Low Detection of Exon Skipping in Mouse Genes Orthologous to Human Genes on Chromosome 22

by

Tzu-Ming Chern

Thesis presented in fulfillment of the requirements for the Degree of Master Scientiae at the South African National Bioinformatics Institute, Department of Biochemistry, Faculty of Natural Sciences, University of the Western Cape

April 2002

Advisor: Professor Winston Hide

ABSTRACT

Alternative RNA splicing is one of the leading mechanisms contributing towards transcript and protein diversity. Several alternative splicing surveys have confirmed the frequent occurrence of exon skipping in human genes. However, the occurrence of exon skipping in mouse genes has not yet been extensively examined. Recent improvements in mouse genome sequencing have permitted the current study to explore the occurrence of exon skipping in mouse genes orthologous to human genes on chromosome 22. A low number (5/72 multi-exon genes) of mouse exon-skipped genes were captured through alignments of mouse ESTs to mouse genomic contigs. Exon-skipping events in two mouse exon-skipped genes (GNB1L, SMARCB1) appear to affect biological processes such as electron and protein transport. All mouse, skipped exons were observed to have ubiquitous tissue expression. Comparison of our mouse exon-skipping events to previously detected human exon-skipping events on chromosome 22 by Hide *et al.*2001, has revealed that mouse and human exon-skipping events were never observed together within an orthologous gene-pair. Although the transcript identity between mouse and human orthologous transcripts were high (greater than 80% sequence identity), the exon order in these gene-pairs may be different between mouse and human orthologous genes. Main factors contributing towards the low detection of mouse exon-skipping events include the lack of mouse transcripts matching to mouse genomic sequences and the under-prediction of mouse exons. These factors resulted in a large number (112/269) of mouse transcripts lacking matches to mouse genomic contigs and nearly half (12/25) of the mouse multi-exon genes, which have matching Ensembl transcript identifiers, have under-predicted exons. The low frequency of mouse exon skipping on chromosome 22 cannot be extrapolated to represent a genome-wide estimate due to the small number of observed mouse exon-skipping events. However, when compared to a higher estimate (52/347) of exon skipping in human genes for chromosome 22 produced under similar conditions by Hide *et al.*2001, it is possible that our mouse exon-skipping frequency may be lower than the human frequency. Our hypothesis contradicts with a previous study by Brett *et al.*2002, in which the authors claim that mouse and human alternative splicing is

comparable. Our conclusion that the mouse exon-skipping frequency may be lower than the human estimate remains to be tested with a larger mouse multi-exon gene set. However, the mouse exon-skipping frequency may represent the highest estimate that can be obtained given that the current number (87) of mouse genes orthologous to chromosome 22 in Ensembl (v1 30th Jan. 2002) does not deviate significantly from our total number (72) of mouse multi-exon genes. The quality of the current mouse genomic data is higher than the one utilized in this study. The capture of mouse exon-skipping events may increase as the quality and quantity of mouse genomic and transcript sequences improves.

April 2002

DECLARATION

I declare that *Low Detection of Exon Skipping in Mouse Genes Orthologous to Human Genes on Chromosome 22* is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Tzu-Ming Chern **April 2002**

Signed:

BIOGRAPHICAL SKETCH

Tzu-Ming Chern was born in Taipei, Taiwan on the 19th of November 1976. She has traveled extensively during her childhood and her family has immigrated to South Africa during 1991 after the gulf war. Tzu-Ming has double-majored in genetics and biochemistry during her BSc degree at Wits, which was completed in 1998. In 1999, she joined a protein-structure group at Wits where she completed her BSc. Honours degree in biochemistry. Recognizing the importance of acquiring computer science skills as a budding scientist, Tzu-Ming has been invited to join a bioinformatics group at the South African National Bioinformatics Institute at UWC where she will complete her MSc degree with Professor Winston Hide. Her MSc research project is entitled "Low Detection of Exon Skipping in Mouse Genes Orthologous to Human Genes on Chromosome 22" and was completed exclusively in an *in silico* laboratory.

ACKNOWLEDGEMENTS

The current study was performed at the South African National Bioinformatics Institute, University of Western Cape.

This thesis would not have been possible without the generous support from the following special people listed below:

1. Mom and Dad- Thank you both for your financial and motivational support
2. Tzu-Mei – Thank you for your free psychological services.
3. Professor Winston Hide – Thank you for your financial, transport, psychological-counselling, strategic-planning, motivational, marketing-skill, career-strategic, writing, accommodations, and rations support. It was a wonderful challenge and opportunity to have worked with you.
4. Tania Broveak Hide – Thank you for your transport, psychological, accommodations, and rations support.
5. Janet, Junaid, Bob, Alan, and Soraya – Special thanks for your innumerable guidance!
6. Thanks to all my friends that I have made during the course of my thesis for their understanding and patience.

TABLE OF CONTENTS

Page

Page

LIST OF FIGURES & TABLES IN THE LITERATURE REVIEW

Page

Tables:

LIST OF FIGURES & TABLES IN THE THESIS

LITERATURE REVIEW

INTRODUCTION

Walter Gilbert originally described alternative RNA splicing as the production of multiple mRNAs with different protein functions derived from a single gene (Gilbert 1978). Tremendous interest has grown in the field of alternative RNA splicing due to several reasons: (1.) The involvement of alternative splicing in both protein functional diversity (Graveley 2001; Black 2000) and in human genetic disorders (Philips & Cooper 2000; Grabowski & Black 2001) marks its biological value in academic and commercial sectors, (2.) Estimates of alternative splicing (Mironov *et al.*1999; Brett *et al.*2000; Hanke *et al.*1999; Modrek *et al.*2001; Clark & Thanaraj 2002; Croft *et al.*2000; Hide *et al.*2001) have suggested that this process occurs frequently in human genes, (3.) Low gene estimates (International Human Genome Sequencing Consortium 2001; Celera Genomics 2001) suggest that alternative splicing may account for the abundant transcript and protein diversity (Graveley 2001; Black 2000; Wright *et al.*2001) that exists in human genes, (4.) Exon skipping, which belongs to one of the five known alternative splicing processes, has recently emerged as the most frequent alternative splicing pattern observed in human genes (Modrek *et al.*2001; Clark & Thanaraj 2002; Croft *et al.*2000; Hide *et al.*2001), and (5.) The roles and mechanisms of alternative splicing in eukaryotic species are still unclear.

Current research efforts (Goldstrohm *et al.*2001; Smith & Valcárcel; Collins & Guthrie 2000) are focused on elucidating biochemical mechanisms and identification of relevant components involved in alternative splicing using laboratory methods. Computational approaches can also assist in the rapid identification and characterization of alternatively spliced transcripts that may otherwise prove to be time-consuming if laboratory techniques are solely employed. Alignments of expressed sequence tags (ESTs) to genomic sequences have been utilized by many studies (Mironov *et al.*1999; Modrek *et al.*2001; Clark & Thanaraj 2002; Croft *et al.*2000; Hide *et al.*2001;

International Human Genome Sequencing Consortium 2001; Kan *et al.*2001; Wolfsberg & Landsman 1997; Thanaraj 1999; Kan *et al.*2000) to determine estimates of alternatively spliced events in human genes. ESTs are a useful and abundant resource for capturing alternatively spliced events. Current mouse and human EST counts (Genbank release 125) at NCBI (US National Center for Biotechnology Information) are estimated to be ~2.3 and ~4.3 million respectively.

Recent sequencing of the mouse genome provides the opportunity to explore alternative splicing in mouse genes. Previous alternative splicing studies in mouse utilizing in silico methods have been few (Kan *et al.*2001; Brett *et al.*2002) due to the lack of complete mouse genomic sequences. Sufficient mouse genomic sequences are currently available at the Ensembl website (http://www.ensembl.org/Mus_musculus/), which permits mouse alternative splicing studies to be performed.

Thesis Rationale

The availability of mouse genomic sequences and the prevalence of exon skipping in human genes have motivated the current study to investigate the occurrence of mouse exon skipping in mouse genes orthologous to human genes on chromosome 22. The capture of mouse exon-skipping events in mouse genes orthologous to human genes on chromosome 22 will be performed through alignments of mouse ESTs to their respective genomic sequences. Results of this investigation will provide the first preliminary estimate of mouse exon skipping, which has yet to be quantified with the completion of mouse genomic sequencing. The current study will also provide the ground work in which hypotheses can be constructed and tested in future mouse alternative splicing studies.

A breakdown of subsequent chapters reviewed in the dissertation literature review are listed below:

(1) The fundamentals of alternative RNA splicing will be described in Chapter 1, which include the following subsections:

- Description of the splicing reaction and the spliceosome.
- The proposed mechanisms leading to alternative RNA splicing will be discussed.

(2) Chapter 2 will discuss the importance of alternative splicing

(3) Chapter 3 will highlight the relevance of exon skipping

(4) Chapter 4 will discuss the utilization of expressed sequence tags to detect alternatively spliced transcripts

(5) Chapter 5 will address the current state of mouse genomic sequencing and the utilities of mouse and human gene comparisons.

Chapter 1: Fundamentals of Alternative RNA Splicing

1.1 What is Alternative RNA splicing?

Most eukaryotic genes undergo RNA splicing, which is a co-transcriptional process in which introns (non-coding sequences) are removed and exons (coding sequences) are ligated from pre-mRNA transcripts to form mature mRNA transcripts. RNA splicing is one of the three RNA processing steps, which prepares eukaryotic RNA transcripts for translation in the cytoplasm. However, for each eukaryotic gene, more than one mature transcript can be generated, hence more than one protein product can be produced from a single gene. The different transcripts that are generated from a single gene also have different exon arrangements and may have different functions. Therefore, alternative RNA splicing is the process by which different transcripts can be generated from a single gene, which may have different functional properties. The proposed mechanisms by which these transcripts can have different exon arrangements are discussed in section 1.3 of this chapter. The following sections prior to section 1.3 will describe the RNA splicing reaction in detail.

1.2.1 The Major Spliceosome

The splicing reaction is catalyzed by a protein-RNA complex known as - the spliceosome. Two types of spliceosomes are known to exist. The major spliceosome recognizes the frequent, GT-AG consensus splice sites, whereas the most recently discovered, low-abundant spliceosome catalyzes rare introns containing AT-AC splice sites (Tarn & Steitz 1997). The major spliceosome consists of five types of small nuclear RNAs (U1, U2, U4, U5, and U6 snRNAs) that interact with at least 50 proteins to form small nuclear ribonucleoproteins (snRNPs). Four snRNPs (U11, U12, U4, and U6) of the minor splicesome are structurally and functionally similar to the U1, U2, U4, and U6 snRNPs of the major spliceosome. Both major and minor spliceosomes have

U5 snRNP in common (Tarn & Steitz 1997). Much research has been performed on the major spliceosome and progress in this area will be the focus of this literature review.

Assembly and catalytic activation of the major spliceosome are necessary to initiate the splicing reaction. Assembly of the spliceosome (see Fig. 1a) occurs with the sequential binding of U1 and U2 snRNPs, followed by the addition of the triple snRNPs (U4/U5/U6). Splicing factors such as SF1 (splicing factor or branch point binding protein) and U2AF (U2 snRNP auxiliary factor), also bind to the conserved intronic sequences during the assembly. Specifically, U1 snRNP binds to the 5' splice site (Hertel *et al.*1997), splicing factor SF1 or branch-point-binding protein binds to the branch point (Berget 1995), and the 65 and 35 kDa subunits of U2 snRNP auxillary factor (U2AF) recognizes the polypyrimidine tract (Lewis *et al.*1995) and 3' splice site (Bentley 1999) respectively (see Fig.1b).

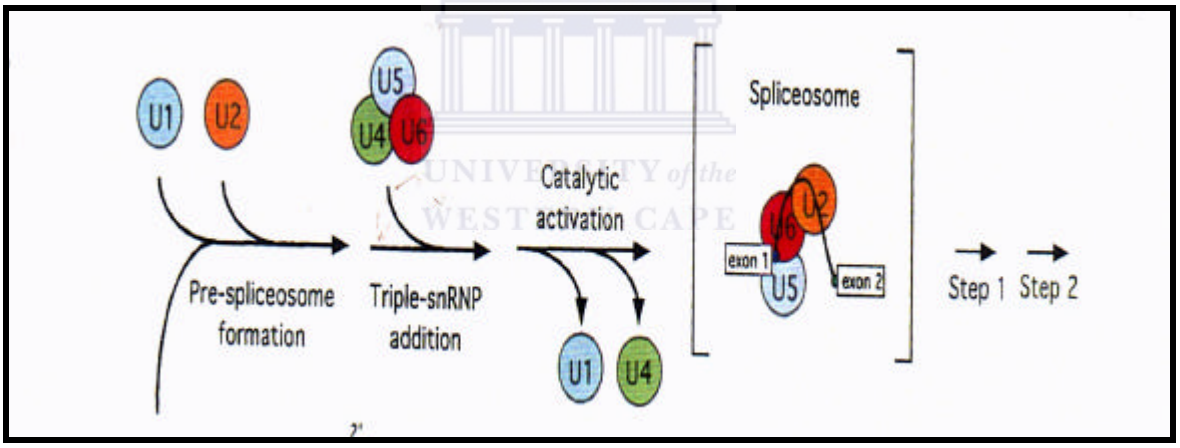


Figure 1a: Assembly and activation of the major spliceosome

Assembly of the major spliceosome requires that U1 and U2 snRNPs bind to their 5' and 3' splice sites respectively followed by the addition of the triple snRNP complex (U4, U5, U6 snRNPs). Activation of the spliceosome is initiated by the release of U1 and U4 snRNPs. The spliceosome inside the large brackets represent the catalytically active spliceosome consisting of U2, U5, and U6 snRNPs (Figure extracted from Collins & Guthrie 2000). Steps 1 and 2 are steps in the splicing reaction as illustrated in Fig. 1c.

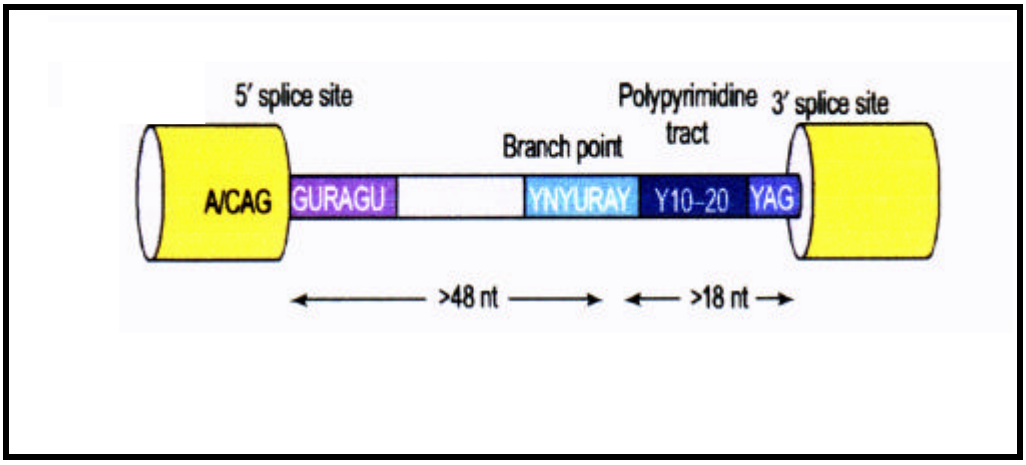


Figure 1b: Intron structure of a pre-mRNA

The diagram illustrates an intron flanked by exons (yellow boxes) in a pre-mRNA. The consensus splice site elements and other intronic splicing regulatory sequences are depicted. Distances from the branch point adenine nucleotide to the splice sites are shown (Figure extracted from Smith & Valcárcel 2000).

Catalytic activation of the spliceosome results in the destabilization of U1 and U4 snRNP (see Fig. 1a), to produce a catalytically active spliceosome that consists of U2, U5, and U6 snRNPs. The U2, U5, and U6 snRNAs forms the catalytic core, which have been proposed to assist in the excision and ligation processes in the splicing reaction. Many studies have supported the evidence that the catalytic steps of pre-mRNA splicing are mediated by spliceosomal RNA. Recent evidence has suggested that a protein, Prp8, plays an important role in spliceosome catalysis by functioning as a cofactor to the spliceosome. Prp8 forms extensive cross-linking interactions with 5' and 3' splice-sites as well as with U5 and U6 snRNAs. Prp8 has been proposed to stabilize tertiary RNA interactions and facilitate the formation of the RNA catalytic core (Collins & Guthrie 2000).

1.2.2 The Splicing Reaction

Once the mature spliceosomal complex becomes activated, initiation of the pre-mRNA splicing reaction and progression through two nucleophilic attacks on the 5' and 3' splice-sites occurs. Fig. 1c describes the splicing reaction in detail. Essentially the splicing reaction can be summarized into 3 sequential steps: (1.) Cleavage of the 5' splice site, (2.) Formation of an intermediate lariat structure, and (3.) Cleavage of 3' splice site coupled with exon ligation and release of excised introns.

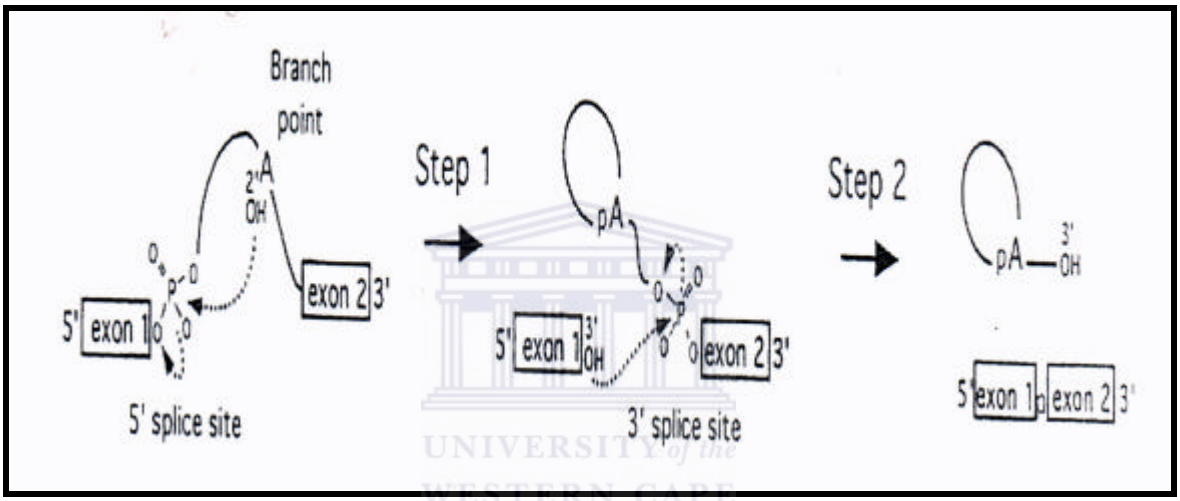


Figure 1c: Pre-mRNA splicing reaction

The splicing reaction consists of three sequential steps: 1.) Cleavage of the 5' splice site is accomplished through a nucleophilic attack by the 2' hydroxyl group of the branch point adenine nucleotide, 2.) Formation of a lasso-shaped structure resulting from the nucleophilic attack by the 2' hydroxyl group on the phosphate group at the 5' splice site, and 3.) Cleavage of the 3' splice site is achieved through a nucleophilic attack by the 3' hydroxyl group of the cleaved 5' exon. Ligated exons and an excised lariat intron are the products that are generated from a splicing reaction (Figure extracted from Collins & Guthrie 2000).

1.3 A Diversity of Mechanisms Affecting Splice-site Choice

The process that allows for the selection of different combinations of splice sites within a pre-mRNA is termed alternative RNA splicing. It is precisely the selection of different splice-sites that result in the diverse number of mRNA variants with different alternative splicing patterns. Alternative splicing patterns result from the usage of alternative 5' splice sites, 3' splice sites, mutually exclusive exons, and intron retention (see Fig. 2). Tissue- and developmentally-specific transcripts are characteristic consequences of alternative splicing. Smith & Valcárcel 2000, have proposed that tissue-specificity may be achieved through a combination of diverse mechanisms that affect splice site choice (Smith & Valcárcel 2000).

Splice-site selection may be influenced by a combination of the following five factors listed below:

- Intrinsic strength of splice sites – discussed in section 1.3.1
- Exon size restriction limits – discussed in section 1.3.2
- Exonic and intronic splicing enhancer or silencer elements – discussed in section 1.3.3.1-1.3.3.2
- Splicing factors that include the serine-arginine (SR) proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs)-discussed in section 1.3.4.1- 1.3.4.2
- Signal transduction pathways and post-translational modifications – discussed in section 1.3.5

The roles of how each factor affects splice-site choice will be discussed in the following subsections.

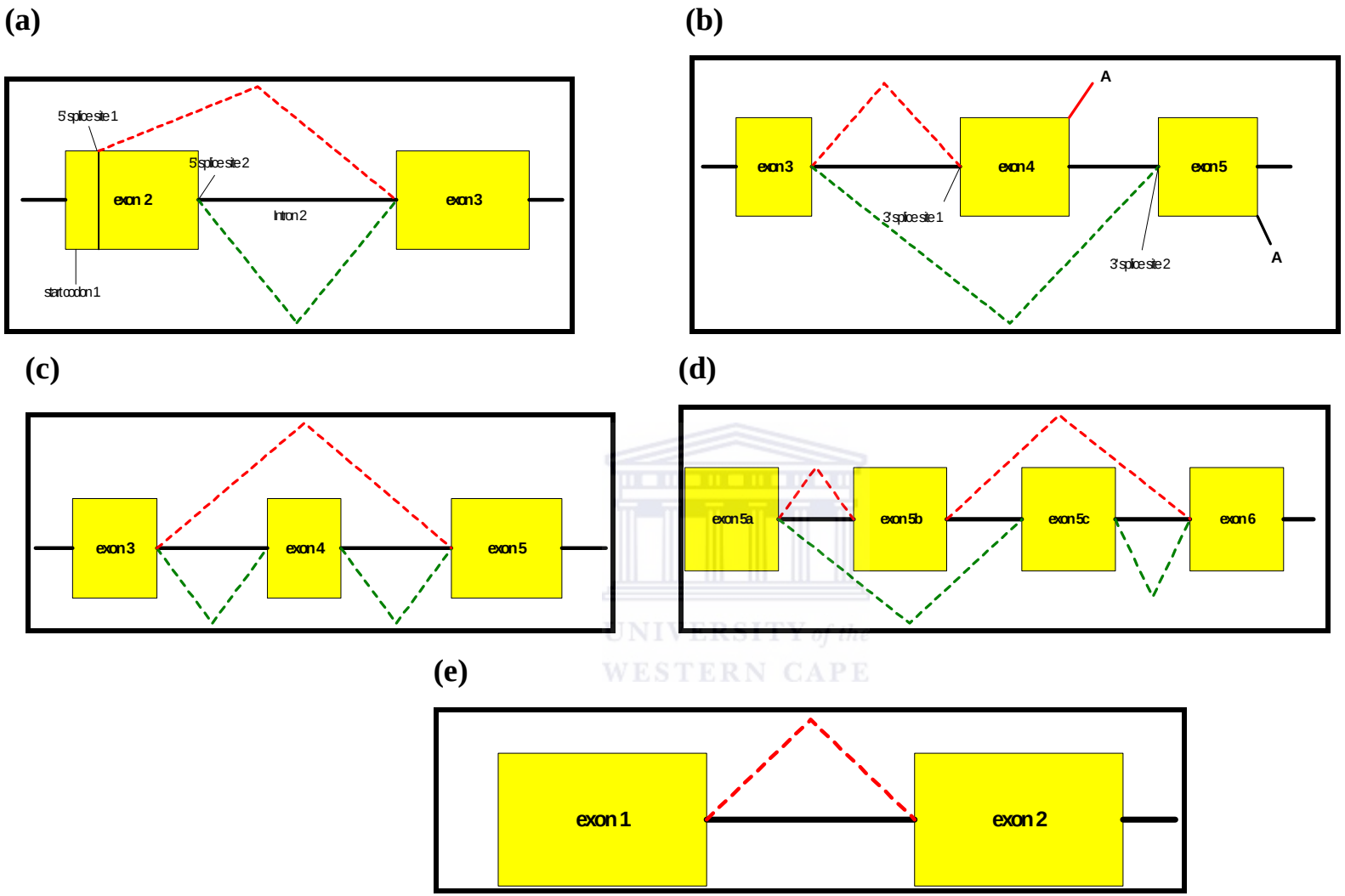


Figure 2: Five known alternative splicing patterns

Alternative splicing patterns are displayed as a.) Alternative 5' splice sites, b.) Alternative 3' splice sites, c.) Cassette exon, d.) Mutually exclusive exons, and e.) Intron retention. Dark black solid lines represent introns. Dotted lines (green and red) represent the alternative splicing of introns. Mutually exclusive exons refer to alternative exons that are never found together in a transcript isoform.

UNIVERSITY of the
WESTERN CAPE

1.3.1 Intrinsic Strength of Splice-sites

Splice sites that follow the GT-AG rule (general consensus) are usually preferred by the spliceosome as compared to those splice sites that do not follow the general consensus sequences. Consensus splice sites have been observed to enhance exon recognition (López 1998).

1.3.2 Exon Size Restriction Limit

The splicing machinery has an exon size restriction limit whereby the optimal exon length should be between 50 to 300 nucleotides. Observations have shown that if the exon length exceeds 300 nucleotides, the spliceosome assembly is inhibited. If the exon length is below 50 nucleotides, the exon may not be recognized, resulting in the skipping of the exon. Although, long vertebrate exons do exist, the mechanisms by which the exon length restrictions are bypassed are unknown (Berget 1995).

1.3.3 Splicing Regulatory Elements

The splicing regulatory elements that assist in splice-site recognition occur either in introns or exons. These short, degenerate splicing elements should not be confused with transcriptional enhancers or silencer elements that activate or repress transcription (Ogbourne & Antalis 1998) respectively.

1.3.3.1 Intronic Splicing Enhancers

Splicing elements occurring in introns are known as intronic splicing enhancers. Intronic splicing enhancers are known to promote the inclusion of tissue-specific exons. In the *cTNT* gene, muscle-specific intron splicing enhancers flank the alternative exon 5, which results in the inclusion of exon 5 in embryonic striated muscles and the exclusion of exon 5 in adult striated muscles (Ryan & Cooper 1996).

1.3.3.2 Exonic Splicing Silencers and Enhancers

Splicing regulatory elements that are found in exons are known as exon splicing silencers and enhancers. An example of a known exon-splicing silencer is found in the P-element transposase gene, which is produced in the germline cells of *Drosophila*. The exon-splicing silencer occurs in exon 3 and promotes intron retention in somatic cells, which further leads to the production of a repressor of transposase (Adams *et al.*1996).

The exon splicing enhancers (ESEs) functions to enhance the use of specific splice sites. Two general classes of enhancers have been identified and are known as the purine-rich and A/C rich enhancers. An example of a purine-rich enhancer is found in exon 4 of the *Drosophila* gene, doublesex (*dsx*). The purine-rich, enhancer element promotes the female-specific inclusion of exon 4 (Ryner & Baker 1991). A/C-rich enhancers or "ACE" also promote the splicing of exons found in the human calcitonin gene (exon4) (van Oers *et al.*1994).

It is important to note that splicing silencer and enhancer elements alone do not affect splice-site choice directly. Instead, these elements are often associated with the binding of specific splicing factors that are essential to the splicing machinery or are present in specific tissues or cell-types.

Most of the known splicing enhancers function together with the binding of serine-arginine (SR) family of proteins. SR proteins are discussed under section "1.3.4.2".

1.3.4 Splicing Factors

Splicing regulatory proteins that are found in mammalian cells include the SR family of RNA-binding proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs).

Examples of how SR-proteins and hnRNPs regulate splice-site selection will be described in the subsequent sub-sections.

1.3.4.1 The Role of HnRNPs in Alternative Splicing

HnRNPs are nuclear RNA-binding proteins that form complexes with nascent RNA polymerase II transcripts. HnRNPs contain RNA-binding motifs and protein domains that have unusual amino-acid distributions. These RNA-binding proteins have been depicted as dynamic structures with abundant structural transitions occurring during mRNA biogenesis and transport. Immunopurification and two-dimensional gel electrophoretic techniques have identified numerous, uncharacterized hnRNPs. A diverse number of putative functions have been proposed for hnRNPs and these include:

- DNA/RNA strand displacement and annealing activities, which allow for the organization of polynucleotide structures
- Involvement in nucleocytoplasmic transport
- Implication in cytoplasmic mRNA trafficking pathways
- Action as a transcriptional activator or repressor
- Involvement in telomere length maintenance
- Involvement in mRNA translation and turnover
- Involvement in alternative splicing

The functions of hnRNPs have been reviewed recently and will not be discussed here (Krecic & Swanson 1999 and references therein). However, the role of hnRNP proteins in regulating splice-site selection is discussed as follows. HnRNPs can negatively regulate splice-site choice either by blocking access to splice sites from other splicing factors or binding to splicing silencer elements thereby promoting the exclusion of alternative exons. On the other hand, hnRNPs can positively regulate splice-site choice by binding to splicing enhancer elements, thus promoting the inclusion of alternative exons.

Examples of negative regulation include the *Drosophila* SXL (sex-lethal), hnRNP A/B, and hnRNP H proteins. An example of positive regulation is described for PTB (polypyrimidine tract binding protein).

Negative regulation of splice-site choice by blocking access to splice sites

The *Drosophila* SXL protein is an hnRNP-like protein that is produced only in female flies. The SXL protein is involved in the *Drosophila* sex determination pathway. The SXL protein initiates the female-specific splicing of the *tra* pre-mRNA by binding to the more upstream of the two 3' competing splice sites (see Fig. 4). Blocking of the 3' splice site upstream of exon 2 causes U2AF^{65} to bind to the polypyrimidine tract, located just upstream of the female-specific 3' splice site of exon 3. Selection of the female-specific 3' splice site results in the production of the female-specific functional TRA protein, which is only produced in female flies (Chabot 1996). The *Drosophila* SXL protein is also an example of how tissue-specific transcription factors can influence splice-site selection.

Negative regulation of splice-site choice by utilizing exonic splicing silencers

The HIV1-tat exon 2 contains exonic splicing silencers (ESSs) that are bound by the hnRNP A/B proteins. The binding of hnRNP proteins to ESSs results in the skipping of the HIV-1 tat exon 2. The skipping of tat exon 2 is essential for the production of a functional tat protein. It has been suggested that the binding of hnRNP A/B proteins to ESSs may contribute towards the assembly of a non-functional complex containing U2 and U1 snRNPs, thereby resulting in the skipping of tat exon 2 (Caputi *et al.*1999)

In the rat -tropomyosin gene (-*TM*), the 5' end of the skeletal muscle-specific exon 7 contains an ESS that is bound by hnRNP H. Binding of hnRNP H to the silencer element results in the skipping of a muscle-specific exon in non-muscle cells. The

skipping of the muscle-specific exon 7 has been proposed to occur through the blocking of the upstream 3' splice-site of exon 7, thereby preventing the spliceosomal components from assembling near the 3' splice-site (Chen *et al.*1999).

Positive regulation of splice-site choice by utilizing intronic splicing enhancers

Polypyrimidine tract-binding proteins (PTBs), which are vertebrate hnRNP proteins, are known to bind to the polypyrimidine tracts near 3' splice-sites. PTBs are often known to cause inhibition of exon recognition. However, PTBs can also positively regulate the inclusion of an alternative 3'-terminal exon 4 of the human calcitonin/calcitonin gene-related peptide (*CT/CGRP*) gene by binding to a complex enhancer element located in intron 4 (Lou *et al.*1999).

1.3.4.2 SR proteins

Human SR proteins are composed of amino-terminal RNA-binding and C-terminal protein-binding domains. The C-terminal protein-binding domains are rich in alternating serine and arginine residues (SR domain). The RNA-binding domain consists of one or two repeats of the ribonucleoprotein (RNP-CS) motifs that are responsible for the RNA-binding specificity of SR proteins. SR domains have been proposed to function in the mediation of protein-protein interactions (Wu & Maniatis 1993; Kohtz *et al.*1994), RNA-binding (Cáceres & Krainer 1993), promotion of RNA-RNA annealing (Lee *et al.*1993), and function as subcellular localization signals (Li & Bingham 1991). The SR domains can be used to interact with other SR-family and SR-related proteins. SR-related proteins are splicing factors that do not interact with snRNPs directly, but also possess the SR domains.

SR proteins have been implicated in almost every step of the spliceosomal assembly and have been reviewed elsewhere in Valcárcel & Green 1996. Of particular interest,

SR proteins can also regulate 5' splice-site selection. It has been observed that certain SR proteins (SF2 and SC35) promote the selection of 5' splice sites proximal to 3' splice sites (Cáceres *et al.*1994; Wang & Manley 1995). On the other hand, some SR proteins such as, SRp40 and SRp55, promote the utilization of distal 5' splice-sites (Zahler & Roth 1995). But the most important, characteristic feature of SR proteins is their ability to bind to specific exonic splicing enhancers through their RNA-binding domains.

An example of how the binding of an SR protein to an exonic splicing enhancer (ESE) can affect splice-site selection is observed with the *Drosophila dsx* gene. The female-specific exon 4 of the *dsx* gene contains an ESE that consists of six 13-nucleotide repeats, also known as the *dsx* repeat element. A multi-protein complex consisting of TRA, TRA2, and SR-family proteins assembles on the ESE. TRA and TRA2 proteins interact with SR-family proteins through their RS-domains. The multi-protein complex promotes the usage of the upstream weak 3' splice site, which results in the inclusion of exon 4 (Lynch & Maniatis 1996).

The binding of SR proteins to splicing enhancers has also been proposed to enhance the binding of $U2AF^{65}$ to the polypyrimidine tract (Zuo & Maniatis 1996; Bouck *et al.*1998). Enhanced $U2AF^{65}$ binding could be achieved through protein-protein interactions between SR proteins and $U2AF^{35}$ through their RS-domains. Thus, the enhancer-bound SR and $U2AF^{35}$ protein complex is thought to increase $U2AF^{65}$ binding to its polypyrimidine tract (Zuo & Maniatis 1996).

Antagonism between SR and hnRNP proteins affect splice-site selection

The activities of SR proteins in alternative splicing can also be antagonized by the activities of members of the hnRNP A/B family of proteins (Zhu *et al.*2001). An example of antagonism of splicing factors during the selection of splice sites has been observed with the SV40 virus large T and small t proteins. These proteins are produced

through the alternative selection of 5' splice sites. High concentrations of SF2 or ASF splicing factors promote the usage of proximal 5' splice sites whereas increase in concentrations of hnRNP A1 promotes the usage of distal 5' splice sites (Chabot 1996). Thus the different concentrations of antagonizing splicing factors can also determine the different patterns of alternative splicing.

Antagonism between enhancer and silencer splicing elements influences splice-site choice

The splicing of mouse immunoglobulin (*IgM*) exons, has been shown to be regulated by splicing enhancer and silencer elements. Splicing of *IgM* exons requires a purine rich enhancer element located within exon 2 of mouse *IgM* gene. The mouse silencer element is also found in exon 2 near the enhancer sequence. The silencer sequence acts as an inhibitor that binds specifically to U2 snRNP. It has been proposed that the U2 snRNP-inhibitor complex interacts with U1 snRNP to form a non-functional complex thereby disrupting spliceosome assembly. The enhancer element has been proposed to counteract the activity of the inhibitor complex via an unknown mechanism, thereby resulting in the splicing of exon 2. Clearly, this is an example of how the activity of splicing regulatory elements can antagonize each another to achieve a specific pattern of splicing (Kan & Green 1999).

1.3.5 Biological Pathways and Post-translational Modifications Affecting Splice-site Choice

Changing the nuclear ratio of splicing factors can be an important mechanism for splice site selection. It has been shown that varying nuclear ratios of splicing factors, hnRNP A1 and SF2/ASF, influences the selection of alternative 5' splice sites. High concentrations of hnRNP A1 relative to the concentrations of splicing factor, SF2/ASF, results in the selection of distal 5' splice sites. Proximal 5' splice sites are selected when SF2/ASF are in excess (Eperon *et al.*1993). Stress-activation of the $MKK_{3/6}$-p38

kinase pathway (and activation of p38 kinase) has been shown to decrease nuclear hnRNP A1 and increase cytoplasmic hnRNP A1 in the cell. The change in the subcellular distribution of hnRNP A1 also affects the selection of alternative 5' splice sites in the adenovirus E1A pre-mRNA splicing reporter gene. Three isoforms (9S, 12S, and 13S) of the E1A reporter gene are generated through varying concentrations of cytoplasmic hnRNP A1. Increasing cytoplasmic hnRNP A1 will result in the inhibition of the 9S transcript isoform since hnRNP A1 favours the selection of proximal 5' splice sites (see Fig. 3) (van der Houven van Oordt *et al.*2000).

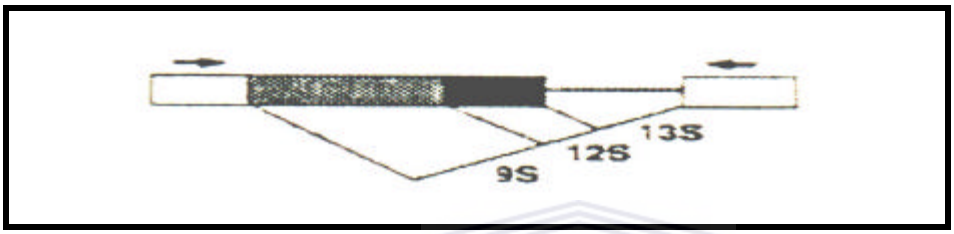


Figure 3: E1A transcript isoforms generated through alternative 5' splice-site usage

White boxes indicate exons and other shaded boxes represent extended exons. The intron sequence is represented by a solid line, which connects the exons. 9S transcript isoforms are generated by selection of the most distal 5' splice site whereas selection of the most proximal 5' splice site results in production of 13S transcript isoforms (Figure extracted from van der Houven van Oordt *et al.*2000).

Activation of the signaling components in the Ras-Raf-MEK-ERK signaling pathway also induces retention of alternative exon 5 in CD44 mature mRNAs. Although the precise molecular interactions of the signaling components between the cell surface and the nuclear splicing machinery have not yet been elucidated, a link between signaling pathways and alternative splicing has been established (Weg-Remers *et al.*2001).

Post-translation modification of SR proteins can also affect splice-site choice. The adenovirus protein E4-ORF4 dephosphorylates SR proteins in adenovirus-infected cells.

In uninfected cells, SR proteins favour the selection of proximal alternative 3' splice sites by binding to intronic repressor elements. In adenovirus-infected cells, dephosphorylated-SR proteins favour selection towards the more distal alternative 3' splice-site. Dephosphorylated-SR proteins also lose their RNA binding capacity for intronic repressor elements (Kanopa *et al.*1998).

Chapter 2: Importance of Alternative Splicing

2.1 The Purpose of Alternative Splicing

Alternative splicing has been observed in genes that are involved in receptor function, developmental processes, and human genetic diseases. In sections 2.1.1-2.1.2, well-studied examples from *Drosophila* and *Homo* sapiens are used to demonstrate the importance of alternative splicing in developmental processes and in generating receptor diversity. Alternative splicing can also result in the production of aberrant transcripts leading to disease. Although, aberrant alternative splicing is not directly relevant to the current work, the importance of aberrant alternative splicing is relevant to understanding why such tremendous interest has been focused on alternative splicing. The relationship between alternative splicing and disease provides another avenue for researchers involved in combating disease. Section 2.1.3 addresses the role of alternative splicing in disease and strategies for treating diseases arising from aberrant alternative splicing.

2.1.1 The Role of Alternative Splicing in Development

2.1.1.1 *Drosophila* sexual development

In the *Drosophila* sex determination pathway in somatic cells (see Fig.4), alternative splicing is utilized extensively to control the expression and functional properties of transcriptional regulators. These regulators are responsible for the sex-specific splicing of pre-mRNA transcripts that are present in both male and female flies. Three important genes encoding transcriptional factors that are crucial in the *Drosophila* sex determination pathway are *sxl*, *tra*, and *dsx* genes, which are present in both male and female flies (Chabot 1996).

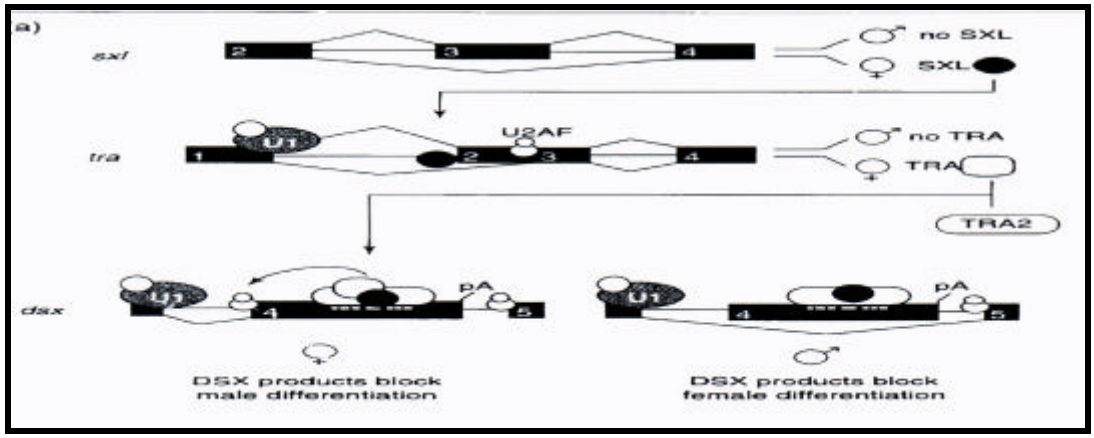


Figure 4: Sex-specific splicing in *Drosophila*

The SXL protein initiates the female-specific splicing by binding to the more upstream of the two competing 3' splice sites to produce a functional female TRA protein. A multi-protein complex consisting of TRA, TRA2, and SXL proteins assembles on the enhancer element present in exon 4 of the *dsx* pre-mRNA and causes U2AF to bind to the female-specific 3' splice site in intron 3. Retention of exon 4 in females results in DSX proteins that inhibit male differentiation (Figure extracted from Chabot 1996).

The sex-specific splicing of the *sxl* gene is the first alternative splicing event to occur in the sex identity pathway. Exon 3 of *sxl* gene is skipped in female *sxl* mRNAs leading to the production of a functional SXL protein. A termination codon is present in exon 3 of *sxl* gene in males and inclusion of this exon in the male mRNAs results in a non-functional SXL protein. SXL protein regulates the splicing of its own transcripts and *msl-2* transcripts (functional MSL-2 proteins are present in males) (MacDougall *et al.*1995). But most importantly, SXL proteins regulate the splicing of *tra* gene, which encodes for the second transcriptional regulator implicated in the sex-determination pathway.

The SXL protein binds to the 3' non-sex-specific splice site upstream of exon2 thereby blocking $U2AF^{65}$ access and activating $U2AF^{65}$-binding to the downstream 3' female-specific splice site. Selection of the female-specific site prevents the inclusion of the

upstream translational stop codon and results in production of functional TRA proteins, which are absent in male flies.

Female-specific splicing of *dsx* transcripts are directed by the formation of a multi-protein complex consisting of TRA, TRA2, and SR proteins. The multi-protein complex assembles on the ESE present in exon4 of the *dsx* pre-mRNA, thereby promoting inclusion of exon 4 by selecting the 3' upstream, female-specific splice site of exon 4.

Active DSX proteins are both functionally different in male (DSX^M) and female (DSX^F) flies and their primary functions are to inhibit the genes required for either male or female differentiation. The products of the *dsx* genes in male and females, maintain the sex-specific state throughout embryonic and larval growth until differentiation. The DSX proteins also function after differentiation, by switching on the expression of yolk protein genes produced only in female adult fat bodies and the glucose dehydrogenase gene. The glucose dehydrogenase genes are expressed in specific patterns in male and female flies (MacDougall *et al.*1995). Thus in every step of the sexual identity pathway, alternative splicing is used to produce male and female-specific transcriptional regulators that drive and maintain the sex-specific splicing during embryonic and larval development.

2.1.1.2 Modulation of binding affinities of the 4.1R protein during erythrocyte morphogenesis

An example of how alternative splicing modulates the binding affinities during erythrocyte morphogenesis, can be observed with the 4.1R gene. The 4.1R protein is a structural protein involved in stabilizing red blood cell membrane skeleton. The interaction of 4.1R protein with other structural proteins such as, spectrin and actin, provide the mature red blood cells with the ability to withstand pressures when traveling through large vessels and narrow capillaries. Exon 16 of the 4.1R gene encodes for the

spectrin-actin binding domain (SABD). The SABD of the 4.1R protein is utilized during its interactions with spectrin and actin. The differential exclusion and inclusion of exon 16 observed in early progenitors and late erythroblasts (immature, nucleated red blood cells) respectively, results in the differential spectrin/actin binding properties, which are characteristic of immature red blood cells during its development into mature erythrocytes. Early progenitors display a low binding affinity towards spectrin/actin and have fragile red blood cell membranes whereas late erythroblasts show a high binding affinity and possess strong, red blood cell membranes (see Fig. 5) (Conboy 1999).

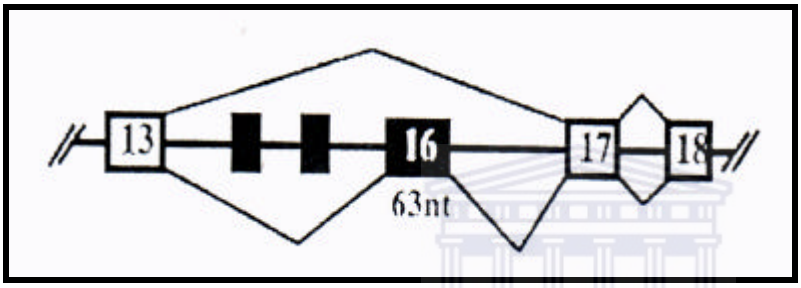


Figure 5 Exon skipping in 4.1 pre-mRNA

Inclusion of exon 16 results in both a high binding affinity to spectrin/actin and strengthening of red cell membranes in situ. Skipping of exon 16 leads to a low affinity binding to spectrin/actin and ineffective red cell membranes in situ (Figure extracted from Conboy 1999). Black solid boxes indicate alternative exons and non-shaded boxes represent constitutive exons. Lines connecting exons that point outward represent RNA splicing of introns.

2.1.1.3 Genes involved in apoptosis

Alternative splicing has also been observed in murine genes that are involved in regulating programmed cell death. Of particular interest is the *Bcl-x* gene, which plays important roles in the central nervous system and in the development of hematopoietic

processes (Jiang & Wu 1999). The five known, variant isoforms that are generated from alternative splicing of *Bcl-x* gene, have either inhibitory or activating functions in cell death. Four isoforms of the *Bcl-x* gene, *Bcl-xL* (Boise *et al.*1993), *Bcl-x* (González-García *et al.*1994; Shiraiwa *et al.*1996), *Bcl-x TM* (Fang *et al.*1994), *Bcl-x* (Yang *et al.*1997), inhibit apoptosis in a variety of tissues whereas the *Bcl-xS* (Boise *et al.*1993) isoform promotes apoptosis. Although examples of alternatively spliced genes (Subramaniam, *et al.*1994; Ffrench-Constant 1995; Celotto & Graveley 2001) involved in development, are abundant in the literature, the functions of their protein isoforms are poorly characterized but have proposed roles in developmental regulation.

2.1.2 The role of Alternative Splicing in Generating Receptor Diversity

Extensive alternative splicing has been observed in many receptors in the nervous system. These receptors regulate the transport and release of specific ions or neurotransmitters that are essential to the process of neurotransmission and neural development. Several receptors that bind neurotransmitters, such as dopamine and GABA (gamma aminobutyric acid), are discussed to illustrate the importance of alternative splicing in regulating receptor functional properties.

2.1.2.1 Dopamine D2 receptor isoforms differ in their inhibitory effects on adenylyl cyclase

The dopamine receptors belong to a family of seven-transmembrane domain G-protein coupled receptors that bind dopamine. An important feature of G-protein coupled receptors is that upon binding of their ligands, G-protein coupled receptors must couple to the heterotrimeric G-proteins in order for transduction of a signal to be successful. When dopamines are bound to their receptors on postsynaptic neurons, either inhibition or excitation of the action potential can occur. At least five known dopamine receptor subtypes exist and they differ in their pharmacological properties, localization, and mechanisms of action. Dopamine D2 receptors are the predominant subtypes found in

the brain and are expressed at high levels compared to other subtypes. Alternative splicing of the D2 receptor pre-mRNA results in the production of long and short protein isoforms, D2L and D2S, which differ in their coupling to G-proteins as well as their inhibitory effects on adenylyl cyclase. In order to inhibit adenylyl cyclase more efficiently, the D2L isoform, requires the -subunit of the inhibitory G-protein (G i2) whereas the D2S does not have this requirement for inhibition. The D2L isoform contains a unique insert of 29 amino acids that is located on the third intracellular loop of the D2L protein. The 29 amino-acid insert, which is lacking in D2S isoforms, has been proposed to mediate the interaction of D2L with G i2, in order to achieve a greater inhibition with adenylyl cyclase than G-protein coupling with D2S isoforms (Guiramand *et al.*1995).

2.1.2.2 Binding affinities for benziodiazepines differ in $GABA_A$ receptor isoforms

Three major classes of receptors that bind GABA, an inhibitory neurotransmitter, are known as the $GABA_A$, $GABA_B$, and $GABA_C$ receptors. $GABA_A$ receptors are essentially chloride channels that mediate synaptic inhibitions. Hetero-oligomeric $GABA_A$ receptor subtypes are generated from combinations of 16 different subunits (1-6, 1-4, 1-4, , and subunits). Functional GABA $_A$ receptors require the assembly of , , and one other subunit type (Chebib & Johnston 1999). The GABA $_A$ 2 subunit undergoes alternative splicing to generate two variant isoforms, 2L (long isoform) and 2S (short isoform). These isoforms differ through the inclusion of an 8 amino acid exon present in 2L, which is absent in the shorter isoform (Wafford *et al.*1993). The affinity for benzodiazepine agonists also differs between the isoforms with the shorter isoform having an increased affinity for benziodiazepines (Quinlan *et al.*2000).

2.1.2.3 Glutamate receptor 5 isoforms have differing effects on neuronal differentiation

Alternative splicing of the mGluR5 gene (metabotropic (metabolically coupled) glutamate receptor 5) generates two structurally different protein isoforms, named mGluR5a and mGluR5b. The mGluR5b receptor differs from the mGluR5a receptor by the additional insertion of a 32 amino acid cassette in the C-terminal tail. It has been shown that mGluR5a receptor inhibits neurite branching and mGluR5b promotes the neuronal growth (Mion *et al.*2001).

2.1.3 The Role of Alternative Splicing in Disease

Human genetic disorders can arise as a result of aberrant alternative splicing. Aberrant alternative splicing occurs mostly through point mutations in splice sites and splicing regulatory elements. These point mutations can result in either the down- or up-regulation of the gene-of-interest, which could alter the ratio of spliced isoforms commonly observed in cancer-associated genes. Current efforts into the treatment of human genetic diseases have utilized antisense oligonucleotide technology to modify the splicing patterns of disease-causing genes (van Deutekom *et al.*2001; Sierakowska *et al.*1996; Schmajuk *et al.*1999; Friedman *et al.*1999).

This section aims to demonstrate examples of human genetic disorders arising from mutations in splice sites and splicing regulatory elements. Modifying splice-variant ratios using antisense oligonucleotide technology will also be discussed.

2.1.3.1 Mutations in Splicing Regulatory Elements

A well-known example in which point mutations affecting splicing regulatory elements has been observed with the *tau* gene. Missense ($^{N}279^{K}$) and silent ($^{L}284^{L}$) mutations are

known to occur in exon 10 of the *tau* gene. The effects of these mutations are one of the causes of frontotemporal dementia with Parkinsonism linked to chromosome 17 (FTDP-17). In FTDP-17, tau proteins aggregates in the brain to form abnormal protein deposits that lead to neuronal cell death. Tau normally promotes microtubule assembly and stability during neuronal development. The alternative tau exon 10 encodes for one of the four microtubule-binding domains also present in tau exons 9, 11, and 12. Inclusion and skipping of tau exon 10, leads to the production of two types of protein isoforms: 4R and 3R isoforms. Tau transcripts containing exon 10 results in 4R isoforms consisting of 4 microtubule-binding domains whereas exclusion of exon 10 results in production of 3R isoforms containing only 3 microtubule-binding domains. Overproduction of 4R isoforms relative to 3R isoforms is a characteristic feature of FTDP-17 (D'Souza *et al.*1999).

It has been proposed that the missense mutation ($^{N}279^{K}$) increases exon 10 inclusion by either enhancing an existing or creating an exon splicing enhancer element whereas the silent mutation ($^{L}284^{L}$) increases exon 10 inclusion by destroying an exon splicing silencer element (D'Souza *et al.*1999; Kalbfuss *et al.*2001).

Other human diseases such as, Menkes disease, porphyria,, encephalomyelopathy, hereditary tyrosinemia I, and metachromatic leukodystrophy that arise from mutations in ESEs have been extensively reviewed elsewhere (Cooper & Mattox 1997).

2.1.3.2 Mutations in Splice-sites

Many splice-site mutations have been implicated in a variety of human diseases (Penzel *et al.*2001; Hutchinson *et al.*2001; Sakamoto *et al.*2001; Noack *et al.*2001). An example of how a splice-site mutation could cause disease has observed with lysosomal neuraminidase gene (*NEU1*). *NEU1* is an enzyme that catabolizes sialylated glycoconjugates. A splice-donor mutation involving a transversion (G->C) in intron 5 results in the inclusion of exon 5. Skipping of exon 5 has been predicted to result in a

frameshift leading to a putative truncated protein lacking a C-terminal region. The C-terminal region consists of asparagine residues and is assumed to be essential during catalysis. The disruption of neuraminidase activity has been proposed to result from the lack of C-terminal regions in predicted, truncated proteins (Penzel *et al.*2001).

Other intronic mutations that activate cryptic splice sites or create aberrant splice sites, which are responsible for human genetic disorders such as Marfan syndrome (Hutchinson *et al.*2001), Barth syndrome (Sakamoto *et al.*2001), and chronic granulomatous disease (Noack *et al.*2001). Generation of aberrant splice sites, are usually preferred over consensus splice sites thereby resulting in the production of aberrant transcripts in affected individuals.

Altered Isoform Ratio

Table 1: Apoptosis-associated genes with altered isoform ratios during cancer

Gene	Isoforms Generated	Isoform type altered in cancer	Up- or down-regulation of the altered isoform type	References
1. Bcl-2	Bcl-2 Bcl-2	Bcl-2	Up-regulation	Adams & Cory 1998
2. Bcl-x	Bcl-xS Bcl-xL	Bcl-xS	Down-regulation	Han *et al.*1998
3. Caspase 2	Ich-1L Ich-1S	Ich-1L	Up-regulation	Fujimura *et al.*1999
4. CD44	Can generate up to 30 different splice variants	CD44v6 (inclusion of variable exon 6)	Up-regulation	Harn *et al.*1996; Herrera-Gayol & Jothy 1999
5. BRCA2	Not provided	12-BRCA2(exclusion of exon 12)	Up-regulation	Bieche & Lidereau 1999

Table 1 summarizes six cancer-associated genes that have altered protein isoform ratio.

Another factor that may contribute towards disease is the alteration of protein isoform ratio. In cancer-associated genes (see Table 1) involved in apoptosis, in which over-expression or under-expression of a particular isoform may contribute towards cancer and observations have shown that cancerous cells with altered protein ratios causes the de-sensitization of cells to chemotherapy treatments. Although the mutations in these cancer-associated genes are not clear, the alteration of splice-variant ratios in these genes have been observed to be characteristic of most cancerous cells. Alterations of splice-variant ratios are currently feasible using antisense oligonucleotide technology.

Antisense Oligonucleotide Technology

Antisense oligonucleotide technology (AOT) has been used extensively in human diseases caused by aberrant alternative splicing (van Deutekom *et al.*2001; Sierakowska *et al.*1996; Schmajuk *et al.*1999; Friedman *et al.*1999). The technology can be used to modify the splicing patterns of diseased genes to generate either a complete or partially functional protein. AOT is commonly utilized in gene therapy to inhibit gene expression on an RNA level.

Antisense oligonucleotides are short nucleic acid fragments that are complementary to their target RNA sequences. The hybridization of antisense oligonucleotides to specific RNA sequences results in a DNA-RNA duplex, which inhibits mRNA translation. The DNA-RNA duplex has been proposed to inhibit translation through the following three mechanisms: 1.) Blocking RNA-interaction with ribosomes, 2.) Arresting translation through hybridization to specific RNA targets, 3.) The endogenous enzyme, ribonuclease H (RNAse H), causes the hydrolysis of the RNA strand in target DNA/RNA duplexes. The efficacies and potential usage of antisense oligonucleotides as potential therapeutic agents in cancer, viral, and other human genetic diseases have been reviewed elsewhere (Alama *et al.*1997; Stein 1999). However, the molecular pathways with regards to how the inhibition of targeted gene expression is mediated, are still unknown.

Duchenne Muscular Dystrophy (DMD)

An example of a mutation causing aberrant splicing that can be restored by antisense oligonucleotides have been observed with Duchenne muscular dystrophy. A frequent, mutation commonly observed in Duchenne muscular dystrophy (*DMD*) patients, is the deletion of exon 45, which results in a frame-shift mutation leading to dystrophin deficiency. Absence of dystrophin results in an eventual loss of muscle fibres leading to premature death in adolescence. Deletion of exon 45 introduces a stop codon in exon 46, which leads to a non-functional dystrophin protein (see Fig. 6). In human *DMD*-affected muscle cells transfected with antisense oligoribonucleotides targeted against exonic splicing enhancers present in exon 46, induction of exon 46 skipping resulted in the restoration of the reading frame and generation of a truncated, functional dystrophin protein (van Deutekom *et al.*2001).

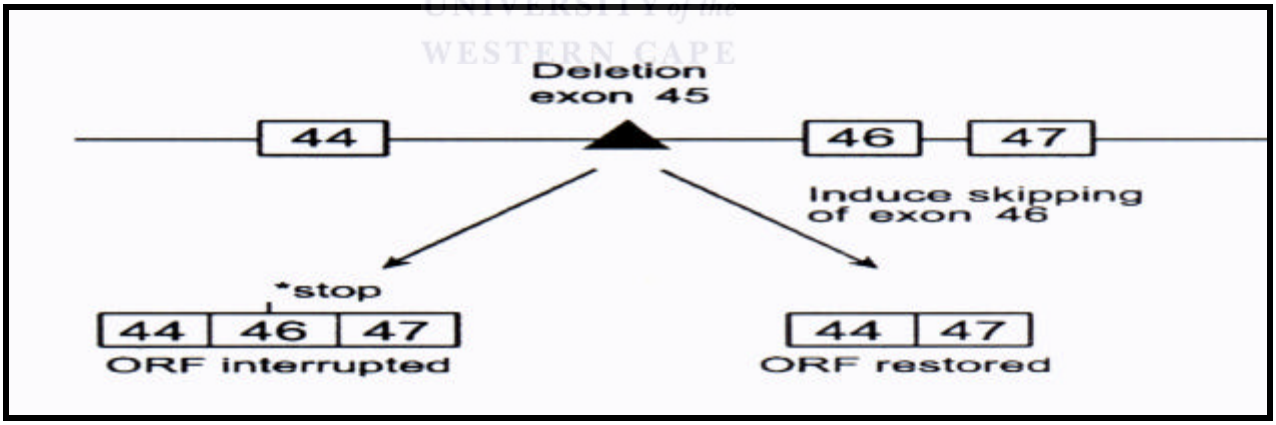


Figure 6: Induction of exon skipping restores reading frame of *DMD* patients

The diagram illustrates exon 46 skipping strategy. Skipping of exon 46 restores the reading frame of *DMD* patients and leads to a milder phenotype rather than premature death among *DMD* children (Figure extracted from van Deutekom *et al.*2001).

Antisense oligonucleotides have also been used to block aberrant acceptor and donor splice-sites generated through intronic mutations in the human -globin (Sierakowska *et al.*1996; Schmajuk *et al.*1999) and cystic fibrosis transmembrane conductance receptor (Friedman *et al.*1999) genes. Targeting antisense oligonucleotides to aberrant splice sites, led to the production of functional proteins in both of the genes mentioned previously.

Chapter 3: High Incidence of Exon Skipping

Recent studies (Modrek *et al.*2001; Clark & Thanaraj 2002; Croft *et al.*2000; Hide *et al.*2001) have observed a frequent occurrence (~10-25%) of human exon-skipping events as compared to other types of alternative splicing patterns (see Table 2).

Table 2: A summary of recent human exon-skipping studies

References	Data set size	Exon-skipping frequency (%)	Method of Detection
Croft *et al.*2000	2698 unique, human gene set	10	* ESTs were searched against human intron sequences using BLAST * ESTs were assembled into contigs using Phrap * EST contigs were searched against exons from the gene of the intron hit using BLAST * Aligned ESTs to genomic sequences
Hide *et al.*2001	347 multi-exon genes on human chromosome 22	15	* Exon-junctions were created from annotated consecutive exon boundaries and searched against human dbEST using BLAST * Aligned ESTs to human genomic sequences
Modrek *et al.*2001	A total of 6201 alternative splice relationships	50*	* Mapped EST and mRNA sequences to annotated genomic sequences using the GeneMine software system
Clark & Thanaraj	2793 human genes	25	* Mapped EST or mRNA sequences to annotated genomic

2002	sequences using BLAST

The table compares the exon-skipping detection methods and frequencies for each study. The exon-skipping frequency is calculated by dividing the total number of exon-skipped events by the total number of genes sampled. The asterisk indicates that the exon-skipping frequency was over-estimated since the total number of genes sampled has not been provided by Modrek *et al.*2001. Thus, the exon-skipping frequency was calculated by dividing the total number of exon-skipped events (704) by the total number of alternatively spliced events (6201). The exon skipping frequencies appears to increase with every new study.

Exon-skipping events have also been widely observed in both development and human genetic disorders. Exon-skipping events are known to occur in the *Drosophila* sex determination pathway and erythrocyte morphogenesis, which was previously described in section 2.1.1.

The majority of exon-skipping events that arise from point mutations are known to occur in many human genetic disorders. However, the negative effects of exon-skipping events causing the disease are not well documented due to the lack of functional protein isoform characterizations. Some well-characterized examples of exon skipping affecting the function of protein isoforms are described in table 3.

Table 3: Human genetic diseases associated with exon skipping

Disease description	Gene	Description of mutation	Exons skipped	Effect of exon-skipping on disease	References
Congenital Fiber-Type Disproportion Myopathy (Novel syndrome of insulin resistance)	Insulin receptor	Point mutation in the last base pair of exon 17, which affects position –1 of the donor splice site of exon 17	Exon 17	Exon 17 encodes for the tyrosine kinase domain. The -subunit of tyrosine kinase is responsible for the activation of signaling pathways stimulated by insulin	Vorwerk *et al.*1999
Human nonpolyposis colorectal cancer (HNPCC)	MLH1 (mismatch repair gene)	Transversion (T-to-A) mutation at position –11 of the MLH1 intron 1	Exon 2	Unknown functional consequences	Clarke *et al.*2000

		splice acceptor site		Skipping of exon 2 results in frame shift mutation leading to truncation of the protein	
Sialidosis (lysosomal storage disease)	Lysosomal neuraminidase	Transversion in splice donor site	Exon 5	Loss of enzyme activity	Penzel *et al.*2001
Dopa-responsive dystonia	GTP cyclohydrolase I	Point-mutation in donor splice site of intron 5	Exon 5	Reduction of enzyme activity	Skrygan *et al.*2001

The prevalence of exon skipping detected in human genes and in human genetic disorders is intriguing. It is still uncertain how much functional and non-functional exon skipping is being utilized in generating functional diversity and in causing diseases. The extent of functional and non-functional exon skipping remains to be explored in order to understand the frequent occurrence of exon skipping.

Chapter 4: Utilization of Expressed Sequence Tags to Detect Alternatively Spliced Transcripts

4.1 The feasibility of using ESTs to detect alternatively spliced transcripts

Expressed sequence tags (ESTs) are partial cDNA sequences that are sequenced from both or either the 5' or 3' ends of cDNA clones using single run sequencing (Adams *et al.*1991). Expressed sequence tags have been widely used in gene discoveries (Vasmatzis *et al.*1998; Liew *et al.*1994; Adams *et al.*1992), differential gene expression studies (Adams *et al.*1992; Schmitt *et al.*1999; Adams *et al.*1995), construction of gene markers (Hukriede *et al.*2001; Scheetz *et al.*2001; Avner *et al.*2001), and in alternative splicing studies (Mironov *et al.*1999; Brett *et al.*2000; Hanke *et al.*1999; Modrek *et al.*2001; Clark & Thanaraj 2002; Croft *et al.*2000; Hide *et al.*2001; Kan *et al.*2001; Wolfsberg & Landsman 1997; Thanaraj 1999; Kan *et al.*2000; Brett *et al.*2002).

4.1.1 Quality of EST data

Some characteristics of EST data described by Jongeneel 2000 are summarized in point form below:

1. **Biased**- ESTs are mostly derived from the 3' ends of mRNA whle ESTs covering the central regions are sparse. Thus, abundant information can be obtained on 3' untranslated regions.

2. **Average quality is low**- Current submission rules from NCBI require that the submitted ESTs have an error rate of less than 1 per cent. An error rate of less than 1 per cent corresponds to a phred score of 20 or more (Phred is a base-calling program developed by Phil Green (Ewing *et al.*1998). However, most of the EST sequences currently in the public databases have not met these quality

criteria. Thus, sequence quality is low due to frame-shift errors that may result from insertions, deletions, and even artefactual stop codons are common.

3. **Not all cDNA libraries are normalized-** In non-normalized libraries highly expressed genes are represented by many ESTs and vice versa. In normalized libraries, the EST representation of highly expressed genes is reduced and the EST representation of genes with low expression is increased. By comparing normalized and non-normalized libraries, we can identify those genes with low and high expression. Also, comparison of normalized and non-normalized libraries for a specific cell type can be used to quantify the level of expression in genes with high and low expression. The exact number of normalized and non-normalized libraries is currently unknown in the public database. Another complication for genes, which appear to have low expression, may be due to the possibility that the small number of ESTs were derived from cDNA libraries that may have their cell-types under-represented.

4. **Incomplete representation of specific developmental stages, cell types, or tissues-** Only genes from a specific tissue, cell type, or developmental stage that were used to prepare the cDNA libraries that are represented in the EST databases.

5. **Contamination by other sequences**-

    - Partially spliced mRNAs that are indistinguishable from genuine splice variants
    - Chimeras resulting from artefactual ligation of unrelated cDNAs
    - Genomic DNA
    - Bacterial DNA
    - cDNA from other species
    - Vector DNA

- Mitochondrial DNA
- Ribosomal DNA
- Erroneous annotation

It is clear that screening procedures are necessary to improve the quality of raw EST data as well as to increase the biological validity of ESTs. Some screening procedures include processing raw ESTs, masking, and setting software parameters to remove paralogues and spurious matches. Screening procedures evidently do vary across studies utilizing ESTs due to differences in the objectives of the study. This literature review will focus specifically on the use of ESTs to capture alternatively spliced transcripts. Screening procedures pertaining to candidate variant transcripts and feasibility of alternatively spliced detection methods will also be reviewed.

4.2 Detection of Alternatively Spliced Events in Expressed Sequence Tags using Computational Methods

Earlier attempts to discover alternatively spliced transcripts have utilized processed ESTs. The STACKdb (Sequence Tag Alignment and Consensus Knowledgebase) database is a gene index, which consists of processed ESTs extracted from updated versions of Genbank. Initially, ESTs are grouped into tissue-specific categories. Further processing steps such as: clustering, assembly, consensus generation, and clone linking are performed for each tissue-specific category. Details of these processing steps have been reviewed elsewhere (Miller *et al.*1999). During clustering, ESTs are clustered using d2_cluster (Burke *et al.*1999), which allows for the capture of alternatively spliced transcripts in addition to other contaminating sequences. The loose clustering approach allows for the detection of alternatively spliced transcripts. CRAW, a program incorporated as part of STACKPACK, provides the visualization reports for transcript variants. CRAW partitions sequences in a cluster into sub-clusters and illustrates the alignment inconsistencies in the CRAW report (Burke *et al.*1998). Each

sub-cluster would ideally represent an isoform and the alignment inconsistencies would indicate the region of variation between other isoforms within the cluster (representing a gene). Therefore, clusters that have sub-clusters with alignment inconsistencies are potential alternatively spliced candidates (see Fig. 7).

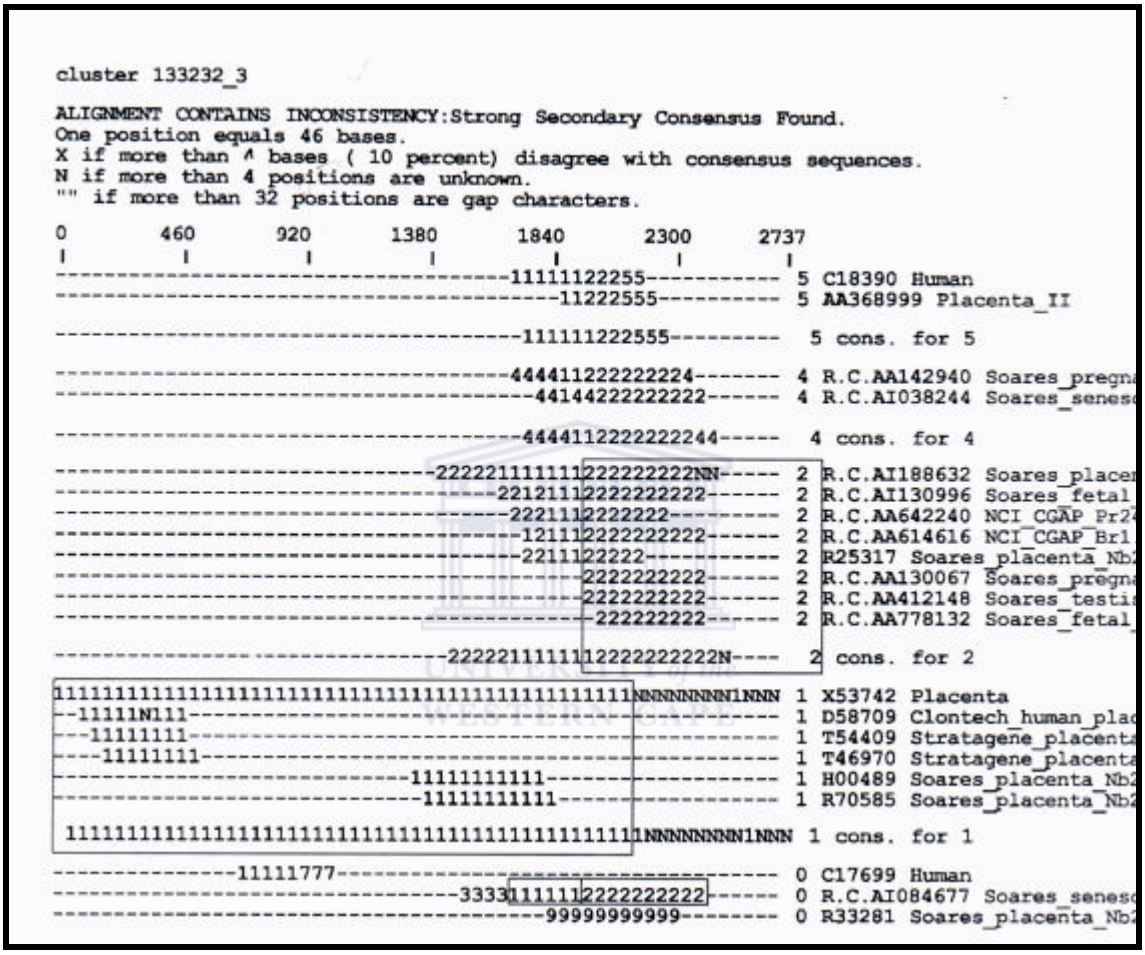


Figure 7: Craw report illustrating alternative transcript isoforms of the fibulin gene

Alternative splicing of the human fibulin gene are known to produce 4 or more isoforms (A-D) (Tran *et al.*1997). Fibulin transcripts are expressed ubiquitously and only two distinct isoforms are distinguished in the craw report. Transcript isoforms that are similar to the B isoform are represented by strings of ones enclosed in boxes and strings of twos enclosed in boxes represents those ESTs that are similar to the C isoform (Figure extracted from Christoffels *et al.*2001).

CRAW reports may be useful in obtaining putative splice variants, but they also have several limitations listed below:

- Splicing patterns cannot be determined accurately
- Location of the alternative splice-site in the EST cannot be described accurately
- Splice-site information cannot be determined
- Exon and intron structure information cannot be obtained since it is uncertain where the transcripts are situated on the genome
- Multiple transcript isoforms derived from a single gene cannot be clearly distinguished if they differ in the same regions since they will be in same sub-cluster.

A more common and efficient approach to detecting alternatively spliced transcripts using ESTs is by aligning ESTs to a genomic sequence. Currently, this approach has been used extensively by many alternative-splicing studies (see Table 4).

Table 4: A comparison between alternative-splicing studies that utilize EST to genome alignments

The table summarizes the parameters, alignment tools, screening procedures, and exon-skipping detection methods for each alternative-splicing study.

References	Method description	Parameters (Extraction of ESTs using BLASTN)	Alignment Tool (ESTs aligned to their genomic counterparts)	Exon-skipping detection method	Additional screening procedures
Wolfsberg & Landsman 1997	Aligned ESTs to genomic sequences	Genomic and cDNA sequences were searched against human dbEST with a P value of $< 10^{-87}$	Sequencher and sim2aln	Not described	Masking was performed by RepeatMasker
International Human Genome Sequencing Consortium 2001	Aligned cDNA/mRNAs to genomic sequences	Not provided	Not provided	Not provided	Not provided
Thanaraj 1999[a]	Aligned ESTs to genomic sequences	* 50bp from consecutive exons were concatenated to form an 100bp exon construct * The 100bp exon construct was searched against ESTs using FASTA (pid > 80%, matching length > 20 nucleotides)	FASTA	Not described	No mismatches were allowed within the –20 to +20 nucleotide region around the consecutive exon junctions
Mironov *et al.*1999	Aligned EST contigs to unannotated genomic sequences	Genomic sequences were searched against EST contigs (from TIGR human gene index)	BLASTN	Loops found in EST contig to genomic sequence alignments represent possible exon-skipping	Masking was performed by RepeatMasker

		with a P value < 10^{-50}		events	
Kan *et al.*2001	Refseq mRNAs were searched against the human contig database using WU-BLAST (>99% identity) * High-scoring mRNA sequences were aligned to human contigs using sim4 to extract genic regions * Human genic regions were aligned to human ESTs	Human genic regions were searched against human ESTs (no parameters available)	Sim4 (alignments) greater than 92% were used)	Not described	Masking was performed by RepeatMasker
Clark & Thanaraj 2002	Aligned human ESTs and mRNAs to human annotated genomic sequences	Genomic sequences were searched against transcript sequences with a P value < $1e^{-10}$	BLAST	* A gap in the transcript to genome alignments represented an exon skip * A skipped exon was defined as a constitutive exon that was absent in the alternative transcript but present in a constitutive transcript	*removal of duplicates (>99% sequence identity) * Hypervariable genes are removed * After manual removal of

				repeats, any match < 95% identity were removed * Transcripts that were associated with more than one gene were also removed
Croft *et al.*2000	Aligned EST contigs to the exons of a gene[b]	*Human EST sequences were BLAST searched against human intron sequences using gapped BLAST ($P<1e^{-10}$) * High-scoring EST hits were assembled into contigs using Phrap * EST contigs were searched against exons from the gene of the intron hit with a P value < $1e^{-10}$	EST contigs that have matches to non-consecutive exons (exon 1-exon3) and were also present in the genomic sequence enabled detection of exon-skipped transcripts	Genes were removed if they have duplicates, have more than one transcript matching other human genes, and genes that fell under MHC and other hypervariable genes

Hide *et al.*2001	Aligned human ESTs to annotated genomic sequences	Exon junctions were created from annotated, consecutive exon boundaries and searched against human dbEST with a P value < $1x10^{-40}$	Sim4	A skipping event was reported when an EST does not contain the skipped exon, but does contain an uninterrupted tag comprised of 50bp from each of the flanking exons	* Paralogues were removed * False positives were removed manually
Modrek *et al.*2001	Aligned unigene clusters (ESTs and mRNAs) to unannotated genomic sequences	Each unigene cluster was searched against a database of human genomic sequences with a P value < $1x10^{-50}$ and a nucleotide mismatch penalty of 11	Dynamic programming	* A candidate splice was detected as a gap in the EST sequences when they were aligned to their respective genomic sequences * Extensive matches aupstream and downstream of the skipped exons were required	Assessed the accuracy of their predicted exon/intron structures by comparing to an independent data produced by NCBI Acembly, a human curated annotation effort
Kan *et al.*2000[c]	Aligned ESTs and mRNA (REFSEQ) to genomic sequences	*cDNA/mRNA sequences were aligned to genomic sequences and genic regions were extracted to produce genomic templates * Genomic templates were searched against human ESTs using WU-BLASTN2 (the match	Sim4	N/A	Masking was performed using RepeatMasker

	must be > 95% identity and the match must cover > 90% of the EST length)	

[a] The objective of this study was to construct a clean data set of EST-confirmed splice sites from *Homo Sapiens*

[b] It is uncertain whether exons and introns were separated for each annotated genomic sequence

[c] The aims of the investigation were to infer UTR sequences from genomically aligned ESTs

UNIVERSITY of the
WESTERN CAPE

The alignment tools, parameters, and additional screening procedures vary between studies. However, the exon-skipping detection methods performed by Croft *et al.*2000, Hide *et al.*2001, Modrek *et al.*2001, and Clark & Thanaraj 2001 are essentially similar to the present study. In all of the previously mentioned exon-skipping studies, an exon-skipping event is identified by a gap in the genomically aligned EST. The gap represents the missing exon present on the genomically aligned EST. The detection method requires access to genomic sequences with well-annotated gene structures in order to identify missing exons in ESTs, when aligned to their corresponding genomic sequences. Furthermore, extensive matches must occur downstream and upstream of the skipped exon between the EST-genome alignments in order for the skipping events to be valid.

Mapping ESTs to genomic sequence has several advantages (Modrek & Lee 2002), which are listed below:

- Removing RT-PCR artifacts in ESTs
- Removing chimeric ESTs
- Removing genomic contamination in ESTs
- Removing some paralogous genes and ESTs (a high level of identity is required – 95-95%)
- Checking for consensus or non-consensus splice sites
- Accurate description of the position of the alternatively spliced event with respect to the genome
- Accurate description of the type of alternative splicing pattern

A major disadvantage of this method is the high rate of false negatives. False negatives are defined as the failure to detect a real splice form. Some factors that would produce a high false negative rate include the requirement for the utilization of stringent parameters (which are used to eliminate paralogues and spurious matches) and the use of incomplete, short genomic contigs (Modrek & Lee 2002). Stringent parameters may

reduce the detection of alternatively spliced events if the spliced events have not met the set criteria. The use of short genomic contigs can also result in a false negative if the contig does not encode for the exons involved in the alternatively spliced event.

Another similar and effective method for detecting alternatively spliced transcripts is the alignment of ESTs to mRNA sequences (Brett *et al.*2000; Hanke *et al.*1999; Brett *et al.*2002). The method has been experimentally confirmed using RT-PCR and DNA sequencing and has a high success rate whereby over 80% of a small number of random clones selected (~10-20) (Brett *et al.*2000; Hanke *et al.*1999) has confirmed the predicted alternatively spliced events. Although partial genomic sequences have been utilized to screen out any DNA contamination in the confirmed alternatively spliced transcript isoforms, the use of genomic sequences was not utilized as part of the main detection pipeline. Some disadvantages of this method may include the following: (i) the total extent of false positive rate is still unknown, and (ii) it may be difficult to describe the pattern of alternative splicing in the absence of genomic sequences. The extent for the different types of known alternative splicing patterns was unknown for studies utilizing this method (Brett *et al.*2000; Hanke *et al.*1999; Brett *et al.*2002).

Inherent alignment tool errors also contribute towards the high frequency of false negatives. Alignment tools such as sim4 (Florea *et al.*1998), est2genome (Mott 1997), and spidey (Wheelan *et al.*2001) have been developed to align transcript sequences (ESTs, mRNAs, cDNAs) to genomic sequences. Comparisons of running times and the extent of accuracy between these programs have revealed that both sim4 and spidey performs better in species-specific alignments than est2genome (Florea *et al.*1998; Wheelan *et al.*2001). Alignments of human annotated mRNAs to human RefSeqs will cause est2genome to have a higher false negative rate (7-fold greater than spidey and 3-fold greater than sim4) than sim4 and spidey. Est2genome also has a longer running time than spidey and sim4. Alignment of a transcript (~5000bp) to a genomic contig (~1MB) will take est2genome one hour and twenty-one minutes. Under the same conditions, spidey and sim4 will run ~300-fold and ~2400-fold faster respectively than Est2genome (Wheelan *et al.*2001).

Chapter 5: Relevance of Mouse and Human Gene Comparisons

5.1 Current State of Mouse Genomic Sequencing

Two major projects, genome-wide and targeted sequencing programs, have been implemented to sequence the mouse genome. The genome-wide program refers to the sequencing of the whole mouse genome and is being undertaken by the Mouse Sequencing Consortium, which consists of a collaboration of three sequencing centers (Washington University Genome Sequencing Center, Whitehead/MIT Center for Genome Research, and the Sanger Centre) and an international database (Ensembl). The targeted sequencing programs (Harvard Partners Genome Center, CSHL Lita Annenberg Hazen Genome Center for Genome Technology, University of Oklahoma's Advanced Center for Genome Technology, and MRC UK Mouse Sequencing Programme) are mainly focused on sequencing chromosomal regions of high biomedical interest. These groups also accept requests from research groups for targeted sequencing of their BACs and small contigs (Mouse Genome Monthly November 2001).

Mouse genomic data are available from ftp sites at both *NCBI* and the Mouse Ensembl Genome Server. The utilization of mouse sequence data from Ensembl has several advantages. Automatic annotations of sequence data are provided and sequence data, which includes mouse cDNA, contig, and peptide sequences, are also easily retrievable from the Mouse Ensembl web site (ftp://ftp.ensembl.org/pub/current_mouse/data/). As the mouse genomic sequencing progresses, the quality of the data also improves. The status of the mouse genomic data that was utilized in this study is illustrated in Fig. 8.

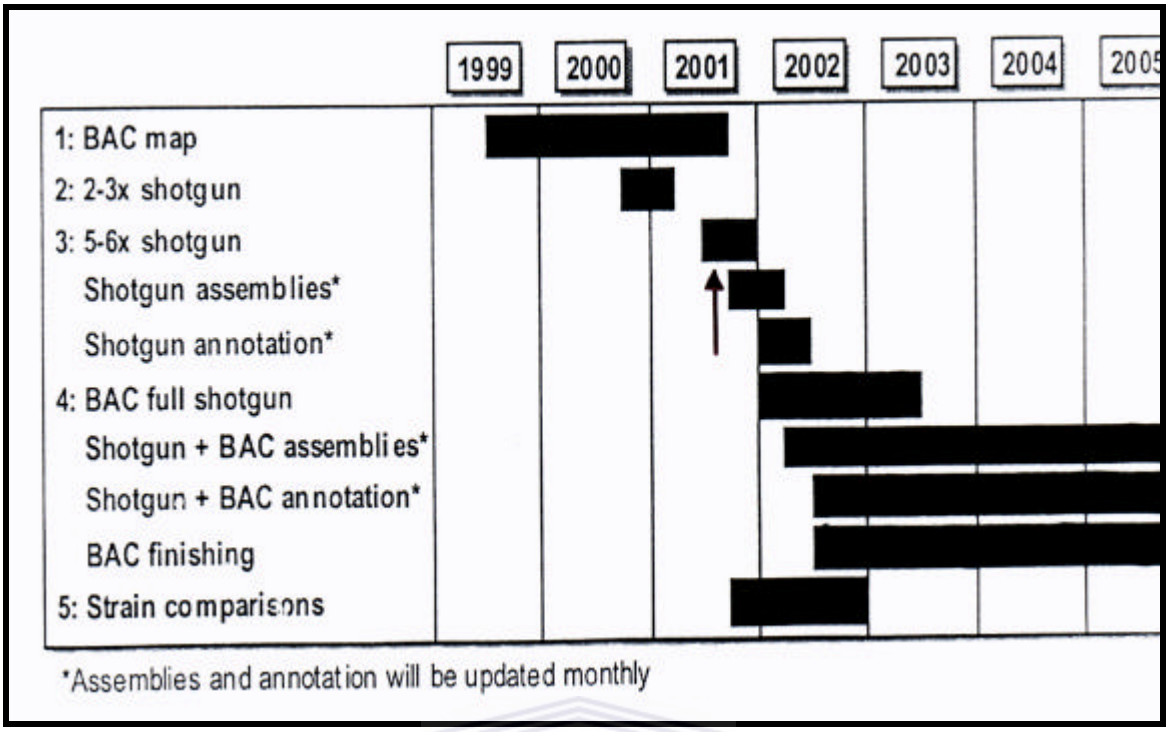


Figure 8: Mouse genome sequencing project timeline

Mouse genomic data (v.0.1.0) consists of unassembled mouse contig data. The total number of contigs in this version is 31568. No positional information or annotation is available for these mouse contigs. The status of mouse ensembl v.0.1.0 corresponds to stage 3 of the mouse sequencing project timeline (see black arrow). The goals at this stage are to achieve greater shotgun coverage (5-6X) and produce longer mouse sequence contigs. Many data versions obtained during this stage are highly variable from each other. As new trace data become available, clone contigs are reassembled and clone contig identifiers may change (Extracted from Mouse Genome Monthly November 2001).

5.2 Comparative Mouse and Human Gene Studies

Comparative studies in general have mostly been aimed at the annotation of the human genome and several advantages are listed below:

- Comparative maps can be constructed. Comparative maps are useful for gene predictions. If a marker is linked to a diseased gene in one species, comparative mapping can predict the location of a candidate diseased gene in another species. Thus, comparative mapping can be utilized to obtain information on candidate diseased genes, patterns of inheritance, and a picture of the immediate environment of individual genes (Clark 1999).
- Comparative gene studies can assist in the functional annotation of the human genome project. By comparing genomes of different species, the organization, evolution, and function of the human genome can be comprehended (The First International Workshop on Comparative Genome Organization 1996).

Comparative mouse and human genome comparisons have often been used to predict coding exons (Kirschner *et al.*2001; Dehal *et al.*2001; Martindale *et al.*2000; Amid *et al.*2001; Lane *et al.*2001) and novel regulatory elements (Flint *et al.*2001; Hardison *et al.*1997; Oeltjen *et al.*1997; Wasserman *et al.*2000; Ansari-Lari *et al.*1998) in the human genome. The frequent sequence conservation that is observed between mouse and human genes has led to the development of many comparative analysis tools such as Alfresco (Jareborg & Durbin 2000), MUMmer (Delcher *et al.*1999), GLASS, ROSETTA (Batzoglou *et al.*2000), PipMaker (Schwartz *et al.*2000), and WABA (Ballie & Rose 2000).

Mouse and human gene comparisons can also be used to examine alternative splicing patterns between these organisms. Conserved alternative splicing patterns have been observed in the *Drosophila* channel gene, *para,* between *D.* melanogaster and *D.* virilis (Thackeray & Ganetzky 1994). Comparing conserved alternative splicing patterns between species can be utilized as a feasible strategy to identify novel alternatively spliced variants in other species. Interestingly, several studies (Laverdière *et al.*2000; Reiter *et al.*2001; Osborne & Tonissen 2001) have observed differences in alternative splicing patterns in mouse and human orthologous transcripts (see Table 5). These

studies suggest that the gene-regulation in mouse and human genes may be controlled by different mechanisms. The extent to which alternative splicing patterns are different between mouse and human orthologous genes are still unknown.

Table 5: Conserved mouse and human orthologous genes display different alternative splicing patterns

These studies appear to suggest that the difference in mouse and human alternative splicing patterns are due to a difference in gene regulation rather than inherent sequence differences, even though a high level of sequence similarity exists between mouse and human genes.

References	Species compared	General description of alternative splicing	Differential splicing description
Laverdière *et al.*2000	Mouse, rat, and human	Alternative selection of 3' splice sites in the p53 tumour suppressor gene	Proximal, 3' alternative splice sites were favoured by mice and rats
Reiter *et al.*2001	Human, mouse, rat, and chicken	Alternative polyadenylation sites and mutually exclusive exons in epidermal growth factor receptor gene	Both mouse and human have transcripts that differ in their inclusion or exclusion of mutually exclusive exons and alternative polyadenylation signals
Osborne & Tonissen 2001	Human and mouse	Alternative selection of species-specific exons and 5' splice sites of the thioredoxin reductase 1 genes	Mouse and human transcript isoforms vary in their selection of alternative exons, which may be species-specific

5.3 Putative Orthologues versus True Orthologues

Comparative gene studies require that the genes, which are being compared between different species, originate from the same ancestral gene. In other words, obtaining orthologous genes is a requirement for comparative gene studies. Orthologues can be referred to as the same gene in different species (Koonin 2001) or more formally defined as homologous genes that have diverged from each other after speciation events

(Eisen 1998). The term, orthology, cannot be used to imply conservation of function between species and has been used incorrectly within the last twelve years (Koonin 2001; Ouzounis 1999). However, the consequence of orthology usually results in the conservation of function whereas the trend in paralogues tends towards the evolution of new functions (Tatusov *et al.*1997).

The most common methods of determining orthologous relationships are the use of both phylogenetic and reciprocal BLAST analyses. The phylogenetic methods consist of three common methods that include parsimony, distance, and maximum likelihood. Details of these methods are reviewed elsewhere (Eisen 1998).

Orthology can also be inferred through levels of sequence similarity that eliminates paralogues or other similar sequences that have arisen as a result of convergence. The reciprocal BLAST analysis method is based on the highest reciprocal sequence similarity match. If a gene from species A is searched against the genome of species B using BLAST, the highest sequence match represents the putative orthologue and vice versa (Tatusov *et al.*1997). Thresholds for inferring orthology using the reciprocal BLAST method have been established by many (Eisen 1998; Remm & Sonnhammer 2000; Xie & Ding 2000; Wheelan *et al.*1999; Mushegian *et al.*1998) to have an estimated cut-off E-value less than $1e^{-05}$ when using protein query sequences. The matching lengths of an alignment pair must be greater than 50% of their total individual sequence length.

The advantages and disadvantages of the reciprocal BLAST and phylogenetic methods are compared in table 6.

Table 6: Comparison of the advantages and disadvantages of phylogenetic and reciprocal BLAST methods

	Advantages	Disadvantages
Phylogenetic Methods	* Gene functions correlate well with gene phylogeny * Most functional predictions are accurate * In phylogenetic analyses, a process called masking can be employed to exclude regions of genes that are highly variable between species. These variable gene regions are biologically insignificant (Eisen 1998). * Necessary for the detection of true orthologues (Remm & Sonnhammer 2000; Xie & Ding 2000; Wheelan *et al.*1999; Mushegian *et al.*1998). * Allow for rate variation and accurate reconstruction of gene history * Inference of gene duplication events	* Time-consuming * Labour-intensive * False positives (by distance based phylogenetic methods (Remm & Sonnhammer 2000) * Different phylogenetic methods produce highly different results. Several different methods will have to be used to assign orthology (Eisen 1998)
Reciprocal BLAST analysis	* Automated easily * Fast method * Accurate for species with short phylogenetic distances	* 73.6% accuracy in detection of true orthologues (26.44% error rate) –(Remm & Sonnhammer 2000) * BLAST assumes a constant molecular clock which is sometimes incorrect * BLAST does not accurately separate orthologues from paralogues * If an alignment contains a large insertion, BLAST reports 2 segments and the overall significance is based on the strongest match only * BLAST uses gap penalties and phylogenetic methods do not.

	* Highest hit method is misleading, different levels of similarity do not correlate with functions * Similarity methods do not cope well when evolutionary rates vary between different species * Similarity based methods perform poorly when variation and gene duplication are combined

Although the disadvantages of using the reciprocal BLAST method exceeds those listed for phylogenetic methods, the reciprocal BLAST method is still an acceptable method of producing candidate orthologues. Reciprocal BLAST methods are mainly chosen for their high-speed performances when compared to phylogenetic methods. However, to further improve the accuracy of orthologue prediction, phylogenetic methods will have to be employed.

REFERENCES

1. Adams, J.M. & Cory, S (1998) The Bcl-2 protein family: arbiters of cell survival. *Science.* **281:** 1322-1326.
2. Adams, M.D. et al (1996) Biochemistry and regulation of pre-mRNA splicing. *Curr. Opin. Cell Biol.* **8:** 331-339.
3. Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., Kerlavage, A.R., McCombie, W.R., and Venter, J.C. (1991) Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project. *Nature.* **252:** 1651-1656.
4. Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D., and White, O., *et al.*(1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature.* **377** (6547 Suppl): 3-174
5. Adams,M.D., Dubnick, M., Kerlavage, A.R., Moreno, R., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C., Venter, J.C. (1992) Sequence identification of 2,375 human brain genes. *Nature.* **355**:632-4.
6. Alama, A., Barbieri, F., Cagnoli, M., and Schettini, G. (1997) Antisense Oligonucleotides As Therapeutic Agents. *Pharmacological Research.* **36**(3)**:** 171-178.
7. Amid, C., Bahr, A., Mujica, A., Sampson, N., Bikar, S.E., Winterpacht, A., Zabel, B., Hankeln, T., and Schmidt, E.R. (2001) Comparative genomic sequencing reveals a strikingly similar architecture of a conserved syntenic

region on human chromosome 11p15.3 (including gene ST50 and mouse chromosome 7. *Cytogenet. Cell Genet.* **93:** 284-290.

8. Ansari-Lari, M.A., Oeltjen, J.C., Schwartz, S., Zhang, Z., Muzny, D.M., Lu, J., Gorrell, J.H., Chinault, A.C., Belmont, J.W., Miller, W., and Gibbs, R.A. (1998) Comparative Sequence Analysisof a Gene-Rich Cluster at Human Chromosome 12p13 and its Syntenic Region in Mouse Chromosome 6. *Genome Res.* **8:** 29-40.
9. Avner, P., Bruls, T., Poras, I., Eley, L., Gas, S., Ruiz, P., Wiles, M.V., Sousa-Nunes, R., Kettleborough, R., Rana, A., Morissette, J., Bentley, L., Goldsworthy, M., Haynes, A., Herbert, E., Southam, L., Lehrach, H., Weissenbach, J., Manenti, G., Rodriguez-Tome, P., Beddington, R., Dunwoodie, S., and Cox, R.D. (2001) A radiation hybrid transcript map of the mouse genome. *Nat. Genet.* **29:** 194-200.
10. Ballie, D.L. & Rose, A.M. (2000) WABA Success: A Tool for Sequence Comparison between Large Genomes. *Genome Res.* **10:** 1071-1073.
11. Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B., and Lander, E.S. (2000) Human and Mouse Gene Structure: Comparative Analysis and Application to Exon Prediction. *Genome Res.* **10:** 950-958.
12. Bentley, D. (1999) Coupling RNA polymerase II transcription with pre-mRNA processing. *Curr. Opin. Cell Biol.* **11:** 347-351.
13. Berget, S.M. (1995) Exon Recognition in Vertebrate Splicing. *J. Biol. Chem.* **270:** 2411-2414.
14. Bieche, I. & Lidereau, R. (1999) Increased level of exon 12 alternatively spliced BRCA2 trancripts in tumor breast tissue compared with normal tissue. *Cancer Res.* **59:** 2546-2550.
15. Black, D.L. (2000) Protein Diversity from Alternative Splicing: A Challenge for Bioinformatics and Post-Genome Biology. *Cell.* **103:** 367-370.
16. Boise, L.H., Gonzalez-Garcia, M., Postema, C.E., Ding, L., Lindsten, T., Turka, L.A., Mao, X., Nunez, G., and Thompson, C.B. (1993) *Bcl-x*, a *Bcl*-2-related gene that functions as a dominant regulator of apoptotic cell death. *Cell.* **74**(4)**:** 597-608.

17. Bouck, J., Fu, X-D., Skalka, A.M., and Katz, R.A. (1998) Role of the constitutive splicing factors U2AF^{65} and SAP49 in suboptimal RNA splicing of novel retroviral mutants. *J. Biol. Chem.* **273:** 15169-15176.
18. Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J., and Bork, P. (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.* **474:** 83-86.
19. Brett, D., Pospisil, H., Valcárcel, J., Reich, J., and Bork, P. (2002) Alternative splicing and genome complexity. *Nat. Genet.* **30:** 29-30.
20. Burke, J., Davidson, D., and Hide.W. (1999) d2_cluster: A validated method for clustering EST and full-length cDNA. *Genome Res.* **9:** 1135-1142.
21. Burke, J., Wang, H., Hide, W., and Davidson, D. (1998) Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.* **8:** 276-290.
22. Cáceres, J.F. & Krainer, A.R. (1993) Functional analysis of pre-mRNA splicing factor SF2/ASF structural domains. *EMBO J.* **12:** 4715-4726.
23. Cáceres, J.F., Stamm, S., Helfman, D.M., and Krainer, A.R. (1994) *Science.* **265:** 1706-1709.
24. Caputi, M., Mayeda, A., Krainer, A.R., and Zahler, A.M. (1999) HnRNP A/B proteins are required for inhibition of HIV-1 pre-mRNA splicing. *EMBO J.* **18:** 4060-4067.
25. Celera Genomics (2001) The Sequence of the Human Genome. *Science* **291:** 1304-1351.
26. Celotto, A.M. & Graveley, B.R. (2001) Alternative Splicing of the Drosophila *Dscam* Pre-mRNA is Both Temporally and Spatially Regulated. *Genetics.* **159:** 599-608.
27. Chabot, B. (1996) Directing alternative splicing: cast and scenarios. *Trends Genet.* **12:** 472-478.
28. Chebib, M. & Johnston, G.A.R. (1999) The 'ABC' of GABA Receptors: A Brief Review. *Clin Exp Pharmacol Physiol.* **26:** 937-940.

29. Chen, C.D., Kobayashi, R., and Helfman, D.M. (1999) Binding of hnRNP H to an exonic splicing silencer is involved in the regulation of alternative splicing of the rat -tropomyosin gene. *Genes & Dev.* **13:** 593-606.
30. Christoffels, A, van Gelder, A., Greyling, G., Miller, R., Hide, T., and Hide, W. (2001) STACK: Sequence Tag Alignment and Consensus Knowledgebase. *Nucleic Acids Res.* **29**(1): 234-238.
31. Clark, F. & Thanaraj, T.A. (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Unpublished paper*.
32. Clark, M.S. (1999) Comparative genomics: the key to understanding the Human Genome Project. *BioEssays.* **21:** 121-130.
33. Clarke, L.A., Veiga, I., Isidro, G., Jordan, P., Ramos, J.S., Castedo, S., and Boavida, M.G. (2000) Pathological Exon Skipping in an HNPCC Proband With *MLH1* Splice Acceptor Site Mutation. *Genes Chromosomes Cancer.* **29:** 367-370.
34. Collins, C.A. & Guthrie, C. (2000) The question remains: Is the spliceosome a ribozyme? *Nature Struct. Biol.* **7:** 850-854.
35. Conboy, J. (1999) The Role of Alternative Pre-mRNA Splicing in Regulating the Structure and Function of Skeletal Protein 4.1. *Proc. Soc. Exp. Biol. Med.* **220:** 73-78.
36. Cooper, T.A. & Mattox, W. (1997) GENE REGULATION '97 The Regulation of Splice-Site Selection, and Its Role in Human Disease. *Am. J. Hum. Genet.* **61:** 259-266.
37. Croft,L., Schandorff, S., Clark, F., Burrage, K., Arctander, P., and Mattick, J.S. (2000) ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome [letter]. *Nat. Genet*. **24:** 340-341.
38. D'Souza, I., Poorkaj, P., Hong, M., Nochlin, D., Lee, V. M-Y., Bird, T.D., and Schellenberg, G.D. (1999) Missense and silent tau gene mutations cause frontotemporal dementia with parkinsonism-chromosome 17 type, by affecting multiple alternative RNA splicing regulatory elements. *Proc. Natl. Acad. Sci. USA* **96:** 5598-5603.

39. Dehal, P., Predki, P., Olsen, A.S., Kobayashi, A., Folta, P., Lucas, S., Land, M., Terry, A., Zhou, C.L.E., Rash, S., Zhang, Q., Gordon, L., Kim, J., Elkin, C., Pollard,M.J., Richardson, P., Rokhsar, D., Uberbacher, E., Hawkins, T., Branscomb, E., and Stubbs, L. (2001) Human Chromosome 19 and Related Regions in Mouse: Conservative and Lineage-Specific Evolution. *Science.* **293:** 104-111.
40. Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O., and Salzberg, S.L. (1999) Alignment of whole genomes. *Nucleic Acids Res.* **27** (11): 2369-2376.
41. Eisen, J.A. (1998) Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. *Genome Res.* **8:** 163-167.
42. Eperon, I.C. *et al.*(1993) Pathways for selection of 5' splice sites by U1 snRNPs and SF2/ASF. *EMBO J.* **9:** 3607-3617.
43. Ewing, B. et al (1998) Base-calling of automated sequence traces using phred. I. Accuracy assessment. *Genome Res.* **8**(3): 175-185.
44. Fang, W., Rivard, J.J., Mueller, D.L., and Behrens, T.W. (1994) Cloning and molecular characterization of mouse *Bcl-x* in B and T lymphocytes. *J. Immunol.* **153**(10): 4388-4398.
45. Ffrench-Constant, C. (1995) Alternative Splicing of Fibronectin – Many Different Proteins but Few Different Functions. *Exp. Cell Res.* **221:** 261-271.
46. Flint, J., Tufarelli, C., Peden, J., Clark, K., Daniels, R.J., Hardison, R., Miller, W., Philipsen, S., Tan-Un, K.C., McMorrow, T., Frampton, J., Alter, B.P., Frischauf, A-M., and Higgs, D.R. (2001) Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the globin cluster. *Hum. Mol. Genet.* **10**(4): 371-382.
47. Florea, L., Hartzell, G., Zhang, Z., Rubin, G., and Miller, W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8:** 967-974.
48. Friedman, K.J., Kole, J., Cohn, J.A., Knowles, M.R., Silverman, L.M., and Kole, R. (1999) Correction of aberrant splicing of CFTR gene by antisense oligonucleotides. *J. Biol. Chem.* **274:** 36193-36199.

49. Fujimura, M., Kasahara, K., Shirasaki, H., Heki, U., Isawa, K., Ueda, A., and Matsuda, T. (1999) Up-regulation of ICH-1L protein by thromboxane A2 antagonists enhances cisplatin-induced apoptosis in non-small cell lung cancer cell lines. *J Cancer Res. Clin. Oncol.* **125:** 389-394.
50. Gilbert,W. (1978) Why genes in pieces? *Nature* **271:** 501.
51. Goldstrohm, A.C., Greenleaf, A.L., and Garcia-Blanco, M.A. (2001) Co-transcriptional splicing of pre-messenger RNAs: considerations for the mechanism of alternative splicing. *Gene.* **277:** 31-47.
52. González-García, M., Perez-Ballestero, R., Ding, L.Y., Duan, L., Boise, L.H., Thompson, C.B., and Núñez, G. (1994) *Bcl*-xL is the major *Bcl*-x mRNA form expressed during murine development, and its product localizes to mitochondria. *Development.* **120:** 3033-3042.
53. Grabowski, P. J. & Black, D.L. (2001) Alternative RNA splicing in the nervous system. *Prog. Neurobiol.* **65:** 289-308.
54. Graveley, B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* **17:** 100-107.
55. Guiramand, J., Montmayeur, J-P., Ceraline, J., Bhatia, M., and Borrelli, E. (1995) Alternative Splicing of the Dopamine D2 Receptor Directs Specificity of Coupling to G-proteins. *J. Biol. Chem.* **270**(13)**:** 7354-7358.
56. Han, J.S., Nunez, G., Wicha, M.S., and Clarke, M.F. (1998) Targeting cancer cell death with a bcl-XS adenovirus. *Springer Semin. Immunopathol.* **19:** 279-288.
57. Hanke, J., Brett, D., Zastrow, I., Aydin, A., Delbruck, S., Lehmann,G., Luft, F., Reich, J., and Bork,P. (1999) Alternative splicing of human genes- more the rule than the exception? *Trends Genet.* **15:** 389-390.
58. Hardison, R.C., Oeltjen, J., and Miller, W. (1997) Long Human-Mouse Sequence Alignments Reveal Novel Regulatory Elements: A Reason to Sequence the Mouse Genome. *Genome Res.* **7:** 959-966.
59. Harn, H.J., Ho, L.I., Shyu, R.Y., Yuan, J.S., Lin, F.G., Young, T.H., Liu, C.A., Tang, H.S., and Lee, W.H. (1996) Soluble CD44 isoforms in serum as potential markers of metastatic gastric carcinoma. *J. Clin. Gastroenterol.* **22:** 107-110.

60. Herrera-Gayol, A. & Jothy, S. (1999) Adhesion proteins in the biology of breast cancer: contribution of CD44. *Exp. Mol. Pathol.* **66:** 149-156.
61. Hertel, K.J., Lynch, K.W., Maniatis T. (1997) Common themes in the function of transcription and splicing enhancers. *Curr. Opin. Cell Biol.* **9:** 350-357.
62. Hide, W.A., Babenko, V.N., van Heudsen, P.A., Seoighe, C., and Kelso, J.F. (2001) The contribution of exon skipping events on Chromosome 22 to protein coding diversity. *Genome Res.* **11**:1848-53
63. Hukriede, N., Fisher, D., Epstein, J., Joly, L., Tellis, P., Zhou, Y., Barbazuk, B., Cox, K., Fenton-Noriega, L., Hersey, C., Miles, J., Sheng, X., Song, A., Waterman, R., Johnson, S.L., Dawid, I.B., Chevrette, M., Zon, L.I., McPherson, J., Ekker, M. (2001) The LN54 radiation hybrid map of zebrafish expressed sequences. *Genome Res.* **11:**2127-32
64. Hutchinson, S., Wordsworth, P., Handford, P.A. (2001) Marfan syndrome caused by a mutation in *FBN1* that gives rise to cryptic splicing and a 33 nucleotide insertion in the coding sequence. *Hum. Genet.* **109:** 416-420.
65. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.
66. Jareborg, N. & Durbin, R. (2000) Alfresco- A Workbench for Comparative Genomic Sequence Analysis. *Genome Res.* **10:** 1148-1157.
67. Jiang, Z. & Wu, J.Y. (1999) Alternative Splicing and Programmed Cell Death. *Proc Soc Exp Biol Med.* **220:** 64-72.
68. Jongeneel, C.V. (2000) Searching the expressed sequence tag (EST) databases: Panning for genes. *Brief. Bioinform.* **1:** 76-92.
69. Kalbfuss, B., Mabon, S.A., Misteli, T. (2001) Correction of Alternative Splicing of Tau in Frontotemporal Dementia and Parkinsonism Linked to Chromosome 17. *J. Biol. Chem.* **276:** 42986-42993.
70. Kan, J.L.C. & Green, M.R. (1999) Pre-mRNA splicing of IgM exons M1 and M2 is directed by a juxtaposed splicing enhancer and inhibitor. *Genes & Dev.* **13:** 462-471.

71. Kan, Z., Gish, W., Rouchka, E., Glasscock, J., and States, D. (2000) UTR Reconstruction and Analysis Using Genomically Aligned EST Sequences. *Intell. Syst. Mol. Biol.* **8:** 218-227.
72. Kan, Z., Rouchka, E.C., Gish, W.R., and States, D.J. (2001) Gene Structure Prediction and Alternative Splicing Analysis Using Genomically Aligned ESTs. *Genome Res.* **11:** 889-900.
73. Kanopa, A., Mühlemann, O., Petersen-Mahrt, S., Estmer, C., Öhrmalm, and Akusjärvi, G. (1998) Regulation of adenovirus alternative RNA splicig by dephosphorylation of SR proteins. *Nature.* **393:** 185-187.
74. Kirschner, R., Erturk, D., Zeitz, C., Sahin, S., Ramser, J., Cremers, F.P.M., Ropers, H.H., and Berger, W. (2001) DNA sequence comparison of human and mouse retinitis pigmentosa GTPase regulator (*RPGR)* identifies tissue-specific exons and putative regulatory elements. *Hum. Genet.* **109:** 271-278.
75. 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.
76. Koonin, E.V. (2001) An apology for orthologs – or brave new memes. *Genome Biology.* **2**(4): comment 1005.1- 1005.2.
77. Krecic, A.M. & Swanson, M.S. (1999) HnRNP complexes: composition, structure, and function. *Curr Opin Cell Biol.* **11:** 363-371.
78. Lane, R.P., Cutforth, T., Young, J., Athanasiou, M., Friedman, C., Rowen, L., Evans, G., Axel, R., Hood, L., and Trask, B.J. (2001) Genomic analysis of orthologous mouse and human olfactory receptor loci. *Proc. Natl. Acad. Sci.* **98:** 7390-7395.
79. Laverdière, M., Beaudoin, J., and Lavigueur A. (2000) Species-specific regulation of alternative splicing in the C-terminal region of the p53 tumor suppressor gene. *Nucleic Acids Res.* **28** (6): 1489-1497.
80. Lee, C.G., Zamore, P.D., Green, M.R., and Hurwitz, J. (1993) RNA annealing activity is intrinsically associated with U2AF. *J. Biol. Chem.* **268:** 13472-13478.
81. Lewis, J.D., Gunderson, S.I., Mattaj, I.W. (1995) The influence of 5’ and 3’ end structures on pre-mRNA metabolism. *J. Cell Sci.* (Suppl.) **19:** 13-19.

82. Li, H. & Bingham, P.M. (1991) Arginine/Serine rich domains of the su(wa) and tra RNA processing regulators target proteins to a subnuclear compartment implicated in splicing. *Cell.* **67:** 335-342.
83. Liew, C.C., Hwang, D.M., Fung, Y.W., Laurenssen, C., Cukerman, E., Tsui, S., Lee, C.Y. (1994) A catalogue of genes in the cardiovascular system as identified by expressed sequence tags. *Proc. Natl. Acad. Sci. U S A.* **91**:10645-9
84. López, A.J, (1998) Alternative splicing of pre-mRNA:developmental consequences and mechanisms of regulation. *Annu. Rev. Genet*. **32:** 279-305.
85. Lou, H., Helfman, D.M., Gagel, R.F., and Berget, S.M. (1999) Polypyrimidine Tract-Binding Protein Positively Regulates Inclusion of an Alternative 3'-Terminal Exon. *Mol. Cell. Biol.* **19:** 78-85.
86. Lynch, K.W. & Maniatis, T. (1996) Assembly of specific SR protein complexes on distinct regulatory elements of the *Drosophila* doublesex splicing enhancer. *Genes Dev.* **10:** 2089-2101.
87. MacDougall, C., Harbison, D., and Bownes, M. (1995) The Developmental Consequences of Alternate Splicing in Sex Determination and Differentiation in *Drosophila. Dev. Biol.* **172:** 353-376.
88. Makalowski, W., Boguski, M. (1998) Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. USA.* **95:** 9407-9412.
89. Martindale, D.W., Wilson, M.D., Wang, D., Burke, R.D., Chen, X., Duronio, V., and Koop, B.F. (2000) Comparative genomic sequence analysis of th Williams syndrome region (LIMK1-RFC2) of human Chromosome 7q11.23. *Mamm. Genome.* **11:** 890-898.
90. Miller, R., Christoffels, A., Gopalakrishanan, C., Burke, J., Ptitsyn, A.A., Broveak, T.R., and Hide, W. (1999) A comprehensive approach to clustering of expressed human gene sequence: The sequence tag alignment and consensus knowledge base. *Genome Res.* **9:** 1143-1155.
91. Mion, S., Corti, c., Neki, A., Shigemoto, R., Corsi, M., Fumagalli, G., and Ferraguti, F. (2001) Bidirectional Regulation of Neurite Elaboration by

Alternatively Spliced Metabotropic Glutamate Receptor 5 (mGluR5) Isoforms. *Mol. Cell Neurosci.* **17:** 957-972.

92. Mironov, A.A., Fickett, J.W., and Gelfand, M.S. (1999) Frequent Alternative Splicing of Human Genes. *Genome Res.* **9:** 1288-1293.
93. Modrek, B. & Lee, C. (2002) A genomic view of alternative splicing. *Nat. Genet.* **30:** 13-19.
94. Modrek, B., Resch,A., Grasso, C., and Lee, C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. **29**(13): 2850-2859.
95. Mott, R. (1997) EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. *CABIOS.* **13:** 477-478.
96. Mouse Sequencing Liason Group. The Latest Progress From the Mouse Genome Sequencing Consortium. *Mouse Genome Monthly,* issue 1 November 2001.
97. Mushegian, A.R., Garey, J.R., Martin, J., and Xiu, L.X. (1998) Large-Scale Taxonomic Profiling of Eukaryotic Model Organisms: A Comparison of Orthologous Proteins Encoded by the Human, Fly, Nematode, and Yeast Genomes. *Genome Res.* **8:** 590-598.
98. Noack, D., Heyworth, P.G., Newburger, P.E., and Cross, A.R. (2001) An unusual intronic mutation in the *CYBB* gene giving rise to chronic granulomatous disease. *Biochim. Biophys. Acta.* **1537:** 125-131.
99. Oeltjen, J.C., Malley, T.M., Muzny, D.M., Miller, W., Gibbs, R.A., and Belmont, J.W. (1997) Large-Scale Comparative Sequence Analysis of the Human and Murine Bruton's Tyrosine Kinase Loci Reveals Conserved Regulatory Domains. *Genome Res.* **7:** 315-329.
100. Ogbourne, S. & Antalis, T.M. (1998) Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochem. J.* **331:** 1-14.
101. Osborne, S.A. & Tonissen, K.F. (2001) Genomic organization and alternative splicing of mouse and human thioredoxin reductase 1 genes. *BMC Genomics.* **2:** 10.
102. Ouzounis, C. (1999) Orthology: another terminology muddle. *Trends Genet.* **15** (11): 445.

103. Penzel, R., Uhl, J., Kopitz, J., Beck, M., Otto, H.F., and Cantz, M. (2001) Splice donor site mutation in the lysosomal neuraminidase gene causing exon skipping and complete loss of enzyme activity in a sialidosis patient. *FEBS Letters.* **501:** 135-138.

104. Philips, A.V. & Cooper, T.A. (2000) RNA processing and human disease. *Cell. Mol. Life Sci.* **57:** 235-249.

105. Quinlan, J.J., Firestone, L.L., Homanics, G.E. (2000) Mice lacking the long splice variant of the gamma 2 subunit of the GABA(A) receptor are more sensitive to benzodiazepines. *Pharmacol. Biochem. Behav.* **66:** 371-374.

106. Reiter, J.L., Threadgill, D.W., Eley, G.D., Strunk, K.E., Danielsen, A.J., Sinclair, C.S., Pearsall, R.S., Green, P.J., Yee, D., Lampland, A.L., Balasubramaniam, S., Crossley, T.D., Magnuson, T.R., James, C.D., and Maihle, N.J. (2001) Comparative Genomic Sequence Analysis and Isolation of Human and Mouse Alternative *EGFR* Transcripts Encoding Truncated Receptor Isoforms. *Genomics.* **71:** 1-20.

107. Remm, M. & Sonnhammer, E. (2000) Classification of Transmembrane Protein Families in the *Caenorhabditis elegans* Genome and Identification of Human Orthologs. *Genome Res.* **10:** 1679-1689.

108. Ryan, K.R. & Cooper, T.A. (1996) Muscle-specific splicing enhancers regulate inclusion of the cardiac troponin T alternative exon in embryonic skeletal muscle. *Mol. Cell. Biol.* **16:** 4014-4023.

109. Ryner, L.C. & Baker, B.S. (1991) Regulation of doublesex pre-mRNA processing occurs by 3'-splice site activation. *Genes Dev.* **5:** 2071-2085.

110. Sakamoto, O., Ohura, T., Katsushima, Y., Fujiwara, I., Ogawa, E., Miyabayashi, S., and Iinuma, K. (2001) A novel intronic mutation of the *TAZ* (G4.5) gene in a patient with Barth syndrome: creation of a 5' splice donor site with variant GC consensus and elongation of the upstream exon. *Hum. Genet.* **109:** 559-563.

111. Scheetz, T.E, Raymond, M.R., Nishimura, D.Y., McClain, A., Roberts, C., Birkett, C., Gardiner, J., Zhang, J., Butters, N., Sun, C., Kwitek-Black, A.,

Jacob, H., Casavant, T.L., Soares, M.B., and Sheffield, V.C. (2001) Generation of a High-Density Rat EST Map. *Genome Res.* **11:** 497-502.

112. Schmajuk, G. Sierakowska, H., and Kole, R. (1999) Antisense Oligonucleotides with Different Backbones. *J. Biol. Chem.* **274**(31): 21783-21789.

113. Schmitt, A.O., Specht, T., Beckmann, G., Dahl, E., Pilarsky, C.P., Hinzmann, B., and Rosenthal, A. (1999) Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acids Res.* **27**: 4251-4260.

114. Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. (2000) PipMaker- A Web Server for Aligning Two Genomic DNA Sequences. *Genome Res.* **10:** 577-586.

115. Shiraiwa, N., Inohara, N., Okaka, S., Yuzaki, M., Shoji, S., and Ohta, S. (1996) An additional form of rat *Bcl*-x, *Bcl*-x[beta], generated by an unspliced RNA, promotes apoptosis in promyeloid cells. *J. Biol. Chem.* **271:** 13258.

116. Sierakowska, H., Sambade, M.J., Agrawal, S., and Kole, R. (1996) Repair of thalassemic human -globin mRNA in mammalian cells by antisense oligonucleotides. *Proc. Natl. Acad. Sci. USA.* **93:** 12840-12844.

117. Skrygan, M., Bartholome, B., Bonafe, L., Blau, N., and Bartholome, K. (2001) A splice mutation in the GTP cyclohydrolase I gene causes dopa-responsive dystonia by exon skipping. *J. Inherit. Metab. Dis.* **24**: 345-351.

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

119. Stein, C.A. (1999) Keeping the biotechnology of antisense in context. *Nat. Genet.* **17:** 209.

120. Subramaniam, V., Bomze, H.M., and Lopez, A.J. (1994) Functional differences between Ultrabithorax protein isoforms in Drosophila melanogaster: evidence from elimination, substitution and ectopic expression of specific isoforms. *Genetics.* **136:** 979-991.

121. Tarn, W-Y. & Steitz, J.A. (1997) Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge. *Trends Biochem. Sci.* **22:** 132-137.

122. Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997) A Genomic Perspective on Protein Families. *Science.* **278:** 631-637.

123. Thackeray, J.R., and Ganetzky, B. (1994) Developmentally Regulated Alternative Splicing Generates a complex Array of *Drosophila para* Sodium Channel Isoforms. *The Journal of Neuroscience* **14:** 2569-2578.

124. Thanaraj, T.A. (1999) A clean data set of EST-confirmed splice sites from *Homo sapiens* and standards for clean-up procedures. *Nucleic Acids Res.* **27**(13): 2627-2637.

125. The First International Workshop on Comparative, Genome Organization. (1996) Comparative Genome Organization of Vertebrates. *Mamm. Genome.* **7:**717.

126. Tran, H., Mattei, M., Godyna, S., and Argraves, W.S. (1997) Human fibulin-1D: molecular cloning, expression and similarity with S1-5 protein, a new member of the fibulin gene family. *Matrix Biol.* **15** (7): 479-493.

127. Valcárcel, J. & Green, M.R. (1996) The SR protein family: pleiotropic functions in pre-mRNA splicing. *Trends Biochem. Sci.* **21:** 296-301.

128. van der Houven van Oordt, W., Diaz-Meco, M.T., Lozano, J., Krainer, A.R., Moscat, J., and Cáceres, J.F. (2000) The $MKK_{3/6}$-p38-signalling Cascade Alters the Subcellular Distribution of hnRNP A1 and Modulates Alternative Splicing Regulation. *J. Cell Biol.* **149:** 307-316.

129. van Deutekom, J.C.T., Bremmer-Bout, M., Janson, A.A.M., Ginjaar, L.B., Baas, F., den Dunnen, J.T., and van Ommen, G-J. B. (2001) Antisense-induced exon skipping restores dystrophin expression in DMD patient derived muscle cells. *Hum. Mol. Genet.* **10**(15): 1547-1554.

130. van Oers, C.C., Adema, G.J., Zandberg, H., Moen, T.C., and Baas, P.D. (1994) Two different sequence elements within exon 4 are necessary for calcitonin-specific splicing of the human calcitonin/calcitonin gene-related peptide I pre-mRNA. *Mol. Cell Biol.* **14:** 951-960.

131. Vasmatzis G., Essand, M., Brinkmann, U., Lee, B., Pastan, I. (1998) Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proc. Natl. Acad. Sci. U S A* **95:** 300-304.

132. Vorwerk, P., Christoffersen, C.T., Müller, J., Vestergaard, H., Pedersen, O., and De Meyts, P. (1999) Alternative Splicing of Exon 17 and a Missense Mutation in Exon 20 of the Insulin Receptor Gene in Two Brothers with a Novel Syndrome of Insulin Resistance (Congenital Fiber-Type Disproportion Myopathy). *Horm. Res.* **52:** 211-220.

133. Wafford, K.A., Bain, C.J., Whiting, P.J., Kemp, J.A. (1993) Functional comparison of the role of gamma subunits in recombinant human gamma-aminobutyric acidA/benzodiazepine receptors. *Mol. Pharmacol.* **44:** 437-442.

134. Wang, J. & Manley, J.L. (1995) Overexpression of the SR proteins ASF/SF2 and SC35 influences alternative splicing in vivo in diverse ways. *RNA.* **1:** 335-346.

135. Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J.W., and Lawrence, C.E. (2000) Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.* **26:** 225-228.

136. Weg-Remers, S., Ponta, H., Herrlich, P., and König, H. (2001) Regulation of alternative pre-mRNA splicing by the ERK MAP-kinase pathway. *EMBO J.* **20:** 4194-4203.

137. Wheelan, S.J., Boguski, M.S., Duret, L., and Makalowski, W. (1999) Hman and nematode orthologs – lessons from the analysis of 1800 human genes and the proteome of *Caenorhabditis elegans. Gene.* **238:** 163-170.

138. Wheelan, S.J., Church, D.M., and Ostell, J.M. (2001) Spidey: A Tool for mRNA-to-Genomic Alignments. *Genome Res.* **11**:1952-7

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

140. Wright, F.A., Lemon, W.J., Zhao, W.D., Sears, R., Zhuo, D., Wang, J-P., Yang, H-Y., Baer, T., Stredney, D., Spitzner, J., Stutz, A., Krahe, R., and Yuan, B. (2001) A draft annotation and overview of the human genome. *Genome Biology.* **2:** 1-30.

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

142. Xie, T. & Ding, D. (2000) Investigating 42 candidate orthologous protein groups by molecular evolutionary analysis on genome scale. *Gene.* **261:** 305-310.

143. Yang, X.F., Weber, G.F., and Cantor, H. (1997) A novel *Bcl*-x isoform connected to the T-cell receptor regulates apoptosis in T cells. *Immunity.* **7:** 629-639.

144. Zahler, A.M. & Roth, M.B. (1995) Distinct functions of SR proteins in recruitment of U1 small nuclear ribonucleoprotein to alternative 5' splice sites. *Proc. Natl. Acad. Sci.* **92:** 2642-2646.

145. Zhu, J., Mayeda, A., Krainer, A.R. (2001) Exon Identity Established through Differential Antagonism between Exonic Splicing Silencer-Bound hnRNP A1 and Enhancer-Bound SR Proteins. *Mol. Cell.* **8:** 1351-1361.

146. Zuo, P. & Maniatis, T. (1996) The splicing factor U2AF^{35} mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing. *Genes & Dev.* **10:** 1356-1368.

THESIS

INTRODUCTION

Exon skipping has recently been established as the most prevalent form of alternative splicing pattern in human genes by several studies (Modrek *et al.*2001; Clark & Thanaraj 2002; Croft *et al.*2000; Hide *et al.*2001). The fundamentals and the importance of alternative splicing have also been discussed in Chapters 1 and 2 of the literature review. Exon skipping has been implicated in the developmental regulation of *Drosophila* (Chabot 1996), generation of protein functional diversity in the human $GABA_A$ receptor family (Quinlan *et al.*2000), and human genetic disorders arising from point mutations (D'Souza *et al.*1999; Kalbfuss *et al.*2001; Cooper & Mattox 1997; Penzel *et al.*2001; Hutchinson *et al.*2001; Sakamoto *et al.*2001; Noack *et al.*2001). Biological explanations for the frequent occurrence of exon skipping in human genes are still unclear.

Recent progress in mouse genomic sequencing has resulted in the availability of sufficient mouse genomic contigs (Ensembl v.0.1.0) to perform a genome-anchored transcript comparison study. Mouse contig sequences are freely available at the mouse Ensembl website (http://www.ensembl.org/Mus_musculus/) at the Sanger centre. The availability of mouse genomic sequences provides the opportunity to investigate alternative splicing in mouse genes as well as perform comparative mouse and human gene analyses. Previous progress in mouse and human alternative splicing gene comparisons may have been hindered due to the lack of mouse genomic sequences. However, exploratory alternative splicing studies in mouse genes have been performed without mouse genomic sequences (Kan *et al.*2001; Brett *et al.*2002).

Human exon-skipping events have been detected through alignments of ESTs (expressed sequence tags) to their corresponding genomic sequences. The approach has been utilized extensively by many studies (Modrek *et al.*2001; Clark & Thanaraj 2002; Hide *et al.*2001; Kan *et al.*2001; Wolfsberg & Landsman 1997; Thanaraj 1999;

Mironov *et al.*1999; Kan *et al.*2000) to detect alternatively spliced events in human genes. Alignments of ESTs to genomic sequences can be utilized as an effective screen to remove errors in ESTs (Modrek & Lee 2002). The utilization of expressed sequence tags to detect alternatively spliced events has been extensively reviewed in Chapter 4 of the literature review.

The current study aims to explore the occurrence of mouse exon skipping in mouse genes orthologous to chromosome 22. Exon-skipping events will be captured through alignments of ESTs to genomic sequences. The study will provide the basis for constructing hypotheses that can be tested in future mouse alternative splicing studies.

METHODS

Data sources

A full listing of data source references is presented in Appendix A. In order to differentiate between mRNA/cDNA and EST sequences, the term 'transcript' was re-defined to represent only mRNA or cDNA sequences. Therefore, EST sequences were not referred to as transcripts in this study unless otherwise mentioned.

Software

- EMBOSS (European Molecular biology Software Suite, v1.11.0) applications such as extractseq and transeq were utilized. All extraction and concatenation of sequences were performed using extractseq and translations of sequences were performed by transeq.
- Sim4 (Florea *et al.*1998) was used for all alignments between transcript (mRNA, cDNA, and EST) and genomic sequences under default parameters unless otherwise specified.
- InterProScan (http://www.ebi.ac.uk/interpro/scan.html) were used to search against the InterPro database (v3.2, July 2001) at EBI for functional protein domains and families.

A. Construction of a Mouse Exon-skipping Detection Pipeline

All scripts used for parsing data were written in PERL and BIOPERL. Parsing scripts are detailed in Appendix Bi.

(i) Generation of 269 mouse transcripts orthologous to human genes on human chromosome 22

Human chromosome 22 mRNAs (Appendix A, i.a) were searched against a pooled mouse transcript database (Appendix A, i.b-e) using BLASTN 2.2.1 program ($P<1x10^{-20}$) (Altschul *et al.*1990). A unique, transcript pair was defined as a human sequence that only has one top BLAST match to a mouse sequence and vice versa. Mouse and human pairs were removed, if a human sequence has more than 1 top BLAST match to a mouse sequence and vice versa. After removal of 34 transcript pairs, a total of 269 unique mouse and human transcript pairs were obtained. The reciprocal BLAST method (Tatusov *et al.*1997) was utilized to predict orthologous mouse and human transcripts. The observed short evolutionary distance between orthologous human and mouse genes (Kumar & Hedges 1998) enabled us to use the reciprocal BLAST method. The feasibility of the detection of putative orthologues using this method has been reviewed in Chapter 5.3 of the literature review. 269 transcript pairs were searched against 524 human chromosome 22 mRNAs and a pooled mouse transcript database using BLASTN 2.2.1 ($P<1x10^{-20}$). Orthologous transcript pairs that have a matching alignment length and a sequence identity of less than 80% were discarded.

(ii) Generation of 72 mouse multi-exon gene set

269 unique mouse transcripts were used to search against ENSEMBL mouse genomic contigs (Appendix A, ii.a) using BLASTN 2.2.1 ($P<1x10^{-20}$). The criteria that were used to predict multi-exon genes are listed below:

1. A multi-exon transcript was defined as containing at least 3 exons.
2. Exons were identified through uniquely matching regions between a mouse multi-exon transcript and its corresponding mouse contig. The exons must be minimally overlapping (less or equal to 8 nucleotides) when compared to other exons on the same mouse transcript. Each exon predicted on the mouse contig must also be discrete when compared to other predicted exons.

3. Multi-exon transcripts were identified as having at least 3 different, minimally overlapping exons matching to discrete, non-overlapping mouse contig regions.

Mouse transcripts with contigs that have unknown chromosomal positions were kept. The ENSEMBL clone identifiers, available for each mouse contig (v.0.1.0), were queried against the ENSEMBL mouse contig database (v.0.2.0, Appendix A, ii.b) to obtain chromosomal positions for each mouse contig. The Ensembl mouse genomic assembly v.0.2.0 contains partial positional information whereas v.0.1.0 lacked positional information. The criteria that were used to remove paralogous transcripts are listed below:

1. A mouse transcript and its corresponding contig belonged to different chromosomal positions.
2. Predicted consecutive exon positions were far apart (>4kb) in alignments between transcript and genomic sequences.

Exon positions for mouse transcript sequences with respect to their genomic sequences (see supplementary table 2 in Appendix C) were obtained for each unique mouse transcript orthologue. These exon positions were defined as **BLAST-predicted exon positions** to avoid confusion.

(iii) Construction of virtual mouse transcript and genomic sequences

Mouse transcript and genomic sequences were reconstructed for each gene in the 72 mouse multi-exon gene set. Reconstructed transcript and genomic sequences were defined as virtual sequences. Virtual sequences were constructed to satisfy two needs: (i) to represent both transcript- and genome-verified exons, (ii) to facilitate the description of transcript- and genome-verified exon positions. Virtual transcript sequences were constructed by concatenating exon sequences extracted from their corresponding mouse orthologous transcripts using the BLAST-predicted exon positions. Virtual genomic sequences were constructed by concatenation of exon

sequences with a 44bp artificial intron sequence (40N's flanked by consensus splice-site signals, GT and AG nucleotides). Sim4 alignments between virtual transcripts and genomic sequences produced exon positions, which were defined as virtual exon positions.

(iv) Aligning mouse ESTs to virtual mouse genomic sequences

The utilization of expressed sequence tags to detect alternatively spliced events has been reviewed in Chapter 4 of the literature review. Seventy-two virtual mouse transcripts were searched against a mouse dbEST database (Appendix A, ii.f) using BLASTN 2.2.1 ($P<1x10^{-20}$). One unique mouse Ensembl transcript accession (ENSMUST00000003623) was eliminated as a result of having no EST representatives in the database. Mouse ESTs obtained in this manner for each virtual mouse transcript were aligned against their corresponding virtual genomic sequence.

(v) Optimization of sim4 and BLAST parameters for the detection of mouse exon-skipping events

Six previously published exon-skipped, mouse and human genes (Mbd1 (Hendrich *et al.*1999), Wbscr1 (Martindale *et al.*2000), prosaposin (Zhao *et al.*1997), GH1 (Palmetshofer *et al.*1995), PTS (Liu *et al.* 2001), and NF2 (Schmucker *et al.*1999)) were used to determine the cut-off parameters for exon-skipping detection. Known exon-skipped transcripts were searched against mouse and human dbEST (Appendix A.i.g for human EST database) at NCBI using BLASTN 2.2.1 at default parameters. Mouse and human ESTs were aligned against their genomic sequences using sim4 under default parameters. The known exon-skipping events were confirmed by mouse and human ESTs with a BLAST P-value of less than $7x10^{-44}$ (lower limit) and were detected in sim4 alignments under default parameters. The approach yielded zero false positives. All predicted mouse exon-skipped events were detected with P-values less than the lower limit (see supplementary table 3 in Appendix C).

(vi) Detection of mouse exon-skipped ESTs

Sim4 reports produced locations and percentage identities for each matching region between alignments of ESTs and genomic sequences. Comparisons of sim4 exon positions (observed exon positions) to virtual exon positions (known exon positions) identified missing exon positions in sim4 alignments. Missing exon positions were treated as exon-skipping events (See Fig. 1).

The criteria for the validation of an exon-skipping event are listed below:

1. Matching exon locations between sim4 and virtual genomic exon locations must flank the missing exon location.
2. If the exons flanking the exon-skipped event have exon lengths greater than 50bp, the sequence identity between the matching regions must be greater than 93%.
3. If the exon lengths were less than 50bp, 100% sequence identity must be obtained.

(vii) Removal of paralogous exon-skipped ESTs

A paralogous EST was identified as having top BLAST matches to mouse contigs belonging to different chromosomal positions or having top BLAST matches to different and distant positions on the same mouse contig. Exon-skipped ESTs were searched against mouse contigs (Ensembl v.0.2.0) using BLASTN 2.2.1 ($P<1x10^{-40}$) in order to search for paralogous ESTs among the exon-skipped ESTs. Five paralogous ESTs with the following Genbank accessions: AW045080, BF322308, AA244369, W80258, and W34728 were removed.

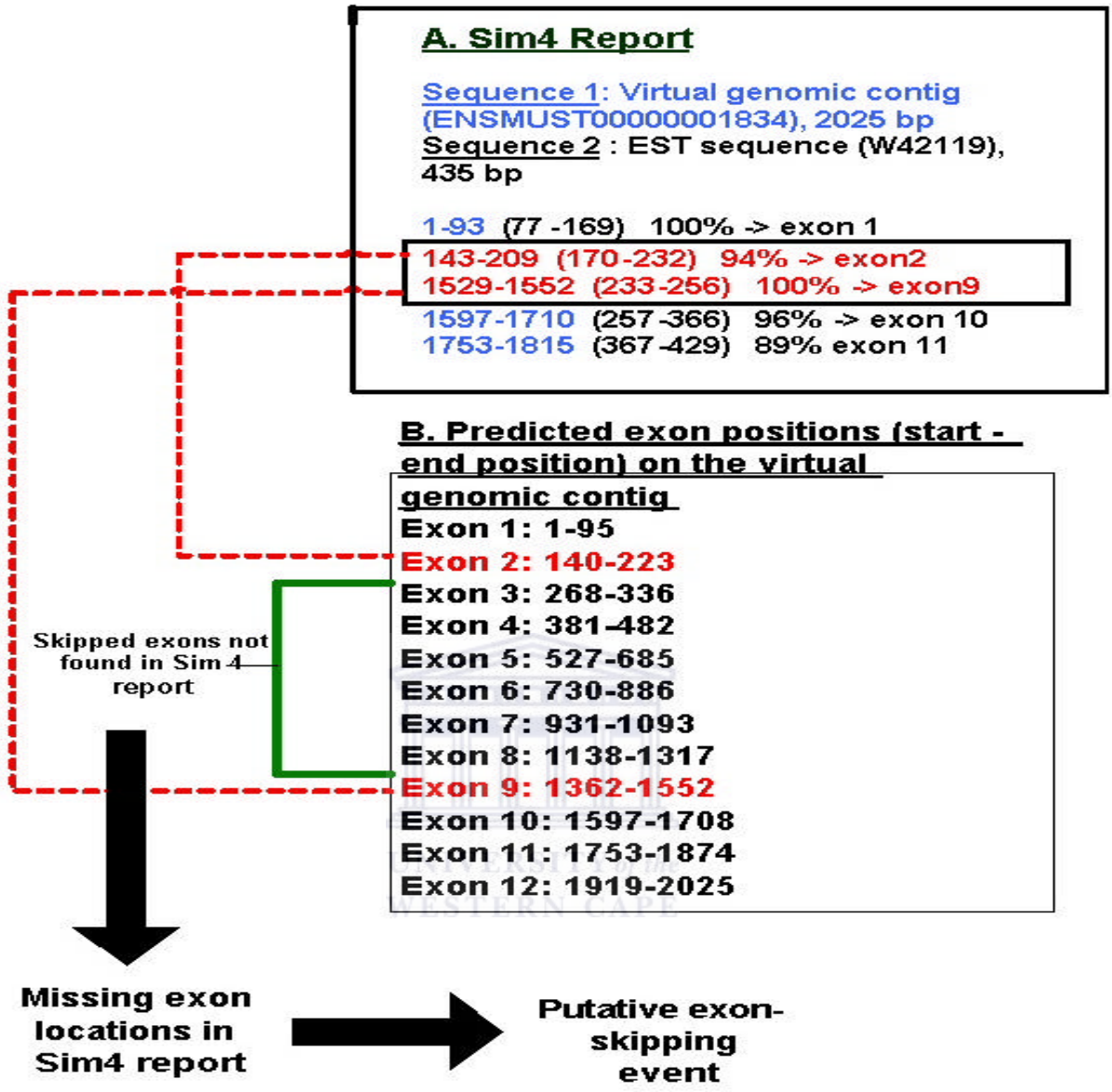


Figure 1: Strategy for the detection of an exon-skipping event

A sim4 generated alignment between an EST and its corresponding genomic sequence (Fig. A) produces locations of similarity between the EST and genomic sequence alignment. These locations of similarity are described numerically with start and end positions. Exon positions of an EST sequence (Genbank nucleotide accession: W42119) can be determined by taking its corresponding virtual genomic positions in the sim4 report (Fig. A) and comparing to known virtual genomic exon positions in Fig. B (see red dotted lines). Missing exon positions were treated as putative exon-skipping events.

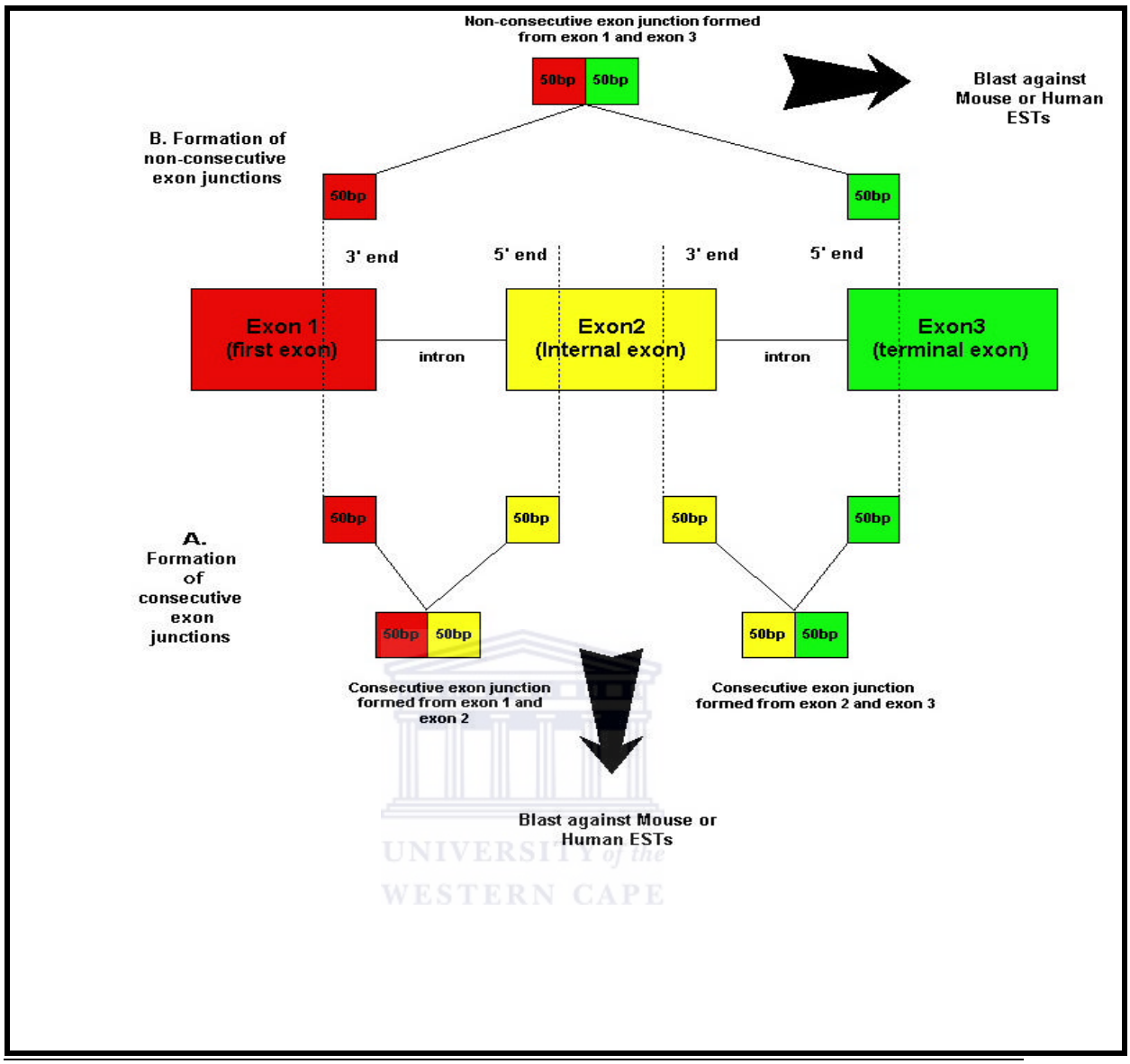
B. Determination of the Rate of Exon skipping Per Gene, Exon-junction Coverage, and EST Representation in Mouse and Human Multi-exon Gene sets

(i) Rate of exon-skipping

The average number of exon-skipping events per gene was calculated by dividing the total number of mouse exon-skipping events by the total number of mouse multi-exon genes sampled (72). The average number of exon-skipping events per exon-skipped gene was calculated by dividing the total number of exon-skipping events by the total number of exon-skipped genes (6). The average rates of mouse exon skipping per gene (exon-skipped gene) were compared to the human exon-skipping rates obtained for chromosome 22 (Hide *et al.*2001).

(ii) Consecutive and non-consecutive exon-junction coverage

Generation of consecutive and non-consecutive exon junctions were produced using j_explorer.pl (Hide *et al.*2001) and are detailed in Fig. 2. The average consecutive exon-junction coverage per gene was calculated by taking the total number of observed, consecutive exon junctions (irrespective of the number of ESTs matching to a given exon junction) and dividing by the known number of consecutive exon junctions. The average consecutive exon-junction coverage was compared between: (i) mouse and human exon-skipped genes, (ii) the total number of mouse and human independent, multi-exon genes. The probability for a given mouse multi-exon transcript to have at least one non-consecutive exon junction was calculated. Details of the probability model have been previously described (Hide *et al.*2001) and calculations are detailed in Appendix B.ii. The human multi-exon gene set was obtained from Hide *et al.*2001 (Appendix A.ii.d).



<u>Figure 2</u>: Insilico formation of mouse consecutive and non-consecutive exon-junctions

Mouse consecutive and non-consecutive exon-junctions were generated by j_explorer.pl, which was previously described by Hide and others [4]. A consecutive exon-junction is defined as extracting 50 base pairs between the ends of consecutive exons (A). A non-consecutive exon-junction is defined as extracting 50 base pairs between the ends of non-consecutive exons (B). Mouse and human exon-junctions were generated using virtual genomic constructs and searched against mouse and human dbEST databases at NCBI using BLASTN 2.2.1 ($P<1x10^{-40}$).

(iii) EST representation

The average EST representation was determined by taking the total number of ESTs that have matches to both consecutive and non-consecutive exon junctions and dividing by the total number of multi-exon genes (or total number of exon-skipped genes). The average EST representation was compared between: (i) mouse and human exon-skipped genes, (ii) the total number of mouse and human independent, multi-exon genes.

C. Effect of Exon-skipping on Tissue-expression, Protein Reading Frames, and Protein Function in Mouse Multi-exon Gene Sets

(i) Determination of the tissue-expression pattern in mouse skipped exons

The tissue-expression pattern for mouse, skipped exons was assessed by examining the BLAST matches of each skipped exon to ESTs derived from different tissues. All tissue information for a given EST was retrieved from its corresponding NCBI EST record. Mouse skipped exons were extracted and isolated from the mouse exon-skipped genes. The skipped exons were searched against mouse dbEST (Appendix A, i.g) databases using BLASTN 2.2.1 ($P<1x10^{-40}$) to retrieve matching ESTs.

(ii) Effect of exon skipping on protein reading frames

Constitutively spliced (represent full-length transcripts) and exon-skipped transcripts were constructed in order to compare the protein reading frames between translated full-length and exon-skipped transcripts. Mouse constitutively spliced transcripts were constructed by concatenating all the predicted exons. Mouse exon-skipped transcripts were constructed by concatenating all the predicted exons except for the skipped exons. Mouse constructed transcripts were searched against a pooled mouse protein database (Appendix A. iii.a-c) respectively using BLASTX 2.2.1 ($P<1x10^{-04}$). Reading frames of the translated transcripts were obtained from BLASTX reports. If the reading frame of the constitutively spliced transcript was different to the reading frame of the exon-

skipped transcript, the difference was concluded to have resulted from a frame-shift mutation.

(iii) Effect of exon skipping on protein function

Protein functions were affected by exon-skipping events if the skipped exons were found to encode for functional protein domains or families. Mouse skipped exons were extracted from the total mouse multi-exon gene set. Mouse skipped exons were searched against a pooled mouse protein database respectively using BLASTX 2.2.1 ($P<1X10^{-04}$), to obtain the correct reading frames for the generation of translated skipped exons. The translated skipped exons were searched against the InterPro database using InterProScan for matches to protein domains and families.

RESULTS & DISCUSSIONS

Generation of the 72 Mouse Multi-exon Gene Set

An exon-skipping detection pipeline was constructed in order to satisfy the following objectives: (i) Obtaining multi-exon, orthologous mouse transcripts on chromosome 22, (ii) Construction of a mouse multi-exon gene set that includes transcript and genomic sequences, and (iii) Detection of mouse exon-skipping events in the multi-exon gene set. The preparation of transcript and genomic sequences that were fed into the detection process forms a major portion of the pipeline. The preparation process consisted of obtaining confirmed orthologous mouse transcripts, prediction of multi-exon transcripts, removal of paralogous transcripts, and reconstructions of transcript and genomic sequences (see Fig. 3).

Two hundred and sixty-nine orthologous mouse transcripts on human chromosome 22 were obtained (see supplementary table 1 in Appendix C). The lack of mouse transcript matches to mouse genomic contigs reduced the 269, mouse transcript data set by 58.4%. Mouse transcripts with matches to mouse genomic contigs that did not meet the multi-exon criteria, further reduced the dataset by 17.8% in addition to the 58.4% reduction. 72 mouse multi-exon orthologous transcripts resulted after removal of 20 paralogous transcripts (see supplementary table 2 in Appendix C). The average alignment coverage for each mouse multi-exon transcript in the 72 mouse multi-exon gene set when aligned to its genomic sequence was ~93.27%. Actual alignment coverages ranged from 50-100% coverage within the 72 multi-exon gene set.

Verification of Internal Mouse Exon Lengths

The primary goal in the prediction of multi-exon transcripts was to rapidly identify approximate locations of coding exons in genomic sequences by alignments of transcripts to their corresponding genomic sequences using BLAST. The accuracy of

the predictions of exon positions was assessed by comparing the observed exon lengths to known exon lengths for a given gene. Any major discrepancies between the observed and known exon lengths would suggest that the exon positions were not predicted correctly. Internal exon lengths (excluding first and last exons) were examined since internal exons mainly undergo exon skipping. The exon lengths in the first and last exons of a gene may vary with different prediction methods and these exon lengths usually depends on whether the 5' and 3' UTR regions are predicted. The accuracy in the prediction of mouse exon positions was examined by comparison to experimentally verified, internal exon lengths and Ensembl internal exon lengths.

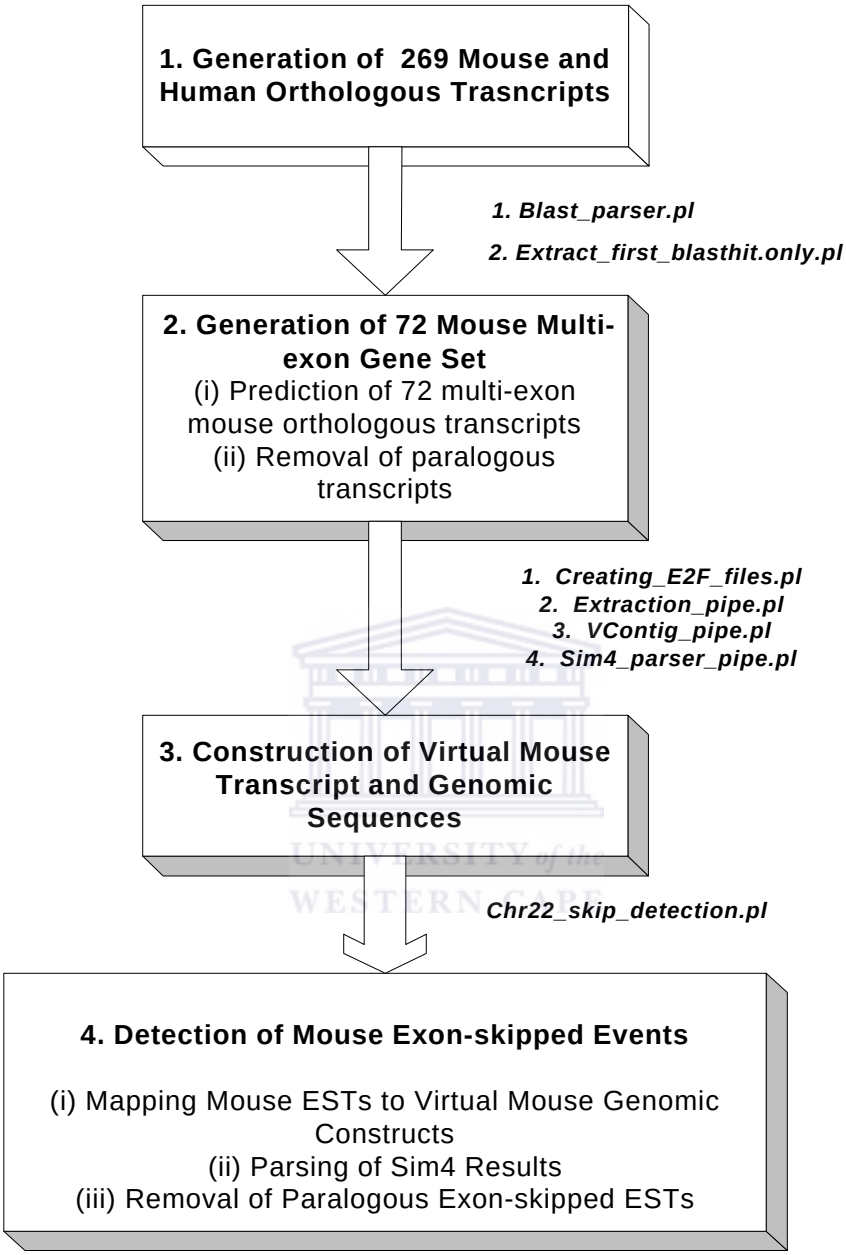


Figure 3: Flow chart of mouse exon-skipping detection pipeline

(i) Comparison to experimentally verified genes

Accurate predictions of exon positions were supported by the evidence of four experimentally verified genes found in literature: ALD (Kennedy *et al.*1996), GGT (Shi *et al.*1995), HO-1 (Alam *et al.*1994), and KBL (Edgar & Polak 2000), which have similar internal exon lengths to those predicted by our method. The experimentally verified internal exon lengths differed to the predicted internal exon lengths by a length of less than 6 nucleotides. The length discrepancy was expected due to our use of mouse exon prediction criteria (see Methods section Aii), which resulted in minimally, overlapping (less than 8 nucleotides) exon boundaries.

(ii) Comparison to Ensembl mouse internal exon lengths

Mouse annotated gene structures were initially unavailable during the release of mouse Ensembl genomic contigs (v.0.1.0) but have recently become available with the release of mouse Ensembl genome assembly version 1 (Jan 30th 2002). The current study utilizes only mouse transcript data to predict mouse exons whereas the Ensembl gene predictions utilized more data to confirm exons. Thus, Ensembl gene predictions were concluded to be more accurate due to their use of additional exon-supporting evidence such as proteins and ESTs. The accuracy of our exon predictions when compared to the Ensembl mouse exon predictions was investigated in multi-exon genes that have matching Ensembl transcript identifiers. The internal exon lengths of the following multi-exon genes: ARVCF, MFNG, GNB1L, and COMT were manually compared to the annotated, Ensembl mouse exon lengths. The maximum exon length difference that was observed for an internal exon belonging to the GNB1L gene was five nucleotides. The maximum exon length difference refers to the difference in the internal exon lengths between our predicted and Ensembl's internal exon lengths. The average maximum internal exon length difference among the four genes (ARVCF, MFNG, COMT, and GNB1L) was found to be four nucleotides. It is also important to note that not all the exons have an exon length difference between our predicted and Ensembl's

internal exon lengths. Due to the small maximum exon length difference, it can be concluded that our predicted internal exon lengths were similar to Ensembl's predicted internal exon lengths.

Identification of Novel Mouse Exon-skipping Events

Five novel mouse exon-skipped genes were identified within the 72 mouse multi-exon gene set (see table 1). The five mouse exon-skipped genes include the hypothetical 55.2 kDa protein, Bid (BH3 interacting domain death agonist), Gnb1l (guanine nucleotide binding protein beta polypeptide 1-like), Smarcb1 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1), and homolog to the hypothetical protein 384D8_6. The functions of the observed mouse exon-skipped genes are still unknown, however the putative functions of their human orthologous counterparts are available. The human BID protein oligomerizes with other apoptosis activators such as, BAK and BAX, to induce apoptosis (Grinberg *et al.*2002). The human SCO2 protein (orthologous to the hypothetical protein 384D8_6) is involved in the transfer of copper to cytochrome c oxidase, in which the transport of copper is crucial to the enzyme's activity (Salviati *et al.*2002). Mammalian cytochrome c oxidases are involved in the transfer of electrons from cytochrome c to molecular oxygen. The human GNB1L protein is involved in signal transduction, transcriptional regulation, and apoptosis (Li & Roberts 2001). The SMARCB1 protein has been proposed to bind to the HIV-1 integrase and promote the integration and targeting of viral DNA to active genes in the host chromosome (Kalpana *et al.*1994). The function of the human hypothetical 55.2 kDa protein is not well-characterized.

Mouse ESTs were aligned to their corresponding genomic sequences and the detection method is illustrated in Fig. 1. The exon-skipping detection method was based on a comparison between known (exon positions obtained from annotated mouse genomic sequences) and observed exon positions (exon positions obtained from alignments of

ESTs to annotated genomic sequences). Missing exon positions were interpreted as exon-skipping events.

(i) Detection of known human exon-skipping events using the current detection method

Eight exon-skipping events in the following human genes: RBX1, GTPBP1, SLC25A17, and GSTT1, which were previously detected by Hide *et al.*2001, were also confirmed by our detection method. The detection of exon-skipping events in the SLC25A17 gene is illustrated in supplementary figures 1 and 2 in appendix C. The results for the rest of the genes that were tested are not shown.

TABLE 1: Identification of *Mus* musculus chromosome 22 orthologous genes with exon-skipped ESTs

Exon-numbering were based on mouse transcript positions with respect to its mouse ENSEMBL genomic contig (see supplementary table 2 in Appendix C). Mouse gene descriptions were obtained from either ENSEMBL or Genbank gene-annotations. The following abbreviations used in column 4 are described as follows: (f/s) frame-shifts have occurred, and (*d*) functional domains were affected. The following symbols are described as follows: (†) Exon-skips detected by j_explorer.pl using altered parameters (150bp, $P<1x10^{-40}$), (‡**)** Exon-skips detected by j_explorer.pl using default parameters (50bp, $P<1x10^{-40}$), (^{del}) Deletion events have occurred in both of the exons flanking the skipped exon, and (^{b}) Gene descriptions obtained from FANTOM annotations (http://fantom.gsc.riken.go.jp/). All mouse Ensembl identifiers are liable to change as new sequence data becomes available. However, mouse Ensembl identifiers used in the study were valid at the time of the release (May 2001 release). See supplementary table 3 in Appendix C for mouse exon-skipped EST annotations.

Mouse Gene description	Mouse RNA or cDNA identifiers (Ensembl transcript and Genbank identifiers)	Skipped exon(s)	Effect of exon-skip	No. of ESTs confirming the skip	Human Locus Name (gene description)
Hypothetical 55.2 kDa protein	ENSMUST00000001834	3-8^{del}	**f/s d**	1	dJ149A16.C22.6 (Similar to Wp:CE09424)
BH3 interacting domain death agonist	BC002031.1	4-5^{†,del}	-	1	BID (BH3 interacting domain death agonist)
Guanine nucleotide binding protein (G protein), beta polypeptide 1-like (Gnb1l)	NM_023120.1	2-6†	**f/s d**	2	GNB1L (Homo sapiens G-protein beta subunit-like protein)
SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1 (Smarcb1)	NM_011418.1	2-4$^{\dagger, del}$	**d**	1	Smarcb1 (SWISNF related matrix associated actin dependent regulator of chromatin subfamily b member 1)
Homolog to HYPOTHETICAL PROTEIN 384D8_6^{b}	AK002487	3-4^{del}	-	1	SCO2 (cytochrome oxidase deficient yeast homolog 2)

(ii) Deletion events occurring in mouse exon-skipping events

Manual examinations of EST to genome alignments for mouse exon-skipped genes revealed that deletion events also exist in addition to exon-skipping events. In human exon-skipped genes, deletion events have not been observed by Hide *et al.*2001. All mouse deletion events were found to occur in both of the exons flanking the skipped exon. Deletion events were detected as having EST matches to partial exons. The length of a deletion for an exon ranged from ~14-300 nucleotides. Four exon-skipped genes have deletion events (see table 1) occurring in both of the exons flanking the exon-skipping event.

In exon-skipped genes that display deletion events, the predicted exon boundaries flanking the skipped exons were also confirmed by mouse ESTs. Although the exon boundaries flanking the skipped exons have been confirmed by ESTs, only one EST was observed to confirm the deletion events for each exon-skipped gene. It can be concluded that the deletion events may be possible alternative donor or acceptor splice-sites, but more evidence and experimental verification is required to confirm the deletion events are genuine.

(iii) Exon-skipping events observed in the region orthologous to human 22q11 region

The human chromosome 22qll region (22q11) is known to undergo translocations, duplications, and deletions leading to human disorders such as velocardiofacial syndromes, DiGeorge syndromes, and tumor diseases (McDermid & Morrow 2002).

The majority (4/5) of the mouse exon-skipping events were found to be orthologous to the 22q11 region suggesting that these exon-skipping events may be implicated in diseases associated within this region. Mouse exon-skipped genes such as, BID and GNB1L, are known to be associated with DiGeorge syndromes in human genes (Gong

*et al.*2000; Funke *et al.*2001), but it is uncertain whether the mouse exon-skipping events detected for these genes are responsible for this disease.

44% (23/52) of the human exon-skipping events observed by Hide *et al.*2001 have also been found to occur in the 22q11 region suggesting that these exon-skipping events may occur quite frequently in this region. Four mouse exon-skipping events were found to be orthologous to the 22q11 region (4/5=80%). Mouse genes that were orthologous to the 22q11 region were located on mouse chromosomes 6,10,16, and 5. The majority of mouse and human exon-skipping events appears to occur quite frequently within the 22q11 region. The frequent occurrence of mouse and human exon-skipping events observed in the 22q11 region may implicate these exon-skipping events in diseases with medical importance or these exon-skipping events may be prone to rearrangements, deletions, and mutations. Although, the frequent occurrence of exon-skipping events in the 22q11 region appears to suggest that the exon-skipping event may be aberrant, the biological validity of whether the exon-skipping event is aberrant or functional remains to be tested in future studies.

(iv) The effect of mouse exon-skipping events on protein function

The effect of each exon-skipped event in protein function and tissue expression was assessed. In both of the exon-skipped genes (GNB1L, SMARCB1), the skipped exons were found to encode for protein domains that were crucial to the primary functioning of the proteins encoded by the gene. Although the protein domain of the hypothetical 55.2 kDa protein was affected by the exon-skipping event, the biological effects on protein function cannot be assessed since the molecular function and the protein domain matching to this protein have not been well characterized.

(a) GNB1L gene

The human GNB1L (guanine nucleotide binding protein (G protein), beta polypeptide 1-like) protein belongs to a family of WD-repeat proteins, in which each member possesses four or more repeating units containing ~40 amino acids ending with tryptophan (W) and aspartic acid (D). The guanine nucleotide binding proteins are known to be involved in signal transduction, transcriptional regulation, and apoptosis (Gong et al. 2000). The skipped exons in the mouse GNB1L gene, encode for protein domains implicated in defense responses and electron transport. The lack of these domains caused by the exon-skipping event may be crucial to the normal functioning (signal transduction, apoptosis, transcriptional regulation) of the GNB1L protein mentioned previously.

(b) SMARCB1 gene

The function of the human SMARCB1 protein has been suggested to bind to the HIV-1 integrase (known to promote the insertion of viral DNA into host chromosomes) and promote the integration and targeting of viral DNA to active genes in the host chromosome (Kalpana *et al.*1994). The skipped exons observed for this gene encodes for the VHS domain, which is involved in the transport of proteins either within or to the outside of the cell. Thus, the absence of this domain may affect the positioning of viral DNA to their destinations in the host DNA.

(v) The occurrence of frame-shift mutations resulting from mouse exon-skipping events

40% (2/5) of the mouse-exon skipping events have resulted in possible frame-shift mutations, which were observed in both the hypothetical 55.2kDa protein and the Gnb1l genes. (see table 1). Frame-shift occurrences in these events may suggest that the protein function in these genes may be destroyed. All frame-shift mutations were also

accompanied by a loss in protein function (inferred from skipped exons encoding for functional protein domains), which further suggests that the exon-skipping events may severely affect the protein function or result in the production of a non-functional protein. Although frame-shift mutations were not observed in the SMARCB1 protein, a functional protein may still be produced, but the protein may have lost some of its functional properties.

(vi) Ubiquitous tissue expression displayed by mouse skipped exons

All mouse, skipped exons were observed to have matches to different EST libraries from different tissues. The observed ubiquitous expression suggests that the skipped exons may not be expressed exclusively in a specific tissue. Perhaps these skipped exons may be expressed in a specific time-period or developmental stage and could not be detected using the current method in the study.

Orthologous gene-pairs that display mutually exclusive mouse and human exon-skipping events

Out of the 20 gene-pairs that displayed either mouse or human exon-skipping events (see table 2), 17 gene-pairs consisted of only human exon-skipping events and the rest displayed only mouse exon-skipping events. Two mouse exon-skipped genes (GNB1L, SCO2) were not compared to human genes since these genes were discarded from the human multi-exon gene set. Pseudogenes, non-multi-exon genes, and some genes that were prone to re-arrangements were discarded during the construction of the human multi-exon gene set.

In all the gene-pairs (BID, SMARCB1, and hypothetical 55.2 kDa protein) that display exclusively mouse exon-skipping events, the EST coverage and EST representation (see Methods Bii and Biii for definition of these terms) were lower in mouse than in the

human counterpart. Therefore, the lack of human exon-skipping events in these genes could not be due to insufficient human EST coverage and EST representation. The

TABLE 2: Exon-junction comparative study between mouse and human multi-exon orthologues that display exon-skipping events

Twenty multi-exon mouse and human transcript orthologues that display exon skipping were compared. Three orthologues have only mouse exon-skipping events and 17 orthologues display only human exon-skipping events. The EST coverage and representation were not compared in the two gene-pairs (GNB1L, SCO2) displaying only mouse exon-skipping events because these genes were not found in the human multi-exon gene set. Previous human exon-skipping events were sampled using Genbank 119 human EST sequences (Hide *et al.*2001). EST representation and consecutive exon-junction coverage were compared as presented in the following table. The following abbreviations are described as follows: (M) only mouse exon-skipping events were observed and (H) only human exon-skipping events were observed.

Human Ensembl Locus Name
Mouse(M)/Human(H) Sequence Identifiers
Predicted Total Exon Number
Number of EST hits to consecutive and non-consecutive exon-junctions using j_explorer.pl (default Parameters)
Consecutive
exon-junction coverage (%)

(M)ENSMUST00000001834
12
105
82

1. dJ149A16.6[M]
(H)dJ149A16.C22.6

12

414

100

(M)BC002031.1

6

8

50

2. BID[M]

(H)AC006285.C22.1

5

67

100

(M)NM_011418.1

5

1

20

3. SMARCB1[M]

(H)AP000349.C22.2

9

226
100

(M)NM_016714.1
7
57
100

4. NUP50[H]
(H)bK217C2.C22.1
5
41
75

(M)NM_016915.1
14
22
69.2

5. PLA2G6[H]
(H)bK228A9.C22.1
17
42

93.8

(M)BC010983
4
11
33.3

6. MFNG[H]
(H)bK390B3.1.1
8
200
100

(M)AK011196
7
108
66.7

7. RAYL[H]
(H)cE132D12.C22.1
7
76
83.3

(M)NM_007968.1

6

237

0

8. EWSR1[H]

(H)bK984G1.C22.4

17

544

93.8

(M)NM_013762.1

9

500

100

9. RPL3[H]

(H)dJ333H23.1.1

10

1002

100

(M)BC003843

3

134

0

10. ST13[H]

(H)dJ408N23.C22.1

12

337

100

(M)NM_013847.1

9

68

75

11. GCAT[H]

(H)dJ466N1.C22.2

9

60

100

(M)ENSMUST00000005549

5

4

75

12. HMG2L1[H]

(H)dJ510H16.C22.2

12

29

90.9

(M)BC006630

10

64

77.8

13. UFD1L[H]

(H)AC000068.C22.2

12

422

100

(M)NM_013711.1

11

51

60

14. TR^{H}

(H)AC000078.C22.1

18

83

88.2

(M)NM_010048.1

10

82

88.9

15. $DGCR2^{H}$

(H)AC000095.C22.1

10

33

100

(M)BC005759.1

13

69

58.3

16. SEC14L2[H]

(H)AC004832.C22.6

12

40

100

(M)BC005800

6

55

40

17. AC005500.4[H]

(H)AC005500.C22.4

6

102

100

(M)ENSMUST0000000334

17

42

50

18. ARVCF[H]

(H)AC005663.C22.1

20

29

73.7

(M)BC003421

9

279

87.5

19. ATP6E[H]

(H)AC006285.C22.6

9

384

100

(M)BC012254.1

5

68

50

20. GSTT1[H]

(H)AP000351.C22.10

5

98

100

majority (15/17) of the gene-pairs that display only human exon-skipping events, the EST coverage was higher in human than its mouse counterpart. In these cases, the lack of mouse exon-skipping events in these gene-pairs may have been due to the low mouse EST coverage. 45% (9/20) of the twenty orthologous gene-pairs that either display exclusively human or mouse exon-skipping events, were observed to be orthologous to the human 22q11 region, which may suggest that the exon order may be different between these gene-pairs. Therefore, conserved exon-skipping events were not observed for mouse and human orthologous genes, which may suggest that the exon order may be different between mouse and human genes. The extent of conserved exon order between mouse and human orthologous genes on chromosome 22 is currently unknown and maybe necessary to detect conserved exon-skipping events.

(i) High transcript identity observed between orthologous mouse and human transcripts

The average transcript identity between mouse and human transcripts in the 72 mouse multi-exon gene set was ~87%. The average transcript identity between mouse and human transcripts in mouse exon-skipped genes was calculated to be ~87%. The transcript identities in total and exon-skipped genes appear to be slightly higher than the established coding sequence identity value (~85%) between mouse and human transcripts (Makalowski & Boguski 1998). Mouse and human transcripts in the 72 multi-exon gene set may include 3' and 5' untranslated regions, which are known to have lower sequence identities (~17%) than coding sequence regions between mouse and human transcripts (Makalowski & Boguski 1998). Therefore, actual coding sequence identities predicted by the current study may be higher than the established value. However, the extent of 5' and 3' UTR regions in the 72 multi-exon gene set was not determined. Perhaps the obtained average transcript identity (~87%) may be too low for the detection of conserved exon-skipping events in mouse and human orthologous genes.

Limitations of j_explorer.pl in the detection of mouse exon-skipping events

J_explorer.pl, is a perl script, that constructs short, exon fragments (or exon junctions) between consecutive and non-consecutive exons (see Fig. 2). These short fragments are used to retrieve potential exon-skipped ESTs whereas the detection approach utilized in the current study does not construct exon junctions. Instead the current approach uses virtual transcripts, which represents all the predicted exons joined together (see Methods Aiii for details on construction of virtual sequences), to retrieve exon-skipped ESTs. Both approaches generate alignments between ESTs and genomic sequences using sim4. In order to assess the ability of our approach to effectively detect exon-skipping events, j_explorer.pl was utilized to detect mouse exon-skipping events.

When using default parameters in the mouse exon-skipped gene set, j_explorer.pl failed to detect all of the mouse exon-skipping events (see table 3). However, the use of altered parameters (150bp exon-junctions were constructed; $P<1x10^{-40}$) improved the detection of exon-skipping events in mouse by 60%. Figure 4 illustrates how the use of default and altered parameters can affect the ability to detect mouse exon-skipping events. Exon-skipped ESTs were detected if the non-deleted portions of exons flanking the exon-skipping event were included in the construction of non-consecutive exon-junctions. J_explorer.pl was unable to detect all mouse exon-skipping events using default parameters due to the exclusion of non-deleted portions of the exons flanking the exon-skipping event during the construction of non-consecutive exon-junctions. Thus, the use of altered parameters would incorporate the non-deleted portions of the flanking exons into the construction of non-consecutive exon junctions.

TABLE 3: Mouse and human statistics on factors that influence the detection of exon skipping on human chromosome 22

The following symbols are described as follows: (*) Calculations are detailed in Appendix B.ii. The values in brackets are representative of only exon-skipped genes whereas the values outside brackets are inclusive of both skipped and non-skipped genes unless otherwise specified. ([a]) Default parameters include: (1.) 50bp were extracted from both consecutive and non-consecutive exon-junctions, (2.) $P<1x10^{-40}$, ([b]) Altered parameters include: (1.) 150bp were extracted from both consecutive and non-consecutive exon-junctions, (2.) $P<1x10^{-40}$.

Factors	Mouse	Human
Total number of multi-exon genes sampled	72	347
Total number of exon-skipped genes	5	52
Total number of exon-skipping events	5	62
Rate of skipping per exon-skipped gene	1.00	1.19
Average number of exons per gene (skipped gene)	8.86(6.67)	8.20(12.9)
Average consecutive exon-junction coverage per gene (skipped gene)	0.53(0.59)	0.82(0.89)
Average number of EST matches to exon junctions per gene (skipped gene)	74.5(151.2)	68.5(166.5)
Probability that a given multi-exon mRNA has at least one non-consecutive exon-junction[a] (altered parameters[b])	0(0.0025)*	0.051[a]
Percentage of exon-skipped events predicted by j_explorer.pl with default parameters[a] (altered parameters[b])	16.6(66.6)	100[a]

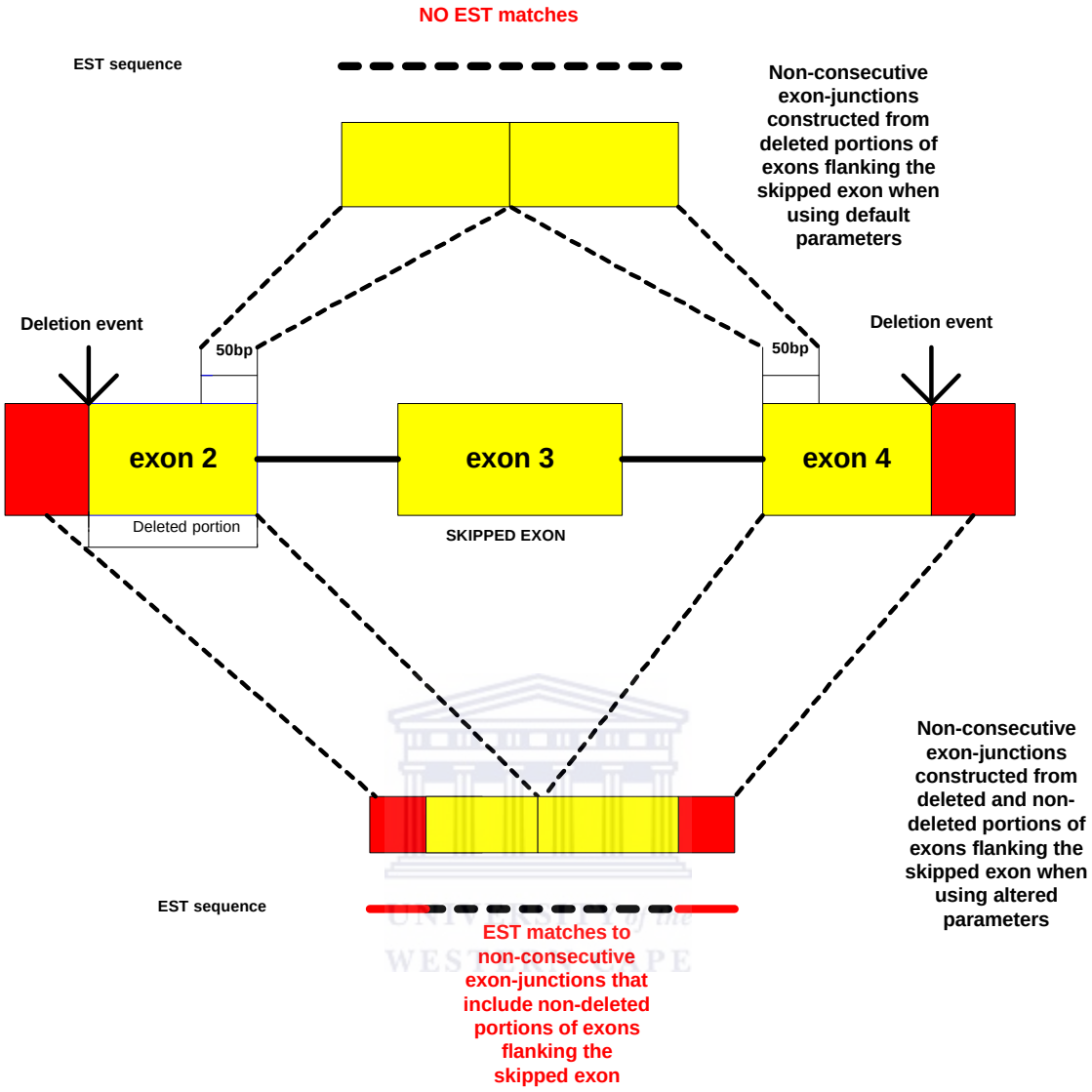


Figure 4: The effect of altered and default parameters on the detection of mouse exon-skipped events

Two mouse exon-skipping events (see Fig.2) that were not detected by j_explorer.pl under altered parameters were due to inherent limitations in j_explorer.pl. These limitations include: (i) the inability to detect exon-skipping events that have more than five exons skipped, (ii) inability to detect exon-skipped ESTs if the EST matches did not meet the cut-off parameters of j_explorer.pl. The limitations in j_explorer.pl may have accounted for the low probability (0.25%) that a multi-exon transcript would have at least one non-consecutive exon-junction confirmed by an EST (see table 1).

Factors contributing towards the low detection of mouse exon-skipping events

(A) Under-predictions of mouse exons when compared to Ensembl mouse exon predictions

46% (33/72) of the 72 mouse multi-exon gene set have annotated gene structures, which were found at the mouse Ensembl website (http://www.ensembl.org/Mus_musculus/). Within the 33 multi-exon genes (see supplementary table 2 in Appendix C) that have Ensembl gene annotations, 76% (25/33) of the genes were found to have matching Ensembl transcript identifiers. 24% (8/33) of the genes have non-matching Ensembl transcript identifiers. Non-matching Ensembl transcript identifiers may arise due to the usage of different transcript release versions and does not suggest that our exon prediction was inaccurate. The exon-skipped gene, BID, was found to have a non-matching Ensembl transcript identifier. However, further analyses have revealed that the transcript (BC002031) utilized in the current study was longer than the Ensembl transcript representative and both transcripts were found to encode for the same gene.

Within the 25 multi-exon genes that have matching Ensembl transcript identifiers: (i) 28% (7/25) have matching Ensembl exon annotations, (ii) 24% (6/25) have over-predicted exons (defined as the number of predicted exons in the current study was greater than the number of Ensembl predicted exons), and (iii) 48% (12/25) have under-

predicted exons (defined as the number of predicted exons in the current study was less than the number of Ensembl predicted exons). Over-predictions of exons may result in falsely predicted exons, which may lead to the prediction of false positives (inability to distinguish a false exon-skipping event from a real exon-skipping event). Under-predictions of exons may result in a large number of false negatives (inability to detect an exon-skipping event). Under-predictions of exons may also arise from the insufficient use of other available evidence such as proteins and ESTs to confirm mouse exons. Over- and under-predictions of exons in all exon-skipped genes were not observed to have affected any of the exon-skipping events.

(B) Low data Availability

(i) Short transcript length

Partial transcripts may contribute towards the low detection of exon-skipping events. The average mouse transcript length in the 72 multi-exon gene set was ~869.7 nucleotides. In mouse exon-skipped genes, the average mouse transcript length was calculated to be ~732 nucleotides. The calculated average transcript length was 2-3 fold lower when compared to the average mouse transcript length (~1976 nucleotides) calculated by Makalowski & Boguski 1998 in a data set of 2,820 mouse transcript sequences, which are represented by the longest available sequences. The low transcript lengths suggest that many partial transcripts may have been used to predict mouse exons.

(ii) Unassembled, incomplete genomic contigs

Mouse genomic contigs (Ensembl version 0.1.0) that were utilized represented ~10% of the finished, draft mouse genomic sequences. A large number of mouse transcripts (112/269) lacked matches to mouse genomic contigs and may have greatly contributed towards the low detection of mouse exon-skipping events. Individual, unassembled

mouse contigs may not have encoded for the full-length gene and may have lacked exons flanking the exon-skipping event. The extent of full-length genes encoded by mouse contigs was unknown.

(C) Possibility of utilizing unknown variant transcripts

A heterogeneous pool of mouse transcript data was used for the prediction of mouse exons. Utilizing different transcript sources may increase the chances of detecting more exon-skipping events. However, increasing the transcript sample size may have increased the number of possible variant transcripts matching to genomic contigs. The extent of variant transcripts was unknown in the utilized transcript sources. Our detection method depends on comparing a set of assumed constitutive exons to observed alternatively spliced exons and searching for discrepancies between these exon sets. The discrepancies were then treated as putative alternatively spliced events. Utilizing variant transcripts to predict exons may lead to many false negatives (inability to predict an alternatively spliced event). Ineffectively, one would be comparing a set of alternatively spliced exons (assumed to be constitutive) to observed, alternatively spliced exons, thus no discrepancies would be noted.

(D) Poor consecutive exon-junction coverage

EST representation and exon-junction coverage maybe important factors in influencing the detection of exon-skipping events. A comparison of these factors between mouse and human independent, total multi-exon gene sets were examined (see table 1). The number of EST matches to exon-junctions (inclusive of consecutive and non-consecutive exon-junctions) was similar between mouse and human total multi-exon gene sets (average of 72 ESTs) and exon-skipped genes (average of 158 ESTs). However, the average consecutive exon-junction coverage in mouse was ~29% lower than human in both the total and in exon-skipped gene sets. The lower exon-junction coverage in mouse exon-skipped genes may have been due to the lack of EST

representatives in the public database (in cases where the number of exons in mouse and human orthologous genes are similar or identical) or under-predictions of mouse exons.

(E) Mouse exon-skipped events were captured by a small number of ESTs

In a previous study performed by Hide and others, human exon-skipping events were captured by a small number of ESTs (highest number of ESTs confirming an exon-skipping event is 11), which is similar to the current study (highest number of ESTs confirming an exon-skipping event was 2). The observation that mouse and human exon-skipping events were captured by a small number of exon-skipped ESTs in the current study may be due to several factors such as: (i) lack of EST isoforms in the current database, (ii) the exon-skipped transcript may not be highly expressed, and (iii) the low abundance of exon-skipped transcript isoforms when compared to the high level of full-length transcripts may be due to an RT-PCR (reverse-transcriptase polymerase chain reaction) artifact. The reduction of the mRNA abundance of the exon-deleted mRNAs in the hamster Hprt gene has been shown to result from an RT-PCR artifact. During the detection of exon-deleted mRNAs using RT-PCR, full-length mRNAs may suppress the amplification of exon-deleted mRNAs by 40-fold either through competition for reagents and primers or rehybridization with exon-deleted mRNAs (Valentine 1998). These factors are not easily manipulatable, but may explain the difficulty in obtaining a large number of mouse and human exon-skipped ESTs.

CONCLUSIONS

A transcript (ESTs and mRNAs) and genome-verified approach towards the detection of mouse exon-skipping events on chromosome 22 has resulted in a low number (5/72) of mouse exon-skipping events. The majority (4/5) of the mouse exon-skipping events were observed to be orthologous to the 22q11 region suggesting that these exon-skipping events may be implicated in the CATCH 22 syndromes (velocardiofacial and DiGeorge syndromes), but remains to be confirmed. The protein function of the two exon-skipped genes (GNB1L, SMARCB1) may have been affected since the protein domains encoded by these genes are implicated in primary biological processes such as electron transport and transportation of proteins to intracellular and extracellur destinations. The protein function of GNB1L proteins may even be more severely affected as frame-shift mutations was also observed for this gene. Ubiquitous tissue expression was observed for all mouse, skipped exons, which suggests that these exon-skipping events may not have the tissue-specific characteristics that are commonly observed in the majority of the alternatively spliced events. Examples of alternatively spliced events displaying tissue-specific isoforms are observed in the *Drosophila* SXL and TRA proteins (MacDougall *et al.*1995), the transcriptional factor Oct-1 (Pankratova *et al.*2001), the Fas apoptosis inhibitory molecule (Zhong *et al.*2001), human LIM-only protein (Ng *et al.*2001), and many more examples are known in literature. However, the possibility of temporally-specific or developmentally-specific splicing may exist within these exon-skipping events, but could not be detected with the current method.

Mouse deletion events were also observed in addition to the exon-skipping events in all mouse exon-skipped genes for the exception of the Gnb1l gene. The deletion events appear to be alternative donor and acceptor splice-sites since, the predicted exon boundaries flanking the skipped exons were also confirmed by mouse ESTs. However, the observed deletion events for each exon-skipped gene were only confirmed by one EST thus, more EST evidence or experimental verification is necessary to confirm that

the deletion events are genuine. The biological significance of deletion events occurring together with exon-skipping events remains to be explored further.

It was interesting to observe that mouse and human exon-skipping events were never simultaneously observed for any of the twenty orthologous gene-pairs that displayed exon-skipping events. It is possible that our dataset may have been biased towards mouse and human orthologous genes with different exon-orders, but this remains to be explored further. The lack of conserved mouse and human exon-skipping events indicates that our approach was not successful in detecting conserved exon-skipping events. The extent of exon-order in all known mouse regions orthologous to chromosome 22 is still unknown and has yet to be determined.

Likely factors that may contribute towards the low detection of mouse exon-skipping events may have resulted from the under-predictions of mouse exons and the lack of mouse transcript matches to genomic sequences. The under-predictions of mouse exons may have resulted in nearly half (12/25) of the mouse multi-exon genes that have matching Ensembl transcript identifiers having under-predicted exons. Similarly, the lack of mouse transcript matches to genomic sequences has also resulted in a large number (112/269) of mouse transcripts lacking matches to mouse genomic contigs. These factors may also be capable of generating many false negatives. Our predictions of mouse exons mainly differed to the Ensembl mouse exon predictions through the insufficient use of other available evidence such as proteins and ESTs. The mouse genomic sequences utilized in the study were in an unannotated state, therefore a rapid approach towards the identification of mouse exons was necessary in order to accomplish the current goal. Perhaps, the utilization of mouse transcripts (cDNAs and mRNAs) to predict mouse exons may not be an adequate approach towards the detection of all possible mouse exons. The current number of annotated mouse genes (known and unknown) orthologous to chromosome 22 in the mouse genome assembly v1 (30th January 2002) is 87. This number does not differ dramatically from our total number (72) of multi-exon genes. It is uncertain how many multi-exon genes are

represented within the 87 mouse orthologous genes. It is evident that even with the current number of possible Ensembl mouse orthologues, this number may still be insufficient for the detection of more mouse exon-skipping events. Perhaps as the mouse genomic data increases, it may become more feasible to detect more mouse exon-skipping events on chromosome 22. It appears that our low frequency (5/72) of mouse exon skipping may represent the maximum number of exon-skipping events that can be detected even with the current mouse genomic data.

Other factors that may result in a low number of mouse exon-skipping events include: (i) the use of partial transcripts, (ii) possible use of unknown variant transcripts, and (iii) the possibility that only a small number of exon-skipped ESTs may exist due to RT-PCR artifacts and lowly expressed genes. These factors are not easily manipulatable when compared to the previously mentioned factors and may require a great deal of time and effort. Obtaining full-length transcripts may require large-scale sequencing and thorough predictions of gene-structures. Another alternative is to use consensus sequences, however updated mouse consensus sequences were not available during our study. New data-mining techniques or identification of protein isoforms that match variant transcripts may be required to identify unknown variant transcripts. The observation that a few mouse exon-skipped ESTs were observed for a given mouse exon-skipped gene also requires further experimental validation in order to determine whether this observation is biological or as a result of an RT-PCR artifact.

A recent study performed by Brett *et al.*2002 has revealed the frequency of alternative splicing between mouse and human transcript (mRNAs and ESTs) data sets to be comparable. The claim provided by Brett and others is doubtful since their detection method is based on aligning mouse mRNAs with ESTs and genomic sequences have not been utilized when compared to the current detection method. It may be quite possible that the presence of pseudogenes, duplications, and paralogues are present in their data sets since genomic sequences have not been utilized in their detection. The alternative splicing estimate obtained by Brett and others is vague and does not

distinguish between the different types of alternative splicing. It may be quite possible that the different types of alternative splicing may differ between species, but this remains to be explored further. Brett and others' use of random data sets to determine alternative splicing between mouse and human genes appears to be weak since these random data sets may be biased towards specific genes that may exhibit frequent alternative splicing. However, the level of the occurrence of alternative splicing in genes that perform different functions remains to be investigated. Although, the mouse exon-skipping frequency obtained in this study cannot be extrapolated to represent all the mouse multi-exon genes in the mouse genome due to our small sample size, the mouse exon-skipping frequency (5/72) obtained in this study appears to be lower than the human exon-skipping frequency (52/347) obtained by Hide *et al.*2001. However, our conclusion remains to be tested on a larger mouse multi-exon dataset. The claim made by Brett and others also contradicts with our conclusion and a genome-verified approach may be necessary to support their claim.

In conclusion, many questions in the field of alternative splicing still persist. The current work has provided the basis in which further interesting hypotheses with regards to mouse alternative splicing can be tested. Successful detection of mouse exon-skipping events requires the use of well-annotated mouse genomic sequences and full-length mouse transcripts. As the quality and the quantity of mouse genomic and transcript data improves, the true exon-skipping frequency in mouse genes can thus be quantified.

REFERENCES

1. Alam, J., Cai, J., and Smith, A. (1994) Isolation and characterization of the mouse heme oxygenase-1 gene. *J. Biol. Chem.* **269:** 1001-1009.
2. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215:** 403-410.
   and Iinuma, K. (2001) A novel intronic mutation of the *TAZ* (G4.5) gene in a patient with Barth syndrome: creation of a 5' splice donor site with variant GC consensus and elongation of the upstream exon. *Hum. Genet.* **109:** 559-563.
3. Brett, D., Pospisil, H., Valcárcel, J., Reich, J., and Bork, P. (2002) Alternative splicing and genome complexity. *Nat. Genet.* **30:** 29-30.
4. Chabot, B. (1996) Directing alternative splicing: cast and scenarios. *Trends Genet.* **12:** 472-478.
5. Liu, T.T., Chang, Y.H., Chiang, S.H., Yang, Y.L., Yu, W.M., and Hsiao, K.J. (2001) Identification of three novel 6-pyruvoyl-tetrahydropterin synthase gene mutations (226C>T, IVS3+1G>A, 116-119delTGTT) in Chinese hyperphenylalaninemia caused by tetrahydrobiopterin synthesis deficiency. *Hum. Mutat*. **18**:83.
6. Clark, F. & Thanaraj, T.A. (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Unpublished paper*.
7. Cooper, T.A. & Mattox, W. (1997) GENE REGULATION '97 The Regulation of Splice-Site Selection, and Its Role in Human Disease. *Am. J. Hum. Genet.* **61:** 259-266.
8. Croft,L., Schandorff, S., Clark, F., Burrage, K., Arctander, P., and Mattick, J.S.(2000) ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome [letter]. *Nat. Genet*. **24:** 340-341.
9. D'Souza, I., Poorkaj, P., Hong, M., Nochlin, D., Lee, V. M-Y., Bird, T.D., and Schellenberg, G.D. (1999) Missense and silent tau gene mutations cause frontotemporal dementia with parkinsonism-chromosome 17 type, by affecting

multiple alternative RNA splicing regulatory elements. *Proc. Natl. Acad. Sci. USA* **96:** 5598-5603.

10. Edgar, A.J. and Polak, J.M. (2000) Molecular cloning of the human and murine 2-amino-3-ketobutyrate coenzyme A ligase cDNAs. *Eur. J. Biochem.* **267:** 1805-1812.
11. Florea, L., Hartzell, G., Zhang, Z., Rubin, G., and Miller, W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8:** 967-974.
12. Funke, B., Pandita, R.K., Morrow, B.E. (2001) Isolation and characterization of a novel gene containing WD40 repeats from the region deleted in velo-cardiofacial/DiGeorge syndrome on chromosome 22q11. *Genomics.* **73:** 264-71.
13. Gong, L., Liu, M., Jen, J., Yeh, E.T. (2000) GNB1L, a gene deleted in the critical region fo DiGeorge syndrome on 22q11, encodes a G-protein beta-subunit-like polypeptide. *Biochim. Biophys. Acta.* **1494:** 185-8.
14. Grinberg. M., Sarig, R., Zaltsman, Y., Frumkin, D., Grammatikakis, N., Reuveny, E., Gross, A.(2002) tBID Homololigomerizes in the mitochondrial membrane to induce apoptosis. *J. Biol. Chem.* **277:** 12237-45
15. Hendrich, B., Abbott, C., McQueen, H., Chambers, D., Cross, S., and Bird,A. (1999) Genomic structure and chromosomal mapping of the murine and human Mbd1, Mbd2, Mbd3, and Mbd4 genes. *Mamm. Genome.* **10:** 906-912.
16. Hide, W.A., Babenko, V.N., van Heudsen, P.A., Seoighe, C., and Kelso, J.F. (2001) The contribution of exon skipping events on Chromosome 22 to protein coding diversity. *Genome Res.* **11**:1848-53.
17. Hutchinson, S., Wordsworth, P., Handford, P.A. (2001) Marfan syndrome caused by a mutation in *FBN1* that gives rise to cryptic splicing and a 33 nucleotide insertion in the coding sequence. *Hum. Genet.* **109:** 416-420.
18. Kalbfuss, B., Mabon, S.A., Misteli, T. (2001) Correction of Alternative Splicing of Tau in Frontotemporal Dementia and Parkinsonism Linked to Chromosome 17. *J. Biol. Chem.* **276:** 42986-42993.

19. Kalpana, G., Marmon, S., Wang, W., Crabtree, G.R., Goff, S.P. (1994) Binding and stimulation of HIV-1 integrase by a human homolog of yeast transcription factor SNF5. *Science.* **266:** 2002-6.
20. Kan, Z., Gish, W., Rouchka, E., Glasscock, J., and States, D. (2000) UTR Reconstruction and Analysis Using Genomically Aligned EST Sequences. *Intell. Syst. Mol. Biol.* **8:** 218-227.
21. Kan, Z., Rouchka, E.C., Gish, W.R., and States, D.J. (2001) Gene Structure Prediction and Alternative Splicing Analysis Using Genomically Aligned ESTs. *Genome Res.* **11:** 889-900.
22. Kennedy, M.A., Rowland, S.A., Miller, A.L., Morris, C.M., Neville,L.A., Dodd, A., Fifield, W.J., and Love, D.R. (1996) Structure and location of the murine adrenoleukodystrophy gene. *Genomics.* **32:** 395-400.
23. Kumar, S. & Hedges, S.B. (1998) A Molecular Timescale for Vertebrate Evolution. *Nature.* **392:** 917-920.
24. MacDougall, C., Harbison, D., and Bownes, M. (1995) The Developmental Consequences of Alternate Splicing in Sex Determination and Differentiation in *Drosophila. Dev. Biol.* **172:** 353-376.
25. Makalowski,W., Boguski, M. (1998) Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. USA.* **95:** 9407-9412.
26. Martindale, D.W., Wilson M.D., Wang, D., Burke, R.D., Chen, X., Duronio, V., Koop, B.F. (2000) Comparative Genomic Sequence Analysis of the Williams Syndrome Region (LIMK1-RFC2) of Human Chromosome 7q11.23. *Mamm. Genome.* **11:**890-898.
27. McDermid, H.E. & Morrow, B.E. (2002) Genomic disorders on 22q11. *Am. J. Hum. Genet.* **70:** 1077-1088.
28. Mironov, A.A., Fickett, J.W., and Gelfand, M.S. (1999) Frequent Alternative Splicing of Human Genes. *Genome Res.* **9:** 1288-1293.
29. Modrek, B. & Lee, C. (2002) A genomic view of alternative splicing. *Nat. Genet.* **30:**13-19.

30. Modrek, B., Resch,A., Grasso, C., and Lee, C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. **29**(13): 2850-2859.
31. Ng, E., Lee, S., Li, H., Ngai, S., Tsui, S., Waye, M., Lee, C., and Fung, K. (2001) Characterisation of tissue-specific LIM domain protein (FHL1C) which is an alternatively spliced isoform of a human LIM-only protein (FHL1). *J. Cell Biochem.* **82:** 1-10.
32. Noack, D., Heyworth, P.G., Newburger, P.E., and Cross, A.R. (2001) An unusual intronic mutation in the *CYBB* gene giving rise to chronic granulomatous disease. *Biochim. Biophys. Acta.* **1537:** 125-131.
33. Palmetshofer, A., Zechner, D., Luger, T.A., Barta, A. (1995) Splicing variants of the human growth hormone mRNA: detection in pituitary, mononuclear cells and dermal fibroblasts. *Mol. Cell Endocrinol.* **113:** 225-234.
34. Pankratova, E., Deyev I., Zhenilo S., Polanovsky, O. (2001) Tissue-specific isoforms of the ubiquitous transcription factor Oct-1. **266:** 239-245.
35. Penzel, R., Uhl, J., Kopitz, J., Beck, M., Otto, H.F., and Cantz, M. (2001) Splice donor site mutation in the lysosomal neuraminidase gene causing exon skipping and complete loss of enzyme activity in a sialidosis patient. *FEBS Letters.* **501:** 135-138.
36. Quinlan, J.J., Firestone, L.L., Homanics, G.E. (2000) Mice lacking the long splice variant of the gamma 2 subunit of the GABA(A) receptor are more sensitive to benzodiazepines. *Pharmacol. Biochem. Behav.* **66:** 371-374.
37. Sakamoto, O., Ohura, T., Katsushima, Y., Fujiwara, I., Ogawa, E., Miyabayashi, S.,
38. Salviati, L., Hernandez-Rosa, E., Walker, W.F., Sacconi, S., DiMauro, S., Schon, E.A., Davidson, M.M. (2002) Copper supplementation restores cytochrome c oxidase activity in cultured cells from patients with SCO2 mutations. *Biochem. J.* **363:** 321-7.
39. Schmucker, B., Tang, Y., and Kressel, M. (1999) Novel alternatively spliced isoforms of the neurofibromatosis type 2 tumor suppressor are targeted to the nucleus and cytoplasmic granules. *Hum. Mol. Genetics.* **8:** 1561-1570.

40. Shi, Z-Z., Habib, G.M., Lebovitz, R.M., and Lieberman, M.W. (1995) Cloning of cDNA and genomic structure of the mouse -glutamyl transpeptidase-encoding gene. *Gene.* **167:** 233-237.
41. Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997) A Genomic Perspective on Protein Families. *Science.* **278:** 631-637.
42. Thanaraj, T.A. (1999) A clean data set of EST-confirmed splice sites from *Homo sapiens* and standards for clean-up procedures. *Nucleic Acids Res.* **27**(13): 2627-2637.
43. Valentine, C.R. (1998) The association of nonsense codons with exon skipping. *Mutat. Res.* **411**: 87-117.
44. Wolfsberg, T.G. & Landsman, D. (1997) A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res.* **25:** 1626-1632
45. Zhao Q., Hay, N., and Morales, C.R. (1997) Structural analysis of the mouse prosaposin (SGP-1) gene reveals the presence of an exon that is alternatively spliced in transcribed mRNAs. *Mol. Reprod. Dev.* **48:** 1-8.
46. Zhong, Z., Schneider, T., Cabral, D., Donohoe, T., and Rothstein, T. (2001) An alternatively spliced long form of Fas apoptosis inhibitory molecule (FAIM) with tissue-specific expression in the brain. *Mol. Immunol.* **38:** 65-72.

Appendix A

Data Sources

Mouse and human transcript, genomic, and protein databases were utilized and are listed under their individual subheadings with data locations in brackets.

(i)Transcript and EST databases

a. 524 human chromosome 22 mRNAs, v.2.3 (http://www.sanger.ac.uk/HGP/Chr22/cwa_archive/Release_2.3_19-05-2000.shtml)
b. Mouse Ensembl cDNAs, v.0.1.0 (ftp://ftp.ensembl.org/pub/mouse-0.1.0/data/cdna/)
c. Mouse Riken cDNAs, Feb 7th 2001 release (http://genome.gsc.riken.go.jp/)
d. Mouse REFSEQ mRNAs, August 2001 release (ftp://ncbi.nlm.nih.gov/refseq/M_musculus/mRNA_Prot/)
e. Mouse Mammalian Gene Collection cDNAs, Feb 7th release (http://mgc.nci.nih.gov/)
f. Mouse dbEST at NCBI, release 123 (ftp://ncbi.nlm.nih.gov/repository/dbEST/)
g. Human dbEST at NCBI, release 125 (ftp://ncbi.nlm.nih.gov/repository/dbEST/)

All mouse transcript sequences with the exception of mouse dbEST were pooled into one database.

(ii) Genomic databases

a. Ensembl Mouse contigs, v.0.1.0 (ftp://ftp.ensembl.org/pub/mouse_0.1.0/data/dna)

b. Ensembl Mouse contigs, v.0.2.0 (ftp://ftp.ensembl.org/pub/mouse-0.2.0/data/dna/)

c. 52 Human EMBL files (http://www.sanbi.ac.za/exon_skipping)

d. 347 human multi-exon genes (http://www.sanbi.ac.za/exon_skipping/input.htm)

(iii) Protein databases

a. Mouse REFSEQ protein sequences, May 18th 2001 (ftp://ncbi.nlm.nih.gov/refseq/M_musculus/mRNA_Prot/)

b. Mouse SWISSPROT, v39.24 (ftp://ftp.ebi.ac.uk/pub/databases/swissprot/)

c. Ensembl mouse peptides, v.0.1.0 (ftp://ftp.ensembl.org/pub/mouse-0.1.0/data/pep/)

All mouse protein sequences were pooled into one database.

Appendix B

(i) Scripts used in mouse exon-skipping detection pipeline

All scripts are made available at http://www.sanbi.ac.za/~tzuming/THESIS_current/. Directions of usage are annotated in all scripts located in the directory.

A summary of the functions performed by each script, are provided below.

1. Extraction of top BLAST hits

***Extract_first_blasthit.only.pl* –** extracts only the first blast hit only in a blast report.

2. Parsing BLAST results

***Blast_parser.pl* –** retrieves query and subject names, total query matching length, percentage identities, E-values, and positions of the query and subject sequence with respect to the subject and query sequences respectively.

3. Extraction and re-naming of mouse sequences

***Split_fasta_extensions.pl* –** Splits a file containing multiple FASTA files and renames the files according to their FASTA headers

4. Virtual construction of exon-constructs

- ***Creating_E2F_files.pl* (**requires ***dots.pl*****) –** produces positions of putative exons in the mouse mRNA or cDNA sequence

- ***Extraction_pipe.pl*** (requires ***flex_extract.pl***) **–** extract exons and concatenate them together to form a virtual mRNA construct
- ***Vcontig_pipe.pl*** (requires ***Concatenation.pl***) **–** extract exons and concatenate them together with 40Ns flanked by consensus splice-site signals, GT and AG, to produce a virtual genomic construct
- ***Sim4_parser_pipe.pl*** (requires ***Sim4_parser2.pl, Chr22_E2F_modify.pl***) **–** produces a list of exon positions from Sim4 results (in which the virtual mouse mRNA was aligned to its virtual genomic sequence)

5. Aligning ESTs to genomic sequences and sim4 parsing

- ***Chr22_skip_detection.pl*** (***Chr22_extract_accessions.pl, Chr22_extract_seqs.pl, FASTA_converter.pl, Chr22_exonskip.py***) – This script takes in the virtual mouse transcript, virtual mouse genomic sequences, and a file containing putative exon positions in the genomic construct as input files. ESTs matching to the virtual mouse construct are retrieved and mapped against the virtual genomic construct using sim4. Sim4 outputs are parsed to create a report containing putative mouse exon-skipping events.

(ii) Calculations of the probability (P) that a given mouse multi-exon mRNA has at least one non-consecutive exon-junction as detected by j_explorer.pl

The equation as described in Hide *et al.*2001:

1. Given the parameters (150bp extracted from consecutive and non-consecutive exon junctions and $P<1x10^{-40}$)

$$\text{Probability}(P) = 1 - (1-f)^m$$
$$= 1 - (1-3/5372)^{326/72}$$
$$= 1-(0.999441)^{4.53}$$
$$= 0.0025$$

f = probability that a given exon junction in a given mRNA is non-consecutive (Number of times an EST spans a non-consecutive junction divided by the total number of times an EST spans a consecutive and non-consecutive junction)

m = average number of observed consecutive exon-junctions per multi-exon gene (Total number of observed consecutive exon-junctions divided by the total number of multi-exon genes)

Appendix C

BLAST Report for the Solute Carrier family 25 gene (SLC25A17)

```
Query= DJ362J20.1.MRNA.EMBL.exons
         (1791 letters)

Database: embl67_humanests
           3,566,182 sequences; 1,673,569,008 total letters

Searching..................................................done

                                                                 Score     E
Sequences producing significant alignments:                      (bits)  Value

BF928481 IL2-NT0200-061200-269-E06 NT0200 Homo sapiens cDNA, mRN...   371  e-100
AW182439 xj42b02.x1 Soares_NFL_T_GBC_S1 Homo sapiens cDNA clone ...   339  2e-90
AA678698 ah01h11.s1 Gessler Wilms tumor Homo sapiens cDNA clone ...   315  3e-83
AI370788 qz89d11.x1 Soares_pregnant_uterus_NbHPU Homo sapiens cD...   313  1e-82
AI274965 ql58g03.x1 Soares_NhHMPu_S1 Homo sapiens cDNA clone IMA...   313  1e-82
AA947261 od86c11.s1 NCI_CGAP_Ov2 Homo sapiens cDNA clone IMAGE:1...   313  1e-82
BF732694 nae14b12.x1 NCI_CGAP_Ov18 Homo sapiens cDNA clone IMAGE...   313  1e-82
AA548666 nj17g12.s1 NCI_CGAP_Pr22 Homo sapiens cDNA clone IMAGE:...   311  4e-82
AW518565 xx98f08.x1 NCI_CGAP_Lym12 Homo sapiens cDNA clone IMAGE...   311  4e-82
AA873759 ob12h05.s1 NCI_CGAP_Kid3 Homo sapiens cDNA clone IMAGE:...   307  6e-81
AI742824 wg46f09.x1 Soares_NSF_F8_9W_OT_PA_P_S1 Homo sapiens cDN...   303  1e-79
R41580 yf88g04.s1 Soares infant brain 1NIB Homo sapiens cDNA clo...   303  1e-79
AI310376 qo66h08.x1 NCI_CGAP_Co8 Homo sapiens cDNA clone IMAGE:1...   301  4e-79
BG505154 602551618F1 NIH_MGC_61 Homo sapiens cDNA clone IMAGE:46...   301  4e-79
BG721061 602692784F1 NIH_MGC_97 Homo sapiens cDNA clone IMAGE:48...   301  4e-79
BG715671 602676938F1 NIH_MGC_96 Homo sapiens cDNA clone IMAGE:47...   301  4e-79
BF979253 602147896F1 NIH_MGC_62 Homo sapiens cDNA clone IMAGE:43...   301  4e-79
AA446906 zw90f03.s1 Soares_total_fetus_Nb2HF8_9w Homo sapiens cD...   301  4e-79
BG700428 602680314F1 NIH_MGC_95 Homo sapiens cDNA clone IMAGE:48...   299  2e-78
AI621013 ts76a02.x1 NCI_CGAP_GC6 Homo sapiens cDNA clone IMAGE:2...   297  6e-78
```

T16466 NIB1349 Normalized infant brain, Bento Soares Homo sapien...	297	6e-78
BF031121 601558804F1 NIH_MGC_58 Homo sapiens cDNA clone IMAGE:38...	295	2e-77
AU131418 Homo sapiens cDNA clone:NT2RP3002540, 5' end, expressed...	293	1e-76
AI468581 tg82e02.x1 Soares_NhHMPu_S1 Homo sapiens cDNA clone IMA...	293	1e-76
BG705697 602668927F1 NIH_MGC_96 Homo sapiens cDNA clone IMAGE:47...	293	1e-76
BG678668 602624465F1 NCI_CGAP_Skn4 Homo sapiens cDNA clone IMAGE...	287	6e-75
BF248397 601821450F1 NIH_MGC_62 Homo sapiens cDNA clone IMAGE:40...	285	2e-74
AW974649 EST386874 MAGE resequences, MAGN Homo sapiens cDNA, mRN...	272	4e-70
BF212676 601813907F1 NIH_MGC_54 Homo sapiens cDNA clone IMAGE:40...	272	4e-70
BE298274 601118144F1 NIH_MGC_17 Homo sapiens cDNA clone IMAGE:30...	270	1e-69
AI565678 tn34g08.x1 NCI_CGAP_Brn25 Homo sapiens cDNA clone IMAGE...	268	6e-69
AI557253 PT2.1_15_E03.r tumor2 Homo sapiens cDNA 3', mRNA sequen...	268	6e-69
BG118520 602348454F1 NIH_MGC_90 Homo sapiens cDNA clone IMAGE:44...	266	2e-68
BF106075 601823119F1 NIH_MGC_77 Homo sapiens cDNA clone IMAGE:40...	258	5e-66
AA356050 EST64542 Jurkat T-cells VI Homo sapiens cDNA 5' end sim...	256	2e-65
AW958597 EST370667 MAGE resequences, MAGE Homo sapiens cDNA, mRN...	254	8e-65
BF238998 601905257F1 NIH_MGC_54 Homo sapiens cDNA clone IMAGE:41...	248	5e-63
AU123445 Homo sapiens cDNA clone:NT2RM2000316, 5' end, expressed...	**242**	**3e-61**
AA385739 EST99419 Thyroid Homo sapiens cDNA 5' end similar to si...	242	3e-61
AA326069 EST29179 Cerebellum II Homo sapiens cDNA 5' end.	**240**	**1e-60**
BF572391 602076451F1 NIH_MGC_62 Homo sapiens cDNA clone IMAGE:42...	238	5e-60
BE389243 601282379F1 NIH_MGC_44 Homo sapiens cDNA clone IMAGE:36...	212	3e-52
BE385685 601275673F1 NIH_MGC_20 Homo sapiens cDNA clone IMAGE:36...	212	3e-52
BE889576 601512493F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:39...	184	6e-44
AA330152 EST34002 Embryo, 12 week II Homo sapiens cDNA 5' end.	184	6e-44

<u>Supplementary Figure 1:</u> Blast report of the human SLC25A17 gene

Virtual human transcripts were constructed by concatenating all the predicted human exons (See Methods A. iii for details) and searched for matching human EST sequences at dbEST (Appendix Ai.g). The red-highlighted ESTs represent previously identified exon-skipped ESTs captured by Hide *et al.*2001.

(A)

seq1 = Human solute carrier family 25 genomic sequence, 50258 bp

seq2 = EST accession: AU123445, 744 bp

(complement strand)

4348-4432 (7-93) 96% <- exon 8

7453-7548 (94-189) 100% <- exon 7

7641-7786 (190-335) 100% <- exon 6

9422-9538 (336-452) 100% <- exon 5

22938-23093 (453-611) 98% <- exon 4

49598-49729 (612-744) 95% exon 1

(B)

seq1 = Human solute carrier family 25 genomic sequence, 50258 bp

seq2 = EST accession: AA326069, 273 bp

(complement strand)

7760-7786 (1-28) 96% <- exon 6

9422-9538 (29-145) 97% <- exon 5

49608-49735 (146-273) 97% exon 1

<u>Supplementary Figure 2:</u> Sim4 report illustrating the detection of exon-skipping events in the human SLC25A17 gene

Alignments of matching human ESTs to the SLC25A17 human genomic sequence were performed using sim4. Positions of the matching regions between ESTs and their genomic sequences are displayed in the sim4 report. Exon positions are annotated in the figure. In report (A), exons 2 and 3 were skipped and in report (B), the skipping of exons 2-4 was observed. The detected exon-skipping events were consistent with previously detected exon-skipping events by Hide *et al.*2001.

Supplementary Table 1: 269 Chromosome 22 mouse and human transcript orthologues

Mouse and human identifiers that are highlighted in **red** represent sequences that have been removed to produce the unique, 269 orthologous dataset. Mouse multi-exon transcript identifiers are highlighted in **blue**.

Human Locus Name	Human Ensembl transcript identifier	Mouse Ensembl transcript or Genbank identifier	E-value	Percentage Identity (%)	Matching mouse transcript length	Total length of mouse transcript
DRG1	bA247I13.C22.1	NM_007879.1	0	91.8	1014	1103
bA247I13.2	bA247I13.C22.2	NM_023743.1	0	88.3	2607	2946
RAD53	bA444G7.C22.1	NM_016681.1	0	86.4	1072	1239
bA494O16.1	bA494O16.C22.1	AK015930	3.00E-92	83.8	411	489
RBX1	bA554C12.C22.1	NM_019712.1	e-131	93.1	301	322
bK57G9.2	bK57G9.C22.2	NM_032396.1	0	88.5	995	1123
bK125H2.1	bK125H2.C22.1	AK016515	e-125	84.4	509	602
bK126B4.2	**bK126B4.C22.2**	**NM_023475.1**	8.00E-64	81.6	378	462
bK126B4.3	**bK126B4.C22.3**	**AK003698**	e-152	86.7	499	574
NHP2L1	bK216E10.2	AK004489	e-124	89.7	341	379
G22P1	bK216E10.C22.1	NM_010247.1	0	82.9	1345	1620
NUP50	bK217C2.C22.1	**NM_016714.1**	2.00E-81	85.6	311	362
CRYBB3	bK221G9.2	NM_021352.1	0	89.1	567	635
CRYBB2	bK221G9.3	NM_007773.1	0	89.3	552	617
TROB2	bK223H9.1	NM_020507.1	0	89.4	515	575
bK223H9.2	bK223H9.2	**AK003520**	e-122	91.5	305	332

PLA2G6	bK228A9.C22.1	**NM_016915.1**	0	89.6	957	1067
bK246H3.2	bK246H3.C22.2	NM_008513.1	1.00E-66	84.2	300	354
NAGA	bK250D10.5	NM_008669.1	e-118	83.9	503	598
bK250D10.2	bK250D10.C22.2	NM_025639.1	3.00E-97	83.8	425	506
bK250D10.3	bK250D10.C22.3	NM_011889.1	0	93	578	620
bK268H5.1	bK268H5.C22.1	**ENSMUST00000004905**	0	90.6	900	992
UPK3	bK268H5.C22.2	NM_023478.1	e-174	83.9	725	863
LARGE	bK282F2.C22.1	**NM_010687.1**	0	88.9	2020	2270
HMOX1	bK286B10.C22.2	NM_010442.1	e-158	85.5	569	664
MCM5	bK286B10.C22.3	**ENSMUST00000005545**	0	88.1	1941	2198
XBP1	bK292E10.1	NM_013842.2	0	88.9	553	615
MFNG	bK390B3.1.1	**BC010983.1**	0	86.4	675	780
bK414D7.1	bK414D7.C22.1	NM_020606.1|	7.00E-44	83.5	219	261
CRYBB1	bK445C9.1	NM_023695.1	e-114	83.6	502	599
CRYBA4	bK445C9.C22.2	NM_021351.1	e-140	85.6	506	590
HMG1L10	bK445C9.C22.3	NM_010439.1	e-174	90.1	458	507
TPST2	bK445C9.C22.4	NM_009419.1	0	88.5	1004	1133
MAFF	bK447C4.C22.1	NM_010755.1	e-177	93.5	394	420
NCF4	bK833B7.C22.1	**NM_008677.1**	0	84.5	848	1002
FBLN1	bK941F9.C22.1	**NM_010180.1**	0	87.8	1416	1610
E46L	bK941F9.C22.2	**AK003530**	0	85.4	726	849
EWSR1	bK984G1.C22.4	**NM_007968.1**	0	93.4	1841	1970
bK445C9.6	bK1048E9.C22.2	NM_018783.3	0	86	1353	1572
bK1048E9.3	bK1048E9.C22.3	AK020319	1.00E-58	88.5	186	209

bK1191B2.3	bK1191B2.C22.3a	ENSMUST00000001222	7.00E-35	82.8	198	238
SLC5A1	cB1E7.C22.1	BC003845.1	0	85.6	1065	1240
cB5E3.2	cB5E3.C22.2	AK008763	1.00E-42	86.3	165	190
cE81G9.2	cE81G9.C22.2	**AK006856**	9.00E-88	84.7	355	418
RAYL	cE132D12.C22.1	**AK011196**	e-125	85.9	464	535
MPST	cE146D10.2	**BC004079.1**	e-176	84	725	862
TST	cE146D10.C22.1	BC005644.1	0	85.1	761	893
PDGFB	cN10C3.C22.1	**NM_011057.1**	0	86	625	725
SYN3	cN28H9.C22.1	**NM_013722.1**	0	87.7	1528	1741
YWHAH	cN44A4.1	NM_011738.1	0	93.2	691	740
CSNK1E	dJ5O6.C22.1	**NM_013767.2**	0	90.4	1128	1246
LGALS1	dJ37E16.3	**NM_008495.1**	5.00E-98	86.1	348	403
SH3BP1	dJ37E16.C22.2	**ENSMUST00000001226**	0	87	739	848
dJ37E16.4	dJ37E16.C22.4	ENSMUST00000006548	3.00E-87	85.9	319	370
dJ37E16.6	dJ37E16.C22.6	ENSMUST00000001228	e-144	88.4	423	477
dJ37E16.7	dJ37E16.C22.7	**ENSMUST00000006546**	e-115	88.4	346	387
RBM9	dJ41P2.1	NM_021477.1	1.00E-60	86.8	217	249
PSCD4	dJ63G5.C22.1	**ENSMUST00000003356**	9.00E-90	81.4	524	642
MYH9	dJ68O2.C22.2	**NM_013607.1**	e-107	81.5	614	752
SLC5A4	dJ90G24.C22.4	**NM_023219.1**	0	85	688	805
dJ102D24.2	dJ102D24.C22.2	AK016311	2.00E-56	83.5	270	322
ADTB1L1	**dJ127L4.2**	**BC008513.1**	1.00E-35	85.4	153	178
ADTB1L2	**dJ127L4.3**	**NM_007454.1**	3.00E-40	84.9	152	178
dJ149A16.6	dJ149A16.C22.6	**ENSMUST00000001834**	0	91.5	1386	1513

FBX07	dJ149A16.C22.8.1	ENSMUST00000001838	3.00E-99	82.4	507	614
SSTR3	dJ151B14.3	NM_009218.1	e-171	86.7	561	646
RAC2	dJ151B14.C22.2	**NM_009008.1**	0	89.2	517	578
CACNA1I	dJ172B20.C22.1	NM_021415.2	e-141	87.4	447	510
ARHGAP8	dJ181C9.C22.2	**AK014171**	0	84.6	738	871
dJ186O1.1	dJ186O1.C22.1	AK005345	3.00E-77	89.1	229	256
DMC1	dJ199H16.1	NM_010059.1	0	91.9	941	1022
dJ215F16.1	dJ215F16.C22.1	AK018615	3.00E-36	83.8	182	216
DIA1	dJ222E13.4	BC004760.1	0	86.9	788	905
dJ222E13.1	**dJ222E13.C22.1**	**NM_023475.1**	1.00E-63	81.6	378	462
NDUFA6	dJ257I20.C22.3	NM_025987.1	e-114	90.2	306	338
CACNG2	dJ293L6.C22.1	**NM_007583.1**	0	93.9	911	969
TIMP3	dJ309I22.C22.1	NM_011595.1	0	89.9	535	594
PACSIN2	dJ323M22.C22.1	**NM_011862.1**	0	89	1298	1457
TTLL1	dJ323M22.C22.2.a	**BC010510.1**	0	87	1091	1253
DNAL4	dJ327J16.1	BC005426.1	e-132	93.7	298	317
NPTXR	dJ327J16.2	NM_030689.1	0	87	1003	1148
CBX6	dJ327J16.C22.3	AK004679	0	93	534	573
RPL3	dJ333H23.1.1	**NM_013762.1**	0	89	1078	1209
SYNGR1	dJ333H23.2.2	AK002972	0	91	642	701
dJ345P10.4	dJ345P10.C22.4b	**AK019850**	6.00E-36	79.6	332	416
PMM1	dJ347H13.3	AK004631	0	90.4	714	788
AC02	dJ347H13.C22.1	BC004645.1	0	89.8	2106	2342
dJ347H13.4	dJ347H13.C22.4	AK005706	0	90.5	614	677

dJ347H13.5	dJ347H13.C22.5	BC010793.1	0	89.5	551	614
PITPNB	dJ353E16.C22.1	NM_019640.1	0	91.6	748	815
SLC25A17	dJ362J20.C22.1	BC011292.1	0	88.7	804	905
dJ366L4.1	dJ366L4.C22.1	AK005942	2.00E-25	90	91	100
GRAP2	dJ370M22.C22.1	NM_010815.1	e-147	87	478	548
UQCRFSL1	dJ370M22.C22.2	NM_025710.1	0	85	678	796
HMG17L-1	dJ388M5.2	**NM_016957.1**	1.00E-57	88.7	181	203
SULTX3	dJ388M5.3	NM_013873.1	0	91.5	783	854
dJ388M5.4	dJ388M5.4	AK014771	2.00E-23	89.5	86	95
MTMR3	dJ394A18.C22.1	AK007036	e-110	94.2	247	261
MAPK12	dJ402G11.C22.1	NM_013871.1	0	85.9	949	1103
MAPK11	dJ402G11.C22.2	NM_011161.2	0	89.5	941	1050
dJ402G11.4	dJ402G11.C22.4	BC005574.1	0	87.3	915	1047
dJ402G11.5	dJ402G11.C22.5	AK005048	e-129	88.8	375	421
dJ402G11.8	dJ402G11.C22.8	NM_031260.1	0	85	1117	1313
TAB1	dJ407F17.C22.1	AK009321	1.00E-39	88.3	137	154
ST13	dJ408N23.C22.1	**BC003843.1**	0	89.9	831	914
OSBPL1	dJ430N8.C22.5a	AK007088	0	89	997	1118
DDX17	dJ434P1.C22.1	NM_007840.1	5.00E-37	87.7	136	154
KCNJ4	dJ434P1.C22.2	AK017299	e-111	82.9	530	638
KDELR3	dJ434P1.C22.3	BC011472.1	0	88.6	555	625
ARFGAP1	dJ437M21.C22.1	**BC004081.1**	e-114	86.6	389	448
GGA1	dJ437O22.C22.1	**ENSMUST00000001225**	0	87.7	1598	1820
H1F0	dJ466N1.1	ENSMUST00000006547	0	91.1	533	584

GALR3	dJ466N1.3	NM_015738.1	0	90.1	575	637
GCAT	dJ466N1.C22.2	**NM_013847.1**	0	87.7	1062	1209
dJ508I15.2	dJ508I15.2	BC006928.1	0	90.1	547	606
GTPBP1	dJ508I15.3	AK004612	0	91.6	1423	1552
dJ508I15.1	dJ508I15.C22.1	BC005733.1	2.00E-60	83.6	282	336
dJ508I15.5	dJ508I15.C22.5	AK013471	e-117	92.2	285	308
TOM1	dJ510H16.C22.1	**NM_011622.1**	0	86.3	720	833
HMG2L1	dJ510H16.C22.2	**ENSMUST00000005549**	0	89.7	516	574
BRD1	dJ522J7.2	ENSMUST00000004985	1.00E-23	83.2	144	172
BZRP	dJ526I14.C22.1b	**BC002055.1**	6.00E-86	84.4	358	423
dJ526I14.2	dJ526I14.C22.2	**ENSMUST00000001217**	0	86.9	1266	1455
dJ526I14.3	dJ526I14.C22.3	**NM_022723.1**	0	87.6	1617	1843
dJ569D19.1	dJ569D19.C22.1	AK015898	0	88.6	710	800
dJ569D19.4	dJ569D19.C22.4	AK010722	2.00E-89	85.2	336	393
dJ579N16.1	dJ579N16.1	AK004079	2.00E-42	82.8	232	279
dJ579N16.3	dJ579N16.3	NM_019977.1	0	86.5	732	845
SBF1	dJ579N16.C22.2	AK020972	0	87.4	914	1044
dJ591N18.1	dJ591N18.C22.1	**NM_025628.1**	2.00E-44	84.2	225	264
TCF20	dJ597B2.C22.1	NM_013836.1	0	88.5	2051	2314
dJ671O14.2	dJ671O14.C22.2	BC011200.1	4.00E-78	82.6	396	478
PPARA	dJ695O20A.C22.1	NM_011144.1	0	85.4	1202	1406
PKDREJ	dJ695O20A.C22.2	NM_011105.1	3.00E-78	82.9	389	468
CBX7	dJ742C19.5	ENSMUST00000000499	e-176	88.8	500	562
RANGAP1	dJ756G23.C22.2	**NM_011241.1**	0	87.7	947	1078

dJ796I17.2	dJ796I17.2	AK003990	0	85	1199	1408
dJ821D11.1	**dJ821D11.C22.1**	**AK016514**	4.00E-84	83.2	397	476
dJ821D11.2	**dJ821D11.C22.2**	**AK016514**	8.00E-66	92.3	324	77
dJ858B16.1	dJ858B16.C22.1	AK019095	5.00E-72	80.5	527	652
PISD	dJ858B16.C22.2	BC003217.1	2.00E-29	86.6	123	141
dJ889J22B.1.1	dJ889J22B.1.1	**ENSMUST00000001230**	e-161	87.3	506	578
dJ930L11.1	dJ930L11.C22.1	AK010756	e-136	85.7	488	568
TEF	dJ979N1.C22.1	NM_017376.1	0	91.3	833	911
dJ1014D13.2	dJ1014D13.C22.2	**ENSMUST00000006545**	0	87	618	709
dJ1039K5.2	dJ1039K5.2	NM_020516.1	e-175	88	523	593
PRKCABP	dJ1039K5.C22.1	NM_008837.1	0	90.8	1016	1116
SOX10	dJ1039K5.C22.4	**ENSMUST00000000579**	e-118	90.8	308	338
POLR2F	dJ1039K5.C22.5	AK007913	e-124	89.5	344	383
dJ1039K5.6	dJ1039K5.C22.6	**AK006539**	1.00E-72	84.8	297	349
ADSL	dJ1042K10.1.2	NM_009634.1	0	89	1013	1136
dJ1042K10.2	dJ1042K10.C22.2	AK007455	0	87.6	767	869
ATF4	dJ1104E15.2	NM_009716.1	0	86.9	621	709
MGAT3	dJ1104E15.C22.1	NM_010795.1	0	90.1	929	1029
dJ1104E15.4	dJ1104E15.C22.4	AK010954	6.00E-27	85.5	124	144
TXN2	dJ1119A7.C22.1	**NM_019913.1**	e-122	86	431	500
EIF3S7	dJ1119A7.C22.2	**NM_018749.1**	0	90.9	1445	1588
dJ1119A7.3	dJ1119A7.C22.3	ENSMUST00000005484	e-111	85.4	436	508
dJ1119A7.4	dJ1119A7.C22.4	NM_019424.1	1.00E-71	85.7	271	315
dJ1163J1.4	dJ1163J1.C22.1	AK005541	0	87.2	1027	1176

dJ1163J1.5	dJ1163J1.C22.5	NM_013882.1	8.00E-36	89.4	119	132
CELSR1	dJ1163J1.C22.6	NM_009886.1	0	84.7	3361	3966
dJ1170K4.1	dJ1170K4.1	AK011637	5.00E-32	84.6	154	181
dJ1170K4.2	dJ1170K4.2	**AK004939**	0	87.3	744	851
dJ1170K4.4	dJ1170K4.C22.4	**ENSMUST00000006876**	8.00E-90	82	485	590
dJ1177I5.1	dJ1177I5.1	ENSMUST00000001237	3.00E-70	89	211	236
MSE55	dJ1177I5.2	**ENSMUST00000001227**	e-107	84.2	433	513
RRP22	Em:AC000026.C22.1	AK008807	e-121	83.9	514	611
ADTB1	**Em:AC000026.C22.2**	**NM_007454.1**	0	90.2	1804	1999
GAR22	Em:AC000026.C22.4	AK019661	e-138	87.5	434	495
GNAZ	Em:AC000029.C22.1	ENSMUST00000000001	0	91.6	964	1051
ACR	Em:AC000036.C22.1	NM_013455.1	e-131	81.9	727	882
AC000036.3	Em:AC000036.C22.3	NM_021423.1	0	88.5	830	936
NLVCF	Em:AC000068.C22.1	NM_010922.1	e-107	85.1	420	490
UFD1L	Em:AC000068.C22.2	**BC006630.1**	0	89.2	825	923
AC000068.3	Em:AC000068.C22.3	AK010889	3.00E-38	86.3	152	175
CDC45L	Em:AC000071.C22.1	**ENSMUST00000005393**	0	88.2	1482	1678
TR	Em:AC000078.C22.1	**NM_013711.1**	0	86.2	626	725
COMT	Em:AC000080.C22.1	**ENSMUST00000000335**	2.00E-93	81.5	542	664
HIRA	Em:AC000085.C22.1	**ENSMUST00000004222**	0	81.5	1579	664
CLDN5	Em:AC000088.C22.1	BC002016.1	e-167	88.6	560	1780
GNB1L	Em:AC000089.C22.2	**NM_023120.1**	e-158	83.1	713	856
TBX1	Em:AC000091.C22.1b	**ENSMUST00000000906**	0	88.7	625	703
PNUTL1	Em:AC000093.C22.1	ENSMUST00000000255	0	89.3	892	997

GP1BB	Em:AC000093.C22.2	NM_010327.1	e-125	87.6	391	445
DGCR2	Em:AC000095.C22.1	**NM_010048.1**	0	87.7	1422	1620
RAB36	Em:AC000102.C22.1	AK018269	0	87	610	700
RABL2B	Em:AC002055.C22.1	AK004012	e-123	90.6	321	353
LIMK2	Em:AC002073.C22.1c	NM_010718.1	0	89.1	1639	1838
AC002073.2	Em:AC002073.C22.2	AK005141	8.00E-77	85.1	304	356
CRKL	Em:AC002470.C22.1	NM_007764.1	0	90.4	825	911
P2RXL1	**Em:AC002472.C22.1**	**NM_011028.1**	0	84.2	906	1074
LZTR1	Em:AC002472.C22.2	NM_025808.1	0	86.5	1436	1658
P2RXL2	**Em:AC002472.C22.4**	**NM_011028.1**	7.00E-44	84.3	200	236
AC002472.7	Em:AC002472.C22.7	AK012958	0	86.9	674	774
AC002472.8	Em:AC002472.C22.8	AK007329	0	88.3	751	849
bK963H5.1	Em:AC003072.C22.1	AK015917	3.00E-60	85.1	246	288
AC004019.4	Em:AC004019.C22.4	ENSMUST00000005026	5.00E-94	82.4	494	592
AC004019.5	Em:AC004019.C22.5	**ENSMUST00000000100**	e-103	87.1	332	380
AC004033.1	Em:AC004033.C22.1	ENSMUST00000006056	3.00E-50	85.8	200	232
LIF	Em:AC004264.C22.2	NM_008501.1	5.00E-93	83.7	416	493
SLC25A1	Em:AC004463.C22.2	**ENSMUST00000003622**	0	88.6	787	887
GSCL	Em:AC004463.C22.3	**ENSMUST00000003623**	2.00E-55	89.1	172	192
AC004471.1	Em:AC004471.C22.1	**ENSMUST00000003251**	0	89.1	869	192
STK22A	Em:AC004471.C22.3	NM_009436.1	0	87.2	948	995
dJ430N8.1	Em:AC004542.C22.1	AK018377	1.00E-90	88	349	1076
AC004832.1	Em:AC004832.C22.1	**ENSMUST00000003683**	e-132	85.3	435	408
AC004832.5	Em:AC004832.C22.5	**ENSMUST00000003682**	9.00E-67	86.8	302	500

SEC14L2	Em:AC004832.C22.6	**BC005759.1**	0	83.8	1077	359
dJ130H16.1	Em:AC004997.C22.1	AK010455	0	85.7	852	125
SF3A1	Em:AC004997.C22.3	**NM_026175.1**	0	86	987	989
ZNF278	Em:AC005003.C22.3	AK018372	0	90.7	613	1087
SMTN	Em:AC005005.C22.2	NM_013870.1	0	93.8	749	652
AC005005.7	Em:AC005005.C22.7	AK018791	9.00E-25	84.2	178	888
PES1	Em:AC005006.C22.1	**NM_022889.1**	0	81.6	1217	217
AC005006.2	Em:AC005006.C22.2	**ENSMUST00000003675**	0	86.2	870	1410
TCN2	Em:AC005006.C22.3	NM_015749.1	3.00E-74	85.7	405	1013
USP18	Em:AC005500.C22.1	NM_011909.1	2.00E-46	82.6	253	305
AC005500.3	Em:AC005500.C22.3	ENSMUST00000006057	e-105	89.1	304	340
AC005500.4	Em:AC005500.C22.4	**BC005800.1**	0	87.3	1676	1917
NF2	Em:AC005529.C22.1	NM_010898.1	0	90.1	1615	1787
NIPSNAP1	Em:AC005529.C22.2	NM_008698.1	0	88.8	701	788
AC005529.5	Em:AC005529.C22.5	NM_026015.1	0	90.8	466	512
AC005529.6	Em:AC005529.C22.6	AK012489	2.00E-46	89.4	144	160
PK1.3	Em:AC005529.C22.7	AK006840	0	89.7	1001	1114
ARVCF	Em:AC005663.C22.1	**ENSMUST00000000334**	0	88.4	787	889
BID	Em:AC006285.C22.1	**BC002031.1**	2.00E-21	82.2	148	179
ATP6E	Em:AC006285.C22.6	**BC003421.1**	0	91	618	678
RANBP1	Em:AC006547.C22.1	**AK002989**	e-179	88.8	510	573
AC006547.2	Em:AC006547.C22.2	**NM_008307.1**	0	84.6	1007	1185
AC006547.3	Em:AC006547.C22.3	**ENSMUST00000005880**	2.00E-64	82.7	354	426
AC006547.4	Em:AC006547.C22.4	**ENSMUST00000005878**	e-174	89.8	468	520

AC006547.5	**Em:AC006547.C22.5**	**ENSMUST00000005875**	0	83.1	1009	1213
AC006548.8	Em:AC006548.C22.8	**NM_010248.1**	2.00E-26	83.7	150	175
AC006946.1	Em:AC006946.C22.1	**ENSMUST00000002977**	3.00E-99	90.9	261	286
IL17R	Em:AC006946.C22.3	**NM_008359.1**	2.00E-89	85.1	350	410
AC007050.2	**Em:AC007050.C22.2**	**BC005632.1**	e-164	86.2	569	653
AC007050.4	**Em:AC007050.C22.4**	**AK017213**	4.00E-43	92.1	129	139
BCRL5	**Em:AC007050.C22.5**	**AK017213**	3.00E-67	89.9	205	227
SNAP29	Em:AC007308.C22.2	NM_023348.1	e-105	84.7	418	492
HCF2	Em:AC007308.C22.3	NM_008223.1	0	85.3	1013	1186
AC007663.2	**Em:AC007663.C22.2**	**NM_011172.1**	2.00E-53	84.2	241	285
AC007663.3	**Em:AC007663.C22.3**	**ENSMUST00000003625**	e-168	87.9	505	573
AC008079.1	Em:AC008079.C22.1	NM_011909.1	5.00E-60	83.8	276	328
TUBAL2	Em:AC008101.C22.1	NM_017379.1	0	89.5	1205	1345
AC008101.3	Em:AC008101.C22.3	AK014598	2.00E-65	90	189	209
HsPOX2	**Em:AC008103.C22.1**	**NM_011172.1**	0	83.6	1399	1671
AC008103.2	**Em:AC008103.C22.2**	**AK017213**	0	88.4	561	632
DGCR6	**Em:AC008103.C22.4**	**ENSMUST00000003625**	e-154	87.9	467	530
AC016026.1	Em:AC016026.C22.1	**ENSMUST00000004559**	0	86.7	894	1030
AC016026.2	Em:AC016026.C22.2	**ENSMUST00000003263**	0	89.7	855	952
AP000344.5	**Em:AP000344.C22.5**	**ENSMUST00000005879**	2.00E-47	81.3	309	379
ASLL	Em:AP000346.C22.1	AK010922	4.00E-25	84.4	130	153
AP000348.4	Em:AP000348.C22.4	ENSMUST00000000928	2.00E-33	92.3	96	103
MMP11	Em:AP000349.C22.1	**NM_008606.1**	e-163	86.1	560	649
SMARCB1	Em:AP000349.C22.2	**NM_011418.1**	0	90.6	1048	1155

AP000350.1	Em:AP000350.C22.1	AK007948	e-173	87.2	554	629
MIF	Em:AP000350.C22.3	NM_010798.1	2.00E-97	87.5	301	343
AP000350.5	**Em:AP000350.C22.5**	**AK016340**	3.00E-23	90.8	79	86
GSTT2	**Em:AP000350.C22.7**	**AK002392**	2.00E-74	82.3	396	480
DDT	Em:AP000351.C22.2	NM_010027.1	2.00E-23	83.4	136	162
AP000351.3	**Em:AP000351.C22.3**	**AK002392**	3.00E-79	82.7	398	480
AP000351.5	**Em:AP000351.C22.5**	**AK016340**	3.00E-27	83.8	145	172
GSTT1	Em:AP000351.C22.10	**BC012254.1**	e-122	82.7	593	716
AP000351.11	**Em:AP000351.C22.11**	**AK016340**	4.00E-25	81.3	187	229
AP000352.1	Em:AP000352.C22.1	**ENSMUST00000001712**	0	88.8	3655	4112
AP000353.2	Em:AP000353.C22.2	**ENSMUST00000006505**	e-129	82.2	686	832
unknown	Em:AP000353.C22.3	**NM_011820.1**	e-115	81.9	626	760
AP000354.2	Em:AP000354.C22.2	AK015541	e-124	90.2	333	368
AP000354.4	**Em:AP000354.C22.4**	**AK017213**	e-121	86.8	424	486
AP000354.6	Em:AP000354.C22.6	NM_025963.1	1.00E-86	87.2	287	328
ADORA2A	Em:AP000355.C22.1	NM_007413.1	6.00E-23	83.6	133	158
GGT1	**Em:AP000356.C22.4**	**NM_008116.1**	0	81.2	1070	1316
BCRL6	**Em:AP000356.C22.5**	**AK017213**	0	88	709	803
SNRPD3	Em:AP000356.C22.7	NM_026095.1	e-144	91.8	350	380
AP000365.2	Em:AP000365.C22.2	NM_011017.1	6.00E-41	83.4	207	247
YME1L2	Em:AP000535.C22.1	NM_013771.1	5.00E-45	92.3	121	130
ABCD1P4	Em:AP000543.C22.3	**NM_007435.1**	e-127	85.7	457	532
AP000550.2	**Em:AP000550.C22.2**	**NM_008116.1**	9.00E-67	80.8	472	582
AP000550.6	**Em:AP000550.C22.6**	**AK017213**	0	88.6	553	622

AP000552.3	**Em:AP000552.C22.3**	**AK017213**	0	88.4	561	632
UBE2L3	Em:AP000553.C22.1	AK007265	0	96.7	450	464
PPIL2	Em:AP000553.C22.2	AK009460	0	85.7	1345	1562
AP000553.3	Em:AP000553.C22.3	NM_023769.1	e-121	90.3	320	353
SDF2L1	Em:AP000553.C22.4	NM_022324.1	e-158	88.3	464	524
AP000553.6	Em:AP000553.C22.6	AK007503	0	84.5	796	939
MAPK1	Em:AP000555.C22.1	NM_011949.1	0	92.6	976	1052
AP000557.1	Em:AP000557.C22.1	NM_010430.1	4.00E-65	85.7	252	293
D86995.1	EM:D86995.C22.1	AK016942	e-174	87	565	647
BCRL4	**Em:D87002.C22.1**	**AK017213**	0	88	653	736
TOP3b	Em:D87012.C22.1	NM_011624.1	0	86.9	2251	2588
IGLC1	Em:D87023.C22.2	BC002129.1	4.00E-24	82	164	199
D88269.1	**Em:D88269.C22.1**	**ENSMUST00000003436**	e-165	86.3	567	650
VPREB1	Em:D88270.C22.1	NM_016983.1	6.00E-70	84.7	289	340
BCR	**Em:U07000.C22.1**	**AK017213**	0	90.5	731	805
U62317.2	Em:U62317.C22.2	**BC003900.1**	e-117	87.5	373	425
SCO2	Em:U62317.C22.3	**AK002487**	e-118	87.5	373	425
CPT1B	Em:U62317.C22.5	NM_009948.1	0	86.6	1310	1510
CHKL	Em:U62317.C22.6	NM_007692.1	0	85.8	973	1132
MAPK8IP2	Em:U62317.C22.7	NM_021921.1	0	89.3	697	779
ARSA	Em:U62317.C22.8	BC011284.1	0	84	1056	1255
U62317.10	Em:U62317.C22.10	AK005702	4.00E-70	88.7	214	240
PVALB	fF24E5.C22.1	**AK013561**	e-103	89.1	297	332

Supplementary Table 2: Mouse predicted exon-positions of chromosome 22 orthologues

The following table describes the genome-confirmed exon positions predicted for each mouse transcript sequence in the 72 multi-exon gene set. Each row in the table represents an exon predicted from its corresponding transcript and genomic contig sequence. The E-value, corresponding mouse genomic contig identifier, percentage identity, length of match, exon positions with respect to both the query mouse transcript and genomic contig sequences were described for each exon. The number in brackets next to the transcript identifier indicates the number of exons predicted by Ensembl mouse exon predictions. Matching Ensembl mouse transcript identifiers are denoted with an asterisk.

Human Locus Name	Mouse ENSEMBL or Genbank cDNA or mRNA Identifier	Mouse ENSEMBL Contig Identifier	E-value	cDNA or mRNA position	Mouse contig position	% identity	Length of match	Exon Number
1. ARVCF	ENSMUST0000000334	AC012399.27.81676.147193	e-104	1=>193	63577=>63385	100	192	1
	*(18)	AC012399.27.81676.147193	3.00E-86	191=>353	55818=>55656	100	162	2
		AC012399.27.81676.147193	0	350=>882	55480=>54948	100	532	3

		AC012399.27.81676.147193	0	879=>1382	54066=>53563	100	503	4
		AC012399.27.81676.147193	e-100	1381=>1566	52929=>52744	100	185	5
		AC012399.27.81676.147193	1.00E-60	1564=>1683	52238=>52119	100	119	6
		AC012399.27.81676.147193	1.00E-91	1682=>1853	51809=>51638	100	171	7
		AC012399.27.81676.147193	1.00E-33	1851=>1925	50327=>50253	100	74	8
		AC012399.27.81676.147193	1.00E-66	1923=>2052	49286=>49157	100	129	9
		AC012399.27.81676.147193	3.00E-80	2052=>2204	48993=>48841	100	152	10
		AC012399.27.81676.147193	e-108	2203=>2402	48567=>48368	100	199	11
		AC012399.27.81676.147193	1.00E-57	2400=>2514	48267=>48153	100	114	12
		AC012399.27.81676.147193	3.00E-43	2512=>2602	48061=>47971	100	90	13
		AC012399.27.81676.147193	8.00E-22	2602=>2656	47758=>47704	100	54	14
		AC012399.27.81676.147193	2.00E-44	2654=>2746	47435=>47343	100	92	15
		AC012399.27.81676.147193	4.00E-67	2747=>2877	47111=>46981	100	130	16
		AC012399.27.81676.147193	4.00E-33	2876=>2949	46286=>46213	100	73	17
2. COMT	ENSMUST0000000335	AC012399.27.81676.147193	e-150	1=>269	40165=>40433	100	268	1
	*(5)	AC012399.27.81676.147193	e-107	267=>464	40800=>40997	100	197	2
		AC012399.27.81676.147193	6.00E-69	463=>595	41308=>41440	100	132	3
		AC012399.27.81676.147193	e-111	588=>795	44075=>44282	99.5	207	4
3. SOX10	ENSMUST0000000579*(							
	3)	AC040983.10.39657.104314	0	1=>432	52752=>53183	100	431	1
		AC040983.10.39657.104314	e-142	430=>686	53974=>54230	100	256	2
		AC040983.10.39657.104314	0	685=>1088	54838=>55241	100	403	3
		AC040983.10.39657.104314	0	1132=>1521	55285=>55674	100	389	4
4. TBX1	ENSMUST0000000906	AC008019.43.12954.57196	5.00E-67	1=>130	161=>290	100	129	1

	(4)	AC008019.43.12954.57196	2.00E-97	173=>353	333=>513	100	180	2
		AC008019.43.12954.57196	2.00E-53	351=>457	1910=>2016	100	106	3
		AC008019.43.12954.57196	1.00E-92	455=>627	2793=>2965	100	172	4
		AC008019.43.12954.57196	1.00E-82	628=>783	3475=>3630	100	155	5
		AC008019.43.12954.57196	3.00E-31	783=>852	4074=>4143	100	69	6
		AC008019.43.12954.57196	2.00E-51	851=>954	4224=>4327	100	103	7
5. GGA1	ENSMUST0000001225	AC026386.20.1.129738	2.00E-36	38=>116	68408=>68330	100	78	1
		AC026386.20.1.129738	7.00E-49	114=>213	67268=>67169	100	99	2
		AC026386.20.1.129738	9.00E-67	201=>338	66463=>66326	98.5	137	3
		AC026386.20.1.129738	3.00E-39	437=>520	64673=>64590	100	83	4
		AC026386.20.1.129738	2.00E-73	518=>658	62426=>62286	100	140	5
		AC026386.20.1.129738	6.00E-40	656=>744	62119=>62031	98.8	88	6
		AC026386.20.1.129738	2.00E-55	738=>848	61534=>61424	100	110	7
		AC026386.20.1.129738	1.00E-81	847=>1001	60820=>60666	100	154	8
		AC026386.20.1.129738	1.00E-28	1000=>1065	59099=>59034	100	65	9
		AC026386.20.1.129738	8.00E-95	1063=>1239	58624=>58448	100	176	10
		AC026386.20.1.129738	e-102	1238=>1427	58083=>57894	100	189	11
		AC026386.20.1.129738	8.00E-92	1427=>1598	57368=>57197	100	171	12
		AC026386.20.1.129738	8.00E-58	1590=>1708	57003=>56885	99.1	118	13
		AC026386.20.1.129738	8.00E-55	1706=>1815	56803=>56694	100	109	14
6. SH3BP1	ENSMUST0000001226*(							
	14)	AC026386.20.1.129738	3.00E-57	44=>157	47682=>47569	100	113	1
		AC026386.20.1.129738	6.00E-37	155=>234	47478=>47399	100	79	2
		AC026386.20.1.129738	4.00E-78	231=>379	46985=>46837	100	148	3

		AC026386.20.1.129738	4.00E-32	376=>447	46268=>46197	100	71	4
		AC026386.20.1.129738	2.00E-43	448=>538	46071=>45981	100	90	5
		AC026386.20.1.129738	1.00E-78	536=>685	45508=>45359	100	149	6
		AC026386.20.1.129738	2.00E-58	683=>798	44766=>44651	100	115	7
		AC026386.20.1.129738	6.00E-40	794=>878	44282=>44198	100	84	8
		AC026386.20.1.129738	2.00E-40	876=>961	44115=>44030	100	85	9
		AC026386.20.1.129738	2.00E-61	957=>1077	43324=>43204	100	120	10
		AC026386.20.1.129738	3.00E-48	1076=>1174	42575=>42477	100	98	11
		AC026386.20.1.129738	2.00E-86	1173=>1335	42184=>42022	100	162	12
		AC026386.20.1.129738	3.00E-45	1333=>1426	40529=>40436	100	93	13
		AC025911.6.9637.16834	0	1427=>1800	4684=>4311	100	373	14
7. MSE55	ENSMUST0000001227	AC026386.20.1.129738	0	1=>464	103178=>102715	100	463	1
		AC026386.20.1.129738	0	463=>992	101371=>100842	100	529	2
		AC026386.20.1.129738	4.00E-99	1044=>1227	100790=>100607	100	183	3
8. dJ889J22B.1.1	ENSMUST0000001230	AC026386.20.201618.217833	6.00E-87	1=>163	15962=>15800	100	162	1
		AC026386.20.201618.217833	1.00E-60	161=>279	15679=>15561	100	118	2
		AC026386.20.201618.217833	2.00E-59	277=>393	15182=>15066	100	116	3
		AC026386.20.201618.217833	e-131	390=>627	14435=>14198	100	237	4
9. AP000352.1	ENSMUST0000001712	AC068241.3.151794.204355	6.00E-45	1=>94	41775=>41682	100	93	1
		AC068241.3.151794.204355	3.00E-59	91=>208	41238=>41121	100	117	2
		AC068241.3.151794.204355	3.00E-71	206=>343	40384=>40247	100	137	3
		AC068241.3.151794.204355	5.00E-98	341=>523	38489=>38307	100	182	4
		AC068241.3.151794.204355	2.00E-66	524=>653	37044=>36915	100	129	5
		AC068241.3.151794.204355	4.00E-80	653=>805	34847=>34695	100	152	6

		AC068241.3.151794.204355	2.00E-91	1090=>1261	32269=>32098	100	171	7
		AC068241.3.151794.204355	9.00E-72	1258=>1396	30211=>30073	100	138	8
		AC068241.3.151794.204355	e-120	1396=>1615	29334=>29115	100	219	9
		AC068241.3.151794.204355	4.00E-37	1613=>1693	26971=>26891	100	80	10
		AC068241.3.151794.204355	e-103	1692=>1882	26386=>26196	100	190	11
		AC068241.3.151794.204355	6.00E-82	1879=>2034	25237=>25082	100	155	12
		AC068241.3.151794.204355	e-106	2035=>2230	24386=>24191	100	195	13
		AC068241.3.151794.204355	e-137	2227=>2474	22082=>21835	100	247	14
		AC068241.3.151794.204355	6.00E-85	2470=>2630	20650=>20490	100	160	15
		AC068241.3.151794.204355	5.00E-61	2628=>2748	19388=>19268	100	120	16
		AC068241.3.151794.204355	3.00E-87	2746=>2910	13890=>13726	100	164	17
		AC068241.3.151794.204355	e-114	2906=>3115	12628=>12419	100	209	18
		AC068241.3.151794.204355	2.00E-78	3113=>3262	12277=>12128	100	149	19
		AC068241.3.151794.204355	e-146	3259=>3522	8412=>8149	100	263	20
		AC068241.3.151794.204355	e-147	3521=>3785	4571=>4307	100	264	21
		AC068241.3.151794.204355	2.00E-81	3782=>3936	2623=>2469	100	154	22
		AC068241.3.151794.204355	5.00E-95	3936=>4113	502=>325	100	177	23
10. dJ149A16.6	ENSMUST0000001834	AC063968.3.1.27688	6.00E-46	1=>95	749=>843	100	94	1
		AC063968.3.1.27688	2.00E-39	91=>174	2657=>2740	100	83	2
		AC063968.3.1.27688	2.00E-30	172=>240	4952=>5020	100	68	3
		AC063968.3.1.27688	4.00E-50	241=>342	7003=>7104	100	101	4
		AC063968.3.1.27688	4.00E-84	339=>497	8833=>8991	100	158	5
		AC063968.3.1.27688	6.00E-83	498=>654	11124=>11280	100	156	6
		AC063968.3.1.27688	2.00E-86	652=>814	12104=>12266	100	162	7

		AC063968.3.1.27688	1.00E-96	812=>991	14464=>14643	100	179	8
		AC063968.3.1.27688	e-103	989=>1179	15058=>15248	100	190	9
		AC063968.3.1.27688	4.00E-56	1180=>1291	15859=>15970	100	111	10
		AC063968.3.1.27688	4.00E-62	1290=>1411	16438=>16559	100	121	11
		AC063968.3.1.27688	4.00E-53	1409=>1515	19335=>19441	100	106	12
11. AC006946.1	ENSMUST0000002977	AC078896.10.54664.87044	e-109	1=>200	23360=>23161	100	199	1
		AC078896.10.54664.87044	5.00E-58	197=>311	21063=>20949	100	114	2
		AC078896.10.54664.87044	1.00E-55	312=>437	18292=>18162	100	125	3
		AC078896.10.54664.87044	1.00E-95	436=>613	14328=>14151	96.1	177	4
		AC078896.10.54664.87044	e-104	609=>800	10577=>10386	100	191	5
		AC078896.10.54664.87044	0	800=>1125	10218=>9893	100	325	6
12. AC004471.1	ENSMUST0000003251	AC079043.14.98198.212950	8.00E-76	1=>145	104005=>104149	100	144	1
		AC079043.14.98198.212950	2.00E-91	143=>313	105052=>105222	100	170	2
		AC079043.14.98198.212950	1.00E-46	314=>409	105310=>105405	100	95	3
		AC079043.14.98198.212950	4.00E-93	409=>582	107387=>107560	100	173	4
		AC079043.14.98198.212950	1.00E-62	581=>703	107671=>107793	100	122	5
		AC079043.14.98198.212950	7.00E-70	700=>834	108067=>108201	100	134	6
		AC079043.14.98198.212950	6.00E-55	830=>939	108932=>109041	100	109	7
		AC079043.14.98198.212950	2.00E-55	937=>1047	110198=>110308	100	110	8
		AC079043.14.98198.212950	1.00E-59	1046=>1163	112375=>112492	100	117	9
		AC079043.14.98198.212950	e-156	1161=>1440	112984=>113263	100	279	10
13. AC016026.2	ENSMUST0000003263*(							
	2)	AC079443.12.8696.17287	4.00E-46	1=>95	257=>351	100	94	1
		AC079443.12.8696.17287	6.00E-51	95=>197	726=>828	100	102	2

		AC079443.12.8696.17287	6.00E-85	195=>354	1236=>1395	100	159	3
		AC079443.12.8696.17287	4.00E-52	350=>454	2261=>2365	100	104	4
		AC079443.12.8696.17287	e-143	454=>711	4550=>4807	100	257	5
		AC079443.12.8696.17287	2.00E-60	710=>828	5022=>5140	100	118	6
		AC079443.12.8696.17287	5.00E-70	826=>960	7712=>7846	100	134	7
14. PSCD4	ENSMUST0000003356	AC073774.2.160678.196882	3.00E-78	1=>149	12035=>11887	100	148	1
		AC073774.2.160678.196882	2.00E-32	146=>217	11795=>11724	100	71	2
		AC073774.2.160678.196882	3.00E-62	215=>336	11178=>11057	100	121	3
		AC073774.2.160678.196882	2.00E-39	334=>417	10489=>10406	100	83	4
		AC073774.2.160678.196882	2.00E-57	416=>529	9845=>9732	100	113	5
		AC073774.2.160678.196882	3.00E-81	525=>678	9602=>9449	100	153	6
		AC073774.2.160678.196882	2.00E-57	678=>791	9117=>9004	100	113	7
		AC073774.2.160678.196882	2.00E-36	790=>868	7328=>7250	100	78	8
		AC073774.2.160678.196882	6.00E-33	867=>939	7127=>7055	100	72	9
		AC073774.2.160678.196882	4.00E-83	939=>1095	6967=>6811	100	156	10
		AC073774.2.160678.196882	4.00E-40	1095=>1179	6727=>6643	100	120	11
15. SLC25A1	ENSMUST0000003622	AC078895.12.59859.213034	4.00E-46	1=>95	94559=>94465	100	94	1
		AC078895.12.59859.213034	3.00E-56	92=>203	94135=>94024	100	111	2
		AC078895.12.59859.213034	6.00E-51	202=>304	93947=>93845	100	102	3
		AC078895.12.59859.213034	5.00E-76	301=>445	93774=>93630	100	144	4
		AC078895.12.59859.213034	5.00E-42	440=>527	93222=>93135	100	87	5
		AC078895.12.59859.213034	4.00E-55	523=>632	93023=>92914	100	109	6
		AC078895.12.59859.213034	2.00E-60	630=>748	92786=>92668	100	118	7
		AC078895.12.59859.213034	1.00E-36	743=>821	92588=>92510	100	78	8

		AC078895.12.59859.213034	4.00E-58	822=>936	92407=>92293	100	114	9
16. GSCL	ENSMUST0000003623	AC079043.14.98198.212950	e-162	1=>290	100268=>100557	100	289	1
		AC079043.14.98198.212950	e-143	289=>545	100690=>100946	100	256	2
		AC079043.14.98198.212950	1.00E-51	542=>645	101590=>101693	100	103	3
17. HIRA	ENSMUST0000004222	AC079831.14.42304.147674	4.00E-91	1=>171	103535=>103365	100	170	1
		AC079831.14.42304.147674	1.00E-57	171=>285	100997=>100883	100	114	2
		AC079831.14.42304.147674	1.00E-32	285=>357	97656=>97584	100	72	3
		AC079831.14.42304.147674	6.00E-53	356=>462	97119=>97013	100	106	4
		AC079831.14.42304.147674	e-119	461=>678	94326=>94109	100	217	5
		AC079831.14.42304.147674	3.00E-42	677=>765	92536=>92448	100	88	6
		AC079831.14.42304.147674	e-106	764=>959	87215=>87020	100	195	7
		AC079831.14.42304.147674	1.00E-88	957=>1123	85067=>84901	100	166	8
		AC079831.14.42304.147674	e-111	1120=>1323	73738=>73535	100	203	9
		AC079831.14.42304.147674	7.00E-56	1322=>1433	72670=>72559	100	111	10
		AC079831.14.42304.147674	6.00E-81	1424=>1577	70951=>70798	100	153	11
		AC079831.14.42304.147674	6.00E-87	1577=>1740	68908=>68745	100	163	12
		AC079831.14.42304.147674	2.00E-25	1739=>1799	68426=>68366	100	60	13
		AC079831.14.42304.147674	1.00E-54	1796=>1905	68021=>67912	100	109	14
		AC079831.14.42304.147674	5.00E-63	1904=>2027	66088=>65965	100	123	15
		AC079831.14.42304.147674	2.00E-87	2028=>2192	65472=>65308	100	164	16
		AC079831.14.42304.147674	6.00E-44	2190=>2281	63915=>63824	100	91	17
		AC079831.14.42304.147674	1.00E-57	2280=>2394	50643=>50529	100	114	18
18. AC016026.1	ENSMUST0000004559	AC083894.9.127619.158322	7.00E-96	1=>192	29467=>29272	97.9	191	1
		AC083894.9.127619.158322	0	256=>827	29208=>28637	100	571	2

		AC083894.9.127619.158322	7.00E-25	850=>916	28614=>28547	98.5	66	3
		AC083894.9.127619.158322	0	973=>1780	28486=>27679	100	807	4
		AC083894.9.127619.158322	2.00E-47	1778=>1875	27234=>27137	100	97	5
		AC083894.9.127619.158322	2.00E-56	1874=>1986	21925=>21813	100	112	6
		AC083894.9.127619.158322	6.00E-47	1985=>2081	20834=>20738	100	96	7
		AC083894.9.127619.158322	4.00E-51	2077=>2180	4714=>4611	100	103	8
		AC083894.9.127619.158322	3.00E-33	2181=>2254	4443=>4370	100	73	9
		AC083894.9.127619.158322	7.00E-99	2253=>2436	4291=>4108	100	183	10
19. CDC45L	ENSMUST0000005393(1							
	9)	AC083895.13.115418.189347	2.00E-27	49=>112	50524=>50461	100	63	1
		AC083895.13.115418.189347	3.00E-48	106=>204	49644=>49546	100	98	2
		AC083895.13.115418.189347	2.00E-73	202=>342	48029=>47889	100	140	3
		AC083895.13.115418.189347	2.00E-77	340=>487	46583=>46436	100	147	4
		AC083895.13.115418.189347	2.00E-24	485=>543	44596=>44538	100	58	5
		AC083895.13.115418.189347	5.00E-28	591=>655	37895=>37831	100	64	6
		AC083895.13.115418.189347	8.00E-61	705=>824	35101=>34982	100	119	7
		AC083895.13.115418.189347	1.00E-68	825=>957	34383=>34251	100	132	8
		AC083895.13.115418.189347	2.00E-49	956=>1056	34018=>33918	100	100	9
		AC083895.13.115418.189347	4.00E-87	1055=>1218	32497=>32334	100	163	10
		AC083895.13.115418.189347	1.00E-74	1215=>1357	26226=>26084	100	142	11
		AC083895.13.115418.189347	2.00E-40	1355=>1440	26009=>25924	100	85	12
		AC083895.13.115418.189347	8.00E-61	1440=>1559	24347=>24228	100	119	13
		AC083895.13.115418.189347	4.00E-35	1560=>1636	24037=>23961	100	76	14
		AC083895.13.115418.189347	1.00E-28	1633=>1698	21122=>21057	100	65	15

20. MCM5	ENSMUST0000005545*(							
	16)	AC084823.7.60726.106658	9.00E-89	1=>167	29859=>29693	100	166	1
		AC084823.7.60726.106658	4.00E-66	166=>294	29479=>29351	100	128	2
		AC084823.7.60726.106658	1.00E-66	294=>423	29268=>29139	100	129	3
		AC084823.7.60726.106658	4.00E-94	422=>597	27025=>26850	100	175	4
		AC084823.7.60726.106658	5.00E-84	597=>755	25402=>25244	100	158	5
		AC084823.7.60726.106658	4.00E-91	750=>920	23718=>23548	100	170	6
		AC084823.7.60726.106658	6.00E-93	919=>1092	22046=>21873	100	173	7
		AC084823.7.60726.106658	2.00E-58	1090=>1205	20312=>20197	100	115	8
		AC084823.7.60726.106658	1.00E-75	1203=>1347	18788=>18644	100	144	9
		AC084823.7.60726.106658	3.00E-30	1346=>1414	18357=>18289	100	68	10
		AC084823.7.60726.106658	1.00E-94	1414=>1590	18029=>17853	100	176	11
		AC084823.7.60726.106658	1.00E-57	1590=>1704	16576=>16462	100	114	12
		AC084823.7.60726.106658	1.00E-66	1704=>1833	15726=>15597	100	129	13
		AC084823.7.60726.106658	1.00E-75	1832=>1976	14767=>14623	100	144	14
		AC084823.7.60726.106658	4.00E-66	1975=>2103	13331=>13203	100	128	15
		AC084823.7.60726.106658	1.00E-50	2100=>2202	12348=>12246	100	102	16
21. HMG2L1	ENSMUST0000005549	AC084823.7.1.60705	5.00E-35	1=>76	16996=>17071	100	75	1
		AC084823.7.1.60705	3.00E-30	74=>141	18552=>18619	100	67	2
		AC084823.7.1.60705	1.00E-53	141=>247	19950=>20056	100	106	3
		AC084823.7.1.60705	6.00E-93	246=>418	20748=>20920	100	172	4
		AC084823.7.1.60705	3.00E-64	416=>540	26242=>26366	100	124	5
22. AC006547.4	ENSMUST0000005878	AC084822.4.61848.69670	5.00E-50	1=>101	4618=>4518	100	100	1
		AC084822.4.61848.69670	6.00E-62	96=>216	4300=>4180	100	120	2

		AC084822.4.61848.69670	1.00E-66	209=>337	2964=>2836	100	128	3
		AC084822.4.61848.69670	1.00E-59	335=>451	1430=>1314	100	116	4
		AC084822.4.61848.69670	2.00E-40	450=>534	250=>166	100	84	5
23. AC006547.3	ENSMUST0000005880	AC084822.4.152300.162083	1.00E-44	1=>92	7307=>7216	100	91	1
		AC084822.4.152300.162083	7.00E-83	90=>245	3085=>2930	100	155	2
		AC084822.4.152300.162083	2.00E-52	246=>350	2037=>1933	100	104	3
		AC084822.4.152300.162083	5.00E-62	348=>468	802=>682	100	120	4
24. AP000353.2	ENSMUST0000006505	AC087540.17.33540.65478	5.00E-32	1=>72	9599=>9528	100	71	1
		AC087540.17.33540.65478	e-116	68=>280	8328=>8116	100	212	2
		AC087540.17.33540.65478	1.00E-81	276=>430	7934=>7780	100	154	3
		AC087540.17.33540.65478	2.00E-90	430=>599	7256=>7087	100	169	4
		AC087540.17.33540.65478	3.00E-95	597=>774	6606=>6429	100	177	5
		AC087540.17.33540.65478	e-109	774=>975	6312=>6111	100	201	6
		AC087540.17.33540.65478	2.00E-40	976=>1061	5929=>5844	100	85	7
		AC087540.17.33540.65478	e-151	1060=>1330	5759=>5489	100	270	8
		AC087540.17.33540.65478	3.00E-76	1330=>1475	5394=>5249	100	145	9
		AC087540.17.33540.65478	4.00E-85	1474=>1634	5156=>4996	100	160	10
		AC087540.17.33540.65478	e-139	1633=>1883	4251=>4001	100	250	11
		AC087540.17.33540.65478	e-152	1883=>2155	3892=>3620	100	272	12
		AC087540.17.33540.65478	7.00E-96	2155=>2333	3514=>3336	100	178	13
		AC087540.17.33540.65478	6.00E-53	2330=>2436	3264=>3158	100	106	14
25. dJ1014D13.2	ENSMUST0000006545	AL589670.3.1.8618	6.00E-75	1=>143	8115=>7973	100	142	1
		AL589670.3.1.8618	1.00E-66	140=>268	7457=>7329	100	128	2
		AL589670.3.1.8618	2.00E-44	268=>359	7210=>7119	100	91	3

		AL589670.3.1.8618	9.00E-34	360=>433	6980=>6907	100	73	4
		AL589670.3.1.8618	1.00E-60	431=>549	6147=>6029	100	118	5
		AL589670.3.1.8618	8.00E-62	594=>714	4920=>4800	100	120	6
26. dJ1170K4.4	ENSMUST0000006876	AL590144.3.105821.121568	e-109	1=>200	15172=>14973	100	199	1
		AL590144.3.105821.121568	1.00E-70	198=>333	11245=>11110	100	135	2
		AL590144.3.105821.121568	4.00E-31	333=>402	10373=>10304	100	69	3
		AL590144.3.105821.121568	2.00E-97	401=>581	9789=>9609	100	180	4
		AL590144.3.105821.121568	e-113	580=>786	4984=>4778	100	206	5
		AL590144.3.105821.121568	4.00E-71	786=>922	4146=>4010	100	136	6
		AL590144.3.105821.121568	1.00E-58	922=>1037	2770=>2655	100	115	7
		AL590144.3.105821.121568	2.00E-54	1035=>1143	2451=>2343	100	108	8
27. BID	BC002031.1(3)	AC006404.24.1.142776	2.00E-43	1=>91	141918=>141828	100	90	1
		AC006404.24.1.142776	2.00E-30	87=>155	126853=>126785	100	68	2
		AC006404.24.1.142776	e-111	152=>364	125431=>125219	99	212	3
		AC006404.24.1.142776	4.00E-75	362=>505	122424=>122281	100	143	4
		AC006404.24.1.142776	e-114	505=>718	120893=>120680	99.5	213	5
		AC006404.24.1.142776	0	718=>1890	119354=>118184	99.5	1172	6
28. GNB1L	NM_023120.1*(6)	AC003066.1.1.176974	3.00E-47	32=>128	69427=>69523	100	96	1
		AC003066.1.1.176974	8.00E-76	126=>270	110817=>110961	100	144	2
		AC003066.1.1.176974	4.00E-65	268=>394	111355=>111481	100	126	3
		AC003066.1.1.176974	4.00E-90	391=>559	114491=>114659	100	168	4
		AC003066.1.1.176974	1.00E-49	556=>656	118475=>118575	100	100	5
		AC003066.1.1.176974	e-119	654=>874	122748=>122968	99.5	220	6
		AC003066.1.1.176974	0	871=>1362	134417=>134908	100	491	7

29. ATP6E	BC003421.1[1*] (9)	AC006447.21.1.140452	7.00E-48	1=>98	107141=>107044	100	97	1
		AC006447.21.1.140452	2.00E-30	97=>165	102863=>102795	100	68	2
		AC006447.21.1.140452	3.00E-56	163=>274	91567=>91456	100	111	3
		AC006447.21.1.140452	6.00E-30	275=>342	90746=>90679	100	67	4
		AC006447.21.1.140452	1.00E-43	342=>432	87224=>87134	100	90	5
		AC006447.21.1.140452	6.00E-33	429=>501	86471=>86399	100	72	6
		AC006447.21.1.140452	3.00E-47	499=>595	83978=>83882	100	96	7
		AC006447.21.1.140452	7.00E-42	596=>683	80805=>80718	100	87	8
		AC006447.21.1.140452	0	681=>1203	78608=>78086	100	522	9
30. SLC5A4	NM_023219.1	AC005302.68.1.66798	e-119	1=>218	63702=>63485	100	217	1
		AC005302.68.1.66798	2.00E-31	219=>289	62772=>62702	100	70	2
		AC005302.68.1.66798	5.00E-53	289=>395	61287=>61181	100	106	3
		AC005302.68.1.66798	2.00E-27	395=>458	59450=>59387	100	63	4
		AC005302.68.1.66798	1.00E-50	454=>560	56613=>56507	99	106	5
		AC005302.68.1.66798	3.00E-48	561=>659	43455=>43357	100	98	6
		AC005302.68.1.66798	5.00E-41	660=>746	42664=>42578	100	86	7
		AC005302.68.1.66798	e-123	746=>969	34208=>33985	100	223	8
		AC005302.68.1.66798	4.00E-72	965=>1103	27791=>27653	100	138	9
		AC005302.68.1.66798	3.00E-54	1104=>1212	25211=>25103	100	108	10
		AC005302.68.1.66798	5.00E-81	1213=>1366	23329=>23176	100	153	11
		AC005302.68.1.66798	3.00E-91	1365=>1535	16796=>16626	100	170	12
		AC005302.68.1.66798	e-120	1532=>1750	15109=>14891	100	218	13
		AC005302.68.1.66798	3.00E-51	1751=>1854	13131=>13028	100	103	14
		AC005302.68.1.66798	e-118	1854=>2069	11577=>11362	100	215	15

31. U62317.2	BC003900.1	AC079487.1.215358.230320	1.00E-69	1=>147	3214=>3360	97.9	146	1
		AC079487.1.215358.230320	0	164=>681	3360=>3877	96.9	517	2
		AC079487.1.215358.230320	e-120	660=>922	3874=>4136	95.8	262	3
		AC079487.1.215358.230320	0	934=>1321	4137=>4524	96.1	387	4
		AC079487.1.215358.230320	2.00E-37	1341=>1445	4520=>4624	94.2	104	5
		AC079487.1.215358.230320	0	1447=>2008	4614=>5175	97.1	561	6
32. MPST	BC004079.1	AL590144.3.45304.78127	0	5=>631	20680=>21306	99.6	626	1
		AL590144.3.45304.78127	0	630=>1134	24021=>24525	100	504	2
		AL590144.3.45304.78127	2.00E-36	1179=>1257	24570=>24648	100	78	3
33. ARFGAP1	BC004081.1	AL583889.2.1.9907	1.00E-51	1=>105	9735=>9631	100	104	1
		AL583889.2.1.9907	2.00E-62	102=>224	2694=>2572	100	122	2
		AL583889.2.89451.105061	3.00E-33	222=>295	2231=>2304	100	73	3
		AL583889.2.89451.105061	2.00E-65	296=>427	4349=>4480	99.2	131	4
		AL583889.2.89451.105061	8.00E-28	428=>513	6189=>6274	91.8	85	5
		AL583889.2.89451.105061	6.00E-44	510=>601	8350=>8441	100	91	6
		AL583889.2.89451.105061	5.00E-26	598=>659	11292=>11353	100	61	7
34. dJ591N18.1	NM_025628.1†	AC079570.1.114235.123330	1.00E-50	2=>103	5992=>6093	100	101	1
		AC079570.1.114235.123330	6.00E-59	101=>216	7527=>7642	100	115	2
		AC079570.1.114235.123330	8.00E-52	216=>319	8626=>8729	100	103	3
		AC079570.1.127807.134884	e-103	317=>506	2404=>2593	100	189	4
35. SF3A1	NM_026175.1†	AC074332.6.1.62799	8.00E-96	2=>180	44967=>45145	100	178	1
		AC074332.6.1.62799	4.00E-64	177=>302	45637=>45762	100	125	2
		AC074332.6.1.62799	e-113	300=>512	51088=>51300	99.5	212	3
		AC074332.6.1.62799	e-144	509=>768	52226=>52485	100	259	4

		AC074332.6.1.62799	4.00E-36	767=>845	55655=>55733	100	78	5
		AC074332.6.1.62799	6.00E-35	919=>995	57750=>57826	100	76	6
		AC074332.6.1.62799	e-103	993=>1183	58625=>58815	100	190	7
		AC074332.6.1.62799	1.00E-60	1181=>1300	59629=>59748	100	119	8
		AC074332.6.1.62799	e-102	1299=>1488	59988=>60177	100	189	9
		AC074332.6.1.62799	9.00E-65	1485=>1611	60863=>60989	100	126	10
		AC074332.6.1.62799	e-137	1607=>1854	61074=>61321	100	247	11
		AC074332.6.1.62799	e-113	1854=>2062	62007=>62215	100	208	12
		AC074332.6.1.62799	1.00E-70	2060=>2196	62663=>62799	100	136	13
		AC079571.1.186896.245130	4.00E-79	2220=>2469	46681=>46927	90.8	249	14
36. SEC14L2	BC005759.1	AC073351.8.73364.109182	1.00E-51	1=>105	14492=>14388	100	104	1
		AC073351.8.73364.109182	1.00E-35	104=>181	12521=>12444	100	77	2
		AC074332.6.158099.169572	2.00E-28	224=>289	6696=>6631	100	65	3
		AC074332.6.158099.169572	e-100	284=>474	6419=>6229	99.4	190	4
		AC074332.6.158099.169572	2.00E-47	473=>570	4438=>4341	100	97	5
		AC074332.6.158099.169572	1.00E-26	569=>631	4218=>4156	100	62	6
		AC074332.6.158099.169572	2.00E-40	630=>715	3774=>3689	100	85	7
		AC074332.6.62820.102469	2.00E-53	715=>822	474=>581	100	107	8
		AC074332.6.62820.102469	6.00E-75	821=>964	661=>804	100	143	9
		AC074332.6.62820.102469	3.00E-92	960=>1132	916=>1088	100	172	10
		AC074332.6.62820.102469	0	1133=>1752	6041=>6660	98.8	619	11
		AC074332.6.62820.102469	0	1767=>2205	6675=>7113	97.4	438	12
		AC074222.8.118396.122551	e-171	2200=>2508	4085=>3777	99.6	308	13
37. AC005500.4	BC005800.1	AC087802.5.93295.187315	e-146	64=>327	23753=>23490	100	263	1

		AC087802.5.93295.187315	7.00E-90	325=>493	16435=>16267	100	168	2
		AC087802.5.93295.187315	0	491=>1215	11737=>11012	99.7	724	3
		AC087802.5.93295.187315	e-107	1211=>1408	3934=>3737	100	197	4
		AC079044.11.18222.24247	e-131	1404=>1642	4439=>4201	100	238	5
		AC079044.11.18222.24247	0	1637=>2594	1151=>194	99.4	957	6
38. UFD1L	BC006630.1(11)	AC083895.13.115418.189347	9.00E-38	1=>81	51494=>51574	100	80	1
		AC083895.13.115418.189347	2.00E-69	81=>214	53986=>54119	100	133	2
		AC083895.13.115418.189347	3.00E-62	248=>369	57060=>57181	100	121	3
		AC083895.13.115418.189347	3.00E-68	369=>500	60195=>60326	100	131	4
		AC083895.13.115418.189347	5.00E-33	501=>573	62380=>62452	100	72	5
		AC083895.13.115418.189347	1.00E-33	572=>645	64145=>64218	100	73	6
		AC083895.13.115418.189347	2.00E-29	643=>709	64912=>64978	100	66	7
		AC083895.13.115418.189347	2.00E-42	757=>845	66178=>66266	100	88	8
		AC083895.13.115418.189347	4.00E-40	844=>928	66794=>66878	100	84	9
		AC083895.13.115418.189347	5.00E-61	926=>1045	73394=>73513	100	119	10
39. MFNG	BC010983.1*(7)	AC026386.20.201618.217833	0	12=>450	11058=>10620	99.7	438	1
		AC026386.20.201618.217833	2.00E-21	447=>500	4796=>4743	100	53	2
		AC026386.20.201618.217833	5.00E-53	498=>604	4524=>4418	100	106	3
		AC026386.20.201618.217833	7.00E-83	600=>756	2073=>1917	100	156	4
40. GSTT1	BC012254.1(4)	AC068241.3.1.36249	7.00E-60	1=>118	3551=>3668	100	117	1
		AC068241.3.1.36249	1.00E-42	115=>203	4920=>5008	100	88	2
		AC068241.3.1.36249	6.00E-82	202=>356	7822=>7976	100	154	3
		AC068241.3.1.36249	3.00E-96	353=>531	15078=>15256	100	178	4
		AC068241.3.1.36249	0	532=>950	17818=>18236	100	418	5

41. LGALS1	NM_008495.1*(4)	AC026386.20.1.129738	5.00E-38	1=>81	23810=>23730	100	80	1
		AC026386.20.1.129738	2.00E-37	81=>160	22565=>22486	100	79	2
		AC026386.20.1.129738	2.00E-95	159=>335	20904=>20728	100	176	3
		AC026386.20.1.129738	e-109	331=>530	20539=>20340	100	199	4
42. MMP11	NM_008606.1*(8)	AC005302.68.66899.205190	e-127	130=>361	119233=>119464	100	231	1
		AC005302.68.66899.205190	5.00E-75	359=>506	120251=>120398	99.3	147	2
		AC005302.68.66899.205190	7.00E-71	503=>639	120479=>120615	100	136	3
		AC005302.68.66899.205190	e-134	638=>880	120753=>120995	100	242	4
		AC005302.68.66899.205190	e-121	879=>1100	121114=>121335	100	221	5
		AC005302.68.66899.205190	e-145	1096=>1356	122064=>122324	100	260	6
		AC005302.68.66899.205190	0	1355=>2186	123627=>124459	99.8	831	7
43. NCF4	NM_008677.1	AL589692.3.40022.166145	e-138	1=>254	30663=>30410	99.6	253	1
		AL589692.3.40022.166145	5.00E-40	253=>341	25117=>25029	98.8	88	2
		AL589692.3.40022.166145	8.00E-79	340=>493	24635=>24482	99.3	153	3
		AL589692.3.40022.166145	2.00E-33	491=>564	22203=>22130	100	73	4
		AL589692.3.40022.166145	7.00E-67	563=>692	20712=>20583	100	129	5
		AL589692.3.40022.166145	2.00E-24	693=>751	20045=>19987	100	58	6
		AL589692.3.40022.166145	4.00E-50	749=>850	19568=>19467	100	101	7
		AL589692.3.40022.166145	3.00E-69	848=>981	15243=>15110	100	133	8
		AL589692.3.40022.166145	3.00E-29	981=>1047	14893=>14827	100	66	9
		AL589692.3.40022.166145	0	1045=>1425	13671=>13291	99.2	380	10
44. AC006547.2	NM_008307.1	AC084822.4.69771.76674	0	42=>1353	6285=>4979	98.1	1311	1
		AC084822.4.69771.76674	8.00E-56	1359=>1470	4554=>4443	100	111	2
		AC084822.4.69771.76674	2.00E-99	1468=>1652	4188=>4004	100	184	3

		AC084822.4.69771.76674	3.00E-58	1652=>1767	3850=>3735	100	115	4
		AC084822.4.69771.76674	5.00E-60	1765=>1883	3649=>3531	100	118	5
		AC084822.4.69771.76674	8.00E-59	1882=>1998	3458=>3342	100	116	6
		AC084822.4.69771.76674	5.00E-63	1991=>2118	2884=>2757	99.2	127	7
		AC084822.4.69771.76674	6.00E-35	2118=>2194	2607=>2531	100	76	8
		AC084822.4.69771.76674	2.00E-59	2194=>2311	2290=>2173	100	117	9
		AC084822.4.69771.76674	7.00E-47	2312=>2408	2084=>1988	100	96	10
		AC084822.4.69771.76674	0	2409=>2810	1908=>1506	99.2	401	11
45. IL17R	NM_008359.1*(12)	AC078896.10.87145.169250	e-128	26=>259	80009=>80242	100	233	1
		AC078896.10.27686.54563	1.00E-76	284=>430	6841=>6987	100	146	2
		AC078896.10.27686.54563	6.00E-57	430=>543	7668=>7781	100	113	3
		AC078896.10.27686.54563	3.00E-65	543=>670	8204=>8331	100	127	4
		AC078896.10.27686.54563	1.00E-88	716=>882	9368=>9534	100	166	5
		AC078896.10.27686.54563	7.00E-41	882=>968	10805=>10891	100	86	6
		AC078896.10.27686.54563	2.00E-41	964=>1051	11309=>11396	100	87	7
		AC078896.10.27686.54563	4.00E-55	1062=>1172	12290=>12400	100	110	8
		AC078896.10.27686.54563	0	1213=>2756	14900=>16443	100	1543	9
		AC078896.10.27686.54563	0	2834=>3264	16521=>16951	100	430	10
46. CACNG2	NM_007583.1*(3)	AL589650.6.1.138700	4.00E-53	1=>107	18351=>18457	100	106	1
		AL589650.6.1.138700	0	157=>601	18508=>18951	98.2	444	2
		AL589650.6.1.138700	4.00E-41	599=>685	124161=>124247	100	86	3
		AC084828.1.44044.60396	1.00E-74	685=>827	11655=>11513	100	142	4
		AC084828.1.44044.60396	0	823=>1443	10351=>9731	100	620	5
47. DGCR2	NM_010048.1*(11)	AC078895.12.59859.213034	e-115	2=>212	58075=>57865	100	210	1

		AC078895.12.59859.213034	4.00E-64	210=>335	39167=>39042	100	125	2
		AC078895.12.59859.213034	2.00E-62	334=>456	26001=>25879	100	122	3
		AC078895.12.59859.213034	e-121	456=>676	25055=>24835	100	220	4
		AC078895.12.59859.213034	2.00E-35	676=>753	23810=>23733	100	77	5
		AC078895.12.59859.213034	5.00E-97	751=>931	23406=>23226	100	180	6
		AC078895.12.59859.213034	e-110	930=>1133	16270=>16067	100	203	7
		AC078895.12.59859.213034	1.00E-76	1132=>1286	11613=>11459	98.7	154	8
		AC078895.12.59859.213034	e-130	1287=>1523	10684=>10448	100	236	9
		AC078895.12.59859.213034	0	1519=>2987	9258=>7786	99.4	1468	10
48. FBLN1	NM_010180.1*(17)	AL583891.6.1.170197	6.00E-72	48=>202	112339=>112187	98.7	154	1
		AL583891.6.1.170197	3.00E-49	201=>309	96499=>96391	98.1	108	2
		AL583891.6.1.170197	6.00E-72	306=>444	94154=>94016	100	138	3
		AL583891.6.1.170197	7.00E-93	443=>616	91435=>91262	100	173	4
		AL583891.6.1.170197	5.00E-26	615=>676	88843=>88782	100	61	5
		AL583891.6.1.170197	4.00E-51	672=>775	87604=>87501	100	103	6
		AL583891.6.1.170197	1.00E-72	774=>913	86986=>86847	100	139	7
		AL583891.6.1.170197	5.00E-60	912=>1051	85864=>85725	95	139	8
		AL583891.6.1.170197	2.00E-75	1051=>1195	80794=>80650	100	144	9
		AL583891.6.1.170197	3.00E-64	1195=>1324	79939=>79810	99.2	129	10
		AL583891.6.1.170197	8.00E-65	1324=>1450	79108=>78982	100	126	11
		AL583891.6.1.170197	8.00E-62	1449=>1570	77783=>77662	100	121	12
		AL583891.6.1.170197	6.00E-69	1569=>1702	76392=>76259	100	133	13
		AL583891.6.1.170197	3.00E-64	1701=>1826	74220=>74095	100	125	14
		AL583891.6.1.170197	2.00E-74	1827=>1969	55144=>55002	100	142	15

		AL583891.6.1.170197	4.00E-70	1969=>2104	53154=>53019	100	135	16
		AL583891.6.1.170197	2.00E-81	2101=>2279	33364=>33187	97.2	178	17
		AL583891.6.1.170197	6.00E-81	2274=>2443	33179=>33010	97.6	169	18
		AL583891.6.1.170197	3.00E-55	2458=>2611	32999=>32848	94.2	153	19
		AL583891.6.1.170197	9.00E-34	2624=>2706	32829=>32747	97.5	82	20
49. AC006548.8	NM_010248.1	AC074048.16.124855.147843	8.00E-83	1=>157	17292=>17448	100	156	1
		AC084021.9.99861.161141	e-169	157=>457	49295=>49595	100	300	2
		AC084021.9.1.99760	e-136	453=>699	27724=>27970	100	246	3
		AC084021.9.1.99760	0	697=>1273	54753=>55329	99.6	576	4
		AC084021.9.1.99760	1.00E-32	1272=>1344	56224=>56296	100	72	5
		AC084021.9.1.99760	e-145	1345=>1609	57298=>57562	99.6	264	6
		AC084021.9.1.99760	1.00E-44	1608=>1700	58294=>58386	100	92	7
		AC084021.9.1.99760	2.00E-52	1701=>1806	58951=>59056	100	105	8
		AC084021.9.1.99760	6.00E-65	1806=>1932	60076=>60202	100	126	9
		AC084021.9.1.99760	2.00E-86	1931=>2093	60801=>60963	100	162	10
50. LARGE	NM_010687.1[†*](4)	AC079554.1.52238.57901	7.00E-32	1063=>1176	177=>59	93.2	113	1
		AC079554.1.17801.21528	e-158	1622=>1904	1591=>1309	100	282	2
		AC079576.1.47763.65447	3.00E-80	1900=>2052	5319=>5167	100	152	3
		AC079576.1.15219.19594	e-108	2048=>2247	1183=>1382	100	199	4
		AC079576.1.15219.19594	0	2245=>3172	3434=>4357	99.4	927	5
51. PDBFB	NM_011057.1[*](6)	AC021061.9.1.22305	6.00E-47	1=>105	10080=>9976	98	104	1
		AC021061.9.1.22305	0	155=>786	9924=>9292	97.6	631	2
		AC021061.9.1.22305	3.00E-95	851=>1040	9227=>9038	98.4	189	3
		AC021061.9.1.22305	4.00E-48	1038=>1136	670=>572	100	98	4

		AC021061.9.168215.173345	1.00E-44	1135=>1227	3550=>3458	100	92	5
		AC021061.9.168215.173345	e-112	1225=>1435	2024=>1814	99.5	210	6
		AC021061.9.168215.173345	3.00E-76	1428=>1577	613=>464	99.3	149	7
		AC021061.9.39522.52254	9.00E-80	1576=>1727	2217=>2368	100	151	8
		AC021061.9.39522.52254	0	1724=>2404	2663=>3349	98.2	680	9
52. RANGAP1	NM_011241.1*(17)	AC074222.8.125450.128857	6.00E-77	1=>265	1396=>1135	89	264	1
		AC074222.8.125450.128857	0	285=>1043	1120=>370	89.3	758	2
		AC074222.8.278501.279947	3.00E-26	1243=>1357	439=>329	88.6	114	3
53. SMARCB1	NM_011418.1*(9)	AC005302.68.66899.205190	e-161	1=>288	126155=>126442	100	287	1
		AC005302.68.66899.205190	7.00E-61	285=>425	130830=>130970	95	140	2
		AC005302.68.66899.205190	5.00E-68	425=>556	132549=>132680	100	131	3
		AC005302.68.66899.205190	2.00E-73	553=>693	136080=>136220	100	140	4
		AC005302.68.66899.205190	3.00E-66	693=>821	137457=>137585	100	128	5
54. RPL3	NM_013762.1*(9)	AC021061.9.184782.191131	e-105	12=>205	1658=>1851	100	193	1
		AC021061.9.184782.191131	8.00E-91	206=>375	2458=>2627	100	169	2
		AC021061.9.184782.191131	4.00E-71	374=>510	3209=>3345	100	136	3
		AC021061.9.184782.191131	e-102	509=>698	3688=>3877	100	189	4
		AC021061.9.184782.191131	3.00E-84	696=>858	4390=>4552	99.3	162	5
		AC021061.9.184782.191131	5.00E-52	857=>961	5432=>5536	100	104	6
		AC021061.9.184782.191131	3.00E-47	960=>1056	5815=>5911	100	96	7
		AC021061.9.184782.191131	9.00E-63	1055=>1177	6134=>6256	100	122	8
		AC021061.9.122739.131408	6.00E-27	1181=>1276	8484=>8388	93.8	95	9
55. SYN3	NM_013722.1	AC063968.3.156279.162887	3.00E-76	968=>1113	362=>507	100	145	1
		AC063968.3.27805.49486	1.00E-35	1113=>1190	20441=>20364	100	77	2

		AC063968.3.27805.49486	5.00E-53	1186=>1292	16270=>16164	100	106	3
		AC063968.3.27805.49486	2.00E-71	1289=>1426	12072=>11935	100	137	4
		AC063968.3.27805.49486	2.00E-40	1425=>1514	11190=>11101	98.8	89	5
		AC063968.3.27805.49486	e-164	1512=>1805	4803=>4510	100	293	6
		AC063968.3.27805.49486	1.00E-75	1806=>1954	2798=>2650	99.3	148	7
56. TR	NM_013711.1(10)	AC003066.1.1.176974	9.00E-61	1=>147	12225=>12368	96.5	146	1
		AC003066.1.1.176974	9.00E-30	145=>212	15617=>15684	100	67	2
		AC003066.1.1.176974	5.00E-44	723=>814	24879=>24970	100	91	3
		AC003066.1.1.176974	1.00E-96	811=>990	26410=>26589	100	179	4
		AC003066.1.1.176974	1.00E-72	988=>1127	39138=>39277	100	139	5
		AC003066.1.1.176974	8.00E-49	1125=>1224	42542=>42641	100	99	6
		AC003066.1.1.176974	7.00E-46	1223=>1317	43184=>43278	100	94	7
		AC003066.1.1.176974	2.00E-33	1314=>1387	45070=>45143	100	73	8
		AC003066.1.1.176974	5.00E-50	1386=>1487	45910=>46011	100	101	9
		AC003066.1.1.176974	e-113	1486=>1693	48041=>48248	100	207	10
		AC003066.1.1.176974	e-116	1690=>1902	49245=>49457	100	212	11
57. Unknown	NM_011820.1	AC087540.17.144500.200123	0	1=>487	17966=>18452	100	486	1
		AC087540.17.144500.200123	2.00E-68	487=>619	31189=>31321	100	132	2
		AC087540.17.144500.200123	9.00E-49	616=>715	31541=>31640	100	99	3
		AC087540.17.144500.200123	e-108	712=>910	32534=>32732	100	198	4
		AC087540.17.144500.200123	4.00E-85	909=>1069	33211=>33371	100	160	5
		AC087540.17.144500.200123	1.00E-78	1067=>1216	33824=>33973	100	149	6
		AC087540.17.144500.200123	5.00E-72	1214=>1352	37319=>37457	100	138	7
		AC087540.17.144500.200123	e-106	1350=>1546	37766=>37962	100	196	8

		AC087540.17.144500.200123	6.00E-53	1547=>1653	38415=>38521	100	106	9
		AC087540.17.144500.200123	3.00E-58	1653=>1772	38682=>38801	99.1	119	10
		AC087540.17.144500.200123	3.00E-58	1769=>1884	38994=>39109	100	115	11
		AC087540.17.144500.200123	0	1881=>2339	43297=>43755	100	458	12
58. NUP50	NM_016714.1[†*](7)	AL513352.3.40623.68090	2.00E-30	1=>69	24186=>24118	100	68	1
		AL513352.3.40623.68090	1.00E-40	68=>153	22194=>22109	100	85	2
		AL513352.3.40623.68090	e-100	154=>339	18029=>17844	100	185	3
		AL513352.3.40623.68090	0	337=>997	16701=>16041	99.6	660	4
		AL513352.3.40623.68090	2.00E-39	996=>1079	14079=>13996	100	83	5
		AL513352.3.40623.68090	6.00E-61	1080=>1199	13206=>13087	100	119	6
		AL513352.3.40623.68090	e-111	1197=>1401	11920=>11716	100	204	7
59. CSNK1E	NM_013767.2[*](9)	AC090533.3.163966.170849	1.00E-34	1=>76	1173=>1098	100	75	1
		AC090533.3.116476.123525	1.00E-43	76=>166	4987=>4897	100	90	2
		AC090533.3.216806.218691	9.00E-54	164=>275	1720=>1831	99.1	111	3
		AC090533.3.241156.241602	2.00E-79	274=>424	369=>219	100	150	4
		AC090533.3.62391.75535	e-124	424=>653	10142=>9913	99.5	229	5
		AC090533.3.62391.75535	6.00E-92	653=>824	9325=>9154	100	171	6
		AC090533.3.62391.75535	8.00E-79	825=>974	8642=>8493	100	149	7
		AC090533.3.62391.75535	e-102	974=>1167	4664=>4471	99.4	193	8
		AC090533.3.62391.75535	5.00E-71	1166=>1306	4378=>4238	99.2	140	9
		AC090533.3.62391.75535	1.00E-28	1305=>1370	3624=>3559	100	65	10
60. PLA2G6	NM_016915.1	AC090533.3.43795.62370	e-114	1=>210	13861=>14070	100	209	1
		AC090533.3.123546.132131	3.00E-98	423=>609	6818=>6632	99.4	186	2
		AC090533.3.123546.132131	e-102	607=>799	3697=>3505	99.4	192	3

		AC090533.3.123546.132131	1.00E-47	797=>894	2248=>2151	100	97	4
		AC090533.3.123546.132131	2.00E-99	893=>1077	1306=>1122	100	184	5
		AC090533.3.123546.132131	1.00E-54	1077=>1186	650=>541	100	109	6
		AC090533.3.202012.208156	3.00E-36	1187=>1265	2449=>2527	100	78	7
		AC090533.3.188774.193447	2.00E-89	1264=>1431	205=>372	100	167	8
		AC090533.3.188774.193447	8.00E-80	1430=>1581	1953=>2104	100	151	9
		AC090533.3.208177.212539	7.00E-71	1581=>1717	4127=>4263	100	136	10
		AC090533.3.97596.106417	2.00E-83	1716=>1873	173=>330	100	157	11
		AC090533.3.97596.106417	2.00E-90	1871=>2040	1675=>1844	100	169	12
		AC090533.3.97596.106417	3.00E-33	2041=>2114	2277=>2350	100	73	13
		AC090533.3.97596.106417	2.00E-77	2112=>2259	2562=>2709	100	147	14
61. EIF3S7	NM_018749.1*(17)	AL589650.6.138801.226259	6.00E-37	9=>88	27967=>28046	100	79	1
		AL589650.6.138801.226259	2.00E-71	84=>221	30135=>30272	100	137	2
		AL589650.6.138801.226259	6.00E-71	266=>402	31288=>31424	100	136	3
		AL589650.6.138801.226259	3.00E-42	400=>488	31957=>32045	100	88	4
		AL589650.6.138801.226259	6.00E-34	487=>561	32515=>32589	100	74	5
		AL589650.6.138801.226259	2.00E-58	559=>674	33709=>33824	100	115	6
		AL589650.6.138801.226259	8.00E-55	675=>804	34575=>34707	96.9	129	7
		AL589650.6.138801.226259	1.00E-78	803=>952	35400=>35549	100	149	8
		AL589650.6.138801.226259	1.00E-68	952=>1084	36156=>36288	100	132	9
		AL589650.6.138801.226259	4.00E-41	1083=>1169	36795=>36881	100	86	10
		AL589650.6.138801.226259	2.00E-67	1169=>1299	36983=>37113	100	130	11
		AL589650.6.138801.226259	3.00E-76	1299=>1444	38630=>38775	100	145	12
		AL589650.6.138801.226259	e-131	1442=>1679	38931=>39168	100	237	13

		AL589650.6.138801.226259	5.00E-53	1735=>1849	39613=>39726	99.1	114	14
62. TXN2	NM_019913.1*(4)	AL589650.6.138801.226259	e-149	50=>317	70771=>71038	100	267	1
		AL589650.6.138801.226259	1.00E-64	315=>440	74061=>74186	100	125	2
		AL589650.6.138801.226259	0	437=>1120	82868=>83551	98.6	683	3
63. SC02	AK002487	AC079487.1.215358.230320	6.00E-60	1=>127	3234=>3360	98.4	126	1
		AC079487.1.215358.230320	0	144=>660	3360=>3877	96.7	516	2
		AC079487.1.215358.230320	e-119	639=>901	3874=>4136	95.8	262	3
		AC079487.1.215358.230320	0	913=>1300	4137=>4524	96.1	387	4
		AC079487.1.215358.230320	3.00E-37	1320=>1424	4520=>4624	94.2	104	5
		AC079487.1.215358.230320	0	1426=>1994	4614=>5182	97.1	568	6
64. RANBP1	AK002989	AC012526.1.1.186272	3.00E-56	3=>114	57216=>57105	100	111	1
		AC012526.1.1.186272	5.00E-73	112=>251	55985=>55846	100	139	2
		AC012526.1.1.186272	1.00E-83	251=>408	53916=>53759	100	157	3
		AC012526.1.1.186272	2.00E-66	409=>537	50431=>50303	100	128	4
		AC012526.1.1.186272	2.00E-29	538=>604	49517=>49451	100	66	5
		AC012526.1.1.186272	2.00E-60	602=>720	48753=>48635	100	118	6
		AC012526.1.1.186272	9.00E-75	866=>1012	48489=>48343	99.3	146	7
65. E46L	AK003530	AL583885.3.112567.117111	5.00E-65	2=>128	4309=>4183	100	126	1
		AL583885.3.89366.99895	e-105	127=>320	4409=>4216	100	193	2
		AL583885.3.71983.89265	2.00E-40	317=>402	15893=>15808	100	85	3
		AL583885.3.71983.89265	2.00E-49	401=>501	5847=>5747	100	100	4
		AL583885.3.71983.89265	3.00E-85	500=>660	4623=>4463	100	160	5
		AL583885.3.62182.71882	4.00E-38	659=>740	7984=>7903	100	81	6
		AL583885.3.62182.71882	5.00E-90	738=>906	3899=>3731	100	168	7

		AL583885.3.58582.62081	2.00E-55	904=>1014	2683=>2573	100	110	8
		AL583885.3.58582.62081	2.00E-92	1014=>1186	983=>811	100	172	9
		AL583885.3.9226.26846	2.00E-27	1185=>1248	5446=>5383	100	63	10
66. dJ1170K4.2	AK004939	AL590144.3.121669.130414	3.00E-21	1=>54	2671=>2618	100	53	1
		AL590144.3.105821.121568	e-131	48=>290	15215=>14973	99.5	242	2
		AL590144.3.105821.121568	4.00E-70	288=>423	11245=>11110	100	135	3
		AL590144.3.105821.121568	1.00E-30	423=>492	10373=>10304	100	69	4
		AL590144.3.105821.121568	2.00E-96	491=>670	9789=>9610	100	179	5
		AL590144.3.105821.121568	e-112	712=>918	4984=>4778	100	206	6
		AL590144.3.105821.121568	1.00E-70	918=>1054	4146=>4010	100	136	7
		AL590144.3.105821.121568	3.00E-58	1054=>1169	2770=>2655	100	115	8
		AL590144.3.105821.121568	8.00E-56	1167=>1278	2451=>2340	100	111	9
		AL590144.3.90945.105720	3.00E-77	1276=>1423	11261=>11114	100	147	10
		AL590144.3.90945.105720	5.00E-48	1424=>1522	10753=>10655	100	98	11
		AL590144.3.90945.105720	3.00E-58	1522=>1637	9849=>9734	100	115	12
		AL590144.3.90945.105720	6.00E-60	1635=>1753	8665=>8547	100	118	13
		AL590144.3.90945.105720	7.00E-78	1753=>1922	8297=>8128	95.8	169	14
		AL590144.3.90945.105720	e-154	1920=>2196	7007=>6731	100	276	15
		AL590144.3.90945.105720	6.00E-72	2194=>2332	5196=>5058	100	138	16
		AL590144.3.90945.105720	0	2330=>3028	4825=>4127	100	698	17
67. dJ1039K5.6	AK006539	AL589670.3.1.8618	3.00E-99	254=>437	1429=>1612	100	183	1
		AL589670.3.1.8618	2.00E-66	437=>565	2153=>2281	100	128	2
		AL589670.3.1.8618	1.00E-49	566=>666	2465=>2565	100	100	3
		AL589670.3.1.8618	0	664=>1130	2693=>3159	100	466	4

68. cE81G9.2	AK006856	AL590144.3.45304.78127	1.00E-46	43=>138	6410=>6315	100	95	1
		AL590144.3.45304.78127	2.00E-29	135=>201	847=>781	100	66	2
		AC087867.2.193495.195983	0	197=>630	843=>1276	100	433	3
		AC087867.2.193495.195983	2.00E-79	625=>775	1959=>2109	100	150	4
		AC087867.2.114736.121529	2.00E-57	773=>886	567=>454	100	113	5
		AC087867.2.114736.121529	4.00E-86	885=>1046	272=>111	100	161	6
69. RAYL	AK011196	AL589692.3.40022.166145	0	5=>339	101396=>101730	100	334	1
		AL589692.3.40022.166145	6.00E-39	336=>418	107503=>107585	100	82	2
		AL589692.3.40022.166145	2.00E-26	418=>479	108368=>108429	100	61	3
		AL589692.3.40022.166145	7.00E-26	478=>538	109464=>109524	100	60	4
		AL589692.3.40022.166145	1.00E-61	538=>658	110233=>110353	100	120	5
		AL589692.3.40022.166145	7.00E-57	654=>766	111196=>111308	100	112	6
		AL589692.3.40022.166145	e-135	766=>1010	115789=>116033	100	244	7
70. PVALB	AK013561(4)	AL589692.3.40022.166145	1.00E-33	2=>75	71339=>71412	100	73	1
		AL589692.3.40022.166145	1.00E-70	75=>210	72843=>72978	100	135	2
		AL589692.3.40022.166145	1.00E-55	208=>318	74161=>74271	100	110	3
		AL589692.3.40022.166145	e-144	317=>583	83801=>84067	99.2	266	4
		AL589692.3.40022.166145	e-158	602=>899	84086=>84383	98.6	297	5
71. ARHGAP8	AK014171	AL513352.3.185476.205053	6.00E-52	465=>569	4576=>4680	100	104	1
		AL513352.3.185476.205053	2.00E-51	567=>678	5261=>5372	98.2	111	2
		AL513352.3.185476.205053	5.00E-34	679=>753	13915=>13989	100	74	3
		AL513352.3.185476.205053	1.00E-37	751=>831	14551=>14631	100	80	4
		AL513352.3.185476.205053	1.00E-68	829=>961	15341=>15473	100	132	5
		AL513352.3.185476.205053	2.00E-52	960=>1065	18239=>18344	100	105	6

		AL513352.3.205154.208242	0	1064=>1484	412=>832	100	420	7
72. dJ345P10.4	AK019850	AL513354.10.8839.226926	5.00E-97	1=>181	138432=>138252	100	180	1
		AL513354.10.8839.226926	9.00E-71	180=>316	135545=>135409	100	136	2
		AL513354.10.8839.226926	e-128	315=>548	126799=>126566	100	233	3
		AL513354.10.8839.226926	3.00E-49	545=>645	120634=>120534	100	100	4
		AL513354.10.8839.226926	3.00E-61	644=>764	116035=>115915	100	120	5
		AL513354.10.8839.226926	5.00E-97	765=>945	106681=>106501	100	180	6
		AL513354.10.8839.226926	1.00E-79	943=>1094	102131=>101980	100	151	7
		AL513354.10.8839.226926	e-111	1093=>1297	98492=>98288	100	204	8
		AL513354.10.8839.226926	e-105	1295=>1488	95293=>95100	100	193	9
		AL513354.10.8839.226926	8.00E-93	1543=>1716	81941=>81768	100	173	10
		AL513354.10.8839.226926	e-135	1715=>1959	74777=>74533	100	244	11
		AL513354.10.8839.226926	4.00E-79	1960=>2110	73639=>73489	100	150	12
		AL513354.10.8839.226926	e-100	2110=>2295	73358=>73173	100	185	13
		AL513354.10.8839.226926	7.00E-81	2295=>2448	70726=>70573	100	153	14
		AL513354.10.8839.226926	0	2444=>2850	69510=>69104	100	406	15

†Mouse mRNA sequences that have unknown chromosomal positions assigned to their corresponding mouse ENSEMBL contigs

Supplementary Table 3: EST annotations of mouse exon- skipping events

Mouse ENSEMBL transcript or Genbank Identifier	Skipped exon(s)	Genbank EST accession	Clone library	E-value*
1. ENSMUST00000001834	3-8	W42119	Soares mouse p3NMF19.5	$9x10^{-81}$
2. BC002031.1	4-5	W63912	Soares mouse embryo NbME13.5 14.5	$1x10^{-159}$
3. NM_023120.1	2-6	W75310,W75333	Soares mouse embryo NbME13.5 14.5, Soares mouse embryo NbME13.5 14.5	0, $1x10^{-143}$
4. NM_011418.1	2-4	AA013500	Soares mouse placenta 4NbMP13.5 14.5	$1x10^{-121}$
5. AK002487	3-4	AA032389	Soares mouse embryo NbME13.5 14.5	$1x10^{-103}$

* E-values were obtained from BLASTN results by querying mouse dbEST databases with mouse virtual transcripts (See Methods Aiii for construction of virtual transcripts).