Regulation of gene expression in human brain using transcriptome sequencing

Author:

Manuel Sebastian Guelfi

Supervisors:

Dr. Mina Ryten

Dr. Michael E. Weale

Prof. John Hardy

UCL Institute of Neurology

Department of Molecular Neuroscience

Thesis submitted in fulfilment of the degree of Doctor of Philosophy

February, 2019

Declaration

I, Manuel Sebastian Guelfi confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm this has been indicated in the thesis.

Publications arising from this thesis:

- **Guelfi S** & D'Sa K et al., (2018). Regulatory sites for known and novel splicing in human basal ganglia are enriched for disease-relevant information. In submission Nature Communications.
- **Guelfi S** & Zhang D et al. Gene regulation and complexity in human hippocampus. In preparation.

Publications not directly related to this thesis:

- Zhang D & **Guelfi S** et al. Misannotation of OMIM genes limits diagnostic yield from genetic tests, particularly for neurogenetic disorders. In advanced preparation.
- **Guelfi S** et al. Functional networks and genetic drivers implicating neuronal and glial mechanisms of intractable partial epilepsy. In review Brain.

- Salih D, Bayram S, **Guelfi S** et al., (2018). Genetic variability in response to Aβ deposition influences Alzheimer's risk. bioRxiv. doi: 10.1101/437657.

- Botia J, **Guelfi S**, et al., (2018). G2P: Using machine learning to understand and predict genes causing rare neurological disorders. bioRxiv. doi: 10.1101/288845.

- Botía JA, Vandrovcova J, Forabosco P, **Guelfi S**, et al., (2017). BMC Syst Biol. doi: 10.1186/s12918-017-0420-6.

- Murthy MN, Blauwendraat C; UKBEC, **Guelfi S**, et al., (2017). Increased brain expression of GPNMB is associated with genome wide significant risk for Parkinson's disease on chromosome 7p15.3. Neurogenetics. doi: 10.1007/s10048-017-0514-8.

- Ferrari R, Wang Y, Vandrovcova J, **Guelfi S**, et al., (2017). Genetic architecture of sporadic frontotemporal dementia and overlap with Alzheimer's and Parkinson's diseases. J Neurol Neurosurg Psychiatry. doi: 10.1136/jnnp-2016-314411.

- Ferrari R, Forabosco P, Vandrovcova J, Botía JA, **Guelfi S**, et al., (2016). Frontotemporal dementia: insights into the biological underpinnings of disease through gene co-expression network analysis. Mol Neurodegener. doi: 10.1186/s13024-016-0085-4.

- Matarin M, Salih DA, Yasvoina M, Cummings DM, **Guelfi S**, et al., (2015). A genome-wide gene-expression analysis and database in transgenic mice during development of amyloid or tau pathology. Cell Rep. doi: 10.1016/j.celrep.2014.12.041.

- Ramasamy A & Trabzuni D & **Guelfi S**, et al., 2014. Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat Neuroscience. doi: 10.1038/nn.3801.

Acknowledgement

I would like to express my gratitude to the people and organisations that have contributed directly to the completion of this manuscript.

Firstly, I would like to thank the Alzheirmer's Research UK for the finanicial support without which this PhD would not have been possible. I also would like to thank Dr Mark Cookson, Dr Raphael Gibbs and Dr Andrew Singleton for giving me the opportunity to visit and work at their laboratories. Warm thanks to my UCL and KCL colleagues, in particular those with whom I work or have worked closely; Hallgeir Jonvik, Dr Adaikalavan Ramasamy, Karishma D'Sa, Dr Jana Vandrovcova, Prof. Juan Botia, Dr Mar Matarin, Regina Reynolds and David Zhang. I could not have been luckier to have met such a delightful group of people. Thank you for your help, support and friendship.

Mostly, I owe gratitude to my supervisors; crossing their career paths has being one of the most enriching experiences of my life. I would like to thank Prof. John Hardy who gave me the opportunity to join his department and start this fantastic journey. I am very grateful to Dr. Michael Weale with who I had the privilege to start my PhD and benefit of his deep scientific knowledge; he has also been extremely helpful and time generous when revising this manuscript. Special thanks go to my principal supervisor and mentor, Dr. Mina Ryten; her constant support, enthusiasm and talent has been the best guidance I could have had – she is my true source of admiration.

I also would like to thank the people that, because of their emotional support, have

encouraged me to finish this manuscript. My close friends, Guillermo, Giacomo, Marco, Juan Camilo, Javier and Matteo. They have always been there for me in these years. My family, especially Abuela Nelly, Zii Francesco and Marisa; their wisdom has been my lighthouse. My Mamma, her humbleness, courage and positive thinking have been a priceless gift during the PhD. Babbo, who I discovered to be a good friend during this PhD and who consistently encouraged me to achieve this goal. Last but not least, I would like to thank Marta. Her constant love and care have accompanied me across this journey; I believe this achievement belongs to you as well as to me.

Abstract

Characterising the molecular mechanisms underlying disease risk variants identified in genome-wide association studies (GWAS) is of major interest. Expression Quantitative Trait Loci (eQTL) mapping studies provide a genome-wide characterisation of the impact of common genetic variation on gene expression and splicing and therefore have the potential to achieve this. In this thesis, I investigated the effect of common genetic variants in human brain through eQTL analysis. As part of the UK Brain Expression Consortium project, the analyses in this PhD thesis were performed on whole transcriptome RNA sequencing data from neuropathologically normal human post-mortem brain. I conducted eQTL analyses on putamen and substantia nigra using different types of quantification in order to interrogate regulation at different stages of RNA processing. This analysis pointed to splicing as an important process for the pathogenesis of Parkison's Disease. Thus, I identify not only disease-relevant regulatory loci but also the types of analyses yielding the most disease-specific information. Due to the limitations of current gene annotation and the complex transcriptomic landscape in human brain, I investigated transcription and splicing in the hippocampus using annotation-agnostic methods. This not only revealed the existence of widespread gene misannotation in the human brain, but also revealed the limitation of current quantification methods to capture transcriptome complexity in brain. Therefore, a reference-free eQTL analysis was performed and by testing for eQTL-GWAS co-localisation I found that incomplete annotation of the brain transcriptome limits the interpretation

of risk loci for neurological disorders. I anticipate that analyses of this kind will have an increasing impact on our understanding of a range of disorders, but are likely to have most impact on neurological and neuropsychiatric disorders because of the high transcriptome complexity of human brain tissue.

Impact Statement

In this thesis, I use expression quantitative trait loci analysis in post-mortem human brain to investigate the effect of genetic variation on gene expression. The aims of this analysis are to provide insights into molecular processes underpinning the association between common genetic variation and neurological and neuropsychiatric phenotypes, and in order to obtain insights into the landscape of genetic regulation of gene expression in human brain. The data generated in this thesis is an important resource for neuroscientists looking for potential target genes, which can be explored in detailed functional analyses both in vitro and in vivo. Additionally, by performing analyses with different biological process in mind this thesis identifies the regulation of splicing as an important mechanism in human brain disorders. In this way, the results of this thesis will aid the identification of the early pathophysiological processes underlying neurological disorders and I hope will ultimately help identify novel therapeutic targets for their prevention and treatment.

The identification of the genetic regulation of splicing as an important process in disease, led me to study transcription and splicing in a reference-free manner. This analysis revealed widespread gene misannotation in human brain. Both the analysis pipelines used in this thesis and the results generated are likely to be of interest to other researchers in the field. The pipeline I built, which uses read depth to define transcribed genomic regions and then integrates this information with reads spanning exon-exon junctions to determine evidence for splicing as well as the associated known

gene if applicable, is applicable to any RNA-seq dataset. The output of this pipeline as applied to hippocampal RNA-seq, could be used by geneticists interested in interpreting both common and rare disease-associated variants. The use of more precise exon and transcript definitions in gene screening panels, may reveal the presence of a pathogenic mutation within a gene or may change the interpretation of genetic variation within current genic boundaries. Therefore, it may increase the diagnostic yield of genetic testing in patients with rare disorders by changing which genomic regions are sequenced and how genetic variants are prioritised. Finally, it is worth noting that most of the data generated in this thesis is already publicly available through the web-based service interface Braineac.v2 (http://braineacv2.inf.um.es/) and a second resource is currently under development, which will incorporate all data. This to facilitate researchers to query in a user-friendly manner the results produced by this thesis.

Contents

Nomenclature

AD	Alzheimer's disease
ALS	Amyotrophic Lateral Sclerosis
Braineac	Brain eQTL Almanac
CAGE-Seq	Cap Analysis Gene Expression Sequencing
cDNA	Complementary DNA
DNA	Deoxyribonucleic acid
DZNE	German Center for Neurodegenerative Diseases
e-eQTLs	Exon eQTLs
eQTL	Expression Quantitative Trait Locus
ER	Expressed Region
ex-ex-eQTLs	Exon-exon junction eQTLs
FDR	False discovery rate
FTCX	Frontal Cortex
ge-eQTLs	Gene-exonic eQTLs
GEUVADIS	Genetic European Variation in Disease

gi-eQTLs	Gene-intronic eQTLs
GTEx	Genotype-Tissue Expression
GWAS	Genome-wide association studies
HD	Huntington's disease
HGP	Human Genome Project
HRC	Haplotype Reference Consortium
lincRNA	Long intergenic non-coding RNA
MCC	Mean Coverage Cut-off
MeSH	Medical Subject Heading
MRC	Medical Research Council
MRG	Max-Region-Gap
MS	Multiple sclerosis
NABEC	North-American Brain Expression Consortium
NGS	Next-generation sequencing
NIH	National Institutes of Health
OMIM	Mendelian Inherited in Man
PCA	Principal Component Analysis
PCR	Polimerase Chain Reaction
PEER	Probabilistic Estimation of Expression Residuals

polyA	Polyadenylated
PUTM	Putamen
RIN	RNA integrity number
RNA	Ribonucleic acid
RNA-Seq	RNA-Sequencing
SCZ	schizophrenia
SHRI	Sun Health Research Institute
SNIG	Substantia nigra
SNP	Single nucleotide polymorphism
SRA	Sequence Read Archive
STAR	Spliced Transcripts Alignment to a Reference
STAR	Spliced Transcripts Alignment to a Reference
STOPGAP	Systematic Target Opportunity assessment by Genetic Association Predictions
SVA	Surrogate Variable Analysis
TCGA	The Cancer Genome Atlas
TES	Transcription End Site
TF	Transcription factor
TPM	Transcripts Per Kilobase Million
TSS	Transcription Start Site

UKBEC	United Kingdom Brain Expression Consortium
WES	Whole Exome Sequencing
WGCNA	Weighted Gene Co-expression Network Analyses
WGS	Whole Genome Sequencing

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

The completion of the Human Genome Project (HGP) led to a new range of challenges - understanding the effect of genetic human variation is amongst the most important. In the last decade, chip-based genome-wide association studies (GWAS), and the rapid expansion of high-throughput sequencing technologies, have both contributed to the discovery of many associations between common genetic variants and observable phenotypic traits, ranging from facial characteristics (Adhikari et al. 2016) to risk of developing neurological disease (Nalls et al. 2014). However, the biological mechanisms underlying these associations are still poorly understood. Furthermore, as loci identified through GWAS often lie in intergenic regions surrounded by different genes, identifying the causal gene is not a trivial task. One approach to addressing this information gap is to consider the impact of genetic risk loci on the most proximal product of deoxyribonucleic acid (DNA), namely ribonucleic acid (RNA) (Cookson et al. 2009). Expression quantitative trait locus (eQTL) studies have been successfully used to investigate how genetic variation influences RNA expression levels. eQTLs can be used to ascertain the causal gene(s) within GWAS loci, which is the fundamental step to understanding the

cascade of effects leading to the trait. In this way, eQTL analysis adds an extra layer of functional information onto the human genome and is one part of a much broader effort to understand the regulatory (as opposed to coding functions of DNA), as exemplified by the collaborative projects ENCODE (The ENCODE Project Consortium 2004) and PsychENCODE (Akbarian et al. 2015).

The aim of eQTL studies is to identify significant associations between common DNA variants and a particular continous trait, in this case RNA expression levels. In order to identify eQTLs, a simple linear regression model is applied to test for correlation using the genotyped data as the independent variable and the RNA expression data as the dependent variable (Figure 1.1). This provides a p-value which indicates the significance of the association and a slope (β) which indicates the strength and direction of the effect, i.e. positive or negative association.

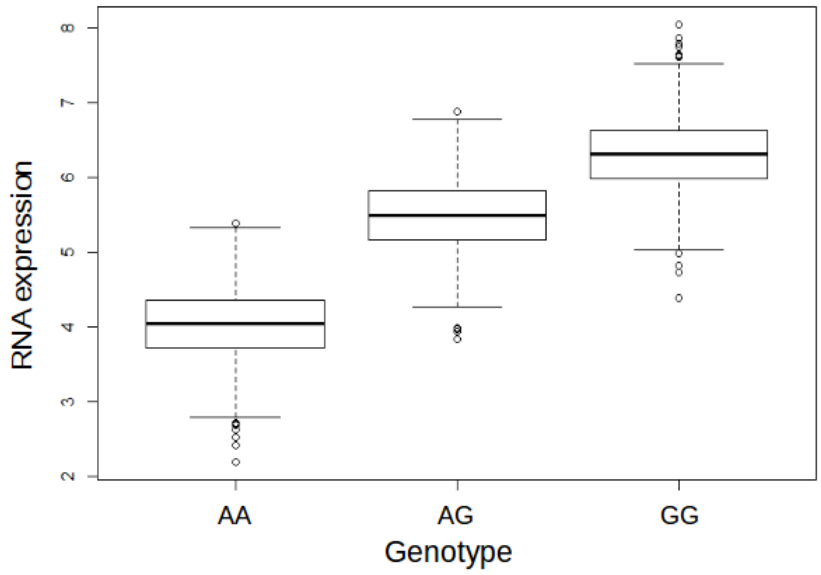

Figure 1.1: **Example of an eQTL plot.** The plot provides an example of a linear relationship between the genotype plotted on the X axis and the RNA expression levels plotted on the Y axis. In this example the dosage of the G allele is positively associated with the gene's expression.

The regulation of gene expression can occur at different points during RNA produc-

tion and processing, such as, transcription factor (TF) binding (McDaniell et al. 2010; Kasowski et al. 2010; Reddy et al. 2012), chromatin accessibility (Degner et al. 2012; Lee et al. 2013), alternative splicing (Montgomery et al. 2010; Pickrell et al. 2010; Li et al. 2016b) and mRNA degradation (Pai et al. 2012) amongst others. However, gene regulation is dictated primarely by TF binding. Transcription factors are proteins that bind to specific DNA sequences, called transcription factor binding sites, and recruit the RNA polymerase II to initiate transcription. Thus, when a variant is located in a transcription factor binding site this affects TF binding, histone alteration and mRNA expression. This has been reported in several studies performing eQTLs in humans as the primary mechanism driving cis-acting eQTLs (Brown et al. 2013; Westra and Franke 2014; Albert and Kruglyak 2015; Pai, Pritchard, and Gilad 2015).

The potential of eQTL analysis with regard to the understanding of regulatory processes as well as loci has increased with the widespread use of RNA-sequencing (RNA-Seq). The intrinsic properties of RNA-Seq allow us to analyse the entire transcriptome, generating a more complete landscape of RNA processing including the transcription of pre-mRNA (as detected by intronic reads), RNA splicing (indicated by exon-exon spanning junctions) and RNA stability (measured by steady state RNA levels (Gaidatzis et al. 2015)). Therefore, eQTL studies performed using RNA-Seq data have the potential to provide additional insights into the mechanisms driving gene expression changes (Pickrell et al. 2010).

Several studies have demonstrated the utility of eQTL analyses in providing insights into a wide range of disorders (Raj et al. 2014). However, the efficacy of brain-related eQTL analyses in the context of neurodegenerative disorders has been harder to demonstrate. This might be because some brain disorders have a systemic nature, affecting global biological processes rather than cell-specific processes. In fact, eQTLs analyses performed using monocytes appear to provide greater insights for Alzheimer's disease (Raj et al. 2014) than brain related eQTL studies. Another alternative explanation is

that the eQTLs of interest are present in brain, but because of current limitations in human brain eQTL studies they are not detected. The most common form of sample collection in brain eQTL studies is at post-mortem, this limits the RNA quality (Sidova et al. 2015) and the sample size. Furthermore, the human brain is a complex organ which consists of many anatomical regions each containing many distinct cell types and with extensive inter-regional differences in expression. Thus, the interpretation of transcriptomic data is complicated and the statistical power to detect regulatory signals is more limited. Finally, human brain shows high levels of splicing and non-coding RNA activity (Briggs et al. 2015), much of which has yet to be fully characterised (Jaffe et al. 2015; Clark et al. 2018) and this further limits eQTL analyses.

1.1.1 Expression quantitative trait loci in human brain: the annotated transcriptome

Given that many disease phenotypes manifest only in certain tissues, to increase the likelihood of finding eQTL and GWAS connections (Moffatt et al. 2007; Emilsson et al. 2008; Grundberg et al. 2012), it is widely believed that eQTL mapping studies are more likely to be successful when performed in tissues relevant to disease of interest.

Several eQTL mapping studies have already been performed in adult human brain tissues using microarray-based technologies. The focus of these studies has been the interpretation of GWAS (Zou et al. 2012; Kim et al. 2012), disease-specific eQTLs (Myers et al. 2007; Webster et al. 2009) and splicing eQTL (Heinzen et al. 2008; Ramasamy et al. 2014). Although these microarray-based eQTL studies significantly improved the understanding of genetic regulation in human brain, they were limited by the technology used. Microarray technologies rely on gene annotation for probe design, which means they rapidly become outdated, particularly with regard to newly identified RNA species (e.g. long non-coding RNAs (Wang, Gerstein, and Snyder 2009)). Furthermore, the use of microarray-based expression measures provides limited information about al-

ternative splicing (Malone et al. 2011; Zhao et al. 2014) and have a restricted dynamic range because of signal saturation (Hsiao et al. 2002).

Next-Generation Sequencing (NGS) technologies overcome many of these limitations and as costs have decreased, RNA-Seq-based eQTL studies in humans have become increasingly common (Montgomery et al. 2010; Pickrell et al. 2010; Lappalainen et al. 2013; Battle et al. 2014). However, eQTL studies using RNA-Seq in human brains are more limited. Fromer and colleagues (Fromer et al. 2016), have performed eQTL analysis on dorsolateral prefrontal cortex from a cohort of schizophrenia and healthy (N=467) individuals which were used to interpret risk loci identified from shizophenia GWAS. This lead to the identification of potential candidate genes for schizophrenia which were experimentally validated.

The only RNA-Seq-based eQTL analysis performed on multiple human brain tissues was performed by the Genotype-Tissue Expression (GTEx) Consortium (Ardlie et al. 2015). While an important resource, to date the GTEx study has focused on regulatory relationships across human tissues rather than regulatory mechanisms specific to brain, which might impact on neurological and neuropsychiatric conditions.

1.1.2 Expression quantitative trait loci in human brain: unannotated transcribed regions

A limitation of current eQTLs studies is that they have largely been performed using existing gene annotation to quantify RNA expression. Gene annotation databases are updated quarterly suggesting that gene annotation is not complete and implying that eQTL analyses progressively become more outdated. Although more comprehensive gene catalogues have been published (Pertea et al. 2018), the high degree of alternative splicing in human brain makes the annotation of brain-expressed genes to be disproportionately inaccurate and incomplete.

This view is supported by recent studies investigating gene expression in human

frontal cortex across development. In their study, Jaffe and colleagues (Jaffe et al. 2015) demonstrated significant differences between the transcriptome as defined by commonly used annotation providers and that defined empirically through RNA-Seq data. Further evidence is provided by a study performed using Long-Read Sequencing showing that the transcript isoform variation in human brain is more abundant than currently described in any gene annotation database (Clark et al. 2018).

Although limited, there are some eQTL analyses which have incorporated information on the regulation of unannotated transcribed regions. Pickrell et al. (Pickrell et al. 2010) used RNA-Seq data to show eQTL variants regulating unannotated non-coding regions and putative protein-coding exons. However, this analysis was performed in human lymphoblastoid cell line samples and not in human brain samples.

eQTL analysis using Cap Analysis Gene Expression Sequencing (Cage-Seq) data from control post-mortem frontal cortex samples has shown regulation of previously unannotated transcribed regions (Blauwendraat et al. 2016). Since this analysis was based on CAGE-Seq measurements, it did not have the potential to identify single nucleotide polymorphism (SNP) regulating intragenic novel splicing of novel 3' exons.

More recently, RNA-Seq data generated from human dorsalateral pre-frontal cortex was used to identify eQTLs targeting transcribed unannotated regions associated risk loci for schizophrenia (SCZ) (Jaffe et al. 2017a). However, these analyses were carried out using RNA-Seq performed on polyadenylated (PolyA) selected fragments, which might miss non-coding RNA species (Tariq et al. 2011) (the majority of which are not polyadenylated) and underestimate 3' UTR variability (Huang et al. 2011).

Long intergenic non-coding RNA (lincRNA) are thought to particularly important in generating tissue specificity in human brain (Derrien et al. 2012; Ward et al. 2015), yet are relatively poorly characterised and not necessarely polyadenylated. Given the regional and cellular specificity of many neurodegenerative disorders capturing this additional complexity might significantly improve the value of eQTL analyses in the

context of brain-related GWAS.

Improving the accuracy of the human brain transcriptome could also have an impact on the detection of rare pathogenic variants. Recent studies have demonstrated that incomplete annotation of alternative splicing may lead to false negatives when performing conventional mutation screening analysis (Gambino et al. 2015). Therefore, accurate gene annotation is an essential component of variant interpretation, and this is likely to be especially important for neurological diseases.

1.1.3 Objectives of this thesis

The first objective of my thesis was to perform eQTL mapping using RNA-seq data derived from post-mortem human putamen and substantia nigra samples originating from neurologically normal individuals with all analyses based on current gene annotation. Having generated eQTL data, in Chapter 3 I not only catalogue eQTL variants, but also aim to identify the molecular process eQTL variants are acting on and their relationship to brain-relevant GWAS hits.

The objective of the second portion of my thesis is to design a pipeline, and then perform annotation-agnostic eQTL mapping using human post-mortem hippocampus samples. As part of this pipeline I consider validation and characterisation of unannotated transcribed regions targeted by eQTL loci. I investigate the efficacy of my approach to eQTL mapping in terms of the generation of novel insights into neurological disorders. Finally, I use this project to provide a proof of principle for annotation-agnostic methods as a means to more effectively capture the full complexity of RNA-seq data.

Chapter 2

Data

2.1 Overview and rationale for tissues examined in this thesis

Substantia nigra (SNIG) and putamen (PUTM) are two of the key components of the subcortical region known as basal ganglia (Figure 2.1). The importance of the basal ganglia for normal brain function and behaviour is emphasised by the wide range of neurological conditions associated with basal ganglia dysfunction. These include: i) hypokinetic movement disorders, such as Parkinson's disease (Dawson 2007), which is characterised by degeneration of the dopamine-producing cells in the substantia nigra pars compacta, ii) hyperkinetic movement disorders, such as Huntington's disease (HD), which primarily involves dysfunction and loss of medium spiny neurons of the striatum (Albin, Young, and Penney 1989), and iii) neuropsychiatric disorders like schizophrenia (Williams et al. 2014). Furthermore, another brain region of clinical interest to human neurodegenerative disorders was analysed, the hippocampus. The hippocampus is part of the limbic system and dysfunction of this regions is a characteristic feature of several forms of dementia, such as Alzheimer's disease (AD), which manifests through formation of plaques and tangles in the hippocampus, disrupting cognitive functions like short-

term memory. Investigating the regulation of gene expression in these brain regions could provide insights into a wide range of neurological diseases.

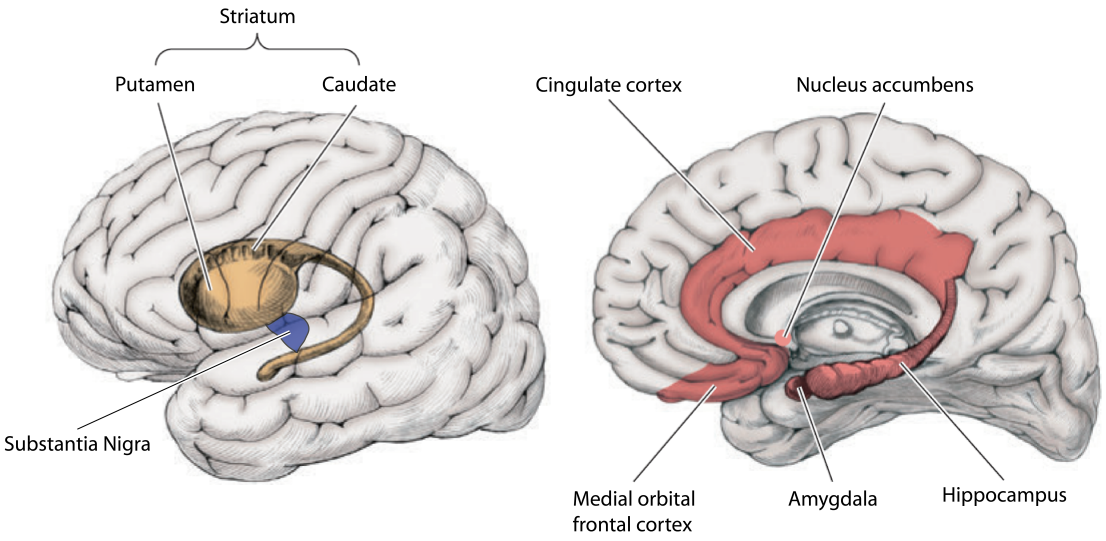

Figure 2.1: **Putamen, substantia nigra and hippocampus**. Figure adapted from figure 14.10 of Neuroanatomy through clinical cases, second edition.

2.2 Data generation

2.2.1 Samples

Human brain samples originating from 129 individuals of European descent were obtained from two donation programs: The Medical Research Council (MRC) Sudden Death Brain and Tissue Bank, Edinburgh, UK (Millar et al. 2007) and the Sun Health Research Institute (SHRI), USA (Beach et al. 2008). In all cases, brain samples originated from individuals with no history of neurological disease and which had been checked through detailed neuropathology before being classified as control samples. Table 2.1 presents a summary of the donor and sample characteristics used in the analyses.

Individuals	129
PUTM Samples	111
SNIG Samples	69
HIPP Samples	101
Age	Range: 16-102, mean: 57.0, median: 57, mode: 60
Gender	74.7% males - 25.3% females
Post-mortem interval	Range: 1-99 hours, mean: 43.8, median: 47, mode: 48
brain bank	84.3% Edinbrugh brain bank 15.7% The Sun Health Research Institute
Major cause of death	Ischemic heart disease (50.2%)

Table 2.1: **Summary of sample characteristics**

A total of 281 brain samples were dissected from three brain regions, namely putamen, substantia nigra and hippocampus. These samples constituted a subset of the United Kingdom Brain Expression Consortium (UKBEC) dataset (Ramasamy et al. 2014). All samples were authorised for ethically approved scientific investigation (Research Ethics Committee number 10/H0716/3) and had fully informed consent for retrieval. Macro-dissection and processing of brain samples was performed by Dr Mina Ryten and Dr Daniah Trabzuni. A detailed description of the sample processing was reported in Trabzuni et al. (Trabzuni et al. 2011). In brief, the miRNeasy 96 sample kit (Qiagen, UK) was used to isolate total RNA and the RNA integrity number (RIN) was assessed for each sample using the RNA 6000 Nano-LabChip kit (Agilent Technologies UK Ltd, UK).

2.2.2 Genotyping data

Macro-dissected samples of human post-mortem brain tissue (generally cerebellum samples) were used to extract DNA. The DNA was genotyped using two BeadChip platforms, the Illumina Infinium Omni1-Quad BeadChip and ImmunoChip (Nalls 2011). Analyses of the data and generation of the SNP calls was performed using GenomeStudio v.1.8.X (Illumina, USA).

Standard quality control on the merged genotyping data was performed. Individu-

als of suspected non-European descent and samples with the percentage of non-missing genotypes of <95% were removed from the analysis. Reported gender status and non-relatedness of samples were confirmed. Monomorphic SNPs, variants with missing position information, variants with a p-value <0.0001 for deviation from Hardy-Weinberg equilibrium, variants with a genotype call rate <95%, variants with less than two heterozygotes present and variants with mismatching alleles from 1000 Genomes Project were removed from the analysis. The genotyped data was used for imputation with the European panel of the 1000 Genomes Project (March 2012: Integrated Phase I haplotype release version 3) using MaCH (Li et al. 2010) and minimac (Howie et al. 2012). Finally, SNPs with allele frequency <5% were removed, resulting in ~5.88 million SNP and ~577 thousand indels for use in downstream analyses. Quality control and imputation of the genotype data was performed by Dr Adaikalavan Ramasamy.

2.2.3 Generation of RNA-Seq data

The work described in this paragraph was carried out by the commercial company AROS Applied Biotechnology A/S (Denmark).

Firstly, 100ng of total RNA was used as input for reverse transcription into complementary DNA (cDNA) using the Ovation RNA-seq system v2 kit (NuGEN, UK). Importantly, in this protocol reverse-transcription was performed using both oligo(dt) and random hexamer primers. This meant that the resulting cDNA could include RNA species, which do not undergo polyadenylation including some lincRNAs. The Covaris S220 Ultrasonicator was used to obtain cDNA fragments, which were used as input for the TruSeq DNA (Illumina, USA) library preparation kit. Adapter and barcode sequences were ligated to the fragments and ten cycles of polymerase chain reaction (PCR) were performed to amplify the cDNA molecules.

Lastly, the cDNA library was sequenced using the HiSeq v3 Flow cell (Illumina, USA) with three samples per lane to obtain an average of 145M paired end reads of

100bp long. Reads were demultiplexed and fastq files generated from the sequencing data using the CASAVA software (Illumina, USA).

2.3 Quality control of putamen and substantia nigra RNA-Seq data

RNA-Seq data was pre-processed to assess data quality.

2.3.1 Pre-alignment quality control

Fastq files contain the probability that a given base is read incorrectly by the sequencer, providing means of base call accuracy. Fastq files also contain adapter sequence, artificial sequences ligated to either end of the cDNA (cDNA) fragment, during library preparation. Trim Galore! (Krueger 2012) (v0.3.1) was used, which incorporates cutadapt (Martin 2011) (v1.2.1), to identify and remove adaptors and FASTQC (Andrews 2010) (v0.10.1) to assess data quality following trimming. Bases with base call accuracy $<= 99\%$ were removed (20 in the Phred scale, Figure 2.2) and all sequences at the 3' of the reads that overlapped with at least 1bp within the adapter sequences were removed.

2.3.2 Alignment

Paired end data was mapped to the human genome (build GRCh37) using tophat2 (Kim et al. 2013) (v2.0.9) with default settings and a transcriptome-guided approach using the Ensembl (Hubbard et al. 2002) reference (v72) based on GENCODE (Harrow et al. 2006) version 18. Reads mapping to mtRNA and rRNA regions were removed from the analysis.

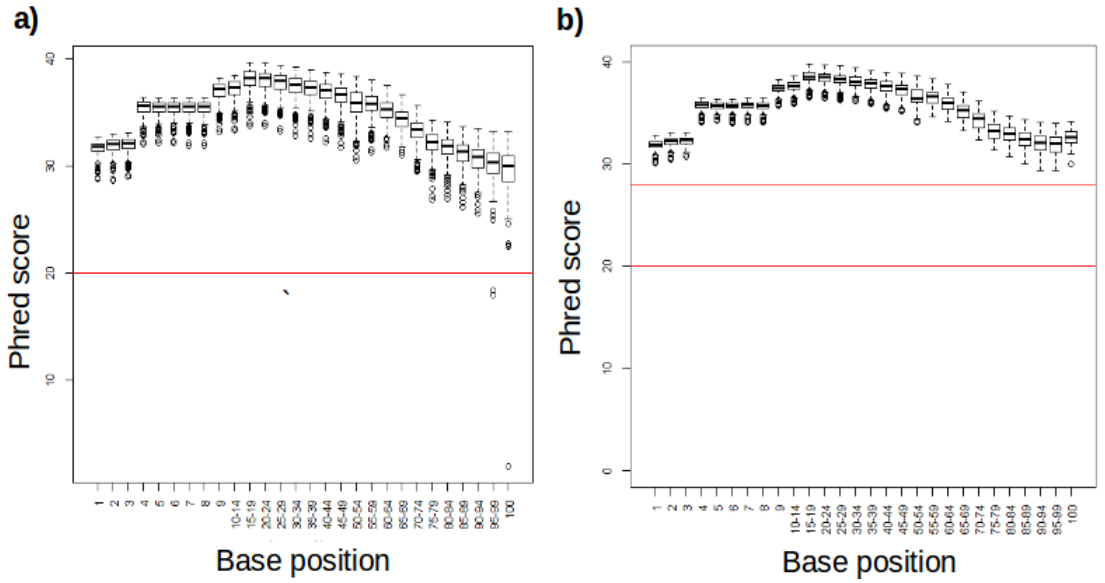

Figure 2.2: **Aggregated base call accuracy before and after base call quality filtering. a)** Aggregated (using means) base call accuracy before filtering. **b)** Aggregated (using means) base call accuracy after base call quality filtering.

2.3.3 Post-alignment quality control

Reads mapping to exonic regions and the number of genes detected were estimated using RNASeQC (Deluca et al. 2012) and formed quality metrics. Samples with less than 15M exonic reads and an exonic mapping rate of less than 10% were considered to have evidence of a failure of library costruction and were excluded entirely from further analysis. Additionally, samples with less than 20M exonic reads and less than 17K genes detected were selected for re-sequencing (Figure 2.3). A total of 17 samples were re-sequenced and processed as in section 3.3.1 and 3.3.2. Mapped sequence data was merged with previous sample data and RNA-SeQC was re-applied to assess the quality of the merged data (Figure 2.3). A total of 170 samples, consisting of 105 putamen and 65 substantia nigra samples, were used in all subsequent analyses.

The resulting aligned files were used as input for the work described in the following sections. The reads added and number of genes detected were further investigated with

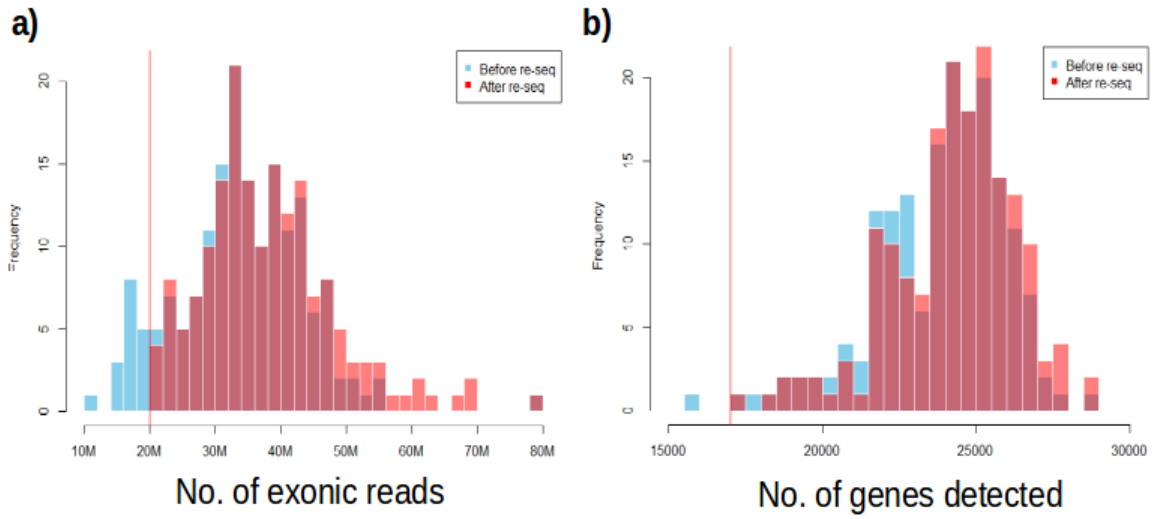

Figure 2.3: **Post-alignment quality control. a)** The Distribution of exonic reads across samples before and after re-sequencing. **b)** The distribution of genes detected across samples before and after re-sequencing.

the intention of assessing whether the number of genes detected had plateaued. Even after geneFastq filesrating 180M reads the number of genes detected does not appear to plateau and the data suggests a linear increase (Figure 2.4).

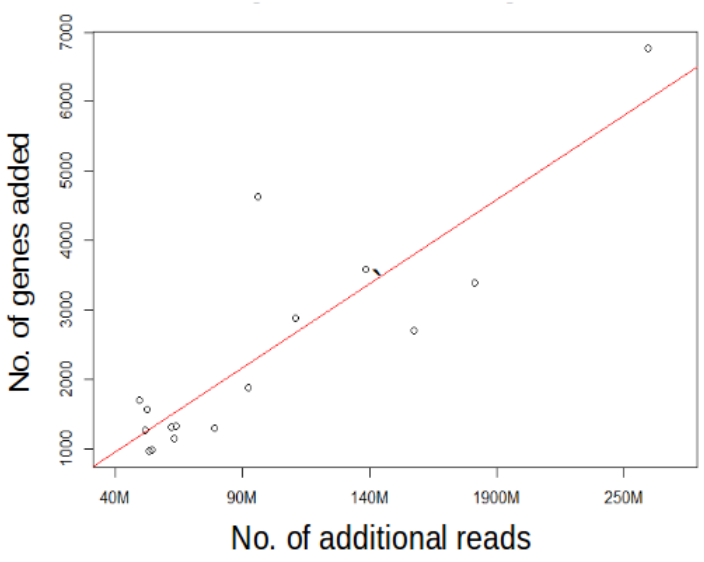

Figure 2.4: **Increment of number of genes detected per reads added.** Plot to show the effect of re-sequencing on gene detection

2.4 Quality control of hippocampus RNA-Seq data

The same sample quality controls as described above were performed on the 101 samples originating from the hippocampus. These quality checks were performed by Karishma D'Sa (PhD in the Department of Genetics and Molecular Medicine KCL) and resulted in the re-sequencing of 9 samples.

In the case of the hippocampal samples, I also performed further analyses to ensure sample identity rather than sequencing quality. I assessed sample data for a potential mismatch between the reported sex of brain donors and the sex as determined by the expression of sex-specific genes, namely *XIST* (X chromosome gene expressed) and *DDX3Y* (Y chromosome gene expressed). Expression in both genes highlighted 5 samples, of which 3 samples were reported as female while the expression suggested they originated from males and 2 samples reported as male, but expression suggested they originated from females. Since no labelling errors could be identified in the sample preparation and sequencing, these samples were removed from the final dataset. Additionally, I performed a PCA analysis to identify outliers on the basis of gene expression. By plotting the first two PC axes I detected a single outlier (Figure 2.5a). Consequently, I ranked and identified the top 10 genes contributing most to variability across samples. Amongst this list I identified several genes known to be markers of cerebellar tissue, including *SMPDL3B* (Bettencourt et al. 2014). Plotting the expression of *SMPDL3B* across all samples suggested that *A653_1278* outlying sample, originated from the cerebellum (Figure 2.5b). This sample was removed from the analysis. Thus, resulting in a final sample set of 95 hippocampus.

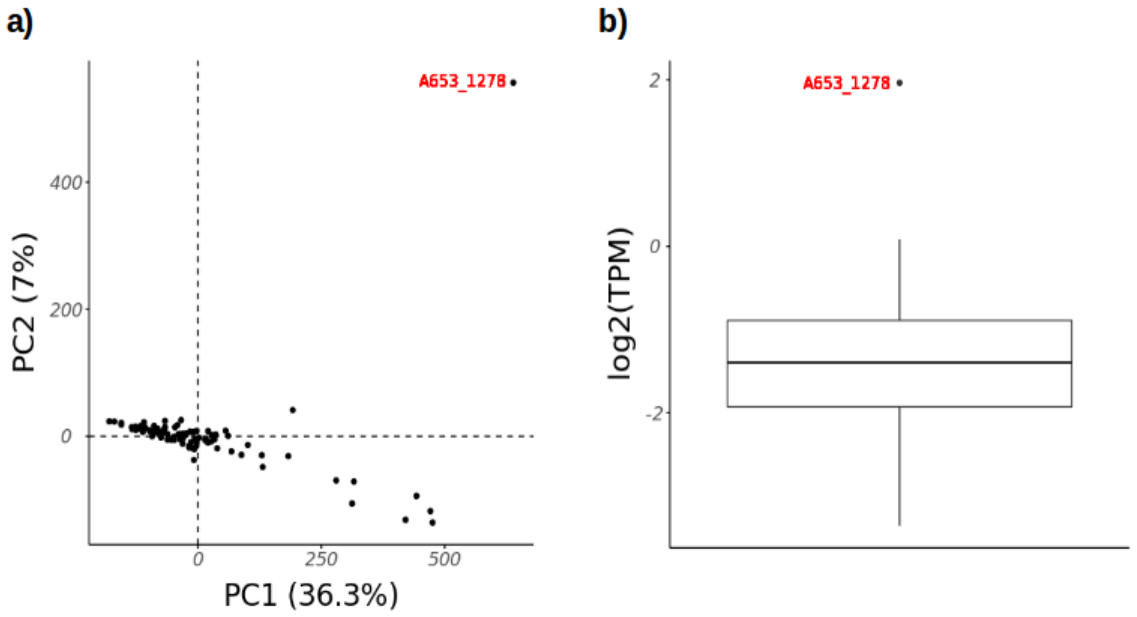

Figure 2.5: **Hippocampus quality control.** Identification of outliers through principal component analysis **(a)** and cell marker *SMPDL3B* gene expression in hippocampus **(b)**

2.5 External datasets

In order to validate the findings presented in this thesis several datasets were used. In choosing datasets I considered the importance of replication across gene expression profiling platforms and independent sample sets. Firstly, I sought to validate brain-expressed unannotated regions using public RNA-Seq datasets as well as curated databases. Secondly, validation of eQTLs was conducted across different quantification platforms, different RNA-Seq library preparations and different human tissues.

Validation of transcribed genomic regions

2.5.1 Recount2

To date Recount2 (Collado-Torres et al. 2017b) is the largest database of expression profiles with over 70,000 RNA-Seq samples from the Sequence Read Archive (SRA), The Cancer Genome Atlas (TCGA) and GTEx.

In this database all sample were processed through Rail-RNA (Nellore et al. 2016b) and derfinder (Collado-Torres et al. 2017a) to generate whole-transcriptome expression profiles for each sample. In this way, Recount2 can be used to provide RNA expression data over any pre-defined genomic region in GRCh38 coordinates. In my thesis, I specifically focused on data processed within Recount2 originating from the GTEx project (Ardlie et al. 2015Aguet et al. 2017). The data contained in recount2 correspond to the GTEx release version 6, which relates to samples dissected from non-diseased post-mortem tissues from 544 individuals covering 53 distinct tissues for a total of 9662 samples. Samples were sequenced to generate RNA-Seq reads from non-stranded 76bp long paired-end reads from polyA-selected libraries. Currently, this is the largest publicy available post-mortem human transcriptomics dataset.

2.5.2 CAGE-Seq datasets

CAGE-Seq (Kodzius et al. 2006) is a technique to measure transcription at the 5' end of capped transcripts. I used two CAGE-Seq datasets to collect more evidence for transcription of unannotated expressed regions. Firstly, a CAGE-Seq dataset containing over 200 control human substantia nigra samples generated at the German Center for Neurodegenerative Diseases (DZNE) and processed by Dr Sanchez-Simon. This dataset is not yet publicly available. Secondly, a public CAGE-Seq dataset of 128 control human frontal lobe samples provided by Blauwenraat and colleagues (Blauwendraat

et al. 2016).

2.5.3 North-American Brain Expression Consortium (NABEC) dataset

An RNA-Seq dataset provided by the North-American Brain Expression Consortium was used to more closely match the library preparation and sequencing technology of our core dataset. Using total RNA as input from 213 human postmortem frontal cortex control samples, the Illumina TruSeq Stranded Total RNA kit was used for library construction, a protocol which includes Ribo-Zero to enable ribosomal RNA reduction. Following cDNA library construction, paired-end reads of 101bp were generated using the Illumina Hi-Seq. Dr. Raphael Gibbs (National Institutes of Health, NIH) processed and checked Fastq files for quality. Sequences were aligned to the genome reference hg19/GRCh37 using the STAR aligner (2-step approach with default parameters). A final bam file dataset with an average of 66M mapped reads per sample (N=213) were provided by Dr. Gibbs and used as input for validation purposes.

2.5.4 NONCODE dataset

To assess whether unannotated transcribed regions were enriched for non-coding RNA, I used the NONCODE (Liu et al. 2005; Zhao et al. 2015) dataset. The NONCODE database contains human non-coding RNA genes identified from public literature (6532 published articles), as well as public gene databases such as, ENSEMBL, GENCODE, RefSeq and lncRNAdb. The collated dataset was then processed by the curators to remove redundancy, filter out protein-coding RNA and assign RNAs with potential functions. The NONCODE catalogue used for the validation analyses was downloaded on the 6th July 2016.

eQTL validation

2.5.5 The Brain eQTL Almanac (Braineac) dataset

To assess robustness, eQTLs were validated across quantification platform. As part of the UKBEC project, the Braineac dataset (Ramasamy et al. 2014) was generated using Affymetrix Human Exon 1.0 ST arrays to produce expression profiles for ten different human brain post-mortem tissues collected from 134 neurologically normal individuals of European ancestry. This is one of the publicly available human brian eQTL resourse.

2.5.6 Genetic European Variation in Disease (GEUVADIS) Consortium dataset

To test the eQTL brain specificity of eQTLs, I used a lymphoblastoid eQTL dataset generated with a large sample size (N=373). The aim of the GEUVADIS project (Lappalainen et al. 2013) was to characterise regulatory variation in different human populations. An average of 49M paired-end reads were generated from total RNA extracted from 465 lymphoblastoid cell lines collected as part of the 1000 Genomes project. Only eQTLs analysed from 373 individuals of European origin were used in this thesis for eQTL validation.

2.5.7 Genotype-tissues Expression Consortium dataset

The GTEx eQTL dataset was used to validate eQTLs across different library preparation protocols. The GTEx Consortium has generated the largest publicly available eQTL dataset for tissue-specific regulatory variants. Sample generation is described in section 2.5.1. I used the GTEx eQTL summary statistics v6.

GWAS interpretation

2.5.8 Systematic Target Opportunity assessment by Genetic Association Predictions (STOPGAP) database

To test for the enrichment of disease risk SNPs amongst eQTLs I used the STOPGAP (Shen et al. 2017) database. STOPGAP, encompass GWAS data from five different resources; GWAS Catalog (Welter et al. 2014), GRASP (Leslie, O'Donnell, and Johnson 2014), GWASdb (Li et al. 2016a), BioVU PheWAS (Denny et al. 2013) and risk loci identified in literature, but not present in the aforementioned resources. STOPGAP contains all trait associated variants with a p-value $\leqslant 10^{-4}$, as well as all variants in linkage disequilibrium (LD) with the associated variant ($r^2 \geq 0.7$ calculated from 1000G phase 1). The trait per variant is classified according to the Medical Subject Heading (MeSH) gathered from pubmed.

2.5.9 GWAS datasets

Colocalisation analyses between known GWAS association signals and eQTL signals were performed using several GWAS. Thus, full-summary statistics for PD (Nalls et al. 2018), AD (Lambert et al. 2013), SCZ (Pardiñas et al. 2018), Multiple Sclerosis (MS) (Beecham et al. 2013), HD progression (Moss et al. 2017), Amyotrophic Lateral Sclerosis (ALS) (Rheenen et al. 2016), intelligence (Savage et al. 2018).

A summary of the datasets analysed in this project either as a "core" dataset or for validation purposes are included in the table 2.2. Furthermore, GWAS datasets used for disease interpretation purposes are included in the table 2.3.

Source	Data types used	Description	Sample Sample	Published/ Unpublished
UKBEC Consortium	total RNA-seq	Multiple human brain tissues	69-111	Unpublished
UKBEC Consortium	Affymetrix array-based eQTL results	Multiple human brain tissues	134	Published
GTeX Consortium	polyA RNA-seq-based eQTL results	Multiple human tissues	82-100	Published
NABEC Consortium	total RNA-seq aligned sequences	Frontal cortex tissue brain tissues	213	Unpublished
Lappanainen (2013)	polyA RNAseq-based eQTL results	lymphoblastoid cell line	373	Published
Blauwendraat (2016)	CAGE-Seq	Human frontal lobe tissue	128	Published
DZNE - Tubingen	CAGE-Seq	Substantia nigra tissue	211	Unpublished

Table 2.2: **Datasets included in this project.**

Source	GWAS Description	Sample size N cases / N controls
Beecham, 2013	MS	14,498 / 24,091
Rheenen, 2016	ALS	12,577 / 23,475
Hensman-Moss, 2017	HD progression	1,989 / NA
Nalls, 2018	PD	37,688 (cases) 18618 (proxy-cases) / 1,417,791
Savage, 2018	Intelligence	269,858 / NA
Pardinas, 2018	SCZ	40,675 / 64,643
Lambert, 2013	AD	17,008 / 37,154

Table 2.3: **GWAS datasets included in this project.**

Chapter 3

Expression Quantitative Trait Loci

3.1 Introduction

This chapter describes the comprehensive set of eQTL analyses performed in two control, adult human brain tissues, putamen and substantia nigra. These analyses were performed in order to improve our understanding of the risk loci for neurological and neuropsychiatric disorders, including those identified for Parkinson's disease (Nalls et al. 2014) and schizophrenia (Ripke et al. 2014).

Mapping eQTL analysis in human brain are particularly challenging, the human brain is a very heterogenous organ composed of a large array of cell types with distinct properties, leading to a similarly distinct transcriptome profile. Brain-expressed genes, not only tend to be longer, but also have a higher proportion of alternative splicing events as compared to all other human tissues (Yeo et al. 2004). Furthermore, RNA-binding proteins show an enrichment of region-specific gene expression in human brain, suggesting more complex gene regulatory mechanisms (Mele et al. 2015). In addition, long non-coding RNAs have been proposed as mechanism to regulate genes (Wang et al. 2011; Guil and Esteller 2012; Engreitz et al. 2016) and the brain transcriptome has preferential long non-coding RNA activity (Briggs et al. 2015) with a significant portion

still uncharacterised. Thus, characterising gene regulation in brain at different RNA processing stages might reveal novel insights into neurological disorders, particularly for splicing which has been previously associated with neurologial disorders (Arnold et al. 2013; Chabot and Shkreta 2016). Furthermore, quantifying transcription outside of annotated genes might yield the identification of regulatory processes that could be important to understand neurological disorders.

Therefore, I performed eQTL analyses using two approaches. Firstly, I quantified transcription using a **reference based approach** (Figure 3.1). Several methods for measuring RNA abundance in annotated regions were implemented with the intention of providing insights into the different regulatory mechanisms. This resulted in four different quantification types and each was used to generate as many eQTL types. eQTL performed on gene-level quantification were split into those either accounting from reads that fell exclusively into intronic or exonic regions of a gene which were termed gi-eQTLs and ge-eQTLs respectively. eQTLs performed on single exons and exon-exon spanning junctions were named e-eQTLs and ex-ex eQTLs. These quantification types were designed to capture pre-mRNA, alternative splicing and steady-state mRNA regulation (Figure 3.2). Secondly, a **non-reference-based approach** (Figure 3.1) quantification was also implemented. In this way, it would be possible to overcome inaccuracies within the current annotations and to invesitigate genetic regulation of novel intergenic transcribed genomic regions in human brain (Figure 3.2).

While in principal a given eQTL locus could potentially regulate genes scattered across the entirety of the human genome, including genes located on different chromosomes, relatively few such eQTLs (termed trans eQTLs) have been identified (Small et al. 2011). eQTLs acting locally (termed cis-eQTLs) form the overwhelming majority of eQTLs identified to date in human tissue. This is likely to be due to the weak effects of such eQTLs and the high burden of multiple testing correction when including genome-wide variants. Given the relatively small numbers of putamen and substantia

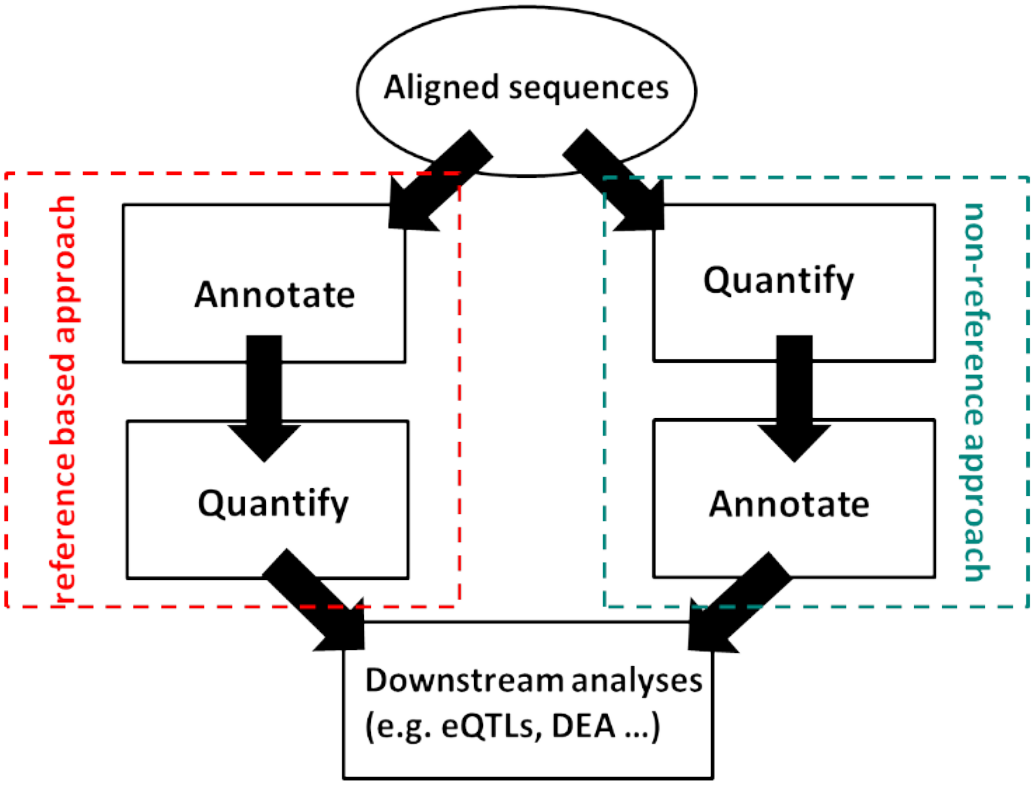

Figure 3.1: **Quantification approaches.** Flowchart to show the overview of the transcription quantification approaches.

nigra samples, I was underpowered for detection of trans-eQTLs and all the analyses presented were focused on the identification of cis-eQTLs (defined as loci within 1Mb of the transcription start or end site of their target gene).

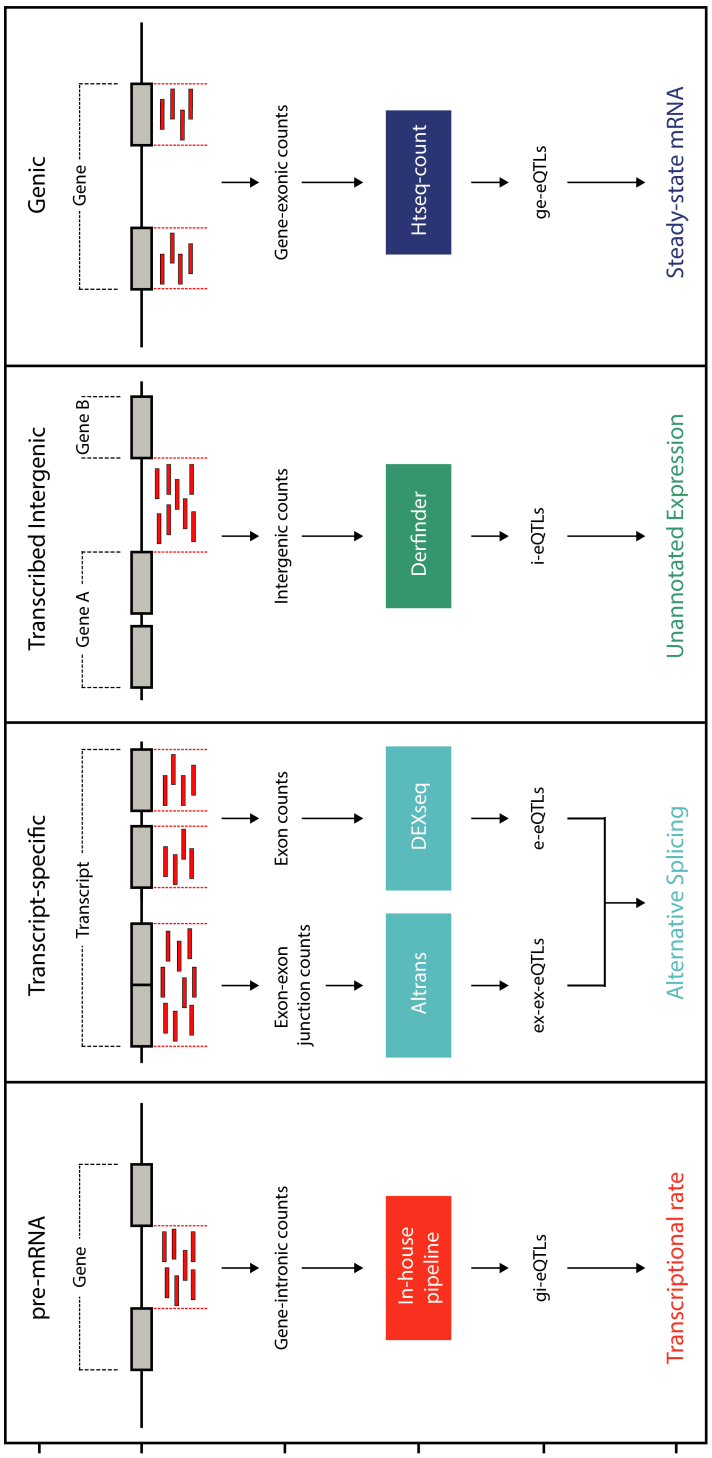

Figure 3.2: **eQTL quantification types**. Diagram to show the approaches used to quantify transcription and generate a range of eQTL classes which reflect different stages of RNA processing.

3.2 Methods

3.2.1 Reference-based quantification

3.2.1.1 Quantification of gene expression considering exonic regions alone

Gene counts were generated using the software HTSeq-Counts (Anders, Pyl, and Huber 2014) (v0.5.4p4) with the "intersection-nonempty" method and using the transcriptome definition from Ensembl (Hubbard et al. 2002) as a reference. This method resolves ambiguous reads by assigning them to the gene that is exclusively and entirely mapped. Moreover, duplicated reads and multi-mapped reads are excluded from this quantification. This approach was used with the unstranded parameter set "on".

3.2.1.2 Quantification of gene expression considering intronic regions alone

Gene counts including exonic and intronic regions were generated using the software bedtools (Quinlan and Hall 2010)(v2.24.0) and using the transcriptome definition from Ensembl (Hubbard et al. 2002)(v72) as a reference. Subsequently, exonic counts as calculated in section 3.2.1.1 were subtracted. Only genes that do not overlap with the genomic location of any other gene were considered. This was done to prevent the inclusion of reads mapping to the exonic regions of other overlapping genes.

3.2.1.3 Quantification of exon expression

Exon counts were estimated using the python DEXSeq (Anders, Reyes, and Huber 2012)(v1.10.6) protocol. This protocol first creates a meta-exon reference collapsing overlapping exons from different isoforms. Reads are then counted for each meta-exon. DEXSeq was performed using the paired-end and unstranded paramerters set "on".

3.2.1.4 Quantification of exon-exon spanning junctions

Reads mapping to exon-exon spanning junctions were estimated using the software Altrans (Ongen and Dermitzakis 2015)(v1.1.02). Altrans uses both paired-end read information and split reads to predict splicing events. Similarly to DEXSeq, Altrans creates an ad-hoc reference to group overlapping exons and identify unique portions. The parameters included in the Altrans analysis were as follows: i) only reads with a minimum read length of 95 bp were included (--read-length) to increase the sensitivity of detecting splice junctions and those over 95bp read long were trimmed to maintain consistency across reads (--trim), ii) unpaired reads (--check-proper-pairing) and reads which were aligned using soft-clipping (--no-clipping) were excluded, iii) an estimated fragment length (--min-exon-length) of 500bp, which is double that expected from our library preparation method, was used as suggested by the authors, iv) split read information (--split-reads) was required to be anchored by at least 2bp either side of splice junctions (--anchor-length), and v) overlapping exons were not quantified (--ignore-prob-groups). The Altrans-based quantification was performed by Karishma D'Sa and parameters were agreed together.

3.2.2 Non-reference-based quantification

3.2.2.1 Quantification of transcribed regions

Identification and quantification of unannotated transcribed regions, was performed using the derfinder (Collado-Torres et al. 2015) R package. Derfinder calculates the coverage at single base-pair resolution (Figure 3.3). Subsequently, it calculates the mean coverage across samples (Figure 3.3), which is then adjusted by the total number of mapping reads. Single base-pairs that did not pass the background-noise threshold (set at 5 reads by default) were filtered out (Figure 3.3) of any further analysis. Transcribed regions were assembled by taking contiguous reads allowing for a maximum gap of 10

bases (Figure 3.3). Any region of less than 100bp was excluded from further analysis. Expression of a region was defined for each sample as the mean coverage of the assembled regions. Intergenic genomic regions that were not annotated in either Ensembl (Hubbard et al. 2002) v72 or the University of California Santa Cruz (UCSC, Karolchik et al. 2003) databases (through the R library TxDb.Hsapiens.UCSC.hg19.knownGene version v3.1.2) were selected. Finally, the sequences of all identified regions were aligned to the human genome using blast (Altschul et al. 1990) and filtered out if they mapped to multiple locations with a minimum (98% alignment). This approach was applied to each tissue separately (putamen and substantia nigra) with the aim of identifying unannotated intergenic tissue-specific transcribed regions.

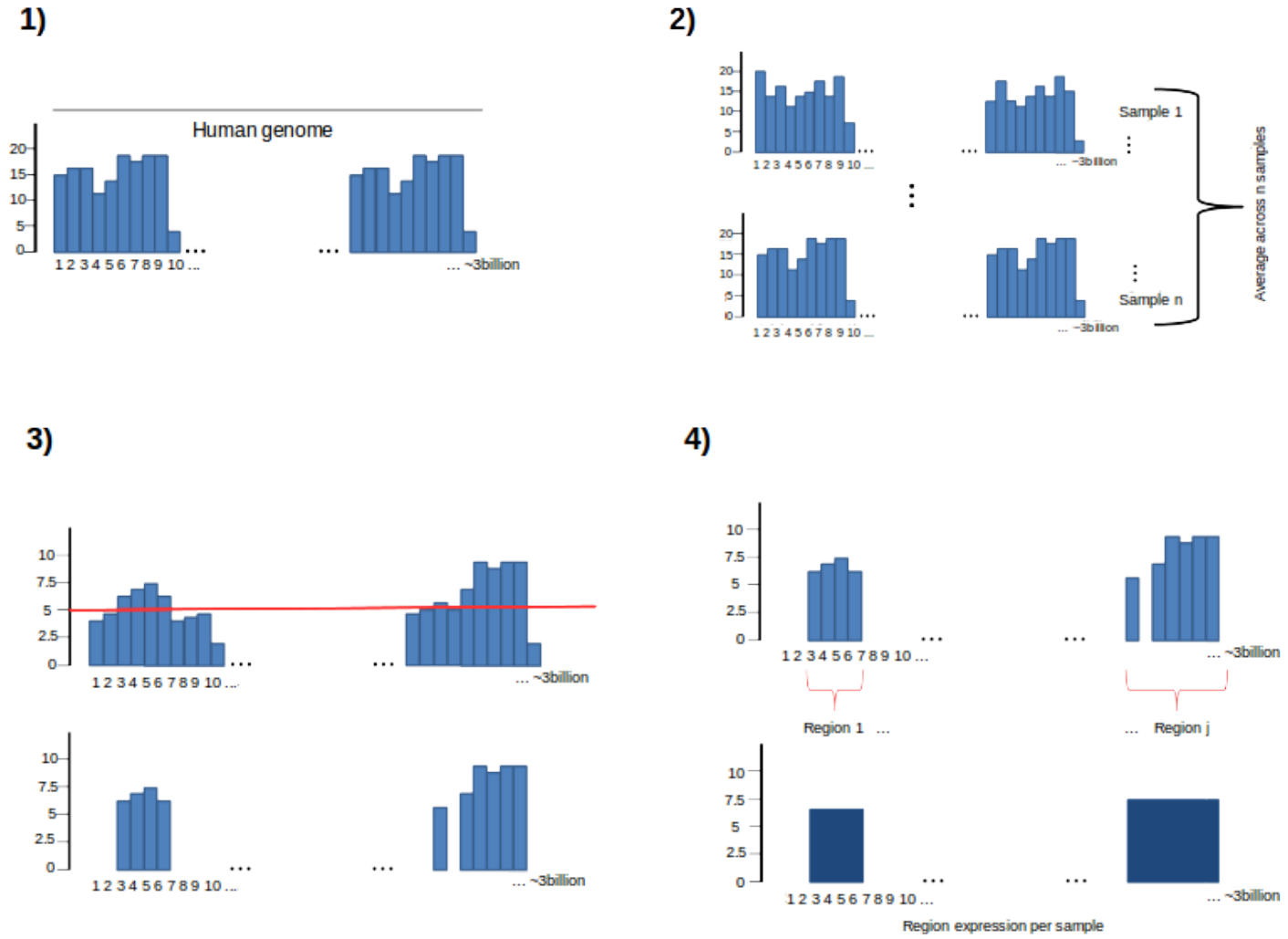

Figure 3.3: **Schema to explain the derfinder algorithm.** **1)** Transcription base-pair level coverage. **2)** Single base-pair mean coverage is calculated across samples. **3)** Cutoff is applied at single base-pair (represented by the horizontal red line) **4)** Regions are defined by unfying adjiacent base-pairs that pass the cutoff with a maximum gap of 10 base-pairs.

3.2.3 Identification of eQTLs

All forms of RNA quantification described above were processed through the same pipeline for eQTL discovery. Putamen and substantia nigra datasets were processed separately.

3.2.3.1 GC correction and normalisation

Pickrell *et al.* (Pickrell et al. 2010) first described guanine-cytosine content (GC-content) as a technical bias present within RNA-Seq data by showing selectivity in the sequencing of genes or exons stratified on the basis of GC-content. This techinical bias seems to be sample specific, and is present when comparing biological replicates (Hansen, Irizarry, and Wu 2012) (Figure 3.4). Therefore, this type of bias would be expected to have a major impact on eQTL detection. For this reason Conditional Quantile Normalisation (CQN,Hansen, Irizarry, and Wu 2012), which corrects per sample GC-content effects and quantile normalises across samples was applied to each quantification method. GC-content was calculated separately for each type of quantification and used as an input for the CQN R-bioconductor package. The resulting normalised expression data was then transformed into RPKM values and converted to a log2 scale.

Following GC correction and normalisation, principal component analysis (PCA) was carried out separately for each quantification type. In the case of gene-exonic or gene-intronic, the major source of variability tissue type (Figure3.5). However, for exon and exon-exon junction quantifications, demonstrated that tissue-specificity correlated best with the second axis. While I investigated all known factors, the most likely source of variability for axis 1 remains unexplained. Interestingly, similar analysis performed for transcribed intergenic regions demonstrated that the first axis was driven by tissue type (Figure 3.5). This suggested that this form of RNA quantification was unlikely to be background noise.

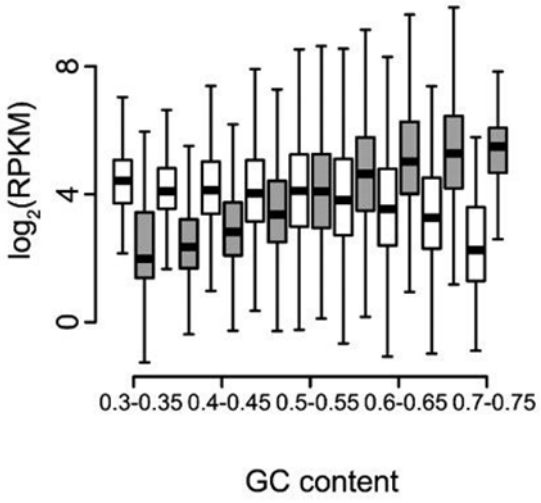

Figure 3.4: **Log2 RPKM expression stratified by GC-content**. White and grey colours represent different samples. Figure adapted from Hansen, Irizarry, and Wu 2012.

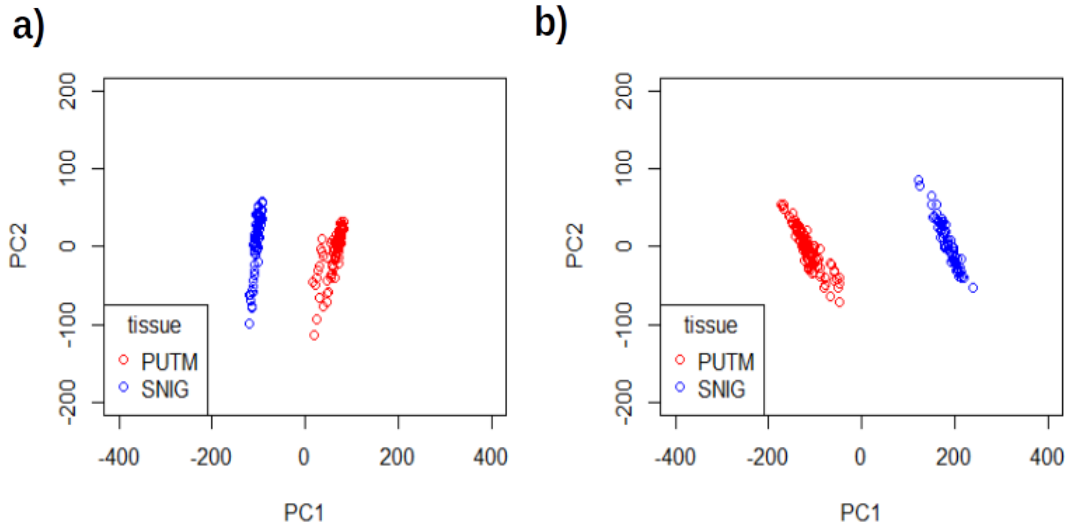

Figure 3.5: **Principal component plots**. Plots of first two principal components coloured by tissue (putamen: red circles and substantia nigra: blue circles) for gene-exonic **(a)** and transcribed intergenic **(b)** quantifications.

3.2.3.2 Removal of batch effects

Unknown factors (i.e. experimental hidden factors) can contribute to non-biological variability in RNA expression levels that can lead to under detection of eQTL signals

(Stegle et al. 2010Listgarten et al. 2010) and to spurious detections (Hyun, Ye, and Eskin 2008).

The Probabilistic Estimation of Expression Residuals (PEER, Stegle et al. 2010) method was used to identify unknown factors affecting expression levels and by correcting for these factors to increase the power of eQTL detection. For all forms of transcriptome quantification (exonic, intronic, and transcribed intergenic) with an RPKM of greater 0.1 in at least 80% of samples in a given tissue, CQN was applied and the resulting expression matrix was used together with gender and age as the input for PEER. PEER was run using default parameters and adding tissue type, age and gender as known covariates. Known characteristics of the samples were correlated with the output of PEER to obtain biological insights into the hidden factors identified. Although, the tissue type was included as covariate in the execution of PEER, one component in the output highly correlated with the tissue type, suggesting non-optimal convergence by PEER.

Thus, an optimisation step was performed which consisted of running PEER twenty times on the dataset, but using a random initialisation of the hidden factors rather than a PCA-based initialisation (as suggested by PEER's authors). The resulting output was further analysed to assess the validity of the optimisation. Correlation analyses were again computed between known characteristics and PEER components. I selected the initialisation procedure and resulting output which produced the minimum correlation to tissue type (based on Pearson's R^2) to obtain the final unknown factor matrix. To assess the validity of the optimisation, correlation analyses were again computed between known characteristics and PEER components. Although, correlation persisted, the maximum correlation diminished considerably to ~0.3 R^2 from ~0.7 R^2 (PCA-based initialisation).

RPKM normalised values were residual corrected using 13 unknown factors from PEER, which were selected because they maximise eQTL identification.

3.2.3.3 eQTL discovery

Variants within ∓ 1Mb span of annotated (i.e. genes, exons, exon-exon spanning junctions) and unannotated (i.e. intergenic regions) expression features were tested using the R package MatrixEQTL (Shabalin 2012) for the presence of an eQTL. In addition gender, age and the first three genetic principal component vectors were added as covariates. The genetic principal component vectors were added to reduce the intra-population variability and therefore increase the power of eQTL detection. The Benjamini–Hochberg method was applied to calculate the false discovery rate (FDR) adjusting for the number of tests performed for each transcriptomic feature.

3.2.3.4 Conditional analysis to obtain independent eQTLs

Stepwise conditional analysis was carried out for each quantification type to filter those SNPs that were in linkage disequilibrium (LD) and detect independent variant effects targeting the same expression feature. The method implemented was first suggested by Cordell and Clayton (Cordell and Clayton 2002) for the analysis of the effects of polymorphisms in the HLA region and involved the following steps: For each significant eQTL (FDR<5%) the linear regression test was re-run for that eQTL association adding the dosage of the significant signal as a covariate. If another eQTL variant had a significant effect and was independent of the previous locus, a second significant eQTL was identified. The conditional analysis was repeated until no significant eQTLs could be identified. This method was able to identify additional signals that have an independent effect on the same gene's expression.

3.2.4 Replication of eQTL signals in independent datasets

Three independent datasets were used to validate the eQTL results. eQTLs generated by Ramasamy et al (Ramasamy et al. 2014), which used the same donor sam-

ples (129 putamen samples and 101 substantia nigra samples), but a hybridisation-based RNA quantification method (namely; Affymetrix Exon arrays, Affymetrix, USA) were investigated in matching brain regions. eQTLs generated by the GTEx consortium (Ardlie et al. 2015), which used RNAseq to measure gene expression, though the library construction method differed (with GTEx using a poly(A)+ RNA-Seq library construction) was also analysed. In this case, eQTL results for putamen (N= 82), nucleus accumbens (N= 93) and caudate (N= 100) were downloaded from the GTEx portal website (http://www.gtexportal.org) on June 10th, 2015. Finally, eQTLs generated by Lappanainen and colleagues (Lappalainen et al. 2013) using 373 lymphoblastoid cell lines from the GEUVADIS consortium were analysed. The summary eQTL results for this dataset were downloaded from the EBI ArrayExpress repository (http://www.ebi.ac.uk/arrayexpress/files/E-GEUV-3/EUR373.gene.cis.FDR5.all.rs137.txt.gz - on the 15th of December 2015) with the accession number E-GEUV-3.

Only ge-eQTLs were considered for replication because either it was not feasible to match genomic locations of the quantification feature or the eQTL results using different quantification methods were not publicly available. Replication of a given ge-exonic eQTL within my study was achieved when the examined SNP-gene pair was significant at 5% FDR in both our dataset and the comparison dataset.

3.2.5 Beta-hetereogeneity testing across different forms of quantification

A mixed model approach was used to identify heterogeneity in eQTL signal strength (beta coefficient or slope), across different form of quantification. To test for beta heterogeneity between ge-eQTL and each other eQTL class (namely gi-eQTLs, e-eQTLs, ex-ex-eQTLs and relevant i-eQTLs) targeting the same gene, two models were fitted: 1) a non-heterogeneous (single slope) model with allele dosage as the main effect and two random effects, namely the index of the individuals and the index of the gene-exonic or

gene-intronic/exon/exon-exon-junction and 2) a heterogeneous (multiple slope) model containing the same terms, but with the addition of a fixed-effect allele dosage × of the gene-exonic or gene-intronic/exon/exon-exon-junction index interaction terms. To avoid any bias driven by the lack of expression, the test was only performed in exon or exon-exon-junctions considered expressed (as defined by as an RPKM of greater than 0.1 in at least 80% of samples). I used the lme4 R package to fit both models. A p-value for the likelihood ratio test comparing the two models was generated with the R function anova(). Finally, p-values were adjusted by the total number of eQTLs included in the beta-heterogeneity analysis using the Benjamini-Hochberg method.

3.2.6 Calculation of eQTL distance from the transcription start and end sites of the target gene

Following conditional analyses the location of each eQTL was calculated with respect to their target gene. These calculations were made taking into account stranding, and therefore eQTLs targeting unannotated transcribed intergenic regions, which did not have this information, were not included in this analysis. In the case of eQTLs located within their target gene, distances were calculated as a percentage proportional to the length of the target gene.

3.2.7 Functional annotation of eQTL signals

Functional annotation of eQTLs was performed using Ensembl Variant Effect Predictor (VEP,McLaren et al. 2016). This analysis was performed with reference to the target gene. If the eQTL was located outside the target gene, it was assigned as up-stream or down-stream as appropriate. If the eQTL was located within the boundaries of the target gene, its annotation was assigned in a conventional manner.

3.2.8 Assessment of brain cell type-specificity of eQTL targets

The RNA-Seq data generated for substantia nigra and putamen samples was used to generate gene co-expression networks analyses and assess the cell type-specificity of eQTL targets. Counts of genes detected at RPKM>0.1 in at least 70% of samples were CQN normalised (as in section 3.2.3.1) and residual corrected for 13 PEER axes (identified as in section 3.2.3.2), gender, age and genetic axes before applying K-means optimised weighted Gene Co-expression Network Analyses (WGCNA) (Botía et al. 2017). Annotation of cell-specific modules was performed using the augmented list of cell-specific gene markers from the userListEnrichment present in the WGCNA R Package (Lein et al. 2007; Oldham et al. 2008; Cahoy et al. 2008; Winden et al. 2009).

Genes assigned to modules significantly enriched for brain-related cell type markers and with a module membership of > 0.3 were allocated a cell type "label" of neuron, microglia, astrocyte, oligodendrocyte and endothelial cell.

Subsequently, each eQTL targeting a known genic region was assigned the same cell-type of the target gene assuming the target gene was a member of a cell type-labelled module. Similarly, for eQTLs targeting unannotated transcribed intergenic regions with moderate and high evidence of being associated to a known gene, the eQTL was assigned to the cell type of the associated gene. For eQTLs targeting unannotated transcribed intergenic regions with low evidence for association with a known gene or which could not be classified, the cell-type was assigned based on its highest module membership (correlation of the expression of the transcribed intergenic region and the first principal component or eigengene of each module), provided the module membership was at least 0.3. Finally, a Fisher's Exact test was applied to test for enrichment of cell type specific information within a given eQTL class. This analysis were carried by Dr Juan A. Botía.

3.2.9 Investigation of GWAS risk variants

3.2.9.1 Enrichment of eQTLs amongst GWAS risk loci for neurological and neuropsychiatric disorders

I investigated the enrichment of eQTL loci and risk loci identified through GWAS and sub-classified by the STOPGAP database. eQTL-GWAS overlap was checked using all eQTLs passing an FDR $< 5\%$. A Fisher's exact test was computed for the eQTLs overlapping GWAS hits classified as "Neurological/behavioral" relative to eQTLs overlapping GWAS hits classified for all other disorders (Aging, Blood, Cardiovascular, Digestive system, Ear, Endocrine, Eye, Infection, Inflammation, Liver & kidney, Metabolic, Miscellaneous, Musculoskeletal, Oncology, Respiratory, Skin & connective tissue, Urogenital). Disorders classification for GWAS hits was performed by STOPGAP using the Medical Subject Heading ontology.

3.3 Results

3.3.1 eQTL signal detection

Association testing was performed between ~6.5 million genetic loci (~5.88M SNPs and ~577K indels - collectively referred as variants) and ~411,000 RNA expression traits in putamen and ~370,000 RNA expression traits in substantia nigra resulting in ~5.3billion eQTL tests. A false discovery rate (FDR) is applied to the p-value to control for the total number of variants tested in $\mp$ 1Mb range of the RNA expression trait. Following conditional analysis and using a FDR of 5%, 21,955 eQTLs were identified in total.

A summary table with the number of signals detected using different quantifications by brain region is presented in Table 3.1. This table also includes the number of independent secondary eQTLs identified following stepwise conditional analysis. The number of secondary independent eQTLs signals identified in this study is small and is

likely to be understimated because detection of secondary effects requires larger sample sizes. While there is a substantial difference in the number of eQTLs detected between the brain regions, this is most likely to be due to the difference in sample size, with putamen having almost double the number of samples available for analysis as compared to substantia nigra.

Tissue	**Quantification**	**Features included**	**eQTL signals**
Putamen (N=105)	Gene-exonic	*19,039*	1,235 (23 ind.sign)
	Gene-intronic	*4,713*	348 (7 ind.sign)
	Exon	*275,703*	8,946 (176 ind.sign)
	Exon-exon Junction	*96,738*	2,180 (52 ind.sign)
	Transcribed intergenic	*14,905*	1,042 (26 ind.sign)
Substantia Nigra (N=65)	Gene-exonic	*17,499*	650 (13 ind.sign)
	Gene-intronic	*4,331*	178 (2 ind.sign)
	Exon	*251,575*	5,436 (87 ind.sign)
	Exon-exon Junction	*88,936*	1,619 (24 ind.sign)
	Transcribed intergenic	*8,135*	314 (13 ind.sign)

Table 3.1: **Summary of significant eQTL at 5% FDR by quantification type.** The term "ind.sign" indicates the number of additional eQTLs which have an independent secondary effect on a gene.

I detected 1,885 gene-exonic eQTLs, corresponding to 1,377 SNPs and 1,273 genes. Thus, eQTLs were identified for 5.1% of the total number of genes tested in this category. This included 36 gene-exonic eQTLs detected through stepwise conditional analyses. Fewer gene-intronic eQTL signals were identified (526 in total), due to the lower number of genes tested. However, the yield of this approach, defined as for both tissues as the total number of eQTL identified divided by the total number of features tested within the same category, remained similar (5.8%).

The largest number of eQTL signals were generated using exon-level quantification. However, this was due to a large proportion of eQTLs tagging exons from the same gene. In fact, 4,943 (out of 8,946) and 2,490 (out of 5,436) eQTL signals targeting exons, respectively for putamen and substantia nigra, shared a target gene.

Exon-exon junction quantification generated 3,799 eQTLs, corresponding to a yield of 2.0%, similar to the yield obtained using exon quantification (2.9%). Finally, testing intergenic transcribed regions had the highest yield with 1,356 eQTLs detected corresponding to 5.9% of all transcribed intergenic regions tested.

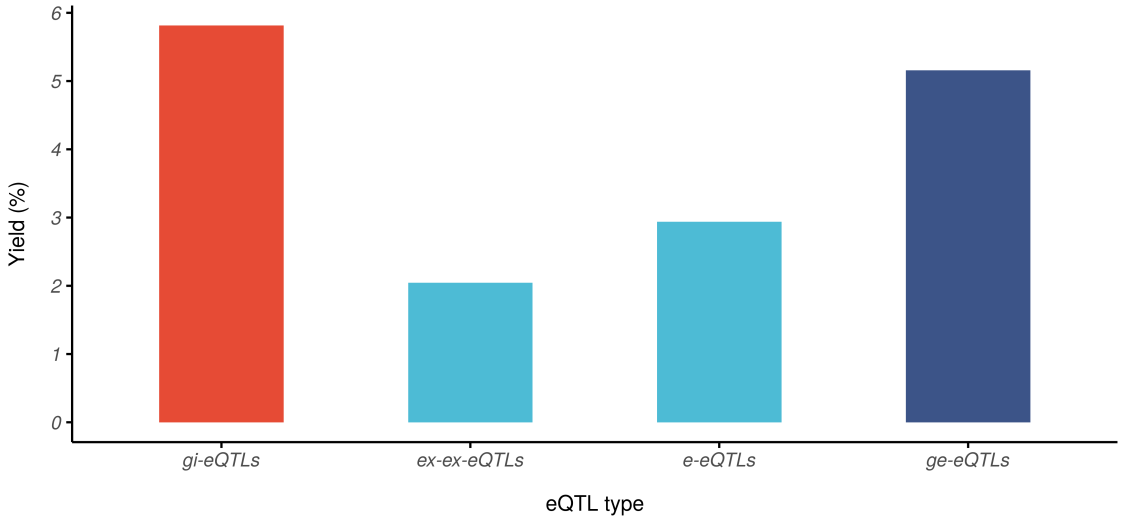

Figure 3.6: **eQTL yields of per quantification type.** Bar chart to show the eQTL yields for both tissues. Yields were calculated as the number of expression features within a category divided by the total number of tested features within the same category.

3.3.2 eQTL signals show high replication rates across platforms and datasets

To determine the quality of the cis-eQTL pipeline, eQTLs detected in this analysis were compared to those identified in putamen (N=129) and substantia nigra (N=101) by Ramasamy et al (Ramasamy et al. 2014). Using this dataset, I found that 50.6% and 50.4% respectively of eQTLs identified using the RNA-Seq-based analysis of putamen and substantia nigra could be detected within the microarray dataset. This replication rate is high considering replication rates across previous eQTL mapping studies (McKenzie et al. 2014; Ardlie et al. 2015), but given that both studies shared the

same biological material it might have been expected to be even higher. This could be attributed to the different technologies used for RNA quantification (sequencing-based versus hybridisation), different analysis pipelines and smaller samples size for the RNA-Seq study.

Replication was also assessed using the RNA-Seq-based eQTLs generated by the GTEx consortium (Ardlie et al. 2015). This demonstrated a replication rate of 38.1% for eQTL signals in putamen and 59.9% for substantia nigra.

Finally, I checked the replication of eQTL signals in the RNA-Seq eQTL study reported by Lappalainen et al. 2013. Using this data set I found that 22.0% and 24.2% of eQTLs identified in putamen and substantia nigra respectively were replicated. Given the significantly larger number of samples in the Lappalainen study (N=373), the higher replication rate of eQTLs within the GTEx datasets which are derived from similar brain regions, emphasizes the importance of validating eQTLs within similar tissues/cell sample datasets and suggests that many of the eQTLs identified are tissue-specific.

3.3.3 Characterisation of eQTL signals

3.3.3.1 Identification of splicing-specific eQTLs

Exon and exon-exon junction quantification has the potential to provide information regarding alternative splicing and could be used to identify transcript-specific eQTLs. However, this is complicated by the usage of a given exon (and corresponding exon-exon junctions within multiple transcripts). Therefore, in order to better determine whether any of the exon and exon-exon junction eQTLs were related to transcript-specific rather than gene level regulation we used beta heterogeneity testing across the gene, as previously described by Ramasamy and colleagues (Ramasamy et al. 2014). The beta-heterogeneity test differentiates between eQTLs that regulate single exons or exon-exon junctions from those that regulate the overall expression of genes indicating

possible alternative splicing regulation. The advantage of this approach is that beta estimation at any given exon or exon-exon junction is independent of power.

I used beta-heterogeneity testing to determine whether the exon and exon-exon eQTL (e-eQTLs and ex-ex-eQTLs) identified were distinct from more standard gene-level eQTLs (ge-eQTLs). This analysis showed that only 13.2% and 6.9% e-QTLs and ex-ex-eQTL were detecteble amongst ge-eQTL (Figure 3.7) suggesting not only that e-eQTL and ex-ex-eQTL are likely to be driven by splicing effects, but that these eQTL classes provide distinct regulatory information as compared to standard gene-level analysis.

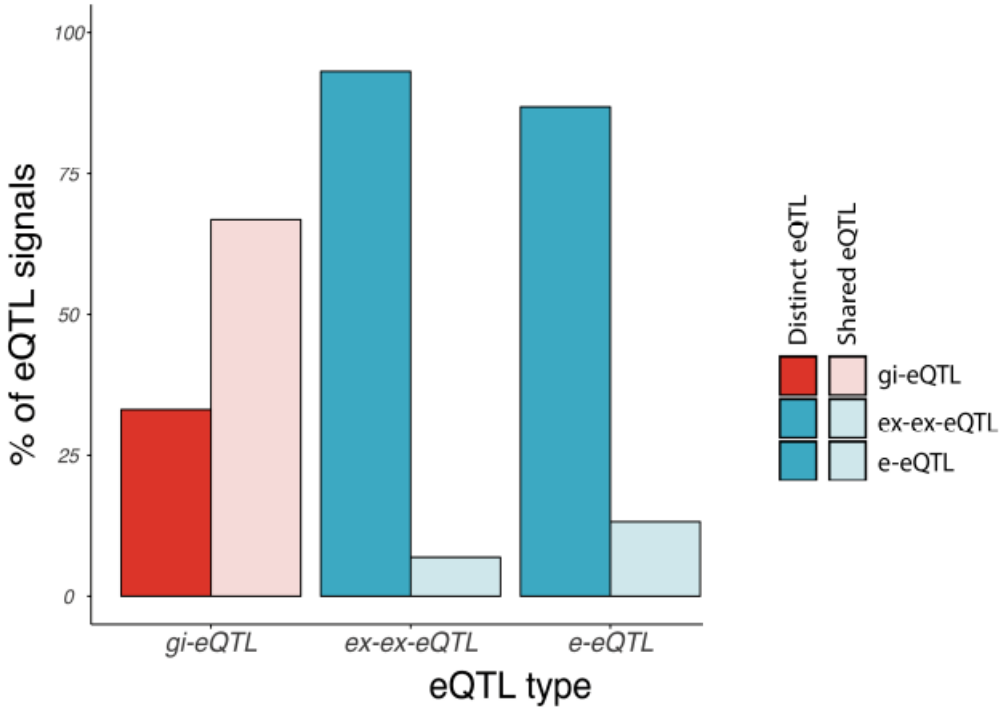

Figure 3.7: **Heterogeneity test comparing gene-level eQTLs to non-standard eQTL.** Bar chart showing the heterogeneity of eQTL signals when comparing gene-level eQTLs to non-standard eQTL classes applied to the same gene. Analysis were perfoermed separately for gi-eQTLs (red), e-eQTLs and ex-ex-eQTLs (turquoise bars). All signals with an FDR-corrected p-value of $< 5\%$ using a beta-heterogeneity test were considered distinct (opaque bars), while an FDR-corrected p-value of $> 5\%$ was taken as evidence of eQTL sharing (transparent bars).

3.3.3.2 Location and functional annotation of eQTLs signals around target genes

Since recent studies have demonstrated that the location of eQTLs with respect to their target gene can provide insights into the molecular basis for their effectsGaffney et al. 2012; Battle et al. 2014, I investigated this for all types of eQTLs identified.

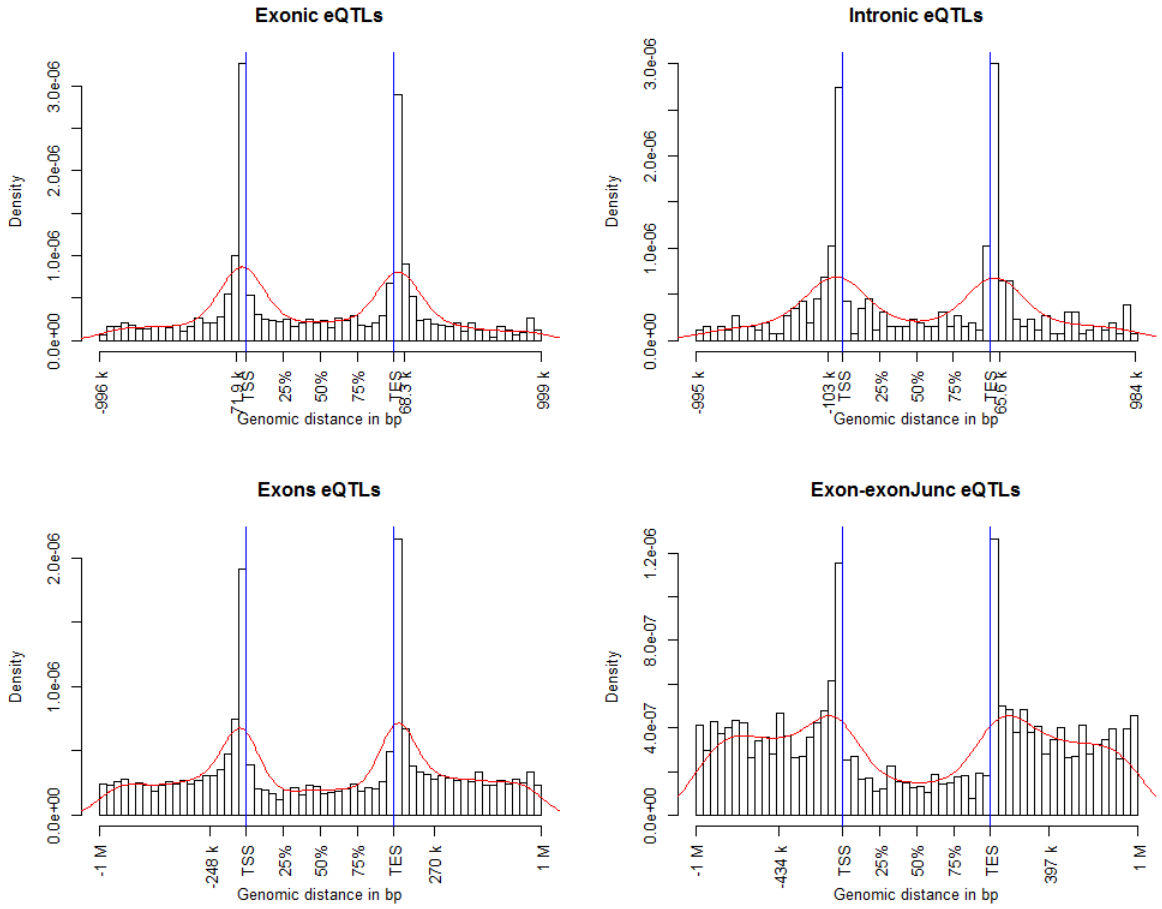

Figure 3.8: **Location of eQTLs with respect to their TSS and TES for the different types of quantification.** Histogram to show the distribution of the location of eQTL variants relative to the target gene. Distance of variants within the gene are epressed in percentage relative to the TSS (0% as the TSS genomic position and 100% as the TES genomic position). The red lines indicate the density of the distribution. The blue lines indicate the TSS and TES of the target gene.

I found that eQTL signals of all types showed a bimodal distribution, with peaks centered around the transcription start site (TSS) and the transcription end site (TES)

for all type of quantification (Figure 3.8). This confirms the findings of previous eQTL studies (Li et al. 2016b). However, I also found a significant association between the location of the variant (namely whether it is was located within the boundaries of its target gene or not) and the type of quantification used to detect the eQTL (P-value chi-square $<2.2 \times 10^{-16}$).

Using the results of the beta-heterogeneity test, in section 3.3.3.1, to confidently separate eQTLs into splicing-specific eQTLs (ex-ex-eQTL and e-eQTL) and gene-level eQTLs (ge-eQTL and gi-eQTL), it was possible to demonstrate that eQTL variants annotated as downstream were enriched amongst splicing-specific eQTLs as compared to gene-level eQTLs (Figure 3.9, Fisher Exact test p-value = 1.4×10^{-3}). Furthermore, there was a significant enrichment of 3' UTR variants (Figure3.9, Fisher's Exact test p-value = 4.9×10^{-7}) amongst eQTLs related to splicing as compared to eQTLs related to changes in overall gene expression. This suggests that splicing-specific eQTL and gene-level eQTL effects potentially underlie different regulatory processes. In particular, enrichment of regulatory variants in 3' UTR have been previously reported as post-transcriptional mechanisms to influence RNA stabilityMatoulkova et al. 2012; Pai et al. 2012.

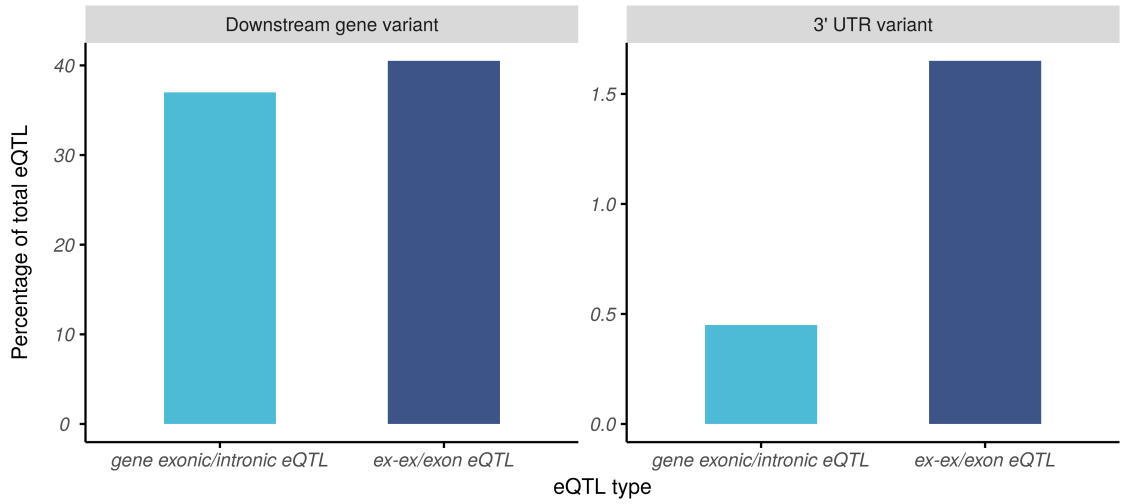

Figure 3.9: **Enrichment of variant annotation for different eQTL classes.** Bar chart to show the percentage of eQTL variants located downstream of the target gene (left panel) and percentage of eQTL variants located at the 3' UTR of the target gene (rigth panel).

3.3.4 Splicing eQTL targets are enriched for neuronal genes

Several studies have demonstrate that eQTL may be generated by specific cell types (Fairfax et al. 2012; Naranbhai et al. 2015; Westra et al. 2015). With this in mind, gene co-expression networks annotated with cell-specific markers were used to assign eQTL target genes (41.5% of all target genes) to 5 broad cell types, namely astrocyte, endothelial, microglia, neuron and oligodendrocyte. Subsequently, cell-specific enrichment was tested within each eQTL quantification category. eQTLs that are more related to splicing, namely e-eQTLs and ex-ex-eQTLs, showed a significant enrichment in neuronal genes (FDR-corrected p-value = 3.89 x 10^{-17} and 9.12 x 10^{-6} in putamen and substantia nigra respectively, Figure 3.10) astrocyte genes (FDR-corrected p-value = 8.74 x 10^{-4} and 1.12 x 10^{-3} for e-eQTLs and ex-ex-eQTLs respectively in substantia nigra, Figure 3.10) and oligodendrocyte genes (FDR-corrected p-value = 1.7 x 10^{-3} and 4.1 x 10^{-2} for e-eQTLs and ex-ex-eQTLs respectively in substantia nigra, Figure 3.10).

This reinforces the importance of capturing different regulatory mechanisms in human brain samples.

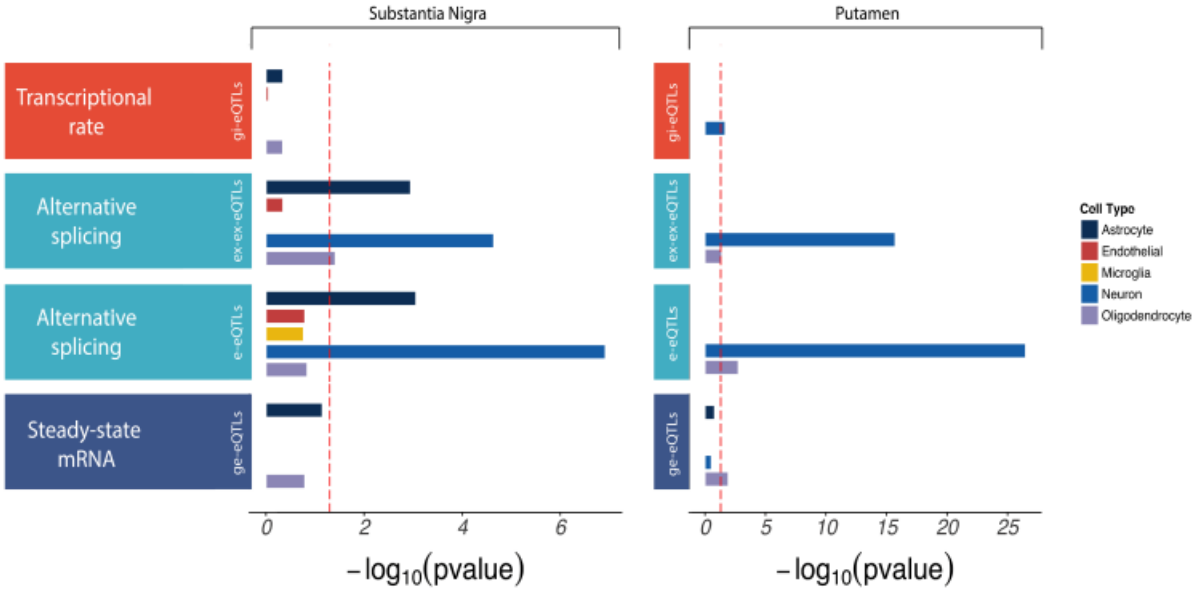

Figure 3.10: **Enrichment of cell-type in eQTL target genes across eQTL classes.** Bar chart to show that the genes and features targeted by different eQTLs classes are variably enriched for genes with cell-biased expression. We performed this analysis separately for eQTLs generated through the analysis of putamen (right panel) and substantia nigra (left panel) RNAseq data. In each case the cut-off for significant cell type-specific enrichment of targeted features is depicted with a dotted red line.

3.3.5 Interpretation of GWAS hits using eQTLs

To evaluate the importance of the eQTLs discovered for complex diseases, the overlap of eQTL signals with known GWAS hits, classified by the STOPGAP database, was explored. Analyses were restricted to GWAS hits passing genome-wide significance (p-value $< 5 \times 10^{-8}$). When considering all GWAS hits, eQTL classes related to changes in overall gene expression had the highest overlap with 7.6% for gi-eQTL and 6.3 ge-eQTL of eQTL variants of these classes also being GWAS hits, followed by e-eQTL and ex-ex-eQTL which have rates of GWAS overlap of 5.1% and 3.7% respectively (Figure 3.11).

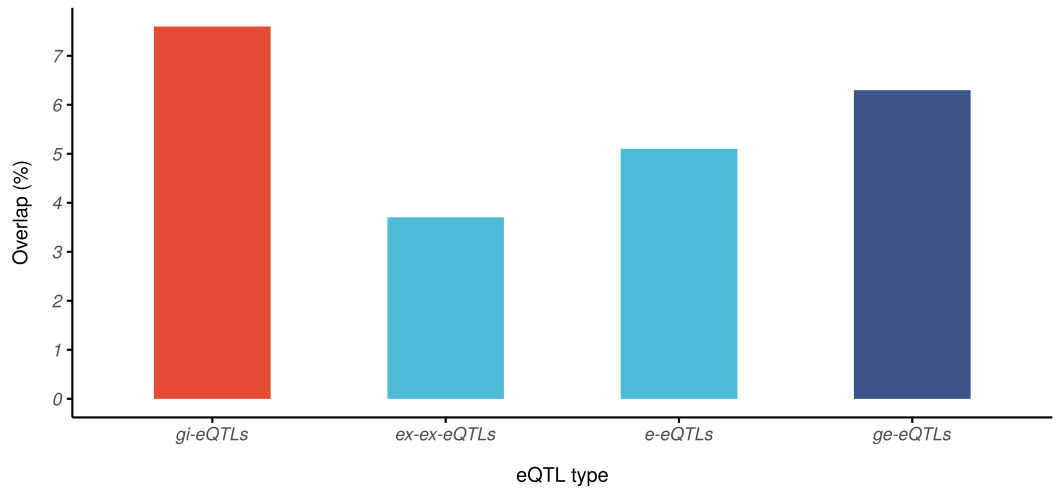

Figure 3.11: **Overlap between GWAS variants and eQTL for the different eQTL classes.** Bar chart to show the percentage of the eQTL variants that overlap with GWAS risk's variants classified by the STOPGAP database stratified by eQTL type.

Interestingly, eQTL categories were enriched for GWAS variants associated with neurological and behavioural disorders (Fisher Exact test p-values for all quantification are <0.05, Figure 3.12) as compared to all other phenotype-associated variants.

Similar, sub-analyses performed on eQTLs of different types, show an enrichment of GWAS risk SNPs for adult neurological disorders was present in ex-ex-eQTLs and e-eQTLs (Figure 3.12). This would suggest that while splicing-specific eQTLs might be harder to detect than gene-level effects, RNA splicing might be an important target for neurological and behavioural risk variants. Furthermore, all eQTL classes show enrichment of GWAS risk SNPs for adult neurological disorders (Figure 3.12), these results highlight the value of performing eQTL analyses within disease relevant tissues despite the challenges.

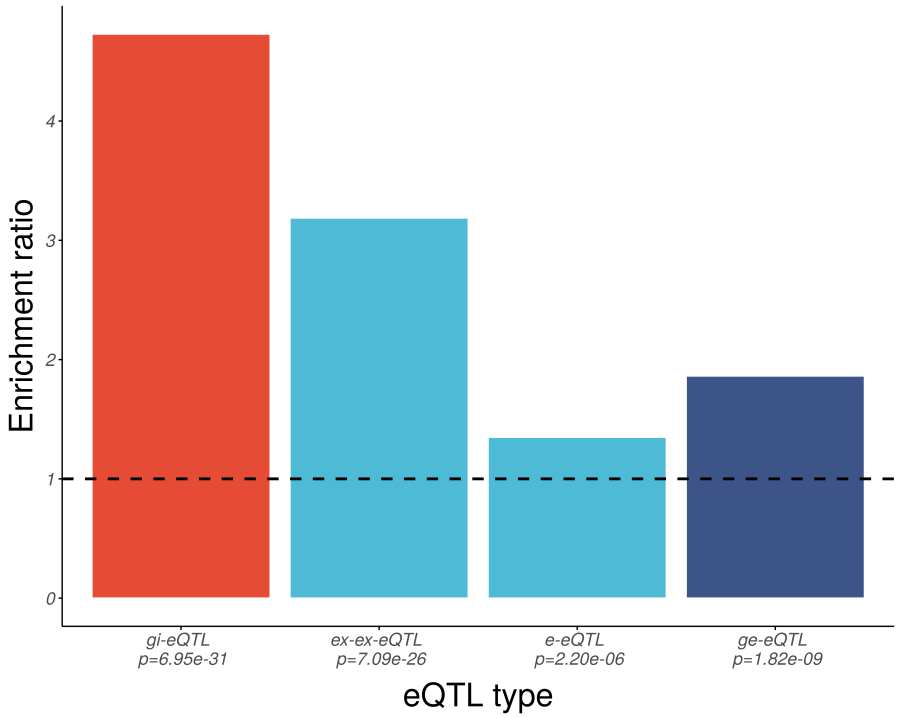

Figure 3.12: **Enrichment of risk variants reported for neurological and behavioural disorders amongst eQTLs categories.** Bar chart to show the enrichment of GWAS risk's variants amongst eQTL types. For each eQTL category Fisher Exact test p-value for the enrichment is displayed on the x axis.

I explored this further for a number of adult neurological disorders for which I also had access to GWAS summary statistics, namely Parksinson's disease (Nalls et al. 2014) and multiple sclerosis (Sawcer et al. 2011). These diseases were selected because while they both have an adult onset, the pathology in multiple sclerosis is largely within white matter, while PD is characterised by neurodegeneration within the basal ganglia and specifically dopaminergic neurons of the substantia nigra. I generated Q-Q plots (as per Li et al. 2016b) to look for evidence of an enrichment of SNPs with low p-values in the disease GWAS amongst the eQTLs detected. As demonstrated by Li et al. 2016b, this approach makes it possible to more sensitively compare different types of eQTLs (relating to different molecular processes) and their contribution to a complex disease. Using this approach an enrichment of SNPs with low p-values in the PD

GWAS was demonstrated amongst exon-exon spanning junctions eQTLs (ex-ex-eQTL) as compared to other types of eQTLs (Figure 3.12a). Given that ex-ex-eQTLs are most closely related to splicing, this suggests that regulation of tissue-specific splicing might be playing a particularly important role in PD. This type of enrichment was not observed within the analysis of the GWAS for MS (Figure 3.12b), which is a disease predominantly of the white matter driven primarely by immune dysfunction.

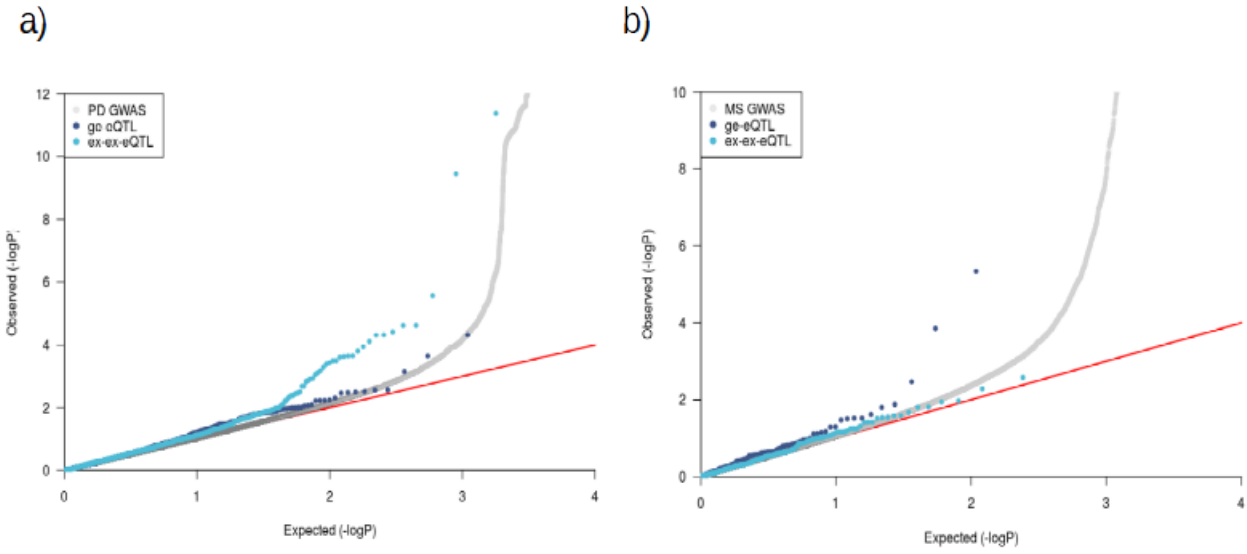

Figure 3.13: **GWAS low p-value enrichment for ex-ex-eQTLs.** a) Quantile-Quantile(Q-Q) plots of PD GWAS p-values for ge-eQTL and ex-ex-eQTL. b) Q-Q plots of MS GWAS p-values for ge-eQTL and ex-ex-eQTL.

3.4 Discussion

Improving the understanding of risk loci for neurological disorders using eQTL analyses has proven to be challenging. Not only is the availability of human brain samples limited, but human brain has various layers of complexity, due to cellular heterogeneity and high regional interconnectivity. This complexity is the motivation for more specific forms of data processing and analyses.

In this chapter, I studied the regulation of different RNA processes, by quantify-

ing transcription with different approaches, in control brain tissues relevant to various neurological and neuropsychiatric disorders. The RNA-Seq dataset used represents a unique resource because of the tissues analysed and the protocols used to generate the data.

Using beta-heterogeneity testing I was able to differentiate variants across different eQTL classes, representing distinct regulatory sites and likely acting through different regulatory mechanisms. While e-eQTLs and ex-ex-eQTLs are likely to be regulating alternative splicing and controling mRNA stability of specific isoforms, gene-level eQTLs are likely to be changing transcriptional rate and this view is supported by the different genomic locations of eQTLs. Furthermore, I found that eQTLs related to splicing (e-eQTL and ex-ex-eQTL) have gene targets enriched for neuronally expressed genes, as evidenced by the cell-type enrichment analysis. This reinforces the importance of different quantification strategies to improve the value of eQTL data.

Finally, I demonstrated a significant enrichment of risk SNPs for neurological and behavioral disorders within the eQTL dataset. This suggests that eQTL analyses performed in disease related tissues are more likely to improve the characterisation of GWAS risk loci. When I investigated the possible molecular mechanisms underlying this enrichment using the different eQTL types in PD (a disease which is characterised by neurodegeneration and dysfunction of the basal ganglia), these findings indicate that splicing should be a focal point in the follow-up of GWAS experiments.

In conclusion, eQTLs not only have the potential to explain how genetic variability can influence particular phenotypes, but by using more complex approaches can provide insights into the molecular mechanisms through which they operate.

Chapter 4

Identification and genetic regulation of transcribed intergenic regions

4.1 Introduction

The analyses presented in this chapter were motivated by growing evidence for incomplete annotation of the adult brain transcriptome. Different components of gene annotation references can be affected by incomplete annotation. Most simply, inaccuracies can be classified as those within known gene boundaries (e.g. intron-exon boundaries) or outside known gene boundaries (e.g. novel genes or novel 3' or 5' exons of known genes). In particular, the former might be more challenging to detect using a total RNA protocol because of the presence of pre-mRNA has the potential to generate false positives. This might not be the case for RNA-Seq generated from poly(A)+ selected libraries because of the significant smaller proportion of pre-mRNA fragments (Zhao et al. 2018). However, total RNA library construction might be more useful for inaccuracies outside known gene known gene boundaries and in particular for non-polyadenylated transcripts. This is important because gene annotation references serve as a framework for most gene expression studies and inaccuracies in these

references have the potential to lead to significant errors.

Several studies have suggested that current gene annotation may be incomplete in human brain (Chen et al. 2011; Zhang et al. 2012; Jaffe et al. 2015; Blauwendraat et al. 2016), not only within gene boundaries but also within intergenic regions. For example Jaffe and colleagues (Jaffe et al. 2015) found evidence for transcription of ~12Mb of intergenic material in adult human frontal cortex. These transcribed intergenic regions were more likely to be differentially expressed during cortical development and the authors postulated that these regions may not have been annotated previously because they are highly cell-specific.

Further evidences for novel intergenic transcription are provided in another study using nuclear fractionation coupled with RNA-Seq (Werner et al. 2015). In this study, Werner and colleagues demonstrate that more than 80% of chromatin-associated RNAs, which promote activation of neighboring genes, were absent in annotation databases. In fact, the advent of RNA-Seq more generally has permitted the discovery of thousands of lncRNAs of which ~40% are thought to be expressed specifically in brain (Derrien et al. 2012). However, given that most of the RNA-Seq studies performed to date have used polyadenylated (poly(A)+) RNA capture libraries and many lncRNAs are not polyadenylated, existing studies may have limited sensitivity when detecting tissue or cell specific lncRNAs. Finally, evidence for incomplete annotation is supported by the almost quarterly updates on gene references, suggesting a continuous refinement of transcriptome annotation.

Inaccuracies of gene annotation might also have an impact in the understanding of complex disease. In the past decade, the improvements of sequencing technologies have generated advances in genetic research through the identification of genetic variants contributing to both complex and rare disease. However, the interpretation of these variants remains a challenging task and more accurate gene annotation could potentially improve the characterisation of variants. At present 43% of GWAS variants map

to intergenic regions (Hindorff et al. 2009) making it very difficult to determine the underlying molecular processes driving increased risk of disease. Therefore, interpretation of GWAS risk loci could potentially be improved through the identification of novel exons or genes lying outside known gene boundaries since intergenic risk SNPs may either lie within these regions or regulate such regions. In particular, undiscovered non-coding RNAs may underlie the function of a proportion of GWAS risk variants.

Determining the pathogenicity of rare variants requires not only the identification of causal genes, but a thorough understanding of the gene's structure (MacArthur et al. 2014). A full characterisation of gene structure will improve the interpretation of a given variant's significance and its impact on gene integrity and function. Moreover, unannotated exons lying outside of gene boundaries are likely to be excluded from targeted genetic screening with the implication that some pathogenic mutations could be missed.

Therefore, improving the accuracy and completeness of the annotation of both coding and non-coding genes improves our understanding of the impact of genetic variation in complex and rare diseases. The complexity of alternative splicing, cellular heterogeneity and difficulty of access to brain tissues increases the impact of misannotation for brain-expressed genes, suggesting this will particularly affect the interpretation of genetic variation for neurological diseases. In this chapter I perform eQTL analysis on the most conservative portion of the data, corresponding to the non-coding portion least interferred with by the coding portion, namely the intergenic transcription. I explored the extent of inaccuracies outside gene boundaries and, subsequently, I investigated the impact of gene annotation incompletness on variant interpretation.

4.2 Methods

4.2.1 Characterisation of transcribed intergenic eQTLs

The approach used to identify eQTLs targeting transcribed intergenic regions was described in the previous chapter (Section 3.2). Given the large quantity of transcribed intergenic regions identified and a limited number of functional gene loci still to be discovered (Pertea et al. 2018), the characterisation of transcribed intergenic regions was designed to prioritise those regions that represent incomplete annotation of existing genes. Therefore, the transcribed intergenic regions were characterised using the steps below (Figure 4.1):

1. Evidence of split reads spanning transcribed intergenic regions and a known gene: These regions were characterised as regions with **strong** evidence of being part of a known gene.

2. Evidence of co-expression (defined as Pearson's R^2>0.2) of a novel transcribed intergenic region with the nearest gene *and* close proximity to that gene (=<5Kb): These regions were characterised as regions with **moderate** evidence of being part of known gene.

3. Evidence of low co-expression (defined as Pearson's $R^2 \leqslant 0.2$) of an unannotated transcribed region with the nearest gene, *and* no split reads linking the region to a known gene *and* distant from any known gene (>5Kb): These regions were characterised as having **weak** evidence of being part of a known gene. This group of regions were further subdivided on the basis of those that have split reads linking one region with "weak" evidence with with another region with "weak" evidence.

4. Some unannotated transcribed regions did not meet criteria for either strong,

moderate or weak evidence of being part of a known gene and so remained uncharacterised.

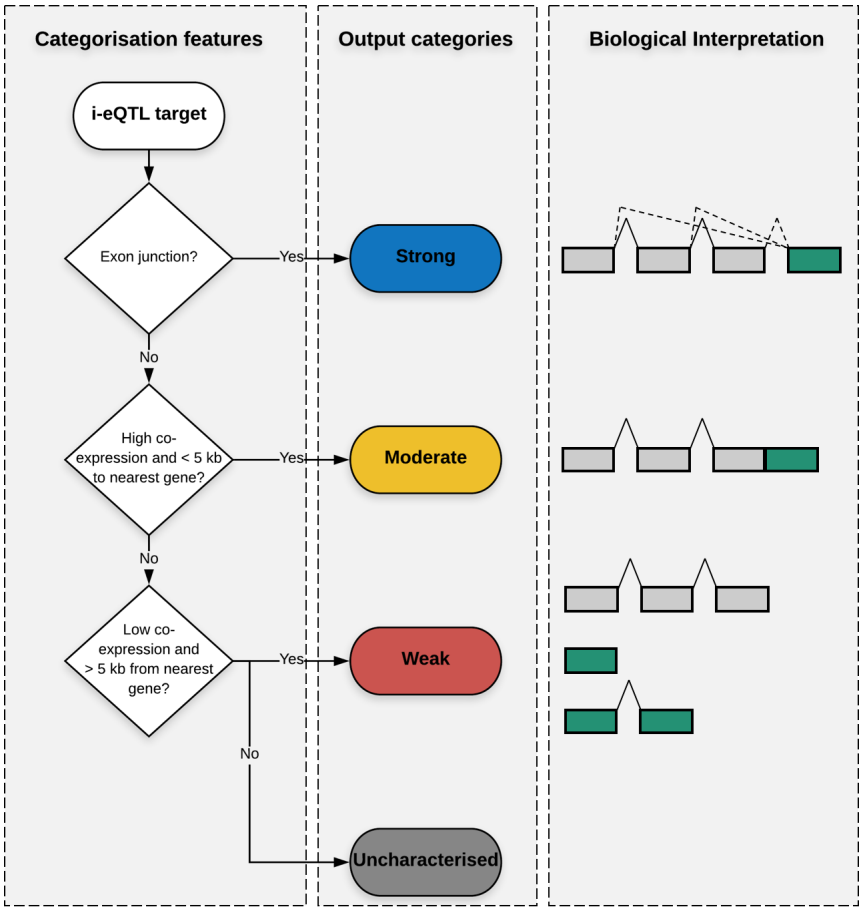

Figure 4.1: **Characterisation of transcribed intergenic regions.** Schematic illustration to show the features used to categorise transcribed intergenic regions targeted by i-eQTLs as those with strong, medium or weak evidence for being part of a known gene.

4.2.1.1 Characterisation of intergenic unannotated regions with strong evidence of being part of a known gene

To determine whether an unannotated transcribed region had strong evidence of being part of a known gene, the information on split reads was collected per sample from the Tophat2Kim et al. 2013 junction output file. Each split read is a short cDNA sequence that aligns to a genomic location with a gap (or multiple gaps) in-between, so providing

accurate information about splicing events (Figure 4.2).

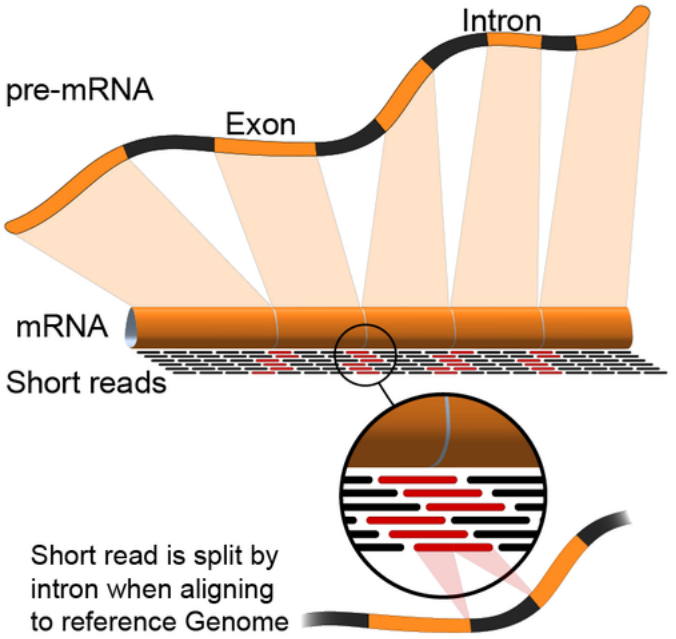

Figure 4.2: **Split read example.** Example of how split reads align in the presence of splicing events. Figure adapted from https://upload.wikimedia.org/wikipedia/commons/0/01/RNA-Seq-alignment.png

Overlaps between transcribed intergenic regions targeted by eQTLs and split reads was assessed using the GenomicRanges (Lawrence et al. 2013) R package. If an unnannotated transcribed region overlapped with a split read, then the same split read was screened for overlap with any surrounding gene. In this way, it was possible to identify split reads that linked the transcribed region with annotated genes and suggested that a novel splicing event had occurred to generate a presumed novel exon. For each region with strong evidence of a connection to a known gene, strand, distance and co-expression was calculated in relation to the relevant annotated gene. Furthermore, the total number of samples in which split reads could be detected was collected in order to provide a measure of confidence. Only transcribed intergenic regions with split reads present in at least 4 separate samples were defined as regions with strong evidence of being part of a known gene.

4.2.1.2 Characterisation of intergenic unannotated regions with moderate evidence of being part of a known gene

Unannotated transcribed intergenic regions were classified as having moderate evidence of being part of a known gene when the regions showed high co-expression and were located in close proximity with their nearest gene. This classification was only used in the absence of relevant split read information. The nearest gene, in genomic distance, from each unannotated transcribed region was selected using the GenomicRanges (Lawrence et al. 2013) R package. The correlation (R^2) was calculated between the expression of each transcribed intergenic region and each exon of the nearest gene.

The distance between the nearest gene and each unannotated transcribed region was also used as a parameter to categorise unannotated transcribed regions. Unannotated transcribed intergenic regions that were more than 5 Kb from the nearest gene were not included within the category. The 5 Kb threshold was determined after calculating the distribution of intron lengths amongst genes expressed in human brain. In this way it was possible to see that while some of the regions with moderate evidence were continuous with the last exons of their nearest gene(Figure 4.1), suggesting an "extension" of a known exon, others were completely separate, suggesting a new splicing event. In summary, transcribed intergenic regions that had an R^2 greater than 0.2 with any of the exons of the nearest gene and which were 5Kb or less from their nearest gene were classified as having moderate evidence for being part of a known gene.

4.2.1.3 Characterisation of intergenic unannotated regions with weak evidence of being part of a known gene

Unannotated intergenic transcribed regions with weak evidence of being part of a known gene were defined based on distance from their nearest gene and co-expression with the nearest gene. Distance and correlation from the nearest gene was calculated as for regions that exhibited moderate evidence for being part of a known gene (section

4.2.1.2). Transcribed intergenic regions were categorised as having weak evidence of being part of a known gene if they had a correlation of $R^2 \leqslant 0.2$ and were more than 5Kb from their nearest gene. These regions were sub-classified using split read information.

4.2.2 Replication of transcribed intergenic regions in independent datasets

4.2.2.1 Replication within the recount2 platform

Replication was assessed using RNAseq data from all 54 tissues generated by GTEx and released within recoung2. Each of the 54 tissues was investigated separately. Using the genome coordinates of all transcribed intergenic regions detected in putamen and substantia nigra, I quantified expression profiles. Quantification was facilitated by Dr Leonardo Collado-Torres at the Lieber Institute. In the case of each transcribed regions, read counts were transformed into RPKM and replication was considered if the region had an RPKM > 0.1 in at least 80% of samples in the tissue of interest.

4.2.2.2 Replication within the NONCODE database

Overlaps between transcribed intergenic regions and regions annotated within the NONCODE (Liu et al. 2005) database were investigated. NONCODE is a comprehensive database of annotated RNA sequences that do not encode for proteins. This database is derived from the published literature and public repositories, such as lncRNAdb (Quek et al. 2015). Transcribed intergenic regions were lifted to the GRCh38 to match the NONCODE genome version using the web-tool liftover from Ensembl. Replication was counted if at least one bp overlap between the region annotated in NONCODE and the intergenic transcribed region. Overlaps were performed using the R Bioconductor package GenomicRanges.

4.2.2.3 Replication within independent RNA-Seq datasets

RNA-seq aligned bam files were obtained for a cohort of 213 human control frontal cortex (FCTX) samples. This data was generated in Dr. Mark Cookson's laboratory as part of the North America Brain Expression Consortium (NABEC). Alignment and standard quality checks were performed by Dr. Raphael Gibbs using the software STAR (Dobin et al. 2013) and version hg19 of the human genome annotation from UCSC (Karolchik et al. 2003). Aligned bam files were analysed using the same approach as that used for analysis of putamen and substantia nigra (section 3.2.2.1). Using the same parameters and filtering steps (RPKM>0.1 in at least 80% of the samples and regions >=100bp), 25021 transcribed intergenic regions were identified in FCTX. Overlaps between the transcribed intergenic regions in putamen/substantia nigra and the frontal cortex samples were performed using the R Bioconductor package GenomicRanges and replication was considered if at least one bp was overlapping.

4.2.2.4 Replication within independent CAGE-Seq datasets

I used CAGE-Seq data from an unpublished human control substantia nigra (N=211) and a publicly available human control frontal lobe (N=128) (Blauwendraat et al. 2016) dataset to assess the replication of unannotated transcribed intergenic regions across different sequencing technologies. CAGE-Seq only captures the 5' end of transcripts and since mRNA degradation takes place from 5' to 3', flanking regions of 500bp were added upstream and downstream to all transcribed intergenic regions. The genomic overlap was performed using the software intersectBed from BEDTools (Quinlan and Hall 2010) with default parameters. In the case of substantia nigra CAGE-Seq data, the genomic coordinates of transcribed intergenic regions were lifted to the GRCh38 to match genome versions. Replication of transcribed intergenic regions was counted if at least one bp was overlapping with a CAGE-Seq identified region. This work was performed in collaboration with Dr. Javier Simon-Sanchez at the German Center for

Neurodegenerative Diseases (DZNE) in Tubingen.

4.2.3 Validation of intergenic transcribed regions using RT-PCR and Sanger sequencing

A small number of transcribed intergenic regions were also validated experimentally by Regina Reynolds (PhD student in Department of Neurodegenerative Disease UCL). The selection of transcribed intergenic regions for validation was guided by the RPKM expression of these regions in putamen and by evidence of split reads supporting these regions. Of the eight transcribed intergenic regions investigated, three regions had strong and three regions had moderate evidence of being part of a known gene and two regions had weak evidence of being part of a known gene. Seven RNA samples, each derived from a different individual, were used for validation.

The High Capacity cDNA RT Kit (Applied Biosystems) in combination with random primers was used to perform reverse transcription with 500µg of total RNA as input. Custom primers were designed to cover the predicted split reads to perform the Polymerase Chain Reaction (PCR) experiment using the FastStart PCR Master (Roche). Gel electrophoresis was used to confirm amplification of bands and by using Exonuclease I (Thermo Scientific) and FastAP Thermosensitive Alkaline Phosphatase (Thermo Scientific), PCR products were removed. The sequencing step was performed using the BigDye terminator kit (Applied Biosystems) and sequences were analysed using the CodonCode Aligner (V. 6.0.2). After sequence quality control, contiguous sequences were aligned to the human genome (GRCh37/hg19) and viewed in the UCSC genome web browser for manual inspection.

4.2.4 Beta-heterogenity testing of eQTLs targeting known annotated regions and intergenic transcribed regions

To identify heterogeneity in eQTL signal strength, beta-heterogeneity testing was performed as previously described in section 3.2.5, but in this chapter was used between eQTL targeting unannotated transcribed regions and eQTL targeting the reference gene. The reference gene is the gene which the intergenic transcribed region have evidence of being part of. With this in mind, I applied the following models using the transcribed intergenic region and the exon of the reference gene with the highest co-expression (for simplicity, referred to as transcribed region and known exon respectively in this paragraph) : 1) a model that assumed a shared eQTL across the transcribed region and the known exon, which contained two main effect terms, allele dosage and an index identifying the transcribed region and the known exon, and one random effect, which indexed each individual in the dataset. 2) a model that assumes no-shared eQTL between the transcribed region and the known exon containing the same terms as in model (1) with the addition of a set of fixed effect allele dosage $\times$ index for the transcribed region and the known exon. The R function lmer() was used to fit both models and a likelihood ratio test was applied to determine the significance. Finally, the Benjamini–Hochberg method was applied to the p-value to adjust for the total number of eQTLs tested.

4.2.4.1 Colocalisation of eQTLs and GWAS risk loci for Neurological disorders

The R software coloc (Giambartolomei et al. 2014) was used to assess the colocalisation of Parkinson's disease and schizophrenia loci amongst eQTL signals. All loci with a GWAS p-value $<10^{-5}$ were tested for colocalisation. The betas and standard errors from the summary statistics of common SNPs across both datasets and within 1MB

of the GWAS SNP of interest were used as input. Coloc was run with default priors and parameters to generate the posterior probabilities of five different hypotheses: **H0:** No colocalisation and no causal SNP identified in both GWAS and eQTL, **H1:** No colocalisation but causal SNP is found within the GWAS set, **H2:** No colocalisation but causal SNP is found within the eQTL set, **H3:** No colocalisation but both sets contain a causal SNP. **H4:** Colocalisation of the causal SNP present in both sets. We defined loci that colocalised as per Li and colleagues (Li et al. 2017b) when H3+H4≥0.8 and H4/H3≥2. While H3+H4≥0.8 represents sufficient test power to reliable call a colocalise locus, the H4/H3≥2 reprensents a greater (at least twice) posterior probability of being colocalised locus. This work was performed by David Zhang (PhD in Department of Neurodegenerative Disease UCL).

4.3 Results

4.3.1 Reference-free approaches enlarge the transcriptome within human substantia nigra and putamen

The number of transcribed bases within unannotated intergenic regions was calculated and compared to the number of bases present within expressed annotated regions. Using this approach an additional 7.09M and 3.48M bases were detected using the reference-free approach, respectively for putamen and substantia nigra. This corresponded to a 12.5% and 6.6% increase in the size of the annotated transcriptome within these tissues (Figure 4.3).

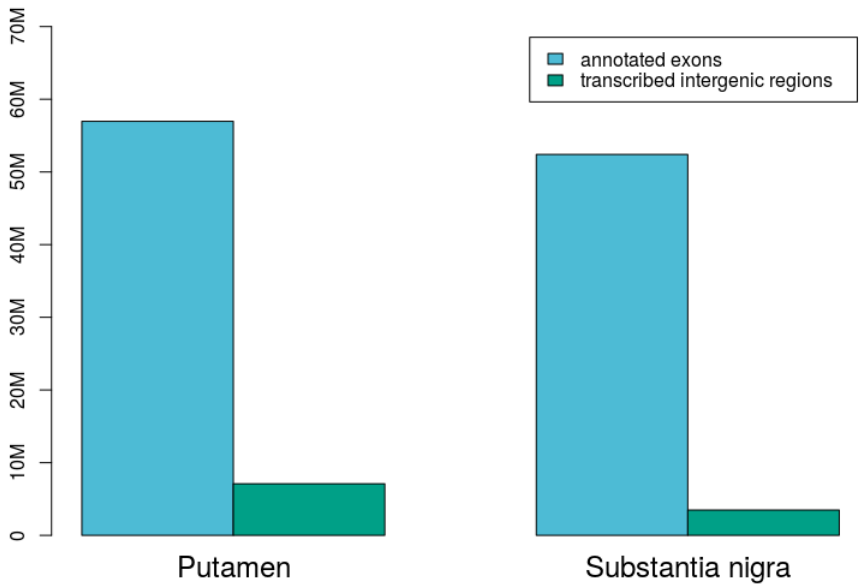

Figure 4.3: **Genomic length of annotated expressed exons compared to transcribed intergenic regions.** Bar chart to show the total size in bp of annotated exons (blue) and transcribed intergenic regions (green) for putamen and substantia nigra.

4.3.2 Replication of transcribed intergenic regions in independent datasets

I recognised that a proportion of the intergenic trancribed regions identified may originate from technical variability, thus I investigated further for evidence of validation across different GENCODE versions and independent datasets. Since GENCODE version 19 was released during the course of this analysis, I checked whether any of the transcribed intergenic (defined as intergenic on the basis of version 18, see 3.2.2.1) regions identified had been incorporated into version 19. Firstly, I measured how many transcribed intergenic regions were annotated as exonic if I had used the the GENCODE version 19. The annotation was performed in both version using the R package annotateRegions with a minimum overlap of 1bp. I found that 1149 (5% of all transcribed intergenic regions identified in GENCODE version 18) were annotated in version

19, but not in the version 18 (Figure 4.4). This not only demonstrates the validity of the reference-free approach to identify novel transcribed regions, but reinforces the underlying rationale for this analysis, namely that transcriptome references for brain are incomplete.

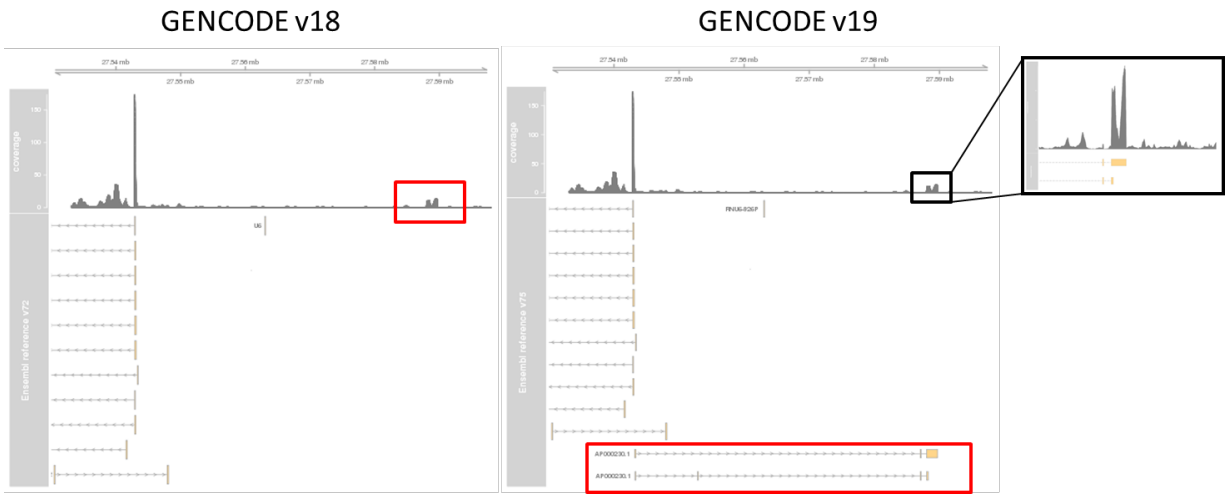

Figure 4.4: **Annotation of transcribed intergenic regions in later version of GENCODE.** An example of a transcribed regions (chr21:27588064-27589052) intergenic that was annotated as intergenic in the GENCODE v.18 (Left). However, the same region was later annotated in the GENCODE v.19 (Right) as long intergenic non-coding RNA: AP000230. The AP000230 gene was manually annotated in the Human and Vertebrate Analysis and Annotation (HAVANA) group as part of the GENCODE project.

Next I investigated all transcribed intergenic regions for evidences of expression in four independent datasets, the GTEx samples within the recount database, the NABEC dataset, the NONCODE database and the Blauwendraat and DZNE CAGE-Seq datasets. Using GTEx data I found that 48.1% of the transcribed intergenic regions were also expressed in matching brain regions; either putamen or substantia nigra. If I considered all central nervous system tissues in GTEx, the replication rates rose to 63.6%. This represents a very high replication rate considering that a polyA selection step was used in the generation of GTEx data, an approach which is expected to exclude non-poly adenylated non-coding RNAs.

With this issue in mind, transcribed intergenic regions were investigated in the

NABEC dataset which was generated with a more similar library preparation of the UKBEC RNA-Seq dataset. Of all transcribed intergenic regions detected there was evidence for transcription in the NABEC dataset in 41.9% and 44.3% respectively in putamen and substantia nigra. This is a high replication rate considering the non-matching tissue types.

Furthermore, to investigate if transcribed intergenic regions were generated through unannotated non-coding RNA transcripts, I used the non-coding RNA catalogue, NONCODE. I found that 22.5% of transcribed intergenic regions detected in putamen and 24.3% regions detected in substantia nigra overlapped with regions catalogued in the NONCODE database. Again, this is a high overlap since it does not take into consideration differences in technologies used to detect non-coding RNA, tissue type or the analysis used to annotate the regions in the NONCODE database.

Finally, I investigated expression of transcribed intergenic regions in both the Blauwendraat (Blauwendraat et al. 2016) and DZNE CAGE-Seq datasets, the replication was 6.4% across both tissues. Although the replication rate in the CAGE-Seq appeared discouraging, the low replication rate may be due to the 5' bias of the CAGE-Seq data, which would not be capable of capturing novel regions identified at the 3' end of transcripts. Furthermore, one would expect to identify a greater number of transcribed intergenic regions at the 3' end of known genes because of the 5' to 3' degradation taking place and the 3' bias generated by oligo(dT) priming (Kuersten 2012).

In summary, considering all transcribed intergenic regions identified across both tissues 75.4% were validated in at least one independent dataset.

4.3.3 Identification and classification of i-eQTLs and their target regions

Association testing was performed between ~6.5 million genetic loci (~5.88M SNPs and ~577K indels - collectively referred as variants) and 14,905 RNA expression traits in

putamen and 8,135 RNA expression traits in substantia nigra. Each RNA expression trait corresponded to a transcribed intergenic region. This resulted in ~0.54billion eQTL tests. Following conditional analysis and using an FDR of 5%, 1363 eQTLs were identified in total. These eQTLs were termed i-eQTLs.

I focused on these i-eQTLs and their target regions in all subsequent analyses on the basis that transcribed intergenic regions with evidence for regulation have the potential to be tested for disease relevance.

This approach led to the prioritisation of 1236 transcribed intergenic regions for further analysis. Of these target regions, 17.4% were classified as regions with strong evidence for being a part of a known gene as indicated by the detection of split reads. This corresponded to 200 transcribed intergenic regions (blue points in Figure 4.5). A further 34.1% of target regions were classified as regions with moderate evidence for being part of a known gene (yellow points in Figure 4.5). These transcribed intergenic regions are likely to represent misannotation of the first/last exons of a known gene, or unannotated novel exons with insufficient read depth to allow detection of a split read. Regions with weak evidence for connection with a neighbouring gene accounted for 17.7% all i-eQTL target regions detected (red points in Figure 4.5). Amongst the 219 of regions with weak evidence of being part of a known gene, were a subset with evidence of split reads linking one intergenic region to another. Finally, 31.9% of i-eQTL targeted regions remained uncharacterised (Table 4.1).

Tissue	**Strong**	**Moderate**	**Weak**	**Unchar.**	**Total**
Putamen (N=105)	158	318	135 (34 with split reads)	301	946
Substantia Nigra (N=65)	42	104	38 (12 with split)	94	290

Table 4.1: **Summary of transcribed intergenic regions classification.** Table to show a summary of i-eQTL target regions divided by classification and tissue.

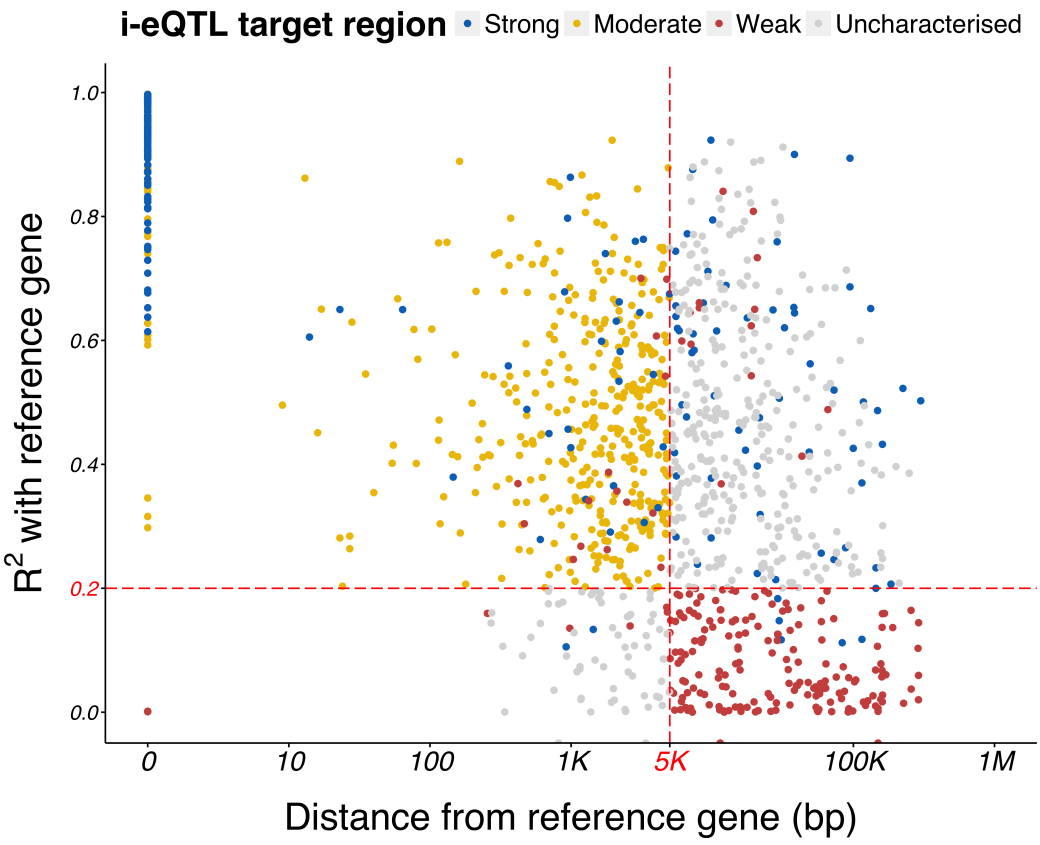

Figure 4.5: **Characterisation of transcribed intergenic regions targeted by i-eQTLs.** Scatter plot to show the characterisation of transcribed intergenic regions. Each point represents a transcribed intergenic region. X axis indicates distance from reference gene and Y axis indicates the Pearson R^2 for the expression between transcribed intergenic region and the nearest exon of reference gene.

4.3.4 Regions with strong evidence for being part of a known gene have higher replication rates in independent datasets.

Replication of i-eQTL target region expression was also considered for each type of classification, namely regions with strong, moderate and weak evidence for being linked to a known gene.

Remarkably, most of the unannotated regions showing strong evidence of being part of a known gene were validated in the recount dataset. The replication rate was 88.9% when considering both putamen and substantia nigra. The validation increased

to 95.4% when considering all GTEx brain tissues contained in recount2 (Figure 4.6). The majority (59.6%) of all i-eQTL target regions with strong evidence of being part of a known gene were also replicated in the NONCODE catalogue. The validation rate was also high in the independent frontal cortex data set made available by NABEC with 65.0% of regions replicating. Furthermore, 15.0% of transcribed intergenic regions upstream of known gene were detected in CAGE-Seq data. Thus, the replication rate was 98.6% if we consider validation in at least one of the independent datasets (Recount2, NONCODE, CAGE-Seq data and NABEC data).

For transcribed intergenic regions classified as having moderate evidence for connection to a known gene in putamen and substantia nigra, recount2 provided support for transcription in 79.0% of cases using either the putamen or substantia nigra GTEx data (Figure 4.6). Furthermore, of these regions 23% were detected in the NONCODE database, 45.3% were identified in the NABEC dataset and 6.9% was found in the CAGE-Seq datasets. When considering all data sets, 92.3% of regions were detected in at least one independent dataset.

Replication of transcription of intergenic regions with weak evidence of being part of a known gene was 41.5% on the basis of GTEx putamen or substantia nigra samples withinin recount2 (Figure 4.6). However, if I considered those regions with split reads support, the replication rate increased to a 93.8%. Across all replication datasets, namely the NONCODE database, NABEC frontal cortex sequencing data and CAGE-Seq data, replication rates were higher when intergenic regions had split read support.

Considering all transcribed intergenic regions with weak support for being part of known gene the validation rate was 79.1% in at least one independent dataset, increasing to 98.0% for those regions supported by split reads. Taken together the higher replication rate in independent datatasets suggests that trranscribed intergenic regions were more likely to replicate when there was evidence not only of transcription, but also splicing (as indicated by the presence of split reads).

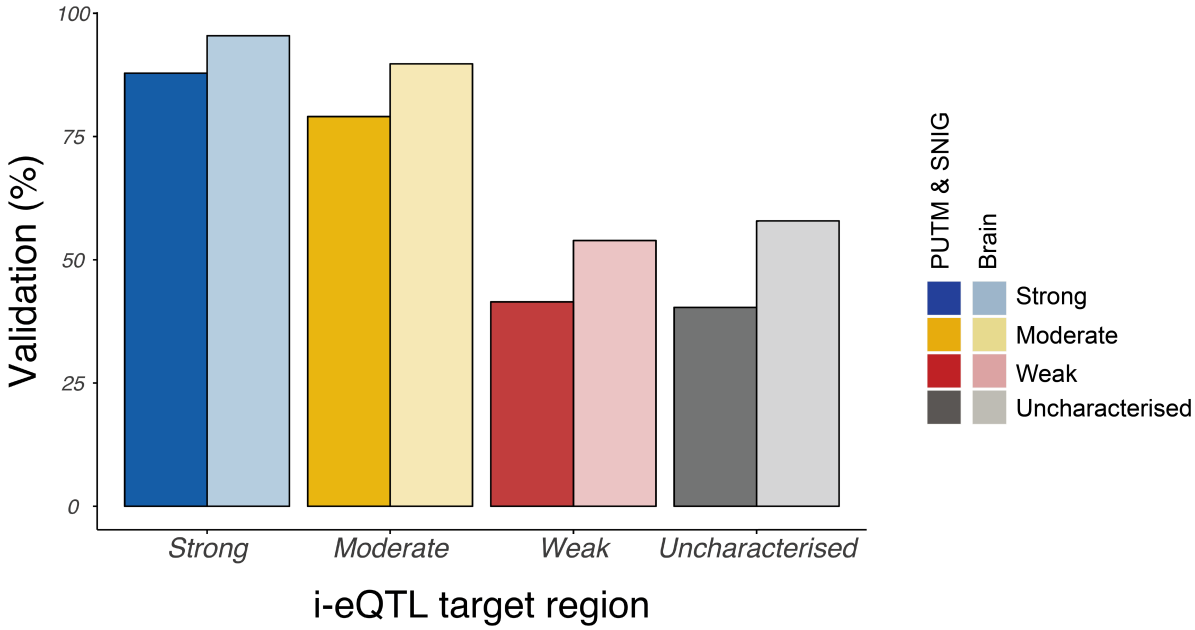

Figure 4.6: **Replication of i-eQTL targets.** Bar plot to show replication of expression of transcribed intergenic regions targeted by i-eQTLs in GTEx data. Replication rates are displayed separately for analyses performed for both putamen (PUTM) and substantia nigra (SNIG) GTEx RNAseq data, separately from RNAseq data generated for all brain tissues within GTEx.

4.3.5 Validation of regions with RT-PCR and Sanger Sequencing

A set of transcribed unannotated regions were selected for experimental validation using RT-PCR and Sanger Sequencing. A final set of eigth unannotated transcribed regions were selected across the different characterisation types (three for region with strong evidence of being part of a gene, three for moderate evidence of being part of a gene, two regions with weak evidence of being part of a gene). Sanger sequencing confirmed the expression of all eight transcribed intergenic regions tested for experimental validation. The validation was independent of the transcribed intergenic regions classification. Interestingly, the Sanger Sequencing not only validate the transcribed intergenic regions, but also the split reads suggesting that the split reads were robustly detected in the

RNA-Seq data.

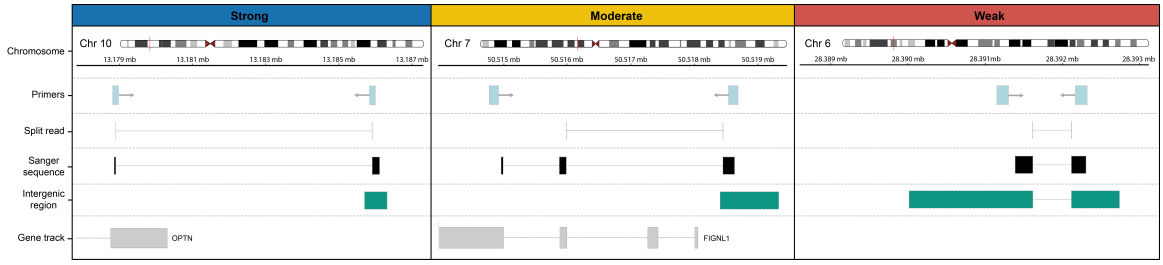

Figure 4.7: **Validation of transcribed intergenic regions using Sanger Sequencing.** Sequencing results for i-eQTL target regions with strong, medium and weak evidence for being part of a known gene. In each case, multiple tracks are provided showing the location of the primers used, the alignment of Sanger sequenced cDNA and the split reads detected by the RNA-Seq data.

4.3.6 Most i-eQTLs represent novel regulatory positions with evidence for functional significance

Beta-heterogeneity testing was used to investigate the effect of i-eQTL signals amongst all characterised regions and their reference gene (as defined by split read information, co-expression or proximity). In all cases, regions targeted by i-eQTLs were separated in two groups based on the beta-heterogeneity test FDR: i) regions that share an eQTL signal with the reference gene (as defined by an FDR $>5\%$), and ii) those that do not share an eQTL signal with their reference gene (as defined by an FDR $=< 5\%$). This is similar to the approach that was used to investigate eQTL sharing in section 3.2.5. This demonstrated that while a portion of known exons share their eQTL signals with other annotated exons (Figure 4.9a), many of the unannotated regions classified as strong, moderate and weak, do not share their associated eQTL signals with their reference gene (Figure 4.9b). As it might have been expected, this difference was most apparent amongst transcribed intergenic regions with weak evidence of being part of a known gene (Fisher Exact test p-value$<$ 2.2 x 10^{-16}, Figure 4.9). This suggests that regions with weak evidence of being part of a known gene are most likely to be functioning independently of the reference gene. However, even amongst target regions with strong

evidence of being part of a known gene, the proportion of i-eQTLs sharing signals with annotated expression features was only 44% (Figure 4.9). Thus, the majority of i-eQTLs represent novel regulatory variants, with those i-eQTL targeting transcribed intergenic regions with strong and moderate evidence of being part of a known gene are probably acting in a transcript-specific manner.

Finally, the transcribed intergenic regions targeted by i-eQTLs were tested for cell-specific enrichment using WGCNA (described in section 3.2.8). Module-membership was used to assign i-eQTL target to modules enriched for gene markers of five different cell types (neuron, oligodendrocyte, astrocyte, microglia and endothelial cell). Using this approach, I found that the targeted transcribed intergenic regions were enriched for regions co-expressed with neuronal genes (FDR-corrected p-value = 1.20 x 10^{-2} in putamen), suggesting that i-eQTLs might be of functional relevance. Thus, this analysis provides evidence of the importance of capturing regulatory information in a reference-agnostic manner.

4.3.7 Implications of transcribed intergenic regions for Mendelian disorders.

In the last decade, Exome-Sequencing (Exome-Seq) has revolutionised the identification of pathogenic mutations for Mendelian disorders, contributing to the improvement of clinical diagnostics. However, this approach depends on the accuracy of gene annotation and if exons are "missing" from references, not all pathogenic variants are captured and sequenced, potentially leading to false negatives. Hence, the identification of unannotated regions with strong and moderate evidence of being part of a known gene could have a direct clinical application in Mendelian disorders since these additional genomic regions may also require screening for disease-causing variants.

The potential importance of transcribed intergenic regions with strong evidence of being part of a known gene was investigated by looking at how many Mendelian

genes would be affected. Of 132 genes associated through split reads with a transcribed intergenic region, 19 genes are already implicated in Mendelian disorders, as catalogued in the Online Mendelian Inherited in Man (OMIM) database.

If transcribed intergenic regions with moderate evidence of being part of a known gene are also included, the number of affected OMIM genes increases to 41 genes. *PEX2* is an example of a Mendelian gene associated with a transcribed intergenic region with strong evidence of being part of a the gene (Figure 4.10). *PEX2* encodes a peroxisomal membrane protein and is associated with the autosomal recessive condition, Zellweger syndrome. Given that a significant proportion of individuals with probable Zellweger syndrome do not receive a molecular diagnosis, the transcribed intergenic region identified could represent a novel genomic region to screen for pathogenic mutations in undiagnosed patients.

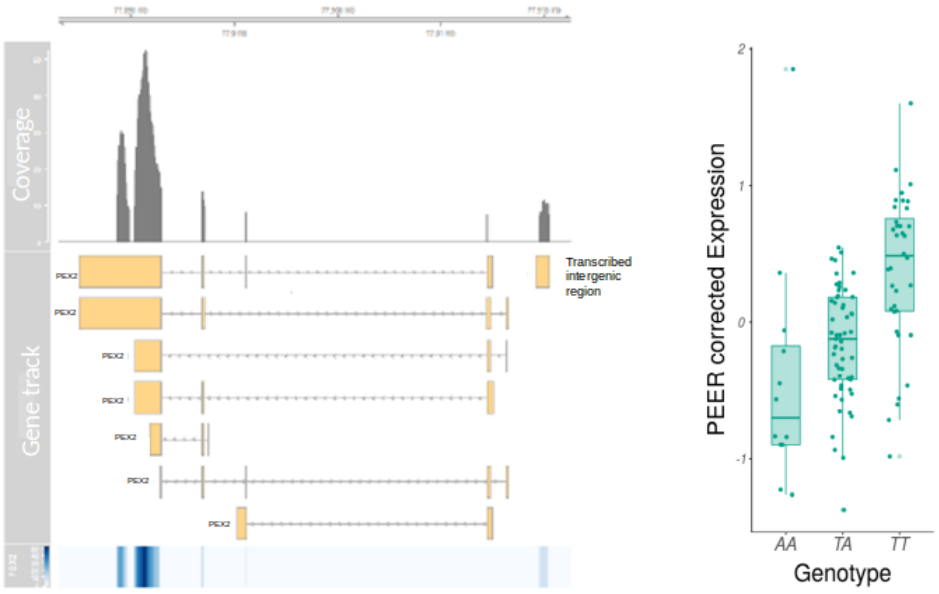

Figure 4.10: **Relevance of transcribed intergenic regions for OMIM genes.** Visualisation of genomic annotations for transcribed intergenic region (DER21123) associated and PEX2. i-eQTL for the rs35877910 tagging the DER21123 region.

4.3.8 Use of i-eQTLs to understand complex diseases

The relevance of i-eQTLs and their target regions was also investigated in the context of complex diseases. First, the overlap between all i-eQTLs and all risk SNPs within the NHGRI catalogue (as in section 3.2.9) was measured. This demonstrated that 7.8% of all i-eQTLs were listed as a risk SNP in the STOPGAP database. Of these i-eQTLs, 6.0% could not be identified through an eQTL tagging an annotated region (i.e. ge-eQTL, e-eQTL and ex-ex-eQTL). Therefore, i-eQTLs provide novel information for interpretation of GWAS hits. Furthermore, using the same approach that was used for ge-eQTLs (section 3.3.5) I found a significant enrichment of risk loci associated with adult neurological disorders (Fisher Exact test p-value 1.29×10^{-7}) as compared with all other phenotypes present in the STOPGAP catalogue amongst i-eQTL. This suggests that i-eQTL are also relevant for the interpretation of adult neurological disorders.

The explanatory potential of i-eQTLs was investigated in more detail in for late onset Alzheimer's Disease (AD, Lambert et al. 2013), Parkinson's Disease (PD, Nalls et al. 2014) and Schizophrenia (Ripke et al. 2014). The aim of this analysis was to compare the information content of i-eQTLs with eQTLs related to annotated regions. I found no evidence of enrichment of low p-values for Parkinson's Disease GWAS (Figure4.11 a). However, i-eQTLs were enriched for low p-values in the Alzheimer's Disease GWAS (figure4.11 b). This enrichment was more apparent amongst i-eQTLs as compared to conventional eQTLs (ge-eQTL, e-eQTL, ex-ex-eQTL). Furthermore, i-eQTLs were also enriched for low p-values in the schizophrenia GWAS. In this case, the enrichment was similar comparing across all types of eQTLs (figure4.11c). This would suggest that i-eQTL target regions might be playing an important role in specific diseases.

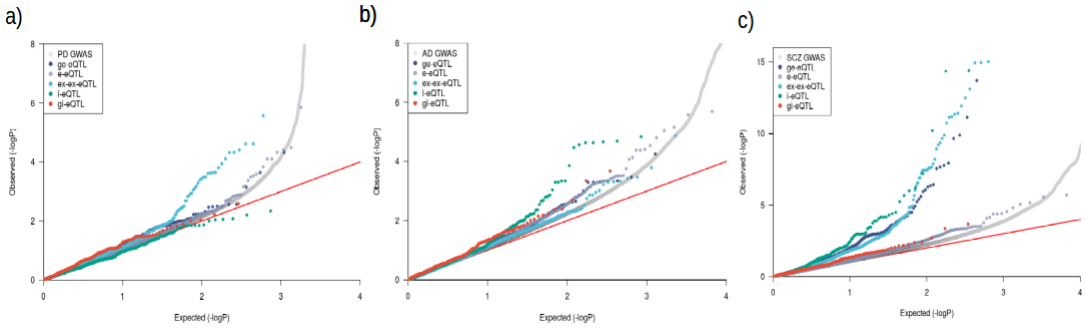

Figure 4.11: a) Q-Q plots of Parkinson's Disease GWAS p-values for ge-eQTL,ex-ex-eQTL,e-eQTL and i-eQTL. b) Q-Q plots of Alzheimer's Disease GWAS p-values for ge-eQTL,ex-ex-eQTL,e-eQTL and i-eQTL. c) Q-Q plots of Schizophrenia GWAS p-values for ge-eQTL,ex-ex-eQTL,e-eQTL and i-eQTL.

I investigated the importance of i-eQTLs in a locus-specific manner using the coloc software (Giambartolomei et al. 2014). I found that twenty-two i-eQTL signals colocalised with risk loci for schizophrenia (Table 4.2). An interesting example is the SNX19 locus, which has already been reported in schizophrenia (Zhu et al. 2016; Fullard et al. 2017) with the lead SNP rs35774874 (GWAS p-value = 1.97 x 10^{-11}). This risk locus colocalised with an i-eQTL targeting a transcribed intergenic region with moderate evidence for being part of SNX19 and located beyond the existing 3'UTR of the gene.

GWAS	Tissue	Region ID	Signif coloc	P-value GWAS	P-value eQTL	PPH3+ PPH4	PPH4/ PPH3	Region location	Gene symbol	Annotation
SCZ	PUTM	DER38322	rs4375688	2.79E-06	3.54E-05	0.81	2.38	chr17:36921941-36926024	PIP4K2B	Strong
SCZ	SNIG	DER19223	rs35774874	1.97E-11	8.58E-06	0.90	4.78	chr11:130743677-130748449	SNX19	Strong
SCZ	PUTM	DER36302	rs8025070	9.41E-10	7.94E-10	1.00	4.58	chr15:84833811-84833975	RP13-262C2.3	Weak
SCZ	PUTM	DER38467	rs113316734	1.53E-05	2.24E-13	0.90	2.40	chr17:44570943-44572432	NSFP1	Weak
SCZ	SNIG	DER24090	rs117124984	1.53E-05	2.66E-06	0.82	2.54	chr17:44570977-44571847	NSFP1	Weak
SCZ	SNIG	DER24091	rs169201	1.53E-05	1.84E-07	0.87	3.03	chr17:44571865-44572339	NSFP1	Weak
SCZ	PUTM	DER14702	rs3733710	8.66E-07	1.27E-06	0.96	2.86	chr5:140086610-140087096	ZMAT2	Moderate
SCZ	PUTM	DER35699	rs572837	4.98E-07	5.27E-08	1.00	10.31	chr15:43823997-43824377	MAP1A	Moderate
SCZ	PUTM	DER38474	rs117124984	1.53E-05	8.26E-10	0.90	2.47	chr17:44578803-44579329	RP11-995C19.2	Moderate
SCZ	PUTM	DER38475	rs117124984	1.53E-05	7.80E-06	0.82	2.58	chr17:44579356-44579762	RP11-995C19.2	Moderate
SCZ	PUTM	DER38461	rs199504	1.53E-05	7.45E-10	0.89	2.24	chr17:44568545-44568754	NSFP1	Uncharacterised
SCZ	PUTM	DER38464	rs117124984	1.53E-05	3.85E-08	0.90	2.54	chr17:44569476-44570032	NSFP1	Uncharacterised
SCZ	PUTM	DER4554	rs56145559	1.01E-09	5.49E-06	0.91	4.62	chr2:73949171-73949724	TPRKB	Uncharacterised
SCZ	PUTM	DER6148	rs908671	6.86E-07	2.33E-09	0.97	2.21	chr2:172619707-172620917	AC068039.4	Uncharacterised
SCZ	SNIG	DER4031	rs13017585	6.86E-07	3.12E-08	0.97	2.30	chr2:172620399-172620883	AC068039.4	Uncharacterised

Table 4.2: Summary of GWAS i-eQTL co-localisation hits.

The co-localisation results also highlighted i-eQTL that were targeting unannotated regions that were not linked to a known genes, fourteen out of twenty-two co-localising i-eQTL were targeting these independent intergenic regions. Amongst these, the region DER36302 (chr15:84165059- 84165223, eQTL p-value = 1.15 x 10^{-10} in putamen) that co-localises with the schizophrenia risk SNP rs950169 (GWAS p-value = 7.62 x 10^{-11}) and has its highest expression in two brain regions relevant to schizophrenia and represents a novel genomic location to be explored. This reinforces the importance of performing eQTL analysis using a reference-free approach to identify novel candidates genes/regions.

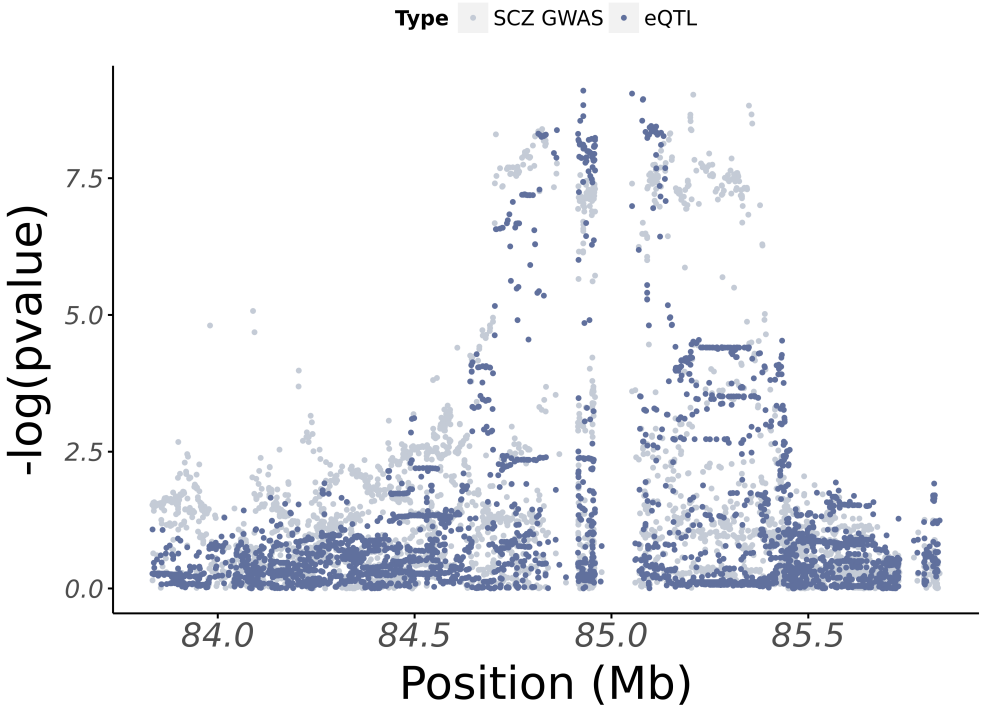

Figure 4.12: **Schizophrenia co-localisation Example.** Co-localisation of the i-eQTL targeting transcribed intergenic region DER36302 with schizophrenia GWAS lead SNP rs950169.

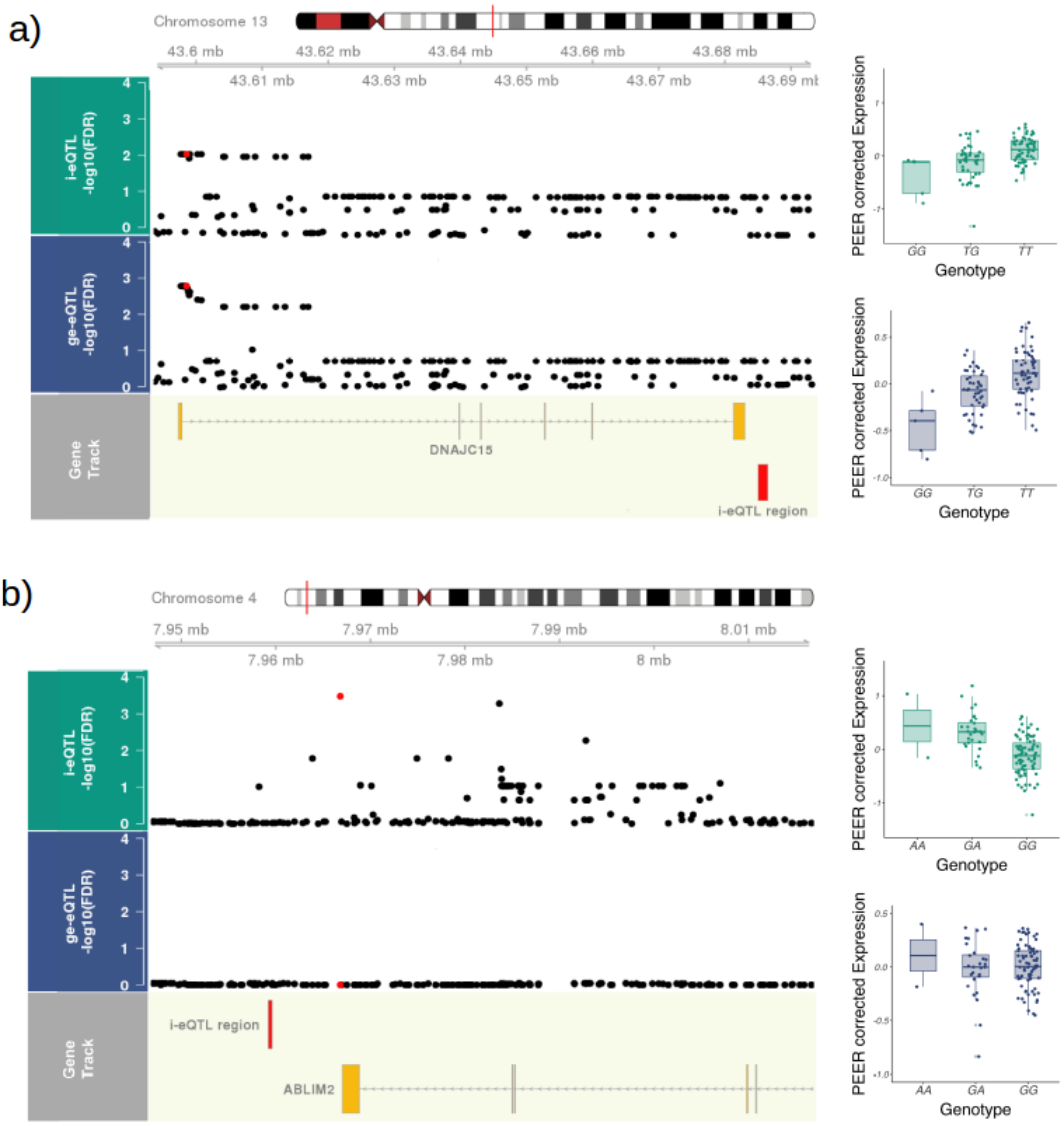

Figure 4.8: **Example of i-eQTL sharing signals with eQTL targeting annotated exons. a)** Association of local variants and rs113317084 (red points), with the expression of the transcribed intergenic region DER32583 (green track) and gene-level expression of DNAJC15 (blue track). **b)** Association of local variants and specifically rs4696709 (red points), with the expression of the transcribed intergenic region DER10633 (green track) and gene-level expression of ABLIM2 (blue track). DER10633 is classified as having strong evidence of being part of ABLIM2 supported by split reads, linking the DER10633 with ABLIM2.

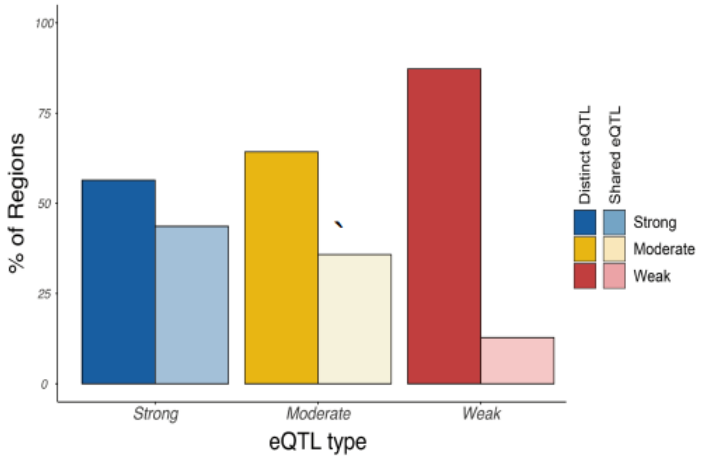

Figure 4.9: **i-eQTL sharing signals with eQTL targeting annotated exons.** Barplot of the beta-heterogeneity test of known exons and all characterised regions separated by significance. All signals with an FDR-corrected p-value of $< 5\%$ using a beta-heterogeneity test were considered distinct (opaque bars), while an FDR-corrected p-value $> 5\%$ was taken as evidence of eQTL sharing (transparent bars).

4.4 Discussion

The complexity of gene expression and splicing in brain make gene annotation in this tissue a challenging task. The recognition of this issue has led me towards the analysis of eQTL effect in a reference-free manner. Using this approach, I identified the existence of transcribed intergenic regions, which represented 6.6-12.5% of the total transcription in the human putamen and substantia nigra. While a comparison with an updated release of annotation reference GENCODE (version 19) indicated that 5% of detected regions were incorporated into the amended reference, replication of these regions in independent datasets was very high (75.4%) suggesting references remain inaccurate. Further validation experiment using RT-PCR and Sanger sequencing, confirmed the expression of a selection of regions and highlighted the reliability of the split read information detected in RNA-Seq data. Toghether these findings suggest that the transcribed intergenic regions identified are most likely to be the result of incomplete annotation of the human brain transcriptome.

Using a reference-free approach I was able to obtain information on the regulation of transcribed intergenic regions, termed i-eQTLs. Of these, a significant proportion represent novel putative regulatory information as compared to reference-based eQTL analysis. This implies that incomplete annotation of the human brain transcriptome is limiting eQTL discovery.

I designed a pipeline to characterise i-eQTL target regions in relation to known genes, on the assumption that I was unlikely to discover entirely new genes. The replication of characterised i-eQTL target regions suggested that transcribed regions with evidence of splicing (as indicated by the presence of split reads) were most reliable. This suggests that split reads while representing a small proportion of the data in a short read RNA-Seq dataset, have an extremely high information content (e.g. splicing events) as well as their reliability across independent datasets. However, 19.9% of i-

eQTL target regions had evidences of splicing information, of which 16.1% were linked to a known gene. This percentage is expected to increase, because I was limited by the quantity of split reads, which form only a small proportion of all RNA-Seq data. In addition my analysis did not include intragenic expressed regions, which might be expected to be more likely to represent novel exons of known genes.

Given that i-eQTL target regions have the potential to be functional, I therefore explored their disease relevance. Currently, there are large numbers of patients suffering from probable Mendelian disorders that do not receive a genetic diagnosis, 50-75% remain without diagnosis (Taylor et al. 2015), by conventional genetic screening, the detection of novel transcribed regions (whether associated with known genes or not) could be important diagnostically. The potential of this approach is demonstrated by the identification of several potential novel exons for genes which are already implicated in Mendelian disorders

Lastly, implications of i-eQTL signals on GWAS hits interpretation was investigated. This showed that GWAS risk loci for neurological and behavioral disorders were enriched within the i-eQTL dataset and that i-eQTL represented additional information for GWAS interpretation that was not identified through reference-based eQTL analyses. Furthermore, by means of colocalisation analysis I have demonstrated that i-eQTL provided insights for specific schizophrenia risk loci. I have highlighted a GWAS locus for schizophrenia that colocalise with an i-eQTL that targets a transcribed intergenic region that appears to be part of an already known candidate gene for shizophrenia.

In conclusion, while the analysis methods utilised to identify i-eQTLs and and characterise target regions, have significant limitations, they overcome some of the current limitations in gene annotation and represent a valuable source of information for the neuroscience community. Particurlarly for those researchers focusing on gene discovery and those focusing on the function of a specific gene.

Chapter 5

Hippocampus analysis

5.1 Introduction

The work described in Chapter 4 suggests that novel transcription can be classified into two major forms: i) novel transcription which appears to be independent of any known gene and could represent a new gene or potentially novel transcription of another form, such as expressed enhancers (Kim et al. 2010), and ii) novel transcription which is associated with a known gene and is likely to be an indication of incomplete annotation of the gene and all its isoforms. The data described in Chapter 4 suggests that the latter, namely incomplete annotation, occurs more frequently. This chapter follows on this work by exploring further the impact of splicing and misannotation of known genes in brain by using both the qualitative and the quantitative information contained in RNA-Seq data.

RNA-Seq data is most commonly used to quantify transcription in pre-defined genomic ranges. This type of analysis heavily relies on the quality of existing annotation and provides limited information on RNA processing, including splicing. While it the most standard way in which RNA-Seq data is used, RNA-Seq data can also provide qualitative information relating to RNA structure as exemplified by the detection of

circular RNAs, gene-fusions and splicing events. However, this information requires more sophisticated forms of analysis and although the methods to capture quantitative information from RNA-Seq data are well established (Anders, Reyes, and Huber 2012; Anders, Pyl, and Huber 2014; Ongen and Dermitzakis 2015; Collado-Torres et al. 2017a), there are fewer methods available to capture qualitative information and they tend to have only been used by a few specific studies.

This disparity in the availability of well-tested methods has arisen because extracting RNA structural information, including splicing information from short-read RNA-Seq data is a challenging task. Several approaches primarily using read depth across exonic regions to infer RNA structure have been developed (Anders, Reyes, and Huber 2012; Trapnell et al. 2013; Pertea et al. 2015; Frazee et al. 2015). While these methods make use of all data available, they tend to be inaccurate when the information to be extracted reaches a higher level of complexity (Hu et al. 2018), as in the case of a gene with multiple expressed transcripts. Other approaches have put greater emphasis on the detection of split reads (Li et al. 2018). As described in section 4.2.1.1, split reads in RNA-Seq data are reads that have a gapped alignment to the genome (e.g. reads mapping to exon-exon junctions). Since they commonly arise due to the process of RNA splicing, they provide key structural information.

While split reads are generally a small fraction of an RNA-Seq dataset, split reads contain accurate and reliable qualitative information. This is well documented by Nellore and colleagues (Nellore et al. 2016a), who investigated splicing diversity by means of split read analysis across 21,504 RNA-Seq samples. They found high replication of split reads across datasets, even using several detection methods. Furthermore, the reliability of split reads has led to this data feature being used to assess the performance of gene reconstruction softwares (Engström et al. 2013).

Split reads have also been used to identify circular-RNAs, detected by using inverted mapped split reads in RNA-Seq samples (Zhang et al. 2014; Gao, Wang, and Zhao

2015), gene-fusions through split reads connecting two known genes (Liu et al. 2013; Zhao et al. 2017), and insertion and deletions (Sun et al. 2016). However, split reads are primarily used to understand variation in splicing (Schreiber et al. 2015; Nellore et al. 2016a; Kremer et al. 2017; Cummings et al. 2015).

In this chapter, I combine an annotation-free approach to quantify transcription with split read data to investigate gene expression and splicing in the human hippocampus. This tissue was selected for this analysis because of its key role in a range of neurological and neuropsychiatric diseases, including Alzheimer's disease and schizophrenia, but also because of its importance in memory and cognition. Given that unannotated transcribed regions might be expected to be enriched for alternatively spliced exons important for complex neurological functions, I hypothesised that hippocampal transcriptomic data could potentially yield more novel findings in this context. I use the expression features I generate as the basis for an eQTL analysis and compare the output to more standard, publicly available approaches. Finally, I investigate whether the annotation-free approach developed in this thesis can generate novel insights into neurological and neuropsychiatric diseases.

5.2 Methods

The generation and quality control of the hippocampus RNA-Seq data is described in detail in section 2.4.

5.2.1 Imputation

The initial standard quality checks of the genotype calls were performed by Dr Adaikalavan Ramasamy (Principal Investigator (Systems Biology), Singapore Institute for Clinical Sciences, A-STAR). Consisted of two types of quality checks. Firstly, quality check on variants were applied; variants containing a score call equals to 0 for all individuals

were filtered out. Secondly, quality check on individuals was applied in the following order; 1) Gender check which identified three mismatched due to mislabeling errors, this was corrected and all individuals were included in the analysis. 2) A familial relatedness check was applied using a identity-by-descent (IBD) threshold of 0.1875, which corresponds to second or third degree of relatives, none of the individuals were removed using this cutoff. 3) Identify population outliers using principal component analysis, three individuals were removed because they were suspected to have Mexican and Chinese ancestry. 4) Genotype call rate filtering of 5%. 5) Since the genotype dataset is the aglomeration of two chip-array platforms (Infinium Omni1-Quad BeadChip and Immunochip) redundant genotype sites were removed. These quality control steps left a total of 786,186 markers.

I then subsetted the genotyped calls to the 101 individuals with RNA-Seq samples and I carried out the following steps, as suggested in the sanger imputation service: 1) I obtained the marker's frequencies using the plink software and then, using Will Rayner's script (http://www.well.ox.ac.uk/~wrayner/tools/#Checking), the reference allele for the markers was harmonised to the match the reference allele of the Haplotype Reference Consortium (HRC) v1.1 Panel. The reference allele of 542,996 SNPs were changed as well as 253 that had the strand inverted. Furthermore, a total of 5,658 were excluded because the allele did not match the refence panel or the SNP was not included in the HRC v1.1 panel.

2) An additional step was added to remove SNPs which can demonstrate strand ambiguity (genotypes G>C, C>G, A>T and T>A). This step removed 21145 SNPs. 3) Retain SNP with MAF>5% and HWE of 0.0001. 4) Sex checks confirmed. The input file used in the Sanger Imputation service contained 759,382 SNPs.

5.2.2 Spliced Transcripts Alignment to a Reference (STAR)

The STAR (v. 2.4.0.1) (Dobin et al. 2013) aligner was used for read alignment because of its sensitivity in identifying split reads (Engström et al. 2013). STAR uses a 2-step approach to increase the split read identification.

In order to accelerate the mapping process to the reference genome I created a STAR index. The index was built using both the genome and transcriptome references downloaded from ENSEMBL database version 87 which corresponds to the human genome reference GRCh38. The first pass of STAR was run with a maximum of 2 mismatches allowed per read and only uniquely mapped reads were retained. Split reads with gaps spanning $\leqslant$ 20bp were considered indels and those with gaps of over 1Mb long were filtered out. Mate paired-end reads aligning more than 1Mb from each other were removed. Furthermore, overhang sequences around the acceptor and donor sites of split reads where selected as per ENCODE standard RNA-Seq pipeline. A minimum of 3 bp overhang was required for split reads matching precisely annotated junctions and a minimum of 8 bp overhang was required for unannotated split reads, with all other split reads removed. I then collected split reads for all samples and removed split reads present $< 5\%$ of samples and all those mapping to the mitochondrial genome. Consequently, a total of 145,751 novel split reads were additionally included to regenerate the STAR index that was used for the second pass alignment.

5.2.3 Quantification

5.2.3.1 Quantification of gene and transcript expression

Transcript quantification was performed using Salmon (v0.8.2) (Patro et al. 2017). I used the "lightweight" alignment approach, which maps the raw sequencing reads directly to the reference transcriptome. Transcripts were quantified using the ENSEMBL reference version 87 and Salmon was run with using options for non-stranded RNA-Seq

libraries, correction for possible 3' fragment positional bias in the RNA-Seq data (Love, Hogenesch, and Irizarry 2016) and withmodeling of GC content bias in order to diminish systematic errors in the transcript quantification estimates (Roberts et al. 2011). I used the Salmon's variational Bayesian Expectation-Maximisation algorithm method to obtain optimised transcript quantification. Finally, a bootstrapping option using a resampling approach was used to account for technical variantion in the abundance estimates. Salmon produced transcript-specific expression values (measured in Transcripts Per Kilobase Million, TPM) values and these were used for the eQTL analysis.

The R package tximport (Soneson, Love, and Robinson 2015) was used to transform Salmon transcript-level quantification to gene-level quantification values. Transcript-level and gene-level quantifications were then filtered by using TPM > 0.3 in at least 80% of the samples, quantile normalisation using the R function normalize.quantiles() was applied to both quantification levels and log transformation was not applied. The TPM-normalised expression profiles of 37,114 transcripts and 20,179 genes were used in subsequent analyses.

5.2.4 Non-reference-based quantification

5.2.4.1 Quantification of alternative splicing

Intron excision ratios were calculated using LeafCutter (Li et al. 2018) or both annotated and unannotated splicing events and used as a measure of splicing. BAM files generated using STAR in section 5.2.2 were converted to an exon-exon spanning junction file format. The exon-exon spanning junction file was used to define intron clusters with alternative intron excision (Figure 5.1) using the threshold of $\geq$50 split reads per intron cluster and maximum intron length of 1Mb. LeafCutter filters out introns present in $<$40% of samples or with small or no variation across samples. Finally, intron excision ratios were standardised across individuals and quantile normalised.

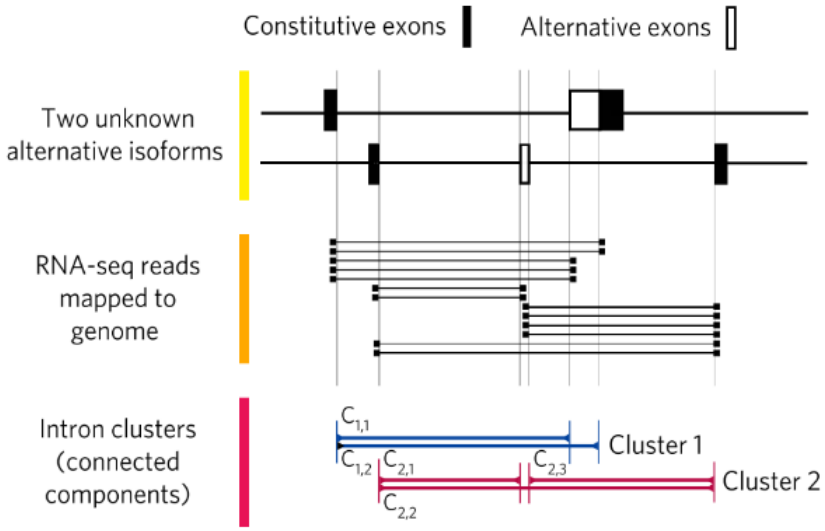

Figure 5.1: **LeafCutter, example intron clusters.** LeafCutter uses split reads to generate clusters of alternative intron excision. In this example, by using five different split reads leafcutter generates two different clusters amongst those split reads that share splice sites. Image adapted from Figure 1a, Li et al. 2018

5.2.4.2 Quantification of transcribed regions

I used the software R derfinder (Collado-Torres et al. 2015) as described in section 3.2.2.1, to quantify transcribed regions. However, this time the the methodology to define the regions was optimised.

Derfinder uses a mean coverage cut-off (MCC), defined as the minimum number of reads for each single base to be considered as transcribed, to define expressed regions (ERs). Once coordinates of ERs are generated, derfinder uses the max-region-gap (MRG), defined as the maximum gap in bases between two adjacent ERs to be merged into one ER. The mean coverage cut-off and the max-region gap parameters were optimised to improve the definition of ER in the hippocampus dataset and this work was performed in collaboration with David Zhang (PhD in Department of Neurodegenerative Disease UCL).

In order to reduce the impact of ambiguos mapping reads across overlapping exons,

a total of 156,339 non-overlapping exons from ENSEMBL v.87 were used as baseline for the optimisation step. Subsequently, the absolute coordinate difference (Δ) between the exon and the identified ER is calculated. Let *ER* be the expressed regions identified by the derfinder that overlaps the exons. Let *start* and *end* be the functions that return, respectively the most left and the most right genomic coordinates, thus for each ER i the Δ is calculated as follows:

$$\Delta i = |start(exon_{\mathrm{i}}) - start(ER_{\mathrm{i}})| + |end(exon_{\mathrm{i}}) - end(ER_{\mathrm{i}})|$$

$$\begin{cases} i = 1 \dots n \quad n \leq 156,339 \end{cases}$$

The Δ was calculated for 1,639 sets of ERs generated using 149 mean coverage cut-offs (from 0.2 to 15 stepping by 0.1) and 11 max-region-gaps (from 0 to 100 stepping by 10). Finally, the accuracy of each set of ER generated is summarised into the median Δ and the total number of ERs with $\Delta = 0$. While the median exon delta represents the overall difference between all ER definitions and what is considereded the "ground truth" (assuming non-overlapping exons are well defined), the number of ERs with $\Delta = 0$ provides the extent to which ERs precisely defined the non-overlapping exons.

The mean coverage cut-off and max-region-gap pair that generated the set of ERs with the lowest median exon delta and highest number of ERs with $\Delta = 0$ was chosen as the most accurate transcriptome definition for each tissue. This analysis suggested a cut-off of 3.3 and a max-region-gap of 10 (Figure 5.2).

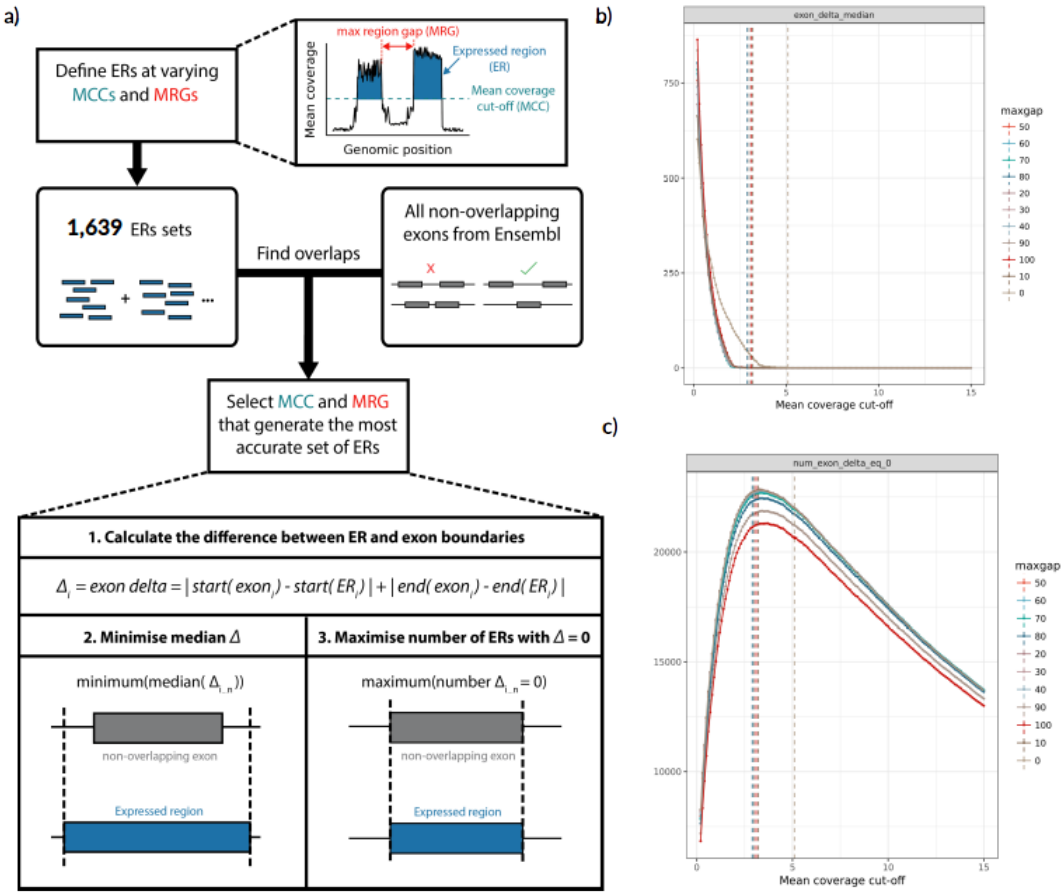

Figure 5.2: **Optimisation mean coverage cutoff (MCC) and max region gap (MRG) for detection of transcription. a)** Transcription is detected for hippocampus in annotation agnostic manner and generating ER. The MCC is the number of reads supporting each base above which that base would be considered transcribed and the max region gap (MRG) is the maximum number of bases between ERs below which adjacent ERs would be merged. The optimisation uses the non-overlapping exons from ENSEMBL v87 reference annotation to optimise the region definition. Courtesy David Zhang. **b)** Line plot to show the selection of the MCC and MRG that minimised the difference between ER and exon definitions (median exon Δ). **c)** Line plot to show the selection of the MCC and MRG that maximised the number of ERs that precisely matched exon definitions ($\Delta = 0$). Dark brown line represents the optimal for MCC (3.3) and MRG (10).

Following ER identification, ERs of < 3 base pairs in length were removed and the expression within the remainder of ERs was quantified. Regions were annotated based on existing annotation using the annotateRegions R function. The GC-content of each ER was calculated using the bedtools (v.2.27.1) and conditional quantile normalisation

was applied to all ERs. Finally, ERs were annotated on the basis of the existing annotation (ENSEMBL v. 87) using the annotateRegions R function.

5.2.5 Pipeline for the identification of eQTLs

5.2.5.1 Removal of batch effects and eQTL pipeline

I tested various approaches to see if it would be possible to improve the estimation and removal of undesiderable variance in expression profiles and so increase the power of eQTL discovery. I used the TPM-normalised gene-level expression profiles generated in section 5.2.3.1 as the input for PCA, Surrogate Variable Analysis (SVA, Leek et al. 2012) and PEER. For SVA and PEER the known factors age, gender, brain bank and the first 3 genetic principal component axes were included.

I also tested the Qualitative Surrogate Variable Analysis (qSVA) method developed by Andrew Jaffe and colleagues (Jaffe et al. 2017b). This method attempts to measure and correct for the effects of RNA degradation on RNA expression profiles specifically in human brain. This is done by measuring the impact of leaving post-mortem brain tissue (grey matter) at room temperature for between 0 and 60 minutes and sampling at regular intervals. At each time point RNA was extracted and sequenced and subsequently differentially expressed genomic regions were identified in a time series analyses. Expressed regions that were differentially expressed over time were considered a proxy for degradation. This experiment was performed using two types of library construction, namely polyA–selected libraries and ribodepletion libraries, the latter generated by using total RNA as input. This generated two sets of differentially expressed genomic regions. I tested both sets by quantifying the expression within these specific genomic regions using the R function getRegionCoverage from derfinder and then applied the R qSVA function and extracted the surrogate variables.

For each method used (PCA, SVA, PEER and qSVA), I performed an eQTL analysis using the MatrixEQTL R package and sequentially adding the axes to identify

the method and the number of axes that maximised eQTL discovery. This approach demonstrated the utility of using PEER with 19 covariates included (Figure 5.3).

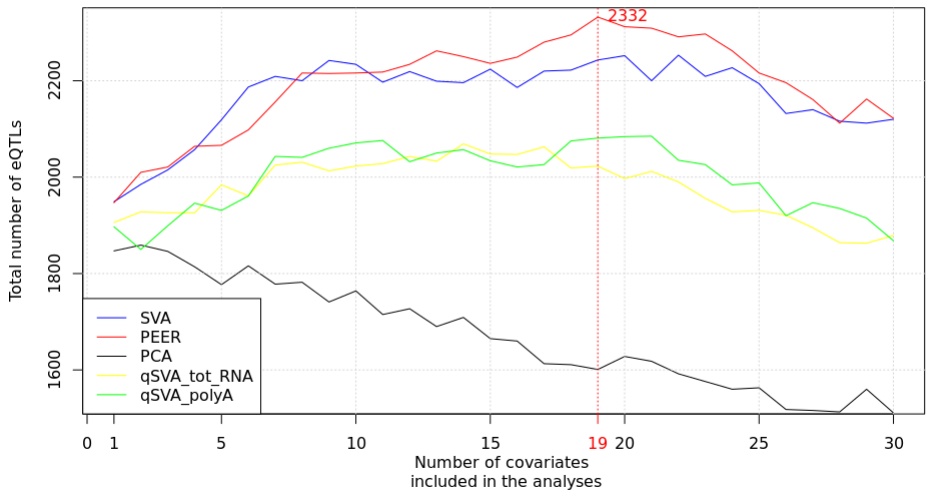

Figure 5.3: **Optimisation of batch effect removal using different methods.** Line plot to show number of eQTL signals identified (y-axis) and number of covariates included in the analysis (x-axis). Each line correspond to a different method to generate the covariates. The red dashed line represents the optimum number of covariates (19 PEER) to include in the analysis to obtain the maximum number of eQTL signals.

Thus, 19 PEER covariates were used in addition to gender, age, brain bank and the first 3 genetic principal component to perform eQTL for all quantification types. Finally, conditional analysis were applied as per section 3.2.3.4 to identify secondary signals.

5.2.6 Replication of eQTL signals in independent datasets

I assessed the replication of eQTL signals in the paired genotyping and RNA-seq data provided by NABEC. While this data set differed in terms of the tissue analysed (frontal cortex versus hippocampus) and the library construction protocol used (Illumina TruSeq Stranded Total RNA kit versus Nugen ovation kit), I had full access to the BAM files which allowed me to re-run the eQTL analysis through the same pipeline as the discov-

ery hippocampus dataset. Imputed genotyping data (using the HRC v.1.1) and aligned sequences RNA-Seq data were provided by Dr Raphael Gibbs (NIH). The aligned sequence data was used to quantify the transcribed regions identified in the hippocampus dataset. The same reference-agnostic eQTL pipeline was then used to test eQTLs passing an FDR threshold of 5% in the discovery hippocampus dataset.

5.2.7 Split read annotation

Split reads were annotated based on the location of the implied donor and acceptor sites and then assessed to identify unannotated split reads. I used the R packages refGenome and GRanges in combination with the transcriptome definition from ENSEMBL v.87 to annotate the split reads. Split reads with both ends present (including the combination of donor and acceptor sites) within ENSEMBL were labelled as "annotated". Split reads with a single end annotated were labelled as either "novel donor site" or "novel acceptor site" as appropriate. Split reads with both ends present within ENSEMBL, but where the precise combination of donor and acceptor sites was not present in the annotation were labeled as "novel splice site usage" reads. Finally, split reads for which neither end was present within existing annotation were labelled as unannotated.

Finally, using the MaxEntScan software (Yeo and Burge 2004) I measured the predicted strength of all donor and acceptor sites implied by the split reads using the 9 and 23 base pair sequences respectively for the donor and the acceptor sites. The MaxEntScan software uses the Maximum Entropy Principle to model the frequency distributions of sequence motifs around a comprehensive dataset of known splice sites. This provides a means of scoring any given sequence motif in terms of the probability of it being a true splice site.

5.2.8 Split read replication

I downloaded split reads relating to all samples included within GTEx v6 (covering 54 human tissues) from the recount2 resource. For each tissue, replication was considered to be successful if 5% of the samples from a given tissue contained split reads with precisely match the same genomic coordinates for the implied intron.

5.3 Results

5.3.1 Misannotation is prevalent in intragenic regions

In order to assess gene misannotation in the human hippocampus, I used three different quantification methods. The output of Derfinder was used to detect genomic regions with transcriptional activity, while the outputs of STAR-derived split reads and LeafCutter were used to detect splicing events. Derfinder identified 348,121 expressed regions (ERs) totaling 130Mb of sequence. STAR detected 351,781 splicing events, defined as unique split reads present in >5% of samples. LeafCutter identified 65,932 splicing events, 91.0% of which were detected by STAR. Unique splicing events identified by STAR and LeafCutter were merged into a single dataset.

Subsequently, ERs and splicing events were annotated using the ENSEMBL v87 gene reference. ERs were subdivided into five different classes (Figure 5.4a): 1) Exonic, ERs exclusively overlapping exons; 2) Intronic, ERs exclusively overlapping introns; 3) Intergenic, ERs exclusively overlapping intergenic regions; 4) Exonic-intronic, ERs overlapping both an exon and an intron; and 5) Exonic-intergenic, ERs overlapping both an exon and an intergenic region. Similarly, splicing events identified using STAR and LeafCutter were divided into three classes (Figure 5.4a): 1) Novel alternative splice site usage, such that both ends of the split read are annotated but the combination of the two exons is not within annotation; 2) Novel alternative donor or novel alternative

acceptor splice sites, such that either the donor or the acceptor sites are unannotated; 3) Novel intergenic splicing, such that both ends of the split read are unannotated.

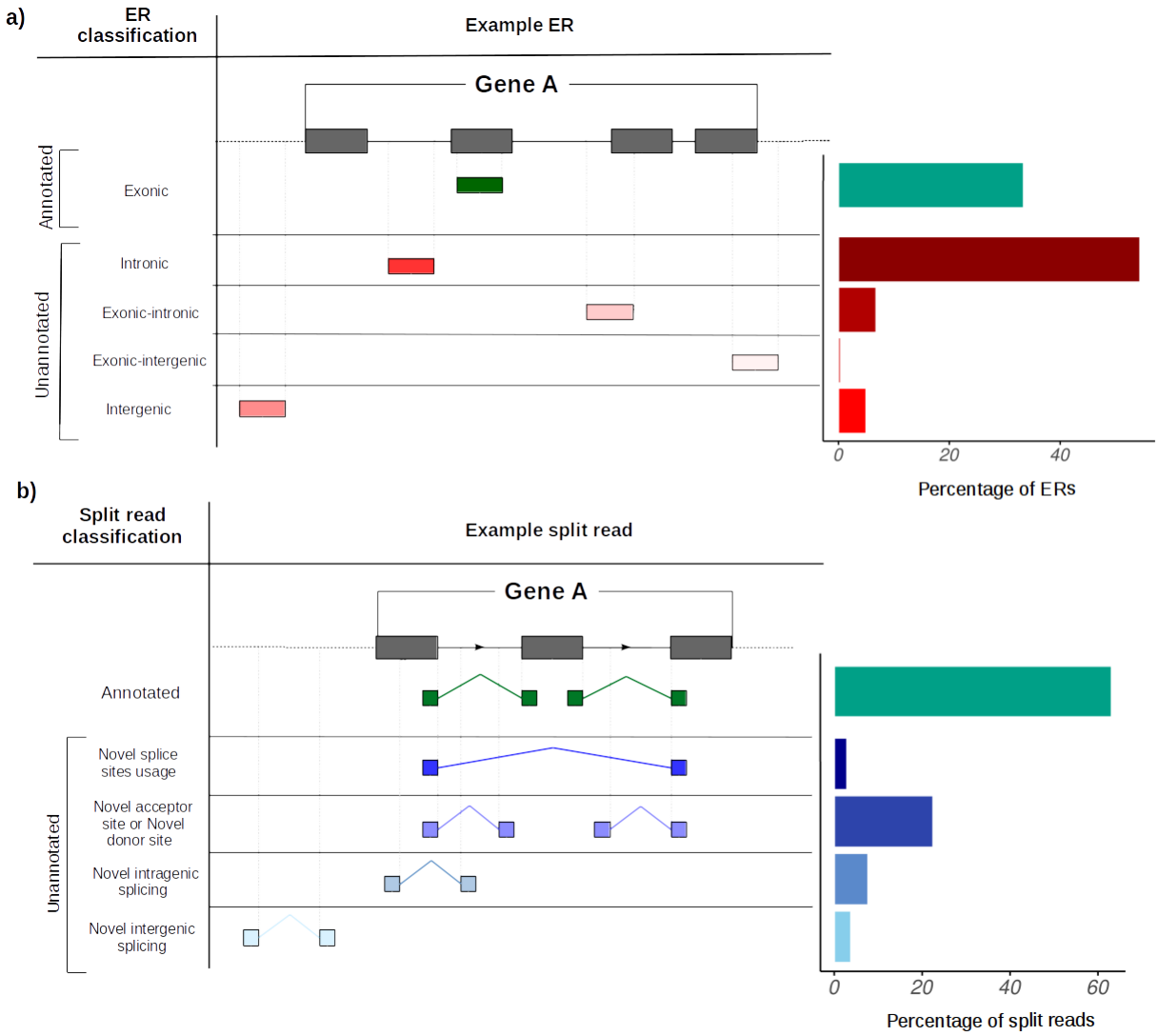

Figure 5.4: **Classification of ER and split reads.a)** Classification of ERs using ENSEMBL v87 reference annotation and bar-chart to show the percentage of ERs for each class. **b)** Classification of split reads using ENSEMBL v87 reference annotation and bar-chart to show the percentage of split reads for each class.

Using this classification system, I was able to measure the extent of unannotated transcription and splicing in the human hippocampus. Unannotated transcription accounted for 66.6% of ERs, corresponding to 99.9 Mb with the majority of unannotated

ERs classified as intronic (54.5% of ERs, Figure 5.4). Of all splicing events, 35.8% were unannotated with most of these events being either alternative donor or alternative acceptor type (Figure 5.4). While I recognise that the information from ERs is harder to interpret given that the library construction method used would be expected to capture pre-mRNA, the analysis of both ERs and splicing events suggests that misannotation is largely within intragenic regions and that alternative splicing may not have been fully characterised in the hippocampus.

5.3.2 ERs identified within introns are largely due to sequencing of pre-mRNA.

ERs were further analysed using split reads. As previously discussed split reads can provide evidence that RNA is not only transcribed from a specific region of the genome but is also processed by the splicing machinery. I found that 31.7% of the annotated ERs (i.e. those ERs which overlapped a known exon) had evidence of splicing. In comparison, 8.4% of unannotated ERs (intronic and intergenic ERs) showed evidence of splicing, of these 94.8% were connected to a known gene. However, only 2.6% of intronic ERs showed evidence of splicing. One explanation for this low level of overlap with splicing events amongst intronic ERs, is that these ERs are generated through reads mapping to pre-mRNA or have been generated by pervasive transcription. To further investigate the molecular basis of intronic ERs, I checked the replication of ERs within the NABEC and GTEx RNA-Seq datasets. While the presence of pre-mRNA fragments would be expected in the NABEC dataset because the library construction used total RNA and ribodepletion, pre-mRNA fragments would be expected to be relatively rare in the GTEx dataset because this data was generated using a polyA-selected library. Focusing on intronic ERs, the replication rate was high within the NABEC dataset (83.9%, Figure 5.5b), but low in the hippocampus GTEx dataset (4.5%, Figure 5.5a). This suggests that the majority of ERs represent pre-mRNA expression, but also

that pre-mRNA expression replicates when considering datasets generated using non-polyA-selected libraries. For all other types of ERs (exonic, intergenic, exon-intronic, exon-intergenic) replication rates were similar in GTEx and NABEC suggesting that these measures were not affected by the library construction method. In keeping with the existing literature, I found that all types of unannotated split reads, including those with no previously reported donor or acceptor splice sites were highly replicable with replication rates of 37.5–65.3% (Figure 5.5a) across all types of novel split reads within GTEx hippocampus. Furthermore, replication was largely robust to the library construction method and replication rates were 42.8-63.0% (Figure 5.5b) in NABEC.

In summary, this analysis demonstrated that while the detection of intronic ERs is due primarily to pre-mRNA sequencing, misannotation of the hippocampal transcriptome is still largely intragenic as evidenced by highly reliable split read data.

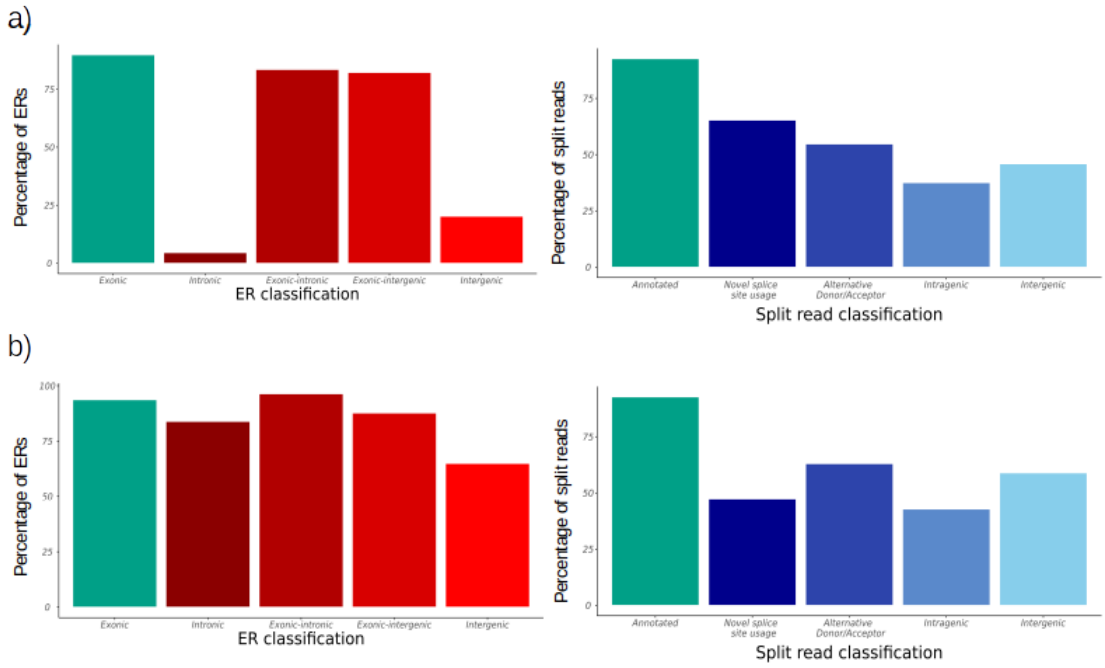

Figure 5.5: **Replication ER and split reads in independent datasets. a)** Bar-chart that illustrates replication of ERs and split reads stratified by classes using the GTEx dataset. **b)** Bar-chart that illustrates replication of ERs and split reads stratified by classes using the NABEC dataset.

5.3.2.1 Annotated and unannotated splice sites are similar in strength

To investigate the splicing efficiency of unannotated splice sites, the 5' and 3' splice site strengths were estimated using the MaxEntScan software (Yeo and Burge 2004) and compared to those estimated for annotated splice sites. The scores produced by the MaxEntScan software (MaxEnt score), were generated using the sequences around the splice site, with higher scores corresponding to higher probabilities for the sequence to be recognised as a splice site. Split reads classed as novel alternative splice site usage were removed from this analysis, because they do not represent novel splice site locations, but only novel usage of already annotated splice sites. For the same reason, the annotated splice sites for split reads classified as novel alternative acceptor and novel alternative donor were not included in this analysis. The distribution of MaxEnt scores for annotated and unannotated splice sites were highly overlapping at both the 5' and 3' sites. The largest differences in the distribution of scores was seen for novel alternative donor and the novel alternative acceptor sites as compared to the equivalent annotated sites (Figure 5.6). Interestingly, amongst unannotated split reads, those annotated as being entirely novel and located within intergenic regions had 5' and 3' MaxEnt score distributions which most closely resembled those of annotated splice sites (Figure 5.6). These findings suggest that the splice sites implied by unannotated split reads have the strength to be recognised by the splicing machinery.

5.3.3 Current quantification softwares do not fully capture the complexity of the hippocampus transcriptome

In order to measure transcriptome complexity, I used a range of different methods, all of which provide transcript-specific information either directly or indirectly. Transcript-specific information is provided by LeafCutter and split reads were detected by the STAR aligner. While Salmon quantifies annotated full-length transcripts, LeafCutter

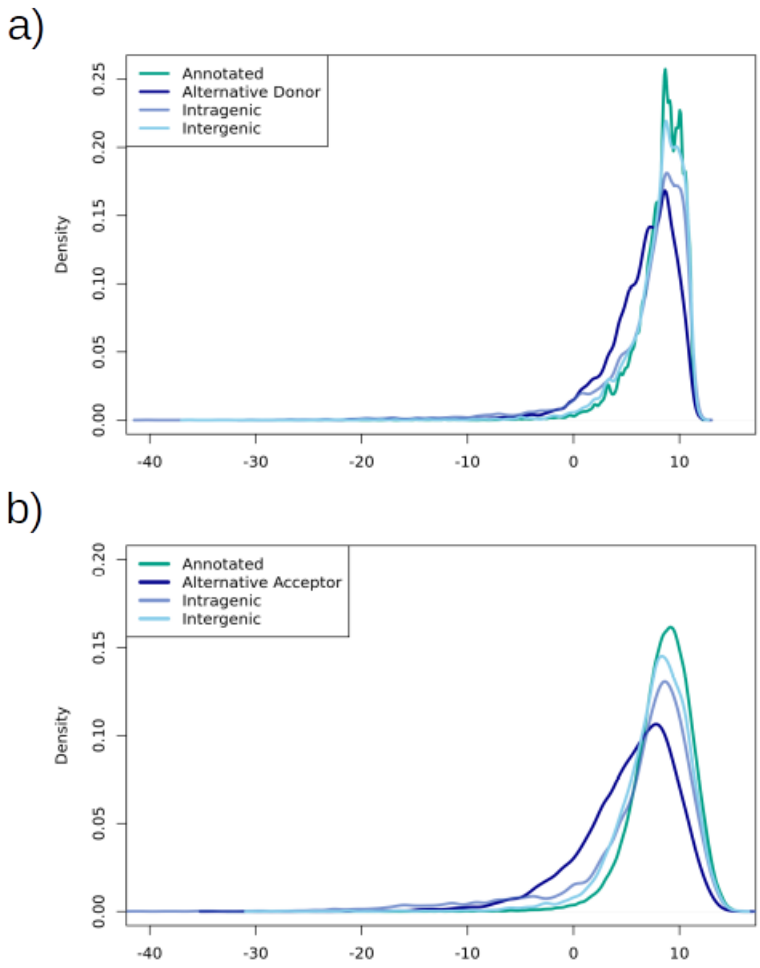

Figure 5.6: **5' and 3' splice site sequence strength. a)** Distributions of 5' splice site MaxEnt scores stratified by split read classes. **b)** Distributions of 3' splice site MaxEnt scores stratified by split read classes.

and split read data provide information about individual splicing events. Since they are annotation-agnostic in their approach, both methods can be used for the detection of novel splicing events.

In the first instance, I focused on the detection of splicing events already within annotation and quantified the number of unique ENSEMBL transcripts detected by each method. While Salmon provides this information directly, in the case of LeafCutter or split reads identified by STAR, this information had to be inferred. To this end, each splicing event (detected by both LeafCutter and the split reads) was annotated using the ENSEMBL v.87 transcript reference and filtered such that only splicing events, which

are unique to a single transcript were retained to generate a minimum set of transcripts that would be compatible with the data (Figure 5.7). This analysis demonstrated that split reads and Salmon detected 3.03 and 2.84 fold more transcripts than LeafCutter respectively (39,386 for split reads, 36,850 Salmon, 12,988 LeafCutter, Figure5.8).

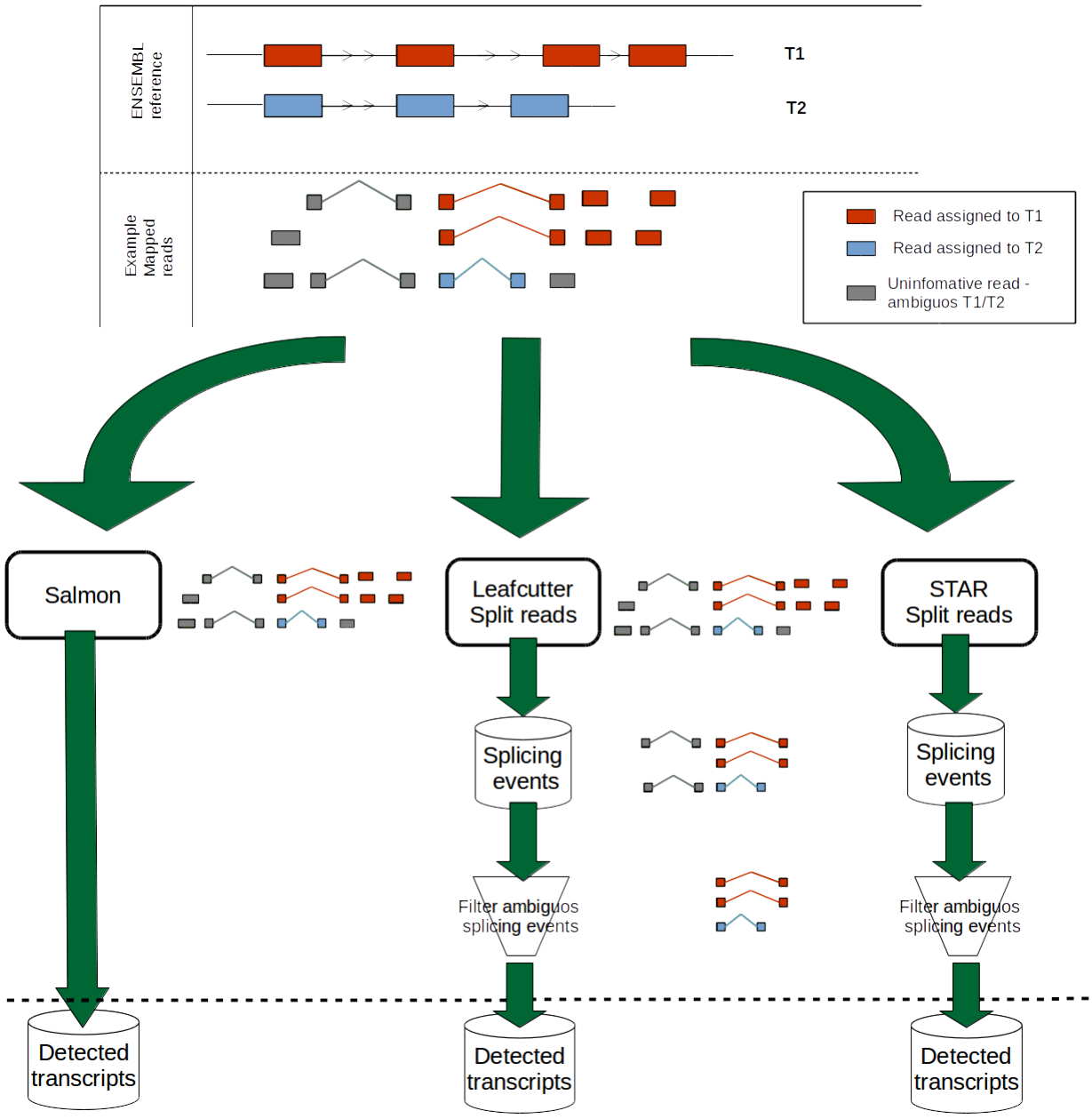

Figure 5.7: **Transcript detection workflow.** Schematic illustration to show the detection of transcript using different methods (Salmon, LeafCutter, split reads).

However, Salmon cannot detect novel splicing, which according to the data generated in Chapter 4 and several recent transcriptome studies is prevalent in human brain (Jaffe et al. 2015; Nellore et al. 2016b; Bartonicek et al. 2017; Clark et al. 2018; Jaffe et al. 2018).

Consistent with the current literature, both LeafCutter and the split read data suggested that novel splicing is widespread in hippocampus. LeafCutter, detected 3,648 novel splicing events, while the split read data identified 42,638 novel splicing events (Figure 5.8). The greater than ten-fold difference in the number of splicing events detected by these methods may be due in part to the fact that LeafCutter only detects splicing when a gene expresses at least two different transcript isoforms within the tissue.

Thus, this analysis suggests that existing methods to quantify splicing events do not capture the full complexity of the transcriptome, even when analyses are restricted to annotation and that while the scarcity of split reads may make quantification problematic, this feature of the RNA-Seq data may provide the most complete snapshot of splicing at present.

5.3.3.1 Evidence for brain-specific misannotation

Given the reliability of unannotated split reads detected in hippocampus, I investigated their tissue specificity with the aim of obtaining insights into functional relevance of the splicing events they capture. Using a subset of the GTEx data I determined the replication of the split reads identified in the hippocampus dataset within all 13 GTEx brain tissue and whole blood RNAseq data sets. A replication matrix of split reads across GTEx tissues was generated, assigning 1 if the split read was detected in the GTEx tissue and 0 otherwise (with replication defined as in section 5.2.8). Unsupervised hierarchical clustering was then performed using the Pearson's linear dissimilarity measure and applied to the replication matrix of split reads across the GTEx tissues.

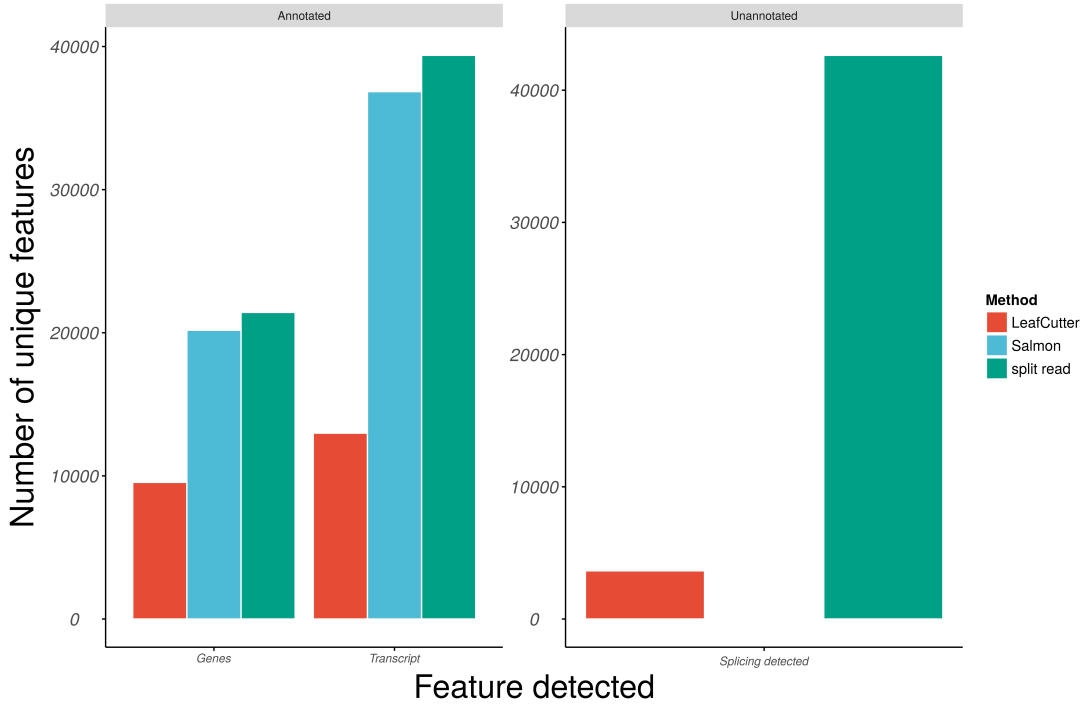

Figure 5.8: **Annotated and unannotated splicing events detected.** Left panel: bar-chart to show the number of detected genes and annotated splicing in the form of transcripts by the different methods. Right panel: bar-chart to show the detection of unannotated splicing events by the different methods.

This approach not only suggested that novel splicing, as measured by the split reads, is often specific to a brain region, but that splicing events also cluster in an anatomically expected manner (Figure 5.9), which would be consistent with functional importance. Furthermore, the low replication rate of split reads in whole blood samples suggests that the split reads are unlikely to be the product of systematic mapping errors that should be independent of the tissue sampled.

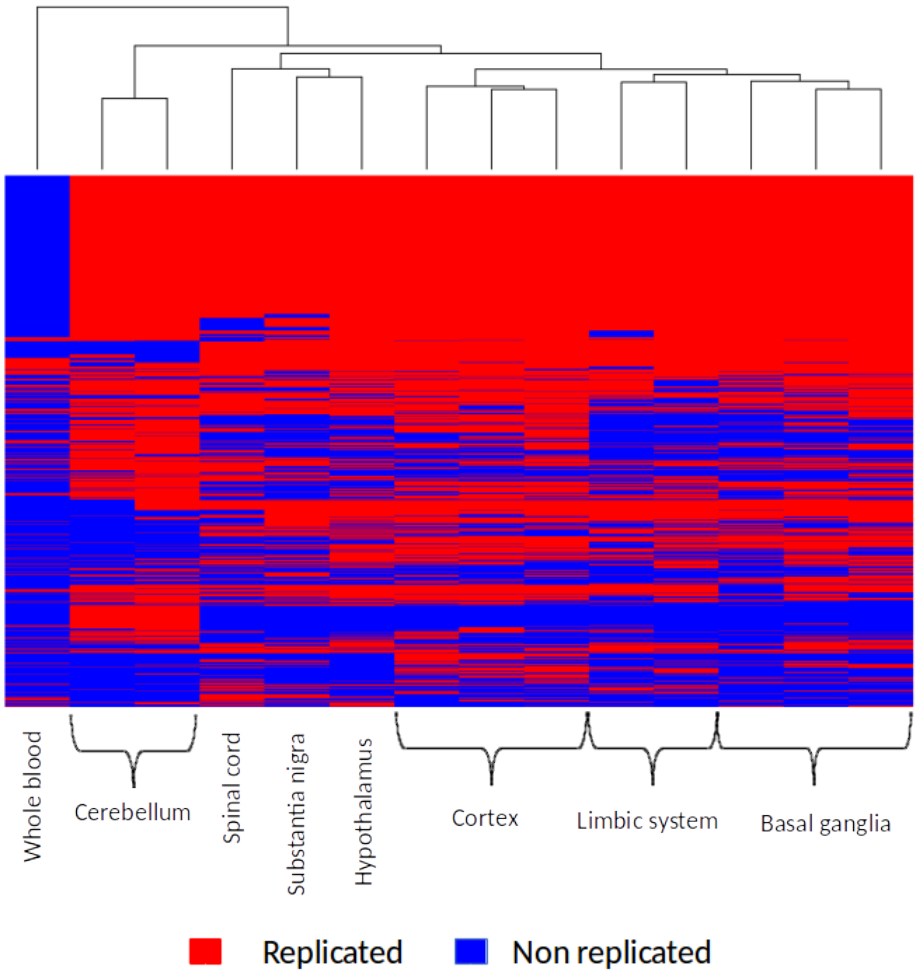

Figure 5.9: **Replication of unannotated split reads cluster in brain anatomically manner.** Heatmap and clustering of replication matrix for split reads in 13 brain samples and whole blood in GTEx.

5.3.4 eQTL discovery

While split reads represent a rich data resource within RNAseq datasets, they are relatively scarce (10.1% of aligned reads) and so an eQTL discovery strategy focused on this portion of the data would discard a significant amount of useful information. By contrast, an eQTL discovery strategy based on derfinder and the quantification of ERs is highly efficient in the sense that all mapping data is used, but risks being inaccurate and difficult to interpret. Therefore, I postulated that an approach, which integrated

both forms of data would be the most informative and could be more explanatory for disease than existing eQTL discovery pipelines based on LeafCutter or Salmon. I tested this by performing eQTL mapping analyses using 5,464,310 SNPs and a total of 471,153 expression features. The expression features were divided into four types, namely gene-level quantification (based on Salmon output), transcript-level quantification (based on Salmon output), splicing quantification (based on LeafCutter output) and ER quantification (based on derfinder output) to generate four different eQTL classes: Gene eQTL (ge-eQTL), transcript eQTL (t-eQTL), splicing eQTL (s-eQTL), expressed region eQTL (er-eQTL). These eQTLs had varying dependence on existing annotation with s-eQTLs having the potential to identify regulatory SNPs targeting novel splicing and er-eQTLs being entirely annotation-agnostic in nature. This resulted in 1,8 billion tests being performed and resulted in the identification of 232,199 eQTLs at 5% FDR after conditional analysis (Table 5.1).

Quantification	**Features tested**	**eQTL signals**	**eQTL unique target features**
Gene (Salmon)	20,179	2,332	2,175
Transcript (Salmon)	37,114	2,419	2,212
Splicing (LeafCutter)	65,739	4,966	4,727
Expressed Region (derfinder)	348,121	24,965	23,749

Table 5.1: **Summary of eQTL discovery.** Table to show a summary of eQTL discovery divided by features tested. Number of unique target features, represents the number of unique eQTL target features for each eQTL class.

Use of entirely reference-based approaches (ge-eQTL and t-eQTL) resulted in the identification of 4,748 independent eQTL signals targeting 2,337 protein coding genes. However, using reference-free approaches (s-eQTL and er-eQTL) resulted in the identification of a substantially larger number of eQTL signals (29,931, an increase of 6.3 fold more eQTLs) targeting a larger proportion of protein coding genes (14,767 genes equating to 66.2% of all protein coding genes). Furthermore, I found that eQTL signals targeting unannotated features represented a large proportion of the eQTL data.

In fact, 68.1% of all er-eQTL were targeting unannotated ERs of which 50.1% were intronic and 7.8% were intergenic (Figure 5.10).

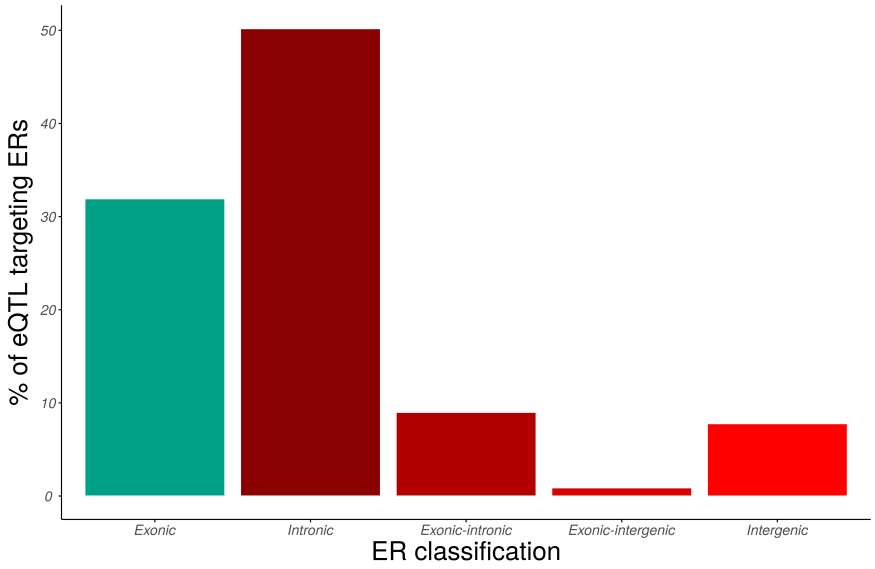

Figure 5.10: **Er-eQTL target ER proportions.** Bar-Chart to show the percentage of er-eQTL targets stratified by ER classes

5.3.4.1 Replication of eQTL

Given that the largest proportion of eQTLs identified in my analysis were generated through the use of the annotation-agnostic quantification method, derfinder, to produce er-eQTLs, I focused my replication analysis on this portion of the data. I used the NABEC dataset to perform this analysis. This is because like the hippocampus RNA-Seq data used to identify eQTLs, it had been generated using a cDNA library construction protocol that would be expected to include pre-mRNA and it was clear from the analysis of ER replication that many of the intronic ERs were likely to be derived from pre-mRNA. However, it is important to note that the NABEC dataset was generated using frontal cortex rather than hippocampus samples and that the library construction protocol though similar in some respects was different (Illumina TruSeq Stranded Total RNA versus Nugen ovation kit). Of the total SNP-ER pairs tested for

replication in the NABEC dataset, 27.3% were replicated using a marginal significance of p-value <0.05. However, replication rate substantially varies depending on the ER annotation. When considering only eQTL targeting annotated ERs (i.e. exonic) the replication rate is 31.1%. Replication rate increases when the eQTL targets ERs that partially overlap exonic regions, 50.6% and 36.0% respectively for regions classified as exonic-intergenic and exonic intronic. Surprinsingly, a high replication rate of 47.6% is observed for eQTL targeting ERs located enterely in intergenic regions. Finally, eQTL targeting ERs classified as intronic show the lowest replication rate of 19.7%. Overall, eQTL that target unannotated ERs reached a 25.2% replication rate.

5.3.5 Reference-free annotation methods yield the improvements GWAS interpretation

Given the evidence described above to suggest that novel ERs are common, can be brain-specific in nature and can be targeted by eQTLs, I investigated the impact of these findings on complex diseases. While I appreciate that novel ERs may be associated with complex diseases either through i) the enrichment of risk SNPs within the novel expressed regions or ii) the regulation of novel ERs by risk SNPs, I focused on the latter possibility.

All eQTL classes were investigated for the enrichment of risk SNPs for neurological and behavioural disorders as compared to all other risk SNPs within the STOPGAP database. This analysis showed that all eQTL types (ge-eQTL, t-eQTL, s-eQTL, er-eQTL) were significantly enriched for neurologically-relevant risk SNPs (Table 5.2). While er-eQTLs had the lowest enrichment for neurologically-risk SNPs, in absolute terms er-eQTLs provided the most information, given the higher number er-eQTLs which were also reported to be risk SNPs (Table 5.2). Furthermore, a similar anlaysis was performed focusing on er-eQTLs and subdividing them into those targeting annotated regions (i.e. exonic ERs) and those targeting unannotated regions (i.e. intronic,

intergenic, exonic-intronic and exonic-intergenic ERs). This analysis showed that both types of er-eQTLs were enriched for risk SNPs for neurological and behavioural disorders (the fisher exact test p-value: annotated = 3.82 x 10^{-21}, unanntoated = 1.10 x 10^{-15}). This would suggest that like annotated ERs, unannotated ERs can contribute to an understanding of neurological and behavioural disorders.

GWAS category	**ge-eQTL**	**t-eQTL**	**s-eQTL**	**er-eQTL**	**GWAS dataset**
Neurological and behavioural disorders	653	672	763	1,130	4,864
All other phenotypes	4,061	4,531	6,600	15,027	80,041
Fisher exact test p-value	4.4e-87	3.2e-77	5.8e-49	1.0e-9	

Table 5.2: **GWAS eQTL enrichment**. Table to show the number of GWAS-eQTL overlap as neurological and behavioural disorders and all other phenotypes with relative Fisher exact test p-value to test the enrichment amongst neurological and behavioural disorders.

I further explored the impact of annotation-free eQTLs on individual risk loci for six different GWASs (PD Nalls et al. 2018, AD Lambert et al. 2013, SCZ Pardiñas et al. 2018, MS Beecham et al. 2013, ALS Rheenen et al. 2016, intelligence Savage et al. 2018 GWAS) using the coloc R software as per section 4.2.4.1. The annotation-free approach implemented using derfinder, produced the highest number of signals colocalising with risk loci (209 risk loci, Table 5.3).

Quantification	**AD**	**ALS**	**Intelligence**	**MS**	**PD**	**SCZ**	**Total**
Gene (Salmon)	0	0	16	1	6	9	*32*
Transcript (Salmon)	0	1	26	1	9	13	*50*
Splicing (LeafCutter)	2	0	19	7	8	11	*47*
Expressed Region (derfinder)	5	0	85	7	35	77	*209*

Table 5.3: **Summary of eQTL-GWAS colocalisation. GWAS eQTL enrichment**. Table to show a summary of eQTL-GWAS colocalisation hits divided by eQTL class.

Of these 209 risk loci, 44.5% (93 risk loci) uniquely colocalised with er-eQTL signals. Furthermore, 44 of the risk loci uniquely colocalised with er-eQTLs targeting unanno-

tated ERs. There was evidence of splicing for eight of the uniquely co-localising eQTLs, of which four targeted intronic ERs, three targeted exonic-intronic ERs and one targeted an intergenic ER. For example, the risk SNP, rs2371214, which is associated with an increased risk of schizophrenia (GWAS p-val = 1.15 x 10^{-7}, Pardiñas et al. 2018), is also an eQTL targeting an intronic ER (chr7:82867064-82867528) within PCLO (Piccolo Presynaptic Cytomatrix Protein) and that these signals colocalise with highest colocalisation at rs10250881 suggested by coloc (PPH3+PPH4=0.93, PPH4/PPH3 = 4.91, the SNPs rs10250881 and rs2371214 are in high LD of R^2=0.94). The intronic ER target by the eQTL appears to be part of a novel isoform of PCLO, as evidenced by the presence of a split read connecting to a coding exon of PCLO (Figure 5.11). PCLO function has been implicated in the regulation of presynaptic proteins and synaptic vesicles trafficking and has been proposed as fontamental for several neural types (Ahmed et al. 2015) and it has been associated with major depressive disorder and bipolar disease (Bochdanovits et al. 2009; Choi et al. 2011). The risk SNP, rs2371214, did not colocalise with any other form of eQTL. Therefore, the annotation-free eQTL analysis performed provides a unique insight into the underlying mechanism through which the risk SNP acts. Thus, these findings suggested that incomplete annotation of the brain transcriptome might be limiting the interpretation of GWAS hits.

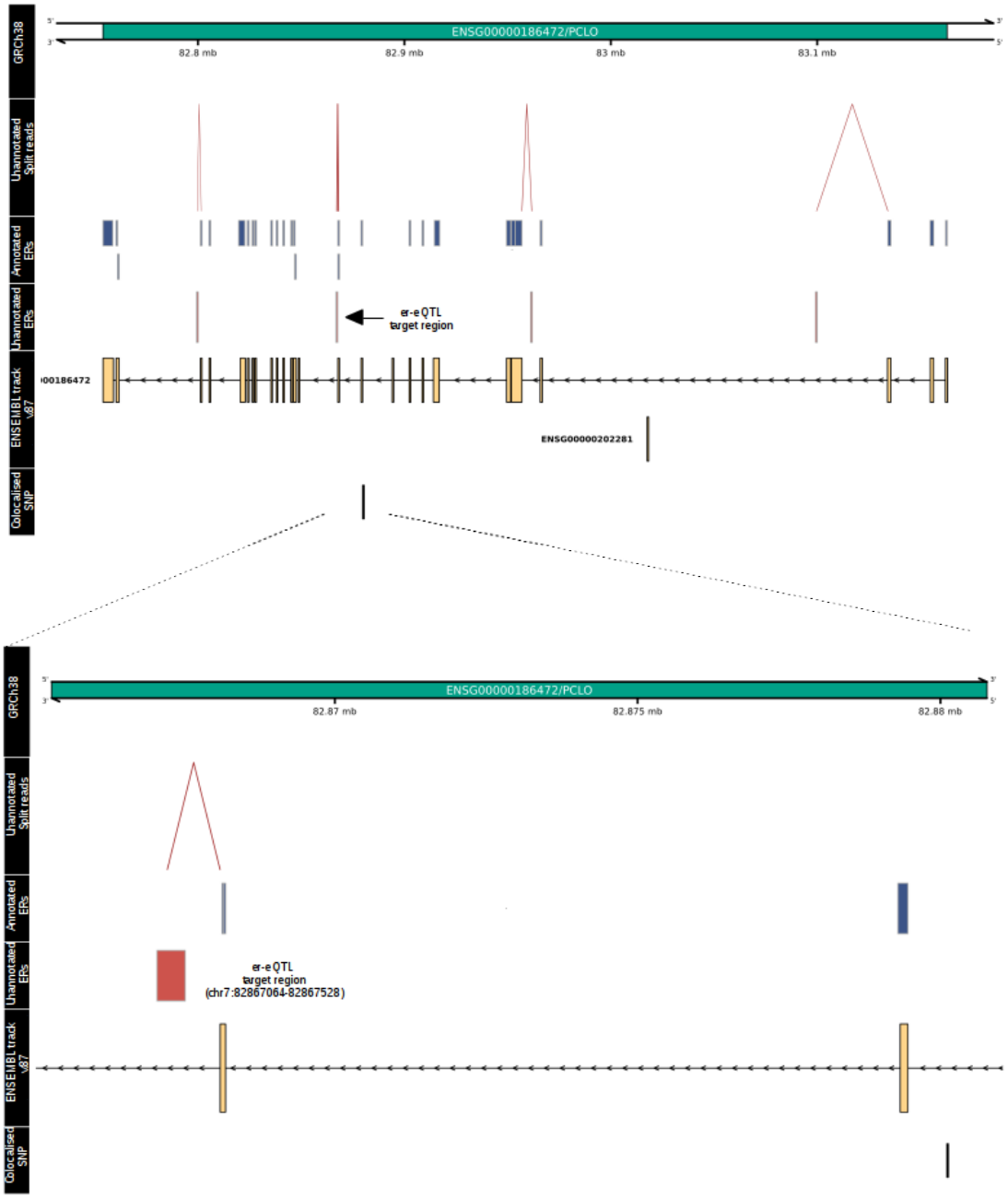

Figure 5.11: **er-eQTL targeting intronic ER in the PCLO gene.** Top panel: Colocalisation of the risk loci for schizophrenia colocalising rs2371214 and regulating the intronic ER (chr7:82867064-82867528) of PCLO (ENSMEBL track). Bottom panel: zoom into the intronic ER, showing a split read linking the coding region of PCLO and the intronic ER. The colocalise SNP is located in an intron of PCLO and upstream of the intronic ER regulated.

5.4 Discussion

It is well-recognised that even now existing gene annotation has limitations (Jaffe et al. 2015; Clark et al. 2018; Pertea et al. 2018) with the evidence suggesting that rather than there being many new genes to find, there is incomplete annotation of the full structural variation in RNA transcripts produced by known genes. Furthermore, in recent years splicing has been increasingly highlighted as an important mechanism underlying disease pathogenesis (Arnold et al. 2013; Li et al. 2016b). However, there has been relatively limited research on the impact of gene misannoation on disease understanding at a genome-wide scale (Kremer et al. 2017; Bartonicek et al. 2017), particularly within human brain. The findings reported in chapters 3 and 4 form the basis for the work described in this chapter, in that these chapters re-enforce the view that splicing is important in human brain disease and that even when considering novel intergenic expression, misannotation of existing genes is likely to be a very important component of the data.

In this chapter, I explore these themes further firstly by using a variety of methods to estimate splicing complexity in human hippocampus. This analysis produced a number of key findings. Most importantly it suggested that some of the most commonly used tools to capture and quantify splicing (Salmon and LeafCutter) are incompatible with split read data (reads spanning exon-exon junctions) and underestimate splicing even when considering annotated splicing. When considering unannotated splicing events, while there is data to suggest the existence of intergenic transcription and post-transcriptional processing, most novel gene expression and splicing relates to known genes. This interpretation of the data is supported by reads spanning exon-exon junctions and to a lesser extent the detection of ERs within introns or close to 3' and 5'UTRs.

Although the reliability of reads spanning exon-exon junctions has been previously

investigated, this feature of RNA-seq data has not been specifically analysed in human brain before. Consistent with previous analyses, I found that, when detected, split read data was extremely reliable. Most importantly, unannotated split reads consistently mapped in identical genomic locations across multiple, independent brain tissue sample sets. The robustness of this portion of the data allowed me to confidently investigate the tissue-specificity of unannotated split reads identified in human hippocampus. The lower detection of unannotated split reads in GTEx whole blood as compared to brain samples suggested not only the brain-specificity of the splicing events tagged by split reads, but that they were also unlikely to be the product of systematic mapping errors. Furthermore, hierarchical clustering of unannotated split read data showed that split read detection clustered across brain regions in an anatomically expected manner, implying that the splicing events captured could be functionally important. This view was supported by the high, predicted splice site strength of the sites implied by unannotated split reads. I found that the distributions of predicted splice site strengths of annotated and unannoated split reads were highly overlapping, suggesting that many of the splice sites underpinning unannotated split reads would be recognised by the splicing machinery and consistent with the functional significance of the splice sites detected.

While reads spanning exon-exon junctions are highly informative, they are also relatively sparse in short read RNA-seq datasets and transcriptome quantification methods based on read density remain important. The most flexible of these approaches is derfinder that can detect and quantify expressed genomic regions irrespective of existing annotation. Using this approach, I detected a large number of novel ERs of which the majority were intronic. Since the library construction method used to generate the hippocampus RNA-seq data was unselective in nature, I recognised that intronic ER detection was likely to be driven by pre-mRNA sequencing.

In fact, comparing ER replication rates between total RNA-seq versus polyA-selected

RNA-seq datasets confirmed this suspicion and as would be predicted the intronic ERs replicated well in the former and poorly in the latter. However, this analysis suggested that there is also likely to be a replicable pattern in pre-mRNA expression, which could be potentially informative. In fact, it is being increasingly recognised that pre-mRNA "noise" within RNA-seq datasets can provide useful information about transcriptional rate (Sibley et al. 2015; La Manno et al. 2018). Given that to date most eQTLs are thought to be driven by regulatory processes operating at the level of transcriptional rate, this might suggest that intronic ERs may be particularly useful expression features within eQTL analyses.

I specifically tested the value of intronic ER detection and quantification together with other forms of novel annotation (exonic-intronic, exonic-intergenic and intergenic ERs) in eQTL discovery. I found that using an annotation-free approach to identify expression features for eQTL analysis considerably increased the total number of unique eQTLs detected as compared to reference-based methods (6.3 fold increase). Furthermore, the eQTLs targeting unannotated ERs had similar replication rates to those targeting annotated ERs (31.1% replication rate for exonic ERs against 25.2% for all unannotated ERs). However, most importantly they were able to provide unique, disease-relevant information. This was most evident using coloc analyses, which tested not only a simple overlap between an eQTL signal and risk SNP, but specifically analysed signals for evidence of colocalisation. This approach showed that 44% of the risk loci analysed from across 6 neurologically-relevant GWASs colocalised only with an eQTL signal targeting an ER. Of the 209 colocalising eQTLs, 74 targeted an ER which had evidence for novel splicing as well as novel transcription. Thus, this analysis suggested that reference-free methods for identifying expression features in RNA-seq data significantly increases the yield of eQTLs from a given data set and can provide novel disease-relevant information. Taken together the analyses presented in this chapter provide compelling evidence for both incomplete annotation of the human hippocam-

pal transcriptome and the value of annotation-free approaches in the identification of expression features of disease-relevance.

Chapter 6

Conclusions and future directions

6.1 Fundamentals of the thesis

In this thesis I presented eQTL analyses performed with the aim of identifying regulatory regions in human brain that could help improve our understanding of neurological disorders. Although discussions were included in each result chapter, this final chapter consists of an overall discussion of the important issues to be considered when performing eQTL analyses in human brain and includes a discussion of the future approaches that could provide further insights into gene regulation in human brain.

A major theme in this thesis is the importance of RNA splicing. RNA splicing is a co-transcriptional process by which non-coding portions of the pre-mRNA sequence are removed and exons are joined together to produce mRNA. While alternative splicing is predominantly viewed as an efficient means of generating mRNA diversity from the same sequence, its role in the regulation of mRNA stability is increasingly being recognised (Xing, Xu, and Lee 2003; Garcia et al. 2004; Li et al. 2017a). Furthermore, there is now considerable evidence to show the biological importance of this process in cell differentiation, including within human brain (Su, D, and Tarn 2018). Finally, and perhaps most importantly, alternative splicing is very common with 89% of multi-exon

genes in the ENSEMBL database having the potential to express multiple transcripts.

Over the course of my PhD, there has been growing evidence to suggest the importance of alternative splicing in complex diseases. Through the identification of splicing eQTLs in lymphoblastoid cell lines, Li and colleagues have shown that loci regulating alternative splicing are enriched for systemic lupus erythematosus risk SNPs (Odhams et al. 2017).

Since the highest fraction of alternatively spliced genes is found in brain (Yeo et al. 2004), alternative splicing might be expected to be particularly important in human brain diseases. In fact, aberrant splicing is already functionally linked to ALS, where mutations in the gene TDP-43 have been shown to result in missplicing of RNA and contribute to neuronal loss (Chabot and Shkreta 2016). Similarly, Raj and colleagues examined alternative splicing in the dorsolateral prefrontal cortex, and found evidence that Alzheimer's Disease risk loci may operate by affecting alternative splicing of key genes (Raj et al. 2018). These findings are consistent with the analysis I presented in chapter 3. By investigating the genetic regulation of gene expression with RNA processing in mind, I found evidence to suggest that alternative splicing may be a key process in mediating PD risk.

Taken together, these findings suggest that alternative splicing and more specifically transcripts rather than genes should be considered as the fundamental unit of information in transcriptomic analyses. This raises two important issues: firstly, the accuracy of current transcript annotation in human brain; and secondly, the extent to which we are currently able to interpret the importance of differential transcript usage biologically. While the latter is outside the scope of this thesis, the former was extensively explored and additional pipelines were implemented to overcome some of the limitations of current transcriptome annotation.

Since the discovery of splicing in the late 1970s, the challenge of generating correct definitions of intron-exon boundaries has been recognised. While missing exons or splic-

ing annotation might seem like a relatively minor problem given that most annotations are correct, the absence of annotation features could result in the mis-assignment of reads to transcripts or splicing events even when those specific annotations are accurate. This will create a wider, systematic bias in splicing/transcript measurements (Soneson et al. 2018).

Thanks to advances in sequencing technologies, in the last decade the identification of correct intron-exon boundaries has become more achievable. The output of commonly used RNA-Seq technologies provides different types of information that can be used to identify intron-exon boundaries. Read-coverage can be used to identify intron-exon boundaries by defining exons. Conversely, split reads provide direct definitions of intron-exon boundaries by defining introns. Both types of data have been studied extensively in this thesis. Although methods relying on read-coverage benefit from a larger proportion of sequence data, my analyses suggest that split reads are more informative and are key for RNA-Seq based analyses of splicing. The nucleotide resolution and reliability provided by split reads allows highly sensitive analyses to be performed. The data generated in section 5.3.3 suggested that even when sophisticated methods are used, the information provided by split reads is not superseded, meaning that split reads represent valuable, and potentially, novel information. While the proportion of split reads in RNA-Seq is currently limited, the increases in the read length even when using short-read sequencing technology will result in a higher proportion of split reads in the data (Chhangawala et al. 2015) and, subsequently more sensitive splicing analyses.

Furthermore, this also implies that RNA-Seq analyses should increasingly prioritise the information contained in split reads. One possibility is to use split reads to first identify precise intron-exon boundaries and then use read-coverage to quantify splicing event, an approach which has driven the creation of LeafCutter.

One of the most interesting findings in this thesis is that misannotation is surprisingly widespread across the genome. Although there is evidence for a significant

quantity of novel transcription and splicing within intergenic regions, it appears to be most prevalent within known genes. A possible explanation for widespread misannotation of alternative splicing might be that modern human biology and genome annotation has largely been performed using too much focus on the gene as the functional unit. For the last 50 years scientists have been attempting to estimate the correct number of protein-coding genes in a given tissue and in fact this task has been complicated by splicing and the complexity this process generates. These considerations led me to study misannotation and, more broadly, transcriptome complexity by using transcripts as units of information. I found that currently the commonly used software tools are very limited to accurately capture let alone quantify transcript information.

As the availability of public genomic datasets continues to increase, there are new challenges for effective and informative data analysis. These new challenges require (besides physical resources) the careful consideration and implementation of two processes. Firstly, to obtain biological insights, analysts need to consider the issues around integration of data of varying types. Whether this is data generated in different fields or by different technologies, integration is required to perform more efficient and comprehensive studies. Secondly, the multiplicity of research questions implies the need for tailored pipelines, which have been designed and optimised to answer a specific hypothesis. Data integration and the use of optimised, tailored pipelines both featured in this thesis. eQTL analysis by definition requires the integration of two types of data, namely genetic and trasncriptomic data. In addition, several independent datasets, such as CAGE-Seq and multiple GWAS datasets, were integrated to produce robust results and identify disease-specific information.

Furthermore, optimisation steps were constantly implemented to overcome the challenges of using inherently noisy data and to maximise the information obtained from RNA-Seq data. For example, this revealed that, by capturing pre-mRNA, we can enhance eQTL yields and improve our understanding of gene regulation. Thus, the work

contained in this thesis is a step towards the generation of more innovative types of RNA-seq analyses, which have the capacity to integrate different types of data more effectively and which take full advantage of the information provided by RNA-Seq data.

6.1.1 Medical implications

One of the major outputs of this thesis was the generation of a large eQTL resource, covering eQTL analyses in three tissues of key importance for human neurodegenerative and neuropsychiatric disease, namely the substantia nigra, putamen and hippocampus. This thesis provides a large resource for the neuroscientist community which can be adopted as a first step to formulate hypothesis of disease mechanisms. While this is not the only eQTL resource available for brain, it is valuable. Replication data sets are important for the assessment of reliability and it is amongst the only reference-agnostic eQTL resource currently available across all human tissues. This feature of the data will mean that despite the inevitable changes in transcriptome annotation it will retain its utility.

Although the generation of eQTL resources has traditionally been driven by the desire to improve the interpretation of disease risk variants identified through GWAS and this was certainly pursued in this thesis, eQTL datasets have other applications. Most importantly, rapid progress in the identification of pathogenic variants contributing to monogenic diseases has made it possible to test the effect of regulatory variants on the penetrance of known pathogenic mutations (Castel et al. 2018). This results in a model of disease which integrates rare and common variation. In fact, combining data relating to complex and rare disease is likely to be a new frontier in clinical genetics and will inevitably lead to increasing attention on modifiers of rare disease.

The identification of widespread transcriptome misannotation is also likely to have medical implications. Over the past ten years next-generation sequencing technologies have reduced sequencing costs and so allowed the introduction of whole exome sequenc-

ing (WES) as a standard approach for clinical diagnostics. However, even with WES only 25-50% patients will receive a genetic diagnosis (Yang et al. 2014; Taylor et al. 2015). One possibility is that the exons where the pathogenic variants lie are excluded from the analysis because of the exome capture kit design. The reference-agnostic approach to RNA-seq analysis pursued in this thesis makes it possible to identify unannotated transcribed regions and particularly those linked to known genes. Thus, this thesis may provide additional information for the design of exome capture kits. There is also reason to believe that this would be particularly valuable for the diagnosis of neurogenetic disorders, which are amongst the disorders with the lowest diagnostic rates. A recent report estimate that the rate of diagnosis for patients of this type is only 26% (Fogel et al. 2016) and it is already well-established that splicing in human brain is particularly complex.

The reduction in cost of using whole genome sequencing (WGS) means that the use of WES will decrease in the next future. The potential of WGS of additionally identify non-coding variants, has motivated projects, such as the UK 100,000 Genomes Project, to use this sequencing approach. However, the advantage of WGS over WES can only be gain if only pathogenic variants outside known exon annotation are correctly identified. Therefore, accurate transcriptome annotation is required to take advantage of WGS potentila, and more importantly, to facilitate diagnosis.

Finally, a complementary approach for diagnostics is the use of transcriptomic data to test the effect of potential pathogenic variants (usually non-coding) on RNA processes (Kremer et al. 2017; Cummings et al. 2017) such as splicing. A direct application for the findings of this thesis is the use of expressed regions and split reads as reference to identify aberrant transcription or splicing in human brain.

6.2 Limitations and future directions

While this study represents a valuable resource in terms of the data and methods generated, there are several limitations of the work.

The analyses I performed in this thesis are restricted to data generated by the UK Brain Expression Consortium, which uses brain samples from a relatively small number of individuals, though that number is comparable to other consortia given the specific brain regions being tested (e.g. GTEx). For this reason, the eQTLs discovered in this thesis represent those with the largest effect sizes. Although the number of samples could be increased as new brain samples become available in brain banks, the batch effects generated by the collection of post-mortem brain samples by different brain banks and/or at different time points will have to be carefully addressed.

Furthermore, the RNA-seq data in this study was generated using a library construction method, which would be expected to capture pre-mRNA. Therefore, a more extensive analysis will be required to improve the understanding of transcriptome complexity and its characteristics in human brain, in particular the extent to which novel transcribed regions are part of mature mRNAs. I anticipate that with the use of 3rd generation sequencing technologies (e.g. Oxford Nanopore Technologies Pacific Biosciences) which have the capability to capture full-length transcript structures, transcripts rather than genes will become the core unit of information in such analyses and will provide a more comprehensive view of transcriptome complexity. The integration of 2nd and 3rd generation sequencing technologies represents an important new phase in transcriptomic analysis, including in the context of eQTLs.

Another limitation of this study is that eQTLs were identified in bulk RNA-Seq data, which means that I was most likely to capture the strongest signals across different cell types rather than regulatory signals specific to a cell type. Analyses included in section 3.3.4, suggested that at least some eQTLs may act in a cell-specific manner and this is

consistent with the literature (Fairfax et al. 2012; Westra and Franke 2014).

eQTL mapping studies have been performed on purified cells (Fairfax et al. 2012; Naranbhai et al. 2015) or using deconvolution methods (Westra et al. 2015). While the former is biased towards easily available cell types precluding brain-relevant analyses, the latter approach depends on the availability of high quality cell-specific transcriptomic data (Zhernakova et al. 2017).

Single-cell and single-nuclear RNA-Sequencing (scRNA-Seq and snRNA-seq) has the potential to address this issue (Lacar et al. 2016; Wijst et al. 2018). The fact that snRNA-Seq protocols can be applied to frozen tissue means that despite its limitations this is likely to dominate analyses for human brain. However, even isolating individual nuclei at scale from adult human brain remains challenging and to date the majority of the single cell studies have been performed using animal models (Zeisel et al. 2018; He et al. 2018). Induced pluripotent stem cell (iPSC) technology represents another way of addressing the problem of cellular purity. Through re-programing fibroblasts (or other cell types) stem cells can be used to generate a range of disease-relevant cell types. Subsequently eQTL mapping studies can be performed and in fact a study of this kind has been published focusing on sensory neurons (Schwartzentruber et al. 2018). This has the potential to make it easier to identify regulatory sites in the cell types where the specific disease phenotype manifests (Alasoo et al. 2015). Furthermore, the use of iPSC-derived cell types may make it easier to analyse state-specific regulation of gene expression. For Example, Salih and colleagues proposed that Alzheimer's disease vulnerability depends an individual's response to amyloid deposition (Salih et al. 2018). Using iPSC-derived cell types, it would be possible to test the association between genetic variability and gene expression within cells in a highly defined state or environment.

In summary, the combination of larger sample sets as well as new technologies from the fields of sequencing to cell biology are likely to overcome many of the limitations of

this project to provide novel insights in the context of brain disease.

6.3 Concluding remarks

While it may be hard to identify a specific landmark, the cumulative progress of genomics and related omics research have changed the entire landscape of biomedical research by shifting the field towards data-driven approaches. In this thesis, the potential of genomics is demonstrated through the flexibility of RNA-Seq technology to capture different aspects of transcriptome complexity and integrate that information with other types of data. This results not only in the ability to rapidly test hypotheses, but also to generate hypotheses which can be investigated in the laboratory. This makes genomics a potentially disruptive force in biomedical research – a force, which has the potential to complicate, but also to improve the lives of individual patients.

Bibliography

Adhikari, Kaustubh et al. (2016). "A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation". In: *Nature Communications* 7, p. 11616. ISSN: 2041-1723. DOI: 10.1038/ncomms11616. URL: http://www.nature.com/doifinder/10.1038/ncomms11616.

Aguet, François et al. (2017). "Genetic effects on gene expression across human tissues". In: *Nature* 550.7675, pp. 204–213. ISSN: 0028-0836. DOI: 10.1038/nature24277. URL: http://www.nature.com/doifinder/10.1038/nature24277.

Ahmed, M. Y. et al. (2015). "Loss of PCLO function underlies pontocerebellar hypoplasia type III". In: *Neurology* 84.17, pp. 1745–1750. ISSN: 0028-3878. DOI: 10.1212/WNL.0000000000001523. URL: http://www.ncbi.nlm.nih.gov/pubmed/25832664http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4424132http://www.neurology.org/cgi/doi/10.1212/WNL.0000000000001523.

Akbarian, Schahram et al. (2015). "The PsychENCODE project". In: *Nature Neuroscience* 18.12, pp. 1707–1712. ISSN: 1097-6256. DOI: 10.1038/nn.4156. URL: http://www.nature.com/doifinder/10.1038/nn.4156.

Alasoo, Kaur et al. (2015). "Transcriptional profiling of macrophages derived from monocytes and iPS cells identifies a conserved response to LPS and novel alternative transcription". In: *Scientific Reports* 5.1, p. 12524. ISSN: 2045-2322. DOI: 10.1038/srep12524. URL: http://www.nature.com/articles/srep12524.

Albert, Frank W. and Leonid Kruglyak (2015). "The role of regulatory variation in complex traits and disease". In: *Nature Reviews Genetics* 16.4, pp. 197–212. ISSN: 1471-0056. DOI: 10.1038/nrg3891. URL: http://www.nature.com/doifinder/10.1038/nrg3891.

Albin, Roger L., Anne B. Young, and John B. Penney (1989). "The functional anatomy of basal ganglia disorders". In: *Trends in Neurosciences* 12.10, pp. 366–375. ISSN: 01662236. DOI: 10.1016/0166-2236(89)90074-X.

Altschul, Stephen F. et al. (1990). "Basic local alignment search tool". In: *Journal of Molecular Biology* 215.3, pp. 403–410. ISSN: 00222836. DOI: 10.1016/S0022-2836(05)80360-2. URL: http://linkinghub.elsevier.com/retrieve/pii/S0022283605803602.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber (2014). "Genome analysis HTSeq - a Python framework to work with high-throughput sequencing data". In: 31.2, pp. 166–169. DOI: 10.1093/bioinformatics/btu638.

Anders, Simon, Alejandro Reyes, and Wolfgang Huber (2012). "Detecting differential usage of exons from RNA-seq data". In: *Genome Research* 22.10, pp. 2008–2017. ISSN: 10889051. DOI: 10.1101/gr.133744.111.

Andrews, Simon (2010). *FastQC - A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Ardlie, K G et al. (2015). "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans". In: *Science* 348.6235, pp. 648–660. ISSN: 0036-8075. DOI: 10.1126/science.1262110. URL: http://www.sciencemag.org/content/348/6235/648.full.

Arnold, E. S. et al. (2013). "ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43". In: *Proceedings of the National Academy of Sciences* 110.8, E736–

E745. ISSN: 0027-8424. DOI: 10.1073/pnas.1222809110. URL: http://www.ncbi.nlm.nih.gov/pubmed/23382207http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3581922http://www.pnas.org/cgi/doi/10.1073/pnas.1222809110.

Bartonicek, N. et al. (2017). "Intergenic disease-associated regions are abundant in novel transcripts". In: *Genome Biology* 18.1, p. 241. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1363-3. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1363-3.

Battle, Alexis et al. (2014). "Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals." In: *Genome research* 24.1, pp. 14–24. ISSN: 1549-5469. DOI: 10.1101/gr.155192.113. URL: http://www.ncbi.nlm.nih.gov/pubmed/24092820http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3875855.

Beach, Thomas G et al. (2008). "The Sun Health Research Institute Brain Donation Program: description and experience, 1987-2007." In: *Cell and tissue banking* 9.3, pp. 229–245. ISSN: 1389-9333. DOI: 10.1007/s10561-008-9067-2. URL: http://www.ncbi.nlm.nih.gov/pubmed/18347928http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2493521.

Beecham, Ashley H et al. (2013). "Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis". In: *Nature Genetics* 45.11, pp. 1353–1360. ISSN: 1061-4036. DOI: 10.1038/ng.2770. URL: http://www.nature.com/articles/ng.2770.

Bettencourt, Conceição et al. (2014). "Insights From Cerebellar Transcriptomic Analysis Into the Pathogenesis of Ataxia". In: *JAMA Neurology* 71.7, p. 831. ISSN: 2168-6149. DOI: 10.1001/jamaneurol.2014.756. URL: http://archneur.jamanetwork.com/article.aspx?doi=10.1001/jamaneurol.2014.756.

Blauwendraat, Cornelis et al. (2016). "Comprehensive promoter level expression quantitative trait loci analysis of the human frontal lobe". In: *Genome Medicine* 8.1, p. 65. ISSN: 1756-994X. DOI: 10.1186/s13073-016-0320-1. URL: http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-016-0320-1.

Bochdanovits, Z et al. (2009). "Joint reanalysis of 29 correlated SNPs supports the role of PCLO/Piccolo as a causal risk factor for major depressive disorder". In: *Molecular Psychiatry* 14.7, pp. 650–652. ISSN: 1359-4184. DOI: 10.1038/mp.2009.37. URL: http://www.nature.com/articles/mp200937.

Botía, Juan A. et al. (2017). "An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks". In: *BMC Systems Biology* 11.1, p. 47. ISSN: 1752-0509. DOI: 10.1186/s12918-017-0420-6. URL: http://bmcsystbiol.biomedcentral.com/articles/10.1186/s12918-017-0420-6.

Briggs, James A et al. (2015). "Mechanisms of Long Non-coding RNAs in Mammalian Nervous System Development, Plasticity, Disease, and Evolution." In: *Neuron* 88.5, pp. 861–877. ISSN: 1097-4199. DOI: 10.1016/j.neuron.2015.09.045. URL: http://www.ncbi.nlm.nih.gov/pubmed/26637795.

Brown, Christopher D. et al. (2013). "Integrative Modeling of eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs". In: *PLoS Genetics* 9.8. Ed. by Greg Gibson, e1003649. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003649. URL: http://dx.plos.org/10.1371/journal.pgen.1003649.

Cahoy, J. D. et al. (2008). "A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function". In: *Journal of Neuroscience* 28.1, pp. 264–278. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.4178-07.2008. URL: http://www.ncbi.nlm.nih.gov/pubmed/18171944http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.4178-07.2008.

Castel, Stephane E. et al. (2018). "Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk". In: *Nature Genetics* 50.9, pp. 1327–1334. ISSN: 1061-4036. DOI: 10.1038/s41588-018-0192-y. URL: http://www.nature.com/articles/s41588-018-0192-y.

Chabot, Benoit and Lulzim Shkreta (2016). "Defective control of pre-messenger RNA splicing in human disease." In: *The Journal of cell biology* 212.1, pp. 13–27. ISSN: 1540-8140. DOI: 10.1083/jcb.201510032. URL: http://www.ncbi.nlm.nih.gov/pubmed/26728853http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4700483.

Chen, Geng et al. (2011). "Revealing the missing expressed genes beyond the human reference genome by RNA-Seq". In: *BMC Genomics* 12.1, p. 590. ISSN: 1471-2164. DOI: 10.1186/1471-2164-12-590. URL: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-12-590.

Chhangawala, Sagar et al. (2015). "The impact of read length on quantification of differentially expressed genes and splice junction detection". In: *Genome Biology* 16.1, p. 131. ISSN: 1474-760X. DOI: 10.1186/s13059-015-0697-y. URL: http://www.ncbi.nlm.nih.gov/pubmed/26100517http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4531809http://genomebiology.com/2015/16/1/131.

Choi, Kwang H. et al. (2011). "Gene Expression and Genetic Variation Data Implicate PCLO in Bipolar Disorder". In: *Biological Psychiatry* 69.4, pp. 353–359. ISSN: 00063223. DOI: 10.1016/j.biopsych.2010.09.042. URL: http://www.ncbi.nlm.nih.gov/pubmed/21185011http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3278480http://linkinghub.elsevier.com/retrieve/pii/S0006322310010115.

Clark, Michael et al. (2018). "Long-read sequencing reveals the splicing profile of the calcium channel gene CACNA1C in human brain". In: *bioRxiv*, p. 260562. DOI:

10.1101/260562. URL: https://www.biorxiv.org/content/early/2018/02/05/260562.

Collado-Torres, Leonardo et al. (2015). "derfinder: Software for annotation-agnostic RNA-seq di erential expression analysis". In: *bioRxiv* 3.1, pp. 41–58. DOI: http://dx.doi.org/10.1101/015370.

Collado-Torres, Leonardo et al. (2017a). "Flexible expressed region analysis for RNA-seq with derfinder". In: *Nucleic Acids Research* 45.2, e9–e9. ISSN: 0305-1048. DOI: 10.1093/nar/gkw852. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw852.

Collado-Torres, Leonardo et al. (2017b). "Reproducible RNA-seq analysis using recount2". In: *Nature Biotechnology* 35.4, pp. 319–321. ISSN: 1087-0156. DOI: 10.1038/nbt.3838. URL: http://www.nature.com/articles/nbt.3838.

Cookson, William et al. (2009). "Mapping complex disease traits with global gene expression". In: *Nature Reviews Genetics* 10.3, pp. 184–194. ISSN: 1471-0056. DOI: 10.1038/nrg2537. URL: http://www.nature.com/doifinder/10.1038/nrg2537.

Cordell, Heather J and David G Clayton (2002). "A Unified Stepwise Regression Procedure for Evaluating the Relative Effects of Polymorphisms within a Gene Using Case / Control or Family Data : Application to HLA in Type 1 Diabetes". In: pp. 124–141.

Cummings, Beryl B et al. "Improving genetic diagnosis in Mendelian disease with transcriptome sequencing". In: DOI: 10.1101/074153.

Cummings, Beryl B et al. (2017). "Improving genetic diagnosis in Mendelian disease with transcriptome sequencing." In: *Science translational medicine* 9.386, eaal5209. ISSN: 1946-6242. DOI: 10.1126/scitranslmed.aal5209. URL: http://www.ncbi.nlm.nih.gov/pubmed/28424332http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5548421.

Dawson, Ted Murray. (2007). *Parkinson's disease : genetics and pathogenesis*. Informa Healthcare, p. 386. ISBN: 9780849336973.

Degner, Jacob F. et al. (2012). "DNaseâI sensitivity QTLs are a major determinant of human expression variation". In: *Nature* 482.7385, pp. 390–394. ISSN: 0028-0836. DOI: 10.1038/nature10808. URL: http://www.nature.com/articles/nature10808.

Deluca, David S. et al. (2012). "RNA-SeQC: RNA-seq metrics for quality control and process optimization". In: *Bioinformatics* 28.11, pp. 1530–1532. ISSN: 13674803. DOI: 10.1093/bioinformatics/bts196.

Denny, Joshua C et al. (2013). "Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data." In: *Nature biotechnology* 31.12, pp. 1102–10. ISSN: 1546-1696. DOI: 10.1038/nbt.2749. URL: http://www.ncbi.nlm.nih.gov/pubmed/24270849http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3969265.

Derrien, Thomas et al. (2012). "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression." In: *Genome research* 22.9, pp. 1775–89. ISSN: 1549-5469. DOI: 10.1101/gr.132159.111. URL: http://www.ncbi.nlm.nih.gov/pubmed/22955988http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3431493.

Dobin, A et al. (2013). "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1, pp. 15–21. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts635. URL: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bts635.

Emilsson, Valur et al. (2008). "Genetics of gene expression and its effect on disease". In: *Nature* 452.7186, pp. 423–428. ISSN: 0028-0836. DOI: 10.1038/nature06758. URL: http://www.nature.com/doifinder/10.1038/nature06758.

Engreitz, Jesse M. et al. (2016). "Local regulation of gene expression by lncRNA promoters, transcription and splicing". In: *Nature*. ISSN: 0028-0836. DOI: 10.1038/nature20149. URL: http://www.nature.com/doifinder/10.1038/nature20149.

Engström, Pär G et al. (2013). "Systematic evaluation of spliced alignment programs for RNA-seq data". In: *Nature Methods* 10.12, pp. 1185–1191. ISSN: 1548-7091. DOI: 10.1038/nmeth.2722. URL: http://www.nature.com/articles/nmeth.2722.

Fairfax, Benjamin P et al. (2012). "Genetics of gene expression in primary immune cells identifies cell type specific master regulators and roles of HLA alleles". In: *Nature Genetics* 44.5, pp. 502–510. ISSN: 1061-4036. DOI: 10.1038/ng.2205. URL: http://www.ncbi.nlm.nih.gov/pubmed/22446964http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3437404http://www.nature.com/doifinder/10.1038/ng.2205.

Fogel, Brent L. et al. (2016). "Clinical exome sequencing in neurogenetic and neuropsychiatric disorders". In: *Annals of the New York Academy of Sciences* 1366.1, pp. 49–60. ISSN: 00778923. DOI: 10.1111/nyas.12850. URL: http://doi.wiley.com/10.1111/nyas.12850.

Frazee, Alyssa C et al. (2015). "Ballgown bridges the gap between transcriptome assembly and expression analysis". In: *Nature Biotechnology* 33.3, pp. 243–246. ISSN: 1087-0156. DOI: 10.1038/nbt.3172. URL: http://www.nature.com/articles/nbt.3172.

Fromer, Menachem et al. (2016). "Gene expression elucidates functional impact of polygenic risk for schizophrenia". In: *Nature Neuroscience* 19.11, pp. 1442–1453. ISSN: 1097-6256. DOI: 10.1038/nn.4399. URL: http://www.nature.com/articles/nn.4399.

Fullard, John F et al. (2017). "Open chromatin profiling of human postmortem brain infers functional roles for non-coding schizophrenia loci". In: *Human Molecular Genetics* 26.10, pp. 1942–1951. ISSN: 0964-6906. DOI: 10.1093/hmg/ddx103. URL:

http://www.ncbi.nlm.nih.gov/pubmed/28335009https://academic.oup.com/hmg/article/26/10/1942/3069851.

Gaffney, Daniel J et al. (2012). "Dissecting the regulatory architecture of gene expression QTLs." In: *Genome biology* 13.1, R7. ISSN: 1474-760X. DOI: 10.1186/gb-2012-13-1-r7. URL: http://www.ncbi.nlm.nih.gov/pubmed/22293038http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3334587.

Gaidatzis, Dimos et al. (2015). "Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation". In: *Nature Biotechnology* 33.7, pp. 722–729. ISSN: 1087-0156. DOI: 10.1038/nbt.3269. URL: http://www.nature.com/doifinder/10.1038/nbt.3269.

Gambino, Gaetana et al. (2015). "Characterization of three alternative transcripts of the BRCA1 gene in patients with breast cancer and a family history of breast and/or ovarian cancer who tested negative for pathogenic mutations." In: *International journal of molecular medicine* 35.4, pp. 950–6. ISSN: 1791-244X. DOI: 10.3892/ijmm.2015.2103. URL: http://www.ncbi.nlm.nih.gov/pubmed/25683334http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4356434.

Gao, Yuan, Jinfeng Wang, and Fangqing Zhao (2015). "CIRI: an efficient and unbiased algorithm for de novo circular RNA identification". In: *Genome Biology* 16.1, p. 4. ISSN: 1465-6906. DOI: 10.1186/s13059-014-0571-3. URL: http://genomebiology.com/2015/16/1/4.

Garcia, Jesus et al. (2004). "A conformational switch in the Piccolo C2A domain regulated by alternative splicing". In: *Nature Structural & Molecular Biology* 11.1, pp. 45–53. ISSN: 1545-9993. DOI: 10.1038/nsmb707. URL: http://www.nature.com/articles/nsmb707.

Giambartolomei, Claudia et al. (2014). "Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics". In: *PLoS Genetics* 10.5.

Ed. by Scott M. Williams, e1004383. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004383. URL: http://dx.plos.org/10.1371/journal.pgen.1004383.

Grundberg, Elin et al. (2012). "Mapping cis- and trans-regulatory effects across multiple tissues in twins". In: *Nature Genetics* 44.10, pp. 1084–1089. ISSN: 1061-4036. DOI: 10.1038/ng.2394. URL: http://www.nature.com/doifinder/10.1038/ng.2394.

Guil, Sònia and Manel Esteller (2012). "Cis-acting noncoding RNAs: friends and foes". In: *Nature Structural & Molecular Biology* 19.11, pp. 1068–1075. ISSN: 1545-9993. DOI: 10.1038/nsmb.2428. URL: http://www.ncbi.nlm.nih.gov/pubmed/23132386http://www.nature.com/articles/nsmb.2428.

Hansen, Kasper D, Rafael A Irizarry, and Zhijin Wu (2012). "Removing technical variability in RNA-seq data using conditional quantile normalization". In: *Biostatistics* 13.2, pp. 204–216. ISSN: 14654644. DOI: 10.1093/biostatistics/kxr054.

Harrow, Jennifer et al. (2006). "GENCODE: producing a reference annotation for ENCODE". In: *Genome Biology* 7.Suppl 1, S4. ISSN: 14656906. DOI: 10.1186/gb-2006-7-s1-s4. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2006-7-s1-s4.

He, Liqun et al. (2018). "Single-cell RNA sequencing of mouse brain and lung vascular and vessel-associated cell types". In: *Scientific Data* 5, p. 180160. ISSN: 2052-4463. DOI: 10.1038/sdata.2018.160. URL: http://www.nature.com/articles/sdata2018160.

Heinzen, Erin L et al. (2008). "Tissue-Specific Genetic Control of Splicing: Implications for the Study of Complex Traits". In: *PLoS Biology* 6.12. Ed. by Edison Liu, e1000001. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1000001. URL: http://dx.plos.org/10.1371/journal.pbio.1000001.

Hindorff, L. A. et al. (2009). "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits". In: *Proceedings of the National Academy of Sciences* 106.23, pp. 9362–9367. ISSN: 0027-8424. DOI: 10.1073/pnas.

0903103106. URL: http://www.ncbi.nlm.nih.gov/pubmed/19474294http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2687147http://www.pnas.org/cgi/doi/10.1073/pnas.0903103106.

Howie, Bryan et al. (2012). "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing". In: *Nature Genetics* 44.8, pp. 955–959. ISSN: 1061-4036. DOI: 10.1038/ng.2354. URL: http://www.nature.com/doifinder/10.1038/ng.2354.

Hsiao, L L et al. (2002). "Correcting for signal saturation errors in the analysis of microarray data." In: *BioTechniques* 32.2, pp. 330–2, 334, 336. ISSN: 0736-6205. URL: http://www.ncbi.nlm.nih.gov/pubmed/11848410.

Hu, Yu et al. (2018). "PennDiff: detecting differential alternative splicing and transcription by RNA sequencing". In: *Bioinformatics* 34.14. Ed. by Janet Kelso, pp. 2384–2391. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty097. URL: http://www.ncbi.nlm.nih.gov/pubmed/29474557http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6041879https://academic.oup.com/bioinformatics/article/34/14/2384/4883492.

Huang, Ru et al. (2011). "An RNA-Seq Strategy to Detect the Complete Coding and Non-Coding Transcriptome Including Full-Length Imprinted Macro ncRNAs". In: *PLoS ONE* 6.11. Ed. by Carlo Gaetano, e27288. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0027288. URL: http://dx.plos.org/10.1371/journal.pone.0027288.

Hubbard, T et al. (2002). "The Ensembl genome database project." In: *Nucleic acids research* 30.1, pp. 38–41. ISSN: 1362-4962. DOI: 10.1093/nar/30.1.38.

Hyun, Min Kang, Chun Ye, and Eleazar Eskin (2008). "Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots". In: *Genetics* 180.4, pp. 1909–1925. ISSN: 00166731. DOI: 10.1534/genetics.108.094201.

Jaffe, Andrew E et al. (2015). "Developmental regulation of human cortex transcription and its clinical relevance at single base resolution." In: *Nature neuroscience* 18.1, pp. 154–161. ISSN: 1546-1726. DOI: 10.1038/nn.3898. arXiv: NIHMS150003. URL: http://dx.doi.org/10.1038/nn.3898.

Jaffe, Andrew E et al. (2017a). "Developmental And Genetic Regulation Of The Human Cortex Transcriptome In Schizophrenia". In: *bioRxiv*, p. 124321. DOI: 10.1101/124321. URL: https://www.biorxiv.org/content/early/2017/11/22/124321.

Jaffe, Andrew E et al. (2017b). "qSVA framework for RNA quality correction in differential expression analysis." In: *Proceedings of the National Academy of Sciences of the United States of America* 114.27, pp. 7130–7135. ISSN: 1091-6490. DOI: 10.1073/pnas.1617384114. URL: http://www.ncbi.nlm.nih.gov/pubmed/28634288http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5502589.

Jaffe, Andrew E. et al. (2018). "Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis". In: *Nature Neuroscience* 21.8, pp. 1117–1125. ISSN: 1097-6256. DOI: 10.1038/s41593-018-0197-y. URL: http://www.nature.com/articles/s41593-018-0197-y.

Karolchik, D et al. (2003). "The UCSC Genome Browser Database." In: *Nucleic acids research* 31.1, pp. 51–4. ISSN: 1362-4962. DOI: 10.1093/NAR/GKG129. URL: http://www.ncbi.nlm.nih.gov/pubmed/12519945http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC165576.

Kasowski, Maya et al. (2010). "Variation in transcription factor binding among humans." In: *Science (New York, N.Y.)* 328.5975, pp. 232–5. ISSN: 1095-9203. DOI: 10.1126/science.1183621. URL: http://www.ncbi.nlm.nih.gov/pubmed/20299548http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2938768.

Kim, Daehwan et al. (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." In: *Genome biology* 14.4, R36.

ISSN: 1465-6914. DOI: 10.1186/gb-2013-14-4-r36. URL: http://genomebiology.com/2013/14/4/R36.

Kim, S et al. (2012). "Association between SNPs and gene expression in multiple regions of the human brain". In: *Translational Psychiatry* 2.5, p. 113. ISSN: 2158-3188. DOI: 10.1038/tp.2012.42. URL: http://www.nature.com/doifinder/10.1038/tp.2012.42.

Kim, Tae-Kyung et al. (2010). "Widespread transcription at neuronal activity-regulated enhancers". In: *Nature* 465.7295, pp. 182–187. ISSN: 0028-0836. DOI: 10.1038/nature09033. URL: http://www.ncbi.nlm.nih.gov/pubmed/20393465http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3020079http://www.nature.com/articles/nature09033.

Kodzius, Rimantas et al. (2006). "CAGE: cap analysis of gene expression". In: *Nature Methods* 3.3, pp. 211–222. ISSN: 1548-7091. DOI: 10.1038/nmeth0306-211. URL: http://www.nature.com/doifinder/10.1038/nmeth0306-211.

Kremer, Laura S. et al. (2017). "Genetic diagnosis of Mendelian disorders via RNA sequencing". In: *Nature Communications* 8, p. 15824. ISSN: 2041-1723. DOI: 10.1038/ncomms15824. URL: http://www.nature.com/doifinder/10.1038/ncomms15824.

Krueger, Felix (2012). *Trim Galore! - A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries. http://www.bioinformatic*. URL: http://www.bioinformatics.babraham.ac.uk/projects/trim{_}galore/.

Kuersten, Scott (2012). "A transposable approach to RNA-seq from total RNA". In: *Nature Methods* 9. DOI: 10.1038/nmeth.f.357.

La Manno, Gioele et al. (2018). "RNA velocity of single cells". In: *Nature* 560.7719, pp. 494–498. ISSN: 0028-0836. DOI: 10.1038/s41586-018-0414-6. URL: http://www.nature.com/articles/s41586-018-0414-6.

Lacar, Benjamin et al. (2016). "Nuclear RNA-seq of single neurons reveals molecular signatures of activation". In: *Nature Communications* 7, p. 11022. ISSN: 2041-1723. DOI: 10.1038/ncomms11022. URL: http://www.nature.com/doifinder/10.1038/ncomms11022.

Lambert, J C et al. (2013). "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease." In: *Nature genetics* 45.12, pp. 1452–8. ISSN: 1546-1718. DOI: 10.1038/ng.2802. URL: http://www.ncbi.nlm.nih.gov/pubmed/24162737http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3896259.

Lappalainen, Tuuli et al. (2013). "Transcriptome and genome sequencing uncovers functional variation in humans". In: *Nature* 501.7468, pp. 506–511. ISSN: 0028-0836. DOI: 10.1038/nature12531. URL: http://www.nature.com/doifinder/10.1038/nature12531.

Lawrence, Michael et al. (2013). "Software for Computing and Annotating Genomic Ranges". In: *PLoS Computational Biology* 9.8. Ed. by Andreas Prlic, e1003118. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003118. URL: http://dx.plos.org/10.1371/journal.pcbi.1003118.

Lee, Kibaick et al. (2013). "Genetic Landscape of Open Chromatin in Yeast". In: *PLoS Genetics* 9.2. Ed. by Hunter Fraser, e1003229. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003229. URL: http://dx.plos.org/10.1371/journal.pgen.1003229.

Leek, Jeffrey T. et al. (2012). "The sva package for removing batch effects and other unwanted variation in high-throughput experiments". In: *Bioinformatics* 28.6, pp. 882–883. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/bts034. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts034.

Lein, Ed S. et al. (2007). "Genome-wide atlas of gene expression in the adult mouse brain". In: *Nature* 445.7124, pp. 168–176. ISSN: 0028-0836. DOI: 10.1038/nature05453. URL: http://www.ncbi.nlm.nih.gov/pubmed/17151600http://www.nature.com/articles/nature05453.

Leslie, R., C. J. O'Donnell, and A. D. Johnson (2014). "GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database". In: *Bioinformatics* 30.12, pp. i185–i194. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu273. URL: http://www.ncbi.nlm.nih.gov/pubmed/24931982http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4072913https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu273.

Li, Jin et al. (2017a). "Roles of alternative splicing in modulating transcriptional regulation." In: *BMC systems biology* 11.Suppl 5, p. 89. ISSN: 1752-0509. DOI: 10.1186/s12918-017-0465-6. URL: http://www.ncbi.nlm.nih.gov/pubmed/28984199http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5629561.

Li, Mulin Jun et al. (2016a). "GWASdb v2: an update database for human genetic variants identified by genome-wide association studies". In: *Nucleic Acids Research* 44.D1, pp. D869–D876. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1317. URL: http://www.ncbi.nlm.nih.gov/pubmed/26615194http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4702921https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1317.

Li, Y. I. et al. (2016b). "RNA splicing is a primary link between genetic variation and disease". In: *Science* 352.6285, pp. 600–604. ISSN: 0036-8075. DOI: 10.1126/science.aad9417. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.aad9417.

Li, Yang I et al. (2017b). "Prioritizing Parkinson's Disease genes using population-scale transcriptomic data". In: *bioRxiv*, p. 231001. DOI: 10.1101/231001. URL: https://www.biorxiv.org/content/early/2017/12/08/231001.

Li, Yang I. et al. (2018). "Annotation-free quantification of RNA splicing using LeafCutter". In: *Nature Genetics* 50.1, pp. 151–158. ISSN: 1061-4036. DOI: 10.1038/s41588-017-0004-9. URL: http://www.nature.com/articles/s41588-017-0004-9.

Li, Yun et al. (2010). "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes." In: *Genetic epidemiology* 34.8, pp. 816–834. ISSN: 1098-2272. DOI: 10.1002/gepi.20533. URL: http://www.ncbi.nlm.nih.gov/pubmed/21058334http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3175618.

Listgarten, Jennifer et al. (2010). "Correction for hidden confounders in the genetic analysis of gene expression." In: *Proceedings of the National Academy of Sciences of the United States of America* 107.38, pp. 16465–16470. ISSN: 0027-8424. DOI: 10.1073/pnas.1002425107.

Liu, Changning et al. (2005). "NONCODE: an integrated knowledge database of non-coding RNAs." In: *Nucleic acids research* 33.Database issue, pp. D112–5. ISSN: 1362-4962. DOI: 10.1093/nar/gki041. URL: http://www.ncbi.nlm.nih.gov/pubmed/15608158http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC539995.

Liu, Chenglin et al. (2013). "FusionQ: a novel approach for gene fusion detection and quantification from paired-end RNA-Seq." In: *BMC bioinformatics* 14, p. 193. ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-193. URL: http://www.ncbi.nlm.nih.gov/pubmed/23768108http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3691734.

Love, Michael I, John B Hogenesch, and Rafael A Irizarry (2016). "Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance es-

timation". In: *Nature Biotechnology* 34.12, pp. 1287–1291. ISSN: 1087-0156. DOI: 10.1038/nbt.3682. URL: http://www.nature.com/articles/nbt.3682.

MacArthur, D G et al. (2014). "Guidelines for investigating causality of sequence variants in human disease". In: *Nature* 508.7497, pp. 469–476. ISSN: 0028-0836. DOI: 10.1038/nature13127. URL: http://www.nature.com/doifinder/10.1038/nature13127.

Malone, John H et al. (2011). "Microarrays, deep sequencing and the true measure of the transcriptome". In: *BMC Biology* 9.1, p. 34. ISSN: 1741-7007. DOI: 10.1186/1741-7007-9-34. URL: http://bmcbiol.biomedcentral.com/articles/10.1186/1741-7007-9-34.

Martin, Marcel (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet.journal* 17.1, p. 10. ISSN: 2226-6089. DOI: 10.14806/ej.17.1.200. URL: http://journal.embnet.org/index.php/embnetjournal/article/view/200/479.

Matoulkova, Eva et al. (2012). "The role of the 3'Â untranslated region in post-transcriptional regulation of protein expression in mammalian cells." In: *RNA Biology* 9.5, pp. 563–576. ISSN: 1547-6286. DOI: 10.4161/rna.20231. URL: http://www.tandfonline.com/doi/abs/10.4161/rna.20231.

McDaniell, Ryan et al. (2010). "Heritable individual-specific and allele-specific chromatin signatures in humans." In: *Science (New York, N.Y.)* 328.5975, pp. 235–9. ISSN: 1095-9203. DOI: 10.1126/science.1184655. URL: http://www.ncbi.nlm.nih.gov/pubmed/20299549http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2929018.

McKenzie, Marna et al. (2014). "Overlap of expression Quantitative Trait Loci (eQTL) in human brain and blood". In: *BMC Medical Genomics* 7.1, p. 31. ISSN: 1755-8794. DOI: 10.1186/1755-8794-7-31. URL: http://bmcmedgenomics.biomedcentral.com/articles/10.1186/1755-8794-7-31.

McLaren, William et al. (2016). "The Ensembl Variant Effect Predictor". In: *Genome Biology* 17.1, p. 122. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0974-4. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4.

Mele, M. et al. (2015). "The human transcriptome across tissues and individuals". In: *Science* 348.6235, pp. 660–665. ISSN: 0036-8075. DOI: 10.1126/science.aaa0355. URL: http://www.ncbi.nlm.nih.gov/pubmed/25954002http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4547472http://www.sciencemag.org/cgi/doi/10.1126/science.aaa0355.

Millar, T et al. (2007). "Tissue and organ donation for research in forensic pathology: the MRC Sudden Death Brain and Tissue Bank." In: *The Journal of pathology* 213.4, pp. 369–75. ISSN: 0022-3417. DOI: 10.1002/path.2247. URL: http://www.ncbi.nlm.nih.gov/pubmed/17990279.

Moffatt, Miriam F. et al. (2007). "Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma". In: *Nature* 448.7152, pp. 470–473. ISSN: 0028-0836. DOI: 10.1038/nature06014. URL: http://www.nature.com/doifinder/10.1038/nature06014.

Montgomery, Stephen B et al. (2010). "Transcriptome genetics using second generation sequencing in a Caucasian population." In: *Nature* 464.7289, pp. 773–7. ISSN: 1476-4687. DOI: 10.1038/nature08903. URL: http://www.ncbi.nlm.nih.gov/pubmed/20220756http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3836232.

Moss, Davina J Hensman et al. (2017). "Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study". In: *The Lancet Neurology* 16.9, pp. 701–711. ISSN: 1474-4422. DOI: 10.1016/S1474-4422(17)30161-8. URL: https://www.sciencedirect.com/science/article/pii/S1474442217301618?via{\%}3Dihub.

Myers, Amanda J et al. (2007). "A survey of genetic human cortical gene expression". In: *Nature Genetics* 39.12, pp. 1494–1499. ISSN: 1061-4036. DOI: 10.1038/ng.2007.16. URL: http://www.nature.com/doifinder/10.1038/ng.2007.16.

Nalls, M A et al (2011). "Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies". In: *The Lancet* 377.9766, pp. 641–649. ISSN: 01406736. DOI: 10.1016/S0140-6736(10)62345-8.

Nalls, Mike A et al. (2014). "Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease". In: *Nature Genetics* 46.9, pp. 989–993. ISSN: 1061-4036. DOI: 10.1038/ng.3043. URL: http://www.nature.com/doifinder/10.1038/ng.3043.

Nalls, Mike A et al. (2018). "Parkinson's disease genetics: identifying novel risk loci, providing causal insights and improving estimates of heritable risk." In: *bioRxiv*, p. 388165. DOI: 10.1101/388165. URL: https://www.biorxiv.org/content/early/2018/08/09/388165.

Naranbhai, Vivek et al. (2015). "Genomic modulators of gene expression in human neutrophils". In: *Nature Communications* 6.1, p. 7545. ISSN: 2041-1723. DOI: 10.1038/ncomms8545. URL: http://www.ncbi.nlm.nih.gov/pubmed/26151758http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4507005http://www.nature.com/articles/ncomms8545.

Nellore, Abhinav et al. (2016a). "Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive". In: *Genome Biology* 17.1, p. 266. ISSN: 1474-760X. DOI: 10.1186/s13059-016-1118-6. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1118-6.

Nellore, Abhinav et al. (2016b). "Rail-RNA: scalable analysis of RNA-seq splicing and coverage". In: *Bioinformatics* 33.24, btw575. ISSN: 1367-4803. DOI: 10.1093/

bioinformatics/btw575. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw575.

Odhams, Christopher A. et al. (2017). "Mapping eQTLs with RNA-seq reveals novel susceptibility genes, non-coding RNAs and alternative-splicing events in systemic lupus erythematosus". In: *Human Molecular Genetics* 26.5, ddw417. ISSN: 0964-6906. DOI: 10.1093/hmg/ddw417. URL: http://www.ncbi.nlm.nih.gov/pubmed/28062664http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5409091https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddw417.

Oldham, Michael C et al. (2008). "Functional organization of the transcriptome in human brain". In: *Nature Neuroscience* 11.11, pp. 1271–1282. ISSN: 1097-6256. DOI: 10.1038/nn.2207. URL: http://www.nature.com/articles/nn.2207.

Ongen, Halit and Emmanouil T Dermitzakis (2015). "Alternative Splicing QTLs in European and African Populations". In: *American Journal of Human Genetics* 97.4, pp. 567–575. ISSN: 15376605. DOI: 10.1016/j.ajhg.2015.09.004. URL: http://dx.doi.org/10.1016/j.ajhg.2015.09.004.

Pai, Athma A, Jonathan K Pritchard, and Yoav Gilad (2015). "The genetic and mechanistic basis for variation in gene regulation." In: *PLoS genetics* 11.1, e1004857. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004857. URL: http://www.ncbi.nlm.nih.gov/pubmed/25569255http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4287341.

Pai, Athma A. et al. (2012). "The Contribution of RNA Decay Quantitative Trait Loci to Inter-Individual Variation in Steady-State Gene Expression Levels". In: *PLoS Genetics* 8.10. Ed. by Greg Gibson, e1003000. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003000. URL: http://dx.plos.org/10.1371/journal.pgen.1003000.

Pardiñas, Antonio F. et al. (2018). "Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection". In: *Nature Genetics*, p. 1. ISSN: 1061-4036. DOI: 10.1038/s41588-018-0059-2. URL: http://www.nature.com/articles/s41588-018-0059-2.

Patro, Rob et al. (2017). "Salmon provides fast and bias-aware quantification of transcript expression". In: *Nature Methods* 14.4, pp. 417–419. ISSN: 1548-7091. DOI: 10.1038/nmeth.4197. URL: http://www.nature.com/articles/nmeth.4197.

Pertea, Mihaela et al. (2015). "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads". In: *Nature Biotechnology* 33.3, pp. 290–295. ISSN: 1087-0156. DOI: 10.1038/nbt.3122. URL: http://www.ncbi.nlm.nih.gov/pubmed/25690850http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4643835http://www.nature.com/articles/nbt.3122.

Pertea, Mihaela et al. (2018). "Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise". In: *bioRxiv*, p. 332825. DOI: 10.1101/332825. URL: https://www.biorxiv.org/content/early/2018/05/29/332825.

Pickrell, J K et al. (2010). "Understanding mechanisms underlying human gene expression variation with RNA sequencing". In: *Nature* 464.7289, pp. 768–772. ISSN: 1476-4687. DOI: 10.1038/nature08872. URL: http://www.ncbi.nlm.nih.gov/pubmed/20220758{\%}5Cnhttp://www.nature.com/nature/journal/v464/n7289/pdf/nature08872.pdf.

Quek, Xiu Cheng et al. (2015). "lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs." In: *Nucleic acids research* 43.Database issue, pp. D168–73. ISSN: 1362-4962. DOI: 10.1093/nar/gku988. URL: http://www.ncbi.nlm.nih.gov/pubmed/25332394http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4384040.

Quinlan, Aaron R. and Ira M. Hall (2010). "BEDTools: A flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6, pp. 841–842. ISSN: 13674803. DOI: 10.1093/bioinformatics/btq033.

Raj, T. et al. (2014). "Polarization of the Effects of Autoimmune and Neurodegenerative Risk Alleles in Leukocytes". In: *Science* 344.6183, pp. 519–523. ISSN: 0036-8075. DOI: 10.1126/science.1249547. URL: http://www.ncbi.nlm.nih.gov/pubmed/24786080http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4910825http://www.sciencemag.org/cgi/doi/10.1126/science.1249547.

Raj, Towfique et al. (2018). "Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility". In: *Nature Genetics*, p. 1. ISSN: 1061-4036. DOI: 10.1038/s41588-018-0238-1. URL: http://www.nature.com/articles/s41588-018-0238-1.

Ramasamy, A et al. (2014). "Genetic variability in the regulation of gene expression in ten regions of the human brain". In: *Nat Neurosci* 17.10, pp. 1418–1428. DOI: 10.1038/nn.3801.

Reddy, Timothy E et al. (2012). "Effects of sequence variation on differential allelic transcription factor occupancy and gene expression." In: *Genome research* 22.5, pp. 860–9. ISSN: 1549-5469. DOI: 10.1101/gr.131201.111. URL: http://www.ncbi.nlm.nih.gov/pubmed/22300769http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3337432.

Rheenen, Wouter van et al. (2016). "Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis". In: *Nature Genetics* 48.9, pp. 1043–1048. ISSN: 1061-4036. DOI: 10.1038/ng.3622. URL: http://www.nature.com/articles/ng.3622.

Ripke, Stephan et al. (2014). "Biological insights from 108 schizophrenia-associated genetic loci". In: *Nature* 511.7510, pp. 421–427. ISSN: 0028-0836. DOI: 10.1038/nature13595. URL: http://www.nature.com/doifinder/10.1038/nature13595.

Roberts, Adam et al. (2011). "Improving RNA-Seq expression estimates by correcting for fragment bias". In: *Genome Biology* 12.3, R22. ISSN: 1465-6906. DOI: 10.1186/gb-2011-12-3-r22. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-3-r22.

Salih, Dervis A et al. (2018). "Genetic variability in response to Aβ deposition influences Alzheimer's risk". In: *bioRxiv*, p. 437657. DOI: 10.1101/437657. URL: https://www.biorxiv.org/content/early/2018/10/08/437657.

Savage, Jeanne E. et al. (2018). "Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence". In: *Nature Genetics* 50.7, pp. 912–919. ISSN: 1061-4036. DOI: 10.1038/s41588-018-0152-6. URL: http://www.nature.com/articles/s41588-018-0152-6.

Sawcer, Stephen et al. (2011). "Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis". In: *Nature* 476.7359, pp. 214–219. ISSN: 0028-0836. DOI: 10.1038/nature10251. URL: http://www.nature.com/doifinder/10.1038/nature10251.

Schreiber, Konrad et al. (2015). "Alternative Splicing in Next Generation Sequencing Data of Saccharomyces cerevisiae". In: *PLOS ONE* 10.10. Ed. by Emanuele Buratti, e0140487. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0140487. URL: http://dx.plos.org/10.1371/journal.pone.0140487.

Schwartzentruber, Jeremy et al. (2018). "Molecular and functional variation in iPSC-derived sensory neurons." In: *Nature genetics* 50.1, pp. 54–61. ISSN: 1546-1718. DOI: 10.1038/s41588-017-0005-8. URL: http://www.ncbi.nlm.nih.gov/pubmed/29229984http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5742539.

Shabalin, Andrey A (2012). "Matrix eQTL: Ultra fast eQTL analysis via large matrix operations". In: *Bioinformatics* 28.10, pp. 1353–1358. ISSN: 13674803. DOI: 10.1093/bioinformatics/bts163. arXiv: 1105.5764.

Shen, Judong et al. (2017). "STOPGAP: a database for systematic target opportunity assessment by genetic association predictions". In: *Bioinformatics* 33.17, pp. 2784–2786. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx274. URL: http://www.ncbi.nlm.nih.gov/pubmed/28472345http://academic.oup.com/bioinformatics/article/33/17/2784/3788008/STOPGAP-a-database-for-systematic-target.

Sibley, Christopher R. et al. (2015). "Recursive splicing in long vertebrate genes". In: *Nature* 521.7552, pp. 371–375. ISSN: 0028-0836. DOI: 10.1038/nature14466. URL: http://www.nature.com/doifinder/10.1038/nature14466.

Sidova, Monika et al. (2015). "Effects of post-mortem and physical degradation on RNA integrity and quality". In: *Biomolecular Detection and Quantification* 5, pp. 3–9. ISSN: 2214-7535. DOI: 10.1016/J.BDQ.2015.08.002. URL: https://www.sciencedirect.com/science/article/pii/S2214753515300048.

Small, Kerrin S et al. (2011). "Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes". In: *Nature Genetics* 43.6, pp. 561–564. ISSN: 1061-4036. DOI: 10.1038/ng.833. URL: http://www.nature.com/doifinder/10.1038/ng.833.

Soneson, Charlotte, Michael I. Love, and Mark D. Robinson (2015). "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences". In: *F1000Research* 4, p. 1521. ISSN: 2046-1402. DOI: 10.12688/f1000research.7563.1. URL: https://f1000research.com/articles/4-1521/v1.

Soneson, Charlotte et al. (2018). "A junction coverage compatibility score to quantify the reliability of transcript abundance estimates and annotation catalogs". In: *bioRxiv*, p. 378539. DOI: 10.1101/378539. URL: https://www.biorxiv.org/content/early/2018/07/28/378539.

Stegle, Oliver et al. (2010). "A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies". In: *PLoS*

Computational Biology 6.5, pp. 1–11. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1000770.

Su, Chun-Hao, Dhananjaya D, and Woan-Yuh Tarn (2018). "Alternative Splicing in Neurogenesis and Brain Development." In: *Frontiers in molecular biosciences* 5, p. 12. ISSN: 2296-889X. DOI: 10.3389/fmolb.2018.00012. URL: http://www.ncbi.nlm.nih.gov/pubmed/29484299http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5816070.

Sun, Zhifu et al. (2016). "Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations". In: *Briefings in Bioinformatics* 18.6, bbw069. ISSN: 1467-5463. DOI: 10.1093/bib/bbw069. URL: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbw069.

Tariq, Muhammad A. et al. (2011). "Whole-transcriptome RNAseq analysis from minute amount of total RNA". In: *Nucleic Acids Research* 39.18, e120–e120. ISSN: 1362-4962. DOI: 10.1093/nar/gkr547. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr547.

Taylor, Jenny C et al. (2015). "Factors influencing success of clinical genome sequencing across a broad spectrum of disorders". In: *Nature Genetics* 47.7, pp. 717–726. ISSN: 1061-4036. DOI: 10.1038/ng.3304. URL: http://www.ncbi.nlm.nih.gov/pubmed/25985138http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4601524http://www.nature.com/articles/ng.3304.

The ENCODE Project Consortium (2004). "The ENCODE (ENCyclopedia Of DNA Elements) Project". In: *Science* 306.5696.

Trabzuni, Daniah et al. (2011). "Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies". In: *Journal of Neurochemistry* 119.2, pp. 275–282. ISSN: 00223042. DOI: 10.1111/j.1471-4159.2011.07432.x. URL: http://doi.wiley.com/10.1111/j.1471-4159.2011.07432.x.

Trapnell, Cole et al. (2013). "Differential analysis of gene regulation at transcript resolution with RNA-seq". In: *Nature Biotechnology* 31.1, pp. 46–53. ISSN: 1087-0156. DOI: 10.1038/nbt.2450. URL: http://www.nature.com/articles/nbt.2450.

Wang, Kevin C. et al. (2011). "A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression". In: *Nature* 472.7341, pp. 120–124. ISSN: 0028-0836. DOI: 10.1038/nature09819. URL: http://www.ncbi.nlm.nih.gov/pubmed/21423168http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3670758http://www.nature.com/articles/nature09819.

Wang, Zhong, Mark Gerstein, and Michael Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature Reviews Genetics* 10.1, pp. 57–63. ISSN: 1471-0056. DOI: 10.1038/nrg2484. URL: http://www.nature.com/doifinder/10.1038/nrg2484.

Ward, Melanie et al. (2015). "Conservation and tissue-specific transcription patterns of long noncoding RNAs." In: *Journal of human transcriptome* 1.1, pp. 2–9. DOI: 10.3109/23324015.2015.1077591. URL: http://www.ncbi.nlm.nih.gov/pubmed/27335896http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4894084.

Webster, Jennifer A. et al. (2009). "Genetic Control of Human Brain Transcript Expression in Alzheimer Disease". In: *The American Journal of Human Genetics* 84.4, pp. 445–458. ISSN: 00029297. DOI: 10.1016/j.ajhg.2009.03.011.

Welter, Danielle et al. (2014). "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations". In: *Nucleic Acids Research* 42.D1, pp. D1001–D1006. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1229. URL: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkt1229.

Werner, Michael S et al. (2015). "Nuclear Fractionation Reveals Thousands of Chromatin-Tethered Noncoding RNAs Adjacent to Active Genes." In: *Cell reports* 12.7, pp. 1089–

98. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2015.07.033. URL: http://www.ncbi.nlm.nih.gov/pubmed/26257179.

Westra, Harm-Jan and Lude Franke (2014). "From genome to function by studying eQTLs". In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1842.10, pp. 1896–1902. ISSN: 0925-4439. DOI: 10.1016/J.BBADIS.2014.04.024. URL: https://www.sciencedirect.com/science/article/pii/S0925443914001112?via{\%}3Dihub.

Westra, Harm-Jan et al. (2015). "Cell Specific eQTL Analysis without Sorting Cells". In: *PLOS Genetics* 11.5. Ed. by Tomi Pastinen, e1005223. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1005223. URL: http://dx.plos.org/10.1371/journal.pgen.1005223.

Wijst, Monique G. P. van der et al. (2018). "Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs". In: *Nature Genetics* 50.4, pp. 493–497. ISSN: 1061-4036. DOI: 10.1038/s41588-018-0089-9. URL: http://www.nature.com/articles/s41588-018-0089-9.

Williams, M. R. et al. (2014). "Neuropathological changes in the substantia nigra in schizophrenia but not depression". In: *European Archives of Psychiatry and Clinical Neuroscience* 264.4, pp. 285–296. ISSN: 0940-1334. DOI: 10.1007/s00406-013-0479-z. URL: http://link.springer.com/10.1007/s00406-013-0479-z.

Winden, Kellen D et al. (2009). "The organization of the transcriptional network in specific neuronal classes". In: *Molecular Systems Biology* 5, p. 291. ISSN: 1744-4292. DOI: 10.1038/msb.2009.46. URL: http://www.ncbi.nlm.nih.gov/pubmed/19638972http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2724976http://msb.embopress.org/cgi/doi/10.1038/msb.2009.46.

Xing, Yi, Qiang Xu, and Christopher Lee (2003). "Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring

domains." In: *FEBS letters* 555.3, pp. 572–8. ISSN: 0014-5793. URL: http://www.ncbi.nlm.nih.gov/pubmed/14675776.

Yang, Yaping et al. (2014). "Molecular findings among patients referred for clinical whole-exome sequencing." In: *JAMA* 312.18, pp. 1870–9. ISSN: 1538-3598. DOI: 10.1001/jama.2014.14601. URL: http://www.ncbi.nlm.nih.gov/pubmed/25326635http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4326249.

Yeo, Gene and Christopher B Burge (2004). "Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals." In: *Journal of computational biology : a journal of computational molecular cell biology* 11.2-3, pp. 377–94. ISSN: 1066-5277. DOI: 10.1089/1066527041410418. URL: http://www.ncbi.nlm.nih.gov/pubmed/15285897.

Yeo, Gene et al. (2004). "Variation in alternative splicing across human tissues". In: *Genome Biology* 5.10, R74. ISSN: 14656906. DOI: 10.1186/gb-2004-5-10-r74. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2004-5-10-r74.

Zeisel, Amit et al. (2018). "Molecular Architecture of the Mouse Nervous System". In: *Cell* 174.4, 999–1014.e22. ISSN: 0092-8674. DOI: 10.1016/J.CELL.2018.06.021. URL: https://www.sciencedirect.com/science/article/pii/S009286741830789X?via{\%}3Dihub.

Zhang, Xiao-Ou et al. (2014). "Complementary Sequence-Mediated Exon Circularization". In: *Cell* 159.1, pp. 134–147. ISSN: 0092-8674. DOI: 10.1016/J.CELL.2014.09.001. URL: https://www.sciencedirect.com/science/article/pii/S0092867414011118?via{\%}3Dihub.

Zhang, Yong E et al. (2012). "Overviews New genes expressed in human brains: Implications for annotating evolving genomes". In: DOI: 10.1002/bies.201200008.

Zhao, Jian et al. (2017). "GFusion: an Effective Algorithm to Identify Fusion Genes from Cancer RNA-Seq Data." In: *Scientific reports* 7.1, p. 6880. ISSN: 2045-2322. DOI: 10.1038/s41598-017-07070-6. URL: http://www.ncbi.nlm.nih.gov/pubmed/28761119http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5537242.

Zhao, Shanrong et al. (2014). "Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells". In: *PLoS ONE* 9.1. Ed. by Shu-Dong Zhang, e78644. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0078644. URL: http://dx.plos.org/10.1371/journal.pone.0078644.

Zhao, Shanrong et al. (2018). "Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion". In: *Scientific Reports* 8.1, p. 4781. ISSN: 2045-2322. DOI: 10.1038/s41598-018-23226-4. URL: http://www.nature.com/articles/s41598-018-23226-4.

Zhao, Yi et al. (2015). "NONCODE 2016: an informative and valuable data source of long non-coding RNAs". In: *Nucleic Acids Research*. DOI: 10.1093/nar/gkv1252.

Zhernakova, Daria V et al. (2017). "Identification of context-dependent expression quantitative trait loci in whole blood". In: *Nature Genetics* 49.1, pp. 139–145. ISSN: 1061-4036. DOI: 10.1038/ng.3737. URL: http://www.ncbi.nlm.nih.gov/pubmed/27918533http://www.nature.com/articles/ng.3737.

Zhu, Zhihong et al. (2016). "Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets". In: *Nature Genetics* 48.5, pp. 481–487. ISSN: 1061-4036. DOI: 10.1038/ng.3538. URL: http://www.ncbi.nlm.nih.gov/pubmed/27019110http://www.nature.com/articles/ng.3538.

Zou, Fanggeng et al. (2012). "Brain Expression Genome-Wide Association Study (eGWAS) Identifies Human Disease-Associated Variants". In: *PLoS Genetics* 8.6. Ed. by Greg Gibson, e1002707. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1002707. URL: http://dx.plos.org/10.1371/journal.pgen.1002707.