Computational identification of regulatory features affecting splicing in the human brain.

Author:

Warren Emmett

Supervisors:

Dr. Vincent Plagnol

Prof. Jernej Ule

UCL Genetics Institute

Department of Genetics, Evolution and Environment

PhD Thesis

September 6, 2016

Acknowledgements

"The most beautiful thing we can experience is the mysterious. It is the source of all true art and all science. He to whom this emotion is a stranger, who can no longer pause to wonder and stand rapt in awe, is as good as dead: his eyes are closed. — Albert Einstein"

As Einstein points out, a sense of wonder is so important to any scientific endeavour. The opportunity I have been given to explore cellular biology using computational methods is truly a great gift. This was largely made possible by the people who have mentored me and given me chances to improve myself such as the option to undertake this PhD. I am incredibly grateful to Dr. Vincent Plagnol for this. I have had the time and space to rediscover my love for bioinformatics and cellular biology which had been lost in years of short term service work. I have gained more in these three years than I could have imagined. I am also very grateful to Prof. Jernej Ule and his lab for the opportunity to collaborate on the recursive splicing project, it was both enjoyable and a great learning experience. I would also like to thank Dr. Chris Sibley who was my mentor on the recursive project, I was very fortunate to have had his help. He is also responsible for most of the great graphs and diagrams in the recursive chapter, I have cited the paper in these cases. Furthermore I'd like to thank all my colleagues in the UGI; Dr. Jon White and Prof. Dave Curtis for their help with the statistics. Lucy van Dorp, Dr. Cian Murphy, Chris Steele and Jack Humphrey for your support and enthusiasm. I have been truly blessed to know such great people and great scientists.

This thesis would not have been written without the support of family, my mother and sister who have both suffered great loss but have remained incredibly supportive, you truly are my heroes. My grandparents who have been a continuous inspiration and support to me throughout my life. And, of course, my girlfriend, Emily Frontain, who kept me company during long hours of thesis writing and revision, making these tasks a true pleasure. I am eternally grateful that I could have these wonderful people in my life.

These three years have not been without sadness. My father passed away suddenly during the first year of my PhD and this was indeed a terrible loss. I did not get the chance to celebrate completion of

this PhD, nor share the successes of the last few years but he is a part of me, both literally (I have half his chromosomes!) and in every delightful moment that comes from being able to follow my passion; exploring cellular biology behind a computer with a cup of coffee. Last year, Konstantin Sofianos, my sister's partner and best friend, passed away in a hospital in Hamburg, Germany as the sun was rising on a crisp Friday morning. The soft, warm, light filled the room as I held my sister's hand in bewildered silence. He died of cancer two weeks before his 32^{nd} birthday. He was a far brighter, more capable man than I and he is sorely missed. I dedicate this body of work to these two fine souls and their love of knowledge and passion for work.

Contents

Abstract

RNA splicing has enabled a dramatic increase in species complexity. Splicing occurs in over 95% of mammalian genes allowing the development of exceptional cellular diversity without an increase in raw gene numbers. This is highlighted by the fact that human and nematodes have the same number of genes (20,000 human genes versus 19,000 genes in Caenorhabditis elegans). Although the mechanistic process of splicing is now well understood there remains a multitude of unexplored dynamics that have only become visible with the power of next generation sequencing (NGS).

The human brain is one of the best examples of an intricate cellular structure. Neuronal cell types are incredibly diverse and specialised, regulated through various transcriptional mechanisms. Recently, long genes (150kb+) have been implicated as crucial to neuronal function and their impairment has been attributed to several neurological disorders. I explore this relationship further by showing that long genes are more highly expressed in the brain than other tissues. Long genes are also distinct in that they are deficient in H3k36me3, a histone mark largely associated with splicing and active transcription. Through analysis of brain RNA-seq data, a novel splicing mechanism known as recursive splicing was identified in long introns. Recursive splice sites (RSS) consist of an intronic 3'splice site followed immediately by a 5' splice site. These sites result in a zero-length exon that regulates the use of cryptic promoters ensuring only the functional isoform is expressed. This discovery lead me to question if other non-canonical forms of splicing are common in the brain.

Backsplicing is a recently discovered splicing mechanism pervasive in the tree of life. This occurs when a 3' end of a downstream exon is spliced onto the 5' end of an upstream exon resulting in a circular RNA molecule (hereafter: circRNA). circRNA are enriched in neuronal genes and mediated by RNA binding factors. I have identified and quantified the presence of circRNA within the brain, identifying a large number of highly expressed novel circRNA. From these findings I identify a subset of highly expressed backsplice junctions that occur between two proximal genes from the same family.

In order to understand the function of these splicing reactions I inspected the splicing features themselves, namely; the 5' and 3' splice sites and the branchpoint. The branchpoint remains a poorly characterised feature and until recently very few have been experimentally validated. I explore these features through the ExAC and UCLex consortia, using cumulative variant ratios to annotate invariant positions within the branchpoint and splice sites. By identifying invariant positions I could then investigate how variation impacts splicing efficiency by integrating whole exome and RNA sequence data from the GEUVADIS consortium. Findings show that exon expression is a poor indicator of splicing dysfunction, showing a three fold lower sensitivity than direct analysis of splice junction reads. I also devise a variant effect score that captures a significant portion of change in splice site efficiency enabling improved prediction of deleterious variants.

Together, this thesis hints at the massive potential of NGS to investigate the diversity of splicing related features while identifying novel features that could be implicated in neurological dysfunction.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Next generation sequencing enables high throughput, nucleotide resolution analysis of cellular processes

Next generation sequencing (hereafter: NGS) has catapulted biological sciences forward by making genome-wide studies possible. The power and versatility of genome-wide sequence cannot be underestimated, massive volumes of NGS data are now freely available to analyse and explore.

There have been major advances over the previous high throughput, hybridization-based microarray technology. NGS provides better quality data, more robust results and lower noise [Buermans and den Dunnen, 2014]. This technology has been applied in large consortia for the systematic evaluation of human polymorphism such as the landmark 1,000 genomes project [Abecasis et al., 2012], the UK10K exome project [Walter et al., 2015] and most recently the staggering 65,000 sample Exome Aggregation Consortium (ExAC) [The EXaC Consortium, 2015]. Recent projects have also aimed at combining DNA information with additional data, in particular RNA-sequencing (hereafter: RNA-seq). These include the GEUVADIS project [Lappalainen et al., 2013], Illumina Bodymap (www.illumina.com; ArrayExpress ID: E-MTAB-513) and Genotype-Tissue Expression Consortium (GTEx) [Genotype-Tissue Expression Consortium, 2015].

Next generation sequencing continues to rapidly accelerate our ability to characterise and discover novel cellular processes. NCBI's public short read sequence archive now hosts over 1,000 TB of data, the equivalent of resequencing the human genome with a 1,000,000 times coverage (Figure 1.1).

It is crucial to find effective ways of analysing these large datasets, particularly given their relevance in understanding cellular biology and ultimately disease etiology. This thesis develops methods by leveraging

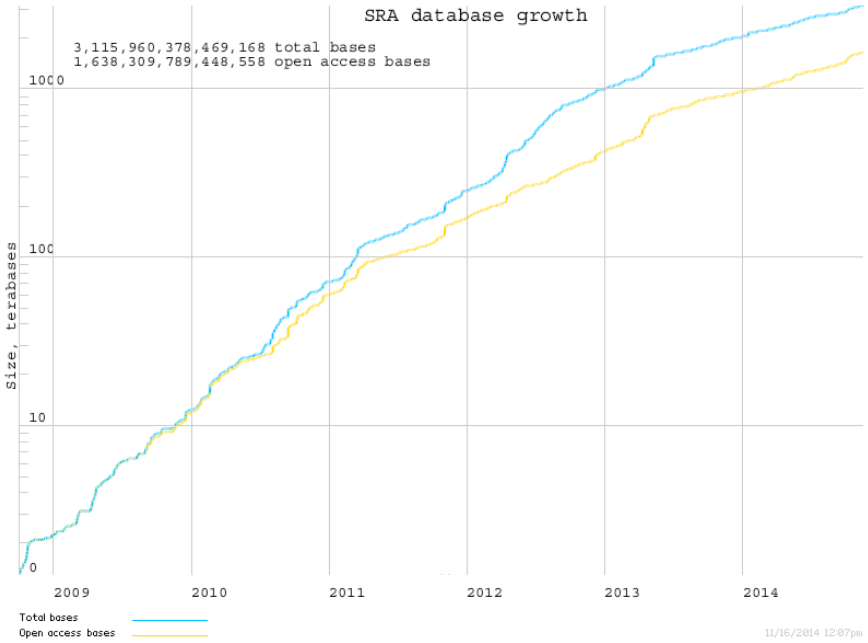

Figure 1.1: Growth of data in NCBI Short Sequence Read Archive (SRA) as a function of time. [NCBI]

publicly available data to explore the non-coding elements of the human brain far more powerfully than was possible before. Primarily, this thesis aims at exploring non-canonical aspects of gene expression, particularly related to RNA splicing.

Compared to microarray technology where a control sample is required to normalize background hybridization [Wang et al., 2009], next generation sequencing can provide absolute expression values. Although there remains known bias in the form of batch effects [Taub et al., 2010; SEQC/MAQC-III Consortium, 2014], especially due to the rapid improvement in equipment and laboratory kits, these samples still provide sequence information which is valuable in its own right.

1.2 Splicing in vertebrate genomes

Splicing was first discovered over thirty years ago in adenovirus and highlighted the alternative use of exons to create multiple mRNA from a single gene locus [Chow et al., 1977; Berget et al., 1977]. With the aid of sequencing we now know that splicing plays a integral role in mRNA diversity affecting over 95% of mammalian genes and controlling regulatory processes such as chromatin modification [Pan et al., 2008; Barash et al., 2010]. This is an essential aspect of the increase in complexity of vertebrates as they share similar gene numbers to invertebrates (20,000 human genes versus 19,000 genes in Caenorhabditis elegans).

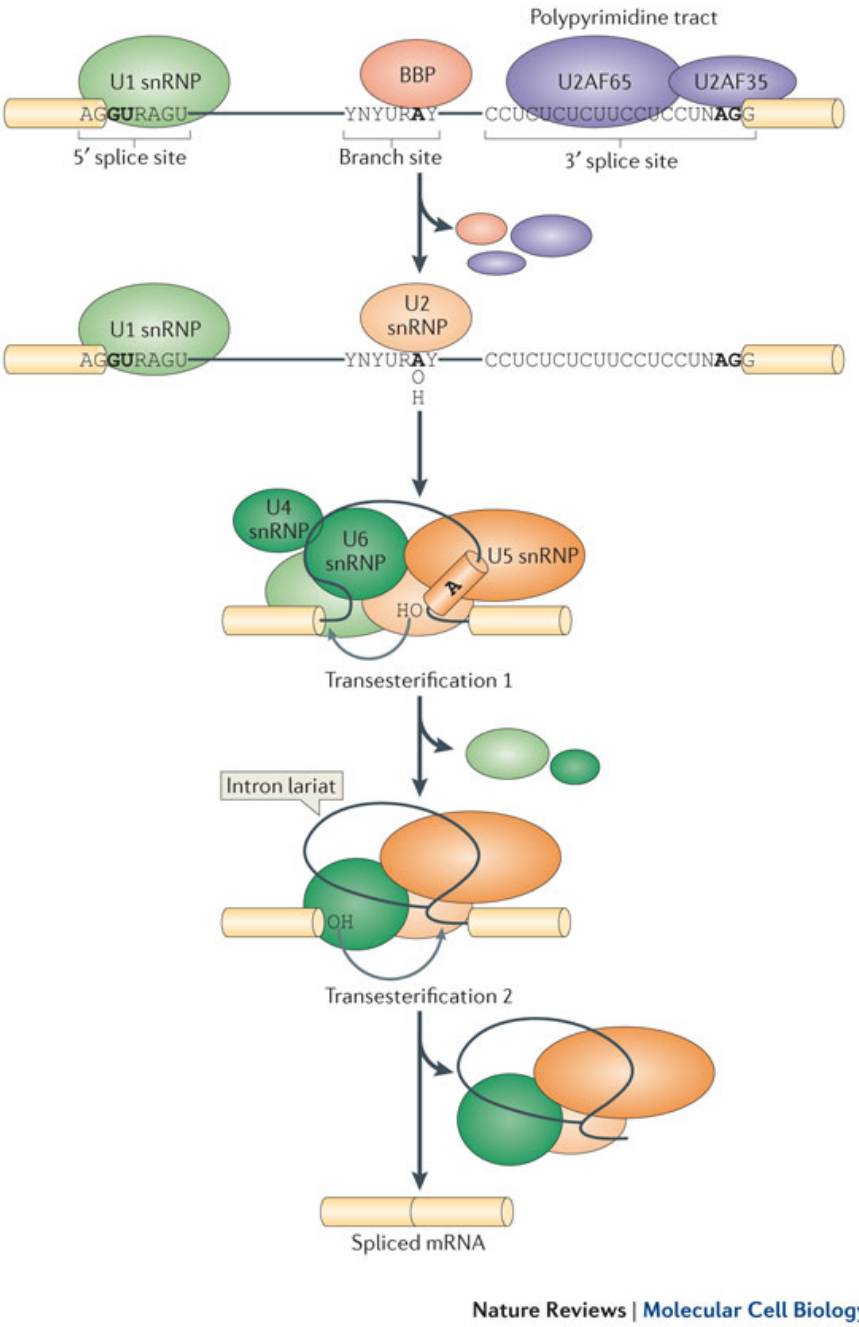

Figure 1.2: Graphical representation of the splicing process. In brief; after binding of splicing factors of snRNPs the splice sites flanking the intron and branchpoint, several rearrangements occur resulting in the transesterification of the severed 5' intron end to the adensoine at the branchpoint creating an intronic lariat. Following this the transesterification of the 5' exon end and the 3' exon start create the final mRNA and release the intronic lariat. [Kornblihtt et al., 2013]

This process is mediated by the spliceosome, a complex of ribonuclearprotein that assembles flanking introns through identification of consensus sequences known as splice sites (the upstream 5' site and

downstream 3' site) and the branchpoint. U1 small nuclear riboprotein (snRNP) binds to the 5' splice site with U2AF35 and U2AF65 binding to the 3' splice site and proximal polypyramidine tract respectively. The upstream branchpoint is bound by the branchpoint-binding protein (BBP). Hereafter, the severed 5' intron end is covalently bonded through transesterification of the adensoine at the branchpoint creating an intronic lariat. The second step is the transesterification of the 5' exon end and the 3' exon start thereby creating the final mRNA and releasing the intronic lariat (Figure 1.2 outlines this process).

The sequence composition of the splice site plays a crucial role in the efficiency of splicing and the inclusion/exclusion of exons. Thus splice sites tend to be highly conserved. Splice sites that diverge from this consensus provide additional variation and these exons may be skipped without the addition of several other modifiers such as cis-regulatory sequences (exonic enhancers/silencers) and trans-acting factors (tissue-specific RNA binding factors such as PTB and NOVA [Ule et al., 2006; Jelen et al., 2007; Kafasla et al., 2012]).

The splicing process has been shown to occur simultaneously during transcription [Beyer and Osheim, 1988; Khodor et al., 2011]. Recently, overwhelming evidence has pointed to the co-transcriptional splicing of the majority of exons while still associated with chromatin [Tilgner et al., 2012]. This is further highlighted by the study of co-transcriptional splicing in the human brain and its use to estimate speed of transcription and provide key insights into exonic definition [Ameur et al., 2011]. An example of co-transcriptional splicing is shown in Figure 1.3, the iconic saw-tooth pattern indicates where splicing occurs and the gradient can be used to infer transcriptional speed.

1.2.1 Circular RNAs are a novel class of non coding RNA created by backsplicing

Circular RNA (hereafter circRNA) are pervasively expressed in mammalian cells and enriched in brain, blood and exosomes [Ashwal-Fluss et al., 2014; Rybak-Wolf et al., 2015; You et al., 2015; Venø et al., 2015]. These non-coding RNA are formed through structural changes within their intronic flanks, often catalysed by RNA binding proteins and palindromic repeat sequences. circRNA can be identified by a back-splice junction which is not present in canonical linear isoforms.

Synthesis of circRNA is reliant on intronic sequences

Advances in the field have confirmed the mechanism that allows the 3' end of a downstream exon to be spliced onto the 5' end of an upstream exon, known as a 'backsplice' (see Figure 1.4). Biogenesis is mediated by

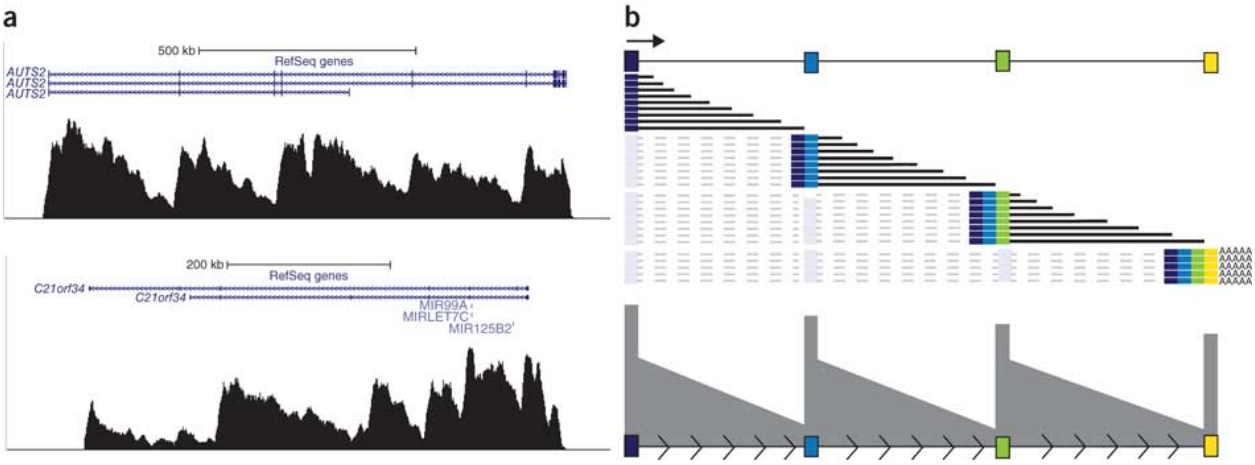

Figure 1.3: (A) Co-transcriptional splicing pattern in human brain in the genes AUTS2 (A. top) and C21orf34 (A. bottom). A clear sawtooth pattern is visible when looking at the histogram of read coverage across introns. (B. top) Diagram showing the formation of pre-mRNA creating the 'sawtooth' pattern across introns. (B. bottom) Extrapolation of the expected pattern to model intronic read coverage. [Ameur et al., 2011]

the spliceosome. circRNAs are generated co-transcriptionally, their production rate closely related to their flanking introns. Canonical mRNA compete with circularization in a tissue-specific manner conserved in vertebrates. This introduces a potential function for circRNAs as regulators of gene expression by competing with linear transcripts. [Ashwal-Fluss et al., 2014]

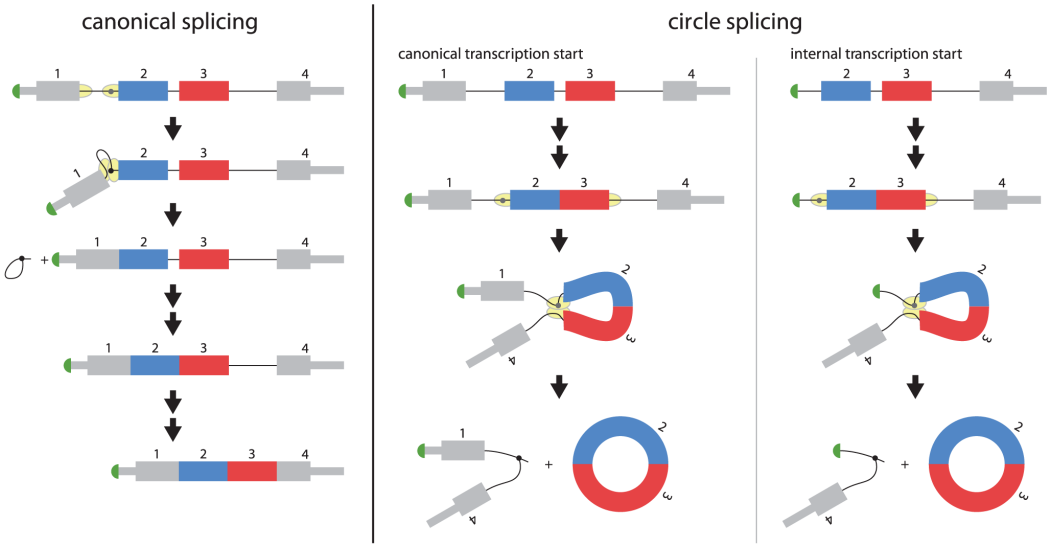

Figure 1.4: Diagrams depicting formation of circular RNA via canonical or internal transcription compared to linear mRNA formation [Salzman et al., 2012]. Canonical and circle splicing are mutually exclusive and are in direct competition.

1.3 Analysis of RNA-seq data and evaluation of splice junctions

This study implements tools and pipelines to exploit an underutilized aspect of RNA sequencing, the splice junctions. A splice junction is inferred from those sequence fragments that overlap exon-exon boundaries in the mRNA and hence map to different exons on the genome (Figure 1.5). This is an incredibly powerful tool to estimate splicing efficiency, provide nucleotide resolution on splicing reactions and reveal patterns not readily evident from gene or exon expression.

1.3.1 Alignment of NGS fragments

Alignment of read fragments produced by NGS technology to the genome/transcriptome is the first step in processing human sequencing data. This can be achieved with a number of different algorithms. The alignment of millions of read fragments requires heuristic approaches for accurate mapping in reasonable wall clock time. This is achieved by sacrificing sensitivity, generally a maximum of 2 mismatches is allowed. Recent advances have improved the level of permutation permitted but this generally results in performance loss. BWA [Li and Durbin, 2009] and BOWTIE [Langmead et al., 2009] both apply the burrows wheeler compression transform to enable fast searching of read space. There is now a second generation of aligners commonly used such as BOWTIE2 [Langmead and Salzberg, 2012].

1.3.2 RNA-seq analysis software

The analysis of RNA-seq data remains highly dependent on different applications of the technology to organism and features of interest. There are multiple methods and approaches to take into consideration. One strategy is the mapping of the sequence fragments to the genome, recovering splice junctions using either predefined exon-exon scaffold reads (such as Tophat2 [Trapnell et al., 2009; Kim et al., 2013]) and/or independent alignment of subsequences (such as GSNAP [Wu and Nacu, 2010], HISAT [Kim et al., 2015], STAR [Dobin et al., 2013]). A second method is direct alignment to the transcriptome and quantification of transcript values through assignment of read counts to isoforms (such as RSEM [Li and Dewey, 2011], Kallisto [Bray et al., 2015]). In this study the focus is on genomic alignment enabling identification of novel features.

There are several common software for quantification of isoforms, a typical example is Cufflinks [Trapnell et al., 2010a, 2012]. Although this does not directly measure splicing change it does enable the building of novel isoforms from the splice junction information. The general steps involve assembling tran-

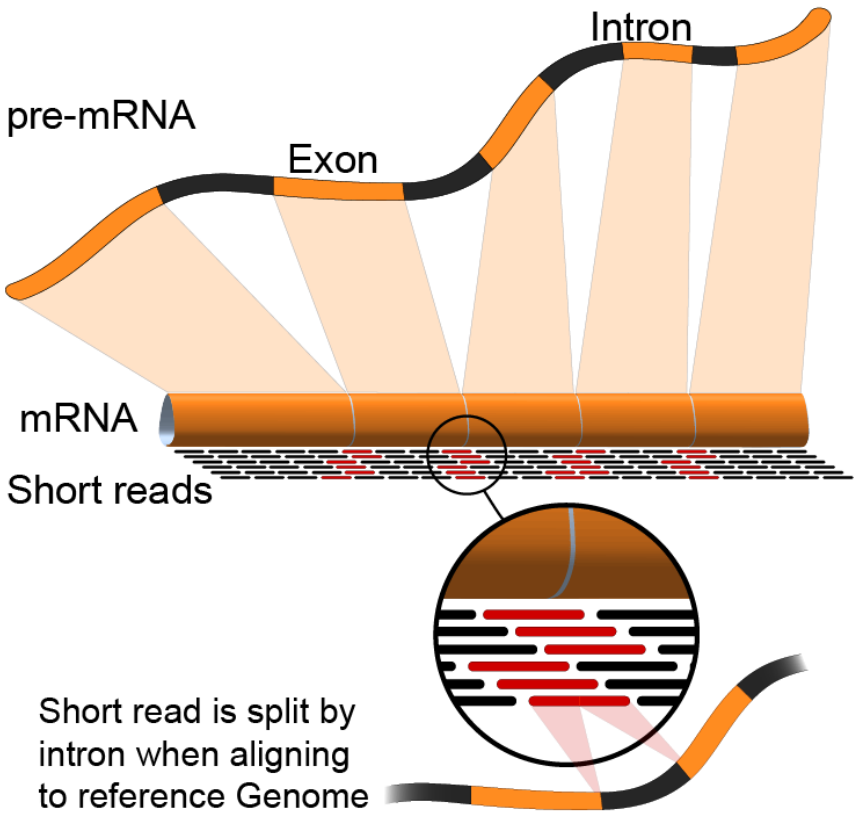

Figure 1.5: Illustration showing how splice junction reads (red) are mapped to two distinct genomic positions based on their constitutive exons (orange). This provides direct evidence of splicing in the cell. [Wikimedia Commons, 2009].

scripts based on splicing information, comparison and merging with known annotations and finally differential expression based on the enhanced annotations.

Splice junction reads are essential to accurate splicing analysis

Splice junction reads are under-utilized as a means to identify and quantify splicing changes. The mixture-of-isoforms (MISO) model [Katz et al., 2010] is probably the most notable exception to this. MISO was the first, popular tool to investigate splice junctions between alternate exons. Figure 1.6 shows the basic outline of the software. It calculates the levels of inclusion of alternate exons using the 'percent spliced in' (PSI) statistic. This is calculated based on the ratio between splice junctions that support its inclusion compared to those that connect the constitutive exons. In order to determine differential splicing a Bayes factor is applied which quantifies the odds of differential exon usage in two sample groups. Posterior probability distributions of PSI are calculated and used to estimate the Bayes factor.

MISO remains one of the most robust tools, one of the few to look at single skipping events rather than complete isoform expression. However, it is not designed to handle low frequency cryptic events as it requires input of all exons to be tested. The inclusion of cryptic events generally requires enough read depth to build an exon structure.

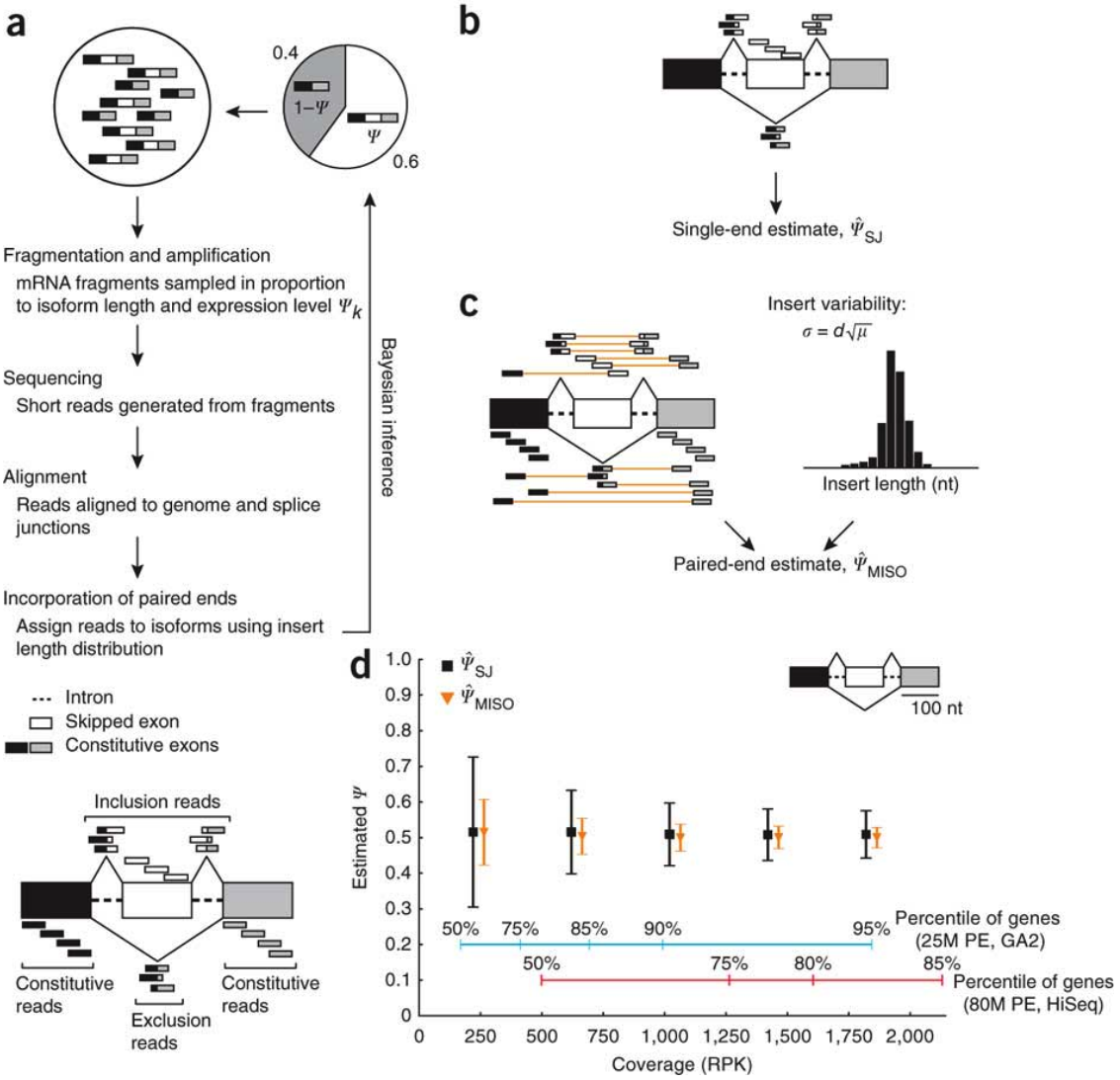

Figure 1.6: MISO work-flow (A) Work-flow showing read fragments from sequencing to alignment and quantification. Fragments aligning to constitutive exons are marked in black and grey, alternative exons in white (B) Psi estimate uses alternate exon reads and splice junctions. (C) Using paired end reads greatly improves results by allowing for insert size to be used (shown in orange) along with the insert distribution (D) A graph showing the estimated psi parameter based on single/paired-end reads based on read coverage. These results were generated by re-sampling at different depths with standard deviations included. Ultimately, paired-end reads have an appreciable effect on expression variance. [Katz et al., 2010]

More recently tools such as Junction-seq have been designed to examine splicing efficiency in a

similar method to exon differential expression requiring two groups to test [Love et al., 2014; Hartley and Mullikin, 2015]. In the following studies, the non-canonical nature of the splicing required the creation of custom algorithms and often required detailed analysis of raw alignment data.

1.4 Sequencing datasets utilized in this thesis

1.4.1 Primary datasets

Two primary datasets are used for multiple analyses in this study (Table 1.1). Each will be discussed below in detail.

Consortium Project	Application	Detail	Number of Samples
UKBEC Consortium	total RNA-seq	Human brain	48
UCLex	Exome	Multiple studies	3,500

Table 1.1: Primary next generation sequencing data analysed in this study.

UKBEC RNA-seq Brain data

Brain samples were collected from the Medical Research Council Sudden Death Brain and Tissue Bank (Edinburgh, UK). Post-mortem human tissue from four individuals of European descent confirmed to be neurologically normal during life. Twelve central nervous system regions were sampled from each individual. The regions studied were: cerebellar cortex, frontal cortex, temporal cortex, occipital cortex, hippocampus, the inferior olivary nucleus (sub-dissected from the medulla), putamen, substantia nigra, thalamus, hypothalamus, intralobular white matter and cervical spinal cord. The libraries were sequenced using Illuminas HiSeq2000 with 100 base pair paired-end reads. Sample data can be found in the Appendix Table 1.

UCLex Exome consortium

UCLex is an in-house consortium of custom capture and whole exome sequencing data. It consists of over 3,500 exomes. These are a collection of rare, mendelian type disease and common disease groups with the addition of healthy controls. As the data are in-house this provides a unique opportunity to explore variation across samples and the ability to inspect individual variants within sample groups. The power of a consortium such as this is the standardised processing and variant calling. This allows for improved

filtering of spurious variation and an estimate of variant frequency in various disease conditions. Details on the distribution of samples between groups can be found in Table 1.2.

Phenotype	#Samples
Inflammatory Bowel Disease	799
Huntington's Disease	48
Ophthalmology, Retinal disorders	371
Dermatology, Inflammatory disorders	63
Sudden Cardiac Death	98
Keratoconus	12
Primary Immunodeficiency	128
Prion Disease	1112
Epilsepsy	164
ARVC	28
Bone Marrow Failure	184
Cone Rod Dystrophy	40
Healthy Controls	892

Table 1.2: UCLex Sample Information. Phenotype and number of samples

1.4.2 Public datasets analysed

Each thesis chapter makes extensive use of public data to expand on hypotheses and reinforce findings. Table 1.3 shows the data consortia analysed and applied in this study. Each dataset will be introduced briefly.

Consortium Project	Application	Detail	Number of Samples
GTEx Consortium	polyA+ RNA-seq	Multiple human tissues	1,749
Illumina Bodymap	total RNA-seq	Multiple human tissues	48
ENCODE	CHIPseq	Histone mark,human and mouse brain	12
ExAC Consortium	Exome	Multiple studies	61,000
GEUVADIS project	Exome, polyA+ RNA-seq	Multiple studies	426

Table 1.3: Public next generation sequencing consortia data analysed in this study.

ENCODE

The Encyclopedia Of DNA Elements (ENCODE) was launched in September 2003 to identify functional elements throughout the human genome. The landmark project aimed at using high-throughput approaches on a variety of functional elements. ENCODE targets range from genes, promoters, enhancers, to transcription factor binding sites, methylation sites and histone modifications. For the purpose of this study I will only focus on a subset of these data, specifically histone modification data taken from mouse and human brain. [The ENCODE Project Consortium, 2004]

ExAC Consortium

Exome Aggregation Consortium (ExAC) is an initiative from a collaboration of groups centred around the Broad Institute and Massachusetts Institute of Technology. The goal is to aggregate vast collections of whole exome data and provide general statistics to the scientific community. The dataset currently spans a staggering 61,000 unrelated individuals. [Samocha et al., 2014; The EXaC Consortium, 2015]

GTEx Consortium

Genotype-Tissue Expression (GTEx) project provides a large resource for interpreting tissue-based gene expression, regulation and its relationship to variation. This project aims to study gene expression differences between multiple human tissue types and compare this to genotype information. This information has been used to calculate expression based quantitative trait loci (eQTL) and provide a large, publicly available dataset for further scientific investigation. This consortium contains over 237 post-mortem donors, with 28 tissue samples per donor spanning 54 distinct body sites. Paired-end mRNA sequencing was performed on a total of 1749 samples, with an average of 82 million reads per sample. This information was processed into files freely downloadable from their website (http://www.gtexportal.org/home/). [The Genotype-Tissue Expression (GTEx) project Consortium, 2015]

Illumina Bodymap

The Illumina Bodymap is a resource of 16 human tissues made available from Illumina sequencing. A total of 48 samples (including biological replicates) were sequenced using a ribosomal RNA depletion protocol to produce paired-end sequenced data (www.illumina.com; ArrayExpress ID: E-MTAB-513).

GEUVADIS project

The GEUVADIS consortium combined RNA-seq from lymphoblastoid cell lines of 465 individuals with variant data from the 1,000 Genomes Project. A subset of 423 samples were analysed as these were part of the 1,000 Genomes Phase 1 dataset. Paired-end, 75bp RNA-seq was performed on total RNA of the 465 EpsteinBarr-virus-transformed lymphoblastoid cell lines. This resulted in an average of 48.9M reads per sample. [Abecasis et al., 2012; Lappalainen et al., 2013]

1.5 Overview of chapters

A brief overview of chapters in this thesis is outlined below.

1.5.1 Long genes, recursive splicing and their relationship to the brain

Lately, several studies have indicated that long genes are linked to several neurological disorders [Lagier-Tourenne et al., 2012; Polymenidou et al., 2011; King et al., 2013]. I investigate the characteristic differences in these long genes by noting their enrichment in brain and distinct epigenetic profile (compared to shorter genes). It is noted that in *Drosophila melanogaster* long introns can contain cryptic elements known as recursive splice sites (hereafter: RSS) that allow for processing of large introns. Recursive splicing is the reconstitution of a 5' splice site after an initial splicing reaction, hence resulting in no exonic inclusion. In this chapter RSS are identified for the first time in long genes within the human brain using a custom designed pipeline. The proposed function of these recursive elements is the maintenance of canonical upstream exons. Alternate cryptic promoters result in inclusion of a 'poison' exon that marks the transcript for nonsense-mediated decay (NMD).

1.5.2 Circular RNA are a pervasive phenomena linked to neuronal genes

The circularization of exons is far more pervasive than first believed. Circular RNAs (hereafter: circRNA) are not only enriched in neuronal genes but their synthesis appears to be partially regulated through RNA binding proteins [Rybak-Wolf et al., 2015; You et al., 2015; Venø et al., 2015; Ashwal-Fluss et al., 2014]. Current research also suggests that circularization partially regulates transcription by reducing the production of canonical mRNA. Here I develop a pipeline to mine a large brain cohort for circular RNA, produce high confidence counts and explore the circRNA landscape in brain. From these findings I identify a subset of highly expressed back-splice junctions that occur between two proximal genes from the same family. In order to explore the presence or absence of these junctions a custom alignment algorithm was implemented.

1.5.3 Exploring polymorphic variation using large exome consortia

Large exome consortia provide a unique opportunity to use variant information in ways never before possible. For instance, it is possible to identify important genes due to significant deficiency in deleterious variation. This relationship has been extrapolated in this study to identify positions within features that are highly invariant. A pipeline was developed to create a nucleotide resolution map of splice sites and branchpoints

based on exome data. This identified specific positions in splice sites and branchpoint motifs that show fewer mutations.

In order to explore the effects of this variation a pipeline was created to integrate splicing variants with gene expression data. Gene/exon expression did not provide significant resolution between variant/wildtype groups highlighting the need for more sensitive measures. Multiple statistics were applied using splice junctions which provided far better resolution of functional change. This change in splicing efficiency can be partially captured through a sequence-based variant effect score.

Chapter 2

Recursive splicing in human brain

2.1 Introduction

Work from this thesis chapter has been published in [Sibley et al., 2015].

2.1.1 Long genes form a subclass with unique characteristics

The human transcriptome contains a great diversity of genes, one of the striking aspects is their length. Gene lengths range from the Tyrosine tRNA (0.2kb) to the Dystrophin gene (2500kb). Over 2,000 genes in the human genome are longer than 150kb, more than ten times the average gene length (10-15kb). This raises questions as to the diverse roles these genes could play and what unique mechanisms are required for correct transcription of this extreme subclass.

In this chapter I will focus on the characteristics of long genes and their relevance to clinical pathology, particularly neurological, and the discovery of cryptic elements known as recursive splice sites.

Long genes as candidates for neurological pathology

Recent literature suggests that disruption of long genes (150kb+) are a key component of several neurological disorders. Long genes differ from shorter genes in transcript processing and key RNA binding proteins. Both TDP-43 and FUS/TLS (Fused in sarcoma/translated in lipsarcoma) are elongation factors enriched in large introns [Polymenidou et al., 2011; Lagier-Tourenne et al., 2012]. Their absence reduces expression of these genes substantially by affecting transcript stability of intronic sequences. Similarly, Topoisomerase 1 (TOP1), a protein that resolves DNA super-coiling, decreases expression of long genes in a dose-dependant fashion in

neurons of both mouse and human [King et al., 2013].

Disruption of the MECP2 gene, involved in methylation and transcriptional repression has been shown to cause Rett syndrome, an autism-like disorder with severe neurological implications [Chahrour and Zoghbi, 2007]. This genes methylation function is widespread but long genes are particularly susceptible due to their increased length and hence the effect of this gene is largely present in brain and neurons [Gabel et al., 2015].

Recently, recurrent DNA double-strand breaks (RDCs) that occur in primary neural stem progenitor cells (NSPCs) have been found within long genes. Almost 90% of these genes are involved in synapse function and/or neural cell adhesion indicating length may play an important role in this process which is essential for neural development. [Wei et al., 2016a]

These findings suggest that long genes are intimately linked to neurological dysfunction. Further investigation into the processing of long genes is essential to improve our understanding of splicing mechanisms within these genes.

Long introns as a peculiar feature of long genes

A common characteristic of long genes is that they often contain one or more long introns. Long introns are generally defined as those longer than 50kb. Over 3,000 human introns are larger than 50 kb, with nearly half being longer than 100 kb [Belshaw and Bensasson, 2006]. Long introns place a high resource burden on the cell and raise several questions regarding their presence in higher eukaryotes. Firstly, the transcription of large introns requires a massive energy commitment to produce pre-mRNA that will be removed and degraded. Secondly, this increases time of transcription for creation of the mRNA (a 150kb intron takes nearly 20min to transcribe!). Thirdly, large introns increase the likelihood of inclusion of cryptic splice sites either through translocation, mutation or RNA binding protein dysfunction [Belshaw and Bensasson, 2006]. These were initially defined as pseudo-exons or cryptic exons [Sun and Chasin, 2000]. One source of cryptic exons are repeat elements which are surprisingly common in higher eukaryotes.

Repeat elements are abundant in long introns and play a role in their processing

One central question to the transcription and correct splicing of long introns is how splice sites can be efficiently connected across such a huge distance. In vertebrate genomes introns are significantly enriched for interspersed repetitive elements (mainly SINEs and LINEs). Figure 2.1 shows the presence of repeat elements across various species. These repeats have enough complementarity to form stems in large introns,

effectively folding the intron into compact structures [Shepard et al., 2009]. Similar RNA hairpins are crucial for splicing of group I and group II introns present in bacteria [Pyle et al., 2007]. These hairpins found in eukaryotes have also been shown to regulate alternative splicing [Rogic et al., 2008]. In human an average of 9.4 possible hairpins were found per 50kb of intron, these mostly formed between oppositely oriented primate-specific Alu-repeats (81.7%) [Belshaw and Bensasson, 2006]. An interesting exception to this case is chicken, where LINE elements replace this functionality [Belshaw and Bensasson, 2006]. In conclusion, these RNA hairpins enable the folding of intronic RNA and would significantly reduce the distance between donor and acceptor splice sites making them central to the correct processing of large introns.

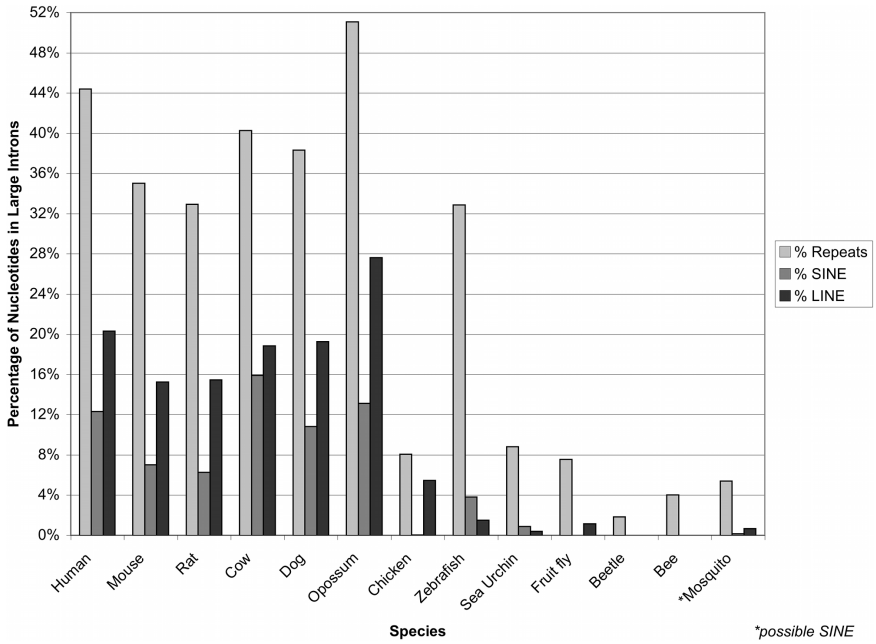

Figure 2.1: The percentage of different interspersed repeats for the complete set of large introns for various species. The light grey bars are for the total percentage of repeats in large introns. The dark grey bars are for the short interspersed element repeats. The black bars are only for long interspersed element repeats. [Belshaw and Bensasson, 2006]

Repeat elements can form cryptic elements if not repressed

An interesting side effect of long vertebrate introns containing many Alu repeats is that their splice sites need to be masked by RNA-binding proteins (RBPs). hnRNP C is an example of an RBP that represses Alu elements [Zarnack et al., 2013]. The uncontrolled expression and splicing of these elements could lead to disease and as such much be tightly regulated by the cell [Dhir and Buratti, 2010]. The evolution of a repeat element as it gets included in a transcript is highlighted in Figure 2.2.

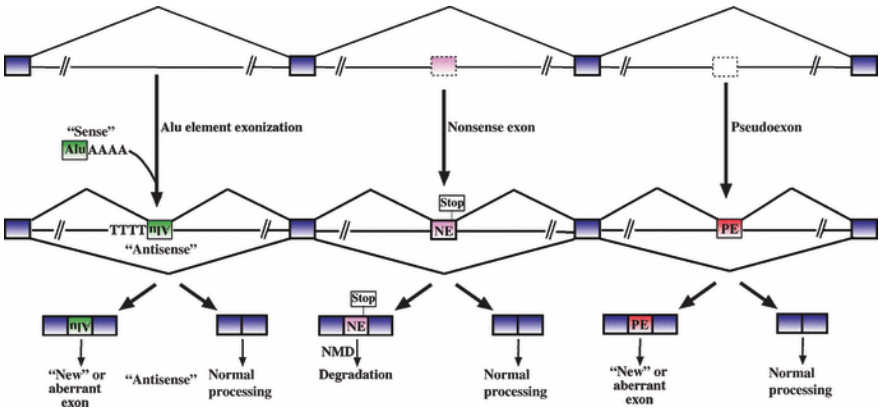

Figure 2.2: Schematic showing exonisation of Au elements and its evolution to non-sense element (NE) and finally to pseudoexon (PE) potentially conveying a novel function often resulting in human disease. [Dhir and Buratti, 2010]

The evolution of Alu exons into cryptic elements indicate the need to carefully explore introns for similar features. Potentially long introns may harbour some of these elements, potential remnants of evolution that may fulfil important functions.

2.1.2 Cryptic elements in long genes

Recursive splicing is a novel mechanism observed in vertebrates to process long introns

Interestingly, invertebrate genomes that contain long introns are almost completely deficient of repeat elements hence stem/hairpin structures are practically absent [Belshaw and Bensasson, 2006]. Hence, another mechanism must be operating to process these long introns. In Drosophila it was discovered that large introns undergo a process called recursive splicing (Figure 2.3) to remove an intron by processing it in multiple steps [Hatton et al., 1998; Burnette et al., 2005].

Recursive splicing was first discovered in Drosophila melanogaster's Ultrabithorax (Ubx) gene as a mechanism to process its long intron [Hatton et al., 1998]. The 73kb Ubx intron was spliced out in four steps, the final step including a recursive site. Recursive sites consist of back-to-back 3' and 5' intronic splice sites thereby creating a zero-length exon. Computational analyses have predicted nearly 200 recursive sites in D. melanogaster, 7 of which were validated by inhibiting lariat de-branching enzymes [Burnette et al., 2005]. A recent breakthrough (and co-publication of [Sibley et al., 2015]) identified nearly 200 recursive sites on a genome-wide scale in Drosophila by leveraging RNA-seq data using splice junctions and co transcriptional splicing patterns [Ameur et al., 2011; Duff et al., 2014].

Given the enrichment of recursive splicing in long genes [Duff et al., 2014] and the established

relationship between long genes and neuronal tissue [Polymenidou et al., 2011; Lagier-Tourenne et al., 2012], I endeavoured to explore the existence of recursive sites in the long genes (150kb+) within human brain.

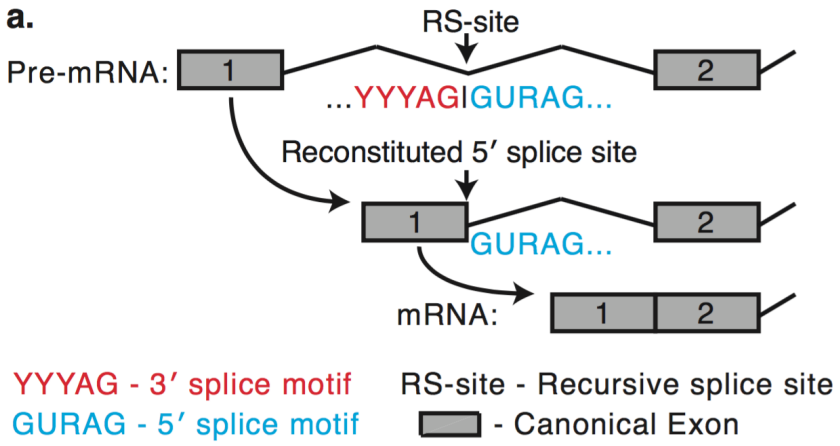

Figure 2.3: Schematic of the recursive site creating a zero length exon in fruit fly needed to process long introns . [Sibley et al., 2015]

2.1.3 Histone modifications are a central characteristic of genes and their cell specific expression

Another crucial aspect of genes is their chromatin state. This is largely determined by histones which are composed of highly conserved proteins (H3, H4, H2A, H2B and H1). These function as building blocks, packaging DNA into nucleosomes that can be folded into chromatin super structures [Luger and Richmond, 1998]. Histones can be posttranslationally modified, most commonly on their tails. The N and C terminals protrude from the core nucleosome and have the potential to be modified and to interact with neighbouring nucleosomes. This can function as binding sites for other proteins and regulate chromatin structure. Histone modifications include acetylation, phosphorylation, methylation, and ubiquitylation.

Histones have been identified as integral components of the machinery that modulates gene transcription, repair, replication and recombination [Strahl and Allis, 2000]. A list of histone marks associated with transcription are shown in Figure 2.4. One of the most well known and well defined histone modifications related to transcription is the trimethylation of H3 lysine 36 (H3k36me3).

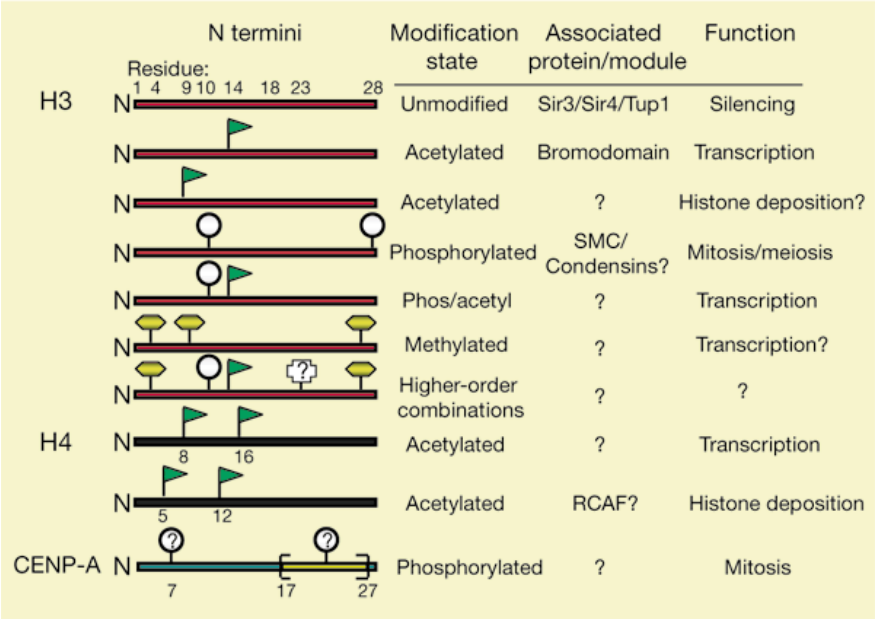

Figure 2.4: Histone modifications known at different residues within the N terminus of histones, some of these have been associated to transcription. [Strahl and Allis, 2000]

H3k36me3 histone modification provides insight into transcription

H3k36me3 is widely recognised as a transcriptional histone mark. It has been linked to active gene bodies, splicing, repair and gene activation [Jenuwein and Allis, 2001; Kolasinska-Zwierz et al., 2009; Sims and Reinberg, 2009]. H3k36me3 is enriched around exons, although alternatively spliced exons tend to show lower levels, indicating a link with transcription and splicing [Sims and Reinberg, 2009]. Transcription can also accumulate further histone marks creating a feedback effect. This may also explain why levels of H3k36me3 also tend to increase toward the 3' end of the gene [Pokholok et al., 2005].

Interestingly, although there are multiple writers of H3K36 methylation only SETD2 is responsible for H3K36 trimethylation. Not surprisingly, SETD2 has been identified as instrumental in dealing with cancer's aberrant transcription [Pfister et al., 2015]. H3k36me3 deacetylation is necessary after transcription to prevent initiation of transcription from aberrant sites within the gene [Pokholok et al., 2005].

Duff et al. in their analysis of recursive sites in fruit fly did not identify any histone marks enriched for recursive sites in Drosophila, however they did not evaluate whether intron length had any impact on histone enrichment [Duff et al., 2014]. The further investigation of H3k36me3 is another key characteristic that can help understand processing differences between long and short genes.

2.2 Methods

2.2.1 Software and tools used in this Chapter

Several software packages were used extensively in this thesis. Python [Cock et al., 2009] (programming language) was used for general file parsing, scripts to wrap and automate other tools and custom analysis on BAM/SAM alignment files. Python packages include Pysam, Biopython and Bedtools libraries [Cock et al., 2009; Quinlan and Hall, 2010; Li et al., 2009].

R statistical [R Core Team, 2016] was applied for normalisation of gene expression (DESeq, DESeq2 [Anders and Huber, 2010]), general matrix manipulation (dplyr,tidyr) and plotting of data (ggplot2) [Wickham, 2009; Wickham and Francois, 2015; Wickham, 2016]).

Bedtools [Quinlan and Hall, 2010] was used for manipulating genomic coordinate data. This tool is without a doubt the most essential to any bioinformaticians kit.

2.2.2 Ancillary public datasets

Both public and primary data were used in this study. The UKBEC Brain consortium was used as primary data. Public RNA-seq datasets used in this study include; GTEx consortium [The Genotype-Tissue Expression (GTEx) project Consortium, 2014] and Illumina Bodymap version 2.0 [Derrien et al., 2012]. Please see Table 1.3 for more information.

Additional datasets analysed for this study alone include C2C12 mouse myoblasts (GSM521256) and myogenic lineage (GSM521259) [Trapnell et al., 2010b] , mouse embryonic stem cells (GSM1346027) and motor neurons (GSM1346035) [Herrera et al., 2014], and differentiation of haematopoietic stem cells (GSM992931) into erythroid lineage (GSM992934) [Madzo et al., 2014].

2.2.3 Expression of long genes in the brain

I quantified the relationship between expression of long genes in neurons and all major tissue types available in the GTEx consortium [Genotype-Tissue Expression Consortium, 2015], Illumina Bodymap version 2.0 (www.illumina.com; ArrayExpress ID: E-MTAB-513) and several mouse cell lines during differentiation (details in Datasets section). For GTEx, gene count data were downloaded, normalised and fold change ratios where calculated using DESeq [Anders and Huber, 2010]. The ratios were correlated to gene length to determine trends in the data.

For the Illumina Bodymap (www.illumina.com; ArrayExpress ID: E-MTAB-513) and ancillary mouse datasets; raw sequence data were downloaded, mapped to their respective genome assemblies (hg19 for human and mm9 for mouse) using Tophat2 [Trapnell et al., 2009]. Aligned reads were summarised into gene counts using HTSeqCount [Anders et al., 2015] and differential expression between relevant groups done with DESeq[Anders and Huber, 2010].

Plots were created using the log fold change calculated by DESeq and length of the longest transcript for each gene. Loess smoothing curves were calculated and graphing was done using ggplot2 [Wickham, 2009].

2.2.4 Identification of recursive splicing in brain

An in-house bioinformatics pipeline was created to process the UKBEC Brain consortium sequence data (Figure 2.5). Raw FASTQ data were aligned to the human genome (build hg19) using the STAR aligner (v2.3 [Dobin et al., 2013]) with enhanced splicing annotations from GENCODE v19 [Steijger et al., 2013]. All aligned BAM files were pooled and only reads within long introns (150kb+, n=943 in 780 genes) were selected. An algorithm then scanned for split reads (also referred to as junction reads) that mapped to a canonical exon and terminated intronically. Junctions were classified as known or novel using the knowngene UCSC annotations [Abe et al., 2015].

These detected junctions were then enumerated and paired if they spanned the intron in an exon-like fashion with a maximum of 400bp gap between the junctions (see PE1 and PE2 from Figure 2.5). For these potential exons, all stop codons in all frames were identified.

Furthermore, all junctions from the 5' end of the upstream exon were identified and classified based on their presence in UCSC, RefSeq and GENCODE databases [Harrow et al., 2012; Abe et al., 2015; O'Leary et al., 2016], all novel junctions were noted as potential cryptic upstream elements.

Another method to identify recursive sites is using the co transcriptional "saw-tooth" pattern created by pre-mrna transcripts mapping to the intron [Ameur et al., 2011]. For more information please refer to Section 1.2. Pooled BAM files were summarised into 5kb bins and these were used to perform linear regression. As transcription occurs at a constant rate, similar gradients should be seen across all introns within a genes. Dividing the large intron based on the above identified splice junctions I could determine whether the new gradients would resemble other introns within the gene (Figure 2.6).

Similarly, an alternate dataset was generated using cross-linking and immunoprecipitation (iCLIP) for fused in sarcoma (FUS) in human brain. iCLIP analysis of FUS binding enables linear regression analysis in a similar way to total RNA-seq data and results in a saw-tooth patten.

Final classification as a recursive splicing element, required adequate splice junction coverage, the significant improvement of co-transcriptional gradient with addition of the splice junction in both RNA-seq and FUS iCLIP datasets.

H3k36me3 histone enrichment in long genes

In order to determine whether long introns have unique histone enrichment patterns compared to short introns, data from both mouse (mm9) and human (hg19) brain was downloaded for H3k36me3 and a control enhancer mark, H3k4me1 (ENCODE project data) [Parkhomchuk et al., 2009]. For mouse, embryonic brain tissue was used (accessions: GSM1000072, GSM1000096) and for human ; Cingulate Gyrus, Hippocampus (Middle) and Mid Frontal Lobe from adult (accessions: GSM669947,GSM773013,GSM773052).

For each dataset, reads were binned into 100bp windows, normalised by total read count. Each exon was flanked by 400bp to compensate for exonic histone signal extending into the intron. Introns were adjusted accordingly. All bins overlapping exons were summed and a mean value was taken across introns. All introns were binned according to size in the following categories: 400nt-2kb, 2-5kb, 5-20kb, 20-50kb, 50-100kb, 100+kb. Only transcripts with at least one intron of 50kb+ were selected.

All introns and exons were normalised using the shortest intron bin (400bp-5kb). RPKM (Reads Per Kilobase of transcript per Million mapped reads) values were then calculated for both exons and introns based on their normalised values. Pearson correlation is then calculated between normalised intron counts and intron length. For exons the normalised exon RPM was correlated to the shortest neighbouring intron.

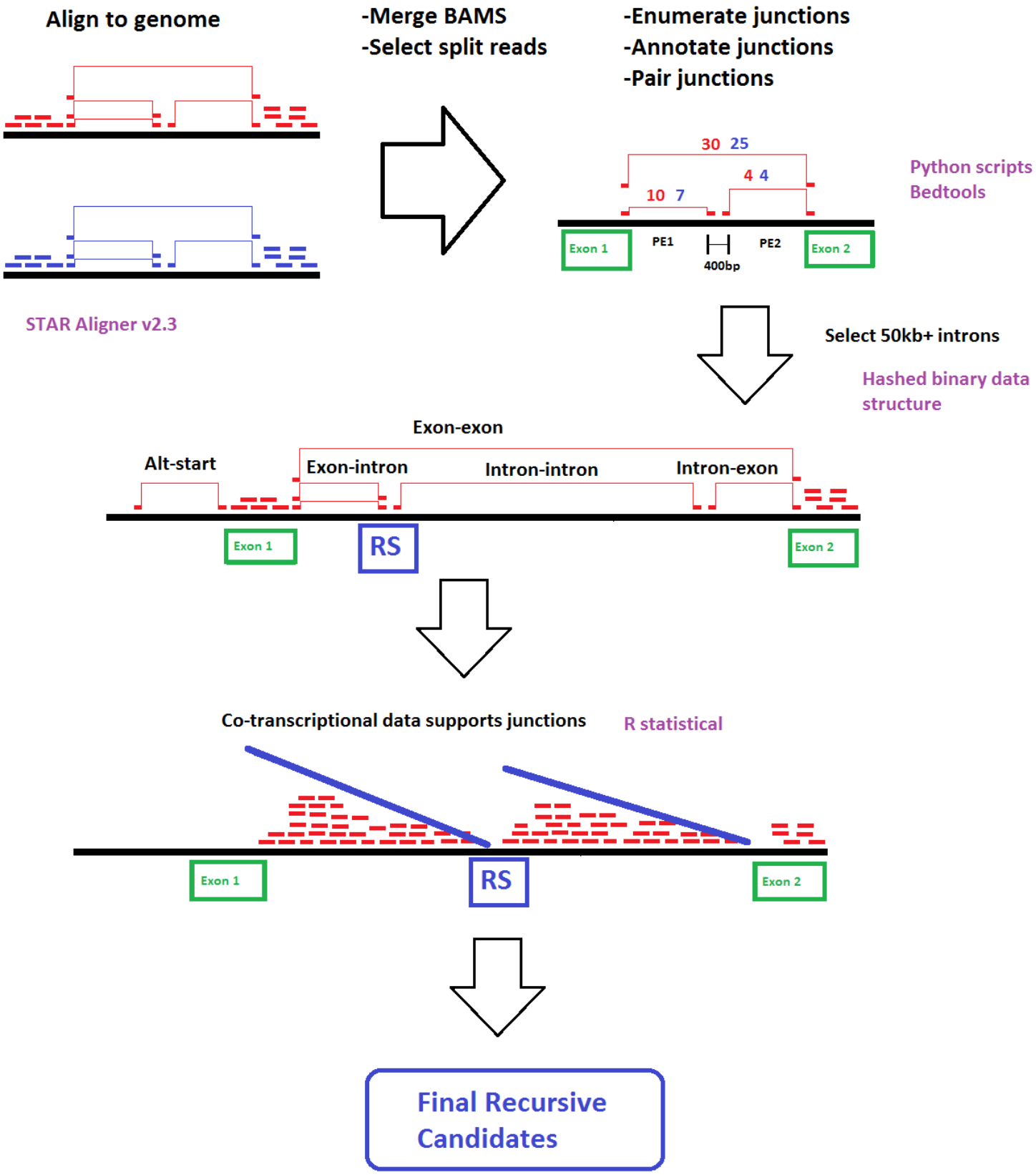

Figure 2.5: Recursive splicing pipeline. Briefly, data are aligned to the human genome, BAM alignment files are parsed for splice junction information which is used to annotate any recursive-like junctions that exist within large introns. If a potential recursive junction is found the co transcriptional pattern of the intron with or without the proposed recursive site is checked.

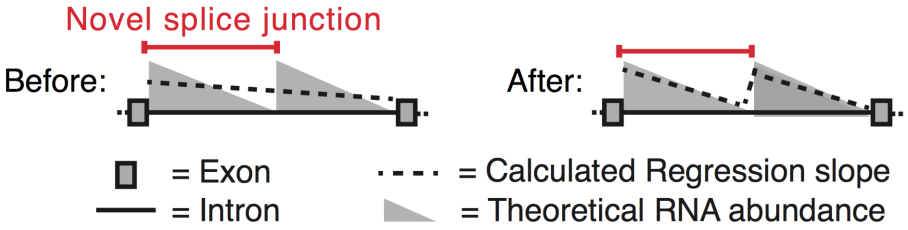

Figure 2.6: The impact of inclusion of the recursive site to modelling the co-transcriptional sawtooth pattern present in RSS genes. Through use of splice junction data, effectively dividing the intron into two. This results in significantly improved goodness of fit of the linear model. [Sibley et al., 2015]

2.3 Results

2.3.1 Expression of long genes is enriched in the brain

It has already been observed in ES cells that long genes, which are mostly silent in an undifferentiated state, become expressed in neurons [Thakurela et al., 2013]. Long genes (150kb+) appear to be consistently more highly expressed in brain in both GTEx (Figure 2.7) and Illumina Bodymap datasets (Figure 2.8). The Illumina Bodymap data includes both Dystrophin and Titin, well known long genes that are highly expressed in muscle. These tend to follow the expected trend, however, Dystrophin's longest intron is only 45kb, this could explain its slightly higher expression pattern.

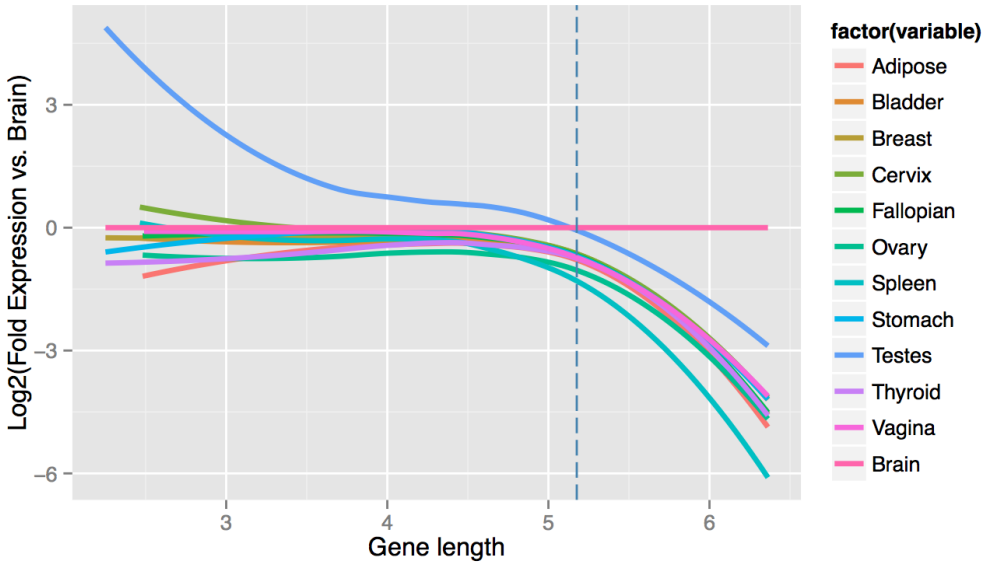

Figure 2.7: GTEx data comparisons by tissue show that long genes are more highly expressed in brain compared to other tissues. Dotted blue line indicates 150kb gene length. Plot shows gene length (log10) as a function of ratio of expression in tissue / expression in brain. Data are represented as Loess smoothing curves. Trendlines indicate an overall enrichment of expression of long genes in the brain compared to all other measured tissues.

For further investigation several other public datasets were explored. Figure 2.9 shows the increase in expression of long genes during differentiation of mouse embryonic stem cells into motor neurons compared to myogenic and erythroid differentiation.

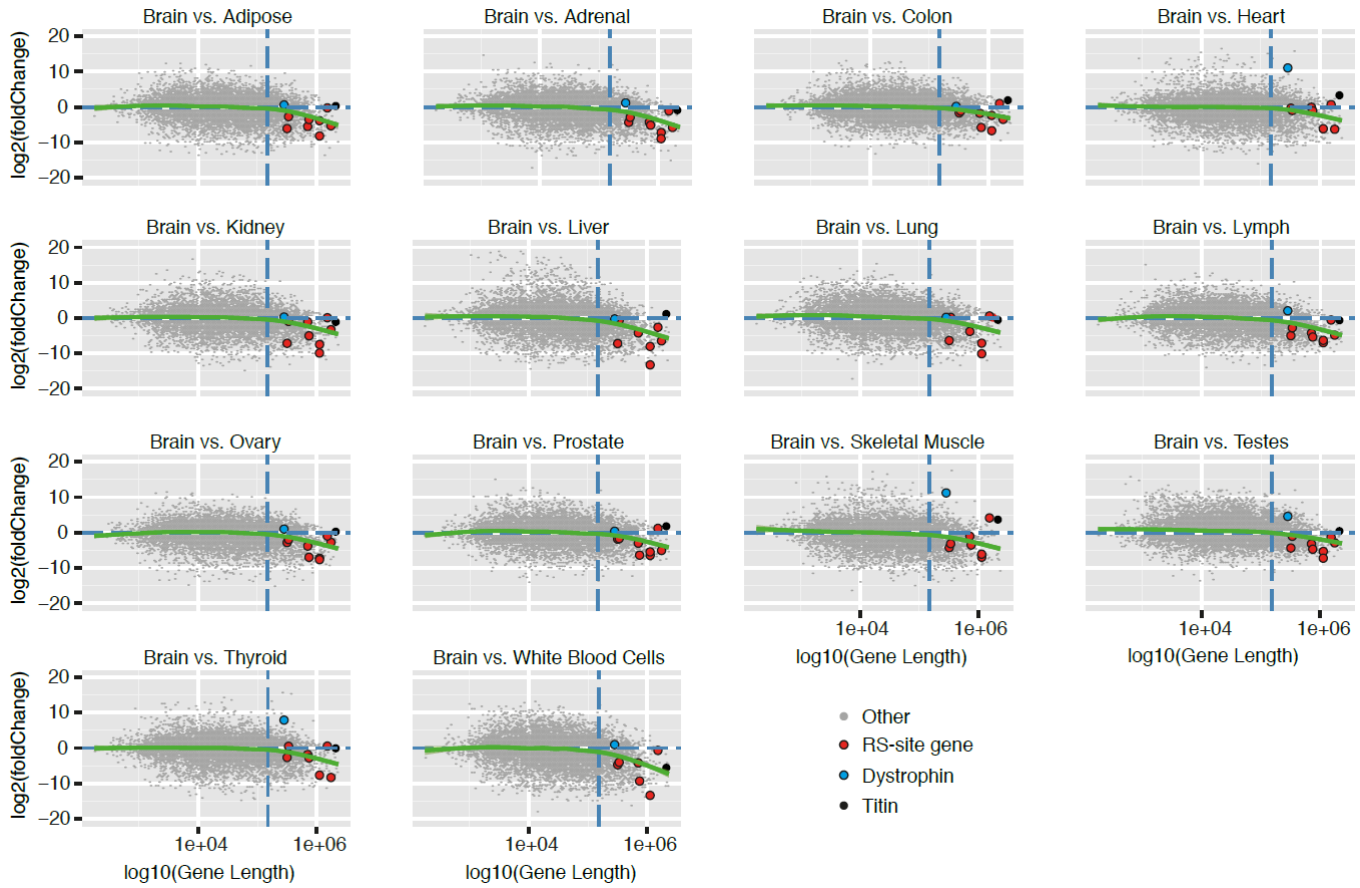

Figure 2.8: Multiple plots from Illumina Bodymap II resource. Graphs show gene expression in each tissue relative to brain (log2foldchange). Genes containing recursive sites (red) and two long genes highly expressed in muscle tissue (dystrophin and titin) are highlighted. All remaining genes are in grey.[Sibley et al., 2015]

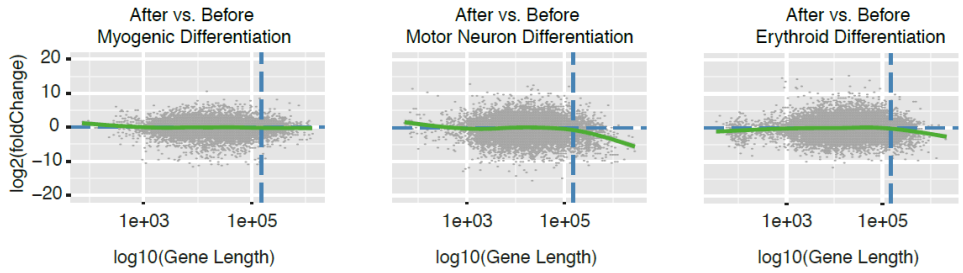

Figure 2.9: Public data showing effects of differentiation on different cell lines (after versus before) as a log fold change. Samples analysed from left to right include; C2C12 mouse myoblasts (GSM521256) into myogenic lineage (GSM521259) [Trapnell et al., 2010b] , mouse embryonic stem cells (GSM1346027) into motor neurons (GSM1346035) [Herrera et al., 2014], and differentiation of haematopoietic stem cells (GSM992931) into erythroid lineage (GSM992934) [Madzo et al., 2014]. [Sibley et al., 2015]

2.3.2 Recursive splice sites identified in human brain

Figure 2.10 highlights the filtering process from 1.5 billion reads, to the 3,000 novel junctions and finally to the 11 confirmed recursive sites in 9 genes (see Table 2.1). Part of the filtering process was modelling the co-transcriptional splicing pattern and determining if the inclusion of the RSS improved regression gradient. This could be done in both RNA-seq and FUS iCLIP data and clearly shows the consistent use of the recursive sites in these genes (Figure 2.11).

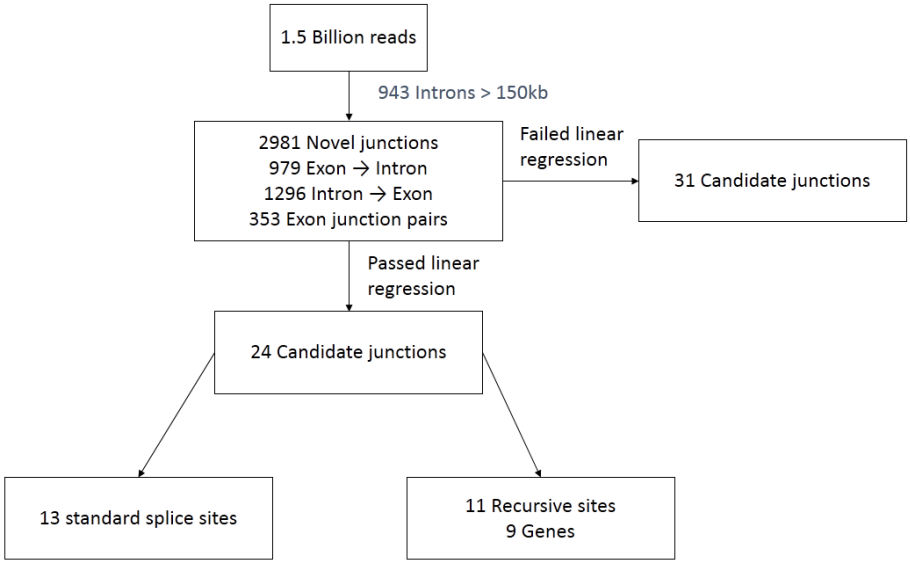

Figure 2.10: Filtering of read junctions through the recursive pipeline. Pooling all brain samples 1.5 billion reads is reduced to 2981 splice junctions within long introns to the final 11 recursive sites found in 9 genes.

Recursive sites showed a strong consensus 3' intronic splice site immediately followed by a 5' splice site (Figure 2.14) comparable to those found in Drosophilia [Hatton et al., 1998]. These sites are highly conserved across all vertebrate species (Figure 2.13). Interestingly, these 9 genes (see Table 2.1) are also some of the longest in human and these elements appear to originate from some of the longest introns across species (see Figure 2.12 A). From looking at the distribution of novel recursive junctions it is clear that the vast majority occur in long genes (Figure 2.12 B).

Alternative 5' splice sites were identified downstream of the recursive site indicating the potential for inclusion of an alternate exon (hereafter: recursive exon, Figure 2.15). The alternative splice sites also appear to be conserved (Figure 2.15 c). These exons contained a stop codon in almost every frame, indicating that these are likely poison exons which would cause transcript degradation through the nonsense-mediated

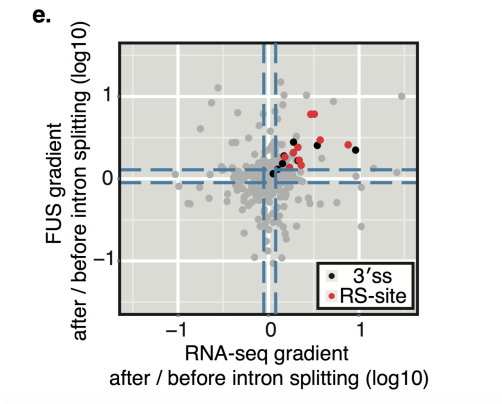

Figure 2.11: Ratio of improvement in gradient before/after adding the recursive site to modelling of the co-transcriptional sawtooth pattern. Red and black dots show junctions that significantly improve the regression gradient and goodness of fit. Grey dots show no significant change. Red dots contact RS-sites and black dots contact sequence of 3' splice sites. [Sibley et al., 2015]

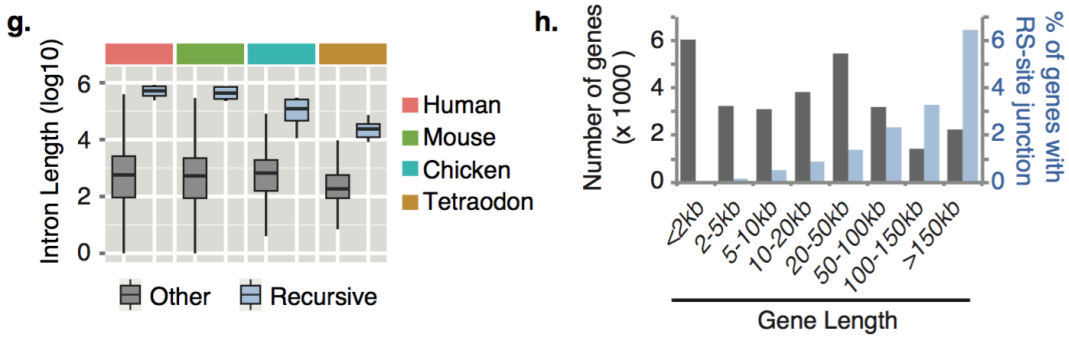

Figure 2.12: (A) Intronic lengths of RSS introns compared to all introns across different species. (B) Histogram of gene lengths (grey bars) with percentage of genes with RSS containing novel junctions (blue bars). [Sibley et al., 2015]

decay pathway. This also reinforces previous studies that found exon recognition is essential for splicing to occur [Ameur et al., 2011]. The recursive exon appears to be a requirement for correct identification of the splice site, as experimentally proved in CADM1 [Sibley et al., 2015]. Interestingly, as compared to Drosophilia, the recursive site does not appear to be necessary for effective splicing of the intron.

The inclusion of a recursive exon, although at very low levels, does occur in all cells. This was investigated by interrogating the upstream gene body of 142 candidate recursive sites (high confidence targets, all cassette exons starting with 5' splice motif GURAG, and novel junctions detected that were consistent with recursive sites but failed to meet significance in linear regression analysis). Several junctions were found between recursive sites and cryptic upstream elements including unannotated promoters and exons. RT-PCR confirmed that an alternative promoter in NTM leads to 100% inclusion of the recursive

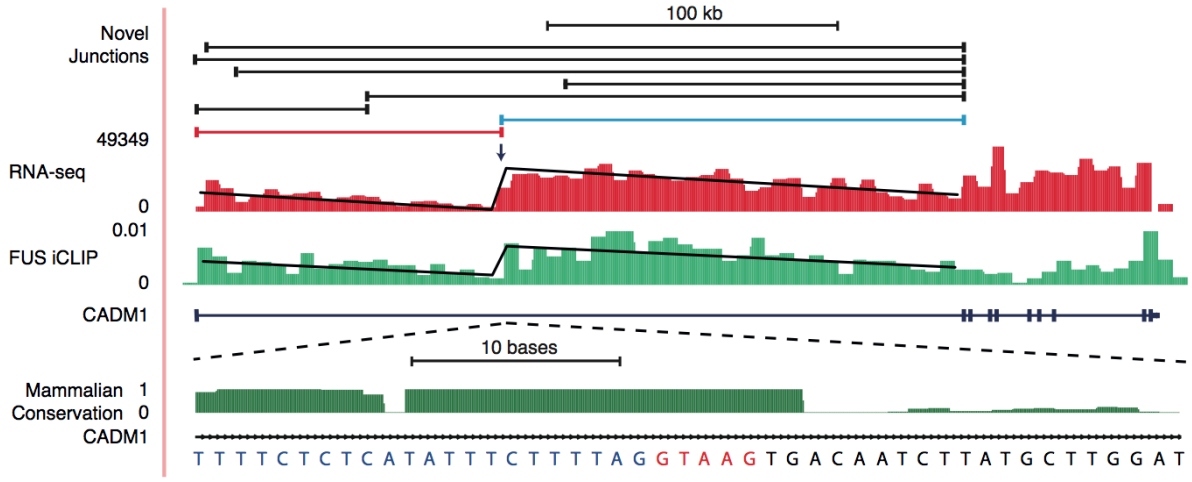

Figure 2.13: (Top) Recursive sites as detected using junction reads (black), the upstream splice junction (red) is abundant while the downstream poison exon junction (blue) is far less prevalent. (Middle) Linear regression of the sawtooth pattern created by binning read coverage across the intron clearly showing splicing to the recursive site. (Bottom) The sequence of the recursive site, showing 3' splice site (blue) head-to-head with a 5' splice site (red). These sites have a high level of species conservation, particularly across mammalian species.

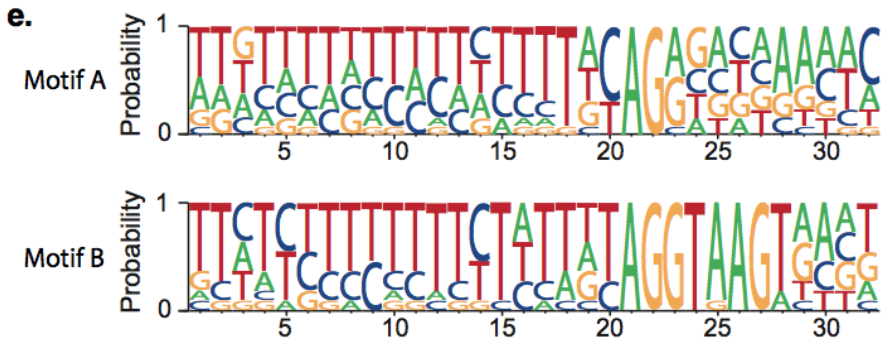

Figure 2.14: Motif of the recursive site showing the polypyramidine tract and 3' splice site followed immediately by a strong consensus 5' splice site.

exon. A similar minor promoter was discovered in CADM2 (Figure 2.17). Figure 2.16 shows the major (P1) and minor (P2) promoters and the associated splice site strength (calculated by MaxEnt [Yeo and Burge, 2004]) for both the reconstituted recursive and alternative RS-exon 5' splice site. Each promoter donates three nucleotides from its upstream exon to reconstitute the RS 5' splice site and this has a large impact on splice site strength. The major promoter reconstitutes a stronger splice site and is preferentially selected by the splicing machinery. However, the minor promoter's reconstituted RSS is weaker than the alternative RS-exon 5' splice site and therefore the RS-exon site is preferentially selected. All MaxEnt scores show a

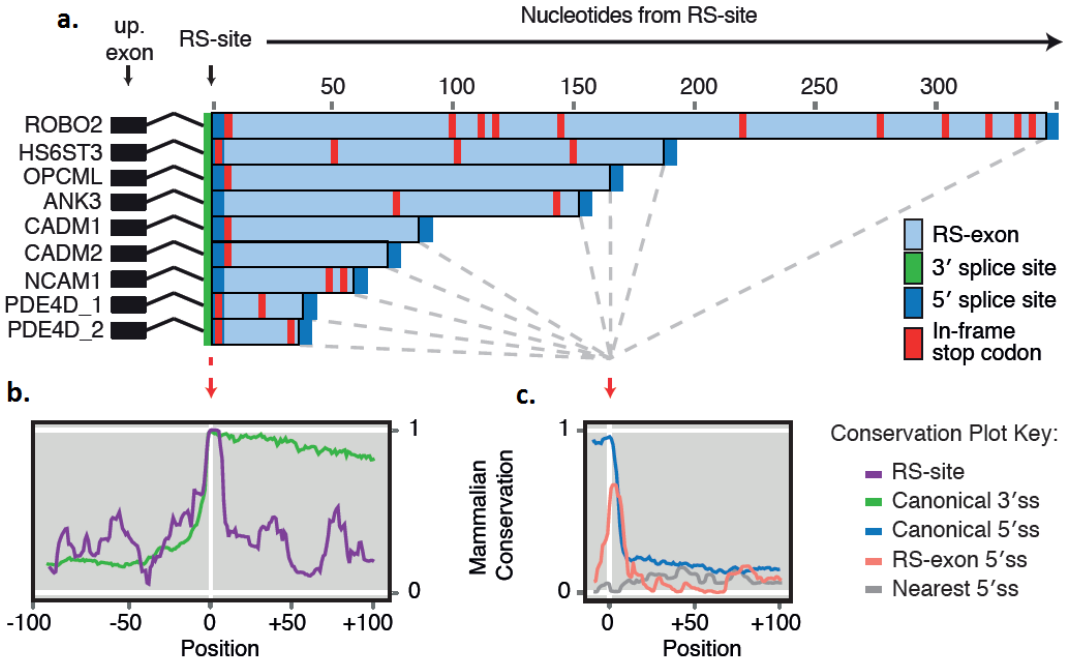

Figure 2.15: (A) Representation of recursive poison exons containing multiple stop codons (red bars) and consensus splice site locations (blue bars). (B) Phylo-P conservation scores aligned at RS-sites and (C) alternate recursive exon 5 splice sites. [Sibley et al., 2015]

similar trend for all RS sites and are present in Table 2.2.

2.3.3 H3k36me3 signal is deficient in long introns

The strong sequence motif of RSS lead to further questions regarding the intronic characteristics of long genes. Such strong conservation may be necessary to distinguish these sites from other background cryptic elements if the environment was not conducive to transcription. I investigated the relationship between intron length and H3k36me3, a well known transcription and splicing-related histone mark. Sequence data from human post-mortem brain tissue and mouse brain tissue were downloaded from ENCODE. H3k36me3 was compared against H3k4me1 (an unrelated enhancer mark) to determine if the effect was specific.

Based on the mouse embryonic brain data (Figure 2.18) a significant decrease in splicing mark H3k36me3 was seen with increase in intron length. This pattern was not seen in the control enhancer mark, H3k4me1. Similarly, exon enrichment for H3k36me3 was inversely proportional to intron length but remained constant in H3k4me1. This systematic decrease in enrichment is clearly visible when looking across the introns grouped in bins according to length (Figure 2.19).

The samples analysed for the human data are shown in Figure 2.20. These data follow the same

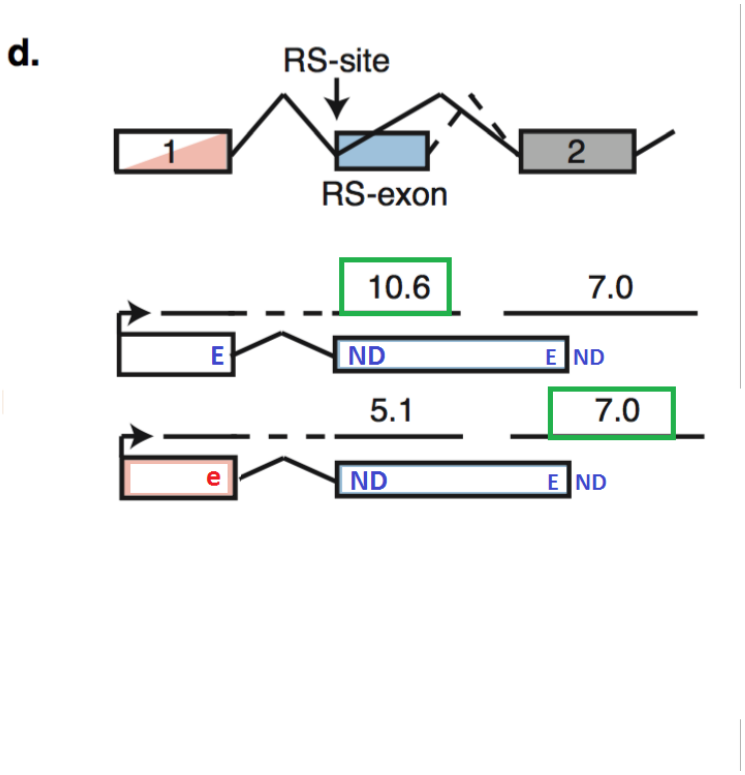

Figure 2.16: MaxEnt scores [Sibley et al., 2015]

trend as dictated by the mouse data but does not have a strongly significant correlation. Reasons for this could include the quality of DNA as human post-mortem tissue tends to be more degraded and hence more variable.

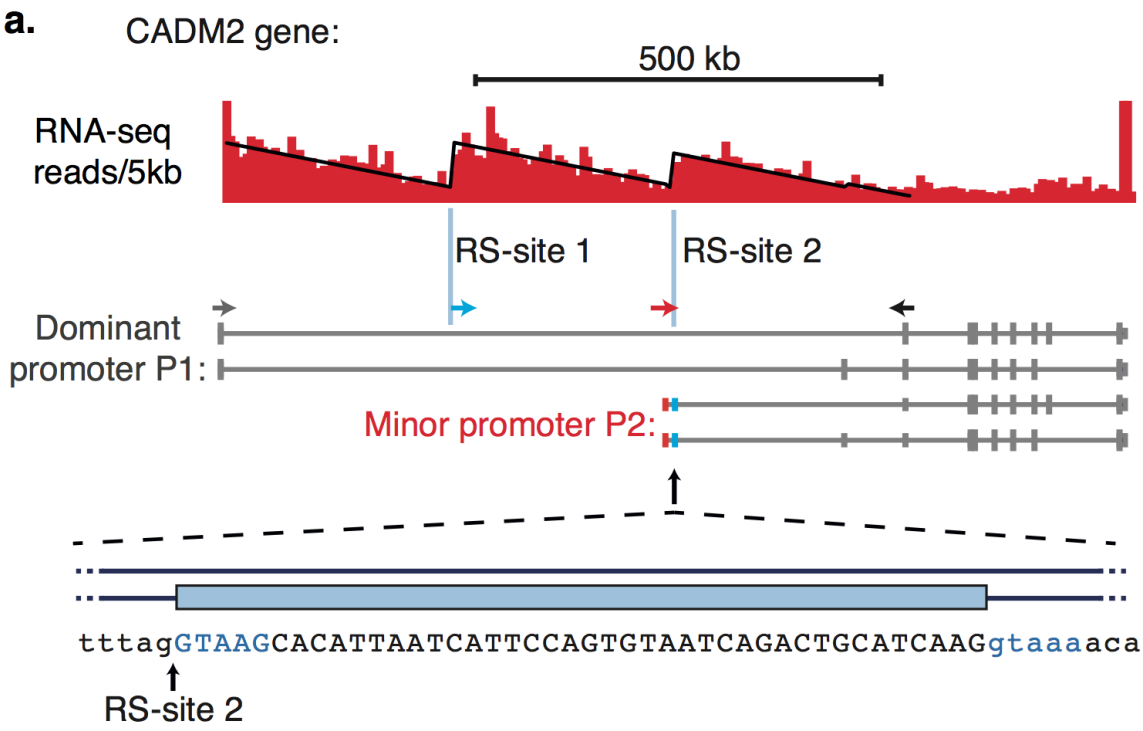

Figure 2.17: Sawtooth co-transcriptional pattern showing the improvements made by linear regression (blue lines). Primers were included in the first RS exon (blue) and second RS exon (red). Zoomed area shows the sequence at the start of the second RS-exon which is also linked to a minor promoter. [Sibley et al., 2015]

Gene name	Strand	Chromosome	RSS location Start	RSS location End	Gene Function	Link to disease
ANK3	-	chr10	62268283	62268483	Required for ion channel localisation	Schizophrenia,mental retardation
CADM1	-	chr11	115270131	115270331	Tumor suppressor	Autism spectrum disorder, cancer
CADM2	+	chr3	85288006	85288206	Expressed in floor-plate	Autism spectrum disorder
CADM2	+	chr3	85560906	85561148		
HS6ST3	+	chr13	97140368	97140568	Modifies heparin sulphate	Obesity
NCAM1	+	chr11	112911547	112911747	Cell to cell adhesion molecule	Bipolar disorder
NTM	+	chr11	131530722	131531042	Cell adhesion molecule	Autism spectrum disorder
OPCML	-	chr11	133166983	133167183	Binds opioid alkaloids	Autism spectrum disorder
PDE4D	-	chr5	58678137	58678337	Phosphodiesterase	Intellectual disability
PDE4D	-	chr5	58981669	58981869		
ROBO2	+	chr3	77389482	77389682	Axonal guidance receptor	Vesico-ureteral reflux

Table 2.1: High confidence recursive sites identified by both junction analysis and co-transcriptional linear regression.

Gene, RS site	Sequence reconstituted 5'ss	Score	RS-exon sequence	Score	RS-site favoured?
PDE4D RS-site 1	TGGGTAAGT	9.23	TGGGTAAGT	9.23	YES
CADM1	CAGGTAAGT	12.75	GCAGTAAGT	7.27	YES
ANK3	AAGGTAAGT	12.19	AGGGTAAGT	10	YES
OPCML	CAGGTAAGT	12.75	GAGGTATGA	7.98	YES
PDE4D RS-site 2	TGGGTAAGT	9.23	GAGGTATGG	7.93	YES
CADM2 RS-site 1	AAGGTGAGT	11.31	TGGGTAAGT	9.23	YES
ROBO2	CATGTAAGT	8.03	ACTGTATGA	3.19	YES
HS6ST3	CAGGTAAGA	11.11	ATAGTATGT	4.33	YES
NCAM1	CAGGTAAGA	11.11	TATGTATGG	1.39	YES
CADM2 RS-site 2	AAGGTAAGC	10.62	AAGGTAAAA	7	YES
NTM	AAGGTAAGT	12.19	CAGGTAGGT	10.73	YES

Table 2.2: MaxEnt splice scores for both RSS reconstituted 5'ss and the RS-exon alternative 5' ss.

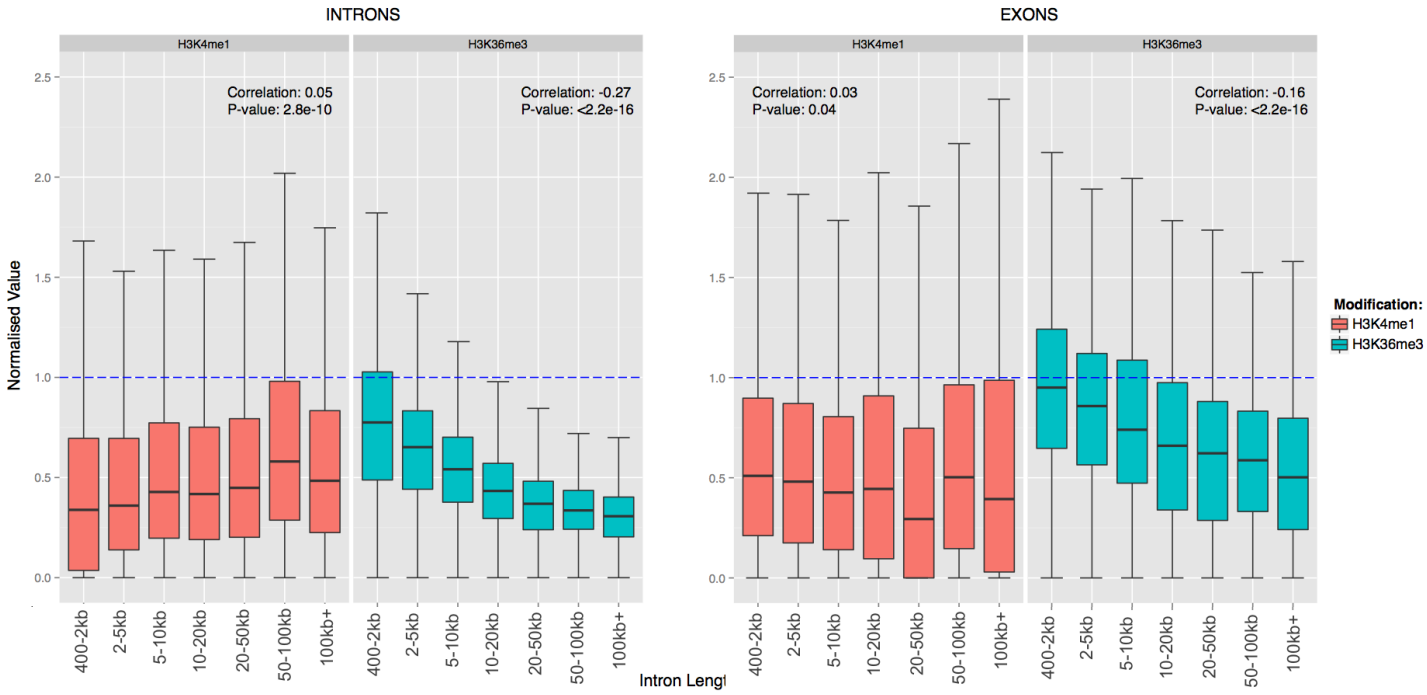

Figure 2.18: Mouse embryonic brain: Relationship of intron length to histone marks H3k36me3 (splicing, repair and active transcription) and H3k4me1 (enhancer mark). Introns are binned and normalised by length.

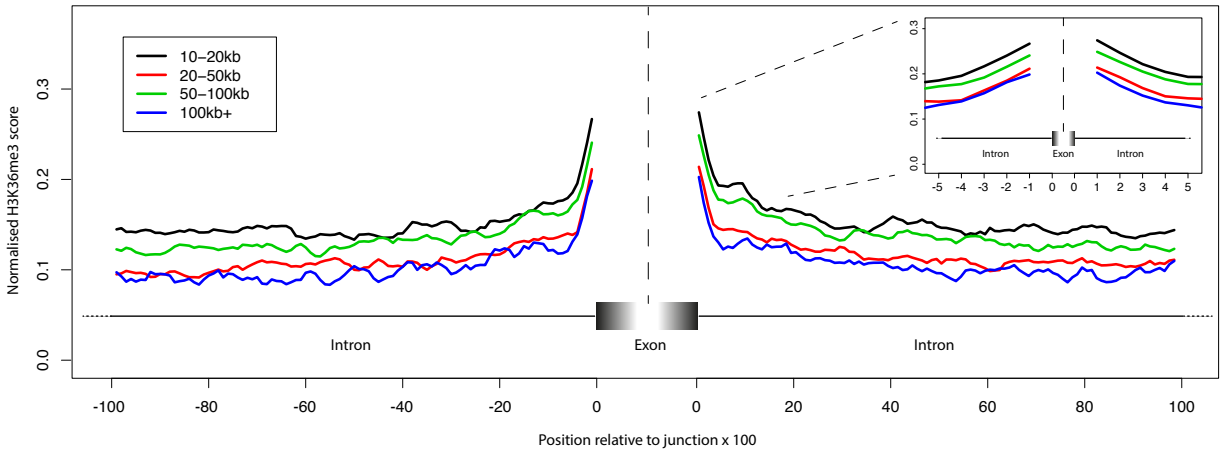

Figure 2.19: Mouse embryonic brain: Relationship of intron length to histone marks H3k36me3 (splicing, repair and active transcription). Introns are binned and read counts are normalised by length.

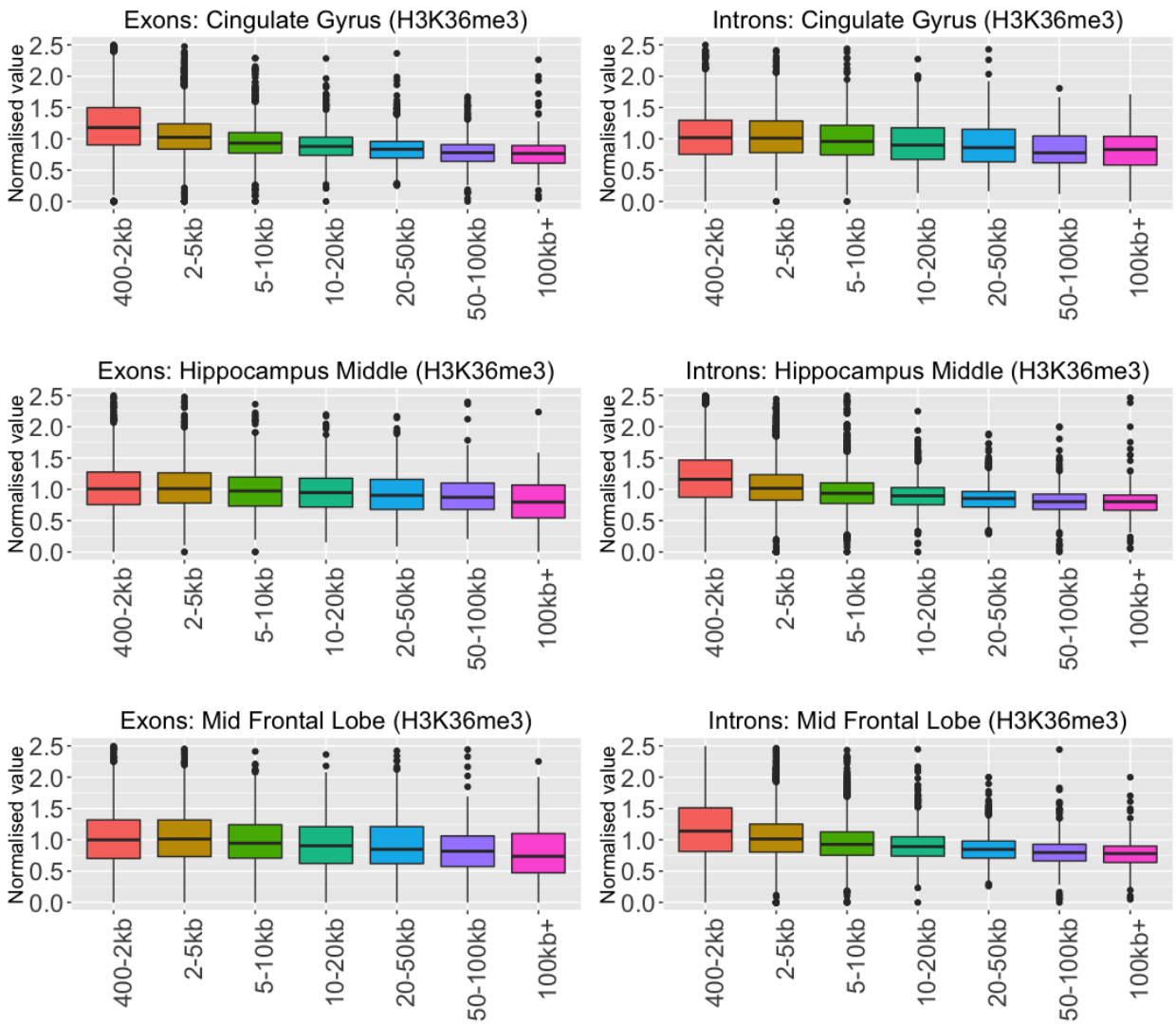

Figure 2.20: Adult human brain: Relationship of intron length to histone mark H3k36me3 (splicing, repair and active transcription) across multiple brain regions. Enrichment of histone marks in exons and introns shown on the left and right respectively.

2.4 Discussion

Long genes have previously been linked to neurological disorders [Lagier-Tourenne et al., 2012; Polymenidou et al., 2011; King et al., 2013]. Here I strengthen that connection by proving that long genes are enriched in brain in multiple datasets. Differentiation appears to have an effect on gene length, showing a slight trend and enrichment in long genes over differentiation. This could be due to the length of time and energy required for transcribing these genes. In actively dividing cells expression of long genes may be impractical and unlikely to complete in a timely fashion.

Here I document the first study to identify recursive splicing in vertebrates. Splicing proteins process these sites in two steps; the recursive exon is detected and the first half of the intron is removed. The recursive 5' splice site is then recognised and the remainder of the intron is removed without inclusion of any exonic nucleotides. A custom pipeline was created to analyse post mortem brain data and 11 high confidence sites were discovered. This pipeline utilised both junction reads and co-transcriptional splicing patterns to effectively characterise these splicing reactions.

Recursive sites are highly conserved and appear to prevent the use of cryptic upstream elements not consistent with the main isoform (Figure 2.21). Although fewer sites are reported than in Drosophilia, the mechanism is also somewhat different, rather than being used to process long introns, recursive sites in human are involved in promoter control. Exon definition is a key element in splicing, even when processing recursive sites an alternate 5' end is required for splicing recognition. This exon contains multiple stop codons and thus its inclusion in a transcript results in its degradation via nonsense mediated decay (NMD).

The inclusion of the recursive exon depends on the strength of the RSS which is in direct competition with the alternate, recursive exon 5' splice site. RSS strength is largely determined by the upstream exon as this provides three crucial nucleotides of the core splice site motif. This implies that inclusion of a non-canonical upstream exon could lead to a weak RSS which would be out-competed by the alternate 5' splice site of the recursive exon. This could be predicted computationally through the splice site scoring program MaxEnt. This opens exciting possibilities to explore splicing competition as a mechanism of transcriptional control.

It is also interesting to note that RSS genes are some of the longest in the human genome. It is striking that the appearance of Alu type SINE repeat elements have provided a way for higher eukaryotes to efficiently process long introns without the help of RSS, which is required in Drosophilia. This may indicate that from this point in evolution RSS were no longer required for intron processing and could differentiate

into other roles. Alternatively, it is also possible that RSS sites have evolved from Alu elements that evaded silencing RNA binding proteins.

Further work will include investigation of shorter introns for potential recursive behaviour and exploration of other non-canonical splicing mechanisms that may operate in long introns. Another avenue to explore is the evolution of these elements. It is clear that more work can be done on the relationship between repeats in long introns and the creation and function of recursive elements. The synergy in this relationship is likely to yield fascinating insights into cellular evolution.

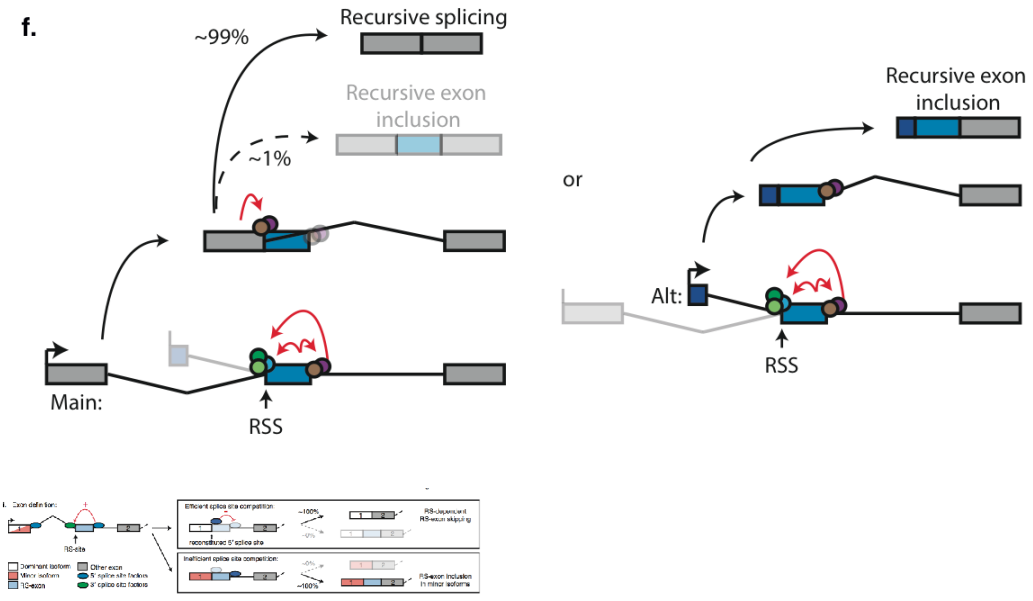

Figure 2.21: (A.) Model for inclusion of recursive exon dependant on promoter usage. (B.) Schematic showing the mechanism of action for recursive splicing resulting in inclusion of the poison recursive exon with use of the minor isoform while it is excluded in the major isoform.

An investigation into histone characteristics of long genes shows a deficiency of H3k36me3 in long introns. This potentially plays a significant role in preventing aberrant transcription of sites within the intron. It also explains why H3k36me3 deacetylation is essential after transcription to prevent interference from within the intron [Pokholok et al., 2005].

This also points to the need for an extremely strong splice signal for the recursive site to be effective. Exploring the relationship between genes with long introns and the constantly growing histone modification

landscape could yield insights into the relationship between processing and chromatin structure. Specifically an investigation into sub groups of long introns that have very few repetitive elements or enrichment of enhancer/silencing marks might pave the way to finding other functionally related elements.

Chapter 3

Characterising circular RNA in the human brain

3.1 Introduction

History of RNA circles

Circular RNA (circRNA) are a recent addition to the growing ranks of non-coding RNA. The circularization of exons within a gene was originally discovered in plants, encoding subviral agents [Sanger et al., 1976] and later in the sex-determining SRY gene as a result of unusual genomic structure [Capel et al., 1993]. CircRNA were also identified at the Fmn locus as creating an inert transcript thereby reducing the expression level of the formin protein [Chao et al., 1998].

Circular RNA are prevalent in all forms of life

Recent publications provided computational and experimental evidence that circRNAs are pervasively expressed throughout the tree of life [Danan et al., 2012; Salzman et al., 2012; Wang et al., 2014]. From these findings a subclass of circRNA was identified as microRNA sponges, such as CDR1as. CDR1as acts a super sponge for mir-7, implicated as an important microRNA in Alzheimers disease. Absence of this circRNA in zebrafish caused a 70% reduction in mid brain size with complete loss in 5% of animals [Memczak et al., 2013]. Similarly, the SRY circle was shown to be a mir-sponge for mir-138 [Hansen et al., 2013].

The functions for the majority of circRNA remain unknown. The majority do not exhibit traits or

capacity to act as miRNA sponges and possess low RNA binding protein density when compared to 3'/5' UTRs. Several studies have been unable to find evidence of translation of circRNA [Guo et al., 2014; You et al., 2015; Memczak et al., 2013; Wang et al., 2014] although a recent publication indicates that a rolling circle amplification mechanism (RCA) can in fact translate circRNA in eukaryotic cells [Abe et al., 2015]

An alternate class of circular RNA, circular intronic RNA (ciRNA), form by circularization of introns. ciRNA impact expression of their parent genes by effecting elongation of the Pol II complex [Zhang et al., 2013]. Knock-down of ci-ankrd52 slightly increased intron retention and a knock-down of downstream splicing events. ciRNA appear to be localized to the nucleus (rather than circRNA which is largely cytoplasmic) and some appear to have alternate roles (other than regulating their parent gene) as they aggregate at different locations in the nucleus. The authors argue that ciRNA may bind RNA binding proteins in a similar way that (after depletion of debranching enzymes) intronic lariats in the cytoplasm sequester TDP-43 thereby suppressing TDP-43 toxicity in ALS disease model [Armakola et al., 2012; Zhang et al., 2013]

Flanking intronic sequences are repeat based or bind splicing factors

Flanking intronic sequences play a key role in circularization. This can be achieved through two known mechanisms; reverse complementary repeat sequences or the binding of splice factors. The splicing factor muscleblind (MBL) has been shown to bind to neighbouring introns drastically increasing the production rate of its circRNA. The proposed mechanism suggests than when MBL protein is in excess, it binds to neighbouring introns increasing the circular isoform and decreasing mRNA production. [Ashwal-Fluss et al., 2014]

Reverse complementary repeats in neighbouring introns allow for pre-mrna folding, creating the loop required for backsplicing (Figure 3.1). Alternative formation of inverted repeated Alu pairs (IRAlus) and competition between them can lead to alternative circularization, meaning several different circRNA can be formed from the same gene (Figure 3.2). On average 3 Alu elements were present in both up- and downstream introns, indicating even partially complementary Alus are enough to promote RNA pairing. Any complementary sequences in flanking introns can promote circularization, an example of this is the SRY gene [Wang et al., 2014; Zhang et al., 2014]

Removal of these sequences dramatically decreases circularization efficiency. Intron length on its own is not a reliable predictor of circularization. The complementary flanking sequences are not conserved between human and mouse and indicate the ability for rapid evolutionary change. [Zhang et al., 2014]

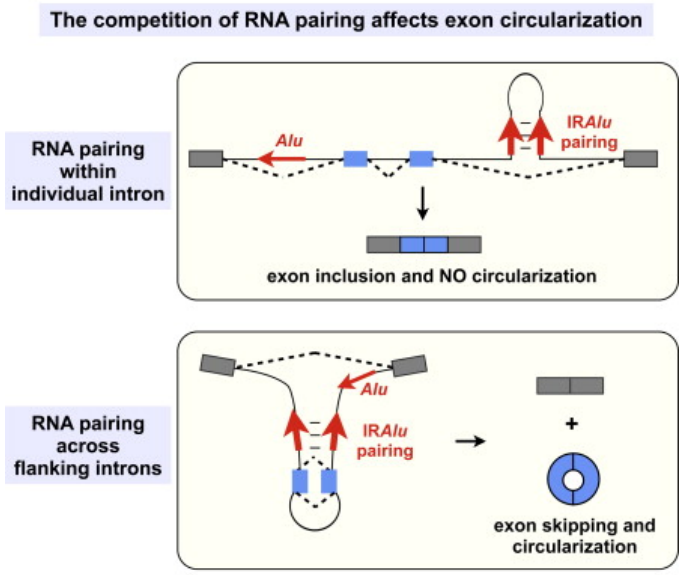

Figure 3.1: Diagrams outlining the proposed mechanism of circularization with inverted repeated Alu pairs (IRAlus) [Zhang et al., 2014].

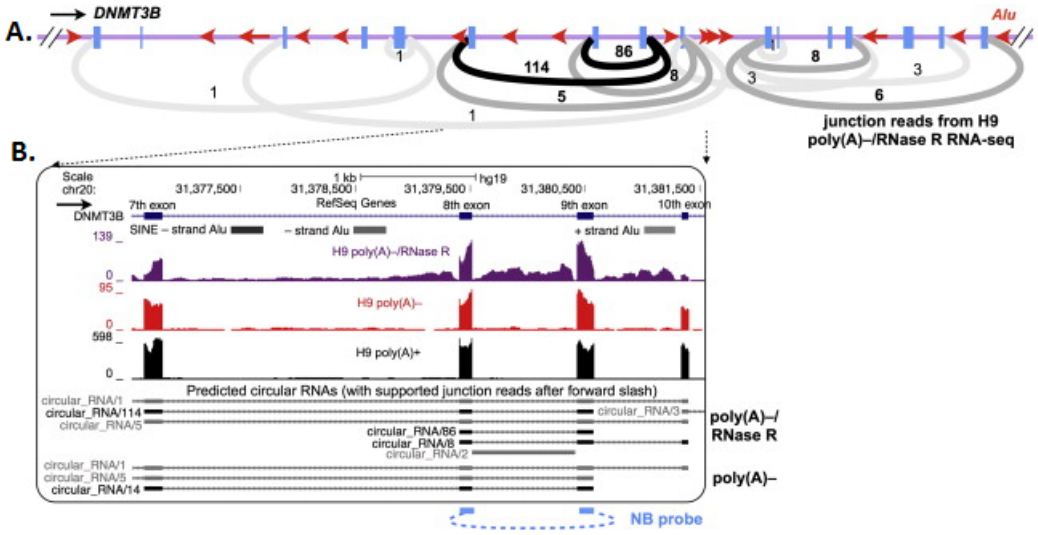

Figure 3.2: (A) Due to multiple Alu elements, there are several conformations pre-mrna can fold into indicating multiple circRNA can be formed from the same gene. (B) Three different tracks using different RNA-seq protocols namely, PolyA+ (inclusion of only mRNA with polyA tails), PolyA- (inclusion of all RNA while depleting rRNA) and PolyA- RNase R (polyA- with digestion of all linear RNA with RNase R). This shows the presence of several circRNA only when depleting linear RNA with RNase R. [Zhang et al., 2014].

circRNA are enriched in mammalian brain and neurological development

Several recent studies have shown that circRNAs are significantly enriched in mammalian brain but even more so in synaptic genes and synaptoneurosomes [Rybak-Wolf et al., 2015; You et al., 2015; Venø et al., 2015; Ashwal-Fluss et al., 2014]. These circRNAs are often well conserved between human, mouse and occasionally Drosophila and are regulated during neuronal differentiation and development [Rybak-Wolf et al., 2015; Venø et al., 2015]. Interestingly, significant expression differences are often observed between the linear and circular isoforms of neural genes, furthermore, the localization of circRNA products (and not their linear counterparts) tend to be higher at the synapse than in the cytoplasm [Rybak-Wolf et al., 2015; You et al., 2015]. This leads to the conclusion that some circRNA may function independently of their linear siblings.

Brain related RNA binding proteins such as TDP-43, FUS and muscleblind have already been implicated in neurodegenerative disease and further work is essential to elucidate mechanisms and effects on circRNA and their relationship to pathology [Lagier-Tourenne et al., 2012; Polymenidou et al., 2011].

circRNA diversity in brain is estimated at 3 circRNA per gene. However, over 2,000 genes show 10 or more isoforms [Rybak-Wolf et al., 2015; Venø et al., 2015]. Their complexity is further increased as there is evidence of differential inclusion of internal exons [You et al., 2015]. There is also evidence that circRNA differential expression is linked to neural plasticity [You et al., 2015; Venø et al., 2015]. This research provides tantalizing clues to the complex cellular processes regulating creation of circRNA and their potential importance in neuronal function.

Potential use of circRNA as biomarkers and their relevance to cancer genomics

circRNAs have been quantified at detectable levels in saliva, blood and within exosomes. Interestingly, 60% of the 327 circles identified in saliva are non-canonical [Bahn et al., 2015]. The enrichment of circRNA in blood is comparable to brain, this provides a unique opportunity to explore their roles as biomarkers [Bahn et al., 2015]. Some circRNA found in blood appear to be more highly expressed than their linear isoforms and may provide a proxy for quantification of expression. [Memczak et al., 2015]

The discovery of exosomes, small membrane vesicles secreted by cells, has provided a unique opportunity to identify biomarkers for disease. Over a 1,000 circRNAs have been identified in human serum exosomes and are enriched in exosomes compared to host cells indicating an active regulation of circRNA transport.[Bahn et al., 2015]

In cancer serum differential regulation of circles is clearly present with 67 missing species and 250

novel cancer-specific circRNA being detected. [Li et al., 2015] A recent publication has shown that novel circRNA can be produced from gene fusions caused by chromosomal rearrangements (see Figure 3.3). These circRNA can contribute to cellular transformation, promote cell viability and confer resistance to therapeutics. They have been shown to have tumor-promoting properties in *in vivo* testing. [Guarnerio et al., 2016]

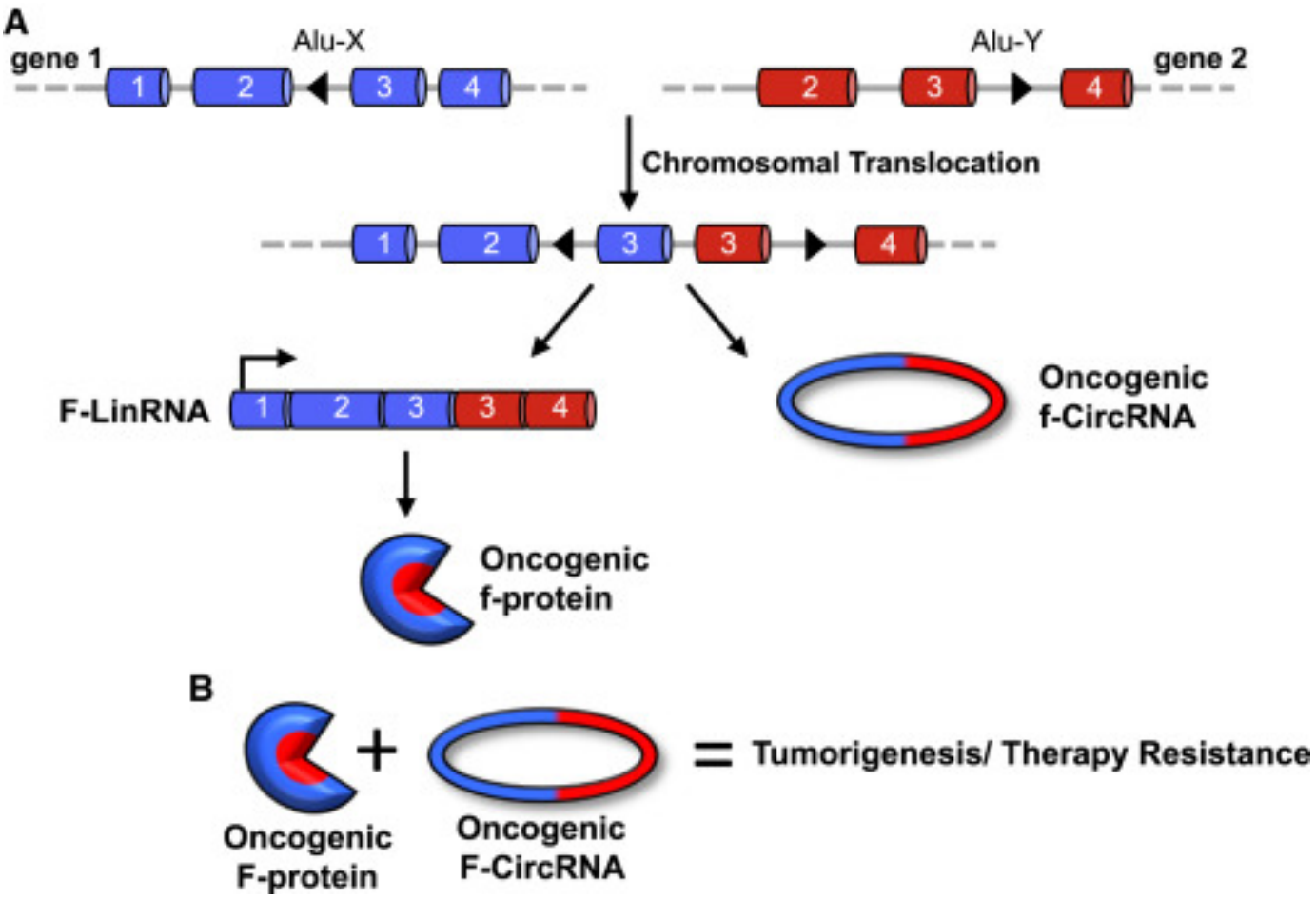

Figure 3.3: (A) Chromosomal translocation in cancer produces a gene fusion which results in novel conformations of complementary Alu elements. These elements promote the circularisation of exons within the fusion gene. (B) The oncogenic fusion proteins (both linear and circular RNA) promote tumourigenesis and resistance to therapeutics. [Guarnerio et al., 2016]

3.1.1 Challenges in identifying Circular RNAs

Detection of circRNA relies on the backsplice junction

CircRNA have remained largely unexplored until now due to difficulties with detection as the only distinguishing feature from the linear transcript is the unexpected backsplice junction. This has lead to several strategies in an attempt to identify and quantify circRNA. Initial approaches focused on validation using custom sequencing kits, mostly exploiting RNase to deplete linear RNA [Jeck et al., 2013; Zhang et al., 2014]. However, total RNA-seq provides a unique opportunity to explore circRNA without requiring specific sample preparation.

Two approaches are currently in use; the first splits unmapped reads into smaller fragments, aligns them to the genome and looks for inverted mapping of read fragments, indicating a potential backsplice site. This approach can successfully identify novel, non-canonical, circRNA but requires reads to overlap the backsplice in a more or less symmetrical way [Memczak et al., 2013; Hoffmann et al., 2014]. This greatly reduces its sensitivity. The majority of new tools embrace this strategy while focusing on different attributes of circRNA [Zhang et al., 2014; Gao et al., 2015].

Recently, an aligner was created using this strategy [Hoffmann et al., 2014], however the resource consumption and time usage on a large dataset (50 high depth RNA-seq samples) is intractable. Average runtime per sample is 15 hours with a memory footprint of 20%. For over a terabyte of data this translates to a minimum of 200GB of memory. This software appears to be most effective when analysing data produced by a circRNA enrichment protocol (such as the use of RNase R to degrade linear mRNA) with far lower depth.

An alternative approach is to create a database of all possible backsplices from annotated exons. Aligning unmapped reads back to this database can accurately quantify all circles. Although this method is more sensitive, it is heavily annotation dependent and cannot identify non-canonical circles. [Salzman et al., 2012; Wang et al., 2014]. All circular RNAs discovered thus far have been uploaded and merged into a public repository called Circbase [Memczak et al., 2013].

3.2 Methods

In addition to the software and tools described in section 2.1.1, this section lists the methods that have been specifically used for this chapter.

3.2.1 Accurate quantification of Circular RNA

Effective identification of circular RNA relies on detection of the backsplice location. Careful quality control is essential to accurately quantify circles. The outline of the bioinformatics pipeline developed for this purpose is shown in Figure 3.4.

Identifying all circRNA in human brain using total RNA sequencing data

All brain samples were mapped to the human genome (build hg19) using the STAR aligner (v2.3 [Dobin et al., 2013]). All unmapped reads were pooled and realigned using the strategy outlined in [Memczak et al., 2013], all reads were divided into smaller seed fragments, aligned to the genome and all inverted fragments were taken as proof of circularization (see Figure 3.4 A). These were then enumerated and filtered to produce a list of potential circular rna. All identified circRNA were merged with all known circRNAs in the Circbase repository.

Creation of a backsplice database and enumeration of circRNA

A backsplice database was created from the discovered circRNA (see Figure 3.4 B); the 5' and 3' ends of each circRNA are joined to create an artificial reference with a total length of 150 nucleotides. This required a 100bp read to overlap with at least 15 nucleotides. Sequence reads from each brain sample were then aligned against the circRNA database and human genome simultaneously using Bowtie2 [Langmead and Salzberg, 2012].

3.2.2 Pitfalls to identification of Circular RNA

One concern when investigating circRNA is to be aware of situations in which reads align to scaffold junctions that do not originate from circular molecules. Although a strategy for paired-end sequencing will be discussed in the following section it is also important to investigate the spurious alignments than can occur across the read scaffold. Figure 3.5 shows various instances were reads can be misaligned to scaffolds.

The first clear concern is mismatches across the read indicating mismapping. Collections of mismatches on a single end, centred around the central splice location or randomly distributed were common (Figure 3.5 A,B,C). These alignments often were primary and may have resulted from similar pseudo genes, other non-canonical splicing or PCR artefacts.

A second consideration is imbalance in the overhang lengths. The scaffolds were created to minimize this as each end was 15bp shorter than a read fragment. This however, did not circumvent minimum overhang

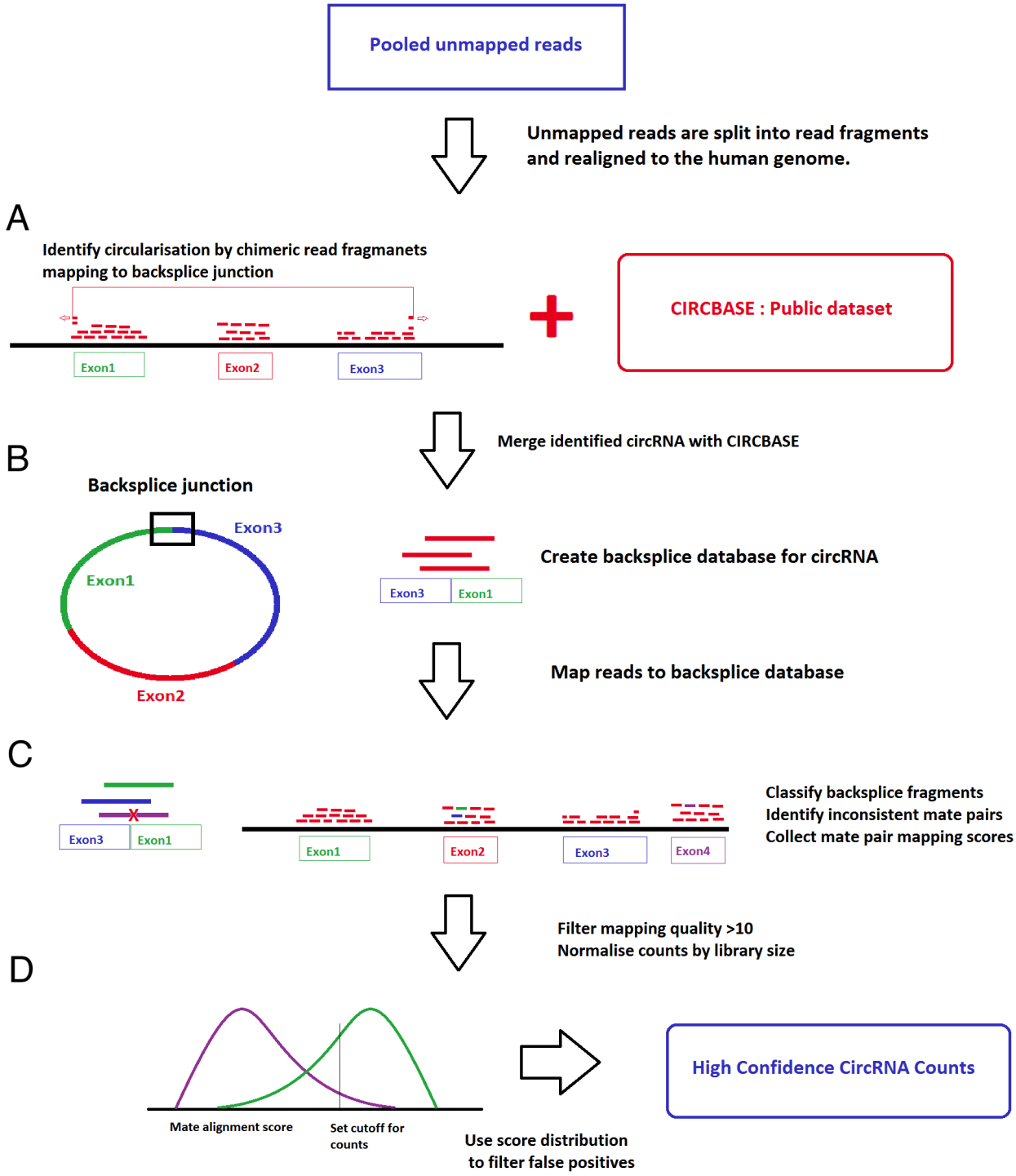

Figure 3.4: Outline of bioinformatics pipeline for processing raw sequence reads to produce high confidence count data for both novel and known circRNA. Briefly, all samples were aligned using STAR, (A) unmapped reads were pooled and initial circRNA discovery was done based on the algorithm from Memzack et. al [Memczak et al., 2013]. (B) All identified circRNA were merged with known circles from Circbase and appropriate backsplice scaffolds were generated for each circRNA. All samples were realigned to the human genome (GRCh37) and backsplice scaffolds. (C) Raw alignment results were filtered using read scaffold alignment. Paired-end information was used to determine which fragments originated from a circular molecule (green, blue read) and which did not (purple read). (D) This information was used to determine differences in distributions of alignment scores for true positives (green line) and false positives (purple line). A threshold could then be assigned to filter reads of interest, this produced a final list of high confidence counts.

reads, with multiple mismatches on the overhang (Figure 3.5 D). More common than this was the aligner labelling the last bases as low quality (a known bias in Illumina reads is their quality drop off at the 3' terminus) thereby soft-clipping these bases without directly affecting alignment score.

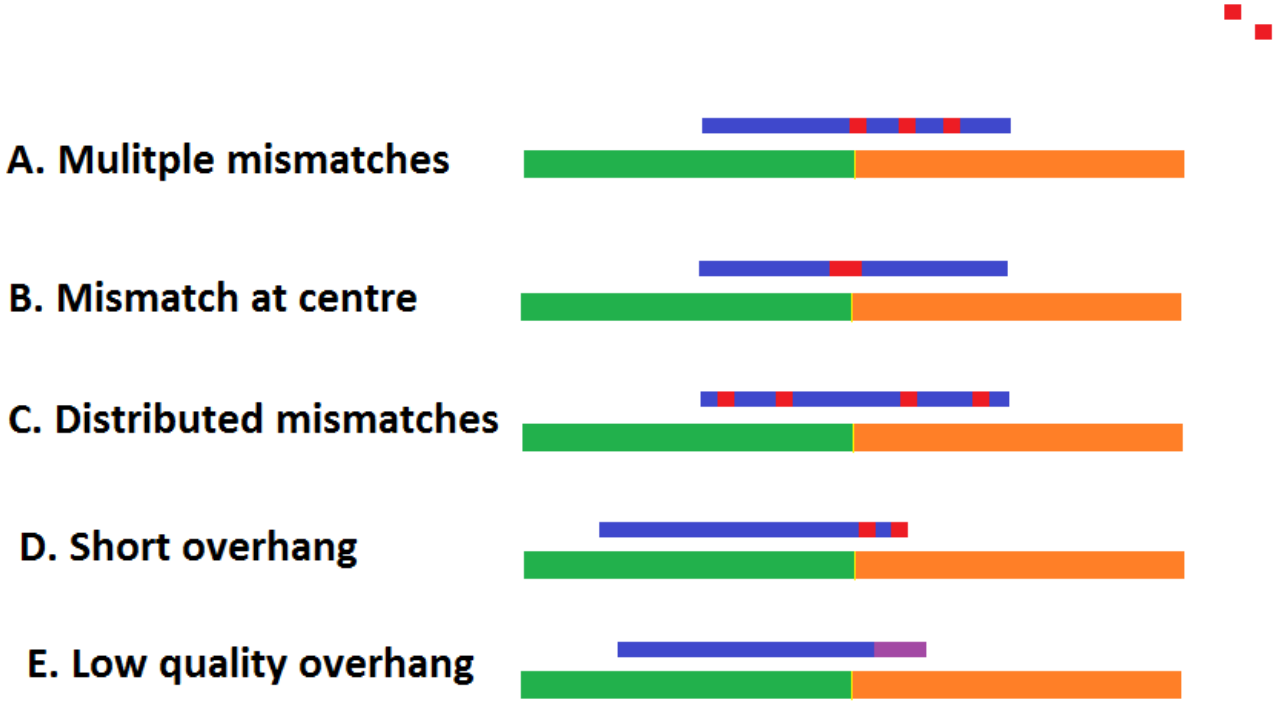

Figure 3.5: Graphical display of several common misalignments that occur when mapping reads to scaffold backsplice junctions. Separate exon ends are shown in orange and green, mismatches are shown in red, low quality bases shown in purple.

It is clear strict filters need to be applied to minimise the occurrence of these false positive alignments. Several filters were implemented to remedy these effects and are described below.

Quality control of backsplice hits, filtering and library normalisation

All samples were pooled to maximize power for the remaining experiments. Each read pair was evaluated based on the alignment score of the read mapping to the backsplice database (hereafter: backsplice/scaffold read) and the location of its mate which aligned to the human genome (hereafter: genome read) (see Figure 3.4 C).

All genome reads falling outside the predicted bounds of the backsplice read (denoting the outer boundary of the circRNA) are considered false positives. The backsplice reads alignment score is recorded and the distribution of true positives and false positives is calculated (Figure 3.4 D). These distributions

allowed determination an appropriate alignment score cut off of -15 (Figure 3.6).

Lastly, the mapping quality (MAPQ) for each backsplice read was required to be higher than 20. Mapping quality grades the uniqueness of the read i.e. the confidence the read aligner has that it has mapped the read to the correct location.

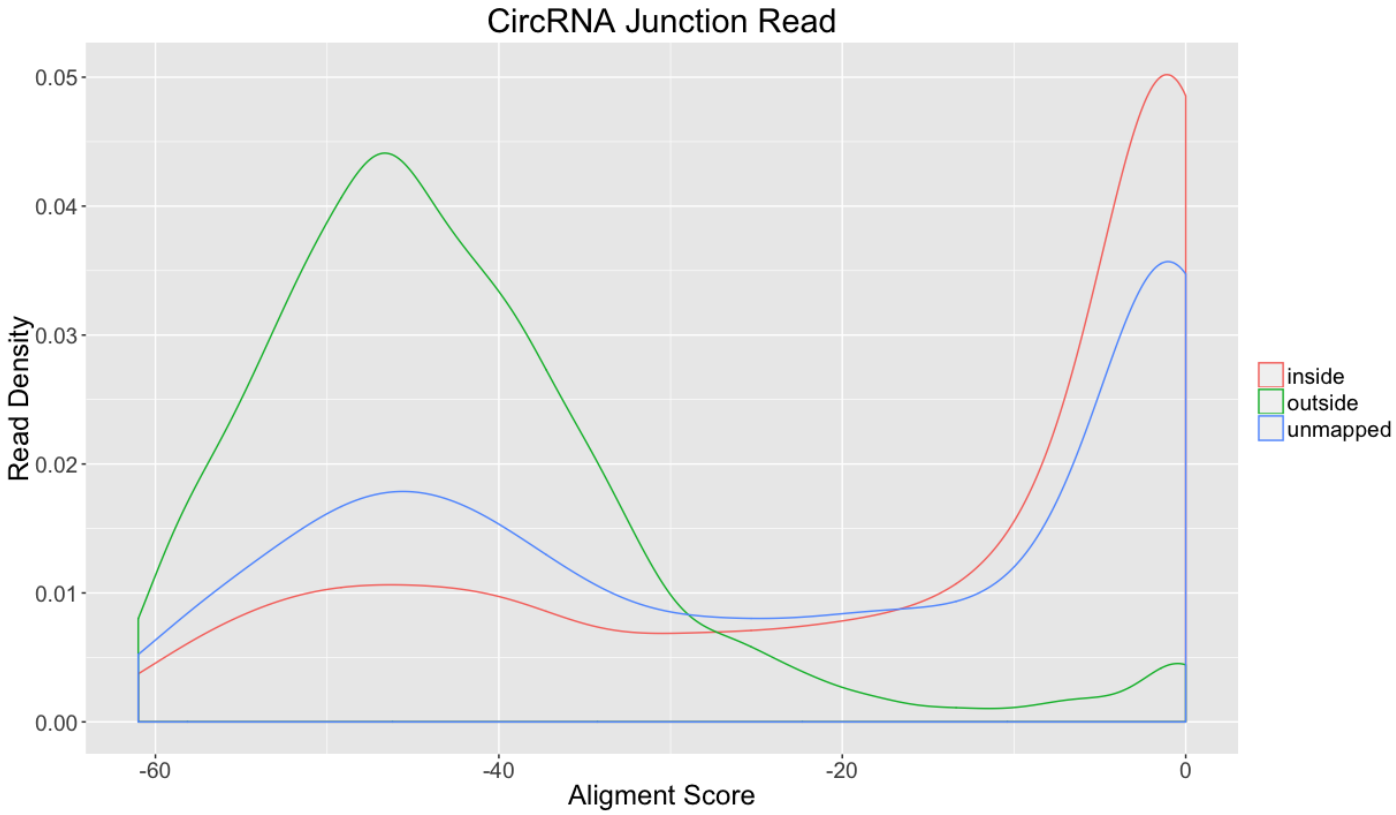

Figure 3.6: Distribution of alignment scores for backsplice reads. Separate categories were created for genome reads located within the backsplice read bounds ("inside", red), false positives were those located outside the backsplice boundary ("outside",green) and those with unmapped mate pairs ("unmapped",blue). A threshold was set to a minimum alignment score of -15 (scale from -60 to 0) to minimise the inclusion of false positive results.

Finally, the raw counts for all circRNAs are library normalised using the bioconductor DESeq package [Anders and Huber, 2010]. A further step is taken to normalise these circRNA by comparing junction counts of backsplices to brain housekeeping gene GAPDH. High confidence circRNA are then annotated accordingly based on overlapping genes, pseudogenes, including a list of recently identified constrained genes intolerant to non-synonymous mutation [Samocha et al., 2014]. This is compiled into a database of circles, hereafter: CircBrDB.

Case study: circRNA differential expression in Bipolar disorder

After the creation of the CircBrDB circRNA catalogue in healthy human brain, these backsplices could be searched for in other datasets. One example is a study done by Akula et. al [Akula et al., 2014] on 4 post-mortem samples of dorsolateral prefrontal cortex from 4 bipolar patients and 4 controls. Data was downloaded from the repository (GEO: GSE53239) , aligned to the database of scaffold backsplices and the human genome (GRCh37). These raw alignment results were then filtered as stipulated in section 3.2.2. The resulting backsplice counts were imported into DESeq [Anders and Huber, 2010], library normalised and low abundance counts were removed. The remaining data were run through the differential expression software to determine fold change across control and bipolar brains. Results were ranked by P value.

3.2.3 Pairwise analysis of highly similar gene pairs

A subgroup of the most highly expressed backsplice junctions were found to connect two proximal genes rather than lie within a single gene. These genes were often from the same family, with clear homology and similar structure. An example of this is the tubulin protein family i.e. *TUBA1A* - *TUBA1B* and *TUBB2B* - *TUBB2A*. These backsplices have been detected in other datasets [Rybak-Wolf et al., 2015; Guo et al., 2014] and were disregarded as alignment artefact. In order to explore this relationship further a pipeline was developed to determine if these backsplice junctions and associated trans-splicing junctions were in fact biological in nature.

A flow diagram covering the computational steps of the pipeline is shown in Figure 3.7. This was implemented using Python, Biopython and custom Python libraries.

Compiling annotations

The CircBrDB backsplice junctions were annotated using GENCODE [Harrow et al., 2012] v19 exon and intron annotations. All information was stored in data structures to allow for ease of access.

Parsing aligned data to retrieve valid alignment pairs for further analysis

Previously aligned data from all 48 brain samples generated by Bowtie 2 was sorted by read name and analysed. For each read pair, the read aligning to the backsplice scaffold in CircBrDB database was interrogated to ensure it was of high mapping quality. This was done to correct for the heuristic nature of the Bowtie2 algorithm, which does not penalise certain low quality mismatches towards the end of the read fragment. Manual analysis was required in order to ensure read integrity. For each read pair, the read mapping to the

backsplice scaffold (backsplice read) and its mate (mate read) were investigated. Each read must achieve a mapping quality score (MAPQ) above 25 or complete sequence match to either the genome reference or the backsplice sequence (see Figure 3.7A).

Creating paired gene annotations for proximal gene pairs

A paired annotation is created for all provided gene pairs using GENCODE transcript annotations. This involves identifying reciprocal exons between the two genes (hereafter: exon siblings) using exon position and collecting sequence information from the exons. The backsplice junction is then annotated according to the transcripts provided, indicating exonic/intronic overlap within each transcript exon/intron. For each exon both up and downstream splice junctions are recorded. This allows for evaluation of the canonical transcript junction vs backsplice junction. This process is demonstrated in Figure 3.7B.

Evaluating the backsplice read and determining minimum overhang

Each read mapping to a backsplice junction is split into its constitutive exons. An overhang distribution is then calculated for each exon, this provides a maximum and minimum value showing how far into the exon all backsplice reads extend (Figure 3.7C). Each exon fragment is then locally realigned to its exon sibling starting with an initial 15 nucleotide fragment from the splice site. The length of the fragment is then increased in a step-wise fashion to the maximum overhang. This provides the minimum overhang required for a satisfactory number of nucleotide differences between backsplice and canonical exon to indicate this read overhang was correctly mapped (Figure 3.7D). A threshold of 2 nucleotide changes was required to define the minimum overhang length. A dynamic programming, local alignment algorithm from the Biopython package with the gap opening penalty -1 and gap extension penalty -4 was used. [Cock et al., 2009]

Each overhang is evaluated in this way to determine if it can be anchored on either end of the backsplice (Figure 3.7E). All reads that passed the minimum overhang length on both exons were flagged and counted. A large subset however only identified a single significant overhang, in order to salvage these reads the mate pair was evaluated.

Evaluating the mate read to validate backsplicing

The mate read was used to determine if the read pair supported the backsplice (Figure 3.7F). Similar to the exon fragment analysis, I need to identify if the mate read maps uniquely to the gene that supports the backsplice junction. The mate sequence was aligned to the exon sibling i.e. Gene A exon 2 vs Gene B

exon2. In order to reduce computational cost a heuristic string matching algorithm was used to determine if there was enough difference to warrant pairwise alignment. This method elucidated high quality backsplice alignments that could be enumerated to determine the prevalence of each backsplice/transplicing event.

Evaluating trans-splicing

In much the same way, this pipeline allows for the evaluation of trans-spliced junctions purported to map between proximal, highly similar genes. These junctions have been largely ignored or considered mapping artefact. This is curious as mapping algorithms are technically biased towards mapping reads to known canonical splice junctions and hence a certain amount of evidence is required for the presence of these non-canonical junctions [Dobin et al., 2013]. Therefore, further investigation of these features alongside the backsplice junctions was undertaken.

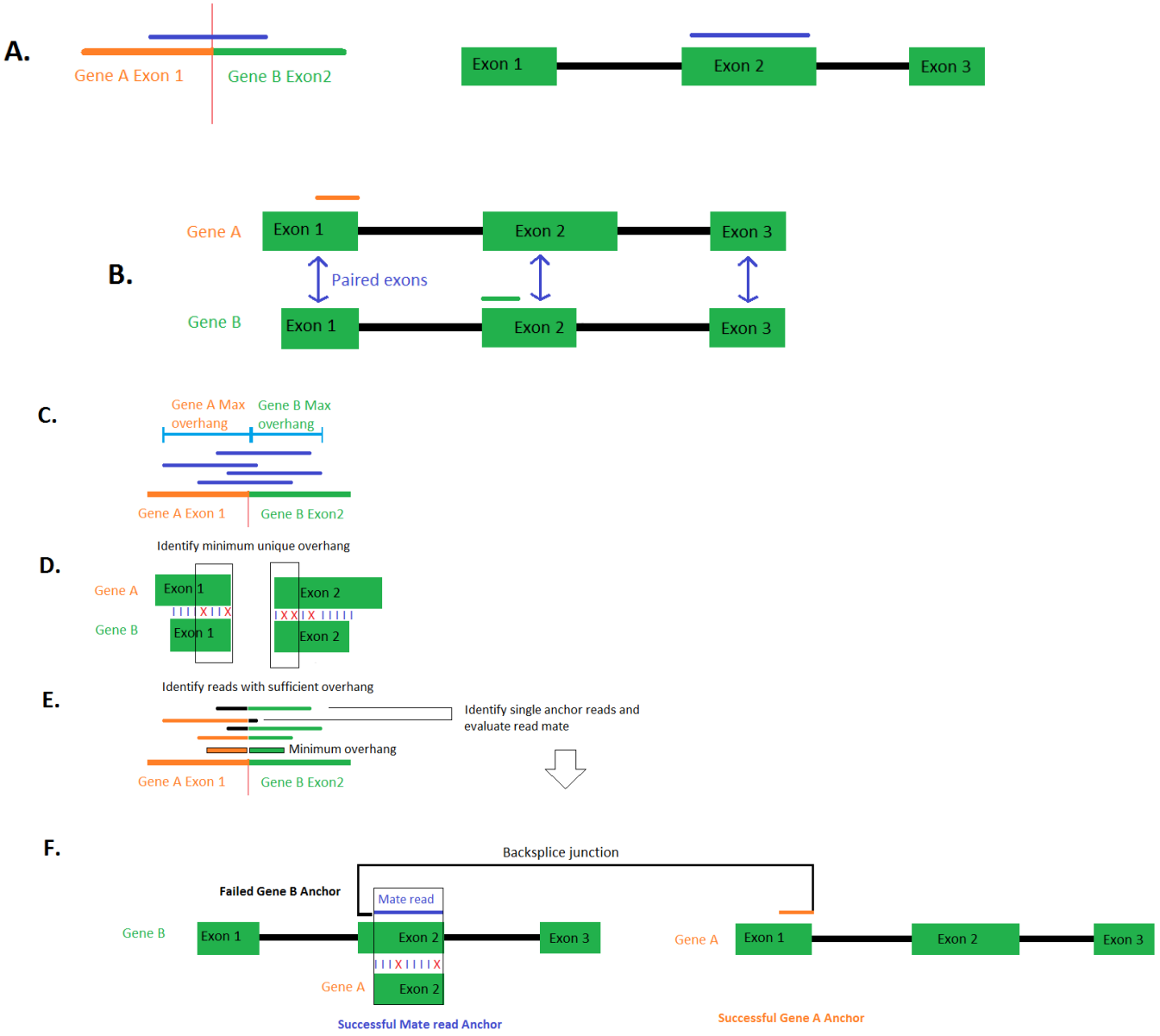

Figure 3.7: Additional pipeline steps to validate backsplices between homologous genes. Annotations are downloaded from GENCODE v19 for each gene. (A) The scaffold read and genomic read are evaluated to ensure they map with satisfactory quality. (B) The genes involved in the backsplice are compared to identify exon pairs with high similarity (hereafter: sibling exons), this provides an exon to compare against the current alignment. (C) Maximum overhang per exon fragment is determined using reads aligned to the scaffold. (D) Exon siblings are pairwise aligned, differences are noted, a minimum overhang length is determined that anchors a read to a specific exon. (E) This minimum overhang can then be used to determine which reads have sufficient evidence to come from a particular exon. In most cases only one side can be uniquely anchored (coloured overhangs) while the other cannot (black overhangs). (F) The mate read can then be pairwise aligned to the exon sibling to ensure if it is sufficiently unique.

3.3 Results

3.3.1 High confidence circRNA

All circRNAs identified in the brain dataset were collected into a database described as "CircBrDB". Table 3.1 outlines the breakdown of processing circRNA, starting with 107,000 circRNA identified during the discovery step of the pipeline, a total of 1,100 circRNA were detected at high levels, above control junctions (standard splice junctions from brain house-keeping genes) see Figure 3.8. I start by investigating the circRNAs with the most read count evidence. The top 35 circRNAs (with average sample expression above 200 high quality backsplice reads per sample) are shown in Table 3.2, annotated by gene, constraint (deficiency of deleterious variation, see Chapter 4), function and disease association. A third of these are novel to CircBrDB while the remaining are also found in Circbase. *CDR1as* is incredibly abundant as it is only expressed in circular form [Memczak et al., 2013].

	circRNA
CircBrDB (Identified in brain)	107,560
All circRNAs in Circbase	92,369
Final merged database (CircBrDB and Circbase)	112,652
Total detected circRNA with >5 high quality reads	103,549
Total above stringent mapping error controls	1,100
High confidence targets (>100 counts)	80
High confidence targets (>200 counts)	35

Table 3.1: Breakdown of circRNA backsplices identified in brain.

3.3.2 Backsplice junctions forming between closely related genes and gene-pseudogenes are abundant in the brain

Several of the top hits in Table 3.2 originate from circRNA formation between two proximal (10-50kb apart) but separate genes. These genes often belong to the same gene family and have identical exon structure. It was also observed that backsplicing occurs between a gene and its proximal pseudogene. Examples include; *CKMT1B-CKMT1A*, *TUBB2A-TUBB2B* and *TUBA1B-TUBA1A*. Interestingly, both tubulin families (*TUBA* and *TUBB*) have a large number of high quality backsplice counts. These backsplice junctions persist in the data even under very strict filters (mapping quality (MAPQ) > 30, alignment score (AS) > -10) designed to correct for homology. The distribution of novel to known circles found between genes is shown in Figure 3.9.

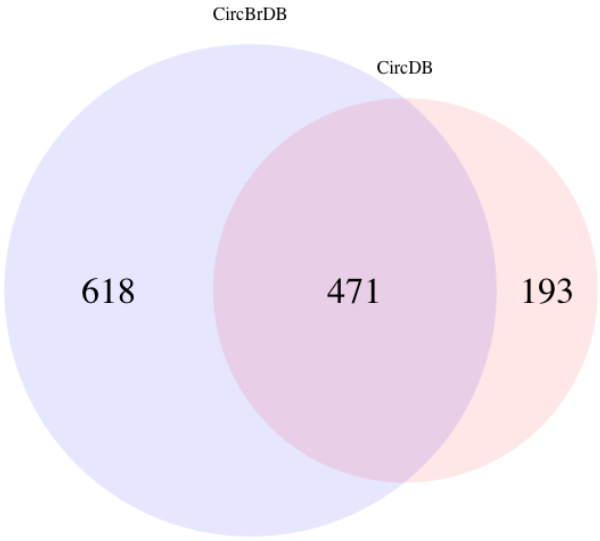

Figure 3.8: Venn diagram outlining the overlap backplice junctions (denoting circRNA) found in CircBase (public repository) and CircBrDB (database used in this study).

Inclusion of gene-pseudogene backsplices in the top 35 indicate these pseudogenes may have a yet unexplored function. Proximal genes (and gene-pseudogene pairs) are always located on the same strand suggesting that transcript read-through could generate one pre-mRNA molecule. A recent publication points to this happening in proximal genes in cancer [Grosso et al., 2015]. Due to the high sequence identity between these genes a method had to be developed to verify these results. To the author's knowledge no tool exists to tackle this specific alignment query.

3.3.3 Pairwise realignment of backsplice junctions

A further processing step was created to determine the validity of backsplice junctions between highly similar genes. Results will be focused on the two most highly expressed backsplice junctions in the Tubulin gene families. Tables 3.3, 3.4 and 3.5 show the backsplice junctions present that pass realignment analysis in the *TUBA* gene pair, *TUBB* gene pair and *TUBB2B* and pseudogene pair respectively. Scaffold reads, that map to the backsplice uniquely, are the strongest evidence followed by mate reads which show that both reads from a mate pair map in a configuration consistent with the backsplice. The majority of backsplice junctions are too similar to distinguish from aligner artefact. However, the *TUBA* gene pair shows expression of several backsplices with scaffold support. The *TUBB* genes show a single strongly supported junction while the high similarity between *TUBB2B* and its pseudogene show inconclusive results.

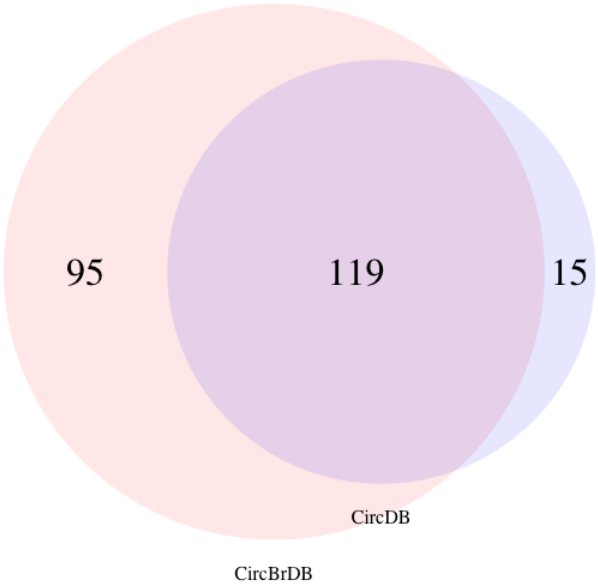

Figure 3.9: A Venn diagram outlining the known occurrences of backsplice junctions between proximal genes in CircBase (public repository) and CircBrDB (database used in this study).

Pairwise realignment of transplicing junctions

Similarly, the results for transplicing realignments are shown in Tables 3.6, 3.7 and 3.8 for *TUBA* gene pair, *TUBB* gene pair and *TUBB2B* - pseudogene pair respectively. These junctions, are more abundant and appear to be more pervasive. Interestingly, very few show scaffold reads indicating the scaffold is not specific enough to anchor a read on both sides. It is noted that *TUBB2B* and pseudogene show scaffold reads in both backsplice and transplicing junctions.

3.3.4 Reciprocal back/transplicing junctions across Tubulin genes

Based on the results from the pairwise realignment, both backsplice and transplicing junctions appear in a reciprocal configuration. Figures 3.10 and 3.11 show the proposed model, whereby the transplice and backsplice products could be created simultaneously from a splicing reaction between the two mRNA transcripts. Figure 3.12 shows the genomic context around these genes with the percent of total transplicing junctions compared to canonical splicing.

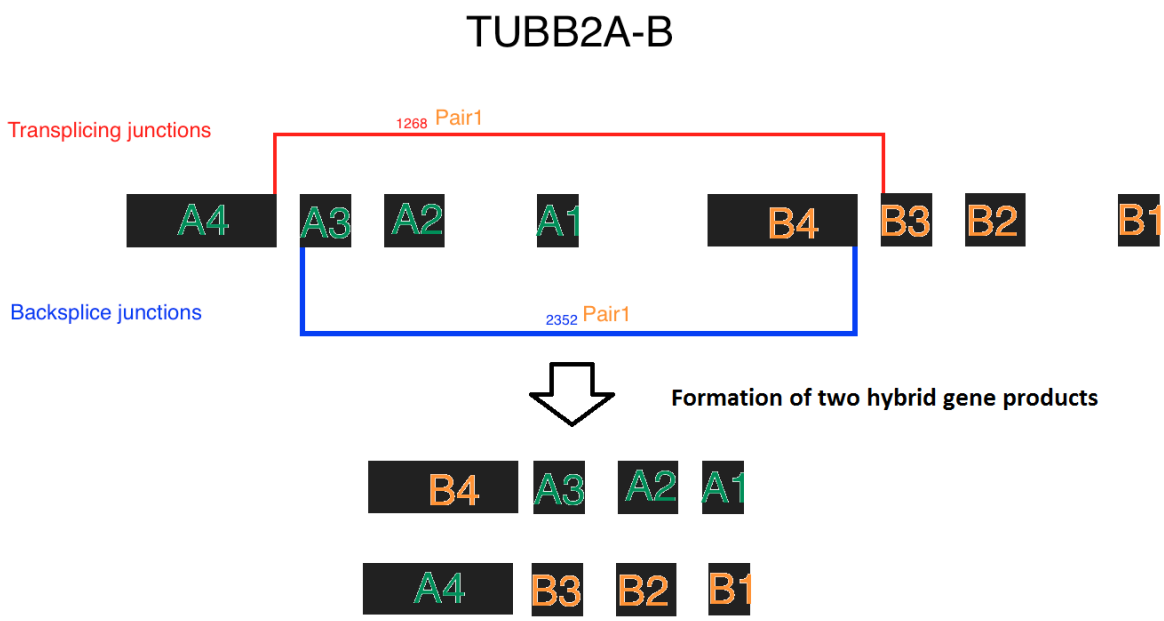

Figure 3.10: Reciprocal splicing in *TUBB* gene pair. *TUBB2A* (green) and *TUBB2B* (orange) show transplicing (red) and backsplicing (blue) junctions. Total numbers of junctions found across all samples are included. Transplice and backsplice junctions are grouped into reciprocal pairs (yellow text).

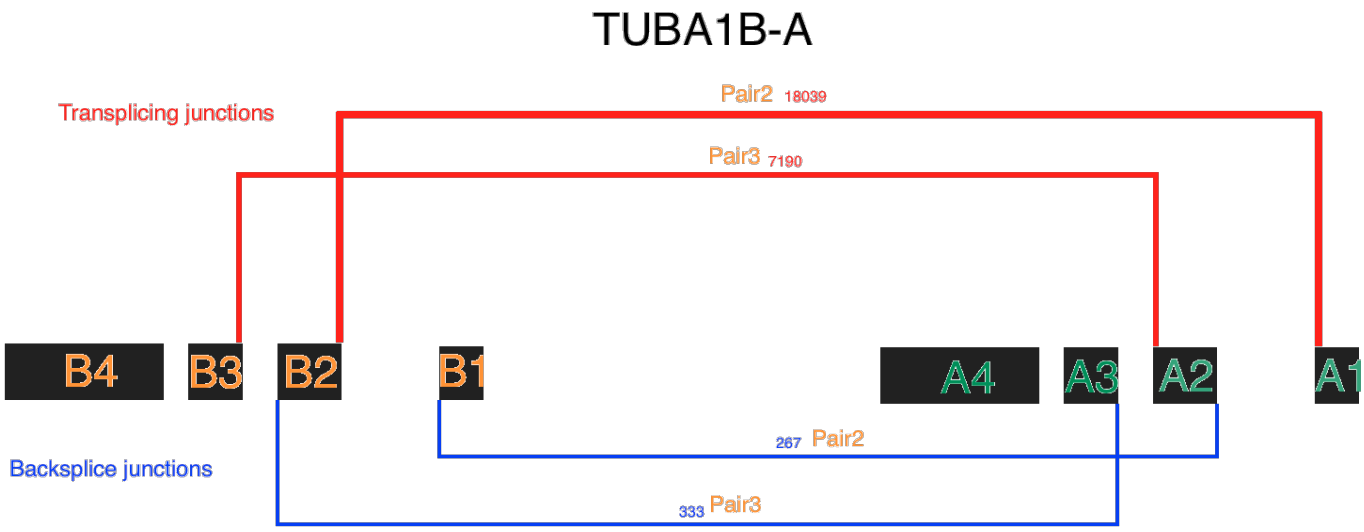

Figure 3.11: Reciprocal splicing in *TUBA* gene pair. *TUBA1B* (orange) and *TUBA1A* (green) show transplicing (red) and backsplicing (blue) junctions. Total numbers of junctions found across all samples are included. Transplice and backsplice junctions are grouped into reciprocal pairs (yellow text).

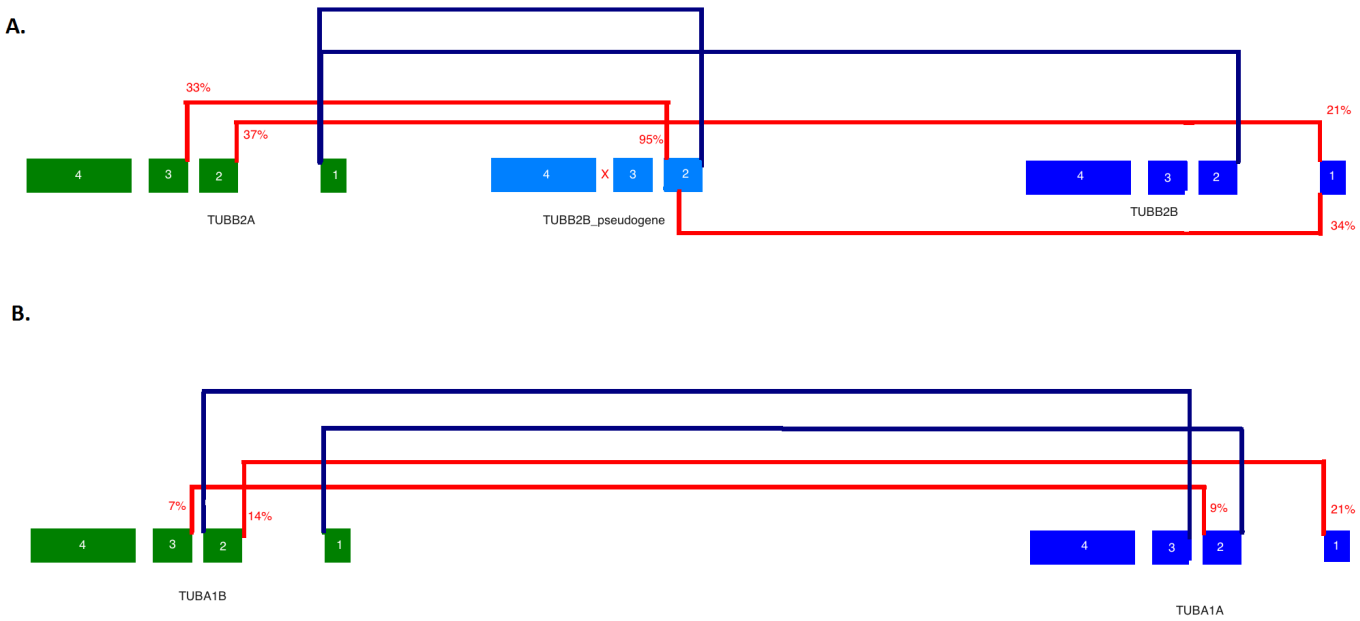

Figure 3.12: Comprehensive illustration of splicing in Tubulin gene pairs within their genomic context. Transplicing (red) and backsplicing (blue) junctions are shown. The percentage of junctions that are identified as transplicing are shown in red text.

3.3.5 Novel circRNA found in 18S rRNA

A brain-specifc circRNA (hg19 location chr21:9827249-9827513) was detected 60kb away from the nearest gene (Figure 3.13), this region is highly enriched for histone marks including H3k36me3. This circle does overlap both spliced ESTs and a Human 18S ribosomal RNA repeat sequence. This could describe an additional level of control over ribosomal genes, or it could be involved in rolling circle transcription identified previously [Hourcade et al., 1973].

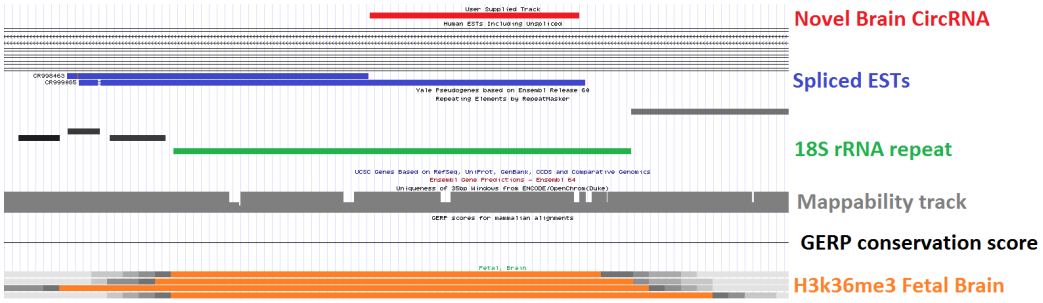

Figure 3.13: A UCSC browser track of the brain specific circle located in an 18S rRNA gene.

3.3.6 Case study: circRNA differential expression in Bipolar disorder

After classifying circRNA across multiple brain regions this database could be applied to other human brain data to sensitively quantify the presence of circRNA in much the same way as protein coding transcripts can be used to estimate expression.

A differential expression was performed on high confidence count data from publicly available, rRNA depleted samples of dorsolateral prefrontal cortex from 4 bipolar patients and 4 controls [Akula et al., 2014].

Interestingly, one of the largest fold-changes occurred in *BPTF* (Table 3.9), a constrained gene implicated as being involved in the development of Bipolar disorder [Li et al., 2013]. As FDR significance cannot be achieved with current counts these results are speculative. The study showed that the linear *BPTF* mRNA did not show any significant changes to transcript or gene expression [Akula et al., 2014]. This circle and others have been observed in several datasets [Salzman et al., 2012; Memczak et al., 2013; Venø et al., 2015] including maternal plasma [Koh et al., 2014]. Recently, it has been reported that several circRNAs in *BPTF* are some of the most highly expressed in brain [Rybak-Wolf et al., 2015]. This opens the possibility of looking at the expression change of circRNA to uncover subtle isoform changes that could provide insight to their function and regulation.

Circ ID	Database type	Gene start	Gene end	Constrained	Average counts	Function	Disease association
083812	Circbase,CircBrDB	CDR1as	CDR1as	.	4,152.3	Mir-7 sponge	Alzheimers
049410	Circbase,CircBrDB	SLC8A1	SLC8A1	.	431.5	Sodium-calcium exchanger	Cardiac Disease
022384	CircBrDB	TUBB2A	TUBB2B	Constrained	383.4	Beta-tubulins	Infantile-onset epilepsy, peripheral neuropathy
007549	Circbase,CircBrDB	PTP4A2	PTP4A2	.	243.6	Protein-tyrosine phosphatase	
103846	Circbase,CircBrDB	ANKRD12	ANKRD12	.	170.7	Ankyrin repeat-containing cofactor 2	
032974	Circbase,CircBrDB	SPECC1	SPECC1	.	104.4	Microtubule stability and actin cytoskeletal reorganization	
100454	Circbase,CircBrDB	BPTF	BPTF	Constrained	102.0	Fetal alzheimer antigen	Alzheimers
078225	CircBrDB	X	UNC13C	.	100.8	Predominantly expressed in brain	Significant motor learning deficit in mouse PubMed:11150314
048866	Circbase,CircBrDB	MAN1A2	MAN1A2	.	95.9	Mannosidase, alpha, class 1A	
103425	CircBrDB	TUBB2A	TUBB2B	Constrained	90.1	Beta-tubulins	
014979	CircBrDB	CKMT1B	CKMT1A	Constrained	87.9	Creatine kinase, mitochondrial 1A	
055782	CircBrDB	X	TULP4	.	76.4		
030742	Circbase,CircBrDB	TJP1	TJP1	.	76.1		
098599	Circbase,CircBrDB	HIPK3	HIPK3	.	73.9	Homeodomain-interacting protein kinase 3;	
090812	Circbase,CircBrDB	APOC1	APOC1P1	.	73.9	APOC1 is a specific inhibitor of CETP	Age-associated memory impairment
106417	Circbase,CircBrDB	PCLO	PCLO	.	73.4	PICCOLO: Involved in presynaptic cytoskeletal matrix	
079705	Circbase,CircBrDB	VCAN	VCAN	.	71.8	Chondroitin sulfate proteoglycan 2	Wagner vitreoretinal degeneration
057992	Circbase,CircBrDB	SLAIN1	SLAIN1	.	71.2	A novel stem cell gene	
0017242	Circbase	AKT3	AKT3	.	70.9	Protein kinase	
014746	Circbase,CircBrDB	ASH1L	ASH1L	Constrained	70.8		
002437	Circbase,CircBrDB	ARID1A	ARID1A	.	66.8		
036629	Circbase,CircBrDB	SLC8A1	SLC8A1	.	66.7	Sodium-calcium exchanger	Cardiac Disease
072371	Circbase,CircBrDB	TUBA1B	TUBA1A	Constrained	64.0	Tubulin, alpha, brain-specific	Lissencephaly type 3 (LIS3)
048643	CircBrDB	TUBA1B	TUBA1A	Constrained	62.8	Tubulin, alpha, brain-specific	Lissencephaly type 3 (LIS3)
074480	CircBrDB	KCNN2	KCNN2	.	55.7	Potassium channel,CA activated subfamily	Frissonnant phenotype
032750	CircBrDB	TUBB2A	PGO243043	Constrained	54.9	Beta-tubulins	Infantile-onset epilepsy, peripheral neuropathy
072389	Circbase,CircBrDB	FMN2	FMN2	Constrained	52.8	Cytoskeletal organization	
035624	Circbase,CircBrDB	RNF220	RNF220	.	51.4		
048988	CircBrDB	X	TULP4	.	49.6		
002661	CircBrDB	RMST	RMST	.	47.1	Noncoding rna in rhabdomyosarcoma	Malignant soft tissue tumor in children
0017252	Circbase	AKT3	AKT3	.	46.4	Protein kinase	
008266	Circbase,CircBrDB	QSER1	QSER1	.	46.0		
018063	CircBrDB	CHD9	CHD9	Constrained	43.4		
100360	Circbase,CircBrDB	CHD9	CHD9	Constrained	42.4		
071213	Circbase,CircBrDB	AFF2	AFF2	.	42.3	FMR2: Colocalizes with the splicing factor SC35	Mental retardation, X-linked, FRAXE type

Table 3.2: Top 35 most expressed circRNA identified in human brain. Average counts are given per sample. Database type indicates if the circRNA is common between Circbase and CircBrDB (unique to brain) . Constrained indicates if the gene is highly intolerant to deleterious mutation (based on [Samocha et al., 2014], see Section 4.1.1).

Circ ID	Database type	Circle counts	Fits circle (mate)	Fits circle (scaffold)
039370	CircBrDB	65	104	104
0026129	Circbase	0	0	0
006692	CircBrDB	126	492	492
0026130	Circbase	0	0	0
072371	Circbase,CircBrDB	267	464	278
048643	CircBrDB	333	870	780

Table 3.3: Backsplice junctions for the *TUBA1A/B* gene pair. Circle counts are the total reads mapping to the scaffold read with Mapping Quality higher than 30. Fits circle (mate) indicates the number of read pairs that map in a configuration consistent with backsplicing. Fits circle (scaffold) indicates the number of scaffold reads consistent with the backsplice. Circbase is a public repository for circRNA while CircBrDB is the database generated during this study from brain data.

Circ ID	Database type	Circle counts	Fits circle (mate)	Fits circle (scaffold)
050637	CircBrDB	4	6	6
001517	CircBrDB	3	19	19
103425	CircBrDB	2	2	2
022384	CircBrDB	2,352	5,306	4,252

Table 3.4: Backsplice junctions for *TUBB2A/B* gene pair. Circle counts are the total reads mapping to the scaffold read with mapping quality higher than 30. Fits circle (mate) indicates the number of read pairs that map in a configuration consistent with backsplicing. Fits circle (scaffold) indicates the number of scaffold reads consistent with the backsplice. Circbase is a public repository for circRNA while CircBrDB is the database generated during this study from brain data.

Circ ID	Database type	Circle counts	Fits circle (mate)	Fits circle (scaffold)
083557	CircBrDB	9	9	8
032750	CircBrDB	196	206	0
043960	CircBrDB	4	5	5

Table 3.5: Backsplice junctions for *TUBB2B* and pseudogene. Circle counts are the total reads mapping to the scaffold read with mapping quality higher than 30. Fits circle (mate) indicates the number of read pairs that map in a configuration consistent with backsplicing. Fits circle (scaffold) indicates the number of scaffold reads consistent with the backsplice. Circbase is a public repository for circRNA while CircBrDB is the database generated during this study from brain data.

Circ ID	Database type	Transplice counts	Fits Transplice (mate)	Fits Transplice (scaffold)
chr12-49523173-49580393	CircBrDB	7,190	7,190	0
chr12-49523356-49580467	CircBrDB	29	30	0
chr12-49521794-49578846	CircBrDB	6	6	0
chr12-49523051-49580119	CircBrDB	18	21	0
chr12-49523116-49580393	CircBrDB	33	33	0
chr12-49521850-49578902	CircBrDB	37	41	24
chr12-49523116-49580184	CircBrDB	16	16	0
chr12-49522721-49580092	CircBrDB	0	0	0
chr12-49523505-49582759	CircBrDB	18,039	18,039	0
chr12-49522647-49579699	CircBrDB	3,368	3,526	3,526

Table 3.6: Transplice junctions for *TUBA1A/B* gene pair. Transplice counts are the total reads mapping to the scaffold read with mapping quality (MAPQ) higher than 30. Fits Transplice (mate) indicates the number of read pairs that map in a configuration consistent with transplicing. Fits Transplice (scaffold) indicates the number of scaffold reads consistent with the transplice. Circbase is a public repository for circRNA while CircBrDB is the database generated during this study from brain data.

Circ ID	Database type	Transplice counts	Fits Transplice (mate)	Fits Transplice (scaffold)
chr6-3156386-3227720	CircBrDB	13	13	0
chr6-3154838-3225726	CircBrDB	110	5,638	0
chr6-3155157-3226392	CircBrDB	1,276	1,278	0

Table 3.7: Transplice junctions for *TUBB2A/B* gene pair. Transplice counts are the total reads mapping to the scaffold read with mapping quality (MAPQ) higher than 30. Fits Transplice (mate) indicates the number of read pairs that map in a configuration consistent with transplicing. Fits Transplice (scaffold) indicates the number of scaffold reads consistent with the transplice. Circbase is a public repository for circRNA while CircBrDB is the database generated during this study from brain data.

Circ ID	Database type	Transplice counts	Fits Transplice (mate)	Fits Transplice (scaffold)
chr6-3179946-3227720	CircBrDB	561	561	561

Table 3.8: Transplice junctions for *TUBB2B* and pseudogene. Transplice counts are the total reads mapping to the scaffold read with mapping quality (MAPQ) higher than 30. Fits Transplice (mate) indicates the number of read pairs that map in a configuration consistent with transplicing. Fits Transplice (scaffold) indicates the number of scaffold reads consistent with the transplice. Circbase is a public repository for circRNA while CircBrDB is the database generated during this study from brain data.

Circ ID	Database type	Gene name	Fold change	log2 Fold change	P value
062023	CircBrDB	CDKL3	27.36066672	4.774031481	0.005401941
051617	Circbase,CircBrDB	BPTF	18.29080292	4.193046502	0.022672423
044167	Circbase,CircBrDB	FAM169A	13.34439136	3.7381616	0.035539273
024169	CircBrDB	EIF2AK4	11.19167739	3.484354376	0.01756694
008046	CircBrDB	ZDHHC11	9.728720098	3.282250018	0.023440319
033998	CircBrDB	BRCA1	7.355875792	2.878897119	0.010804344
089984	CircBrDB	ZFP64	6.557646408	2.713178113	0.000326552
103894	Circbase,CircBrDB	ITGAX,ITGAD	5.388288956	2.42982722	0.000448939

Table 3.9: Differentially expressed circRNA in Bipolar brain. Data was produced using DESeq [Anders and Huber, 2010] differential expression of backsplices identified in the dorsolateral prefrontal cortex from 4 patients and 4 control samples. A circRNA in *BPTF* is identified as differentially expressed and has been shown as a recent candidate for the development of neurodevelopmental disorders [Li et al., 2013].

3.4 Discussion

circRNA are a new class of non-coding RNAs generated by backsplicing of an upstream 3' exon start to a downstream 5' exon end. Formation of circRNA is reliant on these exon splice sites being brought into close proximity. This is catalysed by complementary sequences in flanking introns, either directly through nucleotide hybridization or via RNA binding proteins. A minority of circRNA can function as microRNA sponges. Current evidence suggests that circRNA utilize the spliceosome similarly to linear transcripts thereby controlling gene expression through competition. However, data suggests there are other, undiscovered functions for these noncoding molecules.

This chapter outlines an analysis protocol to produce robust counts for circRNA from RNA-seq data. A custom 2-step pipeline is used to sensitively recover as much data as possible, emphasizing the use of paired-end data. This approach is applied to discover the most in-depth list of backsplice events in the human brain currently available, nearly doubles the number of circRNA backsplice junctions in Circbase when looking at the most robust and highly expressed circRNA.

A pipeline was designed to systematically investigate trans/backsplicing between pairs of proximal genes. This is the first algorithm to investigate splicing between highly similar genes although similar analysis approaches have been created to determine locations of overhangs in repeats using megaBLAST [Wilson and Stein, 2015; Criscione et al., 2014]. This pipeline is both resource and labour intensive as it cannot be applied indiscriminately, it is essential to verify transcript structure and gene annotations.

I find reads that consistently map the back/transplices between these genes. There can be several explanations for this. One possibility is biological or technical noise, high transcription rates of these genes, partial degradation of transcripts and random ligation events could produce these fragments which can then be amplified by PCR. However, the fact that I see these backsplices consistently across brains samples with a distribution of overhang lengths indicate multiple different fragments originate from these backsplices. Paired-end sequencing, which has been undervalued in identification of circRNA, greatly improves resolution by providing consistent mate-pairs mapping uniquely in a fashion concordant with these backsplices.

circRNA / backsplice count data is a gross underestimation of transcript abundance. Detecting circles relies heavily on the backsplice junction although, if paired-end data are available, this could be mitigated to some degree. This is further hampered by the need for unique reads when dealing with highly similar genes. Ultimately, if these splicing events are real they are likely expressed at much higher frequency than stated. These biases will affect accurate quantification as well as statistics applied to this data, specifically

differential expression which is designed for canonical gene expression.

For backsplicing to occur splice sites must be proximal. This implies backsplices between genes requires structural changes in the chromatin. Recent studies have revealed that regions containing highly expressed genes tend to be folded into loops with the help of the transcription factor *CTCF* [Tang et al., 2015]. These regions also tend to be significantly enriched for histone marks which aid transcription. In cancer the role of *CTCF* in preventing the formation of trans-gene products due to insulating loops [Qin et al., 2015] has been investigated.

Another consideration is whether the backsplice would create a circular RNA. This would require transcript read-through between genes. It was recently shown that read-through between proximal genes does occur and can be linked to tumor phenotypes [Grosso et al., 2015]. In the available data no significant evidence of transcript read-through could be detected. This could be due to the low copy number of these events or the instability of such a long RNA fragment. An alternative hypothesis is that two separate RNA molecules could interact post-transcriptionally to produce two hybrid linear fragments. In order to answer these questions laboratory validations will be required.

There is great potential for circRNA based on independent cellular regulation from linear isoforms, enrichment in blood plasma and saliva and their association with neuronal cells and various RNA binding factors implicate their importance as key molecules in cellular processes. The recent finding that gene fusions in cancer can create novel circRNA through complementary Alu elements within the fused introns indicates structural aspects are crucial in circRNA synthesis. The fact that these cancer-specific circRNA are tumourgenic and provide resistance to therapeutics only emphasizes the need for further investigation. [Guarnerio et al., 2016]

Here I expand on the known catalogue of circRNA and related backsplicing in the brain while exploring a peculiar subclass of RNA molecules potentially generated by splicing between transcripts.

Chapter 4

Annotating and functional determination of non-coding features using variant information from exome sequencing

4.1 Introduction

The human genome is a vast landscape containing 3 gigabases of sequence. An estimated 1% resides within genes and other functional units and yet these are so paramount to cellular function they show similarity across various species. One key aspect to understanding the genome is identifying ways to annotate and discover patterns of bias in sequence data. Here I explore how a similarly powerful method can give further insight into cellular function.

4.1.1 Variant conservation as a method to identify constrained sequence

The similarity of genes across diverse species is one of the fundamental observations of evolutionary genetics. This information can be utilized to determine nucleotide conservation across species indicating the importance to cellular function. Sequence conservation across species is widely used to define functional genetic elements. Nucleotide substitution rates that are lower than expected by neutral drift indicate the

cells need to conserve particular sequence motifs. This has spurred computational comparisons of vertebrate genomes to elucidate classes of functional elements including protein-coding genes, RNA genes, enhancers and microRNA target sites [Guigo et al., 2003; Nobrega et al., 2003; Siepel et al., 2007]. These methods have proved highly valuable but lack the ability to identify species-specific conservation or recently evolved mechanisms.

The use of exome sequencing to identify rare, casual variants has paved the way for the analysis of complex, heritable traits. Through using the ExAC exome consortium Samocha et al. were able to estimate the rates of de-novo mutation, produce gene-specific probabilities for different mutation types (such as synonymous, missense, nonsense, essential splice site and frameshift) and apply these to find genes which have significantly fewer mutations than expected (See Figures 4.1 and 4.2) [Samocha et al., 2014]. These "constrained" genes are not only enriched for many disease associated genes but contain hundreds of unknown genes which have yet to be understood [Samocha et al., 2014]. This resource provides an additional level of annotation when analysing variant data as mutations within these regions could be highly relevant to disease causation.

With the creation of large exome consortia such as UCLex and ExAC (See Chapter 1) I am now able to query variants directly to determine the proportion of variation across sequence features. This provides an unprecedented opportunity to confirm human-specific elements with base-pair resolution.

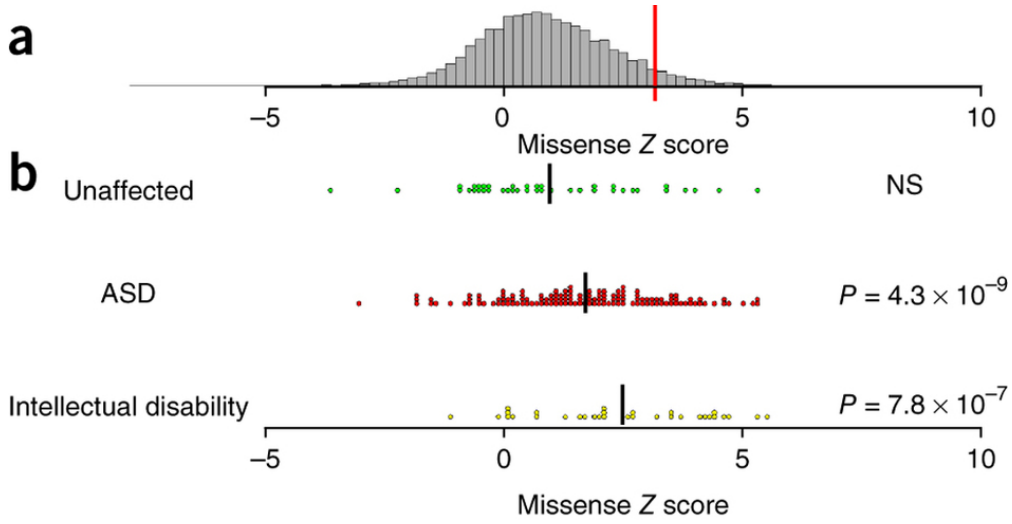

Figure 4.1: (A) Distribution of Z scores for missense mutations across genes in the human genome. Z scores are based on observed vs expected prevalence of single nucleotide polymorphism (SNPs) using ExAC data. A tail of significantly invariant genes is shown beyond the red line. (B) This highlights a higher prevalence of non-synonymous missense variation in Autism spectrum disorder (ASD) and intellectual disability compared to unaffected individuals. Black lines indicate population means. [Samocha et al., 2014]

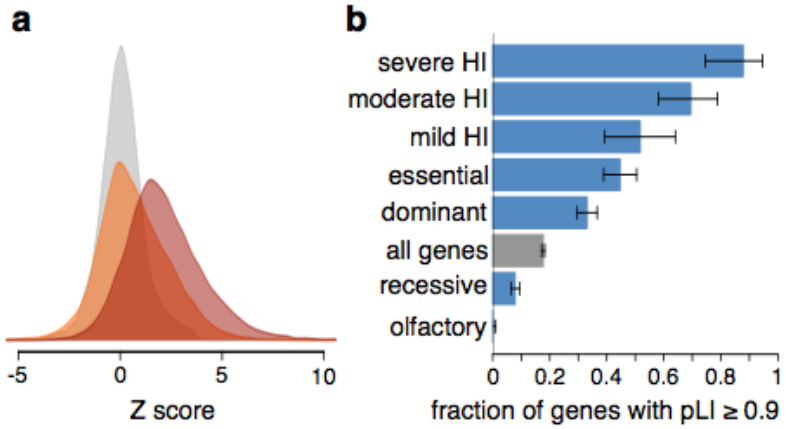

Figure 4.2: (A) Distribution of Z scores across genes in the human genome. Z scores are based on observed vs expected prevalence of SNPs using ExAC data, divided into synonymous(grey), missense(orange) and protein-truncating(red). (B) The proportion of genes that are highly intolerant to deleterious mutation, broken down into categories based on ClinGen annotation. Showing the relationship between cellular importance and probability of genes to be highly intolerant to deleterious mutation. This is showcased by haploinsufficient (HI) genes that consist mostly of constrained genes. [Samocha et al., 2014; The EXaC Consortium, 2015]

4.1.2 Branchpoints are an essential element to exon recognition and splicing

Branchpoints are one of the crucial exonic features required for the formation of the intronic lariat and recognition of the intronic 3' splice site [Reed, 1989]. Disruption of these locations can result in splicing defects attributed to numerous hereditary diseases [Stenson et al., 2003].

Identification of branchpoints is not trivial as they are intronic and poorly understood. *In silico* prediction of these sites have provided a large number of candidates but verifying these sites was intractable [Corvelo and Eyras, 2008]. Alternatively, lariat-spanning junctions can be mined from total RNA-seq data [Taggart et al., 2012] or lariat debranching enzymes can be inhibited during sample preparation [Bitton et al., 2014]. Both approaches are suboptimal in terms of high throughput discovery and have only elucidated a few hundred results.

Recently two studies have identified tens of thousands of branchpoints genome-wide using two different methods. Through the use of individual nucleotide resolution UV cross-linking and immunoprecipitation (iCLIP) [König et al., 2010] more than 64% of branchpoints across 50,000 introns have been resolved [Briese et al., 2016]. Secondly, by enriching for intronic lariats (via RNase R digestion) and reverse transcribing the branched junction (CaptureSeq) Mercer et. al were able to determine branchpoint location (Figure 4.3) [Mercer et al., 2015]. This identified 59,359 high-confidence human branchpoints in >10,000 genes. Their

results supported previous reported studies that the conservation of the U2 binding site; the upstream U and branchpoint A (UnA, referred to as Bbox) remained the most conserved. However, it was also noted that splicing is resistant to Bbox mutations [Berglund et al., 1997; Gao et al., 2008]. Another recent study looked at the impact of branchpoint distance from the 3' splice site. Proximal branchpoints were six times more likely to be spliced [Rosenberg et al., 2015] indicating that clear effects on splicing can be observed at these locations.

These breakthroughs have made the study of branchpoints possible. For the first time, I am able to evaluate these cryptic features using human polymorphism data.

4.1.3 Exploring splice site variation by integrating genomic variation with gene expression data

High-throughput functional interpretation of variation has only recently become feasible thanks to the rise of next generation sequencing. The GEUVADIS consortium combined RNA-seq from lymphoblastoid cell lines of 462 individuals from the 1,000 Genomes Project [Abecasis et al., 2012] with their variant data [Lappalainen et al., 2013] (see Section 1.4.2 for more information). This landmark study illustrated the relationship and effect of exonic variation on gene expression. Analysis performed focused on expression quantitative trait loci (eQTLs), the effect of regulatory and loss of function variants and allele-specific effects. This was succeeded by a second study [Rivas et al., 2015], looking at protein truncating variation (PTV), using GEUVADIS data in combination with gene expression and exome data from the Genotype-Tissue Expression (GTEx) consortium [Genotype-Tissue Expression Consortium, 2015].

Interestingly, their focus on how premature stop codons can trigger nonsense-mediated decay (NMD) supports previous findings that these variants tend to occur roughly 50bp upstream of the 3' exonic splice site [Nagy and Maquat, 1998]. They improve this prediction substantially while noting that rare PTVs are more likely to trigger NMD. They also note the effect of variation around splice sites, defining variants which affect exon expression and how this relates to the allele frequency of the variant (see Figure 4.4) . Overall this study emphasises the importance of nucleotides close to the splice site motif but does not discuss how this information can be used to improve current prediction or how sequence affects the probability of mutation.

An alternative approach to determining variant effects on gene expression, specifically through investigation of alternative exon splicing, was done using a neural network called "SPANR" (Figure 4.5). This algorithm combined variant information with thousands of annotated RNA binding factors, splicing related features and splice site annotations. Each variant was then classified according to the features they dis-

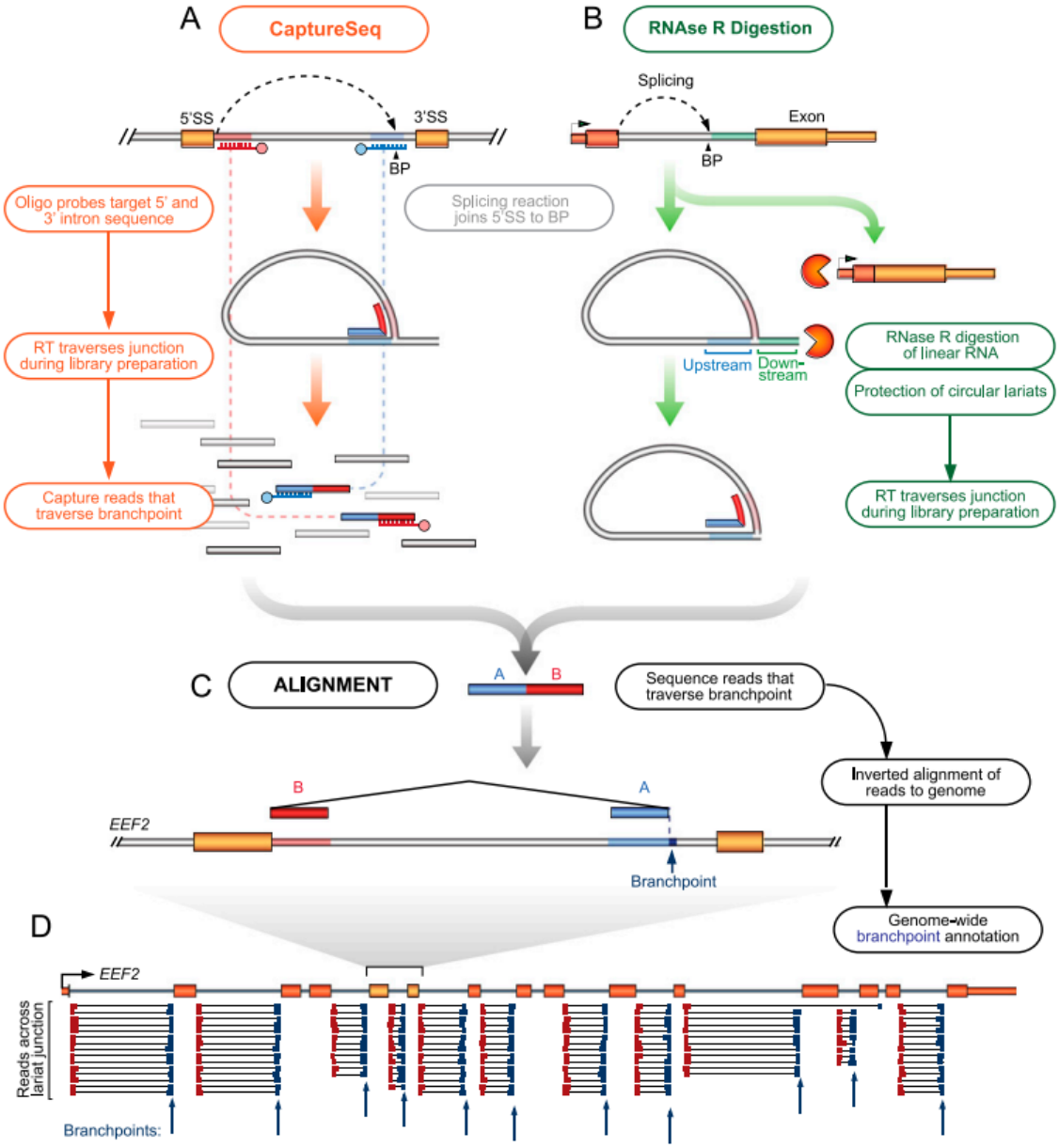

Figure 4.3: Identification of branchpoints using (A) CaptureSeq and (B) RNase R to digest linear mRNAs and selectively enrich circular RNAs including lariats. (C) Reads are aligned to the human genome to identify branchpoint locations with the 3' termini indicating the branching nucleotide.(D) Examples of identified branchpoints in the *EEF2* gene. [Mercer et al., 2015]

rupt and other surrounding RNA features. This information was then combined with splicing information from multiple cell types to train the model to estimate the effect for each variant. This process produced robust and impressive results but remains computationally very expensive and limited to predicting inclusion/exclusion of alternative exons. The question of how much variation can be captured using just splice

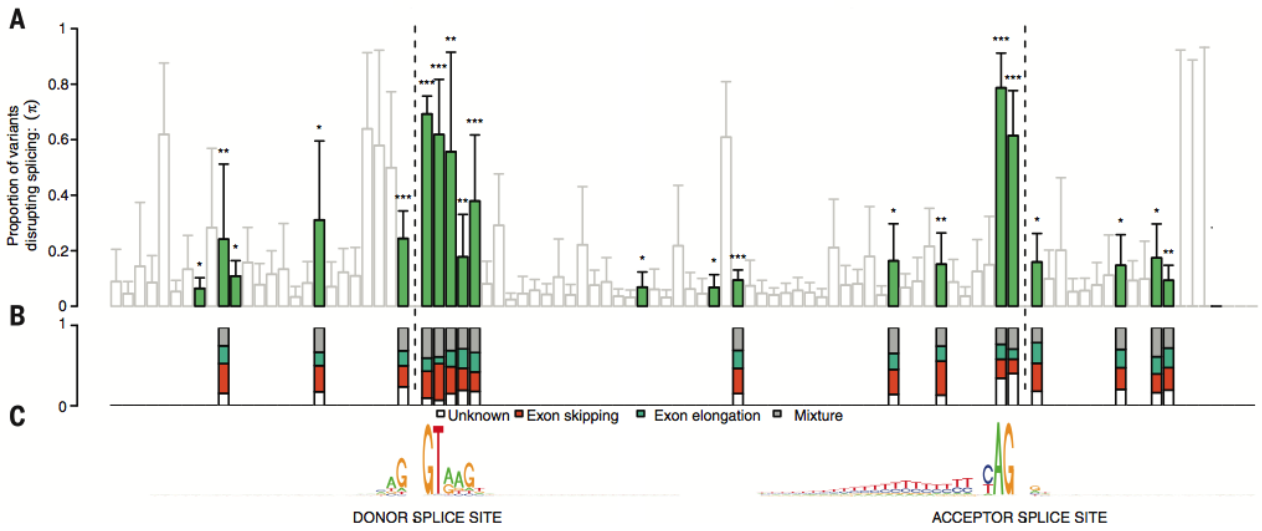

Figure 4.4: An integrated analysis of Genotype-Tissue Expression (GTEx) consortium data. (A) Shows the proportion of variants that effect splicing efficiency and their nucleotide position relative to the splice site. (B) Shows a classification of the types of events that results from splicing. [Rivas et al., 2015]

site information is not addressed. [Xiong et al., 2014]

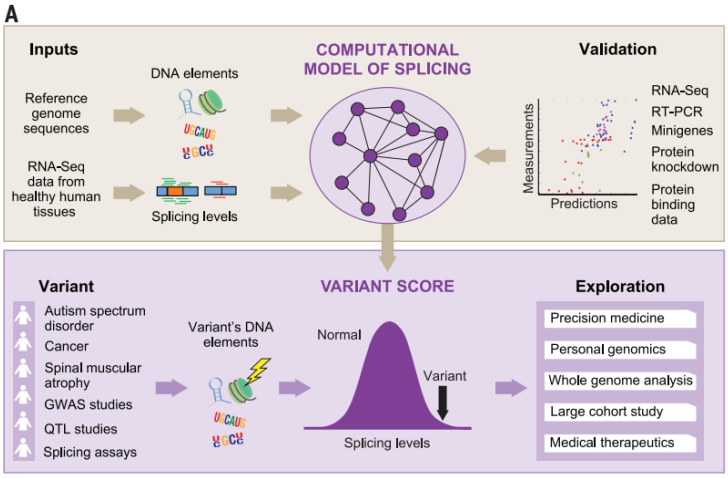

Figure 4.5: (A) Top: A pipeline using machine learning techniques to predict splicing changes by correlating DNA/RNA features with splicing levels in healthy tissues. (A) Bottom: This technique can be applied to filter lists of variants to identify those with a high probability of resulting in splicing changes within genes. [Xiong et al., 2014]

A significant study was recently published that explored alternative exon regulation through introduction of systematic variation in thousands of minigenes transfected into human HEK293 cells. Through the analysis of thousands of data points a model was designed to predict exon skipping effects and variant effects on splice sites. This model used the ratios of exonic hexamers, which appear to function in an additive

way in splicing. These hexamers included core sequences for exonic enhancers/silencers, branchpoints and cryptic splice sites and outperformed the machine learning algorithm "SPANR" mentioned earlier [Rosenberg et al., 2015]

Rosenberg et. al noted that although degenerate sequences in the introns had an impact on splicing, there are indeed many more splicing enhancer/silencing sequences than previously discovered that operate within exons and have a stronger impact than intronic sequence. They were able to predict SNP effects within the alternate exon and splice sites with high accuracy better than that of the state of the art software MaxEnt [Yeo and Burge, 2004]. Similar to SPANR their software resides as a web service, with limited use to predict exon skipping as it requires exon definition for all three exons. This imposes limits on number of calls that can be made at one time and lack of automation as it must be manually submitted. [Rosenberg et al., 2015]

4.2 Methods

In addition to the software and tools described in section 2.1.1, this section lists the methods that have been specifically used for this chapter.

4.2.1 Calculating the cumulative variant ratio across features of interest

In order to determine whether variants could provide insight into feature conservation both UCLex and ExAC [The EXaC Consortium, 2015] datasets were interrogated. For each feature a 20bp flanking region was defined. The cumulative ratio of variants to reference calls was calculated for each position across all features. Features tested include; splice-sites of internal exons, branchpoints and the first 60bp of internal exons for all highly constrained genes [Samocha et al., 2014].

4.2.2 Elucidation of potentially deleterious branchpoint variants

In order to gain the highest resolution branchpoint features from both described high throughput studies [Briese et al., 2016; Mercer et al., 2015] were merged. This resulted in a final set of over 90,000 branchpoints. Schematic of the process is shown in Figure 4.6.

Through visual inspection of the cumulative ratio plots it became clear the 1^{st} motif position (upstream U) and 3^{rd} position (branchpoint A) (UnA) remain the most conserved, in agreement with previous studies [Berglund et al., 1997; Gao et al., 2008]. Through mining these positions in the available consortia,

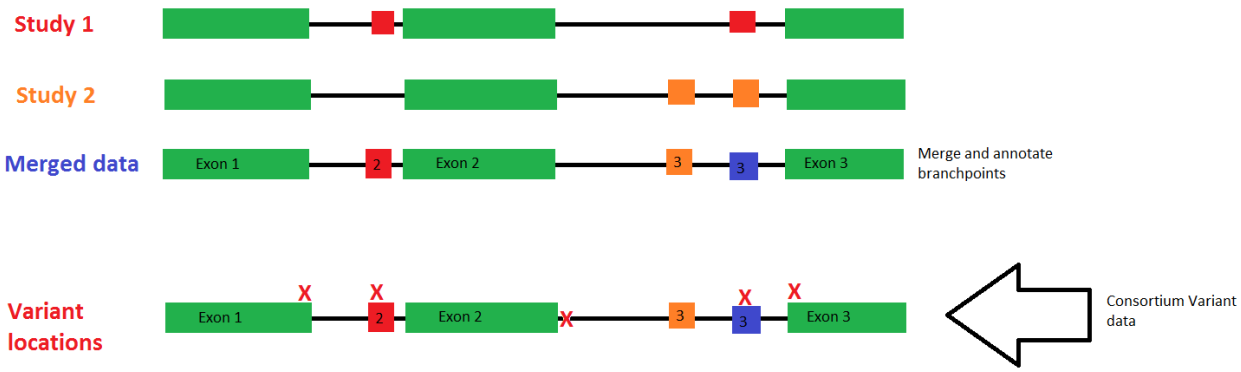

Figure 4.6: Merging branchpoints to achieve an annotated and comprehensive list of all variants. Identified branchpoints from one study (red) are merged with a second (orange), overlapping branchpoints are merged into a single site (blue).

filtering for extremely rare homozygous variants, I was able to select a small number of potentially disease causing candidates present in branchpoints. Ultimately, to determine their effect on splicing further investigation was necessary.

4.2.3 Interpreting splice site variation through integration with gene expression data

A custom pipeline was developed to integrate exome-based genomic variation from the 1,000 genomes project with polyA RNA-seq from the GEUVADIS study [Lappalainen et al., 2013]. This data was then applied to splicing related features to determine the functionally observable effect of mutation.

This pipeline consisted of custom python scripts with extensive use of Pysam, Biopython and Bedtools libraries [Cock et al., 2009; Quinlan and Hall, 2010; Li et al., 2009].

Defining splice site and branchpoint annotations and gathering variant information

A full list of exonic splice sites was retrieved from the GENCODE v19 annotation [Harrow et al., 2012]. All known branchpoints from both studies [Mercer et al., 2015; Briese et al., 2016] were merged into a single database using Bedtools. This merged annotation was used for all further investigation.

A strict sequence definition was then created for both splice sites and branchpoints based on accumulated data from variant graphs and literature. Variants occurring in the last exonic position or two first intronic positions of both 5' / 3' splice sites were recorded. Similarly, for branchpoints only variants occurring at the central A or upstream U (UnA) were recorded and used for further analysis. This strict

definition was chosen to ensure the variants directly affected the features as several other exonic features in close proximity may also overlap extended splice motifs [Xiong et al., 2014]. These variants were grouped into a splicing associated variation database and used for all further analysis.

Exon expression analysis of potentially deleterious mutation

Exon expression data from the GEUVADIS project was analysed to determine if there was a visible effect on gene/exon expression in the data. These data consisted of library depth and Peer-factor normalized read counts. For each variant all samples with at least one allele (heterozygous or homozygous) were grouped and compared to remaining wildtype samples across the gene of interest containing the splicing variant.

Annotation of variants and creation of a variant splice site score

Splice sites annotation

In order to focus on rare variants most likely to result in deleterious change a threshold was set for variations with fewer than 10 homozygote samples.

For each variant the entire splice site sequence was extracted. The variant was applied to the human reference sequence to create a mutated splice site. With small indels care was taken to replace nucleotides in a manner to maintain exon integrity. Table 4.1 shows the bases extracted from each splicing feature.

Splicing feature	Number of exonic bases	Number of intronic bases
5P	2 exonic	7 intronic
3P	2 exonic	20 intronic
BP	0	5 intronic

Table 4.1: Sequence extracted from each splicing feature for further analysis.

Both 3' and 5' splice sequences were scored using the splice site scoring tool MaxEnt [Yeo and Burge, 2004]. MaxEnt models sequence motifs based on the principle of maximum entropy. The maximum entropy distribution is determined using a set of constraints estimated from available splice site data. MaxEnt outperforms naive motif summarising approaches by taking the surrounding sequence of each nucleotide position into account. The difference in splice site score was calculated between wildtype and variant splice sites. This was used to indicate the degree of deviation from the functional splice site.

Branchpoint annotation

A position specific scoring matrix commonly used for motif discovery was applied on the annotated branchpoint database. A degenerate consensus was created. This was concordant with current research indicating the first and third positions were generally "U" and "A" respectively. The score was normalized using

the average intronic GC content (GC 40%). Scoring was done according to a PWM formula originally used for splice sites and other biological sequence [Shapiro and Senapathy, 1987]. The algorithm was implemented using Biopython [Cock et al., 2009], the equation is shown below:

$$score = 100 \times \frac{VariantPWMscore - MinimumPWMscore}{MaximumPWMscore - MinimumPWMscore}$$

Mining BAM files for splice junction data

For each variant position GENCODE v19 annotation was used to identify the gene, exon of interest and upstream exon. All data from the gene of interest were extracted from the 426 BAM files (mapping statistics are available in Appendix Table 2). All junction reads that have at least a 15bp overhang on either side of the intron and were in proximity to the exon of interest were retained. Read counts were extracted from the shores of the exon of interest and the neighbouring exon connected by the junction.

In order to make this analysis tractable optimization was essential as storage and processing of 426 RNA-seq samples was over 2 terabytes of binary compressed data. In order to minimize resource requirements all genes of interest were extracted from all BAM files to create a smaller, easily accessed copy. All junctions and read counts were recorded in hash tables for efficient storage.

After collection of the raw data the canonical splice junction (hereafter; JunctionA) was identified. This was defined as the most highly expressed junction across all samples. Several other statistics were collected based on this junction and are described in Table 4.2 and Figure 4.7. As all splicing starting from the same splice site was in direct competition, this allowed the calculation of canonical junction ratios compared to other non-canonical junctions from the same splice site. These splicing ratios were implemented at both the effected variant splice site and the upstream splice site connecting the junction (UPSTR). Furthermore, it was noticed that in some cases splice junctions were shifted a short distance from the splice site (due to the effect of the variant). In order to capture these a separate statistic (JA Ratio) took these junctions into consideration.

Shifted junctions were defined as junctions that originate within 20bp of the exon of interest splice site and splice in the same direction as the canonical splice junction. A similar approach was taken for the Branchpoint variants as these were associated with the downstream 3' splice site. Filtering was done to remove all samples with low upstream exon expression (< 1.15 normalized exon shore coverage) indicating insufficient expression for analysis.

Statistic	Description
UPST Ratio	A ratio of JunctionA/all junctions from the neighbouring junction exon splice site
Variant exon	Read count of the last 100bp of the exon of interest, normalized by total mapped reads
UPST/EOI exon	Read count of the last 100bp of the neighbouring exon, normalized by total mapped reads
JA Ratio	A ratio of JunctionA/(JunctionA + shifted junctions*)

Table 4.2: Statistics generated from splice junctions and exon expression. *Shifted junctions are defined as junctions that originate within 20bp of the exon of interest splice site and splice in the same direction as the canonical splice junction.

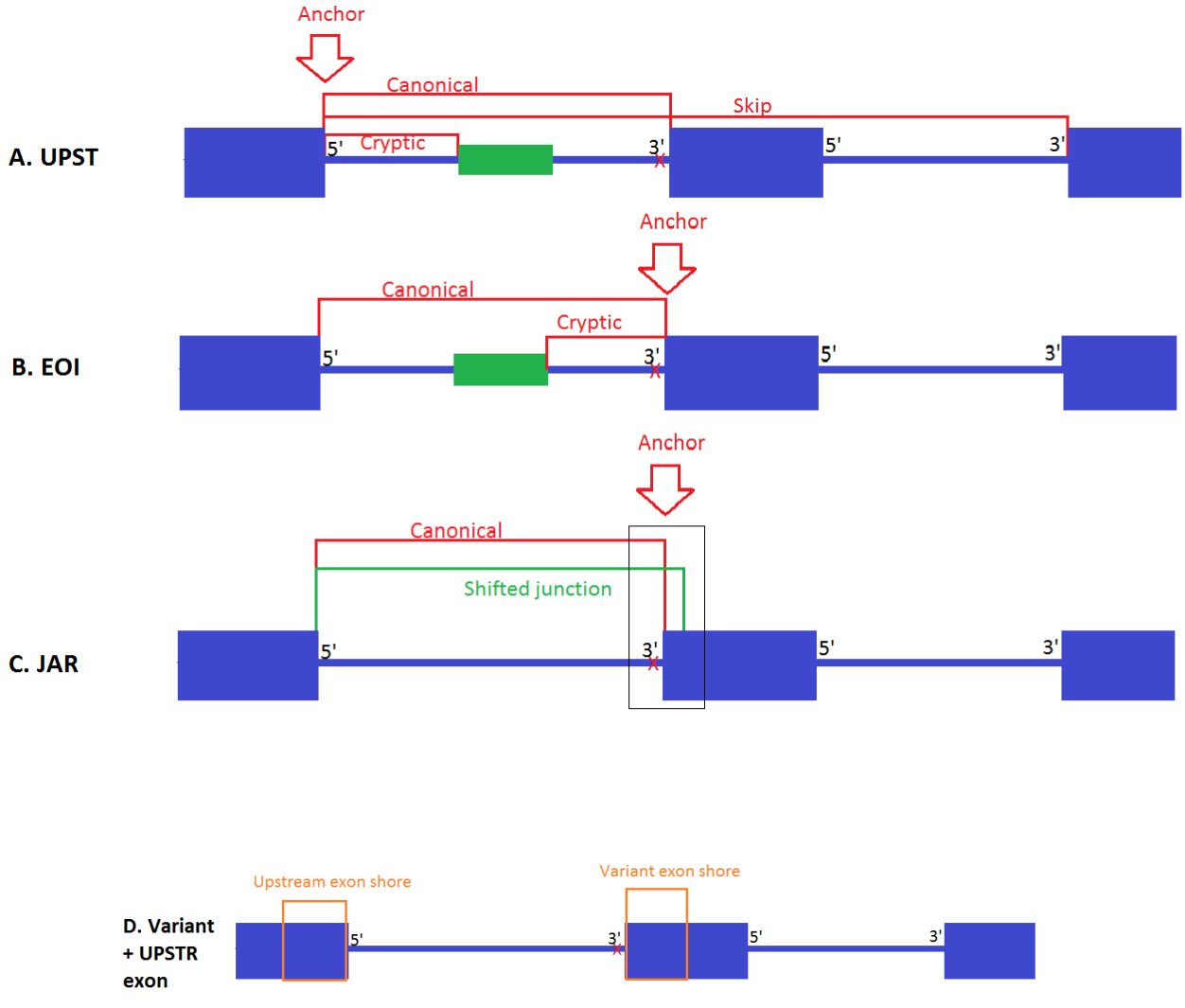

Figure 4.7: Calculation of multiple ratio statistics dependent on which splice site is being investigated. A. Upstream splice site (UPST). B. Variant location (exon of interest i.e. EOI) and C. Looking only at shifted junctions from the variant location (JA Ratio). D. Exonic shores within the variant and upstream exon (Variant exon, UPSTR exon).

Statistical testing to determine significant difference and correlation

A Wilcoxon ranked test was applied to test each statistic; data were divided between wildtype and variant (heterozygous and homozygous were grouped) samples. This provided a P value to gage the difference in expression between these groups indicating whether a change is in fact present.

I correlated the score of each variant to each statistic by fitting a linear model. Initially there was no significant trend in the data. This was largely due to high variance in gene expression leading to inclusion of uninformative samples, this made distinguishing lack of signal from lack of coverage intractable. I filtered aggressively using the P values from the Wilcoxon test (p value < 0.0005) to allow the selection of variants that show differential expression between wildtype and variant alleles. From this filtered data significant correlation between variant score and splicing statistics was obtained.

Given the P value distribution tends towards bimodal/non-normal further verification was necessary. In order to verify results each statistic was bootstrapped 1,000 times with replacement to achieve an average r-squared. In order to get a measure of sensitivity a leave one out cross validation was also performed.

4.3 Results

4.3.1 Variant ratio graphs

Figure 4.8 shows variant frequencies at each codon position closely mimic traditional sequence conservation. Figures 4.9 and 4.10 show overlapping variant graphs of splice sites and branchpoints from UCLex and ExAC respectively. The overlay of splice sites with branchpoints is intended purely as context for comparison of frequency of variation between these two features. It is clear that splice sites have far lower variation than branchpoints (where conserved positions are about as invariant as exons).

Figures 4.9 and 4.10 have similar trends indicating the robust nature of variant summary regardless of individual samples. It is clear that the first two intronic splice site positions are highly conserved. Interestingly, the first exonic position is shown as equally highly conserved in the ExAC data. When evaluating the branchpoint graph it is clear that both the first and third positions are conserved compared to the intronic context. The first position appears to be more highly conserved in both UCLex and ExAC datasets.

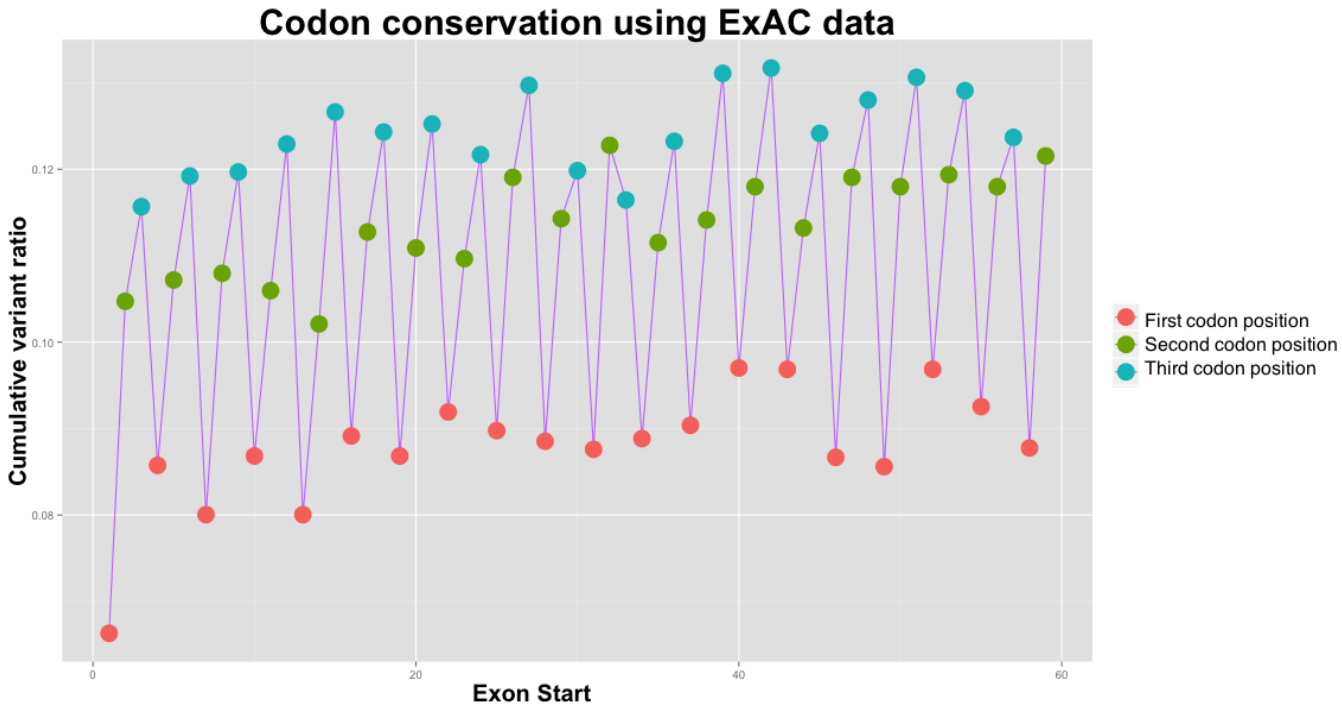

Figure 4.8: The cumulative ratio of variants across internal exons of highly constrained genes. Each codon position is represented by a different colour. Codon conservation is clearly observed indicating the sensitivity of this approach to identify functional conservation.

4.3.2 Potential branchpoint disease variants

By evaluating low frequency changes at the 1^{st} position and 3^{rd} position of all identified branchpoints (within range of the exome capture) it is possible to identify potentially disruptive mutations.

At least 103 (51: 1^{st} position, 52: 3^{rd} position) homozygous branchpoint changes were identified in the UCLex data, 20 % of these fall within the constrained gene category.

For the ExAC data, 191 variants (83: 1^{st} position, 108: 3^{rd} position variants) are found with 5 or less homozygous calls and less than 100 heterozygous calls. 10% of these fall within constrained genes. Interestingly, 59 variants are found on the X chromosome, this introduces the added complexity of the male as hemi-zygous for variants on the X chromosome. However, heterozygous counts for these variants are also extremely low (<8 heterozygous calls) and a similar imbalance is not present in the UCLex data. This phenomenon remains unexplained, possibly indicating a variant calling artefact or an X-linked disease cohort in ExAC.

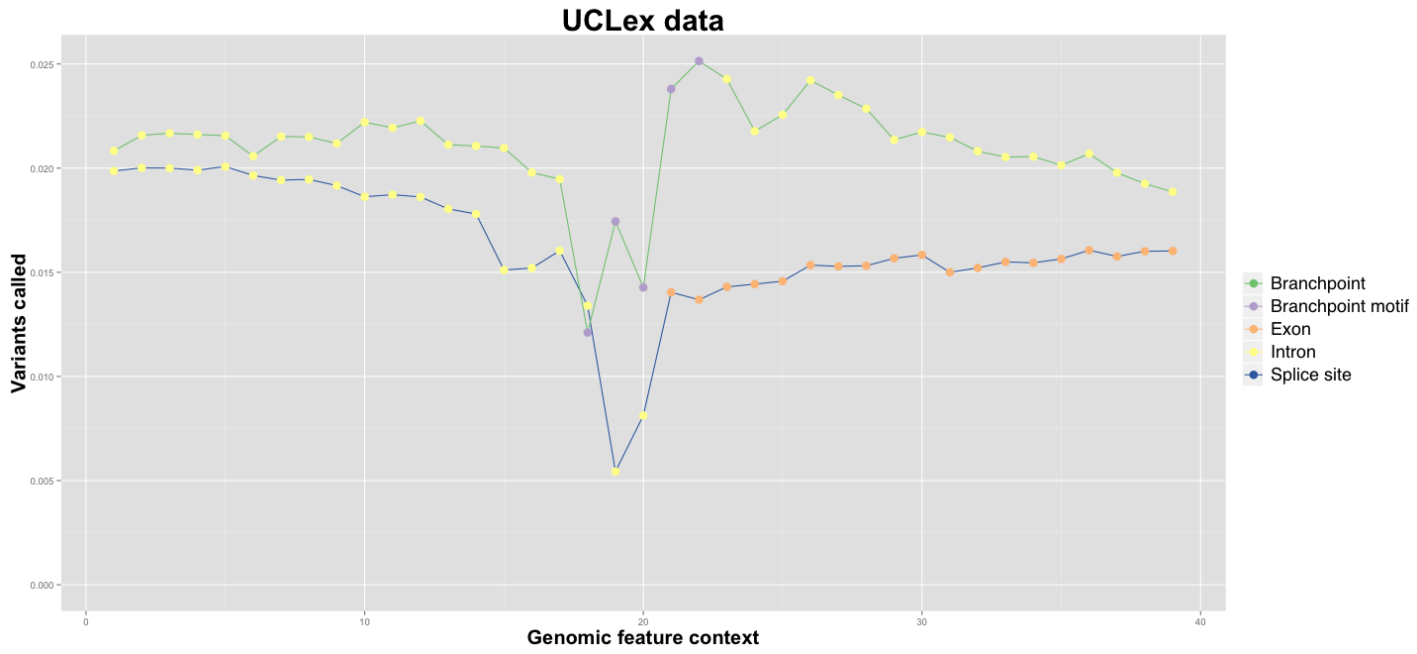

Figure 4.9: UCLex data variant graph showing the ratio of variants across splice sites (blue line) and branchpoints (green line). Yellow positions indicate intronic positions, the branchpoint site is highlighted in purple, exonic nucleotides are shown in orange.

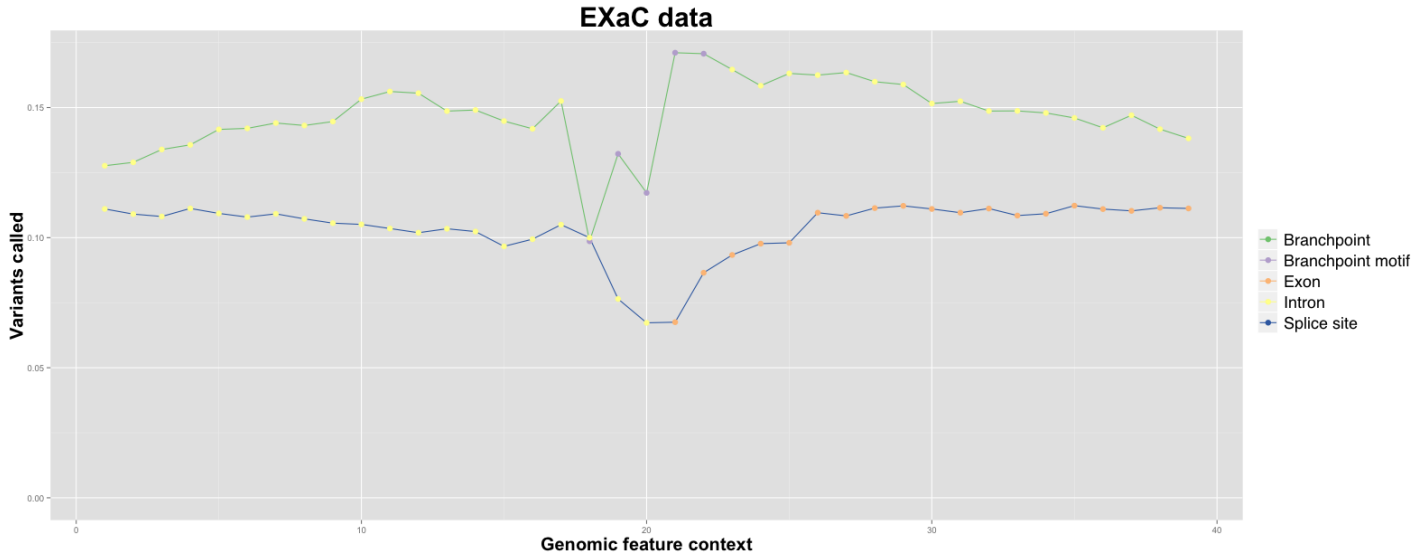

Figure 4.10: ExAC data variant graph showing the ratio of variants both splice sites (blue line) and branchpoints (green line). Yellow positions indicate intronic positions, the branchpoint site is highlighted in purple, exonic nucleotides are shown in orange.

4.3.3 Integrated variant and splice junction analysis

Table 4.3 shows the number of relevant splice feature variants retained after filtering for gene expression. These were filtered further if wildtype and variant groups showed significant splicing difference across any of the statistics. It is clear that splice site mutations occur very infrequently, more than half occur in genes that are not being expressed. Further, due to high levels of noise and expression variation it was essential to select a subgroup of variants that show measurable impact on the splicing phenotype of the gene in question. This reduces usable data to roughly 10% but provides a solid foundation for further testing.

Feature	Total Variants	Expression filtering	Significant variants
5'	1607	770	155
3'	1402	664	130
BP	1001	808	201

Table 4.3: Results from filtering variants associated with splicing features. In order, total rare variation associated with splice features, total variants after filtering for gene expression in the cell line and total variants with highly significant ($p < 0.0005$) change between wildtype/variant groups.

Bar plots in Figure 4.11 provide an overview of general properties associated with variants and their distribution around the splice site. As expected transition/ transversion (Ti/Tv) ratios remain constant across filtering and features. Interestingly, 5' splice sites seem to show the highest conservation at the 2nd intronic position. While 3' splice sites show no significant preference but higher indel occurrence, possibly this can be tolerated more readily by the splicing machinery. Similarly, branchpoints conform to the literature and variant graphs showing fewer effect-variants on the first "U" position.

Exon expression does not capture subtle splicing change

In order to determine the effects of rare splicing variants exon expression levels of the gene were investigated. Figure 4.12 shows three examples of exon expression in genes containing variants within core motifs. It became apparent that very little detectable difference was present in the majority of cases. I concluded that either the majority of these variants have no effect, the effect was being masked due to variance in sample expression or that exon expression does not capture subtle changes to splicing effectively.

Score distributions

If splicing machinery is indeed so robust that the majority of rare variants have no substantial effect a score would be ideal to measure the effect of the variant on the splicing motif. A score was designed using MaxEnt [Yeo and Burge, 2004] to score the wildtype and variant splice sites and the difference between them. This

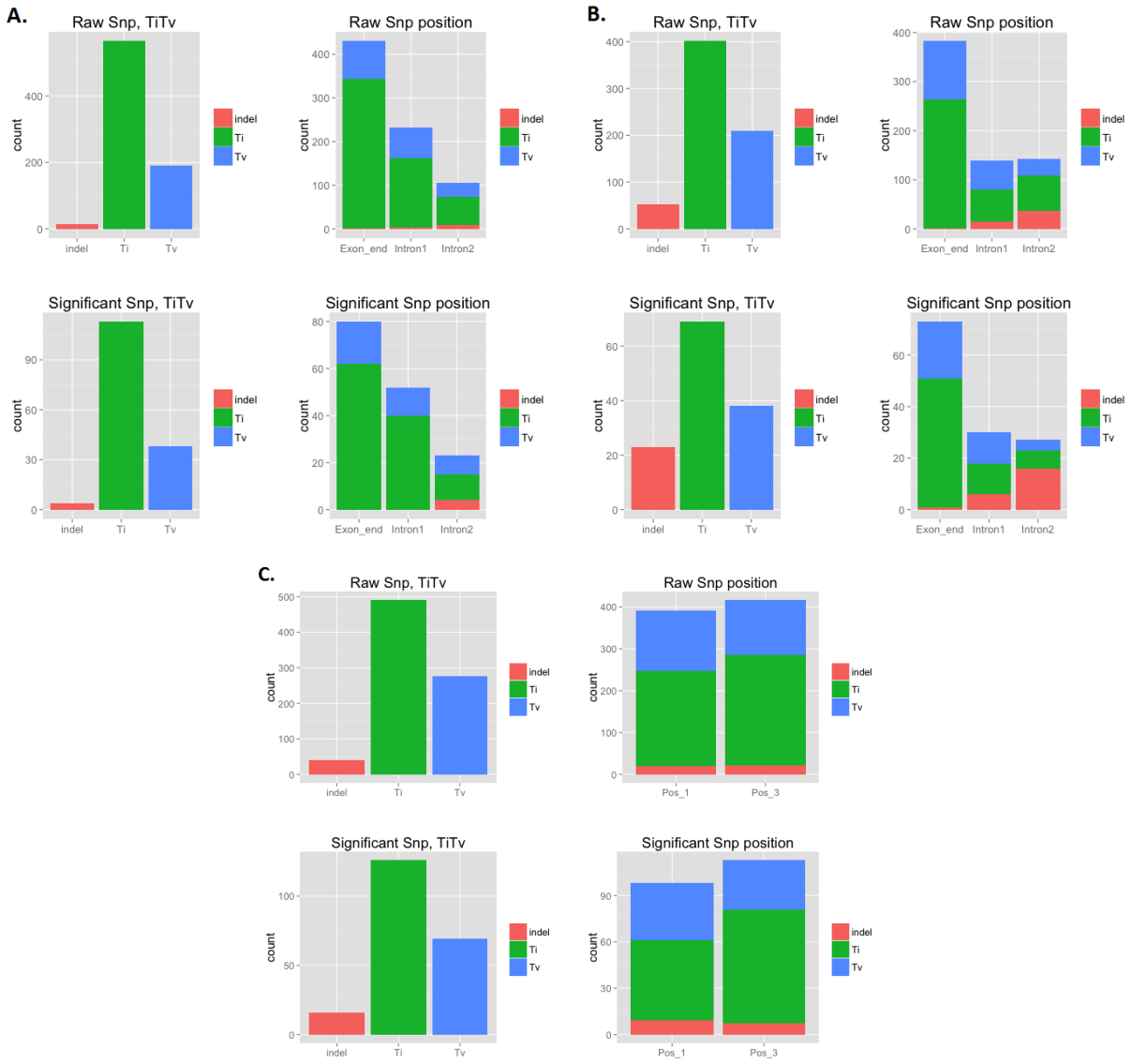

Figure 4.11: Splice site and branchpoint bar plots for both expression filtered (Raw) and significant P value filtered data. Three bar plots show proportion of transitions (Ti) ,transversions (Tv) and indels at each position relative to splice site. (A) 5' splice site variants (B) 3' splice site variants (C) Branchpoint variants.

score compares each splice site against a model created using all known splice sites. Functional splice sites are generally scored above 5. Score distributions were calculated for all variants at 3' and 5' splice sites and are shown in Figures 4.13 and 4.14.

While the distribution shapes are similar it is clear 3' splice sites appear more narrow. This effect could also be due to the difference in model efficiency at capturing nucleotide differences at 3' splice sites. The bimodal distribution of the difference (red) indicates the propensity of variant changes to have very

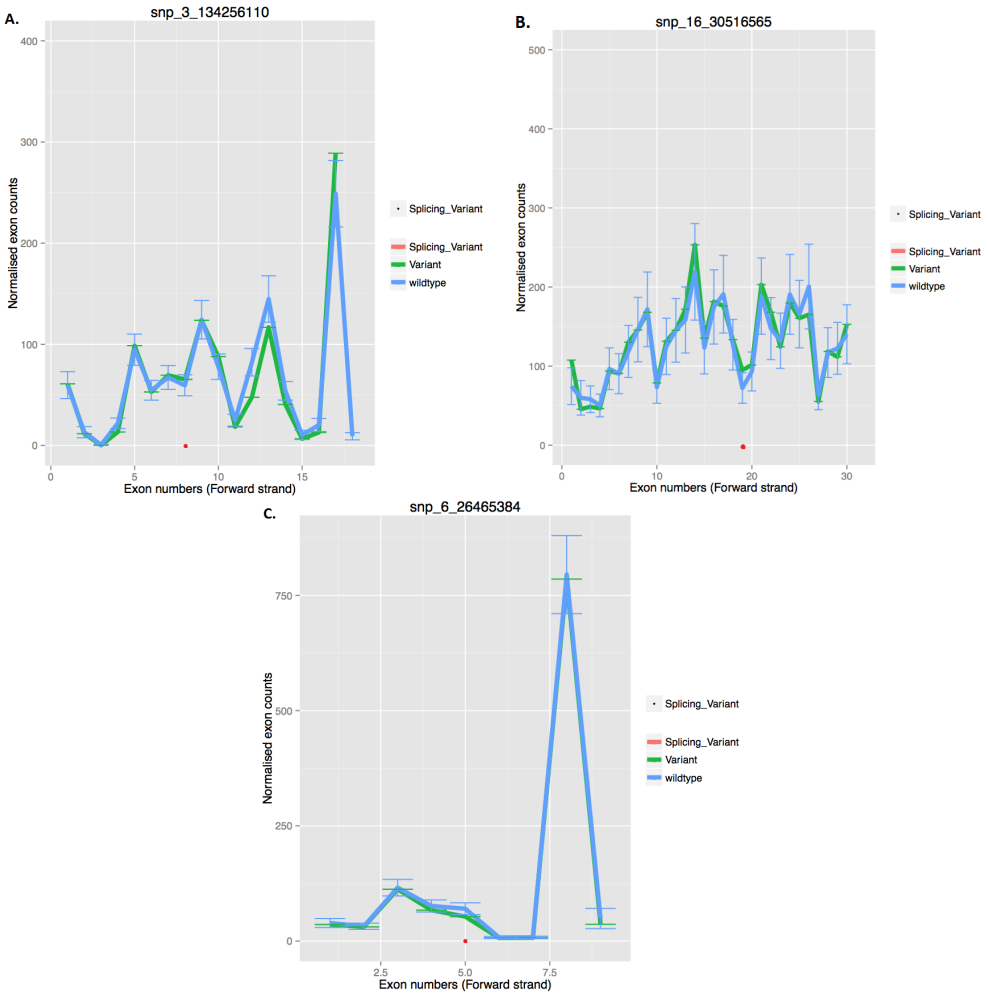

Figure 4.12: Exonic expression for three mutations in core motifs of (A) 5' Splice site (B) 3' Splice site and (C) Branchpoint. Each graph represents exon expression across the length of the gene with consecutive exons arrange from 1^{st} - last exon on the x-axis and normalized exon read count on the y-axis. Wildtype and variant samples are summarized by the blue and green line respectively. Standard deviations are included for each exon. Exon containing the splicing variant is highlighted by a red point on the x-axis.

little (or no) effect or a strong effect.

The score distribution for branchpoints calculated using a position weight matrix is show in Figure 4.15. Although this does show a slight bimodal trend it is clearly far less specific. Figure 4.15 B shows the decomposition of this distribution into the first position (U) and 3^{rd} position (A) of the motif. It is interesting that the distributions indicate a clear difference in score effect. This can be explained by the prominence of the central "A" in the majority of branchpoints. A change to this nucleotide would have a drastic impact on the score. This could be an effect of sampling bias, as all current techniques rely on the central adenine to anchor the branchpoint motif.

The difference between the wildtype and variant scores allowed quantification of motif deviation. It was clear that splicing should be affected based on the variant score even though exon expression showed no effect. The accuracy of this difference score was then tested by looking at splice junctions.

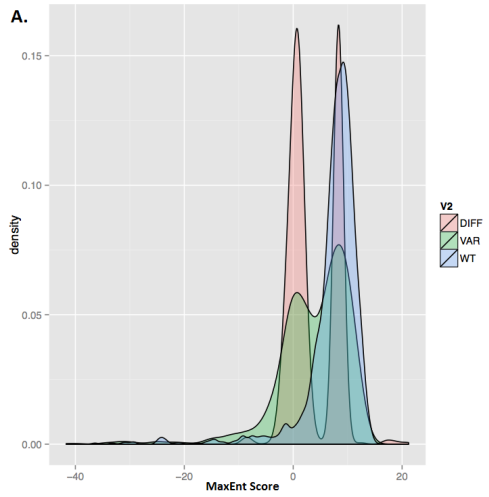

Figure 4.13: Distribution of 3' splice site scores for wildtype (blue), variant (green) and difference (wildtype - variant) (red) categories. MaxEnt score shown on the X axis respectively.

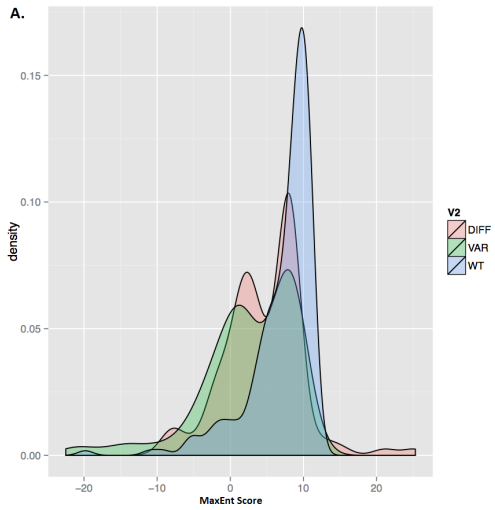

Figure 4.14: Distribution of 5' splice site scores for wildtype (blue), variant (green) and difference (wildtype - variant) (red) categories. MaxEnt score shown on the X axis.

Variant impact on splicing efficiency

In order to investigate the effect on splicing four statistics were created for each splicing feature (5' , 3' and branchpoint) (please refer to Table 4.2 and Figure 4.7). There are several common effects presented here.

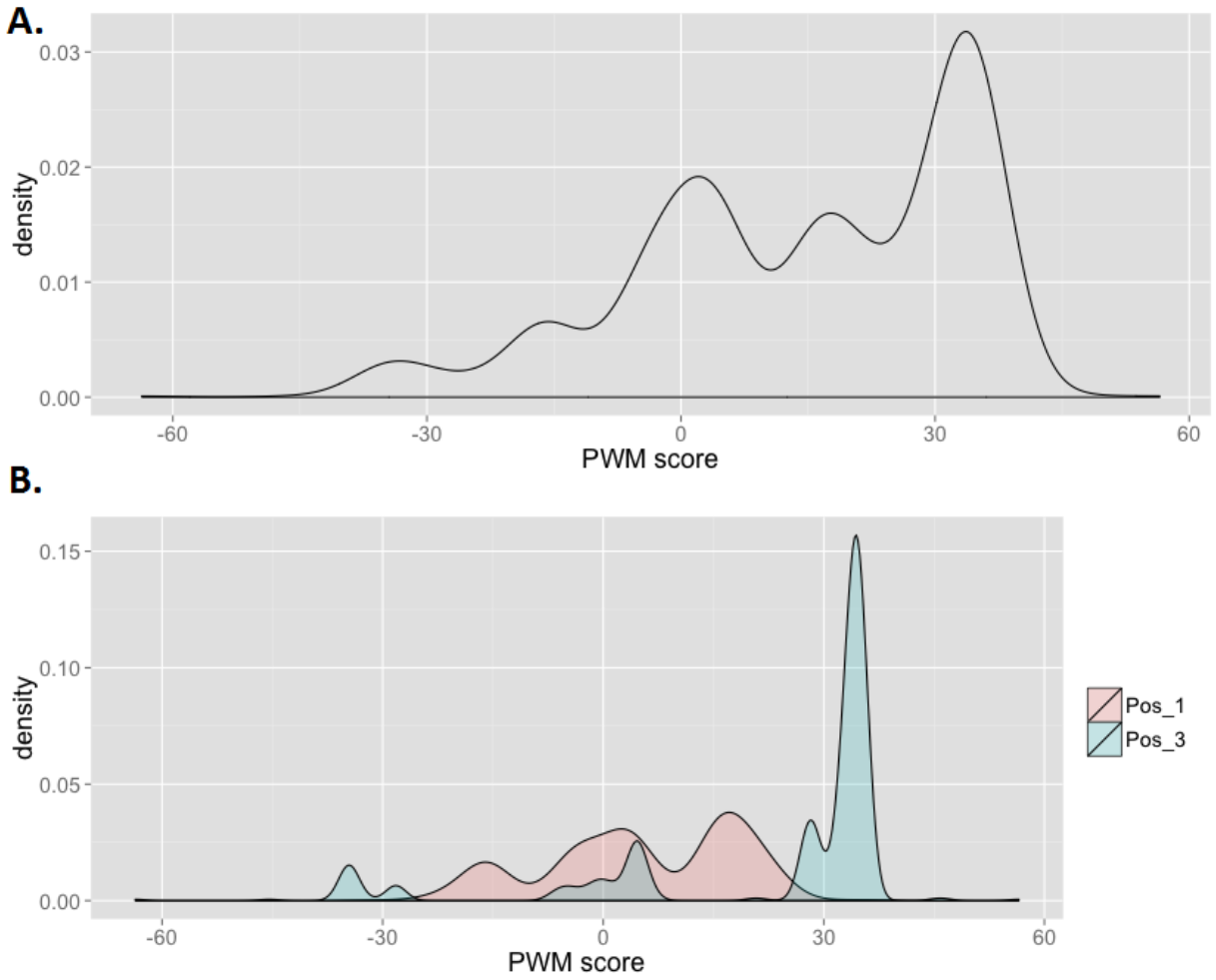

Figure 4.15: A. Distribution of Position weight matrix scores for Branchpoint variants. MaxEnt score shown on the X axis B. Distribution of Position weight matrix scores for both Position 1 (U) and Position 3 (A) branchpoint variants in red and green respectively. MaxEnt score shown on the Y axis.

Overall, in cases where variation has an impact on splicing a dosage effect is clear between homozygous, heterozygous and wildtype. Heterozygous splice site mutation appears to result in a 25-50% change from wildtype. Variant impact on branchpoint mutations appear to be far more muted. This indicates that splicing machinery can recover more efficiently from these changes. Exonic expression is rarely as accurate or distinguishable as the splicing statistics.

Splicing variation decreases efficiency

All splicing features show evidence of variation significantly decreasing splicing efficiency. It is also worth noting that high difference scores for 3' splice site mutations tended to result in a 3 (or multiple of 3) base

pair shift into the exon (Figures 4.16 and 4.17). This points to a potential rescue mechanism that keeps the transcript in-frame.

5' splice site variants that reduce splicing efficiency tend to be compensated by increased expression of alternate exon starts (Figure 4.18), exon skipping and intron retention (Figure 4.19) or exon extension (Figure 4.20).

Branchpoints also show a minor (5-15%) but significant reduction in splicing efficiency (Figures 4.21 and 4.22). However, in one case the dosage effect was more striking, resulting in a 25% and 50% reduction for heterozygous and homozygous respectively (Figure 4.23). This resulted in the creation of an alternate, novel 3' splice site within the exon.

Splicing variation improves efficiency

In rare cases a variant change at a splice site seems to significantly increase its efficiency thereby promoting splicing in an otherwise unused exon. This phenomenon is present in both 3' (Figure 4.24) and 5' splice sites (Figures 4.25 and 4.26) leading to selective use of a single splice site over another (both isoforms are present at equal levels in the wildtype) and inclusion of cryptic exons respectively.

Correlation of score to splicing efficiency

In order to investigate the effectiveness of the variant effect scores a linear regression analysis of the scores against each splicing statistic was undertaken. For this analysis only those variants that showed highly significant (P value < 0.00005) differences between wildtype and variant groups were investigated. This was necessary as variance within unfiltered results was overwhelming due to technical and biological noise.

Figures 4.27 and 4.28 show the linear regression for each statistic versus score for 5' and 3' splice sites respectively. All statistics are significantly (P value < 0.05) correlated with score. In all cases the upstream splice site (UPSTR) and shifted canonical ratio (JAR) statistics performed best. This indicates that, as expected, a significant proportion of splicing variation can be explained by the difference in sequence.

A similar analysis was done for the branchpoint score but none of the statistics showed significant correlation or r-squared above 0.01. This indicates the mutability of sequence is not a major consideration in the majority of cases, and that several unknown contributing factors are involved such as multiple branchpoints per intron.

The above analysis was repeated to determine if variant frequency, like score, can predict splicing efficiency. There is no correlation between damaging splice variants and minor allele frequency (see Figures

4.29 and 4.30). For both 3' and 5' splice sites no significant correlation or r-squared > 0.06 was obtained. This implies that MAF cannot be used to predict splicing pathogenicity in a similar way as other deleterious variation [Rivas et al., 2015]. It is possible that splicing mutations are incredibly rare and may be selected against regardless of effect. Alternately, selection may not operate on these rare, neutral mutations and thus cannot be distinguished from functional mutations on allele frequency alone.

The distribution of data in figures 4.27 and 4.28 is L-shaped indicating a skew in the P values. This is possibly due to the bimodal/non-normal tendency of the distribution. For further assurance, each statistic was bootstrapped a 1,000 times with replacement to achieve an average r-squared. In order to get a measure of sensitivity a leave one out cross validation was done on the data. This is summarized in Tables 4.4 and 4.5. Surprisingly, all statistics appear to be sensitive with the exception of the variant exon. This highlights the lack of sensitivity given by exon expression. This could be due to the inconsistent effects to the variant exon depending on the type of recovery mechanism the cell uses. For example, the inclusion of cryptic elements or subtle splice site shifts is unlikely to result in much change.

Splice statistic	Mean grad	Grad-lowerCI	Grad-upperCI	Min grad	Max grad	Final variants
UPST	0.314925627	0.306135018	0.323716235	0.217787552	0.393927925	46
JAR	0.492104854	0.480462679	0.503747029	0.390713733	0.584601649	20
EOI	0.376046549	0.361683527	0.39040957	0.156183516	0.534347488	24
VE	0.028974493	0.026476701	0.031472284	0.008900012	0.093380182	53

Table 4.4: Cross validation of linear regression on Splicing statistics for 5' splice sites. Slope of regression is correlated to score to determine r-squared. EOI: A ratio of JunctionA/all junctions from the variant exon, UPST : A ratio of JunctionA/all junctions from the neighbouring junction exon ,VE: Read count of the last 100bp of the variant exon, normalized by total mapped reads ,JAR: Shifted junction ratio A ratio of JunctionA/(JunctionA + shifted junctions). Columns in order are; Splicing statistic, Mean gradient, Gradient (lower confidence interval),Gradient (upper confidence interval), Minimum gradient,Maximum gradient and number of final variants used.

Splice statistic	Mean grad	Grad-lowerCI	Grad-upperCI	Min grad	Max grad	Final variants
UPST	0.668573045	0.663165639	0.673980452	0.586792755	0.738014867	30
JAR	0.349764635	0.335178276	0.364350994	0.145883398	0.630156662	17
EOI	0.375049471	0.36380803	0.386290912	0.264749377	0.509700092	18
VE	0.051390732	0.048977608	0.053803856	0.030615054	0.081498466	41

Table 4.5: Cross validation of linear regression on Splicing statistics for 3' splice sites. EOI: A ratio of JunctionA/all junctions from the variant exon, UPST : A ratio of JunctionA/all junctions from the neighbouring junction exon ,VE: Read count of the last 100bp of the variant exon, normalized by total mapped reads ,JAR: Shifted junction ratio A ratio of JunctionA/(JunctionA + shifted junctions). Columns in order are; Splicing statistic, Mean gradient, Gradient (lower confidence interval),Gradient (upper confidence interval), Minimum gradient,Maximum gradient and number of final variants used.

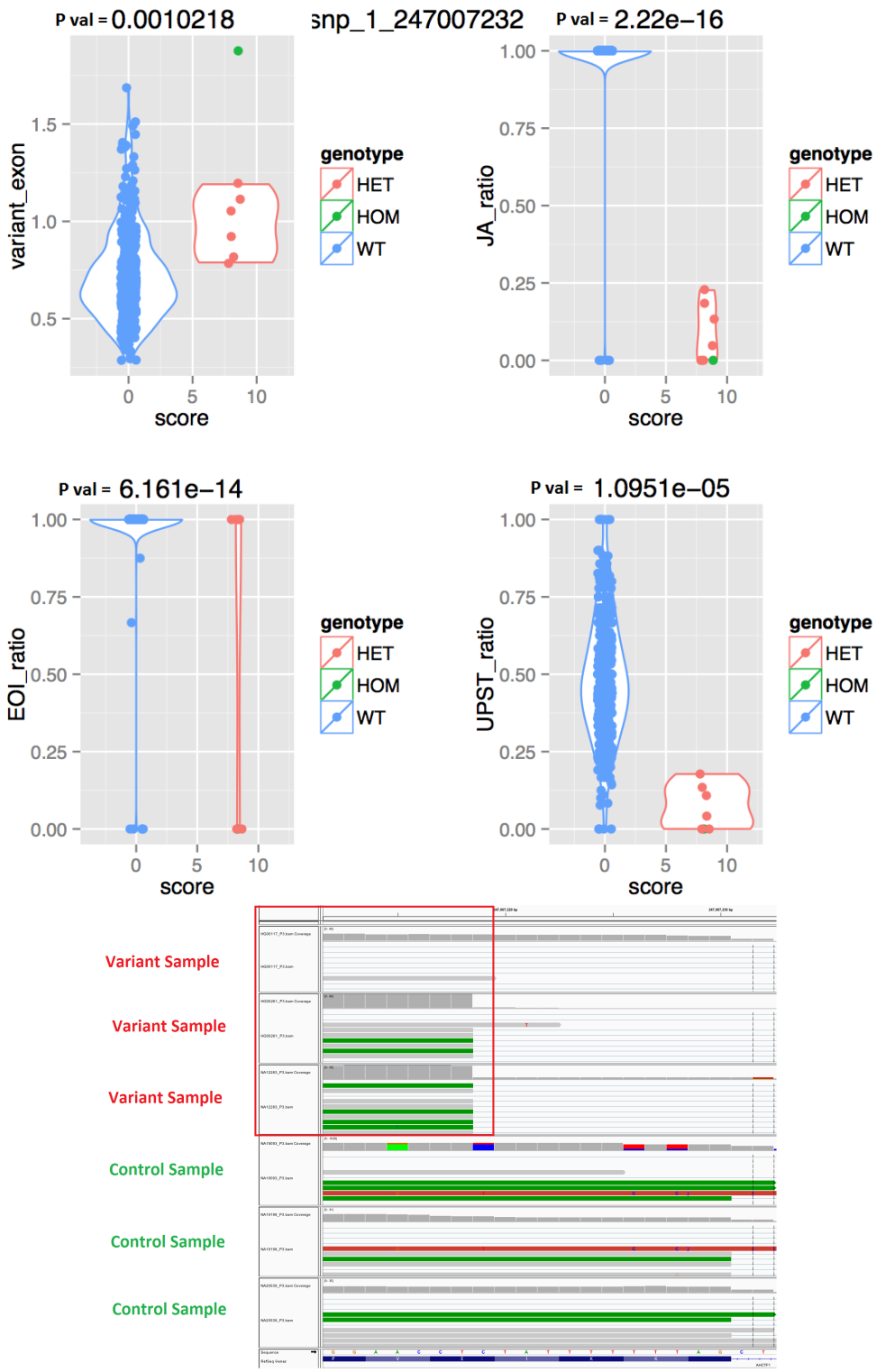

Figure 4.16: Splicing variation decreases efficiency. A. 3' splice site variant significantly decreases canonical splicing (UPST ratio) and creates a shifted junction (JA ratio). B. This results in a shifted exon start by 12 nucleotides.

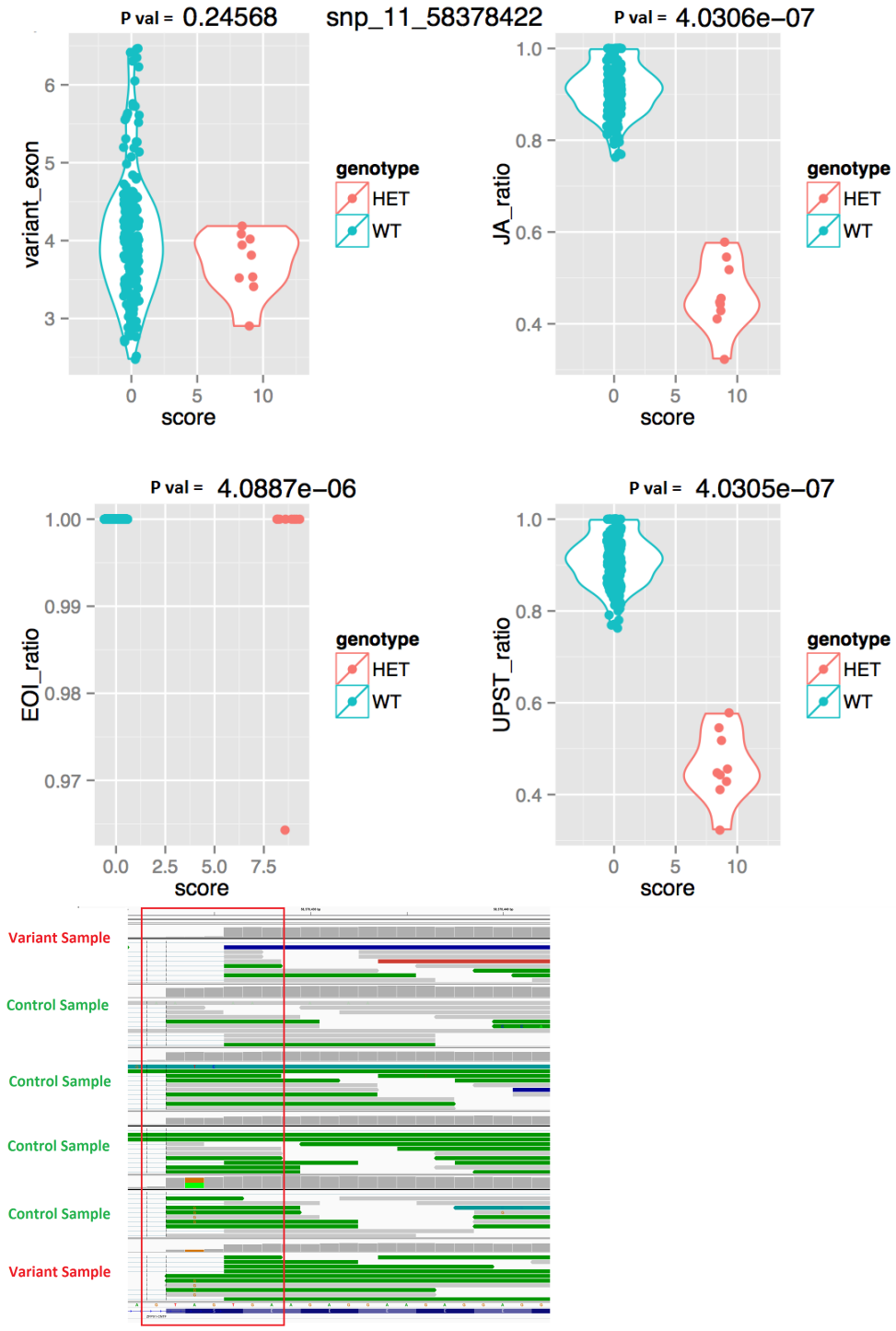

Figure 4.17: Splicing variation decreases efficiency. A. 3' splice site Variant significantly decreases canonical splicing (UPST ratio) and creates a shifted junction (JA ratio). B. This results in a shifted exon start by 3 nucleotides.

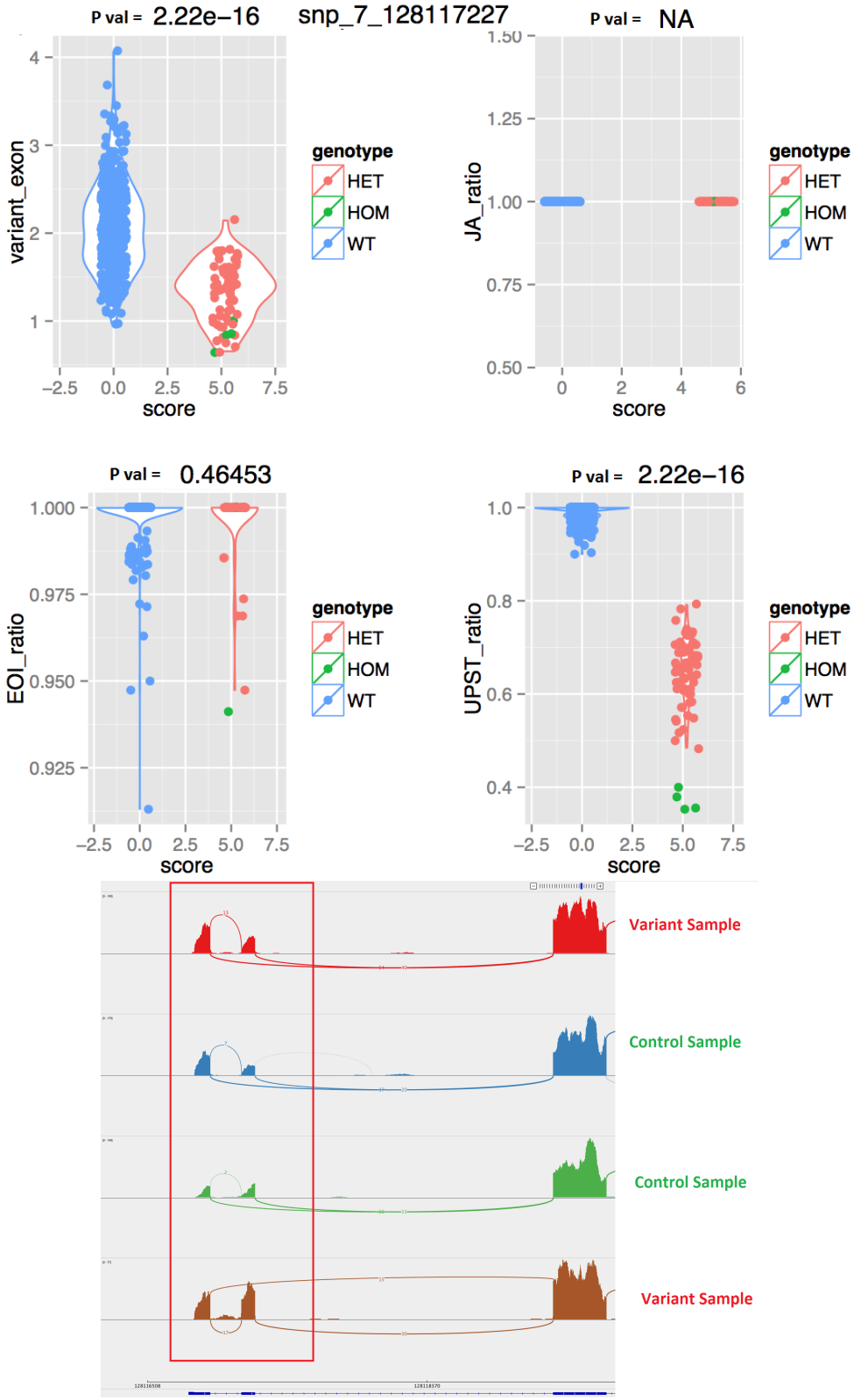

Figure 4.18: Splicing variation decreases efficiency. A. 5' splice site Variant significantly decreases canonical splicing (UPST ratio) and expression of the affected exon (variant exon). B. This results in an increase of alternate junction expression.

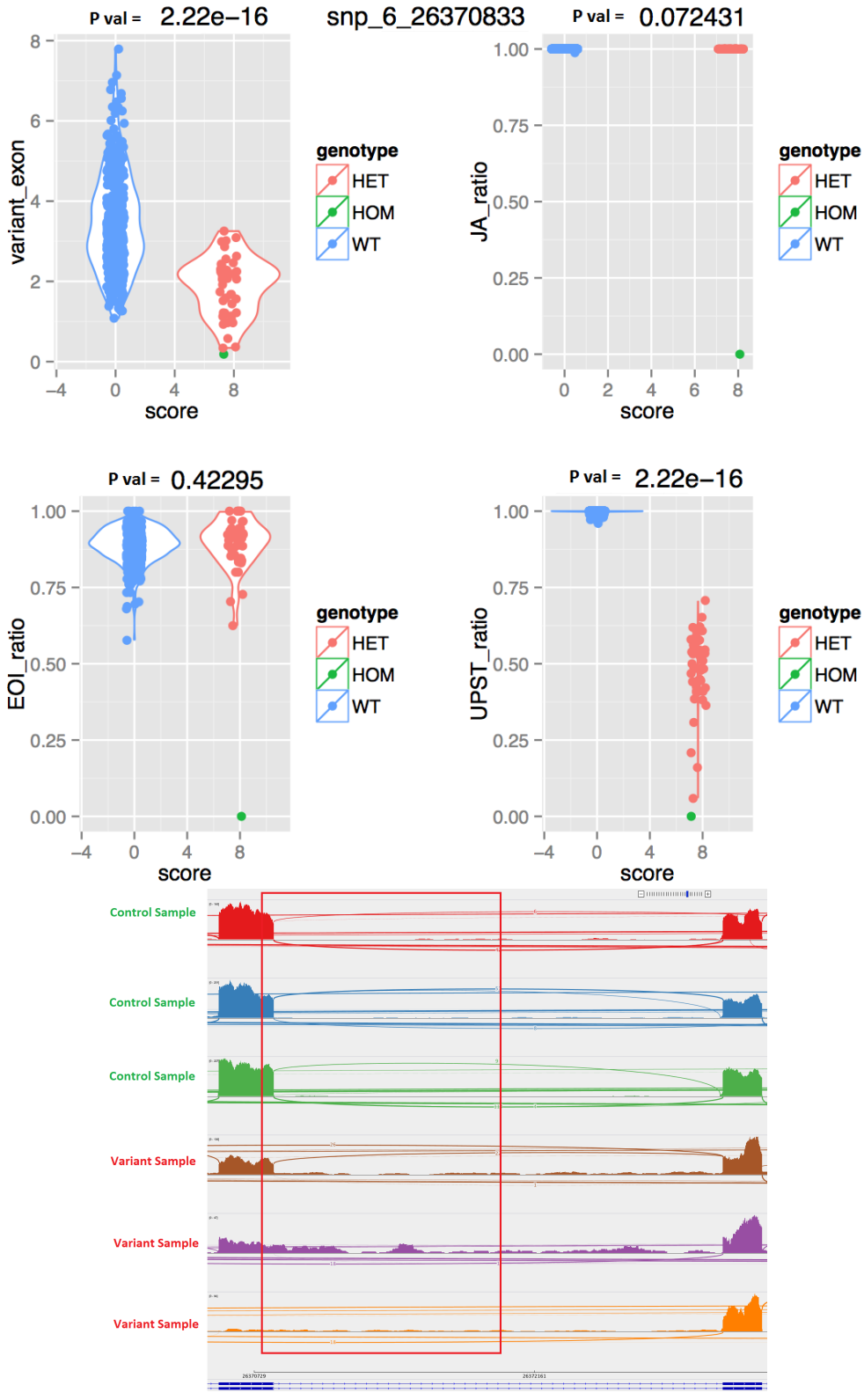

Figure 4.19: Splicing variation decreases efficiency. A. 5' splice site Variant significantly decreases canonical splicing (UPST ratio) and expression of the affected exon (variant exon). B. This results in intronic retention and complete loss of splicing.

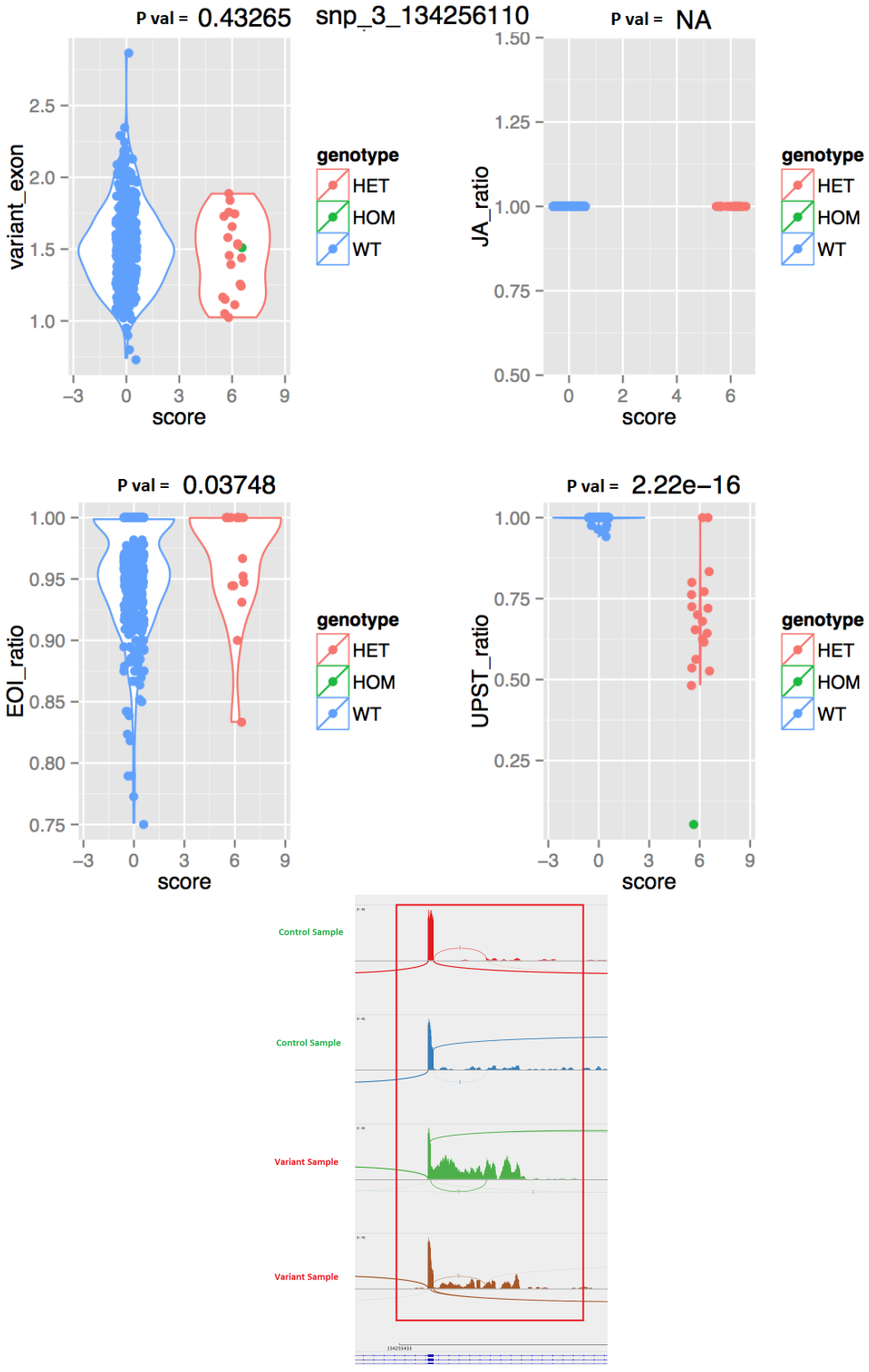

Figure 4.20: Splicing variation decreases efficiency. A. 5' splice site Variant significantly decreases canonical splicing (UPST ratio). B. This results in exon extension and loss of splicing.

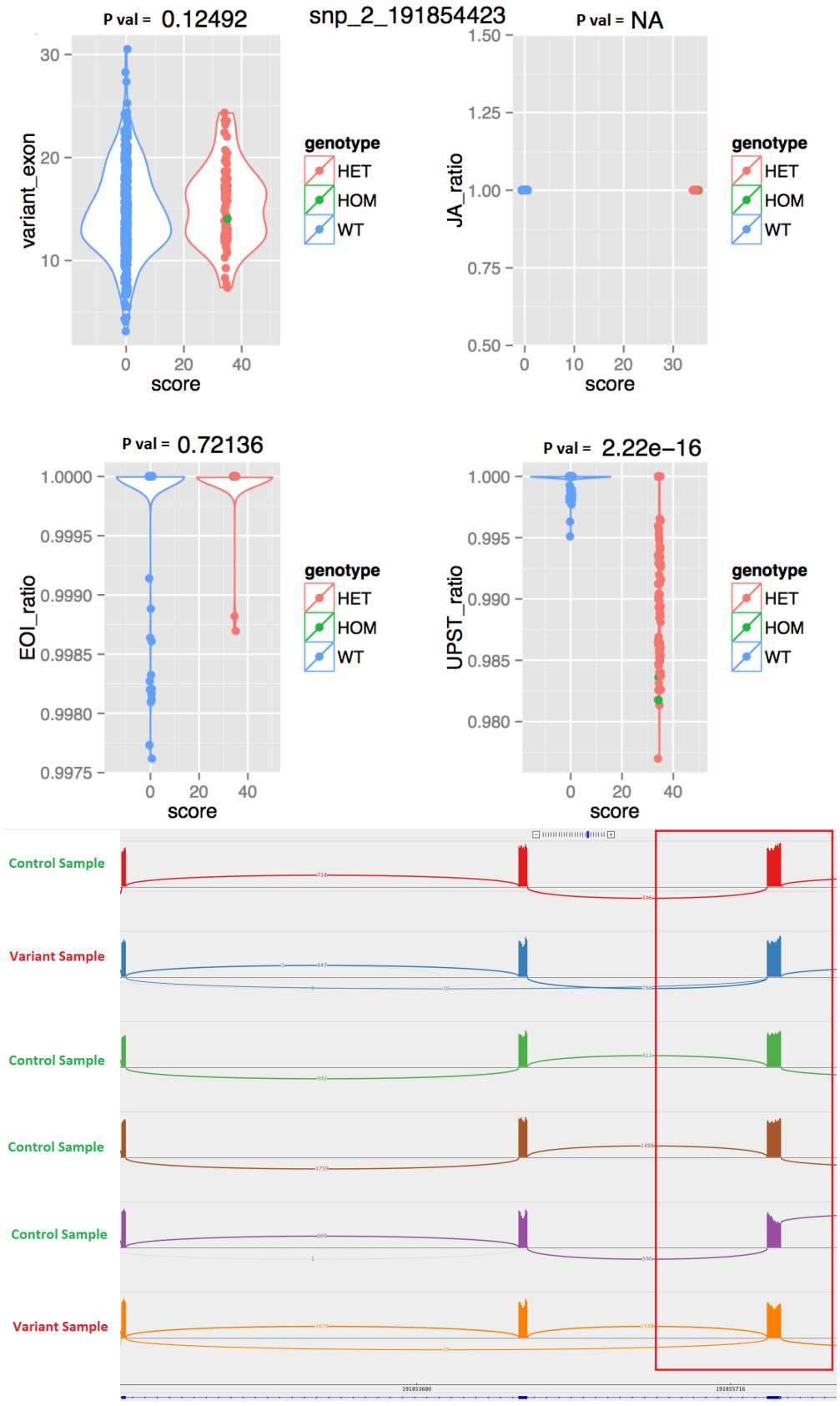

Figure 4.21: Splicing variation decreases efficiency.91A. Branchpoint variant slightly decreases canonical splicing (UPST ratio). B. This results exon skipping in a small percentage of cases.

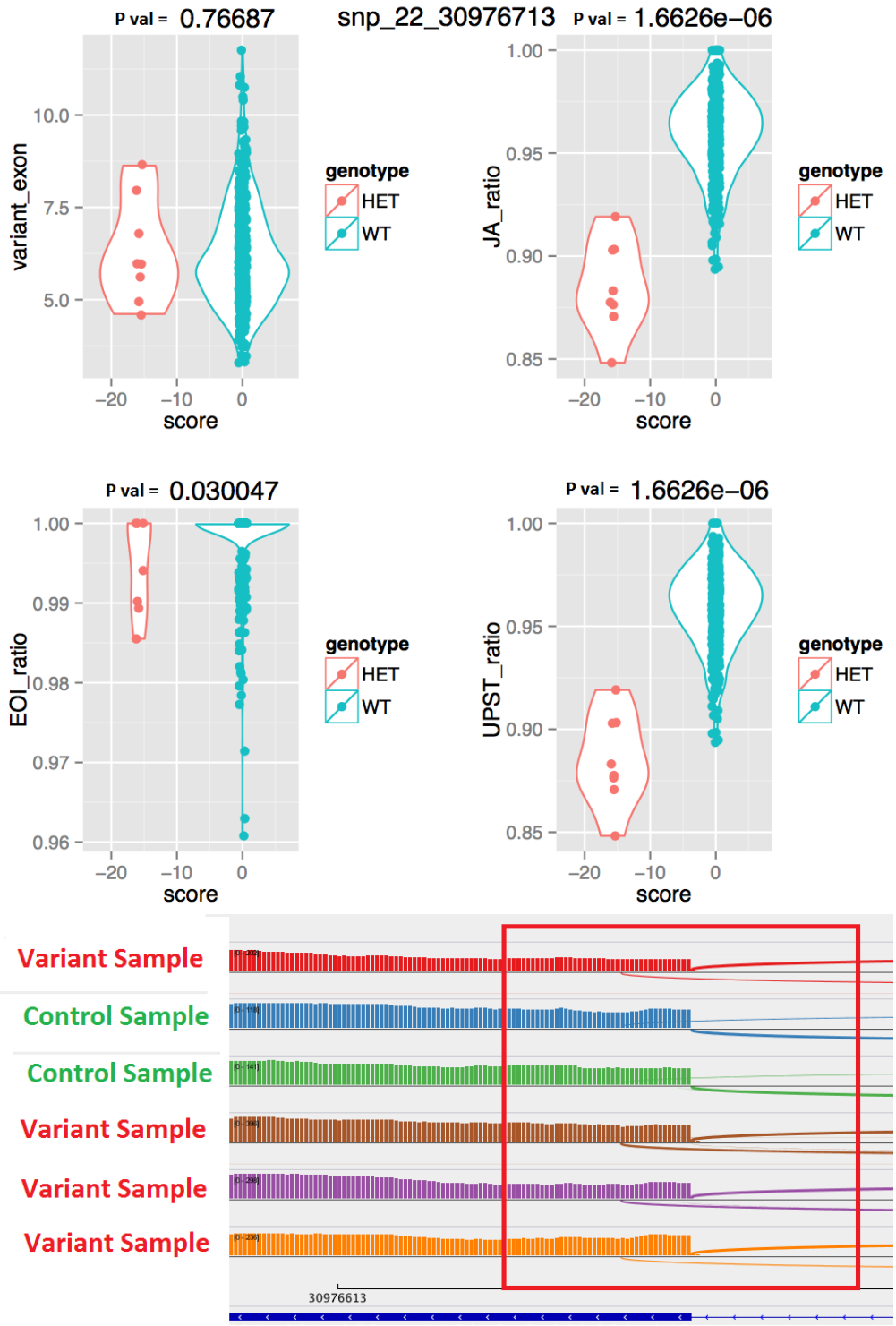

Figure 4.22: Splicing variation decreases efficiency. A. Branchpoint variant slightly decreases canonical splicing (UPST ratio) and shows an increase in 3' splice site shifts (JA ratio). B. This results in use of cryptic 3' splice site in a small percentage of cases.

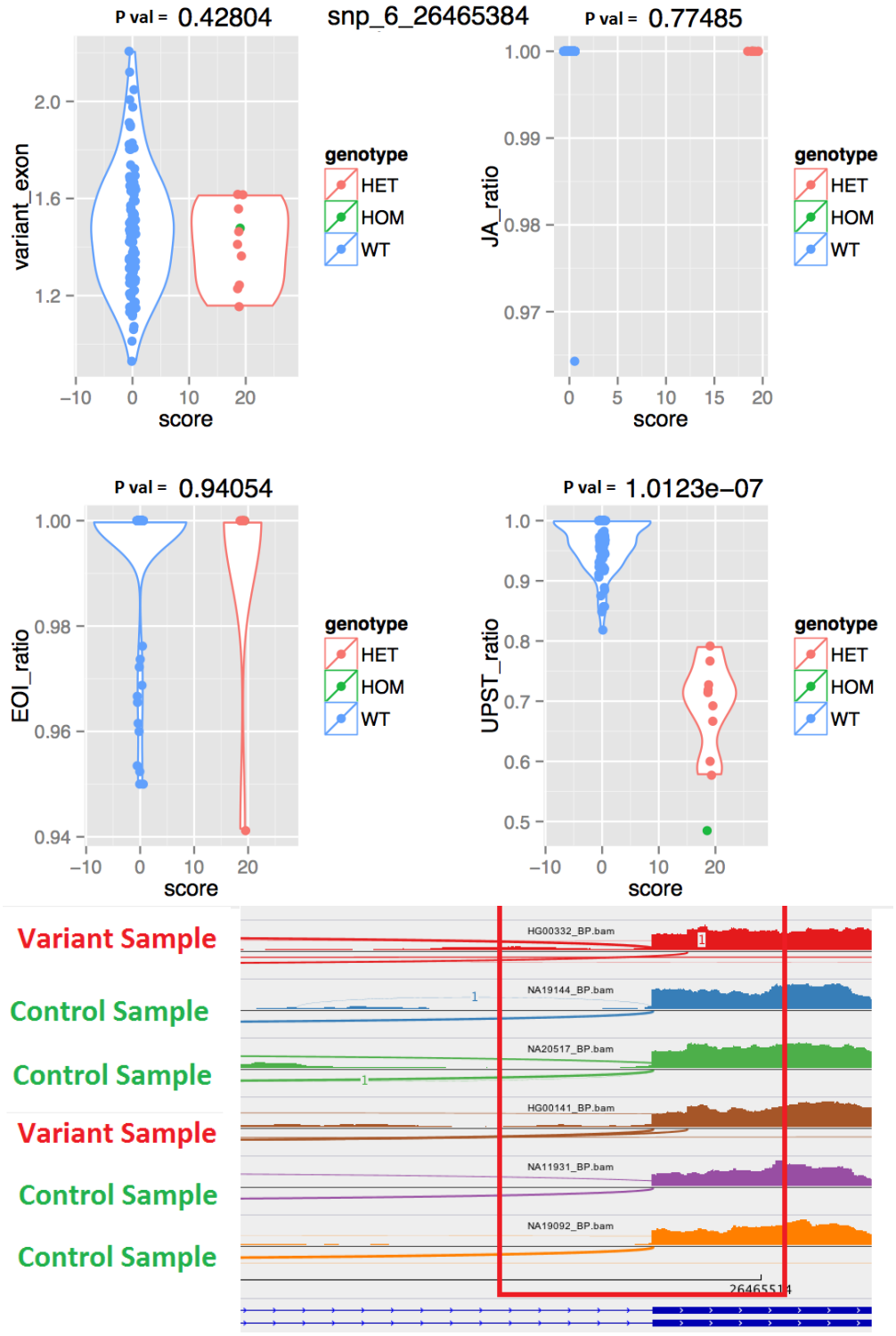

Figure 4.23: Splicing variation decreases efficiency. A. Branchpoint variant slightly decreases canonical splicing (UPST ratio). B. This results in use of the primary splice site, removing an alternate 3' splice site.

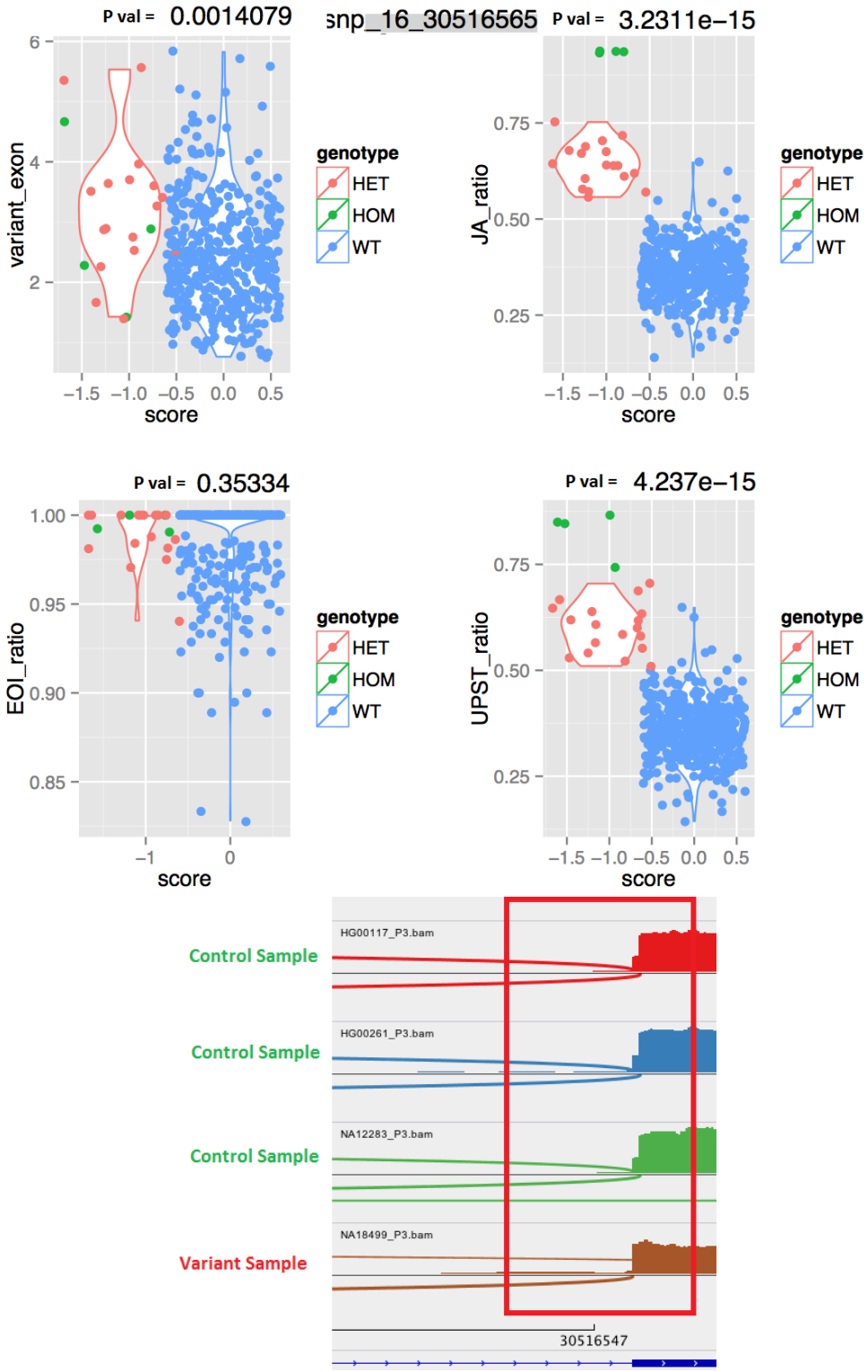

Figure 4.24: Splicing variation improves efficiency. A. 3' splice site variant increases canonical splicing (UPST ratio,JA ratio). B. This results in preferential splicing to this location.

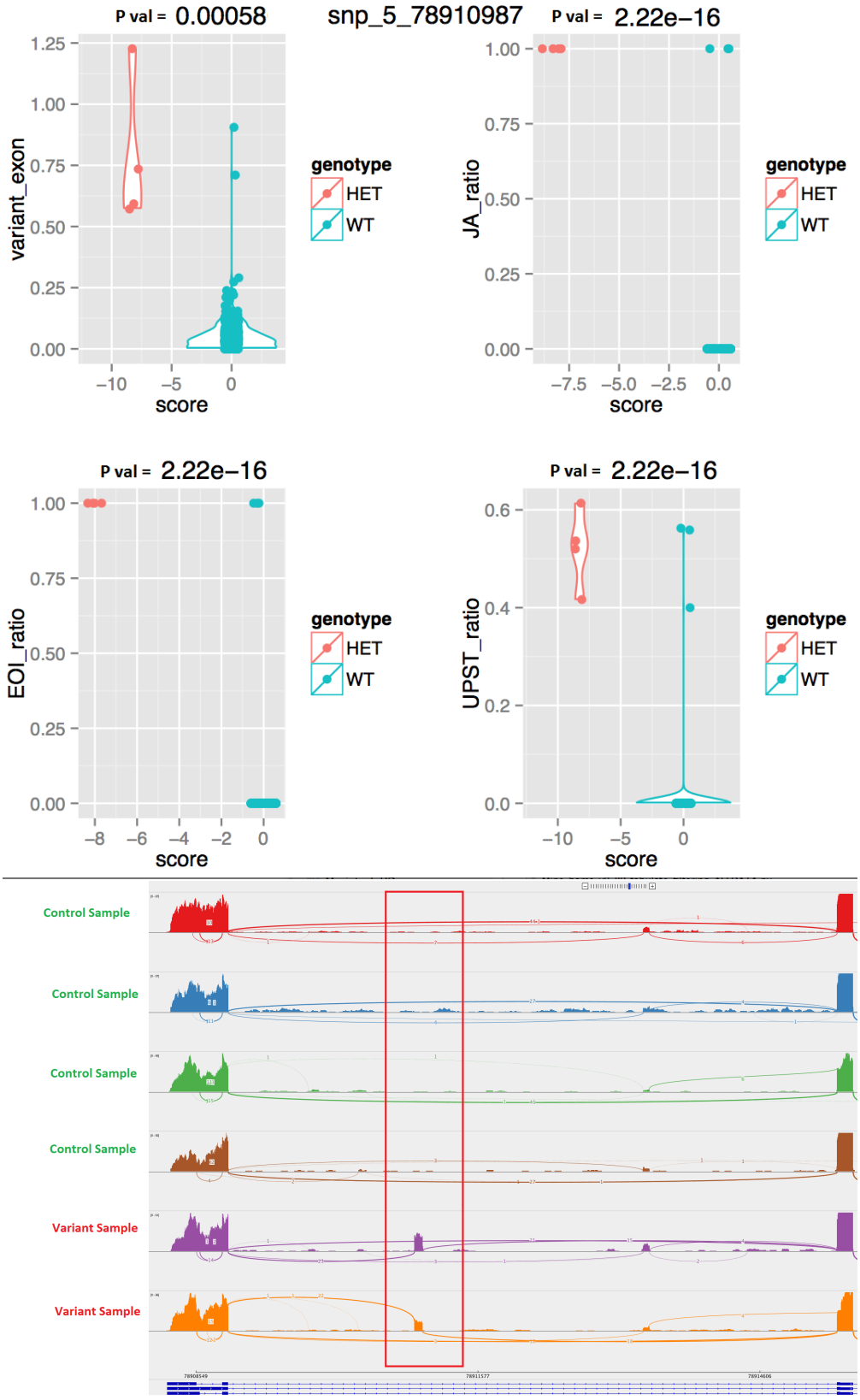

Figure 4.25: Splicing variation improves efficiency. A. 5' splice site variant increases canonical splicing (UPST ratio,JA ratio). B. This results in inclusion of an alternate exon not present in wildtype.

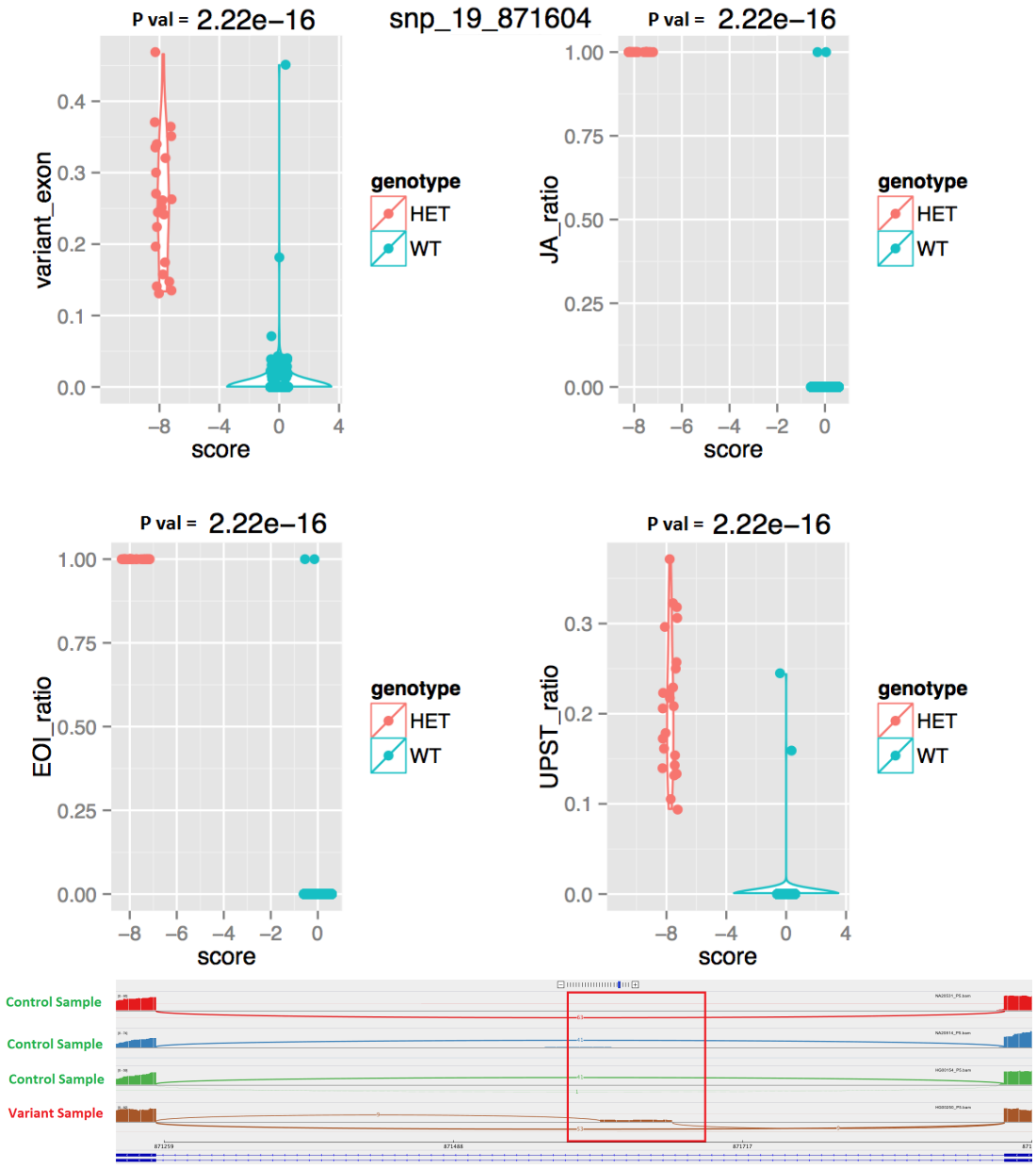

Figure 4.26: Splicing variation improves efficiency. A. 5' splice site variant increases canonical splicing (UPST ratio,JA ratio) and expression of an alternate exon (variant exon). B. This results in inclusion of an alternate exon not present in wildtype.

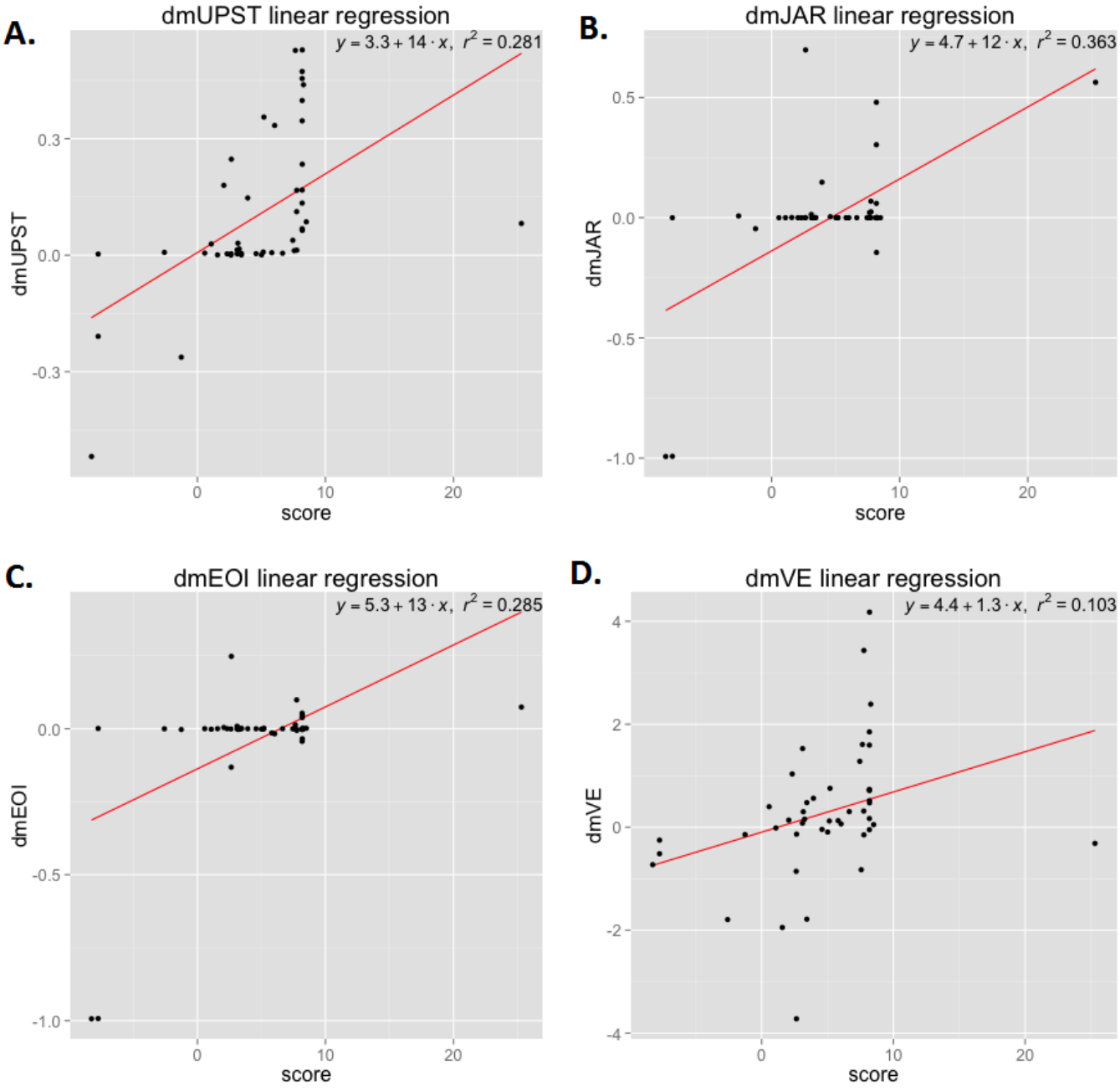

Figure 4.27: Linear regression of 5' splice splice site variant score against difference in means (wildtype group sample mean - variant group sample mean) for; A. Upstream splice site B. Upstream splice site (including only shifted junctions) C. Variant exon splice site ratio D. Variant exon shore.

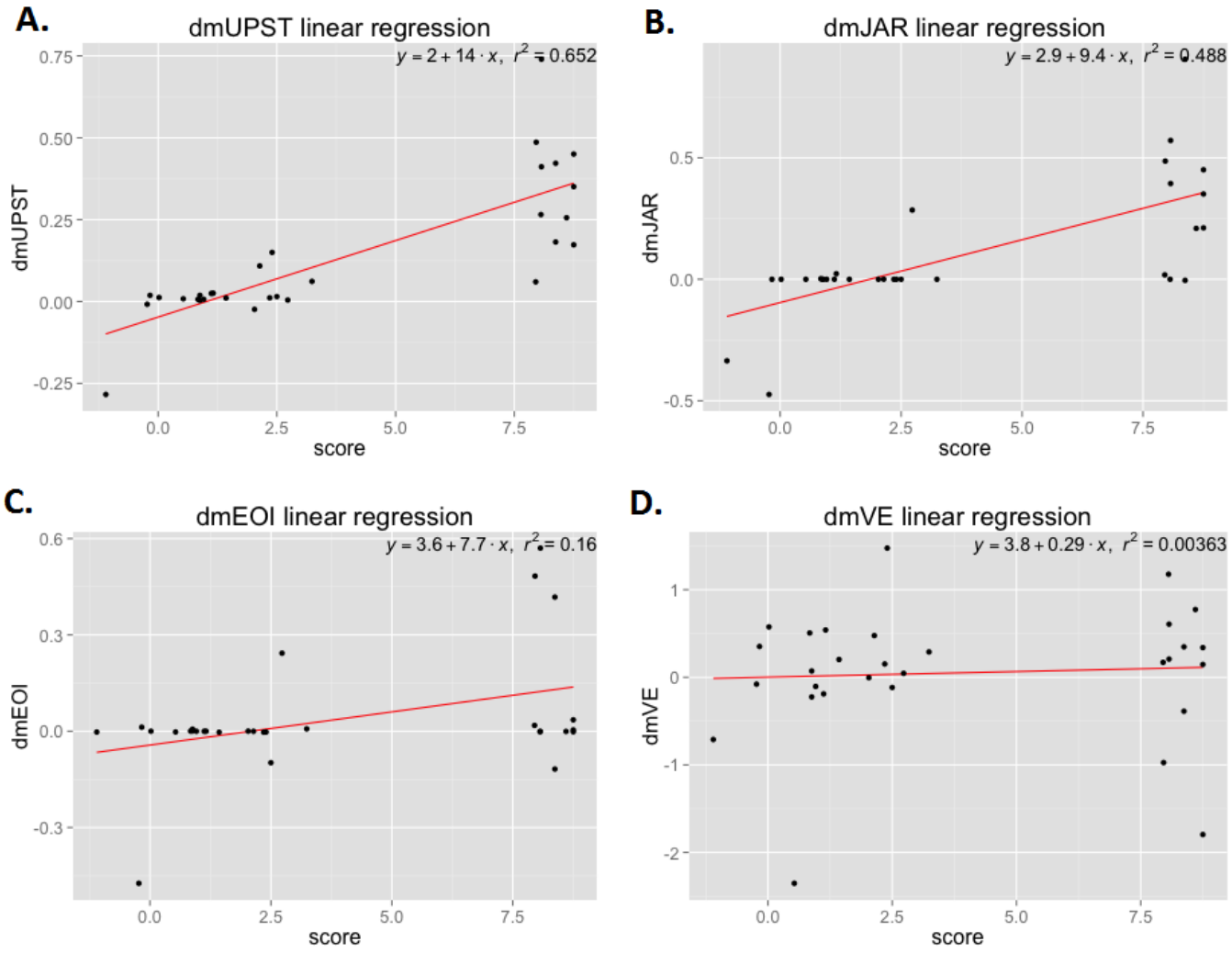

Figure 4.28: Linear regression of 3' splice splice site variant score against difference in means (wildtype group sample mean - variant group sample mean) for; A. Upstream splice site B. Variant splice site (including only canonical and shifted junctions) C. Variant exon splice site ratio D. Variant exon shore .

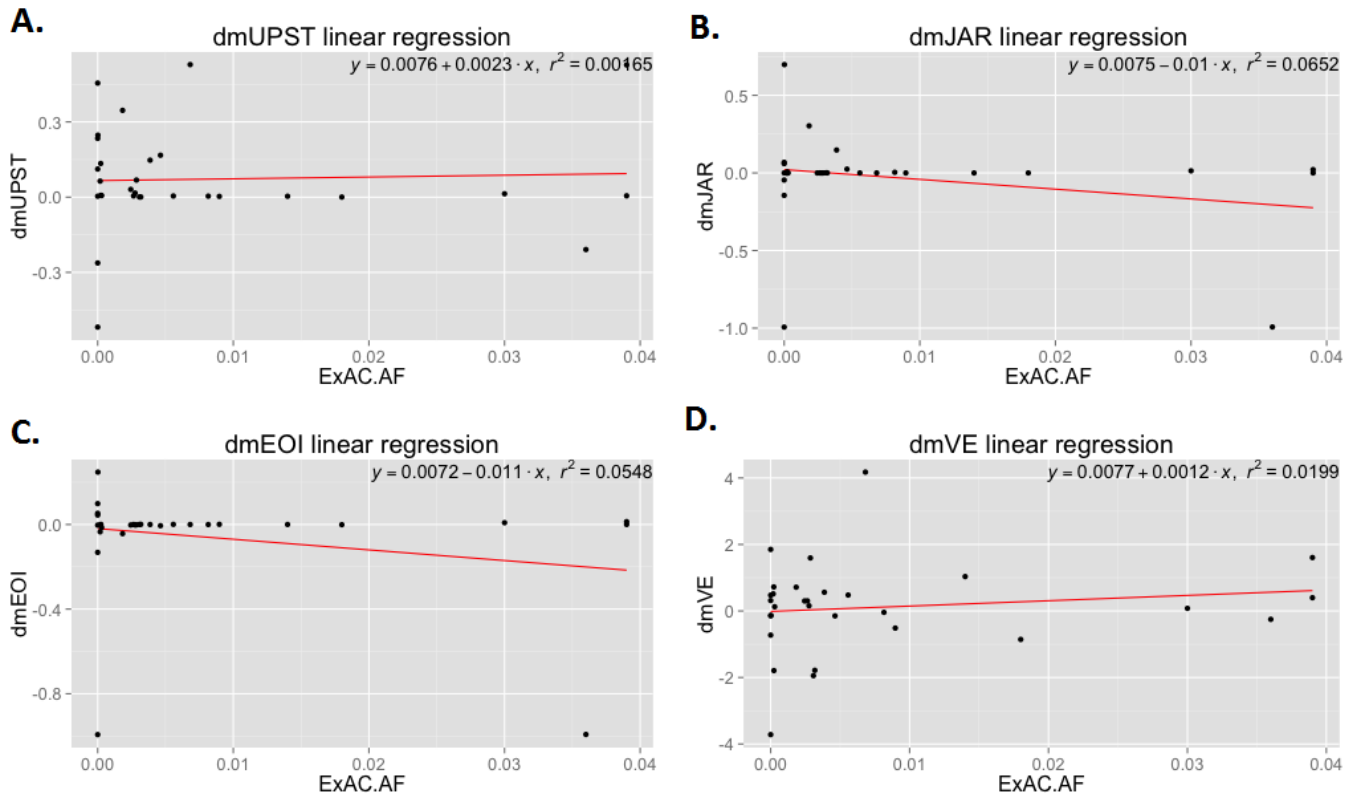

Figure 4.29: Linear regression of 5' splice site variant frequencies from ExAC against difference in means (wildtype group sample mean - variant group sample mean) for; A. Upstream splice site B. Upstream splice site (including only shifted junctions) C. Variant exon splice site ratio D. Variant exon shore .

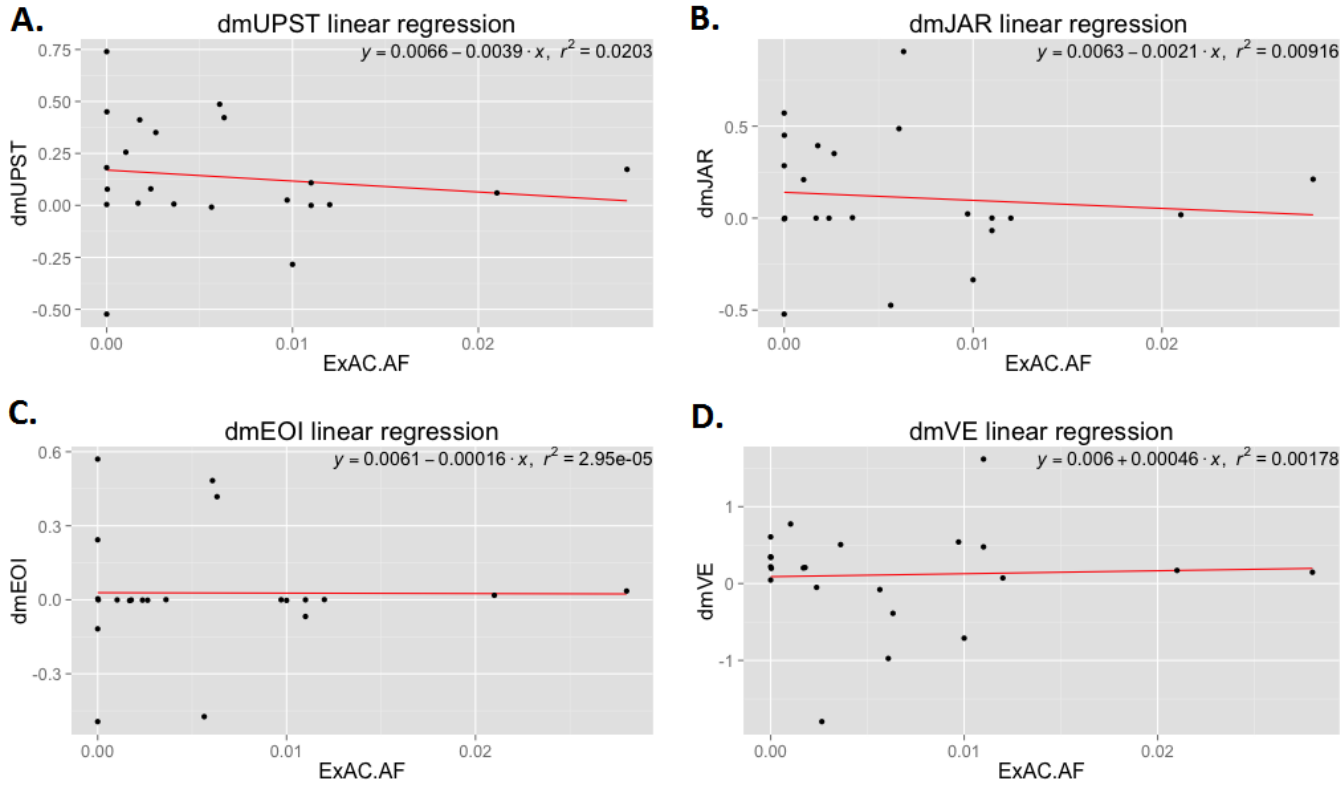

Figure 4.30: Linear regression of 3' splice site variant frequencies from ExAC against difference in means (wildtype group sample mean - variant group sample mean) for; A. Upstream splice site B. Upstream splice site (including only shifted junctions) C. Variant exon splice site ratio D. Variant exon shore .

4.4 Discussion

4.4.1 Variant ratio graphs are a novel method to annotate human specific features

Here I outline an approach to identify 'conserved' genomic positions within splicing features using the frequency of polymorphisms across many individuals. Having successfully identified the expected conservation of the first two intronic splice site nucleotides I have expanded to show that branchpoints also show a pattern of conservation centring on the U and A positions consistent with the literature. This has lead to the identification of over 400 rare, branchpoint mutations in two large consortia which can be investigated further.

The caveat of this approach is the requirement of precise location for genomic features. In order for variant graphs to highlight important nucleotides they must overlap precisely. Using currently available data, a single genomic location cannot be explored in the same way as phylop/GERP species conservation can. This technique provides a unique opportunity to explore nucleotide constraint as a complementary approach to classic species conservation.

4.4.2 More data are necessary to model branchpoints effectively

Branchpoints are emerging as important features that, when disrupted, have a measurable impact on splicing. This study identifies instances where branchpoint variation results in significant (albeit not drastic) change to splicing. Currently, there is not sufficient information to build a successful model for variant effects on branchpoints. Most noticeably, sequence variation appears to be higher than splice sites even at invariant positions. This could indicate that splicing machinery can compensate more effectively to rescue these effects.

Branchpoints may be far more mutable for several reasons; there could be multiple "rescue" branchpoints per intron or other splicing signals (i.e. polypyrimidine tract and 3' splice site) may be sufficient to compensate for weak motifs. At least 40% of branchpoints remain unidentified, this number is likely an underestimate due to lack of knowledge about tissue specific branchpoints. Furthermore, a number of branchpoint variants are missed as they are outside the exome capture range (generally exome baits only capture 50bp around an exon). The use of whole genome sequencing may alleviate this by improving relevant intronic variant information. Lastly, both studies that have identified the largest percentage of branchpoints tend to bias the branchpoint location toward the nearest adenine residue. This could mean that several

sites have been mis-annotated and reinforces the need for further discovery and high-throughput validation methods.

4.4.3 Splice junctions accurately define splice changes

After investigation of exon expression in genes with splicing variants it became clear that the majority had no visible effect. This resulted in two hypotheses; splicing variation often has no effect or the effect is not captured by exon expression. A pipeline was designed for the quantification of splice junctions which showed striking differences in splicing efficiency as well as various splicing compensation mechanisms such as triplicate nucleotide shifts at 3' splice sites.

A score was created to predict the divergence caused by variants from the wildtype splicing feature. This proved effective for splice sites but not for branchpoints. It is clear the distribution of scores for both 3' and 5' splice sites appear to be bimodal, having either a negligible or drastic effect on the splicing motif. However, the score still requires a continuous scale to capture the direction of effect. Interestingly, variation can also enhance splicing efficiency. This effect, although rare, could be just as pathogenic, resulting in the inclusion of cryptic elements thereby creating aberrant transcripts.

4.4.4 Minor allele frequency is not a predictor of splice site pathogenicity

On average the motif score explained roughly 40% of variance. Unmeasured effects such as exonic enhancers/silencers, RNA binding proteins and epigenetic environment (i.e. prevalence of histone marks, methylation) is not be captured. The focus of the study on the nuclear splicing motif is intentionally strict to reduce false positives resulting from altering these features.

This analysis is made possible by significant advances in the field [Lappalainen et al., 2013; Rivas et al., 2015]. These studies showed the power of integrating variation with gene expression. However, their hypothesis that allele frequency can be used to determine deleterious effect does not hold true for splicing variation. Based on current data, minor allele frequency is not an effective predictor of splicing pathogenicity. These mutations are incredibly rare and likely innately selected against. This may change with future expansions of exome consortia if enough rare variation can be captured.

4.4.5 Technical and biological variation impact data quality

Although GEUVADIS is a breakthrough study it does suffer from technological and computational biases. This is particularly evident in gene expression data. The sequenced data was produced on older, less robust

Illumina machines at lower depth with shorter reads. This increases the level of missingness and reduces usable data. Another consideration is the older algorithms used for alignment and variant calls. The recovery of splice junctions is dependent on both the algorithm employed and the length of the reads. Alignment algorithms do not fully recover these reads, especially those with short overhangs. Furthermore, novel junctions are easily missed if read coverage or gene expression is not high enough to pass noise thresholds. For the exome variant data; several insertion/deletions within the data were clearly low quality or incorrectly called and were excluded from this analysis.

Compound variation or multiple proximal SNPs may result in false positive associations as reported by Rivas et al [Rivas et al., 2015]. Although not explicitly tackled by this analysis, is very rare (a single occurrence was noted) and has a minimal effect at locations used for this study.

A large source of variation, exacerbated by lack of read coverage, is the variable expression of several genes across the cell population. This makes distinguishing lack of signal from lack of expression challenging. In order to compensate only variants that showed appreciable change between variant/wildtype groups were selected to estimate predictive value of the variant score.

4.4.6 Variant splicing score captures significant splice junction change

Multiple statistics were generated to describe splicing efficiency. All statistics used a canonical junction to define a ratio of canonical splicing over other, potentially aberrant splicing. Statistics were centred at the splice site where the variant occurs or at the opposite splice site described by the canonical junction. Interestingly, the highest concordance is different between 3' and 5' junctions. 5' splice sites appear to be best described by including shifted junctions (JAR 66%) in the vicinity of the splice site. 3' Splice sites are best described by their upstream canonical splice site (UPST 49%). Potentially the exon may be skipped or a cryptic event occurs within the intron that is not proximal to the original splice site and is not accounted for by the shifted junction statistic.

The poor correlation of variant exon expression to splicing change is truly striking. On average it shows a 2-3 fold weaker correlation than splice junction statistics. This indicates that a great deal of change is currently being missed by focusing on approaches that look at exonic/isoform expression alone.

4.4.7 How the current study compares to recent publications

The recent study by Rosenberg et al. [Rosenberg et al., 2015] attempted to predict variant effect based on variation in and around splice sites. They report similar concordance results based on sequence changes at

splice sites from GEUVADIS data (concordance of 40% from MaxEnt, their software, HAL achieves 60%). Although MaxEnt is outperformed it still achieves an almost equivalent accuracy on heterozygous SNPs ; 81.7% vs 87.1%. This is impressive as HAL is trained on a vast, custom designed library of minigenes.

Rosenberg et al. apply MISO [Katz et al., 2010] to quantify splicing in GEUVADIS data. Although this is a particularly robust tool it does not quantify cryptic events accurately. MISO focuses on known isoforms and expected exon skipping events. This is a major disadvantage as the majority of splicing mutations generate cryptic splice sites which will not be measured. This likely accounts for the higher concordance found in this study as all splice junctions were included in the analysis. Furthermore, HAL includes surrounding exonic and intronic variants making direct comparison difficult.

The pipeline designed here does not require access to a website and can be run very efficiently on thousands of SNPs, potentially as part of a standard annotation pipeline for exome data. Any interesting results from this initial analysis could then be selected manually and validated using the HAL online software.

4.4.8 Optimization is essential for reproducibility of this analysis

A significant part of what made this study feasible was computational optimizations. Analysis of over four hundred BAM files is challenging for several reasons. Firstly, the amount of space required for processing is excessive, well over 2TB just to host the data. This was circumvented by selecting only regions that contained variants of interest across the entire cohort. This was necessary to reduce processing times and storage requirements. Optimizations in terms of analysis focused on running steps in parallel and producing large hash tables (very efficient data structures in Python) to avoid reprocessing unnecessarily. The end result is an analysis which took just over 3 days of wall clock time to process several terabytes worth of raw files. And could be rerun (excluding initial processing and downloading) within hours.

4.4.9 Conclusion

These observations serve as a fundamental starting point for further work into the improved annotation of splicing variants. This will allow for efficient detection of effects on splicing, including currently unexplored instances where variant changes can create novel splice sites/cryptic exons.

Integration of further data may drastically improve variant score prediction. Inclusion of epigenetic environment such as prevalence of histone marks and methylation in a cell type specific manner will further increase correlation. Defining a robust score can be used to rank splicing variation and identify mutations that create novel splice sites within exons or introns. These variants remain largely undetected and result in

various disease phenotypes such as Autosomal recessive bestrophinopathy and X-linked retinitis pigmentosa [Davidson et al., 2010; Webb et al., 2012].

Chapter 5

Conclusion

5.1 Core features and central concepts of the thesis

5.1.1 Splicing as the primary force behind species diversity

Splicing has been known as an integral part of species diversity for over thirty years [Chow et al., 1977; Berget et al., 1977]. It shows the complexity that can be reached without the need to increase raw gene numbers and provides a remarkable way to enhance cellular complexity and transcriptional control. Here, I look at splicing from multiple angles to show we have only scratched the surface of this exciting and dynamic process.

Splicing can be a mechanism to control transcript integrity through the inclusion/exclusion of poison exons. This mechanism, known as recursive splicing, shows how splice sites can be reconstituted by other splicing reactions and splice site competition can play a vital role in guiding the cellular machinery. I then explored how the structure of co-transcribed RNA can create circular molecules through a backsplice junction. This creates a further layer of complexity, showing how RNA structure plays a crucial role in splicing processes. Lastly, I looked at the three core features of the splicing reaction, namely; 5' and 3; splice sites and the branchpoint. The branchpoint has been poorly characterised although it has been shown to be disease causing. I explore the impact of variation at these features, demonstrating that the functional effect of sequence variation can be predicted to some degree. Ultimately, the diversity of splice site strength is crucial to correct cellular function, disruptions of these features can result in deleterious effects within the gene and pathology within the cell. This thesis forms the foundation for further exploration into the dynamics of splicing and the cells unique ability to utilise this mechanism in multiple ways.

5.1.2 Splice junction reads are key to sensitive gene expression analysis

Splice junctions are a key concept and central to sensitive gene expression analysis. The power of detecting splicing events with nucleotide precision has long been underestimated and overlooked. A striking realization from this study is that exon expression is indeed a poor substitute for splice junction analysis. It is 2-3 fold less sensitive than splicing statistics and cannot distinguish subtle splicing changes than can have a drastic effect on the final transcript. Ideally, analysis should integrate expression and splicing in a meaningful way, early attempts (such as Cufflinks [Trapnell et al., 2012] and MISO [Katz et al., 2010]) still miss much of the most interesting non-canonical aspects of splicing. This thesis aims to contribute to the field of RNA processing by demonstrating uses of alignment tools and standard bioinformatic techniques to exploit splice junction reads.

5.1.3 Predicting damaging variation on splicing

MaxEnt [Yeo and Burge, 2004] has been the splice site scoring software used for more than ten years now. This striking fact indicates that sequence clearly plays a substantial role in predicting RNA processing. This tool has been used extensively in this thesis and is recognized as state of the art [Rosenberg et al., 2015]. Investigating variant effects on splicing remains largely unexplored. Several recent large studies [Xiong et al., 2014; Rosenberg et al., 2015] have focused on evaluating alternative splicing by considering all variants in and around these exons without directly answering the question: what effect do variants have on the central splicing motif at a single site? This question was addressed here and shows that each splice site can be evaluated independently and functional effects can be predicted and used to guide annotation. The question that remains is how much of the predictive power of the recent tools (produced by aforementioned studies) can be attributed to this. Expansion and inclusion of variants further from the splice site may in fact be confounded with multiple other features and this needs to be systematically investigated.

5.1.4 The evolution of sequencing technology and its impact on splicing analysis

The acceleration of sequencing technology is at the epicentre of recent bioinformatics innovation. It has provided a unique opportunity for creative methods to be developed and has driven bioinformatics to utilise this technology to the fullest. Although a massive improvement over hybridization-based techniques, sequencing still suffers from common drawbacks such as batch effects and bias due to experimental protocol. Another major remaining drawback is read length. Although improvements are continuously being made reads from

the industry leader, Illumina, remain relatively short.

We are now looking ahead to the next (3rd) generation of sequencing by companies such as Pacific Biosciences [Sakai et al., 2015] and Oxford Nanopore [Wei et al., 2016b]. These technologies promise much longer reads at lower cost. This does not come without its complications and currently the estimated error rate of Oxford Nanopore's sequencer is a staggering 38% [Laver et al., 2015]. This can be mitigated to some degree by high coverage sequencing (as these errors are random). As the technology improves however this would likely be an excellent way to retrieve full transcripts, potentially showing knock-on effects of splicing changes that are not recognizable now. Also, in the case of peculiar splicing patterns, this would once again open a new world to identify and expand on non-canonical transcripts. One exciting application would be to inspect the reciprocal splicing patterns identified in Chapter 3.

5.1.5 Data processing as a crucial skill in bioinformatics

Optimisation is an integral part of any successful genomics undertaking and is likely to become even more paramount in future as data volumes continue to increase.

Genomics implies the mining of huge quantities of raw data. This creates unique challenges as computational demands are diverse. The first consideration is hard drive space. In this study the mining of 50 high depth RNA-seq samples (UKBEC brain data) required several terabytes of space and multiple iterations of pipeline development over the entire study. Similarly, mining 426 GEUVADIS RNA-seq samples (processed BAM files alone are over 2 terabytes) required several workarounds to enable efficient querying of data within reasonable wall clock time. Secondly, both cpu and memory demands are high, for instance, STAR sequence aligner requires at least 10 cores and 50 gigabytes of memory per sample to perform optimally.

This brings us to the first level of optimisation which involves processing raw data; efficient read mappers, highly optimized co-ordinate based tools (such as Bedtools) are essential to overcome basic "heavy lifting" of common data types. These steps often require effective use of a cluster compute system for initial alignment, annotation and further coordinate-based manipulation.

However, in pipeline design the interaction of these tools often result in data bottlenecks that need to be individually addressed. Another challenge, especially in exploratory science, is the need for specialised and tailored algorithms and statistics. These are by nature suboptimal as they are often custom code written in a high-level programming language.

Once initial testing has been completed a second step of optimisations must be applied to the pipeline.

Generally this requires compartmentalising pipeline steps and efficient use of data structures. Compartmentalised processes can be optimised independently and parrallelised which greatly increase efficiency. Data structures can allow for saving of complex, preprocessed data that can be recalled when needed. Although this is seldom discussed in detail I do believe it is an essential (and often overlooked) skill to successfully carrying out any genomics project.

5.1.6 Stepping forward, understanding splicing within the context of exon definition

Exon definition is key to splice site recognition. It plays an essential role in recursive splicing, even when said exon is not included in the final mRNA. A natural expansion of this study is to investigate this relationship further, potentially through cryptic exons. Recently Ling et. al linked the presence of cryptic exons to cellular deficiency of RNA binding protein TDP-43 [Ling et al., 2015]. TDP-43 silences and blocks the expression of these spurious exons but in its absence their expression can disrupt mRNA within neurons resulting in severe pathology. Further work could include investigation of RNA binding factor knockdown experiments (currently available on ENCODE) and their impact on recursive and cryptic exon expression.

A natural extension of the current work on splicing variation (in lieu of cryptic exons) is to start modelling variant changes within introns to determine if these locations are likely to form novel splice sites. Secondly, scanning introns for strong splice motifs (both 5' and 3') can predict locations of 'viable' cryptic exons. Ideally this should take into account RBP data (i.e. TDP-43) as this will help annotate sites.

Splicing remains an intricate process. The sheer elegance of a system that functions so efficiently and yet remains robust to even large sequence change is both impressive and daunting. There are clearly still many secrets and hidden mechanisms to the operation of this system. In order to improve our knowledge of splicing and hence, factors that can disrupt its function, we must integrate data. In this study the integration of variant and expression data is a powerful first step to understanding the effect of variation on transcription. It is clear there is much more than can be explained by polymorphism alone and I believe to design an effective tool to predict splicing function we will need to integrate data from other sources such as histone modification, Dnase hypersensitivity, RNA binding protein density etc. For example, it is clear that H3k36me3 will have a pronounced effect on recognition of splice motifs. In cases with high densities of this mark I would need more drastic sequence change to see effects and vice versa.

5.2 Medical Implications

Whole Exome sequencing has been a core part of the expansion of sequencing technology and its impact on clinical genetics. The promise of exome sequencing comes from the prediction that 85% of the disease-causing mutations are located in coding and functional regions of the genome [Botstein and Risch, 2003; Majewski et al., 2011]. This however, translated into a 20% causal variant discovery in cohorts [Yang et al., 2013], which indicates variants are either being missed or much deleterious variation lies outside of the exome.

While this study aims to improve the former (through enhanced variant splicing prediction), the latter has motivated large-scale whole genome sequencing projects such as 100,000 Genomes Project run by Genomics England Ltd. This will provide a huge improvement to splicing-related features such as deep-intronic variation which could create cryptic exons and disrupt branchpoints. A further advantage would be detection of variation at recursive splice sites that could contribute significantly to brain-related pathology. Change here is likely to have a striking effect as the importance of strong splice signals is exemplified by the strong species conservation and lack of transcription histone marks.

The improved understanding of splice site related variation can highlight potentially damaging polymorphism. These variants can now be classified as having either negative or positive effect on splicing; a negative impact on canonical splice sites is well documented as disease causing but the preferential use of an alternative exon or creation of a new splice site (either exonic or intronic) has not been explored extensively although there are documented cases where this variation is pathogenic [Webb et al., 2012].

Certain branchpoint mutations are disease causing [Stenson et al., 2003], however, the ability to predict branchpoint mutation from sequence remains elusive. The degenerate nature of this feature makes it difficult to characterise, largely due to the lack of known sites and lack of variant information at these positions (as they often fall outside exome capture). Whole genome sequencing will greatly help with the latter but primary identification of these sites will still be essential to successfully modelling this interaction. Factors such as histone modifications and chromatin state may play a key role in how specific this motif needs to be and will need to be studied further.

Circular RNA have great potential as influential, non-coding RNA. Their enrichment in neurons (particularly synaptomes) hints at their importance in neuronal function. circRNA also have a future as disease biomarkers as they are enriched in various easily accessible tissues. In order to use them as biomarkers their detection must be as sensitive and robust as possible and this work will contribute to the methods applicable to accurate quantification from RNA-seq data. This in conjunction with circRNA enrichment

techniques such as RNase treatment (to degrade linear RNA) is essential for accurate quantification.

Another unexplored clinical application is to improve understanding of how cancer deregulates cellular machinery. One example would be the androgen receptor gene which is crucial in prostate cancer. In 2008 it was shown that cryptic exons within these genes signify the presence of cancer [Scott M. Dehm, 2008; Hu et al., 2009], these splice isoforms have recently been further expanded [Lu and Luo, 2013; Krause et al., 2014]. The ability to accurately predict potential cryptic exon presence according to mutation of splice sites could be highly advantageous in identification of causal mutation in cancer.

5.3 Further thoughts and future work

5.3.1 Recursive splicing as a genomic mechanism to control promoter usage

The elucidation of recursive splicing, a process whereby an initial splicing step reconstitutes a 5' splice site, allowing for exclusion of a poison exon, will no doubt yield novel deleterious mutations with significant impact on neuronal function. The occurrence of RS in genes with long (150kb+) introns are significantly enriched in the brain and carry characteristic transcription-related histone marks signatures. Taken together, recursive splicing opens the discussion on another level of transcriptional control using the intrinsic power of exon definition and splice site strength to determine exon inclusion.

The use of co-transcriptional patterns to identify splicing (and hence RS) is only effective in long introns. In order to identify gradients in the data you have to bin read counts by at least one kilobase. Other effects such as expression of the gene, number of samples and heterogeneity between brain regions substantially increase noise levels. For these reasons the co-transcriptional pattern is not detectable in shorter introns.

Another concern is the definition of recursive exon. This is partially defined by the lack of an annotated exon. However, recent improvements in annotation show inclusion of some recursive exons as "alternate" exons. This will require further careful definition of what constitutes a recursive exon. Rather than searching for cases outside of known exons the focus could be shifted to whether the cell can effectively use the head-to-head splice sites in different situations.

The further exploration of RS in shorter genes and other tissues is a logical next step. This will require algorithmic improvements relying heavily on splice junctions and superior sample numbers to identify true positive cases. A natural extension would also focus on analysis of splicesosomal RBPs and their impact on these sites. This will likely expand situations in which this mechanism may fulfil different roles.

5.3.2 Circular RNA as novel, brain enriched RNA molecules

Circular RNA has recently received a great deal of attention. These non-coding RNA remain largely shrouded in mystery. The majority have no reported function, they are independently regulated from their linear isoforms, are significantly enriched in brain, detected in saliva, blood and exosomes. Their characterization depends on their backsplice junction. Here I create a sensitive algorithm combining two well documented analysis procedures to maximise and discover large numbers of circRNA within the human brain.

Although circRNA studies have been done in fetal brain and differentiated neuronal cells [Venø et al., 2015; Rybak-Wolf et al., 2015], this is the first, thorough documentation of circRNA in a large human brain cohort. This would explain the large increase in novel circRNA isoforms.

A novel algorithm was created to examine a subclass of backsplice junctions between proximal, highly homologous gene pairs. Findings indicate potential reciprocal transplicing/backsplicing between transcripts. When looking at highly similar genes the attempt to distinguish bona-fide backsplicing from standard splicing is very challenging. Initial results indicate this could be a novel biological mechanism but without laboratory validation it remains inconclusive.

Quantifying circRNA is notoriously challenging. Relying solely on the backsplice junction results in a significant loss of reads (due to the minimum overhang requirement). This can be somewhat circumvented by creating artificial scaffolds to allow for read recovery, although this also increases redundancy within the search space reducing unique mapping reads. Many of these initially identified backsplice junctions appear to be very lowly expressed, it is unclear if these are noise or low-level biological variation. Currently there is no standard on how to determine expression of circRNA, as such I focus on highly expressed examples.

Identification of transplicing/backsplicing between genes will remain controversial until significant laboratory evidence is available. However, this effect can be further explored in different organisms that many share the same gene structures such as mouse and zebrafish. Due to the difficulty in obtaining high quality RNA from human brain, a comprehensive study of such gene pairs could be undertaken in more easily accessible tissue or within cell lines.

The field of circRNA is still developing rapidly. Future studies will surely provide hints to explore in more detail. One future goal would be to explore transplicing/backsplicing across the transcriptome, possibly in combination with Hi-C and CTCF CHIP-seq data (ideally in human brain), to determine whether chromosomal structure plays a role in their formation.

Currently circRNA are being investigated as potential biomarkers. Taking the opportunity to mine the wealth of publicly available data could reveal further disease markers. This could be particularly inter-

esting in cases where linear isoforms do not show any differential expression, as in the case study of BPTF in Bipolar disorder.

5.3.3 Analysis of splicing variants and their impact on transcription

Large exome consortia provide a unique opportunity to use variant frequencies in novel ways. I outline a method to create a nucleotide resolution map of invariant positions within genomic features. I could then explore the effect of these variants on splicing efficiency by integrating gene expression RNA-seq with whole exome sequencing. In order to do this an estimate of the effect of the variant on the splice site was created. This score captures a significant proportion of variance between variants and non-variants. The effect of variation on splicing efficiency can be either positive or negative, both are potentially deleterious to canonical expression. Surprisingly, the majority of variation has no significant effect on splicing indicating the robust nature of cellular splicing.

Integration of RNA-seq and whole exome is still in its infancy, as such this study remains statistically underpowered. Correlations are drawn from small (20-60) numbers of cases (after filtering) which show significant variant effects. The inability to accurately identify true negative cases is a central concern. Generally, the lack of expression of genes and inadequate sequencing depth increases variability leading to high levels of "missing-ness". Although it is clear the variant score captures a component of sequence variation it remains only part of the story.

Ideally future work would require larger RNA-seq and Exome consortia. With the increase in samples it will be possible to accurately classify variant effects, particularly cases with no effect (or marginal impact). This could also allow potential integration of cell specific epigenetic factors (histone marks, methylation) and RNA binding factors. The expansion of variants of interest deeper into splice sites would also be an important step however this will require careful annotation of variants and their predicted overlap with other exonic/intronic features.

5.4 Final thoughts

Together, this thesis hints at the enormous potential of next generation sequencing to propel our understanding of cellular biology through its sensitivity, re-usability and scale. Central to this is the use of publicly available resources which are becoming increasingly abundant as time passes. It is my opinion that in the near future all cellular studies will benefit directly from the wealth of sequence data, guiding experimental

work and testing hypotheses before entering the lab. Soon computational approaches will not stand separately but be an essential step in every biological study informing experimental design as much as final validation.

Bibliography

Naoko Abe, Ken Matsumoto, Mizuki Nishihara, Yukiko Nakano, Aya Shibata, Hideto Maruyama, Satoshi Shuto, Akira Matsuda, Minoru Yoshida, Yoshihiro Ito, and Hiroshi Abe. Rolling Circle Translation of Circular RNA in Living Human Cells. *Scientific Reports*, 5:16435, nov 2015. ISSN 2045-2322. doi: 10.1038/srep16435. URL `http://www.nature.com/srep/2015/151110/srep16435/full/srep16435.html`.

Goncalo R Abecasis, Adam Auton, Lisa D Brooks, Mark A DePristo, Richard M Durbin, Robert E Handsaker, Hyun Min Kang, Gabor T Marth, and Gil A McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, nov 2012. ISSN 1476-4687. doi: 10.1038/nature11632. URL `http://dx.doi.org/10.1038/nature11632`.

N Akula, J Barb, X Jiang, J R Wendland, K H Choi, S K Sen, L Hou, D T W Chen, G Laje, K Johnson, B K Lipska, J E Kleinman, H Corrada-Bravo, S Detera-Wadleigh, P J Munson, and F J McMahon. RNA-sequencing of the brain transcriptome implicates dysregulation of neuroplasticity, circadian rhythms and GTPase binding in bipolar disorder. *Molecular psychiatry*, 19(11):1179–85, jan 2014. ISSN 1476-5578. doi: 10.1038/mp.2013.170. URL `http://www.ncbi.nlm.nih.gov/pubmed/24393808`.

Adam Ameur, Ammar Zaghlool, Jonatan Halvardson, Anna Wetterbom, Ulf Gyllensten, Lucia Cavelier, and Lars Feuk. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nature structural & molecular biology*, 18(12):1435–40, dec 2011. ISSN 1545-9985. doi: 10.1038/nsmb.2143. URL `http://dx.doi.org/10.1038/nsmb.2143`.

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, jan 2010. ISSN 1465-6914. doi: 10.1186/gb-2010-11-10-r106. URL `http://genomebiology.com/2010/11/10/R106`.

Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 31(2):166–9, jan 2015. ISSN 1367-4811.

doi: 10.1093/bioinformatics/btu638. URL http://www.ncbi.nlm.nih.gov/pubmed/25260700http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4287950.

Maria Armakola, Matthew J Higgins, Matthew D Figley, Sami J Barmada, Emily A Scarborough, Zamia Diaz, Xiaodong Fang, James Shorter, Nevan J Krogan, Steven Finkbeiner, Robert V Farese, and Aaron D Gitler. Inhibition of RNA lariat debranching enzyme suppresses TDP-43 toxicity in ALS disease models. *Nature genetics*, 44(12):1302–9, dec 2012. ISSN 1546-1718. doi: 10.1038/ng.2434. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3510335{&}tool=pmcentrez{&}rendertype=abstract.

Reut Ashwal-Fluss, Markus Meyer, NagarjunaReddy Pamudurti, Andranik Ivanov, Osnat Bartok, Mor Hanan, Naveh Evantal, Sebastian Memczak, Nikolaus Rajewsky, and Sebastian Kadener. circRNA Biogenesis Competes with Pre-mRNA Splicing. *Molecular Cell*, 56(1):55–66, sep 2014. ISSN 10972765. doi: 10.1016/j.molcel.2014.08.019. URL http://www.sciencedirect.com/science/article/pii/S1097276514006741.

Jae Hoon Bahn, Qing Zhang, Feng Li, Tak-Ming Chan, Xianzhi Lin, Yong Kim, David T W Wong, and Xinshu Xiao. The landscape of microRNA, Piwi-interacting RNA, and circular RNA in human saliva. *Clinical chemistry*, 61(1):221–30, jan 2015. ISSN 1530-8561. doi: 10.1373/clinchem.2014.230433. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4332885{&}tool=pmcentrez{&}rendertype=abstract.

Yoseph Barash, John A Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J Blencowe, and Brendan J Frey. Deciphering the splicing code. *Nature*, 465(7294):53–9, may 2010. ISSN 1476-4687. doi: 10.1038/nature09000. URL http://dx.doi.org/10.1038/nature09000.

R Belshaw and D Bensasson. The rise and falls of introns. *Heredity*, 96(3):208–213, mar 2006. ISSN 0018-067X. doi: 10.1038/sj.hdy.6800791. URL http://www.nature.com/doifinder/10.1038/sj.hdy.6800791.

Susan M. Berget, Claire Moore, and Phillip A. Sharp. Spliced segments at the 5 terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences*, 74(8):3171–3175, aug 1977. ISSN 0027-8424. doi: 10.1073/pnas.74.8.3171. URL http://www.pnas.org/content/74/8/3171.

J.Andrew Berglund, Katrin Chua, Nadja Abovich, Robin Reed, and Michael Rosbash. The Splicing Factor BBP Interacts Specifically with the Pre-mRNA Branchpoint Sequence UACUAAC. *Cell*, 89(5):781–787,

may 1997. ISSN 00928674. doi: 10.1016/S0092-8674(00)80261-5. URL http://www.sciencedirect.com/science/article/pii/S0092867400802615.

A L Beyer and Y N Osheim. Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes & Development*, 2(6):754–765, jun 1988. ISSN 0890-9369. doi: 10.1101/gad.2.6.754. URL http://genesdev.cshlp.org/content/2/6/754.

Danny A Bitton, Charalampos Rallis, Daniel C Jeffares, Graeme C Smith, Yuan Y C Chen, Sandra Codlin, Samuel Marguerat, and Jürg Bähler. LaSSO, a strategy for genome-wide mapping of intronic lariats and branch points using RNA-seq. *Genome research*, 24(7):1169–79, jul 2014. ISSN 1549-5469. doi: 10.1101/gr.166819.113. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4079972{&}tool=pmcentrez{&}rendertype=abstract.

David Botstein and Neil Risch. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, 33(3s):228–237, mar 2003. ISSN 10614036. doi: 10.1038/ng1090. URL http://www.nature.com/doifinder/10.1038/ng1090.

Nicolas Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal RNA-Seq quantification. may 2015. URL http://arxiv.org/abs/1505.02710.

Nejc Haberman Zhen Wang Julian Briese, Christopher R Sibley, Gregor Rot Konig, Venkitaraman, Tomaz Curk, and Jernej Ule. Spliceosome iCLIP enables transcriptome-wide characterization of mammalian branch points. *In Submission*, 24(7):1169–79, jul 2016. ISSN 1549-5469. doi: 10.1101/gr.166819.113. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4079972{&}tool=pmcentrez{&}rendertype=abstract.

H P J Buermans and J T den Dunnen. Next generation sequencing technology: Advances and applications. *Biochimica et biophysica acta*, 1842(10):1932–1941, jul 2014. ISSN 0006-3002. doi: 10.1016/j.bbadis.2014.06.015. URL http://www.sciencedirect.com/science/article/pii/S092544391400180X.

James M Burnette, Etsuko Miyamoto-Sato, Marc A Schaub, Jamie Conklin, and A Javier Lopez. Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. *Genetics*, 170(2):661–74, jun 2005. ISSN 0016-6731. doi: 10.1534/genetics.104.039701. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1450422{&}tool=pmcentrez{&}rendertype=abstract.

Blanche Capel, Amanda Swain, Silvia Nicolis, Adam Hacker, Michael Walter, Peter Koopman, Peter Goodfellow, and Robin Lovell-Badge. Circular transcripts of the testis-determining gene Sry in adult mouse

testis. *Cell*, 73(5):1019–1030, jun 1993. ISSN 00928674. doi: 10.1016/0092-8674(93)90279-Y. URL http://www.sciencedirect.com/science/article/pii/009286749390279Y.

Maria Chahrour and Huda Y. Zoghbi. The Story of Rett Syndrome: From Clinic to Neurobiology. *Neuron*, 56(3):422–437, 2007. ISSN 08966273. doi: 10.1016/j.neuron.2007.10.001.

C W Chao, D C Chan, A Kuo, and P Leder. The mouse formin (Fmn) gene: abundant circular RNA transcripts and gene-targeted deletion analysis. *Molecular medicine (Cambridge, Mass.)*, 4(9):614–28, oct 1998. ISSN 1076-1551. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2230310{&}tool=pmcentrez{&}rendertype=abstract.

Louise T. Chow, Richard E. Gelinas, Thomas R. Broker, and Richard J. Roberts. An amazing sequence arrangement at the 5 ends of adenovirus 2 messenger RNA. *Cell*, 12(1):1–8, sep 1977. ISSN 00928674. doi: 10.1016/0092-8674(77)90180-5. URL http://www.sciencedirect.com/science/article/pii/0092867477901805.

Peter J A Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J L de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, 25(11):1422–3, jun 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp163. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2682512{&}tool=pmcentrez{&}rendertype=abstract.

André Corvelo and Eduardo Eyras. Exon creation and establishment in human genes. *Genome biology*, 9 (9):R141, jan 2008. ISSN 1465-6914. doi: 10.1186/gb-2008-9-9-r141. URL http://genomebiology.com/2008/9/9/R141.

Steven W Criscione, Yue Zhang, William Thompson, John M Sedivy, and Nicola Neretti. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC genomics*, 15(1):583, jan 2014. ISSN 1471-2164. doi: 10.1186/1471-2164-15-583. URL http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-583.

Miri Danan, Schraga Schwartz, Sarit Edelheit, and Rotem Sorek. Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic acids research*, 40(7):3131–42, may 2012. ISSN 1362-4962. doi: 10.1093/nar/gkr1009. URL http://nar.oxfordjournals.org/content/40/7/3131.

Alice E Davidson, Panagiotis I Sergouniotis, Rosemary Burgess-Mullan, Nichola Hart-Holden, Sancy Low, Paul J Foster, Forbes D C Manson, Graeme C M Black, and Andrew R Webster. A synonymous codon variant in two patients with autosomal recessive bestrophinopathy alters in vitro splicing of BEST1. *Molecular vision*, 16:2916–22, jan 2010. ISSN 1090-0535. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013070{&}tool=pmcentrez{&}rendertype=abstract.

Thomas Derrien, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec, David Martin, Angelika Merkel, David G Knowles, Julien Lagarde, Lavanya Veeravalli, Xiaoan Ruan, Yijun Ruan, Timo Lassmann, Piero Carninci, James B Brown, Leonard Lipovich, Jose M Gonzalez, Mark Thomas, Carrie A Davis, Ramin Shiekhattar, Thomas R Gingeras, Tim J Hubbard, Cedric Notredame, Jennifer Harrow, and Roderic Guigó. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research*, 22(9):1775–89, sep 2012. ISSN 1549-5469. doi: 10.1101/gr.132159.111. URL http://europepmc.org/articles/PMC3431493.

Ashish Dhir and Emanuele Buratti. Alternative splicing: role of pseudoexons in human disease and potential therapeutic strategies. *The FEBS journal*, 277(4):841–55, feb 2010. ISSN 1742-4658. doi: 10.1111/j.1742-4658.2009.07520.x. URL http://www.ncbi.nlm.nih.gov/pubmed/20082636.

Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1):15–21, jan 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts635. URL http://bioinformatics.oxfordjournals.org/content/29/1/15.

M. O. Duff, S. Olson, X. Wei, A. Osman, A. Plocik, M. Bolisetty, S. Celniker, and B. Graveley. Genome-wide Identification of Zero Nucleotide Recursive Splicing in Drosophila. Technical report, jun 2014. URL http://biorxiv.org/content/early/2014/06/11/006163.abstract.

Harrison W. Gabel, Benyam Kinde, Hume Stroud, Caitlin S. Gilbert, David A. Harmin, Nathaniel R. Kastan, Martin Hemberg, Daniel H. Ebert, and Michael E. Greenberg. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature*, 522(7554):89–93, mar 2015. ISSN 0028-0836. doi: 10.1038/nature14319. URL http://www.nature.com/doifinder/10.1038/nature14319http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4480648{&}tool=pmcentrez{&}rendertype=abstract.

K. Gao, A. Masuda, T. Matsuura, and K. Ohno. Human branch point consensus sequence is yU-

nAy. *Nucleic Acids Research*, 36(7):2257–2267, feb 2008. ISSN 0305-1048. doi: 10.1093/nar/gkn073. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2367711{&}tool=pmcentrez{&}rendertype=abstract.

Yuan Gao, Jinfeng Wang, and Fangqing Zhao. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome biology*, 16(1):4, jan 2015. ISSN 1474-760X. doi: 10.1186/s13059-014-0571-3. URL http://genomebiology.com/2015/16/1/4.

Genotype-Tissue Expression Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, may 2015. ISSN 0036-8075. doi: 10.1126/science.1262110. URL http://science.sciencemag.org/content/348/6235/648.abstract.

Ana Rita Grosso, Ana Paula Leite, Sílvia Carvalho, Mafalda Ramos Matos, Filipa Batalha Martins, Alexandra Coitos Vítor, Joana Mp Desterro, Maria Carmo-Fonseca, and Sérgio Fernandes de Almeida. Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. *eLife*, 4:e09214, nov 2015. ISSN 2050-084X. doi: 10.7554/eLife.09214. URL http://elifesciences.org/content/early/2015/11/17/eLife.09214.abstract.

Jlenia Guarnerio, Marco Bezzi, Jong Cheol Jeong, Stella V. Paffenholz, Kelsey Berry, Matteo M. Naldini, Francesco Lo-Coco, Yvonne Tay, Andrew H. Beck, and Pier Paolo Pandolfi. Oncogenic Role of Fusion-circRNAs Derived from Cancer-Associated Chromosomal Translocations. *Cell*, 165(2):289–302, 2016. ISSN 00928674. doi: 10.1016/j.cell.2016.03.020.

Roderic Guigo, Emmanouil T Dermitzakis, Pankaj Agarwal, Chris P Ponting, Genis Parra, Alexandre Reymond, Josep F Abril, Evan Keibler, Robert Lyle, Catherine Ucla, Stylianos E Antonarakis, and Michael R Brent. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(3):1140–5, feb 2003. ISSN 0027-8424. doi: 10.1073/pnas.0337561100. URL http://www.pnas.org/content/100/3/1140.abstract?ijkey=2c387a1e1aab8460d06e51e1900a2b4408c8044e{&}keytype2=tf{_}ipsecsha.

Junjie U Guo, Vikram Agarwal, Huili Guo, and David P Bartel. Expanded identification and characterization of mammalian circular RNAs. *Genome biology*, 15(7):409, jan 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0409-z. URL http://genomebiology.com/2014/15/7/409.

Thomas B Hansen, Trine I Jensen, Bettina H Clausen, Jesper B Bramsen, Bente Finsen, Christian K Damgaard, and Jørgen Kjems. Natural RNA circles function as efficient microRNA sponges. *Nature*, 495 (7441):384–8, mar 2013. ISSN 1476-4687. doi: 10.1038/nature11993. URL http://dx.doi.org/10.1038/nature11993.

Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J Hubbard. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research*, 22(9):1760–74, sep 2012. ISSN 1549-5469. doi: 10.1101/gr.135350.111. URL http://europepmc.org/articles/PMC3431492.

Stephen W. Hartley and James C. Mullikin. Detection and Visualization of Differential Splicing in RNA-Seq Data with JunctionSeq. dec 2015. URL http://arxiv.org/abs/1512.06038.

A R Hatton, V Subramaniam, and A J Lopez. Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. *Molecular cell*, 2(6):787–96, dec 1998. ISSN 1097-2765. URL http://www.ncbi.nlm.nih.gov/pubmed/9885566.

Francisco J Herrera, Teppei Yamaguchi, Henk Roelink, and Robert Tjian. Core promoter factor TAF9B regulates neuronal gene expression. *eLife*, 3:e02559, 2014. ISSN 2050-084X. URL http://www.ncbi.nlm.nih.gov/pubmed/25006164http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4083437.

Steve Hoffmann, Christian Otto, Gero Doose, Andrea Tanzer, David Langenberger, Sabina Christ, Manfred Kunz, Lesca M Holdt, Daniel Teupser, Jörg Hackermüller, and Peter F Stadler. A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome biology*, 15(2):R34, jan 2014. ISSN 1465-6914. doi: 10.1186/gb-2014-15-2-r34. URL http://genomebiology.com/2014/15/2/R34.

D Hourcade, D Dressler, and J Wolfson. The amplification of ribosomal RNA genes involves a rolling circle intermediate. *Proceedings of the National Academy of Sciences of the United States of America*, 70(10):

2926–30, oct 1973. ISSN 0027-8424. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=427140{&}tool=pmcentrez{&}rendertype=abstract.

R. Hu, T. A. Dunn, S. Wei, S. Isharwal, R. W. Veltri, E. Humphreys, M. Han, A. W. Partin, R. L. Vessella, W. B. Isaacs, G. S. Bova, and J. Luo. Ligand-Independent Androgen Receptor Variants Derived from Splicing of Cryptic Exons Signify Hormone-Refractory Prostate Cancer. *Cancer Research*, 69(1):16–22, jan 2009. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-08-2764. URL http://cancerres.aacrjournals.org/cgi/doi/10.1158/0008-5472.CAN-08-2764.

William R Jeck, Jessica A Sorrentino, Kai Wang, Michael K Slevin, Christin E Burd, Jinze Liu, William F Marzluff, and Norman E Sharpless. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA (New York, N.Y.)*, 19(2):141–57, feb 2013. ISSN 1469-9001. doi: 10.1261/rna.035667.112. URL http://rnajournal.cshlp.org/content/19/2/141.long.

Nejc Jelen, Jernej Ule, Marko Zivin, and Robert B Darnell. Evolution of Nova-dependent splicing regulation in the brain. *PLoS genetics*, 3(10):1838–47, oct 2007. ISSN 1553-7404. doi: 10.1371/journal.pgen.0030173. URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0030173.

T Jenuwein and C D Allis. Translating the histone code. *Science (New York, N.Y.)*, 293(5532):1074–80, aug 2001. ISSN 0036-8075. doi: 10.1126/science.1063127. URL http://www.sciencemag.org/content/293/5532/1074.full.

Panagiota Kafasla, Ian Mickleburgh, Miriam Llorian, Miguel Coelho, Clare Gooding, Dmitry Cherny, Amar Joshi, Olga Kotik-Kogan, Stephen Curry, Ian C Eperon, Richard J Jackson, and Christopher W J Smith. Defining the roles and interactions of PTB. *Biochemical Society transactions*, 40(4):815–20, aug 2012. ISSN 1470-8752. doi: 10.1042/BST20120044. URL http://www.biochemsoctrans.org/content/40/4/815.abstract.

Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods*, 7(12):1009–15, dec 2010. ISSN 1548-7105. doi: 10.1038/nmeth.1528. URL http://dx.doi.org/10.1038/nmeth.1528.

Yevgenia L Khodor, Joseph Rodriguez, Katharine C Abruzzi, Chih-Hang Anthony Tang, Michael T Marr, and Michael Rosbash. Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes & development*, 25(23):2502–12, dec 2011. ISSN 1549-5477. doi:

10.1101/gad.178962.111. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3243060{&}tool=pmcentrez{&}rendertype=abstract.

Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):R36, apr 2013. ISSN 1465-6914. doi: 10.1186/gb-2013-14-4-r36. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4053844{&}tool=pmcentrez{&}rendertype=abstract.

Daehwan Kim, Ben Langmead, and Steven L Salzberg. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–360, mar 2015. ISSN 1548-7091. doi: 10.1038/nmeth.3317. URL http://dx.doi.org/10.1038/nmeth.3317.

Ian F King, Chandri N Yandava, Angela M Mabb, Jack S Hsiao, Hsien-Sung Huang, Brandon L Pearson, J Mauro Calabrese, Joshua Starmer, Joel S Parker, Terry Magnuson, Stormy J Chamberlain, Benjamin D Philpot, and Mark J Zylka. Topoisomerases facilitate transcription of long genes linked to autism. *Nature*, 501(7465):58–62, sep 2013. ISSN 1476-4687. doi: 10.1038/nature12504. URL http://dx.doi.org/10.1038/nature12504.

W. Koh, W. Pan, C. Gawad, H. C. Fan, G. A. Kerchner, T. Wyss-Coray, Y. J. Blumenfeld, Y. Y. El-Sayed, and S. R. Quake. Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. *Proceedings of the National Academy of Sciences*, 111(20):7361–7366, may 2014. ISSN 0027-8424. doi: 10.1073/pnas.1405528111. URL http://www.pnas.org/content/111/20/7361.abstract.

Paulina Kolasinska-Zwierz, Thomas Down, Isabel Latorre, Tao Liu, X Shirley Liu, and Julie Ahringer. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature genetics*, 41(3):376–81, mar 2009. ISSN 1546-1718. doi: 10.1038/ng.322. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2648722{&}tool=pmcentrez{&}rendertype=abstract.

Julian König, Kathi Zarnack, Gregor Rot, Tomaz Curk, Melis Kayikci, Blaz Zupan, Daniel J Turner, Nicholas M Luscombe, and Jernej Ule. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology*, 17(7):909–15, jul 2010. ISSN 1545-9985. doi: 10.1038/nsmb.1838. URL http://dx.doi.org/10.1038/nsmb.1838.

Alberto R Kornblihtt, Ignacio E Schor, Mariano Alló, Gwendal Dujardin, Ezequiel Petrillo, and Manuel J Muñoz. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature*

reviews. Molecular cell biology, 14(3):153–65, mar 2013. ISSN 1471-0080. doi: 10.1038/nrm3525. URL http://dx.doi.org/10.1038/nrm3525.

William C Krause, Ayesha A Shafi, Manjula Nakka, and Nancy L Weigel. Androgen receptor and its splice variant, AR-V7, differentially regulate FOXA1 sensitive genes in LNCaP prostate cancer cells. *The international journal of biochemistry & cell biology*, 54:49–59, sep 2014. ISSN 1878-5875. doi: 10.1016/j.biocel.2014.06.013. URL http://www.ncbi.nlm.nih.gov/pubmed/25008967http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4160387.

Clotilde Lagier-Tourenne, Magdalini Polymenidou, Kasey R Hutt, Anthony Q Vu, Michael Baughn, Stephanie C Huelga, Kevin M Clutario, Shuo-Chien Ling, Tiffany Y Liang, Curt Mazur, Edward Wancewicz, Aneeza S Kim, Andy Watt, Sue Freier, Geoffrey G Hicks, John Paul Donohue, Lily Shiue, C Frank Bennett, John Ravits, Don W Cleveland, and Gene W Yeo. Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. *Nature neuroscience*, 15(11): 1488–97, nov 2012. ISSN 1546-1726. doi: 10.1038/nn.3230. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3586380{&}tool=pmcentrez{&}rendertype=abstract.

Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9, apr 2012. ISSN 1548-7105. doi: 10.1038/nmeth.1923. URL http://dx.doi.org/10.1038/nmeth.1923.

Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25, jan 2009. ISSN 1465-6914. doi: 10.1186/gb-2009-10-3-r25. URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-3-r25.

Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter A C 't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G MacArthur, Monkol Lek, Esther Lizano, Henk P J Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B Montgomery, Peter Donnelly, Mark I McCarthy, Paul Flicek, Tim M Strom, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Angel Carracedo, Stylianos E Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigó, Ivo G Gut, Xavier Estivill, and Emmanouil T Dermitzakis. Transcriptome and

genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–11, sep 2013. ISSN 1476-4687. doi: 10.1038/nature12531. URL http://dx.doi.org/10.1038/nature12531.

T. Laver, J. Harrison, P.A. O'Neill, K. Moore, A. Farbos, K. Paszkiewicz, and D.J. Studholme. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, 3:1–8, 2015. ISSN 22147535. doi: 10.1016/j.bdq.2015.02.001.

Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, jan 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-323. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-323.

Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–60, jul 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp324. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2705234{&}tool=pmcentrez{&}rendertype=abstract.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9, aug 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp352. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723002{&}tool=pmcentrez{&}rendertype=abstract.

Junlin Li, Guifang Zhao, and Xiaocai Gao. Development of neurodevelopmental disorders: a regulatory mechanism involving bromodomain-containing proteins. *Journal of neurodevelopmental disorders*, 5(1):4, jan 2013. ISSN 1866-1947. doi: 10.1186/1866-1955-5-4. URL http://www.jneurodevdisorders.com/content/5/1/4.

Yan Li, Qiupeng Zheng, Chunyang Bao, Shuyi Li, Weijie Guo, Jiang Zhao, Di Chen, Jianren Gu, Xianghuo He, and Shenglin Huang. Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. *Cell research*, 25(8):981–984, jul 2015. ISSN 1748-7838. doi: 10.1038/cr.2015.82. URL http://dx.doi.org/10.1038/cr.2015.82.

Jonathan P Ling, Olga Pletnikova, Juan C Troncoso, Philip CL. Wong, J. Goecks, A. Nekrutenko, J. Taylor, and Team. TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. *Science (New York, N.Y.)*, 349(6248):650–5, aug 2015. ISSN 1095-9203. doi: 10.1126/

science.aab0983. URL http://www.ncbi.nlm.nih.gov/pubmed/26250685http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4825810.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550, jan 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0550-8. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8.

Changxue Lu and Jun Luo. Decoding the androgen receptor splice variants. *Translational andrology and urology*, 2(3):178–186, sep 2013. ISSN 2223-4691. doi: 10.3978/j.issn.2223-4683.2013.09.08. URL http://www.ncbi.nlm.nih.gov/pubmed/25356377http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4209743.

K Luger and T J Richmond. The histone tails of the nucleosome. *Current opinion in genetics & development*, 8(2):140–6, apr 1998. ISSN 0959-437X. URL http://www.ncbi.nlm.nih.gov/pubmed/9610403.

Jozef Madzo, Hui Liu, Alexis Rodriguez, Aparna Vasanthakumar, Sriram Sundaravel, DonneBennettD. Caces, TimothyJ. Looney, Li Zhang, JanetB. Lepore, Trisha Macrae, Robert Duszynski, AlanH. Shih, Chun-Xiao Song, Miao Yu, Yiting Yu, Robert Grossman, Brigitte Raumann, Amit Verma, Chuan He, RossL. Levine, Don Lavelle, BruceT. Lahn, Amittha Wickrema, and LucyA. Godley. Hydroxymethylation at Gene Regulatory Regions Directs Stem/Early Progenitor Cell Commitment during Erythropoiesis. *Cell Reports*, 6(1):231–244, 2014. ISSN 22111247. doi: 10.1016/j.celrep.2013.11.044.

J. Majewski, J. Schwartzentruber, E. Lalonde, A. Montpetit, and N. Jabado. What can exome sequencing do for you? *Journal of Medical Genetics*, 48(9):580–589, sep 2011. ISSN 0022-2593. doi: 10.1136/jmedgenet-2011-100223. URL http://jmg.bmj.com/cgi/doi/10.1136/jmedgenet-2011-100223.

Sebastian Memczak, Marvin Jens, Antigoni Elefsinioti, Francesca Torti, Janna Krueger, Agnieszka Rybak, Luisa Maier, Sebastian D Mackowiak, Lea H Gregersen, Mathias Munschauer, Alexander Loewer, Ulrike Ziebold, Markus Landthaler, Christine Kocks, Ferdinand le Noble, and Nikolaus Rajewsky. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, 495(7441):333–8, mar 2013. ISSN 1476-4687. doi: 10.1038/nature11928. URL http://dx.doi.org/10.1038/nature11928.

Sebastian Memczak, Panagiotis Papavasileiou, Oliver Peters, and Nikolaus Rajewsky. Identification and Characterization of Circular RNAs As a New Class of Putative Biomarkers in Human Blood. *PloS one*,

10(10):e0141214, jan 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0141214. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0141214.

Tim R. Mercer, Michael B. Clark, Stacey B. Andersen, Marion E. Brunck, Wilfried Haerty, Joanna Crawford, Ryan J. Taft, Lars K. Nielsen, Marcel E. Dinger, and John S. Mattick. Genome-wide discovery of human splicing branchpoints. *Genome Research*, 25(2):290–303, feb 2015. ISSN 1088-9051. doi: 10.1101/gr.182899.114. URL http://genome.cshlp.org/content/early/2015/01/05/gr.182899.114.abstract.

Eszter Nagy and Lynne E Maquat. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends in Biochemical Sciences*, 23(6):198–199, jun 1998. ISSN 09680004. doi: 10.1016/S0968-0004(98)01208-0. URL http://www.sciencedirect.com/science/article/pii/S0968000498012080.

NCBI. SRA milestone: Over 2 petabases of sequence data. URL http://www.ncbi.nlm.nih.gov/news/11-25-2013-sra-contains-2-petabases/.

Marcelo A Nobrega, Ivan Ovcharenko, Veena Afzal, and Edward M Rubin. Scanning human gene deserts for long-range enhancers. *Science (New York, N.Y.)*, 302(5644):413, oct 2003. ISSN 1095-9203. doi: 10.1126/science.1088328. URL http://www.sciencemag.org/content/302/5644/413.short.

Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S Joardar, Vamsi K Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M McGarvey, Michael R Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H Rangwala, Daniel Rausch, Lillian D Riddick, Conrad Schoch, Andrei Shkeda, Susan S Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E Tully, Anjana R Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D Murphy, and Kim D Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–45, jan 2016. ISSN 1362-4962. doi: 10.1093/nar/gkv1189. URL http://www.ncbi.nlm.nih.gov/pubmed/26553804http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4702849.

Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative

splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12): 1413–5, dec 2008. ISSN 1546-1718. doi: 10.1038/ng.259. URL http://dx.doi.org/10.1038/ng.259.

Dmitri Parkhomchuk, Tatiana Borodina, Vyacheslav Amstislavskiy, Maria Banaru, Linda Hallen, Sylvia Krobitsch, Hans Lehrach, and Alexey Soldatov. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic acids research*, 37(18):e123, oct 2009. ISSN 1362-4962. doi: 10.1093/nar/gkp596. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2764448{&}tool=pmcentrez{&}rendertype=abstract.

SophiaX. Pfister, Enni Markkanen, Yanyan Jiang, Sovan Sarkar, Mick Woodcock, Giulia Orlando, Ioanna Mavrommati, Chen-Chun Pai, Lykourgos-Panagiotis Zalmas, Neele Drobnitzky, GrigoryL. Dianov, Clare Verrill, ValentineM. Macaulay, Songmin Ying, NicholasB. LaThangue, Vincenzo D'Angiolella, AndersonJ. Ryan, and TimothyC. Humphrey. Inhibiting WEE1 Selectively Kills Histone H3K36me3-Deficient Cancers by dNTP Starvation. *Cancer Cell*, 28(5):557–568, 2015. ISSN 15356108. doi: 10.1016/j.ccell.2015.09.015.

Dmitry K Pokholok, Christopher T Harbison, Stuart Levine, Megan Cole, Nancy M Hannett, Tong Ihn Lee, George W Bell, Kimberly Walker, P Alex Rolfe, Elizabeth Herbolsheimer, Julia Zeitlinger, Fran Lewitter, David K Gifford, and Richard A Young. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, 122(4):517–27, aug 2005. ISSN 0092-8674. doi: 10.1016/j.cell.2005.06.026. URL http://www.ncbi.nlm.nih.gov/pubmed/16122420.

Magdalini Polymenidou, Clotilde Lagier-Tourenne, Kasey R Hutt, Stephanie C Huelga, Jacqueline Moran, Tiffany Y Liang, Shuo-Chien Ling, Eveline Sun, Edward Wancewicz, Curt Mazur, Holly Kordasiewicz, Yalda Sedaghat, John Paul Donohue, Lily Shiue, C Frank Bennett, Gene W Yeo, and Don W Cleveland. Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nature neuroscience*, 14(4):459–68, apr 2011. ISSN 1546-1726. doi: 10.1038/nn.2779. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3094729{&}tool=pmcentrez{&}rendertype=abstract.

Anna Marie Pyle, Olga Fedorova, and Christina Waldsich. Folding of group II introns: a model system for large, multidomain RNAs? *Trends in Biochemical Sciences*, 32(3):138–145, 2007. ISSN 09680004. doi: 10.1016/j.tibs.2007.01.005.

Fujun Qin, Zhenguo Song, Mihaela Babiceanu, Yansu Song, Loryn Facemire, Ritambhara Singh, Mazhar Adli, and Hui Li. Discovery of CTCF-sensitive Cis-spliced fusion RNAs between adjacent genes in hu-

man prostate cells. *PLoS genetics*, 11(2):e1005001, feb 2015. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005001. URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005001{#}pgen.1005001.ref008.

Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–2, mar 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq033. URL http://bioinformatics.oxfordjournals.org/content/26/6/841.short.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.r-project.org.

R Reed. The organization of 3' splice-site sequences in mammalian introns. *Genes & Development*, 3(12b): 2113–2123, dec 1989. ISSN 0890-9369. doi: 10.1101/gad.3.12b.2113. URL http://genesdev.cshlp.org/content/3/12b/2113.

M. A. Rivas, M. Pirinen, D. F. Conrad, M. Lek, E. K. Tsang, K. J. Karczewski, J. B. Maller, K. R. Kukurba, D. S. DeLuca, M. Fromer, P. G. Ferreira, K. S. Smith, R. Zhang, F. Zhao, E. Banks, R. Poplin, D. M. Ruderfer, S. M. Purcell, T. Tukiainen, E. V. Minikel, P. D. Stenson, D. N. Cooper, K. H. Huang, T. J. Sullivan, J. Nedzel, C. D. Bustamante, J. B. Li, M. J. Daly, R. Guigo, P. Donnelly, K. Ardlie, M. Sammeth, E. T. Dermitzakis, M. I. McCarthy, S. B. Montgomery, T. Lappalainen, D. G. MacArthur, A. V. Segre, T. R. Young, E. T. Gelfand, C. A. Trowbridge, L. D. Ward, P. Kheradpour, B. Iriarte, Y. Meng, C. D. Palmer, T. Esko, W. Winckler, J. Hirschhorn, M. Kellis, G. Getz, A. A. Shablin, G. Li, Y.-H. Zhou, A. B. Nobel, I. Rusyn, F. A. Wright, A. Battle, S. Mostafavi, M. Mele, F. Reverter, J. Goldmann, D. Koller, E. R. Gamazon, H. K. Im, A. Konkashbaev, D. L. Nicolae, N. J. Cox, T. Flutre, X. Wen, M. Stephens, J. K. Pritchard, Z. Tu, B. Zhang, T. Huang, Q. Long, L. Lin, J. Yang, J. Zhu, J. Liu, A. Brown, B. Mestichelli, D. Tidwell, E. Lo, M. Salvatore, S. Shad, J. A. Thomas, J. T. Lonsdale, R. C. Choi, E. Karasik, K. Ramsey, M. T. Moser, B. A. Foster, B. M. Gillard, J. Syron, J. Fleming, H. Magazine, R. Hasz, G. D. Walters, J. P. Bridge, M. Miklos, S. Sullivan, L. K. Barker, H. Traino, M. Mosavel, L. A. Siminoff, D. R. Valley, D. C. Rohrer, S. Jewel, P. Branton, L. H. Sobin, M. Barcus, L. Qi, P. Hariharan, S. Wu, D. Tabor, C. Shive, A. M. Smith, S. A. Buia, A. H. Undale, K. L. Robinson, N. Roche, K. M. Valentino, A. Britton, R. Burges, D. Bradbury, K. W. Hambright, J. Seleski, G. E. Korzeniewski, K. Erickson, Y. Marcus, J. Tejada, M. Taherian, C. Lu, B. E. Robles, M. Basile, D. C. Mash, S. Volpi, J. P. Struewing, G. F. Temple, J. Boyer, D. Colantuoni, R. Little, S. Koester, L. J. Carithers, H. M. Moore, P. Guan, C. Compton, S. J. Sawyer, J. P. Demchok, J. B. Vaught, C. A. Rabiner, N. C. Lockhart, M. R. Friedlander, P. A. C. 't Hoen,

J. Monlong, M. Gonzalez-Porta, N. Kurbatova, T. Griebel, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlof, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, E. Lizano, H. P. J. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, P. Flicek, T. M. Strom, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Hasler, A.-C. Syvanen, G.-J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, I. G. Gut, and X. Estivill. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science*, 348(6235):666–669, may 2015. ISSN 0036-8075. doi: 10.1126/science.1261877. URL `http://www.sciencemag.org/content/348/6235/666.abstract`.

Sanja Rogic, Ben Montpetit, Holger H Hoos, Alan K Mackworth, BF Francis Ouellette, and Philip Hieter. Correlation between the secondary structure of pre-mRNA introns and the efficiency of splicing in Saccharomyces cerevisiae. *BMC Genomics*, 9(1):355, 2008. ISSN 1471-2164. doi: 10.1186/1471-2164-9-355. URL `http://www.biomedcentral.com/1471-2164/9/355`.

AlexanderB. Rosenberg, RupaliP. Patwardhan, Jay Shendure, and Georg Seelig. Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences. *Cell*, 163(3):698–711, oct 2015. ISSN 00928674. doi: 10.1016/j.cell.2015.09.054. URL `http://www.cell.com/article/S0092867415012714/fulltext`.

Agnieszka Rybak-Wolf, Christin Stottmeister, Petar Glažar, Marvin Jens, Natalia Pino, Sebastian Giusti, Mor Hanan, Mikaela Behm, Osnat Bartok, Reut Ashwal-Fluss, Margareta Herzog, Luisa Schreyer, Panagiotis Papavasileiou, Andranik Ivanov, Marie Öhman, Damian Refojo, Sebastian Kadener, and Nikolaus Rajewsky. Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Molecular cell*, 58(5):870–85, jun 2015. ISSN 1097-4164. doi: 10.1016/j.molcel.2015.03.027. URL `http://www.ncbi.nlm.nih.gov/pubmed/25921068`.

Hiroaki Sakai, Ken Naito, Eri Ogiso-Tanaka, Yu Takahashi, Kohtaro Iseki, Chiaki Muto, Kazuhito Satou, Kuniko Teruya, Akino Shiroma, Makiko Shimoji, Takashi Hirano, Takeshi Itoh, Akito Kaga, Norihiko Tomooka, M. Margulies, D. R. Bentley, T. P. Michael, R. VanBuren, S. R. Wessler, L. Caporale, C. Alkan, S. Sajjadian, E. E. Eichler, J. F. Denton, J. Eid, H. Lee, K. Berlin, K. Wang, J. C. Dohm, Y. J. Kang, Y. J. Kang, J. Schmutz, R. K. Varshney, R. K. Varshney, S. Liu, K. R. Bradnam, M. D. Bennett, I. J. Leitch, S. Gnerre, O. K. Han, G. Parra, K. Bradnam, I. Korf, J. Schmutz, V. Krishnakumar, J. W. Davey, Y. Honma, T. Y. Seng, S. Koren, A. M. Philippy, H. Funatsuki, Z. Xia, M. G. Murray, W. F. Thompson, Z. Li, H. N. Trick, A. M. Bolger, M. Lohse, B. Usadel, J. R. Miller, H. Li, R. Durbin,

A. McKenna, H. Li, J. T. Simpson, R. Durbin, K. Naito, A. Kaga, N. Tomooka, M. Kawase, M. Boetzer, C. V. Henkel, H. J. Jansen, D. Butler, W. Pirovano., A. C. English, M. Miyamoto, A. J. Adler, G. B. Wiley, P. M. Gaffney, K. W. Broman, H. Wu, Ś. Sen, G. A. Churchill, H. Iwata, S. Ninomiya, S. Kurtz, T. Nussbaumer, O. Kohany, A. J. Gentles, L. Hankus, J. Jurka, F. Cunningham, D. Kim, C. Trapnell, B. J. Haas, N. Rhind, B. J. Haas, H. Numa, T. Itoh, P. Jones, C. Trapnell, A. J. Enright, S. Van Dongen, and C. A. Ouzounis. The power of single molecule real-time sequencing technology in the de novo assembly of a eukaryotic genome. *Scientific Reports*, 5:16780, nov 2015. ISSN 2045-2322. doi: 10.1038/srep16780. URL `http://www.nature.com/articles/srep16780`.

Julia Salzman, Charles Gawad, Peter Lincoln Wang, Norman Lacayo, and Patrick O Brown. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PloS one*, 7(2): e30733, jan 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0030733. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3270023{&}tool=pmcentrez{&}rendertype=abstract`.

Kaitlin E Samocha, Elise B Robinson, Stephan J Sanders, Christine Stevens, Aniko Sabo, Lauren M McGrath, Jack A Kosmicki, Karola Rehnström, Swapan Mallick, Andrew Kirby, Dennis P Wall, Daniel G MacArthur, Stacey B Gabriel, Mark DePristo, Shaun M Purcell, Aarno Palotie, Eric Boerwinkle, Joseph D Buxbaum, Edwin H Cook, Richard A Gibbs, Gerard D Schellenberg, James S Sutcliffe, Bernie Devlin, Kathryn Roeder, Benjamin M Neale, and Mark J Daly. A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, 46(9):944–950, aug 2014. ISSN 1061-4036. doi: 10.1038/ng.3050. URL `http://dx.doi.org/10.1038/ng.3050`.

H. L. Sanger, G. Klotz, D. Riesner, H. J. Gross, and A. K. Kleinschmidt. Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures. *Proceedings of the National Academy of Sciences*, 73(11):3852–3856, nov 1976. ISSN 0027-8424. doi: 10.1073/pnas.73.11.3852. URL `http://www.pnas.org/content/73/11/3852`.

Hannelore V. Heemers Robert L. Vessella Donald J. Tindall Scott M. Dehm, Lucy J. Schmidt. Splicing of a novel AR exon generates a constitutively active androgen receptor that mediates prostate cancer therapy resistance. *Cancer research*, 68(13):5469, 2008.

SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature biotechnology*, 32(9):903–14, sep 2014. ISSN 1546-1696. doi: 10.1038/nbt.2957. URL `http://dx.doi.org/10.1038/nbt.2957`.

M B Shapiro and P Senapathy. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic acids research*, 15(17):7155–74, sep 1987. ISSN 0305-1048. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=306199{&}tool=pmcentrez{&}rendertype=abstract.

Samuel Shepard, Mark McCreary, and Alexei Fedorov. The peculiarities of large intron splicing in animals. *PLoS ONE*, 4(11):e7853, 2009. ISSN 19326203. doi: 10.1371/journal.pone.0007853. URL http://www.ncbi.nlm.nih.gov/pubmed/19924226http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2773006.

Christopher R Sibley, Warren Emmett, Lorea Blazquez, Ana Faro, Nejc Haberman, Michael Briese, Daniah Trabzuni, Mina Ryten, Michael E Weale, John Hardy, Miha Modic, Tomaž Curk, Stephen W Wilson, Vincent Plagnol, and Jernej Ule. Recursive splicing in long vertebrate genes. *Nature*, 521(7552):371–5, may 2015. ISSN 1476-4687. doi: 10.1038/nature14466. URL http://dx.doi.org/10.1038/nature14466.

Adam Siepel, Mark Diekhans, Brona Brejová, Laura Langton, Michael Stevens, Charles L G Comstock, Colleen Davis, Brent Ewing, Shelly Oommen, Christopher Lau, Hung-Chun Yu, Jianfeng Li, Bruce A Roe, Phil Green, Daniela S Gerhard, Gary Temple, David Haussler, and Michael R Brent. Targeted discovery of novel human exons by comparative genomics. *Genome research*, 17(12):1763–73, dec 2007. ISSN 1088-9051. doi: 10.1101/gr.7128207. URL http://genome.cshlp.org/content/17/12/1763.abstract?ijkey=7773628f0a17d5d4e0ff8e995829491ade40cf42{&}keytype2=tf{_}ipsecsha.

Robert J Sims and Danny Reinberg. Processing the H3K36me3 signature. *Nature genetics*, 41(3):270–1, mar 2009. ISSN 1546-1718. doi: 10.1038/ng0309-270. URL http://dx.doi.org/10.1038/ng0309-270.

Tamara Steijger, Josep F Abril, Pär G Engström, Felix Kokocinski, Martin Akerman, Tyler Alioto, Giovanna Ambrosini, Stylianos E Antonarakis, Jonas Behr, and Paul Bertone. Assessment of transcript reconstruction methods for RNA-seq. *Nature methods*, 10(12):1177–84, dec 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2714. URL http://europepmc.org/articles/PMC3851240.

Peter D Stenson, Edward V Ball, Matthew Mort, Andrew D Phillips, Jacqueline A Shiel, Nick S T Thomas, Shaun Abeysinghe, Michael Krawczak, and David N Cooper. Human Gene Mutation Database (HGMD): 2003 update. *Human mutation*, 21(6):577–81, jun 2003. ISSN 1098-1004. doi: 10.1002/humu.10212. URL http://www.ncbi.nlm.nih.gov/pubmed/12754702.

Brian D. Strahl and C. David Allis. The language of covalent histone modifications. *Nature*, 403(6765): 41–45, jan 2000. ISSN 0028-0836. doi: 10.1038/47412. URL http://www.nature.com/doifinder/10.1038/47412.

H. Sun and L. A. Chasin. Multiple splicing defects in an intronic false exon. *Molecular and cellular biology*, 20(17):6414–25, sep 2000. ISSN 0270-7306. doi: 10.1128/MCB.20.17.6414-6425.2000. URL http://mcb.asm.org/cgi/doi/10.1128/MCB.20.17.6414-6425.2000http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=86117{&}tool=pmcentrez{&}rendertype=abstract.

Allison J Taggart, Alec M DeSimone, Janice S Shih, Madeleine E Filloux, and William G Fairbrother. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nature structural & molecular biology*, 19(7):719–21, jul 2012. ISSN 1545-9985. doi: 10.1038/nsmb.2327. URL http://dx.doi.org/10.1038/nsmb.2327.

Zhonghui Tang, OscarJunhong Luo, Xingwang Li, Meizhen Zheng, JacquelineJufen Zhu, Przemyslaw Szalaj, Pawel Trzaskoma, Adriana Magalska, Jakub Wlodarczyk, Blazej Ruszczycki, Paul Michalski, Emaly Piecuch, Ping Wang, Danjuan Wang, SimonZhongyuan Tian, May Penrad-Mobayed, LaurentM. Sachs, Xiaoan Ruan, Chia-Lin Wei, EdisonT. Liu, GrzegorzM. Wilczynski, Dariusz Plewczynski, Guoliang Li, and Yijun Ruan. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell*, 163(7):1611–1627, dec 2015. ISSN 00928674. doi: 10.1016/j.cell.2015.11.024. URL http://www.cell.com/article/S0092867415015044/fulltext.

Margaret A Taub, Hector Corrada Bravo, and Rafael A Irizarry. Overcoming bias and systematic errors in next generation sequencing data. *Genome medicine*, 2(12):87, jan 2010. ISSN 1756-994X. doi: 10.1186/gm208. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3025429{&}tool=pmcentrez{&}rendertype=abstract.

Sudhir Thakurela, Angela Garding, Johannes Jung, Dirk Schübeler, Lukas Burger, and Vijay K Tiwari. Gene regulation and priming by topoisomerase IIα in embryonic stem cells. *Nature communications*, 4: 2478, jan 2013. ISSN 2041-1723. doi: 10.1038/ncomms3478. URL http://www.nature.com/ncomms/2013/130927/ncomms3478/full/ncomms3478.html.

The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)*, 306(5696):636–40, oct 2004. ISSN 1095-9203. doi: 10.1126/science.1105136. URL http://science.sciencemag.org/content/306/5696/636.abstract.

The EXaC Consortium. Analysis of protein-coding genetic variation in 60,706 humans. Technical report, oct 2015. URL http://biorxiv.org/content/early/2015/10/30/030338.abstract.

The Genotype-Tissue Expression (GTEx) project Consortium. GTEx Portal, 2014. URL http://www.gtexportal.org/home/.

The Genotype-Tissue Expression (GTEx) project Consortium. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and biobanking*, 13(5):311–9, oct 2015. ISSN 1947-5543. doi: 10.1089/bio.2015.0032. URL http://online.liebertpub.com/doi/full/10.1089/bio.2015.0032.

H. Tilgner, D. G. Knowles, R. Johnson, C. A. Davis, S. Chakrabortty, S. Djebali, J. Curado, M. Snyder, T. R. Gingeras, and R. Guigo. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research*, 22(9):1616–1625, sep 2012. ISSN 1088-9051. doi: 10.1101/gr.134445.111. URL http://genome.cshlp.org/content/22/9/1616.

Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25(9):1105–11, may 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp120. URL http://bioinformatics.oxfordjournals.org/content/25/9/1105.abstract.

Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–5, may 2010a. ISSN 1546-1696. doi: 10.1038/nbt.1621. URL http://dx.doi.org/10.1038/nbt.1621.

Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, may 2010b. ISSN 1087-0156. doi: 10.1038/nbt.1621. URL http://www.nature.com/doifinder/10.1038/nbt.1621.

Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis

of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7(3):562–78, mar 2012. ISSN 1750-2799. doi: 10.1038/nprot.2012.016. URL http://dx.doi.org/10.1038/nprot.2012.016.

Jernej Ule, Giovanni Stefani, Aldo Mele, Matteo Ruggiu, Xuning Wang, Bahar Taneri, Terry Gaasterland, Benjamin J Blencowe, and Robert B Darnell. An RNA map predicting Nova-dependent splicing regulation. *Nature*, 444(7119):580–6, nov 2006. ISSN 1476-4687. doi: 10.1038/nature05304. URL http://dx.doi.org/10.1038/nature05304.

Morten T. Venø, Thomas B. Hansen, Susanne T. Venø, Bettina H. Clausen, Manuela Grebing, Bente Finsen, Ida E. Holm, and Jørgen Kjems. Spatio-temporal regulation of circular RNA expression during porcine embryonic brain development. *Genome Biology*, 16(1):245, nov 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0801-3. URL http://genomebiology.com/2015/16/1/245.

Klaudia Walter, Josine L. Min, Jie Huang, Lucy Crooks, Yasin Memari, Shane McCarthy, John R. B. Perry, ChangJiang Xu, Marta Futema, Daniel Lawson, Valentina Iotchkova, Stephan Schiffels, Audrey E. Hendricks, Petr Danecek, Rui Li, James Floyd, Louise V. Wain, Inês Barroso, Steve E. Humphries, Matthew E. Hurles, Eleftheria Zeggini, Jeffrey C. Barrett, Vincent Plagnol, J. Brent Richards, Celia M. T. Greenwood, Nicholas J. Timpson, Richard Durbin, Nicole Soranzo, Senduran Bala, Peter Clapham, Guy Coates, Tony Cox, Allan Daly, Yuanping Du, Sarah Edkins, Peter Ellis, Paul Flicek, Xiaosen Guo, Xueqin Guo, Liren Huang, David K. Jackson, Chris Joyce, Thomas Keane, Anja Kolb-Kokocinski, Cordelia Langford, Yingrui Li, Jieqin Liang, Hong Lin, Ryan Liu, John Maslen, Dawn Muddyman, Michael A. Quail, Jim Stalker, Jianping Sun, Jing Tian, Guangbiao Wang, Jun Wang, Yu Wang, Kim Wong, Pingbo Zhang, Ewan Birney, Chris Boustred, Lu Chen, Gail Clement, Massimiliano Cocca, George Davey Smith, Ian N. M. Day, Aaron Day-Williams, Thomas Down, Ian Dunham, David M. Evans, Tom R. Gaunt, Matthias Geihs, Deborah Hart, Bryan Howie, Tim Hubbard, Pirro Hysi, Yalda Jamshidi, Konrad J. Karczewski, John P. Kemp, Genevieve Lachance, Monkol Lek, Margarida Lopes, Daniel G. MacArthur, Jonathan Marchini, Massimo Mangino, Iain Mathieson, Sarah Metrustry, Alireza Moayyeri, Kate Northstone, Kalliope Panoutsopoulou, Lavinia Paternoster, Lydia Quaye, Susan Ring, Graham R. S. Ritchie, Hashem A. Shihab, So-Youn Shin, Kerrin S. Small, María Soler Artigas, Lorraine Southam, Timothy D. Spector, Beate St Pourcain, Gabriela Surdulescu, Ioanna Tachmazidou, Martin D. Tobin, Ana M. Valdes, Peter M. Visscher, Kirsten Ward, Scott G. Wilson, Jian Yang, Feng Zhang, Hou-Feng Zheng, Richard Anney, Muhammad Ayub, Douglas Blackwood, Patrick F. Bolton, Gerome Breen, David A. Collier, Nick Craddock, Sarah Curran, David Curtis, Louise Gallagher, Daniel Geschwind, Hugh Gurling, Peter Holmans, Irene Lee,

Jouko Lönnqvist, Peter McGuffin, Andrew M. McIntosh, Andrew G. McKechanie, Andrew McQuillin, James Morris, Michael C. O'Donovan, Michael J. Owen, Aarno Palotie, Jeremy R. Parr, Tiina Paunio, Olli Pietilainen, Karola Rehnström, Sally I. Sharp, David Skuse, David St Clair, Jaana Suvisaari, James T. R. Walters, Hywel J. Williams, Elena Bochukova, Rebecca Bounds, I. Sadaf Farooqi, Julia Keogh, Gaëlle Marenne, Stephen O'Rahilly, Eleanor Wheeler, Saeed Al Turki, Carl A. Anderson, Dinu Antony, Phil Beales, Jamie Bentham, Shoumo Bhattacharya, Mattia Calissano, Keren Carss, Krishna Chatterjee, Sebahattin Cirak, Catherine Cosgrove, David R. Fitzpatrick, A. Reghan Foley, Christopher S. Franklin, Detelina Grozeva, Hannah M. Mitchison, Francesco Muntoni, Alexandros Onoufriadis, Victoria Parker, Felicity Payne, F. Lucy Raymond, Nicola Roberts, David B. Savage, Peter Scambler, Miriam Schmidts, Nadia Schoenmakers, Robert K. Semple, Eva Serra, Olivera Spasic-Boskovic, Elizabeth Stevens, Margriet van Kogelenberg, Parthiban Vijayarangakannan, Kathleen A. Williamson, Crispian Wilson, Tamieka Whyte, Antonio Ciampi, Karim Oualkacha, Martin Bobrow, Heather Griffin, Jane Kaye, Karen Kennedy, Alastair Kent, Carol Smee, Ruth Charlton, Rosemary Ekong, Farrah Khawaja, Luis R. Lopes, Nicola Migone, Stewart J. Payne, Rebecca C. Pollitt, Sue Povey, Cheryl K. Ridout, Rachel L. Robinson, Richard H. Scott, Adam Shaw, Petros Syrris, Rohan Taylor, Anthony M. Vandersteen, Antoinette Amuzu, Juan Pablo Casas, John C. Chambers, George Dedoussis, Giovanni Gambaro, Paolo Gasparini, Aaron Isaacs, Jon Johnson, Marcus E. Kleber, Jaspal S. Kooner, Claudia Langenberg, Jian'an Luan, Giovanni Malerba, Winfried März, Angela Matchan, Richard Morris, Børge G. Nordestgaard, Marianne Benn, Robert A. Scott, Daniela Toniolo, Michela Traglia, Anne Tybjaerg-Hansen, Cornelia M. van Duijn, Elisabeth M. van Leeuwen, Anette Varbo, Peter Whincup, Gianluigi Zaza, and Weihua Zhang. The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571):82–90, sep 2015. ISSN 0028-0836. doi: 10.1038/nature14962. URL `http://dx.doi.org/10.1038/nature14962`.

Peter L Wang, Yun Bao, Muh-Ching Yee, Steven P Barrett, Gregory J Hogan, Mari N Olsen, José R Dinneny, Patrick O Brown, and Julia Salzman. Circular RNA is expressed across the eukaryotic tree of life. *PloS one*, 9(6):e90859, jan 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0090859. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3946582{&}tool=pmcentrez{&}rendertype=abstract`.

Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, jan 2009. ISSN 1471-0064. doi: 10.1038/nrg2484. URL `http://dx.doi.org/10.1038/nrg2484`.

Tom R Webb, David A Parfitt, Jessica C Gardner, Ariadna Martinez, Dalila Bevilacqua, Alice E Davidson, Ilaria Zito, Dawn L Thiselton, Jacob H C Ressa, Marina Apergi, Nele Schwarz, Naheed Kanuga, Michel Michaelides, Michael E Cheetham, Michael B Gorin, and Alison J Hardcastle. Deep intronic mutation in OFD1, identified by targeted genomic next-generation sequencing, causes a severe form of X-linked retinitis pigmentosa (RP23). *Human molecular genetics*, 21(16):3647–54, aug 2012. ISSN 1460-2083. doi: 10.1093/hmg/dds194. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3406759{&}tool=pmcentrez{&}rendertype=abstract`.

Pei-Chi Wei, Amelia N Chang, Jennifer Kao, Zhou Du, Robin M Meyers, Frederick W Alt, and Bjoern Schwer. Long Neural Genes Harbor Recurrent DNA Break Clusters in Neural Stem/Progenitor Cells. *Cell*, 164(4):644–655, feb 2016a. ISSN 1097-4172. doi: 10.1016/j.cell.2015.12.039. URL `http://www.ncbi.nlm.nih.gov/pubmed/26871630`.

Shan Wei, Zev Williams, P. M. Ashton, S. Nair, T. Dallman, S. Rubino, W. Rabsch, P. R. Brezina, D. S. Brezina, W. G. Kearns, S. Chen, S. Li, W. Xie, X. Li, C. Zhang, S. H. Cheng, P. Jiang, K. Sun, Y. K. Cheng, K. C. Chan, G. M. Cherf, K. R. Lieberman, H. Rashid, C. E. Lam, K. Karplus, M. A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, H. C. Fan, S. R. Quake, M. C. Frith, R. Wan, P. Horton, M. Jain, I. T. Fiddes, K. H. Miga, H. E. Olsen, B. Paten, W. J. Kent, A. Kilianski, J. L. Haas, E. J. Corriveau, A. T. Liem, K. L. Willis, B. Langmead, S. L. Salzberg, N. J. Loman, A. R. Quinlan, N. J. Loman, M. Watson, M. Martin, F. J. Miller, F. L. Rosenfeldt, C. Zhang, A. W. Linnane, P. Nagley, G. E. Palomaki, C. Deciu, E. M. Kloza, G. M. Lambert-Messerlian, J. E. Haddow, J. Quick, A. R. Quinlan, N. J. Loman, J. M. Urban, J. Bliss, C. E. Lawrence, S. A. Gerbi, D. Wells, K. Kaur, J. Grifo, M. Glassner, and J. C. Taylor. Rapid Short-Read Sequencing and Aneuploidy Detection Using MinION Nanopore Technology. *Genetics*, 202(1):37–44, jan 2016b. ISSN 1943-2631. doi: 10.1534/genetics.115.182311. URL `http://www.ncbi.nlm.nih.gov/pubmed/26500254http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4701100`.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL `http://ggplot2.org`.

Hadley Wickham. *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*, 2016. URL `https://cran.r-project.org/package=tidyr`.

Hadley Wickham and Romain Francois. *dplyr: A Grammar of Data Manipulation*, 2015. URL `https://cran.r-project.org/package=dplyr`.

Wikimedia Commons. RNA-seq mapping of short reads in exon-exon junctions., 2009. URL https://en.wikipedia.org/wiki/RNA-Seq{#}/media/File:RNA-Seq-alignment.png.

G. W. Wilson and L. D. Stein. RNASequel: accurate and repeat tolerant realignment of RNA-seq reads. *Nucleic Acids Research*, pages gkv594–, jun 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv594. URL http://nar.oxfordjournals.org/content/early/2015/06/16/nar.gkv594.full.

Thomas D Wu and Serban Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics (Oxford, England)*, 26(7):873–81, apr 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq057. URL http://bioinformatics.oxfordjournals.org/content/26/7/873.full.

H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. C. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, Q. Morris, Y. Barash, A. R. Krainer, N. Jojic, S. W. Scherer, B. J. Blencowe, and B. J. Frey. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347(6218), dec 2014. ISSN 0036-8075. doi: 10.1126/science.1254806. URL http://www.sciencemag.org/content/347/6218/1254806.abstract.

Yaping Yang, Donna M. Muzny, Jeffrey G. Reid, Matthew N. Bainbridge, Alecia Willis, Patricia A. Ward, Alicia Braxton, Joke Beuten, Fan Xia, Zhiyv Niu, Matthew Hardison, Richard Person, Mir Reza Bekheirnia, Magalie S. Leduc, Amelia Kirby, Peter Pham, Jennifer Scull, Min Wang, Yan Ding, Sharon E. Plon, James R. Lupski, Arthur L. Beaudet, Richard A. Gibbs, and Christine M. Eng. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *New England Journal of Medicine*, 369(16):1502–1511, oct 2013. ISSN 0028-4793. doi: 10.1056/NEJMoa1306555. URL http://www.nejm.org/doi/abs/10.1056/NEJMoa1306555.

Gene Yeo and Christopher B Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of computational biology : a journal of computational molecular cell biology*, 11(2-3):377–94, jan 2004. ISSN 1066-5277. doi: 10.1089/1066527041410418. URL http://www.researchgate.net/publication/8423973{_}Maximum{_}Entropy{_}Modeling{_}of{_}Short{_}Sequence{_}Motifs{_}with{_}Applications{_}to{_}

Xintian You, Irena Vlatkovic, Ana Babic, Tristan Will, Irina Epstein, Georgi Tushev, Güney Akbalik, Mantian Wang, Caspar Glock, Claudia Quedenau, Xi Wang, Jingyi Hou, Hongyu Liu, Wei Sun, Sivakumar Sambandan, Tao Chen, Erin M Schuman, and Wei Chen. Neural circular RNAs are derived from synaptic

genes and regulated by development and plasticity. *Nature neuroscience*, 18(4):603–610, feb 2015. ISSN 1546-1726. doi: 10.1038/nn.3975. URL http://dx.doi.org/10.1038/nn.3975.

Kathi Zarnack, Julian König, Mojca Tajnik, Iñigo Martincorena, Sebastian Eustermann, Isabelle Stévant, Alejandro Reyes, Simon Anders, NicholasM. Luscombe, Jernej Ule, S. Anders, W. Huber, S. Anders, A. Reyes, W. Huber, A.L. Beyer, M.E. Christensen, B.W. Walker, W.M. LeStourgeon, E. Buratti, M. Chivers, J. Královicová, M. Romano, M. Baralle, A.R. Krainer, I. Vorechovsky, I.W. Caras, M.A. Davitz, L. Rhee, G. Weddell, S. Martin, D.W., L. Inkel, A. Bramard, E.R. Paquet, V. Watier, M. Durand, and et Al. Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements. *Cell*, 152(3):453–466, jan 2013. ISSN 00928674. doi: 10.1016/j.cell.2012.12.023. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867412015450.

Xiao-Ou Zhang, Hai-Bin Wang, Yang Zhang, Xuhua Lu, Ling-Ling Chen, and Li Yang. Complementary Sequence-Mediated Exon Circularization. *Cell*, 159(1):134–147, sep 2014. ISSN 00928674. doi: 10.1016/j.cell.2014.09.001. URL http://www.sciencedirect.com/science/article/pii/S0092867414011118.

Yang Zhang, Xiao-Ou Zhang, Tian Chen, Jian-Feng Xiang, Qing-Fei Yin, Yu-Hang Xing, Shanshan Zhu, Li Yang, and Ling-Ling Chen. Circular intronic long noncoding RNAs. *Molecular cell*, 51(6):792–806, sep 2013. ISSN 1097-4164. doi: 10.1016/j.molcel.2013.08.017. URL http://www.sciencedirect.com/science/article/pii/S109727651300590X.

Appendix

Sample ID	Sample name	Population	Unmapped reads	Total reads	Uniquely mapped reads
HG00096.1.M1111246	HG00096	GBR	749198	54512330	52490638
HG00097.7.M1202192	HG00097	GBR	2028887	87882216	83657715
HG00099.1.M1202096	HG00099	GBR	689457	42279782	40623431
HG00099.5.M1201313	HG00099	GBR	5281077	79375834	72103090
HG00100.2.M1112158	HG00100	GBR	799233	44683676	42664671
HG00101.1.M1111244	HG00101	GBR	750663	46516462	44750969
HG00102.3.M1202028	HG00102	GBR	947358	42245278	40198976
HG00103.4.M1202083	HG00103	GBR	655952	51371940	49449377
HG00104.1.M1111245	HG00104	GBR	4118192	53789606	48426398
HG00105.1.M1202097	HG00105	GBR	812991	45109250	43153711
HG00105.3.M1202236	HG00105	GBR	3423341	90984450	85252445
HG00106.4.M1202085	HG00106	GBR	642043	55590222	53531102
HG00108.7.M1202192	HG00108	GBR	2191175	81096644	76656384
HG00109.1.M1202094	HG00109	GBR	914189	40426858	38572948
HG00109.3.M1202025	HG00109	GBR	1452343	35339058	32957736
HG00110.2.M1201312	HG00110	GBR	1705761	77273036	73513206
HG00111.1.M1202098	HG00111	GBR	949149	44447978	42327537
HG00111.2.M1112154	HG00111	GBR	1018901	70129972	67086970
HG00112.6.M1201192	HG00112	GBR	781275	51683556	49559920
HG00114.1.M1202093	HG00114	GBR	1036321	53995672	51646881
HG00114.6.M1202171	HG00114	GBR	503591	40109798	38598295
HG00115.6.M1201191	HG00115	GBR	750221	54454188	52356964
HG00116.2.M1201311	HG00116	GBR	1084765	55830456	53239542
HG00117.1.M1111242	HG00117	GBR	2796084	109095664	103900574
HG00117.1.M1202091	HG00117	GBR	1484537	60719248	57771650
HG00117.2.M1112164	HG00117	GBR	819415	51926286	49779807
HG00117.3.M1202026	HG00117	GBR	388899	33075792	31796173
HG00117.4.M1202084	HG00117	GBR	465394	47575136	45949471
HG00117.5.M1201313	HG00117	GBR	648931	61861444	59688047
HG00117.6.M1202171	HG00117	GBR	623404	46317542	44521685
HG00117.7.M1202194	HG00117	GBR	1199228	62145486	59421064
HG00118.4.M1202085	HG00118	GBR	549439	57526386	55538725
HG00119.1.M1202093	HG00119	GBR	611765	39611928	38076495
HG00119.2.M1112166	HG00119	GBR	824561	63762212	61365974
HG00120.3.M1202022	HG00120	GBR	768364	27921432	26454217
HG00121.1.M1111247	HG00121	GBR	555473	41264264	39737825
HG00122.6.M1201191	HG00122	GBR	962321	60682448	58204913
HG00123.4.M1202087	HG00123	GBR	578385	62040106	59807299
HG00124.3.M1202237	HG00124	GBR	1139818	61028462	58346858
HG00125.1.M1111246	HG00125	GBR	3144291	57426696	52967279
HG00126.1.M1111248	HG00126	GBR	1219309	73205592	70182514
HG00127.1.M1111242	HG00127	GBR	1047538	73589032	70884728
HG00128.1.M1111246	HG00128	GBR	1710232	56537920	53499918
HG00129.4.M1202088	HG00129	GBR	668430	48154050	46249311
HG00130.5.M1201317	HG00130	GBR	667930	63452554	61074197
HG00131.1.M1202098	HG00131	GBR	621485	29379134	28095616
HG00131.2.M1112155	HG00131	GBR	806992	59497434	57238075
HG00132.2.M1112154	HG00132	GBR	928224	74637674	71738664
HG00133.1.M1202095	HG00133	GBR	1023518	50696358	48351582
HG00133.2.M1112162	HG00133	GBR	893968	54079816	51726909
HG00134.1.M1202093	HG00134	GBR	596630	32031216	30677000
HG00134.6.M1201196	HG00134	GBR	1114744	79709308	76686408
HG00135.3.M1202028	HG00135	GBR	544287	29310864	28075438
HG00136.4.M1202087	HG00136	GBR	651540	54619716	52584455
HG00137.1.M1202096	HG00137	GBR	1019638	44211128	42094324
HG00137.6.M1202171	HG00137	GBR	791046	48468364	46399910
HG00138.1.M1202092	HG00138	GBR	1401091	55049138	52192193
HG00138.5.M1201315	HG00138	GBR	805132	58861586	56397945
HG00139.7.M1202191	HG00139	GBR	2137325	85762042	81731056
HG00141.5.M1201313	HG00141	GBR	631995	73743558	71482529
HG00142.1.M1202097	HG00142	GBR	707630	38930764	37307190
HG00142.4.M1202086	HG00142	GBR	559264	50277322	48460421
HG00143.1.M1202097	HG00143	GBR	554721	26193256	24973852
HG00143.7.M1202192	HG00143	GBR	1457224	57775998	54880875
HG00145.6.M1201191	HG00145	GBR	2278481	59365444	55577082
HG00146.2.M1112161	HG00146	GBR	847112	57005198	54798442
HG00148.3.M1202026	HG00148	GBR	1748793	41614086	38816013
HG00149.1.M1111246	HG00149	GBR	1395132	61742046	59082919
HG00150.4.M1202087	HG00150	GBR	582078	58932186	56880604
HG00151.3.M1202024	HG00151	GBR	424264	30487154	29208383
HG00152.7.M1202193	HG00152	GBR	2347452	110582784	105688428
HG00154.5.M1201317	HG00154	GBR	742094	65141032	62724775
HG00155.1.M1111242	HG00155	GBR	1270900	42583390	40400356
HG00156.4.M1202081	HG00156	GBR	819157	48969934	46900145
HG00157.5.M1201313	HG00157	GBR	848011	79621746	76740292

HG00158.1.M1111248	HG00158	GBR	3732903	60848970	55738921
HG00159.7.M1202196	HG00159	GBR	1195056	69646436	66524018
HG00160.3.M1202021	HG00160	GBR	414819	32851462	31577762
HG00171.3.M1202022	HG00171	FIN	767933	29521242	27908203
HG00173.3.M1202021	HG00173	FIN	2704658	34700564	31062238
HG00174.4.M1202083	HG00174	FIN	683879	55215260	53082274
HG00176.4.M1202082	HG00176	FIN	578127	49661418	47813290
HG00177.4.M1202088	HG00177	FIN	1142310	55116370	52477167
HG00178.4.M1202088	HG00178	FIN	843399	58419390	56016787
HG00179.1.M1111248	HG00179	FIN	801585	49588788	47487851
HG00180.1.M1111248	HG00180	FIN	1158707	65930310	63149406
HG00181.4.M1202084	HG00181	FIN	601315	49443668	47756284
HG00182.1.M1111244	HG00182	FIN	1112763	53296184	50902769
HG00183.1.M1202094	HG00183	FIN	583462	28937892	27687247
HG00183.5.M1201315	HG00183	FIN	727991	60884944	58555422
HG00185.1.M1111245	HG00185	FIN	1150500	61469182	58971505
HG00186.3.M1202025	HG00186	FIN	390786	30752930	29570016
HG00187.1.M1202092	HG00187	FIN	1543424	75204862	71764523
HG00188.1.M1111243	HG00188	FIN	1385905	73841290	70514248
HG00189.1.M1111245	HG00189	FIN	807370	49911214	47922584
HG00231.2.M1112162	HG00231	GBR	925781	57951478	55419089
HG00232.7.M1202193	HG00232	GBR	810358	37284710	35533448
HG00233.2.M1112158	HG00233	GBR	1446518	65033070	61876887
HG00234.7.M1202194	HG00234	GBR	1598824	98890090	94829074
HG00235.1.M1111241	HG00235	GBR	900107	46554128	44570396
HG00236.5.M1201315	HG00236	GBR	753366	69096566	66719637
HG00237.4.M1202081	HG00237	GBR	903467	94738518	91369321
HG00238.5.M1201315	HG00238	GBR	616380	60169628	58138986
HG00239.7.M1202194	HG00239	GBR	1496057	78605382	75145936
HG00240.2.M1112155	HG00240	GBR	1166380	74518950	71538864
HG00242.3.M1202027	HG00242	GBR	501653	46773516	45122738
HG00243.4.M1202082	HG00243	GBR	457465	53934380	52209855
HG00244.5.M1201317	HG00244	GBR	738074	66702862	64436484
HG00245.4.M1202086	HG00245	GBR	496618	57340796	55497878
HG00246.7.M1202194	HG00246	GBR	1242031	67380572	64462349
HG00247.3.M1202021	HG00247	GBR	539055	51939320	50131559
HG00249.3.M1202024	HG00249	GBR	283753	30945204	29841964
HG00250.1.M1111244	HG00250	GBR	763705	45170548	43394933
HG00251.2.M1112168	HG00251	GBR	1218262	64643794	61693659
HG00252.1.M1111245	HG00252	GBR	968620	52522222	50371724
HG00253.5.M1201317	HG00253	GBR	627597	69198938	66754512
HG00255.2.M1112161	HG00255	GBR	852858	61462332	58977565
HG00256.2.M1112157	HG00256	GBR	1487950	70493240	67121379
HG00257.4.M1202085	HG00257	GBR	451909	48474192	46774497
HG00258.1.M1111246	HG00258	GBR	1142808	58644934	56167007
HG00259.5.M1201313	HG00259	GBR	1040636	106082286	102542292
HG00260.5.M1201317	HG00260	GBR	664703	64797710	62620767
HG00261.6.M1201192	HG00261	GBR	724744	58285418	56115725
HG00262.2.M1112158	HG00262	GBR	1371748	75224892	71855416
HG00263.6.M1201193	HG00263	GBR	1008256	60348520	57757935
HG00264.6.M1201195	HG00264	GBR	644394	44161314	42369179
HG00265.2.M1112154	HG00265	GBR	847342	58321994	55837532
HG00266.6.M1201193	HG00266	FIN	1347520	68981404	65436476
HG00267.4.M1202086	HG00267	FIN	648564	54821964	52754538
HG00268.5.M1201313	HG00268	FIN	665061	59784120	57530749
HG00269.2.M1112167	HG00269	FIN	845209	52600642	50217142
HG00271.2.M1112157	HG00271	FIN	1884590	79845062	75823094
HG00272.7.M1202197	HG00272	FIN	1309631	57956092	55126879
HG00273.3.M1202027	HG00273	FIN	385439	32950518	31687519
HG00274.6.M1201191	HG00274	FIN	1313982	63913242	60900413
HG00275.4.M1202088	HG00275	FIN	469184	50226662	48580170
HG00276.2.M1112158	HG00276	FIN	1734072	79210664	75344869
HG00277.1.M1202095	HG00277	FIN	1260583	26420228	24564578
HG00277.3.M1202026	HG00277	FIN	682802	44028638	42190950
HG00278.1.M1111245	HG00278	FIN	1045086	47568268	45197476
HG00280.1.M1111246	HG00280	FIN	2732752	59825588	55668237
HG00281.1.M1111243	HG00281	FIN	1387406	63903074	60880534
HG00282.2.M1112166	HG00282	FIN	1011084	58880784	56249613
HG00284.1.M1111246	HG00284	FIN	940898	67342242	64882358
HG00285.3.M1202025	HG00285	FIN	725484	52050380	50058220
HG00306.1.M1111241	HG00306	FIN	1597320	48143640	45431834
HG00308.3.M1202023	HG00308	FIN	435523	34143244	32902914
HG00309.7.M1202197	HG00309	FIN	2087048	114200140	109280328
HG00310.4.M1202083	HG00310	FIN	900242	61438062	59050194
HG00311.4.M1202086	HG00311	FIN	578451	51388626	49500197
HG00312.7.M1202196	HG00312	FIN	2172850	106376808	101163286
HG00313.1.M1202092	HG00313	FIN	859508	29162420	27632915
HG00313.2.M1112166	HG00313	FIN	676118	52152128	50202199
HG00315.1.M1202095	HG00315	FIN	872817	21958446	20567151
HG00315.2.M1112155	HG00315	FIN	1403657	73003930	69713221
HG00319.4.M1202087	HG00319	FIN	730102	51673018	49602297
HG00320.1.M1111241	HG00320	FIN	1351027	61802814	59069283
HG00321.1.M1202096	HG00321	FIN	446891	20675140	19755084
HG00321.2.M1112162	HG00321	FIN	918483	63143982	60698266
HG00323.1.M1111244	HG00323	FIN	855723	43631078	41834293
HG00324.1.M1111242	HG00324	FIN	1071064	65534436	62795559
HG00325.1.M1202093	HG00325	FIN	599670	24099402	22913549
HG00325.5.M1201311	HG00325	FIN	969883	71143378	68333523
HG00326.1.M1111247	HG00326	FIN	1449980	65268886	62280910
HG00327.3.M1202024	HG00327	FIN	605889	38607132	37016880
HG00328.1.M1202092	HG00328	FIN	969779	31232236	29480799
HG00328.2.M1112154	HG00328	FIN	1198119	74850434	71622833
HG00329.5.M1201313	HG00329	FIN	3697469	47072576	42215543
HG00330.1.M1202092	HG00330	FIN	1250709	59256250	56616506
HG00331.7.M1202198	HG00331	FIN	2082837	73600510	69501694
HG00332.6.M1202171	HG00332	FIN	912096	52284106	50044943
HG00334.2.M1112165	HG00334	FIN	837351	53791954	51398319
HG00335.2.M1112165	HG00335	FIN	978286	59734656	57271720
HG00336.1.M1202092	HG00336	FIN	1248447	54234874	51707478
HG00337.5.M1201315	HG00337	FIN	1091263	49967898	47508611
HG00338.1.M1202093	HG00338	FIN	966212	26710008	25101274
HG00338.2.M1112158	HG00338	FIN	1431693	91968236	88104947
HG00339.4.M1202081	HG00339	FIN	933999	53479508	51229442

HG00341.2.M1112167	HG00341	FIN	832831	58928796	56217599
HG00342.1.M1111241	HG00342	FIN	1531496	69581996	66411774
HG00343.1.M1111245	HG00343	FIN	1365436	74687016	71577271
HG00344.4.M1202082	HG00344	FIN	1007333	53414964	51060988
HG00345.1.M1202091	HG00345	FIN	7296921	57141914	48671545
HG00346.2.M1201311	HG00346	FIN	1538039	70543656	67191831
HG00349.1.M1202094	HG00349	FIN	1064246	37551542	35572251
HG00349.4.M1202084	HG00349	FIN	726786	52240932	50136711
HG00350.3.M1202023	HG00350	FIN	699656	31012760	29552643
HG00351.6.M1201195	HG00351	FIN	723496	39398264	37608021
HG00353.1.M1111246	HG00353	FIN	2120813	58932298	55142603
HG00355.1.M1111248	HG00355	FIN	1024916	61594360	59127694
HG00355.1.M1202091	HG00355	FIN	1208434	53652850	51202171
HG00355.2.M1112156	HG00355	FIN	1210343	80597632	77360556
HG00355.3.M1202026	HG00355	FIN	751249	39555882	37800017
HG00355.4.M1202083	HG00355	FIN	910434	51909912	49809926
HG00355.5.M1201313	HG00355	FIN	778771	62245906	59973443
HG00355.6.M1201191	HG00355	FIN	1186029	60251304	57617719
HG00355.7.M1202196	HG00355	FIN	1168830	58274530	55694138
HG00356.2.M1112156	HG00356	FIN	1736897	77277652	73380693
HG00358.5.M1201315	HG00358	FIN	2390431	60759656	56909880
HG00359.1.M1111246	HG00359	FIN	1262468	62088610	59550399
HG00360.3.M1202028	HG00360	FIN	790509	45173626	43180239
HG00361.7.M1202195	HG00361	FIN	1571896	72096474	68713859
HG00362.3.M1202022	HG00362	FIN	1180230	49530778	47156283
HG00364.2.M1112158	HG00364	FIN	1601088	71431192	67917815
HG00365.7.M1202195	HG00365	FIN	2464643	117130594	111894206
HG00366.4.M1202081	HG00366	FIN	946746	60585410	58045650
HG00367.2.M1201311	HG00367	FIN	1407901	72678706	69209504
HG00369.5.M1201311	HG00369	FIN	836554	66352376	63792095
HG00371.1.M1111243	HG00371	FIN	1036302	52529480	50335015
HG00372.4.M1202083	HG00372	FIN	617389	54274048	52404123
HG00373.2.M1112156	HG00373	FIN	1781698	78237278	74479259
HG00375.1.M1202092	HG00375	FIN	759917	42326842	40503837
HG00375.4.M1202084	HG00375	FIN	583352	59263738	57107128
HG00376.2.M1112161	HG00376	FIN	917245	61387898	59065023
HG00377.1.M1202096	HG00377	FIN	718624	30247010	28726971
HG00377.2.M1201312	HG00377	FIN	1495105	70377550	66855499
HG00378.1.M1202095	HG00378	FIN	1012921	45581642	43489987
HG00378.5.M1201315	HG00378	FIN	880313	68749508	66143028
HG00379.5.M1201313	HG00379	FIN	722772	67850492	65401089
HG00380.3.M1202028	HG00380	FIN	732489	41777474	39847496
HG00381.3.M1202027	HG00381	FIN	1146762	24796800	22981610
HG00382.4.M1202084	HG00382	FIN	456913	52516778	50830374
HG00383.1.M1111241	HG00383	FIN	1330294	67601226	64675410
HG00384.2.M1112164	HG00384	FIN	1465240	82130466	78546332
HG01334.7.M1202198	HG01334	GBR	1668599	83092828	79571096
HG01789.6.M1201191	HG01789	GBR	961017	51020536	48737578
HG01790.3.M1202023	HG01790	GBR	908438	25332966	23880257
HG01791.2.M1112165	HG01791	GBR	754652	64087010	61858818
HG02215.1.M1111244	HG02215	GBR	1042369	61675318	59182056
NA06984.1.M1111244	NA06984	CEU	2541057	44272764	40664227
NA06985.1.M1111247	NA06985	CEU	1089875	66527668	63524029
NA06986.1.M1111247	NA06986	CEU	1027285	53649706	51220829
NA06986.1.M1202091	NA06986	CEU	1301403	52243706	49594840
NA06986.2.M1112154	NA06986	CEU	916428	66099122	63278699
NA06986.3.M1202021	NA06986	CEU	548421	39459940	37813454
NA06986.4.M1202081	NA06986	CEU	799298	49434086	47330542
NA06986.5.M1201315	NA06986	CEU	610584	58366372	56100684
NA06986.6.M1201194	NA06986	CEU	967160	51258606	48911674
NA06986.7.M1202196	NA06986	CEU	1301654	67915336	64760467
NA06989.3.M1202027	NA06989	CEU	467594	41445550	39910934
NA06994.1.M1202094	NA06994	CEU	514164	25284160	24120407
NA06994.2.M1112157	NA06994	CEU	1693698	75751284	72022673
NA07000.1.M1202092	NA07000	CEU	1617287	54424158	51416839
NA07037.1.M1202093	NA07037	CEU	678159	32308424	30870288
NA07037.6.M1201194	NA07037	CEU	796225	52053342	49969416
NA07048.1.M1202098	NA07048	CEU	768296	32374402	30842352
NA07048.6.M1201196	NA07048	CEU	836344	46896124	44949409
NA07051.1.M1202096	NA07051	CEU	607897	29483640	28024209
NA07051.5.M1201311	NA07051	CEU	2157531	60115850	56262522
NA07056.1.M1111243	NA07056	CEU	1036383	57833950	55235228
NA07346.1.M1202094	NA07346	CEU	564687	24245598	23051670
NA07346.2.M1112157	NA07346	CEU	1433980	66517758	63186357
NA07347.1.M1202097	NA07347	CEU	7352918	37148508	29128350
NA07347.7.M1202193	NA07347	CEU	1661568	67598810	64421511
NA07357.1.M1202098	NA07357	CEU	699607	30542664	29058994
NA07357.2.M1112165	NA07357	CEU	760801	52906598	50696419
NA10847.1.M1202097	NA10847	CEU	948180	31836926	30045031
NA10847.4.M1202084	NA10847	CEU	699053	59587646	57192390
NA10851.1.M1202091	NA10851	CEU	738868	29681638	28164061
NA10851.4.M1202081	NA10851	CEU	657186	49702518	47689333
NA11829.1.M1202097	NA11829	CEU	536534	32059754	30596880
NA11829.6.M1202171	NA11829	CEU	690659	39104832	37226370
NA11830.1.M1202096	NA11830	CEU	881532	40386354	38492170
NA11830.3.M1202024	NA11830	CEU	801426	48715912	46549298
NA11831.1.M1202095	NA11831	CEU	7146741	45612002	37511561
NA11831.7.M1202192	NA11831	CEU	1828048	86943042	82912252
NA11832.1.M1202097	NA11832	CEU	747346	30181644	28679322
NA11832.2.M1201312	NA11832	CEU	1613840	70076616	66589784
NA11840.1.M1202094	NA11840	CEU	701856	24732602	23392390
NA11840.6.M1201195	NA11840	CEU	1139182	53223308	50628261
NA11843.1.M1202095	NA11843	CEU	813859	33349504	31638506
NA11843.4.M1202086	NA11843	CEU	633716	57414544	55146924
NA11881.1.M1111245	NA11881	CEU	1282675	67230954	64247985
NA11892.1.M1111242	NA11892	CEU	978546	48985566	46700450
NA11893.1.M1202094	NA11893	CEU	553256	27479310	26236769
NA11893.7.M1202193	NA11893	CEU	1472326	85643402	81961159
NA11894.1.M1202094	NA11894	CEU	429247	23315654	22292244
NA11894.2.M1112154	NA11894	CEU	1049171	65162588	62282935
NA11918.1.M1111243	NA11918	CEU	829342	48647220	46585815
NA11920.2.M1201311	NA11920	CEU	1672480	62990540	59569087
NA11930.1.M1202098	NA11930	CEU	7371097	42003756	33775440

NA11930.3.M1202028	NA11930	CEU	608595	42256998	40577631
NA11931.1.M1111248	NA11931	CEU	1237330	51250304	48539575
NA11992.1.M1202096	NA11992	CEU	1261294	39641152	37310298
NA11992.6.M1201195	NA11992	CEU	755932	45563128	43559943
NA11993.2.M1112164	NA11993	CEU	1012255	61487248	58799152
NA11994.1.M1202093	NA11994	CEU	808570	39562466	37777612
NA11994.2.M1112167	NA11994	CEU	1164480	84995764	81616942
NA11995.1.M1202095	NA11995	CEU	471356	28791980	27573718
NA11995.7.M1202195	NA11995	CEU	1234135	54004124	51350376
NA12004.1.M1202098	NA12004	CEU	650258	31096898	29635147
NA12004.4.M1202085	NA12004	CEU	2436741	47186642	43560691
NA12005.4.M1202081	NA12005	CEU	541497	55113076	53089896
NA12006.1.M1202095	NA12006	CEU	729594	32191986	30638099
NA12006.5.M1201311	NA12006	CEU	682176	62427064	60033862
NA12043.1.M1202096	NA12043	CEU	813363	37878564	36126962
NA12043.2.M1201277	NA12043	CEU	752794	58858732	56470892
NA12044.4.M1202088	NA12044	CEU	593497	60289002	57971799
NA12045.1.M1202093	NA12045	CEU	784194	32216162	30655830
NA12045.3.M1202023	NA12045	CEU	765840	37476622	35822170
NA12058.1.M1202097	NA12058	CEU	834375	36768270	35068219
NA12058.5.M1201317	NA12058	CEU	628035	63238230	61107459
NA12144.1.M1202093	NA12144	CEU	7524036	37811646	29558208
NA12144.4.M1202081	NA12144	CEU	1094169	63724788	61059790
NA12154.1.M1202093	NA12154	CEU	782529	31569258	30048739
NA12154.5.M1201313	NA12154	CEU	684747	53236140	51226581
NA12155.1.M1202095	NA12155	CEU	797683	36593020	34917613
NA12155.4.M1202088	NA12155	CEU	638052	44337836	42565076
NA12156.1.M1202097	NA12156	CEU	1220027	43855948	41510158
NA12156.4.M1202087	NA12156	CEU	944156	58845272	56345779
NA12234.1.M1202098	NA12234	CEU	505938	33583232	32187333
NA12234.7.M1202196	NA12234	CEU	1136419	68921256	65977024
NA12249.1.M1202091	NA12249	CEU	1215677	57229250	54544449
NA12272.1.M1202095	NA12272	CEU	543693	29563466	28285278
NA12272.3.M1202025	NA12272	CEU	954314	37308646	35379224
NA12273.1.M1111244	NA12273	CEU	1930460	64977326	61362746
NA12275.1.M1202094	NA12275	CEU	986574	24314976	22705231
NA12275.7.M1202195	NA12275	CEU	1787048	62722820	59266182
NA12282.1.M1202096	NA12282	CEU	470209	25766822	24559102
NA12282.3.M1202027	NA12282	CEU	405171	33102104	31675698
NA12283.1.M1202094	NA12283	CEU	727579	22534706	21246734
NA12283.2.M1112165	NA12283	CEU	1009106	52007306	49652716
NA12286.1.M1202096	NA12286	CEU	860663	26830438	25258877
NA12286.2.M1112161	NA12286	CEU	1495991	60585758	57375686
NA12287.1.M1202096	NA12287	CEU	943415	26671348	25123621
NA12287.3.M1202023	NA12287	CEU	463513	34070922	32852090
NA12340.1.M1202093	NA12340	CEU	962514	31851306	30019957
NA12340.5.M1201317	NA12340	CEU	723880	61284172	58771792
NA12341.1.M1111245	NA12341	CEU	1422818	62894450	59837245
NA12342.1.M1111245	NA12342	CEU	5038486	58045918	51613452
NA12347.1.M1111247	NA12347	CEU	1500457	63877988	60733387
NA12348.1.M1202093	NA12348	CEU	1006490	22379028	20842844
NA12348.3.M1202026	NA12348	CEU	675997	28331306	26904703
NA12383.1.M1111243	NA12383	CEU	1005674	55083068	52672640
NA12399.1.M1202094	NA12399	CEU	1375890	29453468	27322198
NA12399.7.M1202191	NA12399	CEU	4650459	146767882	138323147
NA12400.1.M1202097	NA12400	CEU	1159393	26369656	24532465
NA12400.2.M1112162	NA12400	CEU	1140767	50613994	48016199
NA12413.1.M1202096	NA12413	CEU	686352	22990910	21725011
NA12413.2.M1112156	NA12413	CEU	1693506	75916544	72099679
NA12489.1.M1111242	NA12489	CEU	4614731	67642678	61358191
NA12546.1.M1202096	NA12546	CEU	7862860	28236186	19866661
NA12546.3.M1202024	NA12546	CEU	1024630	37335380	35357510
NA12716.1.M1202095	NA12716	CEU	1391067	42124014	39715743
NA12716.7.M1202196	NA12716	CEU	4032998	157239402	148843563
NA12717.1.M1111243	NA12717	CEU	1321021	56071888	53230370
NA12718.1.M1111247	NA12718	CEU	1555797	69986762	66729621
NA12749.1.M1111242	NA12749	CEU	982205	47629086	45588581
NA12750.1.M1202097	NA12750	CEU	878457	25668122	24165460
NA12750.2.M1112162	NA12750	CEU	741336	56406698	54097915
NA12751.1.M1202096	NA12751	CEU	837239	30193600	28628916
NA12751.2.M1112158	NA12751	CEU	1718540	73191558	69507256
NA12760.1.M1202095	NA12760	CEU	656358	27284570	25952227
NA12760.3.M1202023	NA12760	CEU	376643	29995518	28894333
NA12761.1.M1202093	NA12761	CEU	708131	25801218	24493385
NA12761.6.M1201192	NA12761	CEU	925005	59918966	57505186
NA12762.1.M1202095	NA12762	CEU	2222100	36597736	33558082
NA12762.4.M1202083	NA12762	CEU	695218	48182592	46344110
NA12763.1.M1202095	NA12763	CEU	542301	22455894	21380847
NA12763.2.M1112168	NA12763	CEU	1069667	63667750	60888371
NA12775.1.M1202094	NA12775	CEU	6236026	28394024	21532330
NA12775.7.M1202195	NA12775	CEU	3099117	97212116	91357813
NA12776.1.M1202093	NA12776	CEU	1135993	22743282	21054957
NA12776.6.M1201196	NA12776	CEU	1465831	61771810	58639052
NA12777.1.M1202094	NA12777	CEU	641368	29398350	27952492
NA12777.4.M1202086	NA12777	CEU	490979	51273808	49311640
NA12778.1.M1111244	NA12778	CEU	4645786	59227324	53196257
NA12812.1.M1111244	NA12812	CEU	796171	44647732	42822044
NA12813.1.M1202096	NA12813	CEU	1178566	26241158	24439211
NA12813.2.M1112166	NA12813	CEU	843789	51803402	49591230
NA12814.1.M1202093	NA12814	CEU	766696	35313330	33751016
NA12814.7.M1202193	NA12814	CEU	1476289	89433746	85879138
NA12815.1.M1111243	NA12815	CEU	1723726	74284182	70590711
NA12827.1.M1202098	NA12827	CEU	656752	23489612	22244967
NA12827.3.M1202022	NA12827	CEU	422444	30801584	29575941
NA12829.1.M1202094	NA12829	CEU	653417	26240620	24933843
NA12829.7.M1202194	NA12829	CEU	1640141	82367814	78632478
NA12830.1.M1111242	NA12830	CEU	707555	41184128	39460984
NA12842.1.M1202095	NA12842	CEU	578882	28664450	27439826
NA12842.6.M1201194	NA12842	CEU	958363	65530232	63045175
NA12843.1.M1202095	NA12843	CEU	622558	20469766	19364955
NA12843.6.M1201194	NA12843	CEU	728628	44197168	42283889
NA12872.1.M1202093	NA12872	CEU	1185557	32769378	30832771
NA12872.2.M1112162	NA12872	CEU	1299194	64209084	61267310

NA12873.1.M1202097	NA12873	CEU	1090229	23608376	21935062
NA12873.7.M1202191	NA12873	CEU	2193540	69998210	65986258
NA12874.1.M1202093	NA12874	CEU	467501	23604870	22561940
NA12874.3.M1202022	NA12874	CEU	495824	55673256	53690157
NA12889.1.M1202094	NA12889	CEU	741035	27449168	26007988
NA12889.3.M1202237	NA12889	CEU	13057010	81745992	66682723
NA12890.1.M1111243	NA12890	CEU	3812255	56224248	51053484
NA18486.1.M1202096	NA18486	YRI	546872	30091826	28854859
NA18486.7.M1202194	NA18486	YRI	2487061	94159158	89435145
NA18487.1.M1202097	NA18487	YRI	740070	34987964	33451785
NA18487.6.M1201195	NA18487	YRI	761924	44275528	42489000
NA18488.1.M1202098	NA18488	YRI	690450	28457404	27076061
NA18488.4.M1202087	NA18488	YRI	634443	36480658	34908449
NA18489.1.M1202098	NA18489	YRI	576087	27651430	26481968
NA18489.2.M1112161	NA18489	YRI	800135	52162678	50131322
NA18498.1.M1202092	NA18498	YRI	6401837	27781720	20882415
NA18498.2.M1201311	NA18498	YRI	1667741	72699928	69203972
NA18499.1.M1111242	NA18499	YRI	1015040	64316356	61832072
NA18502.1.M1202096	NA18502	YRI	785012	28938892	27501338
NA18502.7.M1202198	NA18502	YRI	3286941	113740072	107761739
NA18505.1.M1202094	NA18505	YRI	752045	32068630	30616719
NA18505.3.M1202021	NA18505	YRI	1369183	31203700	29094659
NA18508.1.M1111241	NA18508	YRI	1073954	51680710	49353199
NA18510.3.M1202027	NA18510	YRI	540956	41294578	39761588
NA18511.1.M1202091	NA18511	YRI	454649	26988772	25928585
NA18511.7.M1202192	NA18511	YRI	2760040	107587174	102297612
NA18517.1.M1202097	NA18517	YRI	787412	25942230	24486268
NA18517.3.M1202021	NA18517	YRI	2603814	30457778	27078338
NA18519.1.M1111243	NA18519	YRI	878229	48033590	46000678
NA18520.1.M1111241	NA18520	YRI	1052477	53217190	51047437
NA18858.1.M1202097	NA18858	YRI	7417591	41246304	32938824
NA18858.2.M1112168	NA18858	YRI	1026927	58240252	55560270
NA18861.1.M1202092	NA18861	YRI	755384	32249850	30771111
NA18861.4.M1202085	NA18861	YRI	722266	50615720	48066771
NA18867.1.M1111242	NA18867	YRI	3635849	56567642	51846383
NA18868.1.M1202097	NA18868	YRI	689232	30762466	29366828
NA18868.5.M1201313	NA18868	YRI	481967	59639732	57708621
NA18870.1.M1202094	NA18870	YRI	781887	31979368	30478087
NA18870.6.M1201194	NA18870	YRI	897463	46007508	44007519
NA18873.1.M1202096	NA18873	YRI	666307	30937178	29526237
NA18873.4.M1202087	NA18873	YRI	685147	52677010	50678941
NA18907.1.M1202091	NA18907	YRI	702851	35090840	33600948
NA18907.6.M1202171	NA18907	YRI	2984775	49814026	45668105
NA18908.1.M1202098	NA18908	YRI	7069507	38656240	30928852
NA18908.7.M1202197	NA18908	YRI	1162491	67258558	64581693
NA18909.1.M1202098	NA18909	YRI	893969	30875610	29286536
NA18909.4.M1202088	NA18909	YRI	991507	56130518	53714052
NA18910.1.M1202095	NA18910	YRI	769182	35553956	34004428
NA18910.3.M1202027	NA18910	YRI	1076429	35873940	33966465
NA18912.1.M1202095	NA18912	YRI	935148	36505036	34721286
NA18912.7.M1202193	NA18912	YRI	1812232	91363368	87325612
NA18916.1.M1202097	NA18916	YRI	765457	36217570	34558999
NA18916.2.M1112156	NA18916	YRI	1316649	82615366	79139060
NA18917.1.M1111245	NA18917	YRI	898028	53942416	51764508
NA18923.1.M1202096	NA18923	YRI	625596	32529660	31095621
NA18923.2.M1112165	NA18923	YRI	739258	62831466	60419340
NA18933.1.M1111242	NA18933	YRI	1277885	46992810	44681519
NA18934.1.M1202093	NA18934	YRI	7120174	41845334	33943370
NA18934.3.M1202237	NA18934	YRI	2001916	97198420	92841927
NA19092.1.M1111243	NA19092	YRI	1169082	69954692	67198906
NA19093.1.M1202097	NA19093	YRI	734898	32406668	30978634
NA19093.7.M1202197	NA19093	YRI	1112907	66525900	63901808
NA19095.1.M1111248	NA19095	YRI	958786	63781962	61378338
NA19095.1.M1202091	NA19095	YRI	1371899	63705330	60914601
NA19095.2.M1201312	NA19095	YRI	1970329	84139054	80132537
NA19095.3.M1202021	NA19095	YRI	599099	38738526	37223429
NA19095.4.M1202082	NA19095	YRI	701006	41379842	39770091
NA19095.5.M1201315	NA19095	YRI	628539	61447466	59308564
NA19095.6.M1201194	NA19095	YRI	835826	53431154	51375555
NA19095.7.M1202198	NA19095	YRI	1181819	49312350	47009267
NA19096.1.M1202098	NA19096	YRI	679613	28342460	27014374
NA19096.4.M1202084	NA19096	YRI	505610	48896484	47197700
NA19098.5.M1201311	NA19098	YRI	788713	79091478	76520463
NA19099.1.M1202094	NA19099	YRI	487662	23438620	22460253
NA19099.2.M1112157	NA19099	YRI	1581226	75752600	72391829
NA19102.1.M1111241	NA19102	YRI	929919	49235462	47210156
NA19107.1.M1202094	NA19107	YRI	775935	37918418	36329958
NA19107.2.M1112165	NA19107	YRI	743966	54005276	52023844
NA19108.1.M1202094	NA19108	YRI	6983460	50855512	42941900
NA19108.5.M1201317	NA19108	YRI	1387580	98151102	94619690
NA19113.1.M1202091	NA19113	YRI	1130215	44703518	42578879
NA19113.3.M1202021	NA19113	YRI	1201208	30804884	28908837
NA19114.1.M1202098	NA19114	YRI	909412	41503732	39622315
NA19114.2.M1112157	NA19114	YRI	1393051	62100390	59256532
NA19116.1.M1202093	NA19116	YRI	715148	34167284	32569306
NA19116.2.M1112157	NA19116	YRI	1609082	77214832	73610780
NA19117.1.M1111244	NA19117	YRI	1067822	66551954	63998574
NA19118.1.M1202095	NA19118	YRI	861450	34496750	32751834
NA19118.7.M1202198	NA19118	YRI	2653966	95610272	90533813
NA19119.1.M1202095	NA19119	YRI	689621	30892858	29496781
NA19119.3.M1202024	NA19119	YRI	313856	32705756	31579620
NA19121.1.M1202096	NA19121	YRI	793864	44170020	42272143
NA19121.3.M1202026	NA19121	YRI	882659	48867404	46626901
NA19129.1.M1202096	NA19129	YRI	6499666	41752726	34373650
NA19129.6.M1202171	NA19129	YRI	931321	47718096	45579140
NA19130.1.M1202091	NA19130	YRI	1015625	36496846	34575684
NA19130.5.M1201317	NA19130	YRI	709294	53927418	51809009
NA19131.1.M1202093	NA19131	YRI	694556	36181184	34674022
NA19131.2.M1112158	NA19131	YRI	1487748	76044278	72764141
NA19137.1.M1111247	NA19137	YRI	3347374	64596650	59692694
NA19138.1.M1202097	NA19138	YRI	567080	25018412	23894997
NA19138.3.M1202028	NA19138	YRI	785540	40445678	38710935
NA19141.1.M1202091	NA19141	YRI	688008	39049606	37324085

NA19141.6.M1201192	NA19141	YRI	868874	57160634	54729755
NA19143.1.M1202094	NA19143	YRI	462505	32056936	30816605
NA19143.2.M1112168	NA19143	YRI	1186604	66746326	63723187
NA19144.1.M1202091	NA19144	YRI	746849	37208028	35589470
NA19144.4.M1202082	NA19144	YRI	806361	48314536	46064324
NA19146.1.M1202098	NA19146	YRI	749022	35299478	33768452
NA19146.5.M1201315	NA19146	YRI	701961	68335568	66007383
NA19147.2.M1201312	NA19147	YRI	1885058	91607896	87478858
NA19149.1.M1111242	NA19149	YRI	1063932	52899822	50658698
NA19150.1.M1202097	NA19150	YRI	677978	38844354	37276872
NA19150.5.M1201317	NA19150	YRI	505172	59069450	57221045
NA19152.1.M1202093	NA19152	YRI	708818	30190086	28739650
NA19152.4.M1202084	NA19152	YRI	657902	55519748	53420523
NA19153.1.M1111247	NA19153	YRI	6972240	71540736	63046762
NA19159.1.M1202098	NA19159	YRI	661279	29193900	27862994
NA19159.3.M1202236	NA19159	YRI	3133144	166643202	159267717
NA19160.1.M1202095	NA19160	YRI	6627986	35955522	28658699
NA19160.3.M1202023	NA19160	YRI	522494	38940964	37565198
NA19171.1.M1111247	NA19171	YRI	1801954	72147472	68743842
NA19172.1.M1202098	NA19172	YRI	974384	35716758	33919148
NA19172.6.M1201191	NA19172	YRI	966576	51461332	49251422
NA19175.1.M1202092	NA19175	YRI	966273	34478264	32696848
NA19175.4.M1202088	NA19175	YRI	1266429	57601384	54886441
NA19184.1.M1111241	NA19184	YRI	1394492	75131880	72020503
NA19185.1.M1202094	NA19185	YRI	599690	28535644	27243396
NA19185.6.M1201194	NA19185	YRI	924237	58016328	55589600
NA19189.1.M1202098	NA19189	YRI	738061	38548800	36963953
NA19189.5.M1201317	NA19189	YRI	699545	66763910	64591812
NA19190.1.M1202098	NA19190	YRI	758232	37325236	35628714
NA19190.4.M1202082	NA19190	YRI	622506	57409818	55340294
NA19197.1.M1202093	NA19197	YRI	848627	34598466	32865875
NA19197.3.M1202022	NA19197	YRI	664168	40357220	38615016
NA19198.1.M1202093	NA19198	YRI	1092953	34767974	32808474
NA19198.2.M1112156	NA19198	YRI	1640335	65593110	62281164
NA19200.1.M1202096	NA19200	YRI	771182	35994658	34426071
NA19200.2.M1112161	NA19200	YRI	759117	56585102	54528843
NA19201.1.M1111243	NA19201	YRI	1238615	60908980	58326549
NA19204.1.M1111246	NA19204	YRI	1118926	51301070	49161636
NA19206.1.M1202092	NA19206	YRI	717078	26310108	24842610
NA19206.2.M1112155	NA19206	YRI	1579214	75803142	71968105
NA19207.1.M1202098	NA19207	YRI	750381	23631760	22306606
NA19207.4.M1202085	NA19207	YRI	633204	57760084	55696912
NA19209.1.M1202092	NA19209	YRI	744334	35587584	34011763
NA19209.6.M1202171	NA19209	YRI	612491	44884530	43195870
NA19210.1.M1202091	NA19210	YRI	1024271	34978590	33148156
NA19210.4.M1202081	NA19210	YRI	526065	43178400	41582911
NA19213.1.M1202091	NA19213	YRI	6605000	40562894	33208194
NA19213.2.M1112156	NA19213	YRI	1549155	71368052	68172822
NA19214.1.M1202097	NA19214	YRI	1038660	43024104	40996641
NA19214.6.M1201192	NA19214	YRI	852827	57983370	55754520
NA19222.1.M1202091	NA19222	YRI	699576	36104426	34424293
NA19222.2.M1112167	NA19222	YRI	935195	64371830	61574224
NA19223.1.M1202091	NA19223	YRI	852803	37198786	35477678
NA19223.7.M1202196	NA19223	YRI	1869615	92228512	88194023
NA19225.1.M1202092	NA19225	YRI	916024	47618624	45565620
NA19225.6.M1201195	NA19225	YRI	834298	57960778	55734819
NA19235.1.M1111246	NA19235	YRI	855494	70264650	67926401
NA19236.1.M1202095	NA19236	YRI	961511	33356244	31604451
NA19236.4.M1202081	NA19236	YRI	734626	42875714	41150748
NA19247.1.M1202092	NA19247	YRI	887882	35936012	34128994
NA19247.3.M1202237	NA19247	YRI	4190710	65714138	59781856
NA19248.1.M1202092	NA19248	YRI	1068769	33725644	31862989
NA19248.5.M1201311	NA19248	YRI	1058983	60364088	57790440
NA19256.1.M1202093	NA19256	YRI	737365	28249682	26802851
NA19256.6.M1201196	NA19256	YRI	826371	57800698	55566909
NA19257.1.M1202091	NA19257	YRI	713380	29762202	28322271
NA19257.6.M1201192	NA19257	YRI	1175300	63081626	60270689
NA20502.3.M1202024	NA20502	TSI	633546	46292322	44530139
NA20503.1.M1111245	NA20503	TSI	890461	50576464	48520549
NA20504.1.M1111247	NA20504	TSI	873840	52388216	50156206
NA20505.1.M1111246	NA20505	TSI	1386864	75876982	72937909
NA20506.1.M1202092	NA20506	TSI	1235196	48177866	45694986
NA20507.1.M1111247	NA20507	TSI	915012	54090610	51788916
NA20508.1.M1111242	NA20508	TSI	1504226	65015656	62135970
NA20509.7.M1202197	NA20509	TSI	3495647	116113638	109766384
NA20510.3.M1202023	NA20510	TSI	406400	31600254	30466351
NA20512.1.M1202098	NA20512	TSI	712457	39175038	37498931
NA20512.6.M1201196	NA20512	TSI	716226	48923668	46941086
NA20513.1.M1202096	NA20513	TSI	1016314	47900866	45822184
NA20513.4.M1202081	NA20513	TSI	614005	54588288	52677025
NA20514.1.M1111244	NA20514	TSI	874892	52609404	50501234
NA20515.5.M1201311	NA20515	TSI	1185628	72288146	68765179
NA20516.5.M1201313	NA20516	TSI	644602	58987376	56768351
NA20517.3.M1202028	NA20517	TSI	651552	42745682	41096770
NA20518.1.M1202096	NA20518	TSI	2328207	37620722	34448089
NA20518.4.M1202083	NA20518	TSI	671751	60964606	58861539
NA20519.1.M1111245	NA20519	TSI	1217384	73240716	70346932
NA20520.3.M1202022	NA20520	TSI	837383	28492330	26885814
NA20521.7.M1202197	NA20521	TSI	1160805	47577258	45157629
NA20524.2.M1112158	NA20524	TSI	1261407	64680104	61686490
NA20525.1.M1111241	NA20525	TSI	1037883	56474686	54020530
NA20527.1.M1111246	NA20527	TSI	1089908	72760118	69963798
NA20527.1.M1202091	NA20527	TSI	1186425	62452370	59758406
NA20527.2.M1112157	NA20527	TSI	1736132	78156530	74353805
NA20527.3.M1202023	NA20527	TSI	367109	27020918	26023917
NA20527.4.M1202082	NA20527	TSI	966546	53374222	51106626
NA20527.5.M1201311	NA20527	TSI	741658	69314238	66800192
NA20527.6.M1201195	NA20527	TSI	691849	42369242	40655587
NA20527.7.M1202198	NA20527	TSI	1475223	57549132	54662278
NA20528.2.M1201311	NA20528	TSI	1669474	64047602	60748869
NA20529.2.M1112156	NA20529	TSI	1599382	62725976	59236410
NA20530.2.M1112156	NA20530	TSI	1148763	73163546	69900982
NA20531.4.M1202081	NA20531	TSI	598596	56589316	54661981

NA20532.1.M1202095	NA20532	TSI	1697847	37047316	34448800
NA20532.3.M1202027	NA20532	TSI	2295902	40565342	37266815
NA20534.2.M1112158	NA20534	TSI	1899371	69522208	65790691
NA20535.3.M1202022	NA20535	TSI	633996	51054612	49085769
NA20536.1.M1111241	NA20536	TSI	1164949	48563226	46297925
NA20537.3.M1202028	NA20537	TSI	904841	42493642	40481686
NA20538.1.M1202095	NA20538	TSI	610884	23633388	22427211
NA20538.3.M1202026	NA20538	TSI	614099	30945634	29451080
NA20539.4.M1202085	NA20539	TSI	736229	54419362	52261367
NA20540.1.M1111242	NA20540	TSI	1003907	43122492	41077962
NA20541.1.M1111244	NA20541	TSI	1286550	70862334	67783961
NA20542.7.M1202192	NA20542	TSI	1485413	40504508	37927522
NA20543.1.M1202098	NA20543	TSI	1138898	38104108	35996270
NA20543.2.M1112155	NA20543	TSI	744267	53774994	51554174
NA20544.4.M1202085	NA20544	TSI	583495	53373484	51281171
NA20581.1.M1111244	NA20581	TSI	759189	46565118	44789200
NA20582.4.M1202082	NA20582	TSI	630883	63922312	61608501
NA20585.3.M1202026	NA20585	TSI	505441	38528754	37087075
NA20586.2.M1112157	NA20586	TSI	1536637	63740492	60150210
NA20588.5.M1201315	NA20588	TSI	823315	61972428	59450006
NA20589.1.M1111243	NA20589	TSI	1248011	60526676	57879685
NA20752.5.M1201317	NA20752	TSI	679320	70510870	68136769
NA20754.1.M1202098	NA20754	TSI	655555	29684412	28376610
NA20754.4.M1202081	NA20754	TSI	529139	49148590	47431969
NA20756.1.M1202094	NA20756	TSI	742828	33035010	31539850
NA20756.2.M1112166	NA20756	TSI	747127	56013648	53752911
NA20757.1.M1111241	NA20757	TSI	1418589	60807078	57864551
NA20758.2.M1112158	NA20758	TSI	1912519	82746652	78492271
NA20759.4.M1202083	NA20759	TSI	637270	46341942	44631092
NA20760.3.M1202025	NA20760	TSI	811560	37635188	35807344
NA20761.1.M1111247	NA20761	TSI	1528755	58661614	55688979
NA20765.1.M1202097	NA20765	TSI	979334	38602064	36656920
NA20765.2.M1112155	NA20765	TSI	1449420	86041348	82389654
NA20766.1.M1111244	NA20766	TSI	1158508	58778104	56234749
NA20768.5.M1201311	NA20768	TSI	1671637	60039132	56819741
NA20769.6.M1201195	NA20769	TSI	782315	42844788	41047256
NA20770.3.M1202021	NA20770	TSI	535786	33257732	31881288
NA20771.1.M1202094	NA20771	TSI	1018433	31960800	30205387
NA20771.2.M1112157	NA20771	TSI	1463564	58780634	55747308
NA20772.3.M1202236	NA20772	TSI	8341509	79218158	68688634
NA20773.1.M1202097	NA20773	TSI	973379	33648318	31854306
NA20773.6.M1201194	NA20773	TSI	848575	48591958	46518112
NA20774.7.M1202191	NA20774	TSI	1393495	42082872	39627515
NA20778.4.M1202081	NA20778	TSI	268583	17279354	16593404
NA20783.4.M1202086	NA20783	TSI	604998	60799330	58712555
NA20785.4.M1202081	NA20785	TSI	733696	54310656	52204325
NA20786.1.M1202098	NA20786	TSI	694653	32119758	30719998
NA20786.2.M1112158	NA20786	TSI	1403649	65654308	62684411
NA20787.6.M1201193	NA20787	TSI	959170	65645640	63047408
NA20790.2.M1112156	NA20790	TSI	928292	54023024	51730387
NA20792.6.M1201196	NA20792	TSI	949584	52821590	50631144
NA20795.5.M1201311	NA20795	TSI	769118	72875624	70416161
NA20796.1.M1202092	NA20796	TSI	2289332	103677232	99030240
NA20797.2.M1112156	NA20797	TSI	1039820	79094216	75634812
NA20798.1.M1202097	NA20798	TSI	830705	39045770	37297027
NA20798.6.M1201196	NA20798	TSI	550778	40597892	39092754
NA20799.1.M1111243	NA20799	TSI	1582489	64787760	61613844
NA20800.1.M1111245	NA20800	TSI	1464204	73763766	70555905
NA20801.7.M1202195	NA20801	TSI	1654123	76258848	72839213
NA20802.1.M1111247	NA20802	TSI	927102	57581676	55162916
NA20803.7.M1202191	NA20803	TSI	3096589	112168988	106102448
NA20804.4.M1202081	NA20804	TSI	774218	54276740	52172136
NA20805.4.M1202087	NA20805	TSI	389860	34905740	33669593
NA20806.3.M1202025	NA20806	TSI	653451	35963954	34295353
NA20807.5.M1201311	NA20807	TSI	982007	67671066	65092361
NA20808.4.M1202086	NA20808	TSI	596061	54180286	52187916
NA20809.6.M1201192	NA20809	TSI	742592	50484026	48464472
NA20810.2.M1112157	NA20810	TSI	1456100	71188628	67774556
NA20811.1.M1111245	NA20811	TSI	880562	48797942	46708256
NA20812.2.M1112166	NA20812	TSI	666723	60337382	58268267
NA20813.5.M1201311	NA20813	TSI	898268	68032742	65399777
NA20814.2.M1112156	NA20814	TSI	953715	65684348	63051875
NA20815.5.M1201315	NA20815	TSI	665495	62647002	60471181
NA20816.3.M1202027	NA20816	TSI	1272231	53197162	50608890
NA20819.3.M1202022	NA20819	TSI	1160829	42423376	40233352
NA20826.1.M1111241	NA20826	TSI	1005035	57621730	55265703
NA20828.2.M1112168	NA20828	TSI	1198492	89680648	85978900

Table 2: Sequence statistics for GEUVADIS samples analysed in Chapter 4.

Sample name	Total reads	Unique mapping reads	Multi mapping Reads	Unmapped reads	Patient	Region
D10	64701312	48700677	905818	15043055	038/09	HYPO
D11	72201033	53594826	902512	17660372	008/10	SPCO
D12	51161381	40161684	793001	10181114	038/09	CRBL
B12	39674488	32699713	499898	6462974	034/09	OCTX
B10	122871958	103175583	1720207	17951593	031/09	FCTXCON
B11	38772376	32355547	562199	5842997	038/09	FCTXCON
C12	60764815	49316723	923625	10506236	008/10	HIPP
C11	74422308	61219790	1056796	12130836	008/10	FCTXCON
C10	58431937	49117886	987499	8297335	008/10	CRBL
A11	45950648	38097682	574383	7273987	038/09	MEDU
A10	80957108	65421438	906719	14612757	034/09	THAL
A12	77120514	62606433	1133671	13372697	038/09	THAL
C8	67743931	54154498	975512	12600371	008/10	PUTM
C9	65525365	53999453	937012	10569241	008/10	TCTX
A1	64270292	53781380	951200	9518430	034/09	CRBL
A3	67397213	54679358	768328	11936046	034/09	FCTXCON
A2	62170292	49817054	789562	11545023	031/09	PUTM
A5	58341045	47343758	700092	10291360	034/09	PUTM
A4	59525146	49149913	714301	9649026	031/09	OCTX
A7	64145906	51983842	718434	11430800	034/09	TCTX
A6	60778108	49242423	784037	10739491	031/09	HIPP
C3	63805190	52065035	1001741	10719271	038/09	WHMT
A8	62254243	51272594	977391	9985580	038/09	PUTM
C1	60503694	49207654	792598	10491340	031/09	TCTX
C7	67246545	56991446	988524	9253124	008/10	THAL
C6	66879537	53122416	936313	12807431	008/10	SNIG
C5	66601578	55685579	959062	9936955	038/09	OCTX
C4	70032920	58547521	1085510	10385882	038/09	HIPP
D8	76677023	60774208	835779	15044031	034/09	HYPO
B3	72673896	56540291	1010167	15123437	038/09	SNIG
A9	59717189	48836717	758408	10116091	034/09	HIPP
C2	59945948	49437423	1438702	9051838	038/09	TCTX
B8	65649228	54849929	833745	9952422	031/09	THAL
D3	73385630	59779934	1042075	12534265	008/10	MEDU
B4	60320675	45759264	621302	13934075	034/09	MEDU
B5	65012072	52607768	864660	11533141	031/09	WHMT
B6	58277446	49611589	967405	7675139	031/09	CRBL
B7	60177454	48075768	728147	11361503	031/09	SNIG
B1	71112243	57259578	760901	13084652	034/09	SNIG
B2	79439930	63218296	929447	15276298	031/09	MEDU
D9	65775337	51140324	934009	13654959	038/09	SPCO
D6	68629712	55466533	967678	12161184	031/09	HYPO
D7	70023053	56935744	945311	12113988	034/09	SPCO
D4	68447238	54874150	965106	12580602	031/09	SPCO
D5	70482205	57548720	1134763	11770528	008/10	HYPO
D2	67810799	53278944	901883	13609627	008/10	WHMT
B9	67515639	54005759	884454	12611921	034/09	WHMT
D1	66632426	52579647	892874	13126587	008/10	OCTX

Table 1: Read statistics from UKBEC post mortem brain data.