Prevalence and frequency spectra of single nucleotide polymorphisms at exon-intron junctions of human genes

By

BUKIWE LUPINDO

B.Sc. (Hons) Medical Bioscience. UWC

Thesis submitted to the Faculty of Science, University of Cape Town, in fulfillment for the Degree of Masters of Bioinformatics.

December 2008

Supervisor: Prof Cathal Seoighe

Department of Molecular and Cellular Biology

PLAGIARISM DECLARATION

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.

2. I have used the **Harvard convention** for citation and referencing. Each contribution to, and quotation in, this thesis from the work(s) of other people has been attributed, and has been cited and referenced.

3. This thesis is my own work.

4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.

Signature Signed by candidate

KEYWORDS

Alternative splicing

Allele specific splicing

Alternatively spliced exons

Constitutively spliced exons

Exon-intron junctions

Single nucleotide polymorphisms

Splicing regulatory SNPs

Allele frequency

Derived allele

Natural selection

SUMMARY

Background

In humans and other higher eukaryotes the observation of multiple splice isoforms for a given gene is common. However it is not clear whether all of these alternatively spliced isoforms are a product of true alternative splicing or some are due to DNA sequence variations in human populations. Genetic variations that affect splicing have been shown to cause variation in splicing patterns and potentially are an important source of phenotypic variability among humans. Furthermore, variation in disease susceptibility and manifestation between individuals is often associated with genetic polymorphisms that determine the way in which genes are spliced. Hence, identification of genetic polymorphisms that might affect the way in which pre-mRNAs are spliced is an area of great interest.

Method

We used the human genome and SNP annotations from the Ensembl and dbSNP databases, to study the prevalence of SNPs close to exon-intron junctions of human genes. We focused on SNPs located at exon-intron junctions because cis-acting motifs around the borders of exons and introns are known to be important for pre-mRNA splicing. In addition, we compared the distribution of SNPs at exon-intron junctions of alternatively spliced exons (ASEs) and those located in exon-intron junctions of constitutively spliced exons (CSEs). We further use SNP allele frequency information from HapMap to look at how natural selection has shaped the distribution of SNPs in the exon-intron junctions.

Results

We used 4 736 096 genotyped SNPs from dbSNP of which 47 036 SNPs map to 5' and 3' exon-intron junctions. These SNPs are mostly located in the non-coding intronic regions of the junction. We observed that exonic positions are significantly less diverse with a higher proportion of rare (frequency ≤ 0.1) derived alleles than intronic positions. We also showed that ASE junctions have lower proportions of SNPs than CSE junctions. These results are consistent with

comparative genomic study which suggests increased conservation of sequences flanking ASEs relative to CSEs.

SNP frequency data from three human populations (European, Asian and African) showed lower frequency of non-synonymous SNPs located near exon boundaries relative to synonymous SNPs, non-coding SNPs and SNPs found in the 5' and 3' splice junctions, as expected, reflecting purifying selection against mutations disrupting amino acid sequence of a protein. We also observed an increased frequency of synonymous SNPs close to boundaries of CSEs relative to corresponding regions of ASEs, suggesting stronger purifying selection acting on intron-exon junctions of ASEs.

Conclusion

The frequency in which SNPs occur and the SNP frequency spectra, provide information about the functional constraints acting on genomic sequences. As we expect, therefore, sequences close to the exon boundaries and splice sites had a low rate of occurrence of SNPs. We observed differences in the prevalence and frequency spectra of SNPs close to intron-exon boundaries of CSEs and ASEs and suggest that these differences reflect differences in the selection pressures affecting these two kinds of exons.

AKNOWLEDGEMENTS

The current study was conducted in the computational biology lab at the University of Cape Town. I wish to express my heart felt gratitude to my supervisor and mentor Professor Cathal Seoighe for his patience, academic insight and personal motivation. I am also thankful to my friend and a colleague Victoria Nembaware for innumerable guidance during the lab work and the collaboration that produced tables 2.1 and table 2.4 in Chapter 2.

I also wish to thank the National Bioinformatics Network (NBN) and UCT for funding this project giving me opportunity to experience bioinformatics research. "You have given me a platform to contribute to medical research ". I am also grateful to the wonderful people at SANBI/UWC who were initially involved in my Bioinformatics training.

A special thank you goes to my family. My mother and my sister, Zukiswa; who have taught me patience, courage and perseverance. To my younger brother Xolela, whose ability to use initiative kept me company during the writing of the thesis. Lastly and most importantly, to my heavenly father "With God's help I made it through."

ABBREVIATIONS

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
SNPs	Single nucleotide polymorphisms
srSNP	splicing regulatory SNP
ASEs	Alternatively spliced exons
CSEs	Constitutively spliced exons
SJ	Splice Junction
ESE	Exonic splicing enhancer
ISE	Intronic splicing enhancer
ESS	Exonic splicing silencers
ISS	Intronic splicing silencers
5'	5 prime end
3'	3 prime end
DAF	Derived allele Frequency

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Theoretical and Conceptual Framework

The 3 billion base pair (bp) that make up the human genome encode between 20 000 and 25 000 protein coding genes (Lander *et al.*, 2001; Venter *et al.*, 2001). The protein complement expressed from these genes is, to a large extent, responsible for the biochemical processes carried out by human cells. The different human cells produce many and diverse transcripts which are translated to proteins that carry out the complex biochemical processes responsible for human physiology. Hence, human biological complexity is carried by fewer genes than expected, with expressed sequences far exceeding the number of genes encoding them (Hillier *et al.*, 1996; Velculescu *et al.*, 1999).

In theory, there are many ways in which a relatively small number of genes could be manipulated so as to generate diversity of gene expression products. However, alternative pre-mRNA splicing of genes has been the most often cited mechanism adopted by eukaryotes to increase protein diversity from relative small number of genes (Gravely, 2001; Kan *et al.*, 2001). The mechanism allows generation of differently spliced transcript from a single gene through alternative use of splice sites. When a gene is alternatively spliced, it produces different transcript patterns; consequently, different protein isoforms from a single gene are produced.

Genome-wide estimates of alternatively spliced genes in human indicate that the mechanism is more prevalent than previously thought; with recent studies estimating more than 60% of human genes as alternatively spliced (Modrek *et al.*, 2001; Johnson *et al.*, 2003). Although numerous alternatively spliced genes have been

discovered, a genome-wide understanding of the regulation of the process is still to be achieved.

1.2 The research Problem

Besides alternative splicing, natural genetic variations between individuals also contribute to transcript diversity (Buckland, 2004; Stranger *et al.*, 2005). This allelic variation in the human population contributes to human protein diversity. The most abundant genetic variations in the human genome are single nucleotide polymorphisms or SNPs (Kruglyak & Nickerson, 2001). SNPs are a useful resource in association studies, gene mapping, pharmacogenomics and evolutionary biology. Hence, several studies are aimed at associating SNPs with varied drug treatment response (Legro *et al.*, 2008), disease susceptibility (Prokunina *et al.*, 2002; Miao *et al.*, 2003) and other phenotypic variation among human populations (Oleksiak *et al*, 2002; Bercovich *et al.*, 2006).

The presence of SNPs within splicing regulatory regions have been associated with occurrence of different transcript isoforms of the same gene (Nembaware *et al.*, 2004; Hull *et al.*, 2007; Kwan *et al.*, 2007). Several examples of SNPs have been associated with subtle or severe phenotypic changes in human genes (Stranger *et al.*, 2005; Kwan *et al.*, 2008) and some are implicated in heritable diseases (Krawczak *et al.*, 1992; Kozyrev *et al.*, 2008). Therefore, genome-wide discovery of SNPs that affect splicing has the capacity to accelerate the association of genetic variants to transcript isoform variation between individuals. It has already been estimated that 21% of alternatively spliced genes are affected by polymorphisms that alter splicing (Nembaware *et al.*, 2004), but no detailed investigation has estimated the prevalence of SNPs that might affect the way in which genes are spliced.

1.3 Potential of Bioinformatics Tools

Most prediction of alternative splice events relies on evidence of expression of alternative combinations of exons from a single gene. The availability of the human genome sequence and gene expression technologies (i.e. EST, cDNA, and microarray) enable bioinformatics analysis of alternatively spliced genes. As a result, to date there are numerous databases of alternatively spliced genes which have been created (See Table 2.2), making it possible to obtain a catalogue of alternatively spliced genes, their genomic sequences and different mRNA isoforms.

A subsequent analysis of the human genome project was the identification of genetic variation within and between human populations (Sherry *et al.*, 2001). The millions of SNPs genotyped and deposited in databases (See Table 2.3) provide us with the opportunity to explore genetic variants that are located within splicing regulatory sequences. These are likely to affect pre-mRNA splicing and can be associated with transcript variants which contribute to phenotype diversity among humans. Hence, a genome-wide survey of splicing regulatory SNPs (srSNPs) will support detection and estimation of the prevalence of allele-specific splicing in the human genome.

Presently, it is not clear whether most of the transcript isoforms found in databases of alternatively spliced genes are true alternative splicing events (same allele can result in different splice patterns within an individual) or some arise from SNPs that affect the way in which mRNAs are spliced. However, there has been reliable identification of genetic variations (i.e. SNPs) from human expressed sequences (Picoult-Newberg *et al.*, 1999; Kenneth *et al.*, 1999; Cargill *et al.*, 1999) which support identification of inherited genetic variation associated with transcript variation. Therefore, the available data sets of alternatively spliced genes and SNPs permit studies aimed at determining the extent of transcript variants which are the result of alternative splicing and those that are associated with genetic polymorphisms.

1.4 General Objectives

It is important to identify and estimate the occurrence of SNPs close to exon-intron boundaries because SNPs in this genomic region are likely to alter transcript splice pattern, thus contributing to phenotypic diversity at a molecular level. This allele-specific splicing may be mistaken for alternative splicing,subsequently overestimating the prevalence of alternatively spliced genes in the human genome. Hence, before we can attribute multiple transcript variants of a gene to alternative splicing, it is necessary to distinguish between transcript variants which are the result of true alternative splicing, possibly serving to increase the coding capacity of a gene, and transcript variants resulting from genetic polymorphisms, which may be responsible for some phenotypic variation between individuals. It is against this background that the current investigation aims to establish the prevalence and allele frequencies of SNPs at exon-intron junction regions of human genes.

1.5 Specific Objectives

In response to the need to identify and generate a catalog of genetic polymorphisms that might affect mRNA splicing, this study looked at the distribution of SNPs in the exon-intron junctions of human genes. We focused on SNPs located at exon-intron junction regions because nucleotide sequences which form the borders of introns and exons contain sequence motifs essential for pre-mRNA splicing (Mount, 1982; Shapiro & Senapathy, 1987; Jackson, 1991).

Exons and introns affected by alternative splicing have distinct features compared to constitutively spliced exons (Sorek & Ast, 2003). Hence, we further compare the distribution of SNPs at exon-intron junctions of alternatively spliced exons (ASEs) and constitutively spliced exons (CSEs). This knowledge will facilitate understanding of sequence conservation in the exon-intron junctions between these two exon categories.

Although many of the SNPs occurring at the exon-intron junctions could be neutral mutations in which different genotypes have the same fitness, a fraction of polymorphisms in this region are expected to affect mRNA splicing. SNPs located at pre-mRNA splicing regulatory motifs may affect the way in which genes are spliced, resulting in different splice patterns which may play an important role in phenotypic diversity and/or genetic disorders between individuals (Faustino & Cooper, 2003; Nissim-Rafinia & Kerem, 2002). For example, SNPs disrupting the splice sites and exonic splicing enhancers (ESEs) might be associated with deleterious transcripts and are expected to be eliminated through purifying selection (Eskesen *et al.*, 2004; Parmely *et al.*, 2006).

It is beyond the scope of our analysis to directly study the transcript expression pattern associated with each SNP located at exon-intron junction; however, we examined the effect of polymorphisms on function indirectly by analyzing the inferred effects of natural selection on the polymorphic sites at the exon-intron junction splicing regulatory sites. The aim is to understand the role of natural selection in shaping the distribution of SNPs at the exon-intron junction. Hence, we used SNP allele frequency information to investigate the influence of natural selection on distribution of SNPs at the exon-intron junctions and compare this between alternatively and constitutively spliced exons.

1.6 Thesis outline

The thesis is composed of six chapters and the main issues discussed in each of the chapters are indicated as follows:

Chapter Two provides an overview of the biology of pre-mRNA splicing and alternative splicing. The current understanding and genomic sequences that are involved in regulation of splicing are also outlined. We also looked into computational or bioinformatics techniques involved in the analysis of alternative

splicing. Finally we considered human genetic variations and the influence of SNPs on the way in which genes are spliced.

Chapter Three describes the methods used to generate the human exon and SNP datasets. The datasets were derived from publicly available genomic information in biological databases. A genome-wide analysis of human SNPs and exons was undertaken. We identified exon-intron junctions and characterized the exons as constitutively and alternatively spliced.

Chapter Four focuses on the analysis of the prevalence of SNPs found at exon-intron regions. SNPs found at exon-intron junctions might affect efficiency of splicing. The chapter also describes a comparison of the prevalence of SNPs at exon-intron junctions of CSEs and ASEs.

In Chapter Five, we assess how natural selection has influenced the distribution of SNP allele frequencies at exon-intron junctions. Using population specific SNP frequency information we carried out a qualitative assessment of natural selection by comparing patterns of derived allele frequencies between SNPs located at non-synonymous sites, synonymous sites, non-coding regions and SNPs located in exon-intron junctions.

Discussions of the main trends, patterns, and connections that may emerge from the results are presented in each chapter. The thesis is concluded by presenting, in Chapter Six, a summary of important points, main conclusions and possible implications of our research. Closely tied to general discussion is the discussion of the possible limitations of our methodology.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Human protein coding genes consist of segments of coding sequences interrupted by segments of non-coding sequences. The non-coding sequences, the introns, get transcribed into precursors of messenger RNAs (pre-mRNAs) but they are not translated into proteins. The coding segments, the exons, are actually the ones that are joined together, forming the mature mRNA sequence that gets translated into the amino acid sequence of a protein molecule. An initial focus of this review is a description on how introns are removed from expressed genes and exons are joined together during a pre-mRNA processing reaction referred as pre-mRNA splicing.

Exons from the same gene can be arranged and joined differently during pre- mRNA splicing to produce different mRNA isoforms through alternative pre-mRNA splicing (Smith & Valcarcel, 2000; Smith & Roberts, 2002). Alternative splicing is a mechanism that can increase the protein-coding capacity of eukaryotic genes (Gravely, 2001; Black, 2000; Kan *et al.*, 2001). However, some of the diversity of human gene products might be a result of genetic variation, which may account for some of the phenotypic variation found in the human population. Natural genetic variation has been shown to control differential expression of gene transcript isoform through their affects on pre-mRNA splicing (Nembaware *et al.*, 2004; Hull *et al.*, 2007; Kwan *et al.*, 2007). Hence, apart from a discussion on phenotypic variation within the individual due to alternative splicing, we will also discuss how genetic variation between individuals can also account for variation found in the human transcriptome.

2.2 pre-mRNA Splicing

Splicing is a pre-mRNA modification process that occurs during gene transcription (Gilbert, 1978). The process involves removal of non-coding introns from a gene and joining adjacent exons to form mature RNA (Figure 2.1). The resulting mature RNA (mRNA) contains coding sequences required for translation of a gene to protein. In general, introns tend to be much longer than exons with an average eukaryotic exon less than 200bp long (Hawkin, 1988; Berget, 1995). Accurate removal of the longer introns and constitutive joining of adjacent exons is accomplished through precise identification of the splice sites found at the intron termini.

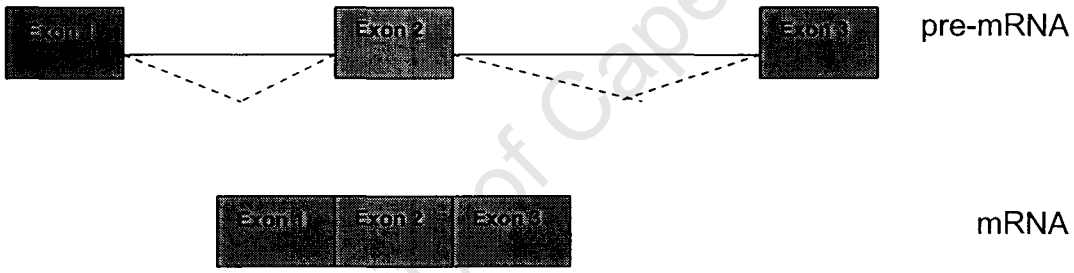

Figure 2.1: Schematic representation of a eukaryotic gene open reading frame (ORF). Exons are shown as boxes and introns as straight lines connecting the exons. The angled doted lines show splice sites where introns are removed from pre-mRNA to produce mRNA.

2.2.1 Mechanisms of pre-mRNA Splicing

2.2.1.1 *Trans*-acting splicing regulatory proteins

The splicing reaction is regulated by the spliceosome, which is a large multi-component protein (Burge *et al.*, 1999; Graveley, 2000). The spliceosome consists of five small nuclear ribonucleoprotein particles (snRNPs). The U1, U2, U4/U6 and U5 snRNP molecules are associated with a large number of proteins, including SR proteins. The spliceosome has two key functions. It identifies the splice sites from

the pre-mRNA and serves as an enzyme which catalyzes the pre-mRNA splicing reaction (Ladd & Cooper, 2002).

SR proteins are splicing factors that are also important in pre-mRNA splicing. These *trans*-regulatory proteins contain an RNA recognition motif (RRM) which binds to distinct pre-mRNA sequences and recruits a number of other splicing factors, including the components of the spliceosome (Wu & Maniatis, 1993; Kohtz *et al.*, 1994; Zuo & Maniatis, 1996). Thus, they promote the assembly of the different inactive components of the spliceosome into a functional and active enzyme (Wu & Maniatis, 1993; Graverly & Maniatis, 1998; Graveley, 2000). It is the snRNP component of the spliceosome, not the SR proteins, which catalyzes the splicing reaction. The SR proteins only recruit the different spliceosome components and assist in activating the enzyme and recognition of splice sites (Norton, 2004; Graveley, 2000).

2.2.1.2 *Cis*-acting splicing regulatory proteins

Although the splicing machinery, the spliceosome, has the role of finding short exon sequences among longer introns, it accomplishes this with high accuracy. Efficient constitutive splicing is achieved through accurate identification of the 5' splice site (at the 5' end of the intron) and 3' splice site (at the 3' end of the intron). It is the recognition of special sequence features found in the junction between the introns and exons by the spliceosome that drives the process. The 5' splice (donor) site and the 3' splice (acceptor) site sequences are key genomic elements required for removal of introns (Mount, 1982; Shapiro & Senapathy, 1987).

The 5' splice site is located at the junction between the upstream exon and the 5' end of an intron whereas the 3' splice site is located at the junction between the 3' end of the intron and the downstream exon (Figure 2.2). Almost all introns are characterized by invariant dinucleotides, the GT at the 5' end and the AG at the 3'

end. In addition, the 3' end of the intron is defined by the branch point located approximately 20 bases from the 3' intron boundary and the polypyrimidine tract (Pyr-rich) near the 3' splice site.

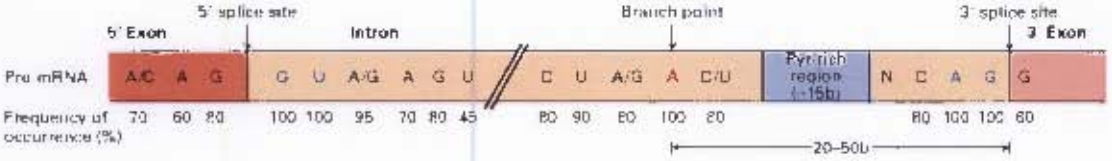

Figure 2.2: *Cis*-acting splicing regulatory sequences. These are located at the junction between an exon and an intron, termed the exon-intron junction. The invariant bases the GU and AG found at the 5' and 3' intron end positions are highlighted in blue. A pyrimidine-rich region (light blue background) near the 3' end of the intron is composed of continuous sequence of approximately 15 uridines. The branch-point adenosine is usually found 20 – 50 bp from the 3' end of the intron. The frequency of the nucleotide is indicated below its position. (Source: Lodish *et al.*, 2003)

The *cis*-acting splicing regulatory sequences are often conserved between vertebrates (Mount, 1982; Burset *et al.*, 2000) and this is an indication of their functional significance. However, these genomic features do not provide sufficient information for recognition of exons and the splice sites located in the exon-intron junctions. These sequences are short and highly degenerate. There are many sequences in the human genome that resemble them, making it difficult to distinguish between real and cryptic splice sites. Besides, mutations in the conserved sequences may occur and this can lead to activation of cryptic splice sites (Treisman *et al.*, 1983; Wieringa *el al.*, 1983). Hence, additional pre-mRNA sequence features are necessary to determine recognition of the correct splice site and ensure efficient pre-mRNA splicing.

In addition to the crucial splice sites, human pre-mRNA sequences contain less conserved sequence features that acts as exonic splicing enhancers (ESE), intronic splicing enhancers (ISE), exonic splicing silencers (ESS) and intronic splicing silencers (ISS) (Baralle & Baralle, 2005). These sequences are found close to the splice sites. The spliceosome makes use of these diverse exonic and intronic splicing enhancer sequences to support recognition of splice site sequence features.

and utilize exonic and intronic splicing silencer sequences to obstruct their recognition (Cartegni, *et al.*, 2002). ESEs in particular are common in most exons. Several studies have identified a range of ESE motifs and concluded that these sequences are short (6-8nt), have high purine content, and sometimes overlap (Liu *et al.*, 1998; Cartegni, *et al.*, 2002; Fairbrother *et al.*, 2002). Besides supporting the recognition of weak splice sites, ESE sequences are binding sites for SR proteins and are highly specific (Graveley, 2000; Blencowe, 2000).

ESS and ISS sequences are less well defined than splicing enhancers. The most described silencers are intronic elements but exonic splicing silencers have also been found (Amendt *et al.*, 1995; Kan & Green, 1999). Obstruction of splice sites by silencers involves binding of heterogeneous ribonucleoprotein particles (hnRNPs) to the silencer sequences (Choi *et al.*, 1986; Lopez, 1998), which then block the activity of SR proteins (Caceres & Krainer, 1997; Smith & Valcarcel, 2000). Binding of hnRNPs to ESS and/or ISS result in inaccessibility of enhancers and splice sites to SR proteins (Del Gatto-Konczak *et al.*, 1999; Cartegni *et al.*, 2002).

.

2.2.1.3 Biochemistry of pre-mRNA splicing

After the recognition of splice sites and the activation of the spliceosome, the pre-mRNA splicing reaction takes place in two trans-esterification reactions (Ruskin *et al.*, 1984; Guthrie, 1989; Lodish *et al.*, 2003). Firstly, the intron is cleaved at the 5' splice site, producing two intermediates product: the upstream exon and an RNA structure composed of the intron and downstream exon called an intron lariat. In the second reaction, the intron is cleaved at the 3' splice site and the two adjacent exons are ligated generating an mRNA, subsequently eliminating the intact intron. In each trans-esterification reaction, one phosphate-ester bond is exchanged for another as shown in Figure 2.3.

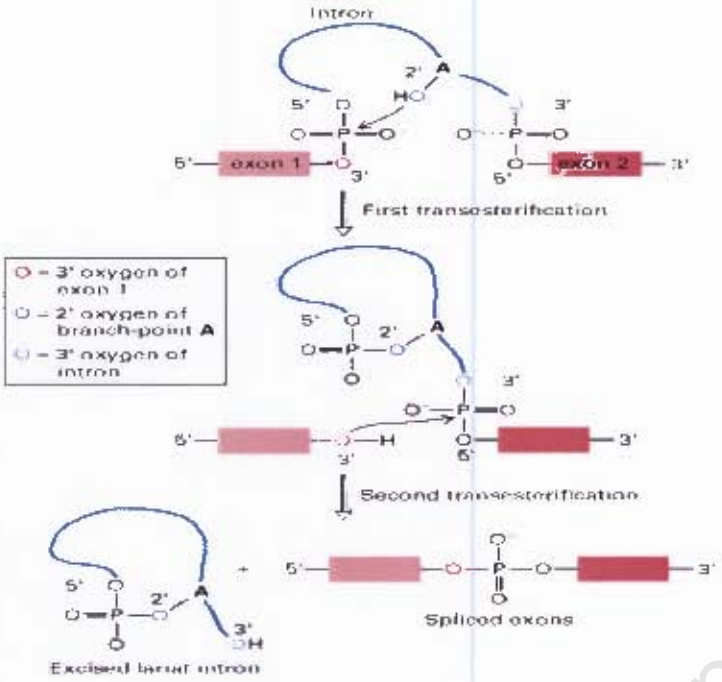

Figure 2.3: Transesterification reactions during pre-mRNA splicing. In the first reaction, the 2'- OH of the branch point adenine (A) residue within the intron acts as a nucleophile and bond with the 5' phosphate of the last nucleoside of the 5' end of the intron. The arrows indicate where the hydroxyl oxygen bond with phosphorus atom. The ester bond between the 5' phosphorus of the intron and the 3' oxygen (red) of exon 1 is exchanged for an ester bond with the 2' oxygen (dark blue) of the branch-point A residue. The first reaction releases exon 1 from the intron. In the second reaction, the ester bond between the 5' phosphorus of exon 2 and the oxygen (light blue) of the 3' nucleoside of the intron is exchanged for an ester bond with the 3' oxygen of exon 1, releasing the intron and joining the two adjacent exons. (Source: Lodish *et al.*, 2003)

2.3 Alternative pre-mRNA splicing

2.3.1 Alternative splicing in the human genome

During pre-mRNA splicing different choices of splice site pairs can be used allowing exons that are arranged and joined in different spliced mRNA patterns, a process referred as alternative pre-mRNA splicing. Figure 2.4 show how introns can be removed and differently spliced exon patterns are joined together through recognition of different splice sites by the spliceosome. During alternative pre-mRNA splicing, an exon may be skipped, shortened, extended or an intron may be retained in the final mRNA product. Intron retention involves failure to remove an intron during pre-mRNA splicing. However, for an intron to be retained it must be properly encoding for amino acids that can be incorporated into the translated protein. Intron retention is one of the rare forms of alternative splice patterns (Ladd & Cooper, 2002) whereas exon skipping is very common in humans (Modrek & Lee, 2003; Yeo *et al.*, 2004).

Although constitutively spliced exons outnumber alternatively spliced exons (Clark & Thanaraj, 2002; Sorek *et al.*, 2006), computational analysis of available transcript sequences infer that more than half of human protein coding genes have one or more exons that are alternatively spliced (Modrek *et al.*, 2001; Johnson *et al.*, 2003). When a gene is alternatively spliced, it can produce different mRNA sequence patterns or isoforms. Accordingly, each gene can produce between two and dozens of transcript patterns depending on the number of alternatively spliced exons. As a consequence this allows phenotypic diversity at a molecular level which can further produce several distinct proteins from a single gene (Lopez, 1998; Smith & Valcarcel, 2000; Smith & Roberts, 2002; Black, 2003). For example, alternative splicing generates multiple SMRT gene transcripts encoding variable protein isoforms which differ in their ability to interact with nuclear receptors (Malartre *et al.*, 2004). Hence alternative splicing is understood to be an evolutionary process used

by mammalian genomes to increase their protein complexity from relative small number of genes (Gravely, 2001; Kan *et al.*, 2001).

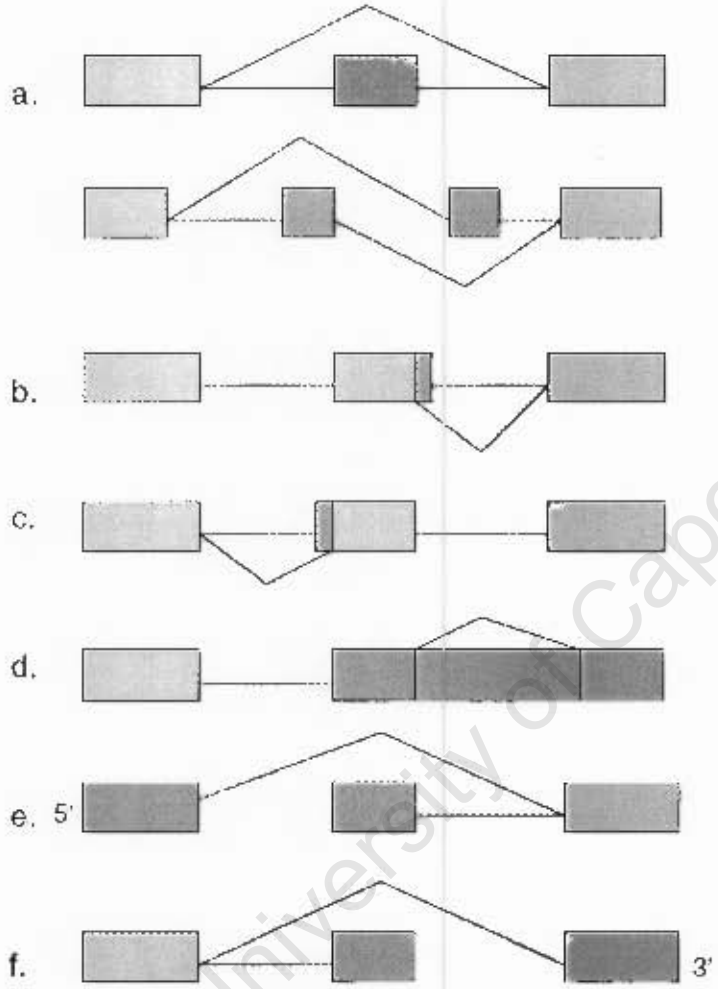

Figure 2.4: Schematics representation of the different patterns of alternatively spliced exons. Exons are shown as boxes and introns as straight lines connecting the exons. Alternatively spliced exons (ASEs) or portion of exons are shown in dark grey and constitutively spliced exons (CSEs) in light color. The angled lines show the exchange of splice sites. CSEs are always included in the same manner in all transcripts whereas ASEs are altered depending on splice site choice. (a) exon skipping or cassette exons and mutually exclusive event; (b) alternative 5' splice site; (c) alternative 3' splice site; (d) Intron retention; (e) 5' alternative untranslated region; (f) 3' alternative untranslated region. (Source: Florea, 2006)

Besides protein diversity, alternative splicing is also responsible for regulation of gene expression (Smith *et al.*, 1989, Nissim-Rafina & Kerem, 2002). For example, alternative splicing can be used to modify or remove a functional domain from a protein. In such cases, alternative splicing operates to up or down regulate gene expression (Smith *et al.*, 1989; Lopez, 1998; Woodley & Valcarcel, 2002). In addition, RNA stability and translation efficiency can be controlled by differential use of alternative untranslated regions (Hughes, 2006). Lastly, alternative splicing can regulate expression of mRNA isoforms that result in premature termination codons, by linking with nonsense-mediated mRNA decay (NMD) (Lewis et al., 2003; Lejeune & Maquat, 2005).

2.3.2 Regulation of alternative pre-mRNA splicing

There is still a lot to be learned about the *trans*- acting regulatory factors that control production of transcript variation based on alternative splice patterns, and about the distribution of *cis*-regulatory elements through which they act. Nevertheless, we know that the general pre-mRNA sequence features and splicing factors involved in the regulation of constitutively spliced exons (CSEs) are also implicated in regulation of alternatively spliced exon (ASEs). Regulation of alternative exon patterns occurs during splice site selection.

The decision of whether to include an exon and how it is included in the mRNA depends on the strength of the splice site elements around it (*cis*-acting) combined with availability and interaction of splicing-associated proteins (*trans*-acting). Splicing-associated proteins are differentially expressed in response to different cellular environment, suggesting a fundamental role of expression levels of these proteins in the regulation of alternative splicing (Zahler et al., 1993). Hence, some studies are aimed at clarifying the molecular mechanism that control changes in splice site choice (Smith & Valcarcel, 2000; Ladd & Cooper, 2002).

Most of alternative splicing events that are analyzed from human genes show that the process is controlled by cellular conditions (Maniatis, 1991; Xu *et al.*, 2002; Black, 2003); that means, pre-mRNA from the same gene produce different mRNA isoforms in different cell types, developmental stage (Lopez, 1998; Xu *et al.*, 2002; Black 2003) or in response to different external stimuli (Stamm, 2002). For example, alternative 3'splice site usage associated with differential use of polyadenylation sites in the calcitonin gene related peptide (CGRP) generates two distinct mRNA isoforms. One isoform is expressed in the thyroid gland and encodes calcium homeostatic hormone whereas the other isoform is expressed in the nervous system and encodes vasodilator neuropeptide (Smith & Valcarcel, 2000).

Cell or tissue specific alternative splicing regulation is due to existence of *trans*-acting splicing factors which bind specifically to exonic or intronic enhancer and silencer elements in cell or tissue regulated manner (Norton, 1994; Graveley, 2000). This is clearly demonstrated by the different members of SR family of proteins that specifically bind with particular ESE sequences in the pre-mRNA. (Wu & Maniatis, 1993; Tacke *et al.*, 1997). Which transcript isoforms are produced in a cell will depend on the availability of both the specific binding sites and the corresponding SR protein. In their review, Faustino and Cooper (2003) concluded that cell or tissue specific transcript isoforms emerge primarily from two features. Firstly, the repression of splice site choice in the unsuitable cell type is combined with activation of splice site in appropriate cell type. Secondly, the relative expression levels of *trans*-acting splicing factors in different tissues controls which c*is*- acting elements to be used during splicing, which in turn enhance or inhibit the recognition of the splice sites by the spliceosome.

Comparative analysis of alternative splicing events in multiple organisms revealed that most conserved alternatively spliced exons have weak splice sites, are comparatively short, and are likely to be multiple of 3nt in length (Sorek *et al.*, 2004; Yeo *et al.*, 2005; Xing & Lee, 2005). In addition, studies have shown that regions flanking orthologous ASEs have fewer SNPs, display higher level of sequence

conservation compared to those of orthologous CSEs (Sorek & Ast, 2003; Yeo *et al.*, 2005) but no function could be assigned. Species (i.e. human or mouse) specific ASEs did not show greater sequence conservation than CSEs (Zheng *et al.*, 2005).

Alternatively spliced exon length allow flexibility in the way in which the exons are arrangement in the mRNA (Xing and Lee, 2005). Consequently, ASEs are arranged in such a way that their inclusion or exclusion does not disrupt or truncate the protein reading frame. This allows inclusion or exclusion of specific functional domains in a protein while maintaining the general function of the gene from which the protein is encoded. Skipping or inclusion of exons occurs in such a way that there is expression of tissue specific isoforms with or without the functional domain.

2.3.3 Alternative pre-mRNA splicing and human diseases

Under normal conditions, alternative pre-mRNA splicing is highly regulated. However, abnormal splice site choices sometimes do occur, resulting in alternatively spliced patterns leading to human pathologies. The control of alternative pre-mRNA splicing can be deregulated and this may result from alteration within splicing regulatory factors (Baralle & Baralle, 2005). Table 2.1 shows a list of genes in which alterations in either cell signaling pathways, *trans*-acting or *cis*-acting splicing regulatory factors which results in aberrantly spliced mRNA isoforms that are involved in disease progression.

Although we have explained pre-mRNA splicing as if it is biochemically separated from other pre-mRNA processing steps (i.e. capping and polyadenylation) that occur during gene transcription, in reality this processes is tightly coordinated and coupled with transcription (Maniatis & Reed, 2002). Therefore, any factors that affect gene transcription regulation are likely to also influence pre-RNA splicing regulation and determine the mRNA isoform product.

Inappropriate cell signals or environmental stimuli can negatively affect regulation and expression of *trans*-acting splicing regulatory factors (Tarn, 2007). SR proteins expression levels are influenced by different cell transduction pathways and cell environments, consequently, these proteins are differentially phosphorylated in a way that directly influences their biochemical properties (Xiao & Manley, 1998; Prasad *et al.*, 1999). Unsuitable phosphorylation status of SR proteins may subsequently affect their cellular localization, their interaction with other proteins, or binding affinity with ESEs. When SR protein expression is not properly regulated, it might affect splice site choice which might direct synthesis of abnormal transcript pattern.

Modifications in spliceosome components are generally associated with serious disease phenotypes in humans. For example, modifications that disrupt the assembly or function of the spliceosomal snRNPs are responsible for Retinitis pigmentosa, and spinal muscular atrophy (Faustino & Cooper, 2003). Four components of the splicing machinery ($U2AF^{35}$, Sm protein D1, SF3b subunit 4, and U1CA) were identified as genes required for early vertebrate developments (Golling *et al.*, 2000). All these four genes are implicated in developmental defects that are thought to result from mutations.

Mutations occurring at *cis*-acting splicing regulatory regions, especially the conserved GT donor splice site or AG acceptor splice positions may create loss of function of the splice site or reduce binding specificity for splicing factors. Mutation occurring at splice site or those creating cryptic splice sites can lead to formation of inappropriate exon pattern (Skrygan *et al.*, 2001) or abnormal intron inclusion in the mRNA. Aberrant mRNA can be unstable (Kinniburgh *et al*, 1982) or translate to lethal proteins (Krawczak *et al.*, 1992; Cartegni *et al.*, 2002). In addition, a study of mutations associated with human inherited diseases observed that 15% of these point mutations are located at the intron-exon junctions and affect splicing (Krawczak *et al.*, 1992; Cartegni *et al.*, 2002; Hudson & Pastinen, 2004). As a further example, point mutations in exonic splicing enhancers have been shown to

cause exon skipping in BRCA1 gene, which subsequently results in cancer (Liu *et al.*, 2001).

Several genetic disorders and many forms of cancer are caused by nonsense mutations which promote premature translation termination (Frischmeyer & Dietz, 1999; Mendell & Dietz, 2001). Nonsense mutations produce shortened non-functional proteins, supposedly targeted for NMD (Hentze & Kulozik, 1999; Lewis *et al.*, 2003). There is evidence of connection between NMD and pre-mRNA alternative splicing (Lejeune & Maquat, 2005), suggesting that aberrantly spliced exons due to nonsense mutations are degraded from the cells. Possible consequences of an aberrant spliced transcript due to nonsense mutation can be serious in the absence of cell surveillance system that helps to avoid the synthesis of abnormal proteins.

Table 2.1: Genes in which a mutation results in aberrant alternative splicing patterns that are implicated in disease

Gene Symbol	***Gene name***	***Mechanism***	***Disease***	***Reference***
ADA	Adenosine deaminase	*Cis* - acting mutation	Adenosine deaminase deficiency	Ozsahin *et al.*, (1997)
ALG3	Asparagine-linked glycosylatior homolog	ESE activated upstream cryptic ss	Congenital disorder	Denecke *et al.*, (2004)
APOA2	Apolipoprotein A-II	5'ss	ApoA2 deficiency	Deeb *et al.*, (1990)
BRCA1	Breast Cancer 1	*Cis* - acting mutation	Breast Cancer	Liu *et al.*, (2001)
CA21HB	Steroid21-hydroxylaseB gene	3'ss	Adrenal hyperplasia	Higashi *et al.*,(1988)
CFTR	Cysticfibrosis transmembra conductance regulator	3'ss	Cystic Fibrosis	Delaney *et al.*, (1993)
COL1A2	collagen, type I, alpha 2	3'ss	Osteogenesis imp. II	Tromp & Prockop (1988)
CYBB	Cytochrome b-245, beta polypeptide	5'ss disruption	Chronic granulomatous	Ishabashi *et al.*, (2001)
DMPK	Dystrophia myotonica-protein kinase	*Trans* - acting mutation	Myotonic dystrophy	Mankodi *et al.*, (2000)
F9	Coagulation factor IX precursor	5'ss	Haemophilia B	Rees *et al.*,(1985)
FBN1	Fibrin 1	*Cis* - acting mutation	Marphan Syndrome	Liu *et al.*, (1997)
GH-1	Growth hormone 1	Cis - acting mutation	Growth hormone deficiency type II	Moseley *et al.*, (2002)
HBA2	Hemoglobin, alpha 2	5'ss created	α thalassaemia	Harteveld *et al.*, (2004)
HBB	Hemoglobin, beta	3'ss	Thalassaemia β	Atweh *et al.*,(1985)
L1CAM	L1 cell adhesion molecule	5'ss created	X-linked hydrocephalus	Du *et al.*, (1998)

Table 2.1 (continued)

Gene Symbol	*Gene name*	*Mechanism*	*Disease*	*Reference*
MAPT	Microtubule-associated protein tau	Cis - acting mutation	Frontotemporal dementia and Parkinsonism	Lee *et al.*,(2001)
OAS1	2',5'-oligoadenylate synthetase	3'ss	Susceptibility in viral infection	Bonnevie-Nielsen *et al.*, (2005)
PAH	Phenylalanine hydroxylase	ESE disrupted	Phenylketonuria	Chao *et al.*,(2000)
PDHA1	Pyruvate dehydrogenase	ESE disrupted	Leigh syndrome	De Meirleir et al. (1994)
PKLR	Pyruvate Kinase liver and red blood cells	5'ss	Pyruvate Kinase deficiency	Kanno *et al.*,(1997)
PMM2	Phosphomanomutase	ESE disruption	Glycoprotein deficiency	Vuillaumier–Barrot *et al.*, (1999)
PRPF31	Pre-mRNA processing factor homolog	*Trans* - acting mutation	Retinitis pigmentosa	Vithana *et al.*,(2001)
PTPRC	Protein-tyrosine phosphata receptor type C	*Cis* - acting mutation	Multiple schlerosis	Jacobsen *et al.*, (2000)
RB1	Retinoblastoma tumor suppresse	5'ss	Retinoblastoma tumor	Dunn *et al.*, (1989)
SMN1	Survivor motor neuron 1	ESE disrupted	Spinal Muscular atrophy	Lorson *et al.*,(1999)
TGFBR2	Transforming growth factor, beta receptor II	5'ss	Marphan syndrome	Mizuguchi *et al.*, (2004)
UROD	Uroporphyrinogen decarboxylate	5'ss	Porphyria, cutaneous	Mendez *et al.*,(1998)

2.3.4 Bioinformatics analysis of Alternative splicing

Prior to high-throughput sequencing of genomes, few alternatively spliced genes were identified using molecular biology techniques. However, with the advent of Genome-wide sequencing it has been possible to identify and characterize novel alternative splice patterns using bioinformatics approaches. As a consequence, a growing number of alternatively spliced genes have been identified. Genome-wide analysis estimates that more than 60% of human genes are alternatively spliced (Modrek *et al.*, 2001; Johnson *et al.*, 2003).

There are many computational methods used to detect alternatively spliced variants of a gene. These include prediction of alternatively spliced exons using comparative genomics (Modrek & Lee, 2003; Sorek & Ast, 2003; Sorek *et al.*, 2004) and *ab initio* prediction (Phillips *et al.*, 2004; Sorek *et al.*, 2005). Nevertheless, the most commonly used methods to identify alternatively spliced genes utilizes expressed sequence Tags (Mironov *et al.*, 1999; Modrek, *et al.*, 2001; Xu *et al.*, 2002) and microarray data (Hu *et al.*, 2001; Shoemaker *et al.*, 2001; Clark *et al.*, 2002; Johnson, 2003; Kampa *et al.*, 2004).

2.3.4.1 cDNA and EST based approach

Complementary DNA (cDNA) and Expressed sequence tags (ESTs) are expressed sequences which represent the expression profile of a gene. cDNA is a single stranded DNA that has been synthesized from the expressed mRNA molecule under specific physiological condition of a cell (Kimmel *et al.*, 1987; Belyavsky *et al.*, 1989; Huang, 2002), whereas, an EST is a fragment of expressed sequences generated by single-pass sequencing from either a 5' or 3' ends of a cDNA clone (Boguski *et al.*, 1993; Boguski & Schuler, 1995). The production of ESTs requires the construction of a cDNA library that represents the transcriptome of the tissue or cell type of interest. Thus, an EST reflects the expression state of a gene in a particular

tissue or cell condition. An EST cluster is a collection of ESTs that provides information about the expression pattern of a gene under different physiological conditions and is constructed to identify structural features of genes, such as gene boundaries, exon-intron junctions and alternatively spliced variants (Mathe *et al.*, 2002).

Alternatively spliced variants of genes are detected by computational methods that align ESTs or cDNA sequences to genomic sequences (Figure 2.5). The aim is to establish the structure of a gene on the genomic sequence, based on the transcript alignment. The genomic regions that match the transcribed sequences are exons and the alignment gaps between them are considered to be introns. During the alignment not only the degree of similarity of the sequences is taken into account, but also the intron-exon junction sequences are checked for the presence of splice signals (GT- AG, polypyrimidine tract, etc). In a widely cited survey, Modrek and colleagues reported alternative splicing events of a gene when two or more transcribed sequences overlap in the genomic sequence (Modrek, *et al.*, 2001).

The use of EST data is extremely helpful in identifying structural features of genes and alternative splicing; however, there are noted limitations to this approach (Mironov *et al.*, 1999; Brett *et al.*, 2000). Firstly, since ESTs are single-pass reads and their sequences are usually between 300-500 bases in length, the quality of the sequence is sometimes low and errors are common. Secondly, EST sequencing covers mostly terminal ends of the transcript; consequently internal exons of long transcripts are poorly represented. This makes it difficult to predict alternatively spliced patterns of internal exons using ESTs. Thirdly, it is not clear how many of the ESTs represent functional transcript or aberrant splicing, hence, further experimental validation of the functional impact of the alternatively spliced variants may be needed. Finally, although some cDNA libraries are normalized (Bonaldo *et al.*, 1996), clones containing rare transcripts are likely to be poorly represented or completely absent from the cDNA library, as a result, rare alternatively spliced patterns are likely not to be found using ESTs prediction.

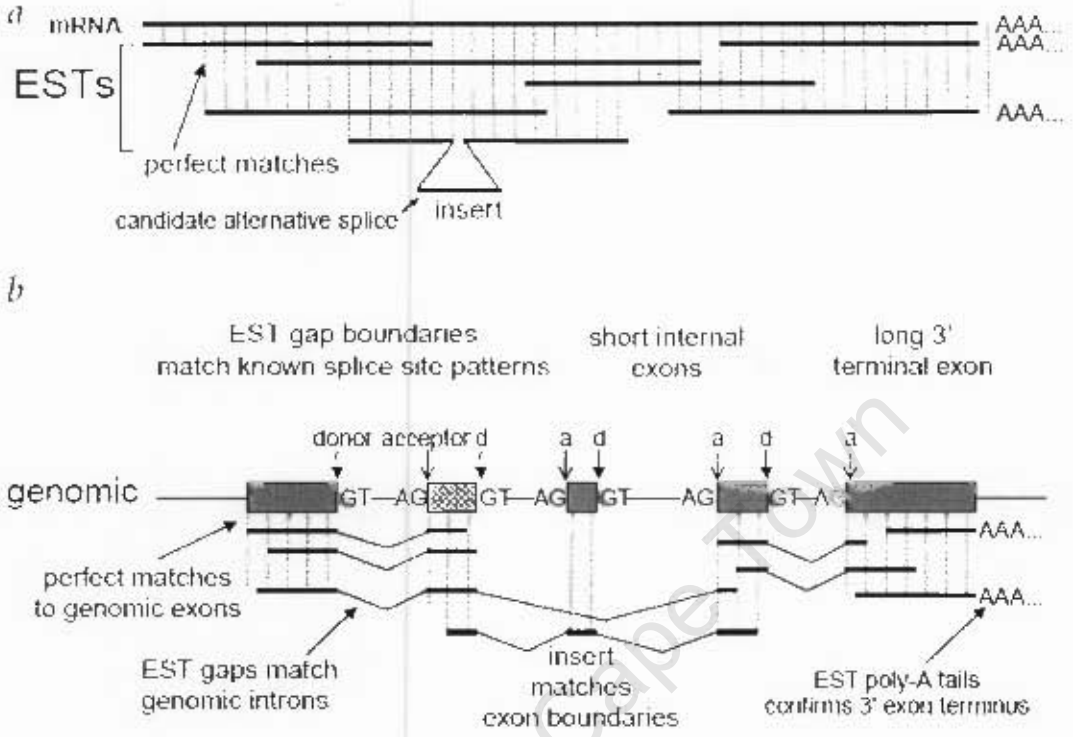

Figure 2.5: Identification of alternatively spliced genes using ESTs. (a) Partial gene structures generated by EST are merged whenever they overlap. Insertion and deletion in ESTs relative to mRNA are identified as potential alternative transcript. (b) Genomic sequence with exons shown as grey boxes and intronic donor (d) and acceptor (a) sites. Expressed EST contigs are aligned below the genomic sequences. Alternative splice variants are detected when two transcripts matching to the same genomic region show difference in the way they align and have boundaries correspond to known splice sites. Types of alternative splicing shown in this diagram are exon skipping, alternate 5' or 3' splice sites, and 3' alternative untranslated region. (Source: Modrek & Lee, 2002)

2.3.4.2 Microarray based approach

cDNA and oligonucleotide microarray technologies were initially developed for parallel analysis of expression profiles of many genes (Schena *et al.*, 1995). These methods are mostly used to detect the changes in gene expression profiles across tissues or different experimental conditions. Lately, there is a growing trend on using microarrays to identify alternative splicing (Shoemaker *et al.*, 2001; Clark *et al.*, 2002; Johnson, 2003). It is possible to investigate whether a transcript isoform is enriched in some cell or tissue types using microarrays.

The common approach to detect alternative splicing is to design oligonucleotide probes for exon-exon junctions. As Lee and Roy (2004) highlighted, it is important to have probes complementary to every exon-exon junction of a given gene since different transcript isoforms have distinct exon-exon junctions. Hybridization data from the junction probe of each transcript are then analysed and compared across tissue samples to detect distribution of various mRNA isoforms. Probe intensity is used to detect changes in exon expression levels (quantitative change) whereas existence of different exon-exon junctions in a gene signifies changes in gene sequence content (qualitative changes). When individual exon-exon junction probes are considerably down regulated relative to the other probe, those with statistical significance above a given threshold level are reported as alternative splicing prediction.

Microarray analysis has made a significant contribution to the discovery of alternatively spliced variants of human genes. However, the method is limited by the fact that prior knowledge of gene structure (transcript) is needed before the experiment is done. Hence exon-exon junction microarray method is mostly used for identification of alteration in exon usage in different cells or tissues but not suitable for discovery of novel splice isoforms (Lee & Roy 2004; Clark *et al.*, 2007). Besides, most gene prediction programs are not performing very well in identifying terminal

exons corresponding to the 5' and 3' UTRs of mRNA. As a result, there could be limited identification of 5' and 3' alternative terminal junction using this method. Predominantly, microarray studies address only one type of alternative splicing, monitoring exon inclusion or deletion events that are tissue specific (Johnson, 2003; Kampa *et al.*, 2004) and the other alternative splicing events (i.e. 5' and/or 3 alternative splice sites usage) are not well studied (Lee & Roy, 2004). Furthermore, when analysing microarray data it might be difficult to separate the difference between transcript changes that are due to general gene expression from changes due to alternative splicing (Le *et al.*, 2004).

2.3.4.3 Alternative splicing databases

The transcript derived alternative splicing predictions have made it possible to identify and characterize alternative splicing events on a genome-wide scale. The comparable nature of microarrays makes them most suitable for monitoring tissue or cell specificity of spliced isoform, while, ESTs are ideal for transcript structure prediction. The rate of alternative splicing recovery depends on available transcript coverage; hence, recent estimates may increase with the increase in transcribed sequence data available.

Besides the prediction of alternatively spliced isoforms using transcripts sequences, a number of approaches use *ab initio* prediction (Phillips *et al.*, 2004; Sorek *et al.*, 2005; Dror *et al.*, 2005; Holste & Ohler, 2008). This computational approach is made possible by the fact that ASEs have different characteristics compared to CSEs. Thus, machine learning can be used to look for special sequence features associated with ASEs and CSEs. On the other hand, recognition of alternatively spliced variants can be achieved through comparative genomics approach by identifying alternative splicing events that are conserved with respect to orthologous genes (Modrek & Lee, 2003; Sorek & Ast, 2003; Kan *et al.*, 2004; Sorek *et al.*, 2004;

Yeo *et al*., 2005). This method is useful because it reduces prediction of non functional alternative splice patterns.

Given the various ways in which datasets of alternative spliced genes are produced, it is now possible to obtain a catalogue of alternatively spliced genes and their mRNA isoforms. The information is stored in specialized alternative splicing databases, some of which are listed in Table 2.2.

Table 2.2: Databases of alternative spliced genes

Tool	*Type of data*	*Species*	*Method*	*Website*	*Authors*
ASAPII	Genomic sequences ASEs and CSEs	15 species	Comparison of DNA and EST-genome alignments	http://www.bioinformatics.ucla.edu/ASAP2	Kim, *et al.*, (2007)
ASD	Alternative splice exons	Human & Mouse	Comparison of cDNA and EST-genome alignments	http://www.ebi.ac.uk/asd/	Stamm *et al.*, (2006)
ASDB	Genomic sequences	Human & Mouse	Clusters corresponding to gene transcript variants	http://hazelton.lbl.gov/~teplitski/alt/	**Dralyuk** *al.*,(2000)
ASG	Splice graphs,combinatorial splice variants	Human	Genome based splice graphs	http://statgen.ncsu.edu/asg/	Leipzig *et al.*, (2004)
AsMamDB	Alternative splicing events	Human, Mouse,rat	Comparison of GeneBank annotate splice variants	http://166.111.30.65/ASMAMDB.html	Ji *et al.*, (2000)
ASPIC	Gene and exon-intron stucture prediction	17 species	Comparison of EST-genome alignments	http://t.caspur.it/ASPIC/	Bonizzoni *et al.*, (2005).
EASED	Alternative splice transcripts	9 species	Comparison of ESTs with mRNAs	http://eased.bioinf.mdc-berlin.de/	Pospisil *et al.*, (2004)
ECgene	Gene and transcript prediction	9 species	Genome based splice graphs	http://genome.ewha.ac.kr/ECgene	Kim *et al.*, (2005).
HOLLYWOOD	Genomic annotations splicing patterns	Human and Mouse	cDNA and genome alignments	http://hollywood.mit.edu/Logo/index.html	Holste *et al.*, (2006)
SpliceInfo	Alternative splicing events	Human	Comparison of mRNA-and protein genomic alignments	http://spliceinfo.mbc.nctu.edu.tw	Huang *et al.*, (2005)

2.4 Human Genetic Variation

Having discussed transcript variation within an individual resulting from alternative splicing, we now turn to a discussion of genetic variation as a cause of transcript variation between individuals. Human genetic variation and our environment are the two basic factors that cause human phenotypic variation. Genetic variation arises through a spectrum of changes in a genome sequence, encompassing point mutations and chromosomal evolution. Genetic variation is natural and governed by biological, demographic and historical processes (Chakravarti, 1999). Notably, the genetic variation that has survived in the human population is non-random; it has been shaped by both genetic drift and natural selection.

Although genetic variation takes many forms, the most common genetic variations found in the human genome are single nucleotide polymorphisms or SNPs (Altshuler *et al.*, 2000). SNPs are heritable single nucleotide substitution, with the rare allele occurring at a frequency of at least 1% within population. Note that the terminology for variation is defined by allele frequency, and in the strictest sense genetic loci are only regarded as polymorphic if the frequency of the uncommon allele is equal to or greater than 1%. When a single base substitution occurs at less than 1% it is considered to be a mutation. However, during the post genomic era this definition is often disregarded; instead "mutations" occurring at less than 1% in general populations might be termed low frequency variations (Collins *et al.*, 1998).

SNPs can be bi, tri, or tetra allelic. SNPs are classified according to their genomic position and the nature of the nucleotide that is affected. Non-coding SNPs are located in either untranslated region (UTR), introns, or they may be intergenic. Coding SNPs are located within protein coding regions and may be non-synonymous (they change amino acid that is encoded) or synonymous (they change the codon but not the amino acid). Non-synonymous SNPs usually influence structural and functional features of the affected proteins (Lucotte, 1998; Ramensky *et al.*, 2002). Although non-coding and synonymous SNPs are thought not to change

the coding capacity of a gene, they may affect gene function through their effect on gene expression regulation (Ramensky *et al.*, 2002). SNPs occurring in coding regions or flanking regulatory sequences of a gene could have deleterious consequence on gene function and possibly cause disease (Stenson *et al.*, 2003).

SNPs are commonly used as genetic markers for finding the genetic basis of phenotypic variation in the human population. It is already known that expression of many genes varies significantly between individuals because of allele specific regulation (Oleksiak *et al.*, 2002; Buckland, 2004; Knight, 2004). For example, Bray and colleagues (2003) showed differential expression in the human brain which occurs due to *cis*-acting genetic variations among individuals. Allele-specific regulation might also produce difference in disease susceptibility or manifestations between individuals (Eriksson *et al.*, 1995; Montagna *et al.*, 2002; Prokunina *et al.*, 2002; Miao *et al.*, 2003,).

Analysis of the effects of SNPs on gene transcripts has assisted in understanding the underlying causes of qualitative or quantitative differences in the way in which genes are expressed, the causes of differences in protein function among individuals. Table 2.3 provides currently available bioinformatics tools which provide analysis of genetic variation and their possible functional implications within the human genome. Similarly, these tools can be further exploited for evaluating the effects of SNPs on pre-mRNA splicing.

Table 2.3: Online SNP databases and related tools

Tool	*Type of data*	*Species*	*Website*	*Reference*
ColiSNPs	Maps nsSNP on the 3D structure of proteins	13 Organisms	http://yayoi.kansai.jaea.go.jp/colisnp/	Kono *et al.*,(2008)
dbSNP	Primary repository for SNP data	16 organisms	http://www.ncbi.nlm.nih.gov/projects/SNP/	Sherry *et al.*, (2001)
F-SNP	Functional effect of SNP	Human	http://compbio.cs.queensu.ca/F-SNP/	Lee & Shantkay, (2008)
HapMap	SNP allele and genotype frequencies and Haplotype block	Human	http://www.hapmap.org/	IHC, (2005)
HGMD	Mutations associated with human inherited disease.	Human	http://www.hgmd.cf.ac.uk/ac/index.php	Stenson *et al.*, (2003).
HGVbase	Genomic variation data of all types (SNP, Indel, and Tandem Repeat). Neutral and disease related mutations	Human	http://www.hgvbase.org/	Fredman *et al.*, (2002)
JSNP	Common genetic variation in Japanese population	Human	http://snp.imsu-tokyo.ac.jp/	Hirakawa *et al.*,(2002)
MutDB	Intergration of genetic variation with molecular features and clinical data	Human	http://mutdb.org/	Singh *et al.*, (2008)
PolyPhen	Predict effect of nsSNP on protein structure and function	Human	www.bork.embl-heidelberg.de/PolyPhen/	Ramensky *et al.*, (2002)
PupaSuit/ SNPeffect	Coding and non coding SNPs and mutations	Human, mouse & rat	http://pupasuite.bioinfo.cipf.es/	Reumers *et al.*, (2008)
SNPtoGo	Assign Go terms to SNPs in addition to genomic positions	Human	https://webtools.imbs.uni-luebeck.de/snptogo	Schwarz *et al.*, (2008)

2.5 Genetic variations that alter alternative pre-mRNA splicing

We have already mentioned that the impact of a SNP on gene expression profile and protein function depends upon its location within the genome (Cargill *et al*, 1999; Hudson *et al*, 2004). Genetic variations located at pre-mRNA splicing regulatory regions may affect the way in which genes are spliced, resulting in differential splice patterns which may play an important role in phenotypic diversity and genetic disorders between individuals (Faustino & Cooper, 2003; Nissim-Rafinia & Kerem, 2005). Some alternatively spliced variants occur between individuals instead of within individuals. For example, Figure 2.6 shows that a G to T single nucleotide polymorphism in donor (5') splice site of exon 14 of human serum albumin A (HSA) gene result in exon 14 skipping in certain individuals of Italian populations (Watkins *et al*., 1991). The protein product of the splice variant differs from normal albumin A.

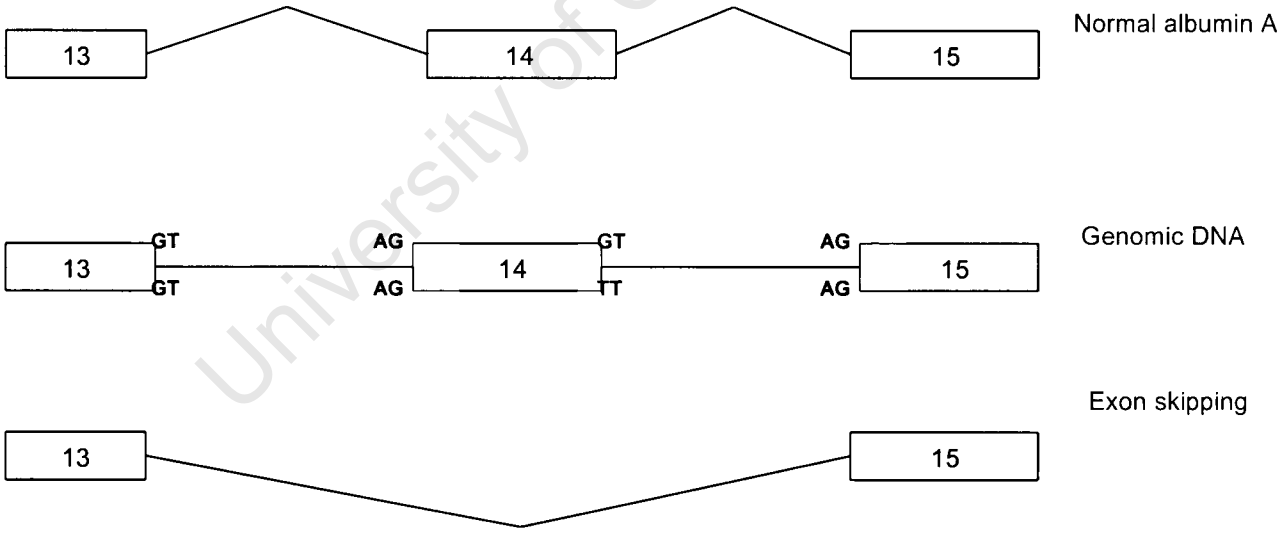

Figure 2.6: Schematic representation of allele dependent alternative splicing. Exons are shown as boxes and introns as straight lines connecting the exons. The canonical donor and acceptor nucleotides for normal albumin A gene are shown above the line on the genomic DNA. The altered Venezia (Italian population) splice junction is shown below the line. The G/T SNP (shown in red) in exon14-intron14 junction inactivates selection of splice site and as a consequence, exon 14 is skipped joining exon 13 and 15.

Cis-acting splicing regulatory single nucleotide polymorphisms (srSNPs) might be the underlying cause of distinct transcript patterns among individuals. Hence, before we attribute multiple transcript variants of a gene to alternative splicing, it is necessary to distinguish between transcript variant which are the result of true alternative splicing (same allele results in different splice patterns), possibly serving to increase the functional repertoire of a gene, and transcript variant resulting from genetic polymorphisms, which may be responsible for some phenotypic variation between individuals.

Although there has been no systematic study showing the prevalence of polymorphism that affects pre-mRNA splicing, Nembaware *et al* (2004) estimated that 21% of alternatively spliced genes found in alternative databases are allele specific alternatively spliced events. Therefore, some of the observed alternative splicing patterns are not a product of alternative splicing of a gene within different tissues of individuals but are due to polymorphisms that may be responsible for phenotypic differences between individuals. Efforts to identify allele dependent spliced transcripts in certain genes have been undertaken. Table 2.4 shows some experimentally validated genes that are affected by allele-specific splicing.

In consideration of the fact that the current bioinformatics methods to detect alternative splicing rely absolutely on the assumption that pre-mRNAs are spliced in the same way in different individuals, it would be of great interest to conduct an investigation of the extent to which genetic polymorphisms affect splicing. This is a topic of Nembaware *et al*. (2004; 2008) work and this thesis will contribute towards developing methods to detect allele-specific splicing in human by looking at how SNPs are distributed at splicing regulatory regions of human genes.

Table 2.4: Allele-Specific Alternatively Spliced Genes and Associated rSNPs

Gene	*Exon affected*	*rSNP*	*Cis-element*	*References*
C3AR1	Exon 2	rs2230318	Unknown, located exonic region	Hasegawa *et al.*, (2004)
CAST	Exon 10	rs7724759	5'ss	Kwan *et al.*, (2007)
CD45	Exon 4	rs12129883	ESS	Jacobsen *et al.*,(2002)
COL5A1	Exon 65	rs13946	3'ss	Wenstrup *et al.*, (1996)
ETV4	Exon 3	rs3765174	NAGNAG acceptor	Hiller *et al.*,(2006)
GABRR1	Exon 2	rs4590242	NAGNAG acceptor	Hiller *et al.*, (2006)
GSTM4	Exon 4 and Exon 5	rs560018	3'ss	Denson *et al.*, (2006)
ITPA	Exon 2 and Exon 3	rs13830	ESS	Arenas *et al.*,(2007)
LDLR	Exon 12	rs688	ESE	Zhu *et al.*, (2007)
MST1R	Exon 19	rs12489386	Unknown, located intronic region	Angeloni *et al.*, (2003)
MUC1	Exon 2	rs4072037	Unknown, located exonic region	Ligtenberg *et al.*, (1991)
OAS1	Exon 7	rs10774671	3'ss	Bonnevie-Nielsen *et al.* 2005)
PARP2	Exon 2	rs2297616	5'ss	Kwan *et al.*, 2008
PMM2	Exon 5	rs28938475	ESE	Vuillaumier-Barrot *al.*,(1999)
RBM23	Exon 6	rs2295682	Unknown, located exonic region	Hull *et al.*,(2007)
TAP2	Exon 12	rs241447	Unknown, located exonic region	Qu *et al.*, (2007)
UROD	Exon 4	rs1804886	5'ss	McManus *et al.*,(1996)

CHAPTER 3

EXON ANNOTATIONS AND SNP DATASETS

3.1 Overview

A major objective of this chapter is to describe the methods used to generate human exon and SNPs datasets, which are further used for analysis. The current investigation is based on the Human genome assembly (NCBI 35). We will exploit exon information from the Ensembl and ASAPII databases whereas the SNP dataset is derived from dbSNP and HapMap. All these datasets are accessed from either interactive website or MySQL database servers via the internet.

3.2 Background

The complete sequence of the euchromatic portion of the human genome is available for analysis (Lander *et al.*, 2004). The genomic data comprises approximately 3 billion base pair sequence and the associated annotations which are stored in database systems (Lander *et al.*, 2001; Venter *et al.*, 2001). As we already cited in Table 2.2 and Table 2.3, databases are core resources of Bioinformatics because they offer a convenient and efficient method of storing and mining vast amounts of genomic data. There are numerous biological databases, depending on the nature of the information being stored. Our challenge is to make sure that the raw genomic data found in the databases is transformed into biological meaningful information applicable in health science research.

The Ensembl database is a system which provides automated genome annotation and tools for the visualisation of annotated genomes. For this project, the Ensembl database is the source of annotated human genome sequence, with confirmed and predicted genes that are often cross referenced with integrated external data

(Hubbard *et al.*, 2005). We chose the Ensembl system for the exon dataset because it provides genomic mapping of exon-intron gene structure together with EST confirmed transcripts. In addition to Ensembl, we used ASAPII to characterize alternatively and constitutively spliced exons. ASAPII is a comprehensive database of alternative splice variants predicted from ESTs and mRNA data (Kim *et al.*, 2007). In ASAPII, the genome-wide manual annotation of alternative splicing events is presented with tissue specific and comparative genomics information.

To develop a genome-wide catalog of sequence polymorphisms that are found close to intron exon boundaries, we used the NCBI database of Single Nucleotide Polymorphisms (dbSNP) which contain approximately 10 millions human genetic sequence polymorphisms (Sherry et al., 2001). dbSNP provides data on human genetic variations that can be easily mapped to exon-intron junctions. We also obtained SNP allele frequency information from the HapMap database (Altshuler *et al.*, 2005). HapMap contains genotype and population specific allele frequency data.

3.3. Materials and Methods

3.3.1 Exon dataset

We obtained pre-computed genomic coordinates (strand, chromosomal start and end positions) for all annotated human exons from the Ensembl MySQL Homo sapiens core database (version 36.35i).
The data can also be downloaded through the Ensembl genome browser, http://dec2005.archive.ensembl.org/index.html, released in December 2005. Ensembl gene annotations can include exons from pseudogenes and very short exons (< 20bp). We removed these from the dataset.

3.3.2 Classifying exons as alternatively and constitutively spliced

We downloaded MySQL tables (files) from a database of alternatively spliced annotated genes, ASAPII http://bioinfo.mbi.ucla.edu/ASAP2/. This data was

released in January 2006. ASAP II provides genomic mappings of alternatively and constitutively spliced exons. To identify Ensembl alternatively spliced exons, we used chromosomal locations of the ASAPII annotated alternatively spliced exon dataset. By default, we consider every Ensembl exon as constitutively spliced unless its corresponding ASAPII exon annotation is alternatively spliced. This method produced a dataset of alternatively spliced exons with both Ensembl and ASAPII identifiers. Using this approach we produced a dataset of 25 086 alternatively spliced exons (ASEs) and 215 833 constitutively spliced exons (CSEs) which we used for further analysis.

For several reasons, the number of alternatively spliced exons identified might be an underestimate. Firstly, the total number of ASAPII annotated exons is a little more than half the total number of Ensembl annotated exons. Ensembl gene structure annotation uses a combination of *ab-initio* gene prediction algorithms, evidence from various external sources (i.e. homologs) together with transcript (i.e. ESTs, cDNA) genomic alignments, whereas ASAPII annotation uses only transcript alignments. Consequently a majority of Ensembl annotated exons are not found in the ASAPII dataset. Secondly, our method of alternatively spliced exon prediction could have overlooked a lot of internal skipped exons due to limitations in the available EST data.

3.3.3 SNP genotype and frequency dataset

We retrieved human SNP data from the Ensembl MySQL server. We used human variation data from the Ensembl Homo sapiens variation database (version 36 35i) which is based on the NCBI dbSNP build 125. Only the RefSNPs (rs) annotated by the NCBI on reference genome sequence contigs were included for analysis. dbSNP contains single and poly nucleotide polymorphisms in addition to short deletion/insertion (indels) polymorphisms. SNP locations on human chromosomes are catalogued together with possible nucleotide character states (alleles). Our analysis include all RefSNPs found in dbSNP, Hence we use the term SNPs for all the different types of genetic sequence variation as annotated in dbSNP.

Although dbSNP is considered a comprehensive database of genomic variation the quality and completeness of dbSNP data has been under scrutiny lately (Reich *et al.*, 2004; Platzer and Hiller, 2006). To avoid including sequencing errors submitted to dbSNP we also considered and used a subset of validated SNPs from the dataset. In dbSNP, a polymorphism is categorized as validated only if the submission was validated by a non-computational method, and/or has frequency or genotype data submitted (i.e. HapMap).

3.4 Results and Discussion

3.4.1 Human exons

We retrieved all available human coding exons from the Ensembl genome database, most of which come from multi-exon protein coding genes. Genome-wide analysis of the human genome sequence indicates that the genome contains approximately 25 000 genes with each gene having an average of 8-10 exons (Thanaraj & Stamm, 2003). This is supported by our dataset of human genes and exons described in Table 3.1. The median exon length is 134 bps although there are shorter and much longer exons in the dataset (Table 3.2). Previous studies of statistical features of human exons indicated that the length of the exon depends on its position on the

genes (Zhang, 1998). According to our distribution of exon lengths we assumed that both internal and terminal exons are included in our analysis.

Table 3.1: Descriptive summary for human genes and exon

Genomic Features	Total No.
Protein coding genes	22 218
RNA coding genes	4 133
Average number of exons per gene	10
Exons	240 918
Exons from protein coding genes	236 785
Exons from RNA genes	4 133

Table 3.2: Distribution of human exon length

Statistic	Min.	Q_1	Median	Mean	Q_3	Max.	SD
Value	21	93	134	274.2	203	19 460	511.27

Q, mean ith quartile, SD mean standard deviation

3.4.2 CSEs vs. ASEs

We further categorized Ensembl exon as either constitutively spliced or alternatively spliced. Using Ensembl and ASAPII annotation we identified 25 086 out of 240 918 exons as alternatively spliced. The two exons categories have distinct lengths. The average length of alternatively spliced exons (ASEs) is not much more than half that of constitutively spliced exons (CSEs) (Figure 3.1). The exon length variation between ASEs and CSEs might be due to either functional evolutionary constraints imposed on internal skipped exons to keep them short and/or existence of partially coding terminal (first and last) exons with alternatively used donor or acceptor sites.

The results concur with previous reports which show that alternatively spliced exons are significantly shorter than constitutively spliced exons (Stamm, 2000; Zheng *et al.*, 2005). Suggestions have been made that the length of alternatively spliced exons (Black, 1991), as well as the length of its flanking introns (Bell *et al.*, 1998) are involved in alternative splicing regulation.

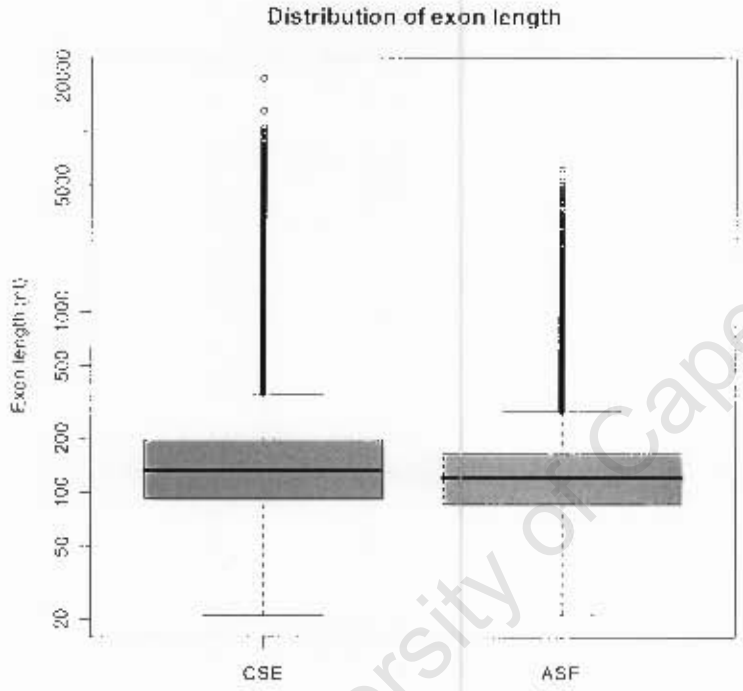

Figure 3.1: Box plot showing distribution of constitutively spliced exon (CSE) and alternatively spliced exon (ASE) lengths. The data consists of 169 061 CSEs and 25 086 ASEs. Alternatively spliced exons are shorter (mean = 158, SD = 220) than constitutively spliced exon (mean = 266, SD = 504). The Mann-Whitney U test indicates a highly significant ($p < 0.0001$) difference between the two medians.

There are several possible explanations for why alternatively spliced exons are shorter than constitutively spliced exons. Firstly, exons that are shorter than an average exon length have a tendency to code in more than one reading frame (Clark and Thanaraj, 2002), implying that short exons can contribute to protein diversity (Liang & Landweber, 2007). Secondly, reduction of the long constitutively spliced exons *in vitro* has been shown to be associated with exon skipping events

(Dominski & Kole, 1991). In addition, Chern (2006) anticipated that the short alternatively spliced exons observed are basically long constitutively spliced exons with cryptic splice sites, which can be alternatively spliced to produce shortened exon of varied length (Figure 3.2).

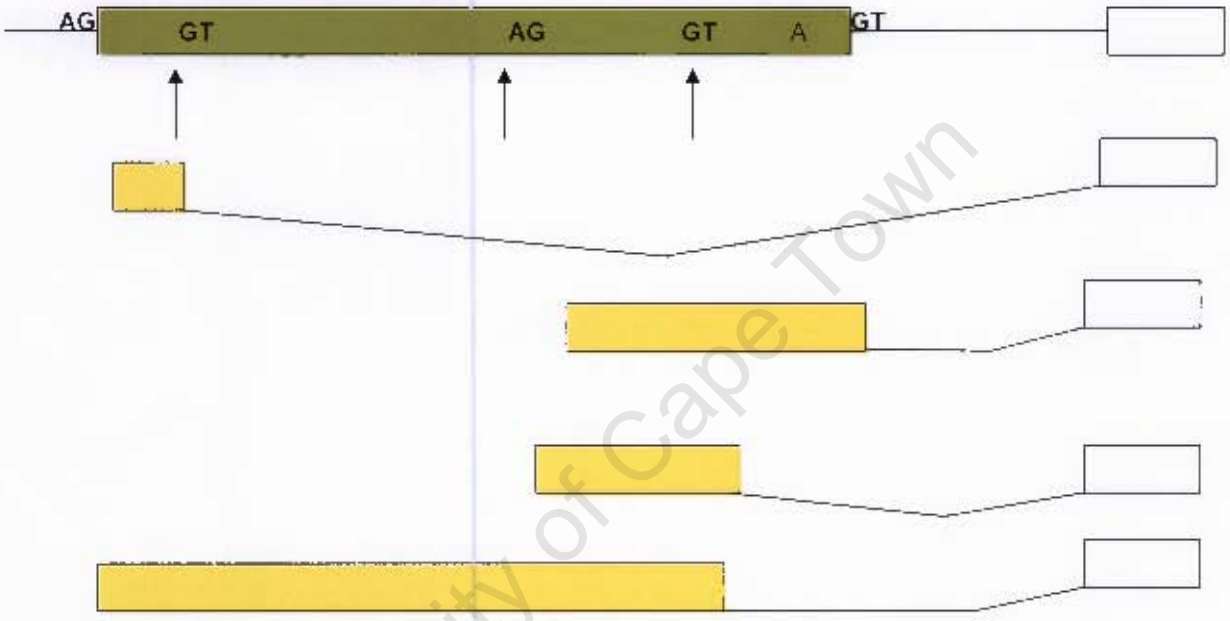

Figure 3.2: graphical representations of possible splice patterns of long constitutively spliced exons containing cryptic splice sites. Long exons (green) are likely to harbor cryptic splice sites (indicated by arrow) and the activation of cryptic donor and acceptor splice sites induce alternative splicing, resulting in shortened exons whose length varies between transcripts.

A recent study demonstrated that alternative 5' and 3' splice sites exons are an intermediate state between constitutive and alternative cassette exons, where the constitutive side resembles constitutive exons, and the alternative side resembles alternative cassette exons (Koren et al., 2007). This model suggests that indeed alternatively spliced exons originated from constitutively spliced exons that acquired a new competing splice site during evolution. This is clearly demonstrated by activation of the cryptic donor splice site in exon 17 of OPA1 gene which results in deletion of the last 40nt of the exon (Schimpf *et al.*, 2006) and consequently a

shortened alternative exon. The observed ASEs sizes support the idea that in addition to the common exon skipping events, many alternatively spliced exons are truncation events rather than extensions (Clark & Thanaraj, 2002). The decreased alternatively spliced exon size might be a reflection of high prevalence of short skipped and truncated exon events.

3.3.2 SNP dataset

We retrieved human variation data across the genome and obtained genotypes and allele frequencies. The majority of polymorphisms found in dbSNP are bi-allelic SNPs, although poly- (tri and tetra) allelic SNPs and small insertions-deletions (indels) polymorphisms also exists (Table 3.2). In the human genome tri-allelic and tetra-allelic polymorphisms are said to be rare almost to the point of non-existence (Brooks, 1999), yet we found 42 296 poly-allelic RefSNPs including 12 280 validated SNPs in dbSNP. Our working definition of SNP is single base pair polymorphism at which different alleles exists in the population. Hence, indels were not included in the validated SNP dataset.

Table 3.3: Descriptive summary of human SNPs dataset

SNP Category	Datasets			
	RefSNPs entries		*Validated SNPs*	
	Number	%	Number	%
Bi-allelic SNPs	9 200 659	93.90	4 723 816	99.74
Indels	555 151	5.67	0	0
Poly-allelic SNPs	42 296	0.43	12 280	0.26
Total	*9 798 106*	*100*	*4 736 096*	*100*

Frequencies of the different SNP categories catalogued in dbSNP expressed in total number and percentage (%).

3.5 Conclusions

Data obtained from the Ensembl, ASAPII, dbSNP and HapMap databases allowed us to create a catalog of human exons and SNPs. All datasets used are based on NCBI Genome assembly 35, making it easy to integrate them. The information includes the genes to which the exons and SNPs are mapped, including the chromosomal positions and strand on which the exons and SNPs are located. The generated datasets allowed the survey on the prevalence of SNPs in the exon-intron junctions of human coding genes feasible, and this is carried out in the next chapter.

CHAPTER 4

PREVALENCE OF SNPS AT EXON-INTRON JUNCTIONS

4.1 Overview

This chapter aims to estimate the prevalence of SNPs at exon-intron junctions of human genes. We address the objective by carrying out a survey of the occurrence frequency of SNPs at exon-intron junctions using the human genome annotations described in chapter three. We also compared the distribution of SNPs at coding exonic regions and non-coding intronic regions flanking the exon boundaries. The data sets of ASEs and CSEs allowed us to estimate and compare the frequency of SNPs found in exon-intron junctions of the two exon categories. This genomic mapping of SNPs is likely to produce a dataset of SNPs which might affect mRNA splicing.

4.2 Background

Alternatively spliced exon patterns from a single gene might be a result of either tissue specific alternative splicing events within an individual (Gravely, 2001; Kan *et al.*, 2001) or due to occurrence of splicing regulatory polymorphisms that affect the way in which genes are spliced in different individuals (Hull *et al.*, 2007; Kwan *et al.*, 2007; Kwan *et al.*, 2008). Earlier studies estimated that 21% of alternatively spliced genes are affected by polymorphisms that affect splicing (Nembaware *et al.*, 2004). Besides, several experimentally validated examples of genes with alternatively spliced exon variants that are affected by SNPs have already been published (table 2.4).

Most of the carefully studied splicing regulatory SNPs that are associated with alternatively spliced exons are *cis*-acting (Angeloni *et al.*, 2003; Královicová *et al.*, 2004; Hiller *et al.*, 2006). For example a C to an A single nucleotide polymorphism found in intron 18, 10bp upstream of acceptor splice site of exon 19, of the MST1R gene is associated with skipping of exon 19 (Angeloni *et al.*, 2003). There are instances where the SNPs are found at undefined positions and operating through *trans*-splicing regulatory factors (Vithana *et al.*, 2001; Horiuchi *et al.*, 2003). Some of the splicing regulatory SNPs are found in generally healthy populations; however there are cases of aberrantly spliced exon patterns leading to diseases that have been shown to arise from SNPs that modulate alternative splicing (Mukai *et al.*, 2004; Krawczak *et al.*, 1992; Hiller *et al.*, 2006; Krawczak *et al.*, 2007).

Identifying human genetic variations that are the basis for heritable phenotypic variation is still a developing research field. In line with aims of the International HapMap Consortium (Altshuler *et al.*, 2005), identifying functionally important variation in the human genome has the potential for increasing our understanding of gene function and for providing markers to study the effect of genetic variations in gene expression and human disease risk. Similarly identifying genetic variation found at splicing regulatory regions that have been associated with phenotypic variation is important for understanding the relationship between splicing and human disease susceptibility (Hull *et al.*, 2007).

A critical examination on how SNPs are distributed in the exon-intron junction can assist in finding regulatory polymorphisms that affect splicing. However, it is difficult to figure out how pre-mRNA splicing is affected by genetic polymorphisms without availability of expression data. We expect that identification and estimation of the prevalence of SNPs at exon-intron junctions will provide insight into polymorphisms responsible for heritable phenotypic differences within the human population and also assist to detect and estimate the prevalence of allele specific splicing in the human genome.

4.3 Materials and Methods

4.3.1 Defining genomic exon-intron junctions

Genomic mapping of exon-intron boundaries in human genes were based on Ensembl gene annotations which provided the chromosomal start and end positions of the exons. Here, we use the term exon-intron junction or splice junction interchangeably, to refer to the 60bp region flanking the 3' and 5' intron boundaries (Figure 4.1). Our investigation focused on this region because it is where most of the known *cis*-acting splicing regulatory sequences are sited (Mount, 1982; Tian & Maniatis, 1994; Burge *et al.*, 1999; Brudno *et al.*, 2001).

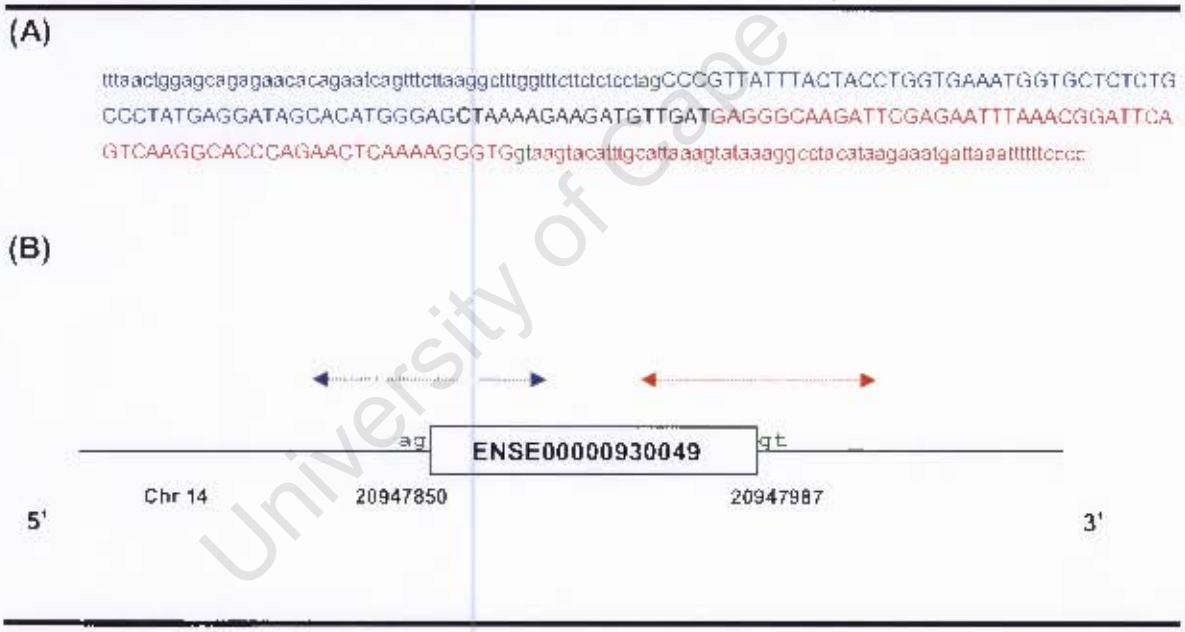

Figure 4.1: Illustration of genomic mapping of exon-intron junctions (A) Genomic sequence of exon 10 (ENSE00000930049) and 60nt flanking intronic sequences of the CHD8 gene. Exon sequences (138 nt) are shown in upper cases and intronic sequences are shown in lower cases with 5' and 3' splice sites dinucleotides highlighted in green. Blue highlights the exon-intron junction near acceptor (3') splice sites and red highlights exon-intron junction near donor (5') splice sites. (B) Graphical representation of the exon (box) and flanking introns (line) with 5' and 3' splice sites dinucleotide highlighted in green. Blue and red double arrow shows genomic mapping of exon-intron junctions flanking the 3' and 5' intron boundaries, respectively. Chromosomal location of the exons is shown according to the positive strand (5' to 3' direction)

4.3.2 Frequency of SNPs in the exon-intron junctions of human exons

Both exon and SNP annotations were based on NCBI 35 genome assembly, which enabled us to easily map the SNPs to the splice junctions. We initially used all annotated SNPs (RefSNPs) to estimate the distribution of polymorphisms as a function of distance from exon-intron junctions; however, we later considered a SNP dataset with only validated entries. We inspected the human genomic DNA for occurrence of polymorphisms in the entire ±60bp sequence flanking the exon boundaries. We further compared the distribution of SNPs between the exon-intron junctions of CSEs and ASEs. A combination of SQL and PERL scripts were used to download and store the data. PERL scripts were also used to compute the frequency of occurrence of SNPs as a function of distance from the exon boundaries.

4.3.3 Statistical analysis of SNP distribution at ASEs and CSEs

We compared the frequency of SNPs at exon-intron junctions of alternatively spliced exons (ASEs) and constitutively spliced exons (CSEs). Data from the survey were entered into Microsoft Excel and then imported into R statistical computing software version 2.1.1 (Ihaka & Gentleman, 1996) for analysis. We used chi-square test to compare the proportions of SNPs in the two groups, to test whether is there a difference in the proportion of SNPs at splice junctions of CSEs and ASEs.

4.4 Results and Discussion

4.4.1 Frequency of SNPs in the exon-intron junctions of human exons

To estimate the prevalence of SNPs at exon-intron junctions of human genes we computed the SNP frequency distribution based on the observed occurrence of polymorphisms as a function of distance from the exon boundaries. A total of 47 036 SNPs were found at splice junctions of approximately 40 000 exons from 14 921 Ensembl genes. This is approximately 1% of total validated SNPs found in dbSNP (version 125). This number might be an under estimate because of stringent filtering during SNP identification and genotyping. Of the identified SNPs, 23 367 were found in the exon-intron junctions near the donor splice sites (Figure 4.2a and Figure 4.2b) whereas 23 669 were found in the exon-intron junctions near the acceptor splice sites (Figure 4.2c and Figure 4.2d). There are few cases where a SNP maps to both 5' and 3' splice junctions of different exons possible due to it being found in splice junctions of two exons separated by an intron shorter than 60bp in length.

The splicing regulatory sequences flanking the 3' splice site are more complex than near the 5' splice sites, largely due to additional sequence constraints required for recognition of acceptor splice sites compared to donor splice sites. We would then expect lower prevalence of SNPs around the genomic region near acceptor splice sites compared to the regions near splice donor sites. However, we detected no significant difference in the average number of SNPs located in the splice junction regions near 5' splice sites compared to near 3' splice sites even when only validated SNPs were used for the analysis ($p = 0.1591$). The reason might be even though the genomic regions harboring the 5' splice sites are shorter and less complex than regions defining 3' splice sites they are similarly important in the regulation of splicing. According to the established "exon definition" model of splice site selection during pre-mRNA splicing, interactions of the spliceosome components with the 5' end of the exon initiate the splicing reaction in vertebrate pre-mRNAs (Robberson *et al.*,1990). This might help to explain the similar

distributions of SNPs in the splice junction regions near 5' splice site compared to near 3' splice site

We observed that the average number of SNPs at exonic regions of the splice junctions near the donor splice sites is lower than the average number of SNPs at the intronic regions when using both validated SNPs and all RefSNPs datasets ($p= 7.269 \times 10^{-13}$, $p =6.785 \times 10^{-13}$, respectively).The same results were observed when considering the splice junction near the acceptor splice site using both validated SNPs and RefSNPs entries ($p= 1.251 \times 10^{-13}$, 5.882×10^{-12} respectively). The significance of the observed difference between the average numbers of SNPs at the intronic verses exonic regions was assessed using a student t-test. These results are in agreement with previous findings showing that coding regions have less sequence variation than non-coding regions due to functional constraints in coding regions (Balasubramanian *et al.*, 2002; Zhao *et al.*, 2003).

We extended the analysis by looking at the distribution of SNPs within the coding parts, the exonic region, of the splice junction. We observed that SNPs occur less frequently near the splice sites positions and the frequency of SNPs become relatively higher as we move away from the splice site positions. The number of SNPs per site portrays the site variability. The higher the number of SNPs in genomic position, the less conserved is the position. Our observations support findings which suggests occurrence of conserved exonic splicing regulatory sequences (i.e. ESE) at highest frequency near exon boundaries, and less frequently as you move away from the exon boundaries (Majewski & Ott 2002; Fairbrother *et al.*, 2004). In addition, there is also evidence of codon usage bias for amino acids that occur frequently in exonic splicing regulatory motifs (i.e. ESE) near the exon boundaries (Willie & Majewski 2004; Parmley & Hurst, 2007). Therefore, it is likely that many nucleotide substitutions that occur close to exon boundaries are under selective pressure to conserve the pattern and function of splicing regulatory elements found close to the exon boundaries.

We also observe that the donor (GT) and acceptor (AG) splice sites dinucleotides at positions +1, +2 (Figure 4.2b) and -2, -1 (Figure 4.2d) respectively have the lowest occurrences of SNPs compared to the other positions in the splice junctions. Approximately 333 validated SNPs are located in the canonical 5' and 3' splice sites from 330 Ensembl annotated exons were identified. A number of these SNPs might alter RNA splicing patterns, consequently causing disease. For example the acceptor splice SNP (rs10774671) found in the 2',5'-oligoadenylate synthetase (OAS1) gene is strongly associated with susceptibility to viral infection (Bonnevie-Nielsen *et al.*, 2005). The decreased frequency of SNPs in the intronic dinucleotides positions compared to exonic regions reflects the more severe effect of splice site mutations on gene function compared to non-synonymous mutations that alter individual amino acids. As a result, the occurrence of polymorphisms in the coding regions is better tolerated than in the splicing regulatory sites.

While there is significant evidence of nucleotide conservation in splice site regions, deviation from consensus splice sites naturally exists among human introns (Burset *et al.*, 2000; Tanaraj & Clark, 2001). For example, an analysis of splice sites in mammalian genomes showed consistent exception to the GT-AG rule where AT-AC and GC-AG splice site pairs are observed (Burset *et al.*, 2000). These cases demonstrate normal transcript expression patterns. We observed an increased occurrence of SNPs at guanine positions of both splice sites. This is mostly observed when all dbSNP entries are used although the validated SNP data also shows the same SNP distribution. These observations were not expected since these positions determine the splice sites marking the intron boundaries and are believed to be highly conserved. According to Vorechovsky (2006) the higher number of polymorphisms in the G than A splice acceptor position is mainly due to the higher proportion of AG creating mutations (cryptic splice sites). We assume the increased occurrence of SNPs at guanine (G) positions of the splice donor and acceptor sites suggest that the most frequently introduced base is guanine during splice site creation. However, we also acknowledge that there might be no biological explanation for the observed increased frequency of SNPs at G positions at the

intron ends rather than sequencing errors in some of the SNP entries submitted in dbSNP as suggested by Platzer and Hiller (2006).

Out of the 47 036 SNPs located in the exon-intron junctions, we identified several examples where the SNP alleles have already been associated with alternative splice patterns, thus contributing to transcript variation among individuals. These include a SNP (rs4072037) in exon 2 of episialin gene (MUC1) which results in activation of cryptic splice site eight nucleotides downstream of the canonical acceptor splice site (Ligtenberg *et al.*, 1991). The naturally occurring allelic differences of the SNPs dictates the choice between the two acceptor splice sites, consequently two splice variants of the exons are produced. Other examples include alleles at exon-intron junctions of RNA binding motif protein 23 (RBM23), transporter 2 ATP binding (TAP2), Poly [ADP-ribose] polymerase 2 (PARP2) CD 46 antigen (CD46), endoplasmic reticulum aminopeptidase 2 (ERAP2), and interferon regulatory factor 5 (IRF5) genes which also favors different splicing isoforms (Hull et al., 2007; Kwan *et al.*, 2008). All these examples are in contrast with other alternative splice events that are either developmental or tissue regulated. Therefore, the term allele specific splicing is more appropriate for the splicing of these genes.

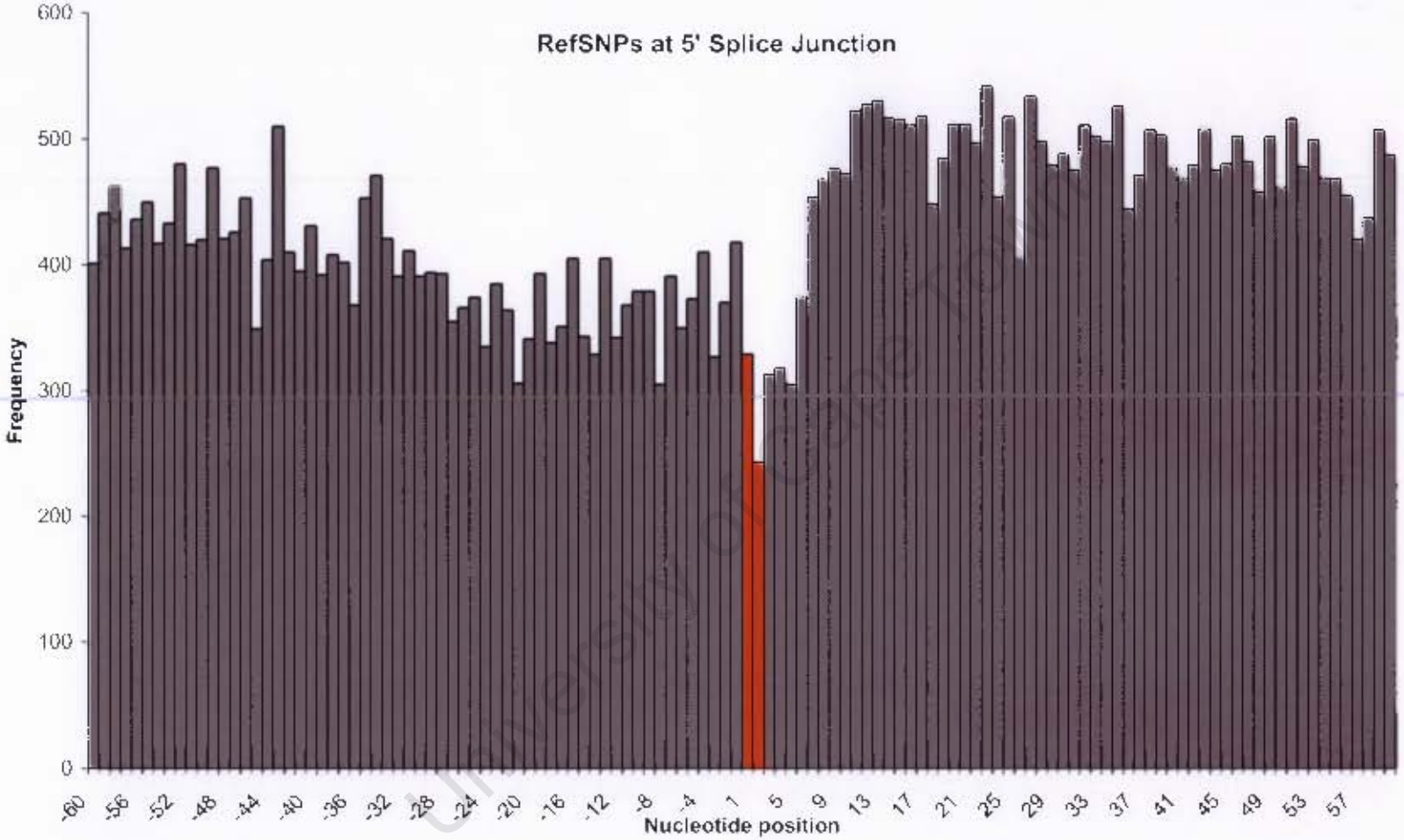

Figure 4.2a : Distribution of RefSNPs at 5' splice junction. RefSNPs include all genetic variation submitted in dbSNP. Distance from the exon boundaries are indicated by negative (upstream) and positive (downstream) values. Red highlights the position of the splice sites dinucleotides located at the intron termini.

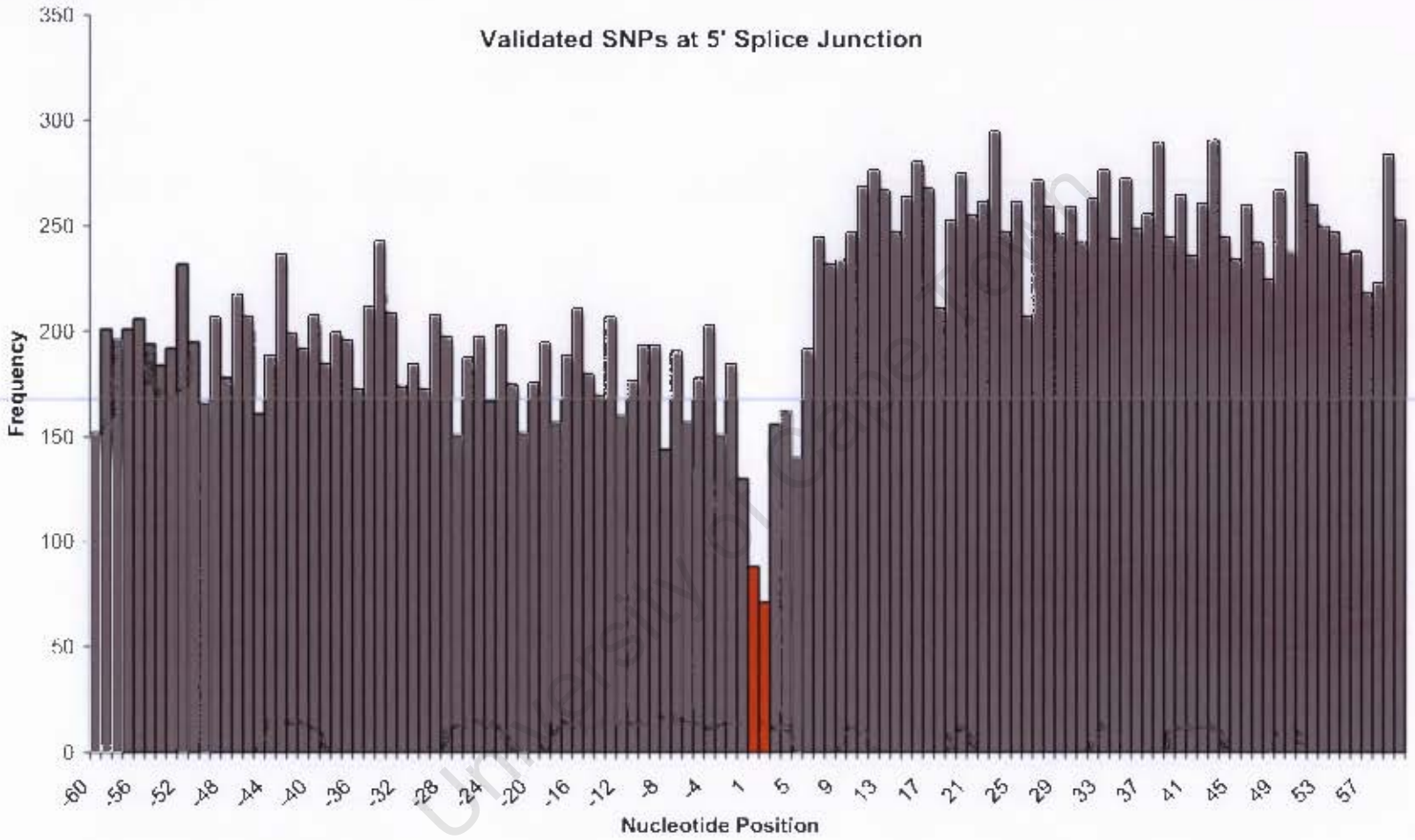

Figure 4.2b : Distribution of validated SNPs at 5' splice junction Validated SNPs only include SNPs confirmed using non-computational methods or has allele frequency data available. Distance from the exon boundaries are indicated by negative (upstream) and positive (downstream) values. Red highlights the position of the splice sites dinucleotides located at the intron termini

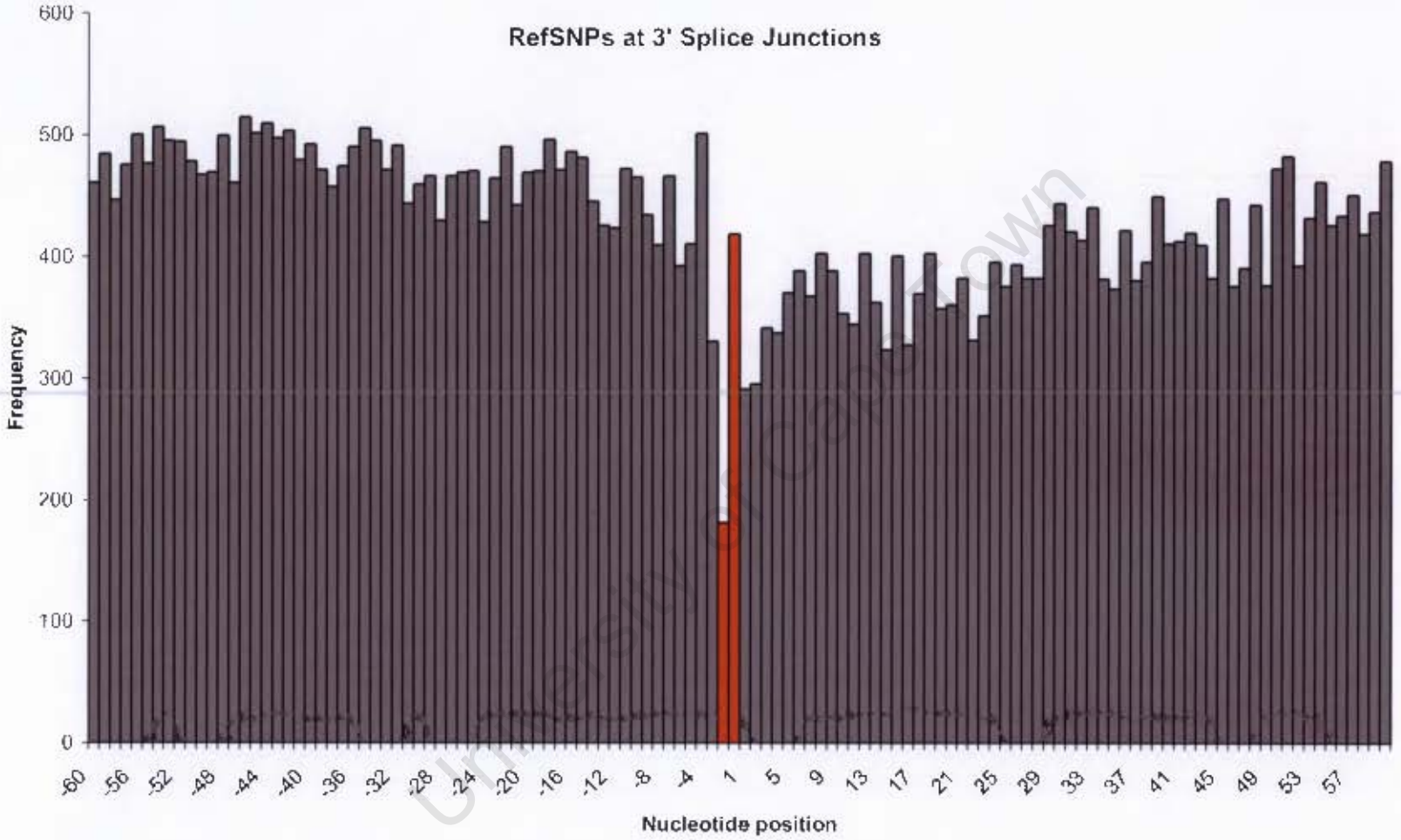

Figure 4.2c : Distribution of RefSNPs at 3' splice junction. RefSNPs include all genetic variation submitted in dbSNP. Distance from the exon boundaries are indicated by negative (upstream) and positive (downstream) values. Red highlights the position of the splice sites dinucleotides located at the intron termini.

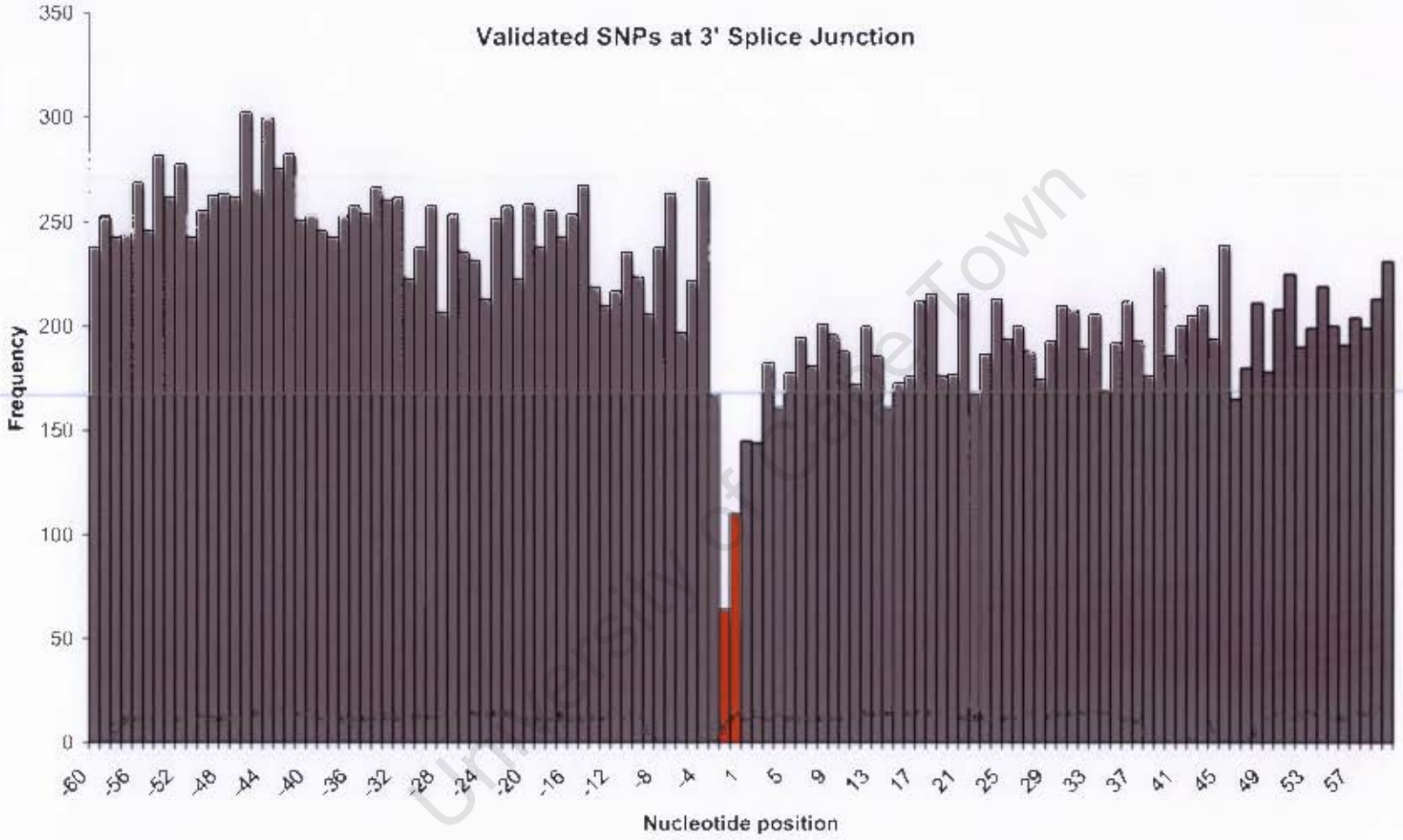

Figure 4.2d : Distribution of validated SNPs at 3' splice junction. Validated SNPs only include SNPs confirmed using non-computational methods or has allele frequency data available. Distance from the exon boundaries are indicated by negative (upstream) and positive (downstream) values. Red highlights the position of the splice sites dinucleotides located at the intron termini.

4.4.2 SNP distributions in the exon-intron junctions of CSEs and ASEs

We categorized the Ensembl exon dataset into constitutively spliced exons (CSEs) and alternatively spliced exons (ASEs), and then computed the frequency distribution of SNPs as a function of distance from the exon boundaries. There were no SNPs which mapped to exon-intron junctions of both CSEs and ASEs in the analysis. We observed that generally ASEs have lower frequencies of occurrence of SNPs at their exon-intron boundaries with higher conservation of splice site sequences compared to CSEs (Figure 4.3a and Figure 4.3b). P-values, with Yates' continuity correction, were determined using chi-square test (Table 4.1a and Table 4.1b). We considered p-values less than (<) 0.05 as indicative of statistical significance. Nevertheless there may possibly be no significant difference between the frequencies of SNPs in the exon-intron junctions of these exons category. There are genomic positions which show a higher rate of occurrence of SNPs at ASE junctions than the corresponding regions in CSEs. Such observations are seen in genomic positions found more than 40bp away from the donor splice site (Table 4.1a).

Our findings are in some way contradictory to what has already been suggested for ASEs, that ASEs are relatively short and are more tolerant of polymorphisms at their splice site and are influenced by weak selective constraints than CSEs (Dye *et al.*, 1998; Stamm *et al.*, 2000; D'Souza & Schellenberg, 2002; Clark & Thanaraj, 2002). Indeed our dataset of ASEs have relatively shorter exon length, however they have fewer SNPs on splice sites and displaying increased exonic and flanking intronic sequence conservation compared to CSEs. The latter is a characteristic that has only been associated with conserved alternatively spliced exons (Sorek & Ast, 2003; Baek & Green, 2005, Yeo *et al.*, 2005). These results might be an indication that our ASEs dataset is composed mostly of conserved ASEs with relatively few human specific ASEs. Alternatively, the results might be an indication that generally ASEs are flanked by conserved sequence features subjected to more functional constraints than CSEs. The low frequency of SNPs close to exon boundaries of

ASEs compared to boundaries of CSEs supports enrichment of conserved exonic splicing regulatory motifs in ASEs compared to CSEs. Intronic splicing regulatory elements associated with tissue specific splicing factors have been observed at higher density near conserved ASEs than CSEs (Brudno *et al*, 2001; Voelker & Berglund, 2007; Yeo *et al*., 2007).

Our observations draw attention to the possibility that the various types of ASE have diverse regulatory mechanisms and evolutionary paths (Mondrek & Lee, 2003; Zheng *et al*., 2005; Chern et al., 2007); consequently they are characterized by different biological properties and features. It is possible that the discrepancy in the observed frequency of SNPs between the splice junction of ASEs and CSEs from what has already been found (Stamm *et al*., 2000; Clark & Thanaraj, 2002; Yeo *et al*., 2005) is due to the different types of ASEs having different patterns of SNP distribution in their splice junctions. For example, complex and simple ASEs might have contrasting frequency of SNPs in their exon-intron junctions compared to CSEs, as they have opposing selection pressure to preserve protein reading frame (Chern & Chaung, 2007). Apart from the two specified exon categories, CSEs and ASEs, it was necessary to further specify the splicing pattern of ASEs studied before we conduct a conclusive analysis of the frequency of SNPs at exon-intron junctions of ASEs compared to corresponding regions of CSEs. Therefore, we suggest that without consideration of the distinct ASE patterns comparison of the prevalence of SNPs at exon-intron junctions of ASEs with CSEs will allow only tentative interpretation of results. Therefore, further analysis is required.

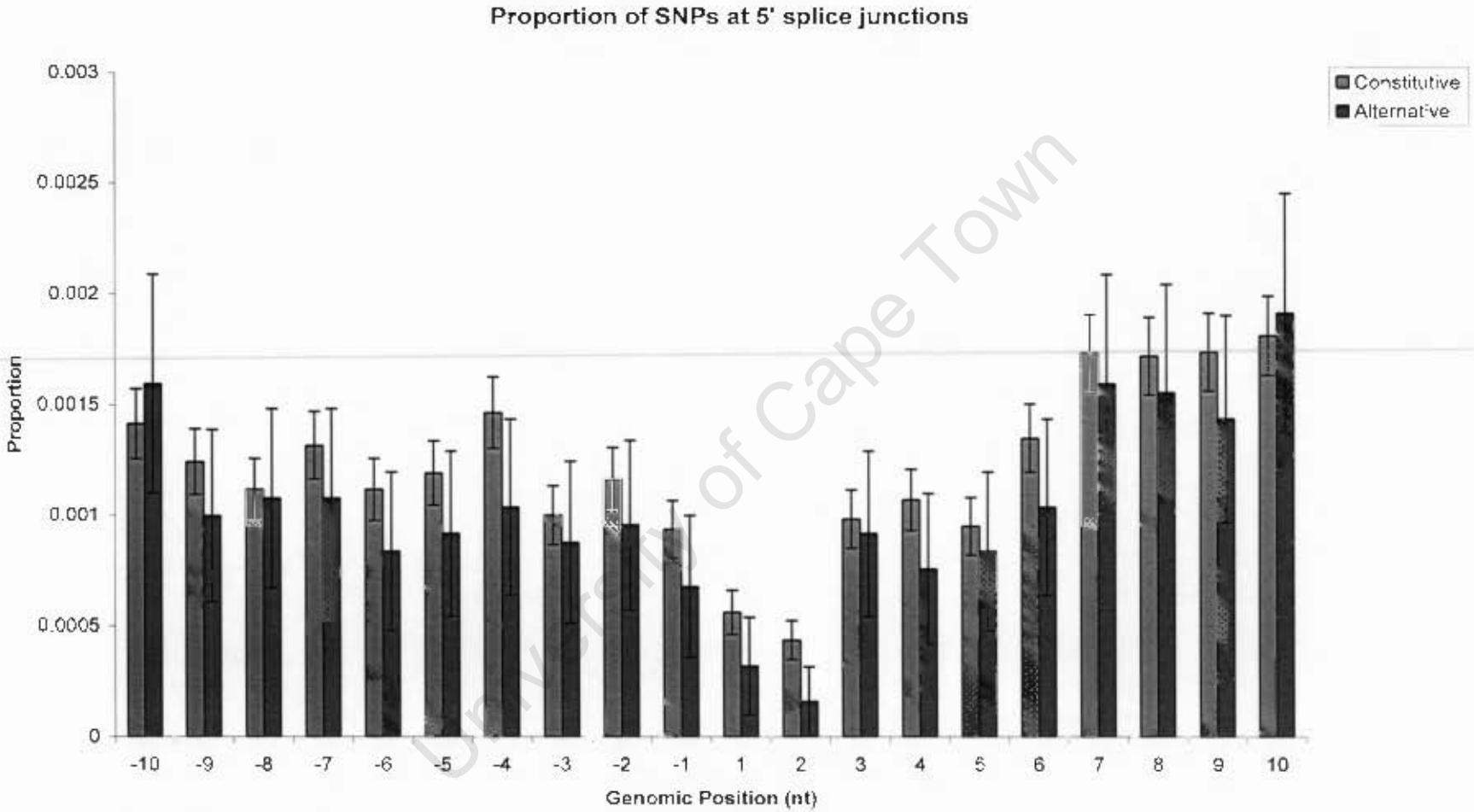

Figure 4.3a: Distribution of validated SNPs at 5' splice junctions of constitutively and alternatively spliced exons.The height of bars represent proportion and error bars represents standard error of proportion (95% confidence limit). ASE have lower frequency of SNPs than CSE.

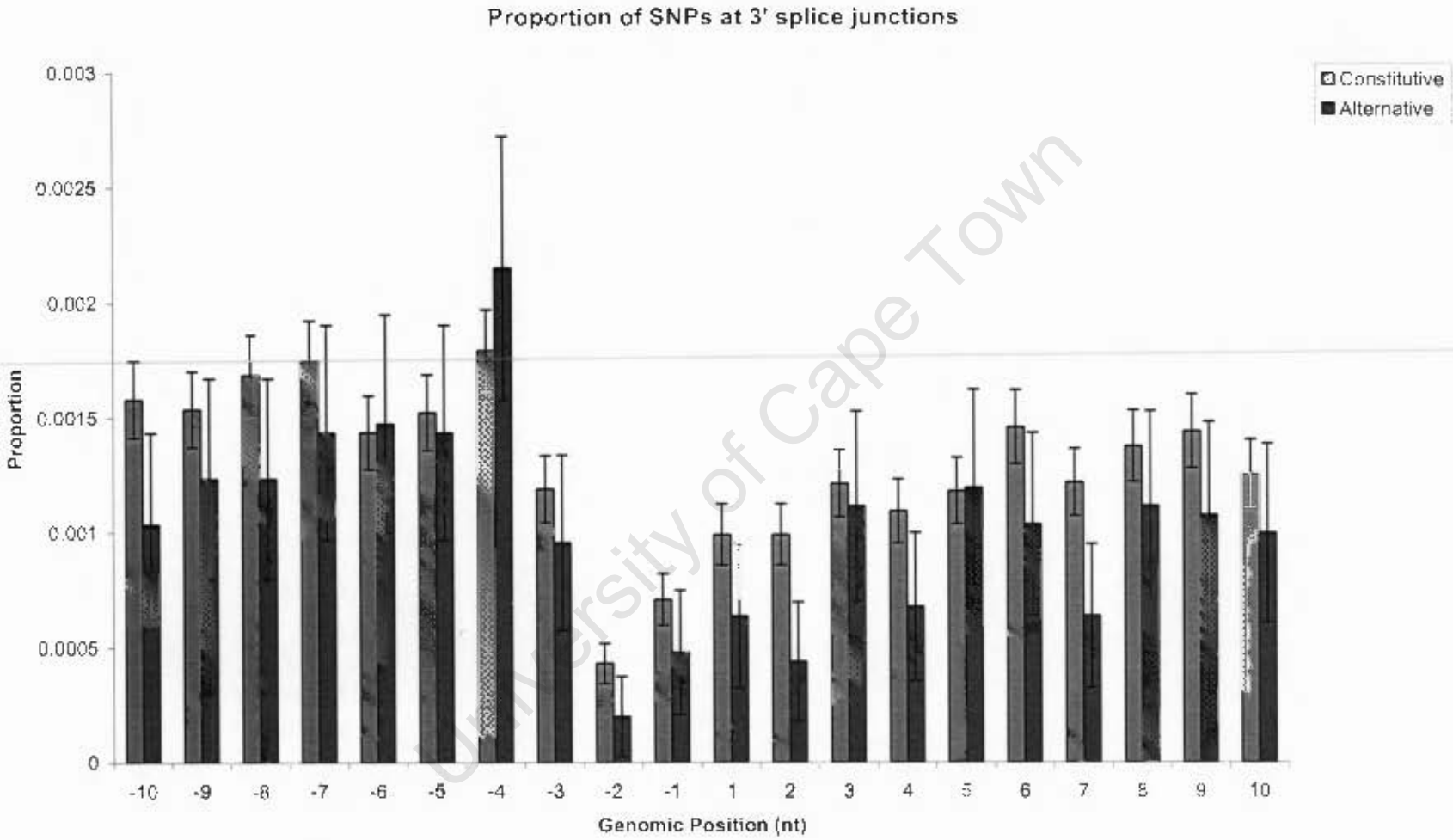

Figure 4.3b: Distribution of validated SNPs at 3' splice junctions of constitutively and alternatively spliced exons. The height of bars represent proportion and error bars represents standard error of proportion (95% confidence limit). ASE have lower frequency of SNPs than CSE.

Table 4.1a: Analysis of SNP density at 5' splice junctions of ASEs and CSEs

Genomic Position (nt)	Exon Type	*RefSNPs*		*Validated SNPs*	
		Proportion	p-value	Proportion	p-value
-60 to – 41	Constitutive	0.0609	< 0.05	0.0279	< 0.05
	Alternative	**0.0503**		**0.0212**	
-40 to – 21	Constitutive	0.0556	< 0.05	0.0270	< 0.05
	Alternative	**0.0468**		**0.0208**	
-20 to -1	Constitutive	0.0516	< 0.05	0.0253	< 0.05
	Alternative	**0.0428**		**0.0214**	
1 to 20	Constitutive	0.0642	< 0.05	0.0324	< 0.05
	Alternative	**0.0550**		**0.0308**	
21 to 40	Constitutive	0.0707	< 0.05	0.0374	< 0.05
	Alternative	**0.0643**		**0.0353**	
41 to 60	Constitutive	0.0689	< 0.05	**0.0365**	< 0.05
	Alternative	**0.0677**		**0.0384**	

Table 4.1b: Analysis of SNP density at 3' splice junctions of ASEs and CSEs

Genomic Position (nt)	Exon Type	*Ref SNPs*		*Validated SNPs*	
		Proportion	p-value	Proportion	p-value
-60 to – 41	Constitutive	0.0703	< 0.05	0.0384	< 0.05
	Alternative	**0.0627**		**0.0340**	
-40 to – 21	Constitutive	0.0677	< 0.05	0.0353	< 0.05
	Alternative	**0.0606**		**0.0317**	
-20 to -1	Constitutive	0.0626	< 0.05	0.0317	< 0.05
	Alternative	**0.0536**		**0.0280**	
1 to 20	Constitutive	0.0513	< 0.05	0.0265	< 0.05
	Alternative	**0.0399**		**0.0191**	
21 to 40	Constitutive	0.0570	< 0.05	0.0281	< 0.05
	Alternative	**0.0451**		**0.0192**	
41 to 60	Constitutive	0.0597	< 0.05	0.0288	< 0.05
	Alternative	**0.0513**		**0.0223**	

4.5 Conclusions

Approximately 1% of the total validated SNPs in dbSNP are located within ±60bp of exon - intron boundaries of human genes. The SNPs are mostly located in the non-coding intronic regions rather than the coding exonic regions. We expect the conserved exonic regions have fewer SNPs due to purifying selection which prohibits occurrence of deleterious SNPs in functional regions. The donor (GT) and acceptor (AG) splice site dinucleotide positions, defining the intron terminals; have the lowest prevalence of SNPs in the exon-intron junctions.

We observed that ASEs have lower proportions of SNPs in their exon-intron junctions than the CSEs. The difference in the proportion of SNPs within exon-intron junctions of ASEs compared to CSEs is highly significant. Nevertheless, these results might be influenced by sampling bias of data from the survey. It is almost impossible to directly predict the alternatively spliced exon patterns associated with each SNP in the exon-intron junction, using only genomic sequences. A number of SNPs in our dataset have already been experimentally validated and associated with allele specific splicing in human protein coding genes. These include RefSNPs, rs2297616 and rs2295682, found in exon-intron junctions of PARP2 and RBM23 genes, respectively. It is highly possible that numerous SNPs found in exon-intron junctions of ASEs affect the way in which mRNAs are spliced, thus, contributing to alternatively spliced transcript patterns.

We suspect the difference between proportions of SNPs in splice junctions of CSEs and ASEs might be, to a certain extent, due to limitations in our method of annotating the exons which contribute to large difference between sample sizes of the two exon categories. Alternatively spliced exons are annotated based on availability and/or representation of exons in sequenced transcripts and our method did not distinguish between the various types of ASEs. It is possible that many of the constitutively spliced exons are actually alternatively spliced exons but there are no

currently available transcripts (i.e. ESTs or mRNA) in the ASAPII dataset to indicate that they are indeed alternatively spliced.

CHAPTER 5

ANALYSIS OF FREQUENCY SPECTRA OF SNPS AT EXON-INTRON JUNCTIONS

5.1 Overview

This chapter aims to understand the role of natural selection in shaping the distribution of SNPs at exon-intron junctions of human genes. It addresses the objective by using the derived allele frequency (DAF) distribution. We used SNP allele frequencies from the HapMap database to compare the derived allele frequency distribution between SNPs located in exon-intron junctions, coding non-synonymous sites, coding synonymous sites and SNPs located in non-coding regions. We further compare the DAF spectrum of SNPs located in exon-intron junctions of constitutively spliced exons (CSEs) and alternatively spliced exons (ASEs). Analysis was done using SNP frequency data from populations of Asian, European and African descent.

5.2 Background

Advances in the human genome project (Lander *et al.*, 2001; Venter *et al.*, 2001) and its subsequent analysis (Zhoa *et al.*, 2003; Majewski & Ott, 2002) have shown that SNP density varies in different regions across the human genome. The rate of accumulation of genetic variation has been shown to vary significantly between the various regions of the human genome, such as protein coding regions, non-coding regions, exon, introns, and flanking gene expression regulatory regions (Chakravarti, 1999; Serre & Hudson, 2006). Non-coding intronic DNA harbors a much higher number of SNPs than the functionally important coding regions. Similarly, this idea is shared by the results we obtained when we looked at the

frequency of SNPs at exon-intron junctions (Figure 4.2a-d). A recent study on conserved verses non-conserved nucleotide positions between human and non-human primates showed that nucleotide diversity within conserved non-coding sequences is reduced compared with species specific non-coding sequences (Asthana *et al.*, 2007). Therefore, SNP distribution in various regions of the human genome is not random, it follows specific patterns, depending on functional constraints on sequences, with coding non-synonymous sites reported to be less variable than non-coding regions (Cargill *et al.*, 1999; Hughes *et al.*, 2003).

More than one allele (nucleotide variant) may be present at a locus, and the proportion of an allele in a population is referred to as the **allele frequency** (Li, 1997; Nei & Kumura, 2000). Allele frequencies change over time due to random drift and/or the relative fitness of the alleles (Hartl & Clark, 1997). The latter means that a SNP allele associated with disease is expected to be kept at low frequency in a population (assuming that the disease phenotype leads to reduction of fitness) whereas an allele that serves to increase the fitness of the organism is likely to increase in frequency over time. The strong conservation of functionally important genomic regions and selection of alleles based on fitness in the genomic DNA is evidence that various regions of the human genome are subjected to natural selection more than other regions, and this contributes to variation in SNP distribution among the different regions of DNA.

Random mutation and natural selection are both believed to be determinants of pattern of SNP distribution in the human genome by altering the allele frequency distribution in a population (Lercher *et al.*, 2002; Fay & Wu, 2003; Zhoa *et al.*, 2003). The neutral theory of molecular evolution suggests that the majority of DNA variations in a population are effectively neutral and the fate of neutral alleles is governed by genetic drift of random occurring mutations (Kimura, 1968; King & Jukes, 1969). In contrast, the selectionist position suggests that most new mutations have an effect on fitness and are affected by natural selection (Nei & Kumura, 2000). Mutations that have deleterious effect on fitness are eliminated from the

population through purifying/negative selection, and are thus found in reduced frequencies in natural populations. In contrast, natural selection of advantageous mutations also occurs and this increases the frequency and speed of fixation of a new allele (positive selection). Natural selection can also maintain several alleles over period of time (balancing selection).

If we know which SNP allele is derived (new mutation that has arisen in a population) and which one is ancestral at each site we can make inferences about the type of selection that has affected a set of locii using the allele frequency spectrum. An allele frequency spectrum is a count of the number of mutations in a given frequency class (Nielsen, 2005). The histogram of the number of derived alleles within the different frequency classes is called the **derived allele frequency spectrum** (Figure 5.1). Allele frequency spectra combine information across a large number of polymorphic loci, providing a summary of the allele frequencies of the various polymorphisms. We expect that given a set of SNPs, an excess of rare derived alleles is a signature of purifying selection (Chen & Rajewsky, 2006). Conversely, positive selection will tend to increase the proportion of high frequency alleles.

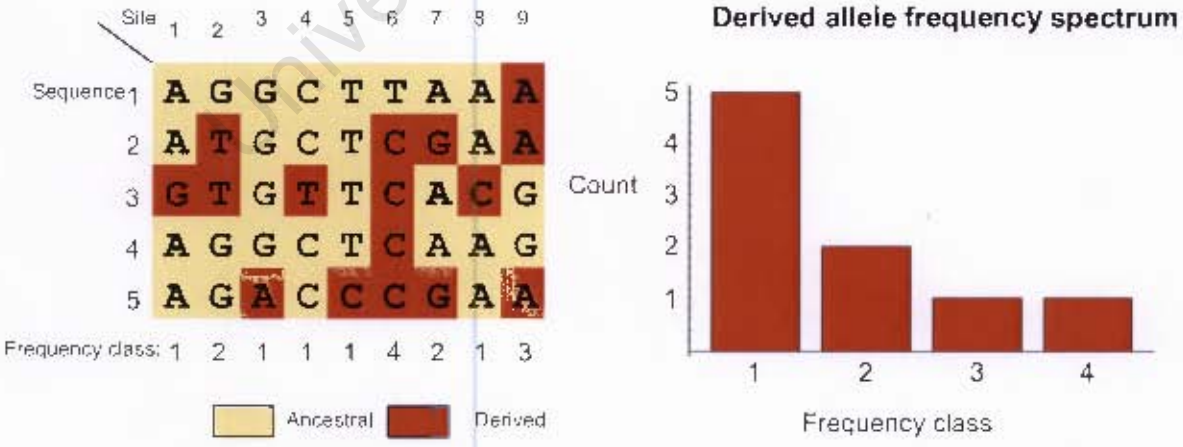

Figure 5.1: An example of a derived allele frequency (DAF) spectrum.

Since natural selection results in changes of allele frequency, changes in a population are usually described by the changes in allele frequencies. Hence, allele frequency has been a fundamental parameter in population genetics in the study of the generation and maintenance of genetic polymorphisms and to understand the mechanisms of evolution at the population level. Here we use allele frequency to study the role of natural selection in shaping the distribution of SNPs at exon-intron junctions of human genes. SNPs around the borders of exons and introns are more likely than other SNPs to have an effect on splicing; and nucleotide substitutions that deleteriously affect splicing are expected to be kept at low frequencies.

5.3 Material and Methods

5.3.1 SNP allele frequency data

SNP allele frequencies compiled from genotype data from European, Asian and African populations were obtained from HapMap (http://www.hapmap.org) release #22. This data is also available in dbSNP. For each human SNP in the initial dbSNP dataset we inferred the ancestral character state of the alleles from genomic alignments of human polymorphism with corresponding chimpanzee nucleotide. We defined the 'ancestral allele' as the human allele identical to the chimpanzee allele and the 'derived allele' as the other human allele. Only bi-allelic, validated SNPs were considered for this analysis. We did not filter for monomorphic SNPs. We were able to define a derived allele for 8 426 200 genotyped SNPs in dbSNP.

Using Ensembl functional annotations, the SNP data were classified according to their genomic positions: coding non-synonymous SNPs, coding synonymous SNPs, and non-coding SNPs. SNPs functional annotations also included bi-allelic SNPs found at the 5' and 3' splice junctions. Non-coding SNPs are located in intergenic, intronic, or UTR regions and SNPs not annotated as any of the other four functional SNP categories.

5.3.2 Derived Allele Frequency (DAF) Analysis

Using SNP frequency information from populations of Asian, European and African ancestry in the HapMap database, we computed the average frequencies of the derived alleles as a function of distance from the exon-intron boundaries. Secondly, we compared the average frequency of the derived alleles of SNPs located in exon-intron junctions with SNPs located in coding non-synonymous sites, coding synonymous sites and SNPs located in non-coding regions. Thirdly, we compare the derived allele frequency distribution of SNPs found at exon-intron junctions of CSEs with those found at exon-intron junctions of ASEs. Finally, we then look at the derived allele frequency spectra of SNPs located in the different SNP categories. The analysis was done using a combination of PERL scripts and the R-statistical package.

5.4 Results and Discussions

5.4.1 Derived allele frequency data

HapMap provides validated DNA sequence variations and allele frequencies from populations with ancestry from Africa, Asia, and Europe. We classified the SNPs into functional categories according to genomic context. Table 5.1 shows that by far the most common SNPs available in the dataset are non-coding SNPs followed by non-synonymous, synonymous and then SNPs located in the exon-intron junctions. Non-coding SNPs include sets of intergenic, intronic, and UTR SNPs which are located within the large genomic regions between the coding regions. The high number of non-coding SNPs in the database is a reflection of the fact that the vast number of SNPs in the human genome lie within the intergenic DNA sequence (Hagmann, 1999; Venter *et al.*, 2001).

Table 5.1: Number of SNPs with derived allele frequency.

SNP Category	Number of SNPs		
	European	Asian	African
5' splice junction	13 220	13 136	13 667
3' splice junction	13 247	13 222	13 795
Non-synonymous Coding	19 345	19 413	19 229
Synonymous Coding	28 657	28 944	28 860
Non-coding SNPs	139 944	140 646	137 723

Numerous studies have highlighted the possibility that SNPs outside the coding regions of the genes might be regulatory SNPs (rSNPs) mostly involved in gene expression variation (Buckland, 2004; Buckland, 2006; Montgomery *et al.*, 2007). On the other hand, studies have shown that the fraction of polymorphisms at different functional regions of the human genome depends on the strength of selection on that region (Gorlov *et al.*, 2006). Thus, a higher rate of occurrence of SNPs in the non-coding regions reflects the weaker or absence of selective pressure affecting much of the non-functional intergenic DNA.

The higher number of synonymous relative to non-synonymous polymorphisms reflects the fact that the human coding sequences have more non-synonymous than synonymous sites. Many nucleotide substitutions in the third codon position are synonymous whereas the first and second codon positions have mostly non-synonymous nucleotide substitutions. Hence, a random new mutation in a coding sequence has a much higher probability of being synonymous than non-synonymous.

5.4.2 Derived allele frequency distribution

Throughout the rest of this chapter we will strive to understand how natural selection affects SNP allele frequencies in the exon-intron junctions of human genes. We examine the derived allele frequency distribution of SNPs located in the entire ±60bp exon-intron junctions (Figure 5.2a and Figure 5.2b). On average the derived allele frequency in the exon-intron junctions is 0.41 (SD = 0.03) in Asian, 0.40 (SD = 0.03) in Europeans, and 0.33 (SD = 0.03) in African populations. Although the number of SNPs observed in the different population groups examined is similar, on average, Asian and European populations have similar distributions frequency of derived alleles, which is significantly higher ($p < 2.2 \times 10^{-16}$) than the frequency of derived alleles in the African population. We observe equal distribution of the average frequency of the derived alleles between the coding exonic and non-coding intronic positions, however, the exonic positions have significantly higher proportion of rare (frequency ≤ 0.1) derived alleles than intronic positions ($p = 0.0003$, Figure 5.3a;$p = 1.96 \times 10^{-7}$, Figure 5.3b). This is shown using data from European population.

The distribution of derived allele frequency observed does not reveal the influence of natural selection on polymorphic sites in the exon-intron junctions. However it does show that derived alleles usually have lower frequency than ancestral alleles. In addition, we observed that the exonic regions of the splice junction are probably enriched with rare derived alleles compared to the intronic regions. Studies on SNP frequencies in human genes showed that non-degenerate regions have low nucleotide diversity and an excess of rare allele compared to degenerate sites (Sunyaev *et al.*, 2000). The excess of rare alleles most likely reflects the presence of deleterious alleles under purifying selective pressure, but could also be explained by the presence of recent selective sweeps (Chen & Rajewsky, 2006).

Derived alleles arise by mutations and the effect of new mutations on gene sequence depends on their location within the human genome. Within the protein

coding regions, synonymous mutations do not change the amino acid sequence of the protein whereas non-synonymous mutations result in a different amino acid being placed in the sequence. Since silent changes normally affect the characteristics of the individual far less than would non-synonymous changes, non-synonymous changes would be more frequently subjected to natural selection (both positive and negative) than would synonymous changes. Most mutations outside the coding regions do not affect the amino acid sequence of genes; however, those changes that are located within gene expression regulatory sequences are more likely to influence gene function. Hence we suspect new mutations in the splicing regulatory regions (exonic and intronic) are also subjected to natural selection.

Harmful mutations rather than beneficial mutations are likely to occur in the splicing regulatory motifs found in the exon-intron junctions, and natural selection is expected to act against these new deleterious mutations (Nei & Kumura, 2000). Hence we suspect negative selection is much more common than positive selection in the exon-intron junctions as this is generally the case anywhere in the functional regions of the human genome.

To test if negative selection operates against new mutations found in the splice sites positions we compared their average frequency of derived alleles with those at

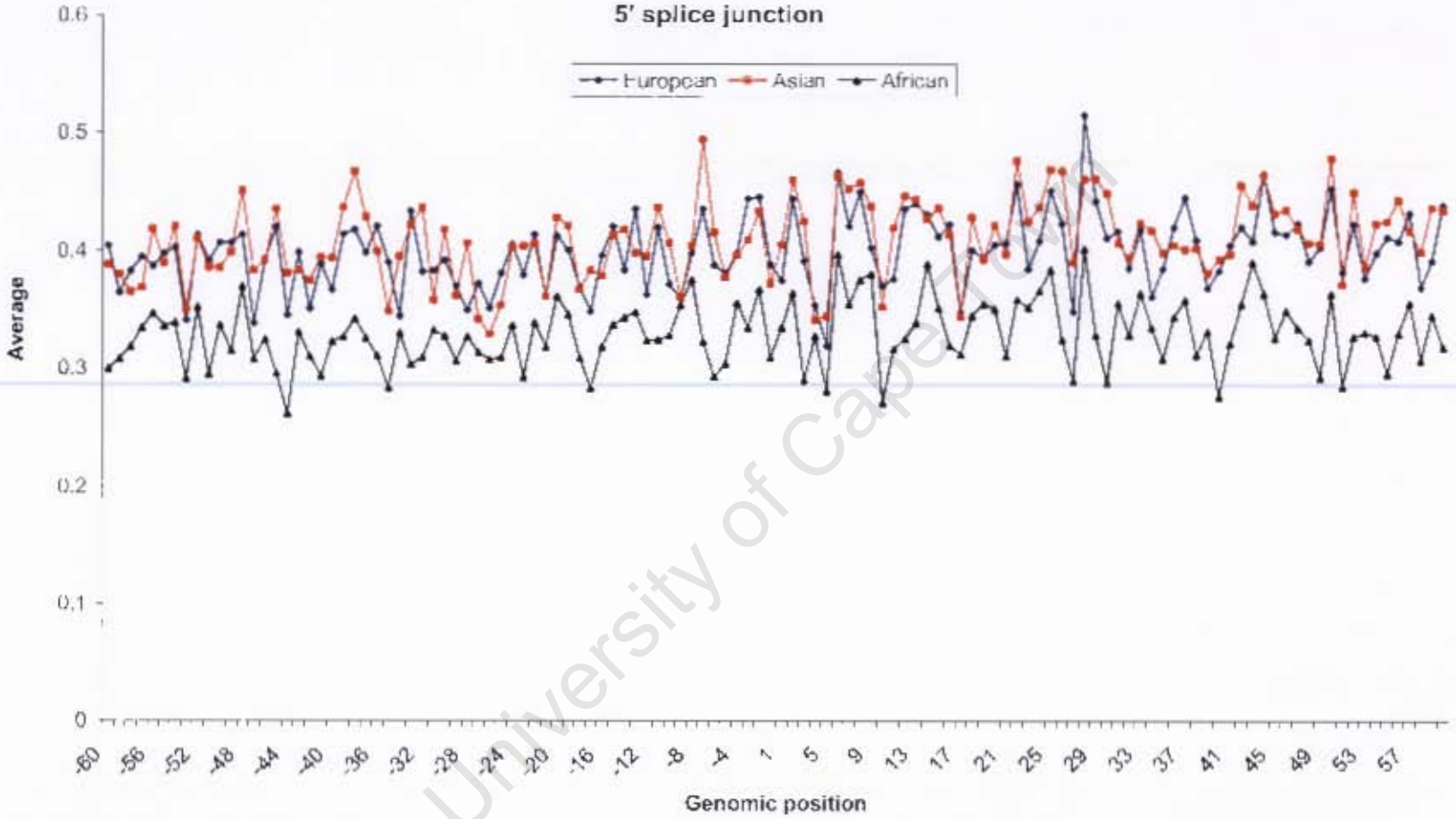

Figure 5.2a: Average frequencies of the derived alleles for SNPs located at 5' splice junctions. Distance from the exon boundaries are indicated by negative (upstream) and positive (downstream) values. The derived allele frequencies are from Europeans, Asians (Chinese and Japanese) and Africans (Yorubans) populations

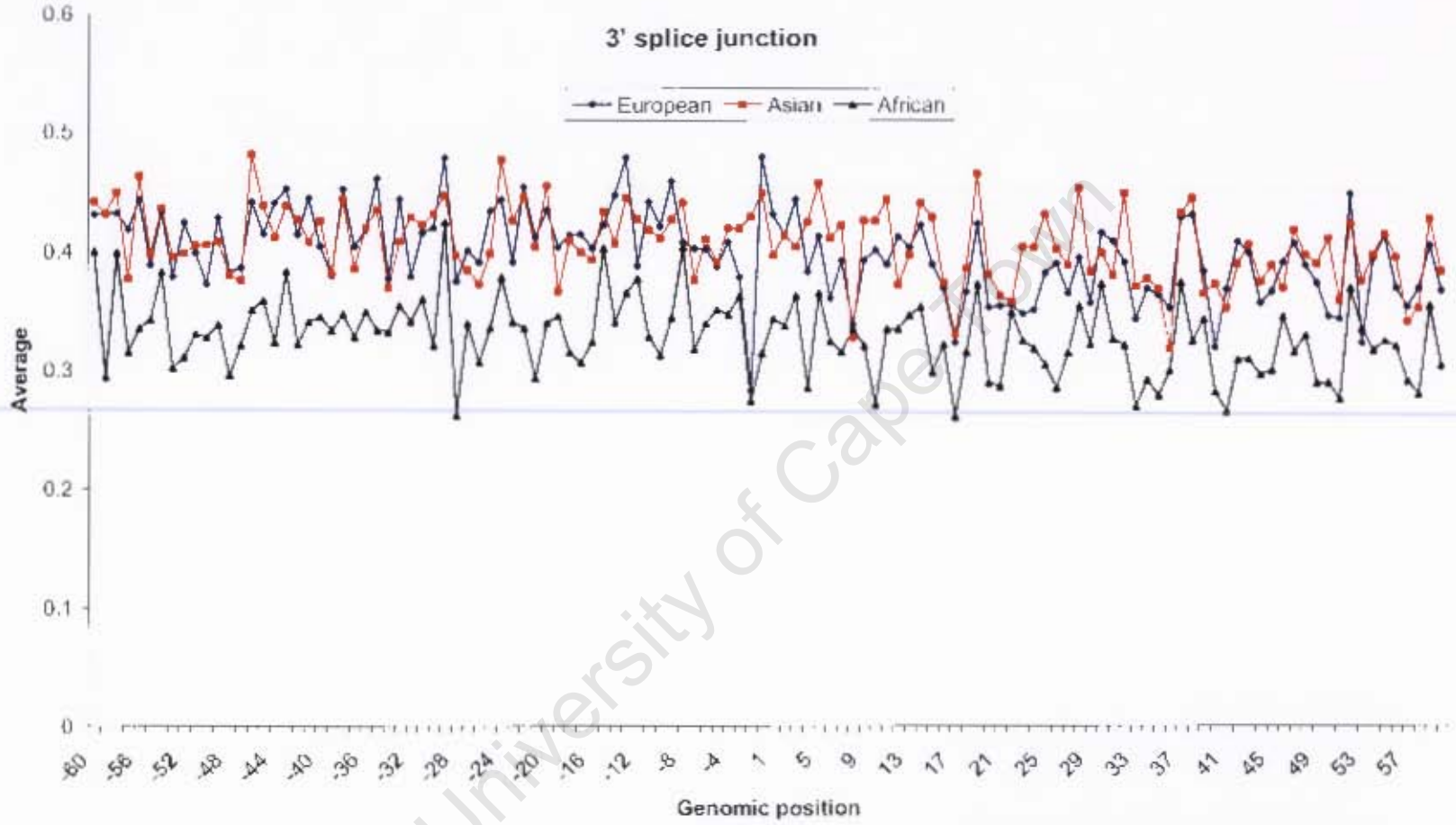

Figure 5.2b: Average frequencies of the derived alleles for SNPs located at 3' splice junctions. Distance from the exon boundaries are indicated by negative (upstream) and positive (downstream) values. The derived allele frequencies are from Europeans, Asians (Chinese and Japanese) and Africans (Yorubans) populations

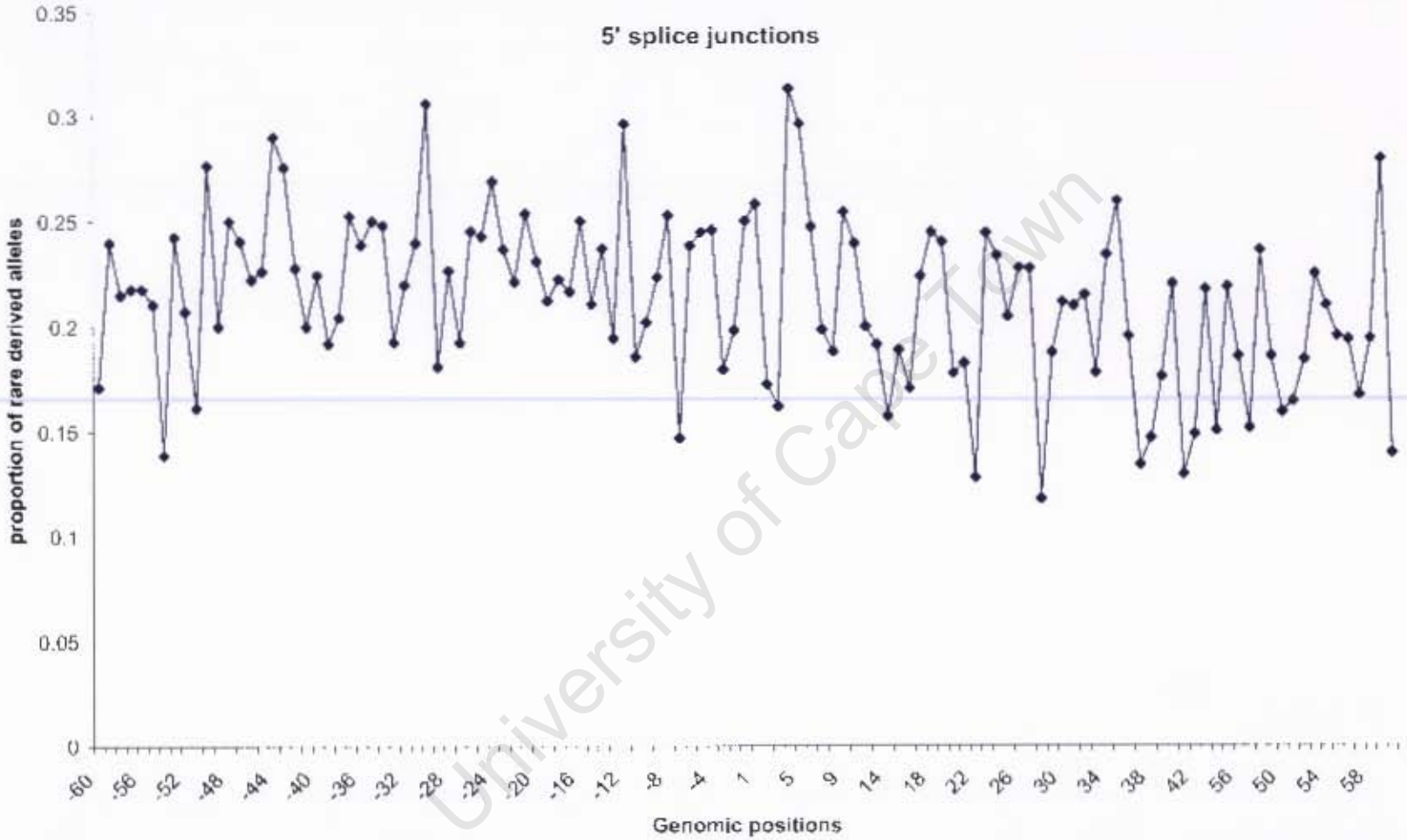

Figure 5.3a: Proportion of rare derived alleles in the European population for SNPs located at 5' splice junctions. Distance from the exon boundaries are indicated by negative (upstream) and positive (downstream) values.

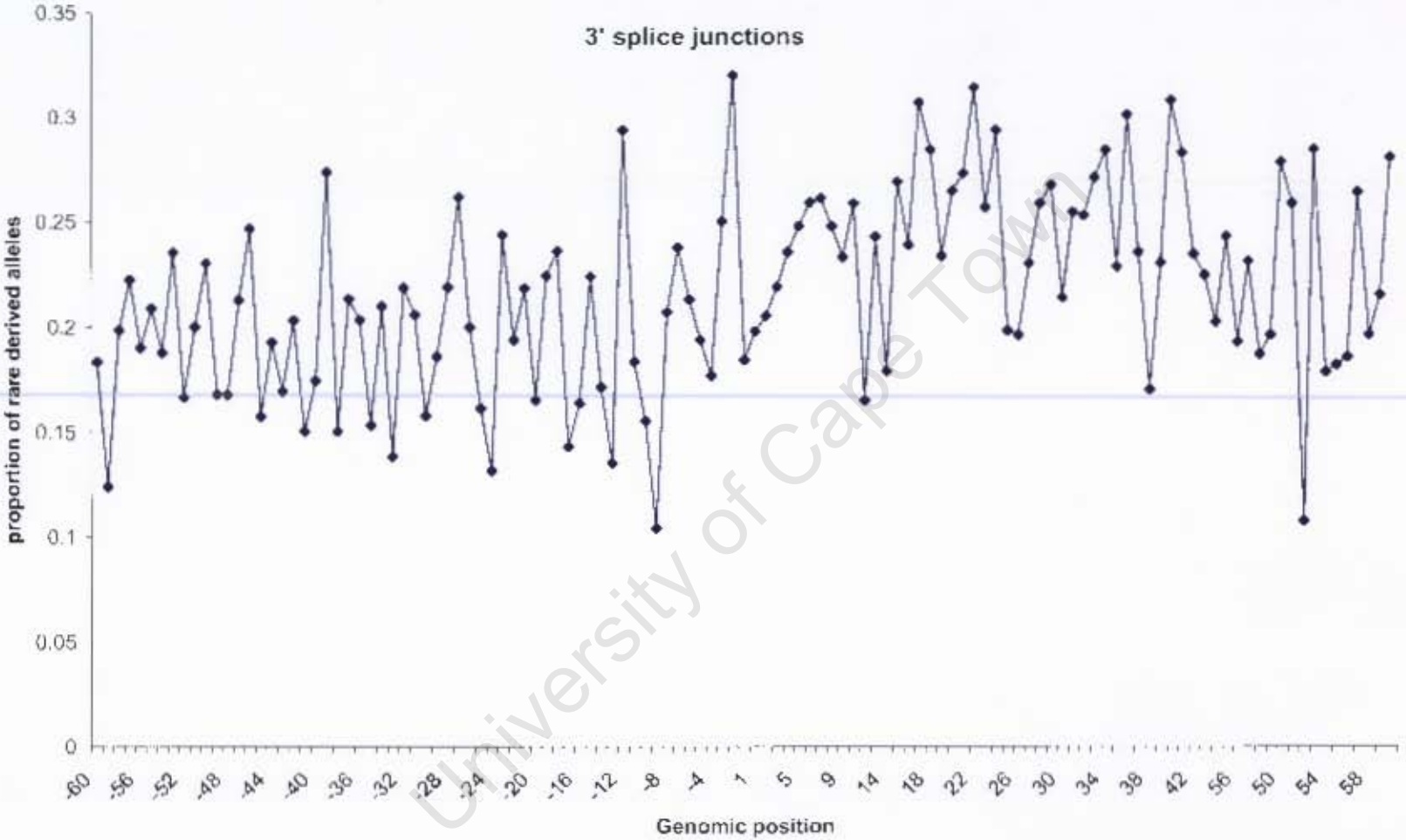

Figure 5.3b: Average frequencies of rare derived alleles in the European population for SNPs located at 3' splice junctions. Distance from the exon boundaries are indicated by negative (upstream) and positive (downstream) values.

coding non-synonymous sites, coding synonymous sites, and SNPs located at non-coding regions of the genome and the results are shown in figure 5.4. Non-synonymous coding SNPs have the lowest average frequencies of derived alleles (p-values in Table 5.2a-c). This is not surprising since new non-synonymous SNPs disrupting the coding capacity of a gene are expected to be kept at low frequencies through purifying selection whereas those that are found at non-coding regions are expected to be under the influence of weaker or no selective pressure.

SNPs located in the splice sites flanking the exon boundary have, on average, increased frequency of derived alleles compared to SNPs located in non-synonymous and synonymous sites (Figure 5.4) but have lower average frequency of derived allele of SNPs in non-coding regions (p-values in table 5.2a-c). SNPs in non-coding regions are unlikely to be under strong selection compared to SNPs in the splicing regulatory regions; hence the observed average frequency of derived alleles of SNPs in the splice sites is understandable. Figure 5.4 suggest the low occurrence of SNPs near the exon boundaries is generally accompanied by higher frequency of the derived alleles than the coding non-synonymous and synonymous SNPs. Unless, positive selection serves to increase frequency of new alleles to improve splicing efficiency, we could not come up with a biological explanation for the increased frequency of the derived alleles in the splicing regulatory sites compared to neutral (synonymous) sites.

The different types of exon in the human genome have been shown to exhibit different evolutionary pressure (Mondrek & Lee, 2003; Zheng *et al.*, 2005; Chern *et al.*, 2006; Chern *et al.*, 2007); consequently they might display different patterns of derived allele frequency. We proposed that the increased frequency of derived allele at exonic positions near splice sites is the indication that new mutations are tolerated in the exon-intron junctions of some exons, especially those exons that are alternatively spliced. In theory, efficient splicing is

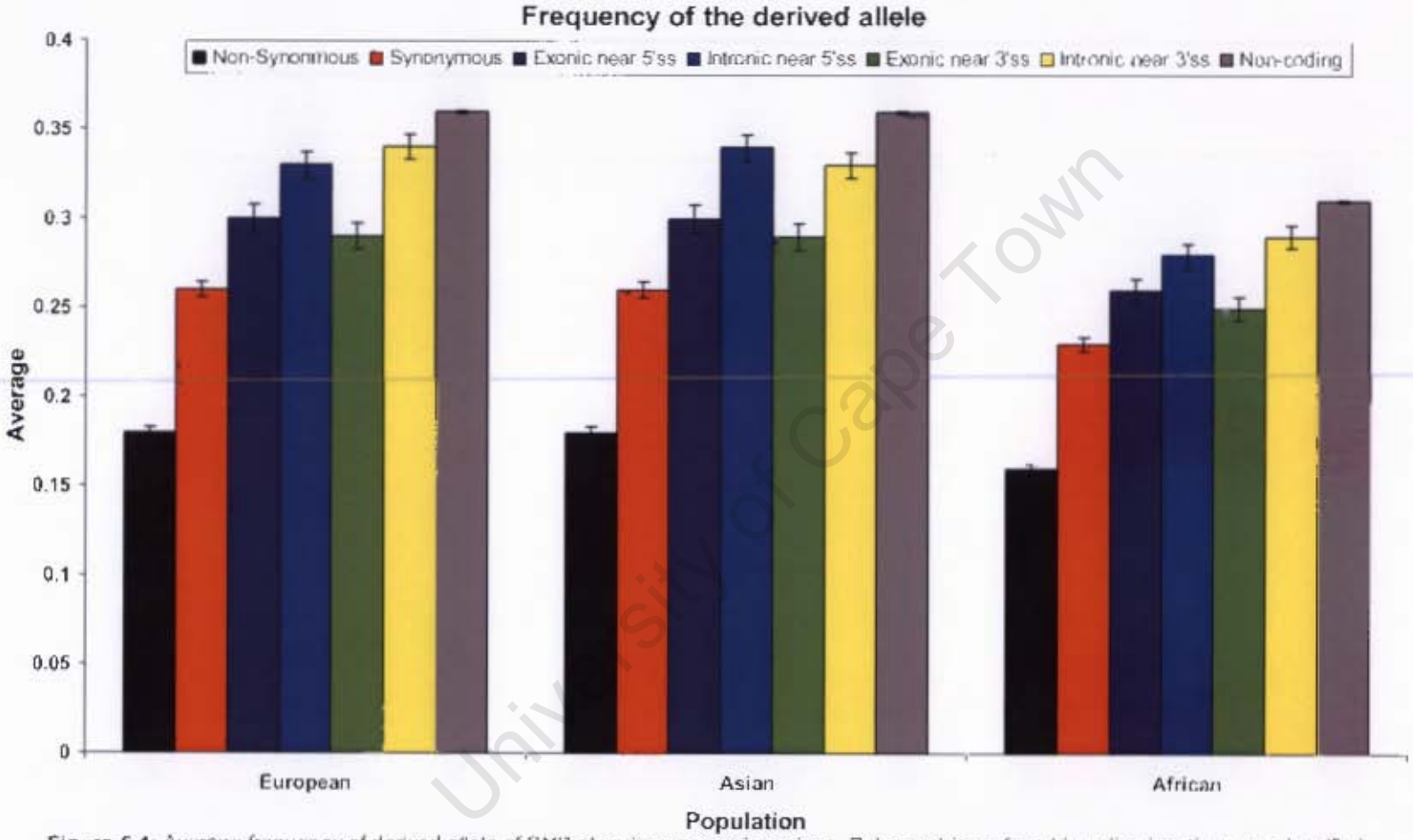

Figure 5.4: Average frequency of derived allele of SNP at various genomic regions. Polymorphisms found in splice junctions are classified according to whether they are located at exonic or intronic regions. Bar height shows average frequency and error bars indicates standard error of a mean

selected against in ASEs. Hence, over evolutionary time new alleles might increase in frequency (and eventually may become fixed if they offer selective advantage in the population), allowing occurrence of SNPs which affect the way in which genes are spliced between individuals. SNPs have already been associated with alternative spliced patterns, allowing transcript diversity among humans (Kwan *et al.*, 2007; Kwan *et al.*, 2008). Alternatively, it is also possible that most SNPs in exon-intron junctions are evolving in an almost complete neutral manner, and those SNPs found at high frequency within a population are mostly due to genetic drift and few are positively selected.

Table 5.2a: Significance (P-values) of comparisons of average frequency of derived alleles using European frequency data.

Genomic Position	NS	S	Exonic 5' SJ	Intronic 5' SJ	Exonic 3' SJ	Intronic 3' SJ	NC
NS	-						
S	**< 0.05**	-					
Exonic 5'SJ	**< 0.05**	**< 0.05**	-				
Intronic 5'SJ	**< 0.05**	**< 0.05**	0.5655	-			
Exonic 3'SJ	**< 0.05**	**< 0.05**	0.6367	0.885	-		
Intronic 3'SJ	**< 0.05**	**< 0.05**	0.3647	0.7987	0.6598	-	
NC	**< 0.05**	**< 0.05**	**0.000184**	**0.0102**	**0.001123**	**0.006626**	-

NS=non-synonymous, S=Synonymous, SJ = splice junction, NC = Non-coding

Table 5.2b: Significance (P-values) of comparisons of average frequency of derived alleles using Asian frequency data.

Genomic Position	NS	S	Exonic 5' SJ	Intronic 5' SJ	Exonic 3' SJ	Intronic 3' SJ	NC
NS	-						
S	**< 0.05**	-					
Exonic 5'SJ	**< 0.05**	**< 0.05**	-				
Intronic 5'SJ	**< 0.05**	**< 0.05**	0.9722	-			
Exonic 3'SJ	**< 0.05**	**< 0.05**	0.9471	0.978	-		
Intronic 3'SJ	**< 0.05**	**< 0.05**	0.5385	0.5345	0.4847	-	
NC	**< 0.05**	**< 0.05**	**0.000055**	**0.000191**	**0.000015**	**0.000781**	-

NS=non-synonymous, S=Synonymous, SJ = splice junction, NC = Non-coding

Table 5.2c: Significance (P-values) of comparisons of average frequency of derived alleles using African frequency data.

Genomic Position	NS	S	Exonic 5' SJ	Intronic 5' SJ	Exonic 3' SJ	Intronic 3' SJ	NC
NS	-						
S	**< 0.05**	-					
Exonic 5'SJ	**< 0.05**	**< 0.05**	-				
Intronic 5'SJ	**< 0.05**	**< 0.05**	0.541	-			
Exonic 3'SJ	**< 0.05**	**< 0.05**	0.7147	0.7804	-		
Intronic 3'SJ	**< 0.05**	**< 0.05**	0.596	0.8968	0.8695	-	
NC	**< 0.05**	**< 0.05**	0.1513	**0.04434**	0.05015	**0.02793**	-

NS=non-synonymous. S=Synonymous, SJ = splice junction, NC = Non-coding

5.4.3 Derived allele frequency (DAF) distribution of SNPs at splice junctions of CSEs and ASEs

The datasets of constitutively spliced exons (CSEs) and alternatively spliced exons (ASEs) allowed us to compare the average frequencies of derived alleles found in splice junctions of these two exon categories. We categorize SNPs in splice junctions according to whether they are located in exonic (non-synonymous sites, synonymous sites, UTR) or intronic regions and compute the average frequencies (Table 5.3a and Table 5.3b). We observed similar distributions of the average frequencies of the derived alleles in the intronic positions between these two exon categories; but, exonic regions show distinct characteristics of derived allele frequencies (DAF). P-values indicating the significance of the difference between DAF in exon-intron junctions of CSEs and ASEs are shown in Table 5.4 (a, b, c).

Generally non-synonymous sites have, on average, a lower frequency of derived alleles compared to the other regions in the splice junctions; however this difference is more often statistically significant in SNPs found in CSEs junctions than corresponding regions of ASEs. Synonymous SNPs in the splice junctions of ASEs

show decreased average frequency of derived alleles, suggesting stronger purifying selection, compared to synonymous SNPs located in the corresponding regions in CSEs. These observations are statistically significant in the 3' splice junction. However, the significance of the difference also depends on population data used. It is important to note the inconsistency in significance of the difference between the different population groups, highlighting differences of allele frequency between human populations. The differences between populations could be explained by mutations providing a selective advantage or disadvantage within a specific population but not in others. However, given that the differences are relatively small they are more likely to be the result of sampling (the p-values presented in the table are not corrected for multiple testing).

As opposed to non-synonymous SNPs, synonymous mutations are usually thought to evolve neutrally in mammals. Nonetheless, when these silent mutations disrupt splicing regulatory sites they can cause abnormal splicing (Montera *et al.*, 2001; Mankodi & Ashizawa, 2003) or activate cryptic splice sites (Deshler & Rossi, 1991; Denecke *et al.*, 2004). There has been evidence suggesting selection acting on synonymous substitutions to ensure efficient pre-mRNA splicing (Willie & Majewski, 2004; Gorlov *et al.*, 2005; Chamary & Hurst, 2005a, Parmley & Hurst, 2007). A recent comparative genomics study characterizing the evolutionary behavior of splicing regulatory motifs showed strong evidence that synonymous substitutions in CSEs tend to create exonic splicing enhancers and disrupt exonic splicing silencers, implying positive selection for ESEs (Ke *et al.*, 2008). This positive selection is said to be compensating for aberrant splicing and mutations that disrupt splice sites. It is possible that the decreased average frequency of derived alleles of synonymous SNPs located in the splice junctions of ASEs compared to the corresponding regions of CSEs is may be a consequence of positive selection acting on splicing promoting synonymous SNPs located in exonic splicing regulatory motifs near the exon boundaries of CSE compensating for mutations that disrupt splice sites.

Table 5.3a: Average frequency of derived alleles of SNPs located within splice junctions of constitutively spliced exons (CSEs).

Population	CSEs							
	5'splice junctions				3'splice junctions			
	Exonic			Intronic	Exonic			Intronic
	NS	S	UTR		NS	S	UTR	
European	0.26 (0.30)	0.34 (0.33)	0.37 (0.35)	0.32 (0.32)	0.26 (0.31)	0.34 (0.32)	0.33 (0.30)	0.32 (0.32)
Asian	0.26 (0.31)	0.33 (0.34)	0.36 (0.35)	0.32 (0.33)	0.27 (0.32)	0.35 (0.32)	0.34 (0.32)	0.32 (0.33)
African	0.23 (0.28)	0.29 (0.28)	0.30 (0.30)	0.27 (0.28)	0.21 (0.27)	0.30 (0.29)	0.29 (0.27)	0.28 (0.29)

NS=non-synonymous, S=Synonymous, UTR=untranslated region. Numbers placed within brackets show standard deviation

Table 5.3b: Average frequency of derived alleles of SNPs located within splice junctions of alternatively spliced exons (ASEs).

Population	ASEs							
	5'splice junctions				3'splice junctions			
	Exonic			Intronic	Exonic			Intronic
	NS	S	UTR		NS	S	UTR	
European	0.22 (0.30)	0.27 (0.30)	0.26 (0.28)	0.31 (0.31)	0.27 (0.34)	0.26 (0.28)	0.25 (0.34)	0.31 (0.32)
Asian	0.29 (0.34)	0.29 (0.32)	0.33 (0.36)	0.31 (0.31)	0.29 (0.35)	0.30 (0.31)	0.21 (0.30)	0.32 (0.33)
African	0.19 (0.28)	0.25 (0.29)	0.24 (0.24)	0.27 (0.28)	0.24 (0.28)	0.22 (0.26)	0.20 (0.25)	0.28 (0.28)

NS=non-synonymous, S=Synonymous, UTR=untranslated region. Numbers placed within brackets show standard deviation

We further compare the DAF spectrums of SNPs located in exon-intron junctions with SNPs located in non-synonymous, synonymous and non-coding regions of the human genome (Figure 5.5a-c). We observe similar derived allele frequency spectra between SNPs located in exon-intron junctions of ASEs and CSEs. Compared to SNPs located in the splice sites, non-synonymous SNPs have an excess of low frequency (rare) derived alleles reflecting purifying selection acting more frequently on non-synonymous mutations that alter amino acid sequences (Fay & Wu, 2003; Chen & Rajewsky, 2006) than polymorphic sites in the splice junctions.

The differences in derived allele frequencies between CSEs and ASEs are marginally statistically significant in the case of a small number of SNP categories. However, the differences between the derived allele frequency distributions are relatively small and the significant comparisons may be the result of multiple hypothesis testing. Statistical significance depends on the magnitude of the difference between the two groups compared and the sample size. The lack of statistical significance in these comparisons may be due to relatively small sample sizes of SNPs in exon-intron junctions of ASEs. However, statistics cannot be used to prove that there is exactly zero difference between the two groups.

Table 5.4a: Significance (P-values) of average frequency of derived alleles in exon-intron junctions of CSEs and ASEs using European frequency data.

SJ Position	5' CSE NS	5' CSE S	5' CSE UTR	5' CSE I	3' CSE NS	3' CSE S	3' CSE UTR	3' CSE I	5' ASE NS	5' ASE S	5' ASE UTR	5' ASE I	3' ASE NS	3' ASE S	3' ASE UTR	3' ASE I
5' CSE NS	-															
5' CSE S	**5.31E-05**															
5' CSE UTR	**1.30E-05**	0.2185														
5' CSE I	**0.0002001**	0.2452	0.02748													
3' CSE NS	0.9655	**3.65E-05**	**1.04E-05**	**0.0001182**	-											
3' CSE S	**1.74E-05**	0.8888	0.2535	0.1683	**1.09E-05**	-										
3' CSE UTR	**0.001273**	0.9101	0.2354	0.4344	**0.00113**	0.8176	-									
3' CSE I	**8.57E-05**	0.2967	**0.03359**	0.8504	**4.58E-05**	0.2081	0.4951	-								
5' ASE NS	0.3502	**0.004435**	**0.0007877**	**0.01331**	0.3365	**0.003443**	**0.007911**	**0.011116**	-							
5' ASE S	0.6718	0.07018	**0.01407**	0.1815	0.6856	0.05724	0.1036	0.159	0.2928	-						
5' ASE UTR	0.9475	0.4393	0.281	0.5647	0.9542	0.4224	0.4599	0.5485	0.663	0.9287						
5' ASE I	0.07706	0.2751	0.05392	0.665	0.07735	0.2282	0.3691	0.5993	0.05456	0.4	0.6551					
3' ASE NS	0.8026	0.1359	**0.04014**	0.2702	0.8152	0.1193	0.1685	0.2476	0.3949	0.9467	0.9596	0.4464	-			
3' ASE S	0.881	0.05998	**0.01317**	0.147	0.8965	**0.04978**	0.0855	0.1298	0.4077	0.8506	0.997	0.3204	0.9215	-		
3' ASE UTR	0.935	0.3412	0.2054	0.4544	0.928	0.3261	0.3608	0.4394	0.7553	0.814	0.9155	0.5427	0.85	0.8912	-	
3' ASE I	**0.04787**	0.3331	0.06635	0.7816	**0.04758**	0.2785	0.4357	0.71	**0.04165**	0.3375	0.623	0.899	0.3928	0.2698	0.512	-

NS=non-synonymous, S=Synonymous, UTR=untranslated region, I = intronic. Numbers in bold shows values ≤ 0.05

Table 5.4b: Significance (P-values) of average frequency of derived alleles in exon-intron junctions of CSEs and ASEs using Asian frequency data

SJ Position	5' CSE NS	5' CSE S	5'CSE UTR	5'CSE I	3'CSE NS	3'CSE S	3'CSE UTR	3' CSE I	5' ASE NS	5' ASE S	5' ASE UTR	5' ASE	3' ASE NS	3' ASE S	3' ASE UTR	3'ASE I
5'CSE NS	-															
5'CSE S	**1.39E-04**															
5'CSE UTR	**6.42E-05**	0.2979														
5'CSE I	**0.0002337**	0.3414	0.06476	-												
3'CSE NS	0.5929	**7.03E-04**	**2.48E-04**	**0.001486**	-											
3'CSE S	**2.95E-06**	0.4997	0.5859	0.06562	**2.03E-05**	-										
3'CSE UTR	**0.0009901**	0.838	0.458	0.3374	**0.003339**	0.7361	-									
3'CSE I	**9.73E-05**	0.4218	0.08117	0.826	**6.77E-04**	0.08972	0.3985	-								
5' ASE NS	0.4904	0.3005	0.1215	0.4884	0.6422	0.1783	0.2731	0.4487	-							
5' ASE S	0.7063	0.3048	0.1074	0.5343	0.4482	0.1619	0.277	0.4842	0.8888	-						
5' ASE UTR	0.5396	0.9695	0.7862	0.9182	0.5964	0.8777	0.9346	0.9368	0.7346	0.7775	-					
5' ASE I	**0.03294**	0.5108	0.1557	0.9401	0.07369	0.2432	0.4538	0.854	0.5625	0.6274	0.9068					
3' ASE NS	0.4694	0.3608	0.1575	0.5593	0.61	0.2272	0.3275	0.5191	0.9576	0.9384	0.7564	0.6247				
3' ASE S	0.3012	0.4383	0.1852	0.685	0.4232	0.2713	0.3956	0.6361	0.8192	0.915	0.8156	0.7512	0.8673			
3' ASE UTR	0.597	0.1523	0.08736	0.2093	0.5142	0.1132	0.1412	0.1979	0.4464	0.3543	0.4129	0.2332	0.4023	0.3539	-	
3' ASE I	**0.0288**	0.5954	0.1945	0.9623	0.06409	0.3059	0.5248	0.954	0.5233	0.5799	0.9279	0.9267	0.5845	0.7032	0.2189	-

NS=non-synonymous, S=Synonymous, UTR=untranslated region, I = intronic. Numbers in bold shows values ≤ 0.05

Table 5.4c: Significance (P-values) of average frequency of derived alleles in exon-intron junctions of CSEs and ASEs using African frequency data

SJ Position	5' CSE NS	5' CSE S	5' CSE UTR	5' CSE I	3' CSE NS	3' CSE S	3' CSE UTR	3' CSE I	5' ASE NS	5' ASE S	5' ASE UTR	5' ASE I	3' ASE NS	3' ASE S	3' ASE UTR	3' ASE I
5' CSE NS	-															
5' CSE S	**4.45E-04**	-														
5' CSE UTR	**1.63E-03**	0.6762	-													
5' CSE I	**0.005414**	0.1546	0.1364	-												
3' CSE NS	0.3134	**2.52E-06**	**6.19E-05**	**2.18E-05**	-											
3' CSE S	**1.16E-04**	0.7606	0.8543	0.3895	**3.67E-07**	-										
3' CSE UTR	**0.01124**	0.7814	0.5521	0.451	**0.000672**	0.6212	-									
3' CSE I	**1.25E-03**	0.3083	0.2316	0.5754	**2.06E-06**	0.16	0.6663	-								
5' ASE NS	0.2735	**0.005942**	**0.004886**	**0.0228**	0.5145	**0.003803**	**0.01418**	**0.01444**	-							
5' ASE S	0.4723	0.2679	0.2006	0.6389	0.2196	0.2051	0.386	0.4832	0.1656	-						
5' ASE UTR	0.9439	0.4786	0.4166	0.6517	0.7769	0.439	0.5314	0.5971	0.5909	0.822	-					
5' ASE I	0.1067	0.4051	0.2925	0.997	**0.01994**	0.2949	0.5993	0.7898	**0.04753**	0.6669	0.6624	-				
3' ASE NS	0.8223	0.153	0.1157	0.3666	0.4959	0.1159	0.2309	0.2833	0.322	0.7344	0.9734	0.4332	-			
3' ASE S	0.6756	**0.03572**	**0.02766**	0.1112	0.9647	**0.02473**	0.06743	0.07742	0.6037	0.3939	0.8067	0.168	0.63	-		
3' ASE UTR	0.6839	0.1978	0.164	0.3146	0.8715	0.1736	0.2358	0.2746	0.8673	0.4816	0.7435	0.3343	0.6321	0.8669	-	
3' ASE I	**0.04072**	0.655	0.4722	0.6645	**0.005349**	0.5363	0.859	0.8679	**0.02522**	0.4862	0.5758	0.745	0.3024	0.1022	0.2694	-

NS=non-synonymous, S=Synonymous, UTR=untranslated region, I = intronic. Numbers in bold shows values ≤ 0.05

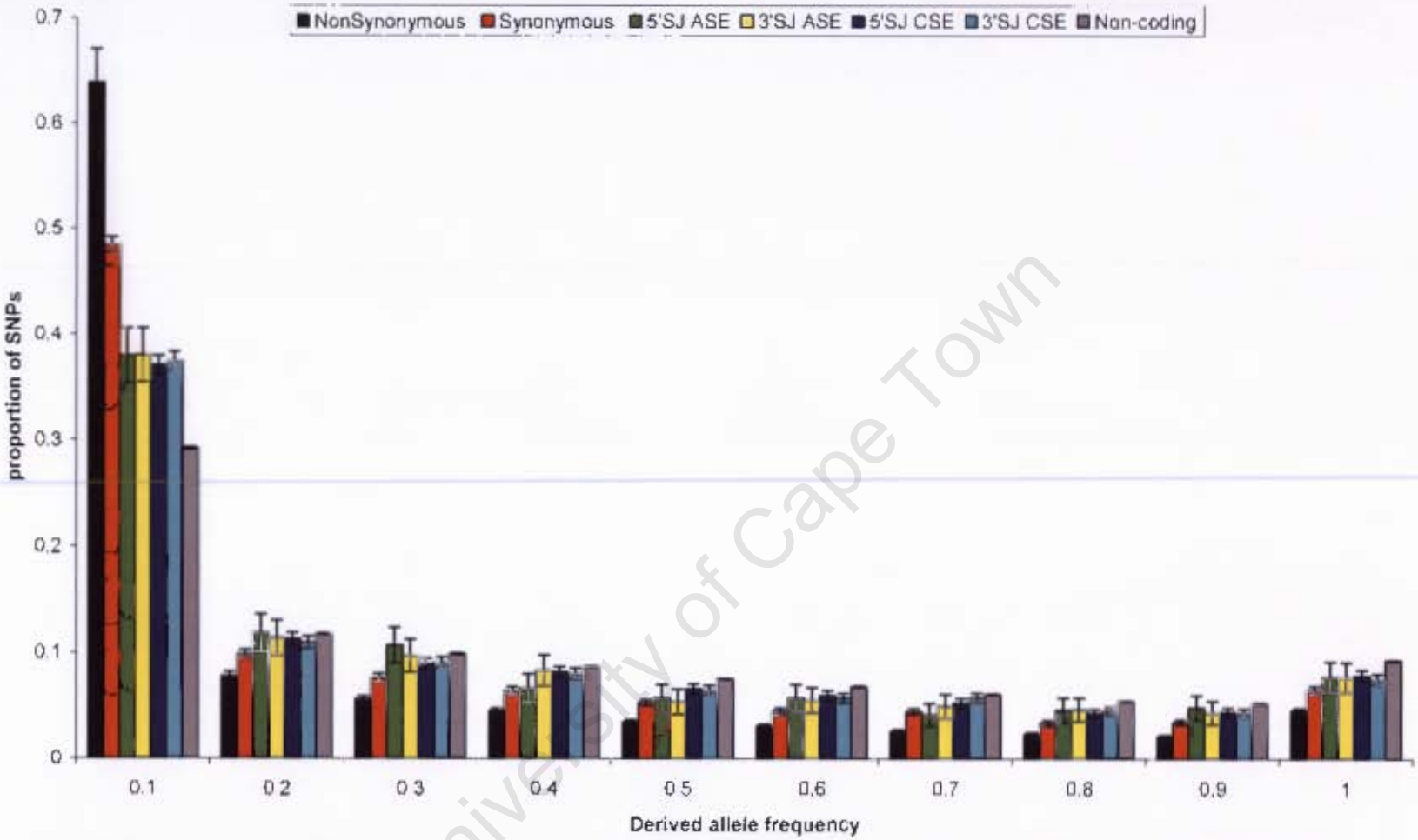

Figure 5. 5a: DAF spectra of SNPs found in functional genomic sites using European population data. The horizontal axis show the derived allele frequency. The vertical axis show the proportion of sites that have a particular frequency. For example, in the European population the proportion of SNPs in the 5 splice junction (SJ) of ASEs with derived allele frequency of 0.1 is 0.379. Error bars represent standard error of proportion

Figure 5.5b: DAF spectra of SNPs found in functional genomic sites using Asian (Chinese and Japanese) population data. The horizontal axis show the derived allele frequency. The vertical axis show the proportion of sites that have a particular frequency. For example, in the Asian population the proportion of SNPs in 5'splice junction (SJ) of ASEs with derived allele frequency of 0.1 is 0.402 . Error bars represent standard error of proportion

Figure 5.5c: DAF spectra of SNPs found in functional genomic sites using African (Yorubans) population data. The horizontal axis show the derived allele frequency. The vertical axis show the proportion of sites that have a particular frequency. For example, in the African population the proportion of SNPs in 5'splice junction (SJ) of ASEs with derived allele frequency of 0.1 is 0.389. Error bars represent standard error of proportion

5.5 Conclusions

The use of population specific SNP frequency data allowed us to get a qualitative understanding of selection by comparing patterns of derived allele frequency between SNPs located at splice sites near exon boundaries with SNPs located at synonymous sites, non-synonymous sites and non-coding regions. Based on their genomic locations these different SNP categories have varying consequences of spliced transcript product and are subjected to different selective forces. As expected under selection, we observed that the frequency of derived alleles for non-synonymous polymorphisms is lower than the average frequencies of derived alleles of the other SNP categories. In addition we showed that SNPs in exon-intron junctions of CSEs and ASEs may be subjected to different evolutionary patterns.

CHAPTER 6
GENERAL CONCLUSIONS

In this work, we directed our efforts to investigate the extent to which SNPs occurred in the exon-intron junctions of human genes. We identified 47 036 SNPs in 5' and 3' exon-intron junctions (comprising 60bp of exon and 60bp of intron) of more than half of Ensembl annotated genes. These SNPs are more often located in the non-coding intronic regions than the coding exonic regions. The exonic regions have higher proportions of rare (frequency ≤ 0.1) derived alleles than intronic positions, with non-synonymous coding SNPs having the lowest average frequencies of derived alleles. These observations reflect the fact that new mutations disrupting the coding sequence of a gene are kept at low frequencies through purifying selection whereas those that are found at non-coding regions are expected to be under the influence of weaker or no selective pressure.

In contrast to what has already been observed, the identified alternatively spliced exons (ASEs) have lower proportions of SNPs in their exon-intron junctions than constitutively spliced exons (CSEs). Nevertheless, there are specific genomic positions where analysis shows a higher prevalence of SNPs at spliced junctions of ASEs than spliced junctions of CSEs. This difference depends on genomic positions (exonic vs. intronic) of the SNP and also which SNP dataset is used. We suspect the dataset of ASE used in the study has diverse pattern of SNP distribution, however without categorization of ASEs into the distinct alternative splice patterns, comparison of the prevalence of SNPs at exon-intron junctions of ASEs with CSEs is not likely to produce a conclusive analysis. The different types of alternative splice isoforms have diverse features and this is highlighted by Zheng and his colleagues (2005). It might be possible that one type of splice pattern (i.e. cassette exon) has lower frequencies of SNPs in their exon-intron junctions compared to CSEs, whereas another type (i.e alternative 3' splice site) might have higher frequencies of SNPs than the corresponding regions in CSEs. Hence, further analysis of ASEs is needed for understanding of sequence conservation in the exon-intron junctions between ASEs and CSEs.

SNPs are fewer at positions close to the exon boundaries than any other region in the exon-intron junctions. However, these positions are accompanied by increased average frequency of the derived alleles compared to non-synonymous and even synonymous sites. The higher average frequency of derived alleles of SNPs in the splicing regulatory sites compared to coding synonymous sites is surprising, since we expected polymorphisms located within splicing regulatory regions to alter efficiency of splicing regulation and consequently to be subjected to purifying selection more often than neutral SNPs. Aberrant spliced transcript patterns are the underlying cause of many heritable human diseases; hence, we expect natural selection would act to reduce the frequency of new mutations altering regulatory splice sites. Unless natural selection frequently serves to increase the frequency of SNPs located in splice sites to improve splicing efficiency in CSEs, the alternative is that mutations that affect splicing are more strongly selected against in ASEs. The different types of exons in the human genome are characterized by different features, including different patterns of SNP distribution and derived allele frequency.

A major limitation on our research methodology is the way in which we have identified ASEs and CSEs. We used ASAPII annotations, which only uses EST and cDNA sequence to identify alternatively spliced exons. EST data provide information on structure of ASEs and include data from different gene expression contexts, but this information is highly biased towards terminal exons and highly expressed genes. Consequently we might have missed rare alternatively spliced patterns and many internal ASEs; as a result these can be mistaken as CSEs. Furthermore, the bias towards terminal exons being identified as alternatively spliced may cause us to mistake characteristics of terminal exons for the characteristics of alternatively spliced exons. Analysis using the latest version of Ensembl database (i.e. version 50) might greatly improved identification of ASEs due to availability of new information. In addition, although exon array methods also have their limitations, inclusion of this information for ASEs identification can also be useful for defining a less biased ASE dataset.

We have identified SNPs which have already been implicated in splicing regulation, in our SNP dataset. These srSNPs allow production of transcript variants in an allele specific manner. For example, an exonic SNP (rs2295682) found in exon-intron junction of the RBM23 gene is associated with exon 6 skipping. This synonymous coding SNP has a G or an A allele, and it is the A allele which is associated with increased exon skipping event (Hull *et al.*, 2007). Another example includes an A/G SNP at the acceptor splice site (AG/AA) of the OAS1 gene which causes shortening of exon 7. The A allele destroys the splice site and is associated with decreased enzyme activity which has been implicated in individuals susceptible to viral infection (Bonnevie-Nielsen *et al.*,2005*)*. This suggests that a fraction of SNPs at exon-intron junctions directly affect the pattern of alternatively spliced exons. Many more SNPs that have an effect on phenotypic variability, including disease susceptibility, are likely to be present in our database. Although we cannot estimate the extent to which all of these are involved in phenotypic diversity, the frequent occurrence of SNPs in critical sites within intron-exon junctions suggests a contribution from population variation to the observed diversity of splicing events.

A significant goal of the research group in which this study was carried out is to establish the extent to which human alternatively spliced genes are affected by *cis*-acting splicing regulatory polymorphisms, which might either completely determine which isoform of the gene is produced or alter the relative proportions of alternative isoforms. This will enable us to determine whether the alternatively spliced transcript isoforms observable in the public database are the result of alternative splicing to increase the functional repertoire of a gene or a result of mutations that may be responsible for some phenotypic differences between individuals. As a means of contributing to this goal, we surveyed the occurrence of SNPs at exon-intron junctions of human genes and provided a description of the way in which SNPs are distributed within the exon-intron junctions, consequently producing a dataset of SNPs that may affect pre-mRNA splicing.

BIBLIOGRAPHY

Altshuler, D., *et al.* (Int'l HapMap Consortium). (2005). A haplotype map of the human genome *Nature*. **437**:1299-1320.

Altshuler, D., Pollara, V.J., Cowles, C.R, Van Etten, W.J., Baldwin, J., Linton, L, *et al.* (2000). A SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*. **407**: 513 – 516.

Amendt, B., Si, Z and Stoltzfus, C.M. (1995). Presence of exon splicing silencers within human immunodeficiency virus type 1 tat exon 2 and tat-rev exon 3: evidence for inhibition mediated by cellular factors. *Mol. Cell. Biol.* **15**. 4606–4615.

Angeloni, D., Duh, F.M., Moody, M., Dean, M., Zabarovsky, E.R., Sentchenko, V., Braga. E and Lerman, M.I. (2003). C to A single nucleotide polymorphism in intron 18 of the human MST1R (RON) gene that maps at 3p21.3. *Mol. Cell. Probes*.**17**:55-57.

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.

Asthana, S., Noble, W.S., Kryukov, G., Grant, C.E and Sunyaev, S. (2007). Widely distributed non-coding purifying selection in the human genome. *PNSA*. **104**: 12410- 12415.

Atweh, G.F., Anagnou, N.P., Shearin, J., Forget, B.G and Kaufman, R.E. (1985).ß-Thalassemia resulting from a single nucleotide substitution in an acceptor splice site. *Nucl. Acids Res*. **13**: 777-790.

Baek, D and Green, P. (2005). Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *PNAS*. **102**: 12813-12818.

Balasubramanian, S., Harrison. P., Hegyi H., Bertone.P., Luscombe,N., Echols,N., McGarve, P., Zhang, Z.L and Gerstein, M. (2002). SNPs on human chromosomes 21 and 22 - analysis in terms of protein features and pseudogenes.*Pharmacogenomics*.**3**:393-402.

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

Bell, M.V., Cowper, A.E., Lefranc, M., Bell, J.I., and Screaton, G.R. (1998). Influence of Intron Length on alternative splicing of CD44. *Mol. Cell. Biol*. **18**: 5930-5941.

Belyavsky, A., Vinogradova, T and Rajewsky, K. (1989). PCR-based cDNA library construction: general cDNA libraries at the level of a few cells. *Nucleic Acids Res.* **17**: 2919–2932.

Bercovich, D., Friedlander, Y., Korem, S., Houminer, A., Hoffman, A., Kleinberg, L., Shochat, C., Leitersdorf, E. and Meiner, V. (2006).The association of common SNPs and haplotypes in the CETP and MDR1 genes with lipids response to fluvastatin in familial hypercholesterolemia. *Atherosclerosis*. **185**: 97-107.

Berget, S.M. (1995). Exon Recognition in Vertebrate Splicing. *J. Biol. Chem.* **270**:2411-2414.

Black, D.L. (1991). Does steric interference between spice sites block the splicing of short c-src neuron specific exon in non-neuronal cells? *Genes Dev.* **5**: 389 – 402.

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

Black, D.L (2003). Mechanism of Alternative Pre-Messanger RNA Splicing. *Annu. Rev. Biochem.* **72**: 291 – 336.

Blencowe, B.J. (2000). Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *TIBS.* **25**: 105 -110.

Boguski, M. S., Lowe, T. M., and Tolstoshev, C.M. (1993). dbEST-database for "expressed sequence tags". *Nat. Genet.* **4**: 332 -333.

Boguski, M.S. and G.D. Schuler. (1995). ESTablishing a human transcript map. *Nature Genet.* **10**: 369-371

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., Pesole, G. (2005) ASPIC: a novel method to predict the exon-intron structure of a gene that is optimally compatible to a set of transcript sequences *BMC Bioinformatics*. **6**: 244.

Bonnevie-Nielsen, V., Field, L.L., Lu,S., Zheng, D, Li,M., Martensen, P.M., Nielsen, T.B., Beck-Nielsen, H., Lau, Y.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.

Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbrück, S., Krueger, S., Reich J and Bork, P. (2000). EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Letters*. **474**: 73 – 76.

Bray, N.J., Buckland, P.R., Owen, M.J. and O'Donovan, M.C. (2003). Cis-acting variation in the expression of a high proportion of genes in human brain. *Hum Genet*. **113**:149-153.

Brookes, A.J. (1999). The essence of SNPs. *Gene*.**234**: 177 – 186.

Brudno, M., Gelfand, M.S., Spengier, S., Zorn, M., Dubchak,I and Conboy,J.G. (2001). Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. *Nucleic Acids Res*. **29**:2338 – 2348.

Buckland, P.R. (2004). Allele-specific gene expression differences in humans. *Hum. Mol. Genet*. **13**: 255 – 260.

Buckland, P.R. (2006). The importance of identification of regulatory polymorphisms and their mechanism of action. *BBA*.**1762**:17- 28.

Burge, C. B.,Tuschl, T., Sharp, P.A. (1999). Splicing of precursors to mRNAs by the spliceosome, p.p 525-560. In R. F. Gesteland, T. R. Cech, and J. F. Atkins (ed.), The RNA world. Cold Spring Harbor Laboratory Press, New York.

Burset, M, Seledtsov, I. A., and Solovyev, V. V.(2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res*. **28**: 4364–4375.

Caceres, J.F. and Krainer, A.R. (1997). Mammalian pre-mRNA splicing factor. In Eukaryotic mRNA Processing (Krainer, A.R., ed.) pp. 174 – 212, Oxford University press.

Cargill, M., Altshuler, D., Ireland, J., Sklar.P., Ardlie, K., Patil, N., Shaw, N., Lane, C.R., Lim, E.P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley,G.Q and Lander, E.S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet*. **22**:231-238.

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

Cavalli-Sforza, L.L., Menozi, P. and Piazza, A. (1994). The History and Geography of Human Genes. Princeton University Press. Princeton.

Chakravarti, A. (1999). Population genetics – making sense out of sequence. *Nature Genet.* **21**: 56 – 60.

Chamary, J.V., and Hurst, L.D (2005a). Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else ? *Trends Genet.* **21**: 256 -259.

Chao, H.K., Hsiao, K.J. and Su, T.S.(2001). A silent mutation induces exon skipping in the phenylalanine hydroxylase gene in phenylketonuria. *Hum.Genet.* **108**: 14 -19.

Chen, F.C., Wang, S.S., Chen, C.J., Li, W.H, Chuang, T.J. (2006). Alternatively and constitutively spliced exons are subject to different evolutionary forces. *Mol Biol Evol.* **23**:675-82.

Chen, K. and Rajewsky, N. (2006). Natural selection on human microRNA binding sites inferred from SNP data. *Nature. Genet.***38**:1452 - 1456.

Chen, F.C and Chaung, T.J (2007). Different alternative splicing patterns are subjected to opposite selection pressure for protein reading frame preservation. *BMC Evol Biol.* **7**:179.

Chen, F.C., Chaw, S.M., Tzeng., Y.H., Wang, S.S and Chaung, T.J. (2007) Opposite Evolutionary effects between different alternative splicing patterns. *Mol. Biol. Evol.* **24**: 1443 – 1446.

Choi, YD., Grabowski, P.J., Sharp, P.A and Dreyfuss, G.(1986). Heterogeneous nuclear ribonucleoproteins: role in RNA splicing. *Science.* **231**:1534-1539.

Clark, F and Thanaraj, T.A. (2002). Categorization and characterization of transcript-confirmed constitutively and alternatively spliced intron and exons from human. *Hum. Mol. Genet.* **11**: 451 – 464.

Clark, T.A., Sugnet, C.W., Ares Jr, M. (2002). Genome-wide Analysis of mRNA Processing in Yeast Using Splicing-Specific Microarrays. *Science.* **296**: 907 – 910.

Clark, T.A., Schweitzer, A.C., Chen, T.X., Staples, M.K., Lu, G., Wang, H.,Williams, A and Blume, J.E. (2007). Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol.* **8**:R64.

Collins, F.S., Brooks, L.D., Chakravarti, A (1998). A DNA polymorphisms discovery resource for research on human genetic variation. *Genome Res.* **8**: 1229 – 1231.

D'Souza, I. and Schellenberg, G.D. (2002). *tau* exon 10 expression involves a bipartite intron 10 regulatory sequence and weak 5' and 3' spice sites. *J. Biol. Chem.***270**: 5346 – 5352.

De Meirleir L, Lissens W, Benelli C, Ponsot G, Desguerre I, Marsac C, Rodriguez D, Saudubray J-M, Poggi F, Liebaers, I. (1994). Aberrant splicing of exon 6 in the pyruvate dehydrogenase-E1α RNA linked to silent mutation in a large family with Leigh's encephalomyelopathy, *Pediatr.Res.* **36**: 707 – 712.

Deeb, S. S ., Takata, k., Peng, R.L., Kajiyama, G and Albers, J.J. (1990). A splice-junction mutation responsible for familial apolipoprotein A-II deficiency. *Am J Hum Genet.* **46**: 822–827.

Del Gatto-Konczak, F. Olive, M., Gesnel, M,C. and Breathnach R. (1999). hnRNP A1 recruited to an exon in vivo can function as an exon splicing silencer .*Mol. Cell. Biol.* **19**: 251 -260.

Delaney, S.J., Rich, D.P., Thomson, S.A., Hargrave, M.R., Loveslock, P.K., Welsh, M.J. and Wainwright, B.J. (1993). Cystic fibrosis transmembrane conductance regulator splice variants are not conserved and fail to produce chloride channels. *Nat.Genet.***4:** 426 – 431.

Denecke, J., Kranz, C., Kemming, D., Koch, H.G. and Marquardt, T. (2004). An activated 5' cryptic splice site in the human *ALG3* gene generates a premature termination codon insensitive to nonsense-mediated mRNA decay in a new case of congenital disorder of glycosylation type Id (CDG-Id). *Hum. Mutat.* **23**: 477 – 486.

Denson, J., Xi, Z., Wu, Y., Yang, W., Neale, G and Zhang, J. (2006). Screening for inter-individual splicing differences in human GSTM4 and the discovery of a single nucleotide substitution related to the tandem skipping of two exons. *Gene.* **379**: 148 – 155.

Deshler, J.O and Rossi, J.J (1991) Unexpected point mutations activeate cryptic 3' splice sites by perturbing a natural secondary structure within a yeast intron. *Genes Dev.* **5**: 1252 – 1263.

Dominski . Z and Kole, R. (1991). Selection of splice sites in pre-mRNA with short internal exons. *Mol. Cell. Biol.* **11**: 6075- 6083.

Dralyuk, I.,Brudno, M.,Gelfand, M. S.,Zorn, M.and Dubchak, I. (2000).ASDB: database of alternatively spliced genes. *Nucleic Acids Res.* **28**: 296-297.

Dror, G., Sorek, R. and Shamir, R. (2005). Accurate identification of alternatively spliced exons using support vector machine. *Bioinformatics.* **21**: 897- 901.

Du, Y.Z., Dickerson, C. Aylsworth, A.S and Swhartz, C.E. (1998). A silent mutation, C924T(G308G), in L1CAM gene result in x linked hydrocephalus (HSAS). *J.Med.Genet.* **35**: 456 – 462.

Dunn, J.M., Phillips, R.A., Zhu, X., Becker, A and Gallie, B.L. (1989). Mutations in the RB1 gene and their effects on transcription. *Mol. Cell. Biol.* **9**: 4596-4604.

Dye, B.T., Buvoli, M., Mayer, S.A., Lin, C.H., and Patton, J.G. (1998). Enhancer elements activate the weak 3' splice site of α-tropomyosin exon 2. *RNA*. **4**: 1523 – 1536.

Eriksson, P., Kallin,B., van 't Hooft, F.M., Bavenholm, P and Hamsten, A. (1995). Allele-specific increase in basal transcription of the plasminogen-activator inhibitor 1 gene is associated with myocardial infarction. *PNAS*. **92**: 1851–1855.

Eskesen, S. T., Eskesen, F. N., and Ruvinsky, A. (2004). Natural selection affects frequencies of AG and GT dinucleotides at the 5' and 3' ends of exons. *Genetics*, **167**: 543 – 550.

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

Fairbrother, W. G., Holste, D., Burge, C. and Sharp, P. (2004). Single nucleotide polymorphism-Based validation of exonic splicing enhancers. *Plos Biol*. **2**: 1388 – 1395.

Faustino, N. A. and Coopper, T. A. (2003). Pre-mRNA splicing and human disease. *Genes Dev*. **17**: 417 – 37.

Fay, J.C and Wu, C. (2003). Sequence divergence, functional constraint, and selection in protein evolution.*Annu. Rev.Genomics. Hum. Genet.***4**: 213 -235.

Florea, L. (2006). Bioinformatics of alternative splicing and its regulation. *Brief Bioinform*. **7**: 55 -69.

Fredman, D., Siegfried,M., Yuan, Y.P., Bork,O.Lehvaslaiho.H and Brookes, A.J.(2002). HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data source. *Nuclei Acid. Res*. **30**: 387 – 391.

Frischmeyer, P.A and Dietz, H.C. (1999). Nonsense-mediated mRNA decay in health and disease. *Hum. Mol. Genet*.**8**: 1893 – 1900.

Garg, K & Green, P. (2007). Differing patterns of selection in alternative and constitutive splice sites. *Genome.Res*. **17**: 1015 – 1022.

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.

Gillespie, J.H. (1991). The causes of Molecular Evolution. Oxford University Press. New York

Golling, G., Amsterdam, A., Sun, Z., Antonelli, M., Maldonado, E., Chen, W., Burgess, S., Haldi, M., Artzt, K., Farrington, S., et al. (2002). Insertional mutagenesis in Zebrafish rapidly identifies genes essential for early vertebrate development. *Nat. Genet.* **31**: 135 – 140.

Gorlov, I.P.,Kimmel, M. and Amos, C.I .(2006). Strength of the purifying selection against different categories of the point mutations in the coding regions of the human genome. *Hum. Mol. Genet.* **15**: 1143 – 1150.

Graur D. & Li,W.-H. (2000). Fundamentals of Molecular Evolution. Sunderland, Massachusetts: Sinauer Assoc.

Graveley, B.R., and Maniatis, T. (1998). Arginine/Serine- rich domain of SR proteins can function as activator of pre-mRNA splicing. *Mol.Cell.* **1**: 765 – 771.

Graveley, B.R. (2000). Sorting out the complexity of SR protein functions. *RNA.* **6**: 1197 – 1211.

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

Guthrie, C. (1989). Catalytic RNA and RNA splicing. *Am. Zool.* **29**: 557 – 567.

Hagmann, W. (1999).'A good SNP may be hard to find'.*Science*.**285**: 21-22.

Harteveld, C.T. et al. (2004). An α- thalassemia phenotype in a Dutch Hindustani, caused by a new point mutation that creates an alternative splice, donor site in the first exon of the α-2 globin gene. *Hemaglobin.* **28**: 255 – 259.

Hartl, D.L and Clark, A.G.(1997). Principles of population genetics. Sinauer and Associates, Sunderland, MA.

Hasegawa, K., Tamari, M., Shao, C., Shimizu, M., Takahashi, N., Mao, X.Q., Yamasaki, A., Kamada, F., Doi, S., Fujiwara, H., Miyatake, A., Fujita, K., Tamura, G., Matsubara, Y., Shirakawa, T and Suzuki, Y. (2004). Variations in the C3, C3a receptor, and C5 genes affect susceptibility to bronchial asthma. *Hum Genet.* **115**: 295-301.

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

Hawkins, J. D. (1988). A survey on intron and exon length. *Nucleic Acids Res.* **16**: 9893-9908.

Hentze, M.W. and Kulozik, A. E. (1999). A perfect message: RNA surveillance and nonsense-mediated decay. *Cell*. **96**: 307 – 310.

Higashi, Y.,Tanae, A., Inoue, H., Hiromasa, T and Fujii-Kuriyama, Y. (1988). Aberrant Splicing and Missense Mutations Cause Steroid 21-Hydroxylase [P-450(C21)] Deficiency in Humans: Possible Gene Conversion Products. *PNAS* . **85**: 7486-7490.

Hillier, L. D., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., Hawkins, M., Hultman, M., Kucaba, M., Lacy, M., Le, M., Le, N., Mardis, E., Moore, B.,Morris, M., Parsons, J., Prange, C., Rifkin, L.., Rohlfing, T., Schellenberg, K. and Marra, M. (1996). Generation and analysis of 280,000 human expressed sequence tags. *Genome Res*. **6**: 807-828.

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.

Hirakawa, M., Tanaka, T., Hashimoto, Y., Kuroda, M., Takagi, T and Nakamura, Y. (2002). JSNP: a database of common genetic variation in Japanese population. *Nuclei Acid. Res*. **30**: 158 – 162.

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

Holste, D., Huo, G., Tung, V. and Burge, C.B. (2006). HOLLYWOOD: a comparative relational database of alternative splicing. *Nucleic Acids Res*. **34**: 56 – 62.

Horiuchi, T., Giniger, E. and Aigaki, T. (2003) Alternative *trans*-splicing of constant and variable exons of a *Drosophila* axon guidance gene, lola. *Genes Dev*. **17**, 2496–2501.

Hu, G.K., Madore, S.J., Moldover, B., Jatkoe, T., Balaban, D., Thomas, J. and Wang Y. (2001) Predicting splice variant from DNA chip expression data. *Genome Res*. **11**:1237-1245

Huang, S.H. (2002). Inverse PCR. cDNA cloning. *Methods Mol Biol*. **192**:293-299.

Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Gilbert J, Hammond M, Herrero J, Hotz H, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Kokocinsci F, London D, Longden I, McVicker G, Melsopp C, Meidl P, Potter S, Proctor G, Rae M, Rios D, Schuster M, Searle S, Severin J, Slater G, Smedley D, Smith J, Spooner W,

Stabenau A, Stalker J, Storey R, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C, Birney E.(2005). Ensembl 2005. *Nucleic Acids Res.* **33**: 447 -453.

Hudson, T.J., and Pastinen, T. (2004). Cis-acting regulatory Variation in the Human Genome. *Science.* **306**: 647- 650.

Hughes, A.L, Packer, B., Welch, R., Bergens, A.W., Chanock, S. (2003) Widespread purification selection at polymorphic sites in human protein –coding loci. *PNAS.* **100**: 15754 – 15757.

Hughes, A.T (2006). Regulation of gene expression by alternative untranslated regions. *Trends. Genet.* **22**:119 – 122.

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

Ihaka, R and Gentleman, R. (1996).R: A Language for Data Analysis and Graphics. *J Comput.Graph. Stats.* **5**: 299-314.

Ishibashi, F. et al. (2001). Improved superoxide generating ability by interferon due to splicing pattern change of transcripts in neutrophils from patients with splice site mutation in CYGG gene. *Blood.* **98**: 436 – 441.

Jackson, I.J. (1991). A reappraisal of non-consensus mRNA splice sites. *Nucleic Acids Res.* **19**: 3795 – 3798.

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

Ji, H., Zhou, Q., Wen, F., Xia, H, Lu, X and Li, Y. (2001). AsMamDB: an alternative splice database of mammals. *Nucleic Acids Res.* **29**: 260 -263.

Johnson J.M., Castle J., Garrett-Engele P., Kan Z., Loerch P.M., Armour C.D., Santos R.,Schadt E.E., Stoughton R. & Shoemaker D.D. (2003).Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays. *Science* **302**:2141-2144.

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., Tammana, H. and Gingeras, T.R. (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14**:331-342.

Kan, J.L., and Green, M.R. (1999). Pre-mRNA splicing of IgM exon M1 and M2 is directed by a justaposed splicing enhancer and inhibitor. *Genes Dev*. **4**: 462 – 471.

Kan, Z., Rouchka, E.C., and Gish, W.R. (2001). Gene Structure Prediction and Alternative Splicing Analysis Using Genomically Aligned ESTs. *Genome Res*. **11**: 889 – 900.

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. Proceedings of the 9th Pacific Symposium on Biocomputing, 42 – 53.

Kanno, H., Fujii, H., Wei, D.C., Chan, L.C., Hirono, A., Tsukimoto, I and Miwa, S . (1997). Frame shift mutation, exon skipping, and a two codon deletion caused by splice site mutations account for pyruvate kinase deficiency. *Blood*. **89**: 4213 – 4218.

Ke, S., Zhang, X.H. and Chasin, L.A. (2008). Positive selection acting on splicing motifs reflects compensatory evolution. *Genome.Res*. **18**: 533 -543.

Kenneth, H., 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.

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**: 93 -98.

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

Kimmel, A.R., Berger, S.L. (1987) Preparation of cDNA and the generation of cDNA libraries: overview. *Methods Enzymol*.**152**:307–316.

Kimura, M . (1968). Evolutionary rate at the molecular level. *Nature*. **217**: 624 – 626.

King, J.L. and Jukes, T.H. (1969). Non-Darwinian evolution. *Science*. **164**: 788 – 798.

Kinniburgh A. J., Maquat L. E., Schedl. T., Rachmilewitz, E, and Ross, J. (1982). mRNA-deficient beta o-thalassemia results from a single nucleotide deletion. *Nucleic Acids Res*. **10**:5421-5427.

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

Kohtz, J.D, Jamison, S.F, Will, C.L, Zuo, P, Luhrmann, R, Garcia-Blanco, M.A, Manley, J.L. (1994). Protein-protein interactions and 5' splice site recognition in mammalian mRNA precursors. *Nature*. **368**: 119 – 124.

Kono, H., Yuasa, T., Nishiue, S. and Yura, K. (2008).coliSNP database server mapping nsSNPs on protein structures. *Nucleic. Acids Res.* **36**: 409-413.

Koren, E., Lev-Maor, G.and Ast, G.(2007). The emergence of alternative 3' and 5' splice site exons from constitutive exons. *PLoS Comput Biol.* **3**:e95.

Kozyrev, S.V., Abelson, A.K., Wojcik, J., Zaghlool, A., Linga Reddy, M.V., Sanchez, E., Gunnarsson, I., Svenungsson, E., Sturfelt, G., Jönsen, A., Truedsson, L., Pons-Estel, B.A., Witte, T., D'Alfonso, S., Barizzone, N., Danieli, M.G., Gutierrez, C., Suarez, A., Junker, P., Laustrup, H,, González-Escribano, M.F., Martin, J., Abderrahim, H., Alarcón-Riquelme, M.E. (2008) Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus. *Nat Genet.* **40**: 211 – 216.

Královicová, J., Houngninou-Molango, S., Krämer, A., Vorechovsky, I.(2004). Branch site haplotypes that control alternative splicing. *Hum. Mol. Genet.* **13**:3189-3202.

Krawczak M., Reiss J., 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.T., Hundrieser, B., Mort, M., Wittig, M., Hampe,J. and Cooper, D.N. (2007). Single base pair substitution in exon-intron junctions of Human Genes: Nature, Distribution and consequences for mRNA splicing. *Human Mutation* **28**:150-158.

Kruglyak, L., Nickerson, D. A. (2001). Variation is a spice of life. *Nature Genet.* **27**: 234.

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

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.

Ladd, A. N., and Cooper, T. A. (2002). Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol.* **3**:1-16.

Lander, E.S *et al*, (Int'l. Human Genome Sequencing Consortium) (2001). Initial sequencing and Analysis of the human genome. *Nature*. **409**: 860 – 921.

Le, K., Mitsouras, K., Roy, M., Wang, Q., Xu, Q., Nelson, S.F. and Lee, C. (2004). Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. Nucleic Acids Res. **32**: e180.

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

Lee, P. H. and Shatkay, H.(2008). F-SNP: Computationally predicted functional SNPs for disease association studies. *Nucleic. Acids Res*. **36**: 825-829.

Lee, V.M., Goedert, M. and Trojanouski, J.Q. (2001). Neurogegenerative tauopathies. *Annu. Rev. Neurosci*. **24**: 1121 – 1159.

Legro, R.S., Barnhart, H.X., Schlaff, W.D., Carr, B.R., Diamond, M.P., Carson, S.A., Steinkampf, M.P., Coutifaris, C., McGovern, P.G., Cataldo, N.A., Gosman, G.G., Nestler, J.E., Giudice, L.C., Ewens, K.G., Spielman, R.S., Leppert, P.C., Myers, E.R, Reproductive Medicine Network. (2008).Ovulatory response to treatment of polycystic ovary syndrome is associated with a polymorphism in the STK11 gene. *J Clin Endocrinol Metab*. **93**:792-800.

Leipzig, J., Pevzner, P., and Heber, S. (2004). The Alternative Splicing Gallery (ASG): bridging the gap between genome and transcriptome. *Nucleic Acids Res*. **32**: 3977 – 3983.

Lejeune, F and Maquat, L.E. (2005). Mechanistic links between nonsense-mediated mRNA decay and pre- mRNA splicing in mammalian cells. *Curr Opin Cell Biol* . **17**: 309 – 315.

Lercher, M.J and Hurst, L.D.(2002) Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet*. **18**: 337 – 340.

Lewis, B.P., Green, R.E., Brenner, S.E. (2003). Evidence for the widespread coupling of alternative splicing with nonsense-mediated mRNA decay in humans. *PNAS*. **100**: 189 – 192.

Liang, H and Landweber, L.F. (2007). A genome wide study of dual coding reasons in the human alternatively spliced genes. *Genome Res*. **16**:190 - 196.

Ligtenberg, M.J., Gennissen, A.M., Vos, H.L., 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, Zhang, M, and Krainer, A.R (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes and Development*.**12**: 1998 – 2012.

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

Liu, W., Qian, C. and Francke, U. (1997). Silent mutation induces exon skipping of fibrillin-1 gene in Morfan syndrome. *Nature Genet*.**16**: 328 – 329.

Lodish, H. F., Baltimore, D., Berk, A., Darnell, J., Matsudaina, P. and Zipursky, L. (2003). Molecular Cell Biology. 4th ed. W.H. Freeman and Company. New York.

Lopez, A. (1998). ALTERNATIVE SPLICING OF PRE-mRNA: Developmental Consequences and Mechanisms of Regulation. *Annu. Rev. Genet*.**32**: 279 – 305.

Lorson, C.L., Hahnen, E., Androphy, E.J., Wirth,B. (1999). A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *PNAS*. **96**: 6307 – 6311.

Lucotte, G. (1998). Celtic origin of the C282Y mutation of hemochromatosis. *Blood. Cells. Mol*. **24**: 433 - 438.

MacDonald, J.H and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*. **351**: 652 – 654.

Majewski, J, and Ott, J. (2002). Distribution and Characterization of regulatory Elements in the Human Genome. *Genome Res*. **12**: 1827 – 1836.

Malartre, M., Short, S. and Sharpe, C. (2004) Alternative splicing generates multiple SMRT transcripts encoding conserved repressor domains linked to variable transcription factor interaction domains. *Nucleic Acids Res*. **32**: 4676-4686.

Maniatis, T and Reed, R. (2002). An extensive network of coupling among gene expression machines. *Nature*. **416**: 499 -506.

Maniatis, T. (1991). Mechanisms of alternative pre-mRNA splicing. *Science*. **251**: 33 -34.

Mankodi, A. and Ashizawa, T.(2003). Echo of silence: silent mutations, RNA splicing, and neuromuscular diseases. *Neurology*.**61**: 1330 – 1341.

Mankodi, A., Logigian, E., Callahan,L., McClain, C., White, R., Henderson, D., Krym, M. and Thornton, C.A. (2000). Myotonic dystrophy in transgenic mice expressing an expanded CUG repeat. *Science*. **289**: 1769 – 1773.

Mathé, C., Sagot, M.F., Schiex, T and Rouze, P. (2002). Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res*. **30**: 4103-4117.

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

Mendell, J. T. & Dietz, H. C. (2001). When the message goes awry: disease-producing mutations that influence mRNA content and performance. *Cell*. **107**: 411–414.

Mendez, M., Sorkin, L., Rossetti, M.V., Astrin, K.H. Batlle, C., Parera, V.E., Aizencang, G and Desnick, R. J. (1998). Familial porphyria cutanea tarda: characterization of seven novel uroporphyrinogen decarboxylase mutations and frequency of common hemochromatosis alleles. *Am J Hum Genet*. **63**: 1363–1375

Miao, X., Yu, C., Tan, W, Xiong, P, Liang G., Lu, W. and Lin, D. (2003). A functional polymorphism in the matrix metalloproteinase-2 gene promoter (-1306C/T)is associated with risk of development but not metastasis of gastric cardia adenocarcinoma. *Cancer Res*. **63**: 3987 – 3990.

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

Mizuguchi et al (2004). Heterozygous TGFBR2 mutations in Morfan syndrome. *Nature. Genet*. **36**: 855 – 860.

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

Modrek, B. and Lee, C. (2003). Alternative splicing in the human, mouse, and rat genomes is associated with an increased rate of exon creation/loss. *Nat. Genet*. **34**: 177 -180.

Montagna , M., Agata, S., De Nicolo, A., Menin, C., Sordi, G., Chieco-Bianchi, L. and D'Andrea, E. (2002). Identification of BRCA1 and BRCA2 carriers by allele-specific gene expression (AGE) analysis. *Int. J. Cancer*. **98**: 732 – 736.

Montera, M., Piaggio, F., Marchese, C., Gismondi, V., Stella, A., Resta, N., Varesco, L., Guanti, G. and Mareni, C.(2001). A silent mutation in exon 14 of the APC gene is associated with exon skipping in a FAP family. *J. Med. Genet.***38**: 863 – 867.

Montgomery, S.B., Griffith,OL., Schuetz, J.M., Brooks-Wilson, A., Jones, S.J.M. (2007). A Survey of Genomic Properties for the Detection of Regulatory Polymorphisms. *PLoS Comput Biol.* **3**:1000 – 1010.

Moseley,C.T., Mullis, P.E ., Prince, M.A., and Phillips III, J.A. (2002). An exon splice enhancer mutations causes autosomal dominant G.H deficiency. *J. Clin. Endocrinol. Met.* **87**: 847 – 852.

Mount, S.M. (1982) A catalogue of splice junction sequences. *Nucleic Acids Res.* **10**:459–72.

Mukai, J., Liu, H., Burt, R.A., Swor, D.E., Lai, W.S., Karayiorgou, M., Gogos, J.A. (2004). Evidence that the gene encoding ZDHHC8 contributes to the risk of schizophrenia. *Nat Genet.* **36**: 674-675.

Nei, M. & Kumar, S. (2000). Molecular Evolution and Phylogenetics. Oxford University press. New York.

.

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

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.

Nielsen, R. (2005) Molecular signatures of Natural Selection. *Ann. Rev.Genet.* **39**: 197 -218.

Nissim-Rafina, M. and Kerem, B. (2002) Splicing regulation as a genetic modifier. *Trends. Genet.* **18**: 123 -12.

Norton, P.A. (1994). Alternative pre-mRNA splicing: factors involved in splice site selection. *J. Cell Sci.* **107**: 1 -7.

Oleksiak, M.F., Gary A. Churchill, G.A. and Crawford, D. L. (2002). Variation in gene expression within and among natural populations. *Nature Genet.* **32**: 261 – 266.

Ozsahin, H., Arredondo-Vega, F.X., Santisteban, I., Fuhrer, H., Tuchschmid, P., Jochum, W., Aguzzi, A., Lederman, H.M., Fleischman, A., Winkelstein, J.A., Seger R.A and Hershfield, M.S. (1997). Adenosine deaminase deficiency in Adults. *Blood.* **89**: 2849 – 2855.

Parmely, J.L and Hurst, L.D. (2007). Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Mol. Biol. Evol.* **24**:1600 – 1603.

Parmley, J. L., Chamary J. V, and Hurst, L.D.(2006). Evidence for Purifying Selection against Synonymous Mutations in Mammalian Exonic Splicing Enhancers. *Mol.Biol.Evol.* **23**.301 – 309.

Phillips, D.L., Park, J.W., Graveley, B.R. (2004). A computational and experimental approach towards a priori identification of alternatively spliced exons. *RNA*. **10**: 1838 – 1844.

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.

Platzer, M. and Hiller, M. (2006). Sequencing error or SNPs at splice-acceptor guanines in dbSNP? *Nat. Biotech.* **24**: 1068 -1070.

Pospisil H, Herrmann A, Bortfeldt R, Reich J. (2004).EASED: Extended Alternatively Spliced EST Database. *Nucleic Acids Res.* **32**. 70 -74.

Prasad, J., Colwill, K., Pawson, T. and Manley, J.L.(1999).The protein kinase Clk/Sty directly modulates SR protein activity: both hyper-and hypophosphorylation inhibit splicing. *Mol.Cell.Biol.* **19**: 6991 – 7000.

Prokunina, L., Castillejo-Lopez, C., Oberg, F., Gunnarsson, I, Berg, L., Magnusson V., Brookes A.J., Tentler, D., Kristjansdóttir, H., Gröndal, G., Bolstad, A.I., Svenungsson, E., Lundberg, I., Sturfelt, G., Jönssen, A., Truedsson, L., Lima, G., Alcocer-Varela, J., Jonsson, R., Gyllensten, U.B., Harley, J.B., Alarcón-Segovia, D., Steinsson, K. and Alarcón-Riquelme, M.E. (2002) A regulatory polymorphism in PDCD1 is associated with susceptibility to systemic lupus erythematosus in humans. *Nat. Genet.* **32**: 666 – 669.

Qu, H., Lu, Y., Marchard, L., Bacot, F., Frechette, R., Tessier, M.C., Montpetit, A and Polychronakos, C. (2007). Genetic Control of alternative splicing in the TAP2 gene possible implication in the genetics of Type 1 Diabetes. *Diabetes*. **56**: 270 – 275.

Ramensky, V., Bork, P., Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **30**: 3894 - 3900.

Rees, D. J. G., Rizza C. R. and Brownlee, G.G (1985). Haemophilia B caused by a point mutation in a donor splice junction of the human factor IX gene. *Nature*.**316**: 643- 645.

Reich, D.E., Gabriel, SB., Altshuler, S.B.(2004), Quality and completeness of SNP databases. *Nat. Genet*.**33**: 457 – 458.

Reumers, J., Conde, L., Medina, I., Maurer-Stroh, S., Van Durme, J., Dopazo, J., Rousseau, F. and Schymkowitz, J. (2008). Joint annotation of coding and non-coding single nucleotide

polymorphisms and mutations in the SNPeffect and pupaSuite Databases. *Nucleic Acids Res.* **36**: 825 – 829.

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.

Ruskin, B., Krainer, A.R., Maniatis, T., Green, M.R. (1984). Excision of the intact intron as a novel lariat structure during pre-mRNA splicing in vitro. *Cell.* **38**: 317 – 331.

Sabeti, P.C. et al (2006). Positive selection in the human lineage. *Science.* **312**: 1614 – 1620.

Schena, M., Shalon, D., Davis, R.W, Brown, P.O.(1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science.* **270**:467-70.

Schimf, S., Schaich, S and Wissinger, B. (2006). Activation of cryptic slice site is a frequent splicing defect mechanism caused by mutations in exon and intron sequences of the OPA1 gene. *Hum. Genet.* **118**: 767 – 771.

Schwarz, D.F., Hädicke, O., Erdmann, J., Ziegler, A., Bayer D and Möller, S. (2008). SNPtoGO: characterizing SNPs by enriched GO terms. *Bioinformatics.* **24**: 146 – 148.

Serre, D. and Hudson, T.J. (2006). Resources for Genetic Variation Studies. *Annu. Rev. Genomics Hum. Genet.* **7**: 443 – 457.

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.

Sherry, S.T., Ward, M .H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M and Sirotkin, K (2001) dbSNP: 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.

Singh, A., Olowoyeye, A., Baenziger, P.H., Dantzer, J. Kann, M.G.,Radivojac, P., Heiland, R. and Mooney, S.D.(2008). MutDB: update on development of tools for the biochemical analysis of genetic variation. *Nucleic. Acids Res.* **36**: D815-D819.

Skrygan, M., Bartholomé B., Bonafé L., Blau N., Bartholomé K. (2001). A splice mutation in the GTP cyclohydrolase I gene causes dopa-responsive dystonia by exon skipping. *JIMD.* **24:** 345-351

Smith,C.W., Patton, J.G. and Nadal-Ginard, B. (1989). Alternative Splicing in the Control of Gene Expression. *Annu. Rev. Genet.* **23**: 527 – 577.

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

Smith, C.W. and Roberts, G.C (2002). Alternative splicing: combinatorial output from the genome. *Curr Opin Chem Biol.***6**: 375 – 383.

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., Dror, G. and Shamir, R. (2006). Assessing the number of ancestral alternatively spliced exons in the human genome. *BMC Genomics.* **7**: 273.

Sorek, R., Shanir, R., and Ast, G. (2004). How prevalent is functional alternative splicing in the human genome? *Trends Genet.* **20**: 68 – 71.

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.

Stamm, S., Zhu, J., Nakai, K., Stoilov,P., Stoss, O and Zhang, M. Q.(2000) An alternative–exon database and its statistical analysis. *DNA Cell. Biol.* **19**: 739 – 756.

Stamm, S.(2002). Signals and their transduction pathways regulating alternative splicing: a new dimension of the human genome. *Hum.Mol. Genet.* **11**: 2409 – 2416.

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

Stranger, B.E., Forrest, M.S., Clark, A.G., Minichiello, M.J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S.E., Tavaré, S., Deloukas, P and Dermitzakis, E.T. (2005) Genome-Wide Associations of Gene Expression Variation in Humans. *PLoS Genetics*. **1**: e78.

Sunyaev., S.R., Lathe III, W.C., Ramensky, V.E. and Bork P. (2000). SNP frequencies in human genes - an excess of rare alleles and differing modes of selection. *Trends Genet*, **16**: 335-337.

Tacke, R., Chen, Y and Manley, J.L. (1997). Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation: Creation of an SRp40-specific splicing enhancer. *PNAS*. **94**:1148–1153.

Tarn, W. (2007). Cellular signals modulate alternative splicing. *Biomed. Sci.* **14**: 517 – 522.

Thanaraj TA, Clark F.(2001). Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids Res.* **29**:2581-93.

Thanaraj, T.A., and Stamm, S. (2003) Prediction and statistical analysis of alternatively spliced exons. *Prog. Mol. Subcell. Biol.* **31**: 1 -31.

The International HapMap Consortium. (2005). The haplotype map of the human genome. *Nature*.**437**: 1299 – 1320.

Treisman, R., Orkin, S, and Maniatis, T. (1983). Specific transcription and RNA splicing defect in five cloned B-thalassemia genes. *Nature*. **302**: 591 – 596.

Tromp, G and Prockop, D.J.(1988).Single Base Mutation in the Proα 2(I) Collagen Gene that Causes Efficient Splicing of RNA from Exon 27 to Exon 29 and Synthesis of a Shortened but In-Frame Proα 2(I) Chain. *PNAS*. **85**: 5254-5258.

Velculescu, V.E., Madden, S.L., Zhang, L., Lash, A.E., Yu, J., Rago, C., Lal, A., Wang, C.J., Beaudry, G.A., Ciriello, K.M., Cook, B.P., Dufault, M.R., Ferguson, A.T., Gao, Y., He, T.C., Hermeking, H., Hiraldo, S.K., Hwang, P.M., Lopez, M.A., Luderer, H.F., Mathews, B., Petroziello, J.M., Polyak, K., Zawel, L., Kinzler, K.W., et al. (1999). Analysis of human transcriptomes. *Nat. Genet.* **23**: 387 - 388.

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. et al. (2001) The sequencing of the human genome. *Science*. **291**: 1304 – 1351

Vithana, E.N., Abu-Safieh, L., Allen, M.J., Carey, A., Papaioanou, M., Chakarova, C., Al- Maghtheh, M., Ebenezer, N.D., Willis, C., Moore, A.T et al. (2001). A human homolog of yeast pre-RNA splicing

gene, PRP31, underlies autosomal dominant retinitis pigmentosa on chromosome 19q13.4 (RP11). *Mol. Cell.* **8**: 375 – 381.

Voelker, R.B and Berglund, J.A. (2007). A comprehensive computational characterization of conserved mammalian intronic sequences reveals conserved motifs associated with constitutive and alternative splicing. *Genome Res.* **17**: 1023 – 1033.

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, 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.

Watkin, S., Madison, J., Davis, E., Sakamoto, Y., Galliano, M., Minchiotti, L and Putman, F.W. (1991). A donor splice mutation and a carboxyl-terminal variants of human serum albumin. *PNAS.* **88**: 5959 – 5963.

Wenstrup, R.J., Langland, G.T., Willing, M.C., D'Souza V.N., 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.

Wieringa, B., Meyer, F., Reiser, J., and Weissmann, C. (1983). Unusual splice sites revealed by mutagenic inactivation of an authentic splice site of the rabbit B-globin gene. *Nature*. **301**: 38 -43.

Willie, E & Majewski, J.(2004).Evidence for codon bias selection at the pre-mRNA level in eukaryotes. *Trends Genet*. **20**: 534 – 538.

Woodley, L and Valcarcel, J. (2002). Regulation of alternative pre-mRNA splicing. *Brief Funct Genomic Proteomic.***3**: 266 – 277.

Wu, J.Y and Maniatis, T. (1993). Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell.***75**: 1061 – 1070.

Xiao, S.H and Manley, J.L. (1998). Phosphorylation-dephosphorylation differentially affects activities of splicing factor ASF/SF2. *EMBO J.* **17**: 6359 – 6367.

Xing, Y and Lee, C.J. (2005). Protein modularity of alternatively spliced exons is associated with tissue specific regulation of alternative splicing. *Plos Genet*. **1**: e34.

Xu, Q., Modrek, B. and Lee, C. (2002). Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res*. **30**: 17 – 20.

Yeo, G. W., Van Nostrand, E., Holste, D., Poggio, T. and Burge, C.B. (2005). Identification and analysis of alternative splicing events conserved in human and mouse. *PNAS*, **102**: 2850 - 2855.

Yeo, G., Holste, D., Kreiman, G, and Burge, C.B. (2004). Variation in alternative splicing across human tissues. *Genome Biol*. **5**: R74.

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

Zahler, A.M., Neugebauer, K.M., Lane, W.S., Roth, M.B. (1993). Distinct function of SR proteins in alternative pre-mRNA splicing. *Science*. **260**: 219 – 222.

Zhang, M.Q.(1998). Statistical features of human exons and their flanking regions. *Hum. Mol. Genet*. **7**: 919 – 932.

Zhang, X. H-F and Chasin, L.A. (2006). Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *PNAS,***103:** 13427 – 13432.

Zheng. C.L, Fu, X. D. and Gribskov, M. (2005). Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA*. **11**: 1777 – 1787.

Zhoa, Z., Fu, Y.X., Hewett- Emmett, D., and Boerwinkle, E.(2003). Investigating Single nucleiotide polymorphism (SNP) density and its implication for molecular evolution. *Gene*.**312**: 207 – 213.

Zhu H, Tucker HM, 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.

Zuo, P. and Maniatis, T.(1996). The splicing factor U2AF35 mediates critical protein-protein interactions in constitutively and enhancer-dependent splicing. *Genes Dev*. **10**: 1356 – 1368.