CATARINA MORAIS SEABRA

DISCOVERING PATHWAYS UNDERLYING AUTISM SPECTRUM DISORDER UPON LOSS-OF-FUNCTION OF CHROMATIN-RELATED GENES

Tese de Candidatura ao grau de Doutor em Biologia Básica e Aplicada submetida ao Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto

Orientador – Doutor James F. Gusella
Categoria – Professor Catedrático
Afiliação – Center for Genomic Medicine, Harvard Medical School, Boston, EUA

Orientador – Doutor Michael E. Talkowski
Categoria – Professor Associado
Afiliação – Center for Genomic Medicine, Harvard Medical School, Boston, EUA

Coorientador – Doutora Patrícia E. Sá Maciel
Categoria – Professora Associada
Afiliação – Instituto de Investigação em Ciências da Vida e Saúde, Universidade do Minho

Coorientador – Doutora Ana Xavier Carvalho
Categoria – Group Leader
Afiliação – Instituto de Investigação e Inovação em Saúde, Universidade do Porto

This work was conducted at the Center for Human Genetic Research at the Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA.

Catarina Seabra was supported by a fellowship (SFRH/BD/52049/2012) awarded by the Doctoral Program in Basic and Applied Biology (GABBA) of the University of Porto and funded by the Portuguese Foundation of Science and Technology (FCT). This work was funded by grants from Autism Speaks and SFARI.

For my Father, in his loving memory.
For my Mother and my Sister.

"Everything is possible. The impossible just takes longer."
— Dan Brown

Acknowledgements

First and foremost, I would like to thank Jim Gusella, for accepting me in his lab and for being an inspiring role model as a scientist, showing me that there is always room for curiosity.

A special thanks to my supervisor Michael Talkowski, for always pushing and encouraging me even when things seemed tough, and for making the lab such a nice group to be a part of. I will never forget your positivity and your ambition.

I would like to thank my mentor Patrícia Maciel who introduced me to Jim and made all this possible and also for her thoughtful and critical insight throughout the PhD.

A special thanks to Ana Xavier Carvalho for being such a caring mentor and for always supporting me in finding a way.

A special acknowledgement to my GABBA family. First of all, to the GABBA scientific committee for selecting me to be a part of this family and giving me this opportunity, and also to GABBA16 for all the support, throughout this period. For the laughs and also the 'tears' when difficult times came up – Joana Nabais, Iliona Wolfowicz, Vera Lemos, Mafalda Azevedo, Raquel Sousa, Ana Luísa Neves, Paula Voabil, Lara Cravo, Ricardo Sousa, André Sousa.

To my true lab mentors – Poornima Manavalan, Derek Tai, Celine de Esch, Claire Redin – thank you for your patience, for the hours you had to spend with me, I will be forever grateful and will promise to continue to apply and pass on the knowledge you taught me. In special, a thank you to Poornima, my "lab wife", for the making countless hours in the lab such fun and rewarding. Thank you to all the Talkowski lab members and in special to: Yu An, for her support in the protein work; Serkan Erdin and Ashok Ragavendran, for teaching me the basics of bioinformatics and for helping me throughout this process; Tatsiana Aneichyk for believing in this project and jumping in in the final step and making such a difference, I will be truly grateful to you. Also, thanks to Catarina Silva and Uma Chandrachud for all your help with the cell culture work.

To João Neto, for always being there in the best and worst times during the PhD. To my Bostonian family, in special to Ana Tellechea, Teresa Capela, Nelma Gomes, Joana Guedes for making the overseas stay such an enjoyable and unforgettable time! To my PAPS mates Ângela Crespo and João Ribas, special

thanks for believing in me and for all the moments shared with the Portuguese-American community in Boston.

To my amazing friends Marisa Oliveira, Inês Martins, Joana Tavares, Cátia Amaral, special thanks for always being there whenever I needed some friendly advice.

Quero também fazer um agradecimento à minha família, em especial aos meus Pais e irmã, por terem sempre apoiado as minhas decisões, e de terem sempre acreditado em mim. Sem vocês e sem o vosso apoio nada disto teria sido possível e, por isso, estou eternamente grata.

Table of Contents

Summary ..... **xi**

Resumo xiii

List of Publications xv

Abbreviations xvii

General Introduction **19**

Autism Spectrum Disorder 21

Genetics of Autism Spectrum Disorder 25

Chromatin and Neurodevelopment 33

Goals 41

Chapter 1 – A Novel Microduplication of *ARID1B*: Clinical, Genetic and Proteomic Findings **55**

Authors 55

Abstract 59

Introduction 61

Clinical Report 63

Materials and Methods 65

Results 69

Discussion 73

Acknowledgements 75

Supplementary Data 77

References 81

Chapter 2 – CRISPR-edited iPSC models of neurodevelopment 83

Authors 85

Abstract 87

Introduction 89

Methods 93

Results 99

Discussion and Conclusions 103

Acknowledgements 105

Supplementary Data 107

References 109

Chapter 3 – Disruption of chromatin remodeler *MBD5* results in dysregulated neuronal-related genes and pathways 113

Authors 115

Abstract 117

Introduction 119

Methods 125

Results 131

Discussion and Conclusions 153

Supplementary Data 167

References 189

Final Remarks 195

Summary of Findings 197

Limitations 200

Conclusions 202

Additional Preliminary Results 205

References 211

Summary

Autism spectrum disorder (ASD) is characterized by persistent deficits in social communication and social interaction. ASD may be caused by an array of different genes and variants that represent risk for ASD. One of these gene classes is chromatin regulators that have shown to be essential for brain development; however, the impact of these genes remains unexplored. Therefore, the main goal of this project was to functionally explore chromatin remodelers that have been previously identified as strong risk factors for ASD and propose a novel system to analyze loss-of-function (LoF) mutations of the genes encoding these remodelers, in the tissue of interest – neuronal progenitors and mature neurons.

We start by reporting a unique case of an intragenic microduplication in the chromatin remodeler, *ARID1B,* in a patient with intellectual disability and show that this caused haploinsufficiency for the encoded protein, adding further evidence that this is a dosage sensitive gene. We also performed proteomic analyses that indicated an enrichment of transcription and cell cycle regulation pathways in this patient. However, patient blood-derived lymphoblasts may not be the best readout to study brain-related phenotypes caused by LoF of the chromatin-remodeling genes.

To overcome this, we propose an improved model to study ASD by using CRISPR-edited induced pluripotent stem cells (iPSC), as we are able to drive the differentiation of these cells toward our tissue of interest. We created an allelic series of isogenic mutants for chromatin remodeling genes (*EHMT1*, *MBD5*, *METTL2A* and *METTL2B*) using CRISPR/Cas9 as a genome editing strategy. CRISPR/Cas9 editing was highly precise, and deletions were created within all targeted genes, regardless of deletion size.

We then show an example of an application of our iPSC-derived cellular models, driving these into the neuronal lineage, to gain insights into the functional role of two distinct regions of the chromatin-remodeling gene *MBD5* during neurodevelopment. The removal of those regions of *MBD5* revealed a short alternative non-coding *MBD5* transcript with the highest expression in neuronal tissue, indicating a promising transcript that may be implicated in neurodevelopment and disease. Genome-wide transcriptomic analysis via RNAseq allowed the identification of the dysregulated genes upon CRISPR editing such as *RAB11FIP1*, *NHLH1-2*, *PLAUR* and *CNTNAP2*; and pathways such as notch signaling

and cell adhesion that gave insight on the protein complexes and pathways that are acting downstream of *MBD5.*

In this study, we added further insight into the functional roles of *ARID1B* and *MBD5* in neurodevelopment and proposed LoF models to study the impact of mutations in ASD-related genes in early neurodevelopment. We identified novel players (genes and pathways) that may be directly implicated in neuronal development and function through the disruption of chromatin remodelers. The analysis of many other chromatin remodeling genes through this approach will allow testing of the hypothesis that the functional consequences of ASD gene defects converge and may identify new avenues of research to ultimately develop potential therapies for ASD and other neurodevelopmental disorders.

Keywords

ASD, CRISPR/Cas9, gene, editing, chromatin, *in vitro* models, haploinsufficiency, *ARID1B*, *MBD5*, neurodevelopment.

Resumo

As perturbações do espetro do autismo (PEA) são caracterizadas por défices persistentes na comunicação e na interação social. As PEA podem ser causadas por um conjunto de diferentes genes e variantes que representam risco para as PEA. Uma dessas classes de genes são os reguladores de cromatina que se mostraram essenciais para o desenvolvimento cerebral, no entanto, o impacto desses genes permanece por explorar. Assim, o objetivo principal deste projeto foi explorar funcionalmente os remodeladores de cromatina, cujas mutações foram previamente identificadas como fortes fatores de risco para as PEA, e propor um novo sistema para analisar mutações de perda de função no tecido de interesse – os progenitores neuronais e neurónios maduros.

Esta tese inicia-se com o relato de um caso único de uma microduplicação intragénica no gene *ARID1B,* num paciente com défice intelectual, em que se demonstra que esta resultou em haploinsuficiência da proteína codificada, adicionando mais evidências de que este é um gene sensível à dosagem. Realizámos também análises proteómicas que demonstram uma perturbação das vias de regulação da transcrição e do ciclo celular neste paciente. No entanto, os linfócitos obtidos a partir do sangue do paciente não são o melhor modelo para estudar fenótipos relacionados com o cérebro causados por perda de função dos genes de remodelação da cromatina.

Para ultrapassar essas desvantagens, propomos um modelo melhorado para estudar as PEA, usando células estaminais induzidas (iPSC), editadas através do sistema de CRISPR/Cas9, uma vez que é depois possível conduzir a diferenciação destas células de modo a transformarem-se no nosso tecido de interesse. Criámos uma série alélica de mutantes isogénicos para genes de remodelação da cromatina (*EHMT1*, *MBD5*, *METTL2A* e *METTL2B*) usando CRISPR/Cas9 como estratégia de edição do genoma e mostramos que as alterações induzidas pelo sistema CRISPR/Cas9 são altamente precisas, tendo sido criadas deleções em todos os gene-alvo, independentemente do tamanho da deleção.

Em seguida, finalizamos mostrando um exemplo de uma aplicação de nossos modelos celulares, derivados das iPSC, transformando-as na linhagem neuronal, para obter evidências acerca do papel funcional de duas regiões distintas do gene de remodelação da cromatina, *MBD5*, durante o neurodesenvolvimento. A remoção dessas regiões do gene *MBD5* revelou um transcrito alternativo que não codifica proteína (lncRNA) que tinha a maior expressão no tecido neuronal. Assim,

o gene *MBD5* poderá ter uma função no neurodesenvolvimento através da ação do lncRNA, regulando-se a si mesmo ou a outros genes envolvidos no processo. A análise transcriptómica de todo o genoma através de RNAseq permitiu a identificação de genes desregulados após a edição com CRISPR/Cas9, nomeadamente *RAB11FIP1*, *NHLH1-2*, *PLAUR* e *CNTNAP2*; e também vias como a de sinalização de Notch e de adesão celular, que vieram contribuir para uma melhor compreensão dos complexos de proteínas e vias que estão a atuar a jusante de *MBD5.*

Neste estudo, apresentámos mais evidências das funções de *ARID1B* e *MBD5* no neurodesenvolvimento e propusemos modelos de perda de função de genes importantes para as PEA para estudar o impacto destes genes ao longo do neurodesenvolvimento. Identificámos novos genes e vias que podem estar diretamente implicados no desenvolvimento neuronal e sua função, através da disrupção dos remodeladores de cromatina. A análise de muitos outros genes de remodelação da cromatina através desta abordagem permitirá testar a hipótese de convergência de vias biológicas e identificar possíveis alvos terapêuticos para as PEA e outras doenças do neurodesenvolvimento.

Palavras-chave

PEA, CRISPR/Cas9, gene, edição, cromatina, modelos *in vitro,* haploinsuficiência, *ARID1B*, *MBD5*, neurodesenvolvimento.

List of Publications

Publication included in this thesis:

Seabra, C.M., Szoko, N., Erdin, S., Ragavendran, A., Stortchevoi, A., Maciel, P., Lundberg, K., Schlatzer, D., Smith, J., Talkowski, M.E., Gusella, J.F. and Natowicz, M.R., 2017. *A Novel Microduplication of ARID1B: Clinical, Genetic and Proteomic Findings.* **Am J Med Gen Part A**.; 9999:1-7. doi:10.1002/ajmg.a.38327

Manuscript in preparation:

Seabra, C.M.*, Aneychyk, T.*, Razaz, P., Erdin, S., Tai, D.J.C., De Esch, C., Manavalan, P., An, Y., Ragavendran, A., Stortchevoi, A., Talkowski, M.E and Gusella, J.F. *The role of MBD5 haploinsufficiency in autism spectrum disorder: Evidences from CRISPR-derived neurons and Mbd5$^{GT/+}$mouse model.*

Other manuscripts to which the candidate contributed during the doctoral work:

Collins, R.L., Brand, H., Redin, C.E., Hanscom, C., Antolik, C., Stone, M.R., Glessner, J.T., Mason, T., Pregno, G., Dorrani, N., Mandrile, G., Giachino, D., Perrin, D., Walsh, C., Cipicchio, M., Costello, M., Stortchevoi, A., An, J.-Y., Currall, B.B., Seabra, C.M., […], Talkowski, M.E., 2017. *Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome*. **Genome Biol**. 18, 36. doi:10.1186/s13059-017-1158-6

Redin, C., Brand, H., Collins, R.L., Kammin, T., Mitchell, E., Hodge, J.C., Hanscom, C., Pillalamarri, V., Seabra, C.M., […], Morton, C.C., Gusella, J.F., Talkowski, M.E., 2016. *The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies*. **Nat. Genet.** 49, 36–45. doi:10.1038/ng.3720

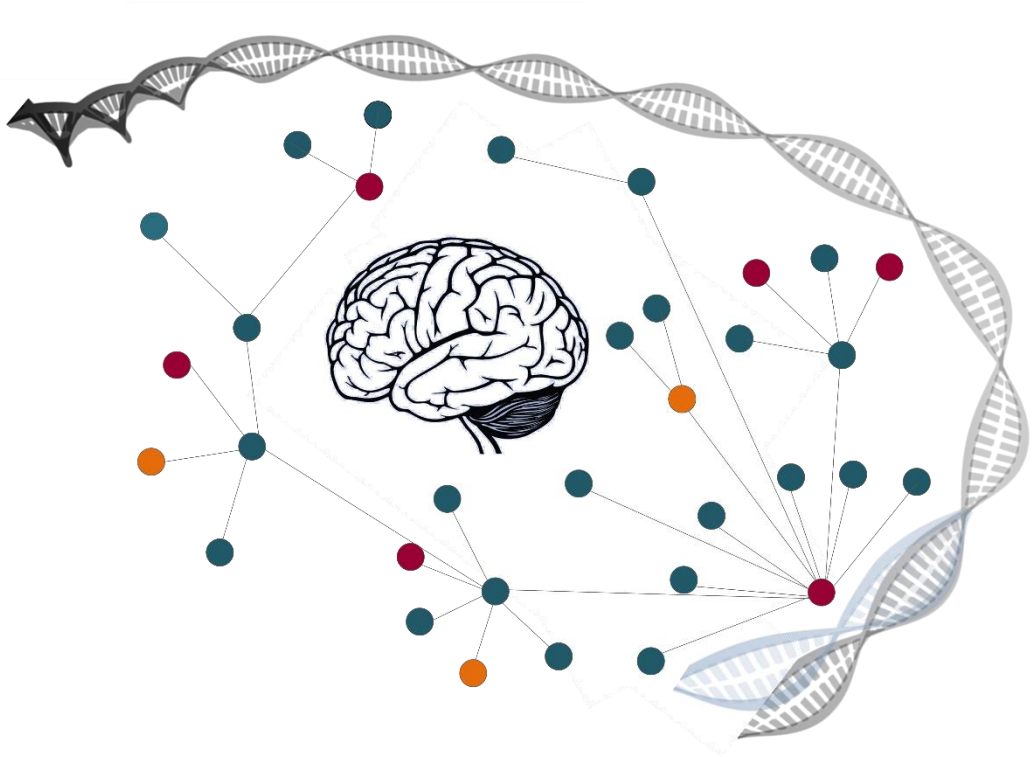
Shaw, N.D., Brand, H., Kupchinsky, Z.A., Bengani, H., Plummer, L., Jones, T.I., Erdin, S., Williamson, K.A., Rainger, J., Stortchevoi, A., Samocha, K., Currall, B.B., Dunican, D.S., Collins, R.L., Willer, J.R., Lek, A., Lek, M., Nassan, M., Pereira, S., Kammin, T., Lucente, D., Silva, A., Seabra, C.M., […] Talkowski, M.E., 2017. *SMCHD1 mutations associated with a rare muscular dystrophy can also cause isolated arhinia and Bosma arhinia microphthalmia syndrome.* **Nat. Genet.** doi:10.1038/ng.3743

Tai, D.J.C., Ragavendran, A., Manavalan, P., Stortchevoi, A., Seabra, C.M., Erdin, S., Collins, R.L., Blumenthal, I., […], Gusella, J.F., Talkowski, M.E., 2016. *Engineering microdeletions and microduplications by targeting segmental duplications with CRISPR*. **Nat. Neurosci.** 19, 517–522. doi:10.1038/nn.4235

Abbreviations

aCGH	Array comparative genomic hybridization
BCA	Balanced chromosomal anomaly
CRISPR	Clustered regularly interspaced palindromic repeats
DEG	Differentially expressed genes
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
EB	Embryoid body
FISH	Fluorescence in situ hybridization
GO	Gene ontology
GSEA	Gene set enrichment analysis
GWAS	Genome-wide association studies
HDAC	Histone deacetylase
HDR	Homology derived repair
HEK	Human embryonic kidney (cells)
HMT	Histone methyltransferase
Indel	Insertion/deletion
iPSC	Induced pluripotent stem cells
KEGG	Kyoto encyclopedia of genes and genomes
lncRNA	Long non-coding ribonucleic acid
LoF	Loss of function
mRNA	Messenger ribonucleic acid
NHEJ	Non-homologous end joining
NPC	Neuronal progenitor cells
qRT-PCR	Quantitative reverse transcription polymerase chain reaction
PCA	Principle component analysis
RNA	Ribonucleic acid
RNAseq	RNA sequencing or whole transcriptome sequencing
ROCK	Rho-associated protein kinase
UTR	Untranslated region
WES	Whole exome sequencing

General Introduction



Highlights

This section summarizes the state-of-the-art work published in the field of autism spectrum disorder, regarding its characterization and genetic etiology, as well as the role of chromatin remodeling and transcriptional regulation in neurodevelopment.

Autism Spectrum Disorder

Clinical Phenotype

Congenital neurodevelopmental disorders are a group of conditions which typically manifest early in development and are characterized by developmental deficits that produce impairments of personal, social, academic, or occupational functioning. Autism spectrum disorder (ASD) is characterized by persistent deficits in social communication and social interaction across multiple contexts, including deficits in social reciprocity, nonverbal communicative behaviors used for social interaction, and skills in developing, maintaining, and understanding relationships. In addition to the social communication deficits, the diagnosis of ASD requires the presence of restricted, repetitive patterns of behavior, interests, or activities Table I. For many individuals the diagnosis of ASD is accompanied by intellectual impairment and/or language impairment (e.g., slow to talk, language comprehension behind production) (American Psychiatric Association 2013). Other common pathological disturbances include gait and motor disturbances, anxiety, epilepsy, sensorial abnormalities, sleep disturbances and comorbidity with psychiatric disorders such as attention deficit hyperactivity disorder, obsessive-compulsive disorder (OCD) and mood disorders (Geschwind 2009).

Prevalence

In recent years, reported frequencies for ASD across U.S. and non- U.S. countries have approached 1% of the population and currently it is estimated that 1 out of 88 children has an ASD in the U.S., representing a 78% increase over the past years (Berg & Geschwind 2012; ADDM Network 2012; American Psychiatric Association 2013). It remains unclear whether higher rates reflect an expansion of the diagnostic criteria of DSM-IV to include subthreshold cases (such as pervasive developmental disorder - not otherwise specified, autistic disorder and Asperger syndrome), increased awareness, differences in study methodology, or a true increase in the frequency of ASD (American Psychiatric Association 2013). In terms of gender incidence, boys are diagnosed four times more often than girls (Abrahams & Geschwind 2008). Females tend to be more likely to show accompanying intellectual disability, suggesting that girls with ASD without accompanying intellectual impairments or language delays may go

unrecognized, perhaps because of subtler manifestation of social and communication difficulties (American Psychiatric Association 2013).

Table I - ASD Diagnostic Criteria (American Psychiatric Association 2013).

ASD Diagnostic Criteria
A. Persistent deficits in social communication and social interaction across multiple contexts, as manifested by the following, currently or by history.
B. Restricted, repetitive patterns of behavior, interests, or activities, as manifested by at least two of the following, currently or by history.
C. Symptoms must be present in the early developmental period (but may not become fully manifest until social demands exceed limited capacities, or may be masked by learned strategies in later life).
D. Symptoms cause clinically significant impairment in social, occupational, or other important areas of current functioning.
E. These disturbances are not better explained by intellectual disability (intellectual developmental disorder) or global developmental delay. Intellectual disability and autism spectrum disorder frequently co-occur; to make comorbid diagnoses of autism spectrum disorder and intellectual disability, social communication should be below that expected for general developmental level.

Etiology

A variety of nonspecific risk factors, such as advanced parental age, low birth weight, or fetal exposure to valproate, may contribute to the risk of developing ASD (American Psychiatric Association 2013). However, several lines of evidence support genetic factors as a predominant cause of ASD.

At least half of ASD is estimated to have its roots in genetic factors (De Rubeis & Buxbaum 2015). Estimates of heritability - the portion of variance in a phenotypic trait among individuals of a population at a given time that can be attributed to genetic differences - have been used to understand the degree to which genetic factors contribute to the difference in susceptibility to ASD among individuals. The genetic variation observed in individuals with ASD is highly heterogeneous and common variants, rare variants, both inherited and *de novo*, can act to increase risk, highlighting the complex risk architecture (De Rubeis & Buxbaum 2015; Iossifov, Ronemus, Levy, Wang, Hakker, Rosenbaum, Yamrom, Y. Lee, et al. 2012). The relative risk of a child being diagnosed with autism is increased at least 25-fold over the population prevalence in families in which a sibling is affected (Jorde et al. 1991). Siblings and parents of an affected child are more likely than controls to show subtle cognitive or behavioral features that are qualitatively similar to those observed in probands (Bolton et al. 1994; Bishop et al. 2004). Indeed, a recent twin study in the UK showed that on all ASD measures, the concordance rates among monozygotic twins (77% - 99%) was significantly higher than those for dizygotic twins (22 – 65%) (Colvert et al. 2015). Several research groups have combined efforts in the recent decades to unravel the genetic factors underlying autism risk and explain its heterogeneity in phenotypical outcomes.

Genetics of Autism Spectrum Disorder

Initial Steps of Genetic Findings in ASD

Before the 1970s, autism was not widely appreciated to have a strong biological basis. Instead, various psychodynamic interpretations, including the role of a cold or aloof style of mothering, were invoked as potential causes (Kanner 1943). The importance of genetic contributions became clear in the 1980s, when the co-occurrence of chromosomal disorders and rare syndromes with ASD was noted (Blomquist et al. 1985). Subsequent twin and family studies provided additional support for a complex genetic etiology, but these were limited by the lack of uniform diagnostic criteria. The development of validated diagnostic and assessment tools in the early 1990s, most notably the Autism Diagnostic Interview - Revised (ADI-R) and the Autism Diagnostic Observation Schedule (ADOS), addressed these concerns and these tools have proved crucial to the advancement of international research into the ASDs. This work, together with important technical advances, made it possible to carry out the first candidate gene association studies and resequencing efforts in the late 1990s. Whole genome linkage studies followed, and were used to identify additional loci of potential interest (depicted in Figure 1) (Abrahams & Geschwind 2008).

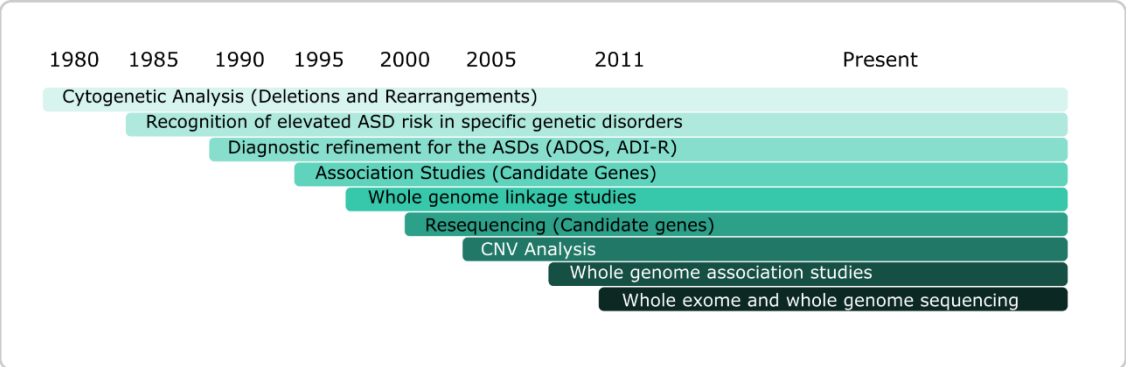


Figure 1 – Methodological changes have revolutionized gene discovery in ASD (adapted from Amaral et al. 2011).

Association Studies and Replication Issues

Genome-wide association studies (GWAS) and linkage analysis have been used to explore the genetic landscape of ASD. These are typically focused on the examination of single-nucleotide polymorphisms (SNPs) and have been estimated to explain 20–50% of the variance in the accountability for ASD (Cross-Disorder Group of the Psychiatric Genomics Consortium et al. 2013; Gaugler et al. 2014; Klei et al. 2012; Devlin & Scherer 2012).

In recent years, six GWAS have been performed for ASD (Anney et al. 2010a; Hussman et al. 2011; Salyakina et al. 2010; Ma et al. 2009; Wang et al. 2009; Weiss et al. 2009). The three largest international studies used family based approaches, each of them pinpointing a promising candidate gene for the disorder: *SEMA5A* at 5p15 (Weiss et al. 2009), *MACROD2* at 20p12.1(Anney et al. 2010) and *CHD10* and *CHD9* (Wang et al. 2009). A common limitation arising from these association studies, as with the linkage analyses, is a lack of replication between studies. Indeed, several research groups have attempted to replicate the association findings of these SNPs that represent the most consistent evidence for association with ASD (Curran et al. 2011; Jonsson et al. 2014; Prandini et al. 2012), but none of them was sufficiently powered to robustly confirm or discard the previous GWAS findings.

Lack of replication in these studies may be due to several reasons, including: (i) genetic and phenotypic heterogeneity between samples due to ascertainment differences and suboptimal sampling that may account for the existence of different pools of common risk variants; (ii) heterogeneity in subject exposure to environmental influences; (iii) data overinterpretation, since the first published study likely represents an overestimate of the true effect size due to a phenomena known as the "winner's curse", suggesting that replication of both association and linkage studies will require larger sample sizes than the initial detection study (Trikalinos et al. 2004; Zollner & Pritchard 2007); and (iv) disparity in sample sizes between research groups leading to false-positive or false-negative results because of differing power to detect real effects (Torrico et al. 2016; Zondervan & Cardon 2004; Meyer et al. 2012).

International collaborative efforts in schizophrenia research made it possible to perform a GWAS study with 36,989 cases that identified 108 significant risk loci for the disorder (Ripke et al. 2014). Similar large-scale projects in ASD would shed light on the contribution of common variants to

autism, confirming or excluding already identified risk alleles and possibly pointing at novel susceptibility loci.

Genotyping Arrays and Massive Parallel Sequencing

Recent advances, including genome-wide copy number arrays and massive parallel sequencing have begun to unravel the genetic complexity in ASD, ranging from large genomic regions to individual nucleotides. Arrays make it possible to detect relative DNA dosage changes (Pinto et al. 2010; Pinto et al. 2014) and whole exome sequencing (WES) can thoroughly survey 1% of the genome comprising the known protein-coding sequences (Iossifov, Ronemus, Levy, Wang, Hakker, Rosenbaum, Yamrom, Y.-H. Lee, et al. 2012; De Rubeis, He, Arthur P Goldberg, et al. 2014). This improved technology, allied with new analysis paradigms and innovative cohorts, is revolutionizing the era of discovery of genomics variants that represent a risk for ASD.

Rare copy-number variation (CNV) is an important source of risk for ASD. CNV examples include *de novo* events observed in 5–10% of ASD cases (Sebat et al. 2007; Christian R. Marshall et al. 2008; Autism Genome Project Consortium et al. 2007), *de novo* or inherited hemizygous deletions and duplications of 16p11.2 (Weiss et al. 2008; Kumar et al. 2007), and exceptionally rare inherited homozygous deletions in consanguineous families (Morrow et al. 2008). Exomes have revealed an excess of genic deletions and duplications in affected patients and an increase in affected subjects carrying exonic pathogenic CNVs overlapping known loci associated with dominant or X-linked ASD and intellectual disability (Pinto et al. 2010; Pinto et al. 2014). Pathogenic CNVs are often associated with variable expressivity, implicating ASD-associated genes previously linked to other neurodevelopmental disorders, as well as novel genes. Consistent with hypothesized gender-specific modulators, females with ASD are more likely to have highly penetrant CNVs (Pinto et al. 2014).

CNV studies allow the identification of novel genes and also the identification of potential biological pathways in the pathogenesis of ASD. Indeed, enrichments have been observed for gene-sets related to cellular proliferation, projection and motility, GTPase/Ras signaling (Pinto et al. 2010), neuronal signaling and development, synapse function, and chromatin regulation (Pinto et al. 2014)

De novo CNVs and large events spanning multiple genes, have been identified as conferring high risk for ASD (Pinto et al. 2014). Although these CNVs provide important leads to underlying biology, they rarely implicate single genes, are rarely fully penetrant, and many confer risk to a broad range of conditions including intellectual disability, epilepsy and schizophrenia (Cross-Disorder Group of the Psychiatric Genomics Consortium et al. 2013). *De novo* coding mutations, being less frequent and potentially more deleterious, could offer insights into risk-determining genes.

Starting in 2009, the technology to selectively sequence all of the protein-coding regions of the genome became widely available. The sequencing of the coding portion of the genome, termed "whole exome sequencing" (WES), allowed for unbiased genome-wide discovery of coding variants or mutations contributing to a disorder's risk at single-base resolution. Several groups began piloting WES in different neurodevelopmental disorders using a trio (father, mother, affected child) or other family design, specifically in families with no previous family history of the disorder, also called simplex or sporadic families (O'Roak et al. 2011; Vissers et al. 2010; Xu et al. 2012). The working hypothesis of these studies was that, in some fraction of these simplex families, there may be a *de novo* mutation (not present in either parent) that coappeared with the disorder in the affected child. These studies showed the feasibility of this approach to detect true *de novo* mutations and a large fraction of possible candidate gene mutations (Veltman & Brunner 2012).

Following those initial studies, large WES studies have taken place (Iossifov, Ronemus, Levy, Wang, Hakker, Rosenbaum, Yamrom, Y. Lee, et al. 2012; Neale et al. 2012; Sanders et al. 2012; O'Roak, Vives, Girirajan, et al. 2012), culminating in two large-scale studies, including 4000 affected children (De Rubeis, He, Arthur P Goldberg, et al. 2014; Iossifov et al. 2014). Data from the family-based Simons Simplex Collection suggests that 30% of all probands have a *de novo* mutation of major effect that contributes to their diagnosis, which may be up to 50% of the girls (Iossifov et al. 2014).

Exome sequencing projects have suggested that severe cases of ASD reveal a higher number of truncating mutations (De Rubeis, He, Arthur P Goldberg, et al. 2014; Iossifov et al. 2014; Krumm et al. 2015; Toma et al. 2014). Thus, it is possible that penetrant rare mutations have a major role in severe ASD phenotypes, whereas common variants may be mainly involved in high functioning autism (Torrico et al. 2016). However, in addition to rare *de novo*

mutations, recent WES studies have also identified a role for rare inherited variants in ASD risk; including maternally transmitted predicted loss-of-function (LoF) variants and recessive/hemizygous LoF variants (Chahrour et al. 2012; Novarino et al. 2012; Lim et al. 2013; Yu et al. 2013). The overall impact of inherited variants on ASD risk will likely be higher when as it does not currently take into account missense variants, whose possible impact is difficult to quantify.

Whole genome sequencing (WGS) provided another major boost to our ability to ascertain point mutations and CNVs (Talkowski et al. 2012; Talkowski, Ernst, et al. 2011; Redin et al. 2017; Brand et al. 2015; Yuen et al. 2015; Michaelson et al. 2012). WGS is an important and efficient tool for the identification of structural variation, particularly balanced chromosomal abnormalities (BCAs). This class of variation includes inversions, translocations, deletions/insertions, and more complex rearrangements consisting of combinations of such events (Redin et al. 2017). Cytogenetic studies estimate a prevalence of 0.2–0.5% for BCAs in the general population (Ravel et al. 2006; Nielsen & Wohlert 1991; Jacobs et al. 1974) and an approximate fivefold increase in the prevalence of BCAs detected by karyotyping has been reported among subjects with neurodevelopmental disorders, particularly for ASD (1.3%) (Christian R Marshall et al. 2008), suggesting that BCAs may represent highly penetrant mutations in these subjects. BCAs might not result in large gains or losses of genetic material at the breakpoint, and therefore they remain undetected by microarray-based genome-wide surveys of genetic variation commonly used in association studies of complex diseases (Talkowski, Ernst, et al. 2011)

The improvement in mutation discovery by WGS comes at a relatively modest increase in sequencing cost since innovative methods such as the "jumping library" strategy developed for WGS which allows for cost-efficient multiplexing of samples and provides a very high yield of fragments with large inserts (Hanscom & Talkowski 2014). Indeed, the degree of resolution that can be obtained through this and other approaches of WGS has enabled the elucidation of the precise breakpoints of BCAs and has facilitated the discovery of numerous pathogenic loci and disrupted genes that represent a risk for ASD (Talkowski et al. 2012).

In summary, the results of all these technologies – from GWAS to WGS - support an extreme polygenicity underlying ASD and the existence of a pool of

risk variants with a wide range of effect sizes. It is now well understood that the genetic contribution to ASD comprises a diversity of sources, including rare de novo single-nucleotide variants, common polymorphic variation, CNVs and structural rearrangements, summarized in Table II below.

Table II - List of major studies of ASD genetics, types of variants identified and genes highlighted. (Abbreviations: GWAS – Genome-wide association studies; BCA- Balanced chromosomal abnormalities.)

Technology	Study	Type of Variant	# ASD Cases	Genes and Loci Found/Highlighted
GWAS	(Weiss et al. 2009)	Common Variants	1553	*SEMA5A*
	(Wang et al. 2009)	Common Variants	4305	*CHD10, CHD9*
	(Ma et al. 2009)	Common Variants	438 families	5p14.1 locus
	(Anney et al. 2010)	Common Variants	1,558 families	*MACROD2*
	(Salyakina et al. 2010)	Common Variants	234 families	3p14.2, 3q25-26, 3p23
	(Hussman et al. 2011)	Common Variants	1293 families	*860 genes - CDH8, SEMA5A, CACNA1G, PTEN, NRXN1, NRP2, CNTNAP2, ZFHX1B*
	(Curran et al. 2011)	Common Variants	1,170	no replication
	(Klei et al. 2012)	Common Variants	3,157 families	-
	(Prandini et al. 2012)	Common Variants	746	*CDH9, CDH10, ATP2B2*
	(Gaugler et al. 2014)	Common Variants	466	-
	(Jonsson et al. 2014)	Common Variants	12,416 + 4,287 twin pairs	no replication
	(Torrico et al. 2016)	Common Variants	7,106	no replication
Genotyping Arrays	(Pinto et al. 2010)	Rare CNV Inherited CNV	996	*SHANK2, SYNGAP1, DLGAP2, DDX53-PTCHD1* locus
	(Sebat et al. 2007)	*Rare De Novo* CNV	118 + 47 families	-
	(Christian R. Marshall et al. 2008)	Structural Variation	427	*SHANK3-NLGN4-NRXN1, DPP6-DPP10-PCDH9, ANKRD11, DPYD, PTCHD1*
	(Glessner et al. 2009)	CNV	859	*NRXN1, CNTN4, NLGN1, ASTN2, UBE3A, PARK2, RFWD2, FBXO40*
	(Pinto et al. 2014)	Rare CNV Inherited CNV	9,050	*CHD2, HDAC4, GDI, SETD5, MIR137, HDAC9*
Whole Exome Sequencing	(O'Roak et al. 2011)	*De Novo* Mutations	20	*FOXP1, GRIN2B, SCN1A, LAMC3*
	(Iossifov, Ronemus, Levy, Wang, Hakker, Rosenbaum, Yamrom, Y. Lee, et al. 2012)	*De Novo* Mutations	343	750 genes
	(Neale et al. 2012)	*De Novo* Mutations	175	*CHD8, KATNAL2*
	(Sanders et al. 2012)	*De Novo* Mutations	238	*SCN2A*
	(O'Roak, Vives, Girirajan, et al. 2012)	*De Novo* Mutations	209	*CHD8, NTNG1, GRIN2B, LAMC3, SCN1A*
	(Chahrour et al. 2012)	Rare Inherited Mutations	19	*UBE3B, CLTCL1, NCKAP5L, ZNF18*
	(Yu et al. 2013)	Rare Inherited Mutations	277	*AMT, PEX7, SYNE1, VPS13B, PAH, POMGNT1*
	(Lim et al. 2013)	Rare Inherited Mutations	933	*USH2A, IFIH1, PKHD1L1, TMLHE, PCDH11X,*

Technology	Study	Type of Variant	# ASD Cases	Genes and Loci Found/Highlighted
				SLC22A14, LUZP4, DGAT2L6, KIAA2022, SRPX2
	(Toma et al. 2014)	Rare Inherited Mutations	21	*COL4A3-MFF, FHIT, MRPL36-NDUFS6, CTNND2, GRM1, ASAH1*
	(De Rubeis et al. 2014)	Rare Coding Variation	3,871	129 genes
	(Iossifov et al. 2014)	*De Novo* Mutations	2,508	27 genes
	(Krumm et al. 2015)	Rare, Disruptive SNV and CNV	2,377	*IMS1, CUL7, LZTR1*
Whole Genome Sequencing	(Talkowski et al. 2012)	BCA	38	*AUTS2, FOXP1, CDKL5, MBD5, SATB2, EHMT1, SNURF-SNRPN, CHD8, KIRREL3, and ZNF507, TCF4, ZNF804A, PDE10A, GRIN2B, ANK3*
	(Michaelson et al. 2012)	*De Novo* Mutations	10 twin pairs	*GPR98, KIRREL3, TCF4*
	(Yuen et al. 2015)	*De Novo* Mutations Rare Inherited Mutations Structural Variation	170	*SCN2A, PTCHD1, SHANK3, DMD, STXBP1*
	(Brand et al. 2015)	Structural Variation	259	-
	(Redin et al. 2017)	BCA	273	*MEF2C*
	(C Yuen et al. 2017)	*De Novo* Mutations	2,620	18 genes
Targeted Sequencing	(O'Roak, Vives, Fu, et al. 2012)	*De Novo* Mutations	2,446	*CHD8, DYRK1A, GRIN2B, TBR1, PTEN, TBL1XR1*

Missing Heritability

The SFARI Gene Database (Basu et al. 2009), an evolving database for the autism research community that is centered on genes implicated in autism susceptibility, contains a total of 725 genes associated with ASD to date. Besides, a massive sequencing study spanning seven countries linked 38 additional genes to autism or developmental delay and intellectual disability (Stessman et al. 2017), expanding the list to close to 1000 genes with evidence of risk for ASD. These known mutations, genetic syndromes, and *de novo* copy number variation probably account for about 10-20% of cases and none of these causes alone accounts for more than 1–2% of ASD cases (Walsh et al. 2008). Despite the great number of genes associated with ASD, there is still a large percentage of idiopathic cases (~70%) for which a genetic source has yet to be assigned, reflecting the missing heritability of complex disorders (Schaaf & Zoghbi 2011).

Several explanations for this missing heritability have been suggested, including: much larger numbers of variants of smaller effect yet to be found; rarer variants (possibly with larger effects) that are poorly detected by available genotyping arrays that focus on variants present in 5% or more of the population;

and structural variants poorly captured by existing arrays (Manolio et al. 2009). WES and WGS projects in ASD have been successfully contributing to uncovering some of the missing heritability of the disorder and novel candidate genes with robust evidence have been proposed (De Rubeis, He, Arthur P. Goldberg, et al. 2014; Iossifov et al. 2014; Krumm et al. 2015; Neale et al. 2012; Sanders et al. 2012; Talkowski et al. 2012; Brand et al. 2015; Redin et al. 2017).

The remaining missing heritability may be due in significant part to genetic interactions (Zuk et al. 2012) as many cases of ASD may result from more complex genetic mechanisms, including co-inheritance of multiple risk alleles or epigenetic modifications (Gupta & State 2007). Understanding how multiple variants act together in a single individual, the risk from noncoding variation and gene vs. environment interactions could shed light on the unexplained portion of ASD cases. The next necessary step is to develop functional validation assays to evaluate the impact of the known variants on their protein function and on neuronal development. Functional validation is especially important for the missense variants that are currently largely ignored unless they occur in known disease-causing genes. Several bioinformatic tools that predict the deleteriousness of genomic variants have been developed, such as SIFT and PolyPhen-2 (Kumar et al. 2009; Adzhubei et al. 2010), however functional validation remains essential to test the biological impact of identified variants in the disease pathogenesis.

Convergence of Biological Pathways in ASD

ASD was thought to be a disorder of the synapse due to the high burden of mutations occurring in synaptic genes. Neuronal development defects include general abnormalities in axon or dendrite growth, synaptogenesis, action potential initiation or propagation, or myelination. It is not known whether these types of abnormalities are found in autism, although there is evidence for connectivity defects from functional imaging studies (Kana et al. 2006; Geschwind & Levitt 2007). Some ASD-associated genes are indeed involved directly in synaptic function as showed by nominal enrichment for postsynaptic density proteins as GRIN2B, GABRB3 and SHANK3 (Iossifov, Ronemus, Levy, Wang, Hakker, Rosenbaum, Yamrom, Y. Lee, et al. 2012; Ben-David & Shifman 2012; Peça et al. 2011). However, recent studies have pointed in a new direction, genes with *de novo* mutations have shown enrichment for multiple molecular functions as global regulation of transcript expression and chromatin modifiers as CHD8, CHD2 and ARID1B (O'Roak, Vives, Girirajan, et al. 2012; O'Roak, Vives, Fu, et al. 2012; Ben-David & Shifman 2013; Talkowski et al. 2012; De Rubeis, He, Arthur P. Goldberg, et al.

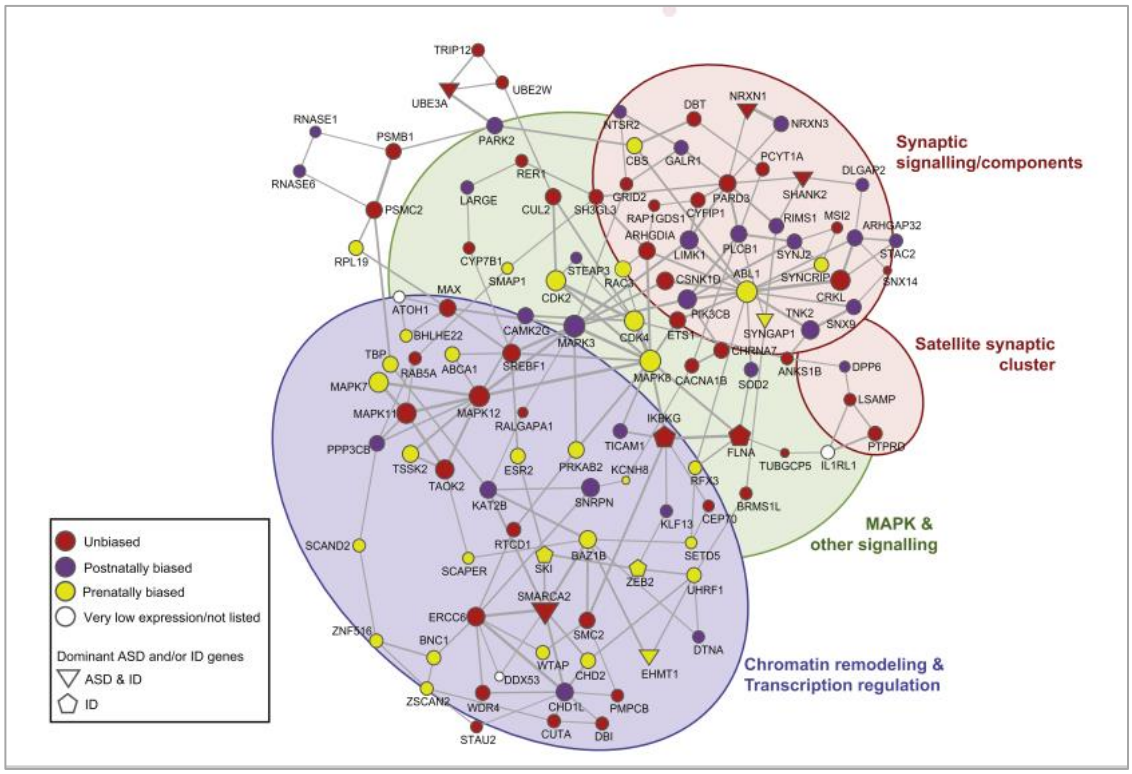


Figure 2 - Network of genes affected by rare de novo CNVs in affected subjects (from Pinto et al., 2014). *This network demonstrates that genes involved in ASD participate in a wide array of processes, from neuronal development and axon guidance to chromatin modification and transcription regulation.*

2014). Therefore, it seems that translational control and chromatin regulation are key players that converge on common downstream biological pathways or brain circuits to give rise to ASD, as confirmed by networks (Figure 2) constructed using these high-confidence ASD risk genes as seeds (O'Roak, Vives, Girirajan, et al. 2012; Parikshak et al. 2015; Willsey et al. 2013; Pinto et al. 2014).

Epigenetic Modifications and Chromatin Remodeling

The chromosomes of eukaryotic cells have the ability to condense and organize their genetic material and control access to genetic information. Chromosomes are comprised of chromatin, a multifaceted and hierarchical nucleoprotein complex containing both histones and non-histone proteins. (Liu et al. 2011). The primary structural unit of chromatin is the nucleosome, which consists of a nucleosome core and linker DNA. The nucleosome core is comprised of ~147 bp of DNA wrapped around a octameric structure containing two molecules each of the core histones H2A, H2B, H3, and H4 (Luger et al. 1997).

The packaging of DNA has a repressive effect on a variety of cellular processes such as gene transcription due to the reduced access of transcription factors to DNA. To overcome this, cells have devised several strategies to modify

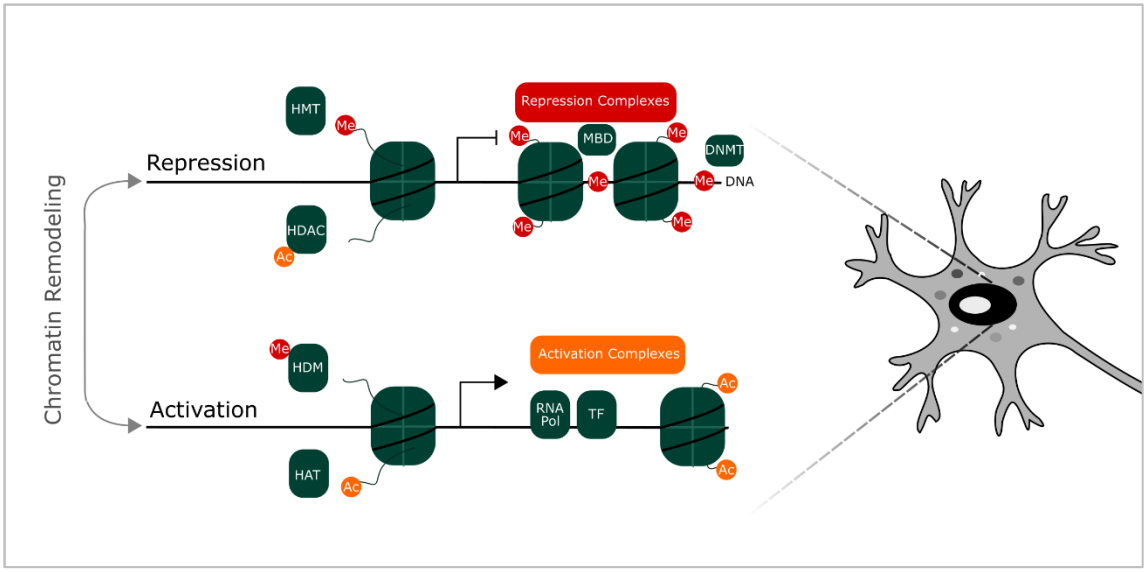


Figure 3 - Chromatin Remodeling. Simplified representation of closed (Repression) and open (Activation) chromatin states.

Top - A closed chromatin state, heterochromatin, is associated with lack of histone acetylation removed by histone deacetylases (HDAC), DNA methylation (Me) by DNA methyltransferases (DNMT), association of methyl binding proteins (MBD) and no gene transcription.

Bottom - The open conformation state, euchromatin, is usually associated with histone acetylation by histone acetyltases (HAT), lack of DNA methylation removed by histone demethylases (HDM), binding of transcription factors (TF) and RNA polymerases (RNA Pol) that define active gene transcription. The interchange between these two DNA states is described as chromatin remodeling.

chromatin structure. These include incorporation of histone post-translational modifications to allow the modification of chromatin structure (Liu et al. 2011; Santos et al. 2006). Indeed, the lysine residues of the core histones are targets for numerous post-translational modifications, by enzymes responsible for acetylation, methylation, phosphorylation, and ubiquitylation, many of which have been directly linked with specific chromatin states (depicted in Figure 3) (Wolffe & Hayes 1999). Closed chromatin, heterochromatin, associated with transcriptional silencing, involves the activity of histone deacetylases (HDACs) that remove acetyl groups from histones and the methylation of histones by histone methyltransferase (HMTs) (Grafodatskaya et al. 2010; Nakao 2001). DNA methylation, the only known modification of DNA itself, also occurs in this state at cytosines followed by guanines (CpG sites) by enzymes of the DNA methyltransferase (DNMT) gene family (Grafodatskaya et al. 2010; Nakao 2001). Open and accessible chromatin, euchromatin, associated with active transcription, is due to the activity of histone acetyltransferases (HATs) that are responsible for histone acetylation and of histone demethylases (HDMs) (Grafodatskaya et al. 2010; Liu et al. 2011; Nakao 2001).

Along with the intrinsic DNA and histone modification, eukaryotic cells also harbor chromatin-remodeling complexes that disrupt chromatin structure to increase access to the underlying DNA. These complexes are typically comprised of multiple subunits and use energy derived from ATP hydrolysis to distort nucleosome structure, mobilize nucleosomes, and possibly to alter higher-order structures. Whereas some remodelers alter chromatin structure to make specific genes more accessible for transcription machinery, others play a role in transcriptional repression. There are 4 classes of ATP-dependent chromatin remodeling complexes that have been described: SWI/SNF, ISWI, CHD and INO80 (Clapier & Cairns 2009).

As supported by the convergence of ASD genes in chromatin regulators, there have been a great number of genes involved in epigenetic modifications identified in ASD cases. These include genes encoding for proteins of all classes of regulation, such as enzymes that affect histone methylation and acetylation, DNA methyltransferases, methyl-binding proteins, as well as genes encoding for members of transcription activation and repression complexes, and are listed in Table III below.

Table III – List of chromatin remodeling genes associated to ASD, to date. Adapted from the Simons Foundation Autism Research Initiative (SFARI) Gene Database, filtered for: ASD, chromatin, histone, transcription. SFARI Db contains a total of 725 genes associated with ASD. Genes with evidence for a certain role but with unknown activity, were placed under a general nomenclature.

Gene	Location	OMIM	Support for Autism	Primary Reference of Support
MBD4	3q21.3	603574	Rare Single Gene variant, Genetic Association	(Cukier et al. 2010)
MBD6	12q13.3		Rare single gene variant	(Cukier et al. 2010)
Histone Acetyltransferase				
ELP4	11p13	606985	Multigenic CNV	(Addis et al. 2015)
EP300	22q13.2	602700	Syndromic	(Vaags et al. 2012)
EP400	12q24.33	606265	Rare single gene variant	(Chahrour et al. 2012)
KAT2B	3p24.3	602303	Rare single gene variant	(Sanders et al. 2015)
KAT6A	8p11.21	601408	Rare single gene variant	(Arboleda et al. 2015)
Histone Deacetylase				
HDAC4	2q37.3	605314	Genetic Association	(Stephen R. Williams et al. 2010)
HDAC6	Xp11.23	300272	Rare Single Gene variant	(Piton et al. 2013)
Histone Demethylase				
KDM4B	19p13.3	609765	Rare single gene variant	(De Rubeis et al. 2014)
KDM5B	1q32.1	605393	Rare single gene variant	(Iossifov et al. 2014)
KDM5C	Xp11.22	314690	Syndromic	(Adegbola et al. 2008)
KDM6B	17p13.1	611577	Rare single gene variant	(Iossifov, Ronemus, Levy, Wang, Hakker, Rosenbaum, Yamrom, Y. Lee, et al. 2012)
PHF2	9q22.31	604351	Rare single gene variant	(Iossifov, Ronemus, Levy, Wang, Hakker, Rosenbaum, Yamrom, Y. Lee, et al. 2012)
PHF8	Xp11.22	300560	Syndromic	(Nava et al. 2012)
Histone Methyltransferase				
EHMT1	9q34.3	607001	Syndromic	(Talkowski et al. 2012)
KMT2A	11q23.3	159555	Syndromic	(De Rubeis et al. 2014)
KMT2C	7q36.1	606833	Rare single gene variant	(De Rubeis et al. 2014)
KMT2E	7q22.3	608444	Rare single gene variant	(Dong et al. 2014)
KMT5B	11q13.2	610881	Rare Single Gene variant	(Sanders et al. 2012)
SETD2	3p21.31	612778	Rare Single Gene variant	(O'Roak, Vives, Girirajan, et al. 2012)
SETD5	3p25.3		Rare Single Gene variant	(Grozeva et al. 2014)
SETDB1	1q21.3	604396	Rare single gene variant	(Cukier et al. 2012)
SETDB2	13q14.2	607865	Syndromic	(Cukier et al. 2012)
Chromatin Remodeling				
ARID1B	6q25.3	614556	Rare single gene variant	(Nord et al. 2011; Hoyer et al. 2012; Yu et al. 2015; Halgren et al. 2012)
ATRX	Xq21.1	300032	Syndromic	(Gibbons 2006)
CECR2	22q11.1-q11.21	607576	Rare single gene variant	(Prasad et al. 2012)
CHD2	15q26.1	602119	Multigenic CNV	(Carvill et al. 2013)
CHD7	8q12.2	608892	Syndromic	(Vissers et al. 2004)
CTCF	16q22.1	604167	Functional	(Gregor et al. 2013)
HMGN1	21q22.2	163920	Genetic Association	(Abuhatzira et al. 2011)

Gene	Location	OMIM	Support for Autism	Primary Reference of Support
MBD3	19p13.3	603573	Rare Single Gene variant, Genetic Association	(Cukier et al. 2010)
TET2	4q24	612839	Rare single gene variant	(Iossifov et al. 2014; Krumm et al. 2015)
TLK2	17q23.2	608439	Rare Single Gene variant	(O'Roak et al. 2011)
Transcription Regulators				
ERG	21q22.2	165080	Genetic association	(Anney et al. 2012)
SATB2	2q33.1	608148	Syndromic	(Talkowski et al. 2012)
WAC	10p12.1	615049	Rare single gene variant	(Iossifov et al. 2014)
Transcription Activators				
AFF2	Xq28	300806	Syndromic	(Abrams et al. 1997)
CREBBP	16p13.3	600140	Syndromic	(Barnby et al. 2005)
MBD5	2q23.1	611472	Rare single gene variant	(Wagenstaller et al. 2007; Jaillard et al. 2009; Williams et al. 2010; Talkowski, Mullegama, et al. 2011; Talkowski et al. 2012)
SMARCA2	9p24.3	600014	Genetic association/functional	(Wolff et al. 2012)
SMARCC2	12q13.2	601734	Functional	(Neale et al. 2012)
Transcription Repressors				
ASXL3	18q12.1	615115	Syndromic	(Dinwiddie et al. 2013)
C11orf30	11q13.5	608574	Rare single gene variant	(De Rubeis et al. 2014)
CHD8	14q11.2	610528	Rare Single Gene variant	(O'Roak, Vives, Girirajan, et al. 2012; Talkowski et al. 2012)
MBD1	18q21.1	156535	Rare Single Gene variant	(Cukier et al. 2010)
MECP2	Xq28	300005	Syndromic	(Zoghbi et al. 1999)
NCOR1	17p12-p11.2	600849	Rare single gene variant	(Iossifov, Ronemus, Levy, Wang, Hakker, Rosenbaum, Yamrom, Y. Lee, et al. 2012)
ZMYND11	10p15.3	608668	Rare single gene variant	(Iossifov, Ronemus, Levy, Wang, Hakker, Rosenbaum, Yamrom, Y. Lee, et al. 2012)

Functional Assessment of Chromatin Remodelers in ASD

Examples of chromatin-related genes that have been followed up and functionally proven to affect neurodevelopment include *CHD8*, *ARID1B*, *SMARCA4*, *ANKRD11*, and *EHMT1*. The chromodomain helicase DNA-binding protein 8 (CHD8) is an ATP-dependent chromatin remodeler of the SNF2 family (Sugathan et al. 2014). CHD8 was identified as one of the genes in the minimal region of overlap of *de novo* 14q11.2 microdeletions in two children with developmental delay and cognitive impairment (Zahir et al. 2007). The first robust hint of CHD8's importance for autism came in 2012, when a study found that people with autism are more than twice as likely to have a harmful spontaneous mutation in CHD8 than in any other gene (O'Roak, Vives, Fu, et al. 2012). Other studies have shown that mutations in CHD8 are indeed highly penetrant, leading to ASD (Bernier et al. 2014) and that CHD8 regulates many functionally distinct genes associated with ASD and members of pathways important to neurodevelopment and Wnt/β-catenin signaling (Wang et al. 2015; Sugathan et al. 2014; Cotney et al. 2015), implicating key pathways in the disorder.

Mutations in the *ARID1B* gene encoding AT-rich interactive domain-containing protein 1B were recently associated with multiple syndromes characterized by developmental delay and intellectual disability, in addition to non-syndromic intellectual disability. While the majority of ARID1B mutations identified to date are predicted to result in haploinsufficiency, the underlying pathogenic mechanisms have yet to be fully understood. ARID1B is a DNA- binding subunit of the Brahma-associated factor (BAF) chromatin remodeling complexes, which play a key role in the regulation of gene activity (Sim et al. 2015). There is now evidence that ARID1B is a repressor of Wnt/β-catenin signaling (Vasileiou et al. 2015). ARID1B was able to associate with β-catenin and repress Wnt/β-catenin-mediated transcription through the BAF core subunit BRG1 (Vasileiou et al. 2015).

In turn, transcription activator Brg1 also known as ATP-dependent helicase SMARCA4, was shown to play an important role in both synapse development and maturation and MEF2-mediated synapse remodeling in mice (Zhang et al. 2016). Indeed, the deletion of *Brg1* in early postnatal hippocampal mouse neurons resulted in reduced dendritic spine density and maturation and impaired synapse activities (Zhang et al. 2016). Additionally, gene expression analyses indicated that Brg1 regulates a significant number of genes known to be involved in synapse

function and implicated in ASD, as seen in CHD8 knockdown experiments (Sugathan et al. 2014; Zhang et al. 2016).

By playing a role in histone acetylation, Ankyrin repeat domain containing protein 11 (ANKRD11) has shown to be crucial for neurodevelopment as deletions or mutation in one allele of *ANKRD11* cause cognitive dysfunction and ASD (Christian R. Marshall et al. 2008; Lo-Castro et al. 2013; Sirmaci et al. 2011). A recent study showed that the knockdown of Ankrd11 in developing murine or human cortical neural precursors caused decreased proliferation, reduced neurogenesis, and aberrant neuronal positioning. Consistent with a role for Ankrd11 in histone acetylation, Ankrd11 was associated with chromatin and colocalized with HDAC3, and expression and histone acetylation of Ankrd11 target genes were altered in neural precursors with point mutations in the Ankrd11 HDAC-binding domain. Thus, Ankrd11 is a crucial chromatin regulator that controls histone acetylation and gene expression during neural development, thereby providing a likely explanation for its association with cognitive dysfunction and ASD (Gallagher et al. 2015).

On the other hand, euchromatin histone methyltransferase 1 (EHMT1) is a histone methyltransferase that is part of the E2F6 complex, capable of histone 3 lysine 9 dimethylation (H3K9me2) in euchromatic regions of the genome, which represses transcription. H3K9 methylation has a fundamental role in heterochromatin formation, transcriptional silencing, X-chromosome inactivation, and DNA methylation. Defects in *EHMT1* are associated with intellectual disability (Bessa et al. 2007) and with 9q subtelomeric deletion syndrome which is due to haploinsufficiency for EHMT1 (Kleefstra et al. 2006; Kleefstra et al. 2012). EHMT1 has shown to play a critical and cell-autonomous role in synaptic scaling by responding to attenuated neuronal firing or sensory drive, suggesting that H3K9me2-mediated changes in chromatin structure govern a repressive program that controls synaptic scaling (Benevento et al. 2016).

Apart from the well described and studied genes, there are still several chromatin regulators for which the specific neurobiological function remains unexplained. *MBD5* (methyl-CpG binding domain protein 5) encodes a member of the methyl-CpG-binding domain (MBD) family that has been implicated as the critical gene responsible for the 2q23.1 deletion syndrome (Talkowski, Mullegama, et al. 2011). Haploinsufficiency of this gene is associated with a syndrome involving microcephaly, intellectual disabilities, severe speech impairment, and seizures. The features associated with a deletion, mutation or duplication of MBD5 and the gene

expression changes observed support MBD5 as a dosage-sensitive gene critical for normal development (Mullegama et al. 2013). Classical methyl-CpG binding proteins contain the conserved DNA binding motif methyl-cytosine binding domain (MBD), which preferentially binds to methylated CpG dinucleotides. These proteins serve as transcriptional repressors, mediating gene silencing via DNA cytosine methylation. Mutations in methyl-CpG binding protein 2 (MeCP2) have been linked to the human mental retardation disorder Rett syndrome, suggesting an important role for methyl-CpG binding proteins in brain development and function (Fan & Hutnick 2005). In contrast, MBD3, MBD5 and MBD6 do not bind methylated DNA, either due to amino acid alterations at critical positions or deletion of key DNA inter- acting residues in the MBD (Laget et al. 2010).

Methyltransferase-like 2B (METTL2B) is a protein of unknown function that is a member of a family of methyltransferases that share homology with the UbiE family of methyltransferases. The *METTL2B* gene which is highly homologous to METTL2A (also of unknown function), was identified as disrupted by the breakpoint of a balanced chromosomal translocation in an ASD subject with a severe phenotype (Talkowski et al. 2012). The patient presented with neuromuscular hypotonia, developmental delay, dysmorphism and previous MRI of brain revealed delayed myelination. Given the severity of the phenotype in the patient bearing no other genetic defects, this gene is also a potential candidate that requires further investigation to pinpoint its role in neurodevelopment and ASD.

In summary, as described throughout this chapter, ASD may be caused by an array of different genes and variants that represent risk for ASD. While this poses a huge problem for understanding and treating ASD, it does suggest a convergence of the developmental and neuronal pathways (anatomical, cellular and molecular) which tie together the known molecular defects causing ASD. Along with this paradigm, it is imperative to further explore the functional roles of the causal variants and genes on protein and cellular function and its impact on phenotype development. Functional validation is essential to test the biological impact of identified genes and will be the focus of this work.

Goals

Chromatin regulators have been shown to be essential for brain development, by controlling processes as neurogenesis and neural differentiation and rely on epigenetic marks such as post-translational modifications of histones and transcriptional regulation of downstream players involved in synaptic function. However, there is a panoply of genes that remain unexplored that can contribute to the theory of convergence of biological pathways in ASD or could reveal downstream players in neurodevelopment. In order to understand this, it is important to first generate models that allow the investigation of the contribution of individual genes to neurodevelopment.

This project will explore chromatin-related genes that were previously identified as strong risk factors for ASD - *EHMT1*, *MBD5*, *METTL2B and ARID1B* - to explore the functional impact of their disruption, either through *in vitro* models or patient samples. This thesis was divided into three chapters where each chapter will address a specific goal. The main goals were to:

Chapter 1. Characterize a unique microduplication of the chromatin regulator *ARID1B* in a patient with intellectual disability and identify the dysregulated pathways contributing to the phenotype.

Chapter 2. Create models of loss-of-function of chromatin-related genes (*EHMT1*, *MBD5*, *METTL2A* and *METTL2B*) using CRISPR/Cas9 genome editing technology in an isogenic cell line of induced pluripotent stem cells to study ASD pathogenesis.

Chapter 3. Generate iPSC-derived neuronal progenitor and mature neuronal cells to investigate the impact of perturbing the chromatin remodeler *MBD5* during neurodevelopment, in terms of genome-wide transcriptomic alterations.

References

Abrahams, B.S. & Geschwind, D.H., 2008. Advances in autism genetics: on the threshold of a new neurobiology. *Nature Reviews. Genetics*, 9(5), pp.341–355.

Abrams, M.T. et al., 1997. Cognitive, behavioral, and neuroanatomical assessment of two unrelated male children expressing FRAXE. *American Journal of Medical Genetics*, 74(1), pp.73–81.

Abuhatzira, L. et al., 2011. The chromatin-binding protein HMGN1 regulates the expression of methyl CpG-binding protein 2 (MECP2) and affects the behavior of mice. *The Journal of Biological Chemistry*, 286(49), pp.42051–62.

Addis, L. et al., 2015. Microdeletions of ELP4 Are Associated with Language Impairment, Autism Spectrum Disorder, and Mental Retardation. *Human Mutation*, 36(9), pp.842–50.

ADDM Network, 2012. ADDM Network Community Report 2012. Available at: http://www.cdc.gov/ncbddd/autism/documents/addm-2012-community-report.pdf.

Adegbola, A. et al., 2008. A novel mutation inJARID1C/SMCX in a patient with autism spectrum disorder (ASD). *American Journal of Medical Genetics Part A*, 146A(4), pp.505–511.

Adzhubei, I.A. et al., 2010. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4), pp.248–9.

Amaral, D.G., Dawson, G. & Geschwind, D.H., 2011. *Autism Spectrum Disorders*,

American Psychiatric Association, 2013. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)* 5th ed.,

Anney, R. et al., 2010. A genome-wide scan for common alleles affecting risk for autism. *Human molecular genetics*, 19(20), pp.4072–82.

Anney, R. et al., 2012. Individual common variants exert weak effects on the risk for autism spectrum disorders. *Human Molecular Genetics*, 21(21), pp.4781–4792.

Arboleda, V.A. et al., 2015. De novo nonsense mutations in KAT6A, a lysine acetyl-transferase gene, cause a syndrome including microcephaly and global developmental delay. *American Journal of Human Genetics*, 96(3), pp.498–506.

Autism Genome Project Consortium, P. et al., 2007. Mapping autism risk loci

using genetic linkage and chromosomal rearrangements. *Nature Genetics*, 39(3), pp.319–28.

Barnby, G. et al., 2005. Candidate-Gene Screening and Association Analysis at the Autism-Susceptibility Locus on Chromosome 16p: Evidence of Association at GRIN2A and ABAT. *The American Journal of Human Genetics*, 76(6), pp.950–966.

Basu, S.N., Kollu, R. & Banerjee-Basu, S., 2009. AutDB: a gene reference resource for autism research. *Nucleic Acids Research*, 37(Database), pp.D832–D836.

Ben-David, E. & Shifman, S., 2013. Combined analysis of exome sequencing points toward a major role for transcription regulation during brain development in autism. *Molecular Psychiatry*, 18(10), pp.1054–1056.

Ben-David, E. & Shifman, S., 2012. Networks of neuronal genes affected by common and rare variants in autism spectrum disorders. G. Gibson, ed. *PLoS genetics*, 8(3), p.e1002556.

Benevento, M. et al., 2016. Histone Methylation by the Kleefstra Syndrome Protein EHMT1 Mediates Homeostatic Synaptic Scaling. *Neuron*, 91(2), pp.341–355.

Berg, J.M. & Geschwind, D.H., 2012. Autism genetics: searching for specificity and convergence. *Genome Biology*, 13(7), p.247.

Bernier, R. et al., 2014. Disruptive CHD8 Mutations Define a Subtype of Autism Early in Development. *Cell*, 158(2), pp.263–276.

Bessa, C., Lopes, F. & Maciel, P., 2007. Molecular Genetics of Intellectual Disability.

Bishop, D.V.M. et al., 2004. Using self-report to identify the broad phenotype in parents of children with autistic spectrum disorders: a study using the Autism-Spectrum Quotient. *Journal of Child Psychology and Psychiatry*, 45(8), pp.1431–1436.

Blomquist, H.Ks. et al., 1985. Frequency of the fragile X syndrome in infantile autism. *Clinical Genetics*, 27(2), pp.113–117.

Bolton, P. et al., 1994. A Case-Control Family History Study of Autism. *Journal of Child Psychology and Psychiatry*, 35(5), pp.877–900.

Brand, H. et al., 2015. *Paired-Duplication Signatures Mark Cryptic Inversions and Other Complex Structural Variation*,

C Yuen, R.K. et al., 2017. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nature Neuroscience*.

Carvill, G.L. et al., 2013. Targeted resequencing in epileptic encephalopathies

identifies de novo mutations in CHD2 and SYNGAP1. *Nature Genetics*, 45(7), pp.825–830.

Chahrour, M.H. et al., 2012. Whole-Exome Sequencing and Homozygosity Analysis Implicate Depolarization-Regulated Neuronal Genes in Autism D. H. Geschwind, ed. *PLoS Genetics*, 8(4), p.e1002635.

Clapier, C.R. & Cairns, B.R., 2009. The biology of chromatin remodeling complexes. *Annual review of biochemistry*, 78(1), pp.273–304.

Colvert, E. et al., 2015. Heritability of Autism Spectrum Disorder in a UK Population-Based Twin Sample. *JAMA psychiatry*, 72(5), pp.415–23.

Cotney, J. et al., 2015. The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. *Nature Communications*, 6, p.6404.

Cross-Disorder Group of the Psychiatric Genomics Consortium, C.-D.G. of the P.G. et al., 2013. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*, 45(9), pp.984–94.

Cukier, H.N. et al., 2010. Novel variants identified in methyl-CpG-binding domain genes in autistic individuals. *Neurogenetics*, 11(3), pp.291–303.

Cukier, H.N. et al., 2012. The expanding role of MBD genes in autism: identification of a MECP2 duplication and novel alterations in MBD5, MBD6, and SETDB1. *Autism Research*, 5(6), pp.385–97.

Curran, S. et al., 2011. No association between a common single nucleotide polymorphism, rs4141463, in the MACROD2 gene and autism spectrum disorder. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 156(6), pp.633–639.

Devlin, B. & Scherer, S.W., 2012. Genetic architecture in autism spectrum disorder. *Current Opinion in Genetics & Development*, 22(3), pp.229–237.

Dinwiddie, D.L. et al., 2013. De novoframeshift mutation in ASXL3 in a patient with global developmental delay, microcephaly, and craniofacial anomalies. *BMC Medical Genomics*, 6(1), p.32.

Dong, S. et al., 2014. De Novo Insertions and Deletions of Predominantly Paternal Origin Are Associated with Autism Spectrum Disorder. *Cell Reports*, 9(1), pp.16–23.

Fan, G. & Hutnick, L., 2005. Methyl-CpG binding proteins in the nervous system. *Cell research*, 15(4), pp.255–61.

Gallagher, D. et al., 2015. Ankrd11 Is a Chromatin Regulator Involved in Autism that Is Essential for Neural Development. *Developmental Cell*, 32(1), pp.31–

42.

Gaugler, T. et al., 2014. Most genetic risk for autism resides with common variation. *Nature genetics*, 46(8), pp.881–5.

Geschwind, D.H., 2009. Advances in Autism. *Annual review of medicine*, 60, pp.367–80.

Geschwind, D.H. & Levitt, P., 2007. Autism spectrum disorders: developmental disconnection syndromes. *Current Opinion in Neurobiology*, 17(1), pp.103–111.

Gibbons, R., 2006. Alpha thalassaemia-mental retardation, X linked. *Orphanet Journal of Rare Diseases*, 1(1), p.15.

Glessner, J.T. et al., 2009. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature*, 459(7246), pp.569–73.

Grafodatskaya, D. et al., 2010. Autism spectrum disorders and epigenetics. *Journal of the American Academy of Child and Adolescent Psychiatry*, 49(8), pp.794–809.

Gregor, A. et al., 2013. De novo mutations in the genome organizer CTCF cause intellectual disability. *American Journal of Human Genetics*, 93(1), pp.124–31.

Grozeva, D. et al., 2014. De Novo Loss-of-Function Mutations in SETD5, Encoding a Methyltransferase in a 3p25 Microdeletion Syndrome Critical Region, Cause Intellectual Disability. *The American Journal of Human Genetics*, 94(4), pp.618–624.

Gupta, A.R. & State, M.W., 2007. Recent Advances in the Genetics of Autism. *Biological Psychiatry*, 61(4), pp.429–437.

Halgren, C. et al., 2012. Corpus callosum abnormalities, intellectual disability, speech impairment, and autism in patients with haploinsufficiency of ARID1B. *Clinical Genetics*, 82(3), pp.248–55.

Hanscom, C. & Talkowski, M., 2014. Design of large-insert jumping libraries for structural variant detection using Illumina sequencing. *Current Protocols in Human Genetics*, 80, pp.7.22.1–9.

Hoyer, J. et al., 2012. Haploinsufficiency of ARID1B, a member of the SWI/SNF-a chromatin-remodeling complex, is a frequent cause of intellectual disability. *American Journal of Human Genetics*, 90(3), pp.565–72.

Hussman, J.P. et al., 2011. A noise-reduction GWAS analysis implicates altered regulation of neurite outgrowth and guidance in autism. *Molecular Autism*, 2(1), p.1.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y., et al., 2012. De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron*, 74(2), pp.285–299.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.-H., et al., 2012. De novo gene disruptions in children on the autistic spectrum. *Neuron*, 74(2), pp.285–99.

Iossifov, I. et al., 2014. The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, 515(7526), pp.216–221.

Jacobs, P.A. et al., 1974. A cytogenetic survey of 11,680 newborn infants. *Annals of Human Genetics*, 37(4), pp.359–76.

Jaillard, S. et al., 2009. 2q23.1 microdeletion identified by array comparative genomic hybridisation: an emerging phenotype with Angelman-like features? *Journal of Medical Genetics*, 46(12), pp.847–55.

Jonsson, L. et al., 2014. Association study between autistic-like traits and polymorphisms in the autism candidate regions RELN, CNTNAP2, SHANK3, and CDH9/10. *Molecular Autism*, 5(1), p.55.

Jorde, L.B. et al., 1991. Complex segregation analysis of autism. *American Journal of Human Genetics*, 49(5), pp.932–8.

Kana, R.K. et al., 2006. Sentence comprehension in autism: thinking in pictures with decreased functional connectivity. *Brain*, 129(9), pp.2484–2493.

Kanner, L., 1943. Autistic Disturbances of Affective Contact. *Pathology*.

Kleefstra, T. et al., 2012. Disruption of an EHMT1-associated chromatin-modification module causes intellectual disability. *American Journal of Human Genetics*, 91(1), pp.73–82.

Kleefstra, T. et al., 2006. Loss-of-function mutations in euchromatin histone methyl transferase 1 (EHMT1) cause the 9q34 subtelomeric deletion syndrome. *American Journal of Human Genetics*, 79(2), pp.370–7.

Klei, L. et al., 2012. Common genetic variants, acting additively, are a major source of risk for autism. *Molecular Autism*, 3(1), p.9.

Krumm, N. et al., 2015. Excess of rare, inherited truncating mutations in autism. *Nature Genetics*, 47(6), pp.582–8.

Kumar, P., Henikoff, S. & Ng, P.C., 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7), pp.1073–81.

Kumar, R.A. et al., 2007. Recurrent 16p11.2 microdeletions in autism. *Human Molecular Genetics*, 17(4), pp.628–638.

Laget, S. et al., 2010. The human proteins MBD5 and MBD6 associate with heterochromatin but they do not bind methylated DNA. S. D. Fugmann, ed. *PloS One*, 5(8), p.e11982.

Lim, E.T. et al., 2013. Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron*, 77(2), pp.235–42.

Liu, N., Balliano, A. & Hayes, J.J., 2011. Mechanism(s) of SWI/SNF-induced nucleosome mobilization. *Chembiochem*, 12(2), pp.196–204.

Lo-Castro, A. et al., 2013. Neurobehavioral phenotype observed in KBG syndrome caused by ANKRD11 mutations. *American journal of medical genetics. Part B, Neuropsychiatric Genetics*, 162B(1), pp.17–23.

Luger, K. et al., 1997. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, 389(6648), pp.251–60.

Ma, D. et al., 2009. A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. *Annals of Human Genetics*, 73(Pt 3), pp.263–73.

Manolio, T.A. et al., 2009. Finding the missing heritability of complex diseases. *Nature*, 461(7265), pp.747–53.

Marshall, C.R. et al., 2008. Structural Variation of Chromosomes in Autism Spectrum Disorder. *The American Journal of Human Genetics*, 82(2), pp.477–488.

Marshall, C.R. et al., 2008. Structural variation of chromosomes in autism spectrum disorder. *American journal of human genetics*, 82(2), pp.477–88.

Meyer, W.K. et al., 2012. Evaluating the Evidence for Transmission Distortion in Human Pedigrees. *Genetics*, 191(1), pp.215–232.

Michaelson, J.J. et al., 2012. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell*, 151(7), pp.1431–42.

Morrow, E.M. et al., 2008. Identifying autism loci and genes by tracing recent shared ancestry. *Science (New York, N.Y.)*, 321(5886), pp.218–23.

Mullegama, S. V et al., 2013. Reciprocal deletion and duplication at 2q23.1 indicates a role for MBD5 in autism spectrum disorder. *European Journal of Human Genetics : EJHG*, (November 2012), pp.1–7.

Nakao, M., 2001. Epigenetics: interaction of DNA methylation and chromatin. *Gene*, 278(1-2), pp.25–31.

Nava, C. et al., 2012. Analysis of the chromosome X exome in patients with autism spectrum disorders identified novel candidate genes, including

TMLHE. *Translational Psychiatry*, 2(10), p.e179.

Neale, B.M. et al., 2012. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, 485.

Nielsen, J. & Wohlert, M., 1991. Chromosome abnormalities found among 34,910 newborn children: results from a 13-year incidence study in Arhus, Denmark. *Human Genetics*, 87(1), pp.81–3.

Nord, A.S. et al., 2011. Reduced transcript expression of genes affected by inherited and de novo CNVs in autism. *European Journal of Human Genetics : EJHG*, 19(6), pp.727–31.

Novarino, G. et al., 2012. Mutations in BCKD-kinase lead to a potentially treatable form of autism with epilepsy. *Science (New York, N.Y.)*, 338(6105), pp.394–7.

O'Roak, B.J. et al., 2011. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genetics*, 43(6), pp.585–9.

O'Roak, B.J., Vives, L., Fu, W., et al., 2012. Multiplex Targeted Sequencing Identifies Recurrently Mutated Genes in Autism Spectrum Disorders. *Science*.

O'Roak, B.J., Vives, L., Girirajan, S., et al., 2012. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 485(7397), pp.246–250.

Parikshak, N.N., Gandal, M.J. & Geschwind, D.H., 2015. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nature Reviews Genetics*.

Peça, J. et al., 2011. Shank3 mutant mice display autistic-like behaviours and striatal dysfunction. *Nature*, 472(7344), pp.437–442.

Pinto, D. et al., 2014. Convergence of Genes and Cellular Pathways Dysregulated in Autism Spectrum Disorders. *The American Journal of Human Genetics*.

Pinto, D. et al., 2010. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466(7304), pp.368–72.

Piton, A. et al., 2013. Analysis of the effects of rare variants on splicing identifies alterations in GABAA receptor genes in autism spectrum disorder individuals. *European Journal of Human Genetics*, 21(7), pp.749–756.

Prandini, P. et al., 2012. The association of rs4307059 and rs35678 markers with autism spectrum disorders is replicated in Italian families. *Psychiatric Genetics*, 22(4), pp.177–181.

Prasad, A. et al., 2012. A Discovery Resource of Rare Copy Number Variations in Individuals with Autism Spectrum Disorder. *G3: Genes|Genomes|Genetics*, 2(12), pp.1665–1685.

Ravel, C. et al., 2006. Prevalence of chromosomal abnormalities in phenotypically normal and fertile adult males: large-scale survey of over 10,000 sperm donor karyotypes. *Human Reproduction (Oxford, England)*, 21(6), pp.1484–9.

Redin, C. et al., 2017. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nature Genetics*, 49(1), pp.36–45.

Ripke, S. et al., 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510), pp.421–427.

De Rubeis, S., He, X., Goldberg, A.P., et al., 2014. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, 515(7526), pp.209–215.

De Rubeis, S., He, X., Goldberg, A.P., et al., 2014. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, 515(7526), pp.209–15.

De Rubeis, S. & Buxbaum, J.D., 2015. Recent advances in the genetics of autism spectrum disorder. *Current Neurology and Neuroscience Reports*, 15(6), p.553.

Salyakina, D. et al., 2010. Variants in several genomic regions associated with asperger disorder. *Autism Research*, 3(6), pp.303–10.

Sanders, S.J. et al., 2012. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, 485.

Sanders, S.J. et al., 2015. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron*, 87(6), pp.1215–1233.

Santos, M., Coelho, P.A. & Maciel, P., 2006. Chromatin remodeling and neuronal function: exciting links. *Genes, Brain and Behavior*, 5, pp.80–91.

Schaaf, C.P. & Zoghbi, H.Y., 2011. Solving the autism puzzle a few pieces at a time. *Neuron*, 70(5), pp.806–8.

Sebat, J. et al., 2007. Strong association of de novo copy number mutations with autism. *Science (New York, N.Y.)*, 316(5823), pp.445–9.

Sim, J.C.H., White, S.M. & Lockhart, P.J., 2015. ARID1B-mediated disorders: Mutations and possible mechanisms. *Intractable & Rare Diseases Research*, 4(1), pp.17–23.

Sirmaci, A. et al., 2011. Mutations in ANKRD11 Cause KBG Syndrome, Characterized by Intellectual Disability, Skeletal Malformations, and

Macrodontia. *The American Journal of Human Genetics*, 89(2), pp.289–294.

Stessman, H.A.F. et al., 2017. Targeted sequencing identifies 91 neurodevelopmental-disorder risk genes with autism and developmental-disability biases. *Nature Genetics*.

Sugathan, A. et al., 2014. *CHD8* regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. *Proceedings of the National Academy of Sciences*, 111(42), pp.E4468–E4477.

Talkowski, M.E., Mullegama, S. V, et al., 2011. Assessment of 2q23 . 1 Microdeletion Syndrome Implicates MBD5 as a Single Causal Locus of Intellectual Disability , Epilepsy , and Autism Spectrum Disorder. *The American Journal of Human Genetics*, pp.551–563.

Talkowski, M.E., Ernst, C., et al., 2011. Next-Generation Sequencing Strategies Enable Routine Detection of Balanced Chromosome Rearrangements for Clinical Diagnostics and Genetic Research. *The American Journal of Human Genetics*, pp.469–481.

Talkowski, M.E. et al., 2012. Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell*, 149(3), pp.525–37.

Toma, C. et al., 2014. Exome sequencing in multiplex autism families suggests a major role for heterozygous truncating mutations. *Molecular Psychiatry*, 19(7), pp.784–790.

Torrico, B. et al., 2016. Lack of replication of previous autism spectrum disorder GWAS hits in European populations. *Autism Research*.

Trikalinos, T.A. et al., 2004. Establishment of genetic associations for complex diseases is independent of early study findings. *European Journal of Human Genetics*, 12(9), pp.762–769.

Vaags, A.K. et al., 2012. Rare deletions at the neurexin 3 locus in autism spectrum disorder. *American Journal of Human Genetics*, 90(1), pp.133–41.

Vasileiou, G. et al., 2015. Chromatin-Remodeling-Factor ARID1B Represses Wnt/β-Catenin Signaling. *The American Journal of Human Genetics*, 97(3), pp.445–456.

Veltman, J.A. & Brunner, H.G., 2012. De novo mutations in human genetic disease. *Nature Reviews Genetics*, 13(8), pp.565–575.

Vissers, L.E.L.M. et al., 2010. A de novo paradigm for mental retardation. *Nature Genetics*, 42(12), pp.1109–1112.

Vissers, L.E.L.M. et al., 2004. Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. *Nature Genetics*, 36(9), pp.955–957.

Wagenstaller, J. et al., 2007. Copy-Number Variations Measured by Single-Nucleotide–Polymorphism Oligonucleotide Arrays in Patients with Mental Retardation. *The American Journal of Human Genetics*, 81(4), pp.768–779.

Walsh, C.A., Morrow, E.M. & Rubenstein, J.L.R., 2008. Autism and brain development. *Cell*, 135(3), pp.396–400.

Wang, K. et al., 2009. Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature*, 459(7246), pp.528–33.

Wang, P. et al., 2015. CRISPR/Cas9-mediated heterozygous knockout of the autism gene CHD8 and characterization of its transcriptional networks in neurodevelopment. *Molecular Autism*, 6, p.55.

Weiss, L.A. et al., 2009. A genome-wide linkage and association scan reveals novel loci for autism. *Nature*, 461(7265), pp.802–808.

Weiss, L.A. et al., 2009. A genome-wide linkage and association scan reveals novel loci for autism. *Nature*, 461(7265), pp.802–8.

Weiss, L.A. et al., 2008. Association between Microdeletion and Microduplication at 16p11.2 and Autism. *New England Journal of Medicine*, 358(7), pp.667–675.

Williams, S.R. et al., 2010. Haploinsufficiency of HDAC4 Causes Brachydactyly Mental Retardation Syndrome, with Brachydactyly Type E, Developmental Delays, and Behavioral Problems. *The American Journal of Human Genetics*, 87(2), pp.219–228.

Williams, S.R. et al., 2010. Haploinsufficiency of MBD5 associated with a syndrome involving microcephaly, intellectual disabilities, severe speech impairment, and seizures. *European Journal of Human Genetics : EJHG*, 18(4), pp.436–41.

Willsey, A.J. et al., 2013. Coexpression Networks Implicate Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism. *Cell*, 155(5), pp.997–1007.

Wolff, D. et al., 2012. In-Frame Deletion and Missense Mutations of the C-Terminal Helicase Domain of SMARCA2 in Three Patients with Nicolaides-Baraitser Syndrome. *Molecular Syndromology*, 2(6), pp.237–244.

Wolffe, A.P. & Hayes, J.J., 1999. Chromatin disruption and modification. *Nucleic Acids Research*, 27(3), pp.711–20.

Xu, L.-M. et al., 2012. AutismKB: an evidence-based knowledgebase of autism

genetics. *Nucleic Acids Research*, 40(D1), pp.D1016–D1022.

Yu, T.W. et al., 2013. Using Whole-Exome Sequencing to Identify Inherited Causes of Autism. *Neuron*, 77(2), pp.259–273.

Yu, Y. et al., 2015. De novo mutations in ARID1B associated with both syndromic and non-syndromic short stature. *BMC genomics*, 16(1), p.701.

Yuen, R.K.C. et al., 2015. Whole-genome sequencing of quartet families with autism spectrum disorder. *Nature Medicine*, 21(2), pp.185–191.

Zahir, F. et al., 2007. Novel deletions of 14q11.2 associated with developmental delay, cognitive impairment and similar minor anomalies in three children. *Journal of Medical Genetics*, 44(9), pp.556–61.

Zhang, Z. et al., 2016. Autism-Associated Chromatin Regulator Brg1/SmarcA4 Is Required for Synapse Development and Myocyte Enhancer Factor 2-Mediated Synapse Remodeling. *Molecular and Cellular Biology*, 36(1), pp.70–83.

Zoghbi, H.Y. et al., 1999. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature Genetics*, 23(2), pp.185–188.

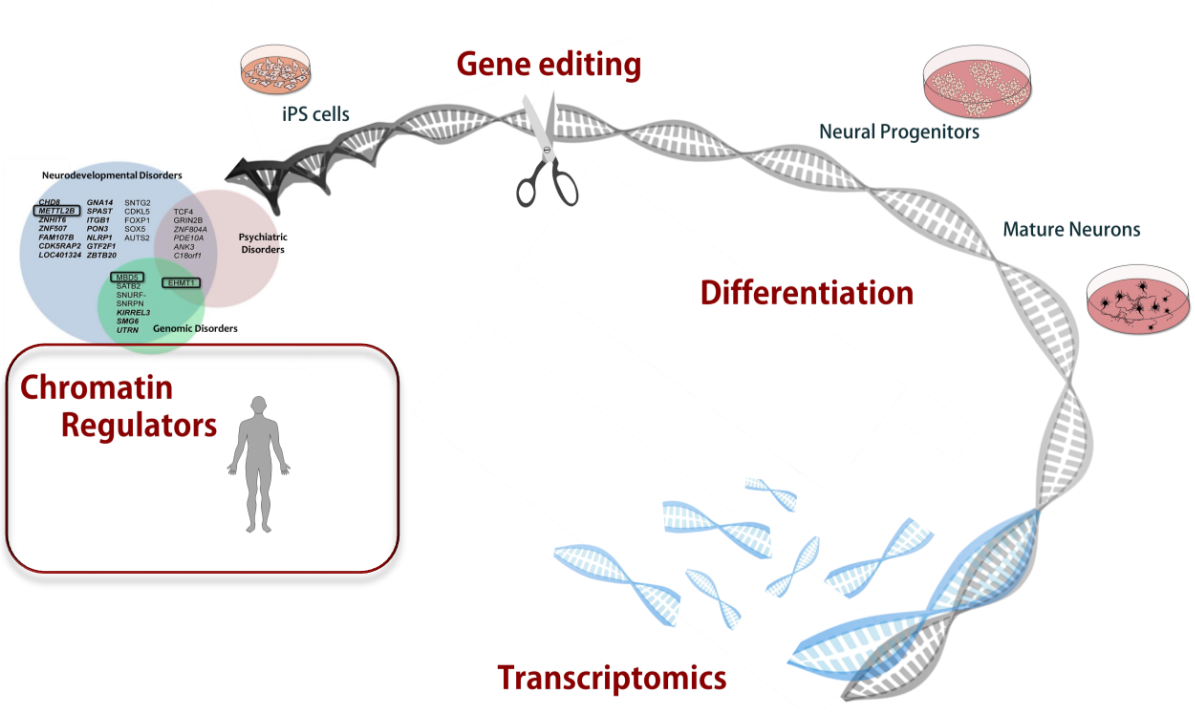
Zollner, S. & Pritchard, J.K., 2007. Overcoming the winner's curse: estimating penetrance parameters from case-control data. *American journal of human genetics*, 80(4), pp.605–15.

Zondervan, K.T. & Cardon, L.R., 2004. The complex interplay among factors that influence allelic association. *Nature Reviews Genetics*, 5(2), pp.89–100.

Zuk, O. et al., 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4), pp.1193–8.

Chapter 1

A Novel Microduplication of *ARID1B*: Clinical, Genetic and Proteomic Findings



Highlights

This chapter focuses on the chromatin remodeler, ARID1B, and the characterization of a unique microduplication of this gene found in a patient with intellectual disability, and proposes haploinsufficiency as the causal feature underlying the phenotype. This chapter was published as a manuscript in Seabra et al. 2017, AJMG Part A. The proteomics analyses were performed by collaborators, Nicholas Szoko and Marvin Natowicz.

Authors

Catarina M. Seabra[1,2,3], Nicholas Szoko[4], Serkan Erdin[2,3], Ashok Ragavendran[2,3], Alexei Stortchevoi[2], Patrícia Maciel[5,6], Kathleen Lundberg[7], Daniela Schlatzer[7], Janice Smith[8], Michael E. Talkowski[2,3,9], James F. Gusella[2,3,10] and Marvin R. Natowicz[4,11]*

Affiliations:

[1]GABBA - Institute of Biomedical Sciences Abel Salazar of the University of Porto, Portugal;

[2]Molecular Neurogenetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA;

[3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA;

[4]Cleveland Clinic Lerner College of Medicine, Cleveland, OH, USA;

School of Medicine, University of Minho, Braga, Portugal;

[6]ICVS/3Bs - PT Government Associate Laboratory, Braga/Guimarães, Portugal;

[7]Center for Proteomics, Case Western Reserve University School of Medicine, Cleveland, OH, USA;

[8]Baylor Genetics Laboratories, Baylor College of Medicine, Houston, TX, USA;

[9]Department of Neurology, Harvard Medical School, Harvard University, Cambridge, MA, USA;

[10]Department of Genetics, Harvard Medical School, Harvard University, Cambridge, MA, USA;

[11]Pathology & Laboratory Medicine, Genomic Medicine, Neurology and Pediatrics Institutes, Cleveland Clinic, OH, USA and Department of Pathology, Case Western Reserve University School of Medicine, Cleveland, OH, USA.

Abstract

Genetic alterations of *ARID1B* have been recently recognized as one of the most common mendelian causes of intellectual disability and are associated with both syndromic and non-syndromic phenotypes. The ARID1B protein, a subunit of the chromatin remodeling complex SWI/SNF-A, is involved in the regulation of transcription and multiple downstream cellular processes. We report here the clinical, genetic and proteomic phenotypes of an individual with a unique apparent *de novo* mutation of *ARID1B* due to an intragenic duplication. His neurodevelopmental phenotype includes a severe speech/language disorder with IQ scores within the normal range and scattered academic skill levels, expanding the phenotypic spectrum of *ARID1B* mutations.

Haploinsufficiency of *ARID1B* was determined both by RNA sequencing and quantitative RT-PCR. Fluorescent in situ hybridization analysis supported an intragenic localization of the *ARID1B* copy number gain. Principal component analysis revealed marked differentiation of the subject's lymphoblast proteome from that of controls. Of 3427 proteins quantified, 1,014 were significantly up- or down-regulated compared to controls ($q<0.01$).

Pathway analysis revealed highly significant enrichment for canonical pathways of EIF2 signaling, protein ubiquitination, tRNA charging and chromosomal replication, among others. Network analyses revealed downregulation of: (1) intracellular components involved in organization of membranes, organelles and vesicles; (2) aspects of cell cycle control, signal transduction and nuclear protein export; (3) ubiquitination and proteosomal function; and (4) aspects of mRNA metabolism. Further studies are needed to determine the detailed molecular and cellular mechanisms by which constitutional haploinsufficiency of *ARID1B* causes syndromic and non-syndromic developmental disabilities.

Keywords

ARID1B, SWI/SNF, SWI/SNF-A, chromatin, regulation, development, intellectual disability, proteome, proteomic

Introduction

Intellectual disability is characterized by significant limitations in cognitive functioning and adaptive behaviors (American Psychiatric Association, 2013) and affects 1–3% of the general population. Mutations of *ARID1B* (AT-rich interactive domain 1B) are an epidemiologically significant subset of mendelian causes of neurodevelopmental disability and are associated with non-syndromic intellectual disability as well as syndromic forms of intellectual disability such as Coffin-Siris syndrome (Santen and Clayton-Smith, 2014; Sim et al., 2015). The product of *ARID1B* is a ubiquitous nuclear-localized protein that is a subunit of SWI/SNF-A, a chromatin remodeling complex that contains over 25 core subunits and that is involved in the regulation of many biological processes, including regulation of transcription (Euskirchen et al., 2012). *ARID1B* mutations associated with intellectual disability include whole gene deletions, intragenic deletions, splice site, nonsense, and frameshift mutations, all of which point to haploinsufficiency as the mechanism causing the phenotype, as well as rare and less well-studied duplications (reviewed in Santen and Clayton-Smith, 2014; Sim et al., 2015). Here, we report the clinical, genetic, and proteomic findings of an individual having a unique loss-of-function mutation of *ARID1B* due to an intragenic duplication.

Clinical Report

This study was approved by the Institutional Review Board of the Cleveland Clinic. The subject is a 14-year-old male born at 39 weeks of gestation to a healthy primagravida 38-year-old mother and unrelated 49-year-old father with non-contributory family histories. There were no medical concerns during infancy. Early developmental milestones were met until 1-year of age but no use of sentences occurred until about 2.5-3 years old.

Physical examinations during early childhood noted language delays, borderline macrocephaly, strabismus, dysarthria, mild hypotonia and mild gross and fine motor incoordination. Diminished physical endurance was also apparent in early childhood and has persisted. He had prolonged recovery times from illnesses, including several developmental regressions that lasted two or more months between ages 7-9 years, as well as two episodes of difficulty recovering from general anesthesia at 3 and 6 years old. Ophthalmologic exam revealed a right optic nerve pit. Growth parameters at 6.5 years included head circumference 54.8 cm (98%), weight 22.1 kg (52%) and height 116.9 cm (38%). Physical exam at 12.4 years showed a non-dysmorphic male with weight 36.6 kg (18%) and height 140.3 cm (7%); neurological exam showed slowed processing to questions or directions, clumsy lateral tongue movements, abnormal gait with bilateral intoeing, mild imbalance, mild dysmetria, slow rapid alternating movements, clumsy fine motor function, and posturing of his arms and hands with stress maneuvers.

The earliest neuropsychological assessment, at 3.5 years, used the Stanford Binet 5 tool and showed full-scale IQ 98, verbal IQ 122 and non-verbal IQ 74. At 5 years, using the WPPSI-III, he was noted to have a full-scale IQ 83, verbal IQ 78 and non-verbal IQ 96, a poor fund of general knowledge and difficulty formulating and expressing verbal concepts. At 8 years, using the WISC-4, there was a full-scale IQ 78, with verbal comprehension 85, perceptual reasoning 102, working memory 86 and processing speed 50. His strongest skills related to nonverbal visual-spatial reasoning. There was slow processing of information, deficiency of working memory and poor visual/graphomotor skills. He was diagnosed with a severe language disorder. Auditory evaluation at 10 years showed an auditory processing disorder with severe difficulty in figure-ground discrimination, integration of words and sentences, temporal integration and phonological processing, and low average auditory comprehension and average auditory short-term memory. At 12.5 years,

using the Wide Range Achievement Test-4, he scored 101, 104, 91 and 67 in word reading, sentence comprehension, spelling and math computation, respectively.

Cranial MRIs at 6 months and 7 years of age showed a mildly dysmorphic corpus callosum. A 48-hour EEG at 5 years was unremarkable. Routine blood tests and urinalysis were normal. Normal metabolic testing included thyroid function tests, plasma amino acids, blood ammonia, and urinary organic acids, acylglycines, guanidinoacetate and creatine. Newborn screening, including for galactosemia, was negative. A fasting global plasma metabolomic analysis was unremarkable. The blood lactate and plasma butyrylcarnitine levels were intermittently increased. The fibroblast lactate:pyruvate ratio was normal, as were activities of fibroblast electron transport chain complexes II, III and IV and pyruvate dehydrogenase. A fibroblast loading study for defects of mitochondrial fatty acid beta-oxidation was negative. A lymphocyte cytogenetic analysis showed a normal 46,XY karyotype at the 550 band resolution. Whole exome sequencing did not reveal pathogenic variants that were likely or definitely related to the clinical phenotype.

Array CGH showed a copy number gain within chromosome band 6q25.3 of approximately 0.361 Mb in size, arr (GRCh37) 6q25.3(157133792-157495187)x3 dn. The duplication, which was not observed in either parent, involves a segment containing exons 2-10 (ENST00000346085) of *ARID1B* (Figure 4B). Array analysis also showed heterozygosity for an approximately 0.003 Mb maternally inherited copy number loss, arr (GRCh37) 9p13.3(34647598-34650608)x1 mat, including part of *GALT*. There was no clinical or biochemical evidence for galactosemia, nor was a mutation of the other *GALT* allele detected.

Materials and Methods

Molecular Cytogenomic Analysis

Array CGH was performed on a 400K oligonucleotide microarray (version 10.2) designed by the Medical Genetics Laboratories at Baylor College of Medicine and manufactured by Agilent Technologies (Santa Clara, CA). It includes exonic coverage of over 4000 candidate and disease genes at an average resolution of 30 kb with 60,000 SNP probes and 670 probes for the mitochondrial genome. Data was extracted using Agilent's Feature Extraction software (version 9.5.3.1) and was analyzed using a web-based software platform (Cheung et al., 2005; Lu et al., 2007).

Transcriptome Analysis

Gene expression was measured by RNAseq and quantitative RT-PCR. Total RNA was extracted from patient-derived EBV-transformed lymphoblastoid cell line (LCL), obtained at nearly 14 years of age, using TRIzol® (Invitrogen) followed by RNeasy Mini Kit (Qiagen) column purification. cDNA was synthetized from 1µg of extracted RNA using SuperScript® II Reverse Transcriptase (ThermoFisher Scientific) with oligo(dT), random hexamers, and RNase inhibitor. The RNAseq library was prepared using the Illumina TruSeq kit and manufacturer's instructions. Libraries were multiplexed, pooled and sequenced on multiple lanes of an Illumina HiSeq2500, generating an average of 30 million paired-end reads of 76 bp. Quality assessment of sequence reads was performed using fastQC (v. 0.10.1) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Sequence reads were then aligned to human reference genome Ensembl GRCh37 (v. 71) using GSNAP (v. 12-19-2014) (Wu and Nacu, 2010) at its default parameter setting. Quality checking of alignments was assessed by a custom script utilizing Picard Tools (http://broadinstitute.github.io/picard/), RNASeQC (DeLuca et al., 2012), RSeQC (Wang et al., 2012) and SamTools (Li et al., 2009). Novel transcript analysis was performed using Cufflinks, and visualized on Integrative Genomics Viewer (Robinson et al., 2011; Thorvaldsdóttir et al., 2013). Counts per Million, generated from gene level counts which were tabulated using BedTools's multibamcov algorithm (v. 2.17.0) (Quinlan and Hall, 2010) on unique alignments for each library at all Ensembl genes (GRCh37 v.71), were calculated to compare expression levels with control samples from six healthy individuals.

Quantitative RT-PCR

qRT-PCR was performed for *ARID1B* using custom designed primers and *ACTB*, *GAPDH* and *POLR2A* were used as endogenous controls. Primers were as described: *ARID1B* (forward - tgcgtcccctcatctctcca, reverse - aggcatctgactacctggga), *ACTB* (forward - tgaagtgtgacgtggacatc, reverse - ggaggagcaatgatcttgat), *GAPDH* (forward - ggacctgacctgccgtctag, reverse - gtagcccaggatgcccttga), *POLR2A* (forward - gcaccacgtccaatgacat, reverse - gtgcggctgcttccataa). Primers (0.75 µM final), cDNA (1:100 final) and nuclease-free water were added to the LightCycler® 480 SYBR Green I Master Mix (Roche) for a final 10µL reaction volume. A LightCycler® 480 (Roche) was used for data acquisition. Values for each subject or control were obtained in at least three technical replicates. Results of technical replicates for the gene of interest were normalized against the average of the three endogenous gene controls. Normalized expression levels were set in relation to seven age- and gender-matched controls, using the ΔΔCt method (Livak & Schmittgen 2001). Results are expressed as fold-change relative to the averaged control individuals. A two-tailed T-test was used to assess statistical significance.

Fluorescent in situ hybridization

Peripheral blood from the proband was collected in a sodium heparin vacutainer tube, cultured for 72 hours with the mitogen phytohemagglutinin, and harvested by standard cytogenetic methods. Slides containing both interphase and metaphase cells were hybridized according to standard protocol with fluorescently labeled BAC clones, RP11-680A17 and RP11-719E16, localized to 6q25.3 and 6q13, respectively. The BAC clones had been grown in broth medium with 20 µg/mL of chloramphenicol followed by DNA extraction using a Plasmid Miniprep kit (Qiagen). The target clone, RP11-680A17, was labeled directly with Spectrum GreenTM dUTP by nick translation with the Abbott Molecular kit. After hybridization and in accordance with the laboratory's standard confirmation protocol for duplications less than 1 Mb, 50 interphase cells were manually scored using a fluorescent microscope to confirm the presence of the duplication and a metaphase cell was examined to confirm the localization of the duplicated segment.

Proteome Analysis

LCLs from the subject, obtained at nearly 14 years of age, and 5 male controls ages 31-40 years were used. 15 μg of protein from each sample was digested with LysC for 1 hour and trypsin overnight at 37°C. Reverse phase LC-MS/MS was performed as described (Tomechko et al., 2015), except that 600 ng of peptide digests was loaded on the HPLC column and the gradient of solvent B ranged from 2 to 40% over 210 min.

Raw LC-MS/MS data files for each sample were processed using Rosetta Elucidator (Rosetta Biosoftware, Seattle, WA) (Version 3.3.01 SP4 25). Automated differential quantification of peptides was performed as previously described (Neubert et al., 2008; Schlatzer et al., 2012; Azzam et al., 2016). Briefly, LC-MS/MS raw data were imported, and for each MS spectrum profile of each LC-MS/MS run, chromatographic peaks and monoisotopic masses were extracted and aligned. Chromatographic peaks were first aligned by retention time and monoisotopic mass. Peaklists with the monoisotopic mass and corresponding MS/MS data were then generated for each sample and searched using Mascot. Resultant peptide identifications were imported into Elucidator and monoisotopic masses annotated with peptide identifications. The false discovery rate for protein identifications was calculated to be 0.02%. The MS/MS peak lists were searched by Mascot (version 2.4.1) (Matrix Science, London, UK) using the human UniProt database. Search settings were as follows: trypsin enzyme specificity; mass accuracy window for precursor ion, 25 ppm; mass accuracy window for fragment ions, 0.8Da; variable modifications including carbamidomethlylation of cysteines, 1 missed cleavage and oxidation of methionine.

Statistical Methods and Bioinformatic Analyses

Raw MS data were obtained for each region in a .csv file; this file contained intensity values, with rows corresponding to peptides and columns corresponding to the sample. Missing values were imputed using a weighted k-nearest neighbors algorithm (Troyanskaya et al., 2001). Next, data were $\log_2$-transformed to achieve normality. Data were visualized with principal component analysis and complete linkage hierarchical clustering. These preprocessing steps were performed using InfernoRDN (formerly DanTE) (Polpitiya et al., 2008). Data were imported into the R statistical programming environment for subsequent analyses.

Because there was one case and multiple controls, we treated individual peptides as observations of a given protein. Therefore, proteins with only one quantified peptide were excluded from our analysis. To compare protein abundance between the case and controls, we constructed a linear mixed effects model that adjusted for subject to account for the non-independence of peptides derived from the same sample. Proteins with $q < 0.01$ were imported into DAVID (https://david.ncifcrf.gov/tools.jsp) for ontology analysis. EASE score threshold was set at a value of 1.0 and minimum count for an annotation term was set to 3. The entire set of proteins with more than one peptide (n = 2,351) was used as background for enrichment analysis. Network and pathway analyses were performed in Ingenuity Pathway Analysis (IPA®, www.qiagen.com/ingenuity). Proteins with q-value < 0.01 were used for enrichment analysis. Network connections were curated based on data from all species and all cell lines and tissues. Enrichment scores and p-values for canonical pathway and network analysis were determined by a one-tailed Fisher's exact test.

Results

Measurement of *ARID1B* expression by quantitative RT-PCR showed a significant decrease of mRNA levels in the patient in comparison to control subjects (p-value 0,02) and the same trend was observed from the RNAseq dataset (Figure 4E). Allele-specific expression could not be assessed since the subject was not heterozygous for SNPs located in coding exons or in the untranslated regions. The most abundant mRNA transcript observed in the RNAseq dataset was ENST00000414678, consistent with the GTEx database for EBV-transformed lymphoblasts; ENST00000350026 and ENST00000346085 were also detected in the patient sample (Figure 4E). IGV Sashimi plots did not reveal novel junctions or transcripts for this gene (Figure 4C), providing no evidence that duplication of this exonic region resulted in alterations in splicing.

Follow-up fluorescent in situ hybridization analyses showed the presence of duplicated ARID1B genome in interphase cells (Figure 4D). The latter, in turn, was localized to chromosome 6q25.3 on analysis of a metaphase cell, the site of the *ARID1B* gene (Figure 4D), confirming that the microduplication occurred within the same chromosome and excluded the hypothesis of complex rearrangements.
Proteomic analyses of the LCLs resulted in the quantification of 15,792 peptides, corresponding to 3,427 proteins. There were 1,014 proteins that were differentially expressed between the case and controls ($q < 0.01$). The differentially expressed proteins and related data are noted in Table IV. Principal component analysis and hierarchical clustering analysis revealed marked separation of the subject's proteome from that of controls (Figure 5).

Ontologic analysis in DAVID revealed enrichment of several annotation clusters with terms relating to ATP binding, mitochondrion and flavin adenine dinucleotide (Supplementary Table 1). Bioinformatic analysis with IPA® revealed enrichment for canonical pathways of EIF2 signaling (p-value 6.43 E-17), protein ubiquitination (p-value 1.61 E-16), regulation of eIF4 and p70S6K signaling (p-value 1.04 E-14), tRNA charging (p-value 4.93 E-13) and cell cycle control of chromosomal replication (p-value 6.63 E-11) Table V). Downregulation of: (1) intracellular components involved in intracellular organization of membranes, organelles and vesicles; (2) aspects of cell cycle control, signal transduction and nuclear protein export; (3) ubiquitination and proteosomal function; and (4) aspects of mRNA metabolism are noted in IPA® network analyses (Supplementary Figure 1).

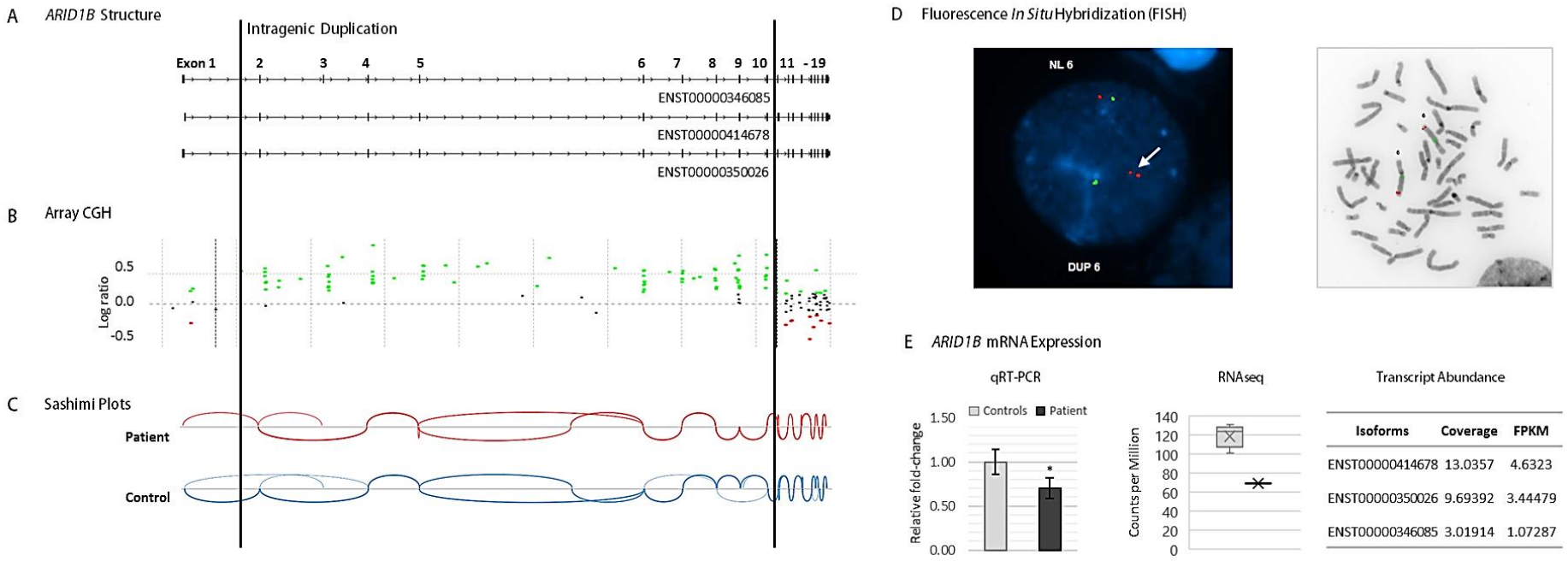


Figure 4 - Molecular and genetic characterization of the ARID1B microduplication.

A – Representation of the ARID1B transcripts expressed in the patient with exon numbering based on ENST00000346085.

B – Array CGH shows the duplicated region within the vertical lines comprising exons 2 to 10. (Green dots represent probes with log ratio above 0.2 and red dots those with log ratio below -0.2. Duplication or deletion is considered when probes are ≥ 0.5 or ≤ -0.5, respectively.)

C – Sashimi plots generated from the RNAseq datasets depict the splice junctions that have a minimum of 3 reads supporting each junction. No novel junctions are observed in the patient.

D – FISH confirms the duplication in interphase and metaphase cells, showing the duplication on chromosome 6 (as indicated by the white arrow) and not inserted into another chromosome.

E – (left to right) Decreased expression levels of ARID1B measured by qRT-PCR (using primers on exon 9 and on the junction of exons 10 and 11), in comparison to 7 age and gender-matched controls ($p_{val} < 0,05$) and by RNAseq in comparison to 5 age-matched controls. Error bars represent standard deviation. Transcript abundance analysis in the patient shows the 3 ARID1B expressed transcripts, measured from the RNAseq dataset using Cufflinks.

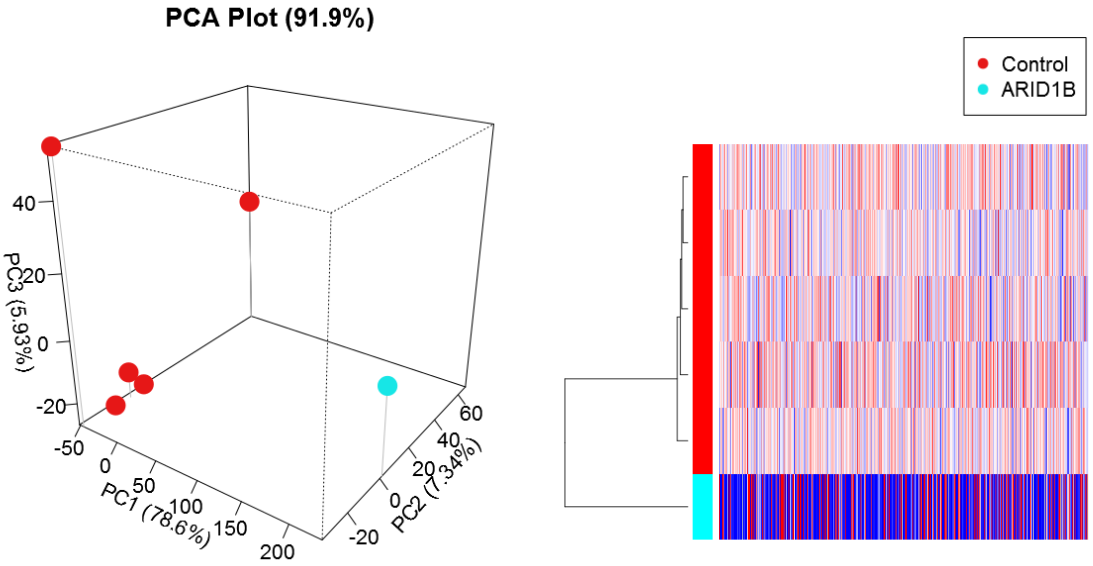


Figure 5 - Three-dimensional principal component analysis (left) and complete linkage hierarchical clustering (right) showing clear separation of ARID1B sample from controls.

Table IV - Bioinformatics analysis of differentially expressed proteins.

UNIPROT ID	GENE NAME	PERCENT EXPRESSION (ARID1B VS. CONTROL)	QVAL
A8MWD9	SNRPGP15	4	9.6E-06
Q6YP21	KYAT3	11	9.6E-06
Q5SRE5	NUP188	< 1	1.53E-05
O75915	ARL6IP5	9	3.08E-05
Q9UBW8	COPS7A	4	4.35E-05
O15260	SURF4	6	5.42E-05
P00395	MT-CO1	< 1	5.42E-05
Q9H061	TMEM126A	6	5.42E-05
P57772	EEFSEC	7	6.94E-05
P22392	NME2	2	8.92E-05
O75607	NPM3	6	9.37E-05
P25787	PSMA2	5	9.37E-05
P49643	PRIM2	4	9.37E-05
P60228	EIF3E	9	9.37E-05
Q00610	CLTC	16	9.37E-05
Q6L8Q7	PDE12	29	9.37E-05
Q8WWC4	C2orf47	9	9.37E-05
Q92616	GCN1	17	9.37E-05

Table V – Most significant pathways identified by Ingenuity Pathway Analysis (top 20 pathways).

INGENUITY CANONICAL PATHWAYS	P-VALUE
EIF2 Signaling	6.31E-17
Protein Ubiquitination Pathway	1.58E-16
Regulation of eIF4 and p70S6K Signaling	1.00E-14
tRNA Charging	5.01E-13
Cell Cycle Control of Chromosomal Replication	6.31E-11
RAN Signaling	3.31E-10
Remodeling of Epithelial Adherens Junctions	1.62E-09
Phagosome Maturation	3.55E-09
Mitochondrial Dysfunction	8.71E-09
mTOR Signaling	1.86E-08
TCA Cycle II (Eukaryotic)	5.13E-07
Apoptosis Signaling	7.76E-07
Oxidative Phosphorylation	3.39E-06
Granzyme B Signaling	4.07E-06
Purine Nucleotides De Novo Biosynthesis II	4.17E-06
VEGF Signaling	6.31E-06
FcÎ³ Receptor-mediated Phagocytosis in Macrophages and Monocytes	6.92E-06
Valine Degradation I	1.05E-05
Fatty Acid beta-oxidation I	1.20E-05
NA Double-Strand Break Repair by Non-Homologous End Joining	2.40E-05

Discussion

The subject has a unique *ARID1B* mutation, a copy number gain involving part of that gene that was initially determined by a chromosomal microarray analysis and which was thought to likely cause a pathologic reduction of *ARID1B* gene expression. Subsequent gene expression data, by both qRT-PCR and RNAseq, indicated haploinsufficiency of *ARID1B* and was consistent with localization of the duplication within the *ARID1B* gene, disrupting expression of the affected allele. Confirmation that there is a large intragenic *ARID1B* duplication was established by a fluorescent in situ hybridization analysis (Figure 1). Haploinsufficiency for *ARID1B* is likely to be the cause of developmental delay in this individual as no additional genetic or metabolic defects were identified. While the measurement of RNA in LCLs does not necessarily reflect the effect of the genetic lesion in the brain, the association of heterozygous inactivating mutations in *ARID1B* with neurodevelopmental phenotypes in other subjects suggests that reduced gene expression also occurs in the central nervous system. As *ARID1B* is expressed at different levels in the brain, namely higher in the cerebellum than in many peripheral tissues and other measured areas of the brain (*http://www.gtexportal.org/*), the consequences of its reduced expression may be even more pronounced there, possibly accounting for aspects of subject's phenotype.

There are two reports of smaller microduplications of *ARID1B*. The clinical and genetic phenotypes of these two individuals and our case are summarized in Supplementary Table 2. Our case has several relatively unique aspects to his phenotype. Few individuals with *ARIB1B* haploinsufficiency are reported with low-normal intellectual function (Santen et al., 2014). Our patient, while having a significant speech/language disorder and cognitive disability, has had several IQ determinations within the normal range but there was considerable (and reproducible) variability in selected skills. In addition, his clinical course has been characterized by easy fatigability and episodes of developmental regression, prompting evaluation for a possible metabolic underpinning of these clinical concerns. Apart from intermittent elevated blood lactate levels, extensive metabolic testing was unrevealing. To our knowledge, other cases of individuals with *ARID1B* haploinsufficiency do not have histories of episodic regression, although the episodic memory dysfunction in one case may be relevant (Yu et al., 2015).

Our results support the view that *ARID1B* is a dosage sensitive gene whose expression can be affected by deletions or duplications. Indeed, the evolutionary

constraint on this gene shows that it is highly intolerant to loss-of-function mutations according to its ExAC pLI score of 1.00 (http://exac.broadinstitute.org/). As we did not identify novel transcripts, the duplicated allele probably resulted in premature termination of transcription. Indeed, if the duplicated exons 2-10 were spliced into the mRNA downstream of the original exon 10, the resultant frameshift after codon 1,008 would add 19 novel amino acids, followed by a premature stop codon, to produce a truncated protein of 1,027 amino acids, that would explain the observed haploinsufficiency. Other chromatin regulators have also been noted to be dosage sensitive causes of neurodevelopmental phenotypes, including *MBD5*, *EHMT1*, *CHD8* and *SATB2* (Talkowski et al., 2012).

The lymphoblast proteomic data are striking and consistent with known roles of ARID1B and the SWI/SNF-A chromatin remodeling complex. Statistically highly significant dysregulation of pathways related to gene transcription were noted, in addition to dysregulation of pathways and networks relating to protein ubiquitination, mTOR signaling, signaling of apoptosis, major metabolic pathways (TCA cycle, mitochondrial oxidative phosphorylation), cell cycle control, mRNA metabolism and intracellular vesicular transport and show considerable overlap with ARID1B-regulated pathways (such as the transcription and cell cycle regulation pathways) and genes noted by others (Euskirchen et al., 2011; Sim et al., 2014; Raab et al., 2015). Yet, although our proteomic analysis indicates widespread and marked differential expression of proteins in the *ARIDIB* haploinsufficient lymphoblasts and the data are consistent with ARID1B-associated observations noted in other systems, there are three important potential limitations of our proteomic findings: first, one cannot generalize these results until similar proteomic analysis of additional cases of ARID1B haploinsufficiency are carried out; second, the proteomic findings, while predictive of marked dysregulation of multiple pathways and networks does not reveal which dysregulated pathways are of greatest pathophysiological significance; and, third, the relevance of lymphoblast findings to brain biology is uncertain.

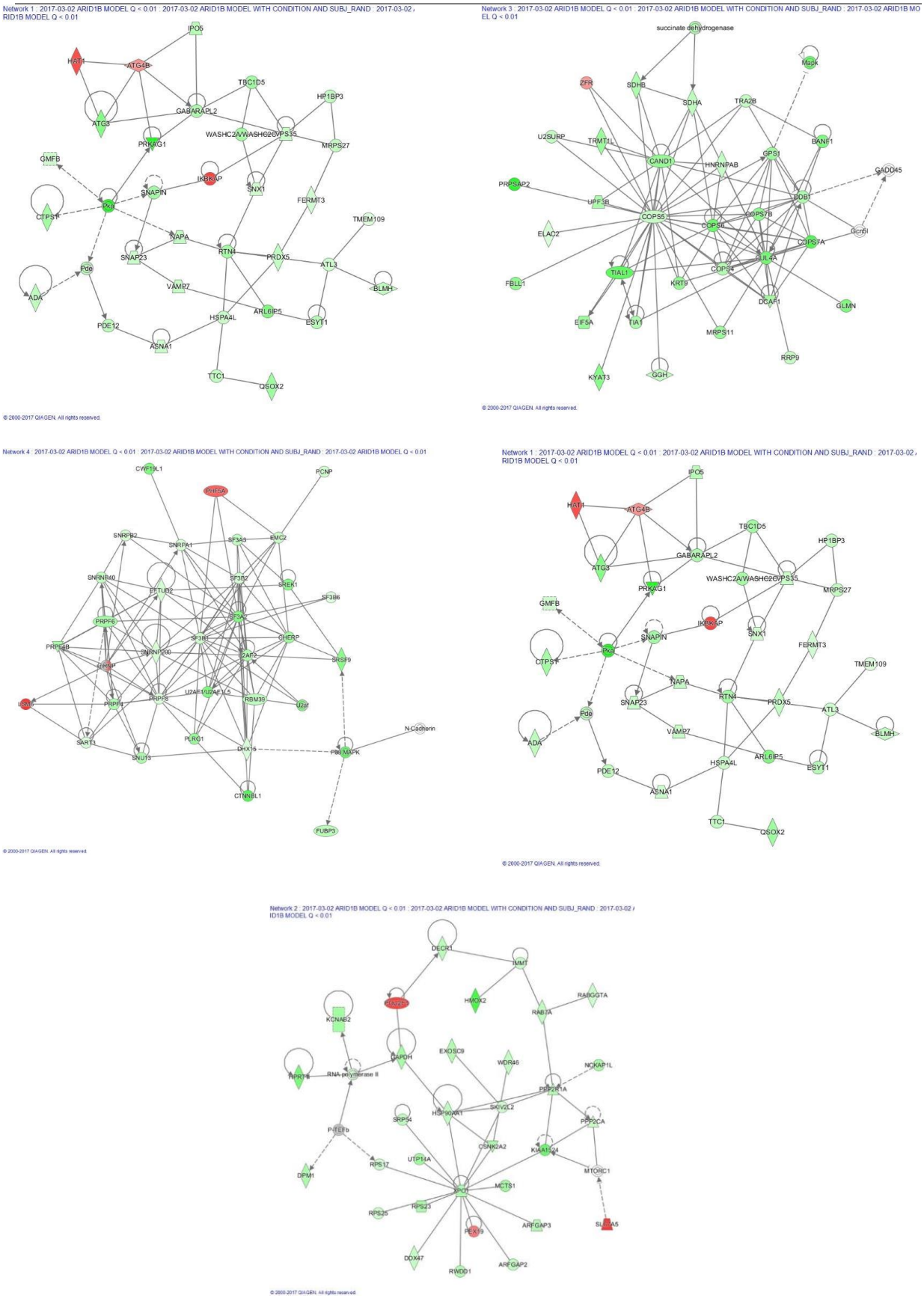
The roles of *ARID1B* in brain development are starting to be understood. Recent *in vitro* studies also support the fact that this protein may be critical in cell proliferation and differentiation and in dendritic arborization and synapse formation (Nagl et al., 2007; Yan et al., 2008; Tuoc et al., 2013; Harmacek et al., 2014). Therefore, ARID1B deficiency may lead to neurodevelopmental defects through the defective differentiation of mature neurons (Ka et al., 2016). This report added further insight on the role of ARID1B haploinsufficiency in the establishment of intellectual disability, yet further studies are required to uncover the precise mechanisms

whereby altered transcriptional and cell cycle regulation pathways lead to impaired brain development and cognitive function.

Acknowledgements

We thank the family for participating in this case report. Funding was provided by FCT Fellowship SFRH/BD/52049/2012 to CMS, NIH grant GM061354 to JFG and MET, SFARI grant 308955 to JFG and R00MH095867 to MET and Autism Research Institute grant to MRN. The authors have no conflicts of interest regarding this work.

Supplementary Data



Supplementary Figure 1 - IPA Networks.

Supplementary Table 1 - DAVID Analysis.

Annotation Cluster 1	**Enrichment Score: 2.97**	
Category	Term	p-value
UP_KEYWORDS	ATP-binding	3.16E-05
GOTERM_MF_DIRECT	GO:0005524~ATP binding	8.14E-05
UP_SEQ_FEATURE	nucleotide phosphate-binding region:ATP	0.001131
UP_KEYWORDS	Nucleotide-binding	0.00171
INTERPRO	IPR027417:P-loop containing nucleoside triphosphate hydrolase	0.285095
Annotation Cluster 2	**Enrichment Score: 2.47**	
Category	Term	p-value
GOTERM_CC_DIRECT	GO:0005739~mitochondrion	7.51E-05
UP_SEQ_FEATURE	transit peptide:Mitochondrion	0.004258
UP_KEYWORDS	Transit peptide	0.005554
UP_KEYWORDS	Mitochondrion	0.008822
GOTERM_CC_DIRECT	GO:0005759~mitochondrial matrix	0.029901
Annotation Cluster 3	**Enrichment Score: 1.82**	
Category	Term	p-value
UP_KEYWORDS	FAD	0.011293
UP_KEYWORDS	Flavoprotein	0.012159
GOTERM_MF_DIRECT	GO:0050660~flavin adenine dinucleotide binding	0.018069
UP_SEQ_FEATURE	nucleotide phosphate-binding region:FAD	0.020543

Supplementary Table 2 - Clinical findings in cases having an exonic duplication of ARID1B.

	THIS STUDY	HOYER *ET AL* 2012	YU *ET AL* 2015
GENDER	Male	Female	Female
INHERITANCE	*De novo*	*De novo*	Maternal
DUP LOCATION (ENST00000346085)	Exons 2 to 10	Exons 5 and 6	Exons 5 to 9
***ARID1B* EXPRESSION**	Decreased	Monoallelic	Not assessed
ID/DD	+	+	+
AUTISM	-	-	-
SPEECH DELAY	+	+	+
SEIZURES	-	-	-
HYPOTONIA	+	+	+
OTHER NEUROLOGICAL	Gross fine and motor incoordination, imbalance, strabismus	Ataxic gait	Esotropia, episodic memory lapse
CRANIAL MRI	Dysmorphic corpus callosum	Delayed myelination	Normal corpus callosum
CRANIOFACIAL	Borderline macrocephaly	Microcephaly, plagiocephaly, frontal bossing, high palate	Macrocephaly
OTHER FEATURES	Optic nerve pit	Heart malformation, sparse hair, brachydactyly hemangiomas, sacral dimple	Short stature, kyphoscoliosis,

References

American Psychiatric Association, 2013. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)* 5th ed.,

Azzam, S. et al., 2016. Proteome and Protein Network Analyses of Memory T Cells Find Altered Translation and Cell Stress Signaling in Treated Human Immunodeficiency Virus Patients Exhibiting Poor CD4 Recovery. *Open Forum Infectious Diseases*, 3(2), p.ofw037.

DeLuca, D.S. et al., 2012. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics (Oxford, England)*, 28(11), pp.1530–2.

Euskirchen, G., Auerbach, R.K. & Snyder, M., 2012. SWI/SNF Chromatin-remodeling Factors: Multiscale Analyses and Diverse Functions. *Journal of Biological Chemistry*, 287(37), pp.30897–30905.

Harmacek, L. et al., 2014. A unique missense allele of BAF155, a core BAF chromatin remodeling complex protein, causes neural tube closure defects in mice. *Developmental neurobiology*, 74(5), pp.483–97.

Ka, M. et al., 2016. Essential Roles for ARID1B in Dendritic Arborization and Spine Morphology of Developing Pyramidal Neurons. *The Journal of Neuroscience,* 36(9), pp.2723–42.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp.2078–2079.

Livak, K.J. & Schmittgen, T.D., 2001. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods*, 25(4), pp.402–408.

Nagl, N.G. et al., 2007. Distinct mammalian SWI/SNF chromatin remodeling complexes with opposing roles in cell-cycle control. *The EMBO Journal*, 26(3), pp.752–63.

Neubert, H. et al., 2008. Label-Free Detection of Differential Protein Expression by LC/MALDI Mass Spectrometry. *Journal of Proteome Research*, 7(6), pp.2270–2279.

Quinlan, A.R. & Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), pp.841–842.

Robinson, J.T. et al., 2011. Integrative genomics viewer. *Nature biotechnology*, 29(1), pp.24–6.

Santen, G.W.E. & Clayton-Smith, J., 2014. The *ARID1B* phenotype: What we have learned so far. *American Journal of Medical Genetics Part C: Seminars in Medical*

Genetics, 166(3), pp.276–289.

Schlatzer, D.M. et al., 2012. A quantitative proteomic approach for detecting protein profiles of activated human myeloid dendritic cells. *Journal of Immunological Methods*, 375(1-2), pp.39–45.

Sim, J.C.H., White, S.M. & Lockhart, P.J., 2015. ARID1B-mediated disorders: Mutations and possible mechanisms. *Intractable & rare diseases research*, 4(1), pp.17–23.

Talkowski, M.E. et al., 2012. Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell*, 149(3), pp.525–37.

Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2), pp.178–92.

Tuoc, T.C. et al., 2013. Chromatin Regulation by BAF170 Controls Cerebral Cortical Size and Thickness. *Developmental Cell*, 25(3), pp.256–269.

Vasileiou, G. et al., 2015. Chromatin-Remodeling-Factor ARID1B Represses Wnt/β-Catenin Signaling. *The American Journal of Human Genetics*, 97(3), pp.445–456.

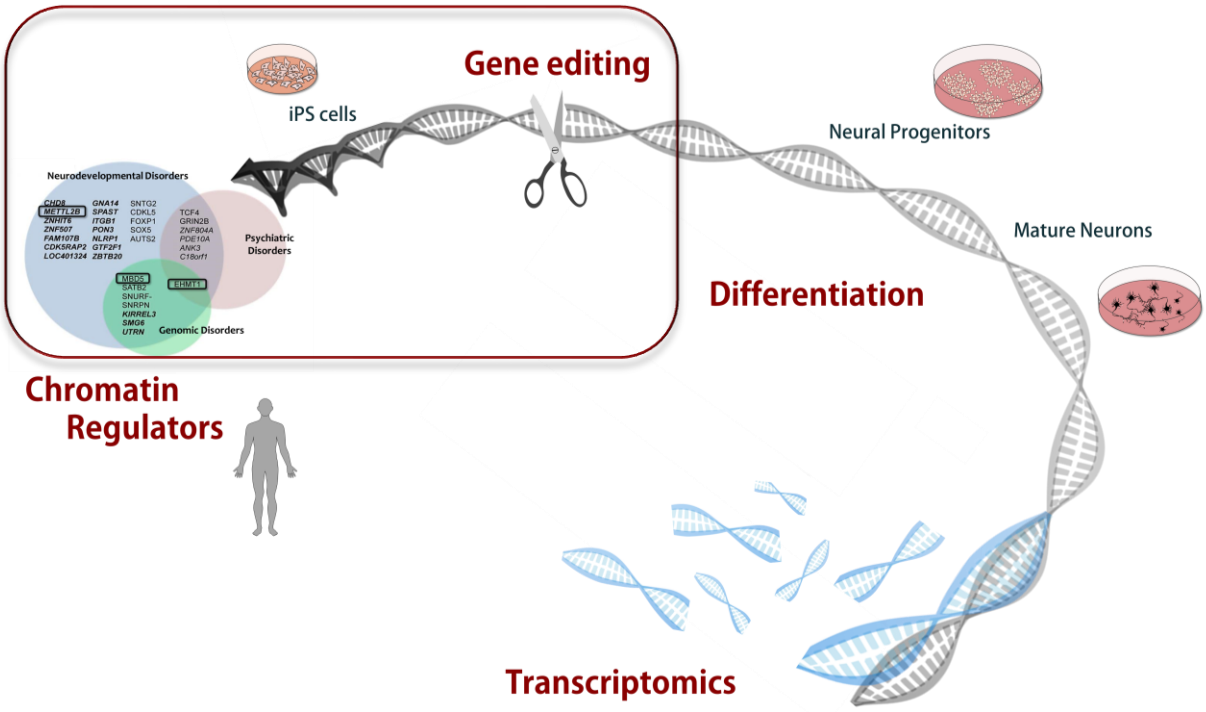
Wang, L., Wang, S. & Li, W., 2012. RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, 28(16), pp.2184–2185.

Wu, T.D. & Nacu, S., 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics (Oxford, England)*, 26(7), pp.873–81.

Yan, Z. et al., 2008. BAF250B-associated SWI/SNF chromatin-remodeling complex is required to maintain undifferentiated mouse embryonic stem cells. *Stem cells (Dayton, Ohio)*, 26(5), pp.1155–65.

Chapter 2

CRISPR-edited iPSC models of neurodevelopment



Highlights

In the second chapter, we highlight the need to develop accurate models to study neurodevelopmental disorders. Using the novel CRISPR/Cas9 genome engineering technology we sought to solve this issue by generating in vitro models of loss-of-function of genes involved in chromatin remodeling that confer substantial risk for autism spectrum disorder and other neurodevelopmental anomalies.

Authors

Catarina M. Seabra[1, 2, 3], Derek J. C. Tai[2, 3], Poornima Manavalan[2], Celine de Esch[2, 3], Patrícia Maciel[4, 5], Michael E. Talkowski[2, 3, 6] and James F. Gusella[2, 3, 7]

Affiliations:

[1]GABBA Program, Institute of Biomedical Sciences Abel Salazar of the University of Porto, Portugal;
[2]Molecular Neurogenetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA;
[3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA;
[4]School of Medicine, University of Minho, Braga, Portugal;
[5]ICVS/3Bs - PT Government Associate Laboratory, Braga/Guimarães, Portugal;
[6]Department of Neurology, Harvard Medical School, Harvard University, Cambridge, MA, USA;
[7]Department of Genetics, Harvard Medical School, Harvard University, Cambridge, MA, USA.

Abstract

Given the rapid pace of discoveries showing that de novo loss-of-function (LoF) mutations in highly conserved genes that are evolutionarily constrained, or intolerant to LoF mutations, represent penetrant sources of genetic risk in autism spectrum disorder (ASD), it is imperative to generate robust models that recreate the human cellular landscape. This is particularly relevant for neurological disorders where brain tissue is not readily available and the relevance of lymphoblast findings (from blood samples) to brain biology is uncertain. Given the limited number of available subjects in most studies, genetic background can introduce significant confounding effects in interpretation. To overcome these barriers, we developed an isogenic *in vitro* modeling approach using an induced pluripotent stem cell (iPSC) line from a healthy male subject to perform precise dual-guide CRISPR/Cas9 gene editing of four independent genes for which LoF mutations represent strong risk factors for ASD (*EHMT1*, *MBD5*, *METTL2A* and *METTL2B*). This will allow the investigation of the role of haploinsufficiency of these chromatin remodelers in the etiology of ASD.

The efficiency of 10 dual guide-RNA combinations to generate deletions was assessed using FACS sorting. Dual guide CRISPR successfully generated deletions in all genes; however, the efficiency varied widely by guide-RNA combination. The overall efficiency of the FACS method was 2,8% to generate predicted ablations, out of 1002 colonies screened (range = 0% - 10.6%). This systematic survey of genome editing approaches suggests that dual-guide deletion generation varies widely by guide-pair. We also found that the increased certainty of deriving a single cell from FACS sorting comes at a significant cost in terms of efficiency and cell viability.

Keywords:
CRISPR, Cas9, models, neurodevelopment, loss-of-function, efficiency, iPSC.

Introduction

The prevailing hypothesis of autism spectrum disorder (ASD) pathology has historically focused on defective synaptic pathways and neuronal circuits. Recently, autism genes that confer substantial risk for ASD have been identified by sequencing the breakpoints of balanced chromosome exchanges in ASD subjects and the coding sequencing in ASD families, pointing the field in a novel direction as they revealed highly penetrant genes involved in chromatin remodeling and transcriptional regulation (Talkowski et al. 2012; O'Roak et al. 2011; Sanders et al. 2012). Indeed, a growing number of transcription factors have been implicated in neurodevelopmental disorders, either through specific syndromes or based on rare mutations in idiopathic cases of common disorders such as ASD. Some of their neural effects may occur from dysregulation of genes controlling early neurodevelopmental processes such as cell migration, axon guidance, or synapse formation, but several also play roles in activity-dependent gene expression and synaptic plasticity. It is now clear that, in addition to direct synaptic disruption, the genetic contribution to ASD acts through alterations in chromatin regulatory mechanisms in human brain development and function (Krumm et al. 2014; De Rubeis et al. 2014; Pinto et al. 2014).

Most of the knowledge about the central nervous system (CNS) and neural function in patients with neurological diseases has been obtained from postmortem tissues that often represent the end stage of the disease. The inability to sample live CNS tissues impedes our progress to understand aspects of the neuropathological abnormalities that develop during the course of diseases. Animal models can mimic genetic forms of human neurological diseases, and our understanding of the mechanisms of neurological diseases has been significantly advanced with transgenic technologies to interrogate synaptic circuits and behavior (Shinoda et al. 2013). Invertebrate models then appeared to provide a more straightforward alternative, due to their rapid generation time, low maintenance cost, simplified genetic manipulation when compared to vertebrates, such as *Xenopus* (Pratt & Khakhalin 2013), *C. elegans* (Bessa et al. 2013), *D. melanogaster* (Furukubo-Tokunaga 2009)

In many cases of neurological disorders with a defined causal gene(s), however, modeling with animal transgenic technology is inadequate due to species differences, genetic backgrounds, or other technical challenges (Cundiff & Anderson 2011; Mattis & Svendsen 2011; Wichterle & Przedborski 2010). The failure of translation to the clinic stems from the complexity of the human brain and the difficulty to model disease specific phenotypes in non-human systems (Wichterle & Przedborski 2010).

This situation indicates that an advancement towards more human relevant models is required to accurately study neurogenetic disorders.

Given the rapid pace of discoveries showing that *de novo* LoF mutations in highly conserved genes represent penetrant sources of genetic risk in ASD, it is imperative to generate robust models that recreate the human cellular landscape. This is particularly relevant as ASD are neurological disorders where brain tissue is not readily available and the relevance of lymphoblast findings (from blood samples) to brain biology is uncertain, it is necessary to eliminate the confounding effect of different genetic backgrounds from patient cell lines. In order to address this issue, efforts have been made towards generating human iPSC models bearing the desired mutations that will allow multiple comparisons, and the contribution of single genes to the pathogenesis of ASD (Wang & Doering 2012).

Seminal work by Takahashi and Yamanaka showed that retroviral expression of a set of four genes (*Oct4*, *Sox2*, *Klf4*, and *c-Myc*) can convert somatic cells into a pluripotent state (Takahashi et al. 2007). Induced pluripotent stem cells (iPSCs) can be driven to differentiate into neurons and glial cells, as well as other terminally differentiated cell types by exposure to a combination of growth factors and cell culture conditions (Denham & Dottori 2011; Dhara & Stice 2008). Therefore, human iPSCs make it possible to study human CNS neuronal lineages. Indeed, disease-specific iPSC lines have been generated from patients with neurodevelopmental diseases including Rett syndrome, Fragile X syndrome, Down syndrome, Angelman syndrome, Prader-Willi syndrome, and Timothy syndrome. The iPSC based models of neurodevelopmental disorders recapitulate the early steps in neural development allowing for isogenic backgrounds that may help to identify the contribution of a single mutation or gene to the underlying neurobiological pathways affected.

Haploinsufficiency of dosage-sensitive chromatin remodelers, as *CHD8* (Sugathan et al. 2014) and *ARID1B* (Sim et al. 2015) and *MBD5* (Talkowski et al. 2011) has been pinpointed as a causal mechanism of the pathogenesis of some cases of ASD. To model loss-of-function (LoF) mutations of chromatin remodelers, it is possible to perform genome editing on iPSC to obtain allelic series of LoF mutations in candidate genes. Gene targeting in human iPSC has been proven to be a challenge (Zwaka & Thomson 2003). Zinc-finger nucleases (ZFNs) and transcription activator-like endonucleases (TALENs) have been applied to gene manipulation of human iPSC (Hockemeyer et al. 2009; Hockemeyer et al. 2011; Zou et al. 2009), however, both technologies require the design of proteins and intricate construction of plasmids for expression of those proteins. These methods were time-consuming and labor-

intensive. The CRISPR/Cas9 gene editing system has shown to be a promising and highly effective technique for modifying the genome of higher eukaryotes, particularly of mammalian cells that have not been readily amenable to gene editing, as human stem cells (Mandal et al. 2014).

In the short period since the initial discovery of efficacy in mammalian cells (Cong et al. 2013; Mali et al. 2013; Jinek et al. 2013; Ran et al. 2013), there have been numerous studies that demonstrate the utility of the CRISPR/Cas9 system for genome engineering, from performing whole-genome-scale knockout screens elucidating gene function in cell culture (Shalem et al. 2014; Wang et al. 2014) and creating mutations in the brains of adult mice (Swiech et al. 2015). The CRISPR system requires the Cas9 nuclease along with two short RNA molecules, the guiding CRISPR RNA (crRNA) and the trans-activating crRNA (tracrRNA), that hybridize to each other and direct Cas9 to the target location for cleavage based on sequence complementarity to the crRNA as well as proximity of the DNA target to a protospacer adjacent motif (PAM), NGG (Figure 6). In 2012, Jinek et al. demonstrated that the two small RNAs, crRNA and tracrRNA,

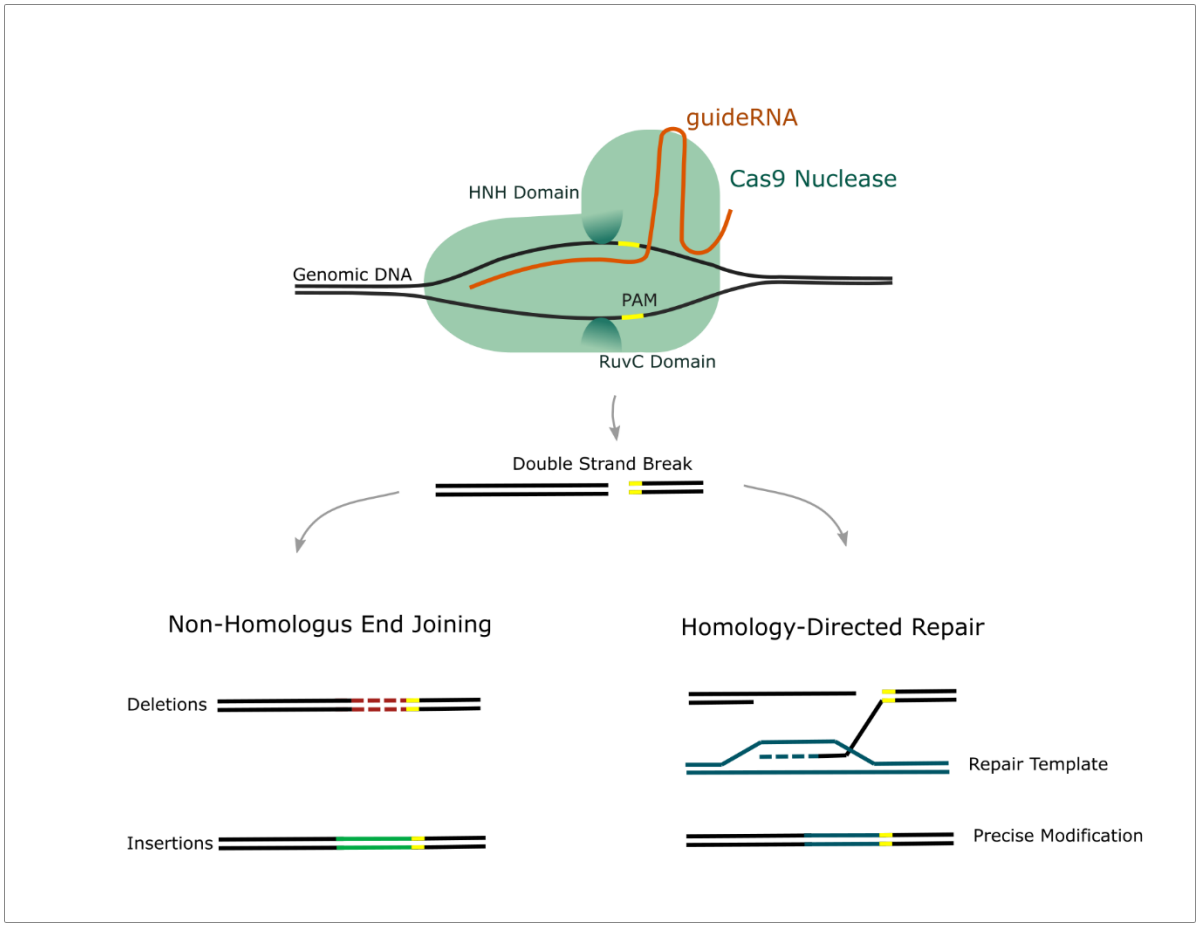


Figure 6 – CRISPR/Cas9 gene editing technology. The Cas9 nuclease introduces double stranded breaks at the target sequence that is located near a PAM motif. The break is repaired by one of two mechanisms: Left) Non-homologous end joining (NHEJ) which creates random insertions or deletions (indels) at the targeted site; Right) Homologous recombination which creates precise changes based on template DNA.

can be combined into a single guide RNA (guideRNA) that can be expressed in mammalian cells. The ability of any given guideRNA to efficiently create a double strand break (DSB) in the target DNA can vary based on the guide RNA sequence and the position in the targeted gene (Shalem et al. 2014; Wang et al. 2014; Koike-Yusa et al. 2014). Besides, the use of small guide RNAs for gene editing makes the CRISPR system attractive, because there is no requirement for protein design or construction of expression plasmids.

Haploinsufficiency, a feature by which chromatin regulators have been found to cause ASD, is often caused by a loss-of-function mutation, in which having only one copy of the wild-type allele is not sufficient to produce enough protein to display the wild type's phenotypic characteristics. To mimic the haploinsufficiency found in ASD patients, we will generate LoF mutations in chromatin-related genes that confer risk for ASD: *EHMT1*, *MBD5*, *METTL2A* and *METTL2B*, by knocking down one of the alleles of each gene in induced pluripotent stem cells using CRISPR/Cas9 technology. This approach will allow the investigation of the mechanisms through which these chromatin remodelers lead to ASD pathogenesis. We show the efficiency and the overall performance of CRISPR in these *in vitro* human cell models. Indeed, the iPSC technology has opened new windows for modeling human diseases, identifying therapeutic targets, developing drug screening systems, and providing continuous autologous cell sources with potential for cell therapies.

Methods

GuideRNA Design and Preparation

The CRISPR/Cas9 gene editing system was used to create iPSC lines with mutations in 4 independent genes for which LoF mutations represent strong risk factors for ASD (*EHMT1*, *MBD5*, *METTL2A* and *METTL2B*). We used a dual-guideRNA strategy for excision of a DNA fragment as this can be easily screened via standard PCR. The guideRNAs were designed, based on genome assembly GRCh37, either outside the exon boundaries in order to excise the entirety of the exon, or within the exon itself to create frameshift or truncating mutations (Figure 8). The exon to target was selected based on proximity to the transcriptional start site. This is expected to prevent any transcription from that allele by disrupting the initial region of the gene, and this would ultimately lead to haploinsufficiency of the chromatin regulators, as observed in ASD patients.

To assure highly specific and effective guideRNAs, online tools were considered: (i) CRISPR design tool (http://tools.genome-engineering.org/) that takes into account off-target predicted sites and gives a score inversely correlated with the number of off-target matches; (ii) sgRNA Designer (http://www.broadinstitute.org/rnai/public/analysis-tools/sgrna-design) which accounts for on-target efficiency (Doench et al. 2014) and (iii) BLAST (NCBI) to query for the guide sequences, including the PAM motif, to determine whether the sequences target uniquely to the desired regions or if there are any potential off-targets (Supplementary Table 3). The position of the guideRNAs designed are represented in Figure 7.

For subsequent transfection in human embryonic kidney cells (HEK293T) and nucleofection in iPSC cells, guideRNAs were cloned into the guideRNA cloning vector *pGuide* (Addgene plasmid 41824), using a *BbsI* restriction site (Ran et al. 2013). To confirm the correct sequence of the guides within the vectors DNA was isolated using the Miniprep Kit (Qiagen), and the guides were Sanger sequenced, using a T7 primer. To obtain transfection-grade DNA, all plasmids were purified from Plasmid Plus Midi Kit according to the manufacturer's instruction (Qiagen), using One Shot® Stbl3™ Chemically Competent *E. coli*.

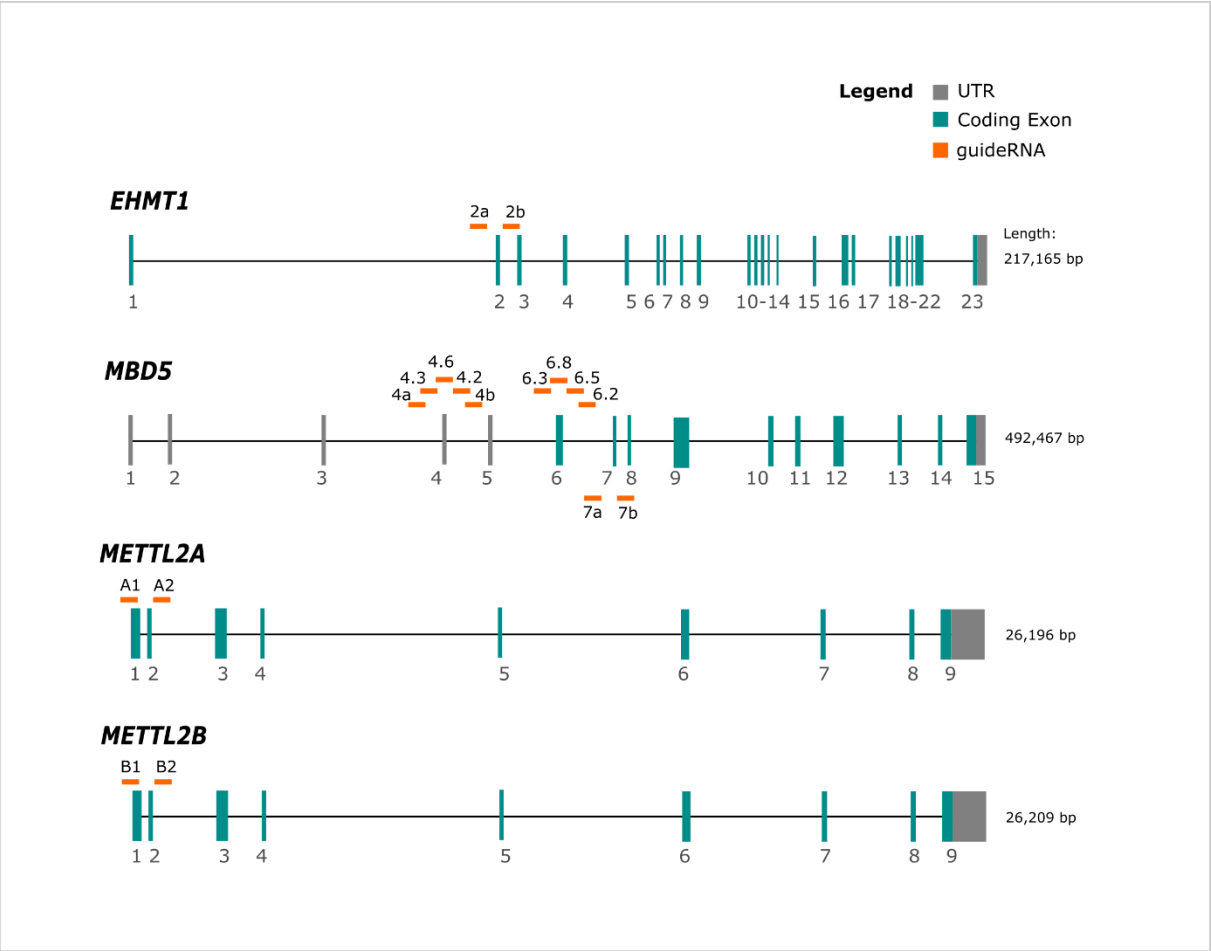


Figure 7 - guideRNA location for each target gene.

Transfection in Human Embryonic Kidney Cells

HEK293T cells were maintained, following standard protocols, with supplemented DMEM medium (Dulbecco's modified Eagle's medium, 10% Fetal bovine serum, 100 units/mL Penicillin G, 100 µg/mL Streptomycin) and incubated at 37 ˚C in a humidified atmosphere with 5% CO2. Cells were plated at a density of 1 – 1.25 x 10^5 cells per well in a 24-well cell culture plate to obtain a 90% confluency after 24 hours. Lipofectamine® 3000 was used for transfection of the guideRNAs inserted onto pGuide along with pCas9_GFP (Addgene 44719), according to the manufacturer's

Figure 8 - GuideRNA design strategy – A) outside the exon boundaries in order to excise the entirety of the exon, or B) within the exon itself to create frameshift or truncating mutations. Scissors represent each Cas9+guideRNA complex that will cleave the DNA in the desired location.

instructions (1 µL of each vector 250 ng/µL) to have a total of 750 ng of DNA per well). Having independent guideRNA cloned onto pGuide, allowed for the combination of different guides, without increasing the DNA load, as the pCas9_GFP vector (9271 bp) is much larger than the pGuide (3915 bp). After 48 hours, the cells were observed under a fluorescence microscope to detect GFP fluorescence (indicative of Cas9-GFP expression) and the medium was supplemented with 3 µg/mL of puromycin for another 48h, after which the medium was switched back to mTeSR1 and DNA was harvested. Puromycin concentration was determined by calculating the puromycin viability curve for HEK293T cells and the optimal concentration at which there was most cell death post-48h was at 3µg/mL (Supplementary Figure 2). Genomic DNA from the pool of cells was extracted using a rapid DNA extraction method (McClive & Sinclair 2001). Briefly, cells were lysed by adding DNA extraction buffer containing Proteinase K (0.2 mg/ml). Samples were digested at 55 ˚C for at least 1 hour followed by Proteinase K inactivation at 95 ˚C for 10 min. Surveyor® Mutation Detection assay (IDTdna) was then used, on the first subset of guideRNAs (indicated on Table VI), following manufacturer's instructions. Both uncut and cut products were ran in an agarose gel to determine the guideRNAs with highest efficiency (Supplementary Figure 3). Band intensity was analyzed using ImageJ Software (http://imagej.nih.gov/ij/).

Nucleofection in Human Induced Pluripotent Stem Cells

Human induced pluripotent stem cells (iPSCs), derived from fibroblasts from a healthy male individual, identified as 8330-8 cells, were generated using standard retroviral vectors and the Yamanaka factors OCT3/4, SOX2, KLF4, and c-MYC (Sheridan et al. 2011). The iPSCs were maintained on Corning® Matrigel® hESC-qualified Matrix-coated dishes with mTeSR™1 medium (STEMCELL Technologies), supplemented with 1% of Penicillin and Streptomycin and incubated at 37 ˚C in a humidified atmosphere with 5% CO_2. The iPSCs (1×10^6cells) were transfected with 1 µg total DNA plasmid, Cas9 expression vector *pX459* (pSpCas9(BB)-2A-Puro plasmid Addgene 48139) along with the chosen guideRNAs (inserted into *pGuide* - Addgene plasmid 41824) and an external EGFP (enhanced green fluorescent protein) vector (Figure 9). For nucleofection of the guideRNAs into the iPSC, the Human Stem Cell Nucleofector Kit 1 (Lonza) and Amaxa Nucleofection II device (Lonza) were used with program B-016, according to the manufacturer's instructions. After nucleofection, the iPSCs were cultured on Matrigel-coated wells using conditioned mTeSR medium supplemented with 10 µM ROCK inhibitor (Y-27632 dihydrochloride, Santa Cruz

Biotech) and 10 ng/ml bFGF (R&D). Treatment with ROCK inhibitor blocks apoptosis of dissociated cultured iPSC, increases survival and cloning efficiency of iPSC, without affecting their pluripotency.

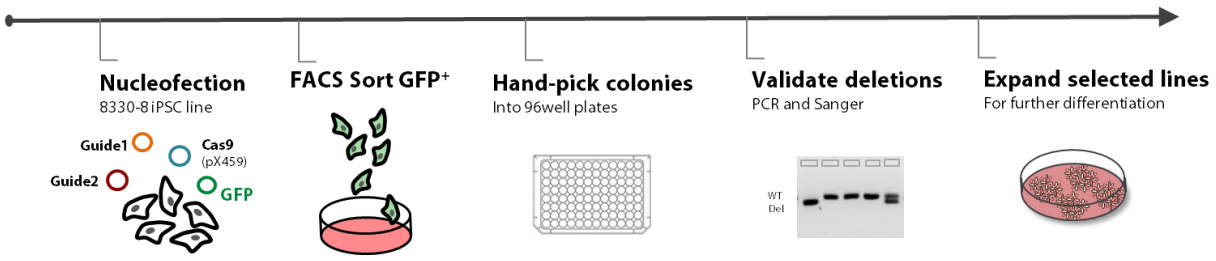


Figure 9 - Workflow of genome engineering and screening in the iPSCs.

Single-cell isolation by fluorescence-activated cell sorting (FACS)

To obtain isogenic iPSC colonies following CRISPR/Cas9 treatment, single cells were isolated by FACS. Around 72 hours post-nucleofection, cells were treated with ROCK inhibitor (Y-27632 dihydrochloride, Santa Cruz Biotech) for 2h and then the iPSCs were dissociated into a single-cell suspension with Accutase and re-suspended in PBS with 10 μM ROCK inhibitor. All samples were filtered through 5-mL polystyrene tubes with 35-μm mesh cell strainer caps (BD Falcon 352235) immediately before being sorted. After adding the viability dye TO-PRO-3 (Invitrogen), the $GFP^{+}TO\text{-}PRO\text{-}3^{-}$ iPSCs were sorted using a gate for high level of GFP expression. The sorted cells were plated either i) with one cell placed in each well of Matrigel-coated 96-well plates by a BD FACSAriaII sorter with a 100-μm nozzle under sterile conditions or ii) onto a 10cm cell culture dish, at a low density allowing colonies to form apart. Cells recovered in conditioned mTeSR medium. Once multicellular colonies from the 10cm dish were clearly visible (2–3 d after sorting), they were collected into individual wells of Matrigel-coated 96-well plates by manual picking. About 14 days after sorting, genomic DNA from all colonies was harvested and used for subsequent validation analyses.

Screening of individual iPSC colonies

To isolate genomic DNA from the iPSC colonies, iPSCs were detached with ReLeSR (STEMCELL Technologies) and then extracted by the Rapid DNA extraction

method described above. For detection of the deletion, the genomic region flanking the CRISPR target site was amplified via standard PCR. Primers were designed using Geneious software and synthesized by Integrated DNA Technologies (IDTdna) (Supplementary Table 3). PCR reactions were performed using 2µl of genomic DNA and Phusion High Fidelity Master Mix (NEB), with the following cycling conditions: 98 ˚C for 2 min and 30 seconds; 98 ˚C for 10 s, 62-64 ˚C for 30 s, 72 ˚C for 30 seconds - 1 min (35-45 cycles); 72 ˚C for 10 min. PCR products were visualized in a 1-2% agarose gel, followed by either gel extraction of the amplicon (QIAquick Gel Extraction Kit, Qiagen) or purification of the PCR product with Illustra ExoProStar (GE Life Sciences) for final Sanger sequencing to determine the exact genomic modifications that occurred via CRISPR/Cas9. In the case of small indels, the PCR amplicons were cloned onto a pCR-Blunt vector (Zero Blunt® PCR Cloning Kit, ThermoFisher Scientific) in order to distinguish each individual allele, according to manufacturer instructions, followed by Sanger sequencing. CRISPR efficiency was then calculated as the sum of heterozygous and homozygous deletions obtained, divided by the total number of colonies screened, for each guide-pair combination.

Results

GuideRNA sequences were confirmed to be correctly inserted onto the *pGuide* vectors, through Sanger Sequencing (Figure 10) prior to transfection. An initial screening using SURVEYOR assay in HEK293T cells to test the efficiency of a subset of individual guideRNAs, showed that the majority of guideRNAs had an efficiency of cutting above 50% (Supplementary Figure 3). This indicates that HEK293T cells are quite tolerant to CRISPR gene-editing. As the indel efficiency shown by this assay was consistent across all guideRNAs in HEK293Tcells, we immediately moved on to transfection in our cells of interest, iPSC, and did not test for the later designed guides in this cell line. This initial analysis gave us confidence for the selection of further guides to use.

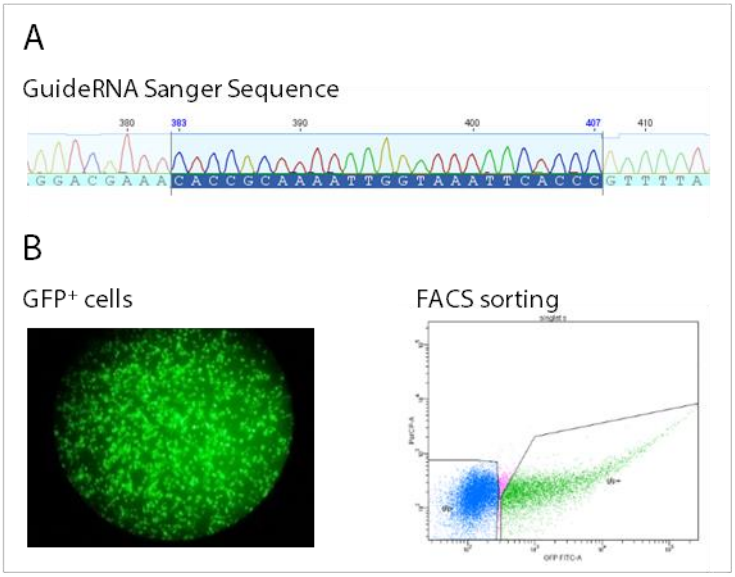
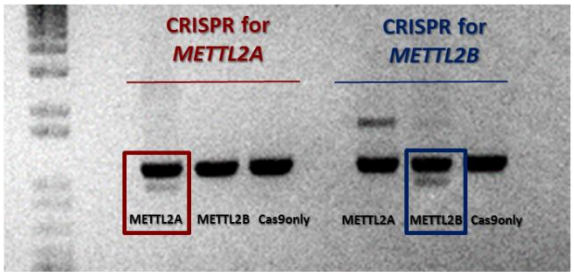


Figure 10 – A) Example of guideRNA 4a inserted onto pGuide; B) GFP signal indicating successful transfection in HEK293T cells via fluorescence microscopy (left) and iPSC via FACS sorting (right).

Upon dual-guide CRISPR transfection in HEK293T cells, GFP could be visualized post 48h, via fluorescence microscope, indicating that the cells were successfully permeated to the entrance of the *pCas9_GFP* vector containing GFP signal and suggestive that they also incorporated the other vectors, containing the guideRNAs (Figure 10). Similarly, after 72h post nucleofection, iPSCs were submitted to cell sorting based on their emission of GFP signal coming from the EGFP vector, also indicative of positive electroporation and high probability of incorporation of *pX459* (Cas9) and *pGuide* vectors (containing guideRNAs) (Figure 10).

Interestingly, for two highly homologous genes targeted (98,41% homology), *METTL2A* and *METTL2B*, CRISPR was able to ablate these genes with specificity to each target. The guideRNAs targeting these genes differed at only 1bp at the 3' end of the first guide and 2bp at the 5' and 3' ends of the second guideRNA (Figure 11).



Gene	GuideRNA 1	GuideRNA 2
METTL2A	AAGCCGGTTCCTGAGAGATC	CTCACGTCCTGGCTGCGGGC
METTL2B	AAGCCGGTTCCTGAGCGATC	TTCACGTCCTGGCTGCGAGC

Figure 11 - CRISPR was highly specific for two homologous genes, targeting either one or the other.

Dual guideRNA CRISPR successfully generated deletions in all four genes targeted in iPSC (*EHMT1*, *MBD5*, *METTL2A* and *METTL2B;* Figure 12), however the efficiency varied by guideRNA combination. The average efficiency using the FACS method was of 2,8% to generate the predicted ablations, out of the total 1002 iPSC colonies screened. The minimum efficiency was of 0% for 5 guideRNA combinations and a maximum of 10.6% (Table VI). For those combinations where there were no deletions detected by PCR, Sanger sequencing was used to determine if either of the individual guides were able to generate indels. In fact, in 3 of these combinations indels were identified. In terms of deletion size, we found no correlation between the deletion sizes (ranged from 13bp to 2200bp) and the efficiency in generating predicted deletions. Although there seemed to be a trend towards a decrease of efficiency with deletion size, it was not significant (R^2=0,951 see in Figure 12).

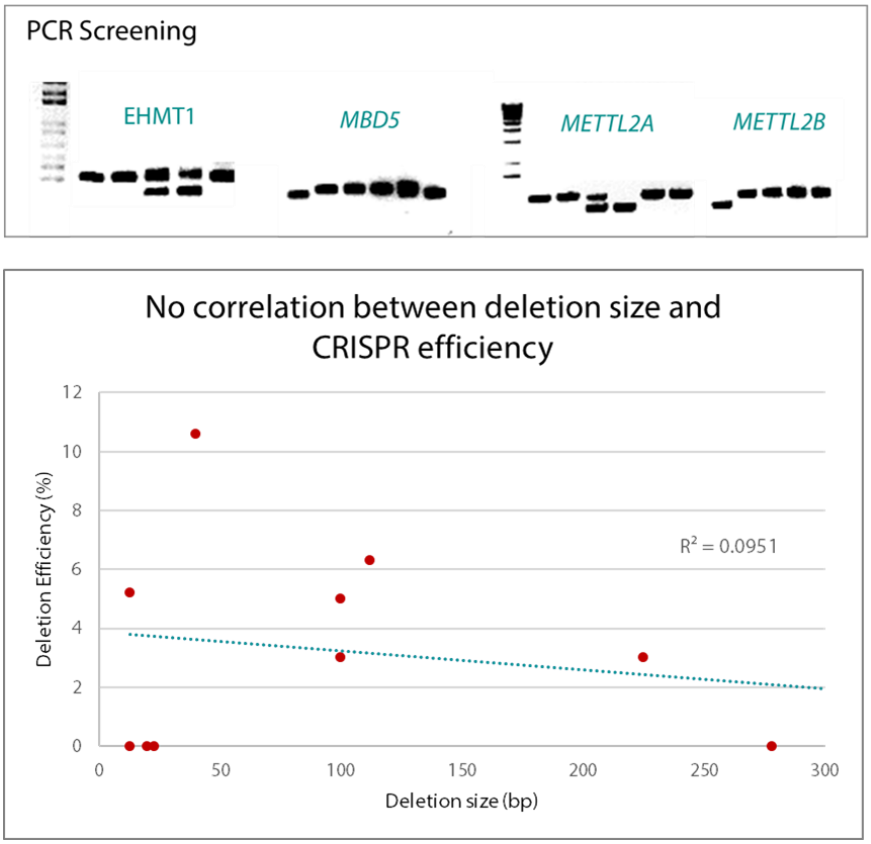


Figure 12 – Top – PCR Screening shows successful deletions in all genes targeted. Top bands – WT allele. Lower bands – homozygous deletions. Double bands – heterozygous deletion.

Bottom - Graph showing the relation between deletion size and CRISPR efficiency shows no correlation.

Table VI – List of genes targeted with CRISPR/Cas9 in iPSCs.

Gene	Guide-pair combination	Exon targeted	Strategy	8330-8 Passage number	Deletion Size (bp)	Colonies screened	Homozygous deletion	Heterozygous deletion	Deletion Efficiency	Screen indels
EHMT1	2a, 2b*	2	whole exon	44	112	95	0	6	6.3%	NA
MBD5	4a, 4b*	4	whole exon	40	278	11	0	0	0.0%	2
MBD5	4.3, 4. 6	4	frameshift	47	23	48	0	0	0.0%	0
MBD5	4.2, 4. 6	4	frameshift	47	20	72	0	0	0.0%	12
MBD5	4a, 6.2	4 to 6	truncation	47	22000	32	1	0	3.1%	NA
MBD5	6.8, 6.3	6	truncation	48	40	264	11	17	10.6%	NA
MBD5	6.8, 6.5	6	frameshift	48	13	96	0	0	0.0%	5
MBD5	7a, 7b	7	frameshift	38	568	192	0	0	0.0%	NA
METTL2A	a1, a2*	1	frameshift	40	100	96	3	2	5.2%	NA
METTL2B	b1, b2*	1	frameshift	40	100	96	3	0	3.1%	NA

NA - not assessed. *Tested using SURVEYOR assay.

Discussion and Conclusions

HEK293T cells have shown to be tolerant to gene modification, induced by CRISPR/Cas9, or other previous technologies as shRNA or TALENs. This human cell type is among the most used in CRISPR reports, due to this facility in transfection and in generating indels (Ran et al. 2013; Mali et al. 2013; Jinek et al. 2013; Cong et al. 2013) and thus is mostly used as a test cell-type to perform experiments prior to the cells of interest. Indeed, different cell types have shown different efficiencies: Mali et al. 2013 performed an experiment of transfecting the same guides in different cell types and showed that in HEK293T, the maximum efficiency of CRISPR-edited indels was of 38% in K562 (myelogenous leukemia cell line), 23% in HEK293T and 4% in iPSC. This difference may be explained due to their different chromatin states and structure and their tolerance to repair double-strand breaks, making some more suitable than others for gene editing.

In this study, CRISPR showed efficiencies in iPSC ranging from 0% to 10,6%. In the case where no deletions occurred, it is possible that either the endogenous cell repair mechanisms corrected the excision in both target sites and did not lead to NHEJ, or the target of the guideRNA combination was deleterious for cell survival. Besides, 4 guideRNAs were electroporated simultaneously (2 *pGuides* with guideRNAs, *pX459* with Cas9, and the EGFP vector) and this could have led to a low probability of internalization of all necessary vectors for correct gene editing and posterior selection. Indeed, it is possible to design a single vector (as *pX459*) and include all guideRNAs in this vector, which could be an approach to decrease the DNA load and increase the probability of gene editing, although the size of this single vector would increase greatly and complicate its entrance in the cell.

After transfection, each cell can be edited differently, depending on its repair mechanisms. Thus, to guarantee an isogenic background within each iPSC colony, without admixture of other edited or wild type cells, FACS was used as an approach to assure the growth of colonies coming from a single cell. However, the increased certainty of deriving a single cell from FACS sorting comes at a significant cost in terms of efficiency and cell viability, as they are removed from their optimal environment and submitted to pressure from the equipment. In order to increase the number of edited cells recovered we could combine an initial puromycin selection followed by FACS sorting based on GFP signal (Tai et al. 2016); however, the use of both selection methods could be even harsher for the stem cells.

It is likely the underlying chromatin structure and epigenetic state of the genomic target loci will impact the efficiency of genome editing in eukaryotic cells. In our study, there was no effect of deletion size on CRISPR efficiency. This might be explained by the 3D structure of chromatin within the nucleus, which can bring in close proximity pieces of DNA that are many base pairs apart, as shown by the formation of chromatin loops and topologically associating domains (TADs) (Rao et al. 2014; Ji et al. 2016).TADs are regions of chromosomes that show evidence of relatively high DNA interaction frequencies based on Hi-C chromosome conformation capture data (Dixon et al. 2015; Dixon et al. 2012; Phillips-Cremins et al. 2013). This proximity could aid in the occurrence of non-homologous end joining (NHEJ) after the double-strand breaks are formed. Chromatin state also is a critical factor in the efficacy of CRISPR, whether it is in a heterochromatin or euchromatin state. In the case of iPSC, this is not as critical, as they are actively proliferating and DNA should be available to the nuclease during replication. However, for post-mitotic cells, it might be more advantageous in some cases when the target is not available to use shRNA instead of CRISPR. In terms of knockdown stability, CRISPR has the advantage of creating permanent alterations which are passed onto the next generation of cells in all divisions with minimal off-target effects and consistent activity (Evers et al. 2016).

CRISPR/Cas9 editing proved to be highly specific for two highly homologous genes (*METTL2A* and *METTL2B*) with the guideRNA sequences differing only by one or two bases. Indeed, studies of CRISPR specificity (Jinek et al. 2012) suggest that target sites must perfectly match the 8-12 base "seed sequence" at the 3′ end of the guideRNA. Cas9 will tolerate single mismatches at the 5′ end in bacteria and *in vitro*, suggesting that the 5′ G is not required (Doench et al. 2014; Mali et al. 2013).

Off-target effects were not directly assessed, however guideRNAs were designed based on off-target prediction software, and also by BLAST of the guideRNA sequence along with the PAM, required for recognition of the Cas9, to rule out homology to other locations in the genome (especially towards the 3' end of the guideRNA). A study in hematopoietic stem cells showed that there were no significant off-target effects by sequencing the bioinformatically predicted sites (Mandal et al. 2014). However, there are reports showing that the off-target prediction tools do not reflect the modification that occur *in vivo* in the cells, and suggest the use additional methods such as Guide-seq (Tsai et al. 2015), which integrates double-stranded DNA oligos whenever there is a double strand break, to address this issue. We sought to solve this problem by targeting each gene with more than one guide-pair combination. This will indicate that there are no off-targets by measuring a readout, such as overall

transcriptional changes (via RNAseq), that should be comparable for all the guideRNA combinations targeting a same locus or gene. Another strategy to minimize the off-target effects would be to use paired Cas9 nickases (Shen et al. 2014), as they only generate nicks in the DNA that will be repaired by endogenous mechanisms unless two Cas9 nickases are brought to close proximity by guideRNAs, and will generate a double strand break. However, using this approach, we would require a total of four guideRNAs to create a double strand break on either side of the target, and this would increase the DNA load at the moment of transfection into the cells, potentially increasing the toxicity.

This systematic survey of genome editing approaches suggests that dual-guide deletion generation varies widely by guideRNA-pair combination and genomic target. The CRISPR-edited human iPSC models have innumerous potential downstream applications and have the ability to be differentiated into the tissue of interest of the disease studied. In our case, we are interested in differentiating these models into the neuronal lineage in order to explore the consequences of LoF of chromatin remodelers in neurodevelopmental disorders, such as ASD. Indeed, this model of human induced pluripotent stem cell will help fill the current knowledge gap in the cellular biology of ASD and can lead to further insights into the molecular mechanisms underlying the disorder.

Acknowledgements

Funding was provided by FCT Fellowship SFRH/BD/52049/2012 to CMS, NIH grant GM061354 to JFG and MET, SFARI grant 308955 to JFG and R00MH095867 to MET.

Supplementary Data

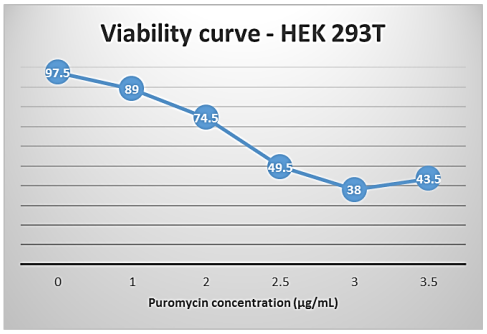
Supplementary Table 3 - Oligos used as guideRNAs and PCR primers for screening.

Type	Gene	Exon	Oligo ID	Sequence (5' to 3')
guideRNAs	*MBD5*	exon 4	sgRNA_ex4_a_sense	CAAAATTGGTAAATTCACCC
			sgRNA_ex4_b_sense	ATGCATACATAAATTCCTAC
			MBD5sg_ex4.6_sense	TTTAGTTCCATCAAACACAG
			MBD5sg_ex4.3_sense	ATGGAACTAAAGAAGCATTA
			MBD5sg_ex4.2_sense	GGAATGACCACCATGGCAGA
			sgRNA_ex6.2_sense	TACAACTTTGCAGGTACCAC
			MBD5sg_ex6.8_sense	CCAGCTATACAAGTTCCTGT
			MBD5sg_ex6.5_sense	GTGGGTTGGCAGCGTCGTGT
			MBD5sg_ex6.3_sense	GAGGCAAAGAGTGTGACGGA
		exon 7	MBD5sg_ex7a _sense	AAGAAACCTGTTTTTACATA
			MBD5sg_ex7b_sense	TGCATGTTCCATCAGTAAGC
	EHMT1	exon 2	sgRNA_ex2_1_sense	GAGAAACAACAGCCGTCAGC
			sgRNA_ex2_2_sense	GGCGGTGTGCACCGAGGGAC
		exon 3	sgRNA_ex3_1_sense	ACTCGGATAGCGGAAAATG
			sgRNA_ex3_2_sense	CTTAAATAAGCCGGCCCTAC
	METTL2A	exon 1	sgRNA_1A_sense	AAGCCGGTTCCTGAGAGATC
			sgRNA_2A_sense	CTCACGTCCTGGCTGCGGGC
	METTL2B	exon 1	sgRNA_1B_sense	AAGCCGGTTCCTGAGCGATC
			sgRNA_2B_sense	TTCACGTCCTGGCTGCGAGC
PCR primers	*MBD5*	exon 4	ex4_fwd_svy	ATCTCCGATCTGCCACTGAC
			ex4_rev_svy	AGGAAAAATGCTGGGCTACC
			MBD5 ex4_rev_seq	GAGAGATATGAAAAAGCCCTGCT
			ex4_fwd_seq	ACAAGCCCTTTCTGTTAGAGTC
		exon 6	ex6_fwd	ACCCCCACTTCAGACAGGTA
			ex6_rev_svy	GCAGAGCCTTCTCCATGACT
			MBD5 ex6_fwd_seq_	TCAGAAGCACTCATTTTTACCC
			ex6_rev_seq	GCCATCAGTCACCATGCTT
		exon 7	mbd5_7fwd	CGATTATTAGCCCGAAGACC
			mbd5_7rev	GGAGGGTTCAGTTTTGTGATTT
	EHMT1	exon 2	ex2_fwd	TCTGCTGGAGGCGACTGTAA
			ex2_rev_2	GAGAGGAAGAGCAGCAGGTTT
	METTL2A	exon 1	fwd_out_A	CTATTAAGAGCTGAATATAG
			intron_rev_A	CTACTAAACATAATTAAAGACAAC
	METTL2B	exon 1	fwd_out_B	CTATGAAGAGCTGACTATAG
			intron_rev_B	CTACTCAACATAATTAAAGACAAA

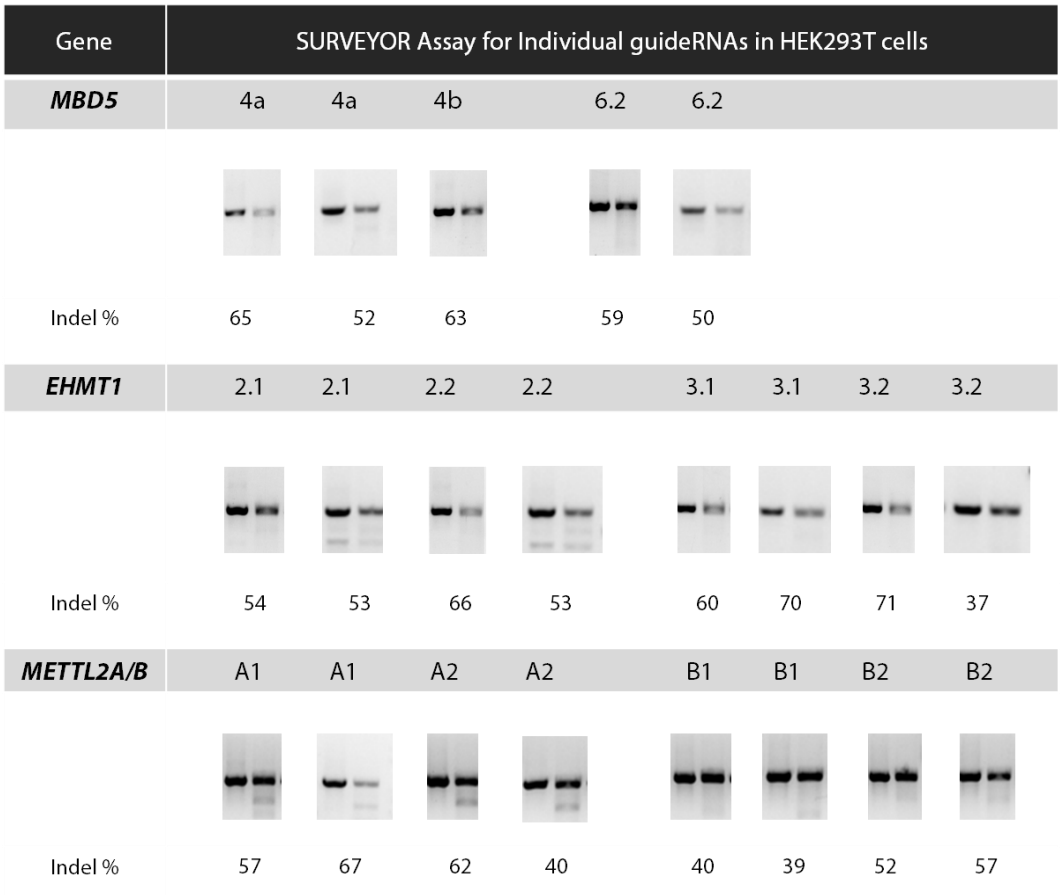
Plate A	Puromycin concentration (µg/mL)					
	0	1	2	2.5	3	3.5
Total number of cells/mL	2.3×10^6	5.4×10^5	2.9×10^5	2.6×10^5	2.0×10^5	2.0×10^5
Live cells/mL	2.2×10^6	4.9×10^5	2.3×10^5	1.9×10^5	1.1×10^5	1.2×10^5
Dead cells/mL	6.0×10^4	5.0×10^4	6.0×10^4	7.0×10^4	9.0×10^4	8.0×10^4
Viability (%)	97	92	79	75	56	60

Plate B	Puromycin concentration (µg/mL)					
	0	1	2	2.5	3	3.5
Total number of cells/mL	3.5×10^6	5.1×10^5	3.1×10^5	4.4×10^5	3.0×10^4	8.0×10^4
Live cells/mL	3.4×10^6	4.4×10^5	2.2×10^5	1.1×10^5	1.0×10^4	2.0×10^4
Dead cells/mL	8.0×10^4	7.0×10^4	9.0×10^4	3.4×10^4	2.0×10^4	6.0×10^4
Viability (%)	98	86	70	24	20	27

Average Viability (%)	97.5	89	74.5	49.5	38	43.5



Supplementary Figure 2 - Puromycin Viability curve for HEK293T cells shows 3µg/mL as the optimal concentration to use to obtain the least number of cells after 48h.



Supplementary Figure 3 – Indel efficiency of individual guideRNAs detected by SURVEYOR assay for a subset of guideRNAs, indicated in Table I. In each image, the band on the left represents the uncut fragment, whilst the band on the right is the one treated by the SURVEYOR assay. The difference of the intensities of the two bands yields the indel percentage, calculated with ImageJ Software.

References

Bershteyn, M. et al., 2017. Human iPSC-Derived Cerebral Organoids Model Cellular Features of Lissencephaly and Reveal Prolonged Mitosis of Outer Radial Glia. *Cell Stem Cell*.

Bessa, C., Maciel, P. & Rodrigues, A.J., 2013. Using C. elegans to Decipher the Cellular and Molecular Mechanisms Underlying Neurodevelopmental Disorders. *Molecular Neurobiology*.

Cong, L. et al., 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science (New York, N.Y.)*, 339(6121), pp.819–23.

Cundiff, P.E. & Anderson, S.A., 2011. Impact of induced pluripotent stem cells on the study of central nervous system disease. *Current Opinion in Genetics & Development*, 21(3), p.354.

Denham, M. & Dottori, M., 2011. Neural differentiation of induced pluripotent stem cells. *Methods in Molecular Biology (Clifton, N.J.)*, 793, pp.99–110.

Dhara, S.K. & Stice, S.L., 2008. Neural differentiation of human embryonic stem cells. *Journal of Cellular Biochemistry*, 105(3), pp.633–40.

Dixon, J.R. et al., 2015. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539), pp.331–336.

Dixon, J.R. et al., 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398), pp.376–380.

Doench, J.G. et al., 2014. Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. *Nature Biotechnology*.

Eiraku, M. & Sasai, Y., 2012. Self-formation of layered neural structures in three-dimensional culture of ES cells. *Current Opinion in Neurobiology*, 22(5), pp.768–777.

Evers, B. et al., 2016. CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. *Nature Biotechnology*, 34(6), pp.631–633.

Furukubo-Tokunaga, K., 2009. Modeling schizophrenia in flies. In pp. 107–115.

Hockemeyer, D. et al., 2009. Efficient targeting of expressed and silent genes in

human ESCs and iPSCs using zinc-finger nucleases. *Nature Biotechnology*, 27(9), pp.851–7.

Hockemeyer, D. et al., 2011. Genetic engineering of human pluripotent cells using TALE nucleases. *Nature Biotechnology*, 29(8), pp.731–4.

Ji, X. et al., 2016. 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. *Cell Stem Cell*, 18(2), pp.262–275.

Jinek, M. et al., 2012. A Programmable Dual-RNA – Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science (New York, N.Y.)*, 337(August), pp.816–821.

Jinek, M. et al., 2013. RNA-programmed genome editing in human cells. *eLife*, 2, p.e00471.

Koike-Yusa, H. et al., 2014. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature Biotechnology*, 32(3), pp.267–73.

Krumm, N. et al., 2014. A de novo convergence of autism genetics and molecular neuroscience. *Trends in Neurosciences*, 37(2), pp.95–105.

Mali, P. et al., 2013. RNA-guided human genome engineering via Cas9. *Science (New York, N.Y.)*, 11(5), pp.367–79.

Mandal, P.K. et al., 2014. Efficient ablation of genes in human hematopoietic stem and effector cells using CRISPR/Cas9. *Cell Stem Cell*, 15(5), pp.643–52.

Mattis, V.B. & Svendsen, C.N., 2011. Induced pluripotent stem cells: a new revolution for clinical neurology? *Lancet Neurology*, 10(4), pp.383–94.

McClive, P.J. & Sinclair, A.H., 2001. Rapid DNA extraction and PCR-sexing of mouse embryos. *Molecular Reproduction and Development*, 60(2), pp.225–226.

O'Roak, B.J. et al., 2011. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genetics*, 43(6), pp.585–9.

Phillips-Cremins, J.E. et al., 2013. Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment. *Cell*, 153(6), pp.1281–1295.

Pinto, D. et al., 2014. Convergence of Genes and Cellular Pathways Dysregulated in Autism Spectrum Disorders. *The American Journal of Human Genetics*.

Pratt, K.G. & Khakhalin, A.S., 2013. Modeling human neurodevelopmental disorders in the Xenopus tadpole: from mechanisms to therapeutic targets. *Disease Models & Mechanisms*, 6(5).

Ran, F.A. et al., 2013. Genome engineering using the CRISPR-Cas9 system. *Nature Protocols*, 8(11), pp.2281–308.

Rao, S.S.P. et al., 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7), pp.1665–80.

De Rubeis, S. et al., 2014. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, 515(7526), pp.209–15.

Sanders, S.J. et al., 2012. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, 485.

Shalem, O. et al., 2014. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science (New York, N.Y.)*, 343(6166), pp.84–7.

Shen, B. et al., 2014. Efficient genome modification by CRISPR-Cas9 nickase with minimal off-target effects. *Nature Methods*.

Sheridan, S.D. et al., 2011. Epigenetic Characterization of the FMR1 Gene and Aberrant Neurodevelopment in Human Induced Pluripotent Stem Cell Models of Fragile X Syndrome M. R. Cookson, ed. *PLoS ONE*, 6(10), p.e26203.

Shinoda, Y., SADAKATA, T. & FURUICHI, T., 2013. Animal Models of Autism Spectrum Disorder (ASD): A Synaptic-Level Approach to Autistic-Like Behavior in Mice. *Experimental Animals*, 62(2), pp.71–78.

Sim, J.C.H., White, S.M. & Lockhart, P.J., 2015. ARID1B-mediated disorders: Mutations and possible mechanisms. *Intractable & Rare Diseases Research*, 4(1), pp.17–23.

Sugathan, A. et al., 2014. *CHD8* regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. *Proceedings of the National Academy of Sciences*, 111(42), pp.E4468–E4477.

Swiech, L. et al., 2015. In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9. *Nature Biotechnology*, 33(1), pp.102–6.

Tai, D.J.C. et al., 2016. Engineering microdeletions and microduplications by targeting segmental duplications with CRISPR. *Nature Neuroscience*, 19(3), pp.517–522.

Takahashi, K. et al., 2007. Induction of pluripotent stem cells from adult human

fibroblasts by defined factors. *Cell*, 131(5), pp.861–72.

Talkowski, M.E. et al., 2011. Assessment of 2q23 . 1 Microdeletion Syndrome Implicates MBD5 as a Single Causal Locus of Intellectual Disability , Epilepsy , and Autism Spectrum Disorder. *The American Journal of Human Genetics*, pp.551–563.

Talkowski, M.E. et al., 2012. Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell*, 149(3), pp.525–37.

Tsai, S.Q. et al., 2015. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology*, 33(2), pp.187–97.

Wang, H. & Doering, L.C., 2012. Induced pluripotent stem cells to model and treat neurogenetic disorders. *Neural Plasticity*, 2012, p.346053.

Wang, T. et al., 2014. Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science*, 343(6166), pp.80–84.

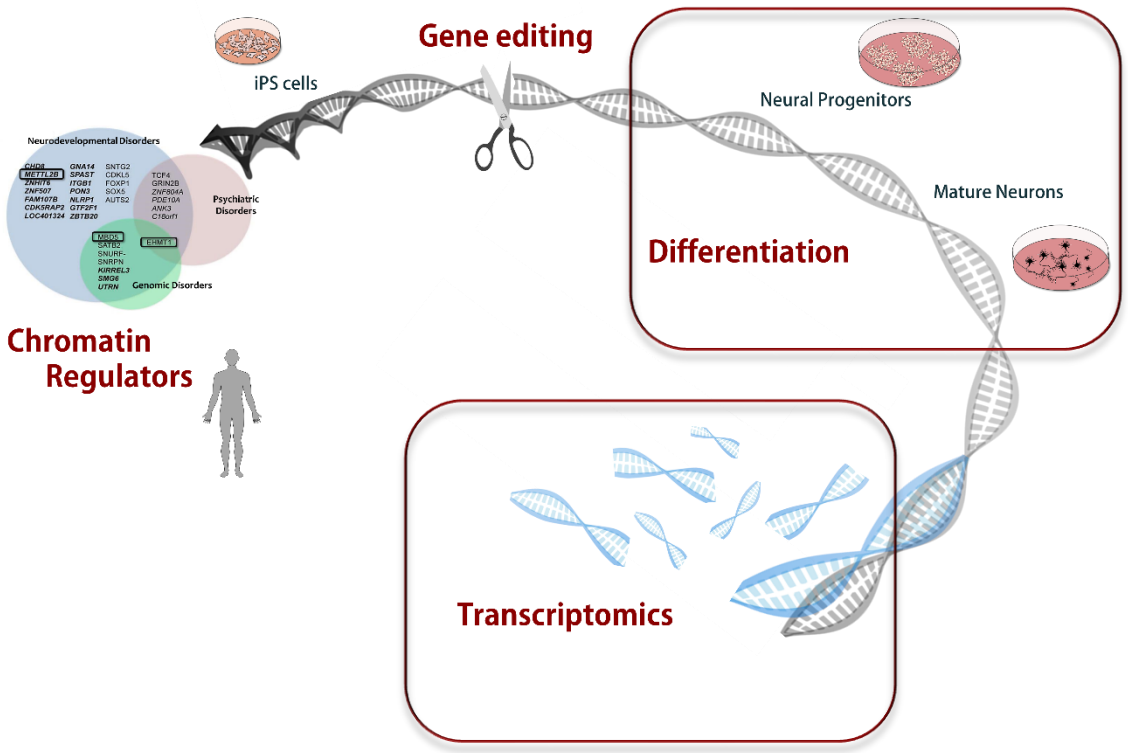
Wichterle, H. & Przedborski, S., 2010. What can pluripotent stem cells teach us about neurodegenerative diseases? *Nature Neuroscience*, 13(7), p.800.

Zou, J. et al., 2009. Gene targeting of a disease-related gene in human induced pluripotent stem and embryonic stem cells. *Cell Stem Cell*, 5(1), pp.97–110.

Zwaka, T.P. & Thomson, J.A., 2003. Homologous recombination in human embryonic stem cells. *Nature Biotechnology*, 21(3), pp.319–21.

Chapter 3

Disruption of chromatin remodeler *MBD5* results in dysregulated neuronal-related genes and pathways



Highlights

The third chapter focuses on a chromatin remodeler, MBD5, and the differentiation of CRISPR-edited iPSC to mature neuronal cells. We investigated the modifications of the transcriptome architecture upon ablating an exon in the 5' UTR and in the MBD Domain, that will give insights on the role of MBD5 during neurodevelopment. The RNAseq analyses of this section were performed in conjunction with bioinformatician Tatsiana Aneichyk.

Authors

Catarina M. Seabra[1, 2, 3], Tatsiana Aneichyk[2, 3], Parisa Razaz[2, 3], Serkan Erdin[2, 3], Poornima Manavalan[2], Celine de Esch[2, 3], Derek J. C. Tai[2, 3], Ashok Ragavendran[2, 3], Yu An[2, 3], Alexei Storchevoi[2], Patrícia Maciel[4, 5], Michael Talkowski[2, 3, 6] and James F. Gusella[2, 3, 7]

Affiliations:

[1]GABBA Program, Institute of Biomedical Sciences Abel Salazar of the University of Porto, Portugal;
[2]Molecular Neurogenetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA;
[3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA;
[4]School of Medicine, University of Minho, Braga, Portugal;
[5]ICVS/3Bs - PT Government Associate Laboratory, Braga/Guimarães, Portugal;
[6]Department of Neurology, Harvard Medical School, Harvard University, Cambridge, MA, USA;
[7]Department of Genetics, Harvard Medical School, Harvard University, Cambridge, MA, USA.

Abstract

MBD5, encoding the methyl-CpG-binding domain 5 protein, has been implicated as the causal locus within the 2q23.1 microdeletion syndrome and represents a significant contributor to the genetic etiology of autism spectrum disorder (ASD) (Talkowski et al. 2011). *MBD5* is a dosage sensitive gene, and haploinsufficiency of *MBD5* mRNA in patients supports it as the major driver gene for phenotypes observed with 2q23.1 microdeletion as well as microduplication syndrome (Mullegama et al. 2013).

In this study, we generated human iPSC-derived neuronal models of CRISPR/Cas9 genome edited iPSC lines bearing mutations in two different regions of the *MBD5* gene, namely exon 4 localized in the 5' UTR and the MBD domain containing exon 6. We ascertained the consequences for local transcript abundance of independent perturbations of these two regions. The canonical transcript MBD5-001 was not the predominantly expressed isoform responsible for *MBD5* expression in the differentiated cell lines. Surprisingly, the transcript that exhibited highest expression overall was a non-protein coding transcript MBD5-010 that comprises only 5' UTR exons 1 and 2. This is a promising non-coding transcript that may be implicated in neuronal development and disease and should be further investigated to determine its potential as a regulatory lncRNA. One transcript, MBD5-014, showed NPC-specific expression, being observed uniquely in the NPC population and absent in the neurons, suggesting a developmental state preference. Genome-wide transcriptomic analysis via RNAseq allowed the identification of the dysregulated genes upon CRISPR editing such as *RAB11FIP1*, *NHLH1-2*, *PLAUR* and *CNTNAP2*; and pathways such as notch signaling and cell adhesion that gave insight on the protein complexes and pathways that are acting downstream of *MBD5*.

In conclusion, we demonstrated that we could generate human loss-of-function neuronal cell lines to study ASD. These indicated genes and pathways that may be directly implicated in neuronal development and function and thus represent promising targets for ASD therapeutics.

Keywords - *MBD5*, CRISPR, differentiation, MBD, neuronal progenitor, neuron, transcript, expression.

The 2q23.1 microdeletion syndrome is a previously described genomic disorder that comprises intellectual disability, severe speech impairment, seizures, behavioral problems, microcephaly, mild craniofacial dysmorphism, small hands and feet, short stature, and broad-based ataxic gait. *MBD5* (MIM 611472), encoding the methyl-CpG-binding domain 5 protein, has now been implicated as the causal locus within the 2q23.1 deletion region (see Figure 15A) and represents a previously unrecognized contributor to the genetic etiology of autism spectrum disorder (ASD) (Talkowski et al. 2011). *MBD5* is a dosage sensitive gene, and haploinsufficiency of MBD5 in patients supports it as the major causative gene for the 2q23.1 microdeletion syndrome, as well as the reciprocal microduplication syndrome (Mullegama et al. 2013). Indeed, *MBD5* has been pinpointed as the single contributor of the 2q23.1 microdeletion syndrome since cases only with *MBD5* deletions presented similar phenotype in comparison to the patients with 2q23.1microdeletion syndrome. Besides deletions, a few coding variants, mostly missense, have been reported in ASD patients (see Figure 15B).

MBD5 belongs to the family of the methyl-CpG-binding domain (MBD) family of proteins, which includes MBD1, MBD2, MBD3, MBD4, MBD5, MBD6, SETDB1, SETDB2 and MECP2, the causative gene for Rett syndrome. The MBD family members have key roles in regulating gene transcription and *in vitro*

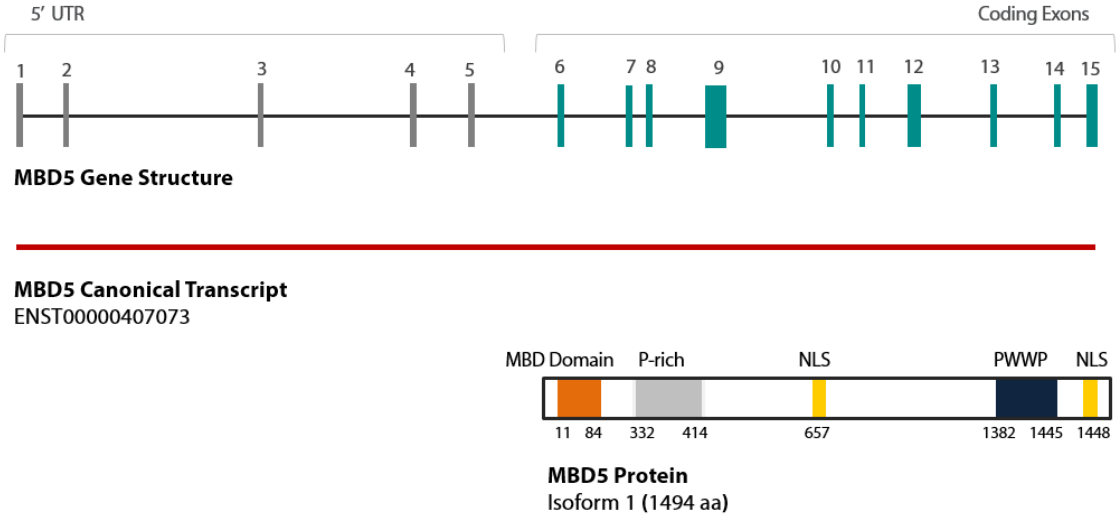


Figure 13 – Top – MBD5 gene structure of 492 467 bp of length (Refseq NM_018328.4). Coding exons of human MBD5 (exons 6–15) are shown in blue. Middle - Translation of all these exons yields protein isoform 1, through the canonical transcript MBD5-001 (ENST00000407073). Bottom - Protein isoform 1, the main described isoform, is composed of a conserved MBD, a proline-rich segment (P-rich), a PWWP domain, and putative nuclear localization signals (NLS).

experiments have led to a model in which MBD1, MBD2, MBD4, and MECP2 recruit chromatin remodelers, histone deacetylases, and methylases to methylated DNA, leading to transcriptional repression (Nan & Bird 2001; Ng et al. 1999). Indeed, their MBD allows the specific recognition of DNA containing methylated cytosine and as a consequence, the proteins serve as interpreters of DNA methylation (Laget et al. 2010). The MBD family members are involved in a variety of functions including DNA damage repair (MBD4), histone methylation (SETBD1 and SETDB2), and X chromosome inactivation (MBD2) (Bogdanović & Veenstra 2009; Roloff et al. 2003), transcript splicing and also gene activation (Young et al. 2005; Yasui et al. 2007; Chahrour et al. 2008). Immunocytochemistry experiments showed that MBD5 localizes in the nucleus to non-heterochromatin regions, which suggests that MBD5 acts as a transcriptional activator (Camarena et al. 2014), and it has been demonstrated that it does not have the ability to bind methylated DNA *in vitro* (Laget et al. 2010).

The *MBD5* gene is composed by 15 exons, of which exons 1 through 5 are part of the 5'UTR, leaving exons 6 through 15 as the coding portion of *MBD5*. In fact, 90% of this gene's length is composed of the 5' UTR alone (~436 kb of the ~492kb total). In terms of transcription, the *MBD5* gene can be transcribed into a total of 33 transcripts (annotated in the Ensembl database), of which 7 are predicted to be protein coding (Supplementary Table 5). The translation of the canonical transcript MBD5-001 (ENST00000407073) leads to the production of the main MBD5 protein isoform reported to date, isoform 1 (UniProtKB ID Q9P267), that has 1494 amino acids and is encoded by exons 6 through15 (Figure 13). This protein has two conserved domains, a MBD and a PWWP (proline-tryptophan-tryptophan-proline) domain, both of which may be found in chromatin-associated proteins. The MBD5 protein isoform 2 (of 851 amino acids) has also been reported and is a shorter version of isoform 1 that excludes the PWWP domain (Laget et al. 2010). Expression studies have shown that isoform 1 is expressed in all tissues but highly expressed in brain and testis, while isoform 2 is expressed in all tissues but highly expressed in brain and

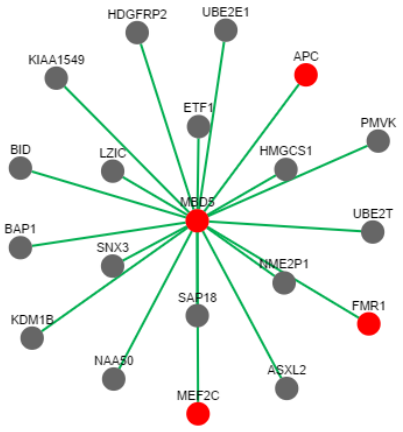


Figure 14 – Interaction nodes with MBD5 (from SFARI Gene Db). Green lines indicate protein binding. Red dots are ASD-related genes.

ovaries (Laget et al. 2010). Although its role is poorly understood, protein-protein interaction studies have shown association of MBD5 protein with products of other ASD genes (Figure 14), indicating it may interact with those via the same pathways.

Mouse models have been effective tools to understand the pathogenesis of 2q23.1 deletion syndrome and will be the key tools in future to aid in designing therapeutic strategies. Several different models alter or abolish Mbd5 expression in mice and include an *Mbd5* knockout (*Mbd5*$^{-/-}$), a brain-specific *Mbd5* conditional knockout (*Mbd5*$^{f/-}$, NestinCre) (Du et al. 2012), and a heterozygous hypomorph (*Mbd5*$^{+/GT}$) (Camarena et al. 2014). The mouse models and more specifically, *Mbd5*$^{+/GT}$, recapitulate aspects of the human condition. The *Mbd5*$^{GT/GT}$ mice exhibit perinatal lethality (Camarena et al. 2014), in contrast to the heterozygous *Mbd5* hypomorphic mouse, *Mbd5*$^{+/GT}$, which was viable. The *Mbd5*$^{+/GT}$ mice exhibited behavioral abnormalities with neuronal function deficits, consistent with the 2q23.1 deletion syndrome human phenotype. *Mbd5*$^{+/GT}$ mice are smaller than wild-type, have abnormal nasal bone development, and exhibit motor coordination impairment, which are present in children with 2q23.1 deletion syndrome. In addition, supporting the theory that MBD5 has a role in autism and ID, *Mbd5*$^{+/GT}$ mice exhibited abnormal social behaviors and clear indication of learning deficits and memory impairment (Camarena et al. 2014). Other models to study MBD5 haploinsufficiency or specific genomic variants have not been described so far.

In this study, we aim to explore one of the iPSC CRISPR-edited models described in the previous chapter and this way we will show an example of an application of those models. To do so, we pursued the iPSC CRISPR-edited models harboring mutations in *MBD5*, as this gene is a compelling candidate whose mutations confer high risk for ASD. Besides, its molecular function remains poorly understood and characterized. Therefore, we will drive the previously described iPSC models into the terminally differentiated neuronal lineage, as this represents the tissue of interest when studying ASD. We are interested in looking at the consequences that occur genome-wide at the transcriptomic level, via RNAseq, upon the perturbation of 2 independent regions of *MBD5*.

The regions individually targeted by CRISPR/Cas9 were the 5'UTR exon 4 of *MBD5* and the MBD-bearing exon 6. These exons were chosen based on previous evidence of their relevance for possible gene function and clinical

reports. Exon 4 is located within the 5' untranslated region of the *MBD5* gene and despite being non-coding, there have been several reports of anomalies in the 5'UTR region in ASD patients that confer disease (Figure 3). Although it represents 90% of the gene's total length, the significance of the 5' UTR is yet unknown.

On the other hand, exon 6 was targeted as is it represents the first coding exon of the canonical transcript MBD5-001 (ENST00000407073) that is the most reported to date. This transcript encodes the MBD5 protein isoform 1 (Q9P267). Therefore, by disrupting this exon, we expected to prevent the transcription of the canonical transcript and, consequently, of the protein isoform 1. This would result in haploinsufficiency of MBD5 when observed in heterozygosity, mimicking what occurs in patients. Additionally, exon 6 contains the initial portion of the MBD domain of MBD5, that is thought to play a role in its function as a chromatin remodeler as this is observed in the other MBD family members.

This strategy will give insights both into the requirement of these 2 regions of *MBD5* during neurodevelopment and also into the transcriptomic impact of ablating these regions during this process. A greater understanding of the role of MBD5 in neurodevelopment will unveil the mechanism by which its mutations confer risk for ASD and may open new avenues of research to ultimately develop potential therapies for neurodevelopmental disorders.

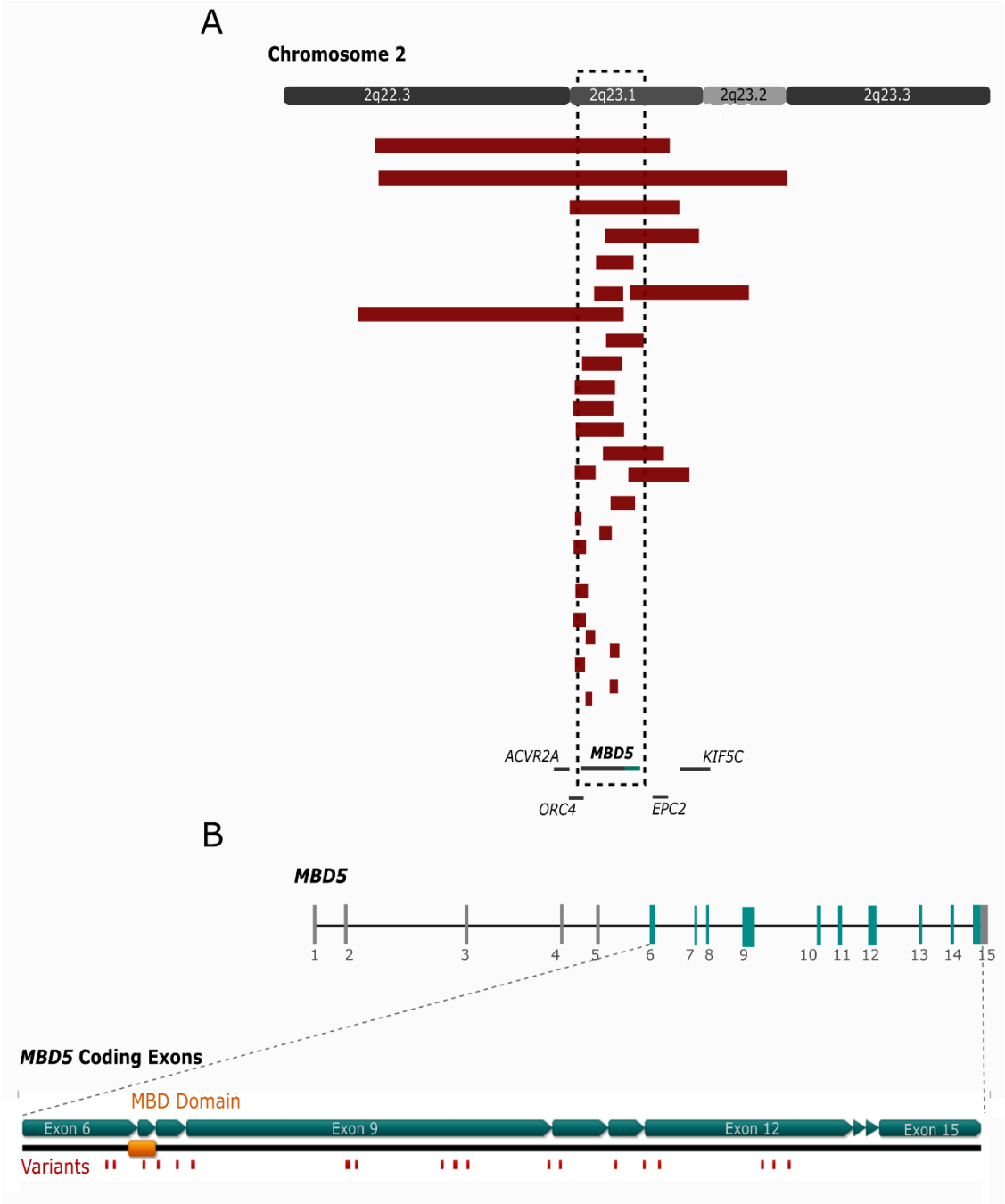


Figure 15 - A) 2q23.1 microdeletion syndrome region was narrowed down to MBD5 as a single causal locus. (patients' deletions represented as red bars); B) MBD5 coding variants (represented in red and MBD Domain in orange). (Adapted from Mullegama & Elsea 2016 and Talkowski et al. 2011.)

GuideRNA Design and Strategy

The CRISPR/Cas9 gene editing system was used to create iPSC lines with mutations in the *MBD5* gene. We used a dual-guideRNA strategy for the excision of a DNA fragment, as this can be readily screened via standard PCR. GuideRNAs for *MBD5* were designed, based on genome assembly GRCh37, and integrated onto *pGuide* vector (Addgene plasmid 41824) as described previously (Chapter 1, Supplementary Table 1). GuideRNAs were designed to either ablate the entire exons or excise a fragment within them. The sequences of guideRNAs are included in Table I of Chapter 1.

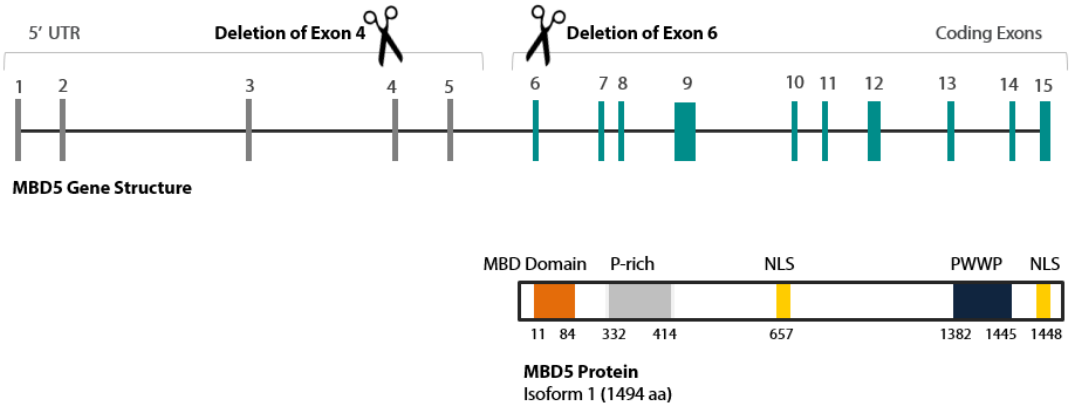


Figure 16 - Guide design strategy, targeting the 5' UTR exon 4 and MBD Domain portion of exon 6 (based on transcript MBD5-001).

Germ Layer Differentiation of iPSC

The germ layer experiment was performed to detect the ability of the iPSC to differentiate into the three germ layers (Itskovitz-Eldor et al. 2000). iPSC were grown under standard conditions and for each line select 2 wells from a 6-wells plate to continue with Detach colonies with 1x Collagenase (1ml; 1mg/ml) for 1 hour at 37ºC. Then transfer the colony pieces to a 6-well low-binding plate (NUNC). In one well for ectodermal differentiation we placed 1/3rd of the pieces in EB medium (conventional iPSC medium without bFGF) supplemented with 10µM SB431542. In a different well, for meso/endodermal differentiation, we put 2/3rd of the pieces in EB medium and added ROCK inhibitor to the medium (for both conditions). The medium was changed every 48 hours, without ROCK

inhibitor. After ~6 days we plated 5 EBs per well for mesoderm, 20 EBs per well for endoderm and 10 EBs per well for ectoderm, using 12-well plates coated with gelatin for Endoderm and Mesoderm, or 12-well plates coated with Matrigel for Ectoderm. From this point on, the medium used for each layer were the following: i) Endodermal – RPMI 1640, 20% FBS, 1x PSG, α-thioglycerol; ii) Mesodermal – DMEM low glucose, 15% FBS, 1x PSG, 1x NEAA; iii) Ectodermal – Neurobasal medium, DMEM/F12, 1x PSG, 1x NEAA, 7,5% BSA, N2, B27 without Vitamin A, SB431542.

Cells were fixed on day 14 with 4% PFA, followed by primary incubation with mouse anti-human β-tubulin III (TUBB3, 1:200 Sigma Aldrich T8578), mouse anti-human smooth muscle α-2 actin (ACTA2, 1:50 DAKO M0851), rabbit anti-human α-fetoprotein (AFP, 1:200, DAKO A0008) and subsequent appropriate fluorochrome conjugated secondary antibodies (1:400 dilution) for microscopic evaluation.

Nucleofection in Human Induced Pluripotent Stem Cells

Human induced pluripotent stem cells (iPSCs), derived from fibroblasts from a healthy male individual, identified as 8330-8 cells, were previously generated using standard retroviral vectors and the Yamanaka factors OCT3/4, SOX2, KLF4, and c-MYC (Sheridan et al. 2011). The iPSCs maintenance and nucleofection were performed as described in Chapter 2. Briefly, iPSCs (1×10^6 cells) were transfected with 1 µg total DNA plasmid, Cas9 expression vector *pX459* (pSpCas9(BB)-2A-Puro plasmid Addgene 48139) along with the chosen gRNAs (inserted into *pGuide* - Addgene plasmid 41824) and an external EGFP (enhanced green fluorescent protein) vector. For nucleofection of the gRNAs into the iPSC, the Human Stem Cell Nucleofector Kit 1 (Lonza) and Amaxa Nucleofection II device (Lonza) were used with program B-016, according to the manufacturer's instructions. After nucleofection, the iPSCs were cultured on Matrigel-coated wells using conditioned mTeSR medium supplemented with 10 µM ROCK inhibitor (Y-27632 dihydrochloride, Santa Cruz Biotech) and 10 ng/ml bFGF (R&D). Single cell FACS sorting and individual colony screening were performed as described in Chapter 1.

Measuring Gene Expression by qRT-PCR

To determine whether MBD5 transcription levels were affected upon CRISPR editing, quantitative real-time polymerase chain reaction (qRT-PCR) was used to measure MBD5 mRNA fold-change relative to housekeeping genes. RNA was obtained by lysing 1-2 million cells using 1mL of Trizol (Invitrogen) then mixed with 1/5th volume of chloroform and centrifuged at 200xg for 5 minutes. The aqueous phase was collected and processed using an RNeasy Mini column (Qiagen). cDNA was synthetized from 1µg of extracted RNA using SuperScript® III Reverse Transcriptase (ThermoFisher Scientific) with oligo(dT), random hexamers, and RNase inhibitor. Quantitative RT-PCR was performed for the target genes using custom designed primers and ACTB, GAPDH and POLR2A were used as endogenous controls (Supplementary Table 4). Primer melting curves and efficiency were verified and only optimal primers were considered. Primers (0.75 µM final), cDNA (1:100 final) and nuclease-free water were added to the LightCycler® 480 SYBR Green I Master Mix (Roche) for a final 10 µL reaction volume. LightCycler® 480 (Roche) was used for data acquisition. Values of expression for each cell line (treated or control) were obtained in at least three technical replicates. Normalized expression levels were set as fold-change in comparison to control cell lines (Table VII), using the ΔΔCt method (Livak & Schmittgen 2001). One-way ANOVA and Tukey post-hoc test was used to assess statistical significance, using IBM SPSS Statistics 24.

iPSC-derived Neuronal Progenitor Differentiation

Expandable neuronal progenitor cells were generated from iPSCs through differentiation by the embryoid body (EB) protocol using STEMdiff™ Neural Induction Medium (STEMCELL Technologies), following the manufacturer's instructions. Briefly, 3×10^6 iPSC were transferred to a micro-patterned culture surface well (AggreWell™800) using centrifugal force, resulting in 10,000 cells per micro-well that would then form embryoid bodies (EBs, day 0). EBs were plated on day 5 onto Corning® Matrigel®-coated plates and expanded for the following days. Around day 12, neural rosette structures were visible (Figure 17) and were manually collected using DMEM-F12 medium and collected into a 15mL tube and plated onto poly-ornithine (PLO, Sigma)/laminin (Sigma) coated culture plates.

PLO was used at a final concentration of 20µg/mL and laminin at 5µg/mL. Isolated cells were expanded in neural expansion medium (70% DMEM (Invitrogen), 30% Ham's F-12 (Mediatech) supplemented with B-27 (Invitrogen), heparin and mitogens EGF (20 ng/mL, Sigma) and bFGF (R&D Systems). After ten passages, cells were collected for subsequent experiments and analyzed for expression of NPC-specific markers.

Immunofluorescence staining was performed after fixation in 4% paraformaldehyde, followed by primary antibody incubation with rabbit anti-human NESTIN (1:500 dilution, Millipore ABD69), mouse anti-human SOX1 (1:200 dilution, Millipore AB15766), rabbit anti-SOX2 (1:200 dilution, Abcam AB59776) and rabbit anti-human PAX6 (1:200 dilution Covance PRB278P) and subsequent appropriate fluorochrome conjugated secondary antibodies (1:400 dilution) for microscopic evaluation. Fluorescence intensity was normalized for the 8330-8 non-treated control sample.

Cell Cycle Analysis

Cell cycle of NPCs was analyzed by propidium iodide (PI) staining that allows the identification of the proportion of cells from a whole cell population that are in one of the three interphase stages (G0/G1, S and G2/M phase) of the cell cycle. Briefly, 1 - 2 × 10^6 cells from each cell line were harvested in 1x PBS and fixed in 70% ethanol for 30min on ice. The pellet was washed twice with PBS and treated with 50µL of Ribonuclease A solution (100µg/mL in 1x PBS) and 400µL of PI solution (50µg/mL in 1x PBS) was added. PI fluorescence of the cells at room temperature was analyzed by flow cytometry by the BD FACSAria II sorter with a 100-µm nozzle under sterile conditions. To determine the cell cycle phase curves the FACSAria Software was used for each cell line.

Neuronal Differentiation

Terminal neuronal differentiation was achieved by plating expanded neural progenitor cells at a seeding density of 2 × 10^6 cells per well on poly-ornithine/laminin-coated plates (coated together overnight) in NPC expansion medium lacking both growth factors EGF and bFGF and heparin, with medium replacement every 3–5 days for 30 days. This will generate a mixed population neuronal subtypes (excitatory and inhibitory). Immunofluorescence staining was performed as described above for neuronal-specific markers, using chicken anti-human MAP2 (1:2500 dilution, EnCor Biotechnology Inc CPCA-MAP2), a microtubule-associated protein, and mouse anti-human SMI312 (1:1000 dilution, Biolegend 837901), a neurofilament axonal marker. Fluorescence intensity was normalized for the 8330-8 non-treated control sample.

Whole Transcriptome Sequencing (RNAseq) and Analyses

RNAseq libraries were prepared using the Illumina TruSeq kit and manufacturer's instructions. Libraries were multiplexed, pooled and sequenced on multiple lanes of an Illumina HiSeq2500, generating an average of 26.67 million paired-end reads of 75 bp.

Quality assessment of sequence reads was performed using fastQC (v. 0.10.1 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Sequence reads were then aligned to human reference genome Ensembl GRCh37 (v. 75) using and gene counts were generated using STAR (v. 2.4.2) (Dobin et al., 2013)

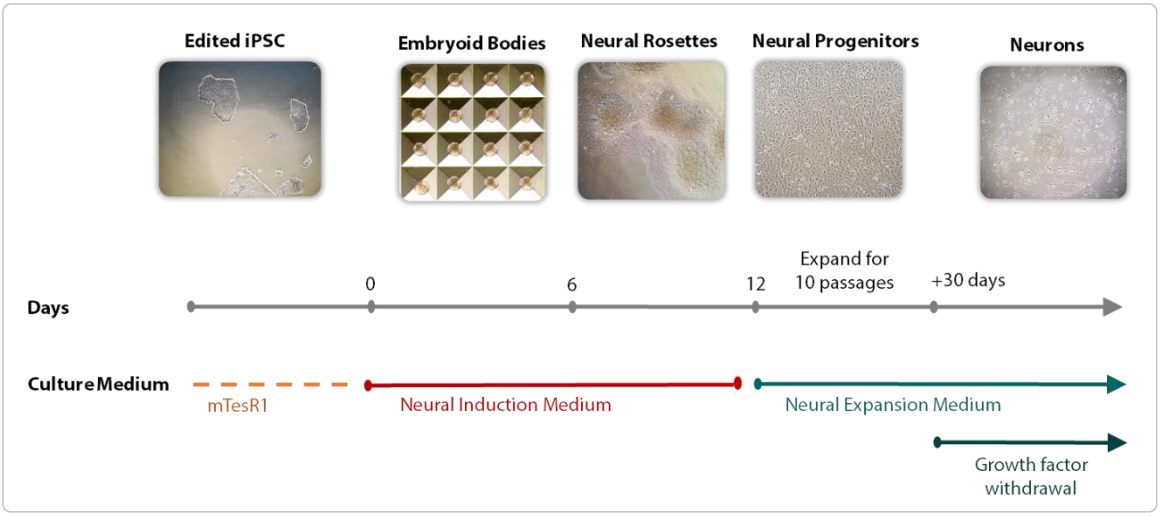


Figure 17 - Differentiation workflow of iPSC into NPC and finally into mature neurons. The medium used in each step is indicated in the figure, along with the timeline required to complete each phase of differentiation.

with following options different from default: --outFilterMultimapNmax 1 --outFilterMismatchNoverLmax 0.1. Quality checking of alignments was assessed by a custom script utilizing Picard Tools (http://broadinstitute.github.io/picard/), RNASeQC (DeLuca et al., 2012), RSeQC (Wang et al., 2012) and SamTools (Li et al., 2009). Differential *MBD5* transcript expression was assessed using RSEM and Bowtie2, and visualized on Integrative Genomics Viewer (Robinson et al., 2011; Thorvaldsdóttir et al., 2013).

Differential Expression Analysis

To account for factors influencing gene counts in each cell type, surrogate variable analyses (SVA) (Leek et al. 2007) were performed detecting two surrogate variables in each cell type. To estimate log2 foldchanges in CRISPR-edited samples vs. control samples, generalized linear models were used to model counts by negative binomial distribution using R package DESeq2 (v 1.10.1) (Love et al. 2014). Estimated log2 foldchanges were tested for significance using Wald test, and corresponding p-values were adjusted for multiple hypothesis testing using Benjamini-Hochberg adjustment method (p-adj).

Network and Enrichment Analyses

Significantly perturbed KEGG pathways and gene ontologies (GO) were analyzed using R package gage (Luo et al. 2009), using log2 foldchanges estimates as gene effects. Differentially expressed genes were tested for enrichment of gene ontologies using package topGO (Alexa & Rahnenfuhrer 2016), with only curated evidence for association of the genes to ontologies, which takes into account a structure of gene ontology tree.

8330-8 iPSC were able to differentiate into all three germ layers

The parental 8330-8 cell line was tested to determine its pluripotency potential of differentiating into all three germ layers. Indeed, the 8330-8 iPSC were able to maturate into the three germ layers, as shown by the expression of relevant markers specific to each embryonic layer. The ectodermal layer-driven cells, stained positively for β-tubulin III (*TUBB3*); the mesodermal layer-driven cells contained smooth muscle α-2 actin (*ACTA2)*, and finally the endodermal layer-driven cells express α-fetal protein (*AFP)* as represented in Figure 18. This assay confirms the potential of this cell line to differentiate into our cell line of interest, namely the ectoderm-derived neuronal progenitor cells and subsequent mature neurons, and therefore was chosen as the cell line to pursue for this study.

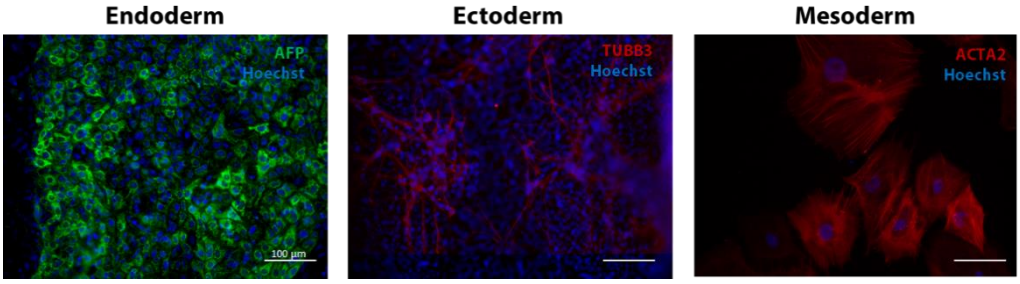


Figure 18 - Germ layer differentiation of the 8330-8 iPSC line shows successful differentiation into all three germ layers.

Edited iPSC showed a range of *MBD5* mRNA expression assessed by qRT-PCR

To understand how the perturbation of certain gene regions affects the local transcript architecture of *MBD5*, iPSC were edited using CRISPR/Cas9 technology. Either the 5'UTR exon 4 or the MBD-containing exon 6 was targeted. Upon CRISPR-editing of the cell lines, an initial screening with standard PCR was used to select edited cell lines, based on the edited amplicon size (shown in Chapter 1). Afterwards, total RNA was extracted from those cell lines in order to assess whether the *MBD5* mRNA levels were affected, via qRT-PCR. The screened iPSC lines showed a decrease in *MBD5* expression ranging from 0% to about 70% (Figure 19). To further pursue differentiation of these models into the neuronal

pathway, we selected the edited cell lines that exhibited the greatest decrease in *MBD5* expression (names shown in boxes in Figure 19). The exon 4 edited cells lines selected were: *4i H6*, *4i H7* and *4i AIID6*. The exon 6-edited cell lines selected were: *6het AIIIB5*, 6het AIVG12 and *6hom AIID2*.

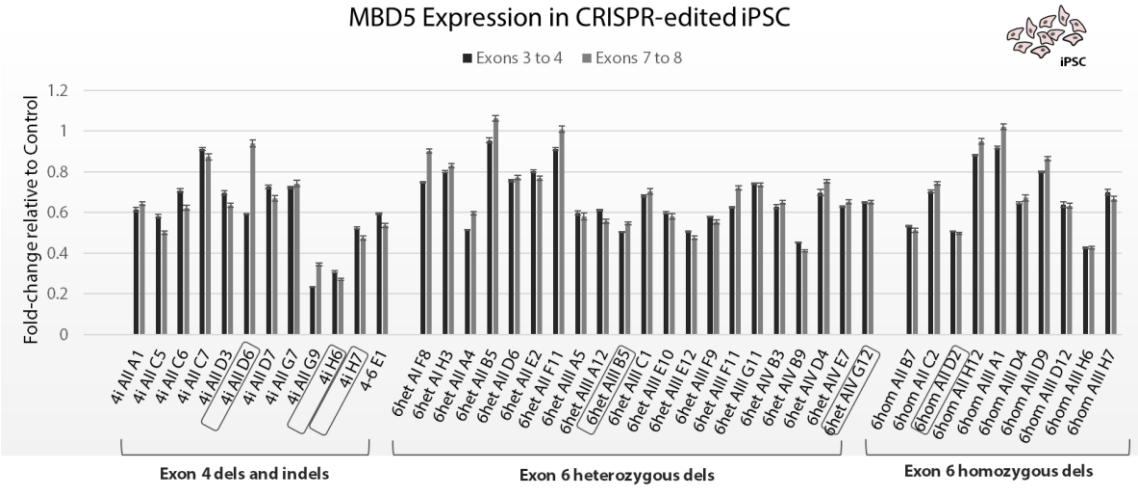


Figure 19 - MBD5 exon mRNA expression in iPSC. Fold-changes were calculated relative to the housekeeping gene ACTB and normalized to the 8330-8 treated control sample.

Sanger sequencing of CRISPR/Cas9-induced mutations indicated repair via non-homologous end joining

To have a complete view of the alterations that occurred in the selected CRISPR-edited cell lines, DNA was extracted for Sanger sequencing of the targeted region. After CRISPR/Cas-9 genome editing the endogenous cell repair mechanisms come into action to repair the double strand breaks induced by this technology, mostly resulting in non-homologous end joining (NHEJ) when using the dual-guideRNA strategy. However, in some cases additional editing can occur in the repaired region by the insertion or deletion of several nucleotides. Indeed, the Sanger sequencing revealed further editing besides NHEJ in some cell lines. The detailed description of each selected cell lined for further differentiation, the mutations occurred in the CRISPR-edited cell lines and their translational consequences are included in Table VII.

The CRISPR lines targeting exon 4, were all heterozygous lines for the mutations induced and there were no complete dual-guide deletions. In these cell lines, only one of the guideRNAs of the pair used for transfection was able to induce a mutation in the genomic DNA. This may be explained either due to a quick repair on the other guideRNA location, that prevented end joining or either because a full exon 4 excision was not tolerated by the iPSC and cells with that mutation were not viable. Therefore, we ended up with 2 cell lines with mutations upstream of exon 4 (*4i H6* and *4i H7*) and 1 cell line with a mutation within the exon 4 (*4i AIID6*) (Figure 20). Cell line *4i H6* contains a 3bp insertion (AAA) and 1bp substitution C>A, while *4i H7* presents a 19bp deletion. Despite having mutated the intron upstream of exon 4, these two lines were kept for follow-up as they showed decreased levels of *MBD5* expression in the qRT-PCR and also as previous reports of *MBD5* mutations in intronic regions have occurred in patients (Talkowski et al. 2011). On the other hand, cell line *4i AIID6* contains a 9bp insertion and G>T substitution within exon 4. As these modifications in the exon 4-targeted cell lines are located upstream of *MBD5* coding sequence, these are not predicted to have a translational consequence, as confirmed by nucleotide-to-protein translation tools as ExPASy Translate Tool (http://web.expasy.org/translate/).

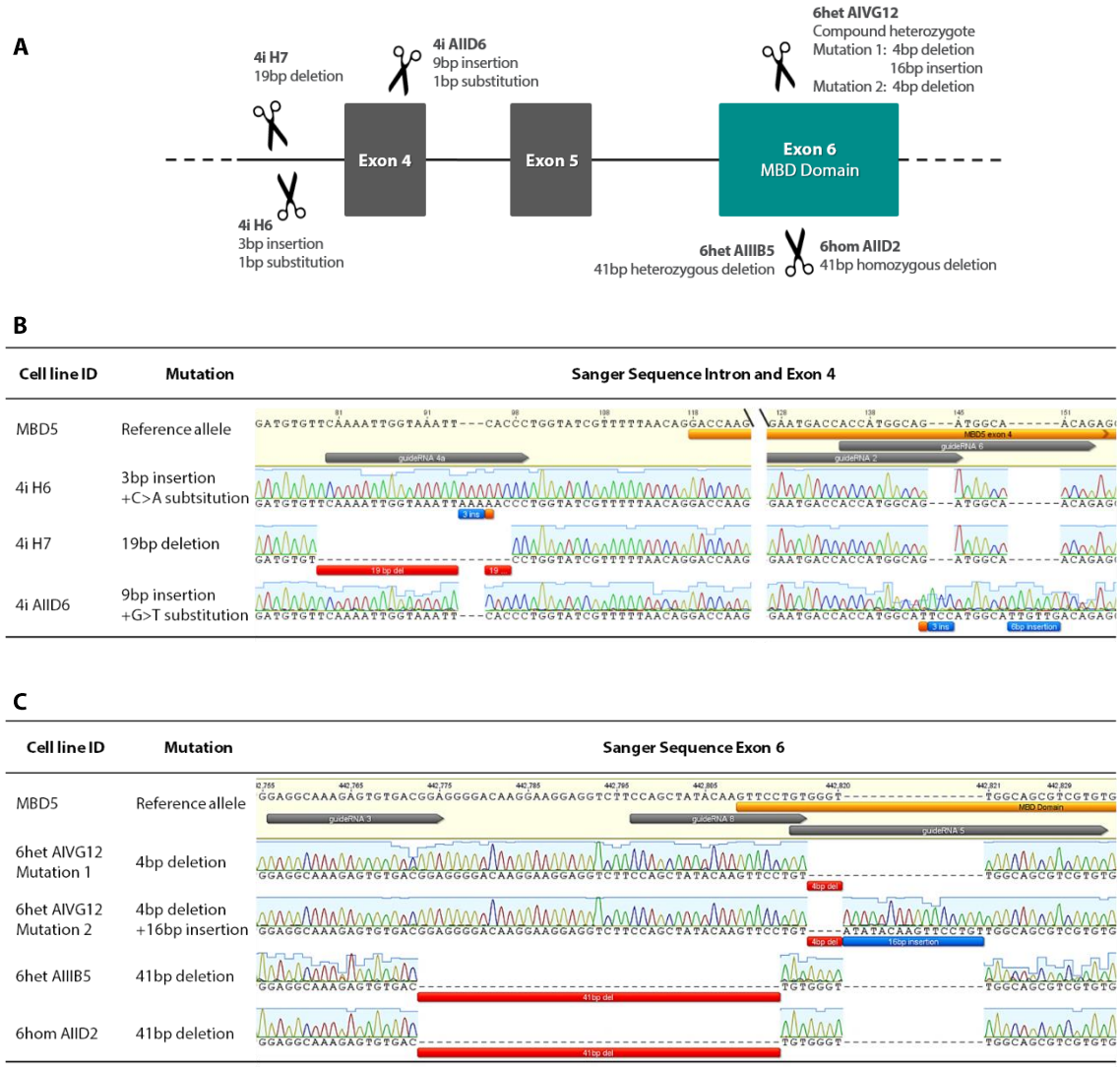


Figure 20 – Characterization of the selected CRISPR-edited cell lines.

A) Scheme of the location of the mutations within the CRISPR-edited cell lines from the intron upstream of exon 4 to exon 6 of the MBD5 gene.

B) Sanger traces of the CRISPR-edited cell lines targeting Exon 4. Top sequence is the reference MBD5 sequence to which the CRISPR-edited sequences were aligned (RefSeq NG_017003). Red – deletion. Blue – insertions. Orange – substitution. Gray – guideRNA sequences. Yellow – exon 4. The two parallel gray lines represent a gap in the sequence to allow the visualization of all modifications.

C) Sanger traces of the CRISPR-edited cell lines targeting Exon 6. Top sequence is the reference MBD5 sequence to which the CRISPR-edited sequences were aligned (RefSeq NG_017003). Red – deletion. Blue – insertions. Orange – substitution. Gray – guideRNA sequences. Yellow – MBD domain.

The selected exon 6-targeted cell lines present modifications within the coding sequence, affecting the MBD domain (Figure 20). Cell line *6het AIVG12* is a compound heterozygote, as it contains a mutation in each of its alleles. On allele 1 this cell line presents a 4bp deletion and a 16bp insertion, leading to a

total gain of 12bp, which makes this modification in-frame. This in-frame mutation leads to a 4 aminoacid insertion and to a 1 aminoacid substitution in the canonical MBD5 protein isoform 1 (Q9P267), and is thus predicted to generate a novel protein of 1498 aminoacids of length. Allele 2 contains only the 4bp deletion, that is predicted to lead to a prematurely truncated peptide of 80 aminoacids of length.

Cell lines *6het AIIIB5* and *6hom AIID2* were edited heterozygously and homozygously, respectively, by NHEJ perfectly repairing cuts at the locations predicted by the guideRNA sequences without additional modifications (Figure 20). This mutation led to the excision of 41bp within exon 6, removing the initial portion of the MBD domain. As this modification causes a frameshift of the open reading frame, it is predicted to result in a premature truncation and generate a 38 amino-acid peptide, when considering the canonical MBD5 protein isoform 1 (Q9P267) (MBD5 protein levels could not be assessed as no suitable antibodies were found upon stringent testing of commercially available reagents).

In summary, of the selected CRISPR-edited cell lines, 2 of them show indels within the intron upstream of exon 4 and therefore are not expected to have a functional impact in *MBD5* usual transcription. Regarding the other 4 cell lines, they harbor mutations in coding regions that may impact *MBD5* transcription and/or translation.

Table VII - CRISPR lines and description.

Cell line ID	*CRISPR vs Control*	*gRNAs Used*	*FACS method*	*Mutation Type*	*Mutation Description*	*Predicted Translational Consequence (ExPASy)*
8330-8 Non-Treated	Non-treated Control	-	10cm dish	-	-	-
8330-8 Treated GM	Treated Control	-	10cm dish	-	-	-
H8-	Negative Control*	4a, 4b	96w	-	-	-
4i H6	CRISPR Exon 4	4a, 4b	96w	Heterozygous	3 bp insertion and 1 bp substitution C>A in intron upstream of exon 4	None (intronic)
4i H7	CRISPR Exon 4	4a, 4b	96w	Heterozygous	19 bp deletion in intron upstream of exon 4	None (intronic)
4i AIID6	CRISPR Exon 4	2, 6	10cm dish	Heterozygous	9bp insertion and G>T substitution in exon	None (5' UTR exon)
6het AIVG12	CRISPR Exon 6	8, 5	10cm dish	Compound Heterozygote	Mutation 1: 4bp deletion + 16bp insertion	In-frame mutation with 4 aa insertion and 1 aa substitution (1498 aa protein)
					Mutation 2: 4bp deletion	Premature Truncation (80 aa peptide)
6het AIIIB5	CRISPR Exon 6	8, 3	10cm dish	Heterozygous	41bp deletion in exon	Premature Truncation (38 aa peptide)
6hom AIID2	CRISPR Exon 6	8, 3	10cm dish	Homozygous	41bp deletion in exon	Premature Truncation (38 aa peptide)

*treated with CRISPR but did not make a mutation (confirmed by Sanger). aa- aminoacid.

Differentiated NPC showed typical morphology during differentiation and expressed NPC-specific markers

To gain further insight into ASD pathophysiology in cell populations such as neurons, we generated and characterized iPSC-derived NPC. These were generated by first differentiating the iPSCs into embryoid bodies (EBs), followed by dissociation, isolation and expansion of neural rosettes in the presence of the mitogens EGF and bFGF (Figure 17). The morphology of EBs was evaluated using light microscopy. EBs exhibited the typical spherical and well-limited appearance of EBs formed from embryonic stem cells (Figure 17). The neural rosettes derived from iPSC properly mimicked the apicobasal organization, forming a radially organized pattern in 2D, much like the neural tube epithelium (Figure 17). Neural rosettes recapitulate many aspects of brain development, including proper lineage progression and timed neuronal specification.

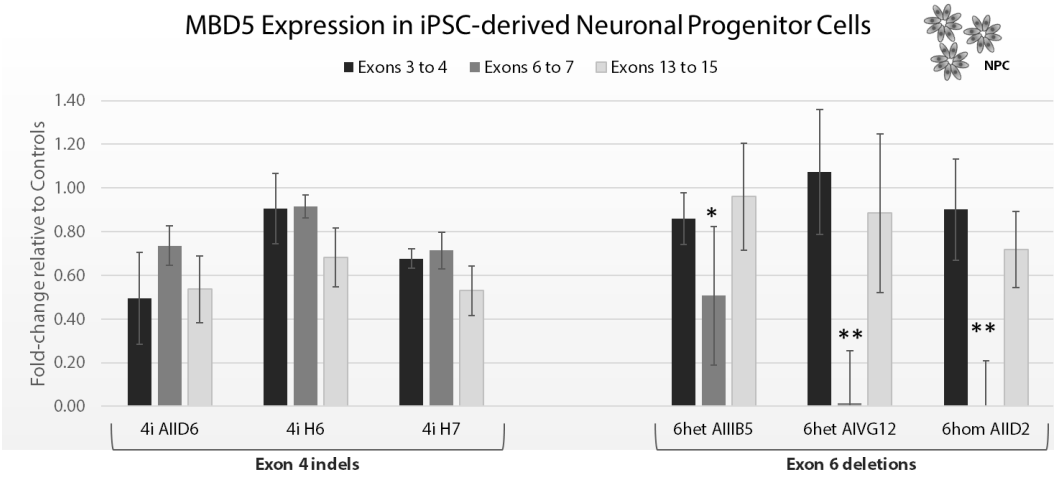


*Figure 21 - MBD5 exon mRNA expression in iPSC-derived NPC. Fold-changes were calculated relative to housekeeping genes ACTB, GAPDH and POLR2A and normalized to the three control samples (see Table I). *pval<0,05. **pval<0,01.*

To confirm that the decrease in *MBD5* mRNA levels in CRISPR-edited cell lines was maintained after differentiation into NPCS, we performed qRT-PCR analysis (Figure 21). This analysis revealed the decreased level was maintained in cell lines *4i AIID6* and *4i H7*, showing levels similar to those seen in haploinsufficient patients (~50% reduction, in comparison to controls). The exon 6-targeted cell lines along with cell line *4i H6* showed less decrease of *MBD5* expression, in comparison to that seen in iPSC. Primer pair targeting exon 6 to 7 confirms the deletions are present in exon 6-targeted cell lines, as the forward primer sits within the deleted regions. Indeed, this shows that the mRNA expression in that region is of about 50% of the controls in the heterozygous

cell line *6het AIIIB5* and is essentially null both in the homozygous line *6hom AIID2* and compound heterozygote *6het AIVG12*, as expected.

To test for cell cycle impairments, since MBD proteins have been described to regulate this process and also as some cell lines seemed to proliferate at different rates, a cell cycle assay using propidium iodide to stain nuclear DNA was used to assess differences in cell proliferation induced upon differentiation of the CRISPR-edited cell lines in comparison to the control cell lines (Figure 22). Indeed, there seemed to be a trend towards decreased proliferation in the treated vs. un-treated cell lines as there was a longer G0/G1 phase. However, the differences were not statistically significant.

The presence of NPC-specific markers NESTIN, PAX6 and SOX1and SOX2 in the differentiated cell lines indicates the differentiation efficiency. The immunocytochemical analysis of the NPC-specific markers in the expanded neuronal progenitor cells showed positive expression all cell lines, with the exception of the SOX2 marker in the compound heterozygote cell line *6het AIVG12* (Figure 23). This indicates that this cell line did not complete the differentiation process entirely.

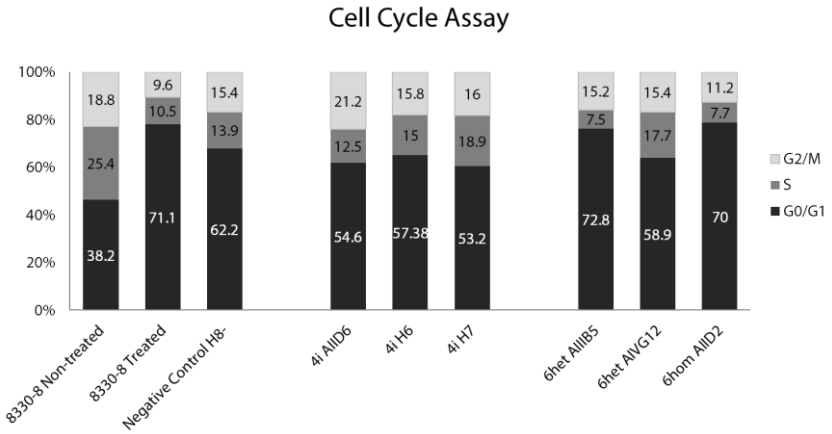


Figure 22 - Cell cycle assay using propidium iodide as a reference of DNA doubling. The treated cell lines show a trend of a larger G0/G1 phase.

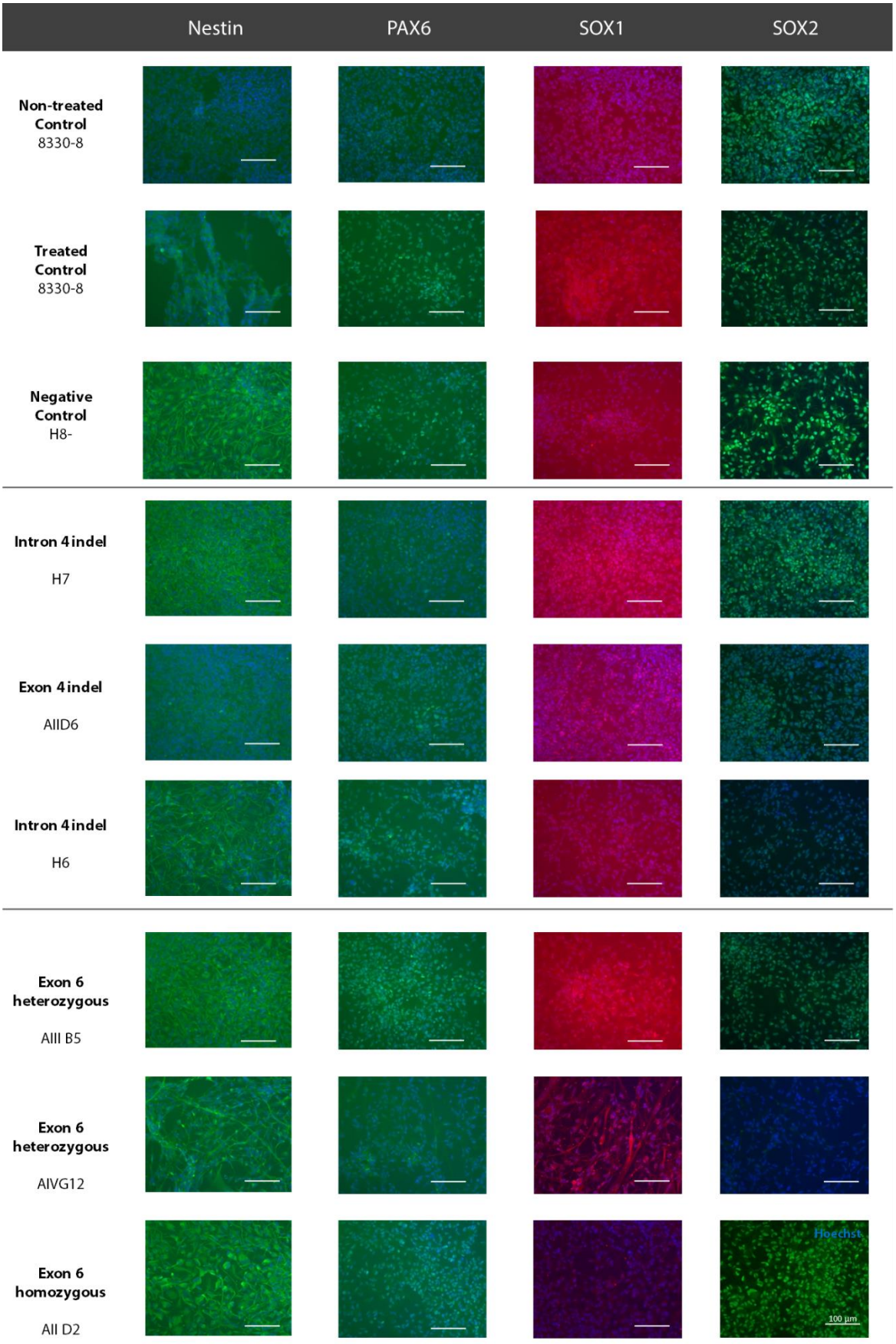


Figure 23 - Immunocytochemical analysis of NESTIN, PAX6, SOX2 (green, nuclei DNA staining overlaid in blue) and SOX1 (red, nuclei DNA staining overlaid in blue) expression in expanded neuronal progenitor cells from iPSC lines expanded in the presence of mitogens EGF and bFGF.

CRISPR-edited neurons displayed typical morphology, expressed terminally differentiated markers and near-regular levels of *MBD5*

Neurons are of great interest as they represent the primary cell type that is thought to be affected in ASD as well as in other neurodevelopmental disorders. We analyzed the differentiated iPSC-derived neuronal progenitor cells after withdrawal of mitogenic factors (EGF, bFGF) for 30 days to guide them into the terminally differentiated neuronal lineage. To assess the *MBD5* mRNA levels in CRISPR-edited neuronal cells after complete differentiation, we performed qRT-PCR analysis (Figure 24). Upon differentiation, all CRISPR-edited cell lines show *MBD5* expression levels similar to the controls. In fact, in exon 4-targeted cell line *4i H7*, the expression is even higher than that observed in controls. Once more, the primer pair targeting exon 6 to 7 confirms the deletions in exon 6-targeted cell lines, indicating there is no mixture with other cells (such as wild-type).

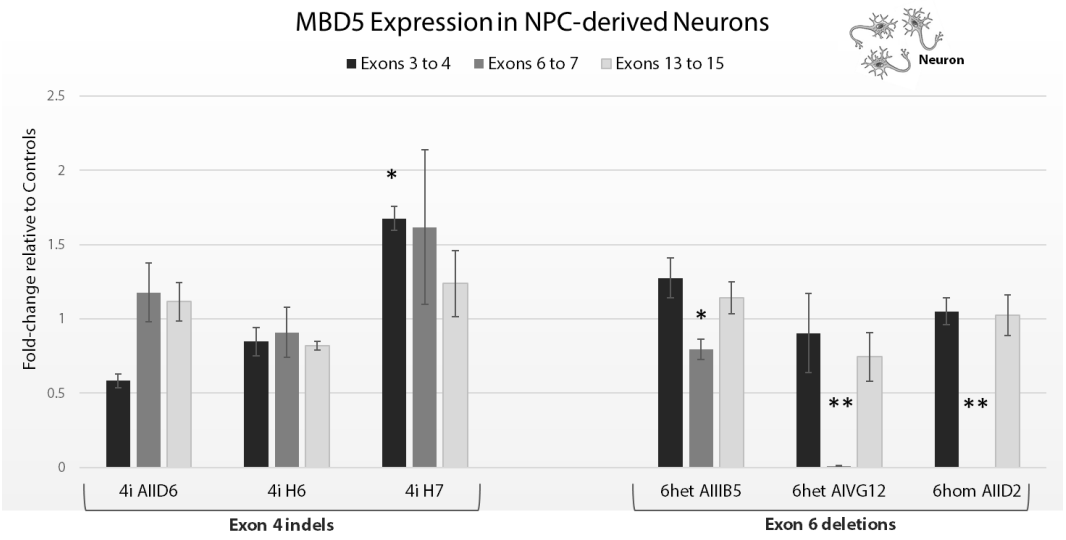


*Figure 24 - MBD5 exon mRNA expression in NPC-derived Neurons. Fold-changes were calculated relative to housekeeping genes ACTB, GAPDH and POLR2A and normalized to the three control samples (see Table I). *p_{val}<0,05. **p_{val}<0,01.*

Immunocytochemistry for neuron-specific markers MAP2 and SMI312 was positive for all cell lines with the exception of the compound heterozygote cell line *6het AIVG12* (Figure 25). In fact, this cell line, exhibited neither a neuronal-like morphology nor the specific markers, indicating that it failed to differentiate into neurons. All other cell lines exhibited both expression of neuronal markers and neuronal-like morphology. MAP2 is a dendrite marker while SMI312 is a neurofilament marker of axons. Neurofilaments are a major component of the neuronal cytoskeleton whose function is to provide structural support for the

axon and to regulate axon diameter. The presence of these markers in the differentiated cell lines are an indicator of success during the maturation process.

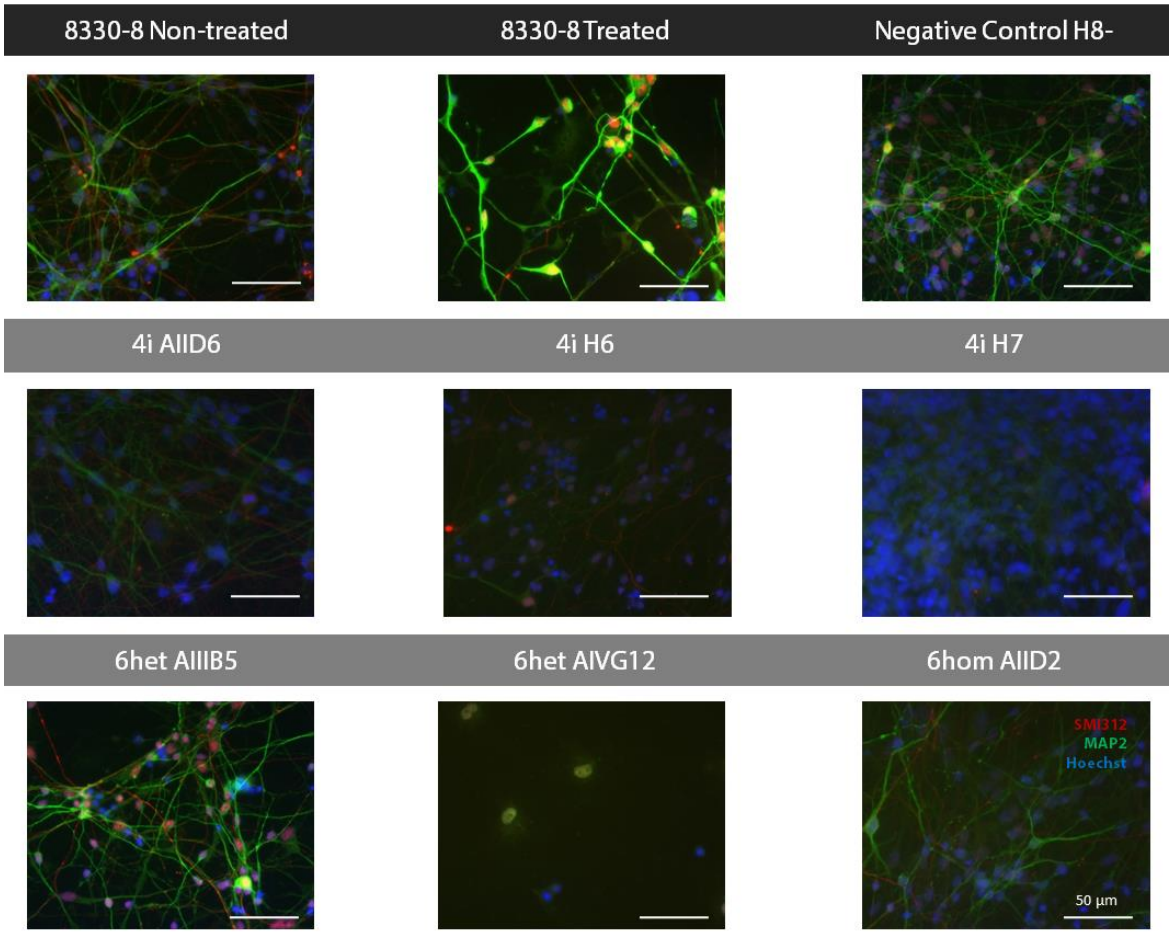


Figure 25 - Immunofluorescence staining for neuronal-specific markers, MAP2, a microtubule-associated protein (green, nuclei DNA staining overlaid in blue), and SMI312 (red), a neurofilament axonal marker in NPC-derived neuronal cells grown in the absence of mitogens EGF and bFGF.

The neuronal differentiation process generated a pool of cellular subtypes

The expression of embryonic, progenitor and neuronal markers and other lineage-specific markers to rule out possible aberrant differentiation were studied by RNAseq. The gene expression is presented in Figure 26 as the logarithmic transformation of counts normalized to library size and gives insight into the cellular maturation during the differentiation process. The results showed that both NPCs and neurons lacked expression of the pluripotency markers, indicating that the iPSC were capable of exiting a stem cell-like state by being induced towards differentiation. NPCs showed expression of NPC-specific markers as previously observed in the immunostaining experiments: Nestin, SOX1, SOX2 and PAX6. The transcriptomic data confirms that NPC cell line *6het AIVG12* does not express SOX1 and SOX2 concordant with what was observed by cell staining (indicated by the black arrows in the panel).

Neurons show expression of terminally differentiated neuronal proteins as: microtubule-associated protein 2 (MAP2), a neuron-specific protein that promotes assembly and stability of the microtubule network; SATB2, FOXG1 and CUX1, expressed by cortical neurons; and of a marker of neuronal migration, doublecortin (DCX). These data show that this differentiation protocol leads to a mixed population of neuronal subtypes such as glutamatergic excitatory (SLC17A6) as well as GABA inhibitory (GAD1, GAD2) neurons.

The growth of glial cells co-occurred along with the neuronal cells. This was noted by the presence of GFAP, present in astrocytes and CNP that is expressed in oligodendrocytes. Indeed, neurons and glia (oligodendrocytes and astrocytes) are originated from the same embryonic ectodermal layer and can both derive from neural progenitor cells (Kriegstein & Alvarez-Buylla 2009). Glial cells support the neuronal cell and are essential for neuronal signaling. Presence of vimentin (VIM), also highlights the presence of these cells. The intermediate filament vimentin is a cytoskeletal component of astroglial cells. However, it should be noted that it has also been reported to be expressed in neurons during developmental periods and under conditions of damage (Tanapat 2013). VIM is initially expressed by nearly all neuronal precursors in vivo, and is replaced by neurofilaments (NFs) shortly after the immature neurons become post-mitotic (Yabe et al. 2003). In our cells, VIM is visibly more elevated within the neuronal progenitor cells and not so pronounced in mature neurons.

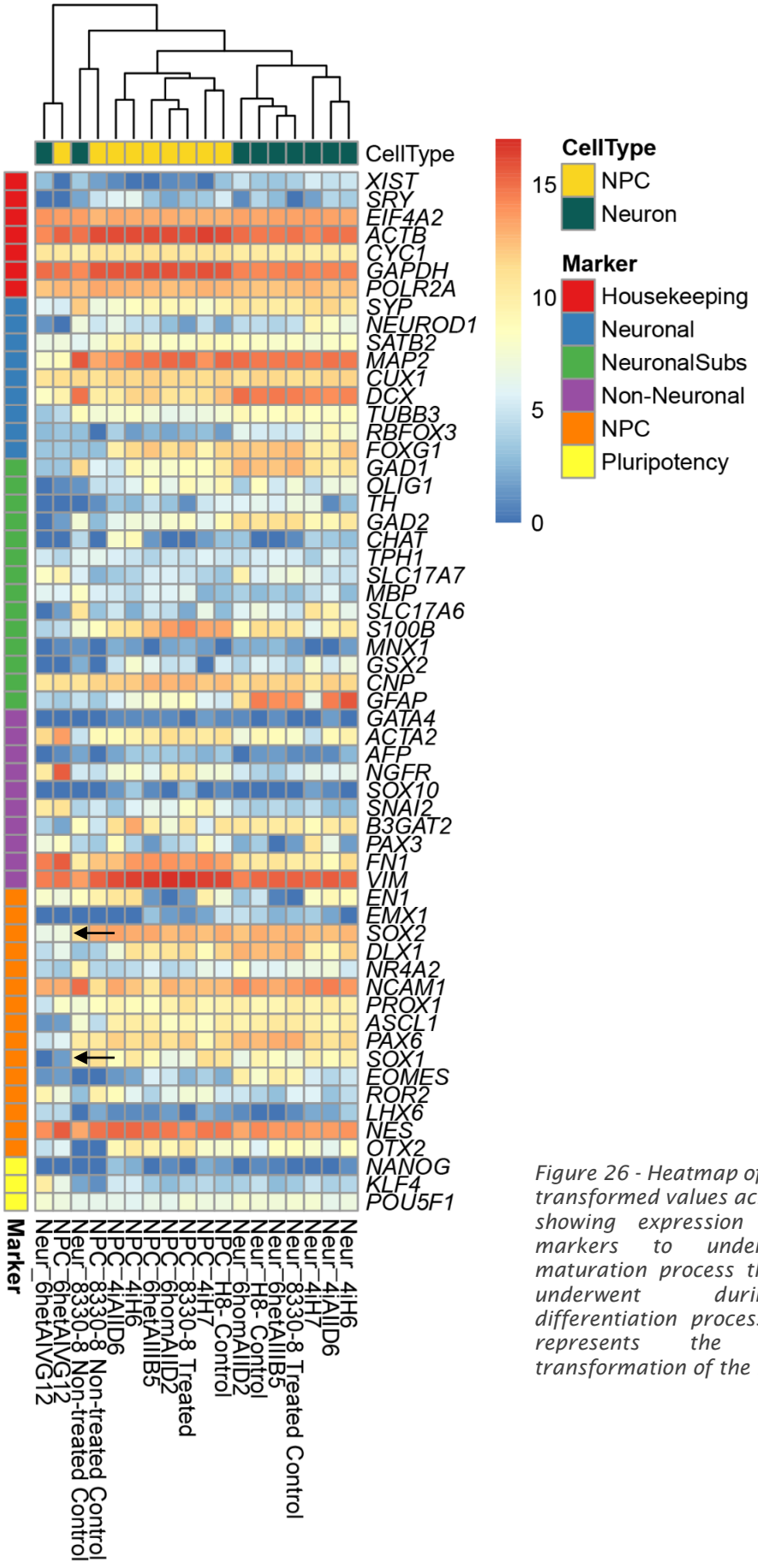


Figure 26 - Heatmap of relative log-transformed values across samples showing expression of selected markers to understand the maturation process that the cells underwent during the differentiation process. The scale represents the logarithmic transformation of the read counts.

Principal component analysis revealed the RNAseq dataset dimensions

To measure the transcriptomic changes that occur upon perturbation of the 5' UTR exon 4 and of the first coding exon that contains the initial portion of the MBD domain - exon 6, we used whole transcriptome sequencing (RNAseq) to sequence all the mature RNA (mRNA) strands present in the cells.

A first assessment used to explore this data was the hierarchical cluster dendrogram. This dendrogram allows us to assess overall similarity between samples as the algorithm behind this method successively pairs together samples showing the highest degree of similarity. A hierarchical cluster dendrogram of this dataset was built and showed that the cell lines cluster into their cell-type categories: NPC and neurons (Figure 27). This dendrogram clearly separates the transcriptome-wide signatures between cell type - NPCs and neurons. Indeed, this is expected as cell types should be clearly demarcated from one another by their individuality as a tissue. Besides, as mentioned previously, the compound heterozygote *6het AIVG12* cell line did not differentiate into the mature neuronal lineage and, indeed, it segregated along with the NPC cell lines in terms of their transcriptional profile. This indicates this cell line preserved a NPC-like phenotype despite the mitogen withdrawal and consequently was excluded from further analyses. Both the non-treated 8330-8 controls cluster within their cell category, separate from the 8330-8 treated cells, indicating the CRISPR treatment induces some level of transcriptomic changes in those cells. For this reason, the 8330-8 non-treated cell lines were not considered for downstream expression analyses to assure the best matched controls.

We then used a principal component analysis (PCA) to visualize sample-to-sample distances, in terms of genome-wide transcriptomic landscape, that reflect the dataset's variance. The percentages that each principal component explain and the PCA plots combining the several components, are represented in Figure 28 for NPCs and Figure 29 for neurons. Principal component 1 (PC1) and PC2 are the directions that separate the data points the most and second most, respectively, and explain most of the variance found in the RNAseq dataset. In this dataset, PC1 explains 35.2% and 46.2% of the variance in NPC and neurons, respectively, while PC2 explains 23.8% and 23.4% of the variance in NPC and neurons, respectively. The first two components when plotted clearly

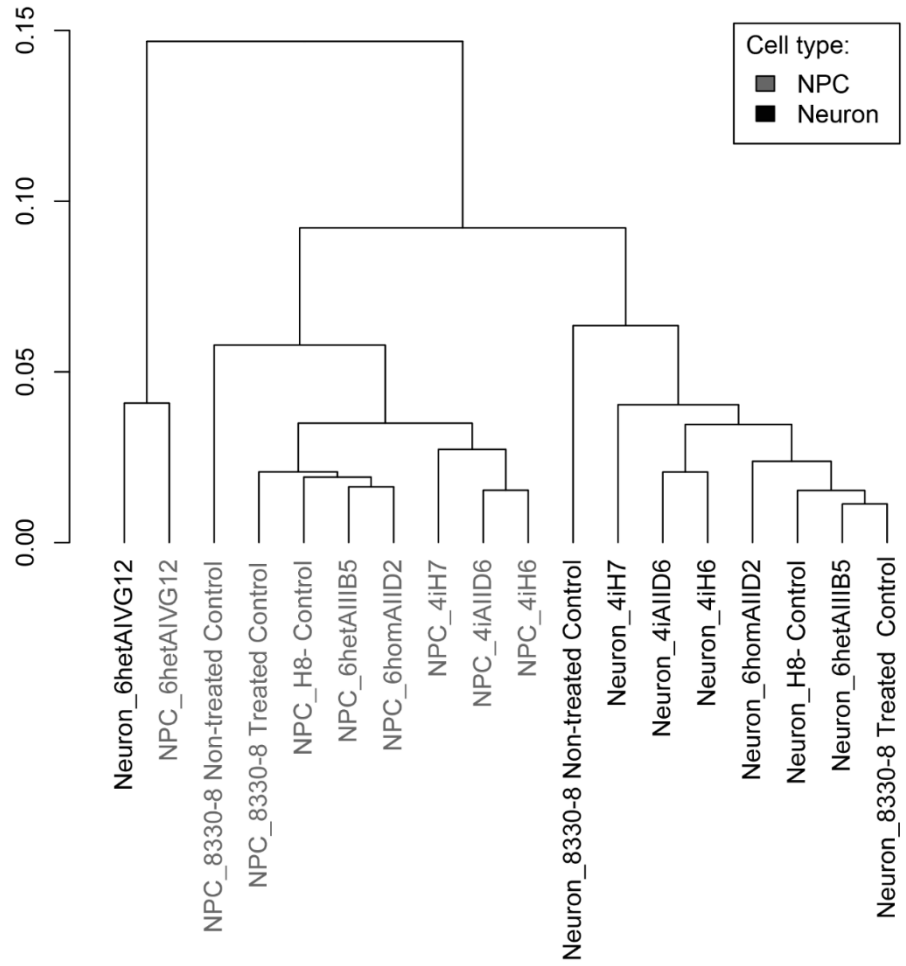


Figure 27- Hierarchical cluster dendrogram shows clustering of the cell lines into their respective categories: NPC and Neurons (Blue – control samples). Cell lines 6het AIVG12 clustered apart, demonstrating its incomplete maturation into the neuronal lineage.

segregated sample *6het AIVG12* from the remaining cell lines, possibly indicating that their differentiation diverged from all other cell lines and did not follow the complete neuroectodermal lineage differentiation. Additional components individually explain less than 15% of the variance found in the dataset, each.

Overall, the PC analysis did not distinguish between wild-type and CRISPR-edited cell lines, indicating that their impact on the transcriptome was not responsible for the variance found in the first components. Despite not identifying differences at a large-scale transcriptomic level, there are possibly other features that will be more prevalent in explaining the variation between the lines such as a modest number of differentially expressed genes or defects concerning specific aspects of neuronal function, e. g. synaptic connectivity.

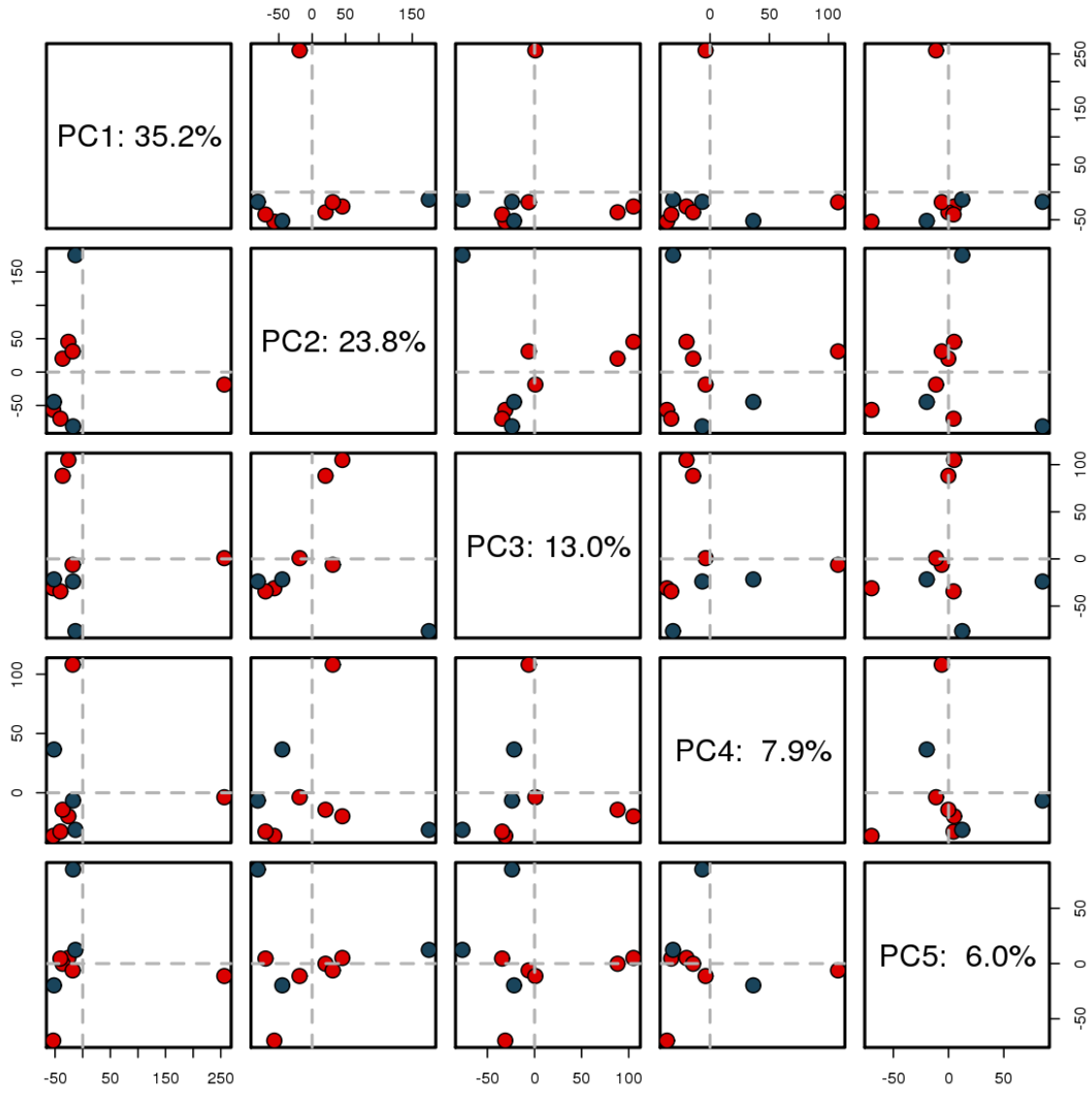


Figure 28 - Principal component analysis of the several dimensions of the RNAseq dataset allow us to determine the variance explained by each component within the NPC - wild-type (blue) and CRISPR-edited cells (red).

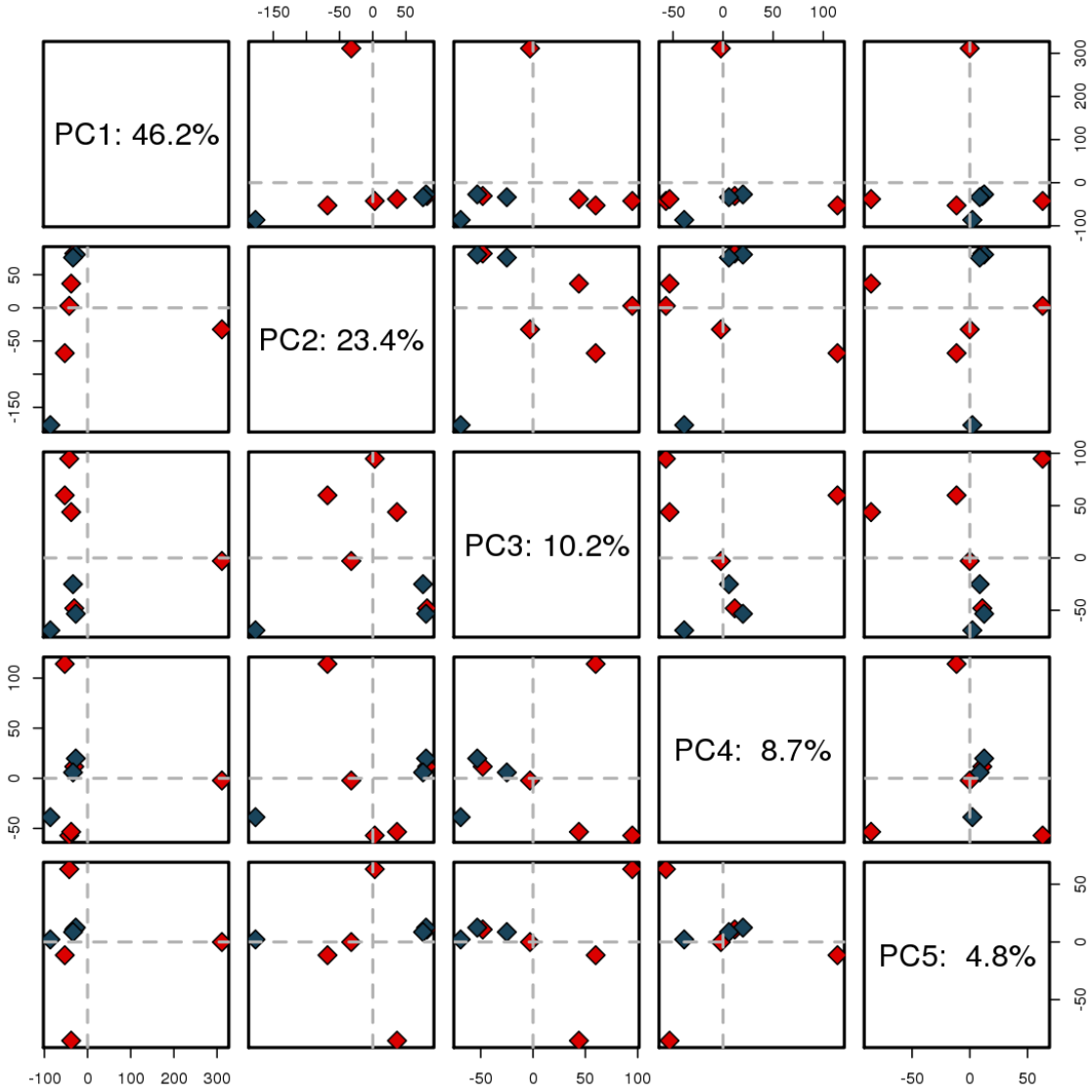


Figure 29 - Principal component analysis of the several dimensions of the RNAseq dataset allow us to determine the variance explained by each component within the neuronal cells - wild-type (blue) and CRISPR-edited cells (red).

MBD5 expression in NPC and neurons was consistent in the RNAseq dataset and in the qRT-PCR

Since the PCA showed that genotype did not explain most of the variance found, we then looked at the *MBD5* mRNA expression levels in the RNAseq dataset, both of the NPC and neuronal cell lines, to confirm whether these levels were maintained in the biological replicates used to generate the RNAseq data. For these analyses, only the exon-targeted cell lines were considered along with the treated controls (*8330-8 treated control* and *H8- negative control* cell lines), using DESeq. We observed that the mRNA expression levels observed in the transcriptomic dataset (Figure 30) recapitulate those detected by qRT-PCR (Figure 30). Regarding the NPC dataset, the CRISPR-edited cell lines show no significant overall difference in *MBD5* expression in comparison to the controls. In contrast, in the neuronal dataset, the CRISPR-edited cells exhibit a higher level of expression of *MBD5* mRNA, when compared to the treated controls. These results support the previous results obtained via qRT-PCR, ruling out any possible technical artifacts and indicating possible compensation mechanisms occurring in the edited cells.

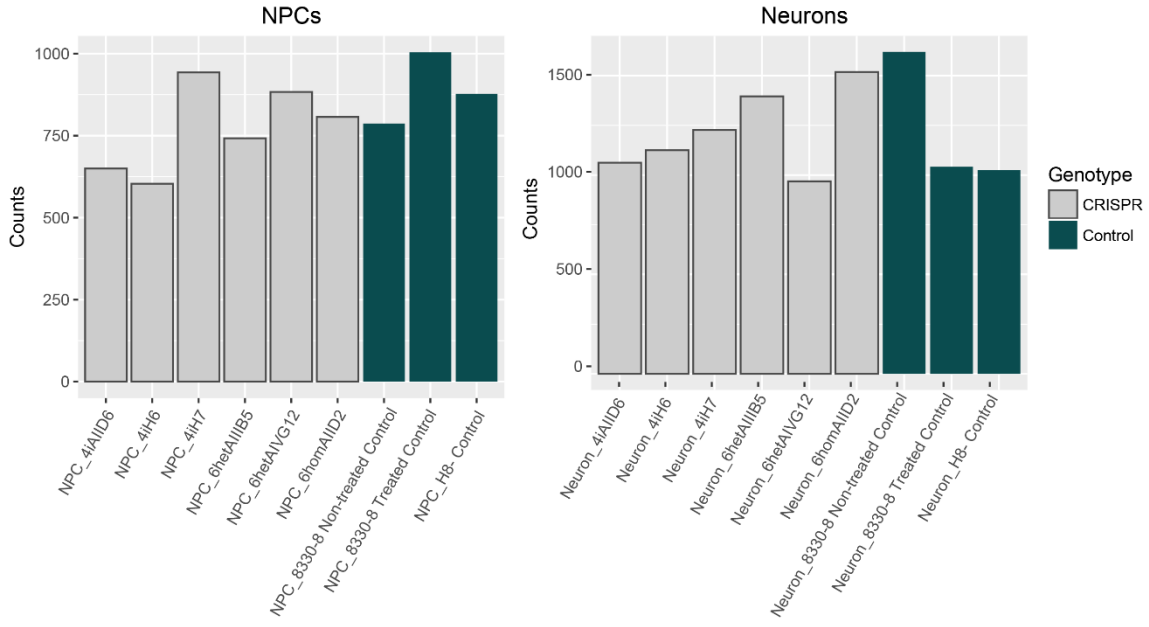


Figure 30 – MBD5 mRNA expression in selected cell lines, from the RNAseq dataset obtained using DESeq2. The expression levels are similar to those observed in the qRT-PCR assay.

Differential *MBD5* transcript expression in NPCs and neurons suggested that MBD might not be crucial for neurodevelopment

To determine what occurred locally within the *MBD5* gene upon editing, the local patterns of *MBD5* gene expression were determined to identify possible changes in its transcriptional pattern during both stages of differentiation, using RSEM and Bowtie2.

There are 33 *MBD5* transcripts annotated in Ensembl, of which 7 are predicted to be protein coding (Supplementary Table 5). Within our RNAseq dataset, 9 *MBD5* transcripts were detected and are listed in Table VIII. Of the 4 protein-coding transcripts identified, including the canonical transcript MBD5-001 that gives rise to MBD5 protein isoform 1 (Q9P267), only 2 of them contain exon 6 (and thus, the MBD domain), indicating there are at least 2 other transcripts that do not require this exon to produce a shorter length protein, not previously described in the literature (Figure 31). All exon-exon junctions from the alternative transcripts were confirmed to be present in the RNAseq reads, using IGV to visualize the individual reads (data not shown). All exonic mutations

Table VIII - MBD5 transcripts detected in the iPSC-derived NPC and Neurons.

Ensembl ID	Alternative Name	*bp*	*Prediction*	*Protein*	*UniProt*
ENST00000407073	MBD5-001	9512	Protein coding	1494aa	Q9P267
ENST00000627651	MBD5-012	5610	Protein coding	851aa	Q9P267
ENST00000416015	MBD5-005	4343	Protein coding	1064aa	H7C066
ENST00000630352	MBD5-014	663	Protein coding	77aa	A0A0D9SEP6
ENST00000478190	MBD5-010	4131	Processed transcript	No protein	-
ENST00000496893	MBD5-003	2788	Retained intron	No protein	-
ENST00000628572	MBD5-015	3595	Nonsense mediated decay	910aa	A0A0D9SF16
ENST00000488372	MBD5-007	548	Processed transcript	No protein	-
ENST00000496158	MBD5-006	544	Processed transcript	No protein	-

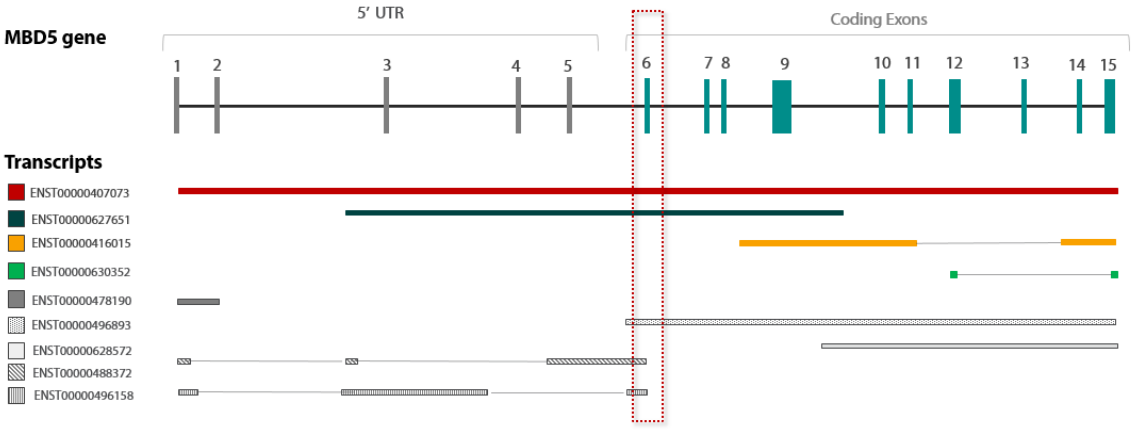


Figure 31 - Transcripts affected by exon 6- and MBD domain-targeted CRISPR deletions. Top: MBD5 gene structure with exons numbered based on transcript MBD5-001.

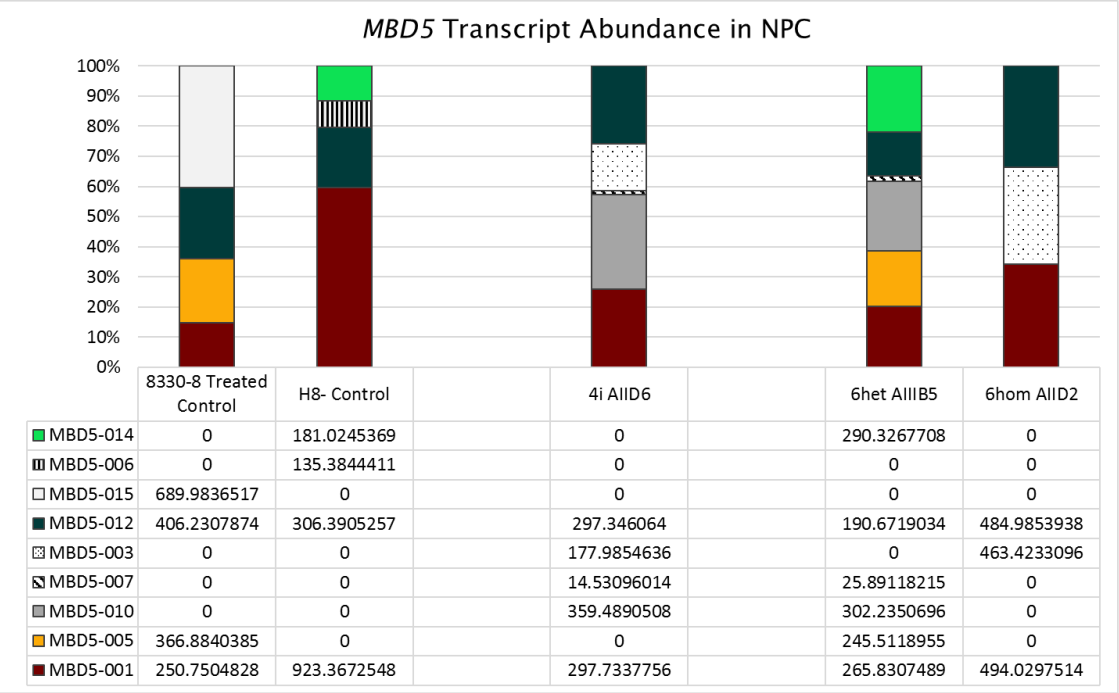
were present in the dataset, with the exception of cell line 4i *AIID6*, where no reads containing the exonic mutation were identified, indicating those transcripts underwent nonsense mediated decay.

Differential transcript abundance analysis revealed that the canonical transcript MBD5-001 was not the most abundant either in the NPC or neuronal cell lines, with the exception of the NPC *H8- negative control* cell line (Figure 32). Surprisingly, the transcript showing highest overall expression across neuronal cell lines was MBD5-010. This short transcript is comprised of only the 5'UTR exons 1 and 2, and is predicted to result in a processed transcript that does not encode a functional protein. This is also verified in the GTEx database, as the second most expressed transcript in the maority of tissues and may suggest an important regulatory role for this transcript in MBD5 function.

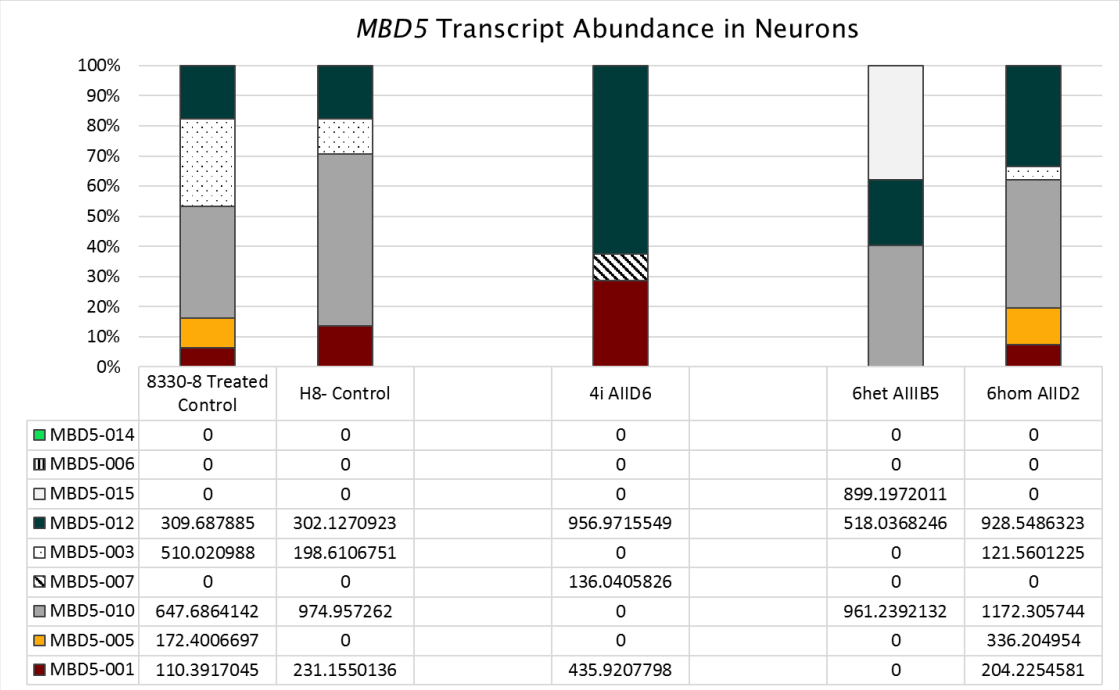
Transcript MBD5-015 was differentially overexpressed in the exon 6-targeted CRISPR neurons in comparison to the matched controls and also in comparison to NPC exon 6 CRISPR-edited cell lines. Transcript MBD5-010 was unaffected by exon 6 CRISPR-targeting in the neuronal cell lines, however it was differentially overexpressed in the NPC CRISPR-edited cell lines in comparison to the matched controls.

One transcript showed NPC-specific expression, MBD5-014, that was observed uniquely in the NPC population and absent in the neurons, suggesting a developmental state preference.

These results indicate that, the *MBD5* ablations had a local effect on the transcript isoforms expressed by this gene during the different stages of development. This shows that the cells were able to maintain normal levels of *MBD5* through the expression of non-canonical and even non-coding transcripts, suggesting that the MBD domain might not be as crucial for neuronal development as previously thought.



	8330-8 Treated Control	H8- Control	4i AIID6	6het AIIIB5	6hom AIID2
MBD5-014	0	181.0245369	0	290.3267708	0
MBD5-006	0	135.3844411	0	0	0
MBD5-015	689.9836517	0	0	0	0
MBD5-012	406.2307874	306.3905257	297.346064	190.6719034	484.9853938
MBD5-003	0	0	177.9854636	0	463.4233096
MBD5-007	0	0	14.53096014	25.89118215	0
MBD5-010	0	0	359.4890508	302.2350696	0
MBD5-005	366.8840385	0	0	245.5118955	0
MBD5-001	250.7504828	923.3672548	297.7337756	265.8307489	494.0297514



	8330-8 Treated Control	H8- Control	4i AIID6	6het AIIIB5	6hom AIID2
MBD5-014	0	0	0	0	0
MBD5-006	0	0	0	0	0
MBD5-015	0	0	0	899.1972011	0
MBD5-012	309.687885	302.1270923	956.9715549	518.0368246	928.5486323
MBD5-003	510.020988	198.6106751	0	0	121.5601225
MBD5-007	0	0	136.0405826	0	0
MBD5-010	647.6864142	974.957262	0	961.2392132	1172.305744
MBD5-005	172.4006697	0	0	0	336.204954
MBD5-001	110.3917045	231.1550136	435.9207798	0	204.2254581

Figure 32 – Differential MBD5 Transcript Expression in NPC (top) and neurons (bottom), generated using RSEM and Bowtie2 (For normalized transcript expression, TPM was multiplied by MBD5 transcript proportions).

MBD family expression levels upon *MBD5* editing remained unaltered

To determine if there was compensation by the MBD family members upon *MBD5* editing, we looked into the expression levels of the MBD family members, in NPC and neurons. Since the MBD proteins share similar roles in chromatin remodeling, the lack of a family member could induce the overexpression of another member, as seen in other systems (Kong et al. 2007). The MBD family gene expression observed in the RNAseq dataset is presented in Figure 33 as the logarithmic transformation of counts normalized to library size. When observing the expression of the other family members, it is noted that there were changes in MBD1 and MECP2 during the transition from NPC to neuron. Indeed, MECP2 is known to be crucial for neuronal function (Chahrour et al. 2008; Nan & Bird 2001). Although these family members were upregulated during differentiation, no meaningful differences were found between controls and the CRISPR-edited cells, reinforcing and supporting the results previously obtained in the PCA.

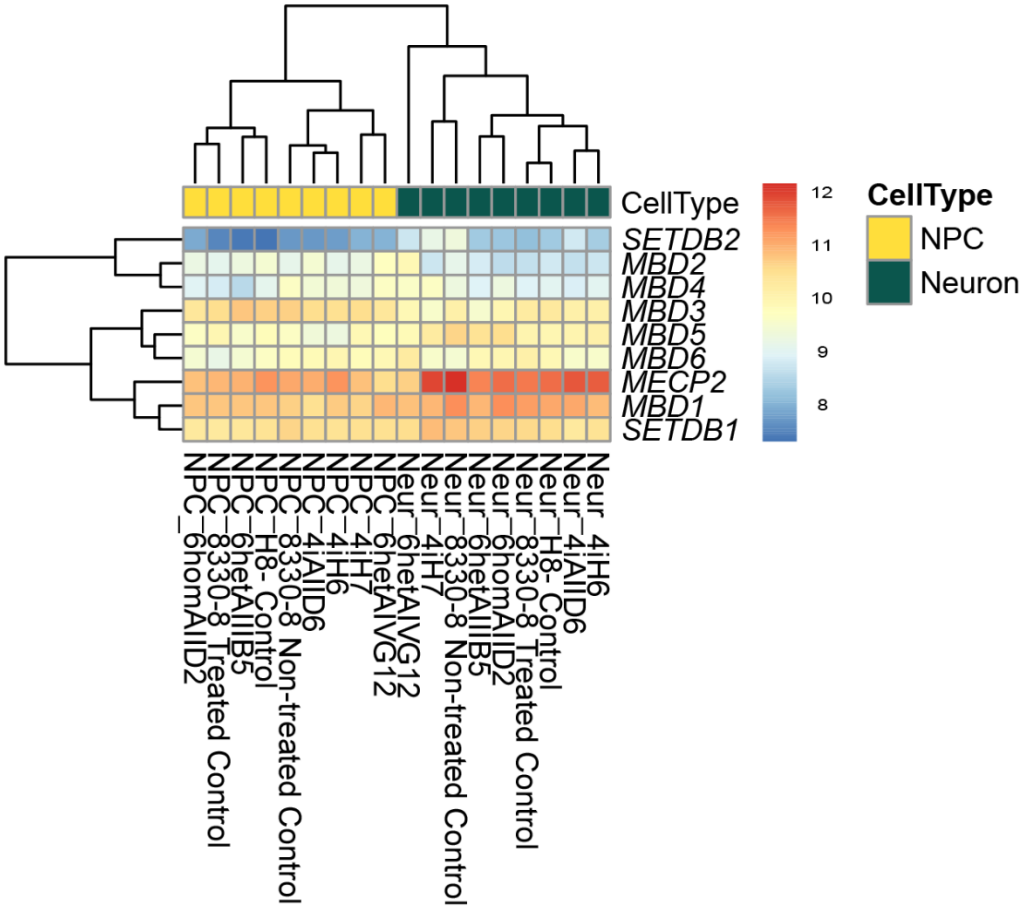


Figure 33 - Heatmap of relative log-transformed values across samples of the MBD family expression upon CRISPR editing.

Genome-wide transcriptomic analyses revealed gene expression changes

The ultimate goal of the whole transcriptome analyses was to determine the genes that were differentially expressed between the wild-type and the CRISPR-edited cell lines (excluding cell line *6het AIVG12*), via quantitative statistical analysis using DESeq. The identification of the dysregulated genes can bring insight on the protein complexes and pathways that are acting downstream of *MBD5* and are directly implicated in neuronal development and function.

Regarding NPCs, there were a total of 53 significantly differential expressed genes (DEG) (adjusted p-value <0.05) and an additional 529 genes that were significant at a nominal level (p-value <0.05). For the neuronal set, there were a total of 6 significantly DEG and an additional 161 that were nominally significant. Table IX lists the top 15 genes that were significantly differentially expressed between control and CRISPR cell lines, both in NPC (top panel) and in neurons (bottom panel). All other differentially expressed genes in NPC and neurons are listed in Supplementary Table 6 and Supplementary Table 7, respectively. Volcano plots resultant from the differential expression analyses, show the most DEG (Figure 34).

The most DEG in the NPC dataset was *RAB11FIP1*, that was found to be significantly upregulated. This gene was also upregulated in the neuronal panel, at a nominally significant level (p_{val} = 0.0004). *RAB11FIP1* is a member of the Rab11-family interacting proteins (Rab11-FIPs) that are critical regulators of intracellular vesicle trafficking and recycling and has been previously associated with axonal development (Eva et al. 2010; Schafer et al. 2016). Indeed, a study RAB11FIP1 has been shown to have a role in the recycling of integrins within axons during cell migration (Eva et al. 2010).

The second most DEG in NPC, was *NHLH1*, of which its family member *NHLH2* is also found to be upregulated in neurons. Nhlh1 and Nhlh2 are neural basic helix-loop-helix (bHLH) genes that have been implicated in mouse neurogenesis (Murdoch et al. 1999). Studies of expression in normal tissues demonstrated expression of NHLH1 and NHLH2 in the developing central and peripheral nervous system, most likely in developing neurons (Lipkowitz et al. 1992) and maintain migration and survival of neuronal precursor cells (Schmid et al. 2007).

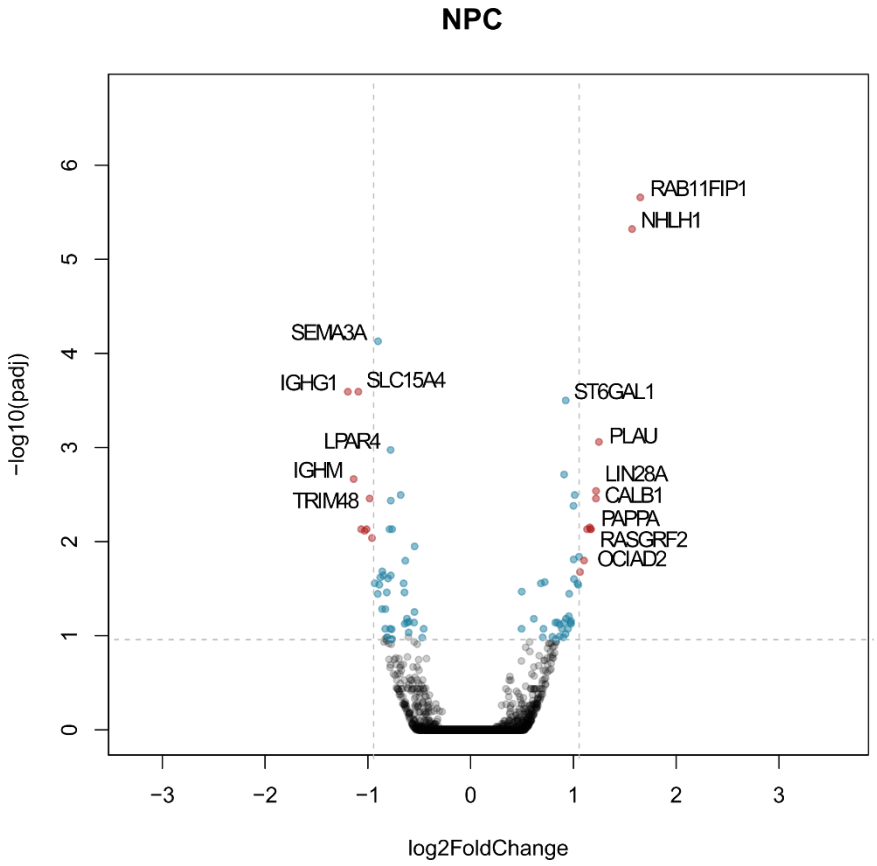
Additionally, 2 DEG present in Table IX have been previously implicated in ASD and are annotated in SFARI Gene Database – *PLAUR* and *CNTNAP2*. PLAUR is a urokinase plasminogen activator receptor which is thought to modulate availability of the MET ligand in the MET signaling pathway and to influence interneuron maturation (Eagleson et al. 2011; Campbell et al. 2008). On the other hand, CNTNAP2, contactin associated protein-like 2, is a member of the neurexins family and was among the first genes with evidence for both rare and common variation contributing to ASD (O'Roak et al. 2011; Alarcón et al. 2008). The human *CNTNAP2* gene is thought to be the largest gene in the genome, spanning approximately 2.3 Mb at chromosomal region 7q35–q36. Studies in mice show that it is involved in neuron–glia interactions in myelinated axons (Poliak et al. 1999) and in the migration of cortical projection neurons (Peñagarikano et al. 2011).

Representations of the gene networks formed by the differentially expressed genes in both NPC and neurons are depicted in Figure 36 and Figure 35, respectively. These networks were generated using the disease association protein-protein link evaluator (DAPPLE), that looks for significant physical connectivity among proteins encoded for by genes in loci associated to disease according to protein-protein interactions reported in the literature (Rossin et al. 2011).

Overall, these results suggest downstream players affected by the perturbation of *MBD5* that might play a role in neuronal development and function. These proteins will be interesting targets for ASD therapeutics if found to be dysregulated in other ASD knockdown models of chromatin remodelers, indicating commonality and convergence of downstream pathways.

Table IX - Top 15 differentially expressed genes using SVA analysis for CRISPR cell lines. (Genes names marked with '' represent those that are annotated in the SFARI gene database as ASD-associated genes.)*

	Gene ID	Name	Foldchange (log2)	P-Value	P-Adj.
NPC	ENSG00000156675	***RAB11FIP1***	1.595	1.1527E-10	2.0041E-06
	ENSG00000171786	***NHLH1***	1.515	5.00303-10	4.3491E-06
	ENSG00000075213	*SEMA3A*	-0.957	1.1685E-08	6.7719E-05
	ENSG00000139370	*SLC15A4*	-1.148	6.1351E-08	0.0002326
	ENSG00000211896	*IGHG1*	-1.250	6.6886E-08	0.0002326
	ENSG00000073849	*ST6GAL1*	0.869	9.9383E-08	0.0002879
	ENSG00000122861	*PLAU*	1.192	3.1958E-07	0.0007938
	ENSG00000147145	*LPAR4*	-0.835	4.4449E-07	0.0009660
	ENSG00000136068	*FLNB*	0.854	9.1171E-07	0.0017612
	ENSG00000211899	*IGHM*	-1.194	1.1323E-06	0.0019686
	ENSG00000131914	*LIN28A*	1.165	1.6674E-06	0.0026355
	ENSG00000117461	*PIK3R3*	-0.737	2.1747E-06	0.0029084
	ENSG00000011422	***PLAUR****	0.957	2.0136E-06	0.0029084
	ENSG00000104327	*CALB1*	1.164	2.5617E-06	0.0031735
	ENSG00000150244	*TRIM48*	-1.038	2.7379E-06	0.0031735
Neurons	ENSG00000164093	*PITX2*	-1.189	3.9039E-07	0.0037424
	ENSG00000197496	*SLC2A10*	-0.840	3.5661E-07	0.0037424
	ENSG00000224597	*PTCHD3P1*	-1.039	1.2410E-06	0.0079312
	ENSG00000100884	*CPNE6*	-1.083	3.2672E-06	0.0156607
	ENSG00000178401	*DNAJC22*	-1.069	1.0397E-05	0.0398677
	ENSG00000156675	***RAB11FIP1***	1.076	1.4576E-05	0.0465777
	ENSG00000131094	*C1QL1*	-0.894	2.216E-05	0.0606964
	ENSG00000177551	***NHLH2***	0.991	3.5457E-05	0.0841996
	ENSG00000173376	*NDNF*	1.006	4.4831E-05	0.0841996
	ENSG00000211445	*GPX3*	-0.819	6.4306E-05	0.0841996
	ENSG00000174469	***CNTNAP2****	0.841	6.5552E-05	0.0841996
	ENSG00000147246	*HTR2C*	1.021	4.6033E-05	0.0841996
	ENSG00000223508	*RPL23AP53*	-0.832	6.2236E-05	0.0841996
	ENSG00000109991	*P2RX3*	0.966	5.7608E-05	0.0841996
	ENSG00000101098	*RIMS4*	0.777	6.5874E-05	0.0841996



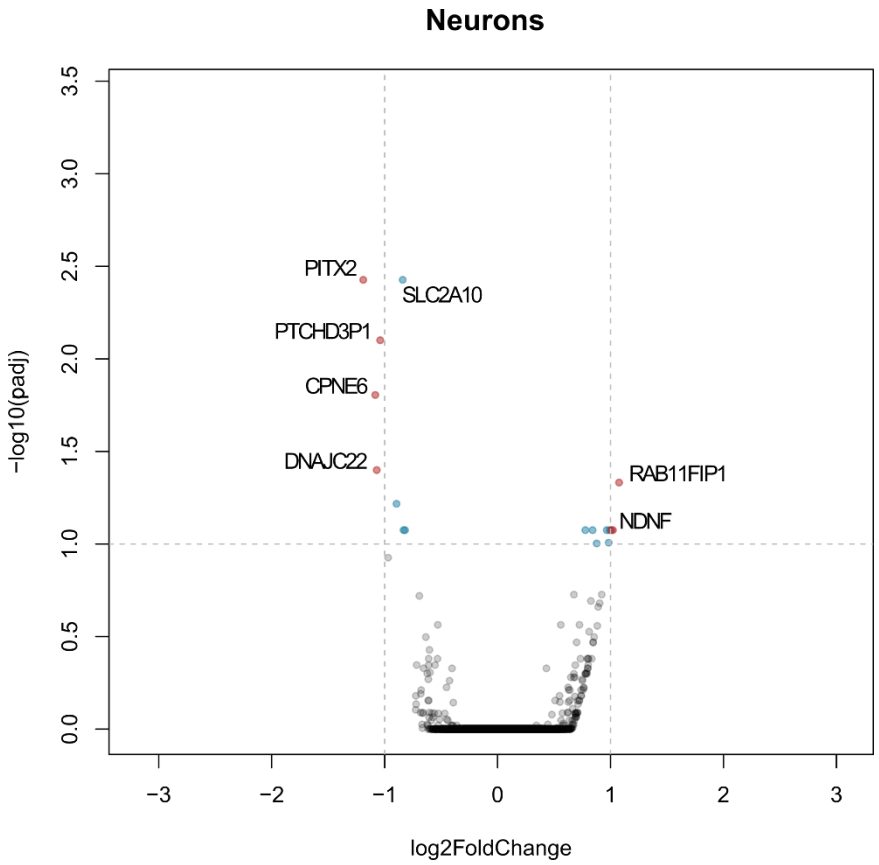


Figure 34 - Volcano Plots showing the DEG for NPC (top) and neurons (bottom). Genes that either have absolute log2 foldchange >1 or p-value <0.001 in NPC and 0.01 in neurons are represented in colored dots (Red dots – CRISPR cell lines; Blue dots – Controls). On the left side of the panels are the downregulated genes and on the right-hand side are the upregulated genes.

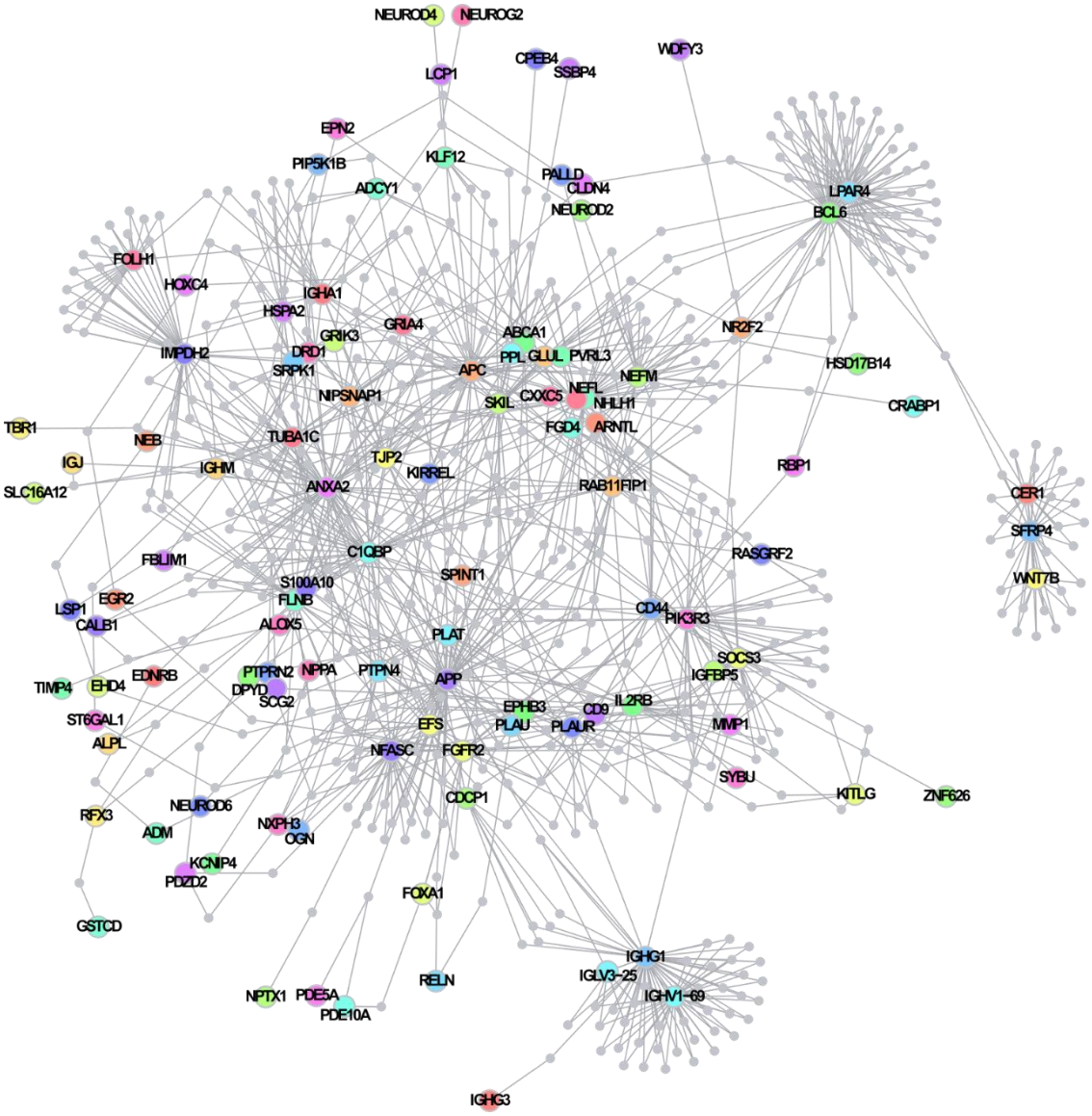


Figure 35 - Gene network formed by the top 200 nominally differentially expressed genes in NPC. Generated using the DAPPLE module within GenePattern.

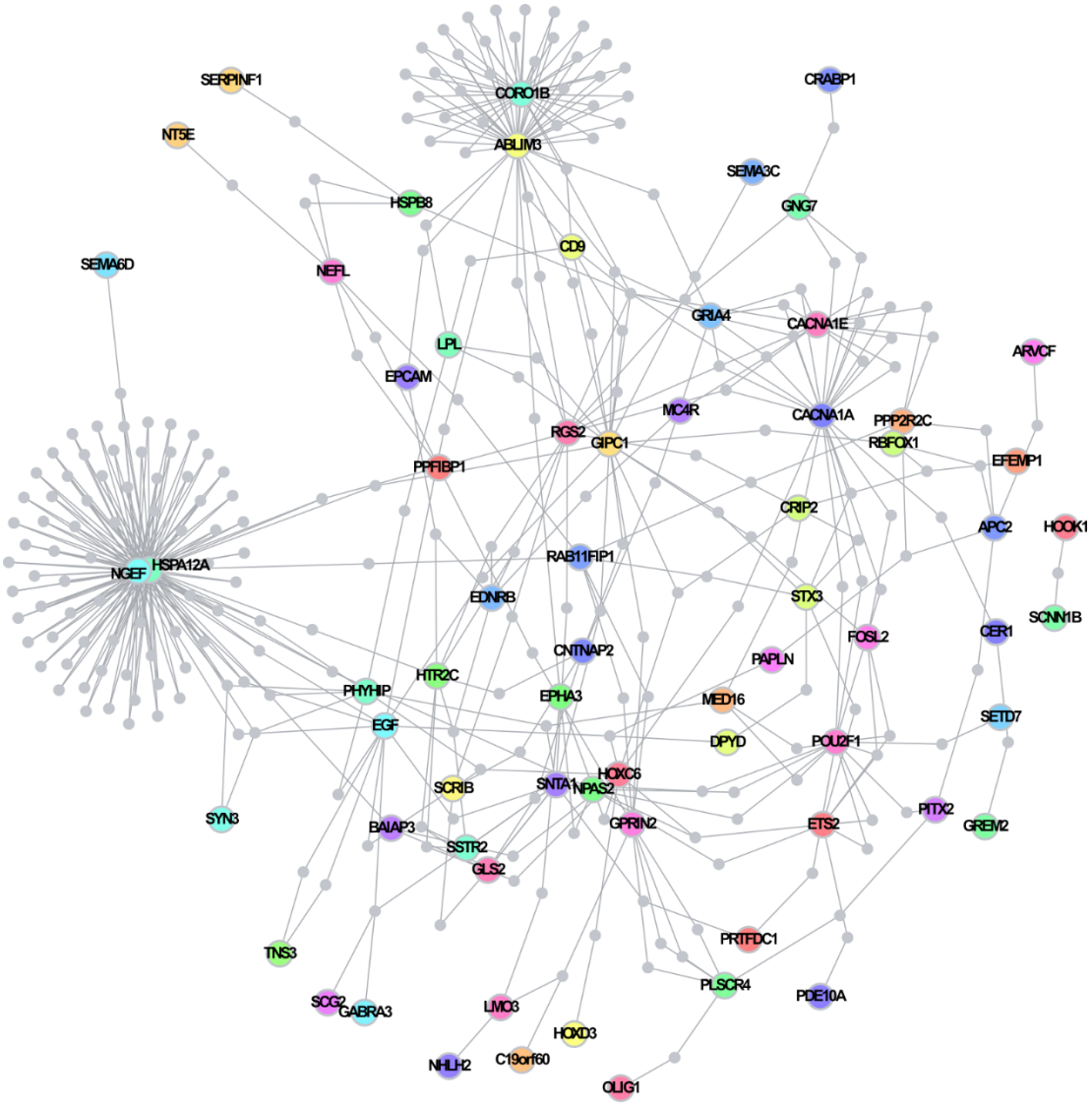


Figure 36 – Gene network formed by the 167 nominally differentially expressed genes in neurons. Generated using the DAPPLE module within GenePattern.

Pathway analysis of DEG showed enrichment of neuron, translation, and cell adhesion-related terms

Once the differentially expressed genes have been identified, a gene set enrichment analysis (GSEA) was performed to determine whether a priori defined gene sets show statistically significant differences between the control and CRISPR cell lines. This will pinpoint the biological significance of the observed expression changes using either a gene ontology (GO) or biological pathway-driven analysis (KEGG - Kyoto Encyclopedia of Genes and Genomes). Gene set approaches are based on the idea that complex diseases such as ASD can be better understood from the perspective of dysregulated gene sets than at the individual gene level. We performed enrichment analysis on all of the differentially expressed genes in both the differentiated NPC and neurons sets.

GO analysis of NPC revealed terms related to neuronal function such as: neuron differentiation, neuron fate commitment, dopaminergic synaptic transmission, glutamatergic synaptic transmission (p-value <0.05; Figure 37). On the other hand, the neuron set also included terms related to brain development: telencephalon and hindbrain development, myelination, Schwann cell differentiation and beta-catenin binding (p-value <0.05; Figure 37). An altered balance between excitation and inhibition has been postulated as a biological mechanism for ASD; this imbalance could arise from different risk genes differentially affecting either or both elements. Besides those, the most significant up-regulated biological processes in NPC were related to translation initiation, elongation and termination and ribosome (q-value <0.01). In fact, KEGG analysis also identified upregulated pathways involving ribosome and oxidative phosphorylation in NPC (q-value <0,01).

Regarding the neuronal set, some of the nominally significant pathways identified in KEGG analysis were the notch signaling pathway, RNA transport and cell adhesion (p-value <0.05). Notch is known to be a key regulator of adult neural stem cells, and Notch signaling also has a role in the regulation of migration, morphology, synaptic plasticity and survival of immature and mature neurons (Ables et al. 2011). On the other hand, cell adhesion genes have previously been associated with other chromatin remodelers that represent a risk for ASD, such as CHD8 (Sugathan et al. 2014). A full listing of GO terms and KEGG pathway enrichments are found in Supplementary Table 9 and Supplementary Table 8, respectively.

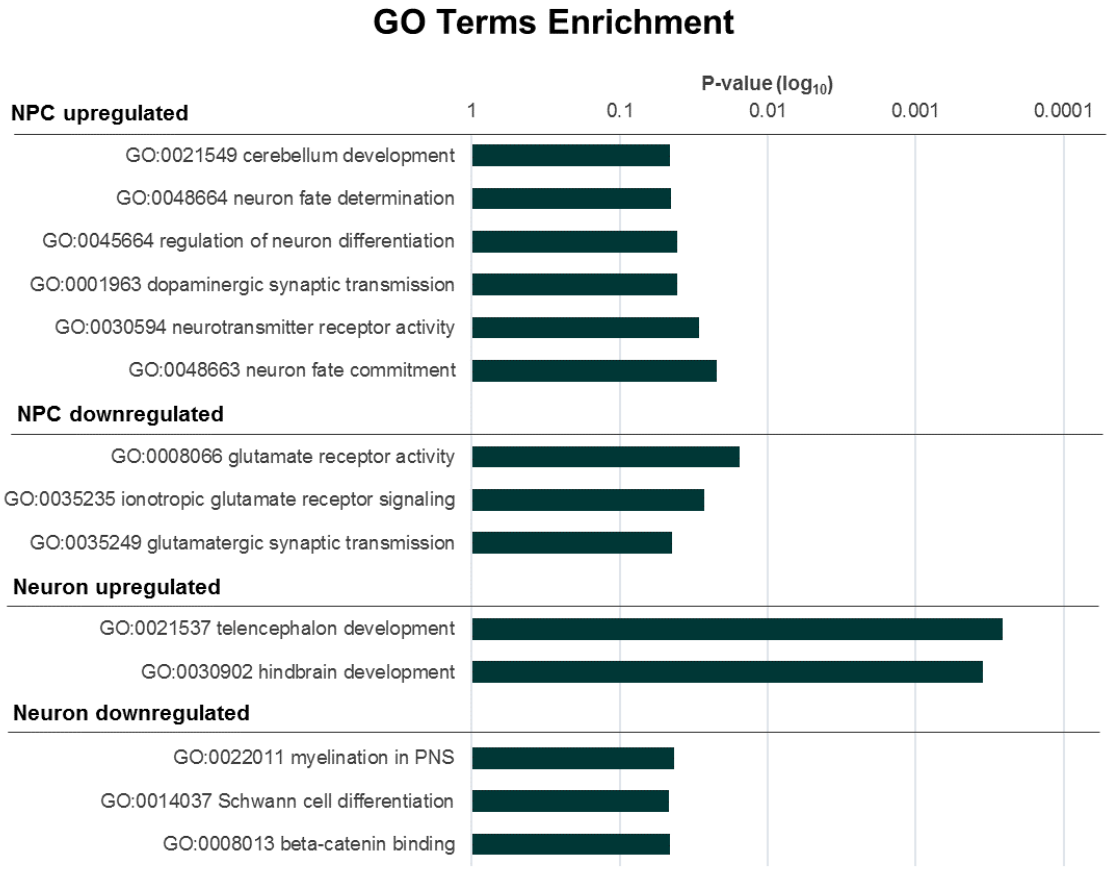


Figure 37 - GO terms show enrichment of neuron-related terms (p-value <0.05).

MBD5 is a prominent example among several genes, including other chromatin remodelers and transcriptional regulators (*ARID1B, CHD8*, *EHMT1*, *MECP2*), that have implicated the disruption of chromatin regulation as a precipitating factor in ASD (Sim et al. 2015; Talkowski et al. 2012; Santos et al. 2007). The *MBD5* gene is highly conserved and extremely intolerant to loss-of-function (LoF) mutations, as indicated by the ExAC browser pLi score of 1.00 (probability of LoF). Indeed, understanding the role of MBD5 in neurodevelopment is of great interest in order to link it to the consequences of its haploinsufficiency in ASD patients.

In this study, we aimed to generate CRISPR/Cas9 genome edited iPSC-derived neuronal cell lines bearing mutations in different regions of the *MBD5* gene, namely the 5' UTR exon 4 and the MBD Domain located in exon 6, to determine the genome-wide transcriptomic effects that occur upon these perturbations. Exon 4 represents a region within the 5' UTR where several mutations have been reported in patients, indicating its role for MBD5 haploinsufficiency, despite its location in a non-coding region of the gene. Thus, the contribution of this region to gene function has not been fully understood and might be regulatory, as an enhancer or modulator of gene expression. On the other hand, exon 6 was chosen as a target of CRISPR/Cas9 as this represents the first coding exon of the canonical and best described transcript of *MBD5*, MBD5-001 and by disrupting the first coding exon we expected to prevent *MBD5* transcription. Exon 6 encodes the MBD domain (shared with exon 7), which has been shown to play a role in regulation of transcription through chromatin remodeling in the other MBD family members (Roloff et al. 2003; Bogdanović & Veenstra 2009), although this has not been proven for MBD5 and MBD6 (Laget et al. 2010).

Initially, we looked at the *MBD5* mRNA expression levels found in the iPSC after deletion by CRISPR/Cas9 nucleofection that revealed a wide range of *MBD5* mRNA expression levels observed upon CRISPR-editing in iPSC. This may be explained due to endogenous cell repair mechanisms subsequent to the ablation or to unknown transcript regulatory mechanisms as *MBD5* presents several alternative transcripts (Supplementary Table 5). On the other hand, RNA levels may not reflect the decrease that occurs in protein levels. To test for this, three polyclonal antibodies against MBD5 were tested to determine protein levels,

however unsuccessfully, due to non-specificity of these antibodies (ab56126 and ab103144 Abcam, sc-107722 SantaCruz). Besides, the differentiation protocol was not driven towards a specific neuronal subtype and the mRNA levels that were assessed are a reflection of the different cell types that can arise from this process, as shown in Figure 26.

The iPSC were successfully driven towards the neuronal lineage, resulting in neuronal progenitor cells and mature neurons. In both differentiated cell lines, *MBD5* expression seemed to be close to normal or even increased in CRISPR-edited cell lines, when compared to controls. These results were consistent between the qRT-PCR and the RNAseq dataset of different biological replicates of the same cell lines, reassuring the validity of these data. This increase in expression post-perturbation in iPSC may reflect some form of compensation in the CRISPR-edited cell lines, as MBD5 seems to be crucial in levels of greater differentiation, showing increasing levels of expression from iPSC to mature neurons in control cells (Figure 38).

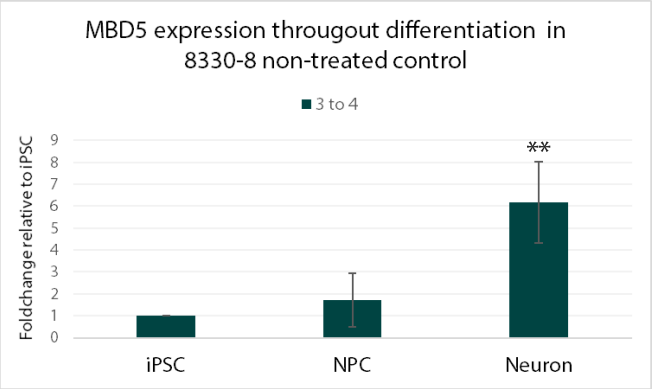


*Figure 38 - MBD5 mRNA expression in control 8330-8 cells, during different developmental stages (qRT-PCR, **p-value$_i$<0,01).*

The deletion of exon 6 and therefore, the MBD domain, generated in our CRISPR-edited cells was expected to prevent transcription of the canonical transcript MBD5-001. The RNAseq results allowed for a deeper appreciation of the intricate nature of *MBD5* transcript achitecture and identified alternative transcripts. As depicted in Figure 32, MBD5-001 was not the main responsible for MBD5 expression, as was initially expected. When looking at the transcripts expressed in the differentiated cell lines, we observe that exon 6 ablations affect the transcription of five different *MBD5* transcripts (Figure 31) and there are 4 transcripts that remain unaffected by this ablation, as they do not contain exon 6 and are initiated before or after this region. Two of the unaffected transcripts are predicted to be protein coding: MBD5-005 (from exons 8-15) and MBD5-014

(exons 12-15), while the other two are predicted to be a processed transcript and nonsense mediated decay, respectively: MBD5-010 (exons 1 and 2) and MBD5-015 (exons 9-15). Regarding this information, the ablation of the MBD domain would not affect the complete expression of *MBD5*, through the expression or compension of alternative transcripts that do not contain this domain. Surprisingly, the transcript that exhibited highest expression overall was MBD5-010 that only comprises 5‘ UTR exons 1 and 2. This is also verified in the GTEx database, as the second most expressed transcript in the maority of tissues and may suggest an important regulatory role for this transcript in MBD5 function. On the other hand, transcript MBD5-014 was observed uniquely in the NPC population and absent in the neurons, suggesting a developmental state preference for this protein-coding transcript and should be confirmed *in vivo* to determine its relevance in neurodevelopment.

Although MBD5-001 contains only 5‘ UTR exons 1 and 2, it may be relevant for disease. Non-coding RNAs are key regulators of gene expression, acting at the individual gene level, regulating cis and trans interactions and contributing to control of transcription and translation, and on a genome wide-scale, regulating accessibility of chromatin and controlling gene pathways (Ulitsky & Bartel 2013; Iyer et al. 2015; Barrett et al. 2013). The use of RNA as a regulatory element has advantages because it can rapidly be synthesized and degraded (Djupedal & Ekwall 2009), has structural plasticity and can modulate gene expression in response to external factors (Ansari 2009) and can act combinatorially to control complex interactions and regulatory pathways (Mattick 2004). Long non-coding RNAs (>200 nucleotides in length) are widely transcribed throughout the genome. An example of a long non-coding RNA is *MSNP1AS*, the expression of which was increased in the postmortem cerebral cortex of individuals with ASD (DeWitt et al. 2016). Elevated expression of *MSNP1AS* decreased neurite number and neurite length in both human neural progenitor cell lines (DeWitt et al. 2016). Previous reports of ASD patient deletions restricted to portions of the large noncoding region that contains multiple exons 5’ to the canonical translational start site in exon 6 (Figure 3, Talkowski et al. 2011; Bonnet et al. 2013; Mullegama & Elsea 2016) and the evidence that MBD5-010 was the highest expressing transcript, suggest that the 5‘UTR region is quite relevant for the neurobiological role of MBD5 in development and illustrates the importance of the non-coding network and the implications of dysregulation in disease.

Several research groups have demonstrated that each of the MBD genes is expressed in the brain, highlighting their relevance in this tissue, however their specific functions have only been determined for a subset of those genes (Bogdanović & Veenstra 2009; Laget et al. 2010; Shahbazian et al. 2002; Jiang et al. 2011). Altogether, our data suggest that the MBD domain of MBD5 may not be critical for the specific differentiation of neurons as the CRISPR-edited cell lines with ablations of this domain were able to successfully differentiate (Figure 25) while overall *MBD5* expression was not affected and even increased in some cases, through the expression of alternative non-coding transcripts. However, we do not have enough evidence to state that the absence of the MBD domain did not affect *MBD5* activity within the cell itself. The neuronal models of disease generated using CRISPR/Cas9 genome editing can be further exploited through a more comprehensive cell biological analysis of the generated neurons, such as the morphological analysis of dendrite length, number of spines and synapses, network connectivity, although this was not within the scope of this project. Considering an alignment of the MBD family MBD regions, the MBD5 MBD has two major differences: a deletion of nine amino acids in the first third of the MBD and an insertion of six amino acids in the last third of the domain (Laget et al. 2010). Further functional studies are be necessary to determine if MBD5 acts as its family members and to assess its MBD capacity of inducing transcription or binding DNA. There are 2 variants associated with ASD that have been reported in the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/). Although being located within the coding portion of exon 6, they are outside of the MBD domain and are annotated as likely benign and of uncertain significance. These data indicate that the canonical transcript MBD5-001 and the exon 6 portion of the MBD domain may not be crucial for the role of MBD5 during neurodevelopment, as previously thought.

Additional studies regarding the other domains in MBD5 such as the Proline-rich domain, asociated with protein-protein interactions (Williamson 1994), and the PWWP domain, would be crucial to understand the role of the other regions in its function. The PWWP domain is often found in DNA-binding proteins that function as transcription factors regulating developmental processes (Stec et al. 2000) and thus may be relevant for MBD5 function. In our study, we did not verify compensatory activity from the MBD family members upon the perturbations targetting *MBD5*, that could be due to the maintenance of *MBD5* normal expression through alternative transcripts.

Whole transcriptome analyses allowed us to determine the genes that were differentially expressed between the wild-type and the CRISPR-edited cell lines. These genes are affected upon perturbation of MBD5, indicating they are acting downstream of *MBD5* and may have a direct impact on neuronal growth and synapse function. Among those genes, *RAB11FIP1*, the most significant DEG in NPC and also upregulated in neurons, seems like a promising candidate as this protein has previously been associated with axonal growth in mice. Other genes to pinpoint from the DEG are those that have been previously found to be disrupted in ASD patients – *PLAUR* and *CNTNAP2*. The fact that genes are dysregulated upon MBD5 ablation suggests they might be interacting via similar pathways or have common players in different pathways.

GSEA of the DEG unveiled the gene families and pathways that were enriched within the CRISPR dataset. These results supported the hypothesis that an imbalance between excitation and inhibition is a biological mechanism for ASD; this imbalance may arise from different genes differentially affecting elements. Indeed, we found upregulated terms related to dopaminergic synapse transmission and downregulated for glutamatergic synapse, indicating a possible imbalance of neurotransmitter activity in these neuronal models. In addition to neuronal terms, we also found an enrichment for translation-related terms that suggest defects in translation of proteins required for normal synaptic function or neuronal growth. Supporting this, it has recently been reported that altering the neocortical excitation/inhibition balance leads to deficits in social behavior and information processing (Yizhar et al. 2011). KEGG pathway analysis identified an enrichment of pathways involved in notch signaling, known to be involved in brain development; and cell adhesion. Enrichment for cell adhesion terms and genes (*NCAM1*) has previously been observed by the knockdown of the ASD-risk gene *CHD8* (Sugathan et al. 2014) in neural progenitors, which is also a chromatin remodeler as *MBD5*. This process (cell adhesion) indicates a possible mechanism by which chromatin remodelers can be acting and impacting neuronal function.

In conclusion, we demonstrated that we could generate human neuronal cell lines to study ASD, from isogenic iPSC models edited using CRISPR/Cas9 technology to create loss-of-function mutations in a chromatin-related candidate ASD gene – *MBD5*. While the individual ablations of the 5' UTR exon 4 and the MBD domain in exon 6 were not sufficient to show a broad genome-wide impact on the PCA, the local investigation of *MBD5* transcript patterns during

neurodevelopment gave insights into the relevance of the different transcripts. We showed that MBD5-010 is a promising non-coding transcript that may be implicated in neuronal development and disease and should be further investigated to determine its potential as a regulatory lncRNA and to clarify the neurobiological role of the *MBD5* transcripts that do not include the MBD domain. On the other hand, genome-wide transcriptomic analysis allowed the identification of the dysregulated genes such as *RAB11FIP1*, *NHLH1-2*, *PLAUR* and *CNTNAP2*; and pathways such as notch signaling and cell adhesion that gave insight on the protein complexes and pathways that are acting downstream of *MBD5* and are directly implicated in neuronal development and function. Those represent promising targets for ASD therapeutics to be able to specifically aim for common and convergent biological pathways that are affected by *MBD5* and may be affected by other chromatin remodelers that represent a risk for ASD (such as *CHD8*).

Future directions of this study will include the transcriptomic analyses of the hypomorphic *Mbd5*$^{CT/+}$ mouse model (see Additional Preliminary Results in the Final Remarks Section) to identify common differentially expressed genes and pathways, that will complement the results obtained from the CRISPR-edited *MBD5* cell models. This combined analysis will allow the interpretation of the results in an *in vivo* context and confirm the role of *MBD5* haploinsufficiency in ASD etiology. These analyses are currently underway and will be submitted for publication, along with the results presented in this Chapter, in the form of an original scientific article.

Supplementary Table 4 - qPCR Primers used in iPSC, NPC and Neuronal cells.

Gene	Target	Sequence (5'-3')	iPSC	NPC	Neuron
ACTB	Exon 5	TGA AGT GTG ACG TGG ACA TC	✓	✓	✓
	Exon 6	GGA GGA GCA ATG ATC TTG AT	✓	✓	✓
GAPDH	Exon 8	GGA CCT GAC CTG CCG TCT AG	·	✓	✓
	Exons 8/9	GTA GCC CAG GAT GCC CTT GA	·	✓	✓
POLR2A	Exon 24	GCA CCA CGT CCA ATG ACA T	·	✓	✓
	Exon 26	GTG CGG CTG CTT CCA TAA	·	✓	✓
MBD5	Exon 3	CAG ATG GCA ACA GAG GATG T	✓	✓	✓
	Exon 4	GCA GTG TAA TGG AGG CAG TT	✓	✓	✓
	Exon 6	CCA GCT ATA CAA GTT CCT GTG G	·	✓	✓
	Exons 6/7	CCA CTG GGA CTG ACA TAA AGC A		✓	✓
	Exon 7	GTG GCT TGG AAT GTC CTC TT	✓	·	·
	Exon 8	TCT GCG GTT CTC TGT TTC AC	✓	·	·
	Exon 13	TTT GGA AGC CTA CAG CCG T	·	✓	✓
	Exons 14/15	TTG GTG TAC AGT CCC AGA CA	·	✓	✓

Supplementary Table 5 - MBD5 Transcripts from Ensembl, showing a total of 33 transcripts, of which 7 are predicted to be protein coding.

Ensembl Transcript ID	Alternative Name	bp	Protein	Biotype	UniProt ID
ENST00000407073.5	MBD5-001	9512	1494aa	Protein coding	Q9P267
ENST00000404807.5	MBD5-004	5920	1727aa	Protein coding	E9PHH0
ENST00000627651.2	MBD5-012	5610	851aa	Protein coding	Q9P267
ENST00000416015.2	MBD5-005	4343	1064aa	Protein coding	H7C066
ENST00000638043.1	MBD5-025	3483	696aa	Protein coding	A0A1B0GW10
ENST00000637159.1	MBD5-016	902	38aa	Protein coding	A0A1B0GUJ9
ENST00000630352.1	MBD5-014	663	77aa	Protein coding	A0A0D9SEP6
ENST00000629878.2	MBD5-013	4130	1086aa	Nonsense mediated decay	A0A0D9SG23
ENST00000628572.2	MBD5-015	3595	910aa	Nonsense mediated decay	A0A0D9SF16
ENST00000478190.3	MBD5-010	4131	No protein	Processed transcript	-
ENST00000638090.1	MBD5-024	2039	No protein	Processed transcript	-
ENST00000637242.1	MBD5-021	1869	No protein	Processed transcript	-
ENST00000637997.1	MBD5-019	1602	No protein	Processed transcript	-
ENST00000635796.1	MBD5-018	1371	No protein	Processed transcript	-
ENST00000637308.1	MBD5-020	1072	No protein	Processed transcript	-
ENST00000638130.1	MBD5-022	1034	No protein	Processed transcript	-
ENST00000637445.1	MBD5-023	958	No protein	Processed transcript	-
ENST00000478804.3	MBD5-002	592	No protein	Processed transcript	-
ENST00000488372.5	MBD5-007	548	No protein	Processed transcript	-
ENST00000496158.5	MBD5-006	544	No protein	Processed transcript	-
ENST00000473478.5	MBD5-008	484	No protein	Processed transcript	-
ENST00000636620.1	MBD5-017	475	No protein	Processed transcript	-
ENST00000470063.5	MBD5-011	324	No protein	Processed transcript	-
ENST00000469438.1	MBD5-009	252	No protein	Processed transcript	-
ENST00000636371.1	MBD5-029	172	No protein	Processed transcript	-
ENST00000637316.1	MBD5-026	100	No protein	Processed transcript	-
ENST00000637830.1	MBD5-027	100	No protein	Processed transcript	-
ENST00000636948.1	MBD5-028	100	No protein	Processed transcript	-
ENST00000637067.1	MBD5-030	100	No protein	Processed transcript	-
ENST00000637835.1	MBD5-031	100	No protein	Processed transcript	-
ENST00000637850.1	MBD5-032	100	No protein	Processed transcript	-
ENST00000637502.1	MBD5-035	100	No protein	Processed transcript	-
ENST00000496893.3	MBD5-003	2788	No protein	Retained intron	-

Supplementary Table 6 - DEG in NPC (P-adj < 0.01)

Gene ID	Name	Log2 Foldchange	P-value	P-adj
ENSG00000156675	RAB11FIP1	1.595	1.15273E-10	2.00413E-06
ENSG00000171786	NHLH1	1.515	5.00303E-10	4.34914E-06
ENSG00000075213	SEMA3A	-0.957	1.16851E-08	6.77191E-05
ENSG00000139370	SLC15A4	-1.148	6.1351E-08	0.000232575
ENSG00000211896	IGHG1	-1.250	6.68856E-08	0.000232575
ENSG00000073849	ST6GAL1	0.869	9.93829E-08	0.000287979
ENSG00000122861	PLAU	1.192	3.196E-07	0.000793796
ENSG00000147145	LPAR4	-0.835	4.44498E-07	0.000966006
ENSG00000136068	FLNB	0.854	9.11712E-07	0.001761225
ENSG00000211899	IGHM	-1.194	1.13228E-06	0.00196859
ENSG00000131914	LIN28A	1.165	1.66745E-06	0.002635483
ENSG00000117461	PIK3R3	-0.737	2.17471E-06	0.002908425
ENSG00000011422	PLAUR	0.957	2.01356E-06	0.002908425
ENSG00000104327	CALB1	1.164	2.5617E-06	0.003173504
ENSG00000150244	TRIM48	-1.038	2.73798E-06	0.003173504
ENSG00000250208	FZD10-AS1	-0.832986125	3.07188E-06	0.003337984
ENSG00000026508	CD44	0.944979857	3.70629E-06	0.003790443
ENSG00000182752	PAPPA	1.104075125	6.65703E-06	0.006429952
ENSG00000248485	PCP4L1	-1.120453523	8.63519E-06	0.006718973
ENSG00000135821	GLUL	-0.844857536	8.09164E-06	0.006718973
ENSG00000145247	OCIAD2	1.108322335	9.0165E-06	0.006718973
ENSG00000113319	RASGRF2	1.11578902	7.83431E-06	0.006718973
ENSG00000106483	SFRP4	-1.067984145	9.28329E-06	0.006718973
ENSG00000139132	FGD4	-0.819710097	9.31688E-06	0.006718973
ENSG00000198796	ALPK2	1.077566296	9.66147E-06	0.006718973
ENSG00000122877	EGR2	-1.087153682	1.04606E-05	0.006994909
ENSG00000136167	LCP1	-1.015496248	1.29416E-05	0.008333453
ENSG00000049130	KITLG	-0.60125141	1.64649E-05	0.010223552
ENSG00000188641	DPYD	0.999136418	2.19517E-05	0.013160443
ENSG00000123307	NEUROD4	0.946225265	2.42897E-05	0.014076685
ENSG00000197747	S100A10	1.047489955	2.57644E-05	0.014449649
ENSG00000047346	FAM214A	-0.691142245	2.67254E-05	0.014520249
ENSG00000130592	LSP1	-0.915785812	3.57575E-05	0.018838786
ENSG00000251129	RP11-734I18.1	1.008875495	3.74154E-05	0.01913246
ENSG00000134982	APC	-0.832308728	4.25323E-05	0.020818274
ENSG00000198739	LRRTM3	-0.904918565	4.3107E-05	0.020818274
ENSG00000183091	NEB	-0.934890052	4.66182E-05	0.021905493
ENSG00000198028	ZNF560	-0.859514674	4.8925E-05	0.022384472
ENSG00000147571	CRH	0.949918518	5.10329E-05	0.022750211
ENSG00000177707	PVRL3	0.665060221	5.6269E-05	0.024457329
ENSG00000106049	HIBADH	0.628338353	6.14492E-05	0.025262695
ENSG00000147894	C9orf72	-0.707525985	6.07812E-05	0.025262695
ENSG00000107242	PIP5K1B	-0.98917524	6.39341E-05	0.025262695
ENSG00000104783	KCNN4	0.983081186	6.27864E-05	0.025262695
ENSG00000196562	SULF2	-0.943987282	6.76207E-05	0.02612564
ENSG00000185774	KCNIP4	0.99159335	6.93422E-05	0.02620832
ENSG00000142192	APP	0.441695337	8.35802E-05	0.030917572
ENSG00000160111	CPAMD8	-0.871854745	8.70648E-05	0.031535598
ENSG00000260343	LINC01043	-0.699272171	8.89064E-05	0.031545439
ENSG00000105996	HOXA2	0.902991981	9.38072E-05	0.03261864
ENSG00000138650	PCDH10	-0.959888842	9.59263E-05	0.032701447
ENSG00000126803	HSPA2	-0.917529279	0.000141607	0.047345638

ENSG00000173208	ABCD2	-0.885879589	0.000144432	0.047379118

Supplementary Table 7 - DEG in Neurons (p-Value <0.01)

Gene ID	*Name*	*Log2 Foldchange*	*P-value*	*P-adj*
ENSG00000164093	PITX2	-1.189	3.90387E-07	0.003742448
ENSG00000197496	SLC2A10	-0.840	3.56613E-07	0.003742448
ENSG00000224597	PTCHD3P1	-1.039	1.241E-06	0.007931221
ENSG00000100884	CPNE6	-1.083	3.26724E-06	0.0156607
ENSG00000178401	DNAJC22	-1.069	1.03968E-05	0.039867745
ENSG00000156675	RAB11FIP1	1.076	1.4576E-05	0.046577741
ENSG00000131094	C1QL1	-0.894	2.216E-05	0.060696357
ENSG00000177551	NHLH2	0.991	3.5457E-05	0.084199568
ENSG00000173376	NDNF	1.006	4.48261E-05	0.084199568
ENSG00000211445	GPX3	-0.819	6.43058E-05	0.084199568
ENSG00000174469	CNTNAP2	0.841	6.55521E-05	0.084199568
ENSG00000147246	HTR2C	1.021	4.60333E-05	0.084199568
ENSG00000223508	RPL23AP53	-0.832	6.22362E-05	0.084199568
ENSG00000109991	P2RX3	0.966	5.76079E-05	0.084199568
ENSG00000101098	RIMS4	0.777	6.58735E-05	0.084199568
ENSG00000066248	NGEF	0.983570256	8.21308E-05	0.098418389
ENSG00000078328	RBFOX1	0.877295573	8.80731E-05	0.099330885
ENSG00000184221	OLIG1	-0.968603937	0.000111278	0.118529824
ENSG00000116741	RGS2	0.676090164	0.000195811	0.187713957
ENSG00000170485	NPAS2	0.922083949	0.000195014	0.187713957
ENSG00000250049	RP11-348J24.2	-0.692328363	0.000208869	0.190697725
ENSG00000251129	RP11-734I18.1	0.826114073	0.000233119	0.203163395
ENSG00000180875	GREM2	0.903533937	0.000250536	0.208849065
ENSG00000133110	POSTN	0.8902025	0.000273909	0.218819181
ENSG00000171951	SCG2	0.725038582	0.000375723	0.27341359
ENSG00000048540	LMO3	0.559768429	0.000383548	0.27341359
ENSG00000101400	SNTA1	-0.52936662	0.000385029	0.27341359
ENSG00000180616	SSTR2	0.882635754	0.000404466	0.27695819
ENSG00000173210	ABLIM3	0.810706528	0.000450524	0.29785867
ENSG00000189229	AC069277.2	-0.635005317	0.000514746	0.318361922
ENSG00000074211	PPP2R2C	0.85553226	0.000505476	0.318361922
ENSG00000196533	C1orf186	0.69976501	0.000594233	0.340042766
ENSG00000145451	GLRA3	0.847308627	0.000577916	0.340042766
ENSG00000185666	SYN3	0.84478075	0.000603007	0.340042766
ENSG00000215912	TTC34	-0.603470406	0.000681677	0.373422849
ENSG00000134709	HOOK1	0.802164676	0.000824996	0.416903595
ENSG00000254369	HOXA-AS3	0.733814673	0.000828365	0.416903595
ENSG00000075223	SEMA3C	0.805703341	0.000913261	0.416903595
ENSG00000231725	VN1R110P	0.792853406	0.000897912	0.416903595
ENSG00000198739	LRRTM3	-0.610107081	0.000875612	0.416903595
ENSG00000167281	RBFOX3	0.83381186	0.000837112	0.416903595
ENSG00000176533	GNG7	-0.531997102	0.000875025	0.416903595
ENSG00000116014	KISS1R	-0.715924978	0.001034364	0.450723914
ENSG00000126243	LRFN3	-0.553765899	0.001023089	0.450723914
ENSG00000218336	TENM3	0.687982637	0.001106723	0.451472283
ENSG00000011677	GABRA3	0.801969855	0.001065018	0.451472283
ENSG00000171222	SCAND1	-0.61043684	0.00110462	0.451472283
ENSG00000243449	C4orf48	-0.65508796	0.001174989	0.469334581
ENSG00000152969	JAKMIP1	0.799577553	0.001200531	0.469750634

ENSG00000112541	PDE10A	0.432171618	0.001265723	0.469935328
ENSG00000157570	TSPAN18	0.797023991	0.001273121	0.469935328
ENSG00000182809	CRIP2	-0.40323275	0.001274534	0.469935328
ENSG00000144227	NXPH2	0.79038291	0.001391163	0.494320735
ENSG00000180938	ZNF572	-0.597864459	0.001392235	0.494320735
ENSG00000261786	RP4-555D20.2	0.777446109	0.001461467	0.500369812
ENSG00000152578	GRIA4	-0.622282572	0.001444507	0.500369812
ENSG00000186446	ZNF501	0.778655936	0.00157311	0.50268735
ENSG00000181195	PENK	0.676091776	0.001497472	0.50268735
ENSG00000130720	FIBCD1	0.781855433	0.001554856	0.50268735
ENSG00000260248	RP11-143K11.1	0.791487987	0.001553447	0.50268735
ENSG00000177468	OLIG3	0.648570252	0.001725811	0.525221813
ENSG00000184344	GDF3	0.684874393	0.001723348	0.525221813
ENSG00000166426	CRABP1	0.674274423	0.001672879	0.525221813
ENSG00000249803	RP11-114H21.2	-0.61287651	0.001809645	0.538157125
ENSG00000124212	PTGIS	0.749932495	0.001824452	0.538157125
ENSG00000204175	GPRIN2	0.750802605	0.001914801	0.54794752
ENSG00000172725	CORO1B	-0.425542184	0.001911085	0.54794752
ENSG00000119888	EPCAM	0.763151376	0.002174456	0.595583603
ENSG00000180900	SCRIB	-0.451899451	0.002163002	0.595583603
ENSG00000250654	RP11-834C11.7	0.62497773	0.002162141	0.595583603
ENSG00000115380	EFEMP1	0.763540975	0.002222993	0.600302099
ENSG00000198216	CACNA1E	0.756997478	0.00233739	0.613901061
ENSG00000144583	MARCH4	0.628744193	0.002330902	0.613901061
ENSG00000197106	SLC6A17	0.73907441	0.002423615	0.615440368
ENSG00000147256	ARHGAP36	0.639727359	0.002402537	0.615440368
ENSG00000099256	PRTFDC1	-0.67617607	0.002439549	0.615440368
ENSG00000218739	CEBPZ-AS1	-0.679913973	0.002587856	0.64437622
ENSG00000145391	SETD7	0.548390758	0.002791037	0.660997407
ENSG00000124785	NRN1	0.741459549	0.002743742	0.660997407
ENSG00000118407	FILIP1	0.740030509	0.002741034	0.660997407
ENSG00000100767	PAPLN	-0.723649932	0.00279251	0.660997407
ENSG00000113248	PCDHB15	0.684138896	0.00285125	0.666670997
ENSG00000159208	C1orf51	0.638754421	0.003184106	0.701216581
ENSG00000136205	TNS3	-0.614497055	0.003073771	0.701216581
ENSG00000231764	DLX6-AS1	0.722584229	0.003248191	0.701216581
ENSG00000173275	ZNF449	0.509328812	0.003255008	0.701216581
ENSG00000168447	SCNN1B	-0.609018052	0.003216454	0.701216581
ENSG00000132386	SERPINF1	0.725561135	0.003044602	0.701216581
ENSG00000157557	ETS2	0.709191374	0.003174585	0.701216581
ENSG00000249992	TMEM158	0.551739561	0.003381986	0.714189946
ENSG00000211829	TRDC	0.626572013	0.00338973	0.714189946
ENSG00000175221	MED16	-0.391612211	0.003453585	0.719734717
ENSG00000144821	MYH15	-0.721122425	0.00359239	0.733763019
ENSG00000140945	CDH13	0.712319698	0.00359744	0.733763019
ENSG00000138798	EGF	0.664431138	0.003848212	0.776650188
ENSG00000152137	HSPB8	-0.724659422	0.003941408	0.787173107
ENSG00000134121	CHL1	0.621268916	0.004179473	0.814437073
ENSG00000250682	LINC00491	0.624129062	0.004322807	0.814437073
ENSG00000172748	ZNF596	-0.603318464	0.004371646	0.814437073
ENSG00000227896	RP11-77P6.2	0.707585769	0.004213454	0.814437073
ENSG00000165868	HSPA12A	0.70818392	0.004375268	0.814437073
ENSG00000198732	SMOC1	-0.681377808	0.004178712	0.814437073
ENSG00000205922	ONECUT3	-0.657801707	0.004357137	0.814437073

ENSG00000075426	FOSL2	0.70382411	0.004744819	0.820643411
ENSG00000145247	OCIAD2	0.690787898	0.004647874	0.820643411
ENSG00000183166	CALN1	0.69141052	0.004696038	0.820643411
ENSG00000178209	PLEC	-0.561189879	0.004523053	0.820643411
ENSG00000180806	HOXC9	0.573749	0.004751026	0.820643411
ENSG00000115266	APC2	-0.581569257	0.004733156	0.820643411
ENSG00000267481	CTC-559E9.5	0.687619577	0.00459481	0.820643411
ENSG00000268041	CTD-2575K13.6	-0.664694636	0.004535632	0.820643411
ENSG00000124074	ENKD1	-0.523406005	0.004850165	0.822939947
ENSG00000269067	ZNF728	-0.46749907	0.004841869	0.822939947
ENSG00000178038	ALS2CL	0.689564258	0.004893218	0.82296207
ENSG00000139517	LNX2	0.480949118	0.005010339	0.835332448
ENSG00000168490	PHYHIP	0.692869412	0.005102174	0.843310215
ENSG00000077616	NAALAD2	0.689526935	0.005253289	0.86086588
ENSG00000229847	EMX2OS	0.69748128	0.00530059	0.861256103
ENSG00000044524	EPHA3	-0.573058962	0.005377662	0.863636653
ENSG00000166900	STX3	0.691690421	0.005450375	0.863636653
ENSG00000226741	CTA-929C8.6	-0.605948668	0.005428733	0.863636653
ENSG00000142327	RNPEPL1	-0.450374374	0.00565056	0.885296152
ENSG00000103260	METRN	-0.513422407	0.005679415	0.885296152
ENSG00000272455	RP4-758J18.13	0.660449375	0.005863358	0.896839765
ENSG00000185818	NAT8L	-0.440189394	0.005987351	0.896839765
ENSG00000145358	DDIT4L	0.679361736	0.005894421	0.896839765
ENSG00000110841	PPFIBP1	0.635468059	0.00596454	0.896839765
ENSG00000230148	HOXB-AS1	0.561979211	0.005975609	0.896839765
ENSG00000114698	PLSCR4	-0.585302604	0.006037879	0.897397335
ENSG00000115194	SLC30A3	0.581688147	0.006334102	0.934182566
ENSG00000110090	CPT1A	0.675857903	0.006664283	0.94370853
ENSG00000135423	GLS2	0.660816757	0.00646727	0.94370853
ENSG00000066629	EML1	0.445362974	0.006652648	0.94370853
ENSG00000259905	PWRN1	-0.668029421	0.006694016	0.94370853
ENSG00000007516	BAIAP3	-0.623642975	0.006503279	0.94370853
ENSG00000141837	CACNA1A	0.667565584	0.006693819	0.94370853
ENSG00000143190	POU2F1	0.343952317	0.007173111	0.955067524
ENSG00000145390	USP53	0.575812776	0.006964324	0.955067524
ENSG00000147869	CER1	-0.616404745	0.007028336	0.955067524
ENSG00000157637	SLC38A10	-0.526820887	0.007222907	0.955067524
ENSG00000166603	MC4R	0.595445691	0.007086435	0.955067524
ENSG00000142303	ADAMTS10	-0.411235242	0.007172589	0.955067524
ENSG00000123159	GIPC1	-0.361543105	0.007192519	0.955067524
ENSG00000160321	ZNF208	-0.400387055	0.007218015	0.955067524
ENSG00000099889	ARVCF	-0.398059803	0.006938433	0.955067524
ENSG00000107731	UNC5B	0.665711763	0.007479013	0.982158313
ENSG00000168993	CPLX1	-0.665126755	0.007628681	0.987026666
ENSG00000197757	HOXC6	0.522803648	0.007722005	0.987026666
ENSG00000159648	TEPP	0.556755164	0.007607685	0.987026666
ENSG00000156413	FUT6	0.52745303	0.007704833	0.987026666
ENSG00000117115	PADI2	0.614925942	0.008327638	0.99041131
ENSG00000188641	DPYD	0.621622411	0.008334159	0.99041131
ENSG00000152061	RABGAP1L	0.372244531	0.00802787	0.99041131
ENSG00000128652	HOXD3	0.637652021	0.007922412	0.99041131
ENSG00000225138	CTD-2228K2.7	-0.428466465	0.008038132	0.99041131
ENSG00000135318	NT5E	-0.602545447	0.008241872	0.99041131
ENSG00000074706	IPCEF1	0.653303468	0.008298918	0.99041131

ENSG00000104725	NEFL	0.655675848	0.008368363	0.99041131
ENSG00000155158	TTC39B	0.576952127	0.008093035	0.99041131
ENSG00000136160	EDNRB	0.391540031	0.008316234	0.99041131
ENSG00000127585	FBXL16	-0.486003399	0.007844801	0.99041131
ENSG00000081665	ZNF506	0.640262244	0.007905841	0.99041131
ENSG00000175445	LPL	0.655136253	0.008645346	0.992558238
ENSG00000269962	RP13-238F13.5	-0.547249228	0.00857762	0.992558238
ENSG00000010278	CD9	0.652628826	0.008542816	0.992558238
ENSG00000137872	SEMA6D	-0.466803556	0.008532825	0.992558238
ENSG00000006015	C19orf60	-0.484351891	0.008616771	0.992558238

Supplementary Table 8 - GO Terms for NPC and Neurons (top 100 terms for each category).

Ontology	*P-value*	*Q-value*
NPC Upregulated Terms		
GO:0003735 structural constituent of ribosome	5.02E-11	3.50E-07
GO:0006415 translational termination	4.59E-10	1.60E-06
GO:0006414 translational elongation	2.22E-09	5.17E-06
GO:0006413 translational initiation	3.94E-09	6.87E-06
GO:0044391 ribosomal subunit	8.46E-09	1.18E-05
GO:0005840 ribosome	4.56E-08	4.80E-05
GO:0005743 mitochondrial inner membrane	4.81E-08	4.80E-05
GO:0019866 organelle inner membrane	9.92E-08	8.65E-05
GO:0006614 SRP-dependent cotranslational protein targeting to membrane	1.25E-07	9.38E-05
GO:0022626 cytosolic ribosome	1.34E-07	9.38E-05
GO:0006613 cotranslational protein targeting to membrane	3.18E-07	0.000201582
GO:0005759 mitochondrial matrix	7.15E-07	0.000415683
GO:0015934 large ribosomal subunit	1.19E-06	0.00063858
GO:0045047 protein targeting to ER	1.90E-06	0.000947616
GO:0022625 cytosolic large ribosomal subunit	5.06E-06	0.002351211
GO:0034660 ncRNA metabolic process	7.35E-06	0.003206579
GO:0005925 focal adhesion	8.61E-06	0.003534648
GO:0030055 cell-substrate junction	9.50E-06	0.003549294
GO:0072599 establishment of protein localization to endoplasmic reticulum	9.67E-06	0.003549294
GO:0005924 cell-substrate adherens junction	1.10E-05	0.00383592
GO:0006520 cellular amino acid metabolic process	2.03E-05	0.006736453
GO:0070972 protein localization to endoplasmic reticulum	2.37E-05	0.007387419
GO:0043624 cellular protein complex disassembly	2.51E-05	0.007387419
GO:0000184 nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	2.54E-05	0.007387419
GO:0032984 macromolecular complex disassembly	3.35E-05	0.009343132
GO:0006399 tRNA metabolic process	4.18E-05	0.011209947
GO:0098800 inner mitochondrial membrane protein complex	4.64E-05	0.011582824
GO:0032543 mitochondrial translation	4.65E-05	0.011582824
GO:0019080 viral gene expression	5.60E-05	0.013036855
GO:0043241 protein complex disassembly	5.61E-05	0.013036855
GO:0045333 cellular respiration	6.46E-05	0.014527403
GO:0042254 ribosome biogenesis	6.72E-05	0.014646788
GO:0044455 mitochondrial membrane part	7.78E-05	0.016452523
GO:0044445 cytosolic part	8.25E-05	0.016929174
GO:0044033 multi-organism metabolic process	9.98E-05	0.019899839
GO:0098798 mitochondrial protein complex	0.000108493	0.020633093
GO:0034470 ncRNA processing	0.000109436	0.020633093
GO:0070125 mitochondrial translational elongation	0.000120001	0.021939713
GO:0019083 viral transcription	0.000122656	0.021939713
GO:0070161 anchoring junction	0.000138399	0.024136742
GO:0022900 electron transport chain	0.000142971	0.024326013

Ontology	*P-value*	*Q-value*
GO:0070126 mitochondrial translational termination	0.000179078	0.029744018
GO:0005912 adherens junction	0.000221176	0.035881871
GO:0022904 respiratory electron transport chain	0.000229135	0.036328241
GO:0070124 mitochondrial translational initiation	0.000247901	0.038430168
GO:0006839 mitochondrial transport	0.000293337	0.044485159
GO:0022613 ribonucleoprotein complex biogenesis	0.000421383	0.062544031
GO:0006418 tRNA aminoacylation for protein translation	0.000450852	0.065523809
GO:0043038 amino acid activation	0.000641703	0.089530417
GO:0043039 tRNA aminoacylation	0.000641703	0.089530417
GO:0008652 cellular amino acid biosynthetic process	0.00066475	0.090166608
GO:0042470 melanosome	0.000685039	0.090166608
GO:0048770 pigment granule	0.000685039	0.090166608
GO:0070469 respiratory chain	0.000703469	0.090877772
GO:0051186 cofactor metabolic process	0.000768522	0.097476493
GO:0015935 small ribosomal subunit	0.00081736	0.101819666
GO:0006612 protein targeting to membrane	0.000938556	0.11426018
GO:0006364 rRNA processing	0.000952993	0.11426018
GO:1901607 alpha-amino acid biosynthetic process	0.000966363	0.11426018
GO:0090150 establishment of protein localization to membrane	0.001078963	0.125447395
GO:0005746 mitochondrial respiratory chain	0.001167387	0.133503133
GO:0016072 rRNA metabolic process	0.001259747	0.141741862
GO:0009141 nucleoside triphosphate metabolic process	0.001466731	0.161605232
GO:0000313 organellar ribosome	0.001507892	0.161605232
GO:0005761 mitochondrial ribosome	0.001507892	0.161605232
GO:1902742 apoptotic process involved in development	0.001545928	0.161605232
GO:0070585 protein localization to mitochondrion	0.001552114	0.161605232
GO:0072655 establishment of protein localization to mitochondrion	0.001712655	0.172401382
GO:0004812 aminoacyl-tRNA ligase activity	0.001754659	0.172401382
GO:0016875 ligase activity, forming carbon-oxygen bonds	0.001754659	0.172401382
GO:0016876 ligase activity, forming aminoacyl-tRNA and related compounds	0.001754659	0.172401382
GO:0042273 ribosomal large subunit biogenesis	0.002013632	0.195098603
GO:0009116 nucleoside metabolic process	0.002169965	0.20736539
GO:0098803 respiratory chain complex	0.002503647	0.235646246
GO:0022627 cytosolic small ribosomal subunit	0.002533467	0.235646246
GO:0019058 viral life cycle	0.002607138	0.239307787
GO:0009112 nucleobase metabolic process	0.002721463	0.246557446
GO:0009144 purine nucleoside triphosphate metabolic process	0.002810539	0.251363063
GO:0008033 tRNA processing	0.002858996	0.25209835
GO:0009156 ribonucleoside monophosphate biosynthetic process	0.002891036	0.25209835
GO:0009124 nucleoside monophosphate biosynthetic process	0.002932264	0.252536704
GO:0051084 'de novo' posttranslational protein folding	0.003107197	0.262156924
GO:0009123 nucleoside monophosphate metabolic process	0.003119126	0.262156924
GO:0009161 ribonucleoside monophosphate metabolic process	0.00322715	0.268007114
GO:1901657 glycosyl compound metabolic process	0.003381474	0.277519595
GO:0006144 purine nucleobase metabolic process	0.003517676	0.2853408

Ontology	*P-value*	*Q-value*
GO:0033173 calcineurin-NFAT signaling cascade	0.003639622	0.291839103
GO:0009199 ribonucleoside triphosphate metabolic process	0.003833274	0.303874048
GO:0006626 protein targeting to mitochondrion	0.003994046	0.311152937
GO:0072657 protein localization to membrane	0.004062077	0.311152937
GO:0016655 oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor	0.004089182	0.311152937
GO:0010821 regulation of mitochondrion organization	0.004103508	0.311152937
GO:0006575 cellular modified amino acid metabolic process	0.004228565	0.314843885
GO:0034976 response to endoplasmic reticulum stress	0.004242449	0.314843885
GO:0016741 transferase activity, transferring one-carbon groups	0.004298807	0.31566821
GO:0006091 generation of precursor metabolites and energy	0.00435595	0.316532361
GO:0044452 nucleolar part	0.004480479	0.322224935
GO:0015036 disulfide oxidoreductase activity	0.004684943	0.33349143
GO:0009205 purine ribonucleoside triphosphate metabolic process	0.00486034	0.341881088
GO:0046112 nucleobase biosynthetic process	0.004900818	0.341881088
NPC Downregulated Terms		
GO:0042129 regulation of T cell proliferation	0.001497439	1
GO:0070663 regulation of leukocyte proliferation	0.002037693	1
GO:0010518 positive regulation of phospholipase activity	0.002282749	1
GO:0042102 positive regulation of T cell proliferation	0.003086179	1
GO:0032944 regulation of mononuclear cell proliferation	0.003109151	1
GO:0050670 regulation of lymphocyte proliferation	0.003271006	1
GO:0060193 positive regulation of lipase activity	0.003345664	1
GO:0050865 regulation of cell activation	0.003402878	1
GO:1900274 regulation of phospholipase C activity	0.003588929	1
GO:0010863 positive regulation of phospholipase C activity	0.003594592	1
GO:0098644 complex of collagen trimers	0.004002092	1
GO:0050870 positive regulation of T cell activation	0.004269975	1
GO:0002694 regulation of leukocyte activation	0.004420835	1
GO:0050853 B cell receptor signaling pathway	0.004548778	1
GO:0070665 positive regulation of leukocyte proliferation	0.005422314	1
GO:0051249 regulation of lymphocyte activation	0.00576641	1
GO:0050863 regulation of T cell activation	0.006192526	1
GO:0051251 positive regulation of lymphocyte activation	0.006545198	1
GO:0019783 ubiquitin-like protein-specific protease activity	0.007582867	1
GO:0032946 positive regulation of mononuclear cell proliferation	0.007846317	1
GO:0050671 positive regulation of lymphocyte proliferation	0.008307522	1
GO:0042098 T cell proliferation	0.008313131	1
GO:0007156 homophilic cell adhesion via plasma membrane adhesion molecules	0.008585307	1
GO:1903039 positive regulation of leukocyte cell-cell adhesion	0.010376615	1
GO:0007202 activation of phospholipase C activity	0.01172011	1
GO:0022843 voltage-gated cation channel activity	0.01209884	1
GO:0010517 regulation of phospholipase activity	0.012110645	1
GO:0016339 calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules	0.012663064	1
GO:0016605 PML body	0.013393155	1
GO:0045582 positive regulation of T cell differentiation	0.014164163	1

Ontology	*P-value*	*Q-value*
GO:0070661 leukocyte proliferation	0.014407385	1
GO:0050867 positive regulation of cell activation	0.014693324	1
GO:0036459 ubiquitinyl hydrolase activity	0.014904336	1
GO:0046638 positive regulation of alpha-beta T cell differentiation	0.015067895	1
GO:0002696 positive regulation of leukocyte activation	0.015116582	1
GO:0042611 MHC protein complex	0.0151697	1
GO:0004843 ubiquitin-specific protease activity	0.015286868	1
GO:0070646 protein modification by small protein removal	0.015322416	1
GO:0008066 glutamate receptor activity	0.015438112	1
GO:0032943 mononuclear cell proliferation	0.015618872	1
GO:0060191 regulation of lipase activity	0.015649386	1
GO:0048407 platelet-derived growth factor binding	0.015975824	1
GO:0046651 lymphocyte proliferation	0.016233284	1
GO:0043550 regulation of lipid kinase activity	0.017085852	1
GO:0034112 positive regulation of homotypic cell-cell adhesion	0.017332032	1
GO:1903037 regulation of leukocyte cell-cell adhesion	0.017513779	1
GO:0036314 response to sterol	0.01796228	1
GO:0004702 receptor signaling protein serine/threonine kinase activity	0.018232999	1
GO:0016579 protein deubiquitination	0.018408833	1
GO:0009214 cyclic nucleotide catabolic process	0.019796997	1
GO:1902107 positive regulation of leukocyte differentiation	0.020571606	1
GO:0010225 response to UV-C	0.020784272	1
GO:1902105 regulation of leukocyte differentiation	0.02176659	1
GO:0004112 cyclic-nucleotide phosphodiesterase activity	0.022104412	1
GO:0000910 cytokinesis	0.022232068	1
GO:0032809 neuronal cell body membrane	0.022497945	1
GO:0044298 cell body membrane	0.022497945	1
GO:0098742 cell-cell adhesion via plasma-membrane adhesion molecules	0.023498445	1
GO:0042393 histone binding	0.023647987	1
GO:0043372 positive regulation of CD4-positive, alpha-beta T cell differentiation	0.023954234	1
GO:0004114 3',5'-cyclic-nucleotide phosphodiesterase activity	0.024117709	1
GO:0034110 regulation of homotypic cell-cell adhesion	0.024909213	1
GO:0070723 response to cholesterol	0.026687237	1
GO:0035235 ionotropic glutamate receptor signaling pathway	0.026744711	1
GO:0045670 regulation of osteoclast differentiation	0.026914498	1
GO:0005249 voltage-gated potassium channel activity	0.027453598	1
GO:0051302 regulation of cell division	0.02773917	1
GO:0004385 guanylate kinase activity	0.027816465	1
GO:0005216 ion channel activity	0.027860068	1
GO:0007064 mitotic sister chromatid cohesion	0.028574679	1
GO:0005085 guanyl-nucleotide exchange factor activity	0.029213352	1
GO:1990266 neutrophil migration	0.029275894	1
GO:0019838 growth factor binding	0.029739481	1
GO:0035091 phosphatidylinositol binding	0.03074735	1
GO:0030551 cyclic nucleotide binding	0.031645138	1

Ontology	*P-value*	*Q-value*
GO:0000819 sister chromatid segregation	0.031908202	1
GO:0022838 substrate-specific channel activity	0.032575287	1
GO:0043217 myelin maintenance	0.032690795	1
GO:0043551 regulation of phosphatidylinositol 3-kinase activity	0.032760548	1
GO:0035064 methylated histone binding	0.033774253	1
GO:0004970 ionotropic glutamate receptor activity	0.033937331	1
GO:0006198 cAMP catabolic process	0.034194816	1
GO:0004709 MAP kinase kinase kinase activity	0.03580754	1
GO:0046710 GDP metabolic process	0.035903292	1
GO:0015267 channel activity	0.03615124	1
GO:0022803 passive transmembrane transporter activity	0.03615124	1
GO:0016604 nuclear body	0.036177898	1
GO:0042110 T cell activation	0.036343949	1
GO:0070489 T cell aggregation	0.036343949	1
GO:0010738 regulation of protein kinase A signaling	0.036999136	1
GO:0010737 protein kinase A signaling	0.037732322	1
GO:0071599 otic vesicle development	0.037801391	1
GO:0030593 neutrophil chemotaxis	0.038055149	1
GO:0044381 glucose import in response to insulin stimulus	0.038441978	1
GO:2001273 regulation of glucose import in response to insulin stimulus	0.038441978	1
GO:0007062 sister chromatid cohesion	0.038501184	1
GO:0071600 otic vesicle morphogenesis	0.038581266	1
GO:0006670 sphingosine metabolic process	0.039047895	1
GO:0042613 MHC class II protein complex	0.039056197	1
GO:0050995 negative regulation of lipid catabolic process	0.03925002	1
Neuron Upregulated Terms		
GO:0021537 telencephalon development	0.000257256	0.484733828
GO:0044708 single-organism behavior	0.000299334	0.484733828
GO:0030902 hindbrain development	0.000352632	0.484733828
GO:0034109 homotypic cell-cell adhesion	0.000772096	0.484733828
GO:0000793 condensed chromosome	0.000801169	0.484733828
GO:0002250 adaptive immune response	0.000828314	0.484733828
GO:0044057 regulation of system process	0.000876665	0.484733828
GO:0007059 chromosome segregation	0.000899364	0.484733828
GO:0000779 condensed chromosome, centromeric region	0.001035558	0.484733828
GO:0007159 leukocyte cell-cell adhesion	0.001120123	0.484733828
GO:0098632 protein binding involved in cell-cell adhesion	0.001153061	0.484733828
GO:0042110 T cell activation	0.001235867	0.484733828
GO:0070489 T cell aggregation	0.001235867	0.484733828
GO:0022613 ribonucleoprotein complex biogenesis	0.001326102	0.484733828
GO:0051302 regulation of cell division	0.001422182	0.484733828
GO:0071593 lymphocyte aggregation	0.001481958	0.484733828
GO:0030003 cellular cation homeostasis	0.001581459	0.484733828
GO:0072503 cellular divalent inorganic cation homeostasis	0.001597257	0.484733828
GO:0007626 locomotory behavior	0.001655805	0.484733828

Ontology	*P-value*	*Q-value*
GO:0000777 condensed chromosome kinetochore	0.001688641	0.484733828
GO:0070486 leukocyte aggregation	0.00175109	0.484733828
GO:0072507 divalent inorganic cation homeostasis	0.001759642	0.484733828
GO:0010817 regulation of hormone levels	0.001796708	0.484733828
GO:0000775 chromosome, centromeric region	0.001850978	0.484733828
GO:0023061 signal release	0.00190392	0.484733828
GO:0006873 cellular ion homeostasis	0.001915827	0.484733828
GO:0098813 nuclear chromosome segregation	0.001918494	0.484733828
GO:0045785 positive regulation of cell adhesion	0.001967784	0.484733828
GO:0032844 regulation of homeostatic process	0.00200675	0.484733828
GO:0045787 positive regulation of cell cycle	0.002151859	0.486771055
GO:0006875 cellular metal ion homeostasis	0.002154162	0.486771055
GO:0021543 pallium development	0.002233501	0.488927321
GO:0030098 lymphocyte differentiation	0.00258379	0.548468191
GO:0032609 interferon-gamma production	0.003087322	0.627084092
GO:0002460 adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	0.003272777	0.627084092
GO:0006874 cellular calcium ion homeostasis	0.003308973	0.627084092
GO:0007218 neuropeptide signaling pathway	0.003747538	0.627084092
GO:0000776 kinetochore	0.003753359	0.627084092
GO:0019221 cytokine-mediated signaling pathway	0.003798056	0.627084092
GO:0055074 calcium ion homeostasis	0.003819457	0.627084092
GO:0051781 positive regulation of cell division	0.003819937	0.627084092
GO:0042254 ribosome biogenesis	0.003847852	0.627084092
GO:0021892 cerebral cortex GABAergic interneuron differentiation	0.003995847	0.627084092
GO:0051233 spindle midzone	0.004132507	0.627084092
GO:0090068 positive regulation of cell cycle process	0.004255859	0.627084092
GO:0070838 divalent metal ion transport	0.004262185	0.627084092
GO:0009612 response to mechanical stimulus	0.004350164	0.627084092
GO:0035637 multicellular organismal signaling	0.004703176	0.627084092
GO:0031123 RNA 3'-end processing	0.004767731	0.627084092
GO:0000070 mitotic sister chromatid segregation	0.004811391	0.627084092
GO:0097154 GABAergic neuron differentiation	0.005194752	0.627084092
GO:0072511 divalent inorganic cation transport	0.005290332	0.627084092
GO:0007093 mitotic cell cycle checkpoint	0.005300104	0.627084092
GO:0006401 RNA catabolic process	0.005378556	0.627084092
GO:0030900 forebrain development	0.005424072	0.627084092
GO:0005179 hormone activity	0.006004623	0.627084092
GO:0021877 forebrain neuron fate commitment	0.006128252	0.627084092
GO:0098687 chromosomal region	0.006210592	0.627084092
GO:0042493 response to drug	0.006297133	0.627084092
GO:0042611 MHC protein complex	0.006721848	0.627084092
GO:0048167 regulation of synaptic plasticity	0.006783476	0.627084092
GO:0016072 rRNA metabolic process	0.007048525	0.627084092
GO:0021766 hippocampus development	0.007195134	0.627084092
GO:0008201 heparin binding	0.007219967	0.627084092

Ontology	*P-value*	*Q-value*
GO:0000075 cell cycle checkpoint	0.007251193	0.627084092
GO:0002521 leukocyte differentiation	0.007360891	0.627084092
GO:0000819 sister chromatid segregation	0.007796558	0.627084092
GO:0021542 dentate gyrus development	0.007804275	0.627084092
GO:0051480 cytosolic calcium ion homeostasis	0.008034365	0.627084092
GO:0006397 mRNA processing	0.008182842	0.627084092
GO:0048706 embryonic skeletal system development	0.00824401	0.627084092
GO:0051321 meiotic cell cycle	0.00825006	0.627084092
GO:0031124 mRNA 3'-end processing	0.008535111	0.627084092
GO:0009897 external side of plasma membrane	0.008551188	0.627084092
GO:0042613 MHC class II protein complex	0.008585498	0.627084092
GO:0006364 rRNA processing	0.008601257	0.627084092
GO:0051983 regulation of chromosome segregation	0.008787114	0.627084092
GO:0046651 lymphocyte proliferation	0.008967986	0.627084092
GO:0000956 nuclear-transcribed mRNA catabolic process	0.009027525	0.627084092
GO:0007204 positive regulation of cytosolic calcium ion concentration	0.009056776	0.627084092
GO:0022407 regulation of cell-cell adhesion	0.009566128	0.627084092
GO:0006402 mRNA catabolic process	0.009633908	0.627084092
GO:0006836 neurotransmitter transport	0.009961799	0.627084092
GO:0007596 blood coagulation	0.010002437	0.627084092
GO:0032649 regulation of interferon-gamma production	0.010015952	0.627084092
GO:0021895 cerebral cortex neuron differentiation	0.01019311	0.627084092
GO:0007611 learning or memory	0.010273824	0.627084092
GO:0070661 leukocyte proliferation	0.010327784	0.627084092
GO:0043090 amino acid import	0.010364372	0.627084092
GO:1903037 regulation of leukocyte cell-cell adhesion	0.010495701	0.627084092
GO:0001824 blastocyst development	0.010509038	0.627084092
GO:1903039 positive regulation of leukocyte cell-cell adhesion	0.010560181	0.627084092
GO:0007599 hemostasis	0.010767125	0.627084092
GO:0043092 L-amino acid import	0.010780442	0.627084092
GO:0036464 cytoplasmic ribonucleoprotein granule	0.010811483	0.627084092
GO:0030217 T cell differentiation	0.010846338	0.627084092
GO:0007631 feeding behavior	0.010889989	0.627084092
GO:0046883 regulation of hormone secretion	0.011008784	0.627084092
GO:0009306 protein secretion	0.011023138	0.627084092
GO:0009914 hormone transport	0.0110372	0.627084092
Neuron Downregulated Terms		
GO:0016054 organic acid catabolic process	0.00233247	0.999656023
GO:0046395 carboxylic acid catabolic process	0.00233247	0.999656023
GO:0060271 cilium morphogenesis	0.00273151	0.999656023
GO:0072329 monocarboxylic acid catabolic process	0.002999731	0.999656023
GO:0009062 fatty acid catabolic process	0.00368891	0.999656023
GO:0019048 modulation by virus of host morphology or physiology	0.004177237	0.999656023
GO:0044782 cilium organization	0.006003419	0.999656023
GO:0003995 acyl-CoA dehydrogenase activity	0.006398074	0.999656023

Ontology	*P-value*	*Q-value*
GO:0052890 oxidoreductase activity, acting on the CH-CH group of donors, with a flavin as acceptor	0.006398074	0.999656023
GO:0006635 fatty acid beta-oxidation	0.007834562	0.999656023
GO:0001578 microtubule bundle formation	0.008083997	0.999656023
GO:0042384 cilium assembly	0.009618991	0.999656023
GO:0005930 axoneme	0.010010733	0.999656023
GO:0097014 ciliary cytoplasm	0.010010733	0.999656023
GO:0044282 small molecule catabolic process	0.01001128	0.999656023
GO:0000062 fatty-acyl-CoA binding	0.010525073	0.999656023
GO:0044003 modification by symbiont of host morphology or physiology	0.010569707	0.999656023
GO:0033539 fatty acid beta-oxidation using acyl-CoA dehydrogenase	0.011325057	0.999656023
GO:0005929 cilium	0.013352073	0.999656023
GO:1901565 organonitrogen compound catabolic process	0.013448626	0.999656023
GO:0050662 coenzyme binding	0.014843757	0.999656023
GO:0050660 flavin adenine dinucleotide binding	0.014924624	0.999656023
GO:0016197 endosomal transport	0.014990647	0.999656023
GO:0000038 very long-chain fatty acid metabolic process	0.015126668	0.999656023
GO:0007034 vacuolar transport	0.016028281	0.999656023
GO:0035082 axoneme assembly	0.01611148	0.999656023
GO:0009145 purine nucleoside triphosphate biosynthetic process	0.01650754	0.999656023
GO:0042073 intraciliary transport	0.019351803	0.999656023
GO:0009251 glucan catabolic process	0.022396989	0.999656023
GO:0016798 hydrolase activity, acting on glycosyl bonds	0.022462738	0.999656023
GO:0034260 negative regulation of GTPase activity	0.022484344	0.999656023
GO:0004602 glutathione peroxidase activity	0.022761779	0.999656023
GO:0007041 lysosomal transport	0.022854901	0.999656023
GO:0009206 purine ribonucleoside triphosphate biosynthetic process	0.023021107	0.999656023
GO:0044247 cellular polysaccharide catabolic process	0.024175032	0.999656023
GO:0036159 inner dynein arm assembly	0.024270533	0.999656023
GO:0005980 glycogen catabolic process	0.02495463	0.999656023
GO:0046039 GTP metabolic process	0.026022687	0.999656023
GO:0048037 cofactor binding	0.026939554	0.999656023
GO:0016790 thiolester hydrolase activity	0.027023183	0.999656023
GO:0015929 hexosaminidase activity	0.027060434	0.999656023
GO:0090314 positive regulation of protein targeting to membrane	0.028033678	0.999656023
GO:0030118 clathrin coat	0.028156937	0.999656023
GO:0009201 ribonucleoside triphosphate biosynthetic process	0.028250164	0.999656023
GO:0030031 cell projection assembly	0.030092229	0.999656023
GO:0030132 clathrin coat of coated pit	0.03021197	0.999656023
GO:0000272 polysaccharide catabolic process	0.031437833	0.999656023
GO:0044437 vacuolar part	0.031550553	0.999656023
GO:0031146 SCF-dependent proteasomal ubiquitin-dependent protein catabolic process	0.032074914	0.999656023
GO:0060972 left/right pattern formation	0.0320859	0.999656023
GO:0018345 protein palmitoylation	0.032252619	0.999656023
GO:0046134 pyrimidine nucleoside biosynthetic process	0.032792513	0.999656023
GO:0033540 fatty acid beta-oxidation using acyl-CoA oxidase	0.033337789	0.999656023

Ontology	*P-value*	*Q-value*
GO:0009142 nucleoside triphosphate biosynthetic process	0.034176726	0.999656023
GO:0090311 regulation of protein deacetylation	0.034603562	0.999656023
GO:0046835 carbohydrate phosphorylation	0.035064691	0.999656023
GO:0061371 determination of heart left/right asymmetry	0.035230821	0.999656023
GO:0014044 Schwann cell development	0.03542371	0.999656023
GO:0006241 CTP biosynthetic process	0.035706696	0.999656023
GO:0046036 CTP metabolic process	0.035706696	0.999656023
GO:0001738 morphogenesis of a polarized epithelium	0.036313377	0.999656023
GO:1903010 regulation of bone development	0.03743552	0.999656023
GO:0007031 peroxisome organization	0.037448749	0.999656023
GO:0097576 vacuole fusion	0.037766478	0.999656023
GO:0008093 cytoskeletal adaptor activity	0.039635874	0.999656023
GO:0042147 retrograde transport, endosome to Golgi	0.042188453	0.999656023
GO:0030119 AP-type membrane coat adaptor complex	0.042253215	0.999656023
GO:0044441 ciliary part	0.042460124	0.999656023
GO:0071276 cellular response to cadmium ion	0.042689966	0.999656023
GO:0006623 protein targeting to vacuole	0.043183933	0.999656023
GO:0022011 myelination in peripheral nervous system	0.043242223	0.999656023
GO:0032292 peripheral nervous system axon ensheathment	0.043242223	0.999656023
GO:0017015 regulation of transforming growth factor beta receptor signaling pathway	0.044575622	0.999656023
GO:1903844 regulation of cellular response to transforming growth factor beta stimulus	0.044575622	0.999656023
GO:0005905 coated pit	0.044648145	0.999656023
GO:0008013 beta-catenin binding	0.045614297	0.999656023
GO:0051817 modification of morphology or physiology of other organism involved in symbiotic interaction	0.045637413	0.999656023
GO:0031063 regulation of histone deacetylation	0.045815612	0.999656023
GO:0007033 vacuole organization	0.046043574	0.999656023
GO:0014037 Schwann cell differentiation	0.046390538	0.999656023
GO:0032456 endocytic recycling	0.046496909	0.999656023
GO:0007368 determination of left/right symmetry	0.046570039	0.999656023
GO:0070286 axonemal dynein complex assembly	0.048454895	0.999656023
GO:0055064 chloride ion homeostasis	0.048711802	0.999656023
GO:0009063 cellular amino acid catabolic process	0.049864337	0.999656023
GO:0050431 transforming growth factor beta binding	0.04997413	0.999656023
GO:0005779 integral component of peroxisomal membrane	0.05030793	0.999656023
GO:0031231 intrinsic component of peroxisomal membrane	0.05030793	0.999656023
GO:0097502 mannosylation	0.052018383	0.999656023
GO:0032266 phosphatidylinositol-3-phosphate binding	0.05229823	0.999656023
GO:0004029 aldehyde dehydrogenase (NAD) activity	0.052499884	0.999656023
GO:0030122 AP-2 adaptor complex	0.05273994	0.999656023
GO:0030128 clathrin coat of endocytic vesicle	0.05273994	0.999656023
GO:0009070 serine family amino acid biosynthetic process	0.053077779	0.999656023
GO:0097352 autophagosome maturation	0.053896412	0.999656023
GO:0005759 mitochondrial matrix	0.054383864	0.999656023
GO:0046132 pyrimidine ribonucleoside biosynthetic process	0.054986163	0.999656023
GO:0019054 modulation by virus of host process	0.055006647	0.999656023

Ontology	*P-value*	*Q-value*
GO:0051917 regulation of fibrinolysis	0.055262077	0.999656023
GO:0005777 peroxisome	0.055710631	0.999656023

Supplementary Table 9 - KEGG Pathways for NPC and Neurons (top 50).

Pathway	*P-Value*	*Q-Value*
NPC Upregulated Pathways		
hsa03010 Ribosome	3.51E-08	5.69E-06
hsa00190 Oxidative phosphorylation	6.75E-05	0.005465069
hsa03013 RNA transport	0.003534145	0.145950205
hsa00970 Aminoacyl-tRNA biosynthesis	0.004496138	0.145950205
hsa04260 Cardiac muscle contraction	0.005551884	0.145950205
hsa03008 Ribosome biogenesis in eukaryotes	0.005881066	0.145950205
hsa00620 Pyruvate metabolism	0.00630649	0.145950205
hsa00240 Pyrimidine metabolism	0.010096882	0.189929902
hsa00290 Valine, leucine and isoleucine biosynthesis	0.010551661	0.189929902
hsa00010 Glycolysis / Gluconeogenesis	0.018615854	0.287481947
hsa03410 Base excision repair	0.019520379	0.287481947
hsa00520 Amino sugar and nucleotide sugar metabolism	0.024195142	0.31141154
hsa00020 Citrate cycle (TCA cycle)	0.026230268	0.31141154
hsa00670 One carbon pool by folate	0.026983303	0.31141154
hsa00480 Glutathione metabolism	0.028834402	0.31141154
hsa04141 Protein processing in endoplasmic reticulum	0.030865342	0.312511584
hsa03020 RNA polymerase	0.034264977	0.326525074
hsa00860 Porphyrin and chlorophyll metabolism	0.042570351	0.373536562
hsa00330 Arginine and proline metabolism	0.043809844	0.373536562
hsa00030 Pentose phosphate pathway	0.050583485	0.403614665
hsa04360 Axon guidance	0.05232042	0.403614665
hsa04920 Adipocytokine signaling pathway	0.059142705	0.425897118
hsa03050 Proteasome	0.060466875	0.425897118
hsa04610 Complement and coagulation cascades	0.069263348	0.467527599
hsa04964 Proximal tubule bicarbonate reclamation	0.074554069	0.483110365
hsa00510 N-Glycan biosynthesis	0.089528502	0.557831434
hsa00270 Cysteine and methionine metabolism	0.093331237	0.559987422
hsa00770 Pantothenate and CoA biosynthesis	0.097709398	0.564588026
hsa00230 Purine metabolism	0.101068227	0.564588026
hsa00531 Glycosaminoglycan degradation	0.109280663	0.590115579
hsa03030 DNA replication	0.121535515	0.635121076
hsa00052 Galactose metabolism	0.139937373	0.693133743
hsa04142 Lysosome	0.15870345	0.693133743
hsa03060 Protein export	0.164711918	0.693133743
hsa00760 Nicotinate and nicotinamide metabolism	0.166908368	0.693133743
hsa00910 Nitrogen metabolism	0.176969993	0.693133743
hsa04110 Cell cycle	0.181048242	0.693133743
hsa04340 Hedgehog signaling pathway	0.189493085	0.693133743
hsa00563 Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	0.190808193	0.693133743
hsa04744 Phototransduction	0.199144227	0.693133743
hsa00561 Glycerolipid metabolism	0.200805782	0.693133743
hsa00590 Arachidonic acid metabolism	0.205718011	0.693133743
hsa00565 Ether lipid metabolism	0.208266459	0.693133743

Pathway	*P-Value*	*Q-Value*
hsa04120 Ubiquitin mediated proteolysis	0.217458434	0.693133743
hsa00250 Alanine, aspartate and glutamate metabolism	0.218238658	0.693133743
hsa00980 Metabolism of xenobiotics by cytochrome P450	0.218568686	0.693133743
hsa00630 Glyoxylate and dicarboxylate metabolism	0.226273578	0.693133743
hsa03320 PPAR signaling pathway	0.228736383	0.693133743
hsa00051 Fructose and mannose metabolism	0.23006482	0.693133743
hsa00260 Glycine, serine and threonine metabolism	0.233193758	0.693133743
NPC Downregulated Pathways		
hsa04062 Chemokine signaling pathway	0.018736484	0.999999965
hsa04672 Intestinal immune network for IgA production	0.030820842	0.999999965
hsa04270 Vascular smooth muscle contraction	0.0322204	0.999999965
hsa04622 RIG-I-like receptor signaling pathway	0.055023791	0.999999965
hsa04070 Phosphatidylinositol signaling system	0.073187515	0.999999965
hsa04620 Toll-like receptor signaling pathway	0.112457876	0.999999965
hsa02010 ABC transporters	0.129526953	0.999999965
hsa04912 GnRH signaling pathway	0.135090757	0.999999965
hsa04720 Long-term potentiation	0.159622722	0.999999965
hsa00650 Butanoate metabolism	0.161641392	0.999999965
hsa00900 Terpenoid backbone biosynthesis	0.166424526	0.999999965
hsa04630 Jak-STAT signaling pathway	0.176390209	0.999999965
hsa00920 Sulfur metabolism	0.183843172	0.999999965
hsa04614 Renin-angiotensin system	0.185026984	0.999999965
hsa04670 Leukocyte transendothelial migration	0.185285847	0.999999965
hsa04710 Circadian rhythm - mammal	0.21462873	0.999999965
hsa00360 Phenylalanine metabolism	0.217925714	0.999999965
hsa04664 Fc epsilon RI signaling pathway	0.22136683	0.999999965
hsa04730 Long-term depression	0.235143127	0.999999965
hsa00790 Folate biosynthesis	0.23862334	0.999999965
hsa04971 Gastric acid secretion	0.247418808	0.999999965
hsa04973 Carbohydrate digestion and absorption	0.25023419	0.999999965
hsa00120 Primary bile acid biosynthesis	0.251132438	0.999999965
hsa04974 Protein digestion and absorption	0.256454445	0.999999965
hsa04810 Regulation of actin cytoskeleton	0.260235872	0.999999965
hsa03450 Non-homologous end-joining	0.260988769	0.999999965
hsa00562 Inositol phosphate metabolism	0.266288626	0.999999965
hsa00350 Tyrosine metabolism	0.266347389	0.999999965
hsa00450 Selenocompound metabolism	0.270896305	0.999999965
hsa04621 NOD-like receptor signaling pathway	0.271710617	0.999999965
hsa04020 Calcium signaling pathway	0.273437534	0.999999965
hsa04150 mTOR signaling pathway	0.276501168	0.999999965
hsa00340 Histidine metabolism	0.285595258	0.999999965
hsa04650 Natural killer cell mediated cytotoxicity	0.296274389	0.999999965
hsa04742 Taste transduction	0.302343615	0.999999965
hsa04146 Peroxisome	0.311406551	0.999999965
hsa04612 Antigen processing and presentation	0.320327886	0.999999965

Pathway	*P-Value*	*Q-Value*
hsa04514 Cell adhesion molecules (CAMs)	0.327895166	0.999999965
hsa04660 T cell receptor signaling pathway	0.330689012	0.999999965
hsa04140 Regulation of autophagy	0.337050871	0.999999965
hsa04666 Fc gamma R-mediated phagocytosis	0.338445174	0.999999965
hsa00140 Steroid hormone biosynthesis	0.343615017	0.999999965
hsa04144 Endocytosis	0.362313529	0.999999965
hsa04623 Cytosolic DNA-sensing pathway	0.364144373	0.999999965
hsa04976 Bile secretion	0.366232824	0.999999965
hsa00603 Glycosphingolipid biosynthesis - globo series	0.373172941	0.999999965
hsa04540 Gap junction	0.377741591	0.999999965
hsa04966 Collecting duct acid secretion	0.400340328	0.999999965
hsa03440 Homologous recombination	0.406079611	0.999999965
hsa04970 Salivary secretion	0.417848438	0.999999965
Neuron Upregulated Pathways		
hsa04514 Cell adhesion molecules (CAMs)	0.001633059	0.264555532
hsa03013 RNA transport	0.004261611	0.34519049
hsa03018 RNA degradation	0.010434594	0.400975865
hsa04540 Gap junction	0.012197712	0.400975865
hsa03015 mRNA surveillance pathway	0.014913783	0.400975865
hsa04662 B cell receptor signaling pathway	0.01849575	0.400975865
hsa04020 Calcium signaling pathway	0.02490763	0.400975865
hsa04964 Proximal tubule bicarbonate reclamation	0.026963391	0.400975865
hsa03010 Ribosome	0.027756869	0.400975865
hsa04640 Hematopoietic cell lineage	0.033107754	0.400975865
hsa04971 Gastric acid secretion	0.033307244	0.400975865
hsa03008 Ribosome biogenesis in eukaryotes	0.033879545	0.400975865
hsa04110 Cell cycle	0.034103384	0.400975865
hsa04973 Carbohydrate digestion and absorption	0.036684548	0.400975865
hsa04664 Fc epsilon RI signaling pathway	0.037876024	0.400975865
hsa04974 Protein digestion and absorption	0.039602555	0.400975865
hsa03040 Spliceosome	0.042524335	0.4052319
hsa04976 Bile secretion	0.052726927	0.455684141
hsa04920 Adipocytokine signaling pathway	0.059075542	0.455684141
hsa00100 Steroid biosynthesis	0.060259726	0.455684141
hsa04672 Intestinal immune network for IgA production	0.062006864	0.455684141
hsa04260 Cardiac muscle contraction	0.063999191	0.455684141
hsa04730 Long-term depression	0.064695897	0.455684141
hsa04010 MAPK signaling pathway	0.072151616	0.487023407
hsa04360 Axon guidance	0.080518523	0.49687329
hsa00980 Metabolism of xenobiotics by cytochrome P450	0.082618534	0.49687329
hsa04970 Salivary secretion	0.082812215	0.49687329
hsa04972 Pancreatic secretion	0.088766521	0.513577729
hsa00601 Glycosphingolipid biosynthesis - lacto and neolacto series	0.092441916	0.516399668
hsa04510 Focal adhesion	0.097325721	0.525558894
hsa04114 Oocyte meiosis	0.102064768	0.533370723

Pathway	*P-Value*	*Q-Value*
hsa04710 Circadian rhythm - mammal	0.117580896	0.572147256
hsa00910 Nitrogen metabolism	0.118628206	0.572147256
hsa04145 Phagosome	0.120080288	0.572147256
hsa04660 T cell receptor signaling pathway	0.132673619	0.601019785
hsa00970 Aminoacyl-tRNA biosynthesis	0.135368041	0.601019785
hsa04612 Antigen processing and presentation	0.138097073	0.601019785
hsa00450 Selenocompound metabolism	0.14462952	0.601019785
hsa04270 Vascular smooth muscle contraction	0.154504986	0.601019785
hsa04740 Olfactory transduction	0.155568248	0.601019785
hsa03420 Nucleotide excision repair	0.157860281	0.601019785
hsa04370 VEGF signaling pathway	0.163198571	0.601019785
hsa04916 Melanogenesis	0.163204823	0.601019785
hsa04914 Progesterone-mediated oocyte maturation	0.163239942	0.601019785
hsa04720 Long-term potentiation	0.18888646	0.630323316
hsa04650 Natural killer cell mediated cytotoxicity	0.190373361	0.630323316
hsa04630 Jak-STAT signaling pathway	0.193548581	0.630323316
hsa00230 Purine metabolism	0.19587644	0.630323316
hsa00670 One carbon pool by folate	0.196263511	0.630323316
hsa03430 Mismatch repair	0.19741609	0.630323316
Neuron Downregulated Pathways		
hsa04330 Notch signaling pathway	0.00530333	0.859139457
hsa04146 Peroxisome	0.011293301	0.914757359
hsa04122 Sulfur relay system	0.042563429	0.998366941
hsa00280 Valine, leucine and isoleucine degradation	0.04957277	0.998366941
hsa04142 Lysosome	0.055124171	0.998366941
hsa00511 Other glycan degradation	0.056881333	0.998366941
hsa00650 Butanoate metabolism	0.062478675	0.998366941
hsa00532 Glycosaminoglycan biosynthesis - chondroitin sulfate	0.082769677	0.998366941
hsa01040 Biosynthesis of unsaturated fatty acids	0.09707081	0.998366941
hsa00531 Glycosaminoglycan degradation	0.1342438	0.998366941
hsa00120 Primary bile acid biosynthesis	0.141214948	0.998366941
hsa00071 Fatty acid metabolism	0.146152746	0.998366941
hsa04622 RIG-I-like receptor signaling pathway	0.159044888	0.998366941
hsa00260 Glycine, serine and threonine metabolism	0.163450192	0.998366941
hsa04810 Regulation of actin cytoskeleton	0.169079929	0.998366941
hsa00350 Tyrosine metabolism	0.17549272	0.998366941
hsa00920 Sulfur metabolism	0.185477031	0.998366941
hsa04962 Vasopressin-regulated water reabsorption	0.187123211	0.998366941
hsa00562 Inositol phosphate metabolism	0.190781867	0.998366941
hsa00040 Pentose and glucuronate interconversions	0.211686661	0.998366941
hsa00640 Propanoate metabolism	0.215599728	0.998366941
hsa00340 Histidine metabolism	0.218486432	0.998366941
hsa00760 Nicotinate and nicotinamide metabolism	0.219732886	0.998366941
hsa00380 Tryptophan metabolism	0.22322052	0.998366941
hsa00563 Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	0.244735229	0.998366941

Pathway	*P-Value*	*Q-Value*
hsa03320 PPAR signaling pathway	0.252866447	0.998366941
hsa00240 Pyrimidine metabolism	0.255568234	0.998366941
hsa00310 Lysine degradation	0.255972541	0.998366941
hsa04910 Insulin signaling pathway	0.273937805	0.998366941
hsa00053 Ascorbate and aldarate metabolism	0.280365518	0.998366941
hsa00564 Glycerophospholipid metabolism	0.280613642	0.998366941
hsa04722 Neurotrophin signaling pathway	0.292352051	0.998366941
hsa04140 Regulation of autophagy	0.297520886	0.998366941
hsa00620 Pyruvate metabolism	0.30713467	0.998366941
hsa04070 Phosphatidylinositol signaling system	0.314148488	0.998366941
hsa00360 Phenylalanine metabolism	0.317153198	0.998366941
hsa04210 Apoptosis	0.33431982	0.998366941
hsa04120 Ubiquitin mediated proteolysis	0.336112862	0.998366941
hsa04610 Complement and coagulation cascades	0.344875128	0.998366941
hsa00020 Citrate cycle (TCA cycle)	0.351172329	0.998366941
hsa00790 Folate biosynthesis	0.353175206	0.998366941
hsa00500 Starch and sucrose metabolism	0.37634617	0.998366941
hsa04620 Toll-like receptor signaling pathway	0.376471046	0.998366941
hsa00603 Glycosphingolipid biosynthesis - globo series	0.379499709	0.998366941
hsa03020 RNA polymerase	0.389236331	0.998366941
hsa00982 Drug metabolism - cytochrome P450	0.405226905	0.998366941
hsa00480 Glutathione metabolism	0.424305621	0.998366941
hsa04144 Endocytosis	0.427287288	0.998366941
hsa03410 Base excision repair	0.428277406	0.998366941
hsa00270 Cysteine and methionine metabolism	0.453400034	0.998366941

References

Ables, J.L. et al., 2011. Not(ch) just development: Notch signalling in the adult brain. *Nature Reviews Neuroscience*, 12(5), p.269.

Alarcón, M. et al., 2008. Linkage, Association, and Gene-Expression Analyses Identify CNTNAP2 as an Autism-Susceptibility Gene. *The American Journal of Human Genetics*, 82(1), pp.150–159.

Alexa, A. & Rahnenfuhrer, J., 2016. topGO: Enrichment Analysis for Gene Ontology. R package version 2.28.0.

Anders, S., Pyl, P.T. & Huber, W., 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2), pp.166–169.

Ansari, A.Z., 2009. Riboactivators: transcription activation by noncoding RNA. *Critical Reviews in Biochemistry and Molecular Biology*, 44(1), pp.50–61.

Barrett, L.W., Fletcher, S. & Wilton, S.D., 2013. Untranslated Gene Regions and Other Non-coding Elements. In pp. 1–56.

Bogdanović, O. & Veenstra, G.J.C., 2009. DNA methylation and methyl-CpG binding proteins: developmental requirements and function. *Chromosoma*, 118(5), pp.549–65.

Bonnet, C. et al., 2013. Extended spectrum of MBD5 mutations in neurodevelopmental disorders. *European Journal of Human Genetics : EJHG*, 21(12), pp.1457–61.

Camarena, V. et al., 2014. Disruption of Mbd5 in mice causes neuronal functional deficits and neurobehavioral abnormalities consistent with 2q23.1 microdeletion syndrome. *EMBO Molecular Medicine*, 6(8), pp.1003–15.

Campbell, D.B. et al., 2008. Genetic evidence implicating multiple genes in the MET receptor tyrosine kinase pathway in autism spectrum disorder. *Autism Research*, 1(3), pp.159–168.

Chahrour, M. et al., 2008. MeCP2, a key contributor to neurological disease, activates and represses transcription. *Science (New York, N.Y.)*, 320(5880), pp.1224–9.

DeLuca, D.S. et al., 2012. RNA-SeQC: RNA-seq metrics for quality control and

process optimization. *Bioinformatics (Oxford, England)*, 28(11), pp.1530–2.

DeWitt, J. et al., 2016. Impact of the Autism-Associated Long Noncoding RNA MSNP1AS on Neuronal Architecture and Gene Expression in Human Neural Progenitor Cells. *Genes*, 7(10), p.76.

Djupedal, I. & Ekwall, K., 2009. Epigenetics: heterochromatin meets RNAi. *Cell Research*, 19(3), pp.282–295.

Dobin, A. et al., 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1), pp.15–21.

Du, Y. et al., 2012. The essential role of Mbd5 in the regulation of somatic growth and glucose homeostasis in mice. J. A. Chowen, ed. *PloS one*, 7(10), p.e47358.

Eagleson, K.L. et al., 2011. The autism risk genes MET and PLAUR differentially impact cortical development. *Autism Research*, 4(1), pp.68–83.

Eva, R. et al., 2010. Rab11 and Its Effector Rab Coupling Protein Contribute to the Trafficking of β1 Integrins during Axon Growth in Adult Dorsal Root Ganglion Neurons and PC12 Cells. *Journal of Neuroscience*, 30(35).

Hodge, J.C. et al., 2014. Disruption of MBD5 contributes to a spectrum of psychopathology and neurodevelopmental abnormalities. *Molecular Psychiatry*, 19(3), pp.368–79.

Itskovitz-Eldor, J. et al., 2000. Differentiation of human embryonic stem cells into embryoid bodies compromising the three embryonic germ layers. *Molecular Medicine (Cambridge, Mass.)*, 6(2), pp.88–95.

Iyer, M.K. et al., 2015. The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics*, 47(3), pp.199–208.

Jiang, Y. et al., 2011. Setdb1-mediated histone H3K9 hypermethylation in neurons worsens the neurological phenotype of Mecp2-deficient mice. *Neuropharmacology*, 60(7-8), pp.1088–97.

Kong, L.-J. et al., 2007. Compensation and specificity of function within the E2F family. *Oncogene*, 26(3), pp.321–327.

Kriegstein, A. & Alvarez-Buylla, A., 2009. The Glial Nature of Embryonic and Adult Neural Stem Cells. *Annual Review of Neuroscience*, 32(1), pp.149–184.

Laget, S. et al., 2010. The human proteins MBD5 and MBD6 associate with

heterochromatin but they do not bind methylated DNA. S. D. Fugmann, ed. *PloS One*, 5(8), p.e11982.

Leek, J.T. et al., 2007. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genetics*, 3(9), p.e161.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp.2078–2079.

Lipkowitz, S. et al., 1992. A comparative structural characterization of the human NSCL-1 and NSCL-2 genes. Two basic helix-loop-helix genes expressed in the developing nervous system. *The Journal of Biological Chemistry*, 267(29), pp.21065–71.

Livak, K.J. & Schmittgen, T.D., 2001. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods*, 25(4), pp.402–408.

Love, M.I., Huber, W. & Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), p.550.

Luo, W. et al., 2009. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10(1), p.161.

Mattick, J.S., 2004. Opinion: RNA regulation: a new genetics? *Nature Reviews Genetics*, 5(4), pp.316–323.

Mullegama, S. V et al., 2013. Reciprocal deletion and duplication at 2q23.1 indicates a role for MBD5 in autism spectrum disorder. *European Journal of Human Genetics : EJHG*, (November 2012), pp.1–7.

Mullegama, S. V & Elsea, S.H., 2016. Clinical and molecular MBD5-neurodevelopmental disorder (MAND). *European Journal of Human Genetics*.

Murdoch, J.N. et al., 1999. Sequence and expression analysis ofNhlh1: a basic helix-loop-helix gene implicated in neurogenesis. *Developmental Genetics*, 24(1-2), pp.165–177.

Nan, X. & Bird, A., 2001. The biological functions of the methyl-CpG-binding protein MeCP2 and its implication in Rett syndrome. *Brain & Development*, 23 Suppl 1, pp.S32–7.

Ng, H.H. et al., 1999. MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. *Nature Genetics*, 23(1), pp.58–61.

O'Roak, B.J. et al., 2011. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genetics*, 43(6), pp.585–9.

Peñagarikano, O. et al., 2011. Absence of CNTNAP2 Leads to Epilepsy, Neuronal Migration Abnormalities, and Core Autism-Related Deficits. *Cell*, 147(1), pp.235–246.

Poliak, S. et al., 1999. Caspr2, a New Member of the Neurexin Superfamily, Is Localized at the Juxtaparanodes of Myelinated Axons and Associates with K+ Channels. *Neuron*, 24(4), pp.1037–1047.

Robinson, J.T. et al., 2011. Integrative genomics viewer. *Nature Biotechnology*, 29(1), pp.24–6.

Roloff, T.C., Ropers, H.H. & Nuber, U.A., 2003. Comparative study of methyl-CpG-binding domain proteins. *BMC genomics*, 4(1), p.1.

Rossin, E.J. et al., 2011. Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology T. Gojobori, ed. *PLoS Genetics*, 7(1), p.e1001273.

Santos, M. et al., 2007. Evidence for abnormal early development in a mouse model of Rett syndrome. *Genes, Brain and Behavior*, 6(3), pp.277–286.

Schafer, J.C. et al., 2016. Rab11-FIP1A regulates early trafficking into the recycling endosomes. *Experimental Cell Research*, 340(2), pp.259–273.

Schmid, T., Krüger, M. & Braun, T., 2007. NSCL-1 and -2 control the formation of precerebellar nuclei by orchestrating the migration of neuronal precursor cells. *Journal of Neurochemistry*, 102(6), pp.2061–2072.

Shahbazian, M.D. et al., 2002. Insight into Rett syndrome: MeCP2 levels display tissue- and cell-specific differences and correlate with neuronal maturation. *Human Molecular Genetics*, 11(2), pp.115–24.

Sheridan, S.D. et al., 2011. Epigenetic Characterization of the FMR1 Gene and Aberrant Neurodevelopment in Human Induced Pluripotent Stem Cell Models of Fragile X Syndrome M. R. Cookson, ed. *PLoS ONE*, 6(10), p.e26203.

Sim, J.C.H., White, S.M. & Lockhart, P.J., 2015. ARID1B-mediated disorders: Mutations and possible mechanisms. *Intractable & rare diseases research*, 4(1), pp.17–23.

Stec, I. et al., 2000. The PWWP domain: a potential protein-protein interaction domain in nuclear proteins influencing differentiation? *FEBS letters*, 473(1), pp.1–5.

Sugathan, A. et al., 2014. *CHD8* regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. *Proceedings of the National Academy of Sciences*, 111(42), pp.E4468–E4477.

Talkowski, M.E. et al., 2011. Assessment of 2q23 . 1 Microdeletion Syndrome Implicates MBD5 as a Single Causal Locus of Intellectual Disability , Epilepsy , and Autism Spectrum Disorder. *The American Journal of Human Genetics*, pp.551–563.

Talkowski, M.E. et al., 2012. Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell*, 149(3), pp.525–37.

Tanapat, P., 2013. Neuronal Cell Markers. *Materials and Methods*, 3.

Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2), pp.178–92.

Ulitsky, I. & Bartel, D.P., 2013. lincRNAs: Genomics, Evolution, and Mechanisms. *Cell*, 154(1), pp.26–46.

Wang, L., Wang, S. & Li, W., 2012. RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, 28(16), pp.2184–2185.

Williamson, M.P., 1994. The structure and function of proline-rich regions in proteins. , 260, pp.249–260.

Wu, T.D. & Nacu, S., 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics (Oxford, England)*, 26(7), pp.873–81.

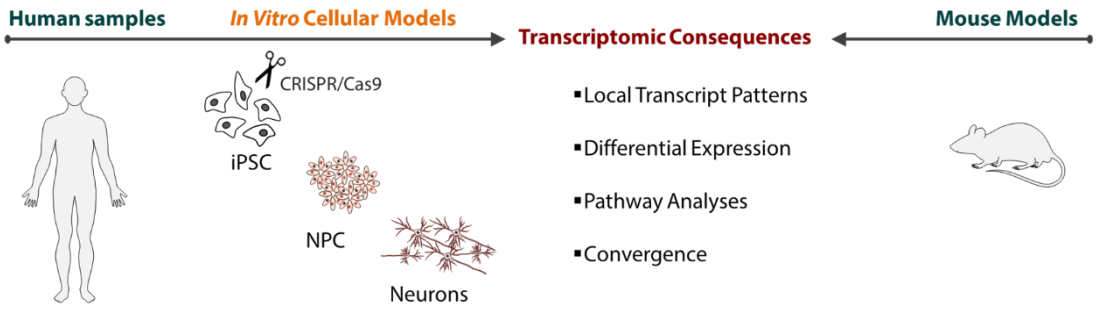
Yabe, J.T. et al., 2003. Regulation of the transition from vimentin to neurofilaments during neuronal differentiation. *Cell Motility and the Cytoskeleton*, 56(3), pp.193–205.

Yasui, D.H. et al., 2007. Integrated epigenomic analyses of neuronal MeCP2 reveal a role for long-range interaction with active genes. *Proceedings of the National Academy of Sciences of the United States of America*, 104(49), pp.19416–21.

Yizhar, O. et al., 2011. Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature*, 477(7363), pp.171–178.

Young, J.I. et al., 2005. Regulation of RNA splicing by the methylation-dependent transcriptional repressor methyl-CpG binding protein 2. *Proceedings of the National Academy of Sciences*, 102(49), pp.17551–17558.

Final Remarks



Highlights

This section summarizes the results obtained in this study, and approaches the limitations of the experimental setup, the main conclusions obtained thus far, and also includes additional preliminary results.

Final Remarks

Summary of Findings

Chapter 1 focuses on a patient with intellectual disability bearing a unique apparent *de novo* mutation of *ARID1B* due to an intragenic duplication. Array CGH showed a copy number gain of chromosome band 6q25.3 of approximately 0.361 Mb in size, arr 6q25.3(157133792-157495187)x3 dn, that involved duplication of a segment containing exons 2-10 (ENST00000346085) of *ARID1B*. Haploinsufficiency of *ARID1B* was determined both by RNA sequencing and quantitative RT-PCR indicating that haploinsufficiency for *ARID1B* is likely to be the cause of developmental delay in this individual, as no additional genetic or metabolic defects were identified. The most abundant *ARID1B* mRNA transcript observed in the patient was ENST00000414678, consistent with the GTEx database for EBV-transformed lymphoblasts and IGV Sashimi plots did not reveal novel junctions or transcripts for this gene in the patient, providing no evidence that duplication of this exonic region resulted in alterations in splicing. Our results support the view that *ARID1B* is a dosage sensitive gene whose expression can be affected by deletions or duplications with impact on phenotype. Other chromatin regulators have also been noted to be dosage sensitive causes of neurodevelopmental phenotypes, including *MECP2* (Temudo & Maciel 2002), *MBD5*, *EHMT1*, *CHD8* and *SATB2* (Talkowski et al., 2012).

Principal component analysis revealed marked differentiation of the subject's lymphoblast proteome from that of controls. Of 3427 proteins quantified, 1,014 were significantly up- or downregulated compared to controls ($q<0.01$). While the measurement of RNA in LCLs does not necessarily reflect the effect of the genetic lesion in the brain, the association of heterozygous inactivating mutations in *ARID1B* with neurodevelopmental phenotypes in other subjects suggests that reduced gene expression also occurs in the central nervous system. Pathway analysis revealed highly significant enrichment for canonical pathways of EIF2 signaling, protein ubiquitination, tRNA charging and chromosomal replication, among others. Network analyses revealed downregulation of: (1) intracellular components involved in organization of membranes, organelles and vesicles; (2) aspects of cell cycle control, signal transduction and nuclear protein export; (3) ubiquitination and proteosomal function; and (4) aspects of mRNA metabolism. This report added further insight on the role of *ARID1B* haploinsufficiency in the establishment of intellectual

disability, yet further studies are required to uncover the precise mechanisms whereby altered transcriptional and cell cycle regulation pathways lead to impaired brain development and cognitive function.

In Chapter 2, we explore the creation of isogenic cellular models to study ASD, as neuronal tissue is not readily available and lymphoblastoid cell lines (generated from blood samples) may not represent the true landscape of the affected cells and additionally, the different patients' genetic background introduces noise in the analyses. Therefore, we show that through dual-guide CRISPR/Cas9 genome editing technology we were able to generate deletions in four ASD candidate chromatin-related genes (*EHMT1*, *MBD5*, *METTL2A* and *METTL2B*) in an iPSC line from a healthy individual. CRISPR/Cas9 showed deletion efficiencies in iPSC ranging from 0% to 10,6%, suggesting that dual-guide deletion strategy efficiency varies widely by guideRNA-pair combination and genomic target. Interestingly, for two highly homologous genes targeted (that share 98,41% homology), *METTL2A* and *METTL2B*, CRISPR was able to ablate these genes with specificity to each target. We found no correlation between the deletion sizes (range from 13bp to 2200bp) and the efficiency in generating the predicted deletions. Although there seemed to be a trend towards a decrease of efficiency with deletion size, it is not significant (R^2=0,0951). These human iPSC models have the potential to be differentiated towards the tissue of interest of the researcher and can represent more comparable models of human disease.

In Chapter 3, we explored the transcriptomic consequences of ablating two different regions in the chromatin-related gene, *MBD5*. We targeted the 5' UTR exon 4 and the MBD domain in exon 6 of *MBD5* to understand the impact of these regions in the role of this protein during neurodevelopment. For this, we were able to differentiate the edited iPSC into the presumed tissue of interest in ASD - neuronal progenitors and mature neurons. The CRISPR-edited *MBD5* mutated iPSC-derived models were described and functionally characterized. The different edited iPSC lines showed *MBD5* mRNA decrease in expression ranging from 0% to 70%. The predicted CRISPR-induced cuts seemed to be repaired via NHEJ, and in some cell lines there were additional insertions or deletions of a few base pairs at the CRISPR-cutting site. As mentioned above, the CRISPR-edited iPSC were successfully differentiated into NPC, followed by withdrawal of mitogenic factors (EGF, bFGF) for 30 days to guide them into the terminally differentiated mature neuronal lineage. Upon differentiation, all CRISPR-edited cell lines showed *MBD5* mRNA expression levels similar to the controls. In fact,

in exon 4-targeted cell line *4i H7*, the expression was even higher than that observed in controls.

We used whole transcriptome sequencing (RNAseq) to determine the genome-wide transcriptomic effects caused upon the perturbation of *MBD5* in both the targeted regions by CRISPR/Cas9. Principal component analysis was used to compare the transcriptomic landscape of the cell lines, based on their principal components. PC1 explained most of the variance found in the RNAseq dataset, ~23%. This PC clearly separated the transcriptome-wide signatures between cell type (NPCs and neurons). Transcriptome sequencing allowed a deeper appreciation of the intricate nature of *MBD5* transcript architecture by identifying alternative transcripts during both stages of differentiation. Transcript MBD5-001 was not the main product of *MBD5* expression in the differentiated cells, as was previously thought. There are 4 transcripts that remain unaffected by this ablation, as they do not contain exon 6, and are initiated before or after this region. Surprisingly, the transcript that exhibited highest expression overall in the neuronal cells was MBD5-010, which only comprises 5'UTR exons 1 and 2, and may represent a regulatory RNA that should be further investigated to elucidate its role in transcriptional regulation. On the other hand, transcript MBD5-014 was observed uniquely in the NPC population and was absent in the neurons, suggesting a developmental state preference for this protein-coding transcript. Altogether, these data suggest that the MBD5 may be relevant for neurodevelopment through the action of alternative non-coding transcripts that do not depend on the presence of the MBD domain. It is also conceivable that the pathogenic effect of MBD5 haploinsufficiency on neurodevelopment occurs outside the neuronal lineage per se or is non-cell autonomous. The role neurobiological of the *MBD5* non-coding transcripts of *MBD5* should be further investigated to understand the importance of the non-coding network and the implications of dysregulation in disease.

Besides the local effects, whole transcriptome analyses allowed us to determine the genes that were differentially expressed between the wild-type and the CRISPR-edited cell lines. Among those genes, *RAB11FIP1*, the most significant DEG in NPC and also upregulated in neurons, seems like a promising candidate as this protein has previously been associated with axonal growth in mice. Other genes to pinpoint from the DEG are those that have been previously found to be disrupted in ASD patients – *PLAUR* and *CNTNAP2*. GSEA of the DEG unveiled the gene families and pathways that were enriched within the CRISPR

dataset. We found upregulated terms related to dopaminergic synapse transmission and downregulated for glutamatergic synapse, indicating a possible imbalance of neurotransmitter activity in these neuronal models. In addition to neuronal terms, we also found an enrichment for translation-related terms that suggest defects in translation of proteins required for normal synaptic function or neuronal growth. KEGG pathway analysis identified an enrichment of pathways involved in notch signaling, known to be involved in brain development; and cell adhesion. Enrichment for cell adhesion terms and genes (*NCAM1*) has previously been observed by the knockdown of the ASD-risk gene *CHD8* (Sugathan et al. 2014) in neural progenitors, which is also a chromatin remodeler as *MBD5*. This process (cell adhesion) indicates a possible mechanism by which chromatin remodelers can be acting and impacting neuronal function.

Limitations

In Chapter 1, we analyzed lymphoblastoid cell lines from blood samples donated by the patient. Lymphoblasts, that arise embryonically from the endoderm, do not originate from the same embryonic germ layer as neurons (ectoderm). Thus, these tissues have divergent maturation processes and mutations in certain genes may not have the same impact on both tissues, and this should be noted when considering transcriptomic and proteomic analyses to identify deregulated pathways of neurodevelopment disorders. To overcome this issue, we felt the need to create human neuronal models harboring LoF mutations of chromatin-related genes. These models solved some of the caveats of analyzing patient blood samples and will allow the direct comparison of the LoF models, without genetic background confounders and allow the analysis of the tissue of interest. While these 2D models do not recapitulate critical features of brain development, as neuronal migration, there has been recent effort in developing iPSC-derived 3D cerebral organoid models (Chailangkarn et al. 2016; Lancaster & Knoblich 2014) that will fill the gap left by 2D models. These organoids, that self-organize through cell sorting and spatially restricted lineage commitment, recapitulate some of the robust regulatory systems of organogenesis in terms of not only cell differentiation, but also spatial patterning and morphogenesis (Eiraku & Sasai 2012).

In Chapter 2 we generate in isogenic series of allelic mutants derived from an iPSC line that was reprogrammed using with retroviral integration (Sheridan et al. 2011). This approach of reprogramming can affect the expression of certain genes if the reprogramming cassettes are integrated within actively transcribed genes or gene promoters, and can ultimately influence RNAseq results in those cases. Alternatives to avoid integrating reprogramming methods have been proposed (reviewed in Malik & Rao 2013), such as adenovirus, Sendai virus as well as non-viral methods such as mRNA or miRNA transfection.

Another important point to consider, from Chapter 3, is that the neuronal maturation protocol used in this work drove the cells into a mixed neuronal population of excitatory and inhibitory neurons. Consequently, we are interpreting events (as the transcriptomic signature) coming from different cell types and might not be targeting the most affected population of neurons in ASD. Despite the fact that the type of neurons that are mostly affected in ASD have yet to be determined, it is believed that a disturbance in the central nervous system excitation and inhibition balance between the glutamatergic and GABAergic systems could underlie the etiology of ASD (Rubenstein & Merzenich 2003), and this is consistent with the observed high comorbidity with epilepsy. The animal model research suggests the primary factor in the excitation/inhibition imbalance is loss of GABAergic inhibitory control over excitatory neurons. This loss of inhibition appears to occur one of two ways: either disruption in GABAergic neurotransmission at the synaptic level or aberrant organization or loss of GABAergic neurons during development. Mutations in several synaptic genes, such as those encoding neuroligins, members of the SHANK family of proteins located at the synaptic density, give rise to ASD-relevant phenotypes in mouse models (Peça et al. 2011), supporting the hypothesis of altered synaptic communication in ASD etiology. Therefore, our model of a mixed population of neurons could mimic the ASD landscape to a certain extent, however, it would be beneficial to distinguish the effects of disrupting the chromatin-related genes separately in each category of neuron – excitatory/glutamatergic and inhibitory/GABAergic – using protocols to drive the differentiation process of iPSC into these cell types (Vazin et al. 2014; Liu et al. 2013; Lin et al. 2015).

Conclusions

As the efforts in ASD research grows, so does the list of genes and chromosomal regions that represent risk factors for ASD and other neurodevelopmental disorders. This project had the goal of exploring functional aspects of genes that have been previously identified to be relevant for ASD risk, as the role of most of the annotated genes remains unknown and could potentially help understand the missing heritability that may involve gene-gene interactions and epigenetics mechanisms.

We characterized a novel unique intragenic microduplication of the chromatin regulator *ARID1B* in an adolescent male with intellectual disability. Previously, duplications of ARID1B had been scarcely described. With this case, we increased the list of ARID1B duplications to a total of 3 cases reported thus far, where haploinsufficiency of this gene is thought to be the causal mechanism underlying the phenotype. This reinforces that ARID1B is dosage sensitive gene, highly constrained that does not tolerate loss-of-function mutations and its haploinsufficiency lead to an enrichment of transcription and cell cycle regulation pathways.

Here, we also demonstrated that it is possible to generate an allelic series of isogenic mutant iPSC from human-derived cells using CRISPR/Cas9 technology, to be able to assess the individual contribution of individual genes for ASD. These cells have the potential to be used in a number of downstream functional analyses, through the differentiation into the tissue of interest, that in our case is the neuronal lineage, to explore the role of those chromatin genes

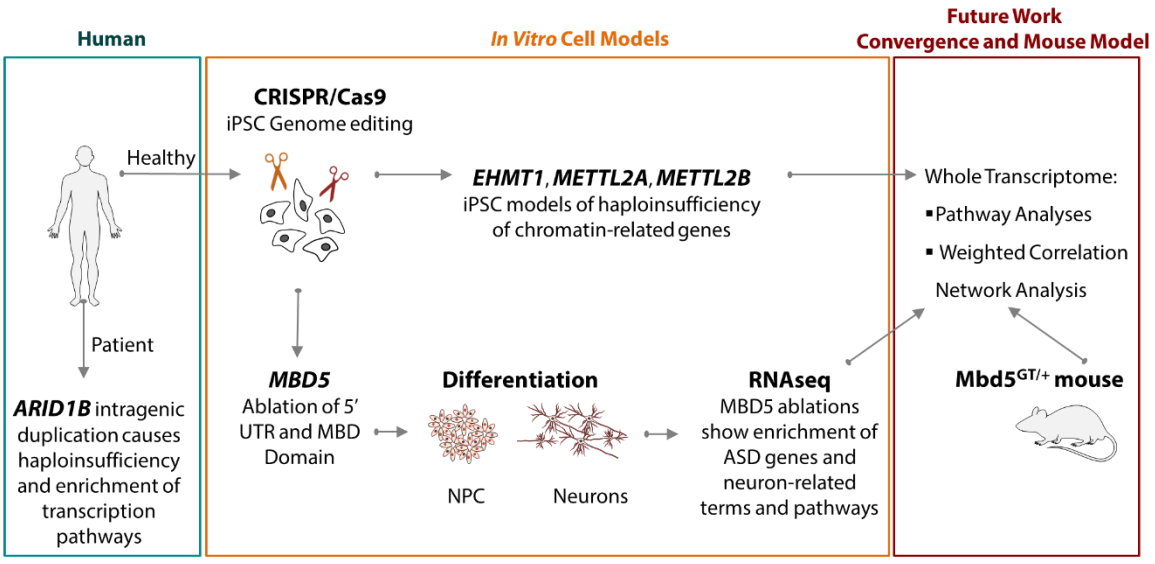


Figure 39 - Summary of this project, featuring current and future work.

(*EHMT1*, *MBD5*, *METTL2B*) in neurodevelopment. Indeed, we were able to pursue the *MBD5* cell lines and drove their differentiation into the intermediate neurodevelopmental stage of NPC and also into terminally differentiated mature neurons. We found that removing the 5' UTR exon 4 and the MBD domain in exon 6 did not affect the cells' ability to develop into mature neurons. We inspected the local *MBD5* transcript pattern upon the removal of both regions and found that the highest expressing transcripts did not only lack the MBD domain but contained only non-coding exons. This brings potential insight into the mechanism of action of MBD5 in neurodevelopment, suggesting that it may not involve the MBD domain, in contrast to what has been seen in the other MBD family members, and instead may rely on non-coding regulatory RNAs to regulate its own activity or that of other proteins.

On the other hand, genome-wide transcriptomic analysis allowed the identification of the dysregulated genes such as *RAB11FIP1*, *NHLH1-2*, *PLAUR* and *CNTNAP2*; and pathways such as notch signaling and cell adhesion that gave insight on the protein complexes and pathways that are acting downstream of *MBD5* and are directly implicated in neuronal development and function. Those represent promising targets for ASD therapeutics to be able to specifically aim for common and convergent biological pathways that are affected by *MBD5* and may be affected by other chromatin remodelers that represent a risk for ASD (such as *CHD8*).

In conclusion, this study was essential to add additional evidence to the functional roles of chromatin-remodeling genes previously associated with ASD and intellectual disability - *ARID1B* and *MBD5* - and to show the potential of human cellular models to study neurodevelopmental disorders. It is possible that the different chromatin-related genes studied are all be related through functional overlap and that would explain how the wide array of the chromatin-related ASD genes can trigger a set of symptoms with a high degree of phenotypic similarity, as found in patients with ASD. Using the cellular models we here propose, in additional studies in this direction, will be crucial to uncover the role of the remaining long list of genes associated with ASD. This way, we have contributed to the ASD field and proposed models to pursue in order to unveil details of the mechanisms by which chromatin remodelers confer risk for ASD that will ultimately lead to broadly effective therapeutic targets for ASD and other neurodevelopmental disorders.

Additional Preliminary Results

Mbd5$^{GT/+}$ Mouse Model

The generation of the Mbd5$^{GT/+}$ mouse model confirmed the causal role of *MBD5* in the 2q23.1 microdeletion syndrome and indicated that neuronal dysfunction is responsible for the observed phenotype (Camarena et al. 2014). This murine model carries an insertional mutation in intron 2 of *Mbd5* generated through gene-trap mutagenesis (Figure 40). This heterozygous hypomorph model recapitulates cardinal phenotypes of 2q23.1 microdeletion carriers including abnormal social behavior, cognitive impairment, and motor and craniofacial abnormalities (abnormal nasal bone). They are small, with reduced body weight, reduced neuromuscular strength and show motor deficits. In addition, neuronal cultures uncovered a deficiency in neurite outgrowth (Camarena et al. 2014). To compare the genome-wide expression changes of MBD5 haploinsufficiency in an animal model organism and in the human-derived CRISPR cellular models, we established a collaboration with the authors of the study Camarena et al., 2014. Tissue samples from cortex, cerebellum and striatum of the heterozygous *Mbd5*$^{GT/+}$ mice at 8-weeks were collected for subsequent RNA extraction (Table X).

The tissues to analyze were selected based on their potential relevance for ASD and the presence of *MBD5* expression. *Mbd5* expression in *Mbd5*$^{GT/+}$ mice was observed throughout the brain in high levels, but predominantly in some structures as the cortex, cerebellum and striatum (Camarena et al. 2014). A variety of clinical studies have reported cerebellar abnormalities in the brains of individuals with ASD, such as increased cerebellar activation during a motor task (Allen et al. 2004) and cerebellar hypoplasia in individuals (Courchesne et al. 1988). The most often reported cerebellar abnormality is a reduction in Purkinje cells, as demonstrated by post-mortem studies (Ritvo et al. 1986; Wegiel et al. 2013). Additionally, according to the GTEx database of human postmortem tissues (http://www.gtexportal.org/), *MBD5* shows the highest mRNA expression levels in cerebellum.

Regarding the striatum, several ASD risk genes have enriched expression in the striatum and have been shown to be important for striatal function. These include the forkhead box transcription factors *FOXP1* (Ferland et al. 2003; Tamura et al. 2004) and *FOXP2* (Takahashi et al. 2003) and the post-synaptic density scaffolding protein, *SHANK3* (Peça et al. 2011). The analysis of *MBD5*

knockdown in these tissues could elucidate upon its role in this structure that is mainly composed of GABAergic medium spiny neurons.

Methods

We thawed the tissues, that were previously frozen at -80ºC, overnight at -20ºC submerged in RNAlater®-ICE Frozen Tissue Transition Solution (ThermoFisher Scientific) in order to transition the frozen tissues to a state enabling easy cutting and extraction of high-quality RNA. RNA was obtained by lysing the tissue in 1mL of Trizol (Invitrogen) using metallic pellets (Qiagen) and a tissue lyser, then mixed with 1/5th volume of chloroform and centrifuged at 200xg for 5 minutes. The aqueous phase was collected and mixed with isopropanol and centrifuged to obtain a pellet that was then washed with 75% ethanol and then air dried and resuspended in RNAse-free water. Each tissue type was collected on the same day, to avoid batch effects. The RNAseq library was prepared using the Illumina TruSeq kit and manufacturer's instructions. Libraries were multiplexed, pooled and sequenced on multiple lanes of an Illumina HiSeq2500, generating an average of 30 million paired-end reads of 70 bp.

Quality assessment of sequence reads was performed using fastQC (v. 0.10.1 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Sequence reads were then aligned to human reference genome Ensembl GRCh37 (v. 71) using GSNAP (v. 12-19-2014) (Wu and Nacu, 2010) at its default parameter setting. Quality checking of alignments was assessed by a custom script utilizing Picard Tools (http://broadinstitute.github.io/picard/), RNASeQC (DeLuca et al., 2012), RSeQC (Wang et al., 2012) and SamTools (Li et al., 2009). Counts per

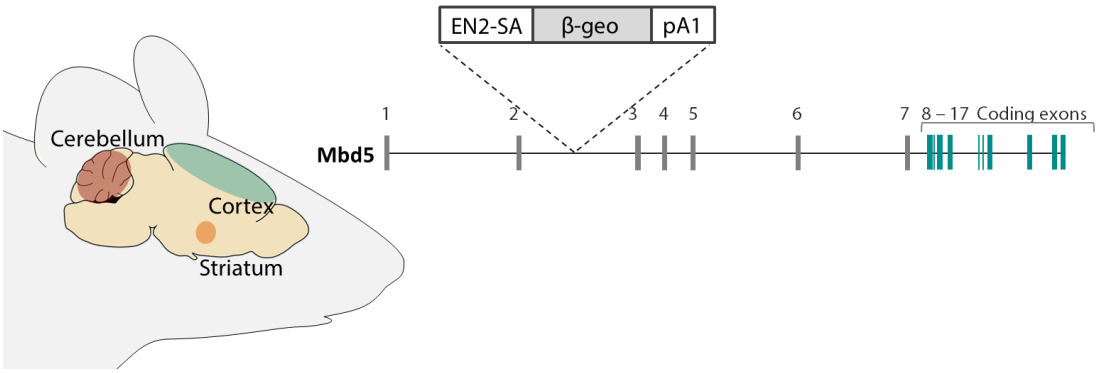


Figure 40 - Mbd5 mouse model characterization. Left - schematic of the brain regions collected for RNAseq analysis. Right - Mbd5 gene structure of 17 exons indicating the gene-trap cassette inserted within intron 2 (adapted from Camarena et al., 2014).

million (CPM) were generated from gene level counts which were tabulated using BedTools' multibamcov algorithm at all Ensembl genes (GRCh37 v.71) and HTseq (Anders et al. 2015) and normalization of counts with DESeq2 (Love et al. 2014) and by housekeeping genes (*ACTB* and *GAPDH*). Differential *Mbd5* transcript expression was assessed using RSEM & Bowtie2.

Preliminary Results

A PCA analysis of the RNAseq mouse dataset revealed that the two principal components that explain most variance separate the cerebellar samples from the remaining brain areas – striatum and cortex, which in turn cluster together. At this point, there is no visible segregation by genotype (Figure 41). Reduction in *Mbd5* expression, shown in the Camarena et al., 2014 et al paper, was recapitulated in all tissues, in variable expression ranges. The tissue showing the greatest difference in *Mbd5* expression in heterozygous mice in comparison to controls was the cortex, followed by cerebellum and striatum (Figure 42).

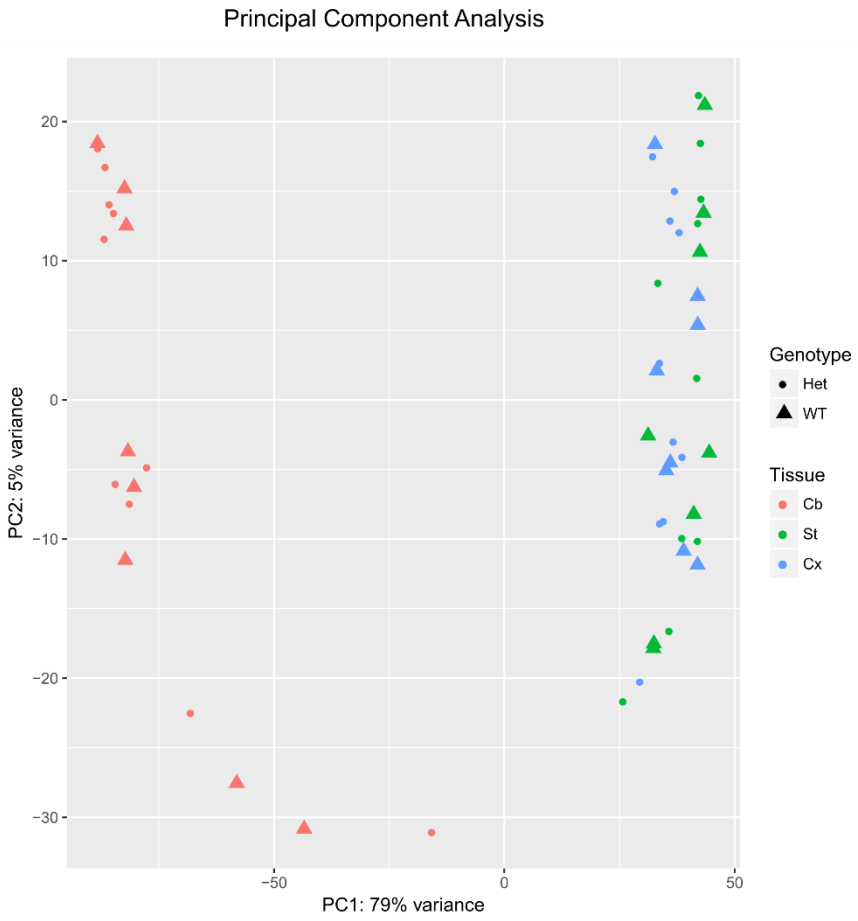


Figure 41 - PCA analysis of Mbd5$^{GT/+}$ mouse tissues.

Cell models lack circuitry organization and, therefore, converging lines of evidence, from mouse models and the cellular human models will help to define the cell types and brain regions that are critical in ASD. The joint analyses of this mouse model along with the human-derived CRISPR cell lines will also allow the comparison of the DEG and pathways to confirm the biological mechanisms involved in the pathogenesis of ASD by haploinsufficiency of Mbd5. A manuscript containing these and future results is currently under preparation.

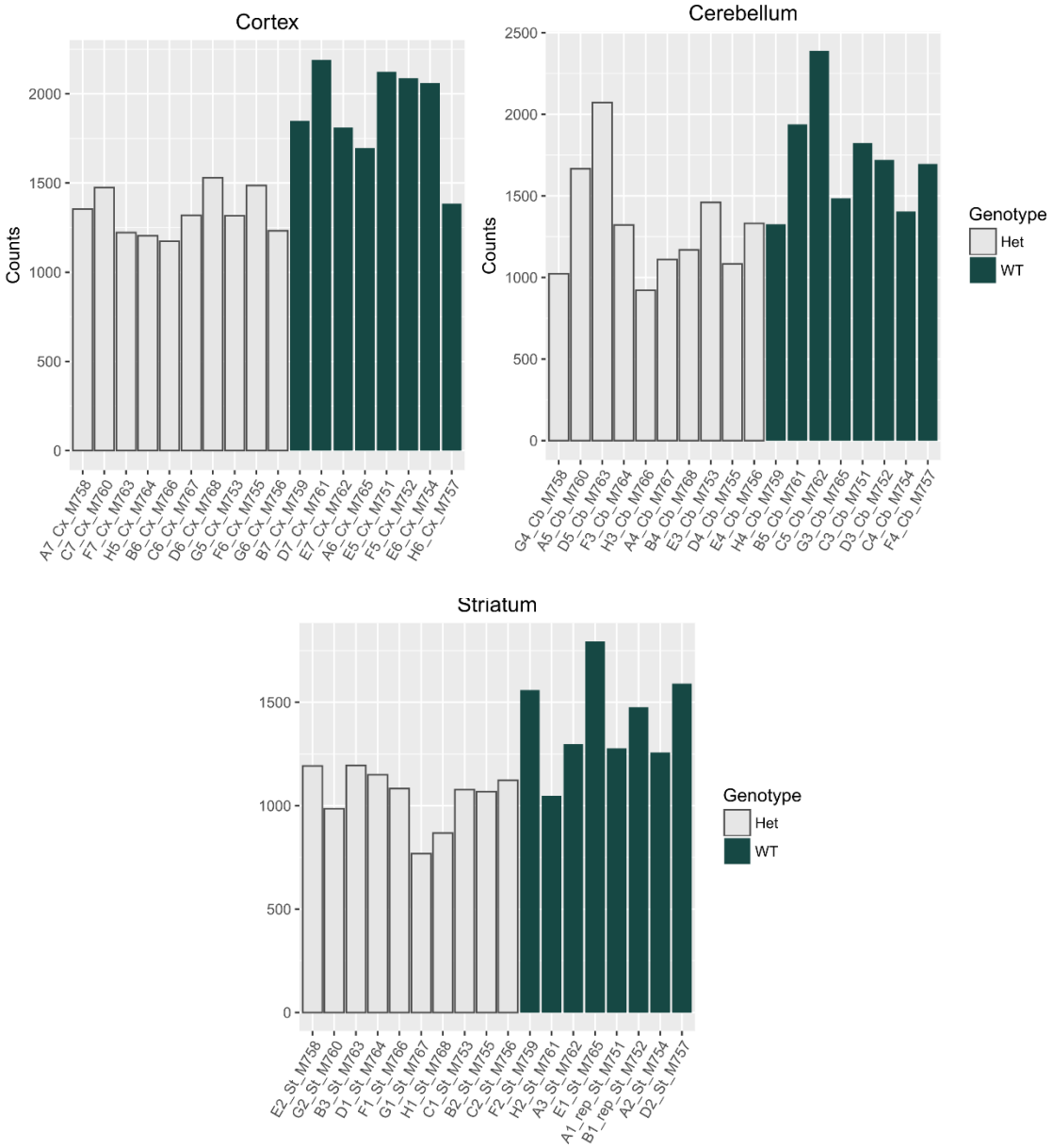


Figure 42 – Mbd5 expression assessed by DESeq2 in mouse samples.

Table X – Information of mouse samples collected.

ID	*Genotype*	*Gender*	*Litter*	*DOB*
M753	Mbd5$^{GT/+}$	M	1	12/31/2015
M764	Mbd5$^{GT/+}$	M	2	12/26/2016
M766	Mbd5$^{GT/+}$	M	2	12/26/2016
M767	Mbd5$^{GT/+}$	M	2	12/26/2016
M768	Mbd5$^{GT/+}$	M	2	12/26/2016
M755	Mbd5$^{GT/+}$	F	1	12/31/2015
M756	Mbd5$^{GT/+}$	F	1	12/31/2015
M758	Mbd5$^{GT/+}$	F	3	12/14/2015
M760	Mbd5$^{GT/+}$	F	3	12/14/2015
M763	Mbd5$^{GT/+}$	F	3	12/14/2015
M751	Wild-type	M	1	12/31/2015
M752	Wild-type	M	1	12/31/2015
M765	Wild-type	M	2	12/26/2016
M754	Wild-type	F	1	12/31/2015
M757	Wild-type	F	1	12/31/2015
M759	Wild-type	F	3	12/14/2015
M761	Wild-type	F	3	12/14/2015
M762	Wild-type	F	3	12/14/2015

References

Allen, G., Müller, R.-A. & Courchesne, E., 2004. Cerebellar function in autism: functional magnetic resonance image activation during a simple motor task. *Biological psychiatry*, 56(4), pp.269–78.

Anders, S., Pyl, P.T. & Huber, W., 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2), pp.166–169.

Camarena, V. et al., 2014. Disruption of Mbd5 in mice causes neuronal functional deficits and neurobehavioral abnormalities consistent with 2q23.1 microdeletion syndrome. *EMBO Molecular Medicine*, 6(8), pp.1003–15.

Chailangkarn, T. et al., 2016. A human neurodevelopmental model for Williams syndrome. *Nature*, In press(7616), pp.1–25.

Courchesne, E. et al., 1988. Hypoplasia of cerebellar vermal lobules VI and VII in autism. *The New England Journal of Medicine*, 318(21), pp.1349–54.

DeLuca, D.S. et al., 2012. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics (Oxford, England)*, 28(11), pp.1530–2.

Ferland, R.J. et al., 2003. Characterization of Foxp2 and Foxp1 mRNA and protein in the developing and mature brain. *The Journal of Comparative Neurology*, 460(2), pp.266–79.

Lancaster, M.A. & Knoblich, J.A., 2014. Generation of cerebral organoids from human pluripotent stem cells. *Nature Protocols*, 9(10), pp.2329–2340.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp.2078–2079.

Lin, L. et al., 2015. In Vitro Differentiation of Human Neural Progenitor Cells Into Striatal GABAergic Neurons. *STEM CELLS Translational Medicine*, 4(7), pp.775–788.

Liu, Y. et al., 2013. Directed differentiation of forebrain GABA interneurons from human pluripotent stem cells. *Nature Protocols*, 8(9), pp.1670–1679.

Love, M.I., Huber, W. & Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), p.550.

Malik, N. & Rao, M.S., 2013. A review of the methods for human iPSC derivation. *Methods in Molecular Biology (Clifton, N.J.)*, 997, pp.23–33.

Peça, J. et al., 2011. Shank3 mutant mice display autistic-like behaviours and striatal dysfunction. *Nature*, 472(7344), pp.437–442.

Ritvo, E.R. et al., 1986. Lower Purkinje cell counts in the cerebella of four autistic subjects: initial findings of the UCLA-NSAC Autopsy Research Report. *American Journal of Psychiatry*, 143(7), pp.862–866.

Rubenstein, J.L.R. & Merzenich, M.M., 2003. Model of autism: increased ratio of excitation/inhibition in key neural systems. *Genes, Brain, and Behavior*, 2(5), pp.255–67.

Sheridan, S.D. et al., 2011. Epigenetic Characterization of the FMR1 Gene and Aberrant Neurodevelopment in Human Induced Pluripotent Stem Cell Models of Fragile X Syndrome M. R. Cookson, ed. *PLoS ONE*, 6(10), p.e26203.

Takahashi, K. et al., 2003. Expression of Foxp2, a gene involved in speech and language, in the developing and adult striatum. *Journal of Neuroscience research*, 73(1), pp.61–72.

Talkowski, M.E. et al., 2012. Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell*, 149(3), pp.525–37.

Tamura, S. et al., 2004. Foxp1 gene expression in projection neurons of the mouse striatum. *Neuroscience*, 124(2), pp.261–7.

Temudo, T. & Maciel, P., 2002. [Rett's syndrome. Clinical features and advances in genetics]. *Revista de Neurologia*, 34 Suppl 1, pp.S54–8.

Vazin, T. et al., 2014. Efficient derivation of cortical glutamatergic neurons from human pluripotent stem cells: A model system to study neurotoxicity in Alzheimer's disease. *Neurobiology of Disease*, 62, pp.62–72.

Wang, L., Wang, S. & Li, W., 2012. RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, 28(16), pp.2184–2185.

Wegiel, J. et al., 2013. Contribution of olivofloccular circuitry developmental defects to atypical gaze in autism. *Brain Research*, 1512, pp.106–122.

Wu, T.D. & Nacu, S., 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics (Oxford, England)*, 26(7), pp.873–81.