Genome-wide Survey and Analysis of Allele-specific mRNA Splicing in Human and Mouse

Victoria Precious Nembaware

This thesis is submitted in fulfilment of the requirements for the degree of *Doctor Philosophiae* at the National Bioinformatics Node, Molecular and Cellular Biology Department, Faculty of Science, University of Cape town

July 2008

Advisor: Professor Cathal Seoighe

Abstract

Accumulating evidence suggest that the mRNA splicing process is tightly governed by complex interactions between numerous *trans* and *cis* acting factors. The importance of mRNA splicing and the transcript diversity generated through alternative mRNA splicing is demonstrated by the high frequency of disease causing mutations that alter splicing through disruption of the splice-regulatory factors. Despite an increasing number of splicing mutations that influence pharmacogenetic responses and hereditary diseases, the effect of genetic polymorphisms on splicing remains a largely under-explored area.

This dissertation aims to examine allele-specific splicing in human and mouse using publicly available datasets. Such datasets, which have been generated from multiple tissue sources and from individuals of diverse backgrounds, are rich and cheap reservoirs of transcript isoforms resulting from alternative splicing as well as isoforms resulting from mutations or polymorphisms (allele-specific isoforms). Published tools were used to analyse microarray and genomic data. However, for the assessment of allele-specific splicing using publicly available high-throughput transcript sequences, we present two novel methods: a heuristic method for quantifying the prevalence of allele-specific splicing and a more sophisticated maximum likelihood method for the detection of individual examples of allele-specific splicing. These methods make use of transcripts that can be mapped to both polymorphisms and computationally predicted mRNA isoforms. Inference of polymorphic alleles from transcripts is laborious hence a pre-computed database was created for the human data and made publicly available for use by the wider research community.

We propose that 20% and 10% of human and mouse genes, respectively, with multiple transcripts are affected by polymorphisms that alter splicing. For the detection of individual genes affected by allele-specific splicing we integrated results from genome-wide microarray, genomic and transcript-based analyses. Such an approach gives more confidence in the predicted candidates. This dissertation is an extensive resource which underscores the prevalence and importance of allele-specific splicing and will support further investigation of polymorphisms that contribute to mRNA isoform diversity observed in human and mouse, some of which are of great medical importance.

Declaration

I declare that "**Genome-wide survey and analysis of allele-specific mRNA splicing in human and mouse**." is my own work, that it has not been submitted for degree or examination at any other university, and that all the resources I have used or quoted, and all work which was the result of joint effort, have been indicated and acknowledged by complete references.

Victoria P. Nembaware

July 2008

"You shall no longer take things at second or third hand, nor look through the eyes of the dead, nor feed on the spectres in books,

...

You shall listen to all sides and filter them from yourself",

Walt Whitman

Acknowledgements

My deeply felt gratitude goes to my advisor Professor Cathal Seoighe. His guidance, encouragement, his ever prompt positive feedback and his most enduring patience have been instrumental in me finishing my thesis. Prof. CS's expertise especially in the simulations (Chapter 4) maximum likelihood models (Chapter 5) and microarrays (Chapter 5), has made it possible for me to complete my thesis. I will always be indebted to Prof. CS, for he went well beyond his call of duty by spending hours on end improving my scientific writing through-out my work. His exceptional enthusiasm for science and particular attention to detail will always be a guide for me.

To my friend, Robert Ndoria Thuku, I am indebted to you, for your most helpful comments on some of my chapters. Enkosi, Bukiwe, for the stimulating discussions and for the collaboration that resulted in Table 3 in Chapter 5. Thank you to Dr. Konrad Scheffeller for your expertise and willingness to collaborate with me in the maximum-likelihood methods used in Chapter 5. A special thanks goes to Rodger for helping me out with my many, "techy" questions.

Thank you to all my colleagues in the CBIO lab who have given me helpful comments throughout my thesis. A special thanks especially to Nobubelo, Venu, Halimah, Natasha, Graham and Sachin for their encouragement and positivity. Thank you, Venu for making sure I was well nourished during my thesis writing period with rotis and mouth-watering curries.

A very special heartfelt thank you, to my Uncle Tony who initiated my academic career by providing me the finances and pushing me out of my comfort zone, your many sacrifices are out of this world. I also feel very privileged to have parents, siblings and friends especially Sandi and Realm who have given me unconditional support and encouragement.

To Justin, I'm speechless, MWAH!

Publications arising from this thesis

Nembaware,V., Lupindo,B., Schouest,K., Spillane,C., Scheffler,K., and Seoighe,C. (2008). Genome-wide survey of allele-specific splicing in humans. BMC. Genomics *9*, 265.

Nembaware,V., Lupindo,B., Scheffler,K., and Seoighe,C. (28-30 January 2007). Identification of allele-specific mRNA transcripts through an integrative analysis of genomic and EST data. Proceedings of the First Southern African Bioinformatics Workshop, Johannesburg, South Africa. *ISBN 978-0-620-38113-0.*

Seoighe,C., **Nembaware,V**., and Scheffler,K. (2006). Maximum likelihood inference of imprinting and allele-specific expression from EST data. Bioinformatics. *22*, 3032-3039.

Nembaware,V., Wolfe,K.H., Bettoni,F., Kelso,J., and Seoighe,C. (2004). Allele-specific transcript isoforms in human. FEBS Lett. *577*, 233-238.

Abbreviations

AS	Alternative Splicing
EST	Expressed Sequence Tag
NMD	Nonsense Mediated Decay
NAS	Nonsense Associated Splicing
ASAPII	The Alternative Splicing Annotation Project II
SNPs	Single Nucleotide Polymorphisms
srSNP	splicing regulatory SNP
mSNP	marker SNP
SI	Splicing Index
BP	Branch points
ESE	Exonic splice enhancer
ESS	Exonic splice silencer
ISE	Intronic splicing enhancers
ISS	Intronic splicing silencers
ANOVA	Analysis of variance
PTC	Premature termination codon
sQTLs	splicing quantitative trait loci
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
qRT-PCR	quantitative Reverse Transcriptase Polymerase Chain Reaction
B6	C57BL/6 and its related strains
nonB6	Mouse strains not related to the C57BL/6 strain

Table of Contents

Abstract……ii
Declaration ……iii
Acknowledgements ……iv
Publications arising from this thesis ……vi
Abbreviations……..xiii

Chapter 1: Thesis Rationale and Objectives
1.1 Introduction 1
Alternative splicing: A single gene produces multiple products 1
1.2 Functional impact of alternative splicing..... 3
Protein-level modifications..... 4
Transcript-level modifications..... 4
1.3 Allele-specific splicing..... 5
1.3.1 Genome-wide detection of allele-specific splicing 5
a) Linkage based analysis..... 7
b) Use of publicly available transcripts 8
1.4 Choice of organisms to study..... 8
1.5. Thesis organization..... 9

Chapter 2: General Literature Review
Abstract..... 12
2.1 pre-mRNA splicing and alternative splicing..... 13
2.1.1 Pre-mRNA splicing 13
2.1.2 Alternative splicing: mechanism and regulation..... 19
2.2 Large-scale detection of alternative splicing..... 24
2.2.1 ESTs..... 24
2.2.2 Microarrays 30
2.3 Databases of alternatively spliced mRNA isoforms 31
2.3.1 Impact of EST limitations on databases of alternative splicing..... 33
False negatives..... 33
False positives..... 35
2.3.2 Mechanisms that spuriously inflate mRNA isoforms in databases of alternatively spliced mRNA isoforms 35
RNA editing..... 35
Random spliceosome errors..... 35
External stress to cells..... 36
Alternative transcription initiation sites.....38
Alternative transcription termination and polyadenylation sites.....38
Allele-specific splicing..... 38
2.4 Allele-specific splicing..... 38
2.4.1 Impact of allele-specific splicing on disease and pharmacogenetics 39
2.4.2 Detection of allele-specific splicing 41
2.5 Concluding remarks 47

Chapter 3: A Database of SNPs Mapped to ESTs
Abstract..... 48

3.1 Introduction 49
3.2 Data and methods 51
3.2.1 Data 51
3.2.2 Quality control 51
3.2.3 Extraction of SNP alleles from ESTs 52
3.2.4 eVOC ontologies 53
3.3 Results 54
3.3.1 A database of SNPs mapped to ESTs 54
3.3.2 Annotation of cDNA libraries using eVOC 55
3.3.3 Web interface 55
3.4 Discussion 56

Chapter 4: Estimation of the Prevalence of Allele-Specific Splicing in Human
Abstract 60
4.1 Introduction 61
4.2 Data and methods 64
4.2.1 Data matrices 64
3.2.2 Simulations 65
4.3 Results 66
4.3.1 Matrices 66
4.3.2 Estimation of allele-specific splicing in human 66
4.3.3 Detecting individual examples of allele-specific splicing 71
4.4 Discussion 71
4.5 Acknowledgements 71

Chapter 5: Identification of Allele-specific mRNA Isoforms in Human Through an Integrative Analysis of Genomic, EST data and Microarrays
Abstract 76
5.1 Introduction 77
5.2 Data and methods 80
5.2.1 Splice site strength prediction 80
5.2.2 Mapping exonic SNP alleles to splice variants 81
5.2.3 Models of regulated and allele-specific splicing 81
5.2.4 Simulations 82
5.2.5 Analysis of Affymetrix exon arrays 83
5.3 Results 84
5.3.1 A genome-wide scan for polymorphisms in splice-regulatory regions 84
5.3.2 A maximum likelihood method to identify allele-specific splicing using EST data 85
5.3.3 Support for srSNPs and mSNPs from publicly-available exon array data 87
5.3.4 Cross-validation of EST and exon array results 92
5.3.5 Splicing index association plots 92
5.4 Discussion 94
5.5 Acknowledgements 102

Chapter 6: An Exploratory Survey of Strain-specific Splicing in Mouse
Abstract 100
6.1 Introduction 101
6.2 Data and methods 104
6.2.1 A database of SNPs mapped to ESTs 104
6.2.2 cDNA library classification 104
6.2.3 Matrices 105
6.2.4 Simulations 105
6.2.5 Detection of individual cases of strain-specific splicing 106
6.3 Results 107
6.3.1 A map of SNPs mapped to ESTs 107
6.3.2 Classification of cDNA libraries as C57BL/6 or non C57BL/6 strain 107
6.3.3 Prevalence of C57BL/6 strain-specific isoforms 110
6.3.4 Gene candidates 110
6.4 Discussion 114

Chapter 7: Conclusion and Future work
7.1 High prevalence of allele-specific splicing 119
7.2 Improvement in the detection of splicing mutations 120
7.3 Development of novel methods for the detection of allele-specific splicing 120
7.4 Summary of resources 121
7.5 Future work 121
7.5.1 Expansion of the snp2estmap database 121
7.5.2 Maximum likelihood Models 122
7.5.3 Further functional characterization of allele-specific isoforms 122
7.6 Concluding remarks 123

Bibliography…… 132

Chapter 1

Thesis Rationale and Objectives

1.1 Introduction

The decoding of information stored in a gene to produce protein, via transcription, is the central dogma of biology (Crick, 1970). Prior to 1977, eukaryotic genes were thought to correspond to continguous stretches of the genome. Based on studies performed on the Adenovirus *HEXON* gene, this understanding of the central dogma of biology was drastically transformed by the Nobel winning discovery of the mRNA splicing process (Berget, Moore et al. 1977; Chow, Gelinas et al. 1977). Viral and eukaryotic genes typically exist in an interrupted manner and undergo mRNA splicing which removes the non-coding interrupting sequences and accurately joins together the remaining sequences to form mature mRNA sequences. These mature mRNA transcripts are then used as templates for the translation of genes into protein.

Alternative splicing: A single gene produces multiple products

Soon after the discovery of mRNA splicing, Walter Gilbert coined the term “intron” to refer to the non-coding segments of genes. Of greater significance is his postulate that a single pre-mRNA transcript could be used as a fixed template to produce multiple distinct mRNA isoforms by varying the demarcations and combinations of introns in a process now commonly known as alternative mRNA splicing (Gilbert, 1978). Experimental evidence supporting Gilbert’s notion soon followed, ending the then-widely accepted cistronic nature of genes or the one-gene one peptide hypothesis. Little did Gilbert know that he had uncovered a versatile mechanism that contributes significantly towards the regulation of gene expression in numerous multi-cellular and viral genes.

Although alternative splicing was discovered in 1978, appreciation of the importance of this biochemical process was only achieved during sequencing of the whole human genome. Computational analysis led to the discovery that alternative splicing is highly prevalent in the human genome (Johnson et al., 2003). From a minor genetic process initially estimated to occur in less than 5% of human genes (Sharp, 1994), alternative splicing is now considered as a highly prevalent mechanism that occurs in approximately 75% of human genes (Johnson et al., 2003) and in many other higher order organisms (Brett et al., 2002). Structural analyses of the computationally predicted alternatively spliced isoforms led to the discovery of the many possible modes through which AS can occur (see Figure 1).

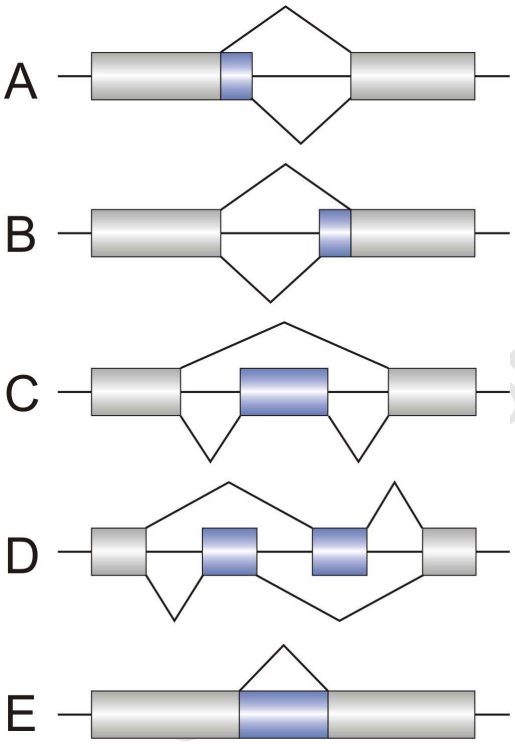


Figure 1: Although more than 20 different splicing modes exist, only the major classes are illustrated in this diagram. This figure was adapted from Graveley, 2001. A) Alternate donor site B) Alternate acceptor site C) Exon skipping D) Mutually exclusive exons E) Intron Retention

Multiple mRNA isoforms can be produced from a single gene, through alternative splicing or in an allele-specific manner. Allele-specific splicing is the expression of distinct mRNA isoforms or amounts from allelic versions of a gene transcript, which contrasts with alternative splicing where distinct mRNA isoforms are processed from identical transcripts. Discerning whether mRNA isoforms result from alternative

splicing or allele-specific splicing based on publicly available transcripts alone, presents a major challenge. Therefore, computational studies have largely disregarded allele-specific splicing and categorized all transcript isoforms as originating from alternative splicing (see Chapter 3).

Allele-specific splicing underlies important phenotypic variations in human including disease susceptibility and drug response (Wang and Cooper, 2007). It is critical to understand the relevance of alternative splicing in order to fully appreciate the pathological effects that allele-specific splicing can have. The importance and specific functional roles of alternative splicing provide a platform for making inferences with regard to the potential impact of allele-specific splicing.

1.2 Functional impact of alternative splicing

The ability of cells to switch functions in a highly coordinated manner is merely a manifestation of alterations occurring at the gene expression level in a timely fashion. A complete collection and description of mRNA transcript expression levels and proteins in a specific cell at a particular time-point can be described as the transcriptome and proteome respectively. Numerous complex regulatory networks operate at the transcription and translation levels of gene expression pathways to control the transcriptome and proteome. Among gene expression regulatory mechanisms, alternative splicing has emerged as a vital component in the regulation of the transcriptome in higher eukaryotes (Graveley, 2001; Lopez, 1998).

Detailed molecular studies of many biological processes critical for an organism's viability, such as neural development have shown that alternative splicing is an important component of their regulatory systems (Graveley, 2001). However, the complete impact of alternative splicing on all biochemical processes is unknown. For example, although alternative splicing of the human fibronectin gene was discovered more than 20 years ago (Kornblihtt et al., 1984) the precise biological roles of the 20 main mRNA isoforms are not as yet fully known. Endeavors towards a detailed and thorough characterization of the global impact of alternative splicing are likely to last

several decades. Modifications to the transcript caused by alternative splicing can be classified broadly into two categories: protein-level or transcript-level modifications.

Protein-level modifications

Alternative splicing contributes significantly to proteome diversity. 74% of alternatively spliced transcripts are thought to be subjected to changes that affect the coding sections of genes (Modrek et al., 2001). The inclusion or exclusion of alternative coding segments gives rise to protein isoforms which can perform distinct biological functions (Black, 2000; Graveley, 2001). Changes in the peptide sequences can modify protein binding affinities, catalytic activities, protein solubility and localization (Black, 2000) and in some cases even produce proteins with opposing functional roles (Graveley, 2001).

Frequently, alternative splicing events tend to preserve the overall integrity of the peptide sequence (Homma et al., 2004; Kriventseva et al., 2003). Proteome diversity through alternative splicing can thus be achieved through a gradient of changes that occur on a single peptide template, from large changes that include removal of peptide segments of more than 100 amino acids in length to subtle amino acid changes. Subtle changes are likely to be widespread, for example, single amino acid alterations are found in 5% of the human genes with NAGNAG acceptors (Hiller et al., 2004a).

One might expect such subtle protein modifications to have negligible impact on protein function, but this has been shown to be untrue (Wen et al., 2004). The functional implications of the modifications in the coding region are not necessarily controlled by the size of the alternatively spliced region but depend on the functional and structural relevance of the affected region (Wen et al., 2004).

Transcript-level modifications

The untranslated regions of mRNAs are rich sites for gene regulatory elements (Mossner and Riederer, 2007; Cowles et al., 2002; Rockman and Wray, 2002). Approximately 26% of alternative splicing has been estimated to affect untranslated

mRNA regions and can thus affect gene regulatory mechanisms (Modrek et al., 2001). Alternative splicing of such sites can control the inclusion or exclusion of important *cis*- regulatory elements thus leading to significant alterations in translation efficiency, mRNA stability, mRNA localization and even transcriptional efficiency of the gene, thus affecting mRNA expression levels.

1.3 Allele-specific splicing

DNA polymorphisms that exist in splicing regulatory elements can disrupt the mRNA splicing pathways and alter splicing patterns in an allele-specific manner. Given the enormous impact of alternative mRNA splicing on the regulation of gene expression in important biological processes through modifications at the protein and transcript levels (Graveley, 2001; Black, 2000a), allele-specific splicing has huge potential to disrupt the tightly controlled transcriptomes and proteomes of any cell through similar sequence alterations. Such inter-individual changes in the transcriptomes and proteomes can result in both subtle and even life-threatening phenotypes (Wang and Cooper, 2007).

Numerous examples of allele-specific splicing that contribute to disease progression and susceptibility have already been reported in literature (Wang and Cooper, 2007). A thorough understanding of the impact of allele-specific splicing on human phenotypes opens up immense possibilities of mapping out medically relevant phenotypic variations and disease aetiology (Wang and Cooper, 2007). A genome-wide catalogue and analysis of allele-specific splicing would afford a more holistic view and a starting point in unpacking the biological impact of such individual-specific mRNA splicing events.

1.3.1 Genome-wide detection of allele-specific splicing

Despite the growing appreciation of genome-wide analysis of intra-individual variation in alternative splicing, only modest efforts have been made towards the study of allele-specific splicing at this level. These, contrasts significantly with the level of research into genome-wide detection of global gene expression variations in

an allele-specific manner (see Figure 2 for comparison of allele-specific splicing and allele-specific gene expression). In part, this could be due to the fact that transcription has been known and studied for a longer period, evident from the establishment of a comprehensive database of transcriptional regulatory systems, TRANSFAC, which catalogues large volumes of well characterised transcriptional regulatory elements and factors (Matys et al., 2006). Databases of similar magnitude for the mRNA splicing process are still under development.

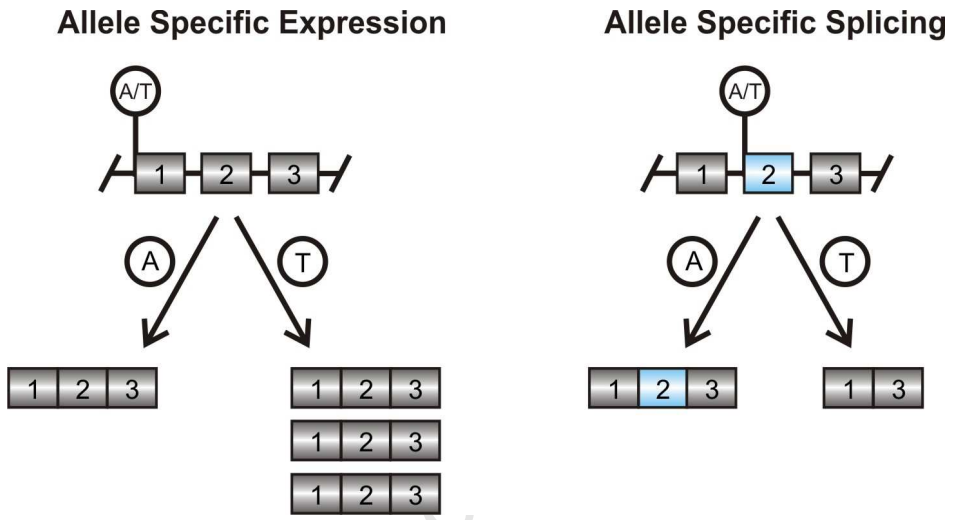


Figure 2: Differences between allele-specific gene expression and allele-specific splicing, illustrated using a gene transcript with three exons. Allele-specific expression involves quantitative changes that affect the transcription of whole gene transcript, while allele-specific splicing is quantitative or qualitative changes that affect only isoforms of mRNA transcripts. We have illustrated an allele-specific splicing event that affects the inclusion or exclusion of an exon in mRNA isoforms.

Altering pre-existing methods and strategies used on similar systems provides a much needed basic frame-work for genome-wide analyses of allele-specific splicing. The mRNA splicing and transcriptional gene regulation, are parallel (Table 1), therefore hurdles faced in the detection of allele-specific expression are similar to those of allele-specific mRNA splicing. Thus many lessons on how to detect genome-wide allele-specific splicing can be learnt based on previous work on allele-specific expression.

Table 1: Comparison of transcriptional control of gene expression and mRNA splicing

Attribute	Transcription	mRNA splicing
Transcript	Global	Only segments of mRNA transcripts which are alternatively spliced
Regulatory system	Core regulatory elements Enhancers, silencers	Core regulatory elements Enhancers and silencers
Location of *cis*-acting regulatory elements	Variable positions and many are yet to be discovered	Less variable locations and many are yet to be discovered
Publicly available data	Serial Analysis of Gene Expression (SAGE) Expressed sequence tags (ESTs) Microarrays	Expressed sequence tags (ESTs) Microarrays

For this current study aimed at genome-wide detection and analyses of allele-specific splicing, two major lessons were learnt from previous studies on allele-specific gene expression a) basing detection on the principle of linkage disequilibrium, and b) use of publicly available transcripts. Arguments for using these two approaches in the detection of allele-specific expression are pointed out below.

a) Linkage based analysis

Revealing the extent of *cis*-acting regulatory mutations on allele-specific expression presents a great challenge. Most of the sequence variants that disrupt the regulatory networks are hard to predict due to the complex nature of the regulatory networks. Genome-wide detection of allele-specific expression could thus be incomplete if directed only at mutations which disrupt known transcriptional regulatory elements.

To circumvent this difficulty, one approach that has gained in popularity for the detection of allele-specific expression is based on identifying associations between gene expression variations and genotypes that are linked to unknown causal variants. This procedure greatly increases the power to detect a comprehensive set of allele-specific expression differences and offers the advantage of simultaneously analyzing allele-specific variations in expression on a genome-wide scale (Knight, 2004).

b) Use of publicly available transcripts

Genome-wide analysis of allele-specific expression requires an extensively sampled transcriptome and genotyped sequence polymorphisms. Traditionally, genome-wide analyses have been performed on experimental platforms which fulfill both these requirements (Cowles et al., 2002; Pastinen et al., 2000). However, these are expensive and time-consuming and other cheaper avenues had to be explored to facilitate research in this field.

The availability of publicly available transcriptomes that harbor genotype information has sparked a paradigm shift in efforts to accelerate the detection and deciphering of allele-specific gene expression; from a purely experimental approach to a combination of experimental and computational approaches based on publicly available datasets. Unintentionally, due to their nature, databases from transcriptome studies have also captured allele-specific expression (Ge et al., 2005). Such cheap and rich reservoirs have inevitably attracted the interests of many research groups studying allele-specific expressing, and produced promising results (Knight, 2004).

1.4 Choice of organisms to study

Allele-specific mRNA splicing can occur in any organism that has multi-exon genes and may contribute in the determination of a significant amount of important phenotypic variation across individuals. In viruses, allele-specific splicing has been implicated in affecting their virulence and pathogenicity (Purcell and Martin, 1993). In human allele-specific splicing has been implicated in disease and cancer progression (Wang and Cooper, 2007).

The huge potential impact of allele-specific splicing on phenotypic variation necessitates its detection in all multi-exon genomes. However, in this thesis we have based our study only on the human and mouse genomes. The importance of studying mouse is underscored by the observation that it is the preferred and most widely used organism for modeling human diseases (Peters et al., 2007). Mouse individuals from a single inbred strain are isogenic (i.e. genetically identical at all loci). Therefore, allele-

specific splicing in mouse is equivalent to strain-specific mRNA splicing, thus this term is used throughout this thesis.

This current work is based purely on computational approaches and consequently the choice of organism to study is immensely influenced by the availability of data. dbEST (Wheeler et al., 2007), the largest public repository of transcriptome data, contains transcripts from more than 1500 species. Human and mouse dominate with their transcriptomes making up collectively 25.65% of the data in the dbEST database (Figure 3). Therefore, a further reason for studying human and mouse is the extensive sampling of the mouse and human transcriptomes in the dbEST database.

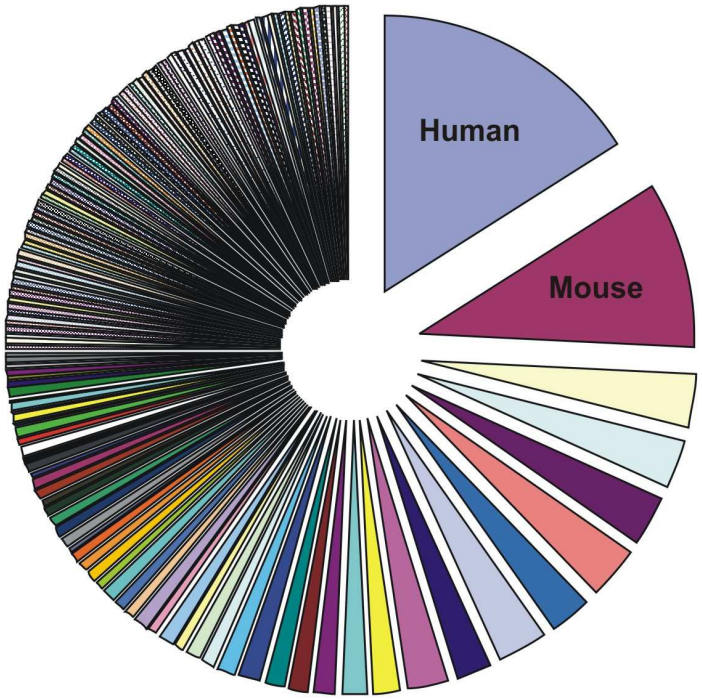


Figure 3: Biased transcript representation towards the human and mouse genomes in the dbEST database version 031408. A total of 50622371 transcripts from 1516 organisms are represented in dbEST and in the pie-chart above. For clarity, only data from human and mouse are labelled.

1.5. Thesis organization

A major goal in the field of genetics is to define, globally, the relationship between genotype, gene expression and phenotypic variations. Towards this end, efforts have been focused largely on genome-wide allele-specific protein changes and allele-

specific global gene expression patterns (Knight, 2004; Knight, 2006). This thesis instead aims to emphasize allele-specific splicing, which has been largely disregarded. Publicly available transcript datasets are exploited to perform a comprehensive genome-wide survey and characterization of allele-specific splicing in human and mouse.

A general literature review is presented in **Chapter 2.** This chapter highlights the important role and usefulness of ESTs in the detection of splicing. The high prevalence of alternative splicing has motivated a large body of work aimed at understanding the regulation of alternative splicing. This knowledge of splicing regulatory elements has been instrumental in the detection and understanding of the impact of mutations that disrupt splicing. In conclusion, tools and methods that can be used for the detection and characterization of splicing mutations are also reviewed.

Chapters 3-6 are based on specific research questions that involved extensive data analyses. Each chapter is thus presented as a separate entity.

The recent recognition that studies on allele-specific expression and allele-specific splicing could be facilitated by use of publicly available data prompted the development of a pre-computed database of polymorphisms that map to expressed sequence tags (ESTs). This database is presented in **Chapter 3**.

Before an in-depth analysis of allele-specific splicing could be performed, we quantified the extent to which the observed transcript variants in a publicly available database of presumed alternatively spliced mRNA isoforms is influenced by allele-specific mRNA splicing. For this analysis, a novel heuristic method was developed based on the principle of linkage disequilibrium between unknown *cis*-acting mutations and observable and genotyped mutations. The approach applied to human data in **Chapter 4** was also used for the estimation of the prevalence of mouse strain-specific splicing in **Chapter 6**.

Chapter 5 presents an integrated analysis by three publicly available datasets (microarray, expressed sequence tags and genomic sequences), for the detection of allele-specific splicing. Publicly available tools were used for analysis of microarray

and genomic data. A maximum likelihood approach was implemented that makes better use of the EST data for detecting allele-specific splicing than the heuristic method presented in **Chapter 4.** Putative *cis*-acting mutations that alter splicing by disrupting well established *cis*-acting regulatory elements were detected by applying *ab inito* prediction tools to genomic sequence data.

Chapter 7 is a concluding chapter and it highlights the main findings and gives a future perspective of this work.

Chapter 2

General Literature Review

Abstract

Considerable effort has been made towards understanding networks that contribute to the regulation of mRNA splicing and alternative splicing of pre-mRNA transcripts. Much of the progress made towards understanding mRNA splicing is attributed to Expressed Sequence Tag (EST) and microarray technologies which permit large-scale detection and subsequent cataloguing of alternatively spliced pre-mRNA isoforms. This review provides a general overview of alternative splicing, its detection and cataloguing into databases. Although these databases of alternatively spliced pre-mRNA transcripts are useful resources, the impact of noise from other mechanisms such as allele-specific splicing is largely unknown and these are discussed in this review. Pointers to the multiple splicing regulatory elements and other gene expression mechanisms which if disrupted by mutations can result in allele-specific splicing are also highlighted. The huge impact of splicing mutations on human diseases and pharmacogenetics has made the characterisation and the detection of allele-specific splicing prominent within the spheres of human genetics.

2.1 pre-mRNA splicing and alternative splicing

2.1.1 Pre-mRNA splicing

A significant number of eukaryotic and virus genes are interrupted by introns (see Chapter 1). Therefore, accurate removal of the introns through pre-mRNA splicing, plays a major role in the expression of genes in higher order organisms. The importance of the pre-mRNA splicing is underscored by the high prevalence of disease-causing mutations that affect the splicing process (Krawczak et al., 1992; Lopez-Bigas et al., 2005).

The mRNA splicing reaction occurs in well defined, highly catalytic *trans*-esterification reactions (Figure 1). This process occurs within a dynamic structure known as the spliceosome which is made up of five U-rich small ribonucleoproteins (snRNPs), namely U1, U2, U4, U5 and U6, and more than 50 proteins (Lopez, 1998). The spliceosome is largely controlled by complex interactions between spliceosome factors and core splicing *cis*-acting elements encoded within the mRNA transcript, which include the donor, acceptor, polypyrimidine tract and the branch point as shown in Figure 1. The splicing process results in the release of introns via formation of lariat structures and is concluded by accurate annealing of the 5' exon and 3' exon of the spliced out introns (Figure 1a).

2.1.1.1 Splice site recognition

Before catalyzing the splicing process, the spliceosome has to identify the introns or splice sites. The spliceosome defines splice sites based on the terminal di-nucleotide bases of the introns, which are usually GT and AG for donor and acceptor sites, respectively (Lopez, 1998). Although these are the only four bases in the splice sites that are highly conserved, additional nucleotides that flank these di-nucleotides make up the consensus sequence are required for the U1 snRNP to bind to the donor site and the SF1/U2snRNP to bind to the acceptor site (Figure 1). About 9 bases define the donor site and approximately 23 bases define the acceptor sites (Yeo and Burge, 2004). Precision in the recognition of the splice-sites depends on the binding affinity

of the U1 factor to the donor site, and that of the U2AF *trans*-acting factor to the acceptor site (Lopez, 1998; Smith and Valcarcel, 2000).

A major obstacle faced by the spliceosome in accurately demarcating splice sites is the abundance of cryptic splice sites in the genomes that can attract the spliceosome with equal or even better efficiency than authentic splice sites (Baralle and Baralle, 2005). For example, in the human *HPRT* gene, the spliceosome has to identify 9 authentic exons from over 100 potential donor sites and over 600 potential acceptor sites (Sun and Chasin, 2000). The information within the core splicing signals (donor, acceptor and polypyrimidine tract) is insufficient for the spliceosome to precisely distinguish weak authentic splice sites amidst such large numbers of potential splice sites and thus, additional auxiliary *cis*-elements are required (Sun and Chasin, 2000).

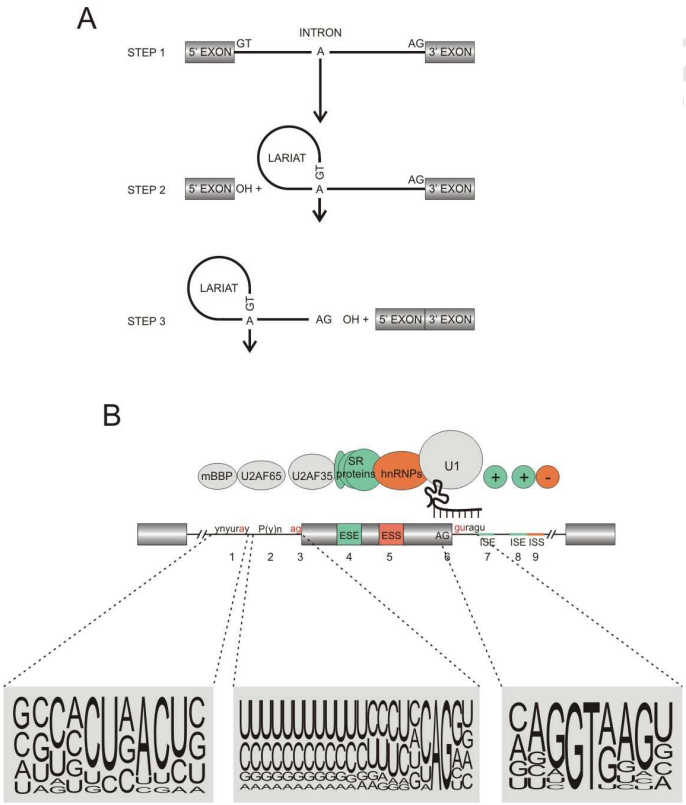


Figure 1: *Cis*-regulatory elements involved in the detection of splice sites and the splicing out process 1: Branch point, 2 and 3: Acceptor sites these include the polypyrimidine tract, 4: Exon Splicing Enhancers (ESE), 5: Exon Splicing Silencers (ESS), 6: Donor site, 7 and 8: Intronic Splicing Enhancers (ISE) and 9: Intron Splicing Silencers (ISS). Figure was adapted from Cartegni et al., 2002.

Two types of auxiliary elements exist, enhancers and silencers which are found both in exons and introns and are named accordingly; Exon Splicing Enhancers (ESEs), Exon Splicing Silencer Elements (ESSs), Intron Splicing Enhancers (ISEs) and Intron Splicing Silencers (ISSs) (Lopez, 1998). As the names suggest, enhancer elements promote the recognition of authentic splice sites by the spliceosome while silencers encourage the spliceosome to by-pass cryptic splice sites.

Enhancer elements

Enhancer elements facilitate recruitment of spliceosome factors to splice sites. *Trans*-acting factors that can bind to the enhancer elements through their RNA-binding domains recruit components of the spliceosome to splice-sites through protein-protein interactions (Lopez, 1998). Strong ESEs are commonly found near exons that are bordered by introns with weak splice signals (Dewey et al., 2006). Therefore, a compensatory relationship exists between ESEs and splice sites.

Extensive computational studies have revealed the existence of many different classes of enhancer elements in constitutively spliced exons (Zhang et al., 2005; Cartegni et al., 2002). For example, G-rich elements have previously been identified that support the recognition of small constitutively spliced exons (McCullough and Berget, 1997). Different splicing enhancer elements are characterised by different nucleotide compositions with different substrate specificities (Ladd and Cooper, 2002). These different characteristics are likely to be the foundation of inconsistencies in the distance between ESEs and splice sites. Although there is a general enrichment of ESEs within 125 base pairs of splice junctions (Fairbrother et al., 2004; Majewski and Ott, 2002), some strong ESEs can still exert that function more than 1000 bases away from the splice junctions (Hull et al., 2007).

Of all the enhancer elements, the best studied are exonic enhancer elements from the serine-arginine (SR) family of splicing factors. The most significant work on SR-specific ESEs is from the Krainer Laboratory (Liu et al., 2000; Liu et al., 1998). Instead of using computational methods to detect ESEs, they used an experimental approach that functionally selects for ESE motifs (6-8 nucleotides in length) enriched in exons. This approach involved modifying the functional SELEX (Systematic

Evolution of Ligands by Exponential enrichment), which is commonly used to identify functional ligands. To facilitate the use of ESEs identified using the SELEX method, Cartegni and co-workers (Cartegni and Krainer, 2003), used their ESE motifs to develop scoring matrices and implemented these in the ESEfinder web-based program. Users can then predict putative SR-specific ESEs by scanning and scoring their sequences using ESEfinder.

Silencer elements

Silencer sequences when bound by *trans*-acting factors hinder the binding of the spliceosome to the mRNA transcript. Of all the silencer elements, the best known are those bound by the hnRNP factors (Cartegni et al., 2002). Interestingly, the polypyrimidine tract can also act as a silencer element when bound to the hnRNP splicing factor in addition to other splicing factors that hinder the interaction of the U2AF spliceosome factor to the acceptor site (Cartegni et al., 2002). An example is that of the SXL (Sex Lethal) protein in Drosophila which binds to the polypyrimidine tract of a specific acceptor site and reroutes the U2AF factor to another site (Lopez, 1998).

2.1.1.2 Splice site recognition mechanisms

One pivotal question in the mRNA splicing process is the exact process by which the splicing machinery is assisted by auxiliary regulatory elements to recognize splice sites. *In vitro* splicing assays have shown that the exon/intron architecture of genes determines how the spliceosome recognizes splices sites (Berget, 1995; Sterner et al., 1996; Robberson et al., 1990; Talerico and Berget, 1994). Depending on intron and exon size, splice site recognition by the spliceosome can occur via the intron or exon. The length is important because for the splicing process to occur, interactions between the different splicing factors on the 5’ and 3’ ends of introns are required (see Figure 2).

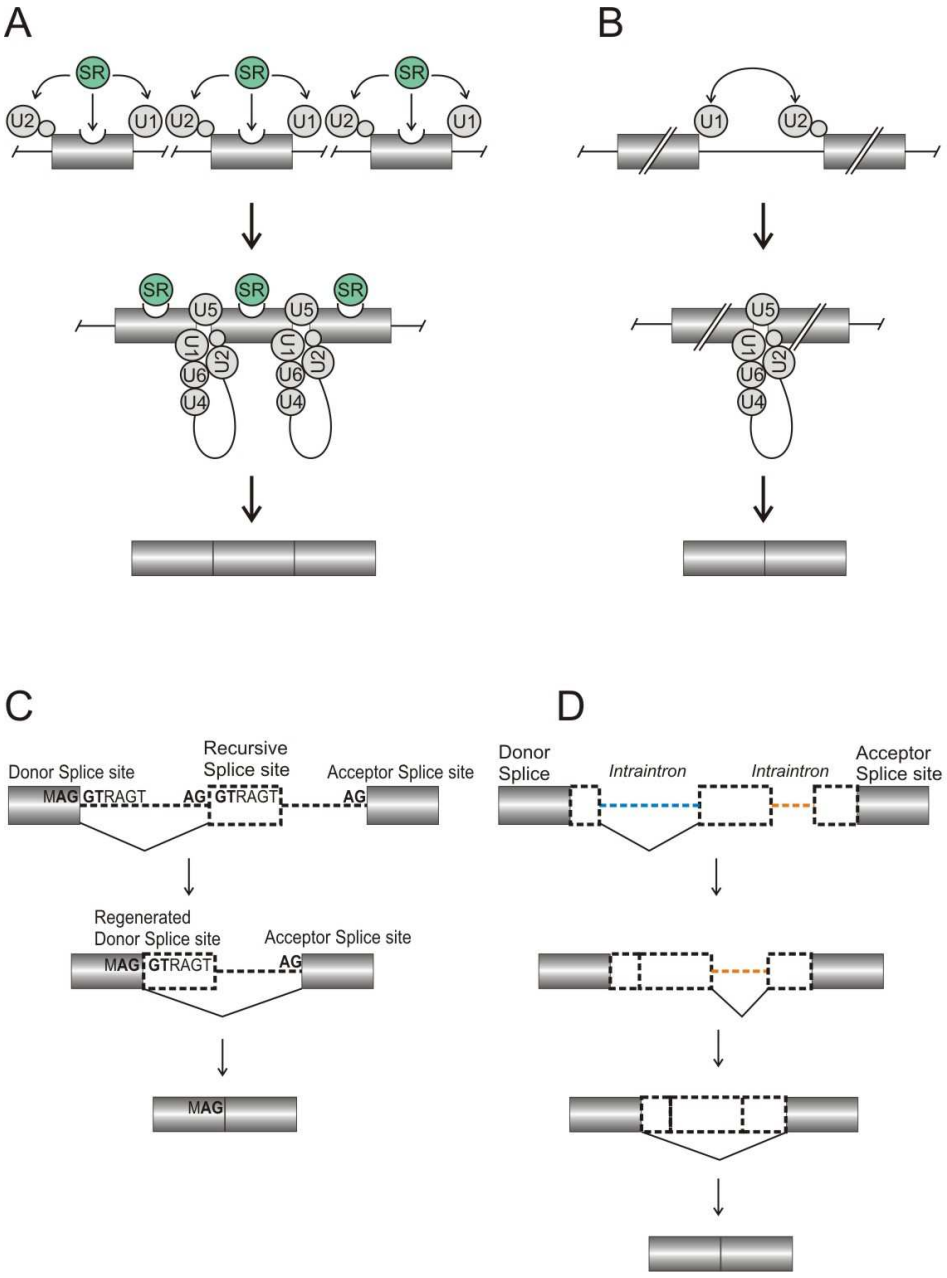


Figure 2: Splice site recognition mechanisms. A) Exon recognition is facilitated across exons. The enhancer molecules bind to their *cis*-acting elements and recruit the spliceosome factors B) Intron recognition is facilitated via the introns when introns are short enough to allow interactions between the U1 and U2 splicing factors which promote the progression of the splicing process C) Recursive splicing; Long introns can have splice sites embedded within them which can be used by the spliceosome. These sites are then used to splice out sections of the introns in a recursive manner D) Intra-splicing; In this model, "intraintrons" (colored blue and orange) and "intraexons" (dotted boxes) can exist within long introns " (Ott et al., 2003). The spliceosome recognizes and splices out the intraintrons and hence shortening the overall intron length. When the introns length is short enough and composed only of intraexons, it is then spliced out via one of the less complex mechanisms such as the exon and introns recognition. The figure is adapted from (Berget, 1995), (Burnette et al., 2005) and (Ott et al., 2003).

Exon Recognition

The average length of introns is much longer than that of exons in the human genome (Venter et al., 2001). Therefore, the recognition of splice sites in human occurs via the exon recognition mechanism since most human exons are short enough to allow for interactions between the 3' and 5' splice junctions via SR proteins (Berget, 1995; Robberson et al., 1990) (Figure 2). Consistent with this mechanism lengthening of small exons that are flanked by long introns results in exon skipping (Sterner et al., 1996). The optimal exon length for the exon recognition splice site mechanism has been reported to be 50-300 bps (Berget, 1995). This optimal length was established using exon lengthening and shortening mutations (Berget 1995).

Intron Recognition

In other lower organisms such as Drosophila and in plants, the introns are on average much shorter than the exons, and this observation supports splice site recognition via intron recognition (Talerico and Berget, 1994) (Figure 2). An upper limit of intron length of 200-250 bps has been suggested for the intron-recognition model (Fox-Walsh et al., 2005). The most prevalent form of regulated splicing in plants and lower order animals is intron retention, which is also consistent with the intron recognition mechanism.

Recursive Splicing

For short exons embedded within long introns, the most likely splice site recognition mechanism would be via the exon definition mechanism (Berget, 1995). However, the splicing out of very long introns would take a long time. For genes such as the dystrophin gene which has very long introns (over 10kb), the mRNA splicing process via exon recognition takes 36 hrs in comparison to genes with small introns which take minutes (Tennyson et al., 1996). Such slow excision of the introns would also encourage the formation of secondary structures through complementary base pairing. A recently proposed and intriguing mechanism is that of recursive splicing (Hatton et al., 1998; Burnette et al., 2005). In this model, very long introns (longer than 10kb) are recognized and spliced out in sub-segments through a series of recursive splicing

steps (Burnette et al., 2005; Hatton et al., 1998). Burnette et al., (2005), identified 124 introns in Drosophila that are spliced out via the recursive splicing mechanism and performed experimental validations for some of these introns through *in vitro* experiments. Because Drosophila has a much lower percentage of introns above 10kbs than humans, recursive splicing could be a common process used to splice out approximately 10% of long introns found in the human genome (Deutsch et al., 2008).

Intra-splicing

An alternative model has been proposed to explain how very long introns are spliced out, namely the "intrasplicing" model (Ott et al., 2003). This model proposes that the splicing machinery first processes the long introns while ignoring the donor sites, and then uses the donor site right at the end (Figure 2). However, this model is based on computational predictions and is yet to be confirmed experimentally (Ott et al., 2003).

2.1.2 Alternative splicing: Mechanism and Regulation

The pre-mRNA splicing process is a much more complicated process as a single pre-mRNA can be alternatively spliced to give multiple distinct mature transcripts (Graveley, 2001). Such AS is often highly regulated and is thus critical for proper functioning of major human physiological and cellular processes as it greatly expands the protein diversity and contributes to the overall complexity of gene expression (Smith and Valcarcel, 2000). Based on large-scale AS detection technologies such as ESTs and microarrays, over 70% of human genes are estimated to be alternatively spliced (Johnson et al., 2003). The detection of such a high frequency of AS in human as well as other eukaryotic organisms sparked a renewed interest in the field of mRNA splicing and alternative splicing (Brett et al., 2002).

Significant efforts are currently underway to unravel the complex systems involved in mRNA splicing and AS (Cartegni et al., 2003). The shift from simple mRNA splicing to alternative splicing is largely enabled by complex regulatory systems that control the ability of the spliceosome to switch between different alternative splicing pathways, and resulting in the generation of multiple mature RNA isoforms from a

single mRNA transcript (Graveley, 2001). The discovery of a 115 exon gene in Drosophila melanogaster which encodes the axonal guidance cell receptor molecule and has the potential to generate over 38 000 mRNA transcripts through alternative splicing of its pre-mRNA transcript, provides one example, albeit extreme of the potential contribution of alternative splicing to transcript diversity (Schmucker et al., 2000).

Alternative inclusion/exclusion of exons/introns is influenced by splice-site strength, pre-mRNA secondary structure (Coleman and Roesser, 1998), splice site recognition mechanisms, and complex interactions between *cis* and *trans*-acting factors. Alternatively spliced exons are enriched with enhancer and silencer molecules to aid in their regulation (Smith and Valcarcel, 2000). Differential interactions of *trans*-acting factors with enhancer and silencer elements control the spliceosome's ability to switch between different alternative splicing pathways. In comparison to constitutively spliced exons, alternatively spliced exons (Baek and Green, 2005; Itoh et al., 2004) are weak and are flanked by highly conserved intron sequences, suggestive of an enrichment of auxiliary regulatory elements (Sorek and Ast, 2003).

2.1.2.1 *Cis*-regulatory elements

Considerable investigation has gone into deciphering the regulatory elements that govern alternative splicing; however, the list is far from being exhaustive (Ladd and Cooper, 2002). Our poor knowledge of the splicing process is highlighted by the continual discovery of splice regulatory elements (Yeo et al., 2007). Identifying *cis*-elements that regulate AS in so many context-specific splicing events is a major task. A difficulty in the detection of enhancer and silencer elements that form part of the alternative splicing regulatory network is that alternative splicing can be regulated by the core *cis*-elements and also by other *cis*-acting elements which are not part of the basal splicing regulatory system.

Unknown *cis*-acting elements have been highlighted by different methods which include natural and directed mutagenesis at random locations in the mRNA transcripts and assessing whether they cause aberrant splicing events (Pagani et al., 2005; Raponi et al., 2007). Using this approach, point mutations were created at various positions

on the cystic fibrosis gene and more than 50% of the point mutations were observed to cause splicing defects (Pagani et al., 2005).

Computational and comparative genomics approaches have also led to the detection of several alternative splicing-specific enhancer and silencer elements. However, the variable location of enhancer and silencer sequences relative to the splice junction complicates the computational detection of enhancers and silencers (see section 2.1.1). Furthermore, *cis* elements are usually characterised by unique consensus sequences. For example, brain-specific alternative exons are flanked by a splicing regulatory element UGCAUG (Minovitsky et al., 2005). Although this element was detected computationally, its conservation in brain-specific exons in mouse suggests that it is an important functional regulator of brain-specific alternative splicing events.

Variations in the manipulation of activities or amounts of *trans*-acting factors during development or in different tissues can lead to the enhancer and silencer elements involved in splice site recognition contributing to the regulation of AS. For example, high concentrations of SR proteins in comparison to hnRNP proteins favors the use of proximal sites and exon inclusion (Ladd and Cooper, 2002), while a higher abundance of hnRNP factors promotes the use of distal splice sites and exon skipping (Ladd and Cooper, 2002).

Splicing factors such as SR proteins have the ability to auto-regulate their own transcript levels (Lareau et al., 2007). Auto regulation of *trans*-acting factors has been shown to occur through coupling of the unproductive splicing with degradation (Lareau et al., 2007). However, a commonly reported means of regulating the SR protein concentration is at the activation stage. SR protein factors require activation through phosphorylation in order to facilitate the splicing *trans*-esterification reactions (Faustino and Cooper, 2003). Numerous extra-cellular stimuli such as hormones, immune response, neuronal activities or external factors such as food can control the relative abundance of enhancer and silencer splicing factors in a cell-specific or developmental specific manner (Faustino and Cooper, 2003).

Such extra-cellular stimuli that trigger a cascade of reactions which lead to a kinase phosphorylating specific enhancer and silencer elements can indirectly control

inclusion or exclusion of an exon or intron (Faustino and Cooper, 2003). The activation of an SRp40 enhancer protein by insulin for the exon inclusion of a protein kinase C (PKC) isoform is an ideal case for illustrating a hormone regulated alternative splicing event (Chalfant et al., 1995). Phosphorylation of the SRp40 enhancer protein that promotes inclusion of an exon of the PKC mRNA is catalyzed by an unknown kinase enzyme. This phosphorylation reaction is activated by a cascade of reactions which are triggered when an insulin receptor is activated by insulin.

Manipulation of *trans*-acting factors may give rise to novel therapeutic approaches for correcting aberrant splicing (Faustino and Cooper, 2003). A noteworthy achievement of the post-genomic era would be a database that documents all *trans*-acting factors, the motifs they bind to and the genes they affect. Such an undertaking is however, not a trivial task as some of the splicing factors themselves are subject to alternative splicing and their splicing is governed by other *trans*-acting factors (Nakahata and Kawamoto, 2005; Wu et al., 2002). An added complexity in the detection and cataloguing of *trans*-acting factors is that some *trans*-acting factors can act both as enhancers and silencers.

2.1.2.2 Coupling of alternative splicing regulation to other regulatory systems

AS occurs concurrently with other gene expression regulatory mechanisms that control transcription (Chern et al., 2008), nonsense mediate decay (NMD) (Green et al., 2003; Lewis et al., 2003), and imprinting. Although these mechanisms are clearly distinct from AS, with some exceptions (Bhasi et al., 2007), it is surprising that researchers still do not make a clear distinction between alternative polyadenylation and transcription start sites from alternative mRNA splicing (see section 2.3.1). Adding to this confusion, alternative splicing regulation of some mRNA isoforms seems interconnected to these gene expression regulatory systems (Kornblihtt, 2007; Zavolan et al., 2003).

Using full-length cDNAs, Zavolan and coworkers (2003), showed that alternative use of transcription start sites is linked to AS patterns (Zavolan et al., 2002; Chern et al.,

2008). The position at which RNA pol II initiates transcription or positions at which the RNA pol II stalls could lead to different splice patterns (Kornblihtt, 2007; Kornblihtt, 2005). Therefore, variations in activity or binding sites of the RNA pol II during transcription can induce mRNA splicing variations.

Premature termination codons (PTCs) which elicit the NMD system can be introduced by AS events (Lewis et al., 2003; Green et al., 2003). This coupling of the NMD and splicing results in the down-regulation of gene expression and such coupling has been reported for many genes including drug transporters such as the *ABCC4* gene (Lamba et al., 2003). Interestingly, the two exons that are inserted through AS, introducing a PTC in the *ABCC4* gene are highly conserved across human, mouse and monkey. This suggests that the regulation of this gene via NMD is evolutionarily conserved and therefore, functionally important. The coupling of the NMD and splicing in regulating gene expression is now commonly referred to as RUST (regulated unproductive splicing and translation) as coined by Lewis et al., (2003).

Based on EST and cDNA analysis, Lewis and coworkers (2003) reported that at least 35% of alternatively spliced transcripts contain PTCs. This discovery led to assumption that RUST may be a widespread mechanism in the human genome for regulating tissue-specific gene expression. However, this speculation was recently interrogated. Using a quantitative microarray platform, Pan and coworkers (2006) compared levels of splice variants with PTC versus non-PTC containing splice variant in 10 different adult mouse tissues. Their results showed that the steady state levels of a majority of PTC-containing splice variants rarely varied according to tissue-type, which implies that RUST might not be as widespread as previously estimated by Lewis et al., (2003).

Imprinting of splice variants in an isoform-specific manner can occur, thus regulating expression of specific mRNA isoforms. For example, out of the five known *GRB10* isoforms, one is expressed in skeletal muscle from the maternal allele only, while all five *GRB10* mRNA splice variants are expressed from both parental alleles in numerous other fetal tissues (Blagitko et al., 2000). The reason why this coupling occurs is not yet clear and thus further complicating our understanding of the regulation of AS.

2.2 Large-scale detection of alternative splicing

2.2.1 ESTs

As a result of whole genome sequencing efforts, the human genome sequence and genome-wide Expressed Sequence Tags (ESTs) were made publicly available (Venter et al., 2001). The genome sequence captures the 3 billion DNA bases that make up the human genome while ESTs target 10% of the expressed section of the genome that contain gene units. Their primary use was intended for gene discovery and annotation, however, the realisation that ESTs are instrumental for the detection of alternative splicing patterns (Wolfsberg and Landsman, 1997), captured the attention of the genomics research community.

ESTs are single pass unedited sequences that are targeted at capturing the transcriptomes from specific tissues at a particular stage (Rezvani et al., 2000; Bonaldo et al., 1996; Rezvani and Liew, 2000; Wolfsberg and Landsman, 1997). Due to their nature and the manner in which they are generated, ESTs form the basis of numerous AS detection protocols. ESTs are also continually mined to further characterise alternative splicing events and have thus become invaluable in understanding the AS process (Lee and Wang, 2005).

2.2.1.1 Generation and storage of ESTs

The initial stage in the production of ESTs involves harvesting of mRNA from the tissue of interest. mRNA transcripts are generated in the nucleus within which they are free from degradation by RNAase cytoplasm-bound enzymes. However, outside of the nucleus, mRNAs are highly unstable molecules characterised by short half-lives. It is therefore unfeasible, if not impossible to produce ESTs directly from the mRNA species. After extraction of the mRNA transcripts from tissue samples, the mRNA transcripts are converted into complementary cDNA from which ESTs are generated. The general production of ESTs is illustrated in the flow diagram in Figure 3.

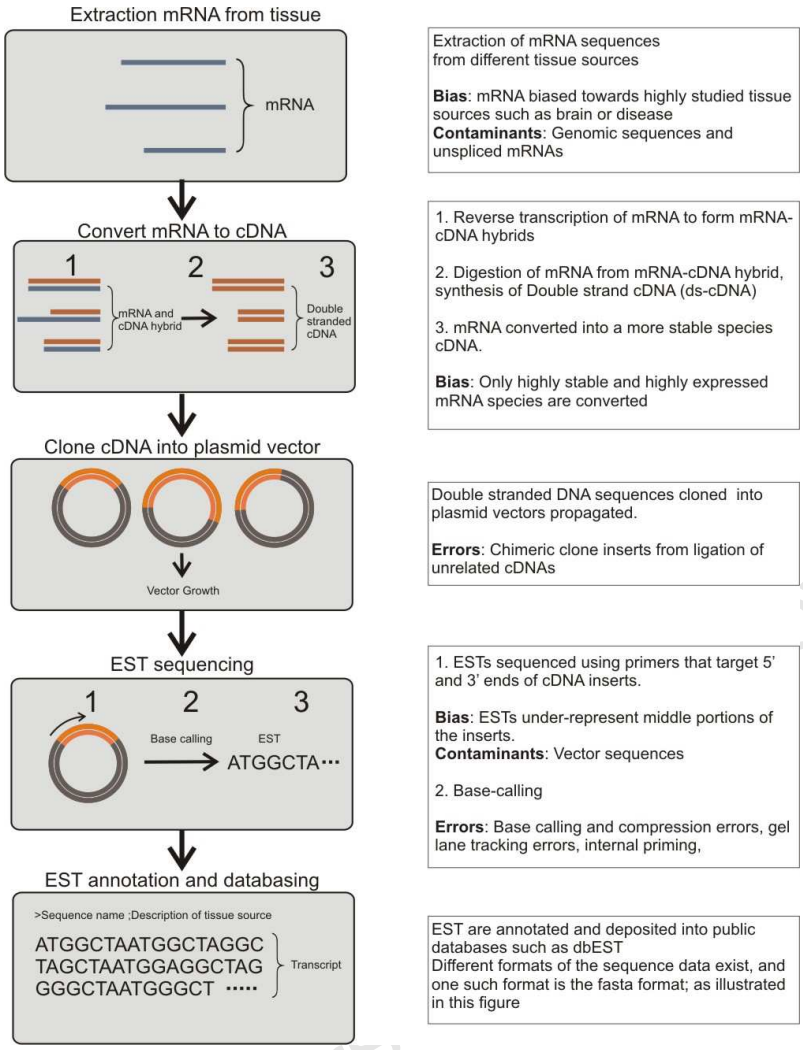


Figure 3: A general overview of the production of ESTs. The bias, contaminants and errors that can occur at each stage are highlighted.

The EST generation protocol allows for rapid generation of large quantities of transcripts. Hence there was a need to develop a number of databases aimed at housing and making ESTs accessible to the public via the internet. One of the largest databases is dbEST and the current version includes data from more than 1500 species with more than 50 million ESTs (see Chapter 1). Alternative splicing and many other EST-based studies are now possible because of these publicly available databases.

2.2.1.2 EST based detection of alternative splicing

Excluding house-keeping mRNA transcripts, genes express specific mRNA isoforms in different tissues or cell-types according to their physiological or biochemical requirements. The different research centres which sequence ESTs, extract mRNA sequences from a wide range of tissue-types and from different expression states which makes ESTs a rich resource for tissue-specific (Xu et al., 2002), disease-specific (Aouacheria et al., 2006), and developmental specific mRNA isoforms. What makes EST sequences even more valuable for research is that the tissue of origin, developmental stage and disease states are included in the annotations of ESTs. Ontologies such as eVOC have even been developed to make full use of such information provided by the sequencing centres (Kelso et al., 2003).

Identification of alternatively spliced mRNA isoforms

Prior to the whole genome sequencing efforts, alternative splicing studies were few because the commonly used RT-PCR method, could not paint a global picture of the prevalence of alternative splicing. Currently, there are several other large-scale non-EST based datasets for the detection of alternative splicing such as protein domains (Hiller et al., 2004b) and processed pseudogenes (Shemesh et al., 2006). However, ESTs have remained the major contributors and focal points in large-scale detection and verification of alternative splicing events.

Numerous algorithms have since been developed for the detection of alternative splicing using ESTs however the underlying strategy is based on aligning ESTs to a genomic/mRNA/intronic template sequence which allows introns to be identified as gaps and exons as the aligned parts (Modrek and Lee, 2002). Differences in insertions and deletions of at least two overlapping ESTs from the same gene could be interpreted as evidence of alternative splicing (Figure 4).

False negative and positives can occur as a result of bias, contaminants and sequencing errors as described in Figure 3, and thus verification of AS detected using ESTs is required. The main advantage is that ESTs provide the structural changes that alternative splicing imparts on mRNA isoforms. Experimental methods such as

microarrays and RT-PCR methods (Figure 4) that are commonly used for verification require *a priori* knowledge of the gene structure and can thus be performed much more readily. Exons and alternatively spliced splice sites detected via ESTs now frequently form the basis of large-scale microarray based AS detection and analysis (Johnson et al., 2003; Pan et al., 2006).

Verification of putative mRNA isoforms based on ESTs has largely thus far shown a high degree of accuracy in the EST based methods (Modrek et al., 2001; Brett et al., 2000). For example, out of 20 EST based putative mRNA isoforms, 16 were verified using the RT-PCR based approach (Brett et al., 2000). Furthermore, several alternatively spliced isoforms detected using EST data analysis had previously been reported by other independent researchers (Mironov et al., 1999; Brett et al., 2000).

ESTs are, on average, 300-500bps in length and are thus limited in the characterization of full length of alternatively spliced isoforms as they capture mainly partial gene segments (Nagaraj et al., 2007). Full-length cDNAs are more suited for the characterization of all splice variations occurring in a gene. Although the number of full-length cDNAs is increasing, they are still much fewer than ESTs (Modrek and Lee, 2002). Given the recent increase in sophisticated methods such as splicing graphs that are able to stitch together and permit visualisation of the full-length transcript isoforms from EST fragments (Malde et al., 2005; Eyras et al., 2004; Xing et al., 2004), ESTs are likely to remain an important source of data for the detection of alternatively spliced isoforms.

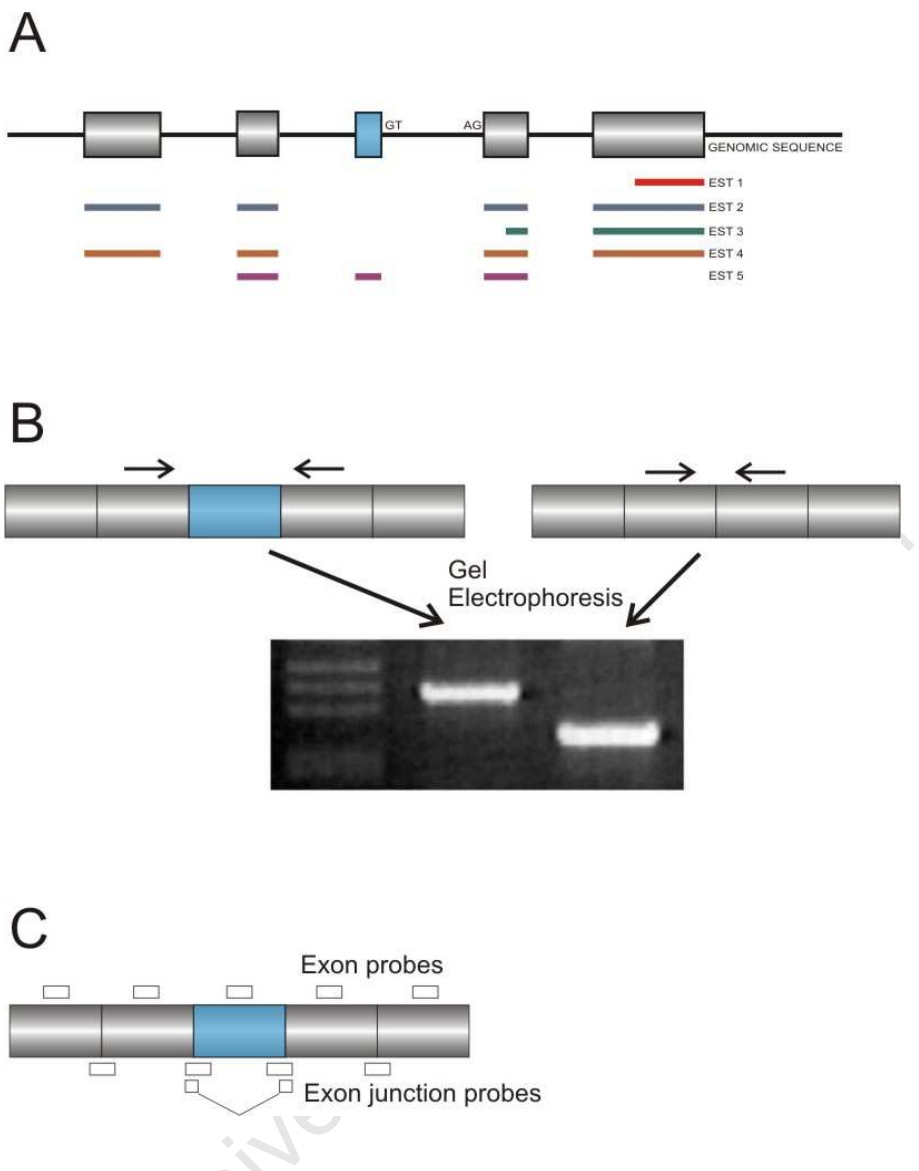


Figure 4: A hypothetical exon skipping event. A) Detection of AS using ESTs. EST 5 supports an exon inclusion which is alternatively skipped in all the other transcripts. B) RT-PCR primers are designed that are amplified to include the alternatively spliced exon. Use of a restriction enzyme to digest the mRNA isoforms followed by gel electrophoresis will clearly illustrate the differences in size of the mRNA transcripts. Adapted from (Modrek and Lee, 2002), and (Wang and Cooper, 2007). C) Exon arrays. Exon-probes are designed to target exon or/and exon junctions. Statistical tools are then used to determine the alternative splicing events based on the hybridisation of these probes from a single gene on spotted arrays or commercial arrays.

While mRNAs (Brett et al., 2000; Venter et al., 2001) and introns can be used as templates to which EST are aligned for the detection of AS, genomic sequences are the currently preferred templates (Kan et al., 2004; Kim et al., 2007; Modrek et al., 2001) (Figure 4). The use of genomic sequences as templates has several advantages. ESTs are assigned to their loci of origin (Modrek et al., 2001; Lee and Roy, 2004), thus

controlling false positives that could be introduced by paralogous sequences, sequence errors and contaminants. Furthermore, the exon/intron structures of isoforms can be confirmed and the AG-GT canonical donor and acceptor sites in genomic sequences that mark the bulk of functional intron termini can be used to substantiate intronic structures (Modrek et al., 2001).

2.2.1.3 Cross-species comparison

The identification of genome-wide AS in an organism using ESTs relies on the completeness of the ESTs in capturing all the transcriptomes (Modrek and Lee, 2002). However, ESTs under-sample the transcriptomes, even in the most highly represented organism such as human, resulting in the under-representation of EST-detected AS events. This has motivated an extension in the use of ESTs to include cross-species based detection of alternative splicing because the splice-sites are conserved across different genomes and in closely related genomes (Kan et al., 2004). Furthermore, transcriptomes sampled using ESTs from different organisms are never identical nor are they of the same magnitude (Brett et al., 2002). Therefore, aligning ESTs from one organism to genomes of closely related organisms can be used to detect novel isoforms missed during the commonly used transcript to genome comparison, within the same organism.

Human and mouse diverged approximately 100 million years ago (Kumar and Hedges, 1998) and therefore still share considerable high percentages of sequence similarity ranging between 70-88% at the nucleotide levels, between pairs of orthologs, in most functionally important genes (Makalowski et al., 1996; Makalowski and Boguski, 1998). Kan et al., (2004) exploited the conservation that exists between orthologous mouse and human genes by performing cross-species transcript to genome comparisons to detect AS. ESTs from human were aligned to the mouse genomic sequences for over 7000 orthologous gene pairs. The same approach was performed for the detection of AS in human. The main limitation of this method is that some orthologs are likely to have diverged to such an extent that ESTs from one organism cannot be aligned accurately to the genomic sequences of another. However, the study managed to positively identify previously published alternatively spliced exons and also discovered novel AS that could not have been identified

through transcript to genome comparisons of the same genome. This cross-species comparison approach promises to provide valuable insights of AS in species that are represented by very little EST data (see Figure 1 in Chapter 1).

2.2.2 Microarrays

Microarrays are platforms that permit characterisation of the transcriptome based on hybridisation between immobilised nucleic acids and nucleic acids from a tissue source under investigation. These were primarily designed for capturing gene expression levels and the immobilized DNA probes were designed to be complementary to the 3' untranslated regions (Srinivasan et al., 2005). After numerous exploratory studies it became clear that microarrays can be used to detect alternative splicing (Castle et al., 2003; Shoemaker et al., 2001; Srinivasan et al., 2005; Clark et al., 2002). When probes are designed for exons in a transcript (Figure 4), alternative splicing could result in either a loss or gain of hybridisation for specific probes targeting the affected regions relative to the other probes in the gene (Wang and Cooper, 2007). For example, the exon skipping event illustrated in Figure 4 would result in loss of signal from the probe set targeting the affected exon as well as the exon junction probes.

There are two ways in which probes can be designed, either by creating tilling arrays or creating probes for known exons/exon junctions (Pan et al., 2004; Johnson et al., 2003). While both methods require *a priori* knowledge of the gene structure, the exon junction method requires further knowledge of the splicing patterns of exons (Lee and Roy, 2004). Hence, the AS events detected using ESTs have frequently formed the basis of large-scale microarrays aimed at analysing alternative splicing (Lee and Roy, 2004; Modrek and Lee, 2002). These include tissue-specific splicing, and even quantification of alternatively spliced isoforms (Johnson et al., 2003; Pan et al., 2004; Pan et al., 2004).

Exon junctions greatly improve the coverage and reliability of microarrays in detecting subtle alternative splicing events and therefore, microarrays are increasingly becoming popular for the detection and analysis of alternative splicing events (Lee

and Roy, 2004; Lee and Wang, 2005; Wang and Cooper, 2007). However, in comparison to ESTs, microarrays are restricted in their coverage of transcriptomes (Johnson et al., 2003). Although microarrays have some well defined advantages, these platforms are unlikely to replace ESTs but rather compliment these versatile transcripts.

2.3 Databases of alternatively spliced mRNA isoforms

Large-scale detection of AS has led to the creation of numerous databases to facilitate research in this area (Table 1). The different methodologies and datasets used result in considerable differences in the databases dedicated to AS and these include the organisms analyzed, the number of mRNA isoforms, database structure and annotations (Bonizzoni et al., 2006). A unifying characteristic of most AS databases is that they are constructed from computational comparisons of ESTs to genomic DNA of known genes (Holste and Ohler, 2008). Thus, most of these AS databases retain the nomenclatures and annotations used in publicly available genome databases such as ENSEMBL (Birney et al., 2006), UNIGENE and SWISS-PROT (Boeckmann et al., 2003).

Most databases of computationally predicted mRNA isoforms have extra database-specific features in addition to the mRNA isoforms such as splicing conservation, putative splicing regulatory elements, predicted coding sequences and even include evidence from published literature to authenticate their computationally predicted AS splicing events (Table 1). These extra features aid in reducing the impact of nonfunctional AS errors associated with the technologies used in the detection of AS and the shortcomings of the algorithms used. This integration of AS databases to other resources which gives them extra features, makes it easier and less timing consuming for researchers to characterize AS events (Holste and Ohler, 2008).

Table 1: Databases of alternatively spliced mRNA isoforms

Database	Data type	Organisms	Extra Features	Other genome databases integrated with AS databases	References
ASAPII	ESTs and mRNAs	Human plus 13 other eukaryotic organisms	-conservation exons and introns. -predicted CDS	Unigene Genbank	(Kim et al., 2007)
HOLLWOOD	ESTs and mRNAs	Human Mouse	-Annotation of putative ESEs and ESS -Splice site scores	Ensembl Genbank	(Holste et al., 2006)
MAASE	Manually curated from published literature	Human Mouse	--	Ensembl Hugo	(Zheng et al., 2005)
SpliceInfo	Based on **PROSPLICER**	Human	-Secondary structure prediction tool -Other regulatory elements discovery tools	Ensembl Swiss-Prot RefSEq UniGene	(Huang et al., 2005)
PASLDB	EST, protein and mRNA alignment to genomic sequence	Human Mouse Nematode	--	GO terms	(Huang et al., 2002)
PROSPLICER	EST, Protein and mRNA aligned to genomic sequence	Human	--	Ensembl Hugo SwissProt Protein UniGene	(Huang et al., 2003)
Eusplice	Transcript annotations in GenBank	23 Organisms	Hypelinked to OMIM Etc	SwissProt: EMBL : UniGene Hugo Ensembl RefSEq OMIM	(Bhasi et al., 2007)

2.3.1 Impact of EST limitations on databases of alternative splicing

Microarray and EST technologies are each associated with their own limitations and these can influence the quality and reliability of mRNA isoforms in databases of AS. Currently, most AS databases are based on ESTs (Lee and Wang, 2005) and hence limitations of ESTs on AS databases will be discussed. Several contaminants, sequencing errors and biases innate in the EST generation protocols contribute to EST quality degradation (Figure 3). Such impurities and biases associated with ESTs can cause spurious detection of AS and this can have a significant impact on the interpretation of results from the databases of alternatively spliced genes. Therefore, it is essential to eliminate any contaminants and artefacts as well as take into account most of the biases. However, there is currently no way to guarantee that all the problematic sequences are filtered out. False positives and negatives can arise in the publicly available transcripts (Modrek and Lee, 2002).

False negatives

ESTs under-represent alternatively spliced isoforms due to the bias that exists in the EST generation protocol and also due to the nature of the mRNA transcripts. The biases in the EST data are reflected in studies which illustrate that only a fraction of genes are represented in the EST data. For example, only about half of the annotated genes on human chromosome 22 are represented by ESTs (de Souza et al., 2000). The likelihood of detecting AS using EST data seems to be positively correlated with the gene expression levels and the isoform expression levels. Genes with more than 300 ESTs are characterized by much higher numbers of mRNA splicing isoforms than lowly expressed genes (Kan et al., 2001). A study of genes that are highly represented by transcript data (that is, > 700 ESTs), showed that 99% of the genes are alternatively spliced (Modrek et al., 2001). Although some of this AS evidence could be false positives from sequencing or cloning artefacts, it is tempting to speculate that as more ESTs which capture the

transcriptome more comprehensively become available, it is possible that all multi-exon genes are indeed subjected to alternative splicing.

False negatives are also attributed to features of specific mRNA transcripts. The unstable nature of mRNA sequences contributes to false negatives in the detection of alternatively spliced mRNA isoforms since ESTs are biased towards the genes with more stable mRNAs. The less stable mRNAs are poorly represented in the EST databases, especially if they are extracted from samples such as pancreatic tissues in which numerous RNAase enzymes are found to occur.

The EST manufacturing protocol is commonly designed such that the start point of the EST generation protocol is controlled whereas the termination of the sequencing reaction is random. This creates a bias towards capturing of expression in the 3' and 5' ends of the genes and largely omitting the middle sections. In comparison to the 5' end, the poly-A tail of the 3' ends of mRNA sequences is an ideal sequence for targeting primers and hence most of the data is heavily biased towards ESTs that capture the 3' ends. Full-length cDNAs would be the most accurate in the detection of mRNA isoforms along the full length of the transcripts; however the number of cDNAs currently available is much less in comparison to ESTs. A recent study proposed a method that designs 5' and 3' primers and reiterates the process of designing primers and re-sequencing based on where the previous ESTs terminated resulting in the capturing of the cDNA clone insert along its entire length (Imanishi et al., 2004). This method holds great promise in making EST transcripts an even more attractive resource in effective deciphering of the transcriptome.

ESTs only capture AS isoforms of genes that are expressed in tissues and cell-types and developmental stages that were used in constructing cDNA libraries. For example, despite the *INI1* gene having over 100 ESTs in dbEST, novel mRNA transcripts are still being discovered (Favre et al., 2003). This is likely to be because ESTs in public databases did not sample all expression states in which this gene is expressed.

False positives

False positives could also be caused by contaminants, sequencing errors and ESTs capturing unspliced or partially spliced pre-mRNA which are still under-going the splicing process. For example, in unspliced or partially spliced pre-mRNA, introns could be falsely detected as intron retention events. In addition, various mechanisms discussed in the subsequent sub-section can also cause false positives.

2.3.2 Other causes of false positives

There are numerous other mechanisms and processes besides AS that result in the production of multiple mRNA isoforms. The AS spliced databases do not take into account all these false positives and therefore, their impact on these databases is currently unknown. These mechanisms and processes are discussed below.

RNA editing

RNA editing is a ubiquitous mechanism that occurs in all living organisms. It occurs post-transcriptionally and modifies the transcript by a single or multiple base insertions, deletions, conversions or substitutions (Niswender, 1998; Seeburg et al., 1998). The insertions and deletions can be erroneously detected as alternatively spliced introns and exons.

Random spliceosome errors

Just like any other biological enzyme or molecule, the spliceosome can also make errors (Graveley, 2001). Such random errors in splice site recognition by the spliceosome can result in the expression of novel mRNA transcripts. The presence of numerous cryptic AG and GT di-nucleotides that exist in the vicinity of authentic splice sites are a major factor in exacerbating spliceosome errors. When these AG and GT sites are found close

to authentic splice sites, the chances of slipping are increased substantially. Using mouse alternatively spliced cDNAs Chern et al., (2006) showed that the spliceosome slips occasionally particularly when AG or GT nucleotides exist close to the authentic splice sites. Such aberrant isoforms account for approximately 5% of AS isoforms in the public databases (Chern et al., 2006).

External stress to cells

Cells stressed as a result of heat, drastic pH changes or lack of oxygen *in-vivo* or *in-vitro* just before extraction of tissues can also lead to aberrant splicing. pH changes and lack of oxygen *in-vivo* can cause aberrant splicing by altering the distribution and sub-cellular localization of the enhancer and silencer *trans* acting factors such as the hnRNP A1 and tra2 (Daoud et al., 2002). The *ICH-1* gene is alternatively spliced depending on the concentration of the tra2 (Trasfomer-2) *trans*-acting factor. Brain ischemia (that is, restricted blood flow to the brain), results in a lack of oxygen which then alters the concentration of tra2 causing aberrant splicing in the brain (Daoud et al., 2002).

Changes in temperature *in-vitro* can also result in aberrant splicing (Colot et al., 2005). Heat can interfere with mRNA conformation and thus can make cryptic splice-sites more accessible to the spliceosome in comparison to the authentic splice sites (Varani et al., 1999). For example, a study on the *CAD* pre-mRNA transcript in hamster cells established that cryptic sites were activated and preferred when cells were heat shocked (Miriami et al., 1994). Cold has also been reported to induce retention of an intronic segment in the neurofibromatosis type 1 (*NF1*) mRNA (Ars et al., 2000).

Such sensitivity of the mRNA splicing process to external physiological changes such as temperature and pH highlights the need to mimic physiological temperatures and to perform fast extraction of mRNA transcripts from tissue samples. Prolonged storage of tissue cultures or cell-lines and slow processing of mRNA transcripts from fresh tissue increases the chances of aberrant splicing and the capturing of such mRNA isoforms by EST transcripts or microarrays. Disregard of such *in-vitro* induced aberrant splicing could inflate mRNA isoforms in databases of publicly and cause misinterpretation of alternative

splicing especially of medical importance. For example in a genetic screening of the *NF1* gene, a cold induced partial intron retention event was erroneously interpreted as being associated to the *NF1* disease (Ars et al., 2000).

Alternative transcription initiation sites

RNA polymerases initiate transcription after recognition of promoter sites. Approximately 18% of all human genes have been estimated to show evidence of alternative transcription initiation based on the presence of alternative promoter site usage (Landry et al., 2003). These alternative transcription initiation sites result in the generation of multiple mRNA isoforms in different cellular conditions such as tissue and developmental stages. The multiple mRNA isoforms generated from alternative transcription initiation are difficult to filter from mRNA isoforms produced through AS and thus inflate mRNA isoforms in the databases of alternatively spliced mRNA transcripts.

Alternative transcription termination and polyadenylation

After the termination of transcription during gene expression, the mRNA transcripts are cleaved followed by the addition of a poly (A) in a process known as polyadenylation. The poly (A) tail added to the mRNA transcript by the polyadenylation protects the transcripts from degradation and is required for export of the mRNA from the nucleus (Hunt et al., 2008; Ford et al., 1997). Variations in the poly (A) tail results in the production of mRNA isoforms from the same gene unit of variable length and activities. Numerous precursor mRNA transcripts have multiple alternative termination and polyadenylation sites (Shen et al., 2008). Alternative transcription, termination and polyadenylation are thus mechanisms that can spuriously inflate the number of mRNA isoforms in the databases of alternatively spliced isoforms.

Allele-specific splicing

Alternative splicing is the splicing of mRNA in different ways from identical transcripts. The production of distinct mRNA isoforms from different alleles of the same gene is known as allele-specific splicing, detailed in the next section.

2.4 Allele-specific splicing

Accuracy in the AS process depends on precision in the determination of exon-intron boundaries, splicing out of introns and annealing on the exons (Krawczak et al., 2006). As indicated in section 1, a myriad of different regulatory *trans*- and *cis*-acting factors are required. Mutations that disrupt any of the *trans*- and *cis*-acting factors involved in splicing causing inaccurate recognition of splice sites or alterations in splicing efficiency can affect both mRNA splicing and alternative splicing (Cartegni et al., 2002), leading to allele-specific splicing (Nembaware et al., 2004).

Cis-acting mutations can disrupt any of the core auxiliary elements specific to splicing (Figure 5). RNA structures have been shown to affect the distance between donor and acceptor sites, mutations that affect these structures can cause an increase in the distance between the splice site and enhancers reducing cross-talk between splicing factors that is required for the splicing process (see section 2.1) (Libri et al., 1995). The number of elements which if mutated could result in allele-specific splicing events is also increased by the coupling of the AS regulation to the regulation of mechanisms such as transcription (Kornblihtt, 2007). Unlike mutations that disrupt *trans*-acting factors, *cis*-acting mutations are easier to detect because they can be linked with much greater ease to the allele-specific isoforms they generate (see Chapter 1).

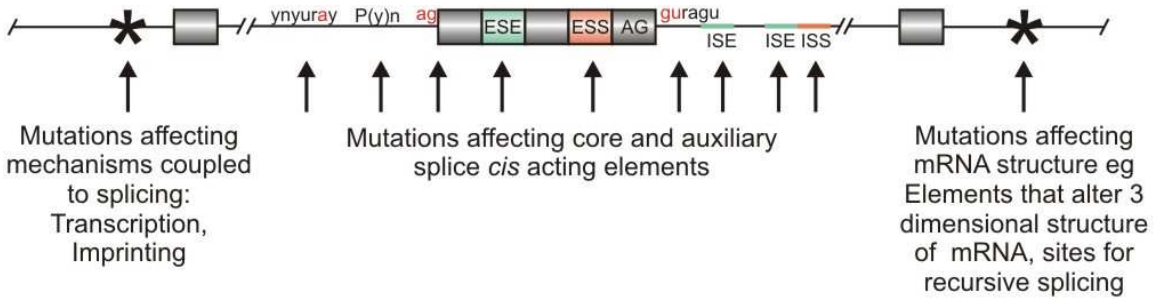


Figure 5: An Illustration of possible positions at which mutations can disrupt splicing. Part of the figure was adapted from Baralle and Baralle, 2005.

Cis-acting mutations have several possible outcomes on splicing patterns, which include exon skipping, activation of cryptic splice sites and intron retention. Important determinants of the outcome of *cis*-acting mutations include, the presence of a strong acceptor or donor site in the close proximity of a destroyed splice site (Krawczak et al., 2006), the type of *cis*-acting element affected (Cartegni et al., 2002; Krawczak et al., 2006; Berget, 1995), and the splice site recognition model that is used to recognize the exon-intron junctions (Berget, 1995). In the absence of strong cryptic splice sites in close proximity to authentic splice sites, mutations that disrupt the exon definition could cause exon-skipping, while errors in intron definition are likely to cause intron retention. Mutations that affect recursive splice sites and *intra-splicing* could lead to a variety of consequences which include truncated mRNA sequences. Knowledge of the consequence of a splicing mutation could be useful in predicting the phenotypic effect of an allele-specific splicing event.

2.4.1 Impact of allele-specific splicing on disease and pharmacogenetics

Cells have evolved complexes that recognize and eliminate aberrantly spliced mRNA isoforms. If translated into protein, mRNA transcripts can have a negative gain of function which is likely to lead to disease. A PTC located at least 50 bps upstream of the last exon-exon junction, triggers the NMD pathway (Lewis et al., 2003) which then degrades the truncated mRNA transcripts. Yet another pathway exists, which instead of

degrading the truncated transcripts, causes skipping of the PTC containing exon. This pathway is now commonly referred to as nonsense associated splicing (NAS) (Wang et al., 2002). Interestingly, in addition to biological screens for PTC containing transcripts, there seems to be yet another surveillance system that eliminates transcripts that have lost their stop codons (Frischmeyer et al., 2002; Maquat, 2002).

However, even with such proofreading mechanisms, some of the allele-specific splicing events slip through and become part of the cell's transcriptome. Allele-specific splicing can cause both qualitative and quantitative changes to splicing (Wang and Cooper, 2007) (Chapter 1). Both qualitative and quantitative allele-specific splicing changes can lead to individual-specific variations in disease severity, alter viral susceptibility and pharmacogenetic effects (Table 2) (Wang and Cooper, 2007).

Mutations that cause splicing changes have even been suggested to be the most frequent cause of human disease (Lopez-Bigas et al., 2005). Most of these changes are toxic to the cell (Wang and Cooper, 2007; Faustino and Cooper, 2003). However, some disease-causing qualitative changes induced by aberrant splicing mutations might not be novel mRNA splicing variants but could be AS events that get expressed in the wrong context. For example, when embryonic and neonatal isoforms of the dystrophy gene are expressed in the adult myotonic dystrophy tissues, diseases such as myotonia and insulin resistance can result (Ranum and Copper, 2006).

Drug efficacies, clearance rates, and responses even for approved drugs are highly variable between individuals with different genotypes (Bracco and Kearsey, 2003). Adverse drug responses are a leading cause of death particularly in developed countries. One of the aims of the pharmacogenetics field is to define the functional consequences of polymorphisms that have been linked to pharmacogenetic variations. Such information is required to design drugs with improved specificity and sensitivity. Numerous polymorphisms of pharmacogenetic relevance affect drug response by causing the expression of allele-specific isoforms (Table 2).

Table 2: Splicing SNPs that affect disease and pharmacogenetics. We have tabulated several SNPs from published literature that lead to disease or pharmacogenetic variations through an effect on splicing based on the naming convention of SNPs in dbSNP. Currently, the dbSNP naming convention is widely used to document and annotate any sequence variants that have recently been discovered (Wang and Cooper, 2007).

Gene	SNP	Cis-element	Quantitative /Qualitative	Description of Disease/ Pharmacogenenetic	References
CD45	rs12129883	ESS	Quantitative	Multiple Sclerosis	(Jacobsen et al., 2002)
COL5A1	rs13946	Acceptor site	Qualitative	Ehlers-Danlos syndrome	(Wenstrup et al., 1996)
PTPRC	rs17612648	NAGNAG acceptor	Qualitative	Multiple sclerosis	(Lynch and Weiss, 2001)
ITPA	rs13830	Exonic splicing silencer element in exon 2	Qualitative	An increased risk of adverse drug reactions in patients treated with the thiopurine drug azathioprine.	(Arenas et al., 2007)
OAS1	rs2660	Acceptor	Qualitative	Susceptibility to type 1 diabetes and viral susceptibility	(Field et al., 2005; Bonnevie-Nielsen et al., 2005)
PMM2	rs2072688	ESE	Qualitative	Carbohydrate-deficient glycoprotein syndrome	(Vuillaumier-Barrot et al., 1999)
LDLR	rs688	ESE	Qualitative	Increased cholesterol in pre-menopausal women	(Zhu et al., 2007)
SCN1A	rs3812718	Donor site	Quantitative	Possibly influence dosage requirement of anti-epileptic drugs carbamazepine and phenytoin	(Tate et al., 2005)
BRCA2	rs41293511	ESE	Qualitative	Breast cancer	(Fackenthal et al., 2002)
BTNL2	rs2076530	Donor site	Qualitative	Sarcoidosis	(Valentonyte et al., 2005)

2.4.2 Detection of allele-specific splicing

Detection of allele-specific splicing events and/or their causal mutations is not always trivial. The number of splicing regulatory elements which if disrupted can lead to aberrant splicing is augmented by the coupling of splicing events to other gene regulatory

systems. Furthermore, since some splicing events appear to be coordinated, the mutations that affect splicing in one region of the gene can also affect splicing patterns in a different region (Fededa et al., 2005) and this complicates the detection of splicing mutations even more. The complex nature and myriad of regulatory elements involved in splicing greatly increases the number of locations at which mutations can affect precision and efficiency of the spliceosome.

2.4.2.1 *Ab initio* prediction tools

One of the most widely used databases of heritable mutations is dbSNP (Sherry et al., 2001). Currently, dbSNP houses more than 10 million human SNPs. The availability of such large sequence variants coupled with the growing appreciation of the need to discover splicing regulating *cis*-acting elements prompted an increase in the of use *ab initio* prediction tools (Table 3). Such tools can be applied to publicly available databases of genomic variants such as dbSNP, which greatly enhances the characterisation of the functional impact of SNPs on disease. Numerous databases have taken this approach by scanning for splicing mutations in dbSNP using *ab initio* tools such as PupaSuite (Conde et al., 2006) and PolyMapr (Freimuth et al., 2005). More recently, the SNAP database has performed a very comprehensive scan by using six different *ab initio* tools on dbSNP entries (Li et al., 2007).

For each allelic version of the transcript, the *ab initio* tools predict the likely impact of the sequence variants on the strength of the splicing regulatory element. Usually a score is assigned which is supposed to correlate with splicing strength (Baralle and Baralle, 2005). Scoring systems are commonly based on comparing allelic versions of a given sequence to a *cis*-element consensus sequence. A splicing mutation that changes the regulatory element from its consensus altering the splicing scores is likely to lead to aberrant splicing.

Although consensus sequences were highly popular, dependencies and compensatory relationships between adjacent and non-adjacent sites exist (Yeo and Burge, 2004).

Matrices based on nucleotide frequencies do not always include such dependencies. Improvements on the *ab initio* algorithms and tools include incorporating dependencies between the nucleotides within regulatory elements (Churbanov et al., 2006; Yeo and Burge, 2004). Such an approach seems promising, one of these algorithms, the maximum entropy, that takes into account site to site dependencies in donor and acceptor sites was recently reviewed as one of the best performing tools among five algorithms in predicting the impact of a mutation on 3'splice sites (Vorechovsky, 2006).

Table 3: *Ab initio* tools used in the detection splicing mutations

Cis-acting element	Tool Name	URL	References
Donor and acceptor	NNSplice	http://www.fruitfly.org/seq_tools/splice.html	(Reese et al., 1997)
Donor and acceptor	MaxEnt	http://genes.mit.edu/burgelab/maxent/	(Yeo and Burge, 2004)
Donor and acceptor	GeneSplicer	http://www.cbs.dtu.dk/services/NetGene2/	(Pertea et al., 2001)
ESEs: SR SF2/ASF, SC35, SRp40, SRp55	ESEFinder	http://rulai.cshl.edu/tools/ESE/	(Cartegni et al., 2003)
ESE	RescueESE	http://genes.mit.edu/burgelab/rescue-ese/	(Fairbrother et al., 2002)
ESS	FAS-Hex2	http://genes.mit.edu/fas-ess/	(Wang et al., 2004)
Branch site	Branch Site Analyzer	http://ast.bioinfo.tau.ac.il/BranchSite.htm	(Kol et al., 2005)

The computational tools developed for predicting the impact of a sequence variant on splicing have had numerous successes (Eng et al., 2004; Vorechovsky, 2006). However, *ab initio* tools should only provide the first point of exploration when accessing the impact of a mutation on splicing (Baralle and Baralle, 2005). There are many instances where splicing scores fail to predict accurately the impact of a mutation on splicing (Carothers et al., 1993; Buchroithner et al., 2004). For example, a mutation in the *LAMB3* gene that creates a new splice site with a splice score of 68.6, is preferentially used over the wild type splice site, even though the wild type splice score has a much higher splice score of 92.2 (Buchroithner et al., 2004) using the scoring system of Shapiro and Senapathy (1987). Low scoring splice sites could still be functional if a strong ESE is

located in close proximity to the splice site. Furthermore, some changes that abolish one ESE could create a functionally compensating ESE at the same location (Cartegni et al., 2002).

Currently, there is no perfect method to estimate the false discovery rate of the *ab initio* tools and hence functional assays are thus required to confirm putative splicing mutations inferred from these computational analyses. The SNAP curators have realized the importance of experimental validations and have thus designed a web browser that facilitates the designing of primers for further experimental validation based on the putative splicing mutations in their database (Li et al., 2007).

2.4.2.2 Hybrid minigene assays

Hybrid minigenes are constructs of plasmids that contain a short RNA fragment of a gene under-study. About 20 years ago a minigene was first used to provide concrete evidence of alternative splicing and the need for the extra *cis*-regulatory elements such as enhancers and silencers in controlling splicing in addition to the canonical donor and acceptor sites (Vibe-Pedersen et al., 1984). Minigenes have since been altered for assaying splicing *cis*-acting mutations. For biallelic mutations two minigenes are created, for the wild-type and mutated alleles (Figure 6). The hybrid minigenes are then transfected into appropriate cell lines followed by RT-PCR and gel electrophoresis. Minigenes are suitable for the capturing and visualization of all allele-specific splicing events, but exon skipping seems to be the most suitable.

The minigene assays are quite popular since the affected tissues are not required and one can identify the causal mutation of observed allele-specific event without the need for human samples. However, this lack of tissue is one of the main disadvantages of using minigenes. As highlighted in section 1, AS is sometimes coupled to transcription and translation. Therefore, minigene assays are unsuitable for analysing allele-specific splicing events that are caused by complex interactions between multiple mutations or elements located far from the affected site. Minigenes are also not ideal for quantitative analysis of splicing mutations however some studies have modified minigene

experiments to suit this task (Hull et al., 2007). By using florescence to illuminate the exon skipped event, minigenes were used to detect the levels of skipping taking place (Hull et al., 2007). Furthermore, instead of performing normal RT-PCR (see Figure 6), qRT-PCR can also be performed.

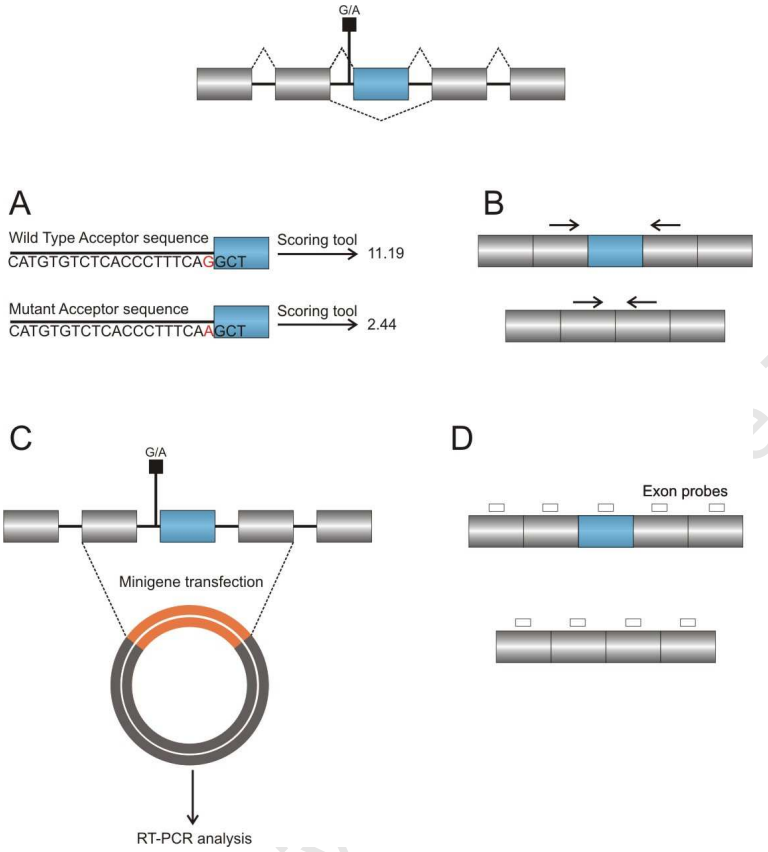


Figure 6: Detection of *cis*-acting mutations that cause splicing differences or their allele-specific products. A hypothetical example of a splicing mutation that disrupts a splice acceptor leading to an exon skipping event is used to illustrate how it can be detected or predicted using different methods A) *ab initio* prediction. B) Primers can be designed that target the allele-specific skipping event. Primers are shown as arrows in the figure. Gel electrophoresis can clearly indicate the skipping of the exon from the different alleles. C) Minigene which encompasses the splicing mutation and exon skipping events can be designed to prove that the G to A mutation leads to aberrant splicing. D) Exon arrays can be designed to illustrate the allele-specific skipping event. Probes are created for the exons of the affected gene. Cells from different individuals are then genotyped. Differences in exon expression from the different individuals can then be noted on the array platforms based on the binding affinities of the cell samples to the array platforms. The figure was adapted from (Baralle and Baralle, 2005) and (Wang and Cooper, 2007).

2.4.2.3 Microarrays

The first study that used microarrays for the detection of aberrant splicing was performed on the *PTCH* gene (Nagao et al., 2005). Probes for exon junctions and exons were used to

measure expression intensity in normal and wild type tissue and used to predict allele-specific splicing events. However, the main disadvantage was that the arrays could only be used for small-scale studies until the availability of genotyped cell lines from the HapMap dataset (The International HapMap Consortium, 2005).

The HapMap project was borne out of the need to further characterize the millions of SNPs discovered from the Human Genome Project (The International HapMap Consortium, 2005). The HapMap consortium genotyped 269 individuals, whose ancestors originated in Africa, Asia and Europe. The availability of the genotyped cell-lines from the HapMap consortium has marked a turning point in the use of arrays for the detection of allele-specific splicing (Nembaware et al., 2008; Kwan et al., 2007; Kwan et al., 2008; Hull et al., 2007). Using the HapMap genotyped cell-lines on the Affymetrix exon array platforms allows for large-scale association of alternative splicing events to polymorphisms and thus, the detection of allele-specific splicing (Chapter 5). However, such analyses do not prove a causal relationship between the SNPs and the allele-specific splicing events. Further experiments such as minigene assays are required to prove such causality.

2.4.2.4 Evaluation of functional assaying methods

The use of each method is associated with its own advantages and disadvantages (Table 4). Some of the methods do not actually detect the casual mutation; they can only prove association of a mutation to an allele-specific event. The most appropriate allele-specific detection method would be one that replicates the cell's environment as accurately as possible while allowing for the detection of allele-specific qualitative and quantitative splicing changes as well. However, this is exceptionally difficult for human studies. Because each method is not perfect, the most ideal would be an integrated analysis of some complementary methods to get the best results.

Table 4: Evaluation of methods for the detection of *cis*-acting splicing mutations using binary notation

Methods	Association	Causality	Qualitative	Quantitative	Tissue source required
RT-PCR	1	0	1	1 (only for qRT-PCR)	1
Minigene	1	1	1	0	0
Ab initio	0	1	1	0	0
Exon-Arrays	1	0	1	1	1

2.5 Concluding remarks

ESTs and microarrays have radically transformed the pace at which alternatively spliced mRNA isoforms are detected and the rate of discovery and characterization of AS regulatory networks. The growing number of characterized splice site recognition models and splicing regulatory networks allows for a greater ease in the discovery of allele-specific splicing events. Although there is a substantial improvement in the methods used for the detection and characterization of allele-specific splicing, there is currently no perfect method. Continual improvements in methods that detect allele-specific splicing will bring an improvement in our understanding of human variation in disease, pharmacogenetic responses and even an understanding of the general AS process.

Chapter 3

A Database of SNPs Mapped to ESTs

Abstract

Discovery of sequence variants associated with inter-individual phenotypic differences remains one of the most long-standing challenges in genetics. Studies based on transcribed Single Nucleotide Polymorphisms (SNPs), are increasingly becoming popular in genotype-phenotype associations since such SNPs can potentially alter protein sequences and/or affect gene expression patterns. In addition, transcribed SNPs are also ideal markers for unknown causal mutations that lead to allele-specific gene expression or allele-specific splicing. A clean dataset of transcribed sequence polymorphisms and their genotypes would therefore greatly facilitate studies aimed at characterising allele-specific gene expression/ allele-specific splicing. However one major drawback is the large number of chromosomes required for genotyping in-order to obtain reliable information. The use of sequence variants encoded in millions of publicly available Expressed Sequence Tags (ESTs) sequenced from a diverse range of ethnic groups, offers an affordable and accelerated strategy for genotyping and studying sequence variants. The scarcity of resources that allow users to harvest and visualize corresponding alleles from ESTs is a key obstacle to studying allele-specific gene expression and allele-specific spling. We have used genomic locations of SNPs and ESTs from the UCSC genome databases to determine the location of the SNPs on ESTs. We report a database, snp2estmap, of the nucleotides present at the polymorphic positions on EST sequences for 532 860 SNPs that could be mapped to 4 461 202 ESTs. To facilitate other studies, an interface was developed to make snp2estmap searchable via the internet at http://mancala.cbio.uct.ac.za/splicing. In addition, potential applications for this resource are discussed briefly.

3.1 Introduction

As a direct consequence of large-scale sequencing projects of cDNA clone libraries, publicly available databases of Expressed Sequence Tags (ESTs) have grown in size exponentially. Comparative evaluation of the current version of dbEST (February 2008) and the December 1991 version showed almost a 500 fold increase in the total number of dbEST transcripts. According to the pioneers of the EST sequencing protocol, the primary use of EST data was to facilitate rapid and inexpensive human gene discovery (Adams et al., 1992). The usefulness of such transcripts continues to expand mainly due to the diversity of genotypes, tissue-types and disease states represented within EST databases. Due to these attributes ESTs have become invaluable for the identification of gene expression variation across tissues (Megy et al., 2002), cancers (Aouacheria et al., 2006) detection of alternatively spliced isoforms (Lee and Wang, 2005; Kim et al., 2007) and for the discovery of transcribed single nucleotide polymorphisms (SNPs) (Buetow et al., 1999; Picoult-Newberg et al., 1999).

It is well known that the accuracy of SNP analysis based on ESTs could be greatly compromised by the error prone nature of these single pass transcripts (Nagaraj et al., 2007). Contrary to this argument, after implementation of pre-processing stages that reduce noise in transcript data, several lines of evidence have established EST data as a robust source of SNP allelic information (Buetow et al., 1999; Hayes et al., 2007). Experimental validation has indicated high accuracy levels of at least 80% in SNPs mined from ESTs (Buetow et al., 1999; Kota et al., 2003). Further validation arises from the consistency in the distribution of SNP allele frequencies between publicly available transcript data and other independent studies (Sunyaev et al., 2000), including the CEU population in the HapMap datasets (Ge et al., 2005). Using 2678 SNPs genotyped from EST data, a correlation co-efficient of 0.7386 was obtained when allele-frequencies estimated from the EST data and the CEU population were compared (Ge et al., 2005). Due to the geographical locations of the largest EST sequencing projects, it is not surprising that the EST data reflect the CEU population more than any of the other HapMap populations (African and Asian) (Ge et al., 2005).

The prevalence of allelic differences in gene expression in human is high (Lo et al., 2003). Several studies have suggested that the regulation of gene expression is mainly controlled from within the 3' and 5' untranslated regions (Rockman and Wray, 2002; Mossner and Riederer, 2007; Cowles et al., 2002). Due to the pre-existing bias of ESTs towards 3' and 5' gene regions, there is a high probability that most SNPs that cause allelic differences in gene expression are over-represented in the transcript data. Beyond, the EST-based SNPs actually causing allelic variations in gene expression, their usefulness can be extended to ascertain the presence of an unknown *cis*-acting regulatory variant based on the principle of linkage disequilibrium (Pastinen et al., 2004; Knight, 2004). This attribute of SNP information derived from ESTs has led to several investigations aimed at understanding allele-specific expression for cases where *cis*-acting variants are difficult to locate (Seoighe et al., 2006; Nembaware et al., 2004; Ge et al., 2005), or even unknown epigenetic factors that lead to imprinting (Seoighe et al., 2006; Yang et al., 2003).

The versatile nature of EST-derived SNP data is further demonstrated by their use in mutational and selection studies (Sunyaev et al., 2000). In fact, the first large-scale investigation of the pattern of selection and mutation in human was conducted on EST-based SNP data (Sunyaev et al., 2000). Allele frequencies from EST data were successfully used to measure selection at degenerate, non-degenerate sites and in the 5' and 3' untranslated regions of human genes (Sunyaev et al., 2000).

For genome-wide studies based on EST-derived SNPs, it is often difficult and time-consuming to determine SNP alleles from publicly available databases. A resource of pre-extracted SNP alleles from EST data would greatly facilitate further research. There are several databases that present SNP alleles extracted from ESTs (Hawken et al., 2004; Guryev et al., 2005). However such studies focus only on transcripts from model organisms such as rat (Huntley et al., 2006) zebrafish (Guryev et al., 2005; Huntley et al., 2006), and cow (Hawken et al., 2004). EST-derived SNP databases that focus on human data are not comprehensive. They do not make use of all publicly available datasets (Irizarry et al., 2000) or they focus solely on a single chromosome (Deutsch et al., 2001).

Currently no genome-wide resource exists that has pre-extracted SNP allele information from ESTs. By utilizing SNP and EST alignments to the human genome from the University of California Santa Cruz (UCSC) genome browser database (Hinrichs et al., 2006), we have created a resource that enables users to query a database using SNP accessions from dbSNP or EST Genbank accessions, returning among other things potentially useful annotations in addition to the SNP alleles represented on the ESTs. We also highlight how this resource can be useful to the scientific community at large by indicating the various ways in which this resource has been extensively used throughout this thesis.

3.2 Data and methods

3.2.1 Data

The UCSC genome browser database provides annotation tracks (Hinrichs et al., 2006), consisting of tables of objects pre-aligned to genomic sequence. We downloaded SQL tables of pre-computed EST and SNP genomic locations from the UCSC annotation tracks version hg17, which were based on NCBI genome assembly version 36 (Hinrichs et al., 2006). dbSNP has multi-nucleotide substitutions, indels and bi-allelic polymorphisms commonly referred to as SNPs. We choose to focus only on SNPs as they form the majority of sequence variants in the dbSNP database and are easiest to genotype from transcript data.

Human EMBL flat files release 89 was downloaded from the EBI ftp server (ftp://ftp.ebi.ac.uk). A total of 7,222,889 million transcript data with cDNA clone library annotations were extracted from the EMBL flat files using Perl scripts.

3.2.2 Quality control

dbSNP is the largest and the most widely used database of genomic variation. In addition to entries deposited by individual researchers, smaller population-specific databases such

as JSNP (http://snp.ims.u-tokyo.ac.jp/), also contribute to dbSNP. Although the overall quality of dbSNP entries is considered to be very high, some of the SNPs are of poor quality due to sequencing artifacts (Platzer et al., 2007). Furthermore, many genes in the genome are part of large gene families. SNP sequences may have multiple high-quality alignments to the genome. To reduce the impact of paralogous sequences and the impact of sequencing errors we removed SNP entries that mapped multiple times to the genome.

ESTs can also map ambiguously to the genome due to the many paralogs that exist in the human genome, sequencing errors and sequence contaminants. To alleviate the impact of paralogous sequences and sequencing errors, ESTs that mapped multiple times to the genome were also discarded. In order to lower the influence of chimeric sequences and contaminants from non-human sequences, we only considered ESTs that mapped with at least 90% of their total length to the genome.

3.2.3 Extraction of SNP alleles from ESTs

Perl scripts were designed to detect all SNPs found within genomic boundaries of ESTs. We then used SNP and EST genomic coordinates from UCSC to extract the corresponding SNP alleles from each polymorphic site along the EST. A flow-diagram in Figure 1 illustrates the general procedure used to extract SNP alleles from ESTs.

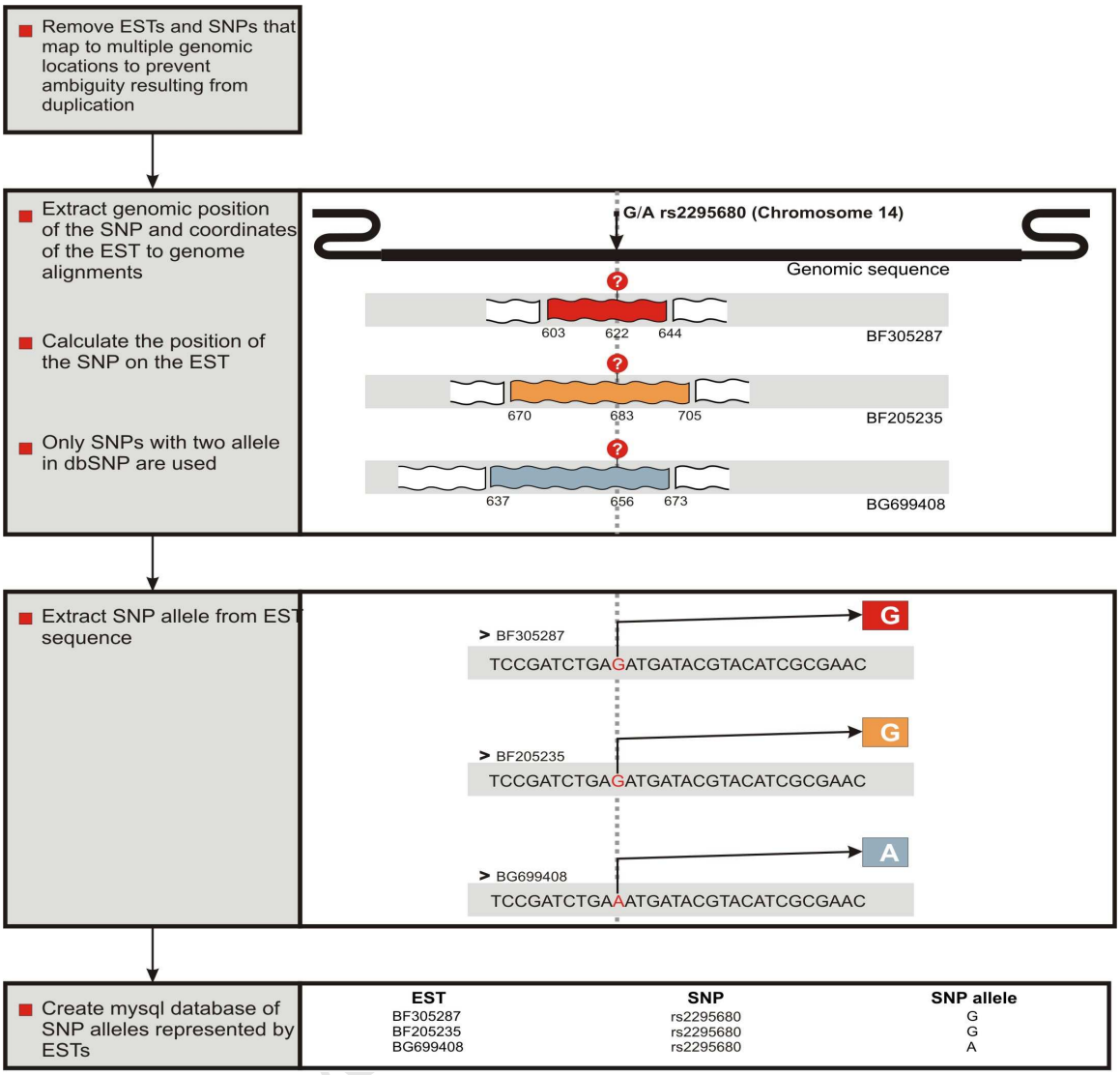


Figure 1: Flow diagram of our method for the extraction of SNP alleles from expressed transcripts. We have illustrated using SNP rs2295680 and only 3 transcripts how the snp2estmap mysql database was created. Entries that map ambiguously to the genome in the UCSC EST and SNP tracks are discarded. Genomic positions of SNPs located within ESTs are extracted and used to determine the position and allele of the SNP on the EST.

3.2.4 eVOC ontologies

The curators of eVOC (Kelso et al., 2003) developed four ontologies using a highly comprehensive and controlled vocabulary to describe human gene expression data. The ontologies include Anatomical System, Cell Type, Pathology and Developmental Stage. Thus far, the eVOC database has annotated 7016 cDNA libraries from dbEST.We downloaded over 6000 cDNA libraries annotated according to Pathology from eVOC website www.**evoc**ontology.org (Kelso et al., 2003). Only cDNA libraries clearly

annotated as originating from cancer or normal tissues were considered further. These cancer annotations were then used to classify ESTs into categories cancer or normal tissues.

3.3 Results

3.3.1 A database of SNPs mapped to ESTs

Based on UCSC transcript alignments to the human genome, we created a database of 532,860 SNPs that map to ESTs. Though deletions/insertion variants and multi-nucleotide substitutions exist, our analysis was based only on bi-allelic SNPs The EST-based SNPs represent approximately 5% of all SNPs that mapped to the genome (Table 1). Some SNPs that were located within ESTs could not be allocated alleles because of masking at the exact SNP locations in the EST transcripts or unavailability of the transcript data due to incongruence in the updating of the EMBL database of transcript data and UCSC genome browser database. ESTs are contaminated with vector sequences during the experimental procedure. Repetitive elements are also frequent in EST data, these include ALU, SINE and LINE elements. EMBL transcript data is put through a masking process to screen ESTs for such contaminants. This masking step is evident in the transcript data where strings of NNNs' have replaced the repeats and vector sequences.

A significant limitation of this study and many other similar studies is that many SNP alleles extracted from EST data potentially result from poor quality sequence data that is characteristic of single-pass sequences (see Chapter 1). To reduce the impact of transcript noise on our results, only ESTs that mapped with at least 90% of their sequence length to the genome were considered. Part of the noise in EST data is from chimeric sequences which result from sequence fragments from two separate cDNA libraries that erroneously ligate into one transcript. Non-human sequences are also likely to be filtered out using the criterion of enforcing that most of the transcript maps to the genome.

Paralogous genes occur frequently in human and can cause SNPs and ESTs to map multiple times to the genome. About 6% of SNPs and 5% of ESTs mapped ambiguously

to the genome (Table 1). The fragmentary nature of the EST and SNP data could also result in these sequences having multiple high similarity matches to the genome purely by chance alone, leading to inaccurate EST or SNP to genome mappings.

Table 1: Summary of the processing of data for snp2estMap

	ESTs	SNPs
Total	7385922	11647909
Duplicates	365088	704032
Mappings (with allele data)	4461202	532860

3.3.2 Annotation of ESTs using eVOC

Based on the eVOC Pathology cDNA library annotations (Kelso et al., 2003) we categorised 4729 libraries as cancer libraries and 2232 as non-cancer libraries. We used the pathology annotated cDNA libraries to integrate the eVOC gene expression information with the SNP to EST mappings. The integration of the database of SNPs mapped to ESTs with eVOC enables researchers to test for associations between cancer expression states and SNP alleles represented on ESTs.

3.3.3 Web interface

We created a simple interface to our MySQL database which is powered by CGI and Perl scripts, available at http://mancala.cbio.uct.ac.za/~victoria. Using SNP rs2295680 found on the *RBM23* gene, as an example, a screen shot of the web-server is shown in Figure 2.

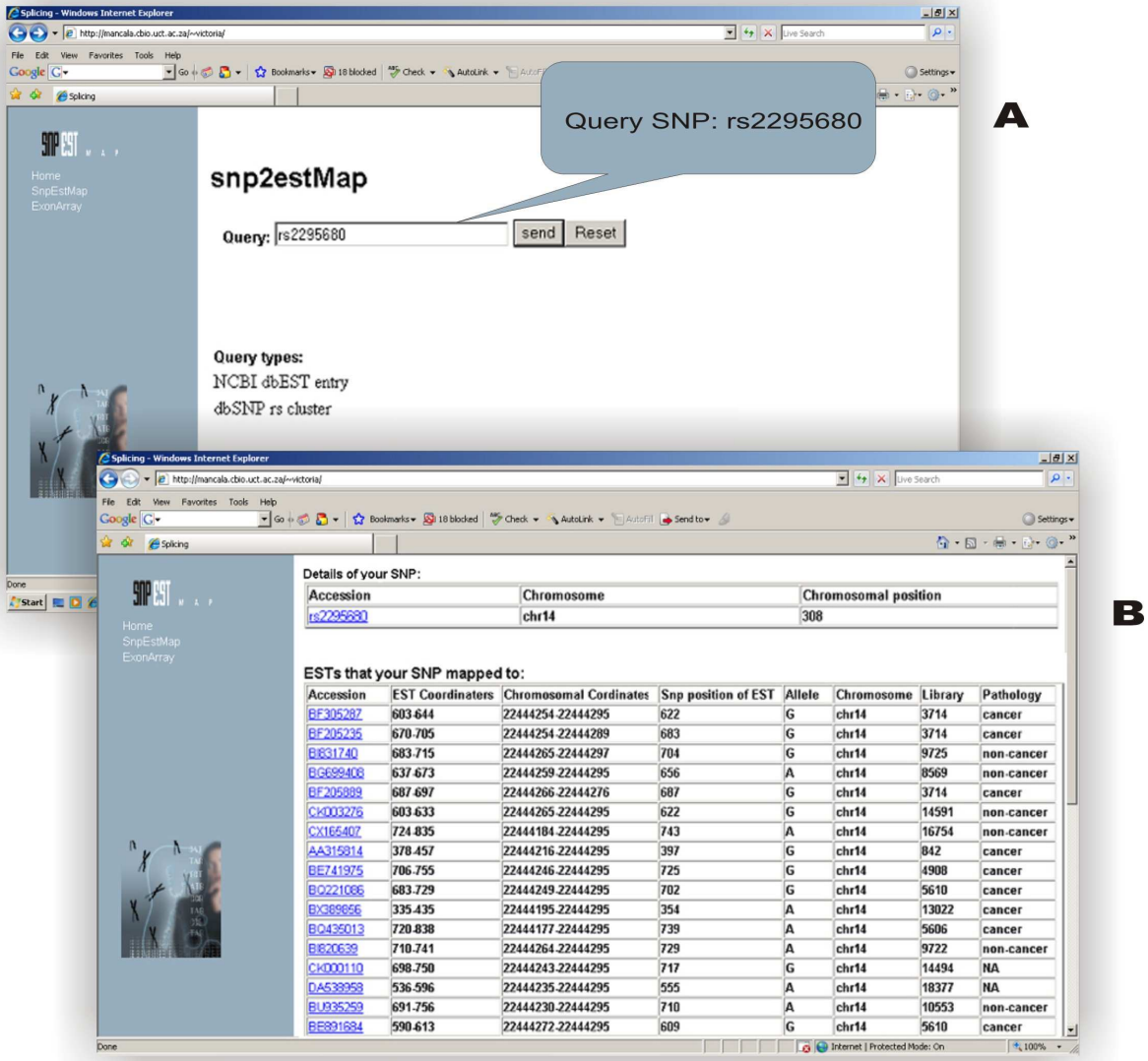


Figure 2: A screen-shot illustrating the web-pages that can be used to query the snp2est database and the output page using the SNP rs2295680. The SNP and EST identifiers are linked to their original sources dbSNP and dbEST respectively.

3.4 Discussion

Major bottlenecks in the application of SNP allele information encoded within publicly available EST data are the integration of at least two databases dbSNP (Sherry et al., 2001) and dbEST (Wheeler et al., 2007) and the SNP extraction process. To facilitate

such investigations we have provided a resource whereby researchers can obtain pre-computed allelic information of SNPs that map to ESTs.

Snp2estmap is likely to contribute significantly to allele-specific gene regulation studies. Recently, it has become well established that SNPs located within coding regions can have dual effects; modifying protein sequences and/or causing allelic differences in gene expression (Cartegni et al., 2002). It has now become imperative to assess all coding SNPs with regard to their impact on gene regulation regardless of their impact on protein sequences. The EST-based SNPs could disrupt regulatory elements located within genic regions. Such SNPs can influence gene expression by modifying mRNA transcription binding sites (Mossner and Riederer, 2007) or even by altering mRNA splicing patterns (Fairbrother et al., 2004; Cartegni et al., 2002).

Allelic imbalances in expression are highly prevalent in human (Lo et al., 2003), and contribute to a wide range of phenotypic differences which include depression in the highly prevalent Parkinson's disease patients (Mossner et al., 2001), and susceptibility to cancers (He et al., 2005). Besides actually causing allelic differences in gene expression, the use of transcribed SNPs as genetic markers in detecting allelic differences in gene expression is now well established. Some *cis*-acting variations that alter gene expression are likely to be located in intronic regions making transcribed SNPs potential highly informative markers if they are in linkage disequilibrium with the unknown regulatory variants genetic or epigenetic in nature (Ge et al., 2005; Pastinen et al., 2005).

In comparison to EST-based SNP analysis, there are several high-throughput methods capable of high precision and accuracy for studying allelic imbalances in gene expression (Kwok, 2001). Some of these methods include use of oligonucleotide arrays that are specifically designed to quantify allelic gene expression differences (Pant et al., 2006. However, microarray based studies can be expensive, laborious and can be limited by the unavailability of tissue samples. ESTs offer an affordable and convenient alternative for quantifying allelic gene expression differences using existing publicly-available data. ESTs have been shown to be ideal for genotyping SNPs due to the redundancy of

transcripts sampled from many cDNA libraries originating from different individuals. Unlike other captured data types which only capture a small number of tissue and disease states, human EST databases currently represent expression states in thousands of different tissues and disease states. Tissue specificity in allele-specific expression is possible (Cowles et al., 2002), and ESTs offer an opportunity to detect this. There are also several expression states represented in EST data which can be accessed through carefully designed controlled vocabularies (Kelso et al., 2003).

Transcribed SNPs detected from EST data have emerged as important means to assess relationships between genotypes and disease phenotypes (Aouacheria et al., 2007). The normalization and subtraction (Bonaldo et al., 1996) procedures that were developed to improve transcript sampling of lowly expressed genes may increase the chances of capturing rare alleles which are likely to be deleterious mutations. ESTs are also sequenced from different pathological states, with a substantial bias towards cancerous tissues. The integration of cDNA libraries annotated according to the cancer status of the tissue from which they were sequenced using eVOC ontologies (Kelso et al., 2003), with the snp2estmap database enables researchers to search for associations between SNP alleles and cancer expression states.

The use of snp2estmap has led to a wide-range of studies specifically for this thesis. Through integration of the EST data with other annotations such as splicing and the pathology state of cDNA libraries we enhanced the utility of snp2estmap. Table 1 lists some of the studies in which we have applied data from snp2estmap. This resource will also allow for similar studies to be performed with greater ease and speed.

Table 1: Studies in this dissertation that were based on the snp2estmap database

Aim of study	Additional Data	Chapter	Publication
Detecting allele-specific expression and imprinting	Pathology annotations from eVOC	Chapter 3	(Seoighe et al., 2006)
Estimating prevalence of allele-specific splicing events	mRNA splicing patterns from ASAPII (The Alternative Splicing Annotation Project II) (Kim et al., 2005)	Chapter 4	(Nembaware et al., 2004)
Detection of allele-specific splicing events	mRNA splicing patterns from the ASAPII database	Chapter 5	(Nembaware et al., 2008)

A future prospect is to expand this resource to encompass as many organisms as possible. In Chapter 6, a similar database was also created and used to analyse strain-specific splicing in mouse. Given that most organisms have varying densities of SNPs some of which are linked to their virulence in the case of disease causing micro-organisms (Spatz and Silva, 2007) or variations in milk quality in sheep (Pirisi et al., 1999), identifying SNP alleles represented on the EST transcripts from many different organisms would provide a useful resource for researchers to harness important sequence variants.

Chapter 4

Estimation of the Prevalence of Allele-Specific Splicing in Human

Abstract

Alternative pre-mRNA splicing is common in multi-cellular organisms, and is estimated to affect 70% of human multi-exon genes. However, high estimates of the prevalence of alternative splicing are based on methods that are incapable of discriminating between mRNA isoforms due to alternative splicing and mRNA isoforms that result from polymorphisms that affect splicing. Although many examples of genes that are spliced in an allele-specific manner have been reported in the literature, no comprehensive genome-wide estimates of the proportion of alternatively spliced genes that are affected by such polymorphisms have been carried out. Based on an integrated analysis of the dbSNP, dbEST and ASAP databases, we find that alternative transcript isoforms are non-randomly associated with closely linked single nucleotide polymorphisms. From the observed level of association between transcript isoforms and single nucleotide polymorphisms, we estimate that 21% of alternatively spliced genes are affected by polymorphisms that either completely determine which form of the transcript is observed or alter the relative abundances of some of the alternative isoforms. We provide a conservative lower bound of 6% on this estimate and point out that alternative splicing of a gene cannot be confirmed with certainty unless more than one alternative mRNA isoform is observed from the same allele.

4.1 Introduction[1]

The widely accepted definition of alternative splicing is that it is a process which produces different mature mRNA sequences from a single pre-mRNA transcript (Lopez, 1998; Graveley, 2001). However, polymorphic versions of genes can also give rise to different mRNA isoforms, (which we shall refer to this as allele-specific splicing here-after), which can easily be erroneously categorized as products of alternative splicing. Numerous examples of allele-specific splicing have been reported in the literature (Krawczak et al., 1992; Khan et al., 2002; Cartegni et al., 2002; Liu et al., 2001). Even transcripts that differ by just one single nucleotide polymorphism (SNP), can potentially produce different mRNA isoform if the mutation disrupts crucial splicing regulatory elements. Heritable point mutations that abolish the splice donor and acceptor sites can lead to activation of cryptic splice sites or promote exon skipping events (Spritz et al., 1981). More intriguing are recent reports that have shown that coding polymorphisms whether nonsense, missense or silent can alter splicing patterns by causing exon skipping events if they disrupt splicing enhancer or create silencer elements (Cartegni et al., 2002). The participation of coding sequence variants in the regulation of splicing adds an extra class of previously disregarded mutations when evaluating allele-specific mRNA isoforms. Given the growing number of publications that report allele-specific splicing events, it is surprising that the contribution of polymorphisms to transcript variation is frequently overlooked and remains unquantified.

Allele-specific splicing can involve not only qualitative changes to mRNA isoforms but also quantitative differences in relative isoform abundances (Buchner et al., 2003; Cartegni et al., 2002). Qualitative changes involve introduction of novel mRNA isoforms with altered exons, whereas quantitative changes, involve alteration of the relative abundances of pre-existing alternatively spliced mRNA isoforms (Buchner et al., 2003). Such structural and dosage alterations often have drastic effects on protein functions. In fact, candidate genes influencing susceptibility to complex diseases such as cancers, have

[1] This chapter is presented in the context of the literature as it was when we undertook and published the study in 2004. Refer to (Nembaware et al., 2004).

been found to produce allele-specific mRNA isoforms which differ between normal and affected individuals (Khan et al., 2002). The contribution of natural genetic variations to the diversity of splice isoforms should therefore be investigated and not overlooked as this holds great promise in contributing to our understanding of diseases and enhancement of human health.

Most databases of alternatively spliced transcripts are detected based on inconsistencies in alignments of ESTs (Expressed Sequence tags) (Modrek et al., 2001; Kim et al., 2005), which are sampled from a diverse range of populations. ESTs are rich in polymorphisms - 50% of ESTs have been reported to be highly polymorphic (Picoult-Newberg et al., 1999). However, the current databases of AS transcripts have so far not distinguished between true alternative splicing and transcripts that arise from polymorphic versions of the genes. In consequence, the contribution of such allele-specific splicing to the current estimates of human multi-exon genes that are alternatively spliced is unknown.

Krawczak et al., (1992), estimated that 15% of human genetic disorders attributable to a single gene involve aberrant splicing caused by single nucleotide mutations. This estimate was based on mutations located in the canonical donor and acceptor sites. Adapting the method employed by Krawczak et al., 1992 to estimate the prevalence of allele-specific splicing in human, would lead to a gross underestimate. Mutations that affect splicing are not restricted to the highly conserved GT and AG donor and acceptor sites, and these represent only a small fraction of mutations that could potentially alter splicing efficiency (Chapter 2). Therefore, estimates which are not based on counts of splicing mutations are needed to quantify the influence of sequence variants splicing.

We developed an alternative approach for the estimation of allele-specific splicing which is based on the principle of linkage disequilibrium between exonic SNPs observed on ESTs and unknown splicing regulatory SNPs. We have used ESTs that can be mapped both to SNPs and alternatively spliced isoforms to estimate prevalence of allele-specific splicing in ASAP, a database of alternatively spliced genes (Modrek et al., 2001). Although exonic SNPs might not be the cause of allele-specific splicing in most

instances, they can be in tight linkage disequilibrium with a mutation that affects splicing patterns, which in most instances is located within intronic regions. However the approach is valid even if the observed exonic SNP is the cause of the allele-specific splicing. We propose a sequential mutation model see Figure 1, in which the two separate linked mutations occurred in an unknown order. In this model, the more recent polymorphism will always be found in association with a single form of the earlier polymorphism. If one polymorphism results in qualitative allele-specific splicing and the other is detectable as an exonic SNP, then this would mean that one combination of transcript and SNP should never be present in EST databases. On the other-hand if polymorphisms do not affect the transcript isoform that is observed, then the choice of transcript isoform should be independent of the SNP allele and all four isoform/SNP combinations could occur in ESTs.

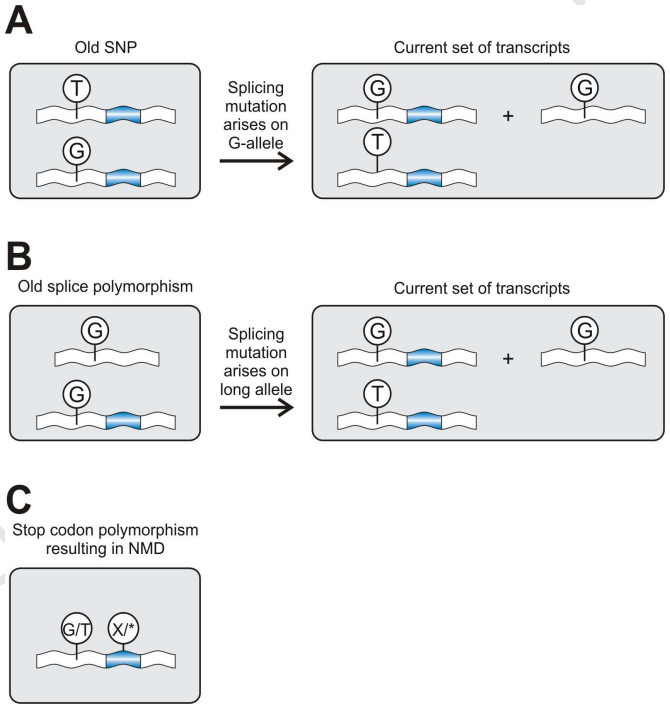


Figure 1: The sequential model of mutations that could give rise to associations between a SNP and alternate isoforms of a gene. One isoform has a skipped exon shown in blue while the major isoform retains the exon **A).** The SNP mutation occurred first, followed by a mutation that leads to an exon skipping event. **B)** In this scenario, a splicing mutation that causes skipping of the exon occurs first followed by the SNP mutation which is observed in the EST transcripts. **C).** A premature termination codon (PTC) in a cassette exon (blue). The PTC polymorphism is linked to an exonic SNP elsewhere on the gene, resulting in nonsense-mediated decay of the longer isoform from the allele containing the stop codon.

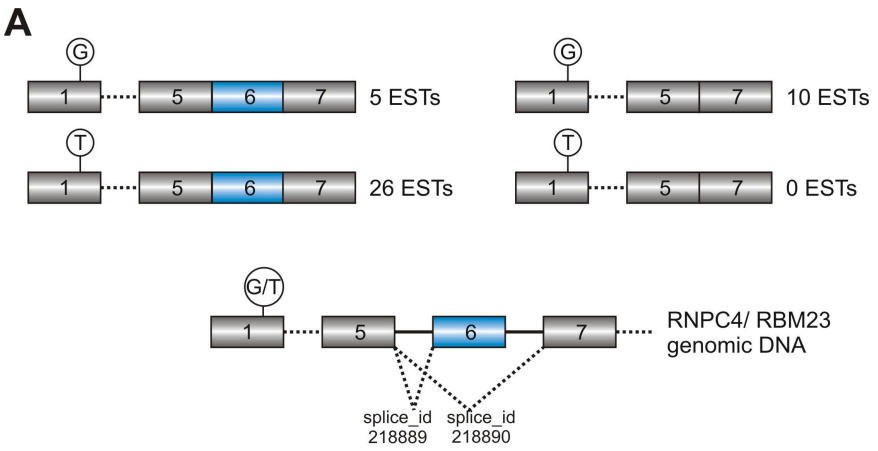
4.2 Data and methods

Pre-computed EST and SNP genomic locations were downloaded from the UCSC Genome Browser (Karolchik et al., 2003) which is based on the NCBI genome assembly 35 (Hubbard et al., 2002). ESTs and SNPs that mapped onto more than one genomic position were discarded to avoid spurious EST to SNP matches from paralogous sequences and sequence contaminants. If multiple alleles were observed at a particular SNP position, only the two most common alleles were considered further. The SNP-EST mapping procedure has been described in detail in Chapter 3, based on a more recent version of the UCSC genome data.

ASAP (Modrek et al., 2001) a database of alternatively spliced gene clusters, was downloaded. This data included short sequence fragments from alternatively spliced exon junctions associated with Unigene clusters. We mapped splice junctions involved in alternative splicing to EST sequences from the same Unigene cluster by scanning the ESTs for exact matches to the splice junctions.

4.2.1 Data matrices

We counted the numbers of EST representing each isoform/SNP combination in a 2X2 data matrix where the columns correspond to transcript isoforms and the rows correspond to SNP alleles for a particular cDNA library. Figure 2 illustrates how matrices were created with data from the *RPNC4/RBM23* gene. In order to eliminate potential bias resulting from multiple ESTs derived from the same tissue sample, we restricted the matrices to a single EST per clone library per SNP allele. To estimate the prevalence of allele-specific isoforms, only one SNP and one splice junction pair per Unigene cluster was considered as multiple splice junctions and SNPs from the same gene may not be independent. The chance of detecting polymorphic splice isoforms which are expressed at very low frequency was increased by selecting the rarest splice junction and a highly represented SNP in each case and the restriction to a single EST per allele per library was maintained.



B

SNP	SPLICE VARIANT 218889	SPLICE VARIANT 218890
G	5	10
T	26	0

Figure 2: A) Illustrated example of allele-specific isoforms detected in the *RNPC4/RBM23* gene. ESTs supporting either the major or minor isoforms are shown above the genomic sequences. ESTs consistent with the major transcript isoform are shown on the left while ESTs consistent with the minor isoform with a skipped exon 6 are shown on the right. 10 ESTs support the skipping of exon 6, all of which have a G at a SNP site in exon 1. Of the ESTs from transcripts that include exon 6, five have a G at the polymorphic site and 26 have T at this site. **B**) A data matrix corresponding to the *RNPC4/RBM23* example in **A).**

4.2.2 Simulations

We carried out simulations to model the distribution of transcript isoforms at each polymorphic site in the absence of association between transcript isoforms and SNP alleles. The computer simulations were performed by constructing random replicates of the data under a null model of no association between SNP allele and mRNA isoform, but with the restriction that row and column sums for each matrix were conserved. The number of matrices in the simulated datasets with a zero cell were counted and used to estimate the expected number of matrices with a zero cell in the absence of an association between transcript isoforms and SNPs. Confidence intervals were determined and these included the values obtained from 95% of the replicates. We also constructed randomized datasets with a bias such that a strong, but not exclusive, association between a row and a

column of the matrix was introduced in a randomly selected proportion of the matrices. Using these simulated datasets we estimated the proportion of affected matrices required to reproduce the observed number of matrices with a zero cell.

Fisher's exact test, a non-parameteric test, was used to test the null hypothesis of no association between the transcript isoforms and the SNP alleles for individual matrices. The Bonferroni method was used to correct for multiple testing.

4.3 Results

4.3.1 Matrices

ASAP (Modrek et al., 2001), is one of the many databases of alternatively spliced isoforms based on gene expression evidence from dbEST (Wheeler et al., 2007). To harness allele-specific splicing information in the ASAP database, our approach relies on SNPs and transcript isoforms that can be mapped to transcript sequences from the public databases. We created a database of 1295, 2X2 matrices from ESTs that can be mapped to SNPs and alternative mRNA isoform pairs. Figure 2 illustrates how the matrices were constructed using the *RPNC4/RBM23* gene as an example. Matrices that have a column or row sum of 1 are forced to at least have a zero cell, which we cannot be certain to be caused by an allelic effect. Such matrices could simply be reflecting insufficient data. From this database of matrices, we observed 139 matrices that contained rows and columns each summing to at least two, thus representing at least two different mRNA isoforms and two different alleles. We considered these 139 matrices to have sufficient EST data to analyse allele-specificity, hence our analysis was based only on these matrices.

4.3.2 Estimation of allele-specific splicing in human

In the case of qualitative allele-specific isoforms, every matrix that is affected necessarily has at least one zero cell. The maximum number of zeros possible (i.e two zeros), are

expected where an exonic SNP has a direct qualitative effect on splicing and each SNP allele is completely associated with exactly one transcript isoform. However, matrices with two zero cells are not always expected for exonic SNPs that actually cause qualitative allele-specific splicing. For example, one SNP allele may give rise to the wild type isoform while the newly acquired SNP allele causes the expression of a mixture of both alternate isoforms. A concrete example of a one zero cell matrix from qualitative allele-specific splicing is shown in Figure 2B. Among the observed matrices, 85 had a zero cell (i.e. one combination of SNP allele and mRNA isoform that did not occur) and were thus consistent with qualitative allele-specific splicing. In equivalent sets of randomized matrices, on average, 71 matrices had a zero element (Figure 3). Only four out of 1000 replicates had as many or more matrices with a zero cell as found in the observed data (i.e 85 matrices). There are 14 extra matrices in the observed data (85 -71 matrices), with at least one zero cell is not likely to have resulted from random data with no relationship between SNP allele and transcript isoforms and instead indicates the presence of allele-specific isoforms in the data.

The numbers of matrices (with 95% confidence individual confidence intervals), with a zero cell from 1000 simulated replicates are shown in Figure 4A. The horizontal line in Figure 4A shows the number of matrices with a zero cell in the observed data. If we consider qualitative allele-specific splicing as the only alternative to random association (ie quantitative allele-specificity not allowed), the proportion of allele-specific isoforms can be inferred from the intersection of the horizontal line with the simulated data points. This intersection, shows additional matrices with a zero cell in the observed data which correspond to the average number of matrices without a zero cell in the randomized datasets. Hence, we can deduce from Figure 4A that the proportion of allele-specific splicing most consistent with 14 additional matrices with a zero cell in the observed data is 21%. The lower (6%), and upper bound (36%), of our estimate were easily inferred from the span of the confidence interval inferred as illustrated in Figure 4A.

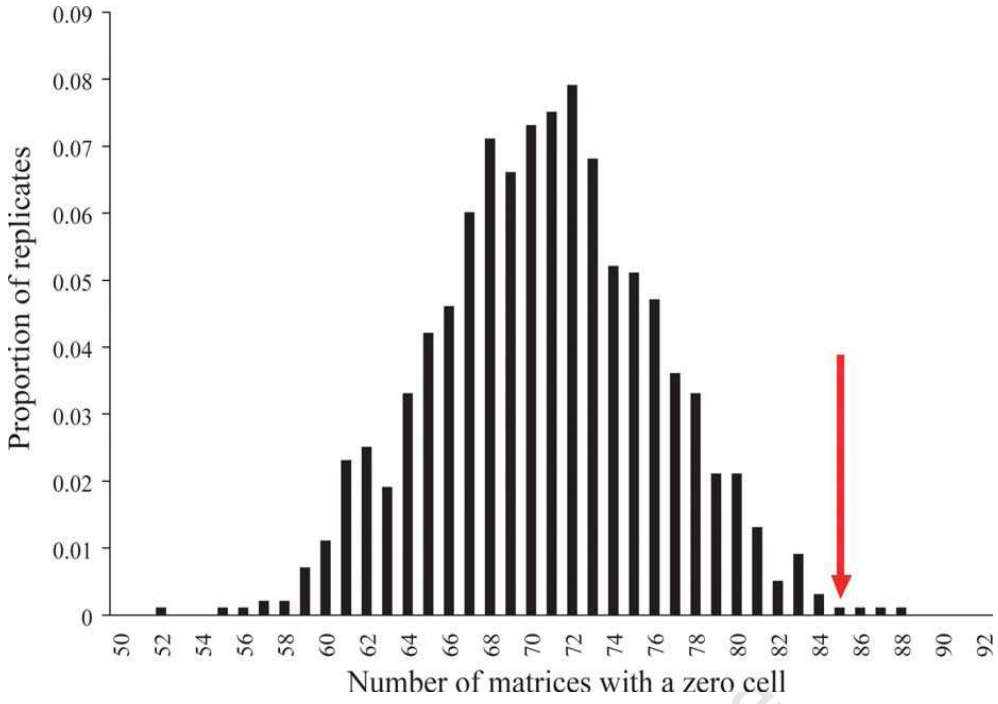


Figure 3: A histogram showing the proportion of matrices with a zero cell from 1000 randomized replicates of the dataset.The arrow indicates the number of matrices with a zero cell in the observed data

Quantitative allele-specificity, can also increase the number of matrices with a zero cell in the data but it does so less efficiently than qualitative allele-specific splicing. We modelled the situation where a proportion of the matrices in the dataset showed a strong, but not exclusive, association between transcript isoform and nucleotide polymorphism and used simulation to estimate the expected number of matrices with a zero cell. The association between rows and columns for the affected matrices was such that, for an individual EST, the probability of it being isoform k given that it is SNP allele i was four times the probability of being isoform k given that it is SNP allele j. With this fourfold association, at least 14% of the isoforms in the dataset would need to be quantitative allele-specific in order to explain the observed number of matrices with a zero cell. The real data are likely to contain a mixture of qualitative and quantitative allele specific isoforms. Combinations of proportions of qualitative and quantitative allele-specific forms that would most likely result in the number of zeros in the observed data are shown in Figure 4B.

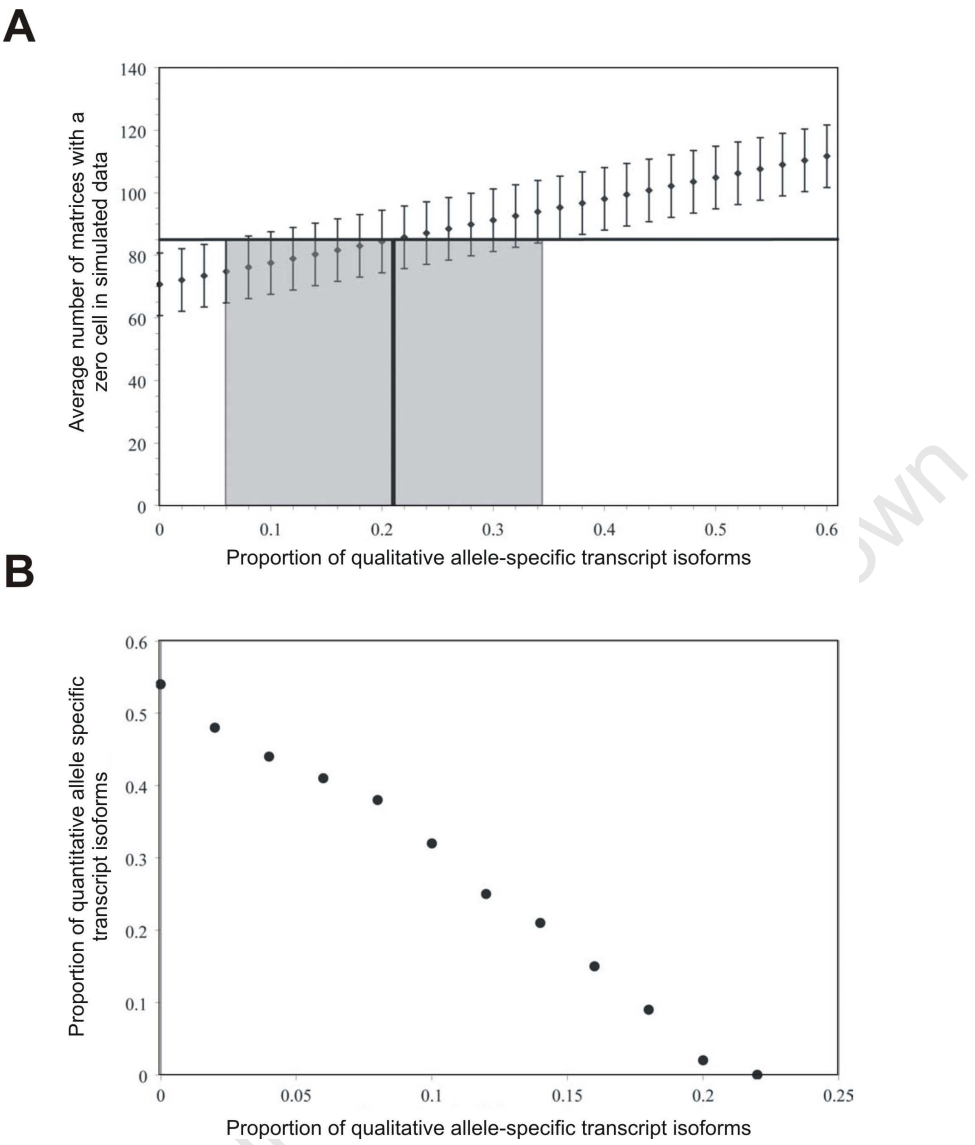
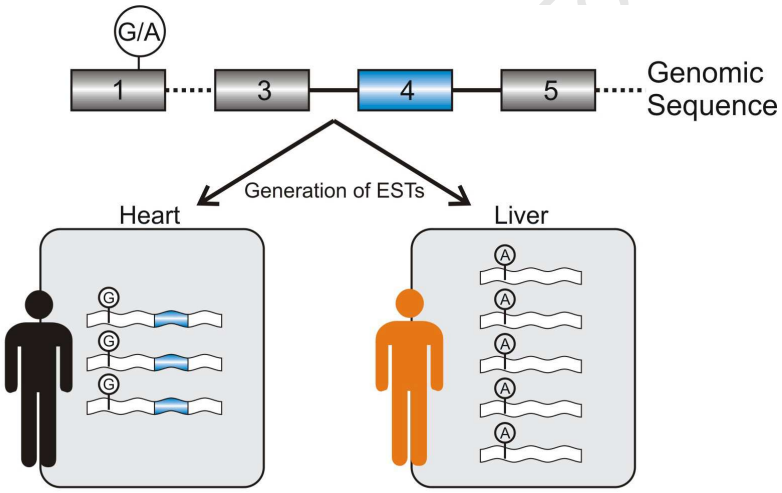


Figure 4: Simulations of qualitative and quantitative allele-specific transcript isoforms. A) Average numbers of matrices with a zero cell from 1000 simulated replicates of the dataset in which a proportion of the simulated matrices are derived from qualitative allele-specific transcript isoforms. The horizontal line shows the number of matrices with a zero cell in the observed data. The shaded area shows the confidence interval and the dark line the proportion of qualitative allele-specific isoforms most consistent with the observed number of matrices with a zero cell. Error bars were derived from the simulation as described in the method section. B) Combinations of qualitative and quantitative allele-specific transcript isoforms that could explain the observed number of matrices with a zero cell in the data. The points on the graph show the proportions of qualitative (y-axis) and quantitative (x-axis) allele-specific isoforms in the simulations that result in approximately the same number of matrices with a zero as we have observed in the data. In this example the strength of the association between rows and columns of the simulated matrices was such that, for an individual EST, the probability of it being in column k given that it was in row i was a factor of four higher than the probability of it being in column k given that it was in row j.

A significant weakness in our estimation of the proportion of alternatively spliced genes that are allele-specific results from restricting to a single transcript sequence per cDNA library. This greatly reduced our dataset and consequently the accuracy of our estimation. This restriction is imposed to prevent inflating our estimate of allele-specific splicing in human as a result of tissue-specific isoforms. For simplicity we have illustrated in Figure 4 how tissue-specific splicing could be falsely detected as allele-specific splicing. Such associations can be highly significant if there are many ESTs of the gene in the two cDNA libraries in which it occurs. We therefore sampled a maximum of two ESTs per cDNA library (one for each allele of the SNP from heterozygous libraries and just one from homozygous libraries).



Allele Specific Splicing

SNP	SPLICE VARIANT: Exon 4	SPLICE VARIANT: Exon 4 skipped
G	3	0
A	0	5

Tissue Specific Splicing

Tissue	SPLICE VARIANT: Exon 4	SPLICE VARIANT: Exon 4 skipped
Heart	3	0
Liver	0	5

Figure 5: Two different cDNA libraries created from different tissues and two homozygous individuals. An exon skipping event that occurs in a tissue-specific manner and which is not in any way allele-specific is used for illustration. If all transcripts from these two cDNA libraries are sampled, the resulting data matrices (bottom), indicate an association between isoform and tissue and the same association is also falsely detected for isoform and allele.

4.3.3 Detecting individual examples of allele-specific splicing

Given a 2X2 matrix that represents a set of transcripts that can be mapped to two alternative transcript isoforms and alternate SNP allele we could test for non-random associations between isoforms and SNPs using Fisher's Exact Test. We restricted the tests to only those matrices that could achieve a maximal p-value of 0.001 based on their row and column totals, to reduce the correction for multiple testing. An example of a data matrix is shown in Figure 2 for the gene *RNPC4/RBM23*. The P-value for the association between the rows and columns of the matrix shown is 0.0002 (from a 2-tailed Fishers Exact Test). The highest statistical significance ($p < 2 X10^{-8}$) of any matrix in the dataset was achieved for the association between this same isoform and another SNP (rs2295680) that was located closer to the alternatively spliced exon junction. Further lines of evidence seem to suggest that this gene does indeed have polymorphic splice variants that are found at sufficiently high frequency in the population to be detectable using different methods (Hull et al., 2007). Firstly the association remains highly significant after correction for multiple testing, and secondly we have detected the same exon 6 skipping using another independent method. We detected the same allele-specific splicing event using an integrated analysis of ESTs, microarrays and genomic data (Chapter 5). For most matrices there are insufficient data to establish allele-specificity with any confidence, especially when correction for multiple testing is applied.

4.4 Discussion

We report the first comprehensive genome-wide estimation of the prevalence of allele-specific splicing in human using an integrated analysis of ESTs, SNPs and alternatively spliced gene transcripts. We propose a lower bound of 6% for the proportion of alternatively spliced genes for which either qualitative allele-specific or quantitative allele-specific transcript isoforms are present in the dataset. Our estimate of the prevalence of allele-specific isoforms took account of just one alternative isoform per gene, while there are multiple potentially independent alternatively spliced junctions for many of the genes represented in the ASAP database. Therefore the total proportion of

genes from the alternative splicing databases for which at least one allele-specific isoform exists may be far higher than the lower bound presented here. Furthermore, this analysis is based on association between transcript isoforms from qualitative and quantitative allele-specific splicing and exonic SNPs. However, only qualitative allele-specific splicing will cause the strongest association. As a consequence, our estimate of the proportion of qualitative allele-specific transcript isoforms required to explain the number of observed matrices with a zero cell is likely to also contribute to our underestimation of the prevalence of allele-specific splicing.

There are several caveats associated with EST data which are also likely to contribute to an underestimation of the prevalence of allele-specific splicing. ESTs are biased towards gene ends, approximately 80% of EST are sampled from the 3' and 5' untranslated gene regions. ESTs are also sparse for all but the most highly expressed genes. Some genes of great medical importance such as the *BRCA1* gene that is known to be spliced in an allele-specific manner (Fackenthal et al., 2002) are not represented even by a single EST. This under-representation of genes in the EST data has also been highlighted by an investigation that quantified the number of chromosome 22 genes represented in the EST database. Only half of the genes on chromosome 22 are represented in EST databases (de Souza et al., 2000). Furthermore, even for genes that are represented by more than 100 ESTs in dbEST such as the *INII* gene, both novel mRNA transcript isoforms and allele-specific isoforms are still being discovered (Favre et al., 2003). These observations highlight the insufficiency of ESTs in capturing all genes and mRNA isoforms.

The density of genetic variation in functionally relevant regulatory elements is expected and has been reported to be low, as a result of purifying selection (Fairbrother et al., 2004). Despite this purifying selection, genetic polymorphisms are still amongst a multitude of factors that play a role in affecting splicing of mRNA in an individual specific manner (Hiller et al., 2006; Fairbrother et al., 2004). We propose that the allele-specific splicing reported is likely to result from *cis*-acting allelic variations that affect the strength of splice signals or their regulatory elements. However, it is possible, that other factors contribute to the observed allele-specific isoforms. An apparent alternatively

spliced isoform could result directly from a polymorphic deletion spanning an entire exon. An example of such a variant was reported for the human growth hormone receptor (Pantel et al., 2000). In some instances, allele-specificity could also result from a polymorphism that determines whether nonsense mediated decay of one isoform occurs (Figure 1C). The SNP allele that leads to NMD could either be a nonsense mutation or a frame-shift mutation that results in an exonic premature termination codon. This could result in an under-representation of the transcript isoform derived from the allele containing the stop codon. In this case, both of the transcript isoforms are produced from each allele (true alternative splicing) but the likelihood of observing both isoforms would be reduced for the allele with the stop codon.

Numerous studies have used public transcript data to investigate tissue-specific splicing (Xu et al., 2002; Minovitsky et al., 2005; Aouacheria et al., 2006). However, use of ESTs to estimate either tissue-specificity or allele-specificity is complicated by biases that exist during the generation of ESTs. In some instances it is impossible to conclude from EST data alone whether significant associations between mRNA isoforms and SNP genotypes are due to true allele-specific splicing or could be attributed to tissue-specific splicing hence tissue-specific splicing can be falsely detected as allele-specific splicing (Figure 5) and *vice versa*. We restricted to one EST per isoform, per allele in any cDNA library to prevent inflating our allele-specific estimate as a result of tissue-specific isoforms. For example, the case illustrated in Figure 5 would have attained a p-value = 0.007937. After restricting to one EST per library the p-value is not significant. This EST sampling restriction was done at the expense of greatly reducing our power to detect genes that showed evidence of allele-specific splicing. Thus we have reanalysed the problem of predicting allele-specific isoform from EST data by developing a more robust maximum likelihood approach. This maximum likelihood approach explicitly models both regulated and allele-specific splicing with extra power from integrating information from all ESTs within a given cDNA library and across libraries (Chapter 5).

Recombination events can reduce the signal in the data and could cause underestimation of the prevalence of allele-specific splicing but cannot introduce false positive results.

However, given the short distances between associated exonic SNPs and alternatively spliced isoforms in most cases, intervening recombination events are likely to be rare. It is possible that transcript isoform choice is also affected by an individual's genotype at other unlinked loci, but our method is unable to detect this. Similarly, although EST sequences are known to be error-prone, random sequencing errors cannot introduce false positive associations between splice isoform and allele, though they are also a potential source of noise.

The existence of polymorphic transcript isoforms within the human population is consistent with the recent discovery that exons found only in minor splice forms are often completely absent from the orthologous gene in human/rodent comparisons (Modrek et al., 2001). This is consistent with a high prevalence of allele-specificity among minor splice isoforms, because evolutionary changes in gene structure must originate as polymorphisms within species. An allele-specific isoform may become fixed in the population if it adds advantageous splice variants or, at least, does not severely compromise the function of the wild-type gene. It is also likely that most such variants persist because their allele-specific splicing events do not directly cause disease but only modulate severity of certain disease phenotypes caused by separate mutation (Buchner et al., 2003).

We found a high prevalence of allele-specific splicing. Interestingly, no natural upper bound on the proportion of alternatively spliced genes that are affected by polymorphism emerges from our analysis and we argue that the contribution of allele-specific transcript isoforms should be stated as a caveat in future estimates of the prevalence of alternative splicing. Our results emphasize that true alternative splicing cannot be confirmed unless more than one transcript is observed from the same allele and also caution that any inference of tissue-specific splicing must take account of allele-specificity.

4.5 Acknowledgements

This chapter was perfomed in collaboration with Janet Kelso and Cathal Seoighe. I was responsible for the collection of data and most of the analysis. CS formulated the statistical methods and JK assisted with the analysis.

Chapter 5

Identification of Allele-specific mRNA transcripts in Human Through an Integrative Analysis of Genomic, EST Data and Microarrays

Abstract

Accurate mRNA splicing depends on multiple regulatory signals encoded in the transcribed RNA sequence. Many examples of mutations within human splice regulatory regions that alter splicing qualitatively or quantitatively have been reported and allelic differences in mRNA splicing are likely to be a common and important source of phenotypic diversity at the molecular level, in addition to their contribution to genetic disease. However, because the effect of a mutation on the efficiency of mRNA splicing is often difficult to predict, many mutations that cause disease through an effect on splicing are likely to remain undiscovered. We have combined a genome-wide scan for sequence polymorphisms likely to affect mRNA splicing with evidence from publicly available Expressed Sequence Tag (EST) and exon array data. The genome-wide scan uses published tools and identified 30,977 Single Nucleotide Polymorphisms (SNPs) located within donor and acceptor splice sites, branch points and exon enhancer elements. For 1,085 candidate splicing polymorphisms the difference in splicing between alternative alleles was corroborated by publicly available exon array data from 166 lymphoblastoid cell lines. We developed a novel probabilistic method to infer allele-specific splicing from EST data. The method makes use of SNPs and alternative mRNA isoforms mapped to EST sequences and models both regulated alternative splicing as well as allele-specific splicing. We report a set of genes showing evidence of allele-specific splicing from an integrated analysis of genomic polymorphisms, EST data and exon array data including several examples for which there is experimental evidence of polymorphisms affecting splicing in the literature. We also present a set of novel allele-specific splicing candidates and discuss the strengths and weaknesses of alternative technologies for inferring the effect of sequence variants on mRNA splicing**.** Our results provide an extensive resource

that can be used to assess the possible effect on splicing of human polymorphisms in putative splice-regulatory sites.

5.1 Introduction

One of the key tasks of the post-genome era is to determine the functional implications of genomic variants. The development of high throughput genotyping technologies and the use of these technologies in large-scale studies has enabled the identification of increasing numbers of human loci that are associated with common genetic disorders (e.g.(The Wellcome Trust Case Control Consortium, 2007)); however, the mechanisms through which genetic variants at many disease-associated loci affect disease susceptibility remain to be determined. Mutations or polymorphisms that affect mRNA splicing can have a profound effect on the function of the spliced product, but these effects are often difficult to predict from the primary genomic sequence. The medical and biological significance of such variants is evident from the large and rapidly increasing volume of literature reporting examples of aberrant mRNA splicing associated with human cancers and genetic diseases (Faustino and Cooper, 2003; Wang and Cooper, 2007). Indeed, point mutations leading to aberrant splicing are thought to be among the most important contributors to human genetic diseases (Lopez-Bigas et al., 2005).

Sequence variants found on the pre-mRNA can affect a number of different, and in some cases imperfectly characterized, *cis*-acting sequences that control splicing. Polymorphisms that occur at the highly conserved donor and acceptor di-nucleotides are an obvious case in which we expect an effect on splicing (Krawczak et al., 1992) and these genomic variants, when they occur close to verified exon boundaries, tend be annotated in databases of sequence polymorphisms, such as dbSNP (Sherry et al., 2001). A much larger proportion of variants are likely to occur at sites where the effect on splicing is less obvious, for example at less conserved sites close to intron/exon boundaries, close to the intronic branch-point (Kralovicova et al., 2006b), or within intronic or exonic splicing enhancer or suppressor sequences (Cartegni et al., 2002). In some cases, such sequence variants will disrupt the normal splicing of the gene and cause

aberrant splicing of a proportion or of all of the transcripts produced. However, if the gene is alternatively spliced to begin with, then sequence variants that affect sites that are involved in controlling isoform abundance may be affected, causing allelic differences in the regulation of alternative splicing, with potentially important biological consequences (Buchner et al., 2003).

The contribution of heritable variation to the observed diversity of mRNA splice isoforms is well established (Nembaware et al., 2004; Hull et al., 2007; Kwan et al., 2007). Using the ASAP database of alternatively spliced mRNA isoforms (Modrek et al., 2001) and transcribed SNPs, we previously estimated that approximately 20% of alternatively spliced genes show evidence of allele-specific splicing either complete allele-specific splicing, in which one allele gives rise to one isoform and another results in the alternative form, or partial allele-specific splicing in which different alleles result in distinct relative isoform abundance (Nembaware et al., 2004) (see Chapter 2). Earlier large-scale studies of alternative and allele-specific splicing relied primarily on Expressed Sequence Tag (EST) sequences. More recently, both exon-junction and exon tiling arrays have been used for genome-wide studies of alternative splicing (Johnson et al., 2003; Pan et al., 2004; Pan et al., 2004). The Affymetrix GeneChip Human Exon 1.0 ST Array has probe-sets targeting approximately 1.4 million known and predicted exons. Alternatively spliced mRNA isoforms detected using the Affymetrix exon array in cell lines genotyped as part of the HapMap project (The International HapMap Consortium, 2005), has given rise to opportunities for high-throughput discovery of alleles that affect mRNA splicing (Kwan et al., 2007; Hull et al., 2007). Though exon arrays are arguably a superior technology, with better exon coverage than ESTs (Kwan et al., 2007), they are also affected by a range of caveats (Lee and Wang, 2005). Integration of results from ESTs and microarrays is likely to increase power to detect allele-specific splicing as both arrays and ESTs have different limitations and advantages for the analysis of alternatively spliced isoforms.

Though for the present it remains a distant goal, a complete description of the effect of human sequence variants on mRNA splicing would be a powerful resource for

understanding human genetic diseases and phenotypes. One option for evaluating the potential effect of *cis*-acting mutations on splicing is to use *ab initio* prediction algorithms that make use of the availability of the complete genome sequence (Yeo and Burge, 2004; Cartegni et al., 2003). In several previous studies, computational tools have been effective in helping to shed light on the impact of a mutation on splicing (Buchner et al., 2003; Kralovicova et al., 2006a; Kralovicova et al., 2006b) and databases of mutations that may affect splicing have been made available (Conde et al., 2006; Li et al., 2007). However, because of the difficulty of predicting all splice regulatory elements from genomic sequence and the even greater difficulty of determining accurately the effect of mutations in these regions on splicing, genomic analysis of SNPs likely to affect splicing needs to be complemented by expression data that provides information about the splice isoforms that are associated with the alternative alleles of a candidate SNP.

We have performed a genome-wide scan for Single Nucleotide Polymorphisms (SNPs) likely to influence splicing efficiency in *cis* using publicly available tools (ESEfinder, (Cartegni et al., 2003), MaxEntScan (Yeo and Burge, 2004), and Branch Site Analyzer (Kol et al., 2005). We have tested predictions based on genomic sequences using publicly available EST and exon array data. We present a novel probabilistic method to infer allelic differences in mRNA splicing from EST data and used recently published Affymetrix exon array hybridisation data derived from 166 lymphoblastoid cell lines (Huang et al., 2007) for which genome-wide genotype data are available through the HapMap project (The International HapMap Consortium, 2005) to test for association between mRNA isoforms and the genotype of putative *cis*-acting splicing polymorphisms.

5.2 Methods

5.2.1 Splice site strength prediction

We downloaded known transcripts, chromosomal genomic data and SNP and exon tables from Ensembl version 36 (Birney et al., 2006), which is based on NCBI Genome build 35. Genes and SNPs that mapped to multiple locations on the genome were discarded. Introns were inferred from the exon genomic coordinates obtained from Ensembl. SNP positions relative to the Ensembl exons and introns were identified via genomic coordinates. SNP positions relative to exon/intron junctions were also determined for isoforms obtained from the ASAPII database.

Published tools for detecting splicing regulatory elements were either requested from authors or downloaded from their respective sites. We extracted 9 nucleotides from the donor splice sites and 23 nucleotides from the acceptor splice sites as required by the maximum entropy algorithm of Yeo and Burge, (2004). Scores for each pair of alternate alleles were then calculated (Yeo and Burge, 2004). We also identified an inflated frequency of SNPs at the G base of the canonical AG acceptor site which has been previously identified as a sequencing artifact (Platzer et al., 2007). We therefore restricted our analysis to validated SNPs using the information from dbSNP125 in the Ensembl database.

The ESEfinder tool (Cartegni et al., 2003) is designed to predict four ESEs: SC35, ASF2, SRp55 and SRp40. ESEfinder uses a position specific weight matrices. An ESE is considered to have a pre-defined length, *m*, and a recommended minimum score *S*. For each SNP we extracted m-1 nucleotides up- and downstream of the SNP. We then calculated the ESE scores for each of the contiguous length *m* subsequences of this sequence. The highest score for each SNP allele was retained if at least one of the scores was above *S* and the other below *S.* Although some strong ESEs can influence splicing at a distance of several kilobases (Graveley et al., 1998), functional ESEs are most abundant in close proximity to splice junctions of internal exons (Fairbrother et al., 2004). We therefore restricted our analysis to ESEs located within 200 bps of exon-intron junctions

of internal exons. Branch point scores for pairs of alternate SNP alleles were computed using Perl scripts provided by Kol et al., (2005).

5.2.2 Mapping exonic SNP alleles to splice variants

We downloaded pre-computed EST and SNP genomic locations from the UCSC Genome Browser (Karolchik et al., 2003), which is based on NCBI genome assembly 36. ESTs and SNPs that mapped multiple times onto the genomic sequence and ESTs for which less than 90% of the sequence mapped to the genome were discarded. We used SNP and EST genomic coordinates to identify the SNP allele corresponding to each EST overlapping the SNP position. ASAPII (Kim et al., 2007), a database of alternatively spliced gene clusters, was downloaded on 9/11/2006. This data included gene and exon genomic locations based on NCBI genome assembly 35 as well as alternative mRNA isoforms (represented by conflicting exon junction pairs) mapped to ESTs.

5.2.3 Models of regulated and allele-specific splicing

For a given allele, A, of an alternatively spliced gene with alternative splice isoforms S_1 and S_2, let x represent the proportion of isoform S_1 produced from allele A in a cDNA library. We assume that x is constant for a given allele and library, but may vary across alleles and/or libraries. The purpose of the model is to determine, using data from several libraries (in which alternative transcript isoforms may be differentially regulated and have different relative expression levels), whether x shows significant variation across alleles.

Consider cDNA library i with N transcripts from allele A, of which we observe a_i ESTs that map to S_1 and $b_i = N\text{-}a_i$ ESTs that map to S_2. Because a_i is binomially distributed with binomial parameter x, we use the beta distribution (conjugate to the binomial) to describe the probability density of x. We share this distribution across all libraries but not necessarily across the two alleles. Thus the values of x for separate libraries are modeled as independent draws from the distribution $f(x, \alpha_A, \beta_A)$ for allele A and $f(x, \alpha_B, \beta_B)$ for allele B, where $f(x, \alpha, \beta)$ is the beta function with parameters α and β.

The likelihood of the data from allele A observed in library *i* can now be expressed as

$$L(D_i \mid \alpha_A, \beta_A) = \int_0^1 x^{a_i} (1-x)^{b_i} \, f(x, \alpha_A, \beta_A) \, dx \qquad (1)$$

The likelihood of the data observed in all cDNA libraries is a product over terms such as this, and the α and β parameters can be estimated by optimizing the likelihood for the combined data set.

An analytical solution to the integral of equation 1 exists, resulting in the following expression for the likelihood of the complete data for a pair of alternate isoforms and SNP alleles:

$$L(D \mid \alpha_A, \beta_A, \alpha_B, \beta_B) = \prod_i \frac{\Gamma(\alpha_A + \beta_A)\Gamma(a_i + \alpha_A)\Gamma(b_i + \beta_A)\Gamma(\alpha_B + \beta_B)\Gamma(c_i + \alpha_B)\Gamma(d_i + \beta_B)}{\Gamma(\alpha_A)\Gamma(\beta_A)\Gamma(a_i + b_i + \alpha_A + \beta_A)\Gamma(\alpha_B)\Gamma(\beta_B)\Gamma(c_i + d_i + \alpha_B + \beta_B)} \quad (2)$$

where a_i, b_i are the numbers of ESTs in cDNA library *i* that map to allele A and splice junctions S_1 and S_2 respectively and c_i, d_i are the corresponding EST counts for allele B. The maximum likelihood parameter estimates were obtained by optimizing the likelihood using Powell's method (Press et al., 1992).

For the null model, we impose the restriction that $\alpha_A = \alpha_B$ and $\beta_A = \beta_B$, such that both alleles are considered to be sampled from the same distribution (no allele-specific effect). To model allele-specific splicing (the alternative model), we allow $\alpha_A \neq \alpha_B$ and estimate separate beta distributions for the alternate alleles of a SNP (we keep the constraint $\beta_A = \beta_B$ because we found that this model already has sufficient freedom to model the desired effect and adding another degree of freedom was unnecessary). If the null model can be rejected in favour of the alternative model (using the likelihood ratio test) we conclude that there is evidence of allele-specific splicing.

5.2.4 Simulations

We constructed 1000 random replicates of the EST data such that, for every SNP, the number of libraries derived from each genotype of the SNP was identical to the real data. Each library in the simulated data was assigned a genotype, with a probability proportional to the number of libraries of that genotype in the real data (this proportion

was adjusted as each simulated library was assigned a genotype). For each library, the total numbers of ESTs derived from each isoform was constrained to be the same as in the real data. For heterozygous libraries ESTs were assigned to alternative SNP alleles with equal probability.

5.2.5 Analysis of Affymetrix exon arrays

We obtained whole genome exon data from the Gene Expression Omnibus (Barrett et al., 2007), which were generated using the Affymetrix Human Exon 1.0ST array by Huang et al. (Huang et al., 2007). These data were generated from 166 lymphoblastoid cell lines for which genome-wide genotype data are available through the HapMap project (The International HapMap Consortium, 2005). SNPs that overlap with probes can affect binding affinities and potentially result in spurious identification of differential expression (Kwan et al., 2007). We therefore removed all probes that overlapped with SNPs from dbSNP, from further analysis (Kwan et al., 2007).

The exon array data was processed using the Affymetrix Power Tools. For all probesets, we used the Plier Sketch algorithm to estimate expression level in each cell-line, and DABG was used to estimate detection above background probabilities (Affymetrix, 2007). For the meta-probeset (transcript) level expression we used only the high confidence (or 'core') probesets from the array to avoid inaccuracy caused by the inclusion of computationally predicted probesets (Kwan et al., 2007). For each probeset that mapped to a meta-probeset, the splicing index (SI) was calculated by dividing the probeset expression estimate by the estimate of the transcript-level expression in each cell line. The core meta-probeset expression estimate was used for non-core probesets that mapped to core as well as non-core meta-probesets.

We used a robust linear model to test for an association for each srSNP, between the SI of all probes within 1kb of the probe and SNP genotype, treating HapMap population as a covariate. We used Holm correction (with significance level 0.05) to control the family-wise error rate and to establish a high-confidence or conservative set of probes with allele-specific SI. A false detection rate correction (also with cut-off set to 0.05) was also used to generate a larger set of events that includes a small proportion of false positive

inferences. All statistical analyses were performed using the R statistical computing environment (The R Project for Statistical Computing, ; Ihaka and Gentleman, 1996).

Table 1: Summary of srSNPs with supporting evidence from EST and Exon array data.

***Cis* element**		**srSNPs**	**EST evidence (α = 0.05)**	**Exon array (FDR=0.1)**	**Exon array (Holm corrected α = 0.05)**
Donor		1970	47	84	20
Acceptor		7248	156	217	22
Branch		2689	41	75	13
ESEs	SC35	5910	44	257	26
	SF2	8992	82	387	44
	SRp40	7776	94	334	32
	SRp55	5231	64	211	19

5.3 Results

5.3.1 A genome-wide scan for polymorphisms in splice-regulatory regions

We used published computational tools to identify 30,977 polymorphisms that occur within predicted or known splicing regulatory sequences (which we refer to as srSNPs), including donor sites, acceptor sites, branch points (BP) and exonic splice enhancer (ESE) elements. The number of SNPs occurring in putative ESEs is much higher than the number in the other *cis* elements (Table 1). This is likely to be due, at least in part, to the high false positive rate of ESE identification compared to the other splice regulatory elements that are identified using positional information, rather than by matching to sequence patterns alone. For each type of splice-regulatory element, publicly available tools were used to score the sequences associated with alternative SNP alleles (Methods). We used gene structure information from Ensembl (Birney et al., 2006) as well as from ASAPII (Kim et al., 2007), to identify srSNPs. This greatly increased our coverage, for example, of the 9,201 polymorphisms identified in donor and acceptor regions (including

17 in dual-specificity sites (Zhang et al., 2007)), 3,868 occurred within exon-intron or intron-exon boundaries common to both databases while 2,759 were unique to Ensembl and 2,574 were unique to ASAPII.

5.3.2 A maximum likelihood method to identify allele-specific splicing using EST data

We previously used linkage disequilibrium between SNPs mapped to EST sequences and alternative splice isoforms to identify allele-specific mRNA isoforms (Nembaware et al., 2004). However, because alternative splicing can be regulated in a tissue-specific way and because multiple ESTs from the same gene can occur in a single cDNA library, we restricted our previous analysis to just one EST per cDNA library per alternative isoform pair. To make better use of the available data we have now developed a probabilistic model that can be applied to detect allele-specific splicing from SNPs mapped to EST sequences (Methods and an illustration of the data for an example gene in Figure 1).

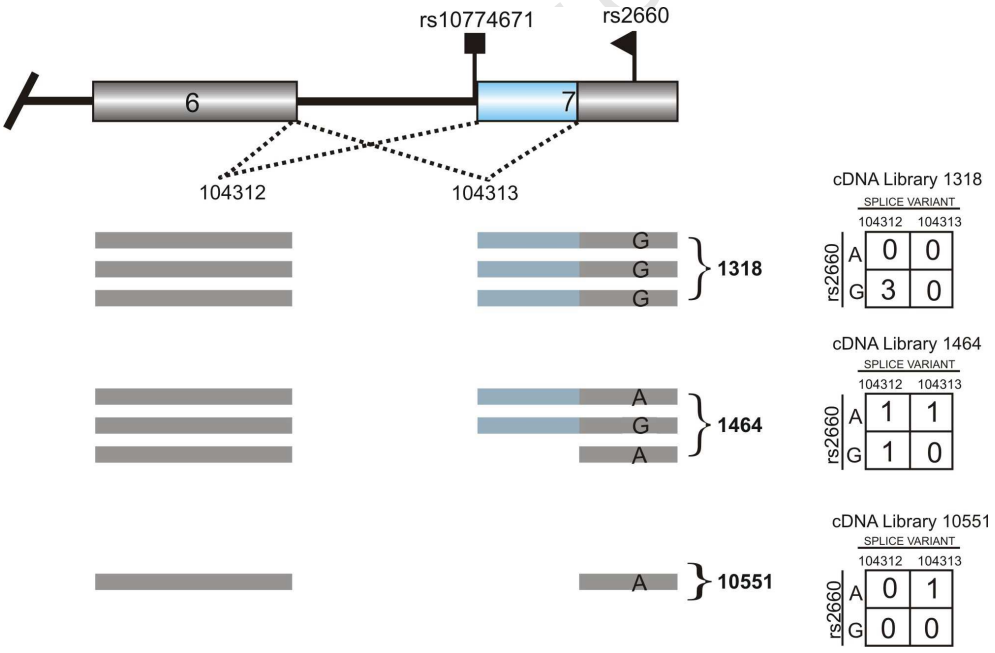


Figure 1: Part of the genomic sequence of the *OAS1* gene showing alternative acceptor site use at exon 7. The putative causative SNP (rs10774671), which occurs at the G site of the canonical acceptor dinucleotide, and an mSNP (rs2660), which was used to infer allele-specific splicing from EST data, are shown. Splice isoforms and mSNP alleles observed in three of a total of 27 cDNA libraries with ESTs that mapped to this region are also depicted. For each library the data are summarized in a two-by-two contingency table, with each EST cross-classified according to mRNA isoform and SNP allele.

The possibility that the isoform is regulated in a tissue-specific way is modeled explicitly. For a given pair of mutually exclusive mRNA isoforms, the proportions of each isoform that occur across different cDNA libraries are modeled using a beta distribution. An allelic effect on splicing is inferred when a model that allows separate beta distributions for two alternative alleles of a SNP (which maps to both isoforms) provides a better fit to the data than a model with a single distribution for both alleles. We found 1,753 marker SNPs (i.e. SNPs in linkage disequilibrium with a splicing event, which we refer to as mSNPs), corresponding to 1,318 genes and 2,283 alternative splice junction pairs, for which the allele-specific mRNA splicing model provided a better fit to the data than the null model at the 5% significance level, using the likelihood ratio test.

The distribution of the likelihood ratio test statistic is asymptotically chi-squared for large sample sizes under the null hypothesis. To test the validity of the test on the observed data, for which the number of data points per test was highly variable, we simulated data identical to the observed data in terms of the numbers of ESTs mapping to alternative alleles and splice isoforms but conforming to the null hypothesis of no association between mRNA isoform and allele. The cumulative distribution of the likelihood ratio statistic on this simulated data was consistently lower than the chi-squared distribution with one degree of freedom (data not shown), which suggests that the likelihood ratio test provides a conservative basis on which to reject the null hypothesis. The distribution of p-values from the simulated data was also not uniform because of the sparseness of the data available for many of the mSNP and splice junctions that were tested. This complicates the application of standard false discovery rate methods to account for multiple testing. Instead we compared the observed and simulated distributions of the likelihood ratio test statistic, which allowed us to estimate the proportion of false discoveries at all levels of the test statistic (Figure 2). There were 91 cases of association between mSNPs and splice isoforms at the true positive rate cut-off of 0.8, shown on the graph (corresponding to approximately 73 true positives and 18 false positives). These came from 54 distinct alternate splice junction pairs and 51 different genes.

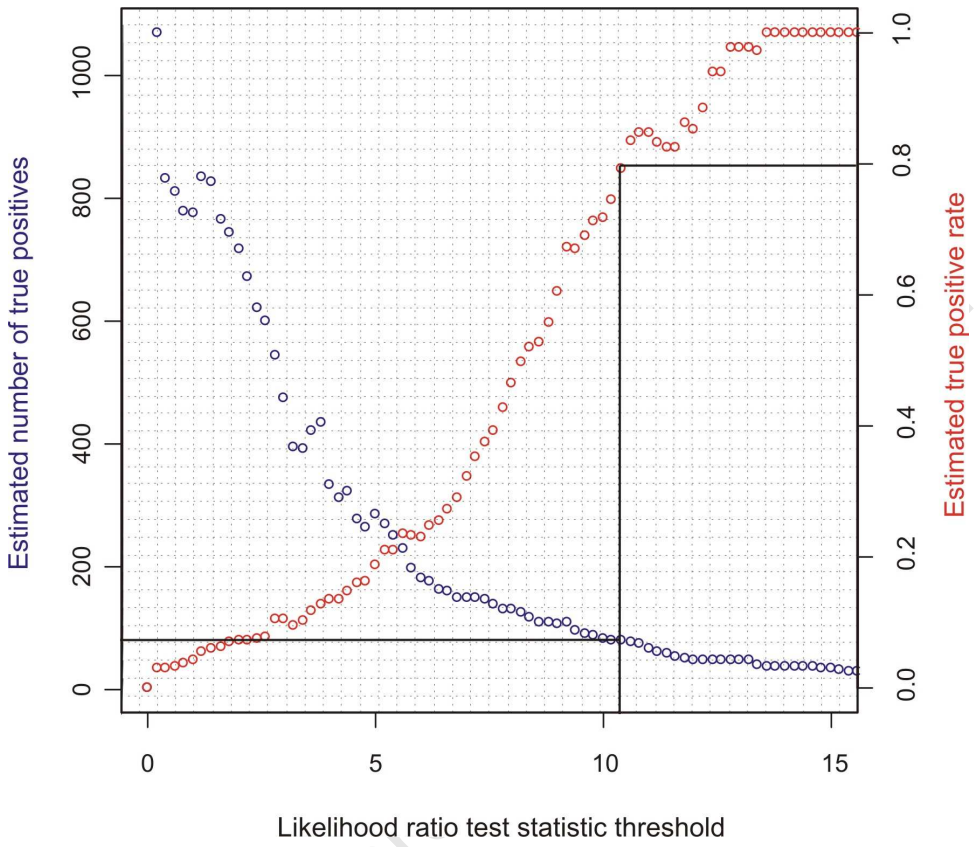


Figure 2: Analysis of simulated EST data. The number of true positives and the true positive rate (equal to one minus the false discovery rate) as a function of the likelihood ratio test statistic were estimated from 1000 randomizations of the matrices of counts of ESTs mapping to alternative SNP alleles and alternative splice isoforms. The solid line shows the number of true positives obtained when the true positive rate is 0.8 (i.e. at a false discovery rate of 0.2).

5.3.3 Support for srSNPs and mSNPs from publicly-available exon array data

Exon array data generated by Huang *et al.* (Huang et al., 2007) from 166 lymphoblastoid cell lines using the Affymetrix Exon 1.0ST were downloaded from the GEO database (Barrett et al., 2007), and processed as described in Methods. The splicing index (SI; (Clark et al., 2002)) was calculated for each probeset, by dividing the probeset-level expression estimate by the meta-probeset (or transcript) level expression estimate.

Probesets analysed included high-confidence core probesets as well as probesets corresponding to predicted exons. The transcript-level expression estimates were inferred using core probesets only, to avoid inaccuracy caused by including spurious probesets in the transcript-level expression estimate (Kwan et al., 2007). Genome-wide genotype data for almost four million SNPs were available for the same cell lines through the HapMap project (The International HapMap Consortium, 2005). For each putative srSNP and mSNP for which genotype data were available we tested for an effect of genotype on SI for each probeset in the region of the mSNP or srSNP, treating the HapMap population from which the sample was derived (Yoruban or Caucasian) as a covariate, and using a robust linear model and robust analysis of variance (ANOVA), implemented in the Insightful Robust Library of the R package (The R Project for Statistical Computing, Ihaka and Gentleman, 1996). In the case of srSNPs, because it is often difficult to predict the impact of the SNP on splicing, all probesets within 1kb of the SNP were tested. For the mSNPs we tested only probesets that fell within the genomic boundaries defined by the alternative exon junctions of the putatively allele-specific splice isoforms. Similarly, to determine whether an srSNP was supported by EST data, we tested whether the srSNP fell within the genomic region defined by the alternative exon junctions (including 3bp of the corresponding exons in the case of putative exonic splice donor and acceptor mutations). Examples of srSNPs for which there was strong evidence of an allelic effect on splicing from the exon array data (Holm-corrected p-value < 0.05) are shown in Figures 3 and 4. Similar diagrams are available for a total of 1,083 putative srSNPs for which there was a probeset SI significantly associated with genotype are available from http://mancala.cbio.uct.ac.za/splicing/ExonArray.

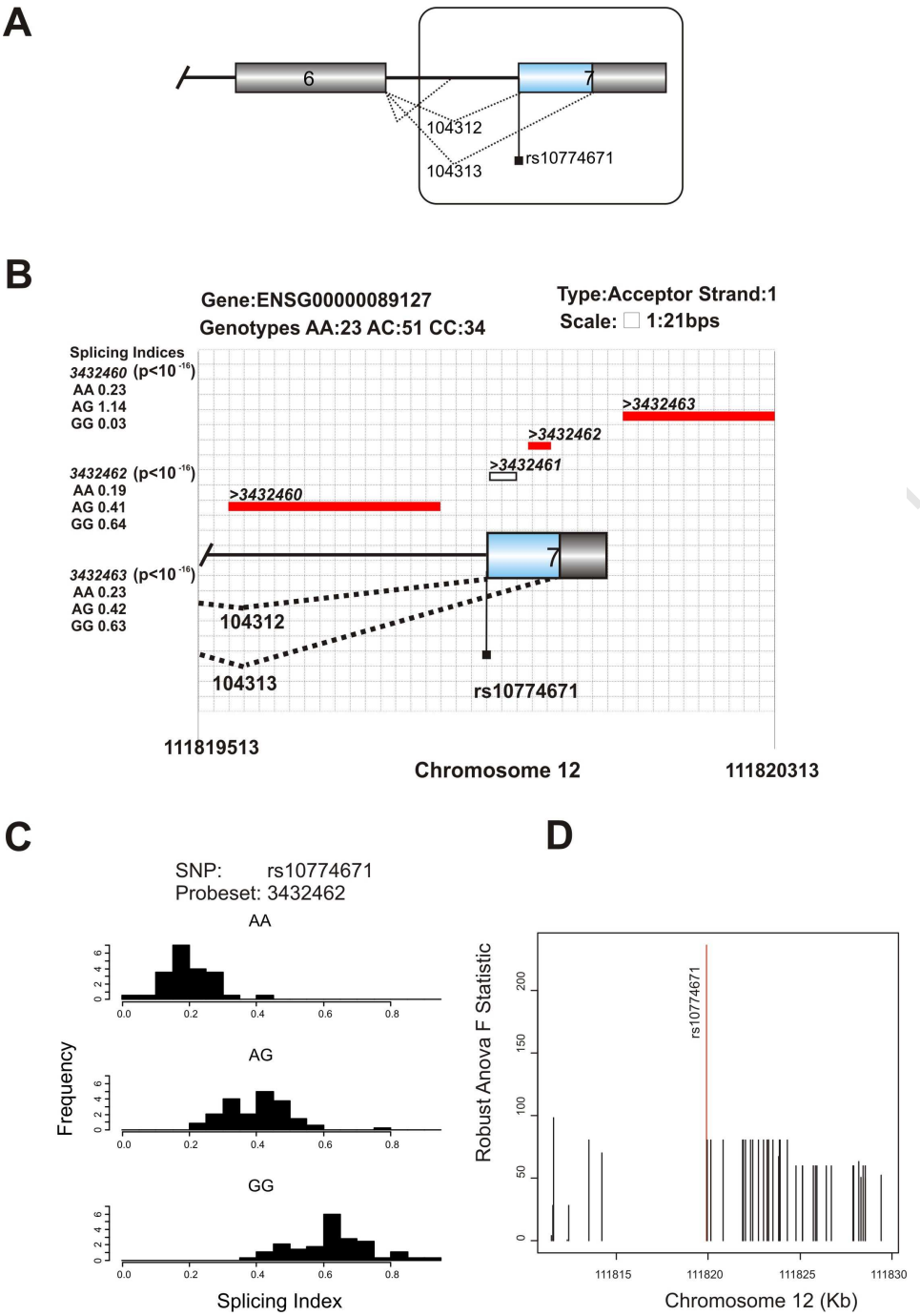


Figure 3: Support for allele-specific acceptor site use in the *OAS1* gene. **A)** Genomic sequence of the *OAS1* gene showing the alternatively spliced exons. The boxed section is magnified and drawn to scale in the next panel. **B)** Relationship between the genotypes of the SNP and the splicing indices of nearby probesets, illustrating that there is likely to be a complex pattern of allele-specific splicing in this gene. Probesets in red are significantly associated with the SNP genotype. The p-values for the association of these probesets to SNP genotypes are also included. Unfilled rectangles represent probesets tha were not tested for association with the genotype because they were not detected above background in a sufficient number of the cell lines or were too distnat from the SNP. **C)** Histograms showing the splicing index distribution as a function of the genotype of a SNP, rs10774671, at the G nucleotide of the canonical splice acceptor site. **D)** Association plot illustrating that rs10774671 is more strongly associated with a probeset between the SNP and an alternative acceptor site than any other SNP in the region for which genotype data were available.

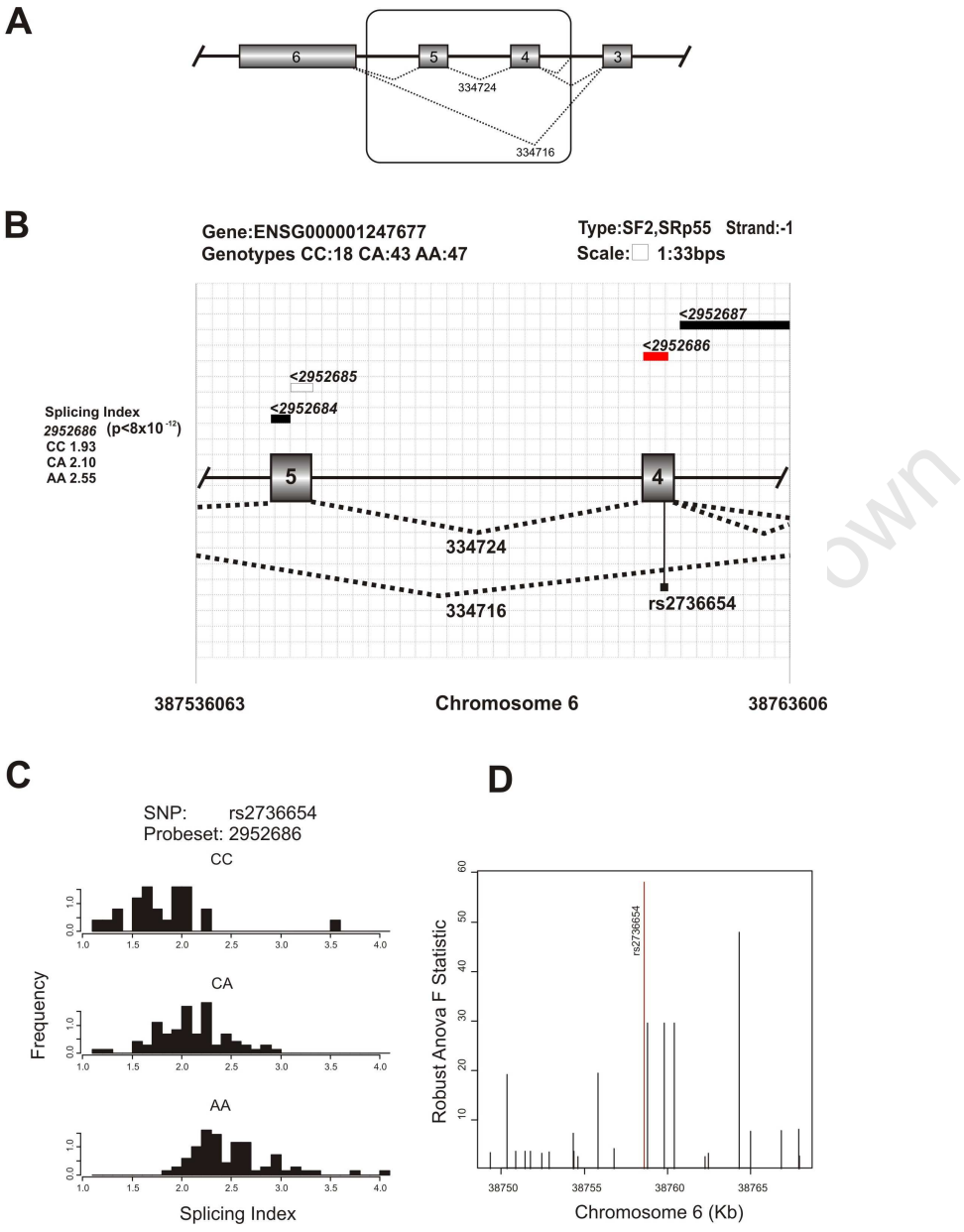


Figure 4: Support for allele-specific exon-skipping in the *GLO1* gene. **A)** Genomic sequence of the *GLO1* gene showing the alternatively spliced exons. **B)** Illustration of the relationship between the genotypes of this SNP and splicing indices of nearby probesets, using the same conventions as in Figure 3. **C)** Histograms showing the splicing index distribution as a function of the genotype of a SNP, rs2736654, predicted to affect an exonic splice enhancer site. **D)** Association plot illustrating that rs2736654 is marginally more strongly associated with a probeset spanning exon 4 than any of the other SNPs in the region for which genotype data were available.

Table 2: A subset of the previously reported allele-specific splice isoforms detected in this study

Gene	Exon	mSNP (p-value)	Exon-array confirmation	srSNP	*Cis*-element	References
CD45	Exon 4	rs12129883 (0.020551)	---	---	ESS	(Jacobsen et al., 2002)
COL5A1	Exon 65	rs13946 (0.046)	---	---	Acceptor site	(Wenstrup et al., 1996)
ETV4	Exon 3	rs3765174 (0.014)	---	---	NAGNAG acceptor	(Hiller et al., 2006)
GABRR1	Exon 2	rs12200969 (0.034)	---	rs4590242	NAGNAG acceptor	(Hiller et al., 2006)
ITPA	Exon 2 and Exon 3	rs13830 (0.030)	YES	---	Exonic splicing silencer element in exon 2	(Arenas et al., 2007)
LDLR	Exon 12	---	YES	rs688	SF2	(Zhu et al., 2007)
MUC1	Exon 2	rs4072037 (0.0011)	---	---	Acceptor site	(Ligtenberg et al., 1991)
OAS1	Exon 7	rs2660 (0.00063)	YES	rs10774671	Acceptor	(Bonnevie-Nielsen et al., 2005)
PMM2	Exon 5	rs2072688 (0.0027)	---	---	ESE	(Vuillaumier-Barrot et al., 1999)
RBM23	Exon 6	rs1951119 ($1.0x10^{-6}$)	YES	rs2295682	SRp40*	(Hull et al., 2007)
UROD	Exon 4	rs1804886 (0.0027)	---	---	---	(McManus et al., 1996)

* Hull *et al.* (2007) did not report that the SNP, rs2295982, disrupts an ESE

Among the classes of splicing regulatory regions analysed, SNPs that occurred in donor sites were slightly more likely to be confirmed by EST and/or exon array data (Table 1). In addition to the srSNPs for which there is supporting evidence from EST and/or exon array data a further 51 mSNPs were supported by exon array data, but no candidate srSNP was identified that could explain the allelic difference in splicing. Some of these may be false positive mSNPs but for the remainder, the causative SNP may be in an intronic splicing element (intronic splicing elements were not included in the genome-wide scan for srSNPs) or in, as yet, uncharacterized splicing regulatory elements. The

possibility also exists that some of the identified putative allele-specific isoforms are caused by mutations located within *trans* regulators of splicing and that association with nearby polymorphisms is a result of population stratification rather than a direct *cis*-acting effect.

5.3.4 Cross-validation of EST and exon array results

15 out of the 54 distinct alternate exon junction pairs with evidence of allele specific splicing from the EST data using a false detection cut-off of 0.2 (above), could be tested for allele-specific splicing using the exon array data. In order to be tested, the 15 mSNPs had to be among the SNPs genotyped in the HapMap populations, a probe or probes had to occur between the genomic coordinates spanned by the alternative exon junction pair and the probe had to be detectable above background in at least some of the lymphoblastoid cell lines. Of these 15, 9 (60%) had at least one probe between the genomic coordinates of the junction pair for which the SI was significantly associated with the genotype of the mSNP ($p < 0.05$, with Bonferroni correction in the case where multiple probes were tested for association with a single mSNP). By comparison, there were 29 (23%) associations from 124 exon junctions that could be tested from a random set of 10,000 alternatively spliced exon junction pairs from ASAPII and nearby exonic SNPs that showed no association with the mRNA isoform. The proportion of allele-specific splicing candidates from the EST data that could be confirmed using the exon array data was significantly higher than for alternatively spliced exon junctions with no evidence of allele-specificity from ESTs ($p = 0.005$ using Fisher's Exact Test). This overlap of allele-specific splicing candidates identified by very different technologies, provides cross-validation for the candidates identified using the two approaches.

5.3.5 Splicing index association plots

A significant association between the SI of a probeset and an srSNP is insufficient to infer a causal relationship between the srSNP and variation in SI. It is possible that the putative srSNP is not causally related to the observed difference in splicing and, instead, that it is in linkage disequilibrium with a nearby SNP that was not predicted to affect

splicing (because of the imperfect understanding of splicing regulation). We can begin to investigate this possibility by testing for an association between the SI and genotype for all of the other nearby SNPs for which genotype data are available for the lymphoblastoid cell lines. For each srSNP we tested for an association between SI and genotype for all genotyped SNPs within 10kb of the srSNP. On average there were 25 such SNPs per srSNP. For the majority (61.8%) of the srSNPs strongly supported by the exon array data, the predicted srSNP showed the most significant or joint most significant association between SI and genotype for at least one of the probesets tested. For the remainder, an alternative SNP, not necessarily predicted to affect splicing, showed a more strongly significant association. The mechanisms through which these alternative SNPs may affect splicing require further investigation. Examples of the association plots are shown in Figures 3 and 4. Similar association plots for all of the srSNPs supported by the lymphoblastoid exon array data are available from our website.

5.3.6 Analysis of allele-specific mRNA splicing candidates

For several of the examples of allele-specific splicing that we identified we were able to find published research articles confirming the same event (Table 2). The mSNP rs2660 ($p = 0.0006$; Figure 1), which we detected in the 2',5'-oligoadenylate synthetase 1 (*OAS1*) gene, for example, has been shown experimentally to be in strong linkage disequilibrium with the srSNP, rs10774671, which occurs at the G of a canonical acceptor site (Bonnevie-Nielsen et al., 2005). Disruption of the canonical acceptor site in intron 6 of the *OAS1* gene promotes the use of two cryptic acceptor sites. Using the EST data we detected one of the cryptic acceptor sites, located 98 bps from the wild type acceptor site (Figure 1). This event was also detectable using the lymphoblastoid exon array data (Figure 3).

There are also many cases of previously unpublished splicing polymorphisms among our results, some of which are likely to be functionally and medically important. For example, the lymphoblastoid exon array data provide strong evidence (robust Anova F statistic: 40.5; $p = 8.72 \times 10^{-11}$) for an association between a probeset in exon 4 of the

GLO1 gene, encoding an enzyme (glyoxalase I) that has been reported to show lower activity in the brains of individuals affected by autism compared to control individuals (Junaid et al., 2004) and the genotype of a SNP in the same exon (C419A or rs2736654; Figure 4). Reduction in enzyme activity has been attributed to the direct effect of this non-synonymous SNP on the amino acid sequence of the protein. The ancestral A allele has been reported to be significantly associated with autism (Junaid et al., 2004) and certain types of panic disorders (Politi et al., 2006). A larger scale study, however, has questioned the association with autism, but has found that the A allele may have a protective effect in the siblings of individuals with autism (Sacco et al., 2007). This non-synonymous SNP occurs in a predicted exon splice enhancer site (the genomic scan for srSNPs predicts that this site acts as an ESE for both SF2 and SRp55 and the A and C alleles have scores 0.44 and 2.96, respectively, for SF2 and 1.39 and 3.53, respectively for SRp55). EST evidence from ASAPII suggests that two exons are skipped (Kim et al., 2007). Skipping of these exons is likely to have a much greater impact on the protein function than the replacement of Alanine by Glutamine at a single site within one of the exons. While the role of *GLO1* in neurological disorders remains controversial (Thornalley, 2006), Sacco *et al.* (Sacco et al., 2007), highlight the need for further investigation of the functional impact of the C419A. Our results suggest that the polymorphism is very likely to impact on splicing. This could have a significant impact on glyoxalase I activity and be the mechanism underlying the disease association.

5.4 Discussion

Large-scale discovery of genomic variants that affect splicing has the capacity to accelerate the association of diseases to causative genomic variants. However, because it is difficult and in many cases, currently not possible to determine the effect of a genomic variant on splicing or on the regulation of alternative splice isoforms from genomic sequence data alone, this remains a challenging task and requires the integration of information from different data types. At present, no single source of data can provide information about all forms of splice variants and each source of data has advantages as well as disadvantages. The publicly available exon array data that we have used here

represents an extremely extensive dataset on isoform abundance in human lymphoblastoid cell lines that can be correlated with the genotype of the cell line. However, this data provides no information on transcripts that are not expressed in lymphoblastoids, or on splicing mutations that affect relative isoform abundance in only a subset of expression contexts. Furthermore, depending on the exact location of probesets in a given gene, many of the transcript isoforms that occur, particularly those that affect donor or acceptor site but do not cause exon skipping or inclusion, are undetectable using exon arrays. When alternative isoforms are distinguishable using the exon arrays, they still provide little information on the nature of the isoforms, and this may need to be inferred either by integrating information from other sources or experimentally.

EST sequences provide information on the structure of alternative isoforms and include data from different gene expression contexts, but this information is highly biased towards ends of genes and is sparse, for all but the most highly expressed genes. The simulations provide ample evidence for frequent allele-specific splicing but also illustrate that there is not enough data to confirm most cases, especially when the effect of a very large number of statistical tests is considered. There were several published examples in the current study of genes known to be spliced in an allele-specific manner, but for which the allele-specific splicing model fits the data no better than the null model. EST data could have a low representation of allele-specific isoform as these are most likely to be minor isoforms, which are generally characterized by low expression levels. However, although most cases of allele-specific splicing will not be detectable using EST sequences alone, ESTs can often be used to elucidate the nature of the allele-specific splicing events detected because ESTs provide information on the actual transcripts that occur.

GLO1 provides an example of a gene with a mutation that is likely to affect splicing, but although there was good coverage of this gene in the EST databases, the allele-specific splicing event was not detectable from the EST data. Because the putative causal SNP is on the skipped exon it is only observed when the constitutive isoform occurs and therefore cannot be tested for association with the skipping event using the EST data. Furthermore, there is only one EST that captures the exon skipping event, denoted by

junction 334716 in ASAPII. There are also several cases of known splicing polymorphisms that could be detected from ESTs but not from the exon array data (Table 2). The gamma-aminobutyric acid (GABA) receptor, rho 1, gene (*GABRR1*), for example, was previously shown to have a SNP (rs4590242), located in the acceptor site that promotes use of an alternate NAGNAG acceptor (Hiller et al., 2006). We detected this srSNP in the genomic data and EST data provided evidence of its effect on splicing with an mSNP rs12200969, (p-value=0.033921). However, due to the lack of a probe that coincides exactly with the end of the exon, the exon arrays were unable to detect this subtle alternative splicing event.

The probability of linkage disequilibrium of an srSNP and mSNP decreases with the distance that separates them. This limitation is highlighted by the failure to associate several transcribed SNPs (rs3093906, rs3093905, rs3093921, rs3093925, rs3093926, rs3093927), located >5000bp away from a putative allele-specific splicing event in the Ribonuclease P RNA component H1 gene (*PARP-2*). The ASAPII database contains the two alternate donor sites at this junction that are 39bp apart and are supported by a total of 28 expressed transcripts, and two *PARP-2* protein isoforms differing by 13 amino acids have been deposited in the SWISSPROT database (Boeckmann et al., 2003). We detected an srSNP (rs2297616) located at position 4 of the corresponding splice donor site and the exon-array data provide strong evidence for an association between the splicing index of a probeset that overlaps the 39bp region between the alternative donor sites and the genotype of this SNP (p-value = 2.00×10^{-57}).

In previous work we used a heuristic method (see Chapter 4) to find associations between SNPs mapped to ESTs and alternatively spliced isoforms in order to detect candidate allele-specific isoforms and to quantify the proportion of alternatively spliced genes that are spliced allele-specifically. However, such associations can also occur because of normal regulation of alternative splicing. For example, consider an alternatively spliced gene for which ESTs occur in just two of the cDNA libraries in dbEST. Assuming that these libraries were constructed from the tissues of single individuals, it is possible that these individuals have different genotypes for an exonic SNP in the gene. If the alternative isoforms of the gene happen to be regulated in a tissue specific way and if the

cDNA libraries are derived from different tissues then this could result in an association between the alleles of the SNP and the mRNA isoforms. This association can be highly significant if there are many ESTs of the gene in the two cDNA libraries in which it occurs. To circumvent this problem in our previous work, we took a maximum of two ESTs per cDNA library (one for each allele of the SNP from heterozygous libraries and just one from homozygous libraries). This caused a substantial loss of data and reduction in power to detect and quantify allele-specific mRNA splicing. In the present work we explicitly model the regulation of alternative splicing and make much better use of the available data.

These results and previous reports (Nembaware et al., 2004; Kwan et al., 2007) suggest that polymorphisms that affect splicing are common. This has important implications, not only for discovering the molecular bases of genetic diseases, but also for the study of alternative splicing. A gene cannot be confirmed to be alternatively spliced unless multiple isoforms are observed from the same allele. Until then the possibility remains that the alternative isoforms observed are polymorphic variants rather than alternatively spliced. Although we have found ample evidence for allelic differences in splicing, isoforms that result entirely from sequence variants might be less common. In the set of examples we report here, there is a relatively small proportion of cases in which the data suggest that the SI might be zero for some variants. Allele-specific splicing may be particularly important in the context of investigations of the regulation of alternative splicing (Sugnet et al., 2006; Xu et al., 2002). Such investigations should ensure that multiple samples from the same tissue source are not treated as independent.

Regulation of splicing is incompletely characterized and additional *cis* elements that regulate splicing are still being discovered (Yeo et al., 2007). A limitation of the current study is that the srSNP candidates are restricted to a subset of well characterized *cis*-acting splice regulatory elements (donor and acceptor sites, polypyrimidine tract, branch points and some exonic splicing elements). The phosphomannomutase 2 gene (*PMM2*), for example, which has allele-specific skipping of exon 5 due to a SNP that disrupts an ESE composed of (GAR)n repeats (Vuillaumier-Barrot et al., 1999), where R is a purine, was detected using the EST data ($p = 0.003$); however, we could not identify an srSNP

because the disrupted ESE is not detected by ESEfinder. Polymorphisms not found in *cis*-regulatory regions can also result in apparent allele-specific splicing if they introduce premature termination codons (PTCs) (Savas et al., 2006) and cause differential nonsense-mediated decay of alternative alleles. Such SNPs are not included in our srSNP database. We have also restricted our analysis to single nucleotide polymorphisms but allele-specific splicing could be due in many cases to other types of polymorphisms such as insertions and deletions (Romano et al., 2002).

In the majority of the examples of allele-specific splicing we have detected, the difference in splicing is quantitative rather than qualitative. This can occur for a gene that is alternatively spliced, but for which a polymorphism exists that affects the proportions of alternative isoforms produced. In some cases, particularly for common polymorphisms, the size of the effect on SI can be relatively small, but still highly significant because of the relatively large number of individuals in each genotype group. In other cases, e.g the alternative isoforms of the *OAS1* gene shown in Figure 3, the SI associated with one genotype may be much greater than for the other genotypes. In general, the size of an effect on SI sufficient for an effect on phenotype is likely to vary substantially from transcript to transcript. Consistent with what has been observed previously for *cis*-acting polymorphisms with a quantitative effect on splicing (Buchner et al., 2003), for the majority of the probesets for which SI was significantly associated with SNP genotype, the SI value of the heterozygote was intermediate to the SI of the two homozygotes. In 888 (77%) of 1,157 associations for which heterozygote and both homozygote cell-lines for the SNP were available, the SI of the heterozygote had an intermediate value. For stronger associations (that remained significant using a family-wise error rate of 0.05), this figure was 242 (96%) from a total of 253.

Using the expression quantitative trait loci (eQTLs) (Morley et al., 2004), as an analogy, loci that affect splicing might be termed splicing quantitative trait loci (sQTLs). In this study we have attempted to identify only *cis*-acting sQTLs. *Trans*-acting sQTLs are also likely to exist, particularly at genes that are involved in regulating alternative splicing. However, the ratio of *trans* to *cis* acting variants may be much smaller for sQTLs than for

eQTLs, because of the relatively more complex regulation of transcription initiation compared to splicing. We have taken a candidate SNP approach to detecting splicing polymorphisms. With the availability of whole-genome exon array data it is also possible to adopt a less directed approach analogous to methods that have been used previously to detect expression quantitative trait loci (Morley et al., 2004). Each probeset could be tested for association with every SNP that overlaps the transcripts to which it belongs. However, because of the multiplicity of probesets per gene this would result in a very large number of tests and would be likely to yield a much larger set of candidates, but potentially a set with lower specificity and for which interpretation is more difficult.

5.5 Acknowledgements

This chapter was published in BMC Genomics and was perfomed in collaboration with Bukiwe Lupindo, Katherine Schouest, Charles Spillane, Konrad Scheffler and Cathal Seoighe. I was in charge of scanning for the mSNPs, srSNPs, data management and quality control and integration of all the outputs. Cathal Seoighe formulated the maximum likelihood method with input from Konrad Scheffler.

Chapter 6

An Exploratory Survey of Strain-specific Splicing in Mouse

Abstract

Distinct inbred mouse strains exhibit a wide range of heritable and naturally occurring phenotypic differences; however, the molecular and functional basis of such variation is largely unknown. Numerous studies have reported detailed examples of strain-specific splicing with prominent consequences for gene function. Increased knowledge of strain-specific splicing promises to be of great value in facilitating the use of mouse as a model for human disease in biomedical research. However, to-date, no genome-wide analysis has been performed, mainly due to the scarcity of genome-wide expression and genomic sequence data for most mouse strains with the exception of the reference strain, C57BL/6J. Hence our analysis explores strain-specific splicing of C57BL/6J and related strains in comparisons to all other unrelated mouse strains that have publicly available transcript data. From our exploratory study, we estimate that differences in splicing between the C57BL/6J and related strain compared to unrelated strains occurs in about 10% of alternatively spliced isoforms in the ASAPII database. Furthermore, we scanned for associations between the strain category from which the transcript is derived, and the alternatively spliced isoforms from the ASAPII database.

6.1 Introduction

Since the early, 1900's, over 100 inbred mouse strains with specific characteristics have been designed and propagated for use as animal models for human diseases (Beck et al., 2000; Wade and Daly, 2005). Unlike the more recent strategy of developing new strains by directly tinkering with genetic material through sequence manipulation and induced mutations, previously, mouse breeders relied heavily on random heritable mutations that could be fixed in mouse colonies (Beck et al., 2000). Most of the commonly used mouse strains were established through phenotype-driven breeding and selection which took place well before the field of genetics was fully developed (Wade and Daly, 2005). Such heritable phenotypes are encoded by genetic variation in the genomes of inbred mouse strains (Wade and Daly, 2005).

The main source of these genetic variations is the different historic protocols that combined breeding and selection of a small but diverse group of ancestral *Mus musculus* species (Beck et al., 2000). A secondary cause of genetic variability in inbred mouse strains are spontaneous mutations that occurred in ancestral strains. The current challenge is to now discover the genetically controlled molecular processes that are likely to mediate phenotypic differences which include differences in mRNA splicing.

Numerous studies have revealed that genetically distinct mouse strains can have mRNAs with unique sets of exons (qualitative strain-specific splicing), or can differ in the rate of inclusion of specific exons in the mature mRNA (quantitative strain-specific splicing) (Buchner et al., 2003; Dolney et al., 2001; Sandilands et al., 2004). The effect of mouse strain-specific mRNA splicing on the mouse phenotype can be striking due to the potential of splicing polymorphisms to drastically alter gene functions. For example strain-specific isoforms of the K-opioid gene responsible for catalyzing alcohol has been implicated as the underlying cause of alcohol preference or avoidance in mice (Dolney et al., 2001). The DBA/2J mouse strain (alcohol avoiding) expresses an additional K-opioid receptor mRNA isoform in comparison to the alcohol preferring strains C57BL/6J and BALB/cj strains (Dolney et al., 2001).

Quantitative strain-specific splicing has also been reported to contribute to mouse phenotypic variability as genetic modifiers of disease. The expression of strain-specific isoforms of the zinc finger protein gene (*SCMN1*) in the C57BL/6J strain and the C3H strain is the underlying cause of variation in disease susceptibility in the B6 *SNC8a med*^{J} and C3H *SNC8a med*^{J} mutants (Buchner et al., 2003). Considering the potential impact of strain-specific mRNA splicing on phenotypic variability in mouse, disregarding strain-specific mRNA isoforms would be a serious oversight.

Important features of inbred mouse strains that make them ideal for biomedical research is their isogenicity (genetic identity) and the fact that each individual mouse is characterized by genome-wide homozygosity, which is achieved after at least 20 consecutive generations of sibling mating (Beck et al., 2000). These characteristics make it easier to compare experimental results from geographically separated labs working on the same inbred strains. With the recent influx of transcripts from a wide range of inbred strains, collating publicly available mouse data offers an accelerated means of studying strain-specific splicing.

To our knowledge no attempt has been made to perform a genome-wide study of the prevalence of strain-specific mRNA isoforms. A preliminary study of the transcripts in EMBL highlighted that transcripts are biased towards the mouse reference genome, C57BL/6J, with ESTs sparsely distributed across approximately 30 mice strains as shown in Figure 1. Given the biased distribution of transcript data towards the C57BL/6J, and the unclear relationships that exist among mouse inbred strains, the process of categorizing the data according to strains required the implementation of a transcript partitioning algorithm. We developed a method that exploits EST-derived SNP alleles to partition clone-libraries into two virtual strains; the B6 and the nonB6 strains. The B6 constitutes the C57BL/6J and its related strains some of which are shown in Figure 2, while all the other inbred strains, not related to the C57BL/6J, fall under the nonB6 strain. We estimated the prevalence of B6 specific splicing, using transcript data that was used to support alternative splicing events in the ASAPII database (Kim et al., 2007). Our premise is that splicing mutations, whether *cis* or *trans*-acting, will generally be shared by

related mice (from the same or similar strains) and this provides an alternative approach for the detection of splicing variants that does not require expressed genetic markers in linkage disequilibrium with the splicing polymorphism (the method used for the human studies in Chapter 3 and 4). We also attempted to detect genes that are spliced in a strain-specific manner by testing for associations between strain category (B6 and nonB6) from which the transcript is derived and the alternatively spliced isoforms from the ASAPII database (Kim et al., 2007).

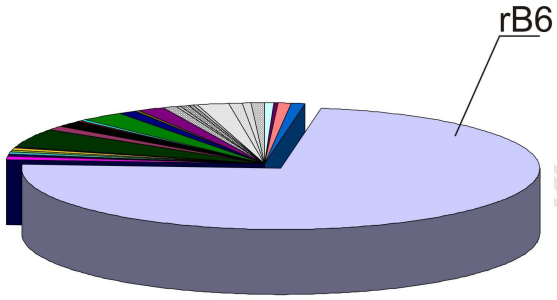


Figure 1: Pie-chart of the distribution of EST sequences according to cDNA library strain annotations. ESTs from cDNA libraries annotated as C57BL/6 case (insensitive) made up 73.59% of the database.

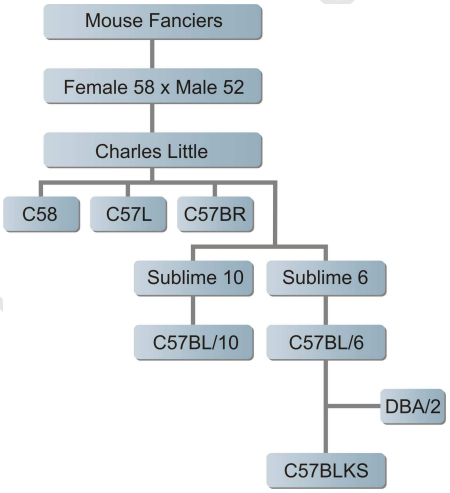


Figure 2: C57BL/6 and related strains (B6)**.** One of the most widely used inbred strains is C57BL/6. Charles Little, an undergraduate student studying under Castle at Harvard University, obtained the ancestor of the C57Bl/6J from A.E.C Lathrop, a retired school teacher who was a mouse fancier. Part of the genealogical history of the B6 strains is shown above highlighting that this strain has contributed to other sub-strains after crossings with other inbred strains such as the DBA/2 strain. The illustration was adapted from two publications Buchner et al., (2003) and Beck et al., (2000).

6.2 Data and methods

6.2.1 A database of SNPs mapped to ESTs

We downloaded the UCSC mouse genome tracks (mm7) which are based on the NCBI genome assembly 36 (Karolchik *et al.*, 2003) in August 2007. Mouse transcripts from the mouse transcript division of EMBL release 93 were downloaded from the European Bioinformatics Institute (www.ebi.ac.uk). All mouse transcripts together with their associated cDNA libraries and strain annotations were extracted from the EMBL flat files. The SNP to EST mapping procedure performed on human data (described in detail in Chapter 2), was applied to the mouse data and a similar MySQL database created.

6.2.2 cDNA library classification

cDNA libraries annotated as C57BL/6J or C57Bl/6J were used to compile a dataset of SNP alleles for what we termed the reference C57BL/6J dataset (ref-B6). Mouse inbred strains are expected to be homozygous at all loci and hence only one SNP allele is expected for each SNP location across the whole mouse genome. However, practically this is not possible due to sequencing errors and also due to at least the 5% heterogosity which exists in some inbred strains. We found a significant number of heterozygous SNPs in the ref-B6 SNP allele dataset. We assigned the SNP alleles to ref-B6 strain by calculating the most prevalent SNP allele. The most prevalent SNP allele for each individual cDNA library, that had transcripts in the SNP to EST dataset were also computed. Based on the percentage of identical alleles between each cDNA library to the ref-B6 SNP allele dataset, we could classify libraries into three different categories, B6, nonB6 and a class of cDNA libraries that could not be profiled with certainty.

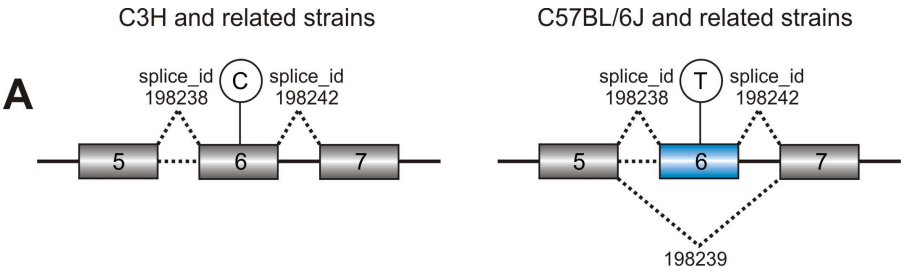
6.2.3 Matrices

Mouse data from the ASAPII database was downloaded in August 2007. ASAPII provides alternatively spliced isoforms as pairs of mutually exclusive introns. The supporting transcript data used in the detection of alternatively spliced isoforms is also available in the ASAPII database. Transcript data that supported alternatively spliced isoforms and which could also be profiled either as B6/nonB6 strains as described in the previous section were used to create 2X2 contingency matrices for each alternatively spliced junction pair (Figure 3). A total of 32455 matrices were created, but only 3987 matrices had row and column sums greater than 1 and were thus informative enough for further analysis (Chapter 4).

We sampled one EST per library for each splice junction for the same reasons discussed previously in Chapter 4. In the ASAPII database (Kim et al., 2007), a single alternatively spliced isoform can be represented by more than one pair of splice junctions. We therefore restricted our matrices to one splice junction per gene and obtained a total of 2149 matrices.

6.2.4 Simulations

In the case of qualitative B6-specific mRNA isoforms, every matrix affected necessarily has at least one zero cell in the nonB6 row. Figure 3 shows a matrix with a zero entry, using the *SCNM1* gene which has a published strain-specific exon 6 skipping event (Buchner et al., 2003). We constructed 1000 randomized matrices of the observed data under the constraint that the rows and columns sums be preserved. For each replicate dataset, the number of matrices with at least one zero cell in the nonB6 row were counted. This is analogous to the method that we used to estimate the proportion of allele-specific transcript isoforms in human, described in Chapter 4.



B

Strain	SPLICE VARIANT 198238	SPLICE VARIANT 198239
rB6	15	4
nonB6	14	0

Figure 3: A) The *SCNM1* has been reported to be spliced in a strain-specific manner. The C to T mutation that occurred in the B6 strain has been predicted to disrupt an ESE causing exon 6 skipping in the B6 mouse strains. **B)** The matrix constructed for the *SCNM1* strain-specific splicing event using the profiled ESTs data that could be mapped to the ASAPII transcripts.

We also constructed 1000 randomized replicates of the observed matrices such that a predefined proportion of the simulated matrices were forced to have at least one zero cell in the nonB6 row. For each of the replicate dataset, the number of matrices with at least one zero cell in the nonB6 row were then counted.

6.2.5 Detection of individual cases of strain-specific splicing

Given the large number of individual matrices being studied means that we expect several false positives purely by chance only. We therefore corrected for multiple testing using the Bonferroni method as well as a false discovery approach (Storey and Tibshirani, 2003).

6.3 Results

6.3.1 A map of SNPs mapped to ESTs

A total of 740 cDNA libraries with 4,653,859 ESTs were extracted from the EMBL data files. However, only 667 cDNA libraries had at least one EST that mapped unambiguously to the genome and with a SNP located within its genomic boundaries.

6.3.2 Classification of cDNA libraries as C57BL/6 or non C57BL/6

To investigate strain-specific splicing in mouse using publicly available transcript data, classification of the cDNA libraries according to the strains they originated from is mandatory. Mouse has a well defined strain nomenclature system (www.jax.org), however there still exists a lack of uniformity in the strain labeling of cDNA libraries deposited in the public databases. For example there are numerous variations in the manner in which C57BL/6J cDNA strain libraries have been deposited; C57BL6J or C57Bl/6J or C57BL6J etc. Mis-breeding, genetic contaminations and strain mix-ups can occur, so animals from the same inbred strains from separate labs may no longer be isogenic. Furthermore, there is a large amount of overlap in mouse genealogical trees which intersect at several locations with unknown sequence contributions and genomic exchange (Wade and Daly, 2005). Therefore, relying on strain names alone to profile and partition transcript data is unreliable.

Improvements in rapid genome-wide SNP genotyping methods has seen recent efforts to untangle mouse inbred strain relationships increasingly focusing on investigating relationships between inbred mouse strains using strain-specific allelic distribution of SNPs (Petkov et al., 2004a; Petkov et al., 2004b). Petkov et al., (2004a) established a SNP-based method to profile inbred mouse using 235 SNPs in 48 mouse strains. A follow up study by the same group, using a set of 1639 SNPs effectively investigated mouse strain relations for a much bigger group of 102 mouse strains (Petkov et al., 2004b). The usefulness of the SNP profiling methods was shown by the ability of the study to verify well known strain contaminations and by successfully reconstructing the genealogical

trees for the mouse strains tested (Petkov et al., 2004b). After a preliminary analysis of sequence data and noting the sparsity of data for most strains (Figure 1), we performed our analysis on strain-specific splicing in the B6 strain in comparison to all the nonB6 strains. We then developed a SNP profiling method to classify libraries as B6 or nonB6 using publicly available SNP and transcript data as summarized in the flow diagram in Figure 5. Of the 661 cDNA libraries that had at a least one polymorphic EST, only 409 cDNA libraries had at least 30 SNPs and thus were informative enough for to use in the classification of cDNA libraries. We have categorized 348 cDNA libraries either as B6 or nonB6 and the remaining libraries we failed to categorise with certainty. An additional advantage to this approach is that even cDNA libraries that were not annotated according to the strain in the EMBL transcript data could be profiled using our method.

Reliability of the profiling method was assessed by sampling and verifying cDNA libraries from the three categories see Table 1. Verification was done using EMBL strain labels. The cDNA libraries with very low similarities to the B6 strain are from the CZECH mouse strain. Based on documented genealogies, the CZECH strain and the B6 do not seem to share a recent common ancestor (Beck et al., 2000). Table 1 also shows some examples of nonB6 and B6 libraries that were accurately profiled.

Table 1: Verification of the cDNA profiling method

cDNA library	Category	Strain-annotation	Percentage of identical alleles between the B6 and nonB6 strains (Standard error of the mean)
NCI_CGAP_Lu29	nonB6	CZECH II	52.98 (0.311)
NCI_CGAP_Lu30	nonB6	CZECH II	53.86% (0.40)
NCI_CGAP_Lu29 lung tumor	nonB6	CZECH II	47.54% (3.20)
Mouse Bone Marrow-derived Mast Cell Expression Library	nonB6	BALB/cJ	50% (14.43)
NIH_BMAP_MAM	B6	C57BL/6J	99.90 (0.096)
Stratagene mouse lung 937302	Uncertain	C57BL/6 x CBA	90.93 (0.48)

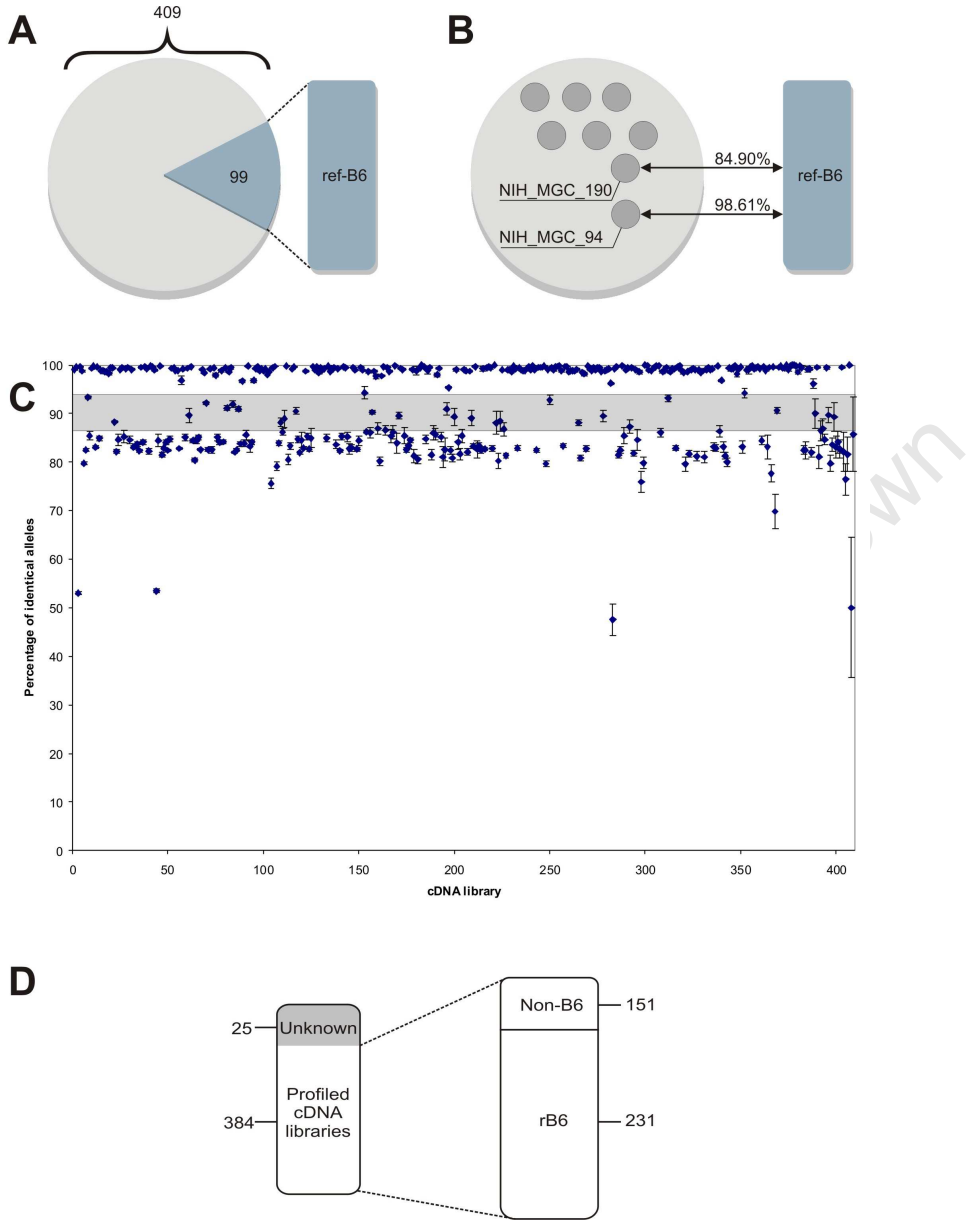


Figure 5: A flow diagram of the method used to partition cDNA libraries as B6 or nonB6. **A)** Out of the 667 libraries that had ESTs and SNPs mapping to the genome, only 409 cDNA libraries had at least 30 SNPs mapped to them and were used for further analysis. 99 clone libraries clearly annotated as C57BL/6J and C57Bl/6J were used to create a dataset of SNP alleles highly prevalent in the C57BL/6J strain (ref-B6) **B)** Percentages of identical alleles were computed between each cDNA library and the ref-B6 library. **C)** A plot of percentage identities and standard error of the mean of identical alleles computed in step B. The grey region shows the region with cDNA libraries which could not be profiled either as B6 or nonB6 with certainty. **D)** The final dataset of cDNAs had 382 cDNA libraries classified either as B6 or nonB6. Out of a total of 409 cDNA libraries 25 could not be classified as B6 or nonB6.

6.3.3 Prevalence of C57BL/6J -specific isoforms

Among the 2149 matrices used for the simulations, we observed 650 matrices that had a zero cell (i.e. one combination of strain and mRNA isoform that did not occur) and were thus consistent with qualitative allele-specific splicing. In equivalent sets of randomized matrices, on average, 497 matrices had a zero element (Figure 6A). The 153 extra matrices in the observed data (650 - 497), with at least one zero cell is not likely to have resulted from random data with no relationship between strain and transcript isoforms and instead indicates the presence of allele-specific isoforms in the data (Figure 5a).

The numbers of matrices (with 95% confidence intervals), with a zero cell from 1000 simulated replicates are shown in Figure 6A. The horizontal line in Figure 6B shows the number of matrices with a zero cell in the observed data. If we consider qualitative allele-specific splicing as the only alternative to random association (i.e. quantitative allele-specificity not allowed), the proportion of allele-specific isoforms in the observed data can be inferred from the intersection of the horizontal line with the simulated data points. This intersection shows additional matrices with a zero cell in the observed data beyond what we would expect to occur by chance. We can infer from Figure 6A that the proportion of allele-specific splicing most consistent with 150 additional matrices with a zero cell in the observed data is 8.4%. The lower (6.6%), and upper bound (10.4%), of our estimate were inferred from the span of the confidence interval inferred (Figure 6B).

6.3.4 Gene candidates

A significant proportion of the candidate strain-specific isoforms are likely to be false positives due to multiple testing and thus Bonferroni correction for multiple testing was performed. However, none of the matrices were statistically significant after the correction for multiple testing using the Bonferonni correction and well as less conservative methods such as the false discovery approach (Storey and Tibshirani, 2003). This observation is likely to be due to a lack of data. Since this is an exploratory study,

we present all the top gene candidates, ranked according to the significance of their p-values (Table 2).

Interestingly, there are also many cases of strain-specific splicing which could be functionally important. The strain-specific splicing of a splicing factor (Splicing factor 3a, subunit 3), could have an impact on the splicing of many other genes. Yet another gene candidate which could be an interesting candidate to follow up further is the Breast cancer antigen gene **(***ERGIC3*). The strain-specific splicing event we detected is an exon 8 skipping event. The exon is 32 bps and since it is located in the coding region of the gene, it has potential to shift the reading frame. Changes in reading frames can alter protein function substantially. Interestingly strain-specific isoforms of this protein have also been reported in the SWISSPROT database (Boeckmann et al., 2003). We checked for possible splicing mutations in the gene. A SNP (rs27325066) is located 8 bps away from the donor site of this exon and could be causing the strain-specific splicing event.

Studies that scan for genes which are differentially expressed in a strain-specific manner have also reported numerous gene candidates (Sandberg et al., 2000; Cowles et al., 2002). However, most of these studies have not gone as far as elucidating the mechanisms that could be causing the differences in expression. In this study we have detected the Adenylate cyclase associated protein gene (*CAP*), highlighted in Table 2, which was previously reported to be differentially expressed in a strain-specific manner. Based on microarray and semi-quantitative RT-PCR, Sandberg et al., (2000), reported that the *CAP* gene is highly expressed in the in 129SvEv strain in comparison to the C57BL/6J strains. The probes used in the microarray experiment could have targeted the strain-specific alternatively spliced segment of the *CAP* mRNA transcript. Hence, the quantitative differential expression of the *CAP* gene in a strain-specific manner reported by Sandberg et al., (2000), could actually be strain-specific splicing.

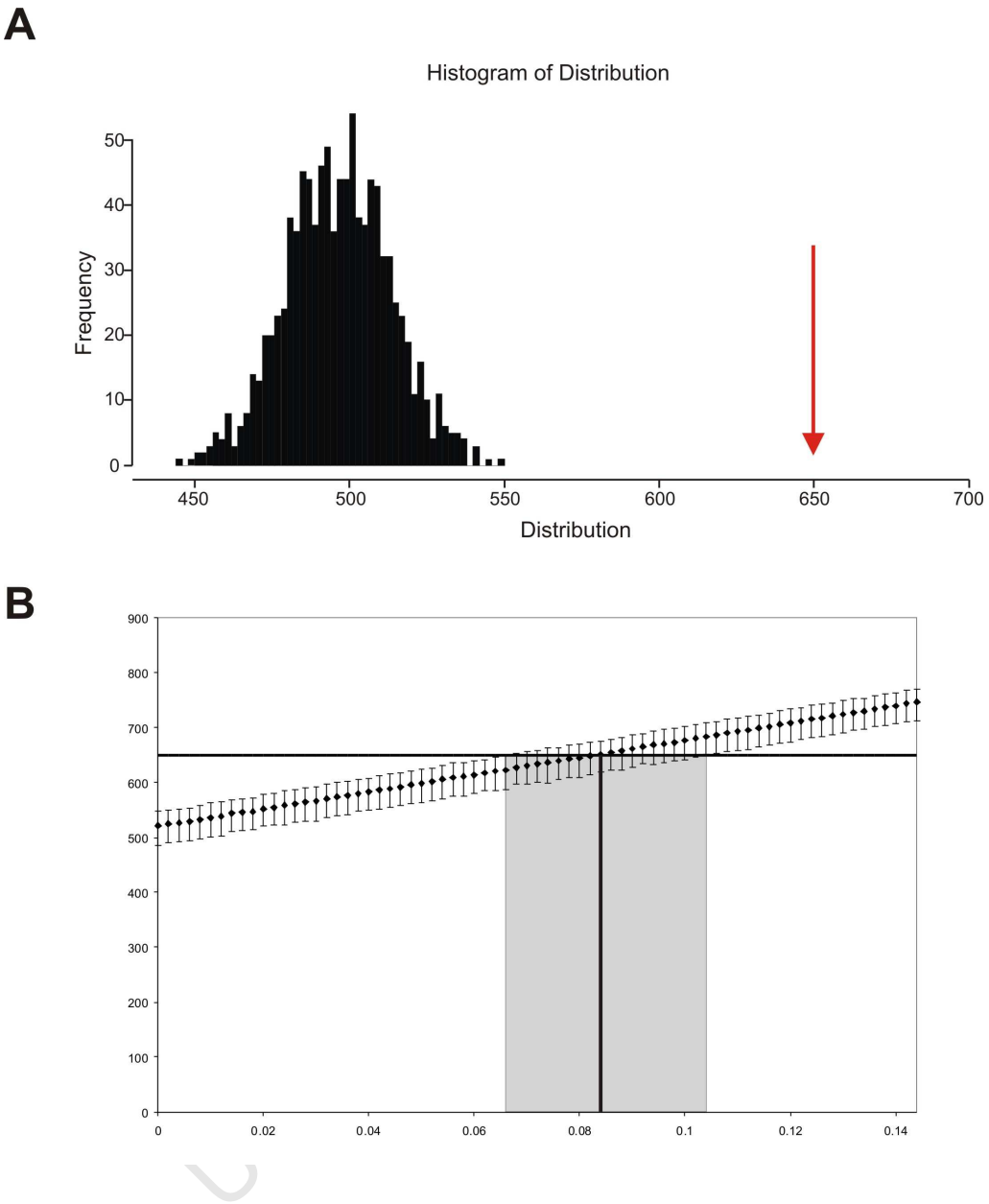


Figure 5: A) A histogram showing the proportion of matrices with a zero cell from 1000 randomized replicates of the dataset. The arrow indicates the number of matrices with a zero cell in the observed data **B)** Average numbers of matrices with a zero cell from 1000 simulated replicates of the dataset in which a proportion of the simulated matrices are derived from qualitative allele-specific transcript isoforms. The horizontal line shows the number of matrices with a zero cell in the observed data. The shaded area shows the confidence interval and the dark line the proportion of qualitative allele-specific isoforms most consistent with the observed number of matrices with a zero cell. Error bars were derived from the simulation as described in the Methods section.

Table 2: List of top candidates detected using the Fishers's Exact test

Unigene Cluster`	Gene name	Uncorrected p-vaue
Mm.293096	Hypothetical gene: Unknown function	0.001262626
Mm.298875	Clathrin, light polypeptide	0.003119719
Mm.211654	Interferon alpha responsive gene	0.004682274
Mm.275720	tRNA methyltransferase 1 homolog (S. cerevisiae)	0.004832414
Mm.298875	Clathrin, light polypeptide	0.005228114
Mm.298875	Clathrin, light polypeptide	0.005847953
Mm.273098	Agrin	0.00595748
Mm.276255	Retired	0.008138144
Mm.25779	Splicing factor 3a, subunit 3	0.009185171
Mm.141276	Breast cancer antigen	0.010013526
Mm.141276	Breast cancer antigen	0.010013526
Mm.29424	RIKEN cDNA 2810012G03	0.010989011
Mm.140761	DnaJ (Hsp40) homolog, subfamily C, member 5	0.011437908
Mm.9239	Zinc finger protein	0.011904762
Mm.22519	Ribokinase	0.011904762
Mm.12967	Inhibitor of kappaB kinase gamma	0.012254902
Mm.215034	Matrin 3	0.013894152
Mm.45367	Zinc finger protein 715	0.013986014
Mm.259567	Mediator of RNA polymerase II transcription, subunit 8 homolog (yeast)	0.014048531
Mm.45602	Bromodomain containing 8	0.015870781
Mm.8687	Adenylate cyclase-associated protein 1 (yeast)	0.016676048
Mm.276255	Transmembrane protein 134	0.016685206
Mm.290791	NADH dehydrogenase (ubiquinone) Fe-S protein 1	0.017722011
Mm.374824	Retired	0.018181818

Table 3: A list of published examples of strain-specific splicing

Gene Cluster	Mutation and splice site	Sequence Variation and description	Strain-specific splicing	Reference
SCNM1: Splicing modifier. Putative splicing factor Mm.182944	SNP Exon Enhancer Element	SNP (**T**/**C**) C57Bl/6J : **T** C3H: **C**	Exon 6 skipping in C57BL/6J strains	(Buchner et al., 2003)
C4: Complement component 4 (within H-2S) Mm.472690	Retrotransposon Intron	Insertion of B2 sequence into an intron of complement C4 gene in H-2k mice resulting in an abnormally spliced B2/C4 transcript plus low expression levels of wild type C4.	Abnormal splicing	(Pattanakit isakul et al., 1992)
MIPP: mouse IAP-promoted placental gene Mm.1350	Retrotransposon Intron	An *IAP* element in the *MIPP* gene in mouse strains that originated from the Bagg strain.	Abnormal splicing	(Chang-Yeh et al., 1991)
BFSP2: beaded filament structural protein 2, phakinin Mm.335403	Deletion 24 bps deletion from splice site	129/SvPas, 129S4/SvJae, CBA and 101 strains : 24 bps deletion C3H, B6, C3HEI, NMRI, SWR, SEC, C57BL/6, Balb/c, DBA, AKR and 102EI : No deletion	Abnormal splicing	(Sandilands et al., 2004)
ADH4: alcohol dehydrogenase 4 Mm.158750	SNP Creates an early polyadenylation sites	C57BL/6J: C3H :	Truncated transcript in the C3H and related strains	(Dolney et al., 2001)

6.4 Discussion

There are over 100 inbred mouse strains with considerable phenotypic and genetic diversity that have been deposited into the most comprehensive database dedicated to mouse at JAX lab (www.jax.com). The genetic variation between strains, coupled with the high isogenicity within strains make mouse ideal as a model for human disease and genetic studies (Beck et al., 2000; Wade and Daly, 2005). To further enhance the use of mouse for biomedical research, it is imperative to fully understand and document the genetic differences and their functional implications that distinguish mouse strains (Wade et al., 2005). Such efforts are already underway with the establishment of the Mouse Phenome Database in 2001, which catalogues strain-specific variations including "concealed" phenotypic variation such as gene expression (Bogue et al., 2007). However little work has been done to detect or document strain-specific splicing on a large-scale. We report the first exploratory

large-scale estimation of the prevalence of strain-specific splicing in mouse. Our results suggest that approximately 8.4% of mouse alternatively spliced in the ASAPII database are spliced differently between B6 and nonB6 strains.

Lack of large-scale comparative studies of strain-specific mRNA splicing is due in part to the scarcity of a comprehensive collection of complete mouse strain transcriptomes and genomic sequences. However, strain-specific information in public repositories of transcript such as EMBL and dbEST has largely been disregarded, mainly due to lack or poor annotations of data. We have developed a method to categorize transcript data according to mouse inbred strains based on EST-derived SNPs. This allows for most ESTs to be categorized into their possible strain of origin without any strain of origin annotation.

A major advantage of this profiling method is that it provides an indirect means to scan for mutations that are likely to cause splicing differences between strains. Unlike the marker SNP based studies that are based on ascertaining the presence of a cis-acting mutation (Nembaware et al., 2004; Ge et al., 2005a), our analysis allows testing for strain-specific splicing caused by either *cis*-acting or *trans*-acting mutations that need not be in linkage disequilibrium with marker SNPs present on the transcript. There are many different types of mutations that can affect splicing (Table 3 for examples). To have a clear picture of the prevalence of strain specific-splicing one should consider all types of mutations that occur in mouse, which include SNPs as well as, for example, retrotransposons, that occur at a high frequency in mouse. A recent review by Maksakova et al., (2006) highlighted the high rates of endogenous retroviral elements in mouse and the impact they can have on the expression of strain-specific transcripts.

When designing a study of strain-specific splicing or gene expression, it is informative to trace the origins of the mutations along the mouse genealogical trees. Splicing mutations acquired and fixed during the propagation of a mouse inbred founder are likely to be detectable in strains that form the genealogical branches thereafter. For example, the presence of a single long terminal repeat (LTR), in the mouse intracisternal A particle (IAP) promoted placental gene (*IPP*), results in the expression of an abberant *IPP* mRNA isoform detectable only in inbred mice strains

C3H/HeJ and BALB/c (Chang-Yeh et al., 1993) that share a common ancestry of the Bagg's Albino mouse ancestor (Beck et al., 2000). An IAP retrotransposition event that could have occurred in a germline cell of a Bagg' s Albino mice followed by subsequent removal of the IAP is suggested to have left the LTR remnant sequence that causes the observed strain-specific splicing event (Chang-Yeh et al., 1993). Using the SNP profiling method allows us to check for strain-specific mutations that occurred along the genealogies of the C57BL/6J and its related mouse strains.

EST sequences provide transcript information for different mouse strains, but ESTs are sparse for most gene segments except for the 3' and 5' gene ends. For the simulations there is enough data in the matrices to make an estimate of the prevalence of B6 strain-specific splicing (Figure 4). However, for the detection of individual gene candidates, there is insufficient data. The *SCNM1* gene example in Figure 4 demonstrates that the EST data is too sparse to confirm most published cases of strain-specific splicing. An alternative approach would be to perform an analogous study to that performed on the Affymetrix exon array platforms (Chapter 5). However, thus far, high quality exon array data is mainly based on single mouse strains and thus not ideal to study strain-specific splicing (Pan et al., 2004). Furthermore, the equivalent of the human lymphoblastoid cell lines genotyped through the HapMap project (The International HapMap Consortium, 2005) and subsequently profiled using affymetrix exon arrays (Huang et al., 2007) is not available for mouse.

Differences in gene expression between mouse strains are frequently investigated with quantitative considerations of transcriptional efficiency in mind (Cowles et al., 2002; Sandberg et al., 2000) A significant number of studies have measured gene expression variation across strains using microarray platforms. However, array studies are based on probes which are normally randomly primed from EST data (Sandberg et al., 2000). A limitation of EST based probes is that they could be targeting strain-specific splicing events which can potentially be detected on array platforms as differential expression in mouse strains. In this current study, we detected one such gene. In light of this result, we propose that studies of strain-specific expression should consider whether there is evidence of strain-specific splicing in the genes inferred to be subject to strain-specific expression.

A genome-wide study of the prevalence of strain-specific isoforms serves to emphasize to mouse researchers, the importance of detection of all possible mRNA isoforms before the designing a mouse knock-out experiment. Inaccurate prediction of alternatively spliced transcripts could lead to poor experimental design primarily during creation of gene knockout strains. A classic example is a mouse strain designed as a gene knockout strain of an estrogen alpha receptor that had a placental specific mRNA isoform which escaped the knockout (Kol et al., 2005). In the eostrogen alpha receptor gene knockout experiment described above, lack of knowledge of alternative isoforms led to years of work with questionable conclusions which could have been avoided had the researchers been aware of all the transcript isoforms. A similar error occurred in the creation of a knockout strain of the Vitamin D Receptor (*VDR*) gene, a truncated form of the supposedly knocked-out *VDR* gene was discovered recently (Bula et al., 2005).

We have used transcript data to partition transcript as having originated from B6-like strains and nonB6 strains. From our estimates we conclude that there is a high prevalence of B6 specific splicing which suggests a high prevalence of strain-specific splicing in general. A recent study, has led to an influx of over 8 million mouse SNPs in dbSNP (Frazer et al., 2007), which greatly increased the number of mouse SNPs in dbSNPs. With such large amounts of sequence variants, a scan for strain-specific splicing using SNP data and splicing scoring tools such as was performed in Chapter 5, could add value to this area of research. However, this study has provided a basis and guide for further research.

Chapter 7

Conclusion and Future work

Completion of the human genome project and improvements in large-scale transcriptome and genotyping capturing technologies has lead to large amounts of data being deposited into publicly available databases. Publicly available transcriptome data and sequence variants are generated from individuals of diverse genetic backgrounds, and hence offer great opportunities to understand the impact of heritable sequence variants on gene expression variability. How best to make use of the publicly available datasets in deducing the effect of a *cis*-acting mutation on the regulation of any of the mechanisms involved in gene expression, remains an open question.

Several studies have developed algorithms that make use of publicly available genome-wide transcript and sequence data and have generally concluded that *cis*-acting variants contribute significantly to allele-specific expression (see Chapter 1 and Chapter 2). However the influence of genotypes on mRNA splicing patterns is a largely unexplored area of research. Work from this thesis has contributed towards understanding the contribution of heritable genetic variations to mRNA transcript diversity through the detection and characterization of allele-specific splicing using publicly available datasets. Simultaneously, work from this thesis has made a significant contribution to the general characterisation of isoforms generated from alternative and allele-specific mRNA splicing (Figure 1). In this chapter, significant conclusions of this work are highlighted and several avenues for future research discussed.

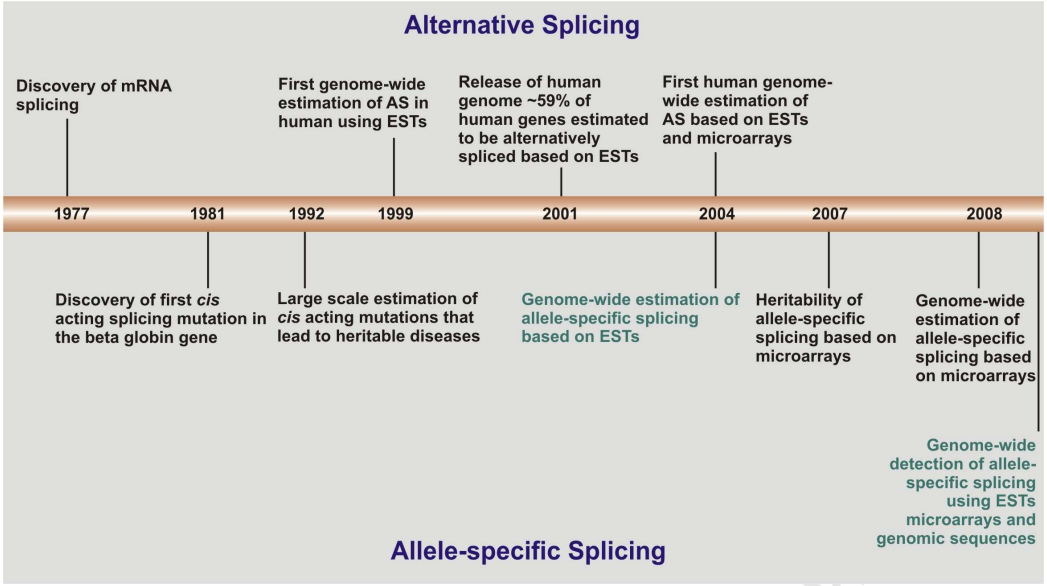


Figure 1: Major contributions in genome-wide detection and estimation of alternatively spliced mRNA isoforms in human. Published contributions from this thesis are highlighted in green.

7.1 High prevalence of allele-specific splicing

Although many attempts have been made to assess the prevalence of allele-specific splicing in human, they are generally biased towards splice variants that cause disease (Lopez-Bigas et al., 2005) or mutations at splice donor and acceptor sites (Krawczak et al., 1992). Work from this thesis was the first to report a genome-wide detection of allele-specific splicing that is not restricted to any disease or to a specific *cis*-regulatory element (Figure 1). Our results in Chapter 4 and 6 indicate a high prevalence of allele-specific splicing in human and mouse and these initial findings were corroborated by the large numbers of individual genes affected by allele-specific splicing reported in Chapter 5.

Such a high prevalence suggests that allele-specific splicing could be a common source of inter-individual transcript variability which ultimately leads to phenotypic differences. Although the functional roles of the allele-specific isoforms may not be obvious from computational analyses alone, we have highlighted the potential impact that allele-specific splicing can have on disease (Chapter 2 and 5). The reported allele-specific splicing candidates are likely to have a large impact on disease and pharmacogenetics and should be investigated further.

7.2 Improvement in the detection of splicing mutations

Splicing mutations when located in coding regions are commonly misclassified as causing phenotypic variations through non-synonymous amino acid changes (Cartegni et al., 2003). The *GLO1* gene discussed in Chapter 5 highlights how such misclassifications are bound to occur. At present, no single source of publicly available datasets can provide information about splicing *cis*-acting mutations and all the resultant allele-specific splicing products. Each source of data has its own advantages as well as disadvantages (Chapter 5). The integrated approach of using three different publicly available datasets (ESTs, microarrays and genomic sequences), produces both putative splicing mutations and their allele-specific splicing products. Therefore we recommend such integrated approaches for more accurate inferences of the impact of mutations on mRNA splicing patterns.

The availability of putative *cis*-acting mutations and their resultant allele-specific mRNA isoforms allows for a less time consuming way for designing minigene assays or any other experiments to validate allele-specific splicing patterns. Therefore, work from this thesis has contributed to the improvement of the detection of splicing mutations by providing an extensive resource that can be used to assess the possible effect on splicing of human polymorphisms located in putative splice-regulatory sites. This allows researchers to make more informed decision when pursuing functional characterisation of mutations that are already associated to disease.

7.3 Novel methods for using publicly available ESTs in allele-specific splicing

Two types of data from large-scale analyses are commonly used; genomic and microarrays and thus computational tools for analysing allele-specific splicing in microarray and genomic datasets are well established. This contrasts sharply with methods for detection of allele-specific splicing based on ESTs since these transcripts have generally been underutilised. A major contribution of this work is the development of two useful methods, a heuristic and maximum likelihood approach that firmly establishes ESTs as a useful data source for the detection of allele-specific splicing.

7.4 Summary of resources

We provide several resources (see Table 1) that can be used to accelerate the investigation of allele-specific splicing as well as other allele-specific gene expression variation. For example the snp2estMap that is presented in Chapter 3 has already been used for studying allele-specific imprinting (Seoighe et al., 2006).

Table 1: Summary of data contributions

Chapter	Resource	Input data	Location
3	snp2estmap	ESTs	http://mancala.cbio.uct.ac.za/splicing
5	Allele-specific isoforms and putative rSNPs	ESTs Microarrays Genomic	http://mancala.cbio.uct.ac.za/splicing

7.5 Future work

7.5.1 Expansion of the snp2estmap database

The snp2estmap database presented in Chapter 3 only contains human data. Expansion of this resource to include SNPs mapped to ESTs of all the highly represented multi-exon organisms in dbEST would greatly accelerate the use of ESTs in the detection of allele-specific splicing across phyla.

7.5.2 Maximum Likelihood Models

The significance of the statistical models developed in a maximum likelihood framework extends beyond the detection of allele-specific splicing, to the detection of allele-specific expression (Seoighe et al., 2006). The maximum likelihood models could also easily be adjusted to detect allele-specific splicing that occurs in a context specific manner such as in a tissue-specific (Cowles et al., 2002) or disease specific manner (Wang and Cooper, 2007). Work is already underway to analyze allele-specific splicing that occurs in a cancer-specific manner, based on the snp2estmap dataset and the maximum likelihood models developed in this study.

The detection of allele-specific splicing in Chapter 4, 5 and 6 is based on the assumption that SNP markers used for the detection of allele-specific splicing are in linkage disequilibrium with *cis*-acting splicing mutations. However, this assumption was not been tested. Based on haplotype blocks that have already been characterized by the HapMap project (The International HapMap Consortium, 2005), statistical models could be developed that estimate the likelihood of marker SNPs being in linkage disequilibrium with putative *cis*-splicing mutations. For future work, such an approach, if applied to results from this study would greatly substantiate and add value to this current analysis and to any other linkage disequilibrium based detection studies.

7.5.2 Further functional characterization of allele-specific isoforms

An exciting and promising direction for further research is characterization of the direct impact of allele-specific splicing patterns identified from this study on human phenotypic variation. Putative *cis*-acting variants discovered in this study that have already been associated to disease, cancers and pharmacogenetics in publicly available database such as the PharmGKB database (Altman, 2007) and the Human Gene Mutation Disease Database (Stenson et al., 2003) would be an ideal starting point. Knowledge of allele-specific splicing patterns caused by mutations already associated to disease could significantly enhance the design and development of more effective therapies.

There are several reports of splicing factors that are modifiers to disease. Some of the allele-specific splicing candidates discovered (see Chapter 5) are involved in the regulation of splicing. Future work that focuses of characterising allele-specific splicing of splicing regulators could be of great medical significance.

Although the computational and microarray approaches presented in this thesis are invaluable for identifying putative srSNPs of medical significance, they are associated with several shortcomings. Work is currently underway by a collaborating group to validate some of the novel candidates presented in this study through the use of RT-PCR and qRT-PCR to confirm the computational results presented in this study.

7.6 Concluding remarks

The medical impact of allele-specific splicing is apparent from the ever-increasing number of genetic diseases and pharmacogenetic effects that are linked to splicing defects (Faustino and Cooper, 2003; Wang and Cooper, 2007). A thorough description and cataloguing of all human genotypes and the quantitative and qualitative effect they exert on splicing would be a powerful resource for understanding human genetic diseases and phenotypes. The role of computational based analysis of publicly available genome-wide datasets promises to be increasingly important towards this undertaking.

Bibliography

Affymetrix. Affymetrix Papers: Exon Probeset Annotations and Transcript Cluster Groupings v1.0; Exon Array Background Correction v1.0; Guide to Probe Logarithmic Intensity Error (PLIER) Estimation; Alternative Transcript Analysis Methods for Exon Arrays v1.1. http://www.affymetrix.com/support/technical/whitepapers.affx . 2007.

Altman,R.B. (2007). PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. Nat. Genet. *39*, 426.

Aouacheria,A., Navratil,V., Barthelaix,A., Mouchiroud,D., and Gautier,C. (2006). Bioinformatic screening of human ESTs for differentially expressed genes in normal and tumor tissues. BMC. Genomics *7*, 94.

Aouacheria,A., Navratil,V., Lopez-Perez,R., Gutierrez,N.C., Churkin,A., Barash,D., Mouchiroud,D., and Gautier,C. (2007). In silico whole-genome screening for cancer-related single-nucleotide polymorphisms located in human mRNA untranslated regions. BMC. Genomics *8*, 2.

Arenas,M., Duley,J., Sumi,S., Sanderson,J., and Marinaki,A. (2007). The ITPA c.94C>A and g.IVS2+21A>C sequence variants contribute to missplicing of the ITPA gene. Biochim. Biophys. Acta *1772*, 96-102.

Ars,E., Serra,E., de la,L.S., Estivill,X., and Lazaro,C. (2000). Cold shock induces the insertion of a cryptic exon in the neurofibromatosis type 1 (NF1) mRNA. Nucleic Acids Res. *28*, 1307-1312.

Baek,D. and Green,P. (2005). Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. Proc. Natl. Acad. Sci. U. S. A *102*, 12813-12818.

Baralle,D. and Baralle,M. (2005). Splicing in action: assessing disease causing sequence changes. J. Med. Genet. *42*, 737-748.

Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M., and Edgar,R. (2007). NCBI GEO: mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res. *35*, D760-D765.

Beck,J.A., Lloyd,S., Hafezparast,M., Lennon-Pierce,M., Eppig,J.T., Festing,M.F., and Fisher,E.M. (2000). Genealogies of mouse inbred strains. Nat. Genet. *24*, 23-25.

Berget,S.M. (1995). Exon recognition in vertebrate splicing. J. Biol. Chem. *270*, 2411-2414.

Bhasi,A., Pandey,R.V., Utharasamy,S.P., and Senapathy,P. (2007). EuSplice: a unified resource for the analysis of splice signals and alternative splicing in eukaryotic genes. Bioinformatics. *23*, 1815-1823.

Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T., Down,T., Durbin,R., Fernandez-Suarez,X.M., Flicek,P., Graf,S., Hammond,M., Herrero,J., Howe,K., Iyer,V., Jekosch,K., Kahari,A., Kasprzyk,A., Keefe,D., Kokocinski,F., Kulesha,E., London,D., Longden,I., Melsopp,C., Meidl,P., Overduin,B., Parker,A., Proctor,G., Prlic,A., Rae,M., Rios,D., Redmond,S., Schuster,M., Sealy,I., Searle,S., Severin,J., Slater,G., Smedley,D., Smith,J., Stabenau,A., Stalker,J., Trevanion,S., Ureta-Vidal,A., Vogel,J., White,S., Woodwark,C., and Hubbard,T.J. (2006). Ensembl 2006. Nucleic Acids Res. *34*, D556-D561.

Black,D.L. (2000). Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell *103*, 367-370.

Blagitko,N., Mergenthaler,S., Schulz,U., Wollmann,H.A., Craigen,W., Eggermann,T., Ropers,H.H., and Kalscheuer,V.M. (2000). Human GRB10 is imprinted and expressed from the paternal and maternal allele in a highly tissue- and isoform-specific fashion. Hum. Mol. Genet. *9*, 1587-1595.

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S., and Schneider,M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. *31*, 365-370.

Bogue,M.A., Grubb,S.C., Maddatu,T.P., and Bult,C.J. (2007). Mouse Phenome Database (MPD). Nucleic Acids Res. *35*, D643-D649.

Bonaldo,M.F., Lennon,G., and Soares,M.B. (1996). Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res. *6*, 791-806.

Bonizzoni,P., Rizzi,R., and Pesole,G. (2006). Computational methods for alternative splicing prediction. Brief. Funct. Genomic. Proteomic. *5*, 46-51.

Bonnevie-Nielsen,V., Field,L.L., Lu,S., Zheng,D.J., Li,M., Martensen,P.M., Nielsen,T.B., Beck-Nielsen,H., Lau,Y.L., and Pociot,F. (2005). Variation in antiviral 2',5'-oligoadenylate synthetase (2'5'AS) enzyme activity is controlled by a single-nucleotide polymorphism at a splice-acceptor site in the OAS1 gene. Am. J. Hum. Genet. *76*, 623-633.

Bracco,L. and Kearsey,J. (2003). The relevance of alternative RNA splicing to pharmacogenomics. Trends Biotechnol. *21*, 346-353.

Brett,D., Hanke,J., Lehmann,G., Haase,S., Delbruck,S., Krueger,S., Reich,J., and Bork,P. (2000). EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. FEBS Lett. *474*, 83-86.

Brett,D., Pospisil,H., Valcarcel,J., Reich,J., and Bork,P. (2002). Alternative splicing and genome complexity. Nat. Genet. *30*, 29-30.

Buchner,D.A., Trudeau,M., and Meisler,M.H. (2003). SCNM1, a putative RNA splicing factor that modifies disease severity in mice. Science *301*, 967-969.

Buchroithner,B., Klausegger,A., Ebschner,U., Anton-Lamprecht,I., Pohla-Gubo,G., Lanschuetzer,C.M., Laimer,M., Hintner,H., and Bauer,J.W. (2004). Analysis of the LAMB3 gene in a junctional epidermolysis bullosa patient reveals exonic splicing and allele-specific nonsense-mediated mRNA decay. Lab Invest *84*, 1279-1288.

Buetow,K.H., Edmonson,M.N., and Cassidy,A.B. (1999). Reliable identification of large numbers of candidate SNPs from public EST data. Nat. Genet. *21*, 323-325.

Burnette,J.M., Miyamoto-Sato,E., Schaub,M.A., Conklin,J., and Lopez,A.J. (2005). Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. Genetics *170*, 661-674.

Camargo,A.A., Samaia,H.P., Dias-Neto,E., Simao,D.F., Migotto,I.A., Briones,M.R., Costa,F.F., Nagai,M.A., Verjovski-Almeida,S., Zago,M.A., Andrade,L.E., Carrer,H., El Dorry,H.F., Espreafico,E.M., Habr-Gama,A., Giannella-Neto,D., Goldman,G.H., Gruber,A., Hackel,C., Kimura,E.T., Maciel,R.M., Marie,S.K., Martins,E.A., Nobrega,M.P., Paco-Larson,M.L., Pardini,M.I., Pereira,G.G., Pesquero,J.B., Rodrigues,V., Rogatto,S.R., da Silva,I.D., Sogayar,M.C., Sonati,M.F., Tajara,E.H., Valentini,S.R., Alberto,F.L., Amaral,M.E., Aneas,I., Arnaldi,L.A., de Assis,A.M., Bengtson,M.H., Bergamo,N.A., Bombonato,V., de Camargo,M.E., Canevari,R.A., Carraro,D.M., Cerutti,J.M., Correa,M.L., Correa,R.F., Costa,M.C., Curcio,C., Hokama,P.O., Ferreira,A.J., Furuzawa,G.K., Gushiken,T., Ho,P.L., Kimura,E., Krieger,J.E., Leite,L.C., Majumder,P., Marins,M., Marques,E.R., Melo,A.S., Melo,M.B., Mestriner,C.A., Miracca,E.C., Miranda,D.C., Nascimento,A.L., Nobrega,F.G., Ojopi,E.P., Pandolfi,J.R., Pessoa,L.G., Prevedel,A.C., Rahal,P., Rainho,C.A., Reis,E.M., Ribeiro,M.L., Da Ros,N., de Sa,R.G., Sales,M.M., Sant'anna,S.C., dos Santos,M.L., da Silva,A.M., da Silva,N.P., Silva,W.A., Jr., da Silveira,R.A., Sousa,J.F., Stecconi,D., Tsukumo,F., Valente,V., Soares,F., Moreira,E.S., Nunes,D.N., Correa,R.G., Zalcberg,H., Carvalho,A.F., Reis,L.F., Brentani,R.R., Simpson,A.J., and de Souza,S.J. (2001). The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. Proc. Natl. Acad. Sci. U. S. A *98*, 12103-12108.

Carothers,A.M., Urlaub,G., Grunberger,D., and Chasin,L.A. (1993). Splicing mutants and their second-site suppressors at the dihydrofolate reductase locus in Chinese hamster ovary cells. Mol. Cell Biol. *13*, 5085-5098.

Cartegni,L., Chew,S.L., and Krainer,A.R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat. Rev. Genet. *3*, 285-298.

Cartegni,L. and Krainer,A.R. (2003). Correction of disease-associated exon skipping by synthetic exon-specific activators. Nat. Struct. Biol. *10*, 120-125.

Cartegni,L., Wang,J., Zhu,Z., Zhang,M.Q., and Krainer,A.R. (2003). ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Res. *31*, 3568-3571.

Castle,J., Garrett-Engele,P., Armour,C.D., Duenwald,S.J., Loerch,P.M., Meyer,M.R., Schadt,E.E., Stoughton,R., Parrish,M.L., Shoemaker,D.D., and Johnson,J.M. (2003). Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing. Genome Biol. *4*, R66.

Chalfant,C.E., Mischak,H., Watson,J.E., Winkler,B.C., Goodnight,J., Farese,R.V., and Cooper,D.R. (1995). Regulation of alternative splicing of protein kinase C beta by insulin. J. Biol. Chem. *270*, 13326-13332.

Chang-Yeh,A., Mold,D.E., and Huang,R.C. (1991). Identification of a novel murine IAP-promoted placenta-expressed gene. Nucleic Acids Res. *19*, 3667-3672.

Chern,T.M., Paul,N., van Nimwegen,E., and Zavolan,M. (2008). Computational Analysis of Full-length cDNAs Reveals Frequent Coupling Between Transcriptional and Splicing Programs. DNA Res. *15*, 63-72.

Chern,T.M., van Nimwegen,E., Kai,C., Kawai,J., Carninci,P., Hayashizaki,Y., and Zavolan,M. (2006). A simple physical model predicts small exon length variations. PLoS. Genet. *2*, e45.

Churbanov,A., Rogozin,I.B., Deogun,J.S., and Ali,H. (2006). Method of predicting splice sites based on signal interactions. Biol. Direct. *1*, 10.

Clark,T.A., Sugnet,C.W., and Ares,M., Jr. (2002). Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. Science *296*, 907-910.

Coleman,T.P. and Roesser,J.R. (1998). RNA secondary structure: an important cis-element in rat calcitonin/CGRP pre-messenger RNA splicing. Biochemistry *37*, 15941-15950.

Colot,H.V., Loros,J.J., and Dunlap,J.C. (2005). Temperature-modulated alternative splicing and promoter use in the Circadian clock gene frequency. Mol. Biol. Cell *16*, 5563-5571.

Conde,L., Vaquerizas,J.M., Dopazo,H., Arbiza,L., Reumers,J., Rousseau,F., Schymkowitz,J., and Dopazo,J. (2006). PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. Nucleic Acids Res. *34*, W621-W625.

Cowles,C.R., Hirschhorn,J.N., Altshuler,D., and Lander,E.S. (2002). Detection of regulatory variation in mouse genes. Nat. Genet. *32*, 432-437.

Crick,F. (1970). Central dogma of molecular biology. Nature *227*, 561-563.

Daoud,R., Mies,G., Smialowska,A., Olah,L., Hossmann,K.A., and Stamm,S. (2002). Ischemia induces a translocation of the splicing factor tra2-beta 1 and changes alternative splicing patterns in the brain. J. Neurosci. *22*, 5889-5899.

de Souza,S.J., Camargo,A.A., Briones,M.R., Costa,F.F., Nagai,M.A., Verjovski-Almeida,S., Zago,M.A., Andrade,L.E., Carrer,H., El Dorry,H.F., Espreafico,E.M., Habr-Gama,A., Giannella-Neto,D., Goldman,G.H., Gruber,A., Hackel,C., Kimura,E.T., Maciel,R.M., Marie,S.K., Martins,E.A., Nobrega,M.P., Paco-

Larson,M.L., Pardini,M.I., Pereira,G.G., Pesquero,J.B., Rodrigues,V., Rogatto,S.R., da Silva,I.D., Sogayar,M.C., de Fatima,S.M., Tajara,E.H., Valentini,S.R., Acencio,M., Alberto,F.L., Amaral,M.E., Aneas,I., Bengtson,M.H., Carraro,D.M., Carvalho,A.F., Carvalho,L.H., Cerutti,J.M., Correa,M.L., Costa,M.C., Curcio,C., Gushiken,T., Ho,P.L., Kimura,E., Leite,L.C., Maia,G., Majumder,P., Marins,M., Matsukuma,A., Melo,A.S., Mestriner,C.A., Miracca,E.C., Miranda,D.C., Nascimento,A.N., Nobrega,F.G., Ojopi,E.P., Pandolfi,J.R., Pessoa,L.G., Rahal,P., Rainho,C.A., Da Ros,N., de Sa,R.G., Sales,M.M., da Silva,N.P., Silva,T.C., da,S.W., Jr., Simao,D.F., Sousa,J.F., Stecconi,D., Tsukumo,F., Valente,V., Zalcbeg,H., Brentani,R.R., Reis,F.L., Dias-Neto,E., and Simpson,A.J. (2000). Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. Proc. Natl. Acad. Sci. U. S. A *97*, 12690-12693.

Deutsch,S., Iseli,C., Bucher,P., Antonarakis,S.E., and Scott,H.S. (2001). A cSNP map and database for human chromosome 21. Genome Res. *11*, 300-307.

Deutsch,S.I., Rosse,R.B., Mastropaolo,J., Long,K.D., and Gaskins,B.L. (2008). Epigenetic therapeutic strategies for the treatment of neuropsychiatric disorders: ready for prime time? Clin. Neuropharmacol. *31*, 104-119.

Dewey,C.N., Rogozin,I.B., and Koonin,E.V. (2006). Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. BMC. Genomics *7*, 311.

Dolney,D.E., Szalai,G., Duester,G., and Felder,M.R. (2001). Molecular analysis of genetic differences among inbred mouse strains controlling tissue expression pattern of alcohol dehydrogenase 4. Gene *267*, 145-156.

Eng,L., Coutinho,G., Nahas,S., Yeo,G., Tanouye,R., Babaei,M., Dork,T., Burge,C., and Gatti,R.A. (2004). Nonclassical splicing mutations in the coding and noncoding regions of the ATM Gene: maximum entropy estimates of splice junction strengths. Hum. Mutat. *23*, 67-76.

Eyras,E., Caccamo,M., Curwen,V., and Clamp,M. (2004). ESTGenes: alternative splicing from ESTs in Ensembl. Genome Res. *14*, 976-987.

Fackenthal,J.D., Cartegni,L., Krainer,A.R., and Olopade,O.I. (2002). BRCA2 T2722R is a deleterious allele that causes exon skipping. Am. J. Hum. Genet. *71*, 625-631.

Fairbrother,W.G., Holste,D., Burge,C.B., and Sharp,P.A. (2004). Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS. Biol. *2*, E268.

Fairbrother,W.G., Yeh,R.F., Sharp,P.A., and Burge,C.B. (2002). Predictive identification of exonic splicing enhancers in human genes. Science *297*, 1007-1013.

Faustino,N.A. and Cooper,T.A. (2003). Pre-mRNA splicing and human disease. Genes Dev. *17*, 419-437.

Favre,M., Butticaz,C., Stevenson,B., Jongeneel,C.V., and Telenti,A. (2003). High frequency of alternative splicing of human genes participating in the HIV-1 life cycle: a model using TSG101, betaTrCP, PPIA, INI1, NAF1, and PML. J. Acquir. Immune. Defic. Syndr. *34*, 127-133.

Fededa,J.P., Petrillo,E., Gelfand,M.S., Neverov,A.D., Kadener,S., Nogues,G., Pelisch,F., Baralle,F.E., Muro,A.F., and Kornblihtt,A.R. (2005). A polar mechanism coordinates different regions of alternative splicing within a single gene. Mol. Cell *19*, 393-404.

Field,L.L., Bonnevie-Nielsen,V., Pociot,F., Lu,S., Nielsen,T.B., and Beck-Nielsen,H. (2005). OAS1 splice site polymorphism controlling antiviral enzyme activity influences susceptibility to type 1 diabetes. Diabetes *54*, 1588-1591.

Ford,L.P., Bagga,P.S., and Wilusz,J. (1997). The poly(A) tail inhibits the assembly of a 3'-to-5' exonuclease in an in vitro RNA stability system. Mol. Cell Biol. *17*, 398-406.

Fox-Walsh,K.L., Dou,Y., Lam,B.J., Hung,S.P., Baldi,P.F., and Hertel,K.J. (2005). The architecture of pre-mRNAs affects mechanisms of splice-site pairing. Proc. Natl. Acad. Sci. U. S. A *102*, 16176-16181.

Frazer,K.A., Eskin,E., Kang,H.M., Bogue,M.A., Hinds,D.A., Beilharz,E.J., Gupta,R.V., Montgomery,J., Morenzoni,M.M., Nilsen,G.B., Pethiyagoda,C.L., Stuve,L.L., Johnson,F.M., Daly,M.J., Wade,C.M., and Cox,D.R. (2007). A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature *448*, 1050-1053.

Freimuth,R.R., Stormo,G.D., and McLeod,H.L. (2005). PolyMAPr: programs for polymorphism database mining, annotation, and functional analysis. Hum. Mutat. *25*, 110-117.

Frischmeyer,P.A., van Hoof,A., O'Donnell,K., Guerrerio,A.L., Parker,R., and Dietz,H.C. (2002). An mRNA surveillance mechanism that eliminates transcripts lacking termination codons. Science *295*, 2258-2261.

Ge,B., Gurd,S., Gaudin,T., Dore,C., Lepage,P., Harmsen,E., Hudson,T.J., and Pastinen,T. (2005). Survey of allelic expression using EST mining. Genome Res. *15*, 1584-1591.

Gilbert,W. (1978). Why genes in pieces? Nature *271*, 501.

Graveley,B.R. (2001). Alternative splicing: increasing diversity in the proteomic world. Trends Genet. *17*, 100-107.

Graveley,B.R., Hertel,K.J., and Maniatis,T. (1998). A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. EMBO J. *17*, 6747-6756.

Green,R.E., Lewis,B.P., Hillman,R.T., Blanchette,M., Lareau,L.F., Garnett,A.T., Rio,D.C., and Brenner,S.E. (2003). Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. Bioinformatics. *19 Suppl 1*, i118-i121.

Guryev,V., Berezikov,E., and Cuppen,E. (2005). CASCAD: a database of annotated candidate single nucleotide polymorphisms associated with expressed sequences. BMC. Genomics *6*, 10.

Hatton,A.R., Subramaniam,V., and Lopez,A.J. (1998). Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. Mol. Cell *2*, 787-796.

Hawken,R.J., Barris,W.C., McWilliam,S.M., and Dalrymple,B.P. (2004). An interactive bovine in silico SNP database (IBISS). Mamm. Genome *15*, 819-827.

Hayashizaki,Y. (2003). RIKEN mouse genome encyclopedia. Mech. Ageing Dev. *124*, 93-102.

Hayes,B.J., Nilsen,K., Berg,P.R., Grindflek,E., and Lien,S. (2007). SNP detection exploiting multiple sources of redundancy in large EST collections improves validation rates. Bioinformatics. *23*, 1692-1693.

He,H., Olesnanik,K., Nagy,R., Liyanarachchi,S., Prasad,M.L., Stratakis,C.A., Kloos,R.T., and de la,C.A. (2005). Allelic variation in gene expression in thyroid tissue. Thyroid *15*, 660-667.

Hiller,M., Backofen,R., Heymann,S., Busch,A., Glaesser,T.M., and Freytag,J.C. (2004b). Efficient prediction of alternative splice forms using protein domain homology. In Silico. Biol. *4*, 195-208.

Hiller,M., Huse,K., Szafranski,K., Jahn,N., Hampe,J., Schreiber,S., Backofen,R., and Platzer,M. (2004a). Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. Nat. Genet. *36*, 1255-1257.

Hiller,M., Huse,K., Szafranski,K., Jahn,N., Hampe,J., Schreiber,S., Backofen,R., and Platzer,M. (2006). Single-Nucleotide Polymorphisms in NAGNAG Acceptors Are Highly Predictive for Variations of Alternative Splicing. Am. J. Hum. Genet. *78*, 291-302.

Hinrichs,A.S., Karolchik,D., Baertsch,R., Barber,G.P., Bejerano,G., Clawson,H., Diekhans,M., Furey,T.S., Harte,R.A., Hsu,F., Hillman-Jackson,J., Kuhn,R.M., Pedersen,J.S., Pohl,A., Raney,B.J., Rosenbloom,K.R., Siepel,A., Smith,K.E., Sugnet,C.W., Sultan-Qurraie,A., Thomas,D.J., Trumbower,H., Weber,R.J., Weirauch,M., Zweig,A.S., Haussler,D., and Kent,W.J. (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. *34*, D590-D598.

Holste,D., Huo,G., Tung,V., and Burge,C.B. (2006). HOLLYWOOD: a comparative relational database of alternative splicing. Nucleic Acids Res. *34*, D56-D62.

Holste,D. and Ohler,U. (2008). Strategies for identifying RNA splicing regulatory motifs and predicting alternative splicing events. PLoS. Comput. Biol. *4*, e21.

Homma,K., Kikuno,R.F., Nagase,T., Ohara,O., and Nishikawa,K. (2004). Alternative splice variants encoding unstable protein domains exist in the human brain. J. Mol. Biol. *343*, 1207-1220.

Huang,H.D., Horng,J.T., Lee,C.C., and Liu,B.J. (2003). ProSplicer: a database of putative alternative splicing information derived from protein, mRNA and expressed sequence tag sequence data. Genome Biol. *4*, R29.

Huang,H.D., Horng,J.T., Lin,F.M., Chang,Y.C., and Huang,C.C. (2005). SpliceInfo: an information repository for mRNA alternative splicing in human genome. Nucleic Acids Res. *33*, D80-D85.

Huang,R.S., Duan,S., Bleibel,W.K., Kistner,E.O., Zhang,W., Clark,T.A., Chen,T.X., Schweitzer,A.C., Blume,J.E., Cox,N.J., and Dolan,M.E. (2007). A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. Proc. Natl. Acad. Sci. U. S. A *104*, 9758-9763.

Huang,Y.H., Chen,Y.T., Lai,J.J., Yang,S.T., and Yang,U.C. (2002). PALS db: Putative Alternative Splicing database. Nucleic Acids Res. *30*, 186-190.

Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T., Durbin,R., Eyras,E., Gilbert,J., Hammond,M., Huminiecki,L., Kasprzyk,A., Lehvaslaiho,H., Lijnzaad,P., Melsopp,C., Mongin,E., Pettett,R., Pocock,M., Potter,S., Rust,A., Schmidt,E., Searle,S., Slater,G., Smith,J., Spooner,W., Stabenau,A., Stalker,J., Stupka,E., Ureta-Vidal,A., Vastrik,I., and Clamp,M. (2002). The Ensembl genome database project. Nucleic Acids Res. *30*, 38-41.

Hull,J., Campino,S., Rowlands,K., Chan,M.S., Copley,R.R., Taylor,M.S., Rockett,K., Elvidge,G., Keating,B., Knight,J., and Kwiatkowski,D. (2007). Identification of Common Genetic Variation That Modulates Alternative Splicing. PLoS. Genet. *3*, e99.

Hunt,A.G., Xu,R., Addepalli,B., Rao,S., Forbes,K.P., Meeks,L.R., Xing,D., Mo,M., Zhao,H., Bandyopadhyay,A., Dampanaboina,L., Marion,A., Von Lanken,C., and Li,Q.Q. (2008). Arabidopsis mRNA polyadenylation machinery: comprehensive analysis of protein-protein interactions and gene expression profiling. BMC. Genomics *9*, 220.

Huntley,D., Baldo,A., Johri,S., and Sergot,M. (2006). SEAN: SNP prediction and display program utilizing EST sequence clusters. Bioinformatics. *22*, 495-496.

Ihaka,R. and Gentleman,R. (1996). R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics *5*, 299-314.

Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M., Yura,K., Miyazaki,S., Ikeo,K., Homma,K., Kasprzyk,A., Nishikawa,T., Hirakawa,M., Thierry-Mieg,J., Thierry-Mieg,D., Ashurst,J., Jia,L., Nakao,M., Thomas,M.A., Mulder,N., Karavidopoulou,Y., Jin,L., Kim,S., Yasuda,T., Lenhard,B., Eveno,E., Suzuki,Y., Yamasaki,C., Takeda,J., Gough,C., Hilton,P., Fujii,Y., Sakai,H., Tanaka,S., Amid,C., Bellgard,M., Bonaldo,M.F., Bono,H., Bromberg,S.K., Brookes,A.J., Bruford,E., Carninci,P., Chelala,C., Couillault,C., de Souza,S.J., Debily,M.A., Devignes,M.D., Dubchak,I., Endo,T., Estreicher,A., Eyras,E., Fukami-Kobayashi,K., Gopinath,G.R., Graudens,E., Hahn,Y., Han,M., Han,Z.G., Hanada,K., Hanaoka,H., Harada,E., Hashimoto,K., Hinz,U., Hirai,M., Hishiki,T., Hopkinson,I., Imbeaud,S., Inoko,H., Kanapin,A., Kaneko,Y., Kasukawa,T., Kelso,J., Kersey,P., Kikuno,R., Kimura,K., Korn,B., Kuryshev,V., Makalowska,I., Makino,T., Mano,S., Mariage-Samson,R., Mashima,J., Matsuda,H., Mewes,H.W., Minoshima,S., Nagai,K., Nagasaki,H., Nagata,N., Nigam,R., Ogasawara,O., Ohara,O., Ohtsubo,M., Okada,N., Okido,T.,

Oota,S., Ota,M., Ota,T., Otsuki,T., Piatier-Tonneau,D., Poustka,A., Ren,S.X., Saitou,N., Sakai,K., Sakamoto,S., Sakate,R., Schupp,I., Servant,F., Sherry,S., Shiba,R., Shimizu,N., Shimoyama,M., Simpson,A.J., Soares,B., Steward,C., Suwa,M., Suzuki,M., Takahashi,A., Tamiya,G., Tanaka,H., Taylor,T., Terwilliger,J.D., Unneberg,P., Veeramachaneni,V., Watanabe,S., Wilming,L., Yasuda,N., Yoo,H.S., Stodolsky,M., Makalowski,W., Go,M., Nakai,K., Takagi,T., Kanehisa,M., Sakaki,Y., Quackenbush,J., Okazaki,Y., Hayashizaki,Y., Hide,W., Chakraborty,R., Nishikawa,K., Sugawara,H., Tateno,Y., Chen,Z., Oishi,M., Tonellato,P., Apweiler,R., Okubo,K., Wagner,L., Wiemann,S., Strausberg,R.L., Isogai,T., Auffray,C., Nomura,N., Gojobori,T., and Sugano,S. (2004). Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS. Biol. *2*, e162.

Irizarry,K., Kustanovich,V., Li,C., Brown,N., Nelson,S., Wong,W., and Lee,C.J. (2000). Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. Nat. Genet. *26*, 233-236.

Itoh,H., Washio,T., and Tomita,M. (2004). Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes. RNA. *10*, 1005-1018.

Jacobsen,M., Hoffmann,S., Cepok,S., Stei,S., Ziegler,A., Sommer,N., and Hemmer,B. (2002). A novel mutation in PTPRC interferes with splicing and alters the structure of the human CD45 molecule. Immunogenetics *54*, 158-163.

Johnson,J.M., Castle,J., Garrett-Engele,P., Kan,Z., Loerch,P.M., Armour,C.D., Santos,R., Schadt,E.E., Stoughton,R., and Shoemaker,D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science *302*, 2141-2144.

Junaid,M.A., Kowal,D., Barua,M., Pullarkat,P.S., Sklower,B.S., and Pullarkat,R.K. (2004). Proteomic studies identified a single nucleotide polymorphism in glyoxalase I as autism susceptibility factor. Am. J. Med. Genet. A *131*, 11-17.

Kan,Z., Castle,J., Johnson,J.M., and Tsinoremas,N.F. (2004). Detection of novel splice forms in human and mouse using cross-species approach. Pac. Symp. Biocomput. 42-53.

Kan,Z., Rouchka,E.C., Gish,W.R., and States,D.J. (2001). Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res. *11*, 889-900.

Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J., Weber,R.J., Haussler,D., and Kent,W.J. (2003). The UCSC Genome Browser Database. Nucleic Acids Res. *31*, 51-54.

Kelso,J., Visagie,J., Theiler,G., Christoffels,A., Bardien,S., Smedley,D., Otgaar,D., Greyling,G., Jongeneel,C.V., McCarthy,M.I., Hide,T., and Hide,W. (2003). eVOC: a controlled vocabulary for unifying gene expression data. Genome Res. *13*, 1222-1230.

Khan,S.G., Muniz-Medina,V., Shahlavi,T., Baker,C.C., Inui,H., Ueda,T., Emmert,S., Schneider,T.D., and Kraemer,K.H. (2002). The human XPC DNA repair gene: arrangement, splice site information content and influence of a single nucleotide polymorphism in a splice acceptor site on alternative splicing and function. Nucleic Acids Res. *30*, 3624-3631.

Kim,N., Alekseyenko,A.V., Roy,M., and Lee,C. (2007). The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. Nucleic Acids Res. *35*, D93-D98.

Kim,N., Shin,S., and Lee,S. (2005). ECgene: genome-based EST clustering and gene modeling for alternative splicing. Genome Res. *15*, 566-576.

Knight,J.C. (2004). Allele-specific gene expression uncovered. Trends Genet. *20*, 113-116.

Kol,G., Lev-Maor,G., and Ast,G. (2005). Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation. Hum. Mol. Genet. *14*, 1559-1568.

Kornblihtt,A.R. (2005). Promoter usage and alternative splicing. Curr. Opin. Cell Biol. *17*, 262-268.

Kornblihtt,A.R. (2007). Coupling transcription and alternative splicing. Adv. Exp. Med. Biol. *623*, 175-189.

Kornblihtt,A.R., Vibe-Pedersen,K., and Baralle,F.E. (1984). Human fibronectin: cell specific alternative mRNA splicing generates polypeptide chains differing in the number of internal repeats. Nucleic Acids Res. *12*, 5853-5868.

Kota,R., Rudd,S., Facius,A., Kolesov,G., Thiel,T., Zhang,H., Stein,N., Mayer,K., and Graner,A. (2003). Snipping polymorphisms from large EST collections in barley (Hordeum vulgare L.). Mol. Genet. Genomics *270*, 24-33.

Kralovicova,J., Gaunt,T.R., Rodriguez,S., Wood,P.J., Day,I.N., and Vorechovsky,I. (2006a). Variants in the human insulin gene that affect pre-mRNA splicing: is -23HphI a functional single nucleotide polymorphism at IDDM2? Diabetes *55*, 260-264.

Kralovicova,J., Lei,H., and Vorechovsky,I. (2006b). Phenotypic consequences of branch point substitutions. Hum. Mutat. *27*, 803-813.

Krawczak,M., Reiss,J., and Cooper,D.N. (1992). The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. Hum. Genet. *90*, 41-54.

Krawczak,M., Thomas,N.S., Hundrieser,B., Mort,M., Wittig,M., Hampe,J., and Cooper,D.N. (2006). Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. Hum. Mutat.

Kriventseva,E.V., Koch,I., Apweiler,R., Vingron,M., Bork,P., Gelfand,M.S., and Sunyaev,S. (2003). Increase of functional diversity by alternative splicing. Trends Genet. *19*, 124-128.

Kumar,S. and Hedges,S.B. (1998). A molecular timescale for vertebrate evolution. Nature *392*, 917-920.

Kunne,C., Lange,M., Funke,T., Miehe,H., Thiel,T., Grosse,I., and Scholz,U. (2005). CR-EST: a resource for crop ESTs. Nucleic Acids Res. *33*, D619-D621.

Kwan,T., Benovoy,D., Dias,C., Gurd,S., Provencher,C., Beaulieu,P., Hudson,T.J., Sladek,R., and Majewski,J. (2008). Genome-wide analysis of transcript isoform variation in humans. Nat. Genet. *40*, 225-231.

Kwan,T., Benovoy,D., Dias,C., Gurd,S., Serre,D., Zuzan,H., Clark,T.A., Schweitzer,A., Staples,M.K., Wang,H., Blume,J.E., Hudson,T.J., Sladek,R., and Majewski,J. (2007). Heritability of alternative splicing in the human genome. Genome Res. *17*, 1210-1218.

Kwok,P.Y. (2001). Genomics. Genetic association by whole-genome analysis? Science *294*, 1669-1670.

Ladd,A.N. and Cooper,T.A. (2002). Finding signals that regulate alternative splicing in the post-genomic era. Genome Biol. *3*, reviews0008.

Lamba,J.K., Adachi,M., Sun,D., Tammur,J., Schuetz,E.G., Allikmets,R., and Schuetz,J.D. (2003). Nonsense mediated decay downregulates conserved alternatively spliced ABCC4 transcripts bearing nonsense codons. Hum. Mol. Genet. *12*, 99-109.

Landry,J.R., Mager,D.L., and Wilhelm,B.T. (2003). Complex controls: the role of alternative promoters in mammalian genomes. Trends Genet. *19*, 640-648.

Lareau,L.F., Inada,M., Green,R.E., Wengrod,J.C., and Brenner,S.E. (2007). Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature *446*, 926-929.

Lee,C. and Roy,M. (2004). Analysis of alternative splicing with microarrays: successes and challenges. Genome Biol. *5*, 231.

Lee,C. and Wang,Q. (2005). Bioinformatics analysis of alternative splicing. Brief. Bioinform. *6*, 23-33.

Lewis,B.P., Green,R.E., and Brenner,S.E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc. Natl. Acad. Sci. U. S. A *100*, 189-192.

Li,S., Ma,L., Li,H., Vang,S., Hu,Y., Bolund,L., and Wang,J. (2007). Snap: an integrated SNP annotation platform. Nucleic Acids Res. *35*, D707-D710.

Libri,D., Stutz,F., McCarthy,T., and Rosbash,M. (1995). RNA structural patterns and splicing: molecular basis for an RNA-based enhancer. RNA. *1*, 425-436.

Ligtenberg,M.J., Gennissen,A.M., Vos,H.L., and Hilkens,J. (1991). A single nucleotide polymorphism in an exon dictates allele dependent differential splicing of episialin mRNA. Nucleic Acids Res. *19*, 297-301.

Liu,H.X., Cartegni,L., Zhang,M.Q., and Krainer,A.R. (2001). A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. Nat. Genet. *27*, 55-58.

Liu,H.X., Chew,S.L., Cartegni,L., Zhang,M.Q., and Krainer,A.R. (2000). Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. Mol. Cell Biol. *20*, 1063-1071.

Liu,H.X., Zhang,M., and Krainer,A.R. (1998). Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev. *12*, 1998-2012.

Lo,H.S., Wang,Z., Hu,Y., Yang,H.H., Gere,S., Buetow,K.H., and Lee,M.P. (2003). Allelic variation in gene expression is common in the human genome. Genome Res. *13*, 1855-1862.

Lopez,A.J. (1998). Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. Annu. Rev. Genet. *32*, 279-305.

Lopez-Bigas,N., Audit,B., Ouzounis,C., Parra,G., and Guigo,R. (2005). Are splicing mutations the most frequent cause of hereditary disease? FEBS Lett. *579*, 1900-1903.

Lynch,K.W. and Weiss,A. (2001). A CD45 polymorphism associated with multiple sclerosis disrupts an exonic splicing silencer. J. Biol. Chem. *276*, 24341-24347.

Majewski,J. and Ott,J. (2002). Distribution and characterization of regulatory elements in the human genome. Genome Res. *12*, 1827-1836.

Makalowski,W. and Boguski,M.S. (1998). Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc. Natl. Acad. Sci. U. S. A *95*, 9407-9412.

Makalowski,W., Zhang,J., and Boguski,M.S. (1996). Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. Genome Res. *6*, 846-857.

Maksakova,I.A., Romanish,M.T., Gagnier,L., Dunn,C.A., van de Lagemaat,L.N., and Mager,D.L. (2006). Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS. Genet. *2*, e2.

Malde,K., Coward,E., and Jonassen,I. (2005). A graph based algorithm for generating EST consensus sequences. Bioinformatics. *21*, 1371-1375.

Maquat,L.E. (2002). Molecular biology. Skiing toward nonstop mRNA decay. Science *295*, 2221-2222.

Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K., Voss,N., Stegmaier,P., Lewicki-

Potapov,B., Saxel,H., Kel,A.E., and Wingender,E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. *34*, D108-D110.

McCullough,A.J. and Berget,S.M. (1997). G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. Mol. Cell Biol. *17*, 4562-4571.

McManus,J.F., Begley,C.G., Sassa,S., and Ratnaike,S. (1996). Five new mutations in the uroporphyrinogen decarboxylase gene identified in families with cutaneous porphyria. Blood *88*, 3589-3600.

Megy,K., Audic,S., and Claverie,J.M. (2002). Heart specific genes revealed by EST sampling. Genome Biol. *3*, REPRINT0008.

Minovitsky,S., Gee,S.L., Schokrpur,S., Dubchak,I., and Conboy,J.G. (2005). The splicing regulatory element, UGCAUG, is phylogenetically and spatially conserved in introns that flank tissue-specific alternative exons. Nucleic Acids Res. *33*, 714-724.

Miriami,E., Sperling,J., and Sperling,R. (1994). Heat shock affects 5' splice site selection, cleavage and ligation of CAD pre-mRNA in hamster cells, but not its packaging in InRNP particles. Nucleic Acids Res. *22*, 3084-3091.

Mironov,A.A., Fickett,J.W., and Gelfand,M.S. (1999). Frequent alternative splicing of human genes. Genome Res. *9*, 1288-1293.

Modrek,B. and Lee,C. (2002). A genomic view of alternative splicing. Nat. Genet. *30*, 13-19.

Modrek,B., Resch,A., Grasso,C., and Lee,C. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res. *29*, 2850-2859.

Morley,M., Molony,C.M., Weber,T.M., Devlin,J.L., Ewens,K.G., Spielman,R.S., and Cheung,V.G. (2004). Genetic analysis of genome-wide variation in human gene expression. Nature *430*, 743-747.

Mossner,R., Henneberg,A., Schmitt,A., Syagailo,Y.V., Grassle,M., Hennig,T., Simantov,R., Gerlach,M., Riederer,P., and Lesch,K.P. (2001). Allelic variation of serotonin transporter expression is associated with depression in Parkinson's disease. Mol. Psychiatry *6*, 350-352.

Mossner,R. and Riederer,P. (2007). Allelic variation of a functional promoter polymorphism of the serotonin transporter and depression in Parkinson's disease. Parkinsonism. Relat Disord. *13*, 62.

Nagao,K., Togawa,N., Fujii,K., Uchikawa,H., Kohno,Y., Yamada,M., and Miyashita,T. (2005). Detecting tissue-specific alternative splicing and disease-associated aberrant splicing of the PTCH gene with exon junction microarrays. Hum. Mol. Genet. *14*, 3379-3388.

Nagaraj,S.H., Gasser,R.B., and Ranganathan,S. (2007). A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief. Bioinform. *8*, 6-21.

Nembaware,V., Lupindo,B., Schouest,K., Spillane,C., Scheffler,K., and Seoighe,C. (2008). Genome-wide survey of allele-specific splicing in humans. BMC. Genomics *9*, 265.

Nembaware,V., Wolfe,K.H., Bettoni,F., Kelso,J., and Seoighe,C. (2004). Allele-specific transcript isoforms in human. FEBS Lett. *577*, 233-238.

Niswender,C.M. (1998). Recent advances in mammalian RNA editing. Cell Mol. Life Sci. *54*, 946-964.

Ott,S., Tamada,Y., Bannai,H., Nakai,K., and Miyano,S. (2003). Intrasplicing--analysis of long intron sequences. Pac. Symp. Biocomput. 339-350.

Pagani,F., Raponi,M., and Baralle,F.E. (2005). Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc. Natl. Acad. Sci. U. S. A *102*, 6368-6372.

Pan,Q., Saltzman,A.L., Kim,Y.K., Misquitta,C., Shai,O., Maquat,L.E., Frey,B.J., and Blencowe,B.J. (2006). Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. Genes Dev. *20*, 153-158.

Pan,Q., Shai,O., Misquitta,C., Zhang,W., Saltzman,A.L., Mohammad,N., Babak,T., Siu,H., Hughes,T.R., Morris,Q.D., Frey,B.J., and Blencowe,B.J. (2004). Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Mol. Cell *16*, 929-941.

Pant,P.V., Tao,H., Beilharz,E.J., Ballinger,D.G., Cox,D.R., and Frazer,K.A. (2006). Analysis of allelic differential expression in human white blood cells. Genome Res. *16*, 331-339.

Pantel,J., Machinis,K., Sobrier,M.L., Duquesnoy,P., Goossens,M., and Amselem,S. (2000). Species-specific alternative splice mimicry at the growth hormone receptor locus revealed by the lineage of retroelements during primate evolution. J. Biol. Chem. *275*, 18664-18669.

Pastinen,T., Ge,B., Gurd,S., Gaudin,T., Dore,C., Lemire,M., Lepage,P., Harmsen,E., and Hudson,T.J. (2005). Mapping common regulatory variants to human haplotypes. Hum. Mol. Genet. *14*, 3963-3971.

Pastinen,T., Raitio,M., Lindroos,K., Tainola,P., Peltonen,L., and Syvanen,A.C. (2000). A system for specific, high-throughput genotyping by allele-specific primer extension on microarrays. Genome Res. *10*, 1031-1042.

Pastinen,T., Sladek,R., Gurd,S., Sammak,A., Ge,B., Lepage,P., Lavergne,K., Villeneuve,A., Gaudin,T., Brandstrom,H., Beck,A., Verner,A., Kingsley,J., Harmsen,E., Labuda,D., Morgan,K., Vohl,M.C., Naumova,A.K., Sinnett,D., and Hudson,T.J. (2004). A survey of genetic and epigenetic variation affecting human gene expression. Physiol Genomics *16*, 184-193.

Pattanakitsakul,S., Zheng,J.H., Natsuume-Sakai,S., Takahashi,M., and Nonaka,M. (1992). Aberrant splicing caused by the insertion of the B2 sequence into an intron of the complement C4 gene is the basis for low C4 production in H-2k mice. J. Biol. Chem. *267*, 7814-7820.

Pertea,M., Lin,X., and Salzberg,S.L. (2001). GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res. *29*, 1185-1190.

Peters,L.L., Robledo,R.F., Bult,C.J., Churchill,G.A., Paigen,B.J., and Svenson,K.L. (2007). The mouse as a model for human biology: a resource guide for complex trait analysis. Nat. Rev. Genet. *8*, 58-69.

Petkov,P.M., Cassell,M.A., Sargent,E.E., Donnelly,C.J., Robinson,P., Crew,V., Asquith,S., Haar,R.V., and Wiles,M.V. (2004a). Development of a SNP genotyping panel for genetic monitoring of the laboratory mouse. Genomics *83*, 902-911.

Petkov,P.M., Ding,Y., Cassell,M.A., Zhang,W., Wagner,G., Sargent,E.E., Asquith,S., Crew,V., Johnson,K.A., Robinson,P., Scott,V.E., and Wiles,M.V. (2004b). An efficient SNP system for mouse genome scanning and elucidating strain relationships. Genome Res. *14*, 1806-1811.

Picoult-Newberg,L., Ideker,T.E., Pohl,M.G., Taylor,S.L., Donaldson,M.A., Nickerson,D.A., and Boyce-Jacino,M. (1999). Mining SNPs from EST databases. Genome Res. *9*, 167-174.

Pirisi,A., Piredda,G., Papoff,C.M., Di Salvo,R., Pintus,S., Garro,G., Ferranti,P., and Chianese,L. (1999). Effects of sheep alpha s1-casein CC, CD and DD genotypes on milk composition and cheesemaking properties. J. Dairy Res. *66*, 409-419.

Platzer,M., Hiller,M., Szafranski,K., Jahn,N., Hampe,J., Schreiber,S., Backofen,R., and Huse,K. (2007). Sequencing errors or SNPs at splice-acceptor guanines in dbSNP? Nature Biotechnology *24*, 1068-1070.

Politi,P., Minoretti,P., Falcone,C., Martinelli,V., and Emanuele,E. (2006). Association analysis of the functional Ala111Glu polymorphism of the glyoxalase I gene in panic disorder. Neurosci. Lett. *396*, 163-166.

Press,W.H., Flannery,B.P., Teukolsky,S.A., and Vetterling,W.T. (1992). *Numerical Recipies in C: The Art of Scientic Computing*.

Purcell,D.F. and Martin,M.A. (1993). Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. J. Virol. *67*, 6365-6378.

Raponi,M., Baralle,F.E., and Pagani,F. (2007). Reduced splicing efficiency induced by synonymous substitutions may generate a substrate for natural selection of new splicing isoforms: the case of CFTR exon 12. Nucleic Acids Res. *35*, 606-613.

Reese,M.G., Eeckman,F.H., Kulp,D., and Haussler,D. (1997). Improved splice site detection in Genie. J. Comput. Biol. *4*, 311-323.

Rezvani,M., Barrans,J.D., Dai,K.S., and Liew,C.C. (2000). Apoptosis-related genes expressed in cardiovascular development and disease: an EST approach. Cardiovasc. Res. *45*, 621-629.

Rezvani,M. and Liew,C.C. (2000). Role of the adenomatous polyposis coli gene product in human cardiac development and disease. J. Biol. Chem. *275*, 18470-18475.

Robberson,B.L., Cote,G.J., and Berget,S.M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol. Cell Biol. *10*, 84-94.

Rockman,M.V. and Wray,G.A. (2002). Abundant raw material for cis-regulatory evolution in humans. Mol. Biol. Evol. *19*, 1991-2004.

Romano,M., Marcucci,R., Buratti,E., Ayala,Y.M., Sebastio,G., and Baralle,F.E. (2002). Regulation of 3' splice site selection in the 844ins68 polymorphism of the cystathionine Beta -synthase gene. J. Biol. Chem. *277*, 43821-43829.

Sacco,R., Papaleo,V., Hager,J., Rousseau,F., Moessner,R., Militerni,R., Bravaccio,C., Trillo,S., Schneider,C., Melmed,R., Elia,M., Curatolo,P., Manzi,B., Pascucci,T., Puglisi-Allegra,S., Reichelt,K.L., and Persico,A.M. (2007). Case-control and family-based association studies of candidate genes in autistic disorder and its endophenotypes: TPH2 and GLO1. BMC. Med. Genet. *8*, 11.

Sambrook,J. (1977). Adenovirus amazes at Cold Spring Harbor. Nature *268*, 101-104.

Sandberg,R., Yasuda,R., Pankratz,D.G., Carter,T.A., Del Rio,J.A., Wodicka,L., Mayford,M., Lockhart,D.J., and Barlow,C. (2000). Regional and strain-specific gene expression mapping in the adult mouse brain. Proc. Natl. Acad. Sci. U. S. A *97*, 11038-11043.

Sandilands,A., Wang,X., Hutcheson,A.M., James,J., Prescott,A.R., Wegener,A., Pekny,M., Gong,X., and Quinlan,R.A. (2004). Bfsp2 mutation found in mouse 129 strains causes the loss of CP49' and induces vimentin-dependent changes in the lens fibre cell cytoskeleton. Exp. Eye Res. *78*, 875-889.

Savas,S., Tuzmen,S., and Ozcelik,H. (2006). Human SNPs resulting in premature stop codons and protein truncation. Hum. Genomics *2*, 274-286.

Schmucker,D., Clemens,J.C., Shu,H., Worby,C.A., Xiao,J., Muda,M., Dixon,J.E., and Zipursky,S.L. (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell *101*, 671-684.

Seeburg,P.H., Higuchi,M., and Sprengel,R. (1998). RNA editing of brain glutamate receptor channels: mechanism and physiology. Brain Res. Brain Res. Rev. *26*, 217-229.

Seoighe,C., Nembaware,V., and Scheffler,K. (2006). Maximum likelihood inference of imprinting and allele-specific expression from EST data. Bioinformatics *22*, 3032-3039

Sharp,P.A. (1994). Split genes and RNA splicing. Cell *77*, 805-815.

Shapiro,M.B. and Senapathy,P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. *15*, 7155-7174.

Shemesh,R., Novik,A., Edelheit,S., and Sorek,R. (2006). Genomic fossils as a snapshot of the human transcriptome. Proc. Natl. Acad. Sci. U. S. A *103*, 1364-1369.

Shen,Y., Ji,G., Haas,B.J., Wu,X., Zheng,J., Reese,G.J., and Li,Q.Q. (2008). Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation. Nucleic Acids Res. *36*, 3150-3161.

Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M., and Sirotkin,K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308-311.

Shoemaker,D.D., Schadt,E.E., Armour,C.D., He,Y.D., Garrett-Engele,P., McDonagh,P.D., Loerch,P.M., Leonardson,A., Lum,P.Y., Cavet,G., Wu,L.F., Altschuler,S.J., Edwards,S., King,J., Tsang,J.S., Schimmack,G., Schelter,J.M., Koch,J., Ziman,M., Marton,M.J., Li,B., Cundiff,P., Ward,T., Castle,J., Krolewski,M., Meyer,M.R., Mao,M., Burchard,J., Kidd,M.J., Dai,H., Phillips,J.W., Linsley,P.S., Stoughton,R., Scherer,S., and Boguski,M.S. (2001). Experimental annotation of the human genome using microarray technology. Nature *409*, 922-927.

Smith,C.W. and Valcarcel,J. (2000). Alternative pre-mRNA splicing: the logic of combinatorial control. Trends Biochem. Sci. *25*, 381-388.

Sorek,R. and Ast,G. (2003). Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. Genome Res. *13*, 1631-1637.

Sorek,R., Shemesh,R., Cohen,Y., Basechess,O., Ast,G., and Shamir,R. (2004). A non-EST-based method for exon-skipping prediction. Genome Res. *14*, 1617-1623.

Spatz,S.J. and Silva,R.F. (2007). Sequence determination of variable regions within the genomes of gallid herpesvirus-2 pathotypes. Arch. Virol. *152*, 1665-1678.

Spritz,R.A., Jagadeeswaran,P., Choudary,P.V., Biro,P.A., Elder,J.T., deRiel,J.K., Manley,J.L., Gefter,M.L., Forget,B.G., and Weissman,S.M. (1981). Base substitution in an intervening sequence of a beta+-thalassemic human globin gene. Proc. Natl. Acad. Sci. U. S. A *78*, 2455-2459.

Srinivasan,K., Shiue,L., Hayes,J.D., Centers,R., Fitzwater,S., Loewen,R., Edmondson,L.R., Bryant,J., Smith,M., Rommelfanger,C., Welch,V., Clark,T.A., Sugnet,C.W., Howe,K.J., Mandel-Gutfreund,Y., and Ares,M., Jr. (2005). Detection and measurement of alternative splicing using splicing-sensitive microarrays. Methods *37*, 345-359.

Stenson,P.D., Ball,E.V., Mort,M., Phillips,A.D., Shiel,J.A., Thomas,N.S., Abeysinghe,S., Krawczak,M., and Cooper,D.N. (2003). Human Gene Mutation Database (HGMD): 2003 update. Hum. Mutat. *21*, 577-581.

Sterner,D.A., Carlo,T., and Berget,S.M. (1996). Architectural limits on split genes. Proc. Natl. Acad. Sci. U. S. A *93*, 15081-15085.

Storey,J.D. and Tibshirani,R. (2003). Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. U. S. A *100*, 9440-9445.

Sugnet,C.W., Srinivasan,K., Clark,T.A., O'Brien,G., Cline,M.S., Wang,H., Williams,A., Kulp,D., Blume,J.E., Haussler,D., and Ares,M., Jr. (2006). Unusual intron conservation near tissue-regulated exons found by splicing microarrays. PLoS. Comput. Biol. *2*, e4.

Sun,H. and Chasin,L.A. (2000). Multiple splicing defects in an intronic false exon. Mol. Cell Biol. *20*, 6414-6425.

Sunyaev,S., Hanke,J., Brett,D., Aydin,A., Zastrow,I., Lathe,W., Bork,P., and Reich,J. (2000). Individual variation in protein-coding sequences of human genome. Adv. Protein Chem. *54*, 409-437.

Talerico,M. and Berget,S.M. (1994). Intron definition in splicing of small Drosophila introns. Mol. Cell Biol. *14*, 3434-3445.

Tate,S.K., Depondt,C., Sisodiya,S.M., Cavalleri,G.L., Schorge,S., Soranzo,N., Thom,M., Sen,A., Shorvon,S.D., Sander,J.W., Wood,N.W., and Goldstein,D.B. (2005). Genetic predictors of the maximum doses patients receive during clinical use of the anti-epileptic drugs carbamazepine and phenytoin. Proc. Natl. Acad. Sci. U. S. A *102*, 5507-5512.

Tennyson,C.N., Dally,G.Y., Ray,P.N., and Worton,R.G. (1996). Expression of the dystrophin isoform Dp71 in differentiating human fetal myogenic cultures. Hum. Mol. Genet. *5*, 1559-1566.

The International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299-1320.

The R Project for Statistical Computing. http://www.r-project.org.

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661-678.

Thornalley,P.J. (2006). Unease on the role of glyoxalase 1 in high-anxiety-related behaviour. Trends Mol. Med. *12*, 195-199.

Valentonyte,R., Hampe,J., Huse,K., Rosenstiel,P., Albrecht,M., Stenzel,A., Nagy,M., Gaede,K.I., Franke,A., Haesler,R., Koch,A., Lengauer,T., Seegert,D., Reiling,N., Ehlers,S., Schwinger,E., Platzer,M., Krawczak,M., Muller-Quernheim,J., Schurmann,M., and Schreiber,S. (2005). Sarcoidosis is associated with a truncating splice site mutation in BTNL2. Nat. Genet. *37*, 357-364.

Varani,L., Hasegawa,M., Spillantini,M.G., Smith,M.J., Murrell,J.R., Ghetti,B., Klug,A., Goedert,M., and Varani,G. (1999). Structure of tau exon 10 splicing

regulatory element RNA and destabilization by mutations of frontotemporal dementia and parkinsonism linked to chromosome 17. Proc. Natl. Acad. Sci. U. S. A *96*, 8229-8234.

Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A., Gocayne,J.D., Amanatides,P., Ballew,R.M., Huson,D.H., Wortman,J.R., Zhang,Q., Kodira,C.D., Zheng,X.H., Chen,L., Skupski,M., Subramanian,G., Thomas,P.D., Zhang,J., Gabor Miklos,G.L., Nelson,C., Broder,S., Clark,A.G., Nadeau,J., McKusick,V.A., Zinder,N., Levine,A.J., Roberts,R.J., Simon,M., Slayman,C., Hunkapiller,M., Bolanos,R., Delcher,A., Dew,I., Fasulo,D., Flanigan,M., Florea,L., Halpern,A., Hannenhalli,S., Kravitz,S., Levy,S., Mobarry,C., Reinert,K., Remington,K., Abu-Threideh,J., Beasley,E., Biddick,K., Bonazzi,V., Brandon,R., Cargill,M., Chandramouliswaran,I., Charlab,R., Chaturvedi,K., Deng,Z., Di,F., V, Dunn,P., Eilbeck,K., Evangelista,C., Gabrielian,A.E., Gan,W., Ge,W., Gong,F., Gu,Z., Guan,P., Heiman,T.J., Higgins,M.E., Ji,R.R., Ke,Z., Ketchum,K.A., Lai,Z., Lei,Y., Li,Z., Li,J., Liang,Y., Lin,X., Lu,F., Merkulov,G.V., Milshina,N., Moore,H.M., Naik,A.K., Narayan,V.A., Neelam,B., Nusskern,D., Rusch,D.B., Salzberg,S., Shao,W., Shue,B., Sun,J., Wang,Z., Wang,A., Wang,X., Wang,J., Wei,M., Wides,R., Xiao,C., Yan,C., Yao,A., Ye,J., Zhan,M., Zhang,W., Zhang,H., Zhao,Q., Zheng,L., Zhong,F., Zhong,W., Zhu,S., Zhao,S., Gilbert,D., Baumhueter,S., Spier,G., Carter,C., Cravchik,A., Woodage,T., Ali,F., An,H., Awe,A., Baldwin,D., Baden,H., Barnstead,M., Barrow,I., Beeson,K., Busam,D., Carver,A., Center,A., Cheng,M.L., Curry,L., Danaher,S., Davenport,L., Desilets,R., Dietz,S., Dodson,K., Doup,L., Ferriera,S., Garg,N., Gluecksmann,A., Hart,B., Haynes,J., Haynes,C., Heiner,C., Hladun,S., Hostin,D., Houck,J., Howland,T., Ibegwam,C., Johnson,J., Kalush,F., Kline,L., Koduru,S., Love,A., Mann,F., May,D., McCawley,S., McIntosh,T., McMullen,I., Moy,M., Moy,L., Murphy,B., Nelson,K., Pfannkoch,C., Pratts,E., Puri,V., Qureshi,H., Reardon,M., Rodriguez,R., Rogers,Y.H., Romblad,D., Ruhfel,B., Scott,R., Sitter,C., Smallwood,M., Stewart,E., Strong,R., Suh,E., Thomas,R., Tint,N.N., Tse,S., Vech,C., Wang,G., Wetter,J., Williams,S., Williams,M., Windsor,S., Winn-Deen,E., Wolfe,K., Zaveri,J., Zaveri,K., Abril,J.F., Guigo,R., Campbell,M.J., Sjolander,K.V., Karlak,B., Kejariwal,A., Mi,H., Lazareva,B., Hatton,T., Narechania,A., Diemer,K., Muruganujan,A., Guo,N., Sato,S., Bafna,V., Istrail,S., Lippert,R., Schwartz,R., Walenz,B., Yooseph,S., Allen,D., Basu,A., Baxendale,J., Blick,L., Caminha,M., Carnes-Stine,J., Caulk,P., Chiang,Y.H., Coyne,M., Dahlke,C., Mays,A., Dombroski,M., Donnelly,M., Ely,D., Esparham,S., Fosler,C., Gire,H., Glanowski,S., Glasser,K., Glodek,A., Gorokhov,M., Graham,K., Gropman,B., Harris,M., Heil,J., Henderson,S., Hoover,J., Jennings,D., Jordan,C., Jordan,J., Kasha,J., Kagan,L., Kraft,C., Levitsky,A., Lewis,M., Liu,X., Lopez,J., Ma,D., Majoros,W., McDaniel,J., Murphy,S., Newman,M., Nguyen,T., Nguyen,N., and Nodell,M. (2001). The sequence of the human genome. Science *291*, 1304-1351.

Vibe-Pedersen,K., Kornblihtt,A.R., and Baralle,F.E. (1984). Expression of a human alpha-globin/fibronectin gene hybrid generates two mRNAs by alternative splicing. EMBO J. *3*, 2511-2516.

Vorechovsky,I. (2006). Aberrant 3' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. Nucleic Acids Res. *34*, 4630-4641.

Vuillaumier-Barrot,S., Barnier,A., Cuer,M., Durand,G., Grandchamp,B., and Seta,N. (1999). Characterization of the 415G>A (E139K) PMM2 mutation in carbohydrate-deficient glycoprotein syndrome type Ia disrupting a splicing enhancer resulting in exon 5 skipping. Hum. Mutat. *14*, 543-544.

Wade,C.M. and Daly,M.J. (2005). Genetic variation in laboratory mice. Nat. Genet. *37*, 1175-1180.

Wang,G.S. and Cooper,T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. Nat. Rev. Genet. *8*, 749-761.

Wang,J., Chang,Y.F., Hamilton,J.I., and Wilkinson,M.F. (2002). Nonsense-associated altered splicing: a frame-dependent response distinct from nonsense-mediated decay. Mol. Cell *10*, 951-957.

Wang,Z., Rolish,M.E., Yeo,G., Tung,V., Mawson,M., and Burge,C.B. (2004). Systematic identification and analysis of exonic splicing silencers. Cell *119*, 831-845.

Wen,F., Li,F., Xia,H., Lu,X., Zhang,X., and Li,Y. (2004). The impact of very short alternative splicing on protein structures and functions in the human genome. Trends Genet. *20*, 232-236.

Wenstrup,R.J., Langland,G.T., Willing,M.C., D'Souza,V.N., and Cole,W.G. (1996). A splice-junction mutation in the region of COL5A1 that codes for the carboxyl propeptide of pro alpha 1(V) chains results in the gravis form of the Ehlers-Danlos syndrome (type I). Hum. Mol. Genet. *5*, 1733-1736.

Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Geer,L.Y., Kapustin,Y., Khovayko,O., Landsman,D., Lipman,D.J., Madden,T.L., Maglott,D.R., Ostell,J., Miller,V., Pruitt,K.D., Schuler,G.D., Sequeira,E., Sherry,S.T., Sirotkin,K., Souvorov,A., Starchenko,G., Tatusov,R.L., Tatusova,T.A., Wagner,L., and Yaschenko,E. (2007). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. *35*, D5-12.

Wiltshire,T., Pletcher,M.T., Batalov,S., Barnes,S.W., Tarantino,L.M., Cooke,M.P., Wu,H., Smylie,K., Santrosyan,A., Copeland,N.G., Jenkins,N.A., Kalush,F., Mural,R.J., Glynne,R.J., Kay,S.A., Adams,M.D., and Fletcher,C.F. (2003). Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse. Proc. Natl. Acad. Sci. U. S. A *100*, 3380-3385.

Wolfsberg,T.G. and Landsman,D. (1997). A comparison of expressed sequence tags (ESTs) to human genomic sequences. Nucleic Acids Res. *25*, 1626-1632.

multiple transcript isoforms from EST fragment mixtures. Genome Res. *14*, 426-441.

Xu,Q., Modrek,B., and Lee,C. (2002). Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Res. *30*, 3754-3766.

Yang,H.H., Hu,Y., Edmonson,M., Buetow,K., and Lee,M.P. (2003). Computation method to identify differential allelic gene expression and novel imprinted genes. Bioinformatics. *19*, 952-955.

Yeo,G. and Burge,C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377-394.

Yeo,G.W., Nostrand,E.L., and Liang,T.Y. (2007). Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. PLoS. Genet. *3*, e85.

Zavolan,M., Kondo,S., Schonbach,C., Adachi,J., Hume,D.A., Hayashizaki,Y., and Gaasterland,T. (2003). Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. Genome Res. *13*, 1290-1300.

Zavolan,M., van Nimwegen,E., and Gaasterland,T. (2002). Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome. Genome Res. *12*, 1377-1385.

Zhang,C., Hastings,M.L., Krainer,A.R., and Zhang,M.Q. (2007). Dual-specificity splice sites function alternatively as 5' and 3' splice sites. Proc. Natl. Acad. Sci. U. S. A *104*, 15028-15033.

Zhang,X.H., Kangsamaksin,T., Chao,M.S., Banerjee,J.K., and Chasin,L.A. (2005). Exon inclusion is dependent on predictable exonic splicing enhancers. Mol. Cell Biol. *25*, 7323-7332.

Zheng,C.L., Kwon,Y.S., Li,H.R., Zhang,K., Coutinho-Mansfield,G., Yang,C., Nair,T.M., Gribskov,M., and Fu,X.D. (2005). MAASE: an alternative splicing database designed for supporting splicing microarray applications. RNA. *11*, 1767-1776.

Zhu,H., Tucker,H.M., Grear,K.E., Simpson,J.F., Manning,A.K., Cupples,L.A., and Estus,S. (2007). A common polymorphism decreases low-density lipoprotein receptor exon 12 splicing efficiency and associates with increased cholesterol. Hum. Mol. Genet. *16*, 1765-1772.