maryland

Supplementary Table 12: Summary of annotation using 5hmC gain and 5hmC loss regions, related to Figure 4.3B

Annotation	5hmC gain				5hmC loss			
	# of peaks	Total size (bp)	Log2 Enrichment	% of total peaks	# of peaks	Total size (bp)	Log2 Enrichment	% of total peaks
3UTR	136	22886309	0.885	1.392586525	29	22886309	0.968	1.475826972
miRNA	0	94147	-13.254	0	0	94147	-10.94	0
ncRNA	49	6695098	1.185	0.501740733	7	6695098	0.691	0.356234097
TTS	163	31230086	0.698	1.669055908	20	31230086	-0.016	1.017811705
pseudo	3	1946287	-1.062	0.03071882	4	1946287	1.666	0.203562341
Exon	200	36298701	0.776	2.04792136	62	36298701	1.399	3.155216285
Intron	5524	1248800035	0.459	56.56358796	1532	1248800035	0.922	77.96437659
Intergenic	3291	1648836077	-0.689	33.69854598	275	1648836077	-1.957	13.99491094
Promoter	370	34875586	1.721	3.788654516	33	34875586	0.547	1.679389313
5UTR	30	2766515	1.752	0.307188204	3	2766515	0.744	0.152671756
snoRNA	0	262	-13.254	0	0	262	-10.94	0
snRNA	0	11	-13.254	0	0	11	-10.94	0

Supplementary Table 13: Motif analysis results of human 5hmC gain

Motif Name	Consensus	P-value	Log P-value	q-value (Benjamini)	# of Target Sequences with Motif (of 9766)	% of Target Sequences with Motif	# of Background Sequences with Motif (of 40149)	% of Background Sequences with Motif
SCL(bHLH)/HPC7-Scl-ChIP-Seq(GSE13511)/Homer	AVCAGCTG	1.00E-13	-3.08E+01	0	4652	47.63%	17616.3	43.87%
Nanog(Homeobox)/mES-Nanog-ChIP-Seq(GSE11724)/Homer	RGCCATTAAC	1.00E-04	-1.06E+01	0.0003	4235	43.36%	16595.4	41.33%
AR-halfsite(NR)/LNCaP-AR-ChIP-Seq(GSE27824)/Homer	CCAGGAACAG	1.00E-02	-6.38E+00	0.0087	3734	38.23%	14775.8	36.80%
Tgif2(Homeobox)/mES-Tgif2-ChIP-Seq(GSE55404)/Homer	TGTCANYT	1.00E-05	-1.17E+01	0.0001	3323	34.03%	12841	31.98%
Ptf1a(bHLH)/Panc1-Ptf1a-ChIP-Seq(GSE47459)/Homer	ACAGCTGTTN	1.00E-03	-8.74E+00	0.0014	3197	32.74%	12463.1	31.04%
Tgif1(Homeobox)/mES-Tgif1-ChIP-Seq(GSE55404)/Homer	YTGWCADY	1.00E-02	-4.64E+00	0.0341	3012	30.84%	11946.1	29.75%
Smad3(MAD)/NPC-Smad3-ChIP-Seq(GSE36673)/Homer	TWGTCTGV	1.00E-03	-8.82E+00	0.0013	2821	28.89%	10938.3	27.24%
Nkx6.1(Homeobox)/Islet-Nkx6.1-ChIP-Seq(GSE40975)/Homer	GKTAATGR	1.00E-08	-1.91E+01	0	2754	28.20%	10296.6	25.64%
HEB(bHLH)/mES-Heb-ChIP-Seq(GSE53233)/Homer	VCAGCTGBNN	1.00E-03	-7.81E+00	0.0029	2465	25.24%	9550.7	23.79%
Eomes(T-box)/H9-Eomes-ChIP-Seq(GSE26097)/Homer	ATTAACACCT	1.00E-02	-6.68E+00	0.0069	2251	23.05%	8743.8	21.78%
Znf263(Zf)/K562-Znf263-ChIP-Seq(GSE31477)/Homer	CVGTSCTCCC	1.00E-02	-4.64E+00	0.0341	1994	20.42%	7818.1	19.47%
Olig2(bHLH)/Neuron-Olig2-ChIP-Seq(GSE30882)/Homer	RCCATMTGTT	1.00E-06	-1.58E+01	0	2038	20.87%	7551.3	18.81%
Isl1(Homeobox)/Neuron-Isl1-ChIP-Seq(GSE31456)/Homer	CTAATKGV	1.00E-04	-1.03E+01	0.0004	1945	19.92%	7362.5	18.34%
Foxo1(Forkhead)/RAW-Foxo1-ChIP-Seq(Fan_et_al.)/Homer	CTGTTTAC	1.00E-05	-1.27E+01	0.0001	1961	20.08%	7342.7	18.29%
BMAL1(bHLH)/Liver-Bmal1-ChIP-Seq(GSE39860)/Homer	GNCACGTG	1.00E-04	-1.03E+01	0.0004	1821	18.65%	6869.5	17.11%
NF1-halfsite(CTF)/LNCaP-NF1-ChIP-Seq(Unpublished)/Homer	YTGCCAAG	1.00E-18	-4.24E+01	0	2007	20.55%	6861.4	17.09%
Ascl1(bHLH)/NeuralTubes-Ascl1-ChIP-Seq(GSE55840)/Homer	NNVVCAGCTGBN	1.00E-11	-2.64E+01	0	1885	19.30%	6689.7	16.66%

LRF(Zf)/Erythroblasts-ZBTB7A-ChIP-Seq(GSE74977)/Homer	AAGACCCYYN	1.00E-02	-5.55E+00	0.0166	1678	17.18%	6496.6	16.18%
Lhx3(Homeobox)/Neuron-Lhx3-ChIP-Seq(GSE31456)/Homer	ADBTAATTAR	1.00E-07	-1.65E+01	0	1756	17.98%	6421.6	15.99%
NPAS(bHLH)/Liver-NPAS-ChIP-Seq(GSE39860)/Homer	NVCACGTG	1.00E-02	-6.72E+00	0.0067	1597	16.35%	6117.5	15.24%
NeuroG2(bHLH)/Fibroblast-NeuroG2-ChIP-Seq(GSE75910)/Homer	ACCATCTGTT	1.00E-09	-2.28E+01	0	1698	17.39%	6043.8	15.05%
Sox6(HMG)/Myotubes-Sox6-ChIP-Seq(GSE32627)/Homer	CCATTGTTNY	1.00E-02	-4.86E+00	0.0291	1513	15.49%	5869.6	14.62%
Smad4(MAD)/ESC-SMAD4-ChIP-Seq(GSE29422)/Homer	VBSYGTCTGG	1.00E-02	-5.82E+00	0.0134	1515	15.51%	5830.1	14.52%
Rbpj1(?)/Panc1-Rbpj1-ChIP-Seq(GSE47459)/Homer	HTTTCCCASG	1.00E-02	-4.76E+00	0.0313	1497	15.33%	5810.5	14.47%
ZFX(Zf)/mES-Zfx-ChIP-Seq(GSE11431)/Homer	AGGCCTRG	1.00E-02	-5.46E+00	0.0176	1497	15.33%	5775.9	14.38%
Smad2(MAD)/ES-SMAD2-ChIP-Seq(GSE29422)/Homer	CTGTCTGG	1.00E-03	-7.60E+00	0.0033	1507	15.43%	5722.3	14.25%
MYB(HTH)/ERMYB-Myb-ChIPSeq(GSE22095)/Homer	GGCVGTTR	1.00E-03	-7.01E+00	0.0052	1498	15.34%	5710.6	14.22%
AMYB(HTH)/Testes-AMYB-ChIP-Seq(GSE44588)/Homer	TGGCAGTTGG	1.00E-02	-5.94E+00	0.0123	1305	13.36%	4986.4	12.42%
TATA-Box(TBP)/Promoter/Homer	CCTTTTAWAGSC	1.00E-02	-4.83E+00	0.0295	1286	13.17%	4962	12.36%
BMYB(HTH)/Hela-BMYB-ChIP-Seq(GSE27030)/Homer	NHAACBGYYV	1.00E-02	-6.80E+00	0.0063	1290	13.21%	4891.8	12.18%
Bcl6(Zf)/Liver-Bcl6-ChIP-Seq(GSE31578)/Homer	NNNCTTTCCAGGAAA	1.00E-06	-1.56E+01	0	1357	13.90%	4888.5	12.17%
Ap4(bHLH)/AML-Tfap4-ChIP-Seq(GSE45738)/Homer	NAHCAGCTGD	1.00E-08	-1.93E+01	0	1322	13.54%	4667.7	11.62%
FOXA1(Forkhead)/LNCAP-FOXA1-ChIP-Seq(GSE27824)/Homer	WAAGTAAACA	1.00E-03	-8.30E+00	0.002	1246	12.76%	4662.3	11.61%
AP-2gamma(AP2)/MCF7-TFAP2C-ChIP-Seq(GSE21234)/Homer	SCCTSAGGSCAW	1.00E-03	-8.63E+00	0.0015	1205	12.34%	4489	11.18%
EBF1(EBF)/Near-E2A-ChIP-Seq(GSE21512)/Homer	GTCCCCWGGGGA	1.00E-05	-1.18E+01	0.0001	1220	12.49%	4452.3	11.09%
Lhx1(Homeobox)/EmbryoCarcinoma-Lhx1-ChIP-Seq(GSE70957)/Homer	NNYTAATTAR	1.00E-06	-1.50E+01	0	1224	12.53%	4387.3	10.93%
RXR(NR),DR1/3T3L1-RXR-ChIP-Seq(GSE13511)/Homer	TAGGGCAAAGGTCA	1.00E-04	-9.25E+00	0.001	1178	12.06%	4363.9	10.87%

Atoh1(bHLH)/Cerebellum-Atoh1-ChIP-Seq(GSE22111)/Homer	VNRVCAGCTGGY	1.00E-11	-2.69E+01	0	1236	12.66%	4194.8	10.45%
MyoG(bHLH)/C2C12-MyoG-ChIP-Seq(GSE36024)/Homer	AACAGCTG	1.00E-04	-9.70E+00	0.0006	1107	11.34%	4071.7	10.14%
FOXM1(Forkhead)/MCF7-FOXM1-ChIP-Seq(GSE72977)/Homer	TRTTTACTTW	1.00E-02	-5.93E+00	0.0123	1061	10.86%	4017.7	10.01%
Lhx2(Homeobox)/HFSC-Lhx2-ChIP-Seq(GSE48068)/Homer	TAATTAGN	1.00E-10	-2.31E+01	0	1166	11.94%	3997	9.95%
Pit1(Homeobox)/GCrat-Pit1-ChIP-Seq(GSE58009)/Homer	ATGMATATDC	1.00E-03	-7.59E+00	0.0033	1041	10.66%	3878.3	9.66%
Tcf12(bHLH)/GM12878-Tcf12-ChIP-Seq(GSE32465)/Homer	VCAGCTGYTG	1.00E-05	-1.25E+01	0.0001	1060	10.85%	3813.3	9.50%
FOXK1(Forkhead)/HEK293-FOXK1-ChIP-Seq(GSE51673)/Homer	NVWTGTTTAC	1.00E-02	-6.38E+00	0.0087	1013	10.37%	3810.7	9.49%
HIF-1b(HLH)/T47D-HIF1b-ChIP-Seq(GSE59937)/Homer	RTACGTGC	1.00E-03	-7.71E+00	0.0031	1016	10.40%	3776.1	9.40%
PPARE(NR),DR1/3T3L1-Pparg-ChIP-Seq(GSE13511)/Homer	TGACCTTTGCCCCA	1.00E-02	-5.63E+00	0.0158	974	9.97%	3684.2	9.18%
Tcf21(bHLH)/ArterySmoothMuscle-Tcf21-ChIP-Seq(GSE61369)/Homer	NAACAGCTGG	1.00E-05	-1.31E+01	0	1028	10.53%	3675.5	9.15%
Foxf1(Forkhead)/Lung-Foxf1-ChIP-Seq(GSE77951)/Homer	WWATRTAAACAN	1.00E-02	-6.68E+00	0.0069	962	9.85%	3598.8	8.96%
ZNF467(Zf)/HEK293-ZNF467.GFP-ChIP-Seq(GSE58341)/Homer	TGGGGAAGGGCM	1.00E-02	-5.45E+00	0.0176	947	9.70%	3584.4	8.93%
AP-2alpha(AP2)/Hela-AP2alpha-ChIP-Seq(GSE31477)/Homer	ATGCCCTGAGGC	1.00E-03	-8.11E+00	0.0023	931	9.53%	3430	8.54%
Otx2(Homeobox)/EpiLC-Otx2-ChIP-Seq(GSE56098)/Homer	NYTAATCCYB	1.00E-02	-6.13E+00	0.0104	914	9.36%	3428.2	8.54%
Pdx1(Homeobox)/Islet-Pdx1-ChIP-Seq(SRA008281)/Homer	YCATYAATCA	1.00E-07	-1.62E+01	0	972	9.95%	3391.2	8.45%
Fox:Ebox(Forkhead,bHLH)/Panc1-Foxa2-ChIP-Seq(GSE47459)/Homer	NNNVCTGWGYAAACASN	1.00E-05	-1.28E+01	0.0001	943	9.66%	3354	8.35%
STAT4(Stat)/CD4-Stat4-ChIP-Seq(GSE22104)/Homer	NYTTCCWGGAAR	1.00E-04	-9.25E+00	0.001	912	9.34%	3323.6	8.28%
NFAT(RHD)/Jurkat-NFATC1-ChIP-Seq(Jolma_et_al.)/Homer	ATTTTCCATT	1.00E-03	-8.73E+00	0.0014	899	9.21%	3288	8.19%
FoxL2(Forkhead)/Ovary-FoxL2-ChIP-Seq(GSE60858)/Homer	WWTRTAAACAVG	1.00E-02	-5.21E+00	0.0216	855	8.75%	3229.5	8.04%

NeuroD1(bHLH)/Islet-NeuroD1-ChIP-Seq(GSE30298)/Homer	GCCATCTGTT	1.00E-09	-2.15E+01	0	947	9.70%	3197.3	7.96%
Zic(Zf)/Cerebellum-ZIC1.2-ChIP-Seq(GSE60731)/Homer	CCTGCTGAGH	1.00E-02	-5.14E+00	0.0228	806	8.25%	3038.2	7.57%
MyoD(bHLH)/Myotube-MyoD-ChIP-Seq(GSE21614)/Homer	RRCAGCTGYTSY	1.00E-03	-7.15E+00	0.0048	778	7.97%	2861.5	7.13%
Foxa2(Forkhead)/Liver-Foxa2-ChIP-Seq(GSE25694)/Homer	CYTGTTTACWYW	1.00E-04	-1.00E+01	0.0005	764	7.82%	2730.2	6.80%
Foxo3(Forkhead)/U2OS-Foxo3-ChIP-Seq(E-MTAB-2701)/Homer	DGTAAACA	1.00E-03	-7.07E+00	0.0051	722	7.39%	2645.4	6.59%
Unknown(Homeobox)/Limb-p300-ChIP-Seq/Homer	SSCMATWAAA	1.00E-02	-5.60E+00	0.0161	706	7.23%	2628.7	6.55%
Myf5(bHLH)/GM-Myf5-ChIP-Seq(GSE24852)/Homer	BAACAGCTGT	1.00E-03	-9.18E+00	0.001	719	7.36%	2577.9	6.42%
MafA(bZIP)/Islet-MafA-ChIP-Seq(GSE30298)/Homer	TGCTGACTCA	1.00E-02	-5.46E+00	0.0176	687	7.03%	2558.3	6.37%
OCT:OCT-short(POU,Homeobox)/NPC-OCT6-ChIP-Seq(GSE43916)/Homer	ATGCATWATGCATRW	1.00E-02	-6.18E+00	0.0102	687	7.03%	2535	6.31%
n-Myc(bHLH)/mES-nMyc-ChIP-Seq(GSE11431)/Homer	VRCCACGTGG	1.00E-04	-1.13E+01	0.0002	688	7.04%	2409.4	6.00%
Egr1(Zf)/K562-Egr1-ChIP-Seq(GSE32465)/Homer	TGCGTGGGYG	1.00E-02	-4.85E+00	0.0292	641	6.56%	2398.5	5.97%
HOXA9(Homeobox)/HSC-Hoxa9-ChIP-Seq(GSE33509)/Homer	GGCCATAAATCA	1.00E-02	-5.65E+00	0.0156	636	6.51%	2352.9	5.86%
Unknown-ESC-element(?)/mES-Nanog-ChIP-Seq(GSE11724)/Homer	CACAGCAGGGGG	1.00E-03	-7.97E+00	0.0025	652	6.68%	2349.5	5.85%
Zic3(Zf)/mES-Zic3-ChIP-Seq(GSE37889)/Homer	GGCCYCCTGCTGDGH	1.00E-05	-1.22E+01	0.0001	674	6.90%	2339.1	5.83%
CHR(?)/Hela-CellCycle-Expression/Homer	SRGTTTCAAA	1.00E-05	-1.28E+01	0.0001	671	6.87%	2315.7	5.77%
Stat3+il21(Stat)/CD4-Stat3-ChIP-Seq(GSE19198)/Homer	SVYTTCCNGGAARB	1.00E-03	-8.95E+00	0.0012	650	6.66%	2317.7	5.77%
KLF10(Zf)/HEK293-KLF10.GFP-ChIP-Seq(GSE58341)/Homer	GGGGGTGTGTCC	0.1	-4.414	0.0412	612	0.0627	2299.3	0.0573
Max(bHLH)/K562-Max-ChIP-Seq(GSE31477)/Homer	RCCACGTGGYYN	0.01	-6.171	0.0102	618	0.0633	2266.9	0.0565
FOXK2(Forkhead)/U2OS-FOXK2-ChIP-Seq(E-MTAB-2204)/Homer	SCHTGTTTACAT	0.0001	-10.66	0.0003	642	0.0657	2248.2	0.056

Arnt:Ahr(bHLH)/MCF7-Arnt-ChIP-Seq(Lo_et_al.)/Homer	TBGCACGCAA	0.001	-7.668	0.0032	578	0.0592	2071.1	0.0516
ZNF264(Zf)/HEK293-ZNF264.GFP-ChIP-Seq(GSE58341)/Homer	RGGGCACTAACY	0.001	-7.049	0.0051	568	0.0582	2048.8	0.051
c-Myc(bHLH)/mES-cMyc-ChIP-Seq(GSE11431)/Homer	VVCCACGTGG	0.01	-5.145	0.0228	503	0.0515	1848.5	0.046
CLOCK(bHLH)/Liver-Clock-ChIP-Seq(GSE39860)/Homer	GHCACGTG	0.001	-7.212	0.0046	513	0.0525	1832.9	0.0456
Nr5a2(NR)/mES-Nr5a2-ChIP-Seq(GSE19019)/Homer	BTCAAGGTCA	0.001	-8.161	0.0022	508	0.052	1791.2	0.0446
Oct4(POU,Homeobox)/mES-Oct4-ChIP-Seq(GSE11431)/Homer	ATTTGCATAW	0.01	-5.141	0.0228	476	0.0487	1743.8	0.0434
Hoxc9(Homeobox)/Ainv15-Hoxc9-ChIP-Seq(GSE21812)/Homer	GGCCATAAATCA	0.001	-7.328	0.0041	478	0.0489	1695.8	0.0422
Hand2(bHLH)/Mesoderm-Hand2-ChIP-Seq(GSE61475)/Homer	TGACANARRCCAGRC	0.01	-5.548	0.0166	465	0.0476	1689.7	0.0421
Stat3(Stat)/mES-Stat3-ChIP-Seq(GSE11431)/Homer	CTTCCGGGAA	0.001	-8.232	0.0021	468	0.0479	1637.9	0.0408
Tlx?(NR)/NPC-H3K4me1-ChIP-Seq(GSE16256)/Homer	CTGGCAGSCTGCCA	1E-13	-30.49	0	546	0.0559	1617.4	0.0403
NF1(CTF)/LNCAP-NF1-ChIP-Seq(Unpublished)/Homer	CYTGGCABNSTGCCAR	1E-08	-20.71	0	491	0.0503	1529.7	0.0381
USF1(bHLH)/GM12878-Usf1-ChIP-Seq(GSE32465)/Homer	SGTCACGTGR	0.01	-5.485	0.0174	417	0.0427	1505	0.0375
PAX5(Paired,Homeobox)/GM12878-PAX5-ChIP-Seq(GSE32465)/Homer	GCAGCCAAGCRTGACH	0.01	-4.978	0.0261	412	0.0422	1499.9	0.0374
FOXP1(Forkhead)/H9-FOXP1-ChIP-Seq(GSE31006)/Homer	NYYTGTTTACHN	0.001	-7.825	0.0029	411	0.0421	1429.9	0.0356
Pax8(Paired,Homeobox)/Thyroid-Pax8-ChIP-Seq(GSE26938)/Homer	GTCATGCHTGRCTGS	0.01	-6.532	0.0077	387	0.0396	1366	0.034
c-Myc(bHLH)/LNCAP-cMyc-ChIP-Seq(Unpublished)/Homer	VCCACGTG	0.0001	-9.982	0.0005	402	0.0412	1355.8	0.0338
Pit1+1bp(Homeobox)/GCrat-Pit1-ChIP-Seq(GSE58009)/Homer	ATGCATAATTCA	0.01	-6.555	0.0076	358	0.0367	1254.9	0.0313
Foxa3(Forkhead)/Liver-Foxa3-ChIP-Seq(GSE77670)/Homer	BSNTGTTTACWYWGN	0.001	-7.986	0.0025	325	0.0333	1102.4	0.0275
STAT5(Stat)/mCD4+-Stat5-ChIP-Seq(GSE12346)/Homer	RTTTCTNAGAAA	0.001	-7.36	0.0041	298	0.0305	1012.7	0.0252

PRDM14(Zf)/H1-PRDM14-ChIP-Seq(GSE22767)/Homer	RGGTCTCTAACY	0.01	-4.742	0.0314	274	0.0281	973.9	0.0243
STAT1(Stat)/HelaS3-STAT1-ChIP-Seq(GSE12782)/Homer	NATTTCCNGGAAAT	0.01	-6.497	0.0078	264	0.027	900.4	0.0224
EBF(EBF)/proBcell-EBF-ChIP-Seq(GSE21978)/Homer	DGTCCCYRGGGA	0.01	-4.611	0.0343	240	0.0246	846.6	0.0211
Tbx20(T-box)/Heart-Tbx20-ChIP-Seq(GSE29636)/Homer	GGTGYTGACAGS	0.01	-6.187	0.0102	192	0.0197	637.9	0.0159
GRE(NR),IR3/RAW264.7-GRE-ChIP-Seq(Unpublished)/Homer	VAGRACAKWCTGTYC	0.01	-5.362	0.0188	186	0.019	628.7	0.0157
Pitx1:Ebox(Homeobox,bHLH)/Hindlimb-Pitx1-ChIP-Seq(GSE41591)/Homer	YTAATTRAWWCCAGATGT	0.0001	-10.79	0.0003	206	0.0211	628.3	0.0156
Rfx1(HTH)/NPC-H3K4me1-ChIP-Seq(GSE16256)/Homer	KGTTGCCATGGCAA	0.01	-6.072	0.0109	188	0.0193	624.9	0.0156
OCT4-SOX2-TCF-NANOG(POU,Homeobox,HMG)/mES-Oct4-ChIP-Seq(GSE11431)/Homer	ATTTGCATAACAATG	0.01	-4.617	0.0343	178	0.0182	611.4	0.0152
Brn2(POU,Homeobox)/NPC-Brn2-ChIP-Seq(GSE35496)/Homer	ATGAATATTC	0.001	-9.155	0.001	107	0.011	300.9	0.0075
Rfx2(HTH)/LoVo-RFX2-ChIP-Seq(GSE49402)/Homer	GTTGCCATGGCAACM	0.0001	-9.833	0.0006	82	0.0084	212.6	0.0053
RFX(HTH)/K562-RFX3-ChIP-Seq(SRA012198)/Homer	CGGTTGCCATGGCAAC	0.00001	-11.8	0.0001	74	0.0076	175.9	0.0044
NF1:FOXA1(CTF,Forkhead)/LNCAP-FOXA1-ChIP-Seq(GSE27824)/Homer	WNTGTTTRYTTTGGCA	0.01	-4.736	0.0314	45	0.0046	126.4	0.0031
OCT:OCT(POU,Homeobox,IR1)/NPC-Brn2-ChIP-Seq(GSE35496)/Homer	ATGAATWATTCATGA	0.01	-5.933	0.0123	21	0.0022	43.3	0.0011

Supplementary Table 14: Summary of MSigPathway results (GREAT analysis), related to Figure 4.3F

ID	Desc	Binom FdrQ	RegionFold Enrich	Hyper FdrQ	GeneFold Enrich
REACTOME_IONOTROPIC_ACTIVITY_OF_KAINATE_RECEPTORS	Genes involved in Ionotropic activity of Kainate Receptors	7.92E-23	3.79	3.89E-03	2.74E+00
REACTOME_SIGNALING_BY_NOTCH	Genes involved in Signaling by NOTCH	1.19E-19	1.98	9.13E-04	1.63E+00
REACTOME_SIGNALING_BY_NOTCH1	Genes involved in Signaling by NOTCH1	1.11E-19	2.14	8.52E-04	1.77E+00
REACTOME_NOTCH1_INTRACELLULAR_DOMAIN_REGULATES_TRANSCRIPTION	Genes involved in NOTCH1 Intracellular Domain Regulates Transcription	1.75E-18	2.27	3.83E-03	1.85E+00
PID_NETRIN_PATHWAY	Netrin-mediated signaling events	1.93E-17	2.55	7.50E-04	2.17E+00
REACTOME_SIGNALING_BY_RHO_GTPASES	Genes involved in Signaling by Rho GTPases	5.44E-16	1.84	7.49E-03	1.48E+00
WNT_SIGNALING	Genes related to Wnt-mediated signal transduction	3.60E-11	1.76	7.55E-06	1.86E+00
PID_WNT_NONCANONICAL_PATHWAY	Noncanonical Wnt signaling pathway	1.32E-10	2.27	1.93E-03	2.07E+00
REACTOME_DCC_MEDIATED_ATTRACTIVE_SIGNALING	Genes involved in DCC mediated attractive signaling	1.62E-08	2.54	5.44E-03	2.55E+00
PID_REELINPATHWAY	Reelin signaling pathway	3.23E-08	2.01	1.01E-03	2.18E+00
ST_INTEGRIN_SIGNALING_PATHWAY	Integrin Signaling Pathway	6.78E-08	1.69	3.65E-03	1.62E+00
PID_FAK_PATHWAY	Signaling events mediated by focal adhesion kinase	7.93E-08	1.74	7.75E-04	1.84E+00
KEGG_ADHERENS_JUNCTION	Adherens junction	1.07E-06	1.56	2.02E-06	2.01E+00

PID_PDGFRBPATHWAY	PDGFR-beta signaling pathway	1.27E-06	1.51	3.64E-03	1.48E+00
REACTOME_PHOSPHOLIPASE_C_MEDIATED_CASCADE	Genes involved in Phospholipase C-mediated cascade	3.48E-06	1.66	8.88E-03	1.71E+00
PID_HIF1_TFPATHWAY	HIF-1-alpha transcription factor network	4.97E-06	1.62	1.61E-04	1.87E+00
PID_MET_PATHWAY	Signaling events mediated by Hepatocyte Growth Factor Receptor (c-Met)	6.02E-06	1.58	2.06E-03	1.66E+00
KEGG_COLORECTAL_CANCER	Colorectal cancer	1.34E-05	1.56	1.49E-03	1.77E+00
BIOCARTA_TGFB_PATHWAY	TGF beta signaling pathway	8.18E-05	1.90	8.76E-03	2.22E+00
REACTOME_CIRCADIAN_CLOCK	Genes involved in Circadian Clock	1.89E-04	1.52	6.49E-05	2.03E+00
PID_BMPPATHWAY	BMP receptor signaling	3.02E-04	1.52	1.82E-05	2.22E+00
BIOCARTA_NO1_PATHWAY	Actions of Nitric Oxide in the Heart	3.22E-04	1.73	5.12E-03	2.01E+00
PID_VEGFR1_2_PATHWAY	Signaling events mediated by VEGFR1 and VEGFR2	4.23E-04	1.51	2.28E-03	1.70E+00
REACTOME_SIGNALING_BY_FGFR1_MUTANTS	Genes involved in Signaling by FGFR1 mutants	1.31E-03	1.67	3.26E-03	2.08E+00
REACTOME_SIGNALING_BY_NODAL	Genes involved in Signaling by NODAL	1.41E-03	1.93	8.76E-03	2.22E+00
PID_SYNDECAN_4_PATHWAY	Syndecan-4-mediated signaling events	3.70E-03	1.54	5.05E-03	1.98E+00

Supplementary Table 15: Summary of annotation using 5hmC gain and 5hmC loss regions, related to Figure S5D

Annotation	5hmC gain				5hmC loss			
	# of peaks	Total size (bp)	Log2 Enrichment	% of total peaks	# of peaks	Total size (bp)	Log2 Enrichment	% of total peaks
3UTR	371	19824330	1.03	1.538589143	918	19834990	0.956	1.413873829
miRNA	0	24854	-14.558	0	1	24854	0.754	0.001540168
ncRNA	72	2930845	1.422	0.298594119	90	2936051	0.362	0.138615081
TTS	439	26293235	0.865	1.8205947	839	26405695	0.413	1.292200591
pseudo	5	496629	0.135	0.020735703	10	502464	-0.262	0.015401676
Exon	1068	33394408	1.803	4.429146104	1022	33478124	0.356	1.574051257
Intron	9676	931524390	0.18	40.12773193	40429	933720660	0.86	62.2674347
Intergenic	11790	1585281725	-0.302	48.89478704	21204	1673583365	-0.913	32.65771316
Promoter	636	28567870	1.28	2.637581388	385	28680226	-0.83	0.592964515
5UTR	56	2112428	1.532	0.232239871	30	2124691	-0.757	0.046205027
snoRNA	0	19	-14.558	0	0	19	-15.987	0
snRNA	0	5631	-14.558	0	0	5631	-15.987	0

Supplementary Table 16: Motif analysis results of human 5hmC gain

Motif Name	Consensus	q-value (Benjamini)	# of Target Sequences with Motif (of 9766)	% of Target Sequences with Motif	# of Background Sequences with Motif (of 40149)	% of Background Sequences with Motif
SCL(bHLH)/HPC7-Scl-ChIP-Seq(GSE13511)/Homer	AVCAGCTG	0.0035	12162	50.44%	12775	49.38%
Nanog(Homeobox)/mES-Nanog-ChIP-Seq(GSE11724)/Homer	RGCCATTAAC	0.0372	10398	43.12%	10955.9	42.35%
Tbx5(T-box)/HL1-Tbx5.biotin-ChIP-Seq(GSE21529)/Homer	AGGTGTCA	0.0058	8564	35.52%	8940.5	34.56%
Olig2(bHLH)/Neuron-Olig2-ChIP-Seq(GSE30882)/Homer	RCCATMTGTT	0.0009	5529	22.93%	5676.1	21.94%
BMAL1(bHLH)/Liver-Bmal1-ChIP-Seq(GSE39860)/Homer	GNCACGTG	0	5134	21.29%	5120.4	19.79%
ZNF711(Zf)/SHSY5Y-ZNF711-ChIP-Seq(GSE20673)/Homer	AGGCCTAG	0.0011	4855	20.13%	4968.7	19.21%
NF1-halfsite(CTF)/LNCaP-NF1-ChIP-Seq(Unpublished)/Homer	YTGCCAAG	0	4743	19.67%	4759	18.40%
COUP-TFII(NR)/Artia-Nr2f2-ChIP-Seq(GSE46497)/Homer	AGRGGTCA	0.0004	4667	19.35%	4754	18.38%
Meis1(Homeobox)/MastCells-Meis1-ChIP-Seq(GSE48085)/Homer	VGCTGWCAVB	0.019	4589	19.03%	4748.2	18.35%
Ascl1(bHLH)/NeuralTubes-Ascl1-ChIP-Seq(GSE55840)/Homer	NNVVCAGCTGBN	0.0022	4557	18.90%	4668.8	18.05%
NeuroG2(bHLH)/Fibroblast-NeuroG2-ChIP-Seq(GSE75910)/Homer	ACCATCTGTT	0.0212	4422	18.34%	4574.5	17.68%
Sox3(HMG)/NPC-Sox3-ChIP-Seq(GSE33059)/Homer	CCWTTGTY	0	4418	18.32%	4237.9	16.38%
ZNF416(Zf)/HEK293-ZNF416.GFP-ChIP-Seq(GSE58341)/Homer	WDNCTGGGCA	0.0004	4102	17.01%	4158.9	16.08%
LRF(Zf)/Erythroblasts-ZBTB7A-ChIP-Seq(GSE74977)/Homer	AAGACCCYYN	0.0323	3934	16.31%	4068.9	15.73%
Sox10(HMG)/SciaticNerve-Sox3-ChIP-Seq(GSE35132)/Homer	CCWTTGTYYB	0	4211	17.46%	4002.4	15.47%

Sox6(HMG)/Myotubes-Sox6-ChIP-Seq(GSE32627)/Homer	CCATTGTTNY	0	3771	15.64%	3691.4	14.27%
ZFX(Zf)/mES-Zfx-ChIP-Seq(GSE11431)/Homer	AGGCCTRG	0	3685	15.28%	3568.1	13.79%
GATA3(Zf)/iTreg-Gata3-ChIP-Seq(GSE20898)/Homer	AGATAASR	0.0034	3286	13.63%	3340.8	12.91%
RXR(NR),DR1/3T3L1-RXR-ChIP-Seq(GSE13511)/Homer	TAGGGCAAAGGTCA	0.0459	3078	12.76%	3173.7	12.27%
Atoh1(bHLH)/Cerebellum-Atoh1-ChIP-Seq(GSE22111)/Homer	VNRVCAGCTGGY	0	3263	13.53%	3161.5	12.22%
AP-2gamma(AP2)/MCF7-TFAP2C-ChIP-Seq(GSE21234)/Homer	SCCTSAGGSCAW	0.0006	2672	11.08%	2672.1	10.33%
Sox15(HMG)/CPA-Sox15-ChIP-Seq(GSE62909)/Homer	RAACAATGGN	0	2583	10.71%	2349.6	9.08%
FOXM1(Forkhead)/MCF7-FOXM1-ChIP-Seq(GSE72977)/Homer	TRTTTACTTW	0.0146	2302	9.55%	2335.7	9.03%
MITF(bHLH)/MastCells-MITF-ChIP-Seq(GSE48085)/Homer	RTCATGTGAC	0	2417	10.02%	2301	8.89%
Sox9(HMG)/Limb-SOX9-ChIP-Seq(GSE73225)/Homer	AGGVNCCTTTGT	0	2418	10.03%	2246.1	8.68%
HOXB13(Homeobox)/ProstateTumor-HOXB13-ChIP-Seq(GSE56288)/Homer	TTTTATKRGG	0	2253	9.34%	2208.6	8.54%
HOXD13(Homeobox)/Chicken-Hoxd13-ChIP-Seq(GSE38910)/Homer	NCYAATAAAA	0	2304	9.56%	2172	8.40%
Gata4(Zf)/Heart-Gata4-ChIP-Seq(GSE35151)/Homer	NBWGATAAGR	0.0095	2134	8.85%	2153.6	8.32%
Six2(Homeobox)/NephronProgenitor-Six2-ChIP-Seq(GSE39837)/Homer	GWAAYHTGAKMC	0.0019	2146	8.90%	2141.5	8.28%
TEAD4(TEA)/Tropoblast-Tead4-ChIP-Seq(GSE37350)/Homer	CCWGGAATGY	0.0016	2030	8.42%	2018.4	7.80%
Sox2(HMG)/mES-Sox2-ChIP-Seq(GSE11431)/Homer	BCCATTGTTC	0	2274	9.43%	1999	7.73%
Pdx1(Homeobox)/Islet-Pdx1-ChIP-Seq(SRA008281)/Homer	YCATYAATCA	0.0041	1992	8.26%	1992.9	7.70%
AP-2alpha(AP2)/Hela-AP2alpha-ChIP-Seq(GSE31477)/Homer	ATGCCCTGAGGC	0.0066	1979	8.21%	1986.2	7.68%

Foxf1(Forkhead)/Lung-Foxf1-ChIP-Seq(GSE77951)/Homer	WWATRTAAACAN	0.0061	1963	8.14%	1968.5	7.61%
Sox4(HMG)/proB-Sox4-ChIP-Seq(GSE50066)/Homer	YCTTTGTTCC	0	2215	9.19%	1901.5	7.35%
Max(bHLH)/K562-Max-ChIP-Seq(GSE31477)/Homer	RCCACGTGGYYN	0	1925	7.98%	1858.3	7.18%
Sox17(HMG)/Endoderm-Sox17-ChIP-Seq(GSE61475)/Homer	CCATTGTTYB	0	1824	7.56%	1593.8	6.16%
ZNF264(Zf)/HEK293-ZNF264.GFP-ChIP-Seq(GSE58341)/Homer	RGGGCACTAACY	0.0034	1590	6.59%	1573.8	6.08%
Unknown-ESC-element(?)/mES-Nanog-ChIP-Seq(GSE11724)/Homer	CACAGCAGGGGG	0.0001	1614	6.69%	1554.5	6.01%
HLF(bZIP)/HSC-HLF.Flag-ChIP-Seq(GSE69817)/Homer	RTTATGYAAB	0.0058	1537	6.37%	1525.2	5.90%
Zic3(Zf)/mES-Zic3-ChIP-Seq(GSE37889)/Homer	GGCCYCCTGCTGDGH	0.0001	1497	6.21%	1442.4	5.58%
TEAD(TEA)/Fibroblast-PU.1-ChIP-Seq(Unpublished)/Homer	YCWGGAATGY	0	1583	6.56%	1432.8	5.54%
ZBTB18(Zf)/HEK293-ZBTB18.GFP-ChIP-Seq(GSE58341)/Homer	AACATCTGGA	0.0061	1422	5.90%	1407.6	5.44%
Hand2(bHLH)/Mesoderm-Hand2-ChIP-Seq(GSE61475)/Homer	TGACANARRCCAGRC	0	1473	6.11%	1347	5.21%
Unknown(Homeobox)/Limb-p300-ChIP-Seq/Homer	SSCMATWAAA	0	1515	6.28%	1310	5.06%
CLOCK(bHLH)/Liver-Clock-ChIP-Seq(GSE39860)/Homer	GHCACGTG	0	1356	5.62%	1244.3	4.81%
TEAD2(TEA)/Py2T-Tead2-ChIP-Seq(GSE55709)/Homer	CCWGGAATGY	0	1405	5.83%	1224.3	4.73%
CEBP(bZIP)/ThioMac-CEBPb-ChIP-Seq(GSE21512)/Homer	ATTGCGCAAC	0	1290	5.35%	1206.5	4.66%
THRa(NR)/C17.2-THRa-ChIP-Seq(GSE38347)/Homer	GGTCANYTGAGGWCA	0.0005	1058	4.39%	1008.5	3.90%
NF1(CTF)/LNCAP-NF1-ChIP-Seq(Unpublished)/Homer	CYTGGCABNSTGCCAR	0	1074	4.45%	970.9	3.75%
RBPJ:Ebox(?,bHLH)/Panc1-Rbpj1-ChIP-Seq(GSE47459)/Homer	GGGRAARRGRMCAGMTG	0	1041	4.32%	968.3	3.74%

Prop1(Homeobox)/GHFT1-PROP1.biotin-ChIP-Seq(GSE77302)/Homer	NTAATBNAATTA	0.0012	1006	4.17%	963.5	3.72%
Atf7(bZIP)/3T3L1-Atf7-ChIP-Seq(GSE56872)/Homer	NGRTGACGTCAY	0.0001	988	4.10%	927	3.58%
GRHL2(CP2)/HBE-GRHL2-ChIP-Seq(GSE46194)/Homer	AAACYKGTTWDACMRGTTTB	0.0091	946	3.92%	922.6	3.57%
Oct4(POU,Homeobox)/mES-Oct4-ChIP-Seq(GSE11431)/Homer	ATTTGCATAW	0	1020	4.23%	904.3	3.50%
Oct6(POU,Homeobox)/NPC-Pou3f1-ChIP-Seq(GSE35496)/Homer	WATGCAAATGAG	0	985	4.08%	867.1	3.35%
ERE(NR),IR3/MCF7-ERa-ChIP-Seq(Unpublished)/Homer	VAGGTCACNSTGACC	0.0001	761	3.16%	699.3	2.70%
Brn1(POU,Homeobox)/NPC-Brn1-ChIP-Seq(GSE35496)/Homer	TATGCWAATBAV	0	759	3.15%	622.4	2.41%
Phox2a(Homeobox)/Neuron-Phox2a-ChIP-Seq(GSE31456)/Homer	YTAATYNRATTA	0	667	2.77%	551.4	2.13%
Pit1+1bp(Homeobox)/GCrat-Pit1-ChIP-Seq(GSE58009)/Homer	ATGCATAATTCA	0	679	2.82%	545.5	2.11%
Oct2(POU,Homeobox)/Bcell-Oct2-ChIP-Seq(GSE21512)/Homer	ATATGCAAAT	0	894	3.71%	527.8	2.04%
EBF(EBF)/proBcell-EBF-ChIP-Seq(GSE21978)/Homer	DGTCCCYRGGGA	0.0459	533	2.21%	516.6	2.00%
Six1(Homeobox)/Myoblast-Six1-ChIP-Chip(GSE20150)/Homer	GKVTCADRTTWC	0	570	2.36%	490.4	1.90%
HRE(HSF)/Striatum-HSF1-ChIP-Seq(GSE38000)/Homer	TTCTAGAABNTTCTA	0.0487	498	2.07%	481.6	1.86%
Tbox:Smad(T-box,MAD)/ESCd5-Smad2_3-ChIP-Seq(GSE29422)/Homer	AGGTGHCAGACA	0	528	2.19%	464.6	1.80%
PAX3:FKHR-fusion(Paired,Homeobox)/Rh4-PAX3:FKHR-ChIP-Seq(GSE19063)/Homer	ACCRTGACTAATTNN	0	571	2.37%	448.9	1.74%
PGR(NR)/EndoStromal-PGR-ChIP-Seq(GSE69539)/Homer	AAGAACATWHTGTTC	0.0022	477	1.98%	436.5	1.69%
HRE(HSF)/HepG2-HSF1-ChIP-Seq(GSE31477)/Homer	BSTTCTRGAABVTTCYAGAA	0.0027	408	1.69%	369.9	1.43%

Tcfcp2l1(CP2)/mES-Tcfcp2l1-ChIP-Seq(GSE11431)/Homer	NRAACCRGTTYRAACCRGYT	0.0338	359	1.49%	337.1	1.30%
GATA(Zf),IR3/iTreg-Gata3-ChIP-Seq(GSE20898)/Homer	NNNNNBAGATAWYATCTVHN	0	300	1.24%	246.7	0.95%
GATA3(Zf),DR8/iTreg-Gata3-ChIP-Seq(GSE20898)/Homer	AGATSTNDNNDSAGATAASN	0.0066	235	0.97%	205.7	0.79%
Rfx2(HTH)/LoVo-RFX2-ChIP-Seq(GSE49402)/Homer	GTTGCCATGGCAACM	0.0304	218	0.90%	196.4	0.76%
ZNF669(Zf)/HEK293-ZNF669.GFP-ChIP-Seq(GSE58341)/Homer	GARTGGTCATCGCCC	0	195	0.81%	148.8	0.58%
DUX4(Homeobox)/Myoblasts-DUX4.V5-ChIP-Seq(GSE75791)/Homer	NWTAAYCYAATCAWN	0	191	0.79%	72.7	0.28%
ZNF41(Zf)/HEK293-ZNF41.GFP-ChIP-Seq(GSE58341)/Homer	CCTCATGGTGYCYTWYTCCCTTGTG	0.0001	95	0.39%	64.1	0.25%
ZNF16(Zf)/HEK293-ZNF16.GFP-ChIP-Seq(GSE58341)/Homer	MACCTTCYATGGCTCCCTAKTGCCY	0.0369	28	0.12%	18.5	0.07%
OCT:OCT(POU,Homeobox,IR1)/NPC-Brn2-ChIP-Seq(GSE35496)/Homer	ATGAATWATTCATGA	0.0338	22	0.09%	13.9	0.05%
ZFP3(Zf)/HEK293-ZFP3.GFP-ChIP-Seq(GSE58341)/Homer	GGGTTTTGAAGGATGARTAGGAGTT	0.0269	10	0.04%	4.1	0.02%

Supplementary Table 17: Summary of MSigPathway results (GREAT analysis), related to Figure 4.4G

ID	Desc	Binom FdrQ	RegionFold Enrich	Hyper FdrQ	GeneFold Enrich
KEGG_PATHWAYS_IN_CANCER	Pathways in cancer	1.59E-73	1.68	2.62E-12	1.39
KEGG_ADHERENS_JUNCTION	Adherens junction	1.11E-68	2.40	7.89E-08	1.62
KEGG_FOCAL_ADHESION	Focal adhesion	1.45E-55	1.81	1.13E-13	1.52
KEGG_BASAL_CELL_CARCINOMA	Basal cell carcinoma	7.26E-45	2.28	1.11E-03	1.47
KEGG_WNT_SIGNALING_PATHWAY	Wnt signaling pathway	1.33E-42	1.75	4.13E-04	1.31
REACTOME_G_ALPHA1213_SIGNALLING_EVENTS	Genes involved in G alpha (12/13) signalling events	2.50E-42	2.30	3.58E-03	1.36
REACTOME_NRAGE_SIGNALS_DEATH_THROUGH_JNK	Genes involved in NRAGE signals death through JNK	1.69E-41	2.52	2.36E-02	1.37
KEGG_COLORECTAL_CANCER	Colorectal cancer	2.05E-41	2.14	1.28E-05	1.57
KEGG_ACUTE_MYELOID_LEUKEMIA	Acute myeloid leukemia	5.67E-38	2.29	2.99E-04	1.51
WNT_SIGNALING	Genes related to Wnt-mediated signal transduction	7.28E-36	1.95	1.86E-04	1.42
PID_DELTANP63PATHWAY	Validated transcriptional targets of deltaNp63 isoforms	3.18E-32	2.31	3.20E-04	1.56
PID_NECTIN_PATHWAY	Nectin adhesion pathway	1.60E-30	2.49	9.38E-05	1.72
PID_ILK_PATHWAY	Integrin-linked kinase signaling	3.29E-30	2.52	1.91E-03	1.50
KEGG_HEDGEHOG_SIGNALING_PATHWAY	Hedgehog signaling pathway	1.56E-28	2.04	1.25E-02	1.37
KEGG_PANCREATIC_CANCER	Pancreatic cancer	7.77E-28	2.02	4.96E-03	1.36
PID_MYC_REPRESSPATHWAY	Validated targets of C-MYC transcriptional repression	1.16E-27	1.99	5.80E-03	1.38
REACTOME_SIGNALING_BY_RHO_GTPASES	Genes involved in Signaling by Rho GTPases	1.74E-25	1.73	2.98E-03	1.30
PID_HIF1_TFPATHWAY	HIF-1-alpha transcription factor network	2.83E-25	1.92	3.93E-07	1.64
PID_CMYB_PATHWAY	C-MYB transcription factor network	2.99E-25	1.74	5.69E-04	1.41
KEGG_MELANOGENESIS	Melanogenesis	7.74E-24	1.68	1.02E-04	1.42
KEGG_PROSTATE_CANCER	Prostate cancer	1.26E-23	1.78	3.11E-04	1.41
PID_ECADHERIN_NASCENTAJ_PATHWAY	E-cadherin signaling in the nascent adherens junction	3.05E-23	2.31	6.86E-03	1.47

KEGG_LEUKOCYTE_TRANSENDOTHELIAL_MIGRATION	Leukocyte transendothelial migration	1.24E-22	1.70	8.67E-05	1.39
REACTOME_INTERFERON_GAMMA_SIGNALING	Genes involved in Interferon gamma signaling	9.60E-22	2.37	2.83E-04	1.58
KEGG_ENDOMETRIAL_CANCER	Endometrial cancer	1.29E-20	1.89	1.46E-05	1.62
REACTOME_INTERFERON_SIGNALING	Genes involved in Interferon Signaling	2.35E-20	1.81	1.89E-03	1.29
PID_VEGFR1_2_PATHWAY	Signaling events mediated by VEGFR1 and VEGFR2	2.39E-20	1.89	1.87E-04	1.47
KEGG_VEGF_SIGNALING_PATHWAY	VEGF signaling pathway	4.37E-20	1.96	2.51E-05	1.51
BIOCARTA_TGFB_PATHWAY	TGF beta signaling pathway	1.64E-18	2.30	1.65E-03	1.75
REACTOME_INTEGRIN_CELL_SURFACE_INTERACTIONS	Genes involved in Integrin cell surface interactions	1.02E-17	1.79	1.84E-04	1.44
KEGG_P53_SIGNALING_PATHWAY	p53 signaling pathway	2.91E-17	1.77	2.35E-02	1.31
PID_ATF2_PATHWAY	ATF-2 transcription factor network	6.32E-17	1.83	3.96E-02	1.29
PID_BCR_5PATHWAY	BCR signaling pathway	6.66E-17	1.79	2.49E-03	1.40
PID_RAC1_REG_PATHWAY	Regulation of RAC1 activity	6.88E-17	1.93	8.95E-03	1.46
KEGG_CHEMOKINE_SIGNALING_PATHWAY	Chemokine signaling pathway	2.79E-16	1.53	7.49E-03	1.22
KEGG_B_CELL_RECEPTOR_SIGNALING_PATHWAY	B cell receptor signaling pathway	5.65E-16	1.79	1.29E-03	1.40
PID_PI3KCIPATHWAY	Class I PI3K signaling events	6.73E-16	2.00	3.44E-04	1.54
BIOCARTA_EGFR_SMRTE_PATHWAY	Map Kinase Inactivation of SMRT Corepressor	9.18E-16	2.73	4.55E-02	1.68
PID_SMAD2_3NUCLEARPATHWAY	Regulation of nuclear SMAD2/3 signaling	2.06E-15	1.59	2.07E-02	1.27
PID_AP1_PATHWAY	AP-1 transcription factor network	2.81E-15	1.65	9.74E-03	1.34
PID_RHOA_REG_PATHWAY	Regulation of RhoA activity	4.30E-15	1.94	1.89E-02	1.38
KEGG_NON_SMALL_CELL_LUNG_CANCER	Non-small cell lung cancer	5.42E-15	1.82	2.00E-02	1.35
KEGG_GLIOMA	Glioma	5.41E-15	1.78	3.17E-04	1.48
BIOCARTA_ALK_PATHWAY	ALK in cardiac myocytes	7.72E-15	1.78	2.48E-05	1.69
KEGG_INSULIN_SIGNALING_PATHWAY	Insulin signaling pathway	1.27E-14	1.54	2.26E-05	1.38
REACTOME_COLLAGEN_FORMATION	Genes involved in Collagen formation	1.70E-14	1.70	2.96E-05	1.58
PID_INTEGRIN1_PATHWAY	Beta1 integrin cell surface interactions	1.70E-14	1.65	5.22E-03	1.37
KEGG_ADIPOCYTOKINE_SIGNALING_PATHWAY	Adipocytokine signaling pathway	2.25E-14	1.79	4.27E-02	1.27

PID_HEDGEHOG_2PATHWAY	Signaling events mediated by the Hedgehog family	3.20E-14	1.99	2.93E-02	1.51
PID_PS1PATHWAY	Presenilin action in Notch and Wnt signaling	4.22E-14	1.75	2.67E-03	1.47
BIOCARTA_IL6_PATHWAY	IL 6 signaling pathway	4.70E-14	2.69	1.98E-02	1.57
REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION	Genes involved in Extracellular matrix organization	5.77E-14	1.61	1.85E-04	1.43
PID_AJDISS_2PATHWAY	Posttranslational regulation of adherens junction stability and dissassembly	7.35E-14	1.71	6.82E-04	1.52
PID_IFNGPATHWAY	IFN-gamma pathway	2.06E-13	1.87	1.65E-04	1.61
PID_GLYPICAN_1PATHWAY	Glypican 1 network	4.88E-13	1.86	8.11E-03	1.56
REACTOME_CLASS_B_2 _SECRETIN_FAMILY_RECEPTORS	Genes involved in Class B/2 (Secretin family receptors)	4.88E-13	1.63	1.58E-02	1.28
PID_MET_PATHWAY	Signaling events mediated by Hepatocyte Growth Factor Receptor (c-Met)	5.25E-13	1.62	1.43E-04	1.45
ST_INTEGRIN_SIGNALING_PATHWAY	Integrin Signaling Pathway	1.09E-12	1.59	2.07E-02	1.27
PID_CDC42_PATHWAY	CDC42 signaling events	2.76E-12	1.65	1.39E-03	1.41
KEGG_SMALL_CELL_LUNG_CANCER	Small cell lung cancer	2.90E-12	1.54	2.13E-05	1.48
PID_WNT_SIGNALING_PATHWAY	Wnt signaling network	4.44E-12	1.90	1.18E-02	1.51
KEGG_ECM_RECEPTOR_INTERACTION	ECM-receptor interaction	5.18E-12	1.51	1.03E-04	1.44
PID_BMPPATHWAY	BMP receptor signaling	1.12E-11	1.65	6.89E-03	1.45
BIOCARTA_WNT_PATHWAY	WNT Signaling Pathway	1.12E-11	1.82	6.97E-04	1.70
REACTOME_REGULATION _OF_IFNG_SIGNALING	Genes involved in Regulation of IFNG signaling	1.41E-11	2.66	2.02E-02	1.70
SIG_PIP3_SIGNALING_IN_B_LYMPHOCYTES	Genes related to PIP3 signaling in B lymphocytes	1.81E-11	1.80	2.06E-02	1.42
REACTOME_NCAM_SIGNALING _FOR_NEURITE_OUT_GROWTH	Genes involved in NCAM signaling for neurite out-growth	2.14E-11	1.60	1.12E-03	1.43
KEGG_CHRONIC_MYELOID_LEUKEMIA	Chronic myeloid leukemia	2.63E-11	1.55	1.34E-03	1.40
PID_S1P_S1P3_PATHWAY	S1P3 pathway	1.01E-10	1.84	1.93E-05	1.78
PID_GMCSF_PATHWAY	GMCSF-mediated signaling events	1.08E-10	1.95	8.99E-03	1.47
KEGG_DORSO_VENTRAL_AXIS_FORMATION	Dorso-ventral axis formation	1.30E-10	2.09	2.54E-03	1.68

KEGG_FC_GAMMA_R_MEDIATED_PHAGOCYTOSIS	Fc gamma R-mediated phagocytosis	1.67E-10	1.54	4.63E-04	1.38
SIG_BCR_SIGNALING_PATHWAY	Members of the BCR signaling pathway	3.03E-10	1.71	8.62E-03	1.41
BIOCARTA_INTEGRIN_PATHWAY	Integrin Signaling Pathway	3.47E-10	1.84	1.68E-03	1.54
PID_INTEGRIN_A9B1_PATHWAY	Alpha9 beta1 integrin signaling events	5.38E-10	1.86	2.92E-02	1.47
ST_STAT3_PATHWAY	STAT3 Pathway	6.19E-10	2.77	1.96E-02	1.84
REACTOME_GLYCEROPHOSPHOLIPID _BIOSYNTHESIS	Genes involved in Glycerophospholipid biosynthesis	9.26E-10	1.60	3.45E-02	1.25
KEGG_LYSOSOME	Lysosome	1.12E-09	1.61	4.80E-02	1.20
REACTOME_NCAM1_INTERACTIONS	Genes involved in NCAM1 interactions	1.59E-09	1.63	2.50E-03	1.51
REACTOME_INTEGRIN_ALPHAIIB _BETA3_SIGNALING	Genes involved in Integrin alphaIIb beta3 signaling	1.86E-09	2.04	3.99E-02	1.43
PID_ECADHERIN_STABILIZATION_PATHWAY	Stabilization and expansion of the E-cadherin adherens junction	2.55E-09	1.71	1.82E-03	1.52
ST_JAK_STAT_PATHWAY	Jak-STAT Pathway	2.98E-09	2.79	3.07E-02	1.84
PID_TAP63PATHWAY	Validated transcriptional targets of TAp63 isoforms	4.09E-09	1.65	4.96E-03	1.42
PID_TCPTP_PATHWAY	Signaling events mediated by TCPTP	5.60E-09	1.65	8.02E-05	1.62
BIOCARTA_NO1_PATHWAY	Actions of Nitric Oxide in the Heart	6.74E-09	1.79	8.54E-03	1.53
BIOCARTA_CELL2CELL_PATHWAY	Cell to Cell Adhesion Signaling	6.88E-09	2.13	1.91E-03	1.84
REACTOME_PRESYNAPTIC_NICOTINIC _ACETYLCHOLINE_RECEPTORS	Genes involved in Presynaptic nicotinic acetylcholine receptors	7.47E-09	3.78	2.99E-02	1.69
BIOCARTA_NTHI_PATHWAY	NFkB activation by Nontypeable Hemophilus influenzae	1.00E-08	1.84	1.51E-02	1.54
KEGG_FC_EPSILON_RI_SIGNALING_PATHWAY	Fc epsilon RI signaling pathway	1.56E-08	1.52	3.24E-04	1.43
REACTOME_FACILITATIVE_NA_ INDEPENDENT_GLUCOSE_TRANSPORTERS	Genes involved in Facilitative Na+-independent glucose transporters	2.13E-08	2.71	8.24E-03	1.84
BIOCARTA_MAL_PATHWAY	Role of MAL in Rho-Mediated Activation of SRF	7.02E-08	2.44	1.73E-02	1.63
ST_B_CELL_ANTIGEN_RECEPTOR	B Cell Antigen Receptor	8.16E-08	1.79	3.56E-03	1.50
REACTOME_SIGNALLING_TO_RAS	Genes involved in Signalling to RAS	1.13E-07	2.18	2.12E-02	1.49

PID_PTP1BPATHWAY	Signaling events mediated by PTP1B	1.15E-07	1.57	2.55E-04	1.54
PID_SYNDECAN_1_PATHWAY	Syndecan-1-mediated signaling events	1.55E-07	1.55	1.43E-02	1.39
PID_P38GAMMADELTAPATHWAY	Signaling mediated by p38-gamma and p38-delta	1.92E-07	2.59	1.25E-02	1.84
PID_AMB2_NEUTROPHILS_PATHWAY	amb2 Integrin signaling	1.93E-07	1.75	4.96E-02	1.34
PID_PDGFRAPATHWAY	PDGFR-alpha signaling pathway	2.21E-07	1.78	2.93E-02	1.51
BIOCARTA_SHH_PATHWAY	Sonic Hedgehog (Shh) Pathway	2.73E-07	1.85	8.81E-03	1.72
REACTOME_PLATELET_AGGREGATION _PLUG_FORMATION	Genes involved in Platelet Aggregation (Plug Formation)	3.41E-07	1.72	3.47E-02	1.38
KEGG_VIBRIO_CHOLERAE_INFECTION	Vibrio cholerae infection	4.02E-07	1.61	4.29E-02	1.30
KEGG_TYPE_II_DIABETES_MELLITUS	Type II diabetes mellitus	4.12E-07	1.55	1.01E-03	1.52
REACTOME_FRS2_MEDIATED_CASCADE	Genes involved in FRS2-mediated cascade	4.52E-07	1.68	2.80E-02	1.41
PID_AVB3_OPN_PATHWAY	Osteopontin-mediated events	7.07E-07	1.66	2.87E-02	1.43
BIOCARTA_CDK5_PATHWAY	Phosphorylation of MEK1 by cdk5/p35 down regulates the MAP kinase pathway	7.24E-07	2.67	1.25E-02	1.84
ST_WNT_BETA_CATENIN_PATHWAY	Wnt/beta-catenin Pathway	7.31E-07	1.59	3.64E-02	1.40
BIOCARTA_TCR_PATHWAY	T Cell Receptor Signaling Pathway	7.80E-07	1.53	1.17E-02	1.41
PID_ECADHERIN_KERATINOCYTE_PATHWAY	E-cadherin signaling in keratinocytes	1.48E-06	1.78	1.98E-02	1.57
BIOCARTA_GSK3_PATHWAY	Inactivation of Gsk3 by AKT causes accumulation of b-catenin in Alveolar Macrophages	1.65E-06	1.56	1.59E-02	1.50
REACTOME_REGULATION_OF_ INSULIN_SECRETION_ BY_GLUCAGON_LIKE_PEPTIDE1	Genes involved in Regulation of Insulin Secretion by Glucagon-like Peptide-1	2.74E-06	1.61	3.96E-02	1.35
PID_CERAMIDE_PATHWAY	Ceramide signaling pathway	2.92E-06	1.52	1.07E-03	1.50
BIOCARTA_MEF2D_PATHWAY	Role of MEF2D in T-cell Apoptosis	3.46E-06	1.79	1.21E-02	1.64
REACTOME_GPVI_MEDIATED_ACTIVATION_CASCADE	Genes involved in GPVI-mediated activation cascade	4.77E-06	1.60	1.27E-03	1.61
PID_TXA2PATHWAY	Thromboxane A2 receptor signaling	5.97E-06	1.51	2.64E-03	1.42

REACTOME_AQUAPORIN_MEDIATED_TRANSPORT	Genes involved in Aquaporin-mediated transport	1.62E-05	1.51	1.37E-02	1.38
SIG_REGULATION _OF_THE_ACTIN_CYTOSKELETON_BY_RHO_GTPASES	Genes related to regulation of the actin cytoskeleton	2.81E-05	1.63	8.99E-03	1.47
BIOCARTA_GATA3_PATHWAY	GATA3 participate in activating the Th2 cytokine genes expression	2.93E-05	1.72	3.67E-02	1.60
KEGG_HISTIDINE_METABOLISM	Histidine metabolism	3.09E-05	1.89	2.90E-02	1.45
KEGG_GLYCOSAMINOGLYCAN_BIOSYNTHESIS _CHONDROITIN_SULFATE	Glycosaminoglycan biosynthesis - chondroitin sulfate	3.34E-05	1.53	2.93E-02	1.51
PID_AURORA_A_PATHWAY	Aurora A signaling	4.25E-05	1.79	2.65E-03	1.59
BIOCARTA_PYK2_PATHWAY	Links between Pyk2 and Map Kinases	4.72E-05	1.62	3.99E-02	1.43
SIG_IL4RECEPTOR_IN_B_LYPHOCYTES	Genes related to IL4 rceptor signaling in B lymphocytes	5.39E-05	1.60	8.11E-03	1.56
REACTOME_AMINE_COMPOUND _SLC_TRANSPORTERS	Genes involved in Amine compound SLC transporters	6.01E-05	1.56	8.11E-03	1.56
REACTOME_ACETYLCHOLINE_BINDING _AND_DOWNSTREAM_EVENTS	Genes involved in Acetylcholine Binding And Downstream Events	1.25E-04	2.08	1.36E-02	1.71
REACTOME_AMINO_ACID_TRANSPORT _ACROSS_THE_PLASMA_MEMBRANE	Genes involved in Amino acid transport across the plasma membrane	1.93E-04	1.51	1.59E-02	1.47
PID_INTEGRIN_CS_PATHWAY	Integrin family cell surface interactions	2.38E-04	1.57	2.12E-02	1.49
REACTOME_REGULATION_OF_KIT_SIGNALING	Genes involved in Regulation of KIT signaling	2.40E-04	1.63	3.67E-02	1.60
REACTOME_HIGHLY_CALCIUM_PERMEABLE _POSTSYNAPTIC_NICOTINIC _ACETYLCHOLINE_RECEPTORS	Genes involved in Highly calcium permeable postsynaptic nicotinic acetylcholine receptors	3.33E-04	2.02	8.24E-03	1.84
REACTOME_BASIGIN_INTERACTIONS	Genes involved in Basigin interactions	7.10E-04	1.61	7.03E-03	1.60
BIOCARTA_IL22BP_PATHWAY	IL22 Soluble Receptor Signaling Pathway	2.05E-03	1.93	2.48E-02	1.61