2005

# Comparative Analysis of Alternative Splicing in Homo sapiens, Mus musculus and Rattus norvegicus Transcriptomes

Bahar Taneri

CELEBRATING 100 YEARS
1901
THE
ROCKEFELLER
UNIVERSITY
PRO·BONO·HUMANI·GENERIS

# Comparative Analysis of Alternative Splicing in *Homo sapiens, Mus musculus* and *Rattus norvegicus* Transcriptomes

A thesis presented to the faculty of The Rockefeller University
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

by

Bahar Taneri

The Rockefeller University
New York
June, 2005

# DEDICATION

This thesis is dedicated to the three most important people in my life... My father Ertug Taneri, my mother Remziye Taneri and my brother Cem Taneri... For their enormous support and encouragement, for always believing in me and for their endless love...

# ACKNOWLEDGMENTS

I would like to thank my advisor Terry Gaasterland for giving me the chance to work for her, for being a great mentor and a great role model. Over the years, Terry has provided me with every opportunity and all the guidance I needed to learn and grow as a scientist. I absolutely could not have asked for a better advisor. I am truly grateful to have been her student.

I would like to acknowledge my faculty advisory committee members, Robert Darnell, Ali Brivanlou and my chairman Shai Shaham for all their advises over the years and Nikolaus Rajewsky for being the external member on my thesis committee.

The Rockefeller University Dean's Office members have assisted my studies and research in many ways. I would like to acknowledge Dean Sidney Strickland, Marta Delgado, Kristen Cullen, Sue Ann Chong and Cris Rosario for their assistance.

Several great scientists outside of Rockefeller have influenced my research. I would like to thank Xiang-Dong Fu from University of San Diego, Peter Byers from University of Washington and John Reed from The Burnham Institute for inspiration to focus on pathological splicing due to cancer and for their guidance in examining disease related alternative splicing. Joe Sorge from Stratagene has been especially helpful with useful discussions on analysis of tissue-specificity of mouse transcription factor isoforms.

Robert Darnell and his student Jernej Ule have been great collaborators on the Nova projects. I would like to thank them for sharing their data with me and providing the opportunity for me to learn from them. Robert Darnell has been a second mentor to me. I truly appreciate his guidance on studies of tissue-specific regulation of alternative splicing.

Over the years, Gaasterland Lab members have made the work environment fun and friendly. I would like to acknowledge Anna Neill, Hannah Schmidt-Glenewinkel, Doug Gaasterland, Jay Kanopka and Xiujie Wang. I would especially like to thank Nancy Sosa for sharing the lab space with me and for greeting me with a smiley face every morning when I came to the lab. Nancy's company has made the workdays very enjoyable. I want to thank Yupu Liang not only for her help with Nova binding-motif visualizations, but also for making the coffee breaks at Java Girl a lot of fun. I am happy to have Nancy and Yupu as my friends.

I want to particularly thank Ben Snyder, an extremely talented programmer and a great coworker, who helped me enormously with my work. I am truly grateful to Ben for his help with my research. I was fortunate to have worked with Alexey Novoradovsky. I would like to acknowledge his

help with development of the web-interfaces for our databases and his help with the analysis of alternative splicing in orthologous genes.

Over the years many friends made my life in New York fun and enjoyable. I would like to thank Gulya Fayzikhodjeava for being a very nice and caring person, Makoto Ishii for throwing fun parties, Ann Tang, Pinar Akpinar, Annita Charalambous and Elenora Market for being good friends. Deniz Yucel, Caglar Tanrikulu and Brenda Heyman have been parts of my Ph.D. from long distances. I would like to thank them for their friendship and I especially thank Deniz for her support.

I am very lucky to have known Agnes Viale, Shirly Pinto, Dvora Schmulewitz, Joe Scandura and Ed Yatco who are absolutely great friends. They have been a big part of my life in New York. I want to especially thank Agnes and Shirly for always thinking about me and being there for me. Shirly and Ed always welcomed me to their house and to their family. Agnes and Joe particularly provided great support for me with our weekly dinners during the last year of my Ph.D. I am grateful for all their friendship and for all the good memories we shared during the past five years.

I would like to thank all my family. My grandmothers Emine Ipar and Feriha Taneri for keeping me in their prayers, my cousins Niyazi Oztoprak and Suphi Taneri for their support and all members of The Taneri Family and The Ipar Family.

Esra Asilmaz and Silvia Novelli have been my family in the States. They were there for me every single day of my graduate studies. I am very lucky to have them as my best friends. Silvia, as a very caring and fun person, has always created good memories for me with her humor and has been there for me to depend on. Esra is the greatest friend who helped me enormously with the ups and downs of my Ph.D., more than anybody else. As my roommate for four years, Esra always understood me and was always there for me when I needed her advise and support. I will always cherish the time we spent in New York and I want to thank her for being the wonderful person she is.

I can't thank enough to my father Ertug Taneri, my mother Remziye Taneri and my brother Cem Taneri for their enormous love, help and support and for always being there for me. My parents have always believed in me and provided me with every single opportunity they could. Cem has been a big brother to me who understood me and helped me out whenever I needed. I am grateful for the best parents and the best brother in the world. It is because of Anne, Baba and Cem that I was able to come this far.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# PUBLICATIONS

Novoradovsky A, **Taneri B**, Snyder B, Gaasterland T. Splicing patterns in human, mouse and rat transcriptomes. A comparative study. *In preparation.*

**Taneri B**, Novoradovsky A, Snyder B, Gaasterland T. Databases for comparative analysis of human-mouse orthologous alternative splicing. *Lecture Notes in Computer Science 2005,* 3388:123-131.

**Taneri B**, Snyder B, Novoradovsky A, Gaasterland T. Alternative splicing of mouse transcription factors affect their DNA-binding domain architecture and is tissue specific. *Genome Biology 2004,* 5(10):R75.

**Taneri B**, Snyder B, Gaasterland T. Effect of alternative splicing on structure and function of mouse transcription factors. *Eighth Annual International Conference on Research in Computational Molecular Biology. (RECOMB 2004) March 2004 San Diego, CA, USA.*

**Taneri B**, Liang Y, Novoradovsky A, Gaasterland T. Comparative alternative splicing forms: DNA and RNA binding proteins in human, mouse and rat. *Pacific Symposium on Biocomputing 2004. Extended Workshop on Alternative Splicing, January 2004, Hawaii, USA.*

Eisman JA, **Taneri B**, Nguyen TV, Ott J. Linkage in extended pedigree studies of bone phenotypes. *Journal of Bone and Mineral Research 2001,* 16: S351.

# ABSTRACT

Analyzing transcriptomes in the context of all available genome and transcript sequence data has the potential to reveal biologically meaningful insight into functional properties of genes and complexity of genomes. Alternative splicing is one of the major mechanisms contributing to the complexity of genomes. This important cellular process generates several different messenger RNA transcripts from a single gene, expression of which produces structurally and functionally different proteins. Regulation of alternative splicing could be tissue-specific, developmental stage and/or physiological condition dependent.

Comprehensive analysis of alternative splicing is essential to understand fully the capacity of genomes and thus proteomes. Comparative analyses of alternative splicing across species can provide significant biological insight not only to evolution of alternative splicing, but also to its regulation and functional significance.

For comprehensive analyses of alternatively spliced genes, we developed and utilized databases of alternatively spliced transcripts in transcriptomes of *Homo sapiens, Mus musculus* and *Rattus norvegicus*. Our databases allow in-depth analyses of alternative and constitutive exons within alternatively spliced genes. Interactive web implementation of our databases brings to end-users the ability to instantly identify orthologous human-mouse, human-rat and mouse-rat gene-pairs with their corresponding exons. A novel visualization method we introduce, provides easy access to conserved alternative splicing data and a tool to explore the evolutionary significance, regulation and function of this important biological process.

Our statistical analysis showed high prevalence of variant loci in human, mouse and rat transcriptomes. 81% of human loci are variant, as are 74% of mouse loci and 58% of rat loci, revealing widespread presence of

alternative splicing in all three transcriptomes. We further showed that alternative splicing events are mainly due to the presence or absence of cassette exons. More than 60% of alternative exons are cassette exons in all three transcriptomes.

Specifically, to analyze the impact of alternative splicing on transcription factor protein structure, we studied the effect of cassette exons on protein domain architectures of mouse transcription factors. We showed that alternative splicing preferentially adds or deletes domains important in DNA-binding function of the transcription factors. 75% of the domains affected by cassette exons are DNA-binding domains. Further, we showed that there is a single transcription factor isoform within a given tissue and isoforms differ across different tissues indicating tissue-specificity of alternatively spliced transcription factors. These results indicate that alternative splicing might contribute to differential gene expression via creation of tissue-specific transcription factor isoforms.

In addition, we showed that in the human transcriptome, there is a high prevalence of transcript sequence data from cancer tissues. More than 80% of human variant loci contain transcripts from cancer tissues. We showed that cancer transcripts introduce variation beyond normal alternative splicing via cancer-specific cassette exons. In the majority of tissues, more than 20% of the cassette exons are from cancer transcripts only. Our results quantitatively validate presence of aberrant alternative splicing in cancer sequence data.

Lastly, through a comparative analysis of alternatively spliced genes in transcriptomes of *Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana* and *Plasmodium falciparum* to those in human, mouse and rat transcriptomes, we showed that there is more alternative splicing in genomes of more complex organisms and that there is an elevation of alternative splicing in mammalian genomes.

# CHAPTER 1

# INTRODUCTION

## 1.1 Eukaryotic Gene Expression, Transcription and Splicing

The initial step of eukaryotic gene expression is the process of transcription, which involves synthesis of RNA molecules from the DNA. Transcribed RNA molecules are further modified via three processes called 5' capping, RNA splicing and 3' polyadenylation. Modified RNA molecules are then transported out of the nucleus for protein synthesis via a process called translation (Alberts *et al.*, 4th ed.).

Transcription is critical in that it specifies which proteins are expressed within a cell. Functional RNA molecules transcribed from DNA molecules are translated into proteins. Pre-mRNA (premature messenger RNA) processing determines the mRNA (mature messenger RNA) molecule's localization, stability and its interaction with other proteins (Cramer *et al.*, 2001). 5' capping prevents degradation of the RNA molecule and enables its interaction with the ribosomes. 3' polyadenylation further stabilizes the RNA molecule (Cramer *et al.*, 2001).

RNA splicing is the process which removes non-coding sequences (introns) from the pre-mRNA molecule and ligates together the coding-sequences (exons). Splicing takes place in the nucleus. Spliced mRNAs are exported out of the nucleus for protein synthesis (Alberts *et al.*, 4th ed.). Splicing and transcription of RNAs are temporally and spatially coordinated within the cell (Cramer *et al.*, 2001). Figure 1.1 illustrates how RNA processing takes place while transcription occurs. (Figure from Cramer *et al.*, 2001).

## 1.2 Alternative Splicing

Alternative splicing is an important cellular process which generates several different mRNA transcripts from a single gene, increasing the functional complexity of genomes (Black, 2000; Brett *et al.*, 2002). This process enables production of structurally and functionally different proteins. These proteins can range from minimal changes in function to absolutely opposite functions. Expression of different splice forms can be tissue-specific, developmental-stage and/or physiological condition dependent (Graveley, 2001).

Alternative splicing is thought to contribute significantly to increasing the complexity of human genome, given that the latest estimate of the number of protein coding genes in human is only 25,000 (International Human Genome Sequence Consortium, 2004). As of 3/1/05, the latest number of Ensembl gene predictions for fugu (*Fugu rubripes*), zebrafish (*Danio rerio*), mouse (*Mus musculus*) and rat (*Rattus norvegicus*) are 22,089, 23,524, 28,069 and 23,751 respectively (Ensembl 2005a; 2005b; 2005c; 2005d). Fruitfly (*Drosophila melanogaster*) and worm (*Caenorhabditis elegans*) currently have

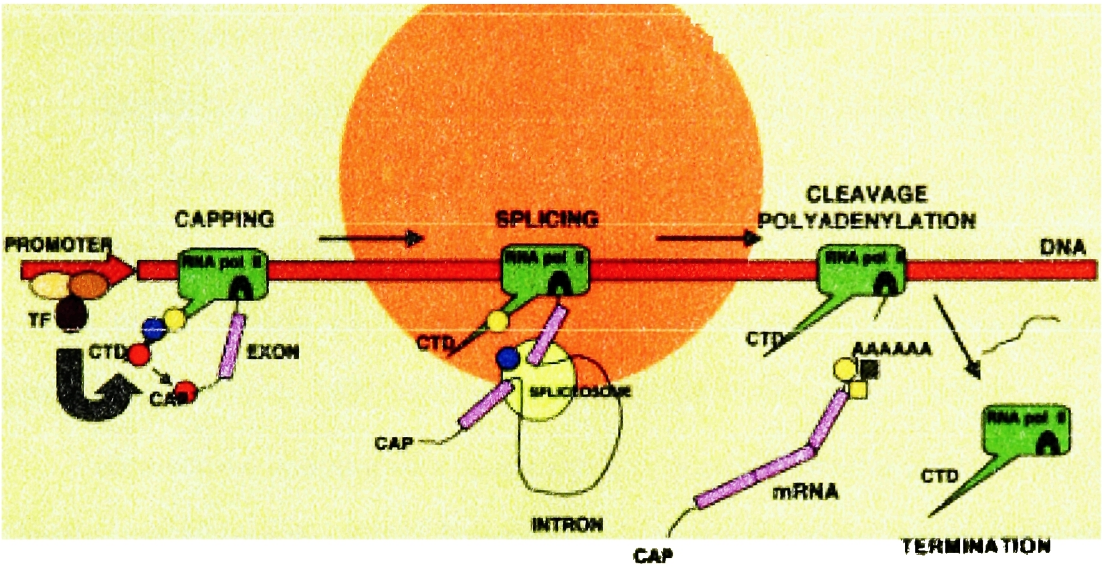


**Figure 1.1** RNA transcription and RNA processing are temporally and spatially coordinated. Red line denotes DNA. Purple boxes denote exons. TF denotes transcription factors. Green box denotes RNA polymerase II. Small circles denote proteins functioning in 5' capping, splicing and 3' polyadenylation. Big yellow circle denotes the spliceosome. (Figure from Cramer *et al.,* 2001).

13,833 and 20,516 genes predicted by Ensembl respectively (Ensembl 2005e; 2005f). The surprisingly low number of human genes leads biologists to search for mechanisms that enable complexity of our transcriptomes and consequently our proteomes.

Alternative splicing is a widespread phenomenon which occurs very frequently within genomes. Earlier bioinformatics analyses estimated up to 65% of the human genes to be alternatively spliced. Examples of alternatively spliced genes include vast majority of immune system and nervous system genes (Mironov *et al.*, 1999; Modrek *et al.*, 2001; Modrek and Lee, 2002).

Alternative splicing of a transcript can occur in several different ways (Smith and Valcarcel, 2000). Figure 1.2 summarizes types of alternatively spliced exons. (Figure from Cartegni *et al.*, 2002). Figure 1.2a shows a *cassette exon* which is skipped in some transcripts and is included in its entirety in other transcripts sequenced from the same gene. Figures 1.2b and 1.2c show *length-variant exons* with alternative 5' and 3' splice sites respectively. These exons are present in all transcripts sequenced from a given gene but they vary in nucleotide length. Mutually exclusive cassette exons are shown in Figure 1.2d. Either one of these exons is included in a given transcript sequenced from a given gene. Figure 1.2e shows an intron retained as an alternative exon.

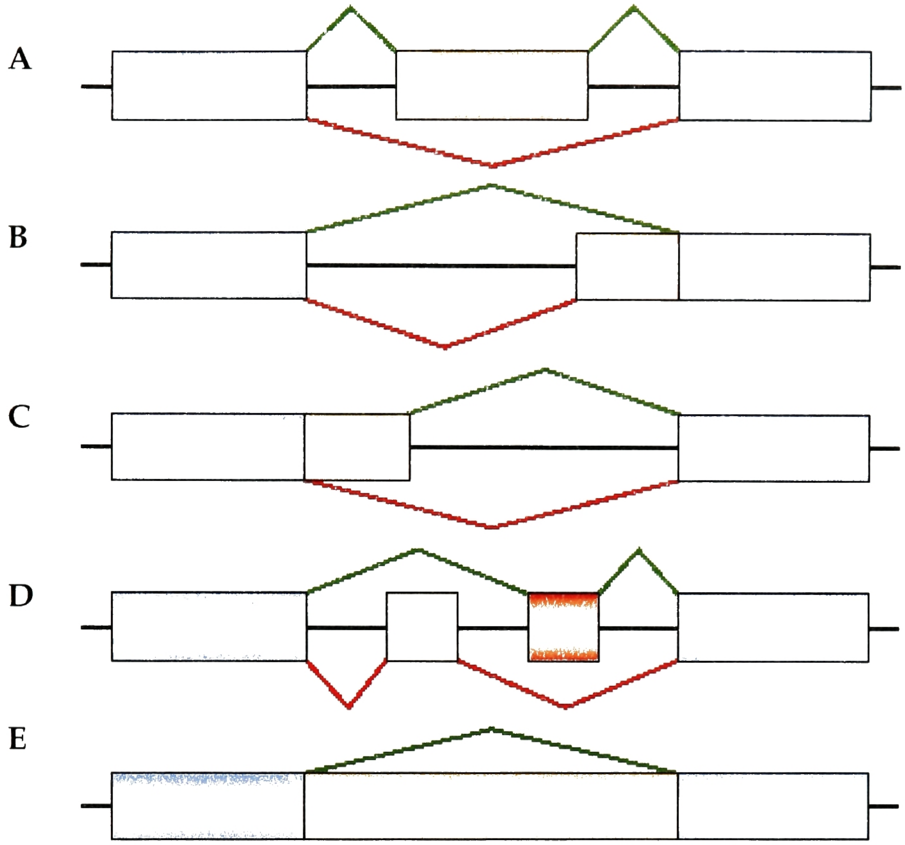


**Figure 1.2** Types of alternative splicing. Blue boxes represent constitutive exons. Red and orange boxes represent alternative exons. **A**. Cassette exon: an alternative exon that is either skipped or included entirely. **B.** 5'-length-variant exon: an alternative exon with alternative 5' splice sites. **C.** 3'-length-variant exon: an alternative exon with alternative 3' splice sites. **D.** Mutually exclusive cassette exons: either one of these alternative exons is included in a given transcript. **E.** Intron retention: intron is retained as an alternative exon. (Figure from Cartegni *et al.*, 2002).

## 1.3 Evolution and Conservation of Alternative Splicing

RNA splicing is thought to have originated from Group II introns with autocatalytic function, which are large, naturally catalytic RNAs (Ast, 2004; Lehman and Schmidt, 2003). The basic splicing machinery is evolutionarily conserved (Ast, 2004) as further discussed in Section 1.4.

Alternative splicing is conserved across species (Thanaraj *et al.*, 2003; Valenzuela *et al.*, 2004; Yeo *et al.*, 2005) and occurs widely in protein-coding genes of multicellular eukaryotes (Ast, 2004). On the other hand, unicellular eukaryotes such as the yeast *Saccaromyces cerevisiae* have not been observed to exhibit regulated alternative splicing (Ast, 2004).

In human and mouse transcriptomes, both alternative exons and their flanking intronic sequences have been computationally observed to be conserved at higher levels compared to constitutive exon conservation (Sorek *et al.*, 2003; Sugnet *et al.*, 2004). Alternative exons which are protein coding and which do not introduce frame-shifts or premature stop codons have been shown to be highly conserved between species (Resch *et al.*, 2004a). However, some reports show lower levels of conservation of human and mouse cassette exons (Nurtdinov *et al.*, 2003). Ast and colleagues point to the similar sequence characteristics shared by human and mouse conserved cassette exons and state that conserved alternative splicing is likely to be functional (Sorek *et al.*, 2004). Thus the remaining cassette exons are either species specific or due to aberrant splicing.

In addition to alternative exons, splice sites tend to be conserved between human and mouse (Carmel *et al.*, 2004). Majority of the mammalian splice junctions have canonical nucleotide sequences (Burset *et al.*, 2000; Weir and Rice, 2004).

## 1.4 Mechanism of Splicing and Alternative Splicing

Splicing involves a network of *cis*-acting and *trans*-acting factors (Smith and Valcarcel, 2000). Proteins of this network interact to facilitate splicing through two processes termed Exon Definition (ED) and Intron Definition (ID) (Berget, 1995).

Splicing is catalyzed by the spliceosome complex which is composed of five small nuclear RNAs (snRNAs) called U1, U2, U4, U5, U6 and as many as 300 other proteins (Will and Luhrmann, 1997; Nilsen, 2003). Spliceosome machinery recognizes exons in the pre-mRNA with high precision, removes introns and ligates the exons to each other forming the mature mRNA to be translated into proteins. Figure 1.3 illustrates assembly of the spliceosome proteins and RNA molecules. (Figure from Smith and Valcarcel, 2000). Assembly of the snRNAs with the proteins form the small nuclear ribonuclear protein complexes (snRNPs), which bind to the splice sites on the pre-mRNA molecule and facilitate removal of introns and ligation of exons (Zhou *et al.*, 2002; Jurica and Moore, 2003).

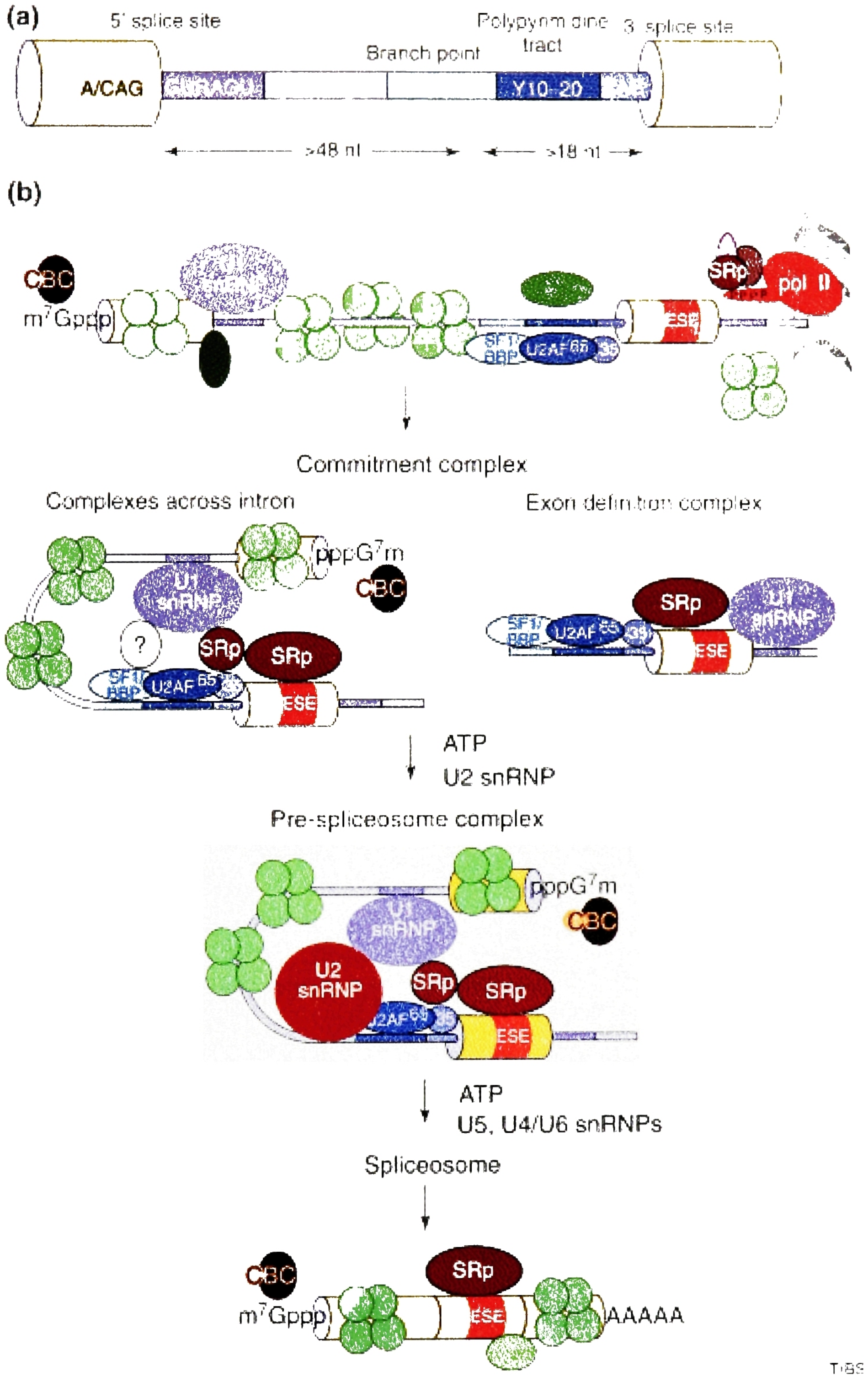


**Figure 1.3** Spliceosome assembly. Green circles denote hnRNPs. CBC denotes cap binding complex. Yellow color denotes exons. Splicing takes place via 5' splice site cleavage and intron lariat formation followed by 3' splice site cleavage and exon ligation. (Figure from Smith and Valcarcel, 2000).

Four types of genomic sequence sites play important roles in intron splicing (Brow, 2002). These sequences, illustrated in Figure 1.4, are the following and they function in two transesterification reactions involved in intron removal. (Figure from Brow, 2002).

(1) 5' splice site: exon-intron junction at the 5' end of the intron.

(2) 3' splice site: exon-intron junction at the 3' end of the intron.

(3) branch point: sequence upstream of the 3' splice site.

(4) polypyrimidine tract: sequence between 3' splice site and branch site.

In addition to conventional splicing signals, several additional splicing signals are involved in alternative splice site selection. Exonic Splicing Enhancers (ESEs), Exonic Splicing Silencers (ESSs), Intronic Splicing Enhancers (ISEs) and Intronic Splicing Silencers (ISSs) are short binding sequences for proteins which regulate alternative splicing.

ESEs function as binding sites for a specific group of proteins called the SR proteins (short for serine-arginine rich proteins) (Stojdl and Bell, 1999). SR proteins are well conserved and contain two distinct motifs as RRMs (RNA recognition motifs) and the RS domain (Arginine-Serine rich domain). SR proteins bind to pre-mRNA via their RRM motifs and their RS domain mediates protein-protein interactions (Cartegni *et al.*, 2002; Ma and He, 2003). SR proteins enable Exon Definition by recruiting the spliceosome to the exons via their RS domain. Figure 1.5a illustrates recruitment of the splicing machinery to ESEs on the pre-mRNA via SR proteins. In addition, SR proteins might function by antagonizing the effects of splicing inhibitory proteins which bind to ESSs neighboring ESEs as shown in Figure 1.5b. (Figure from

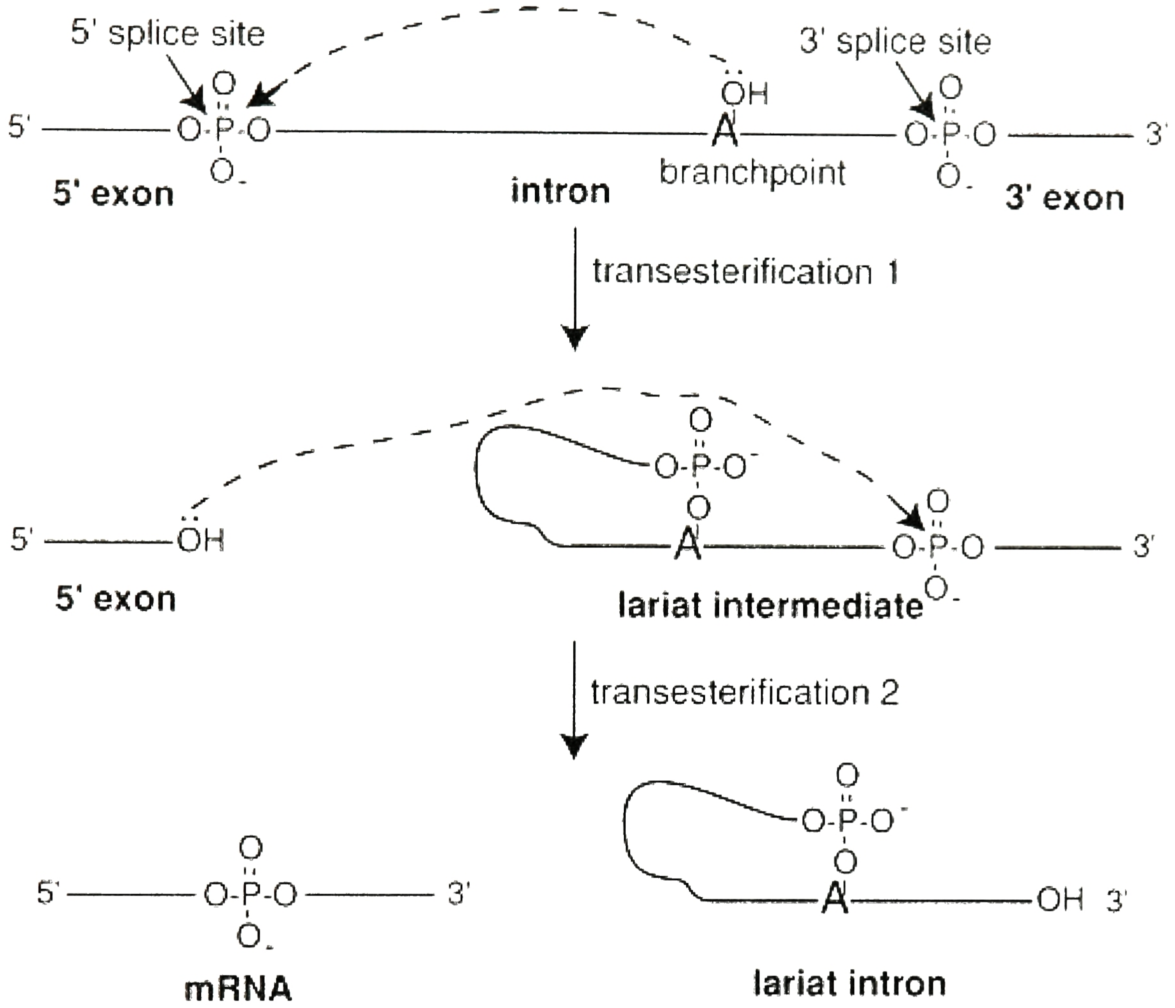


**Figure 1.4** Transesterification reactions catalyzed by the spliceosome. 5' splice site, branchpoint and 3' splice site play important roles in these reactions. (Figure from Brow, 2002).

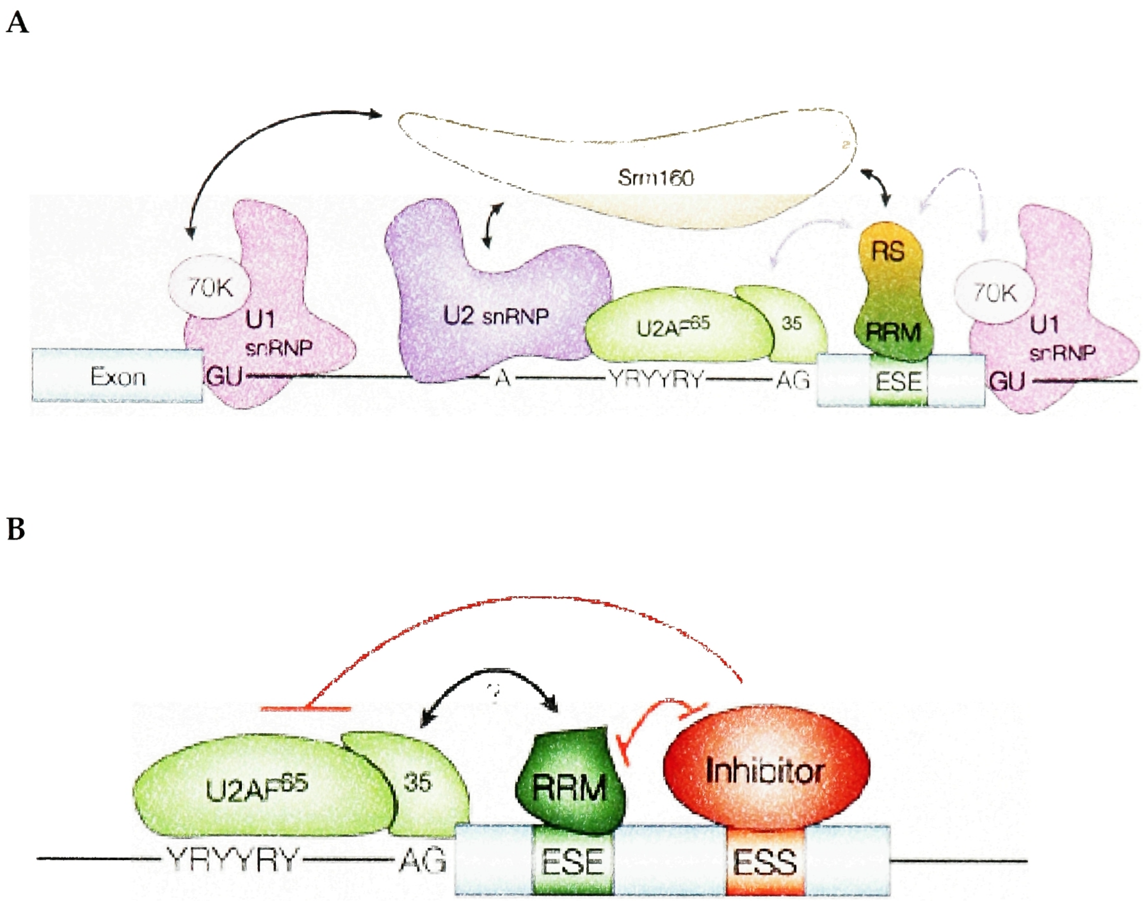


**Figure 1.5** **(A)** Recruitment of splicing machinery to ESEs via SR proteins. SR protein (containing RS and RRM motifs) binds to the ESE via the RRM motif and recruits other splicing proteins via the RS domain. **(B)** Antagonistic effect of SR proteins to splicing inhibitors. SR protein binds to the ESE via the RRM motif and antagonizes the neighboring inhibitory protein which is bound to the ESS. Blue boxes denote exons. Black lines denote intronic sequences. (Figure from Cartegni *et al.*, 2002).

Cartegni *et al.*, 2002). Since SR proteins recruit the basal splicing machinery to the RNA, they are required both for alternative and for constitutive splicing.

Intron Definition is carried out by several splicing proteins binding to the splice sites on both ends of an intron as illustrated in Figure 1.6, which results in removal of the intron. (Figure from Ast, 2004). SR proteins might mediate Intron Definition as well.

ESSs are less well characterized compared to ESEs and their mechanism of action is less well understood (Cartegni *et al.*, 2002). ESSs interact with negative regulators of splicing such as hnRNPs (heterogeneous nuclear ribonucleoproteins) (Wang *et al.*, 2004). hnRNPs contain RNA binding domains as well as protein-protein interaction domains. Depending on the affinity and concentration of positive and negative regulators of splicing such as SR proteins and hnRNPs, an alternative exon can either be included or excluded. Silencing of splicing can also take place via ISSs. Negative splicing factors bind to two ISSs flanking the alternative exon on both sides. Interaction of the negative splicing proteins with each other while bound to ISSs results in skipping of the alternative exon as illustrated in Figure 1.7. (Figure from Cartegni *et al.*, 2002).

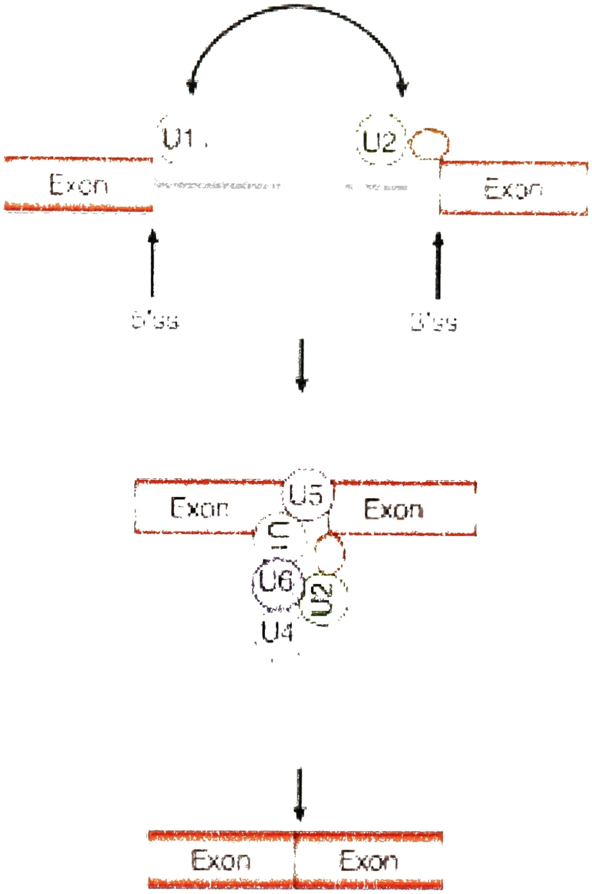


**Figure 1.6** Intron Definition. Red boxes denote exons. Gray line denotes intron. Circles denote splicing proteins, which are bound to splice sites on both ends of the intron. (Figure from Ast, 2004).

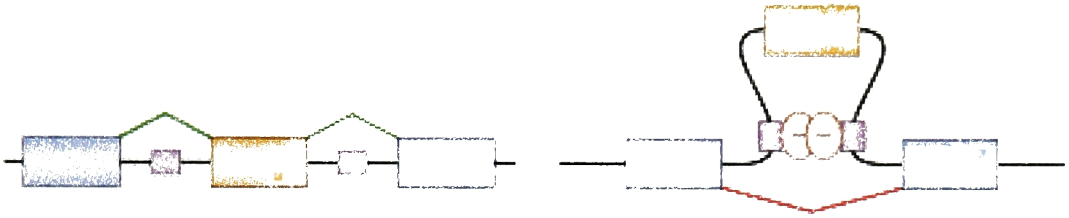


**Figure 1.7** Cassette exon skipping via ISSs. Purple boxes denote ISSs. Orange box denotes the cassette exon. Blue boxes denote constitutive exons. Circles denote negative splicing factors bound to the ISSs. (Figure from Cartegni *et al.*, 2002).

## 1.5 Regulation of Alternative Splicing

Regulation of alternative splicing can occur in a tissue-specific, developmental stage and/or in a physiological condition dependent manner (Graveley, 2001; Woodley and Valcarcel 2002; Black, 2003). Complex interactions between RNA binding proteins and *cis* regulatory elements lead to tissue-specific and/or cell-specific splicing patterns (Lopez, 1998). Tissue-specific alternative splicing plays a very important role in animal development. Defects in regulation of this process lead to severe human diseases (Grabowski and Black, 2001).

Tissue-specificity of alternative splicing is controlled by splicing regulatory proteins; expression of which is restricted to certain cell types, such as neuron-specific splicing regulators Nova-1 and Nova-2 (Jensen *et al.*, 2000a; Yang *et al.*, 1998). PTB (poly-pyrimidine-tract-binding protein) is one of the major proteins shown to play a role in tissue-specific alternative splicing (Singh *et al.*, 1995; Valcarcel and Gebauer, 1997; Zhang *et al.*, 1999). NAPOR (neuroblastoma apoptosis related RNA binding protein) is another RNA-binding protein which regulates alternative splicing in the brain (Zhang *et al.*, 2002). A list of splicing regulatory proteins and their functionally important domains are provided in Figure 1.8. (Figure from Black, 2003). Furthermore, cell specific regulation could depend on regulation of splicing via different isoforms of a given splicing factor (Pacheco *et al.*, 2004). To understand further the regulation of alternative splicing, it is essential to identify and analyze tissue-specific forms of alternatively spliced genes on global levels (Xu *et al.*, 2002).

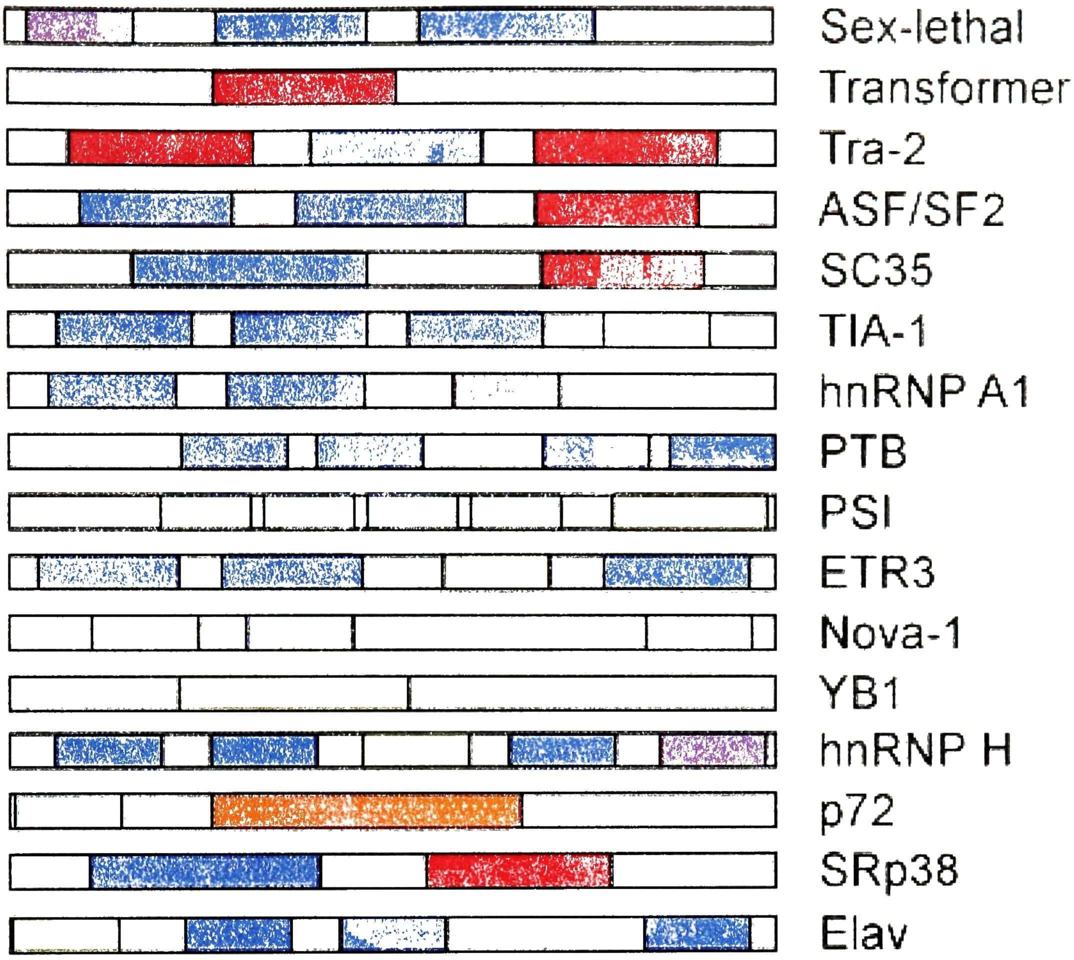


**Figure 1.8** Splicing regulatory proteins. Pink boxes denote glycine-rich domains. Blue boxes denote RRM domains. Red boxes denote either RS or related domains. Yellow boxes denote glycine-rich and other specialized domains. Green boxes denote KH (K-homology) domains. Cyan box denotes RGG domain. Orange box denotes DEAD box domain. (Figure from Black, 2003).

## 1.6 Coupling of Transcription and Alternative Splicing

There is evidence indicating that alternative splicing takes place in coordination with transcription (Caceres and Kornblihtt, 2002). Alternative splicing of a gene can be dependent on the promoter region regulating the transcription of that gene. Figure 1.9 illustrates models explaining promoter regulation of alternative splicing. (Figure from Caceres and Kornblihtt, 2002). Different promoters might recruit different quantities of SR proteins or independent of the SR proteins, different promoters might enable fast or slow acting RNA polymerases (Cramer *et al.*, 1997; 1999; de la Mata *et al.*, 2003; Nogues *et al.*, 2002; 2003).

## 1.7 Implications of Aberrant Splicing in Human Diseases

Proper regulation of splicing is essential to healthy physiology. Aberrant splicing has been documented to lead to pathological conditions including severe diseases in humans such as *osteogenesis imperfecta* and spinal muscular atrophy (Cooper and Mattox, 1997; Schwarze *et al.*, 1999). Defects in alternative splicing lead to a broad spectrum of diseases including genetic diseases such as Ehlers-Danlos Syndrome (Byers *et al.*, 1997; Schwarze *et al.*, 1997; Takahara *et al.*, 2002) and Occipital Horn Syndrome (Qi and Byers, 1998). Table 1.1 shows a range of genetic disorders caused by defects in alternative splicing. (Data from Caceres and Kornblihtt, 2002).

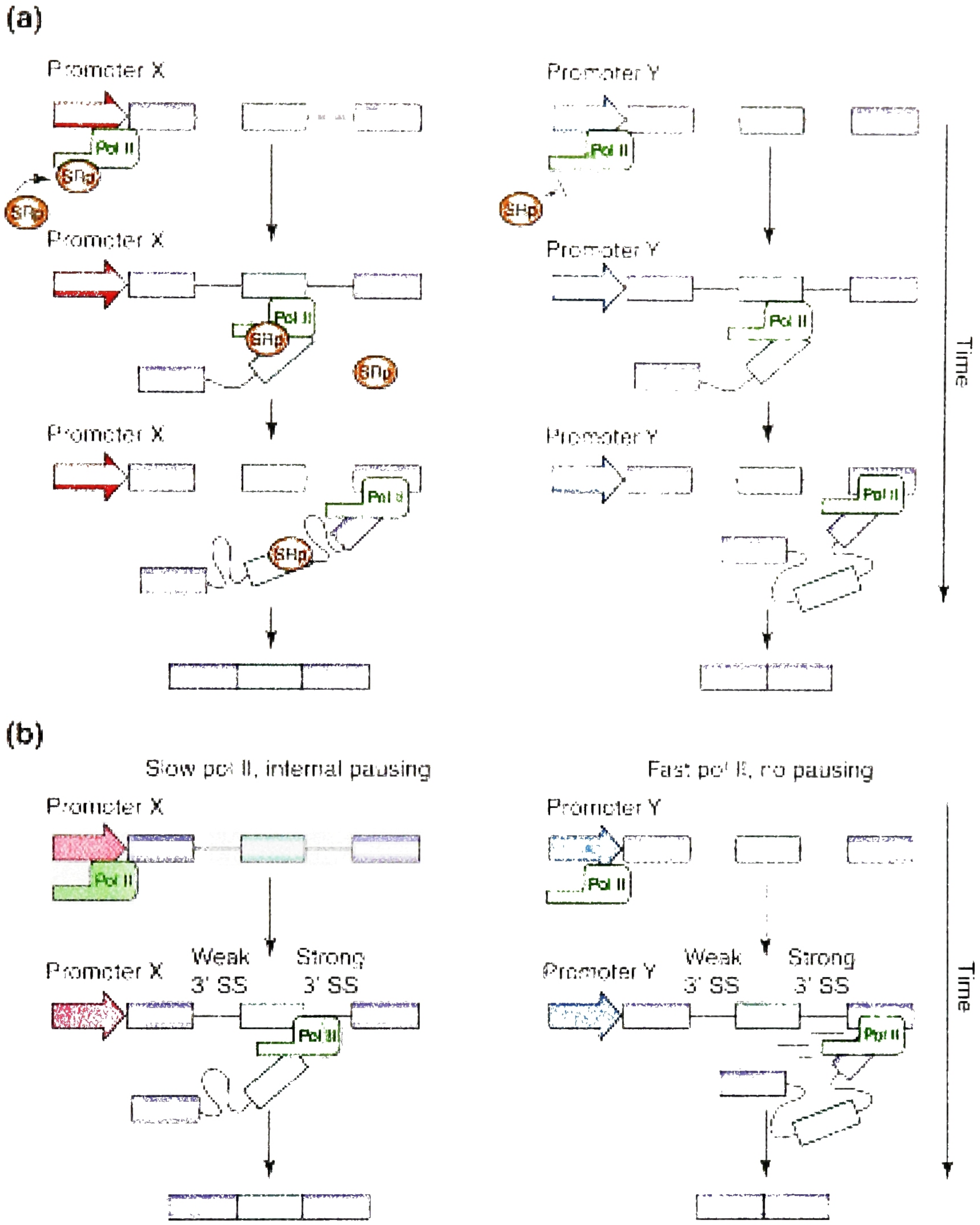


**Figure 1.9** **(a)** Different promoters might recruit different quantities of SR proteins. Promoter on the left recruits more SR proteins compared to the promoter on the right. **(b)** Independent of the SR proteins, different promoters might enable a slow acting RNA polymerase II (on the left) or a fast acting RNA polymerase II (on the right). Purple boxes denote constitutive exons, blue boxes denote alternative exons. Green box denotes RNA polymerase II, orange circles denote SR proteins. Black lines denote intronic sequences. (Figure from Caceres and Kornblihtt, 2002).

**Table 1.1** Genetic disorders caused by defects in alternative splicing.

| Genetic Disorder | Gene |
|---|---|
| Acute intermittent porphyria | Porphobilinogen deaminase |
| Breast and ovarian cancer | BRCA1 |
| Carbohydrate-deficient glycoprotein syndrome type 1a | PMM2 |
| Cerbrotendinous xanothomatosis | Sterol-27-hydroxylase |
| Cystic fibrosis | CFTR |
| Ehlers-Danlos syndrome type V1 | Lysyl hydroxylase |
| Fanconi anemia | FANCG |
| Frontotemporal dementia (FTDP-17) | Tau |
| Hemophilia A | Factor VIII |
| HPRT deficiency | Hypoxanthine phosphoribosyl transferase |
| Leigh's encephalomyelopathy | Pyruvate dehyrogenase E1α |
| Marfan syndrome | Fibrillin-1 |
| Metachromatic leukodystrophy (juvenile form) | Arylsulfatase A |
| Neurofibromatosis type 1 | NF1 |
| OCT deficiency | Ornithine carbamoyltransferase |
| Porphyria cutanea tarda | Uroporphyrinogen decarboxylase |
| Sandhoff disease | Hexosaminidase |
| Severe combined immunodeficiency | Adenosine deaminase |
| Spinal muscle atrophy | SMN1 |
| Spinal muscle atrophy | SMN2 |
| Tyrosinemia type 1 | Fumaryl acetoacetate hydrolase |

(Data from Caceres and Kornblihtt, 2002).

Aberrant splicing has been widely documented in cancers (Brinkman, 2004; Venables, 2004; Kalnina *et al.*, 2005). For example, alternatively spliced forms of cadherin-11 and Bcl-x genes have been implicated in breast and prostate cancers (Mercatante *et al.*, 2001a; Feltes *et al.*, 2002). Valcarcel and colleagues showed correlation of amounts of alternatively spliced isoforms of p16 and p14$^{ARF}$ genes with tumor formation in Hodgkin's lymphoma (Relogio *et al.*, 2005).

RNA binding proteins functioning in tissue specific regulation of alternative splicing, discussed in Section 1.5, are also involved in disease states including myotonic dystrophy, thalassemia and paraneoplastic opsoclonus-myoclonus ataxia (Musunuru, 2003) and cancers such as neuroblastoma (Palm *et al.*, 1999).

Several efforts are being made to correct disease states due to defects in alternative splicing (Sazani and Kole, 2003). Using antisense oligonucleotides, Kole and colleagues showed correction of cystic fibrosis transmembrane conductance regulator (CFTR) gene (Friedman *et al.*, 1999). Misteli and colleagues showed correction of alternative splicing of tau gene implicated in Parkinsons and dementia (Kalbfuss *et al.*, 2001). van Ommen and colleagues showed correction of dystrophin levels in Duchenne muscular dystrophy patients, by inducing exon skipping (van Deutekom *et al.*, 2001). Given that alternative splicing has been implicated in cancers, efforts are being made to develop antisense oligonucleotides to control alternative splicing for chemotherapeutic purposes (Mercatante *et al.*, 2001b; Mercatante and Kole 2002).

## 1.8 Bioinformatics and Alternative Splicing

Availability of extensive full-length transcript sequences and EST data from genomes of different organisms enables detailed bioinformatics analyses of alternative splicing within and across species. Multiple sequence alignment approaches and DNA chip data have been used to search for evidence of alternative exons and to detect alternative splicing events in transcriptomes (Hu *et al.*, 2001; Grasso *et al.*, 2004; Sakai and Maruyama, 2004; Xing *et al.*, 2004).

In recent years, a variety of alternative splicing databases have been developed based on transcript sequence data. Many of these databases have splice forms from one organism only, for example the Alternative Splicing Annotation Project (ASAP), Manually Annotated Alternatively Spliced Events (MAASE) and SpliceNest human databases (Coward *et al.*, 2002; Lee *et al.*, 2003; Zheng et al., 2004). Several databases contain alternative splice forms from two or more different organisms. These include Extended Alternatively Spliced EST Database (EASED) (Pospisil *et al.*, 2004), the Putative Alternative Splicing database (PALSdb) (Huang *et al.*, 2002), database of alternatively spliced genes (ASDB) (Dralyuk *et al.*, 2000), an alternative splice database of mammals (AsMamDB) (Ji *et al.*, 2001), database of canonical and non-canonical mammalian splice sites (SpliceDB) (Burset *et al.*, 2001) and the Alternative Splicing Database (ASD) (Thanaraj *et al.*, 2004).

Bioinformatics has also been applied to comparative studies of alternative splicing. Given that mouse and human genes are highly conserved with about 80% of the mouse genes having human orthologs and more than 90% of the human and mouse genomes being within conserved syntenic regions (Mouse Genome Sequencing Consortium, 2002), most of the comparative studies of alternative splicing focus on these two species. In a recent study, Kan *et al.* employed cross-species analyses of human and mouse gene sets and identified novel alternative splice forms (Kan *et al.*, 2004). Sugnet *et al.* used whole genome alignments to assess conservation of alternative splicing and established human-mouse conserved splicing graphs (Sugnet *et al.*, 2004). As a result of a human-mouse comparative study, Modrek and Lee showed that alternative splicing results from exon creation or loss (Modrek and Lee, 2003). All of the existing alternative splicing databases and the cross-species comparison analyses are informative and facilitate further understanding of alternative splicing. In addition, efforts are underway to analyze alternative splicing utilizing microarray technology (Clark *et al.*, 2002; Castle *et al.*, 2003; Lee and Roy, 2004).

In the work described here, we introduce three comprehensive databases of alternative splicing in human, mouse and rat transcriptomes. In contrast to existing databases, these bring to the end-user the ability to find not only the orthologous alternatively spliced genes but also their corresponding alternative and constitutive exons. In addition, we describe novel computational methods to comprehensively analyze tissue-specificity of alternative splicing and the effect of cassette exons on protein domain architecture.

## 1.9 Organization of the Thesis

The main focus of this thesis is to answer questions about prevalence of alternative splicing in genomes and its effect on transcriptomes. Toward this goal, we developed three alternative splicing databases in human, mouse and rat transcriptomes as presented in Chapter 2. We utilized these databases to study tissue distribution of alternatively spliced mouse transcription factor (TF) transcripts and distribution of all variant mouse transcripts as described in Chapter 3. In addition in Chapter 3, we provide a detailed analysis of the effect of alternative exons on protein domain architecture of mouse TFs. Chapter 4 details pathological splicing in human cancer tissues. Chapter 5 discusses studies of neuron-specific splicing regulator, Nova. Chapter 6 introduces alternative splicing databases in *Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana* and *Plasmodium falciparum.* In Chapter 7, we present the methods used for the above analyses. We discuss our findings and future directions in Chapter 8.

# CHAPTER 2

# USING DATABASES TO ANALYZE ALTERNATIVE SPLICING

## 2.1 Definitions

We developed three alternative splicing databases for human, mouse and rat transcriptomes called HumanSDB3, MouSDB5 and RatSDB2 respectively. HumanSDB3 stands for Human Splicing Database version 3; MouSDB5 for Mouse Splicing Database version 5; and RatSDB2 for Rat Splicing Database version 2. Methods used in developing these databases are described in Materials and Methods, Section 7.1.

For the work described in this thesis, we defined the following terms presented in italics. All terms are applicable to all three databases. A *transcript* is a sequence transcribed from the genomic DNA sequence and spliced. A *locus* is a genomic region that includes a set of overlapping transcripts mapped to the genome such that a given transcript appears only in one locus. Within a locus a *cassette exon* is completely included in some transcripts and completely excluded from others. A *length-variant exon* has alternative 5' or 3' splice sites or both, in different transcripts. An exon can be both length-variant and cassette. A *variant exon* is either a cassette exon or a length-variant exon or both. A *genomic exon* is an uninterrupted series of nucleotides each of which maps to a transcript. A *cluster* is the set of transcripts that map to a locus. A *variant cluster* contains one or more variant exons. An *invariant cluster* has no variant exons (Taneri *et al.,* 2004; Taneri *et al.,* 2005). (In Chapter 1, Figure 1.2a illustrates cassette exons. Figures 1.2b and 1.2c illustrate length-variant exons). Figure 2.1 illustrates the concept of a genomic exon. The *genomic exon map* of a given cluster is constructed by putting together all the genomic exons with intronic sequences for that cluster.

## 2.2 Annotation of Alternative Exons

There are 10 kinds of exons annotated in HumanSDB3, MouSDB5 and RatSDB2 as described below. These exons are annotated based on EST and full-length transcript sequence data as described in Materials and Methods, Section 7.2.

(1) *Length-invariant constitutive exon*:

a constitutive exon which has the same splice sites in all transcripts of a given cluster.

*(2) 5'-variant constitutive exon:*

a length-variant exon with different 5' splice sites in different transcripts of a given cluster.

*(3) 3'-variant constitutive exon:*

a length-variant exon with different 3' splice sites in different transcripts of a given cluster.

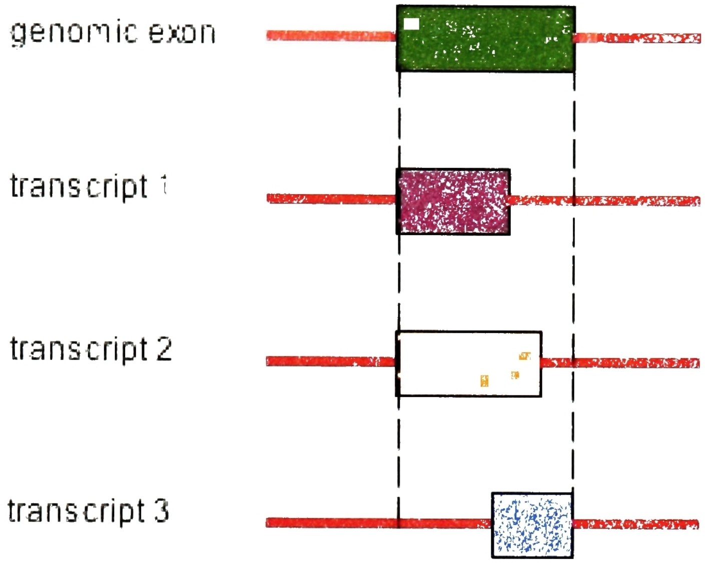


**Figure 2.1** Genomic exon map construction. Green box denotes the genomic exon for a particular cluster at a particular genomic locus. Genomic exon starts with the leftmost nucleotide of the exon in transcript 1 (purple exon), it spans all of the exon in transcript 2 (yellow exon) and ends with the rightmost nucleotide of the exon in transcript 3 (blue exon). Genomic exon is the uninterrupted series of nucleotides spun by the exons in transcripts 1, 2 and 3.

(4) *5' and 3'- variant constitutive exon:*

a length-variant exon which differs both at the 5' splice site and at the 3' splice site among different transcripts of a given cluster.

(5) *Length-invariant cassette exon:*

an exon which is present in some transcripts and completely absent from other transcripts of a cluster. This exon has the same splice sites in all transcripts in which it is present.

(6) *5'-variant cassette exon:*

a cassette exon which has varying 5' splice sites in different transcripts of a given cluster.

(7) *3'-variant cassette exon:*

a cassette exon which has varying 3' splice sites in different transcripts of a given cluster.

(8) *5' and 3'-variant cassette exon:*

a cassette exon which varies both at the 5' splice site and at the 3' splice site among different transcripts of a given cluster.

(9) *Transcript- terminal cassette exon:*

a cassette exon which is either at the 5' end or at the 3' end of the transcript. This exon does not map to any genomic exons of its cluster but rather maps to intronic regions.

(10) *Problem exon:*

an exon which is not completely mapped to the genome.

(Cassette exons and length-variant exons are illustrated in Chapter 1, Figure 1.2).

## 2.3 Statistical Analyses: Variation in Human, Mouse and Rat Genomes

As shown in Table 2.1, a total of 4,635,471 human, a total of 3,322,164 mouse and a total of 552,436 rat UniGene input transcript sequences were used to construct the alternative splicing databases. Of these sequences 31% of human, 35% of mouse and 26% of rat transcripts mapped to their respective genomes with stringent constraints as described in Section 2.4 and further detailed in Materials and Methods, Section 7.1. Final versions of the databases contain a total of 20,707 HumanSDB3 clusters, a total of 20,090 MouSDB5 clusters and a total of 11,330 RatSDB2 clusters. The number of invariant clusters in human, mouse and rat databases are 3,881, 5,221 and 4,721 respectively. In all three transcriptomes, variant loci are in higher numbers than invariant loci. The number of variant clusters in human, mouse and rat databases are 16,826, 14,869 and 6,609 respectively (Table 2.1). Invariant clusters are loci with genes which do not display alternative splicing. Variant clusters are loci with alternatively spliced transcripts. Figure 2.2 shows percent variation in human, mouse and rat transcriptomes to be 81.3%, 74% and 58% respectively. These results reveal widespread presence of alternatively spliced genes in all three transcriptomes. Data used to extract these statistical information and the AutoDB database schema are further detailed in Chapter 7, Section 7.4 and Table 7.2.

**Table 2.1** Statistical Analysis of HumanSDB3, MouSDB5 and RatSDB2 databases.

| | HumanSDB3 | MouSDB5 | RatSDB2 |
|---|---|---|---|
| **Total number of clusters** | 20707 | 20090 | 11330 |
| **Number of invariant clusters** | 3881 | 5221 | 4721 |
| **Number of variant clusters** | 16826 | 14869 | 6609 |
| **Number of input transcripts** | 4635471 | 3322164 | 552436 |
| **Number of mapped transcripts** | 1459966 | 1149658 | 142831 |
| **% of input transcripts mapped** | 31% | 35% | 26% |
| **Average transcripts per cluster** | 70.5 | 57.2 | 12.6 |

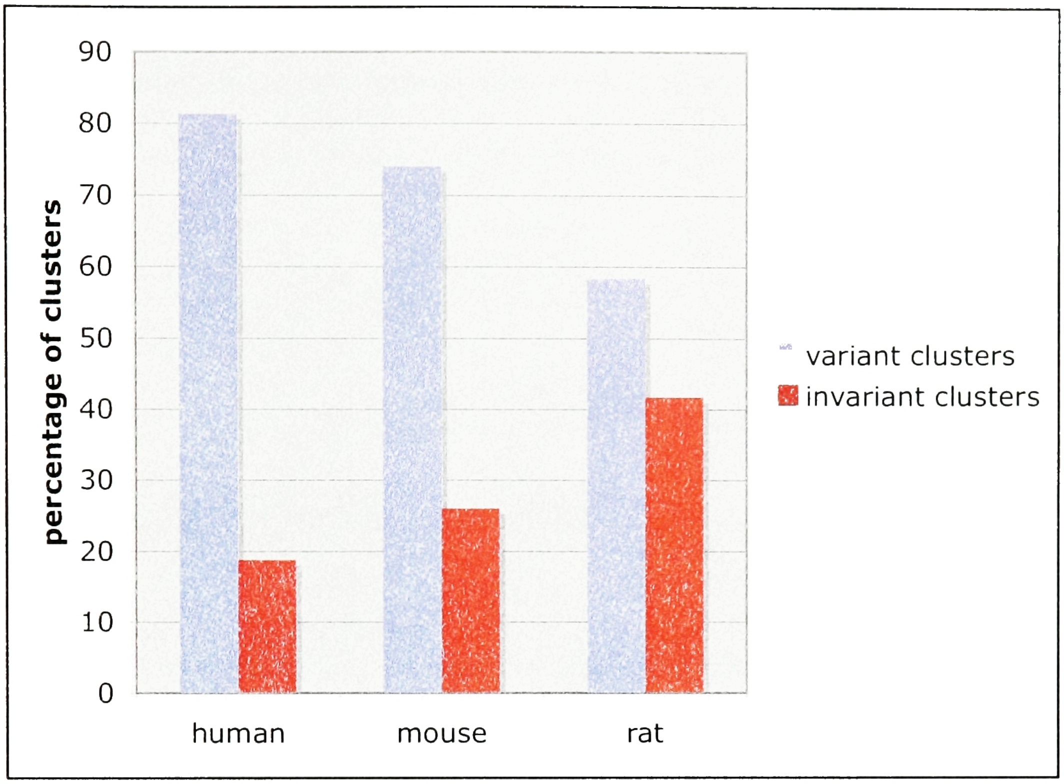


**Figure 2.2** Distribution of variant and invariant clusters in HumanSDB3, MouSDB5 and RatSDB2. Blue bars show percent variant clusters and red bars show percent invariant clusters in each transcriptome. Variant clusters are 81%, 74% and 58% in human, mouse and rat transcriptomes, respectively.

The lower proportion of variation in the rat transcriptome reflects a lower number of rat input transcripts. Figure 2.3 shows that the percentage of variant clusters correlates positively with the number of input transcripts (Fig. 2.3a) and with the average number of transcripts within clusters (Fig. 2.3b).

## 2.4 Database Pipeline and Filters

In this section, we discuss the AutoDB pipeline used in building the databases and provide detailed information about the number of transcripts rejected at each step of this pipeline. AutoDB steps are described in more detail in Chapter 7, Section 7.1. For HumanSDB3, we began with 4,635,471 input transcripts. After the blat step, 8,436 transcripts were rejected (0.18% of input transcripts) and 4,627,035 remained. This number is only 0.27% of all transcripts which were rejected in total by all filters. At the SIM4 step, 625 transcripts were rejected (0.01%) and 4,626,410 remained. This was 0.02% of the total rejected. After SIM4 we filtered for transcript quality. Each transcript had to have at least 2 exons and had to match the genome with at least 75% identity. Each exon had to either match the genome with at least 95% identity or had to have 5 or fewer mismatches. At the quality filter step, 3,150,291 transcripts were rejected and 1,476,119 remained. This was 67.96% of the input transcripts. This fallout constitutes 99.21% of the total fallout for the pipeline. The final step applied filters at the cluster level. Each transcript

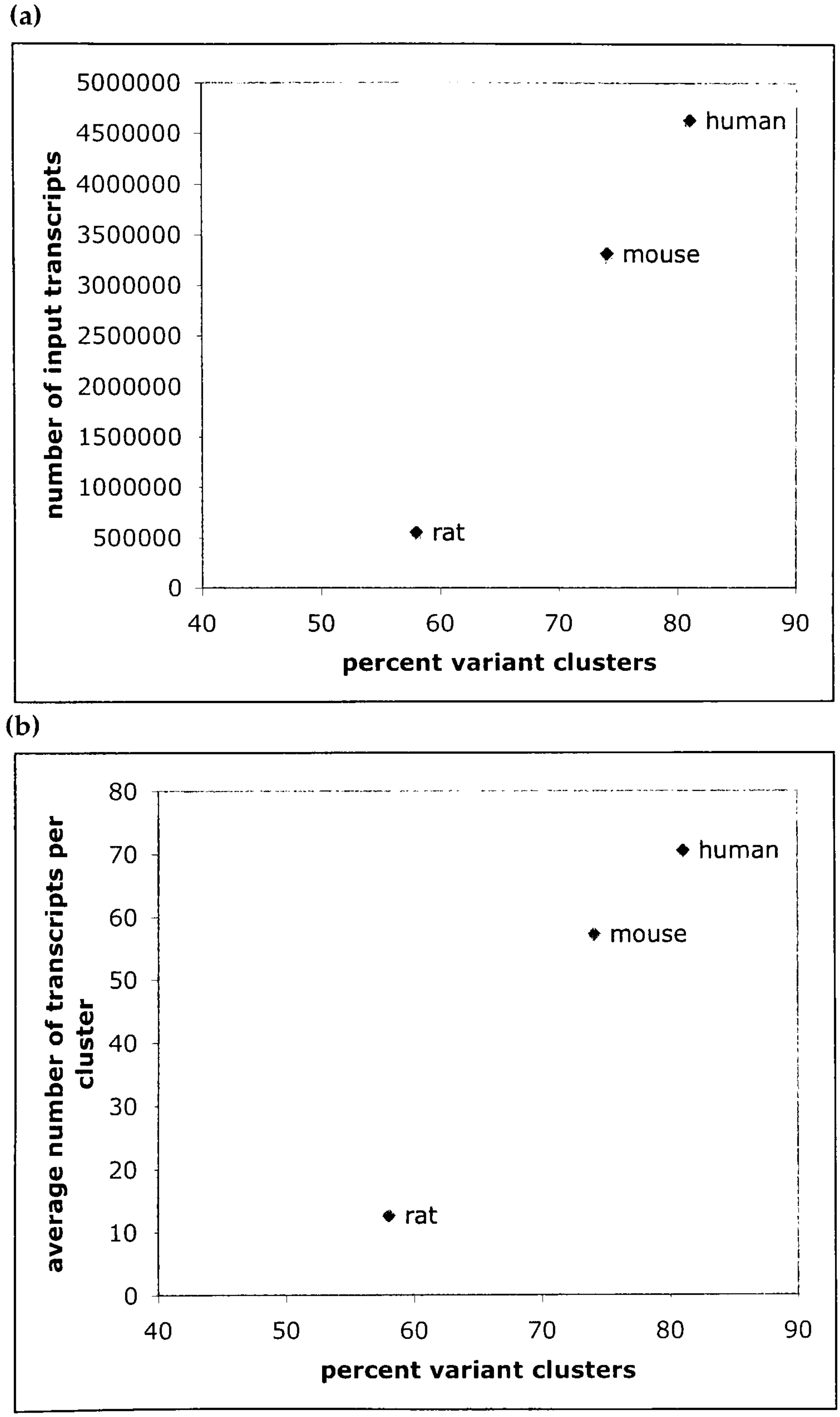


**Figure 2.3** Percent variation in transcriptomes is positively correlated with **(a)** number of input transcripts and **(b)** average number of transcripts per cluster.

had to belong to a cluster with at least 3 transcripts in total. Transcripts included in singleton and doubleton clusters were dropped. At this step, 16,153 transcripts were lost. This was 0.35% of the input transcripts and 0.51% of the total rejected. After the pipeline, total fallout added up to 3,175,505 transcripts, which was 68.5% of the input transcripts. Table 2.2 shows numbers of human input transcripts rejected at each AutoDB filter.

For MouSDB5, we began with 3,322,164 input transcripts. After the blat step, 5,784 transcripts were rejected (0.17% of input transcripts) and 3,316,380 remained. This number is only 0.27% of all transcripts which were rejected in total by all filters. At the SIM4 step zero transcripts were lost. After SIM4 we filtered for transcript quality. Each transcript had to have at least 2 exons and had to match the genome with at least 75% identity. Each exon had to either match the genome with at least 95% identity or had to have 5 or fewer mismatches. At the quality filter step, 2,153,688 transcripts were rejected and 1,162,692 remained. This was 64.83% of the input transcripts. This fallout constitutes 99.13% of the total fallout for the pipeline. The final step applied filters at the cluster level. Each transcript had to belong to a cluster with at least 3 transcripts in total. Transcripts included in singleton and doubleton clusters were dropped. At this step, 13,034 transcripts were lost. This was 0.39% of the input transcripts and 0.6% of the total rejected. After the pipeline, total fallout added up to 2,172,506 transcripts, which was 65.39% of the input transcripts. Table 2.3 shows numbers of mouse input transcripts rejected at each AutoDB filter.

For RatSDB2, we began with 552,436 input transcripts. After the blat step, 3,801 transcripts were rejected (0.69% of input transcripts) and 548,635 remained. This number is only 0.93% of all transcripts which were rejected in total by all filters. At the SIM4 step 11 transcripts were lost (0.002%), 548,624 remained. This was 0.003% of the total fallout. After SIM4 we filtered for transcript quality. Each transcript had to have at least 2 exons and had to match the genome with at least 75% identity. Each exon had to either match the genome with at least 95% identity or had to have 5 or fewer mismatches. At the quality filter step, 395,624 transcripts were lost, 153,000 remained. This was 71.61% of the input transcripts. This fallout constitutes 96.59% of the total fallout for the pipeline. The final step applied filters at the cluster level. Each transcript had to belong to a cluster with at least 3 transcripts in total. Transcripts included in singleton and doubleton clusters were dropped. At this step, 10,169 transcripts were lost. This was 1.84% of the input transcripts and 2.48% of the total fallout. After the pipeline, total fallout added up to 409,605 transcripts, which was 74.15% of the input transcripts. Table 2.4 shows numbers of rat input transcripts rejected at each AutoDB filter. Figure 2.4 shows that the majority of input transcripts are filtered out by the transcript and exon quality filter stage for all three of the transcriptomes.

**Table 2.2** Human input transcript rejection rates at each AutoDB filter.

| Stage | Filter | Number transcript rejected | % total transcript rejected |
|---|---|---|---|
| 1 | BLAT | 8,436 | 0.18% |
| 2 | SIM4 | 625 | 0.01% |
| 3 | Transcript + Exon Quality | 3,150,291 | 67.96% |
| 4 | Cluster Quality | 16,153 | 0.35% |

**Table 2.3** Mouse input transcript rejection rates at each AutoDB filter.

| Stage | Filter | Number transcript rejected | % total transcript rejected |
|---|---|---|---|
| 1 | BLAT | 5,784 | 0.17% |
| 2 | SIM4 | 0 | 0% |
| 3 | Transcript + Exon Quality | 2,153,688 | 64.83% |
| 4 | Cluster Quality | 13,034 | 0.39% |

**Table 2.4** Rat input transcript rejection rates at each AutoDB filter.

| **Stage** | **Filter** | **Number transcript rejected** | **% total transcript rejected** |
|---|---|---|---|
| 1 | BLAT | 3,801 | 0.69% |
| 2 | SIM4 | 11 | 0.002% |
| 3 | Transcript + Exon Quality | 395,624 | 71.61% |
| 4 | Cluster Quality | 10,169 | 1.84% |

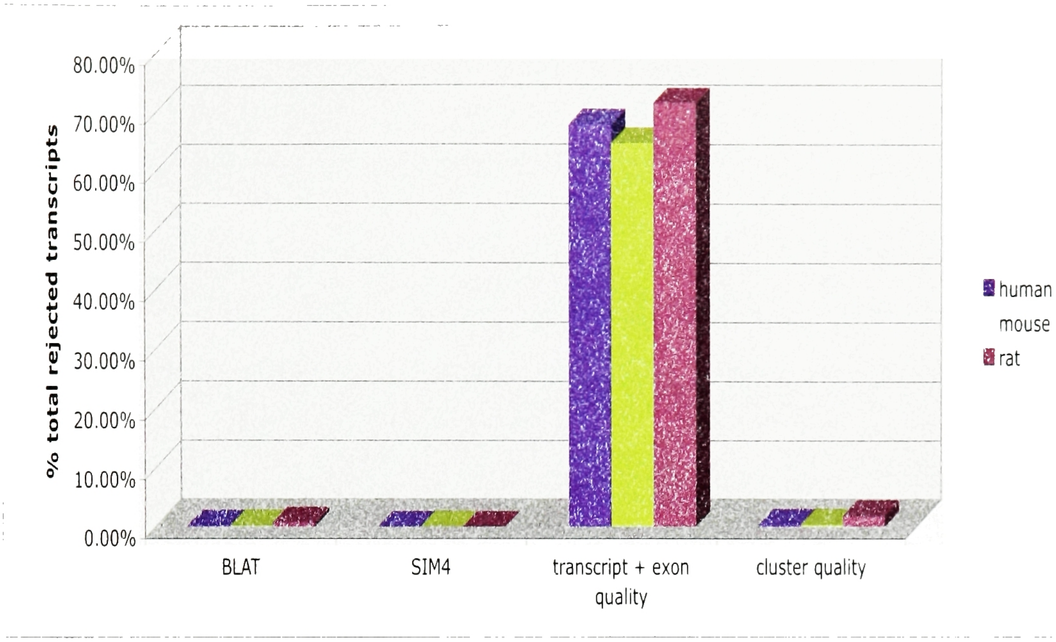


**Figure 2.4** Percent rejection rates of input transcripts by AutoDB filters.

## 2.5 Distribution of Alternative and Constitutive Exons

Table 2.5 shows the distribution of alternative and constitutive exons in human, mouse and rat transcriptomes. The computed human, mouse and rat transcriptomes contain a total of 218,353, 200,274 and 138,659 exons, 37%, 28% and 14% of which are alternative exons respectively. In all three transcriptomes, the majority of alternative exons are cassette exons. In human, 67% of alternative exons are cassette exons, as are 62% of mouse and 62% of rat alternative exons (Fig. 2.5). As shown in Table 2.5, the majority of cassette exons are length invariant. In human, 77% of cassette exons do not vary in nucleotide length. This percentage is 82% for mouse and 93% for rat cassette exons. Given the fact that the majority of alternative exons are cassette exons, our analyses described in the following chapters focus on this type of variant exons.

Length-variant exons generally vary either on the 5' end or on the 3' end of the exon. Very few length-variant exons vary on both the 5' and the 3' ends of the exon. Figures 2.6, 2.7 and 2.8 show a comparison of length-variant exon distribution to length-variant intron distribution. In human, in contrast to both 5' and 3' length variation in 7% of the exons, 68% of the introns are both 5' and 3' variant (Fig. 2.6). In mouse, in contrast to both 5' and 3' length variation in 5% of the exons, 71% of the introns vary at both ends (Fig. 2.7). Likewise, in rat 3% of the exons vary both at 5' and at 3' ends compared to an 81% in the introns (Fig. 2.8). These findings suggest an intron-centric rather

**Table 2.5** Exon Analysis of HumanSDB3, MouSDB5 and RatSDB2

| | HumanSDB3 | MouSDB5 | RatSDB2 |
|---|---|---|---|
| **Total number of exons** | 218353 | 200274 | 138659 |
| **Total number of constitutive length-invariant exons** | 137150 | 143780 | 118595 |
| **Total number of alternative exons** | 81203 | 56494 | 20064 |
| **Total number of length-variant exons** | 26528 | 21489 | 7607 |
| **Number of constitutive 5' and 3' length-variant exons** | 1978 | 1140 | 215 |
| **Number of constitutive 5' length-variant exons** | 12761 | 10721 | 3864 |
| **Number of constitutive 3' length-variant exons** | 11789 | 9628 | 3528 |
| **Total number of internal cassette exons** | 54675 | 35005 | 12457 |
| **Number of length-invariant internal cassette exons** | 42136 | 28817 | 11615 |
| **Number of 5' and 3' length-variant internal cassette exons** | 1804 | 642 | 31 |
| **Number of 5' length-variant internal cassette exons** | 5407 | 2945 | 397 |
| **Number of 3' length-variant internal cassette exons** | 5328 | 2601 | 414 |

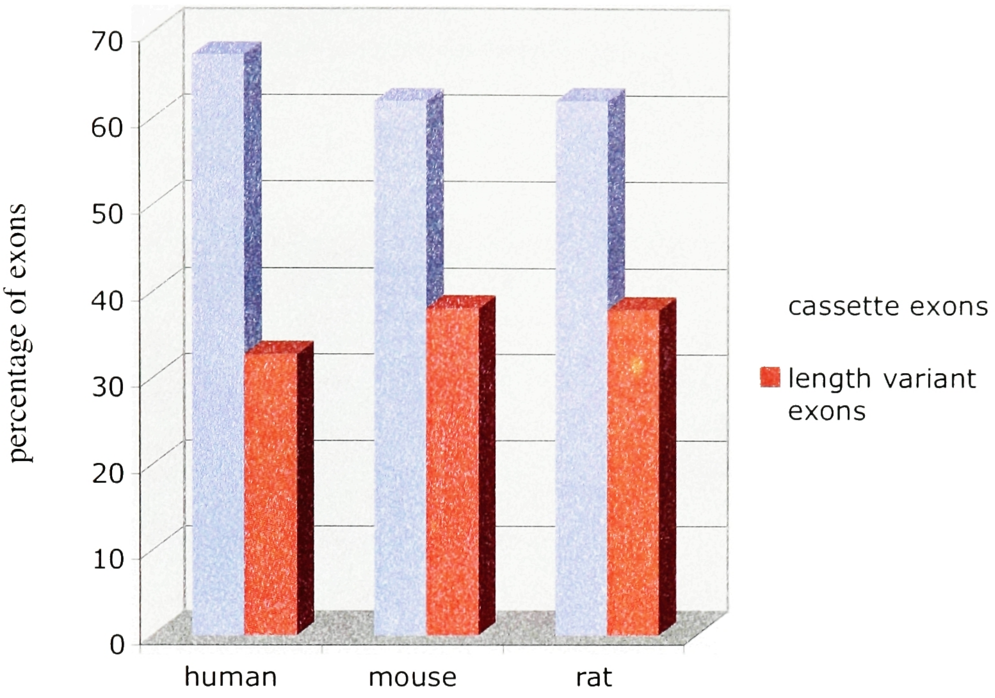


**Figure 2.5** The majority of alternative exons are cassette exons. 67% of human, 62% of mouse and 62% of rat alternative exons are cassette exons.

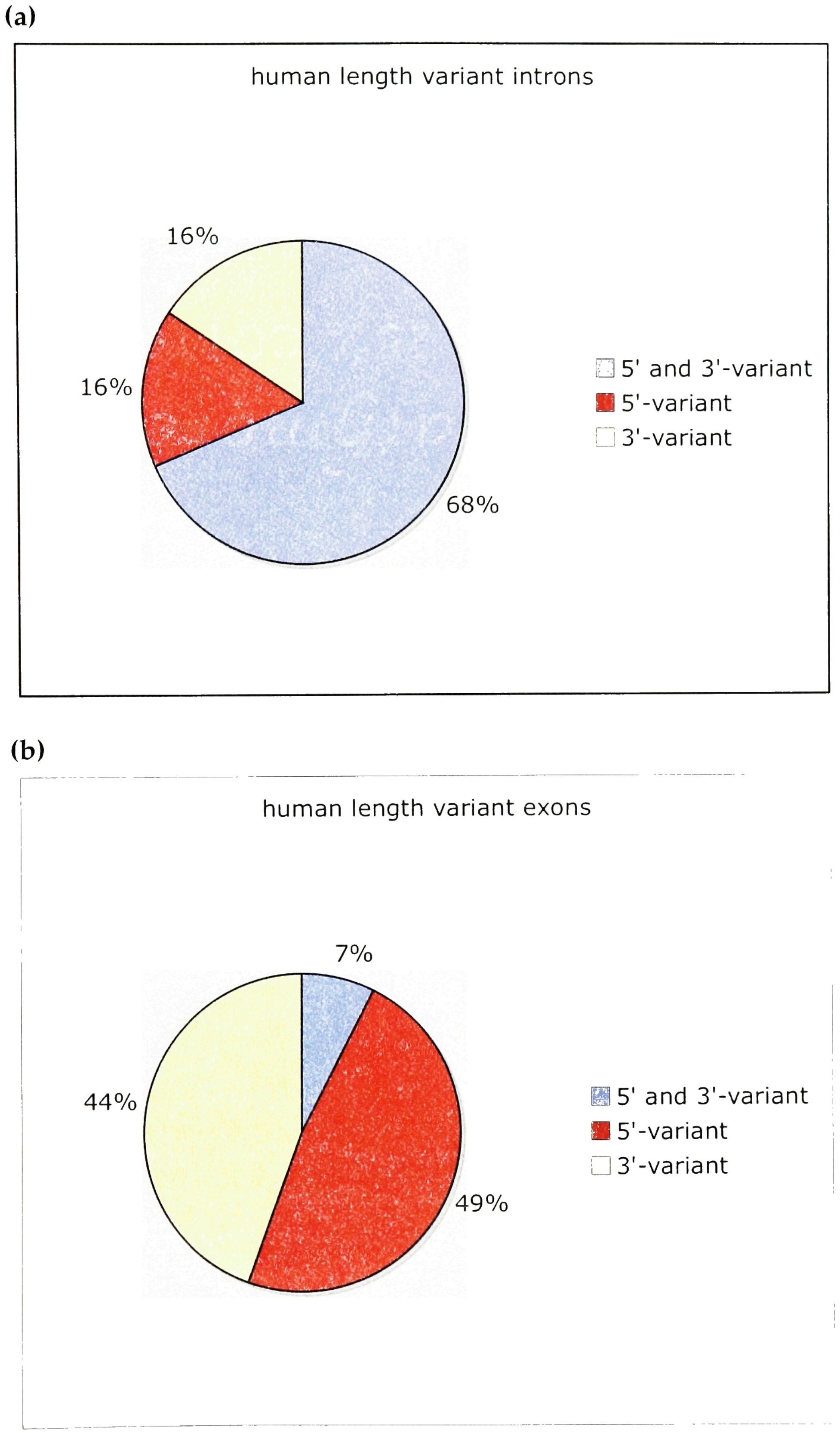


**Figure 2.6** **(a)** Distribution of length variation in HumanSDB3 introns. **(b)** Distribution of length variation in HumanSDB3 exons.

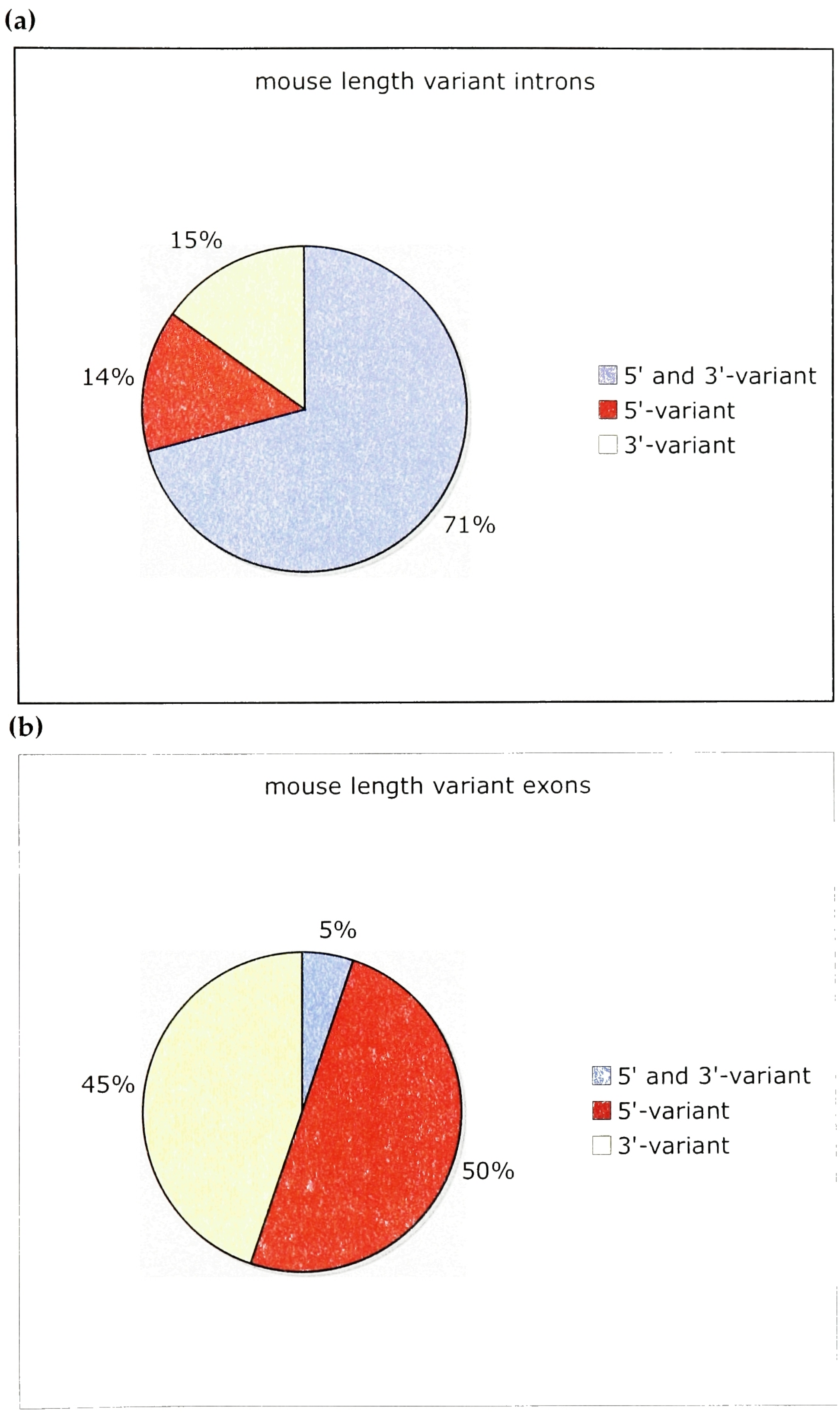


**Figure 2.7 (a)** Distribution of length variation in MouSDB5 introns. **(b)** Distribution of length variation in MouSDB5 exons.

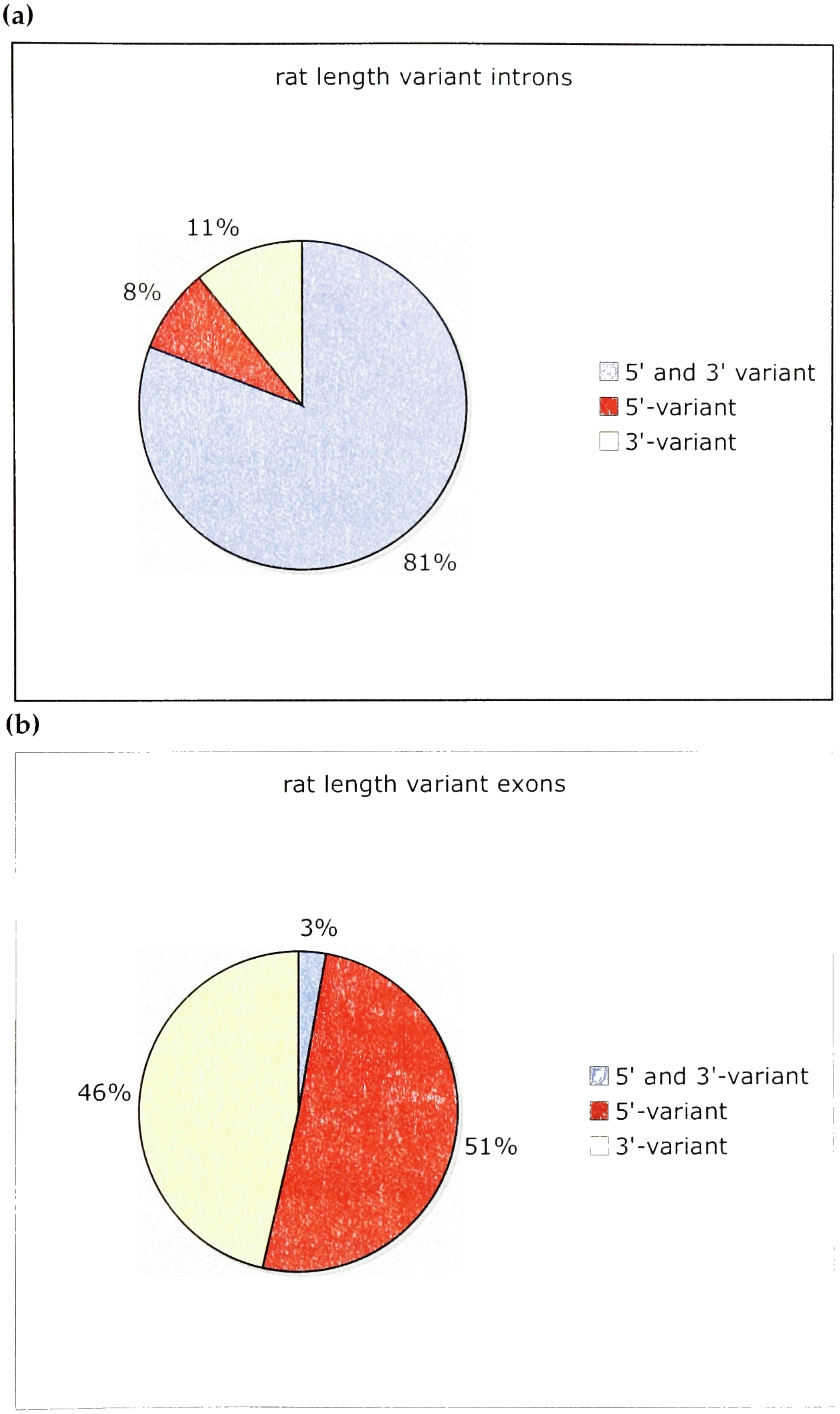


**Figure 2.8 (a)** Distribution of length variation in RatSDB2 introns. **(b)** Distribution of length variation in RatSDB2 exons.

than an exon-centric regulation of length variation in alternative exons. If length variation was controlled in an exon-centric manner, then distribution of 5' and 3' length variation in introns should have been lower or similar in numbers compared to the same distribution in exons.

## 2.6 Transcript Terminal Exons

We observed a specific subset of cassette exons present either at the 5' or at the 3' ends of transcripts, which do not map to any genomic exons of the clusters they are present in. These exons rather map to intronic regions. We termed these exons as *transcript-terminal cassette exons* to distinguish them from the *internal cassette exons* discussed in Section 2.5.

Transcript-terminal cassette exons occur in high numbers in human, mouse and rat transcriptomes. Table 2.6 provides statistics of occurrences of transcript-terminal cassette exons in HumanSDB3, MouSDB5 and RatSDB2. Figure 2.9 shows that transcript-terminal cassette exons are widely present in all three transcriptomes. In human, 30% of all cassette exons are transcript-terminal. In mouse and rat these percentages are 32% and 26% respectively. Transcript-terminal cassette exons elevate alternative exon numbers by 22% in human, 22% in mouse and 18% in rat transcrioptomes. In addition, we show that in 1234 of human variant clusters, 1475 of mouse variant clusters and 817 of rat variant clusters, variation is due to transcript-terminal cassette exons. These constitute 7% of all variant clusters in human, 10% of all variant clusters in mouse and 12% of all variant clusters in rat (Fig. 2.10).

**Table 2.6** Transcript Terminal Exon Analysis of HumanSDB3, MouSDB5 and RatSDB2

| | **HumanSDB3** | **MouSDB5** | **RatSDB2** |
|---|---|---|---|
| **Total number of alternative exons** | 104674 | 72652 | 24500 |
| **Total number of length-variant exons** | 26528 | 21489 | 7607 |
| **Total number of cassette exons** | 78146 | 51163 | 16893 |
| **Total number of internal cassette exons** | 54675 | 35005 | 12457 |
| **Total number of terminal cassette exons** | 23471 | 16158 | 4436 |

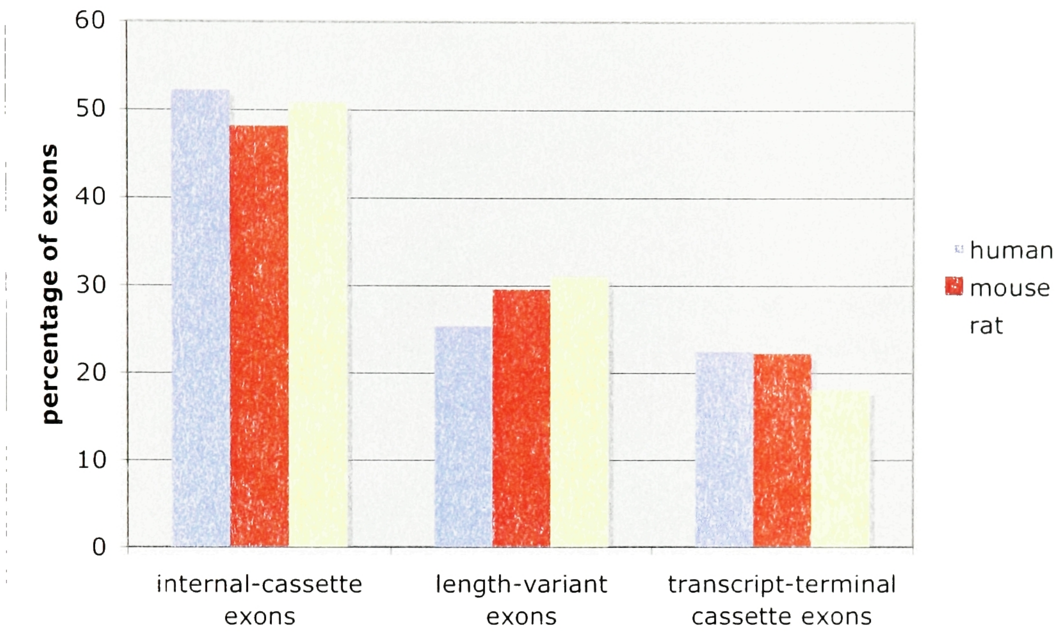


**Figure 2.9** Prevalence of transcript-terminal cassette exons in human, mouse and rat transcriptomes. Blue bars denote human alternative exons. Red bars denote mouse alternative exons. Yellow bars denote rat alternative exons. Percentage of internal cassette exons are highest in all three transcriptomes. Transcript-terminal cassette exons are almost as prevalent as length-variant exons. Transcript-terminal exons elevate the alternative exon numbers in human by 22%, in mouse by 22% and in rat by18% respectively.

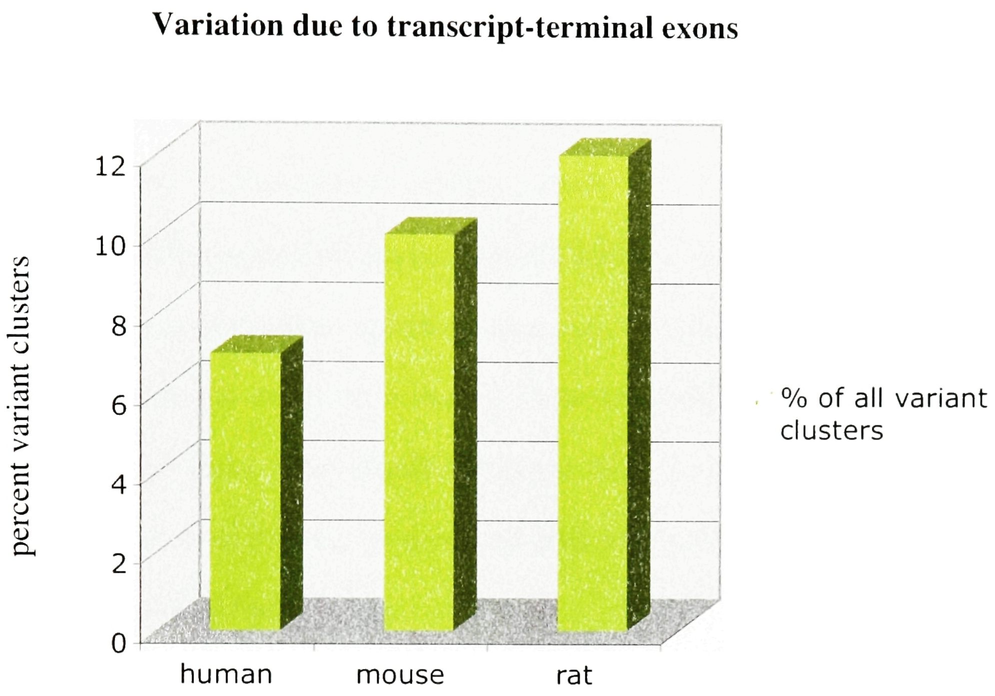


**Figure 2.10** Variant clusters due to transcript-terminal cassette exons. 7% of human variant clusters, 10% of mouse variant clusters and 12% of rat variant clusters are variant due to presence of transcript-terminal cassette exons.

Our analysis showed for the first time high prevalence of transcript-terminal cassette exons in human, mouse and rat transcriptomes. These exons are either due to noise in the sequence data or they might be due to relatively rare transcription events such as alternative initiation or termination of transcription (Zavolan *et al.*, 2003).

In our alternative splicing analyses, presented in the following chapters, we focus only on internal cassette exons to minimize aberrant splicing in the data. Our results indicate that it is important to consider differences between transcript-terminal and internal cassette exons in future analyses of alternative splicing.

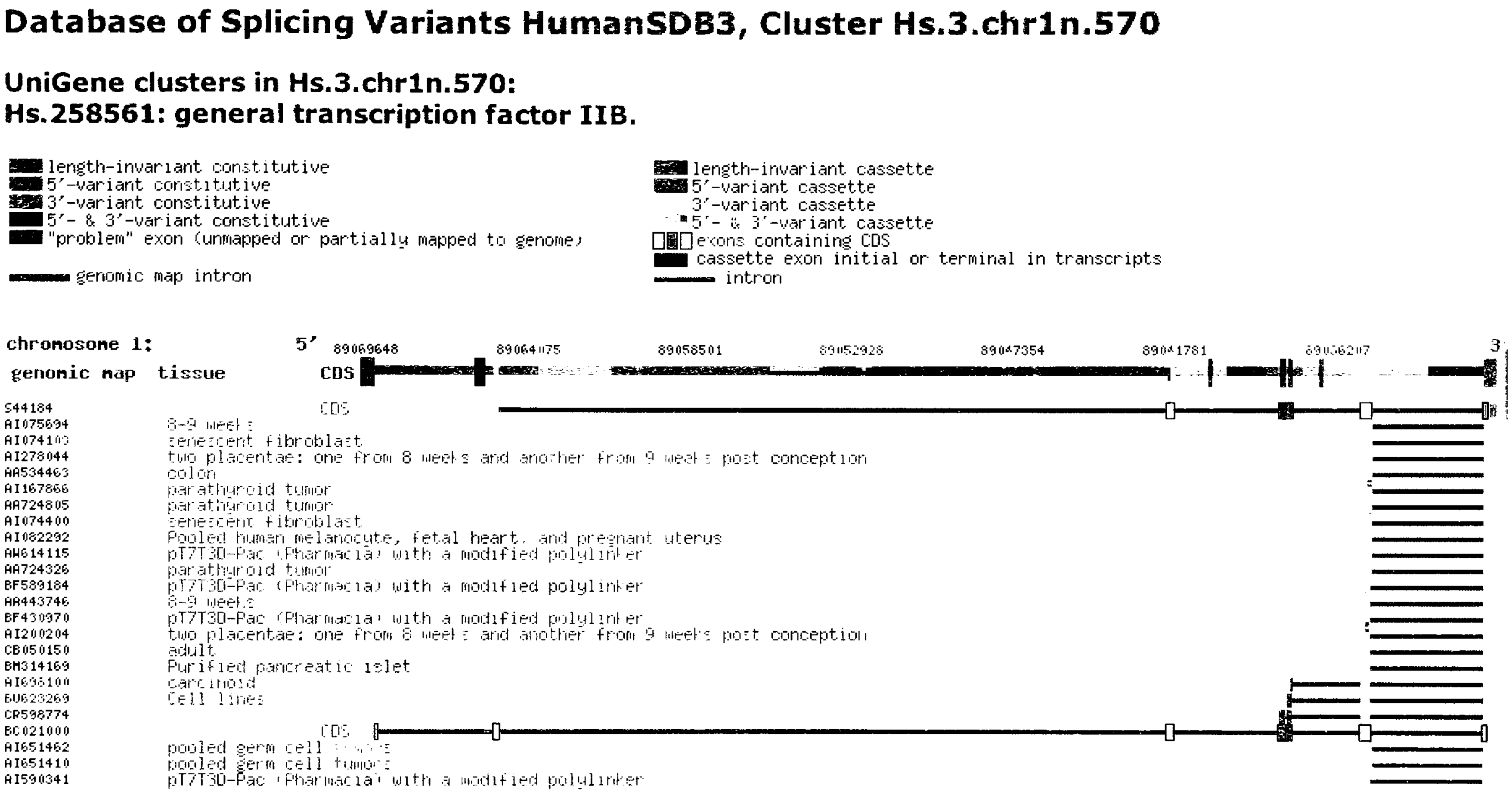
## 2.7 Graphical User Interfaces of Alternative Splicing Databases

HumanSDB3, MouSDB5 and RatSDB2 alternative splicing databases can be browsed through web-based graphical user interfaces. AutoDB hosts all three databases at http://genomes.rockefeller.edu/autodb/sdb.php (also available at http://sgc.ucsd.edu/autodb/sdb.php). HumanSDB3 can be accessed at http://genomes.rockefeller.edu/autodb/sdb.php?db=HumanSDB3 (also available at http://sgc.ucsd.edu/autodb/sdb.php?db=HumanSDB3), MouSDB5 can be accessed at http://genomes.rockefeller.edu/autodb/sdb.php?db=MouSDB5 (also available at http://sgc.ucsd.edu/autodb/sdb.php?db=MouSDB5) and

RatSDB2 can be accessed at http://genomes.rockefeller.edu/autodb/sdb.php?db=RatSDB2 (also available at http://sgc.ucsd.edu/autodb/sdb.php?db=RatSDB2).

Web implementations of the databases provide access to each variant and invariant splicing cluster along with database statistics. Through these interfaces, users can access database statistics on transcripts, chromosomes, clusters and exons. When available, annotation and library information for each transcript are provided. Nucleotide sequences of each alternative and constitutive exon are also given. Splice clusters can be accessed by providing as input any of the following parameters of the user's gene of interest: annotation, chromosome number, cluster ID, cluster type or UniGene ID.

Figures 2.11, 2.12 and 2.13 show screenshots of variant HumanSDB3, MouSDB5 and RatSDB2 clusters respectively. Figure 2.11 shows HumanSDB3 variant cluster representing human general transcription factor IIB. Figure 2.12 shows MouSDB5 variant cluster representing mouse general transcription factor IIB. Figure 2.13 shows RatSDB2 variant cluster representing rat general transcription factor IIB. Web implementation of the databases is further described in Materials and Methods, Section 7.5.



## Database of Splicing Variants HumanSDB3, Cluster Hs.3.chr1n.570

### UniGene clusters in Hs.3.chr1n.570:
### Hs.258561: general transcription factor IIB.

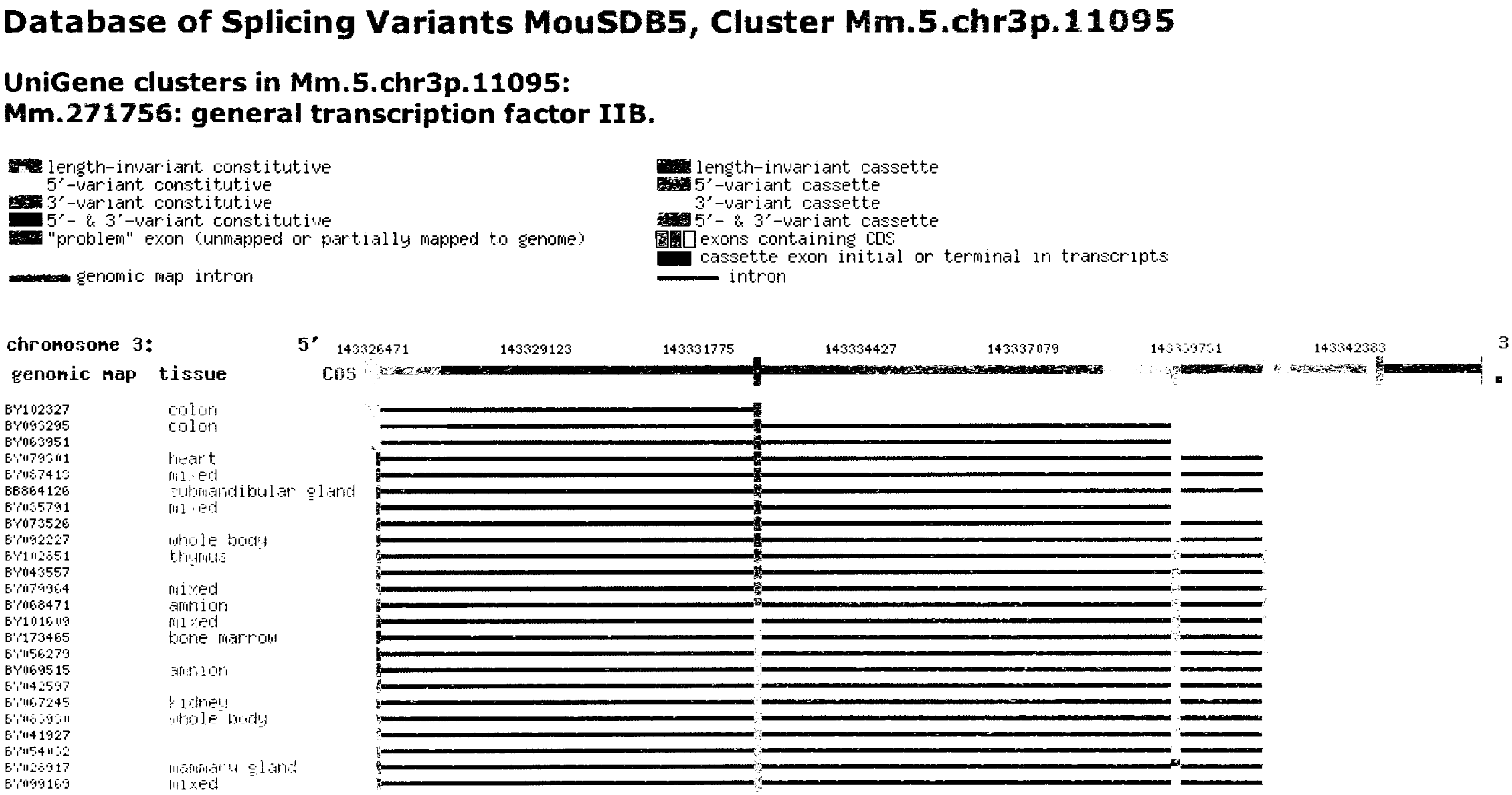
▆ length-invariant constitutive
▆ 5′-variant constitutive
▆ 3′-variant constitutive
▆ 5′- & 3′-variant constitutive
▆ "problem" exon (unmapped or partially mapped to genome)

▬ genomic map intron

▆ length-invariant cassette
▆ 5′-variant cassette
3′-variant cassette
5′- & 3′-variant cassette
□▆□ exons containing CDS
▆ cassette exon initial or terminal in transcripts
▬ intron

| genomic map | tissue |
|---|---|
| S44184 | CDS |
| AI075694 | 8-9 weeks |
| AI074103 | senescent fibroblast |
| AI278044 | two placentae: one from 8 weeks and another from 9 weeks post conception |
| AA534463 | colon |
| AI167866 | parathyroid tumor |
| AA724805 | parathyroid tumor |
| AI074400 | senescent fibroblast |
| AI082292 | Pooled human melanocyte, fetal heart, and pregnant uterus |
| AW614115 | pT7T3D-Pac (Pharmacia) with a modified polylinker |
| AA724326 | parathyroid tumor |
| BF589184 | pT7T3D-Pac (Pharmacia) with a modified polylinker |
| AA443746 | 8-9 weeks |
| BF430970 | pT7T3D-Pac (Pharmacia) with a modified polylinker |
| AI200204 | two placentae: one from 8 weeks and another from 9 weeks post conception |
| CB050150 | adult |
| BM314169 | Purified pancreatic islet |
| AI696100 | carcinoid |
| BU623269 | Cell lines |
| CR598774 | |
| BC021000 | CDS |
| AI651462 | pooled germ cell tumors |
| AI651410 | pooled germ cell tumors |
| AI590341 | pT7T3D-Pac (Pharmacia) with a modified polylinker |

**Figure 2.11** Human General Transcription Factor IIB. Screenshot from HumanSDB3 cluster Hs.3.chr1n.570. This figure displays a partial view of the ESTs and the full-length transcripts mapping to this variant locus on human chromosome 5. Full image can be viewed at http://genomes.rockefeller.edu/autodb/cluster_map.php?cluster_id=Hs.3.chr1n.570&db=HumanSDB3 (also at http://sgc.ucsd.edu/autodb/cluster_map.php?cluster_id=Hs.3.chr1n.570&db=HumanSDB3).



Database of Splicing Variants MouSDB5, Cluster Mm.5.chr3p.11095

UniGene clusters in Mm.5.chr3p.11095:
Mm.271756: general transcription factor IIB.

▨ length-invariant constitutive
5'-variant constitutive
▨ 3'-variant constitutive
■ 5'- & 3'-variant constitutive
▨ "problem" exon (unmapped or partially mapped to genome)

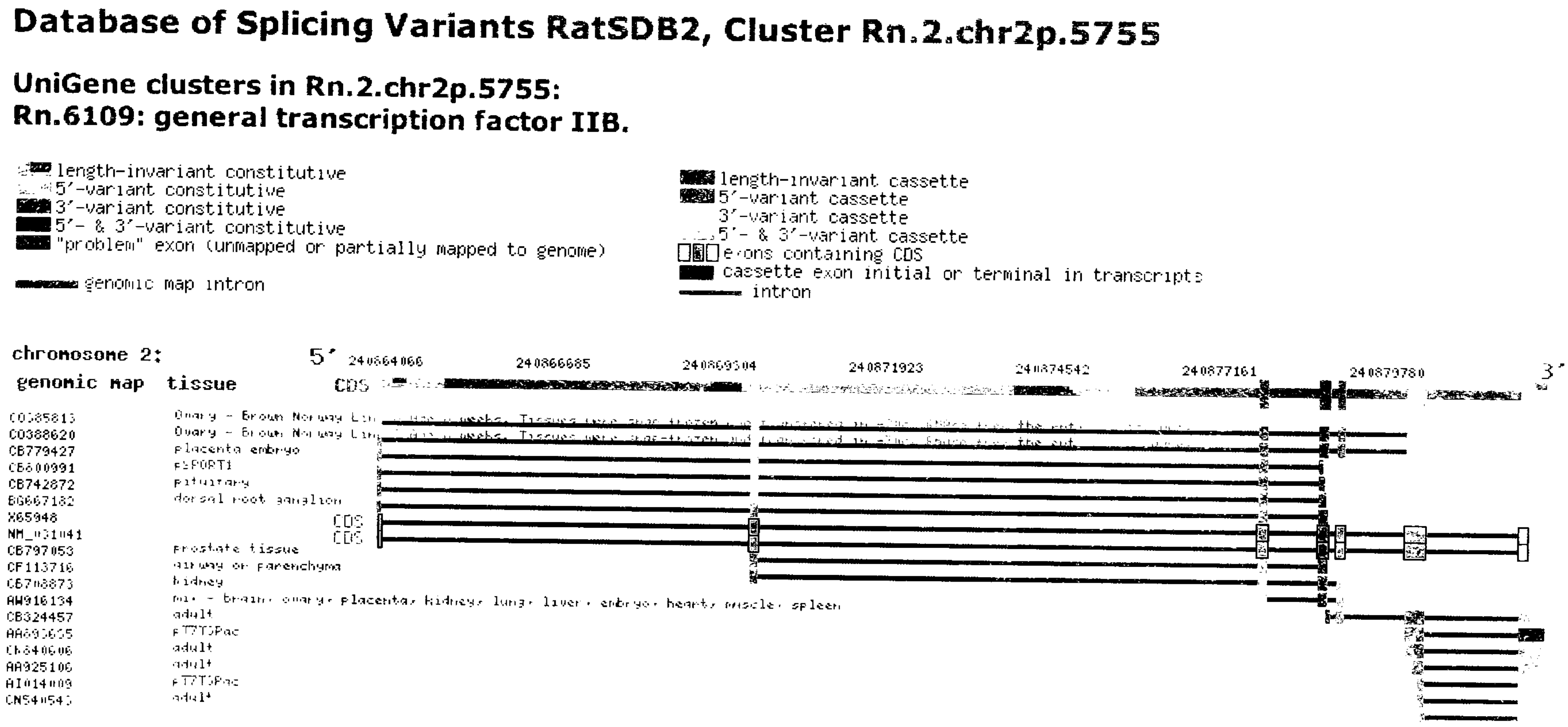
▬ genomic map intron

▨ length-invariant cassette
▨ 5'-variant cassette
3'-variant cassette
▨ 5'- & 3'-variant cassette
▨■□ exons containing CDS
■ cassette exon initial or terminal in transcripts
— intron

chromosome 3:

| genomic map | tissue |
|---|---|
| BY102327 | colon |
| BY093295 | colon |
| BY063951 | |
| BY079301 | heart |
| BY067413 | mixed |
| BB864126 | submandibular gland |
| BY035791 | mixed |
| BY073526 | |
| BY092227 | whole body |
| BY102651 | thymus |
| BY043557 | |
| BY079964 | mixed |
| BY068471 | amnion |
| BY101609 | mixed |
| BY173465 | bone marrow |
| BY056279 | |
| BY069515 | amnion |
| BY042597 | |
| BY067245 | kidney |
| BY063950 | whole body |
| BY041927 | |
| BY054032 | |
| BY026917 | mammary gland |
| BY099169 | mixed |

**Figure 2.12** Mouse General Transcription Factor IIB. Screenshot from MouSDB5 cluster Mm.5.chr3p.11095. This figure displays a partial view of ESTs and full-length transcripts mapping to this variant locus on mouse chromosome 3. Full image can be viewed at http://genomes.rockefeller.edu/autodb/cluster_map.php?cluster_id=Mm.5.chr3p.11095&db=MouSDB5 (also at http://sgc.ucsd.edu/autodb/cluster_map.php?cluster_id=Mm.5.chr3p.11095&db=MouSDB5).



## Database of Splicing Variants RatSDB2, Cluster Rn.2.chr2p.5755

### UniGene clusters in Rn.2.chr2p.5755:
### Rn.6109: general transcription factor IIB.

▨ length-invariant constitutive
▨ 5′-variant constitutive
▨ 3′-variant constitutive
■ 5′- & 3′-variant constitutive
■ "problem" exon (unmapped or partially mapped to genome)
▬ genomic map intron

▨ length-invariant cassette
▨ 5′-variant cassette
3′-variant cassette
▨ 5′- & 3′-variant cassette
□▨□ exons containing CDS
■ cassette exon initial or terminal in transcripts
▬ intron

chromosome 2:

| genomic map | tissue |
|---|---|
| CO385813 | Ovary - Brown Norway Lin… weeks. Tissues were snap frozen … in … the … |
| CO388620 | Ovary - Brown Norway Lin… weeks. Tissues were snap-frozen and … in … the cut |
| CB779427 | placenta embryo |
| CB600991 | FSPORT1 |
| CB742872 | pituitary |
| BG667182 | dorsal root ganglion |
| X65948 | CDS |
| NM_031041 | CDS |
| CB797053 | prostate tissue |
| CF113716 | airway or parenchyma |
| CB708873 | kidney |
| AW916134 | mix - brain, ovary, placenta, kidney, lung, liver, embryo, heart, muscle, spleen |
| CB324457 | adult |
| AA693655 | pT7T3Pac |
| CK840606 | adult |
| AA925106 | adult |
| AI014009 | pT7T3Pac |
| CN540543 | adult |

**Figure 2.13** Rat General Transcription Factor IIB. Screenshot from RatSDB2 cluster Rn.2.chr2p.5755. This figure displays a view of ESTs and full-length transcripts mapping to this variant locus on rat chromosome 2. The figure can be browsed at http://genomes.rockefeller.edu/autodb/cluster_map.php?cluster_id=Rn.2.chr2p.5755&db=RatSDB2 (also at http://sgc.ucsd.edu/autodb/cluster_map.php?cluster_id=Rn.2.chr2p.5755&db=RatSDB2).

## 2.8 *Compare Species* Feature of the Databases

This section introduces the interactive web implementation of our databases which allows queries by the end-user to compare orthologous alternatively spliced gene-pairs. Figure 2.14 shows the flowchart of this feature for HumanSDB3 and MouSDB5 orthologous cluster comparison. Our *Compare Species* feature allows the end-user to compare any two orthologous transcripts and their corresponding exons. As shown in Figure 2.15, our program aligns the human and mouse genomic exon maps. Lines going across the opposite exons indicate the orthologous human-mouse exons. In this figure, top line represents the genomic exon map for human general transcription factor IIH and bottom line represents the genomic exon map for the orthologous mouse general transcription factor IIH.

The *Compare Species* feature of HumanSDB3 and MouSDB5 databases can be accessed at http://genomes.rockefeller.edu/autodb/compare_1.php?db1=HumanSDB3&db2=MouSDB5 (also available at http://sgc.ucsd.edu/autodb/compare_1.php?db1=HumanSDB3&db2=MouSDB5). End-users are able to find their gene-pair of interest by providing as input any of the following parameters: Keyword, Gene Symbol, Splicing Cluster ID, GenBank accession number or UniGene cluster ID. This feature is also available for orthologous human-rat and orthologous mouse-rat genes.

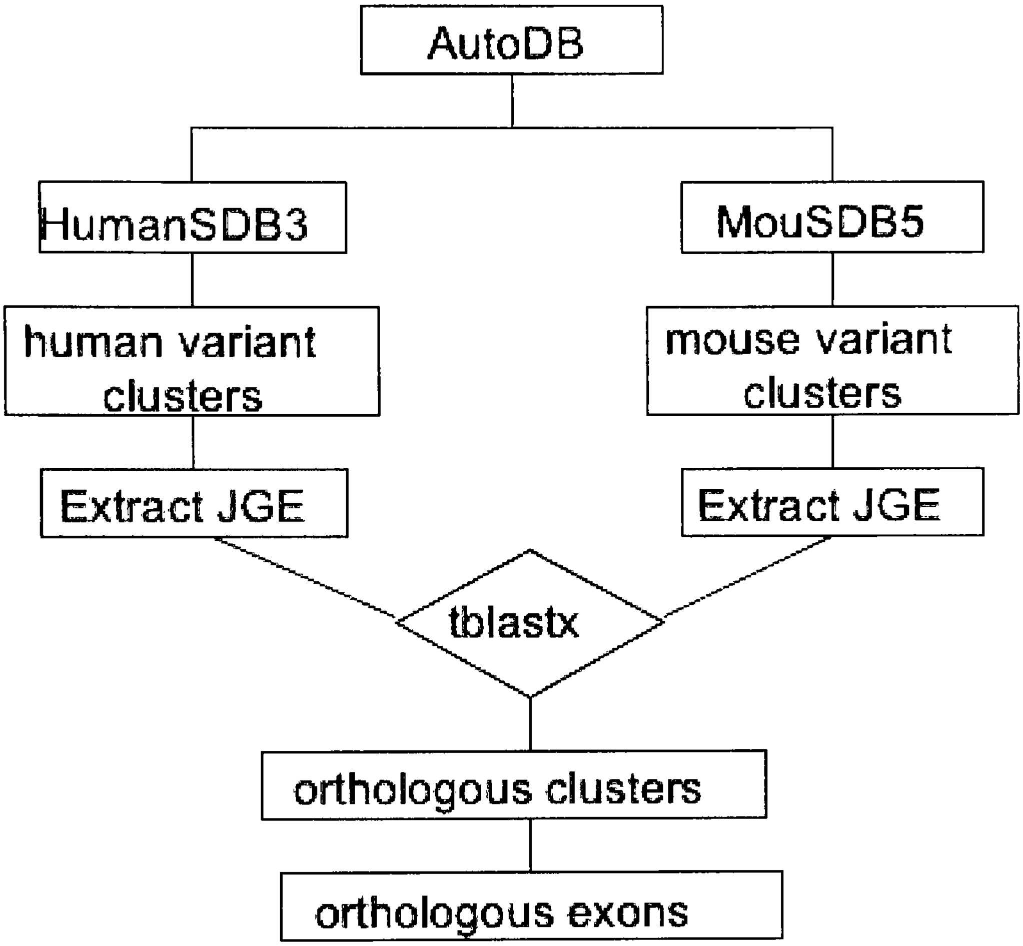
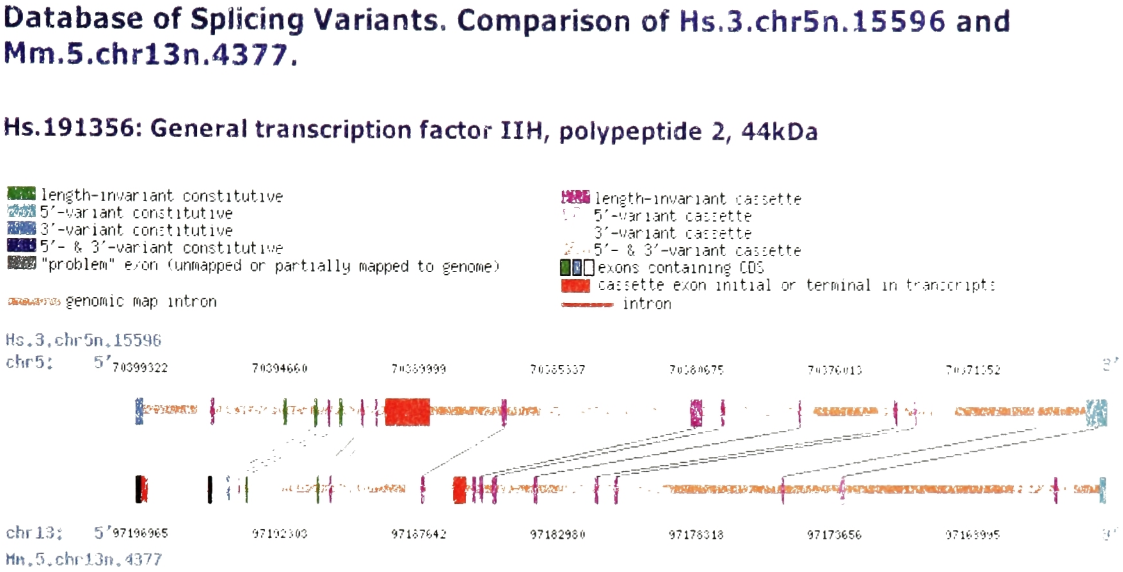


**Figure 2.14** Flowchart of *Compare Species* Feature of the databases. Joined-genomic exons (JGEs) from variant clusters of HumanSDB3 and MouSDB5 are utilized to establish orthology of human-mouse gene pairs.



Database of Splicing Variants. Comparison of Hs.3.chr5n.15596 and Mm.5.chr13n.4377.

Hs.191356: General transcription factor IIH, polypeptide 2, 44kDa

🟩 length-invariant constitutive
▨ 5′-variant constitutive
▨ 3′-variant constitutive
■ 5′- & 3′-variant constitutive
▨ "problem" exon (unmapped or partially mapped to genome)

▨ genomic map intron

▨ length-invariant cassette
▨ 5′-variant cassette
3′-variant cassette
▨ 5′- & 3′-variant cassette
🟩■□ exons containing CDS
🟥 cassette exon initial or terminal in transcripts
▬ intron

**Figure 2.15** Comparison of exons in an orthologous human-mouse gene pair for general transcription factor IIH, polypeptide 2. Top line represents genomic exon map of human variant cluster Hs.3.chr5n.15596. Bottom line represents genomic exon map of mouse variant cluster Mm.5.chr13n.4377. Orthologous exons are indicated by lines going across from the two genomic exon maps.

HumanSDB3 and RatSDB2 comparison feature can be accessed at http://genomes.rockefeller.edu/autodb/compare_1.php?db1=HumanSDB3&db2=RatSDB2 (also available at http://sgc.ucsd.edu/autodb/compare_1.php?db1=HumanSDB3&db2=RatSDB2). MouSDB5 and RatSDB2 comparison feature can be accessed at http://genomes.rockefeller.edu/autodb/compare_1.php?db1=MouSDB5&db2=RatSDB2 (also available at http://sgc.ucsd.edu/autodb/compare_1.php?db1=MouSDB5&db2=RatSDB2 ). Development of the *Compare Species* feature is further described in Materials and Methods, Section 7.6.

## 2.9 Discussion

The databases presented here provide a comprehensive overview of alternative splicing events in human, mouse and rat transcriptomes. These databases identify, classify, compute, store and answer queries about splice variants within these three transcriptomes. They provide the user with information about all transcripts that map to a certain genomic locus. Complete sequences of transcripts, their annotations and when available the tissues from which they are sequenced from are provided. All exons of each transcript within a cluster are annotated as alternative or constitutive and their sequences are provided as links.

Our databases enable detailed computational analyses of alternative splicing. In addition, our databases open up possibilities for experimental design to further study alternative splicing on a single gene level or at more global levels. The *Compare Species* feature of these databases allows users to view genomic maps of orthologous alternatively spliced genes in pairs. Users can visualize the orthologous gene structures and their corresponding exons. This information is readily accessible and along with our novel visualization tool, it facilitates easy comparative analysis of alternative splicing and allows studies of evolution of this important cellular process.

# CHAPTER 3

# ALTERNATIVE SPLICING OF MOUSE TRANSCRIPTION FACTORS

## 3.1 Introduction to Transcription, Transcription Factors and Gene Expression

As introduced in Chapter 1, transcription is a critical process which specifies mRNAs and thus proteins expressed within a cell. RNA polymerases, enzymes which catalyze transcription, bind to promoter sequences and move along the DNA template until they reach a stop codon. In addition to RNA polymerases, transcription requires many other proteins including transcription factors (TFs). Transcription factors help RNA polymerase bind to the regulatory sequences on DNA and facilitate transcription (Alberts *et al.*, 4th ed.).

Expression of a given gene is dependent on the interactions of different transcription factors and their cofactors with the regulatory regions of that gene. Different transcription factors regulate transcription in different ways. Transcription factors in turn are regulated by several processes which include interaction with other proteins and signaling cascades (Lopez, 1995). As discussed in Section 1.6, alternative splicing is thought to occur in coordination with transcription (Caceres and Kornblihtt, 2002). In this chapter, we describe a detailed study of alternative splicing of mouse transcription factors.

## 3.2 Regulation of Transcription Factors by Alternative Splicing

Alternative splicing is one of the mechanisms that regulates transcription factor activity. As noted by Lopez, alternative splicing can affect TF protein structure in two primary ways (Lopez, 1995). Alterations can be in the DNA-binding domains of TFs, affecting their affinity or specificity for regulatory regions on the DNA; or alterations can modulate interactions of TFs with their cofactors (Latchman, 2001). Such changes have been observed experimentally to alter specificity or binding strength or to switch between activator and repressor isoforms of the same transcription factor (Foulkes and Sassone-Corsi, 1992). For example, the cAMP-response element modulator has three different isoforms with entirely different DNA-binding domains, which are all transcription activators. On the other hand, isoforms of the human transcription factor AML1 function both as positive and as negative regulators of transcription (Lopez, 1995).

We used MouSDB3 (a prior release of MouSDB5), a database of mouse alternatively spliced transcripts, to study alternative splicing of transcription factors. As shown in Figure 3.1, in MouSDB3 out of the 461 TF clusters, 62% are variant, posing this group of proteins as a good candidate for alternative splicing studies.

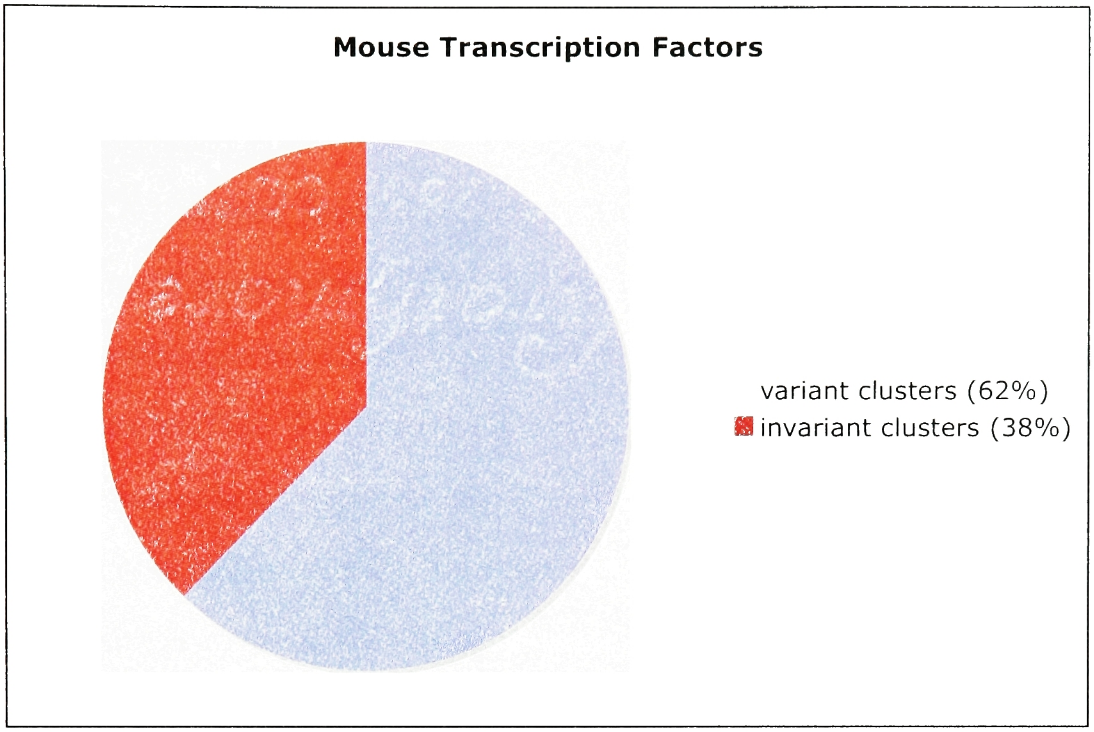


**Figure 3.1** Cluster Analysis of Mouse Transcription Factors. Out of a total of 461 Transcription Factor clusters in MouSDB3, 287 are variant and 174 are invariant.

## 3.3 Tissue Distribution of Alternatively Spliced Transcription Factor Transcripts

We analyzed tissue distribution of alternatively spliced transcription factors in 18 tissues chosen from the existing libraries in MouSDB3. These tissues are chosen because they contain both variant and invariant transcripts annotated as TFs. To account for library ambiguities, libraries which list several tissues or cell types for a single entry such as *mixture of brain and testis* and libraries which provide no tissue information at all such as *embryo* or *carcinoma* are screened out. In addition, for some tissues, several parts of a given tissue are pooled into one term. For example, the term *brain* corresponds to all parts of the brain found in MouSDB3 libraries, including cerebellum, thalamus, hippocampus and 16 other libraries. Methods used for the tissue distribution analysis are described in Chapter 7, Section 7.7.

As shown in Figure 3.2a, our comparative analysis revealed that in 14 out of the 18 tissues analyzed, the proportion of TFs that are variant is higher than the proportion of all genes that are variant. Eight tissues exhibited more than two-fold differences in variant TFs versus variant genes in total (Fig. 3.2b). Note that values in Figure 3.2b are base 2 logarithms of the ratios. Tissues with more than two-fold differences have $\log_2$ values above 1 on the graph. In salivary gland, skeletal muscle, urinary bladder and testis, the fold-differences are 8.7, 5.6, 3.8 and 3.0-fold respectively. These tissues are followed by spinal cord, liver, adipose tissue and eye which also exhibit more than two-fold differences.

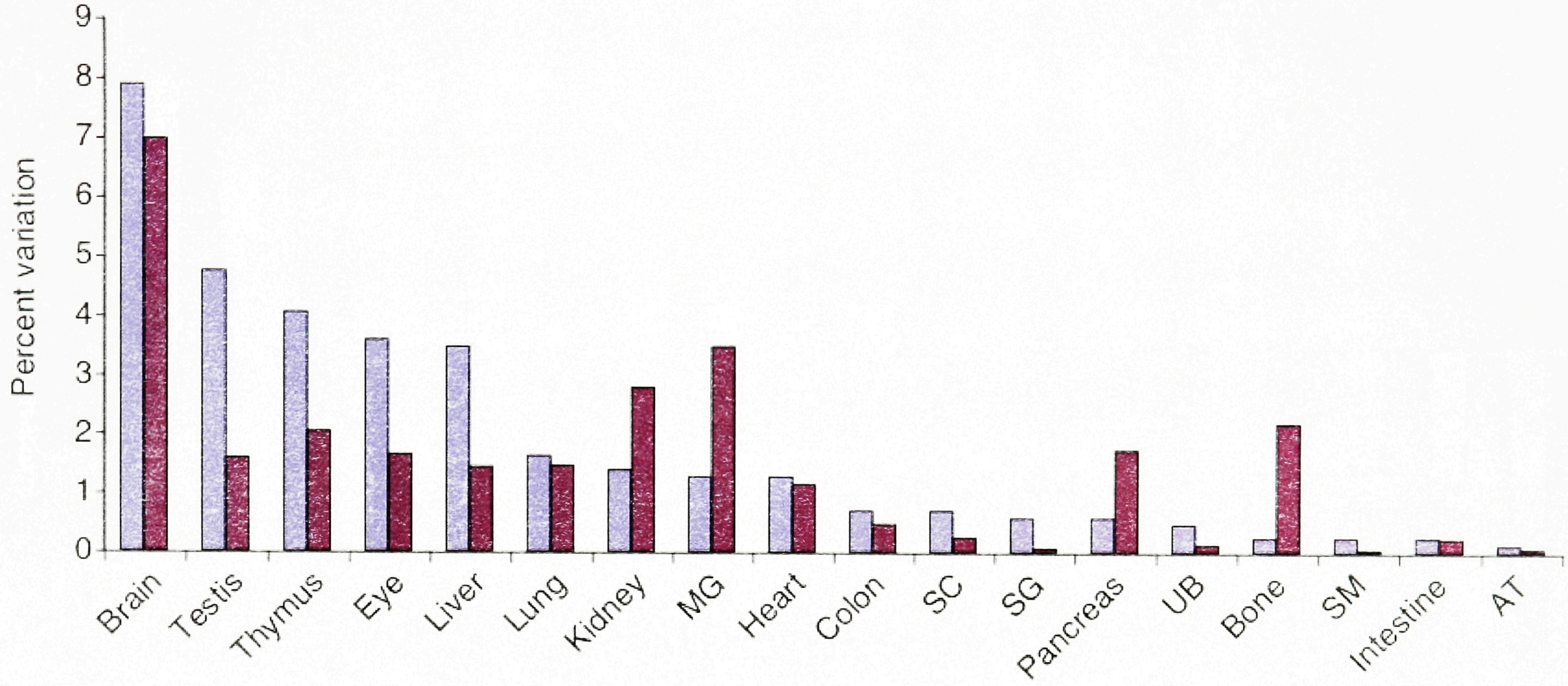


**Figure 3.2a** TF variation is higher in the majority of tissues compared to variation in all genes. For each tissue the number of variant TF transcripts in tissue is normalized by the total number of TF transcripts in MouSDB3 and is represented as a blue bar. This number is computed as follows; t = number of variant TF transcripts in tissue; T = total number of variant TF transcripts; bar value=(t/Tx100). Red bars represent the number of variant transcripts of all genes in the tissue normalized by the total number of variant transcripts in MouSDB3. This number is computed as follows; a = total number of variant transcripts in tissue; A = total number of all variant transcripts in MouSDB3; bar value=(a/Ax100). Tissue abbreviations; MG: mammary gland, SC: spinal cord, SG: salivary gland, UB: urinary bladder, SM: skeletal muscle, AT: adipose tissue.

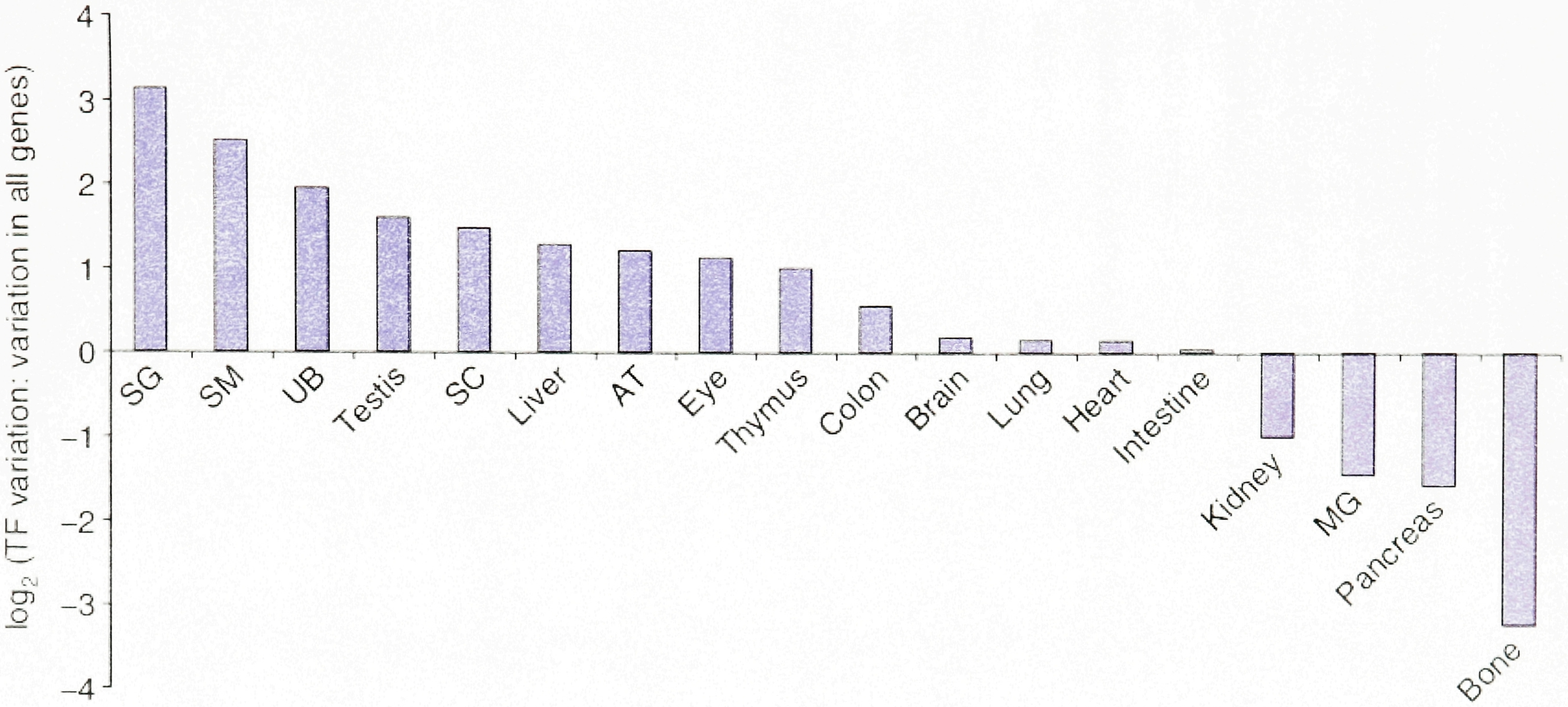


**Figure 3.2b** Fold differences between number of variant TF transcripts and number of all variant genes. This value is computed as follows; bar value=log2((t/T)/(a/A)). Please see Figure 3.3a legend for definitions of t, T, a and A. Tissues are in descending order from highest to lowest fold difference of variation in TF versus variation in all genes. Tissue abbreviations; SG: salivary gland, SM: skeletal muscle, UB: urinary bladder, SC: spinal cord, AT: adipose tissue, MG: mammary gland.

Figures 3.3 and 3.4 show that the values of fold-differences shown in Figure 3.2b are independent of sampling depth of the transcripts from these tissues. Sampling depth is the total number of transcripts sequenced per tissue. Figures 3.3a and 3.4a display absolute numbers of variant TF transcripts and absolute numbers of the entire variant transcripts of the transcriptome per given tissue respectively. The correlation coefficient of the absolute number of TFs and the fold-differences between variant TFs and all genes is –0.13, indicating that they do not correlate. Likewise, the correlation coefficient of the absolute number of all genes and the fold-differences between variant TFs and all genes is –0.46. In addition, the scatter-plots in Figures 3.3b and 3.4b show that there is no correlation between the sampling depth and the fold-differences. These results indicate that TFs exhibit higher levels of alternative splicing compared to the rest of the variant mouse genes.

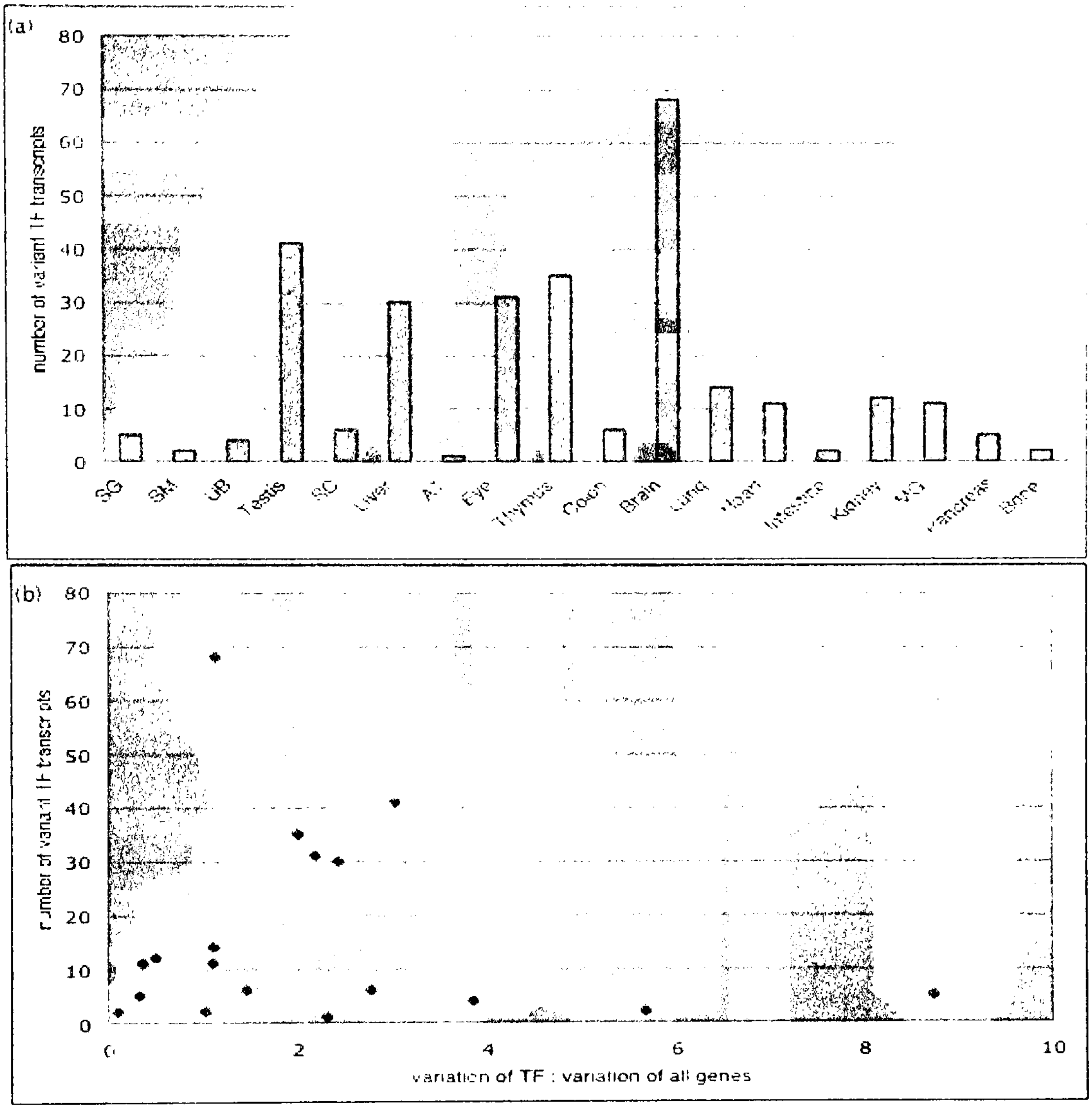


**Figure 3.3** Fold-differences are independent of sampling depths of transcripts from given tissues. **(a)** Absolute number of variant TF transcripts per tissue. y-axis: number of variant TF transcripts; x-axis: tissues. **(b)** For each tissue; x-axis: ratio of variant TF transcripts to all variant transcripts; y-axis: absolute number of variant TF transcripts.

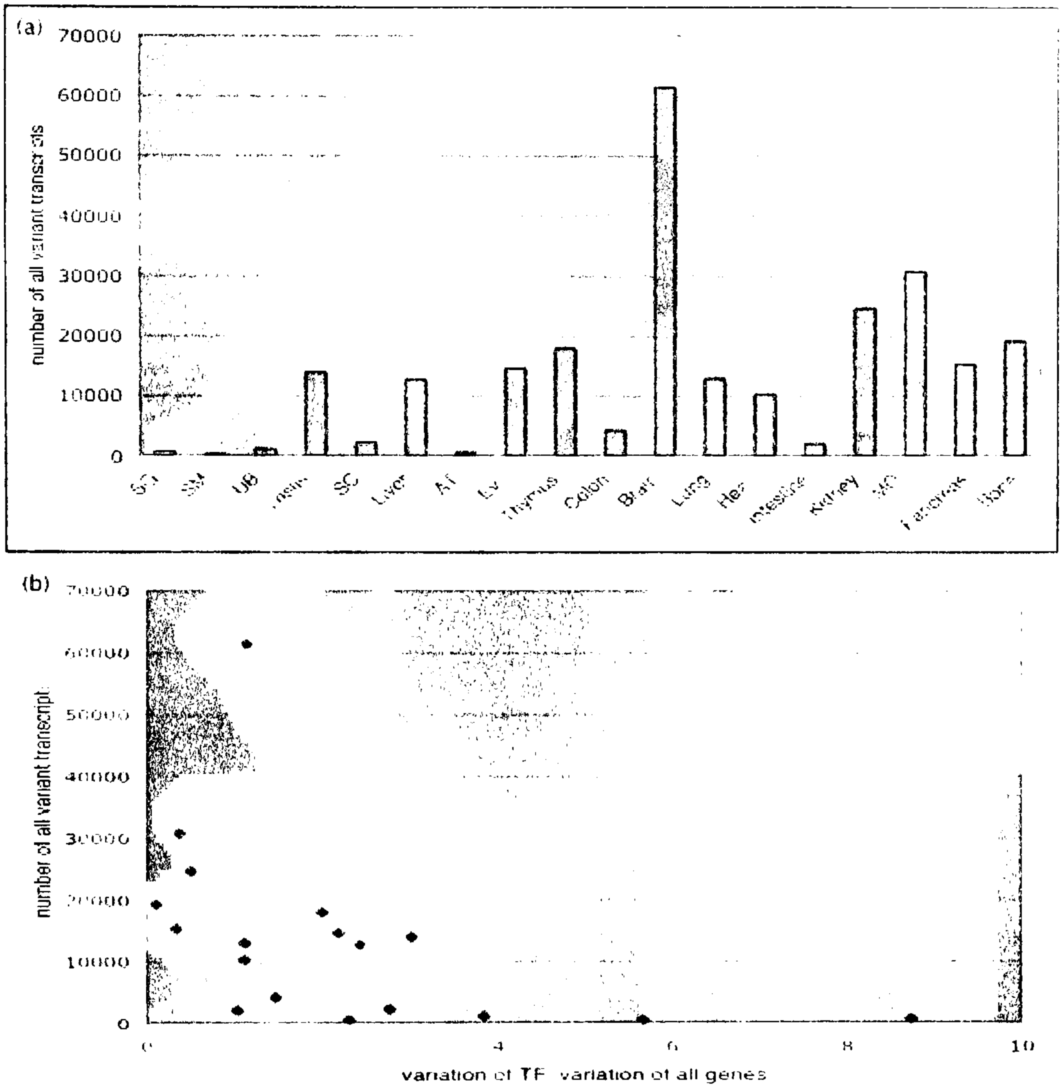


**Figure 3.4** Fold-differences are independent of sampling depths of transcripts from given tissues. **(a)** Absolute number of all variant transcripts per tissue. y-axis: number of all variant transcripts; x-axis: tissues. **(b)** For each tissue; x-axis: ratio of variant TF transcripts to all variant transcripts; y-axis: absolute number of all variant transcripts.

## 3.4 Homogeneity-Heterogeneity of Isoforms

To study tissue-specificity of alternatively spliced TF isoforms and other mouse genes, we defined four terms presented below in italics and further described in Materials and Methods, Section 7.7.

(1) *homogeneity within tissues:* all transcripts sequenced from the same tissue have cassette exons with same splice sites (Fig. 3.5).

(2) *heterogeneity within tissues:* transcripts sequenced from the same tissue have cassette exons with different splice sites (Fig. 3.6).

(3) *homogeneity across tissues:* transcripts sequenced from different tissues have cassette exons with same splice sites (Fig. 3.7).

(4) *heterogeneity across tissues:* transcripts sequenced from different tissues have cassette exons with different splice sites (Fig. 3.8).

We used the ratio of homogeneity-heterogeneity to assess tissue specificity. To determine the homogeneity-to-heterogeneity ratios of isoforms within the 18 tissues studied, we analyzed the presence of different isoforms of TFs within and across these tissues. Differences between isoforms were established based on splice sites of cassette exons within the coding sequences of transcripts. Variations due to 5' and 3' truncations of transcripts were not counted.

As shown in Figure 3.9, when heterogeneity-to-homogeneity ratios are compared within and across tissues, there is significantly higher heterogeneity of isoforms across tissues than within a single tissue (p-value = 0.04). This is true both for variant transcription factors and for all the variant genes in the mouse transcriptome. When single tissues are taken into account, TFs are more homogeneous within each tissue analyzed (Fig. 3.10). Heterogeneity to homogeneity ratios in all tissues are lower than 1, indicating homogeneity of TF isoforms within these tissues. When all genes are considered, heterogeneity to homogeneity ratios are also below 1, indicating homogeneity of isoforms of all variant genes within these tissues. However, TFs are significantly more homogeneous within a given single tissue when compared to all genes (p-value = 0.02). These findings indicate that the majority of TF isoforms and  isoforms of all alternatively spliced genes  differ across tissues; whereas, within a given single tissue there generally is only one isoform.

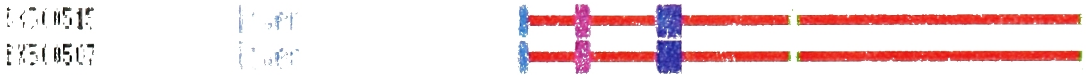


**Figure 3.5 Homogeneity of Isoforms Within Tissues.** Within this variant cluster all liver transcripts have the same cassette exon splice sites (i.e. they are all the same isoform). Pink boxes denote the cassette exons. Other boxes denote constitutive exons. Red lines denote intronic sequences. This cluster is considered to be *homogeneous within liver*.

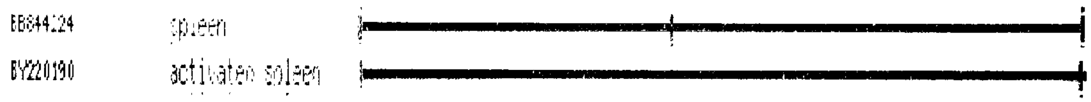


**Figure 3.6 Heterogeneity of Isoforms Within Tissues.** Within this variant cluster different spleen transcripts have different cassette exon splice sites (i.e. they are different isoforms). Pink box denotes the cassette exon. Blue and green boxes denote constitutive exons. Red lines denote intronic sequences. This cluster is considered to be *heterogeneous within spleen.*

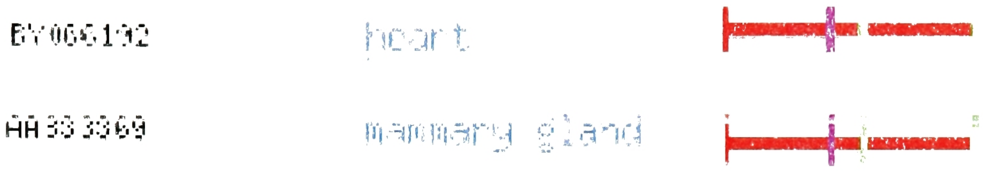


**Figure 3.7 Homogeneity of Isoforms Across Tissues.** Within this variant cluster, all heart transcripts and all mammary gland transcripts have the same cassette exon splice sites (i.e. they are all the same isoform). Pink box denotes the cassette exon. Red lines denote intronic sequences. This cluster is considered to be *homogeneous across heart and mammary gland*. (For a cluster to be considered homogeneous across tissues, all tissues studied have to have the same cassette exon splice sites and for an across tissue comparison, the cluster has to be *homogenous within* all tissues analyzed.)

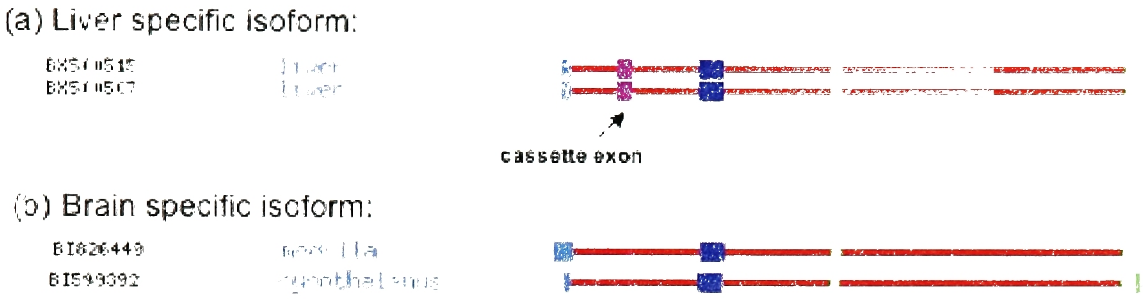


**Figure 3.8 Heterogeneity of Isoforms Across Tissues.** Within this variant cluster, liver transcripts and brain transcripts have different cassette exon splice sites (i.e. they are different isoforms). Pink boxes denote the cassette exon. Other boxes denote constitutive exons. Red lines denote intronic sequences. This cluster is considered to be *heterogeneous across liver and brain*. (For a cluster to be considered heterogeneous across tissues, at least one of the tissues studied has to have different cassette exon splice sites and all tissues studied must be homogeneous within themselves.)

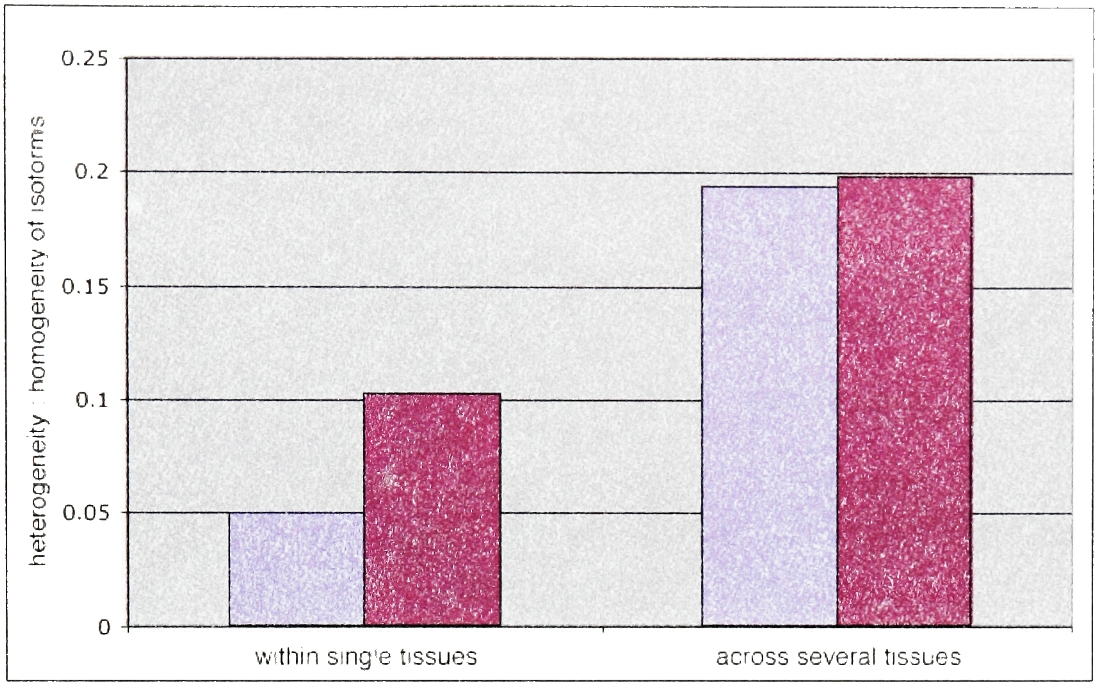


**Figure 3.9** Isoforms of alternatively spliced genes are more homogeneous within single tissues than across different tissues. Blue bars represent ratio of number of TF clusters with multiple isoforms within a given tissue to number of TF clusters with only a single isoform within that tissue. Red bars represent ratio of number of all variant clusters with multiple isoforms within a given tissue to number of all variant clusters with only a single isoform within that tissue.

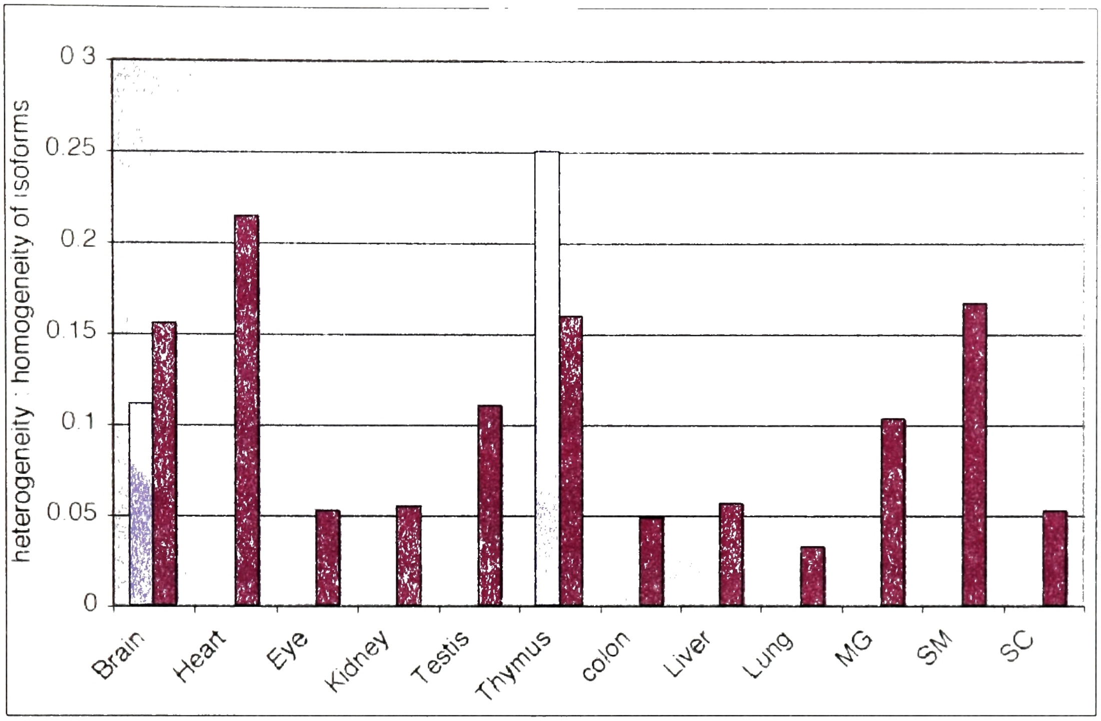


**Figure 3.10** Distribution of heterogeneity versus homogeneity of isoforms within single tissues. Blue bars represent ratio of TF clusters with multiple isoforms within the given tissues to TF clusters with only a single isoform within that tissue. Red bars represent ratio of all variant clusters with multiple isoforms within the given tissue to variant clusters with only a single isoform within that tissue. Tissue abbreviations; MG: mammary gland, SM: skeletal muscle, SC: spinal cord.

## 3.4 Effect of Cassette Exons on Transcription Factor Protein Domain Architecture

Analyzing proteins in the context of all available genome and transcript sequence data has the potential to reveal functional properties not accessible through protein sequence analysis alone. Alternative splicing has been shown to alter domain architectures of proteins (Kriventseva et al., 2003; Liu *et al.*, 2003; Cline *et al.*, 2004; Resch *et al.*, 2004b). To analyze the impact of alternative splicing on transcription factor protein structure, we screened 287 variant TF loci within MouSDB3 (a prior release of MouSDB5) for the presence of cassette exons within coding sequences. As shown in Figure 3.11, the 287 variant TF loci contain a total of 324 cassette exons, of which 23% (76 exons) are in-frame and in codon position 1. Only 11% of the cassette exons are expected to be multiples of three and in codon position 1 randomly. The two-fold difference between expected and observed numbers indicates a bias towards presence of in-frame cassette exons.

Of the transcripts containing the 76 in-frame cassette exons, 66 had domain architectures predicted by SMART (Ponting *et al.*, 1999; Letunic *et al.*, 2002; 2004). SMART is a search tool for domain architectures and can be accessed at http://smart.embl-heidelberg.de/. 80% (53 exons) of these exons induced a domain structure alteration when skipped. Of these 53 structure-altering exons, 68% (36 exons) were within coding regions for domains that are important for TF activity. The remaining 32% (17 exons) were proximal to the computed domain boundaries (Fig. 3.11). When the cassette exon was

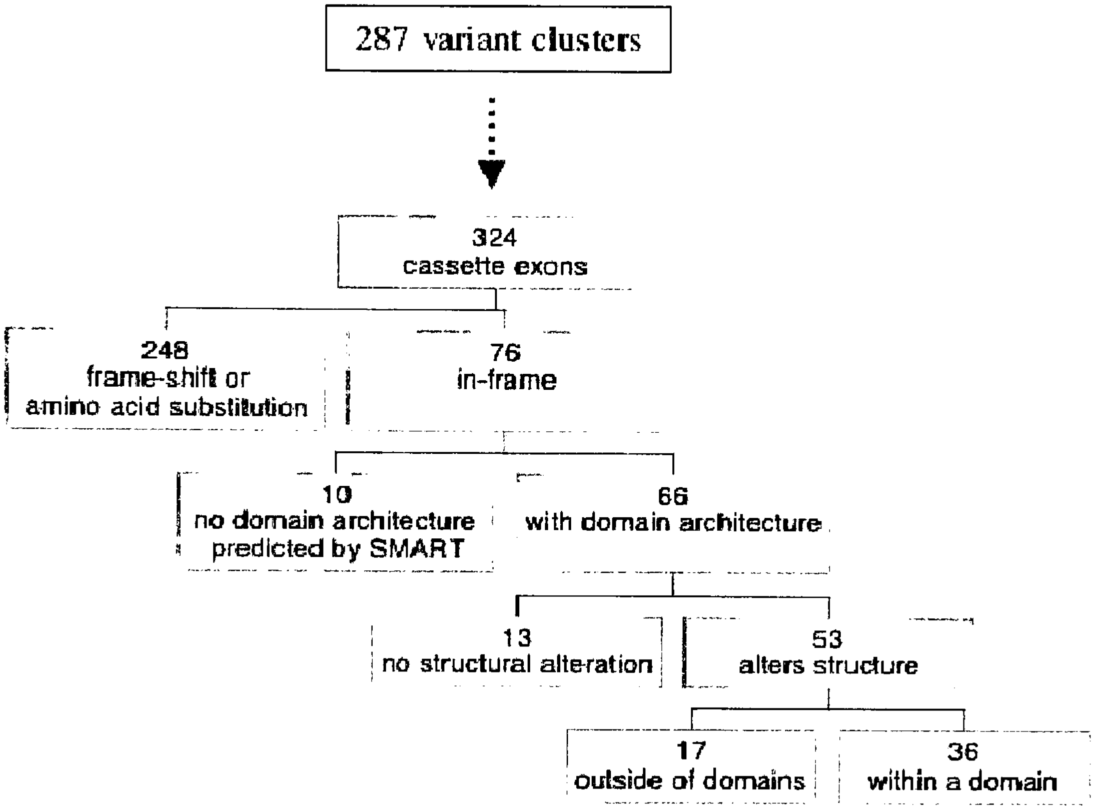


**Figure 3.11** Mouse transcription factor cassette exon analysis. 287 variant TF clusters have a total of 324 cassette exons. When 76 of the 324 cassette exons are skipped, the altered transcripts are in-frame. Exclusion of remaining exons either introduces an amino-acid substitution or causes frame-shifting. Of the in-frame exons, 53 alter domain architecture and 13 do not. Of the exons that cause domain alteration, 36 are in coding regions for domains and 17 are proximal to the coding regions.

removed, the sequence no longer met the computational criteria for the domain.

Figure 3.12 illustrates loss of a zinc finger domain detected by SMART due to deletion of an in-frame cassette exon. Top panel illustrates the transcript, denoted as the *original transcript,* with the cassette exon which codes for a basic leucine zipper domain and a zinc finger domain shown on the right. Bottom panel illustrates the transcript, denoted as the *altered transcript,* which skips the cassette exon. This transcript codes only for the basic leucine zipper domain and is missing the zinc finger domain (shown to the right of the transcript).

SMART (Schultz *et al.,* 1998; 2000) and Pfam (Bateman *et al.,* 2002; 2004) annotations of the altered domains revealed that 75% of the domains affected by alternative splicing with known functions are DNA-binding domains. Pfam is a database of protein families and can be accessed at http://www.sanger.ac.uk/Software/Pfam/index.shtml. Figure 3.13 provides homeobox domain annotation as an example of a DNA-binding domain. Figure 3.14 shows a partial list from the web-page which displays the annotation of transcripts containing domains altered by cassette exons. Links to domain annotations and to GenBank entries for the transcripts are provided on this page. The full list can be accessed at http://genomes.rockefeller.edu/~bahar/TF.html (also available at http://sgc.ucsd.edu/~bahar/TF.html). Sequences of cassette exons can also be downloaded in forms of fasta files. Nucleotide and amino acid sequences of the original and the altered transcripts are also provided as fasta files.

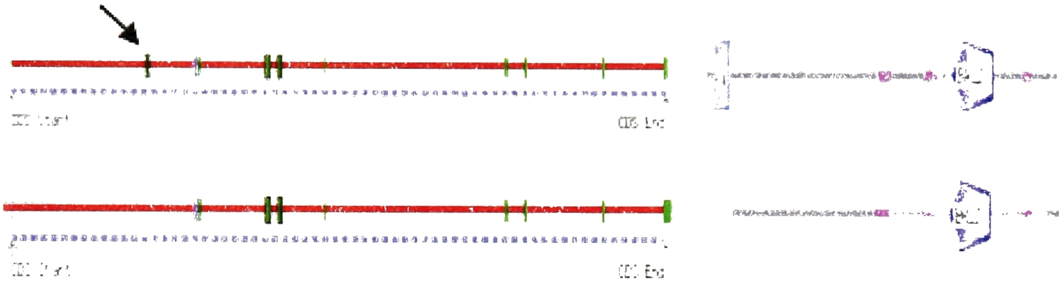
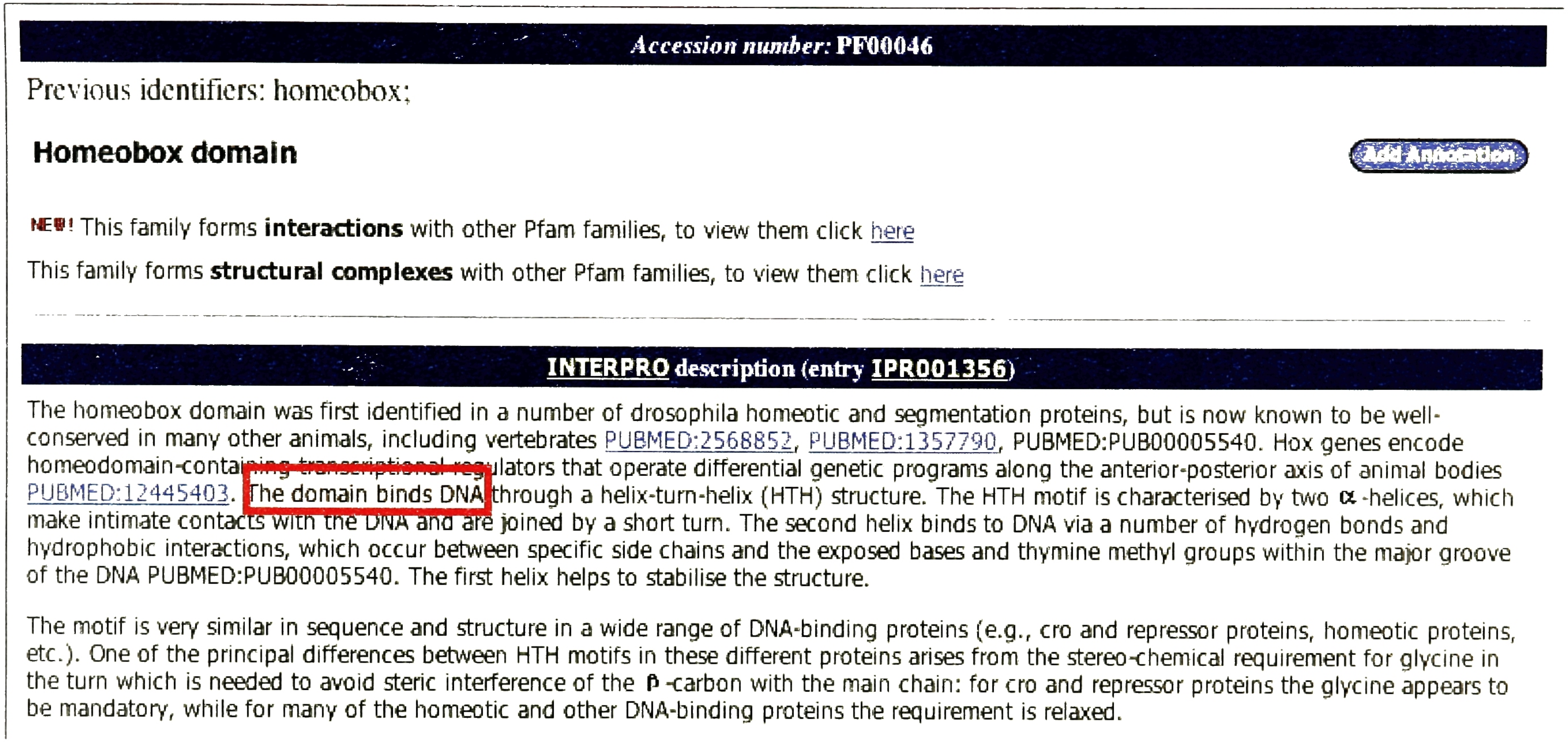


**Figure 3.12** Deletion of in-frame cassette exon, indicated by the arrow, leads to loss of detection of zinc finger domain by SMART. Top left panel shows the original transcript, bottom left panel shows the altered transcript. Right panels show protein domains coded by these transcripts.

BRLZ denotes basic region leucine zipper; ZnF C2H2 denotes zinc finger C2H2 and ▬ denotes low complexity regions.



Accession number: PF00046

Previous identifiers: homeobox;

**Homeobox domain**

NEW! This family forms **interactions** with other Pfam families, to view them click here

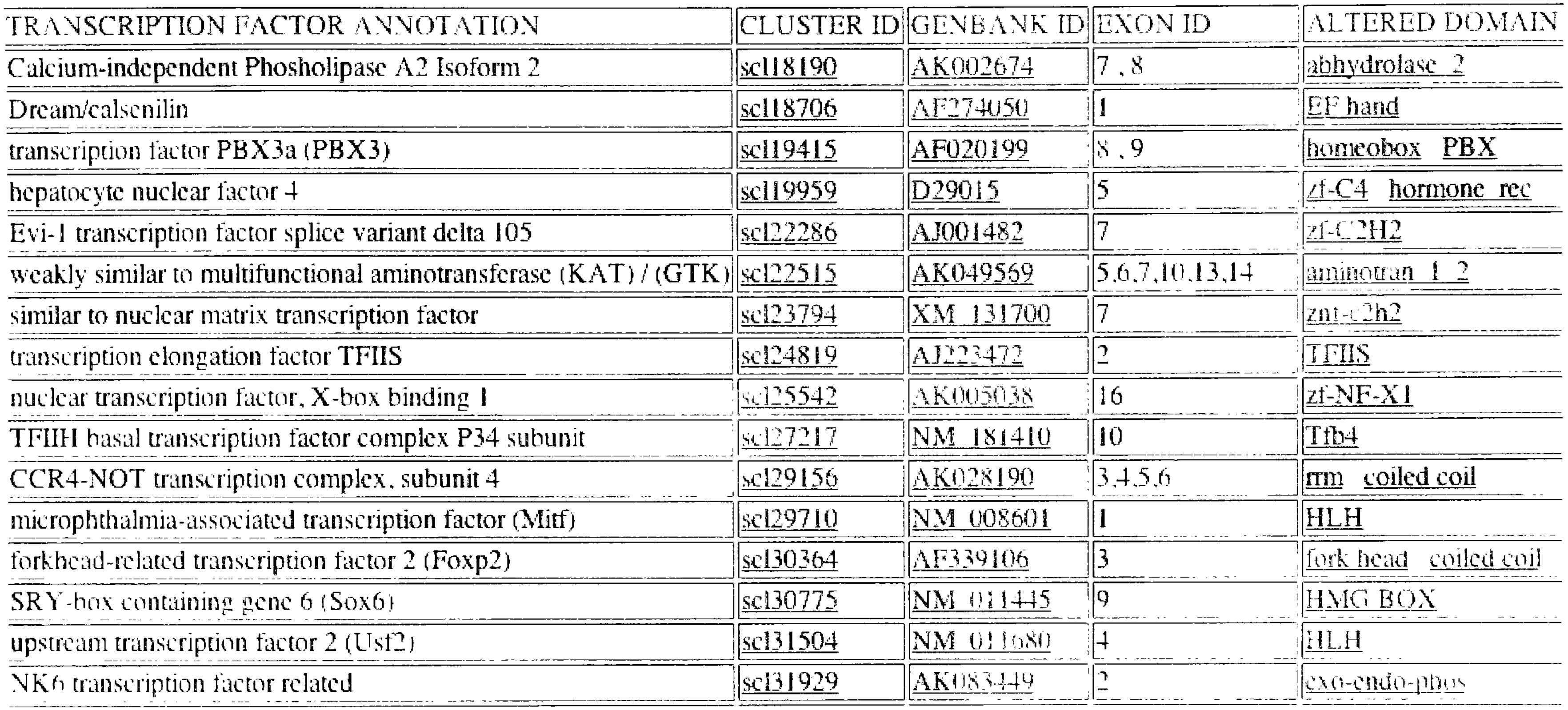
This family forms **structural complexes** with other Pfam families, to view them click here

INTERPRO description (entry IPR001356)

The homeobox domain was first identified in a number of drosophila homeotic and segmentation proteins, but is now known to be well-conserved in many other animals, including vertebrates PUBMED:2568852, PUBMED:1357790, PUBMED:PUB00005540. Hox genes encode homeodomain-containing transcriptional regulators that operate differential genetic programs along the anterior-posterior axis of animal bodies PUBMED:12445403. The domain binds DNA through a helix-turn-helix (HTH) structure. The HTH motif is characterised by two α-helices, which make intimate contacts with the DNA and are joined by a short turn. The second helix binds to DNA via a number of hydrogen bonds and hydrophobic interactions, which occur between specific side chains and the exposed bases and thymine methyl groups within the major groove of the DNA PUBMED:PUB00005540. The first helix helps to stabilise the structure.

The motif is very similar in sequence and structure in a wide range of DNA-binding proteins (e.g., cro and repressor proteins, homeotic proteins, etc.). One of the principal differences between HTH motifs in these different proteins arises from the stereo-chemical requirement for glycine in the turn which is needed to avoid steric interference of the β-carbon with the main chain: for cro and repressor proteins the glycine appears to be mandatory, while for many of the homeotic and other DNA-binding proteins the requirement is relaxed.

**Figure 3.13** Pfam annotation for homeobox domain, confirming that it binds DNA. This annotation can be found at http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00046.



| TRANSCRIPTION FACTOR ANNOTATION | CLUSTER ID | GENBANK ID | EXON ID | ALTERED DOMAIN |
|---|---|---|---|---|
| Calcium-independent Phosholipase A2 Isoform 2 | scl18190 | AK002674 | 7 , 8 | abhydrolase_2 |
| Dream/calsenilin | scl18706 | AF274050 | 1 | EF hand |
| transcription factor PBX3a (PBX3) | scl19415 | AF020199 | 8 , 9 | homeobox PBX |
| hepatocyte nuclear factor 4 | scl19959 | D29015 | 5 | zf-C4 hormone_rec |
| Evi-1 transcription factor splice variant delta 105 | scl22286 | AJ001482 | 7 | zf-C2H2 |
| weakly similar to multifunctional aminotransferase (KAT) / (GTK) | scl22515 | AK049569 | 5,6,7,10,13,14 | aminotran_1_2 |
| similar to nuclear matrix transcription factor | scl23794 | XM_131700 | 7 | znf-c2h2 |
| transcription elongation factor TFIIS | scl24819 | AJ223472 | 2 | TFIIS |
| nuclear transcription factor, X-box binding 1 | scl25542 | AK005038 | 16 | zf-NF-X1 |
| TFIIH basal transcription factor complex P34 subunit | scl27217 | NM_181410 | 10 | Tfb4 |
| CCR4-NOT transcription complex, subunit 4 | scl29156 | AK028190 | 3,4,5,6 | rrm coiled coil |
| microphthalmia-associated transcription factor (Mitf) | scl29710 | NM_008601 | 1 | HLH |
| forkhead-related transcription factor 2 (Foxp2) | scl30364 | AF339106 | 3 | fork head coiled coil |
| SRY-box containing gene 6 (Sox6) | scl30775 | NM_011445 | 9 | HMG BOX |
| upstream transcription factor 2 (Usf2) | scl31504 | NM_011680 | 4 | HLH |
| NK6 transcription factor related | scl31929 | AK083449 | 2 | exo-endo-phos |

**Figure 3.14** Domains altered by in-frame cassette exons (shown in the rightmost column). Partial view of the web-page containing links to variant clusters, annotations of transcripts and altered domains. The complete list can be accessed at http://genomes.rockefeller.edu/~bahar/TF.html (also at http://sgc.ucsd.edu/~bahar/TF.html).

Methods used in this analysis can be found in Chapter 7, Section 7.8. Appendix A provides a full list of domains altered by cassette exons along with GenBank ids for transcripts and variant exon numbers.

## 3.6 Discussion

In this work, we show that the majority of TF isoforms and isoforms of all alternatively spliced genes differ across tissues. Further, we show that within a given single tissue there generally is only one isoform. These findings suggest that alternative splicing creates tissue specific isoforms of TFs and thus can contribute to regulation of gene expression in a tissue specific manner.

Given the finding that in 78% of the tissues studied, TFs display higher proportions of variation compared to proportions of all genes in these tissues and the finding that 62% of all TF loci are variant, indicate the widespread impact of alternative splicing on regulation of gene expression via different TF isoforms.

In addition, the study presented here provides quantitative evidence that alternative splicing preferentially adds or deletes domains important to the DNA-binding function of TFs via inclusion or exclusion of in-frame cassette exons.

We provide quantitative evidence for structural and functional significance of variant exons in control of transcription and regulation of gene expression by alternative splicing via creation of tissue-specific TF isoforms. The work described here implies that future high-throughput screens of gene expression analyses should be sensitive to multiple alternatively spliced forms of TFs. Because gene expression arrays are intended to measure transcription, the next generation of arrays should contain probes specific to all known isoforms of genes represented on the arrays. In addition, this study opens up possibilities of detailed analysis of other groups of proteins which are widely affected by alternative splicing.

# CHAPTER 4

# PATHOLOGICAL SPLICING IN CANCER TISSUES

## 4.1 Introduction

Given that aberrant splicing has been widely documented in cancer cells and cancer tissues (Brinkman, 2004; Hui *et al.*, 2004; Venables, 2004; Kalnina *et al.*, 2005), we investigated the degree to which cancer ESTs add variation beyond normal alternative splicing. For this analysis, we used HumanSDB3, a database of alternative splicing in human transcriptome, which contained a total of 361 DNA sequence libraries from cancer related tissues and cells. Methods used for this analysis are described in detail in Chapter 7, Section 7.9.

## 4.2 Definitions

For this study, we defined the following terms presented in italics. *Normal-only clusters* contain transcripts from normal tissues only. *Cancer-only clusters* contain transcripts from cancer tissues only. *Mixed clusters* contain transcripts both from normal tissues and from cancer tissues. *Potentially pathological clusters* comprise cancer-only clusters and mixed clusters.

## 4.3 Extent of Cancer Transcripts in Human Transcriptome Data

Figure 4.1 illustrates the distribution of normal and cancer transcripts within human variant clusters. 13% of human variant clusters contain transcripts from normal tissues only. 14% of human variant clusters contain transcripts from cancer tissues only. The remaining 73% of clusters contain a combination of transcripts both from normal and from cancer tissues.

Figure 4.2 shows the same values within each tissue. All tissues studied have more than 80% of their clusters to have cancer transcripts with the exception of prostate for which this number is 73%. These clusters either contain all cancer transcripts or are a mixture of normal and cancer transcripts. For breast, brain, colon, lung, kidney, skin and ovary more than 10% of clusters are completely cancer-only (Fig. 4.3). These numbers are exceptionally high for ovary and skin at 67% and 47% respectively. Taken together, these results underscore the high prevalence of transcript sequence data from cancer tissues in the human transcriptome.

## 4.4 Distribution of Cassette Exons in Normal versus Cancer Clusters

In our analyses described in previous chapters, we used the presence and absence of cassette exons to determine distinct splice isoforms. We investigated the average number of cassette exons in normal versus cancer

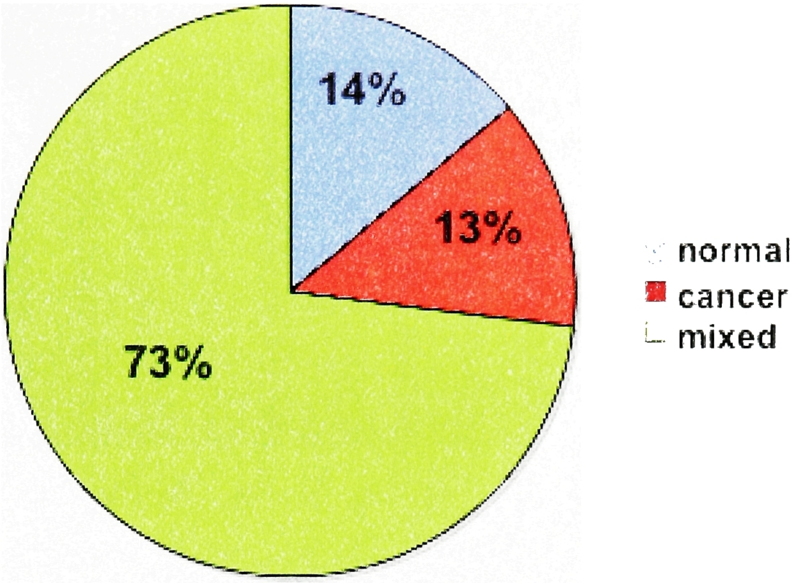


**Figure 4.1** Distribution of normal and cancer transcripts within human variant clusters. 14% of clusters contain transcripts from normal tissues only. 13% of clusters contain transcripts from cancer tissues only. 73% of clusters contain a combination of transcripts both from normal and from cancer tissues.

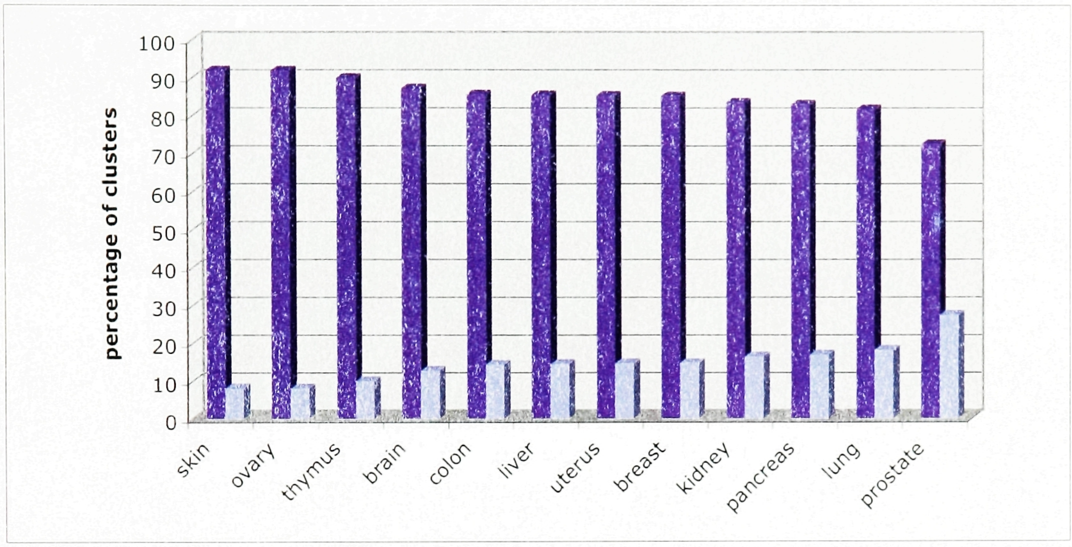


**Figure 4.2** Distribution of human variant clusters across tissues. Purple bars denote sum of cancer-only and mixed clusters i.e. potentially pathological splicing clusters. Blue bars denote normal-only clusters. y-axis: percentage of clusters.

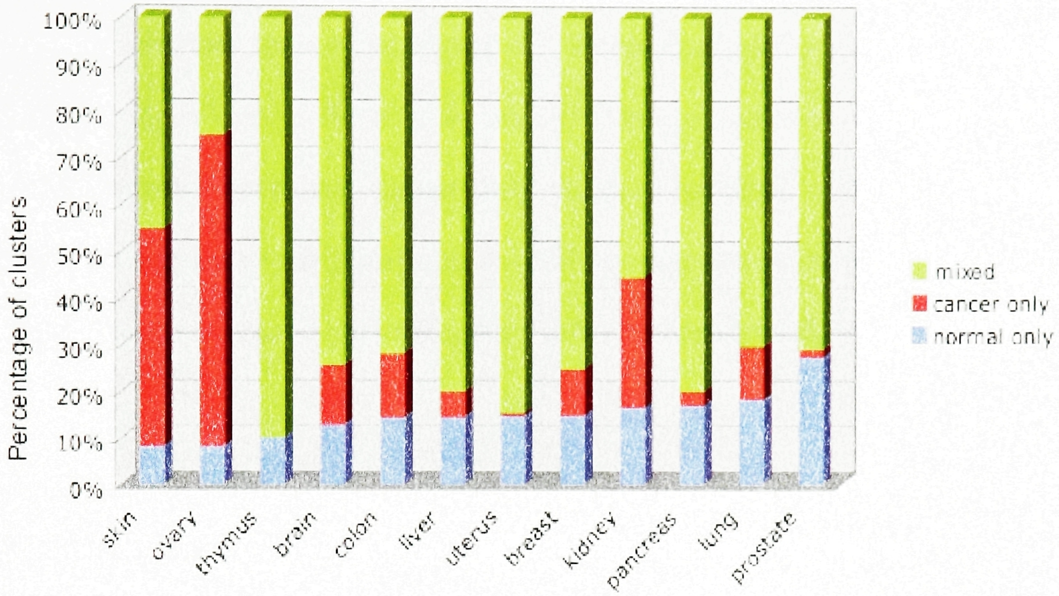


**Figure 4.3** Distribution of human variant clusters with cancer, normal and mixed transcripts across tissues. y-axis: percentage of clusters. Tissues are ordered by sum of cancer only and mixed clusters, in descending order.

transcript-containing variant clusters. With the exception of thymus, the remaining 11 tissues contained transcripts from both *normal-only* and *cancer-only* clusters, as seen in Figure 4.3. When analyzed across these tissues, there is no significant difference in overall average numbers of cassette exons between normal-only and cancer-only clusters (p-value = 0.516). Likewise, there is no significant difference in overall average numbers of length-variant exons between normal-only and cancer-only clusters (p-value = 0.489).

## 4.5 Splicing in Normal versus Cancer Tissue Isoforms

We investigated the cassette exons within HumanSDB3 variant clusters which contain both cancer and normal transcripts, i.e. the mixed clusters as defined in Section 4.2. We wanted to assess the differences between normal and cancer transcripts based on cassette exon splice sites. Our global level cluster analysis revealed that 5.75% of the cassette exons are exclusive to cancer transcripts only and 2.43% of the cassette exons are exclusive to normal transcripts only. The remaining 91.81% of cassette exons are common to both normal and cancer transcripts. Figure 4.4 illustrates introduction of a cancer-specific cassette exon by a cancer transcript.

When we investigated cassette exon distributions within each tissue, we found that a striking number of cassette exons are introduced by cancer transcripts in most of the tissues. Figure 4.5 shows that in 7 out of the 12 tissues studied, more than 20% of all cassette exons are introduced by cancer

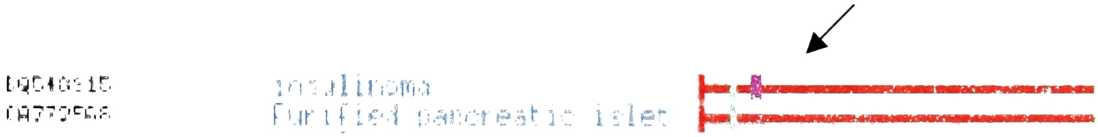


**Figure 4.4** Introduction of a cassette exon by a cancer transcript. Cassette exon shown by the black arrow appears only in cancer transcripts of the variant human cluster Hs.3.chr12p.4960. Normal transcripts do not have this cassette exon. All transcripts of this cluster can be viewed at http://genomes.rockefeller.edu/autodb/cluster_map.php?cluster_id=Hs.3.chr12p.4960&db=HumanSDB3 (also available at http://sgc.ucsd.edu/autodb/cluster_map.php?cluster_id=Hs.3.chr12p.4960&db=HumanSDB3).

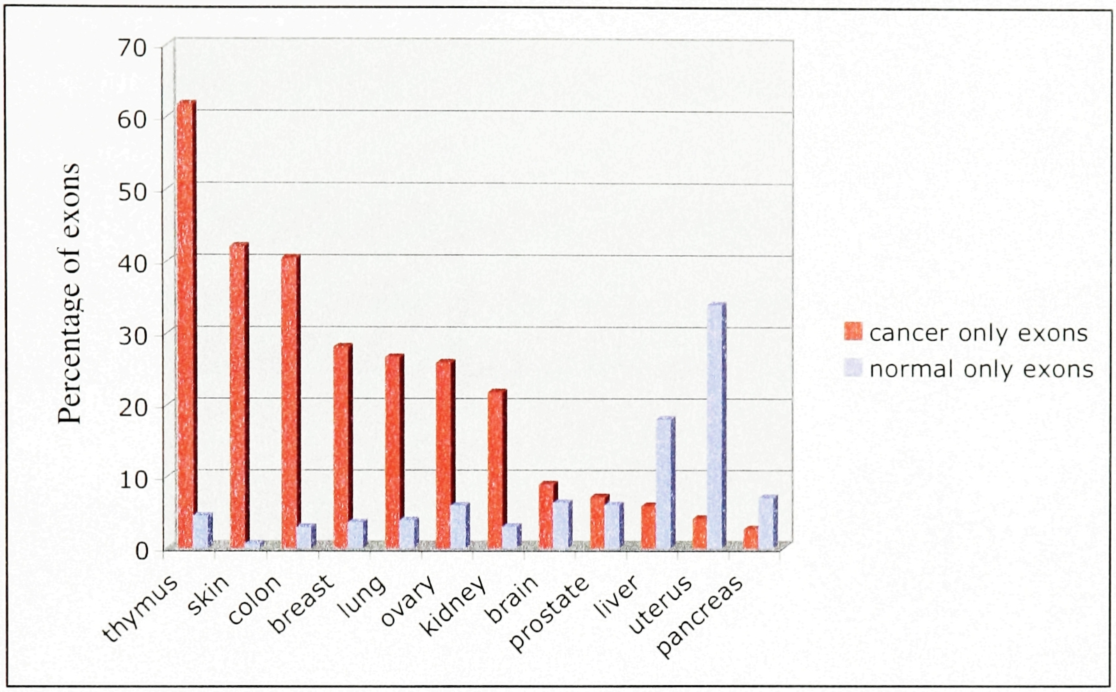


**Figure 4.5** Introduction of cassette exons by cancer transcripts. y-axis: percentage of exons in transcripts of *mixed* clusters. Red bars denote percentage of cassette exons specific to cancer transcripts. Blue bars denote percentage of cassette exons specific to normal transcripts.

transcripts. These numbers are exceptionally high for thymus at 61.8%, for skin at 42.1% and for colon at 40.5%. These tissues are followed by breast, lung, ovary and kidney, which have 28.1%, 26.7%, 25.9% and 21.8% of their cassette exons introduced by cancer transcripts respectively. These findings provide quantitative evidence for introduction of cassette exons by cancer transcripts, which create variation beyond normal alternative splicing and validate presence of aberrant splicing in cancers.

## 4.6 Discussion

Very high numbers of human transcript sequence data heralds from cancer cells and cancer tissues. Since aberrant splicing in cancer tissues has been widely reported and given our finding that cancer transcripts introduce cassette exons beyond normal alternative splicing, it is vital to take into account tissue information about transcripts in studies of alternative splicing.

The vast amount of data from cancer tissues opens up possibilities for comparative studies of normal and pathological splicing, following the methodology we introduced in this chapter. Future studies of cancer versus normal alternative splicing could include investigation of mutually exclusive cassette exons, which are present in normal transcripts one at a time. It would be of interest to analyze whether these exons are present together in cancer transcripts and/or absent from cancer transcripts at the same time. Analysis of functional differences in cancer-specific isoforms compared to normal isoforms will shed light into many cancer related genes and proteins and will open up possibilities for future studies in cancer biology.

# CHAPTER 5

# NEURON-SPECIFIC SPLICING REGULATOR NOVA

## 5.1 Introduction to NOVA

Nova-1 is an RNA binding protein specific to neurons. Nova-1 protein was isolated as a self-antigen in POMA (paraneoplastic opsoclonus myoclonus ataxia), a motor dysfunction autoimmune disorder linked to breast cancer (Buckanovich *et al.*, 1996). Expression of Nova-1 is restricted to the ventral spinal cord and the hindbrain. Nova-2, another member of the Nova protein family, closely associated with Nova-1, is also an RNA binding protein and is expressed at regions of brain where Nova-1 is not expressed including hippocampus and neocortex (Yang *et al.*, 1998).

Nova-1 binds to RNA in a sequence specific manner and regulates alternative splicing (Jensen *et al.*, 2000a; Dredge *et al.*, 2003). KH domain of Nova-1 is necessary for its RNA binding function and recognition of UCAU elements on the RNA (Buckanovich and Darnell, 1997; Jensen *et al.*, 2000b).

Ule *et al.* have developed a novel method called CLIP, which involves UV cross-linking and immunoprecipitation, for identification of novel RNA targets of Nova. They showed that most of the RNA targets identified code for proteins involved in neuronal inhibition and for proteins which function at neuronal synapses (Ule *et al.*, 2003).

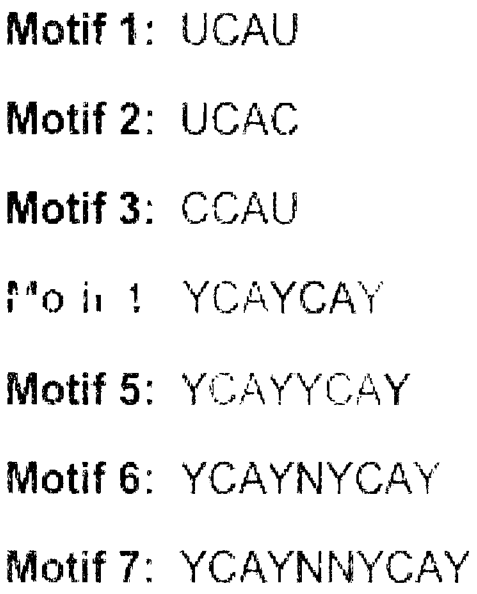
Here we introduce a novel visualization tool that displays Nova binding motifs within target pre-mRNAs for alternative splicing regulation. In addition, we describe a computational method for prediction of novel target RNAs containing alternative exons potentially regulated by Nova.

## 5.2 A Novel Visualization Tool for NOVA Binding Motifs

We developed a novel tool, which enables visualization of binding motifs within sequences of interest. This visualization tool, named TFHunter (short for Transcription Factor Hunter since it was originally developed for TF binding site searches), is a web-based tool where users can access binding site visualizations via a graphical user interface. Partial or full views of pre-mRNAs with distinct exon and intron structures can be browsed for binding motifs. In addition, TFHunter can be made applicable to any target sequences and any binding motifs uploaded by users. Web implementation of this tool can be accessed at http://genomes.rockefeller.edu/yupu-cgi/TFH/main.cgi (Y. Liang private communication).

## 5.3 Application of TFHunter to NOVA Binding Sequences

Using TFHunter we studied the distribution of 7 short sequences which are predicted to be major sequences involved in binding of Nova to its target sequences (J. Ule, R. Darnell private communication). These motifs are illustrated in Figure 5.1.



Motif 1: UCAU

Motif 2: UCAC

Motif 3: CCAU

Motif 4: YCAYCAY

Motif 5: YCAYYCAY

Motif 6: YCAYNYCAY

Motif 7: YCAYNNYCAY

**Figure 5.1** NOVA binding motifs used in TFHunter visualizations. In these sequences, Y stands for U or C. N stands for A, U, C or G.

We studied these sequences in a set of genes which contain alternative exons regulated by Nova. Microarray experiments have verified that expression of each of these exons is either increased or decreased in tissues of Nova knockout mice. Table 5.1 provides the list of genes studied and shows whether their alternative exon is upregulated or downregulated by Nova. Exons which showed an increase in expression are the ones downregulated by Nova. Expected binding sites for Nova are either within and/or upstream of this group of alternative exons. Exons which showed a decrease in expression are the ones upregulated by Nova. Expected binding sites for Nova are downstream to this group of alternative exons (J. Ule, R. Darnell private communication).

Figure 5.2 shows neogenin exon 27, an experimentally identified alternative exon shown to be downregulated by Nova (Ule *et al.*, 2003). This figure illustrates the application of TFHunter to the visualization of Nova binding motifs. Colored boxes denote Nova binding motifs described in Figure 5.1. Numbers on the right illustrate the number of occurrences of each motif within the displayed subsequence of the gene. Numbers in parentheses indicate the number of occurrence of each motif within the entire gene. The clustered presence of motifs 5 (YCAYYCAY), 6 (YCAYNYCAY) and 7 (YCAYNNYCAY) denotes potentially strong binding sites for Nova. Similar to neogenin, for Chl1, Rap1 and Sk1IP genes, TFHunter detected Nova binding motifs within the alternative exon and/or upstream to the alternative exon. Figures 5.3 and 5.4 show the binding motif clusters in Chl1 and Rap1 respectively.

**Table 5.1** Genes with alternative exons regulated by NOVA

| Gene | Alternative Exon Regulation by Nova |
|---|---|
| Neogenin | Downregulated |
| NMDA – exon 5 | Downregulated |
| Chl1 | Downregulated |
| Rap1 | Downregulated |
| Sk1IP | Downregulated |
| NMDA – exon 21 | Upregulated |
| APLP2 | Upregulated |
| Ank3 | Upregulated |
| Cask | Upregulated |
| Dbs | Upregulated |
| Epha5 | Upregulated |
| Fodrin | Upregulated |
| Lar | Upregulated |
| Necl1 | Upregulated |
| PMCA1 | Upregulated |
| Calsyntenin | Upregulated |

(Data configured via private communication with J. Ule and R. Darnell).

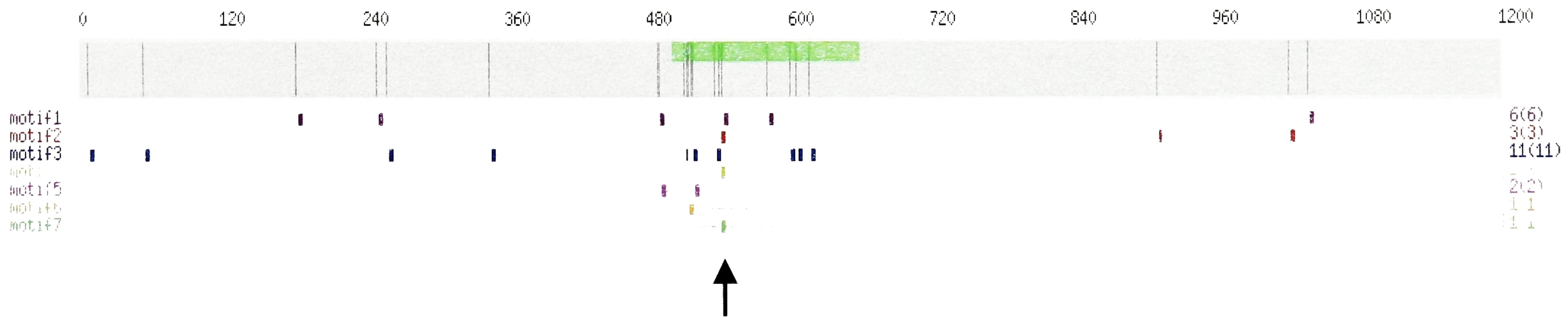


**Figure 5.2** NOVA binding sites within exon 27 of Neogenin. Green box denotes alternative exon 27. Gray areas flanking the alternative exon denote intronic sequences. Small colored boxes denote NOVA binding motifs. (Please see Fig. 5.1 for motif sequences). Numbers on the right denote how many times each motif occurs within the viewed subsequence and numbers in parentheses show how many times each motif is present within the entire gene. Cluster of NOVA binding motifs within the alternative exon is indicated by the arrow.

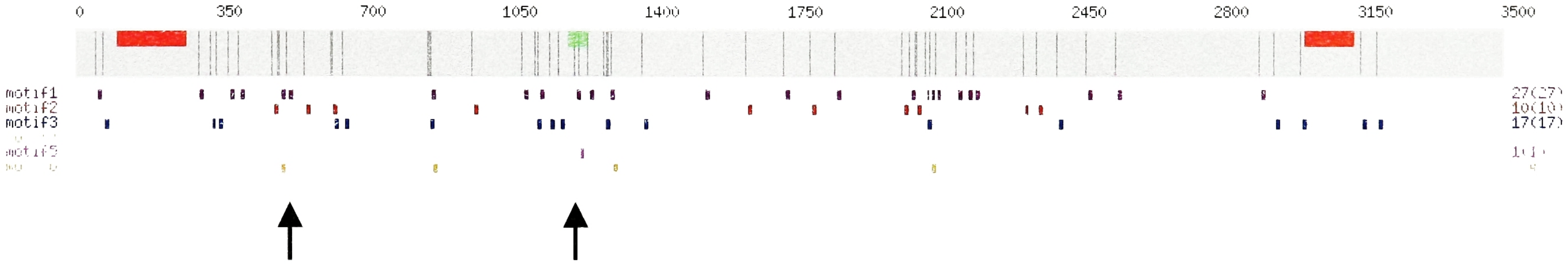


**Figure 5.3** NOVA binding sites for alternative exon of Chl1 gene. Green box denotes the alternative exon. Red boxes denote constitutive exons. Gray areas flanking the exons denote intronic sequences. Small colored boxes denote NOVA binding motifs. (Please see Fig. 5.1 for motif sequences). Numbers on the right denote how many times each motif occurs within the viewed subsequence and numbers in parentheses show how many times each motif is present within the entire gene. Clusters of NOVA binding motifs within and upstream of the alternative exon are indicated by the arrows.

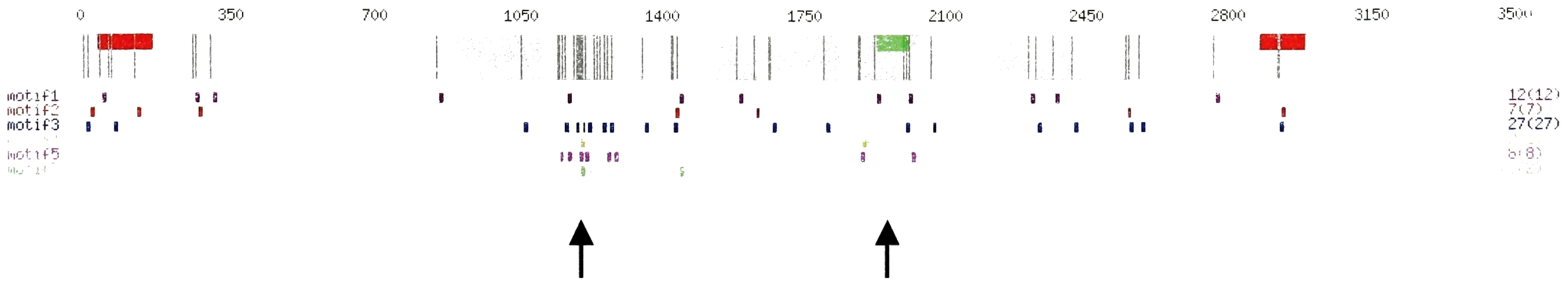


**Figure 5.4** NOVA binding sites for alternative exon of Rap1 gene. Green box denotes the alternative exon. Red boxes denote constitutive exons. Gray areas flanking the exons denote intronic sequences. Small colored boxes denote NOVA binding motifs. (Please see Fig. 5.1 for motif sequences). Numbers on the right denote how many times each motif occurs within the viewed subsequence and numbers in parentheses show how many times each motif is present within the entire gene. Clusters of NOVA binding motifs within and upstream of the alternative exon are indicated by the arrows.

Figure 5.5 shows the TFHunter illustration of Necl1 gene. As expected, Nova-binding motifs cluster downstream of the alternative exon. Similar to Necl1, for APLP2, Cask, Dbs, Epha5, Fodrin, LAR, PMCA1 and Calsyntenin genes, TFHunter detected Nova binding motifs downstream of the alternative exons. Figures 5.6, 5.7, 5.8 and 5.9 show clusters of Nova binding motifs within APLP2, Epha5, LAR and Calsyntenin respectively. TFHunter development is described in Materials and Methods, Section 7.10. All of the above findings of Nova binding motif localizations with respect to alternative exons are in accordance with our predictions and clusters of long motifs in and around cassette exons validate importance of YCAY elements in Nova binding to the target RNA molecules (Jensen *et al.*, 2000b).

## 5.4 Target Prediction for NOVA

Here we introduce a computational target prediction method for Nova using our alternative splicing databases. Using variant clusters in HumanSDB3, MouSDB5 and RatSDB2, we predicted a set of candidate targets which contain alternative exons splicing of which is potentially regulated by Nova. Our prediction program, called NTHunter (Nova Target Hunter), selects a master RNA from each variant cluster. Master RNAs are the transcripts with the highest number of exons within a given cluster. From the master RNA, NTHunter extracts exon sequences with their 500 base-pair upstream and 500 base-pair downstream intronic sequences.

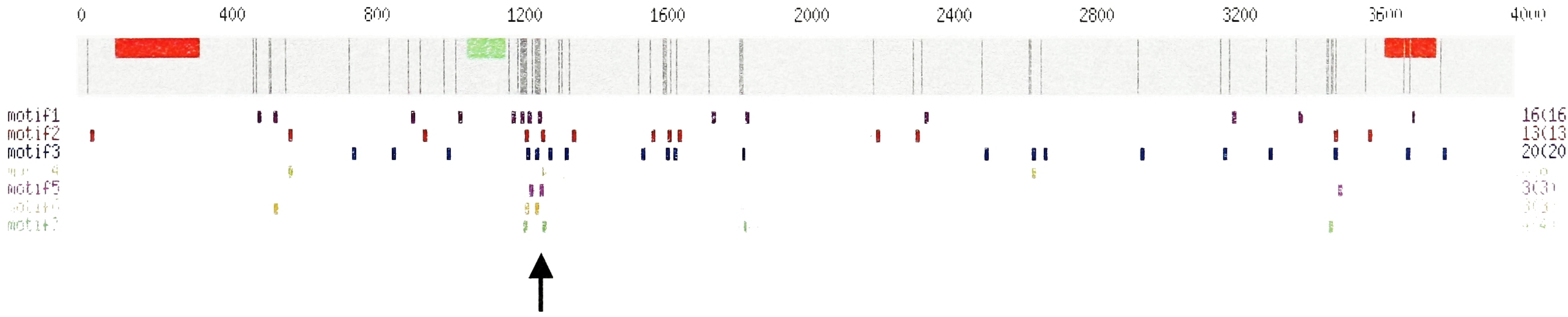


**Figure 5.5** TFHunter illustration of Necl1 gene. Green box denotes the alternative exon. Red boxes denote flanking constitutive exons. Gray areas flanking the exons denote intronic sequences. Small colored boxes denote NOVA binding motifs. (Please see Fig. 5.1 for motif sequences). Numbers on the right denote how many times each motif occurs within the viewed subsequence and numbers in parentheses show how many times each motif is present within the entire gene. Cluster of NOVA binding motifs downstream of the alternative exon is indicated by the arrow.

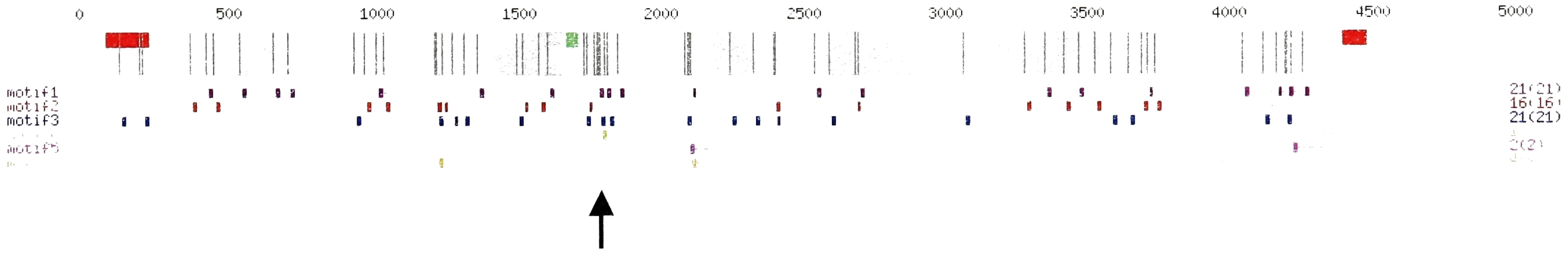


**Figure 5.6** TFHunter illustration of APLP2 gene. Green box denotes the alternative exon. Red boxes denote flanking constitutive exons. Gray areas flanking the exons denote intronic sequences. Small colored boxes denote NOVA binding motifs. (Please see Fig. 5.1 for motif sequences). Numbers on the right denote how many times each motif occurs within the viewed subsequence and numbers in parentheses show how many times each motif is present within the entire gene. Cluster of NOVA binding motifs downstream of the alternative exon is indicated by the arrow.

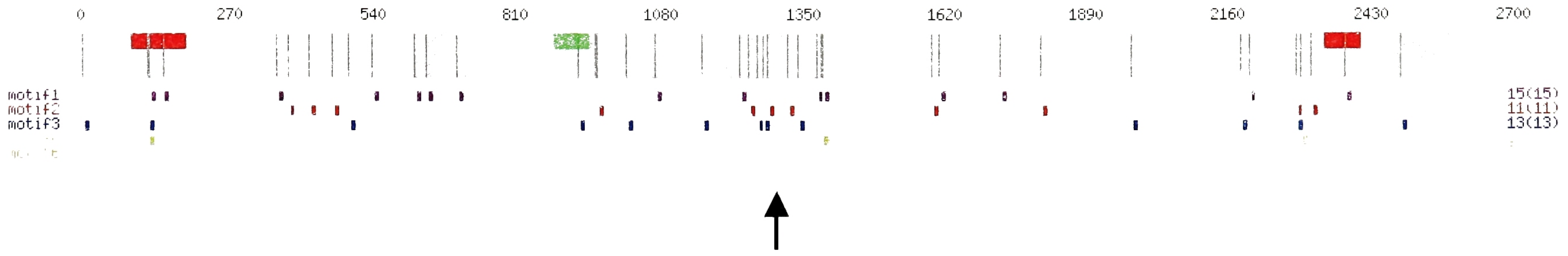


**Figure 5.7** TFHunter illustration of Epha5 gene. Green box denotes the alternative exon. Red boxes denote flanking constitutive exons. Gray areas flanking the exons denote intronic sequences. Small colored boxes denote NOVA binding motifs. (Please see Fig. 5.1 for motif sequences). Numbers on the right denote how many times each motif occurs within the viewed subsequence and numbers in parentheses show how many times each motif is present within the entire gene. Cluster of NOVA binding motifs downstream of the alternative exon is indicated by the arrow.

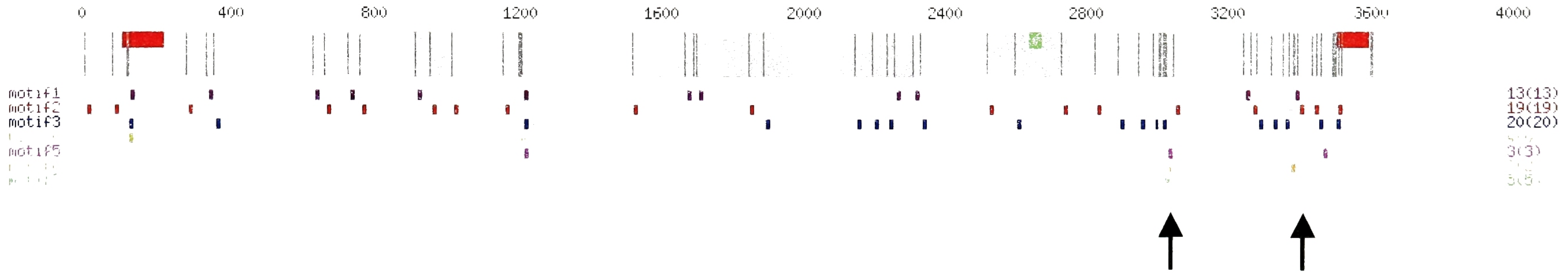


**Figure 5.8** TFHunter illustration of LAR gene. Green box denotes the alternative exon. Red boxes denote flanking constitutive exons. Gray areas flanking the exons denote intronic sequences. Small colored boxes denote NOVA binding motifs. (Please see Fig. 5.1 for motif sequences). Numbers on the right denote how many times each motif occurs within the viewed subsequence and numbers in parentheses show how many times each motif is present within the entire gene. Clusters of NOVA binding motifs downstream of the alternative exon are indicated by the arrows.

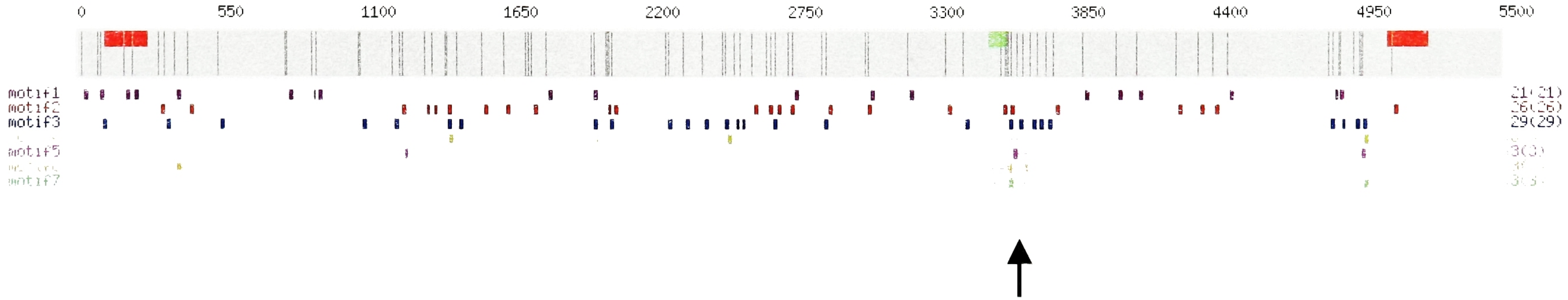
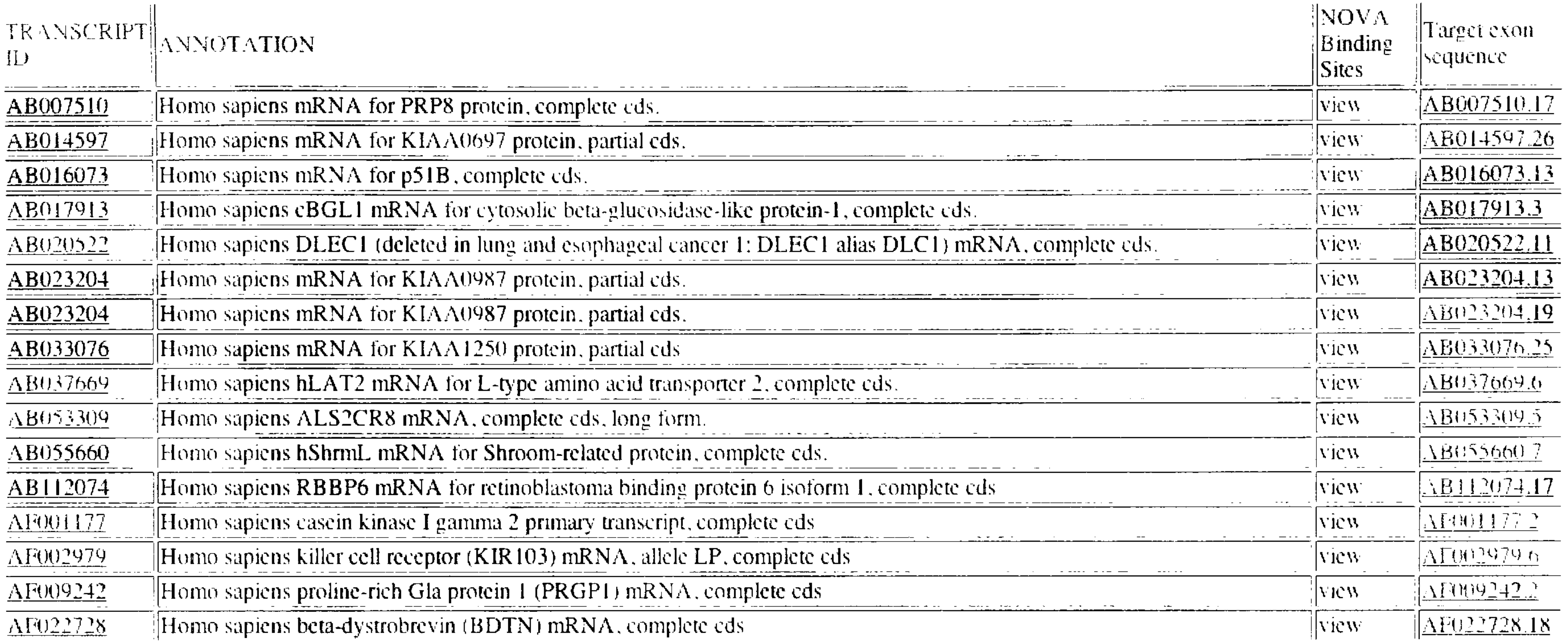


**Figure 5.9** TFHunter illustration of Calsyntenin gene. Green box denotes the alternative exon. Red boxes denote flanking constitutive exons. Gray areas flanking the exons denote intronic sequences. Small colored boxes denote NOVA binding motifs. (Please see Fig. 5.1 for motif sequences). Numbers on the right denote how many times each motif occurs within the viewed subsequence and numbers in parentheses show how many times each motif is present within the entire gene. Cluster of NOVA binding motifs downstream of the alternative exon is indicated by the arrow.

Exons are labeled with respect to their variation types. NTHunter divides the exon into 100 base-pair regions and searches for Nova binding motif clusters within these regions using a sliding window. The window size is the number of nucleotides in a given motif and it is indexed at 1 base-pair. NTHunter searches for the presence of at least four occurrences of motifs 1, 2 and/or 3 and at least two occurrences of motifs 4, 5, 6 and/or 7. (Motif sequences are shown in Fig. 5.1). If any cassette exon contains a combination of these sequences, the transcript containing the given exon is flagged as a candidate gene for Nova regulation. Using the above method, we predicted 471 human, 136 mouse and 36 rat genes to have one or more alternative exons regulated by Nova. Based on the findings discussed in Section 5.3, we predict these exons to be downregulated by Nova since these genes contain clusters of Nova binding sites within their alternative exons. Figure 5.10 provides a partial list of human candidate Nova target genes. The complete list of human target genes along with mouse and rat target gene lists can be accessed at http://genomes.rockefeller.edu/~bahar/NTHunter/NTHunter.html (also available at http://sgc.ucsd.edu/~bahar/NTHunter/NTHunter.html). This web-page provides, GenBank links of the target genes, their annotations and sequences of the potential target cassette exons. Appendices B, C and D show the complete lists of human candidate target genes, mouse candidate target genes and rat candidate target genes respectively. NTHunter algorithm is described in Materials and Methods, Section 7.11.



| TRANSCRIPT ID | ANNOTATION | NOVA Binding Sites | Target exon sequence |
|---|---|---|---|
| AB007510 | Homo sapiens mRNA for PRP8 protein, complete cds. | view | AB007510.17 |
| AB014597 | Homo sapiens mRNA for KIAA0697 protein, partial cds. | view | AB014597.26 |
| AB016073 | Homo sapiens mRNA for p51B, complete cds. | view | AB016073.13 |
| AB017913 | Homo sapiens cBGL1 mRNA for cytosolic beta-glucosidase-like protein-1, complete cds. | view | AB017913.3 |
| AB020522 | Homo sapiens DLEC1 (deleted in lung and esophageal cancer 1: DLEC1 alias DLC1) mRNA, complete cds. | view | AB020522.11 |
| AB023204 | Homo sapiens mRNA for KIAA0987 protein, partial cds. | view | AB023204.13 |
| AB023204 | Homo sapiens mRNA for KIAA0987 protein, partial cds. | view | AB023204.19 |
| AB033076 | Homo sapiens mRNA for KIAA1250 protein, partial cds | view | AB033076.25 |
| AB037669 | Homo sapiens hLAT2 mRNA for L-type amino acid transporter 2, complete cds. | view | AB037669.6 |
| AB053309 | Homo sapiens ALS2CR8 mRNA, complete cds, long form. | view | AB053309.5 |
| AB055660 | Homo sapiens hShrmL mRNA for Shroom-related protein, complete cds. | view | AB055660.7 |
| AB112074 | Homo sapiens RBBP6 mRNA for retinoblastoma binding protein 6 isoform 1, complete cds | view | AB112074.17 |
| AF001177 | Homo sapiens casein kinase I gamma 2 primary transcript, complete cds | view | AF001177.2 |
| AF002979 | Homo sapiens killer cell receptor (KIR103) mRNA, allele LP, complete cds | view | AF002979.6 |
| AF009242 | Homo sapiens proline-rich Gla protein 1 (PRGP1) mRNA, complete cds | view | AF009242.2 |
| AF022728 | Homo sapiens beta-dystrobrevin (BDTN) mRNA, complete cds | view | AF022728.18 |

**Figure 5.10** A partial list of 471 human candidate genes predicted by NTHunter. Leftmost column provides GenBank ids. Rightmost column provides links to NTHunter target exon sequences. The complete list can be accessed at http://genomes.rockefeller.edu/~bahar/NTHunter/NTHunter.html (also at http://sgc.ucsd.edu/~bahar/NTHunter/NTHunter.html).

Even though known targets for NOVA such as Neogenin, KCNQ and Ankyrin (Appendix B) are predicted by NTHunter, when compared to constitutive exon background sequence sets randomly selected from HumanSDB3, there seems to be no statistical difference between the numbers of cassette and constitutive exons predicted by NTHunter. Figure 5.11 shows this comparison. Based on this finding and availability of future experimental data on alternative splicing regulation by Nova, NTHunter algorithm will be modified requiring more stringent clusters of Nova-binding motifs both in terms of motif occurrence and in terms of the relative positions of motifs to each other within alternative exons and within their flanking intronic sequences, as further described in Chapter 8.

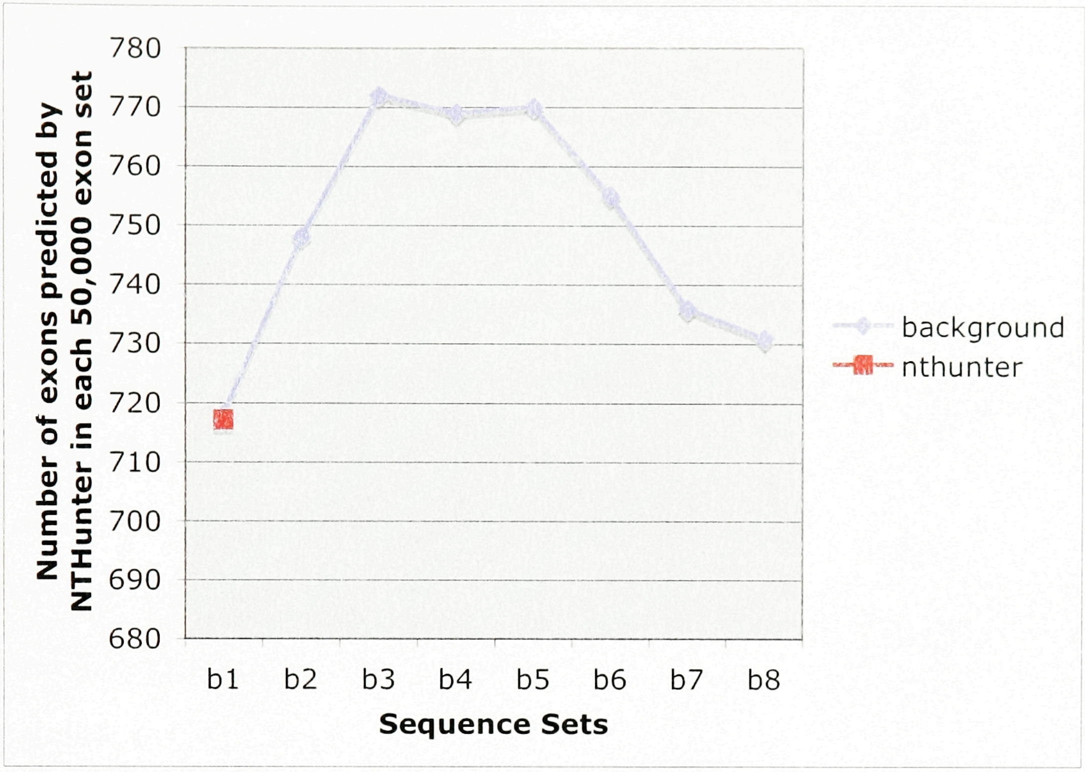


**Figure 5.11** Eight different background sets of human constitutive exons were generated (b1-b8), each of which contained 50,000 exons and had similar length distributions of around 136 nucleotides. These background sets were compared to a cassette exon set of 50,000 human exons (NTHunter set), of a similar length distribution to those of the background exons (around 115 nucleotides). Number of exons predicted by NTHunter within each of these sequence sets are as follows. Cassette exon set (NTHunter set): 717, background set-1: 718, background set-2: 748, background set-3: 772, background set-4: 769, background set-5: 770, background set-6: 755, background set-7: 736 and background set-8: 731.

## 5.5 Discussion

We developed a novel visualization method for sequences specific to Nova binding to pre-mRNA. With this method, we enable detailed analysis of Nova binding sites within target genes. In addition, we computationally predicted human, mouse and rat candidate genes with alternative exons potentially regulated by Nova. These targets are predicted on the basis of the fact that they met a certain threshold for presence of clustered Nova binding motifs within their alternative exons. It is of future interest to rank the candidate genes by establishing motif density in reference to alternative exon position and by taking into account the functional annotation of genes. Validation of target genes by experimental methods will lead to understanding further the mechanism of regulation of alternative splicing by Nova. In addition, our results enable future studies of orthologous Nova targets, which will further add to the knowledge on how Nova functions. For this purpose, studies of orthologous human-mouse, human-rat, mouse-rat gene-pairs and human-mouse-rat gene-triplets for localization of Nova binding motifs will be of interest.

# CHAPTER 6

# ALTERNATIVE SPLICING BEYOND MAMMALIAN TRANSCRIPTOMES

## 6.1 Alternative Splicing and Model Organisms

Alternative splicing has been widely studied in model organisms especially in *Drosophila melanogaster* as discussed in Section 6.2. Since alternative splicing is conserved across species (Brett *et al.*, 2002; Valenzuela *et al.*, 2004), it is of importance to study this process in species beyond human, mouse and rat, in order to understand its evolution and functional significance. To this extent, we developed databases of alternatively spliced forms in transcriptomes of *Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana* and *Plasmodium falciparum.* These databases are developed using the AutoDB pipeline described in Chapters 2 and 7.

Beyond providing useful information on alternative splicing in these organisms, these databases illustrate generality of our approach. AutoDB can be applied to genome and transcript sequence data beyond human, mouse and rat.

## 6.2 *Drosophila melanogaster* Alternative Splicing Database

*Drosophila melanogaster* is one of the major model organisms used to study alternative splicing (Graveley, 2002). Dscam (Down syndrome cell adhesion molecule) gene of *Drosophila melanogaster* contains 95 alternative exons and exhibits striking amounts of alternative splicing with more than 38,000 possible different isoforms. Dscam codes for an axon guidance receptor and has been widely studied (Celotto and Graveley, 2001; Graveley *et al.*, 2004).

We developed a database of alternatively spliced transcripts in *Drosophila melanogaster* transcriptome, termed DmelSDB5, using the AutoDB pipeline described in Chapter 7. Interactive web implementation of this database can be accessed at http://sgc.ucsd.edu/autodb/search_clusters.php?db=DmelSDB5. Tables 6.1 and 6.2 show numbers of input transcripts, mapped transcripts and transcript rejection rates by AutoDB for DmelSDB5. Table 6.3 provides a statistical overview of clusters in this database. Table 6.4 shows alternative exon distributions in DmelSDB5.

**Table 6.1** Mapped *D. melanogaster* transcripts

| | |
|---|---|
| Total number of input transcripts | 418,023 |
| Total number of mapped transcripts | 220,459 |
| % mapped transcripts | 53% |
| Average number of transcripts per cluster | 39.16 |

**Table 6.2** AutoDB rejection rates for *D. melanogaster* input transcripts

| | Total numbers | Percentages |
|---|---|---|
| Blat | 3365 | 1.7% |
| SIM4 | 0 | 0% |
| Transcript + Exon Quality | 189,545 | 96% |
| Cluster Quality | 4,654 | 2.3% |
| Total | 197,564 | |

**Table 6.3** Clusters of DmelSDB5

| | Clusters | Percentage |
|---|---|---|
| Variant | 3018 | 35% |
| Invariant | 5647 | 65% |
| Total | 8665 | |

**Table 6.4** Exons of DmelSDB5

| | Exon Numbers |
|---|---|
| Total number of exons | 47,403 |
| Constitutive exons | 40,257 |
| Alternative exons | 7,146 |
| | |
| Total number of internal cassette exons | 1,972 |
| Length-invariant internal cassette exons | 1,725 |
| 5' length-variant internal cassette exons | 149 |
| 3' length-variant internal cassette exons | 83 |
| 5' and 3' length-variant internal cassette exons | 15 |
| | |
| Total number of terminal cassette exons | 2,225 |
| Length-invariant terminal cassette exons | 1,990 |
| 5' length-variant terminal cassette exons | 25 |
| 3' length-variant terminal cassette exons | 210 |
| 5' and 3' length-variant terminal cassette exons | 0 |
| | |
| Total number of length-variant exons | 2,949 |
| 5' length-variant exons | 1,414 |
| 3' length-variant exons | 1,484 |
| 5' and 3' length-variant exons | 51 |

## 6.3 *Caenorhabditis elegans* Alternative Splicing Database

*C. elegans is* one of the major model organisms which displays alternative splicing (Zhuang *et al.*, 2003). Using AutoDB, we developed a database of alternatively spliced transcripts in *C. elegans* transcriptome, termed CeleganSDB5, which can be accessed at http://sgc.ucsd.edu/autodb/search_clusters.php?db=CeleganSDB5. Tables 6.5 and 6.6 show numbers of input transcripts, mapped transcripts and transcript rejection rates by AutoDB for CeleganSDB5. Table 6.7 provides a statistical overview of clusters in this database. Table 6.8 provides alternative exon distributions in CeleganSDB5.

## 6.4 *Arabidopsis thaliana* Alternative Splicing Database

Available sequence data for *Arabidopsis thaliana* enables studies of alternative splicing in this plant (Iida *et al.*, 2004). Using the AutoDB pipeline described in Chapter 7, we developed a database of alternatively spliced transcripts in *A. thaliana* transcriptome, termed AthalSDB3. Interactive web implementation of this database can be accessed at http://sgc.ucsd.edu/autodb/search_clusters.php?db=AthalSDB3. Tables 6.9 and 6.10 show numbers of input transcripts, mapped transcripts and transcript rejection rates by AutoDB for AthalSDB3. Table 6.11 provides a statistical overview of clusters in this database. Table 6.12 shows exon distributions in AthalSDB3.

**Table 6.5** Mapped *C. elegans* transcripts

| | |
|---|---|
| Total number of input transcripts | 329,181 |
| Total number of mapped transcripts | 122,274 |
| % mapped transcripts | 37% |
| Average number of transcripts per cluster | 24.21 |

**Table 6.6** AutoDB rejection rates for *C. elegans* input transcripts

| | Total numbers | Percentages |
|---|---|---|
| Blat | 3,598 | 1.7% |
| SIM4 | 0 | 0% |
| Transcript + Exon Quality | 175,599 | 85% |
| Cluster Quality | 27,710 | 13.3% |
| Total | 206,907 | |

**Table 6.7** Clusters of CeleganSDB5

| | Clusters | Percentage |
|---|---|---|
| Variant | 2211 | 23% |
| Invariant | 7490 | 77% |
| Total | 9701 | |

**Table 6.8** Exons of CeleganSDB5

| | Exon Numbers |
|---|---|
| Total number of exons | 72,984 |
| Constitutive exons | 68,136 |
| Alternative exons | 4,848 |
| | |
| Total number of internal cassette exons | 1,868 |
| Length-invariant internal cassette exons | 1,762 |
| 5' length-variant internal cassette exons | 43 |
| 3' length-variant internal cassette exons | 52 |
| 5' and 3' length-variant internal cassette exons | 11 |
| | |
| Total number of terminal cassette exons | 851 |
| Length-invariant terminal cassette exons | 835 |
| 5' length-variant terminal cassette exons | 4 |
| 3' length-variant terminal cassette exons | 12 |
| 5' and 3' length-variant terminal cassette exons | 0 |
| | |
| Total number of length-variant exons | 2129 |
| 5' length-variant exons | 1031 |
| 3' length-variant exons | 1052 |
| 5' and 3' length-variant exons | 46 |

**Table 6.9** Mapped *A. thaliana* transcripts

| | |
|---|---|
| Total number of input transcripts | 510,437 |
| Total number of mapped transcripts | 156,880 |
| % mapped transcripts | 31% |
| Average number of transcripts per cluster | 13.97 |

**Table 6.10** AutoDB rejection rates for *A. thaliana* input transcripts

| | Total numbers | Percentages |
|---|---|---|
| Blat | 12,744 | 3.6% |
| SIM4 | 0 | 0 |
| Transcript + Exon Quality | 331,454 | 93.75% |
| Cluster Quality | 9,359 | 2.65% |
| Total | 353,557 | |

**Table 6.11** Clusters of AthalSDB3

| | Clusters | Percentage |
|---|---|---|
| Variant | 3,022 | 21% |
| Invariant | 11,244 | 79% |
| Total | 14,266 | |

**Table 6.12** Exons of AthalSDB3

| | Exon Numbers |
|---|---|
| Total number of exons | 98,308 |
| Constitutive exons | 93,185 |
| Alternative exons | 5,123 |
| Total number of internal cassette exons | 1,082 |
| Length-invariant internal cassette exons | 1,026 |
| 5' length-variant internal cassette exons | 35 |
| 3' length-variant internal cassette exons | 19 |
| 5' and 3' length-variant internal cassette exons | 2 |
| Total number of terminal cassette exons | 380 |
| Length-invariant terminal cassette exons | 364 |
| 5' length-variant terminal cassette exons | 4 |
| 3' length-variant terminal cassette exons | 12 |
| 5' and 3' length-variant terminal cassette exons | 0 |
| Total number of length-variant exons | 3661 |
| 5' length-variant exons | 2185 |
| 3' length-variant exons | 1411 |
| 5' and 3' length-variant exons | 65 |

## 6.5 *Plasmodium falciparum* Alternative Splicing Database

Given that *Plasmodium falciparum* is a major pathogen causing malaria, and that post-transcriptional regulation including alternative splicing has been documented in its transcripts (Singh *et al.*, 2004), we developed a database of splice forms in the transcriptome of this organism.

Database of alternatively spliced transcripts in *P falciparum* transcriptome, termed PfalSDB2, is developed by using the AutoDB pipeline described in Chapter 7. Interactive web implementation of this database can be accessed at http://sgc.ucsd.edu/autodb/search_clusters.php?db=PfalSDB2. Tables 6.13 and 6.14 show numbers of input transcripts, mapped transcripts and transcript rejection rates by AutoDB for PfalSDB2. Table 6.15 provides a statistical overview of clusters in this database. Table 6.16 provides exon distributions in PfalSDB2.

**Table 6.13** Mapped *P. falciparum* transcripts

| | |
|---|---|
| Total number of input transcripts | 22,413 |
| Total number of mapped transcripts | 2,848 |
| % mapped transcripts | 13% |
| Average number of transcripts per cluster | 11.85 |

**Table 6.14** AutoDB rejection rates for *P. falciparum* input transcripts

| | Total numbers | Percentages |
|---|---|---|
| Blat | 579 | 3% |
| SIM4 | 0 | 0 |
| Transcript + Exon Quality | 17,729 | 91% |
| Cluster Quality | 1257 | 6% |
| Total | | |

**Table 6.15** Clusters of PfalSDB2

| | Clusters | Percentage |
|---|---|---|
| Variant | 46 | 11% |
| Invariant | 372 | 89% |
| Total | 418 | |

**Table 6.16** Exons of PfalSDB2

| | Exon Numbers |
|---|---|
| Total number of exons | 1729 |
| Constitutive exons | 1644 |
| Alternative exons | 85 |
| | |
| Total number of internal cassette exons | 22 |
| Length-invariant internal cassette exons | 21 |
| 5' length-variant internal cassette exons | 0 |
| 3' length-variant internal cassette exons | 1 |
| 5' and 3' length-variant internal cassette exons | 0 |
| | |
| Total number of terminal cassette exons | 15 |
| Length-invariant terminal cassette exons | 14 |
| 5' length-variant terminal cassette exons | 1 |
| 3' length-variant terminal cassette exons | 0 |
| 5' and 3' length-variant terminal cassette exons | 0 |
| | |
| Total number of length-variant exons | 48 |
| 5' length-variant exons | 22 |
| 3' length-variant exons | 26 |
| 5' and 3' length-variant exons | 0 |

## 6.6 Results

Figure 6.1a shows overall presence of alternative splicing in *Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana* and *Plasmodium falciparum.* In *Drosophila* 35% of clusters are variant, in *C. elegans* 23% of clusters are variant and in *Arabidopsis* and *P. falciparum,* 21% and 11% clusters are variant respectively.

Figure 6.1b shows that variation levels detected are positively correlated with the average number of transcripts per cluster. This result reveals that the extent of alternative splicing detected in genomes of different organisms is dependent on the availability of expressed sequence data for those organisms.

When we compare all seven species for which we have developed alternative splicing databases, human has the highest prevalence of alternative splicing followed by mouse and rat. *Drosophila, C. elegans, Arabidopsis* and *P. falciparum* display lower levels of variation in descending order (Fig. 6.2a). Figure 6.2b shows that there is a positive correlation between detection of variation and the number of input transcripts in all seven species. However, the fact that rat and *Drosophila* have similar numbers of input transcripts and the finding that there is almost 2-fold higher alternative splicing in the rat transcriptome, suggest that more complex organisms might have more alternative splicing. A similar argument can be made for *Drosophila* and *C. elegans* as well (Fig. 6.2b).

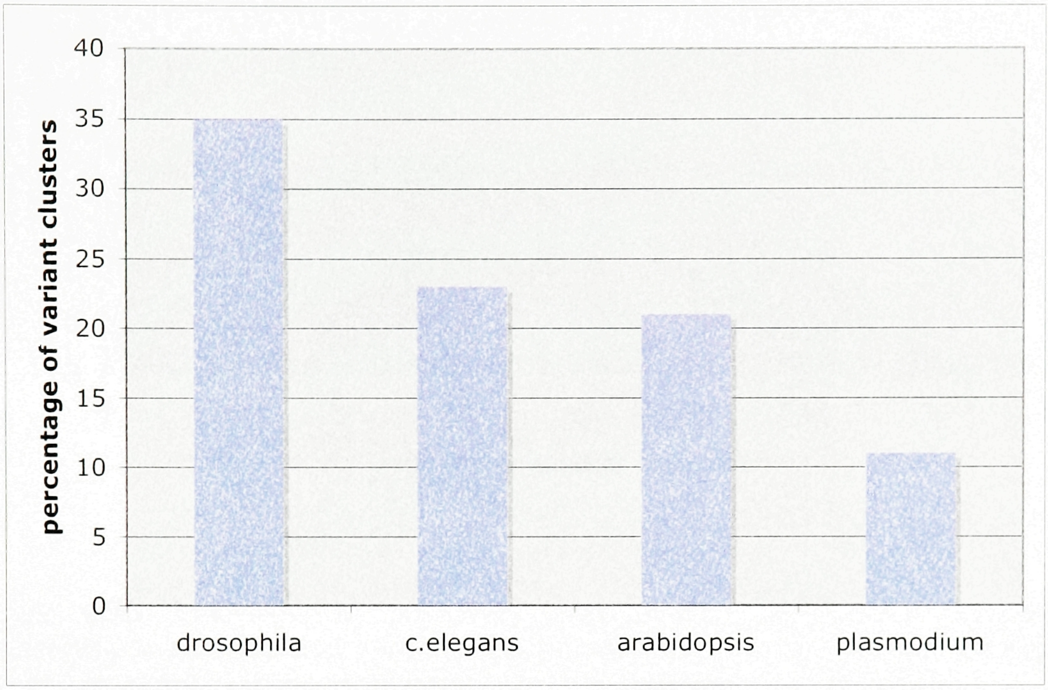


**Figure 6.1a** Variation in *Drosophila, C. elegans, Arabidopsis* and *P. falciparum* genomes. Bar values are percent variant clusters.

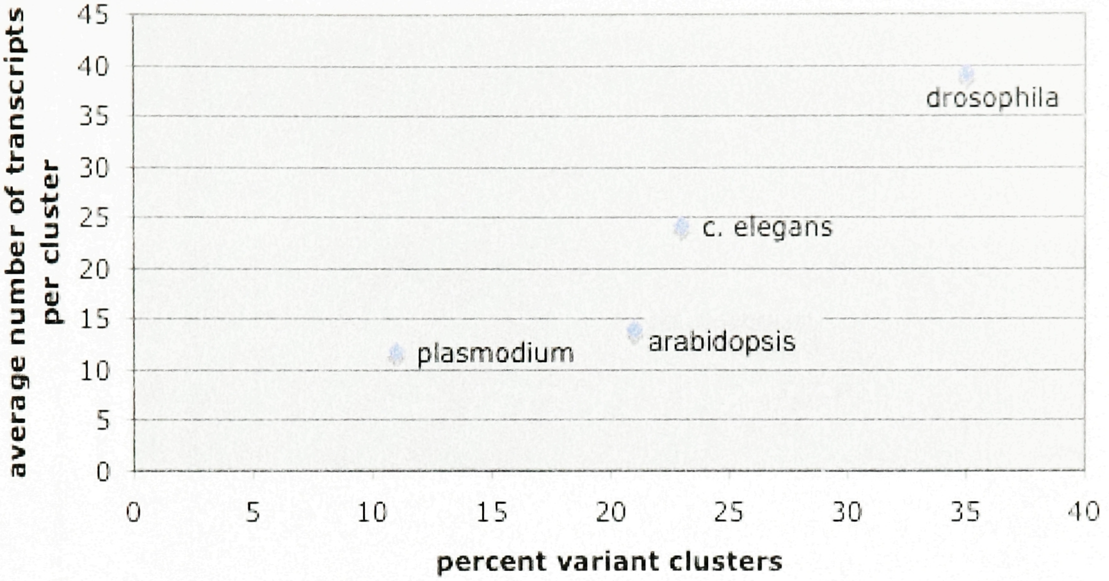


**Figure 6.1b** Variation is positively correlated with percent mapped transcripts. x-axis: percent variant clusters in each genome. y-axis: average number of mapped transcripts per cluster for corresponding genomes.

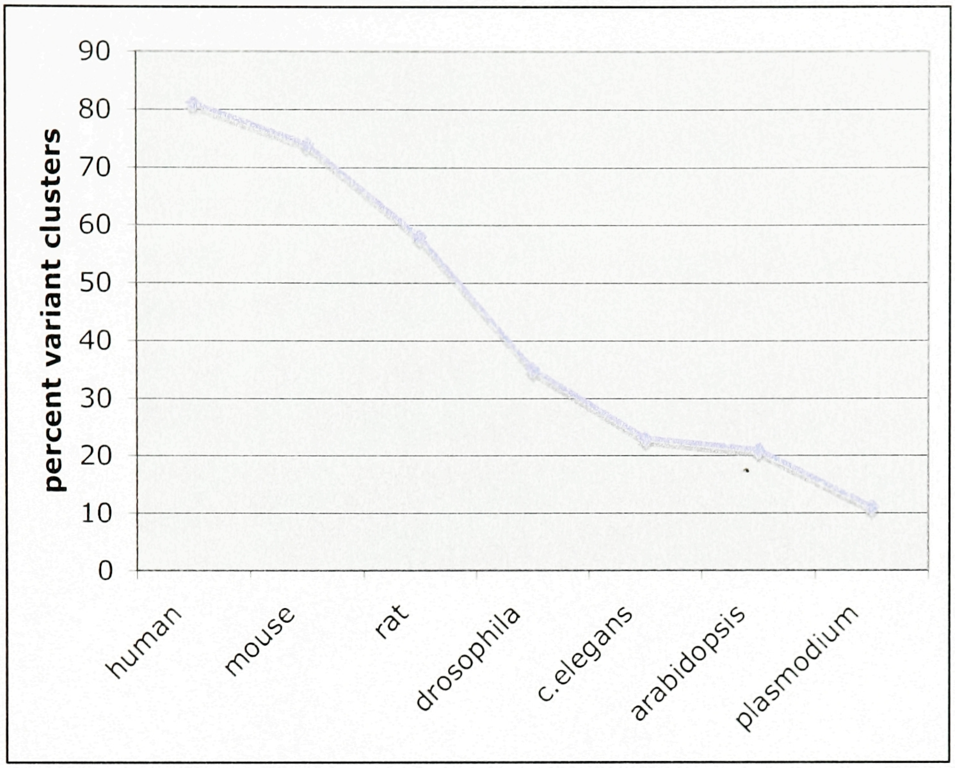


**Figure 6.2a** Alternative splicing across seven different species. Data-points show percentage of variant clusters in each organism.

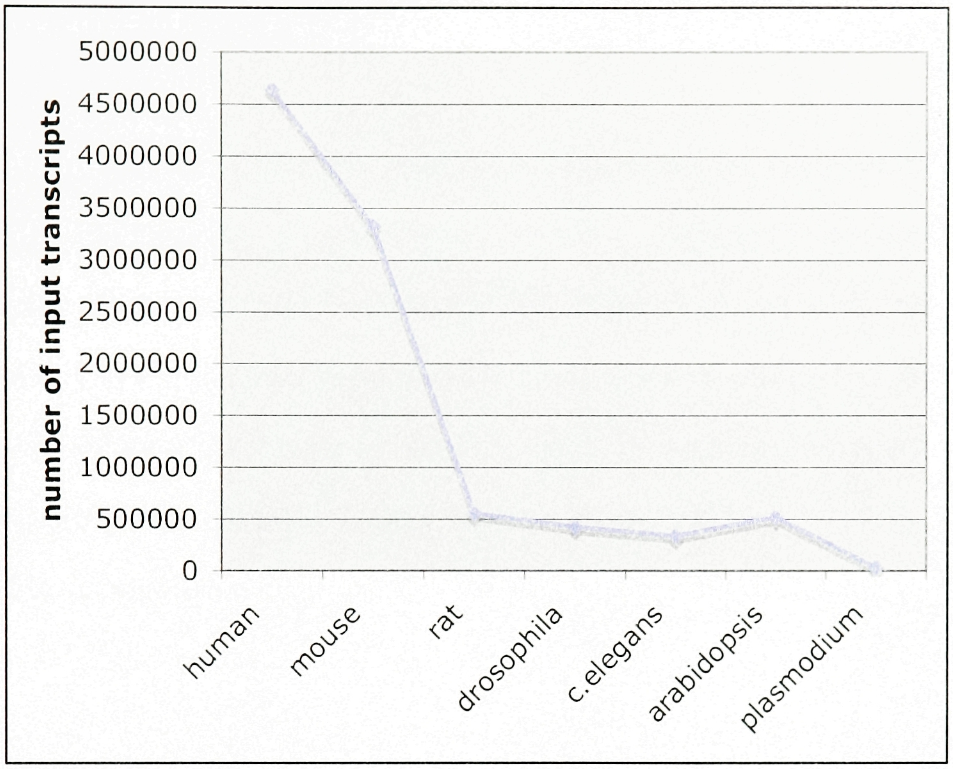


**Figure 6.2b** Detection of alternative splicing is positively correlated with the available expressed sequence data. Data-points show number of input transcripts for each organism.

## 6.7 Discussion

We developed alternative splicing databases for the transcriptomes of *Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana* and *Plasmodium falciparum* and computed the degree of variation in these organisms. These databases show that the method we developed to construct alternative splicing databases, namely the AutoDB pipeline, is applicable to genomes and transcriptomes beyond human, mouse and rat.

Our comparative statistical analyses in the seven species for which we developed alternative splicing, indicate that alternative splicing is more prevalent in genomes of more complex organisms. As further discussed in Chapter 8, development of additional databases using genomes of other organisms will shed light into how alternative splicing increases the complexity of genomes.

Availability of DmelSDB5, CeleganSDB5, AthalSDB3 and PfalSDB2 and any future database constructed using AutoDB, will prove useful in comparative studies of alternative splicing. Cross-species comparative analyses will give insight into functionally important splicing events along with species-specific splicing features.

# CHAPTER 7

# MATERIALS AND METHODS

## 7.1 Development of Alternative Splicing Databases

To study alternative splicing we developed three databases for human, mouse and rat transcriptomes called HumanSDB3, MouSDB5 and RatSDB2 respectively as introduced in Chapter 2. These databases are built based on modified versions of methods earlier described by Zavolan *et al.*, 2002. All three databases were developed using input transcript sequences downloaded from UniGene. We used UniGene human version no. 173, mouse version no. 139 and rat version no. 134 available from ftp://ftp.ncbi.nih.gov/repository/UniGene/. These transcript sequences were aligned to the UCSC genomes hg17 for human available at http://hgdownload.cse.ucsc.edu/goldenPath/hg17/, mm5 for mouse available at http://hgdownload.cse.ucsc.edu/goldenPath/mm5/, and rn3 for rat available at http://hgdownload.cse.ucsc.edu/goldenPath/rnJun2003/ using blat (Kent *et al.*, 2002). Top 10% matches from blat reports were aligned to the genomic regions by SIM4 (Florea *et al.*, 1998) and the top scoring match was considered to be the best alignment. The following criteria were applied to each best-aligned transcript sequence for inclusion in the final database. The transcript had to have a 75% or greater identity to the genome. In addition, every exon of the transcript had to have either a 95% identity to the genome or had to contain 5 or less mismatches. Each transcript had to have a minimum of two exons. Final clusters had to have a minimum of three transcripts each. Figure 7.1 shows the AutoDB pipeline used for creating the alternative splicing databases. In Chapter 2, Section 2.4 we detail the number of input transcripts lost at each step of AutoDB pipeline for HumanSDB3, MouSDB5 and RatSDB2.

DmelSDB5, CeleganSDB5, AthalSDB3 and PfalSDB2 were also developed using the AutoDB pipeline. Transcript data sources for full-length transcripts and ESTs of *Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana* and *Plasmodium falciparum* are GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) and dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) respectively. We used *Drosophila* and *C. elegans* UCSC genome (http://genome.ucsc.edu/) releases dm2 and Ce2 respectively. *Arabidopsis* genome was obtained from GenBank and *Plasmodium* genome was obtained from PlasmoDB (http://plasmodb.org/).

## 7.2 Annotation of Alternative and Constitutive Exons

An exon is considered *cassette* only if one or more transcripts which span the given genomic region are missing the given exon. Length variation is determined by inspecting all exons of a cluster which fall into the same

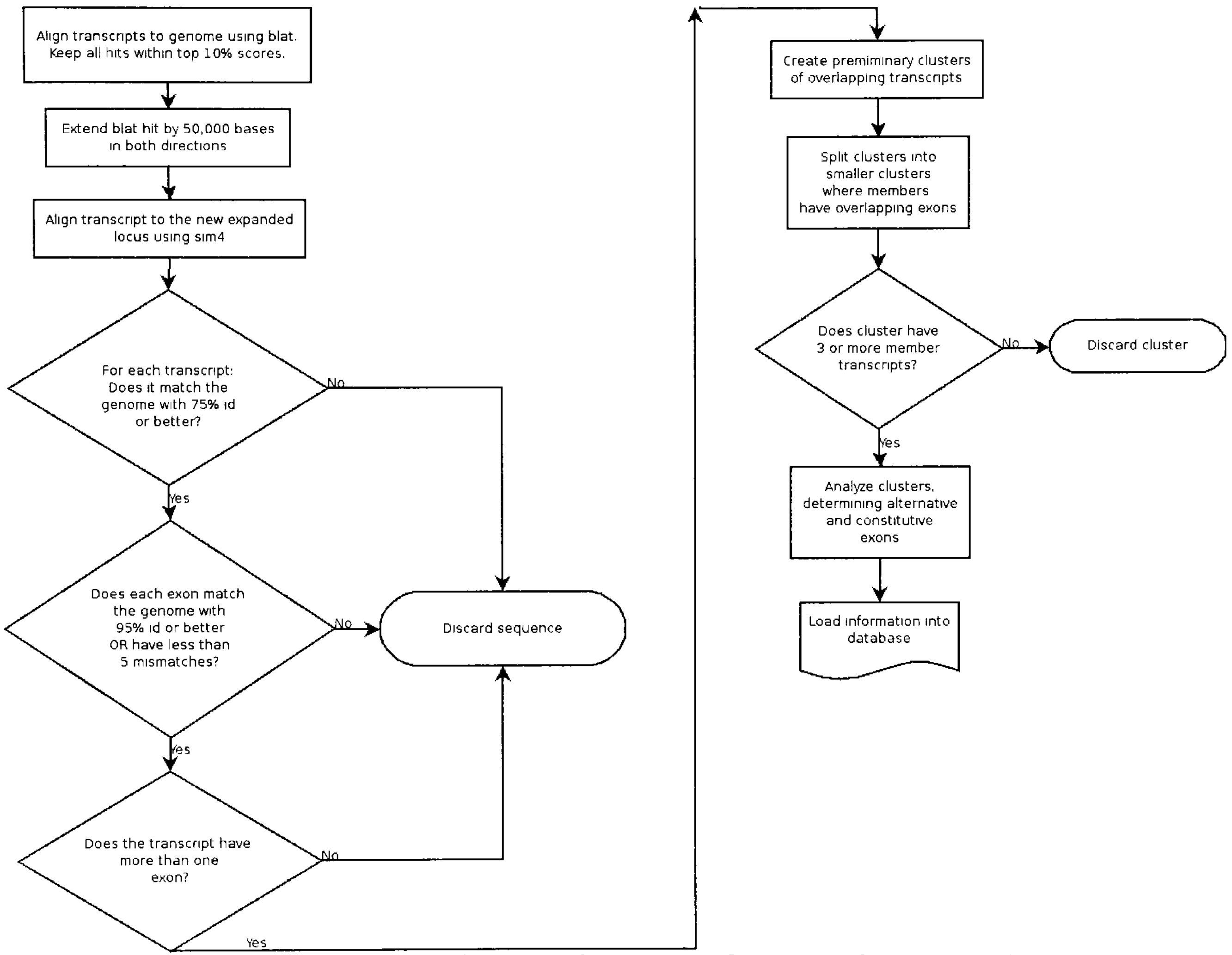

Align transcripts to genome using blat. Keep all hits within top 10% scores.
Extend blat hit by 50,000 bases in both directions
Align transcript to the new expanded locus using sim4
For each transcript: Does it match the genome with 75% id or better?
No
Yes
Does each exon match the genome with 95% id or better OR have less than 5 mismatches?
No
Discard sequence
Yes
Does the transcript have more than one exon?
No
Yes
Create premiminary clusters of overlapping transcripts
Split clusters into smaller clusters where members have overlapping exons
Does cluster have 3 or more member transcripts?
No
Discard cluster
Yes
Analyze clusters, determining alternative and constitutive exons
Load information into database

**Figure 7.1** AutoDB pipeline used in development of HumanSDB3, MouSDB5, RatSDB2, DmelSDB5, CeleganSDB5, AthalSDB3 and PfalSDB2.

genomic region. If the 5' end of these exons differ in nucleotide length between transcripts, then the exon is flagged as a *5′ length-variant* exon. Likewise, if the 3' end of these exons differ in nucleotide length between transcripts, then the exon is flagged as a *3′ length-variant* exon. In Chapter 1, we illustrate cassette and length-variant exons in Figure 1.2. Exons which do not differ in nucleotide length and appear in all transcripts sequenced from the same gene with the same splice sites are *constitutive* exons. Alternative and constitutive exons are further discussed in Chapter 2, Section 2.2.

## 7.3 Definition of Database Terms

For the work described in this thesis, we defined the following terms presented in italics. All terms are applicable to all databases. A *transcript* is a sequence transcribed from the genomic DNA sequence. A *locus* is a genomic region that includes a set of overlapping transcripts mapped to the genome such that a given transcript appears only in one locus. Within a locus a *cassette exon* is completely included in some transcripts and completely excluded from others. A *length-variant exon* has alternative 5' or 3' splice sites or both, in different transcripts. An exon can be both length-variant and cassette. A *variant exon* is either a cassette exon or a length-variant exon or both. A *genomic exon* is an uninterrupted series of nucleotides each of which maps to a transcript. A *cluster* is the set of transcripts that map to a locus. A *variant cluster* contains one or more variant exons. An *invariant cluster* has no variant exons (Taneri *et al.*, 2004; Taneri *et al.*, 2005).

## 7.4 AutoDB Schema and Statistical Analysis

AutoDB is built using a relational database system, PostgreSQL-7.4. AutoDB schema includes creation of the following six tables: Cluster Table, Clone Table, Clone Exon Table, Clone Intron Table, Cds Table and Genomic Exon Table. We used Perl DBI interface to perform SQL queries on these tables to extract statistical information. Cluster Table contains information on types of clusters as variant and invariant. Clone Table contains transcript information including tissues from which transcripts are sequenced. Clone Exon Table contains information on exons of each transcript in all clusters. Clone Intron Table contains information on introns of each transcript in all clusters. Genomic Exon Table contains information on all genomic exons from each cluster. All of the tables are consisted of several data columns, which are shown in Table 7.1. AutoDB database building process in detailed in Appendix E.

**Table 7.1** AutoDB Schema – Data columns included in AutoDB Tables.

| Cluster Table | Clone Table | Clone Exon Table | Clone Exon Table (continued) | Clone Intron Table | Cds Table | Genomic Exon Table |
|---|---|---|---|---|---|---|
| Cluster id | Clone id | Exon id | Splice junction 3site | Intron id | Clone id | Cluster exon |
| Chromosome id | Cluster id | Clone id | Cassette | Clone id | Chromosome id | Cluster id |
| Orientation | Data source | Exon number | Cass initial | Intron number | Orientation | Exon number |
| Chr beg | Chromosome id | Cluster exon | Cass internal | Chromosome id | Chr beg | Chromosome id |
| Chr end | Orientation | Chromosome id | Cass terminal | Orientation | Chr end | Orientation |
| Variant | Chr beg | Orientation | Variation 5end | Position clone | Clone beg | Chr beg |
|  | Chr end | Clone beg | Variation 3end | Chr beg | Clone end | Chr end |
|  | Clone length | Clone end | ss 5end | Chr end | Data source | Chr length |
|  | Number exons | Clone sequence | Map 5end | Splice site 5prime | Synonyms | Variant |
|  | Sequence | Chr beg | ss 3end | Splice site 3prime |  | Chr sequence |
|  | Library | Chr end | Map 3end | Data source |  | Cassette |
|  | Annotation | Chr sequence | Problem exon | Synonyms |  | Cass initial |
|  | Synonyms | Splice junction 5site | Data source |  |  | Cass internal |
|  |  |  | Synonyms |  |  | Cass terminal |
|  |  |  |  |  |  | Variation 5end |
|  |  |  |  |  |  | Variation 3end |

## 7.5 Web Implementation of the Databases

Online access to the PostgreSQL-7.4 splice databases (SDBs) is provided at http://genomes.rockefeller.edu/autodb/sdb.php, also mirrored at http://genomes.ucsd.edu/autodb/sdb.php. HumanSDB3, MouSDB5 and RatSDB2 web pages are dynamically generated by a series of PHP scripts, deployed on the Apache-2.0 web server. PostgreSQL database connections are carried out via built-in PHP database functions.

Each AutoDB pipeline-generated SDB has been supplemented by a number of additional tables that provide faster online access to the SDB statistics and contain information about splice clusters and individual chromosomes.

If a particular splice cluster is accessed the first time through a web interface, the PNG graphic cluster maps are generated either by PHP scripts or by a PERL script that uses a GD library. The generated graphic files are cached for faster subsequent access to the splice cluster. (A. Novoradovsky private communication).

## 7.6 Computation of Orthologous Splicing Clusters for *Compare Species* Feature

Computation of orthologous splicing clusters for the *Compare Species* feature of the databases was carried out as follows. Joined genomic exons (JGEs) were created for each variant cluster by joining all genomic exons from the genomic exon map of each cluster without including any intronic sequences. JGEs from one species were aligned to the JGEs of other species using tblastx. Pairwise synteny information for human-mouse (http://hgdownload.cse.ucsc.edu/goldenPath/hg17/vsMm5), for human-rat (http://hgdownload.cse.ucsc.edu/goldenPath/hg17/vsRn3) and for mouse-rat (http://hgdownload.cse.ucsc.edu/goldenPath/mm5/vsRn3) were used to establish orthologous clusters. Any JGE pair with an e-value of less than 1E-04 across more than 10% of the JGE length was considered orthologous. All exons from orthologous variant clusters were further aligned using tblastx. Exon pairs that matched with an e-value of 1E-04 or lower were considered orthologous exons.

## 7.7 Analyses of Tissue Distribution of Alternatively Spliced Transcripts

To study tissue-specificity of alternatively spliced isoforms, we defined the following four terms associated with each variant cluster.

(1) *homogeneity within tissues:*

all transcripts sequenced from the same tissue have cassette exons with same splice sites.

(2) *heterogeneity within tissues:*

transcripts sequenced from the same tissue have cassette exons with different splice sites.

(3) *homogeneity across tissues:*

transcripts sequenced from different tissues have cassette exons with same splice sites.

(4) *heterogeneity across tissues:*

transcripts sequenced from different tissues have cassette exons with different splice sites.

To analyze the tissue distribution of alternatively spliced mouse transcription factors discussed in Chapter 3, we used 18 tissues chosen from the existing libraries in MouSDB3 (a prior release of MouSDB5). These tissues were chosen on the basis of the fact that they contain both variant and invariant transcripts annotated as TFs. To account for library ambiguities, libraries which list several tissues or cell types for a single entry such as *mixture of brain and testis* and libraries which provide no tissue information at all such as *embryo* or *carcinoma* are screened out. In addition, several parts of a given tissue are pooled into one library. For example, the term *brain* corresponds to all parts of the brain found in MouSDB3 libraries, including cerebellum, thalamus, hippocampus and 16 other libraries.

To assess tissue-specificity of mouse TFs within tissues, we *used homogeneity-to-heterogeneity ratios.* To determine the homogeneity-to-heterogeneity ratio of isoforms within the 18 tissues studied, differences between isoforms are established based on the splice sites of cassette exons within the coding sequences of transcripts. Variations due to 5' and 3' truncations of transcripts are not taken into account.

## 7.8 Analysis of the Effect of Cassette Exons on Protein Domain Architecture

To analyze the effect of cassette exons on mouse TF domain architecture, we computed in-frame cassette exons which start at codon position one within the coding sequences of transcripts and have a nucleotide length of a multiple of three. These exons were computationally deleted from the TF transcripts. Original and altered coding sequences were translated into amino acid sequences and run through SMART to determine the altered domains. We illustrate original and altered transcripts along with the SMART domains computed for their amino acid sequences in Chapter 3, Figure 3.8. More information on SMART is provided in Chapter 3, Section 3.5.

## 7.9 Cancer Tissue Analysis

Cancer tissue analysis, discussed in Chapter 4, consists of several steps. Step one involves extracting all libraries from HumanSDB3 and separating those into normal and cancer categories. SQL query for extracting cancer libraries was constructed to search for words ending in 'oma', and/or beginning with either 'cancer' or 'tumor'. All other library entries were placed in the normal tissue list. Both lists were further manually filtered and yielded to library lists for 12 tissues as follows: brain, breast, colon, kidney, liver, lung, ovary, pancreas, prostate, skin, thymus, and uterus.

Using the normal and cancer library lists, we define variant clusters in HumanSDB3 as *normal, cancer* or *mixed.* A normal cluster contains transcripts sequenced only from normal tissues. A cancer cluster contains transcripts sequenced only from cancer tissues. A mixed cluster contains transcripts sequenced both from normal tissues and from cancer tissues. Clusters which contain transcripts with no library information are not counted. Same cluster analysis was repeated within each of the 12 tissues. In addition, statistical analysis of exons provided minimum, maximum and average number of cassette and length-variant exons for all three types of clusters.

To assess differences of alternative splicing between normal and cancer tissues, we compared cassette exon splice sites from normal and cancer transcripts within the same variant clusters previously defined as *mixed clusters.* Transcript numbers sequenced from each tissue and whether or not they display different splicing patterns were reported.

## 7.10 TFHunter Development

To find Nova binding sites on uploaded input sequences, TFHunter utilizes position weight matrices for each binding motif not allowing for any mismatches. A postscript scores the input sequence using sliding windows based on position weight matrices. Each window size is equal to a given motif size. (Nova binding motif sequences are provided in Chapter 5, Fig. 5.1). Images of binding motif localizations on input sequences are web implemented using GD.pm scripts and can be accessed at http://genomes.rockefeller.edu/yupu-cgi/TFH/main.cgi.

## 7.11 NTHunter Algorithm

As introduced in Chapter 5, our prediction program, called NTHunter (Nova Target Hunter), selects a master RNA from each variant cluster. Master RNAs are the transcripts with the highest number of exons within a given cluster. From the master RNA, NTHunter extracts exon sequences with their 500 base-pair upstream and 500 base-pair downstream intronic sequences. Exons are labeled with respect to their variation types. NTHunter divides the exon into 100 base-pair regions and searches for Nova binding motif clusters within these regions using a sliding window. Window size is the size of a given motif and it is indexed at 1 base-pair. NTHunter searches for presence of at least 4 of motifs 1, 2 and/or 3 and at least 2 of motifs 4, 5, 6 and/or 7. (Motif sequences are shown in Chapter 5, Fig. 5.1). If any cassette exon is picked up by NTHunter, the transcript containing the given exon is flagged as a candidate gene for Nova regulation.

## 7.12 Sampling Depth Controls

Sampling depth denotes the absolute number of transcripts of a given gene sequenced from a given tissue. Since some tissues are more frequently used in DNA libraries, in all our analyses we account for sampling depths from each tissue. To control for any sampling depth bias, the absolute number of transcripts sequenced per tissue is normalized in each analysis.

# CHAPTER 8

# CONCLUSIONS

## 8.1 Discussion of Results

The work described in this thesis provides quantitative evidence for contribution of tissue-specific alternative splicing to the complexity of transcriptomes and proteomes. In this study, along with three new databases of alternative splicing in human, mouse, rat transcriptomes and computational methods to compare variant transcription at gene loci, we introduce a novel web-based visualization method to study comparative alternative splicing. This tool brings to the end-user the ability to analyze alternative splicing in their gene of interest. Users can view all exons of their gene, access their nucleotide sequences and learn about the libraries of the transcripts sequenced for that gene. In addition, users are able to find orthologous human-mouse, human-rat and mouse-rat pairs for their genes of interest and to study pairs of orthologous constitutive and alternative exons in detail.

Our analyses of the human, mouse and rat transcriptomes show that alternative splicing is widespread within all three species. Overall variation in human loci is 81%. Variant mouse and rat loci are 74% and 58% respectively. In all three transcriptomes, alternative splicing is mainly due to the presence or absence of cassette exons. 67% of human alternative exons are cassette exons, as are 62% of mouse and 62% of rat alternative exons. This finding indicates functional significance of this type of alternative exons.

In addition, our results show widespread presence of a specific subset of cassette exons which are present either at the 5' or at the 3' ends of transcripts. These exons, termed transcript-terminal cassette exons, do not map to any other genomic exon within their cluster, but rather map to intronic sequences. Transcript-terminal exons are either due to alternative starts and ends of transcription or they might be due to sequence artifacts.

Through integrated analyses of DNA and protein sequences for transcription factor (TF) genes, we show that alternative splicing of TFs is more prevalent in the entire mouse transcriptome and in specific tissues when compared to alternatively spliced forms of all variant genes. In 78% of the analyzed tissues, a higher proportion of TFs exhibit alternative splicing than does the set of all variant genes in the mouse transcriptome. This result along with the finding that 62% of mouse TF loci are variant indicates the widespread impact of alternative splicing on regulation of transcription factor function.

We show that alternative splicing changes TF structure by adding or deleting domains via cassette exons. Our study reveals that 80% of alternatively spliced mouse TFs have different domain architectures due to introduction of an in-frame cassette exon by alternative splicing. 75% of the altered domains play a role in DNA binding. These findings provide quantitative evidence for the role of alternative splicing in controlling the presence of domains in proteins. They also suggest that alternative splicing might regulate TF activity and thus alter gene expression by changing the DNA-binding domain architecture of these proteins.

Our analyses reveal that within a single tissue there generally is only one TF isoform, and across tissues, TF isoforms differ. This finding indicates tissue specificity of alternatively spliced TFs and suggests that TFs might regulate gene expression in a tissue specific manner by having different isoforms in different tissues. In addition, our study shows that all variant loci in the mouse transcriptome display isoform homogeneity within single tissues and heterogeneity across different tissues, indicating tissue-specificity of alternatively spliced mouse genes. This finding greatly expands the knowledge on contribution of alternative splicing to tissue-specific expression of mouse genes.

In addition, we show that in the human transcriptome, there is a high prevalence of transcript sequence data from cancer tissues. More than 80% of human variant loci contain transcripts from cancer tissues. Further, we show that cancer transcripts introduce variation beyond normal alternative splicing. In the majority of tissues, more than 20% of the cassette exons are from cancer transcripts only. Our results quantitatively validate presence of aberrant alternative splicing in cancer sequence data.

Lastly, by studying alternatively spliced genes in transcriptomes of *Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana* and *Plasmodium falciparum* in comparison to those in human, mouse and rat, we showed that alternative splicing is more prevalent in genomes of more complex organisms.

## 8.2 Implications of Work and Future Directions

The work described in this thesis has implications in several different fields and paves way to numerous future studies as described below.

### (a) Alternative Splicing and Complexity of Genomes

The databases described in this thesis provide detailed information on alternative splicing in three mammalian transcriptomes, *Homo sapiens, Mus musculus* and *Rattus norvegicus,* as well as alternative splicing in three model organisms, *Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana* and a pathogen *Plasmodium falciparum.* Our studies on these organisms indicate that there is more alternative splicing in more complex organisms and that there is an elevation of alternative splicing in mammalian genomes. To further investigate whether alternative splicing increases as organisms get more complex, it is of future interest to run AutoDB on available expressed sequence data and genomes of several other organisms including the following.

*Fugu rubripes* (Fugu)

*Danio rerio* (Zebrafish)

*Caenorhabditis briggsae* (Worm)

*Gallus gallus* (Chicken)

*Pan troglodytes* (Chimpanzee)

*Ciona intestinalis* (Sea squirt)

*Oryza sativa* (Rice)

Statistical analysis of alternative splicing in transcriptomes of the organisms listed above, along with the statistical analysis provided in this thesis will shed light on how alternative splicing increases the complexity of different genomes.

### (b) Comparative Analysis of Conserved Alternative Splicing

The work described here has significant implications in further understanding the evolution, function and regulation of alternative splicing. We provide a novel method for studying conservation of alternative splicing. Our visualization tool for orthologous exon pairs provides a means for comparative studies of alternative exons in human-mouse, human-rat and mouse-rat gene-pairs. Easy access to comparative alternative splicing data together with instant information retrieval about the variant exons and their sequences further aids in experimental design for alternative splicing studies.

The Compare Species tool we developed opens up possibilities for studies of not only the conserved sequences, but also the non-conserved species-specific features such as exons that are present only in human and not in mouse. This feature can also be used in identification of *shadow exons*. Such exons are annotated in one species, they are present in the intronic sequence of the other species but they have not yet been annotated as exons.

In addition, future work using the Compare Species tool could focus on investigating parallelism in splicing patterns across species, particularly on human, mouse, rat gene-pairs and gene-triplets. Concordance of genomic structures of homologous genes with respect to alternative exons and their variation types could provide insight into functionally important splicing events. In addition, parallelism in characteristics of alternative exons could be studied, such as reading frame preservation.

It is of further interest to extend comparative studies of alternatively spliced genes beyond mammalian genomes. For this purpose, the three databases described in this thesis for transcriptomes of *Drosophila melanogaster, Caenorhabditis elegans* and *Arabidopsis thaliana* will prove useful. In addition new splicing databases in organisms mentioned in Section 8.2a of this chapter could provide resources for detailed analysis of conserved alternative exons.

Species analysis could be extended to more closely related organisms such as human, chimpanzee and macaque and compared to more distantly related organisms such as mouse. Identification of splicing events present in human, chimpanzee and macaque but not in mouse can reveal evolutionarily important alternative splicing events and show how alternative splicing increases genome complexity of higher organisms.

### (c) Aberrant Alternative Splicing in Cancer Transcripts

Our studies point to high prevalence of cancer transcripts in the human transcriptome data which lead to introduction of variation beyond normal alternative splicing. These findings imply that future studies of alternative splicing should take into account differences of variation between normal and cancer transcripts. Future analysis of cancer-specific cassette exons should focus on transcripts from tumor tissues only. Exclusion of transcripts from cancer cell lines would provide more accurate information with regards to the cassette exons introduced by cancer transcripts. Normal-only cassette exons would be determined in the same way by comparative studies of transcripts from normal tissues and tumor tissues. In addition to studies of cancer-specific cassette exons described in this thesis, it would be of further interest to investigate differences of length-variant exons between transcripts sequenced from normal and cancer tissues.

### (d) NOVA Regulated Alternative Splicing

Our studies on neuron specific splicing regulator NOVA point to the importance of studying tissue-specific RNA binding proteins and indicate that motif searches in tissue-specific sets of transcripts would lead to identification of novel regulators of splicing specific to given tissues.

To expand the studies on NOVA binding motifs described in this thesis, the program GoMiner (Zeeberg *et al.*, 2003) could be used to get functional annotations of predicted NOVA target genes. Genes that function at the synapses and in axon guidance would be of higher priority to be experimentally tested. In addition, identification of orthologous human-mouse, human-rat and mouse-rat target gene-pairs and human-mouse-rat target gene-triplets would prove useful for NOVA binding site searches. Genes which have orthologous targets and function either as a synaptic protein or in axon guidance would be tested experimentally. RT-PCR could be used as a method to validate regulation of these candidate alternative exons by NOVA.

Furthermore, to learn more about regulation of alternative exons by NOVA, orthologous sets of human, mouse and rat alternative exons can be generated, based on experimentally identified mouse exons, which are known to be either upregulated or downregulated by NOVA. A set of alternative exons which are not regulated by NOVA would be used as a control sequence set. In these sets of sequences, searches for NOVA binding motifs can be developed on variations of YCAY motifs. In addition, overrepresentation of any other motifs could be investigated. Any newly identified motifs could function as binding sites for other RNA-binding proteins which might act antagonistically to NOVA.

NOVA was identified as a self-antigen in an autoimmune motor dysfunction disorder called POMA (Buckanovich *et al.*, 1996). This disorder occurs in patients with certain cancers, such as breast and ovarian cancer, when neuron-specific proteins are expressed in tumor cells. In this thesis, we have shown existence of cancer-specific alternative exons. It would be of special interest to study this notion with respect to NOVA. To this extent, cassette exons in nervous tissues such as brain and cassette exons in cancerous breast and ovarian tissues would be compared to the cassette exons in normal ovary and breast tissues. This approach would lead to identification of neuron-specific cassette exons expressed in cancer tissues.

Understanding tissue-specificity of alternative splicing is key to understanding how this biological process contributes to differential gene expression. Therefore, in addition to studies of tissue-specific RNA binding proteins, it is of particular interest to study global tissue-specificity of distinct isoforms of alternatively spliced genes. To address this question global comparative tissue-distribution analysis in orthologous sets of alternatively spliced genes in human-mouse, human-rat and mouse-rat can be performed. Furthermore, investigation of temporal specificity of distinct isoforms would provide insight into the role of alternative splicing during development and differentiation.

### (e) Global Detection of Alternative Splicing Events via Splicing Arrays

A natural extension of the studies described in this thesis would be the development of robust measurements for tissue-specific splice forms. For this purpose, microarrays which measure tissue-specific splicing events and methods to detect tissue-specific alternative splicing from microarray expression data are being developed. To this extent, Yeakley *et al.* describe a fibre-optic array used to analyze alternative splicing in human genes (Yeakley *et al.*, 2002). Pan *et al.* utilize a quantitative microarray to study tissue-specific alternative splicing in mouse genes (Pan *et al.*, 2004). Wang *et al.* describe a splice array and introduce an algorithm which detects known splice variants (Wang *et al.*, 2003). Le *et al.* develop a method for detection of tissue-specific alternative splicing from microarray data (Le *et al.*, 2004). Yeo *et al.* study distribution of alternative splicing across human tissues utilizing microarray expression data (Yeo *et al.*, 2004). Johnson *et al.* utilize exon junction microarrays to study alternative splice forms of human genes on a whole genome scale (Johnson *et al.*, 2003).

Information in our databases will enable production of splicing microarrays which measure the expression levels of alternatively spliced isoforms of genes. These microarrays can be on a whole genome scale and they can be built for different species such as human and mouse. The work described here implies contribution of alternative splicing to regulation of gene expression via tissue specific TF isoforms and thus indicates that future high-throughput screens of gene expression analyses should be sensitive to multiple alternatively spliced forms of TFs. Since gene expression arrays are intended to measure transcription, the next generation of arrays should contain probes specific to all known isoforms of genes represented on the arrays.

Splicing chip data together with the information in our alternative splicing databases will enable individual experiments such as validation of tissue-specific isoforms by quantitative PCR, which can verify the existence of different isoforms and will lead to functional characterization of isoforms via further experiments. All this effort will lead to a better understanding of proteomes (Fig. 8.1).

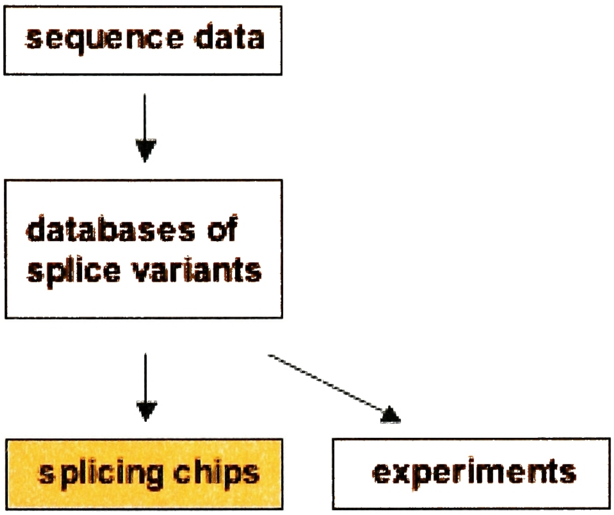


**Figure 8.1** Overview of expansion of alternative splicing research. Information provided in HumanSDB3, MouSDB5, RatSDB2, DmelSDB5, CeleganSDB5, AthalSDB3 and PfalSDB2 will lead to possible splicing microarrays and further wet-lab experiments.

### (f) Effect of Alternative Splicing on Protein Structures

Given the fact that alternatively spliced exons are highly conserved across species (Sorek and Ast, 2003), it would be of further interest to study the effect of cassette exons on proteomes of other organisms. Strong sequence homology between mouse, human and rat exons suggests that a comparative analysis of human, mouse and rat TF variations will be a natural extension of the studies described here.

Other sets of proteins which are widely alternatively spliced are kinases and G-protein coupled receptors. The studies on the effect of cassette exons on protein domain architecture could be expanded to these sets of proteins. It would be of special interest to investigate if alternative splicing regulates these proteins by targeting a specific functional domain of these proteins, as we have shown in this thesis for transcription factor DNA binding domains.

In addition, it would be of further interest to investigate differential effects of major an minor forms of cassette exons in protein domain architectures. Major forms of cassette exons appear in more than 50% of the sequenced transcripts and minor forms of cassette exons appear in less than 50% of the sequenced transcripts of a given gene (Modrek and Lee, 2003). Analysis of effects of species-specific minor cassette exons on protein domain architecture would be of special interest.

### (g) Transcript-initial and Transcript-terminal Cassette Exons

Our results point out to high occurrences of *transcript-initial* and *transcript-terminal* cassette exons. To gain more information on transcript-initial (5'-most exon) and transcript-terminal (3'-most exon) cassette exons, we will study the properties of these exons including nucleotide length distribution of these exons, their frame preservation, tissue distribution and whether or not they are conserved across human, mouse and rat. In addition, average number of constitutive and average number of alternative exons in transcripts containing transcript-initial and transcript-terminal cassette exons could be compared to those numbers in transcripts which do not have these exons. Upstream regions of transcript-initial cassette exons could be analyzed for gene expression regulatory sequences and transcript-terminal cassette exons could be analyzed for presence of polyadenylation signals. These analyses will help determine whether transcript-initial and transcript-terminal cassette exons are real starts and ends of transcription or they are due to errors in splicing and/or sequencing. Furthermore, future studies of alternative splicing should take into account the transcript-initial and transcript-terminal cassette exons when investigating variation within transcriptomes.

### (h) Regulation of Length-Variation in Alternative Exons

Our studies point to an intron-centric regulation of length variation in alternative exons rather than an exon-centric regulation. To further investigate this process, conserved intronic sequence elements flanking the human, mouse and rat length-variant exons could be studied. Searches for splicing regulatory sequence elements within the length-variant exons and comparing these to the conserved sequence elements in their flanking intronic regions will provide insight into regulation of length-variation in alternative exons.

### (i) Investigating Potential Regulation of Alternative Splicing by miRNAs

miRNAs are small non-coding RNA molecules which play an important role in translational repression of mRNAs (Pasquinelli and Ruvkun, 2002). Previous studies indicate possible regulation of alternative splicing by miRNAs in *Arabidopsis thaliana* (X. Wang, Dissertation 2004). To investigate the notion of regulation of alternative splicing via miRNAs in human and mouse, HumanSDB3 and MouSDB5 databases described in this thesis can be utilized. miRNAs could silence splicing by binding to splice sites. To investigate the degree to which this happens in alternatively spliced human and mouse genes, exon-intron junctions and exon-exon junctions of alternative exons could be searched for potential miRNA binding sequences.

# APPENDIX A

## LIST OF PROTEIN DOMAINS ALTERED BY CASSETTE EXONS (FROM CHAPTER 3)

| Cluster ID | Transcription Factor Annotation | Domains Altered by Cassette Exons |
|---|---|---|
| Scl18190 | Calcium-independent Phosholipase A2 Isoform 2 | Abhydrolase_2 |
| Scl18706 | Dream/calsenilin | EF hand |
| Scl19415 | transcription factor PBX3a (PBX3) | Homeobox, PBX |
| Scl19959 | hepatocyte nuclear factor 4 | zf-C4, hormone_rec |
| Scl22286 | Evi-1 transcription factor splice variant delta 105 | zf-C2H2 |
| Scl22515 | weakly similar to multifunctional aminotransferase (KAT) / (GTK) | Aminotran_1_2 |
| Scl23794 | similar to nuclear matrix transcription factor | znf-c2h2 |
| Scl24819 | transcription elongation factor TFIIS | TFIIS |
| Scl25542 | nuclear transcription factor, X-box binding 1 | zf-NF-X1 |
| Scl27217 | TFIIH basal transcription factor complex P34 subunit | Tfb4 |
| Scl29156 | CCR4-NOT transcription complex, subunit 4 | rrm, coiled coil |
| Scl29710 | microphthalmia-associated transcription factor (Mitf) | HLH |
| Scl30364 | forkhead-related transcription factor 2 (Foxp2) | Fork head, coiled coil |

| | | |
|---|---|---|
| Scl30775 | SRY-box containing gene 6 (Sox6) | HMG BOX |
| Scl31504 | upstream transcription factor 2 (Usf2) | HLH |
| Scl31929 | NK6 transcription factor related | exo-endo-phos |
| Scl34454 | Transcription factor IIB | Kinase, DUF667 |
| Scl36910 | transcriptional regulator, SIN3A | PAH |
| Scl38300 | TTF-I interacting protein 5 (TIP5 gene) | MBD, AT_hook, DDT, bromodomain, PHD |
| Scl39514 | transcription factor UBF | HMG |
| Scl40076 | myocardin A (BSAC2A), alternatively spliced | SAP |
| Scl40242 | transcription factor 7, T-cell specific | HMG BOX |
| Scl41218 | flotillin 2 (Flot2) | band-7, flotillin |
| Scl45206 | Kruppel-like factor 12 (Klf12) | znf-c2h2 |
| Scl48808 | activator protein 4 (Ap4) | HLH, coil |
| Scl50663 | similar to zinc finger transcription factor TREP-132 | znf-c2h2, ELM2, myb-DNA binding |
| Scl50814 | cAMP responsive element binding protein-like 1 | BRLZ, BZIP, coil |

# APPENDIX B

## LIST OF HUMAN TARGET GENES PREDICTED BY NTHunter (FROM CHAPTER 5)

| GenBank ID - Exon No | Annotation |
|---|---|
| AB007510.17 | mRNA for PRP8 protein, complete cds |
| AB014597.26 | mRNA for KIAA0697 protein, partial cds |
| AB016073.13 | mRNA for p51B, complete cds |
| AB017913.3 | cBGL1 mRNA for cytosolic beta-glucosidase-like protein-1, complete cds |
| AB020522.11 | DLEC1 (deleted in lung and esophageal cancer 1; DLEC1 alias DLC1) mRNA, complete cds |
| AB023204.13 | mRNA for KIAA0987 protein, partial cds |
| AB023204.19 | mRNA for KIAA0987 protein, partial cds |
| AB033076.25 | mRNA for KIAA1250 protein, partial cds |
| AB037669.6 | hLAT2 mRNA for L-type amino acid transporter 2, complete cds |
| AB053309.5 | ALS2CR8 mRNA, complete cds, long form |
| AB055660.7 | hShrmL mRNA for Shroom-related protein, complete cds |
| AB112074.17 | RBBP6 mRNA for retinoblastoma binding protein 6 isoform 1, complete cds |
| AF001177.2 | casein kinase I gamma 2 primary transcript, complete cds |
| AF002979.6 | killer cell receptor (KIR103) mRNA, allele LP, complete cds |
| AF009242.2 | proline-rich Gla protein 1 (PRGP1) mRNA, complete cds |
| AF022728.18 | beta-dystrobrevin (BDTN) mRNA, complete cds |
| AF026548.10 | branched chain alpha-ketoacid dehydrogenase kinase precursor, mRNA, nuclear gene encoding mitochondrial protein, complete cds |
| AF034771.6 | natural killer cell inhibitory receptor (KIR2DL4) mRNA, variant 1, complete cds |
| AF034773.6 | natural killer cell inhibitory receptor (KIR2DL4) mRNA, variant 3, complete cds |
| AF037333.12 | Eph-like receptor tyrosine kinase hEphB1c (EphB1) mRNA, complete cds |
| AF040965.15 | unknown protein IT12 mRNA, partial cds |
| AF051907.3 | PRUNE-like protein mRNA, complete cds |
| AF055989.2 | Shaw type potassium channel Kv3.3 (KCNC3) mRNA, complete cds |

| | |
|---|---|
| AF075575.54 | dysferlin mRNA, complete cds |
| AF092565.17 | splicing factor Prp8 mRNA, complete cds |
| AF102546.5 | dachshund (DACH) mRNA, complete cds |
| AF109388.2 | P2X2B receptor mRNA, complete cds |
| AF126424.5 | cell cycle checkpoint protein (R24L) mRNA, complete cds |
| AF146074.23 | ABC protein mRNA, complete cds |
| AF165967.2 | DDP-like protein mRNA, complete cds |
| AF170562.14 | ubiquitin-specific processing protease (USP25) mRNA, complete cds |
| AF171669.6 | glycoprotein-associated amino acid transporter LAT2 (LAT2) mRNA, complete cds |
| AF182034.9 | polycystic kidney disease-like 2 protein (PKDL2) mRNA, complete cds |
| AF190823.2 | P2X2B receptor (P2X2) mRNA, complete cds |
| AF208857.8 | BM-015 mRNA, complete cds |
| AF212229.9 | GL008 mRNA, complete cds |
| AF251052.14 | RALBP1 mRNA, complete cds |
| AF260427.2 | purinoceptor P2X2B (P2X2) mRNA, complete cds, alternatively spliced |
| AF285436.6 | clone KIR2DL4v4 killer-cell Ig-like receptor mRNA, complete cds |
| AF301013.11 | regulator of nonsense transcripts 2 (RENT2) mRNA, complete cds |
| AF311326.18 | testis development protein PRTD [Homo sapiens], mRNA sequence |
| AF311326.21 | testis development protein PRTD [Homo sapiens], mRNA sequence |
| AF317840.3 | cytosolic beta-glucosidase mRNA, complete cds |
| AF318574.11 | UPF2 (UPF2) mRNA, complete cds |
| AF323728.19 | OSBP-related protein 6 mRNA, complete cds |
| AF323990.3 | cytosolic beta-glucosidase (CBG) mRNA, complete cds |
| AF325189.10 | NADPH oxidase 5 beta mRNA, complete cds |
| AF325190.10 | NADPH oxidase 5 beta mRNA, complete cds |
| AF350500.6 | Four-span transmembrane protein 1 (4SPAN1) mRNA, complete cds |
| AF356492.5 | dachshund (DACH) mRNA, complete cds |
| AF380181.3 | SON DNA binding protein isoform C (SON) mRNA, complete cds, alternatively spliced |
| AF380183.3 | SON DNA binding protein isoform E (SON) mRNA, complete cds, alternatively spliced |
| AF392448.19 | oxysterol-binding protein-like protein OSBPL6 (OSBPL6) mRNA, complete cds |

| | |
|---|---|
| AF410459.21 | CD109 [Homo sapiens], mRNA sequence |
| AF417489.19 | tensin 3 mRNA, complete cds |
| AF450090.4 | KCCR13L mRNA, complete cds |
| AF467288.30 | BCL8B protein (BCL8B) mRNA, complete cds |
| AF479813.3 | chemokine-like factor super family 3 (CKLFSF3) mRNA, complete cds |
| AF479813.4 | chemokine-like factor super family 3 (CKLFSF3) mRNA, complete cds |
| AF502591.15 | BRCC1 (BRCC1) mRNA, complete cds |
| AF513717.3 | E3 ubiquitin ligase (TRIAD3A) mRNA, complete cds; alternatively spliced |
| AF513718.3 | E3 ubiquitin ligase (TRIAD3B) mRNA, complete cds; alternatively spliced |
| AF520570.7 | unknown mRNA |
| AF523354.10 | prestin (PRES) mRNA, complete cds |
| AJ001443.17 | mRNA for SAP 130 spliceosomal protein |
| AJ005670.5 | mRNA for dachshund protein |
| AJ133767.5 | mRNA for ZASP protein, partial, varient 3 |
| AJ251595.9 | mRNA for transmembrane glycoprotein (CD44 gene) |
| AJ252246.16 | mRNA for kainate receptor subunit (GRIK2 gene) |
| AJ278964.3 | partial mRNA for cytosolic beta-glucosidase (GLUC gene) |
| AJ400850.2 | mRNA for MUC4 protein splice variant sv13 (MUC4 gene) |
| AJ400850.5 | mRNA for MUC4 protein splice variant sv13 (MUC4 gene) |
| AJ488102.21 | mRNA for steerin2 protein (STEERIN2 gene) |
| AJ711055.2 | CMPD01 Homo sapiens cDNA clone CMPD08387, mRNA sequence |
| AK001000.5 | cDNA FLJ10138 fis, clone HEMBA1003148, highly similar to Homo sapiens mRNA full length insert cDNA clone EUROIMAGE 381801 |
| AK026927.2 | cDNA: FLJ23274 fis, clone HEP02623, highly similar to AF161354 Homo sapiens HSPC091 Mrna |
| AK027115.2 | cDNA: FLJ23462 fis, clone HSI08475 |
| AK027255.5 | cDNA: FLJ23602 fis, clone LNG15735 |
| AK027375.13 | cDNA FLJ14469 fis, clone MAMMA1000897, weakly similar to INTER-ALPHA-TRYPSIN INHIBITOR HEAVY CHAIN H3 PRECURSOR |
| AK055726.4 | cDNA FLJ31164 fis, clone KIDNE1000104, weakly similar to SYNTAXIN 7 |
| AK055752.15 | cDNA FLJ31190 fis, clone KIDNE2000403 |
| AK056216.1 | cDNA FLJ31654 fis, clone NT2RI2004230, highly similar to Mus musculus mRNA for homeodomain protein Meis2a |

| | |
|---|---|
| AK056324.3 | cDNA FLJ31762 fis, clone NT2RI2007754, weakly similar to INTESTINAL MEMBRANE A4 PROTEIN |
| AK056324.4 | cDNA FLJ31762 fis, clone NT2RI2007754, weakly similar to INTESTINAL MEMBRANE A4 PROTEIN |
| AK056412.14 | cDNA FLJ31850 fis, clone NT2RP7000600, highly similar to Homo sapiens activating receptor PILRbeta mRNA |
| AK056819.2 | cDNA FLJ32257 fis, clone PROST1000262, highly similar to VASOACTIVE INTESTINAL POLYPEPTIDE RECEPTOR 1 PRECURSOR |
| AK056875.8 | cDNA FLJ32313 fis, clone PROST2003232, weakly similar to BETA-GLUCURONIDASE PRECURSOR (EC 3.2.1.31) |
| AK074370.6 | cDNA FLJ23790 fis, clone HEP21466 |
| AK074532.5 | cDNA FLJ90051 fis, clone HEMBA1002551, weakly similar to PUTATIVE SERINE |
| AK074744.4 | cDNA FLJ90263 fis, clone NT2RM4000997, weakly similar to MONO- AND DIACYLGLYCEROL LIPASE PRECURSOR (EC 3.1.1.-) |
| AK074772.7 | cDNA FLJ90291 fis, clone NT2RP1001031, weakly similar to VEGETATIBLE INCOMPATIBILITY PROTEIN HET-E-1 |
| AK075081.6 | cDNA FLJ90600 fis, clone PLACE1001401, weakly similar to HIGH AFFINITY IMMUNOGLOBULIN EPSILON RECEPTOR BETA-SUBUNIT |
| AK075128.4 | cDNA FLJ90647 fis, clone PLACE1004282, weakly similar to MONO- AND DIACYLGLYCEROL LIPASE PRECURSOR (EC 3.1.1.-) |
| AK075381.13 | cDNA PSEC0071 fis, clone NT2RP2002115, weakly similar to INTER-ALPHA-TRYPSIN INHIBITOR HEAVY CHAIN H3 PRECURSOR (ITI HEAVY CHAIN H3) (SERUM-DERIVED HYALURONAN-ASSOCIATED PROTEIN) (SHAP) |
| AK075413.9 | cDNA PSEC0104 fis, clone NT2RP2004795 |
| AK075560.2 | cDNA PSEC0260 fis, clone NT2RP3004059 |
| AK090431.2 | mRNA for FLJ00348 protein |
| AK091522.10 | cDNA FLJ34203 fis, clone FCBBF3019784, highly similar to Mouse mRNA for seizure-related gene product 6 type 2 precursor |
| AK091990.2 | cDNA FLJ34671 fis, clone LIVER2001099, moderately similar to Rattus norvegicus mRNA for putative integral membrane transport protein |
| AK092358.7 | cDNA FLJ35039 fis, clone OCBBF2017035, highly similar to Mus musculus mRNA for GATS protein |
| AK092757.5 | cDNA FLJ35438 fis, clone SMINT2002884, weakly similar to CMRF35 ANTIGEN PRECURSOR |
| AK093303.2 | cDNA FLJ35984 fis, clone TESTI2014097, highly similar to V_segment translation product |

| | |
|---|---|
| AK093627.2 | cDNA FLJ36308 fis, clone THYMU2004916, highly similar to BRANCHED-CHAIN AMINO ACID AMINOTRANSFERASE, MITOCHONDRIAL PRECURSOR (EC 2.6.1.42) |
| AK095389.15 | cDNA FLJ38070 fis, clone CTONG2015518 |
| AK096516.4 | cDNA FLJ39197 fis, clone OCBBF2005077, moderately similar to CARNITINE O-PALMITOYLTRANSFERASE I, MITOCHONDRIAL LIVER ISOFORM (EC 2.3.1.21) |
| AK096902.6 | cDNA FLJ39583 fis, clone SKMUS2004897, highly similar to ACTIN, ALPHA SKELETAL MUSCLE |
| AK098653.5 | cDNA FLJ25787 fis, clone TST06855, highly similar to CYCLIN-DEPENDENT KINASE INHIBITOR 3 (EC 3.1.3.48) |
| AK098666.2 | cDNA FLJ25800 fis, clone TST07092 |
| AK122673.2 | cDNA FLJ16118 fis, clone ASTRO2013585 |
| AK122747.5 | cDNA FLJ16272 fis, clone NT2NE2018916, weakly similar to POLYPEPTIDE N-ACETYLGALACTOSAMINYLTRANSFERASE (EC 2.4.1.41) |
| AK123905.4 | cDNA FLJ41911 fis, clone PEBLM2008605, highly similar to VERY-LONG-CHAIN ACYL-COA SYNTHETASE (EC 6.2.1.-) |
| AK124140.4 | cDNA FLJ42146 fis, clone TESTI4000434, highly similar to Homo sapiens SPG protein (SPG) mRNA |
| AK124273.3 | cDNA FLJ42279 fis, clone TLIVE2002690 |
| AK125377.10 | cDNA FLJ43387 fis, clone OCBBF2006764, highly similar to Mus musculus seizure related gene 6 (Sez6) |
| AK125436.4 | cDNA FLJ43447 fis, clone OCBBF2032590, weakly similar to H.sapiens mRNA for melanoma-associated chondroitin sulfate proteoglycan |
| AK125848.2 | cDNA FLJ43860 fis, clone TESTI4007404 |
| AK126026.12 | cDNA FLJ44038 fis, clone TESTI4028880, highly similar to Glucose transporter type 3, brain |
| AK126026.4 | cDNA FLJ44038 fis, clone TESTI4028880, highly similar to Glucose transporter type 3, brain |
| AK126026.5 | cDNA FLJ44038 fis, clone TESTI4028880, highly similar to Glucose transporter type 3, brain |
| AK126569.4 | cDNA FLJ44606 fis, clone BRACE2005991 |
| AK126766.4 | cDNA FLJ44813 fis, clone BRACE3044495, weakly similar to Homo sapiens GROS1-L protein |
| AK127030.2 | cDNA FLJ45086 fis, clone BRAWH3028796, highly similar to Homo sapiens makorin, ring finger protein, 1 (MKRN1) |
| AK127853.12 | cDNA FLJ45956 fis, clone PLACE7011072, highly similar to Vacuolar ATP synthase subunit B, kidney isoform (EC 3.6.1.34) |

| | |
|---|---|
| AK127853.13 | cDNA FLJ45956 fis, clone PLACE7011072, highly similar to Vacuolar ATP synthase subunit B, kidney isoform (EC 3.6.1.34) |
| AK128419.9 | cDNA FLJ46562 fis, clone THYMU3040172, highly similar to T-cell differentiation antigen CD6 precursor |
| AK131202.15 | cDNA FLJ16059 fis, clone TESTI2029252, weakly similar to Homo sapiens mRNA for LAK-4p |
| AK131446.24 | cDNA FLJ16586 fis, clone TESTI4000137 |
| AL079278.5 | mRNA full length insert cDNA clone EUROIMAGE 381801 |
| AL079296.9 | mRNA full length insert cDNA clone EUROIMAGE 609395 |
| AL136693.2 | mRNA; cDNA DKFZp564E227 (from clone DKFZp564E227); complete cds |
| AL525037.6 | NEUROBLASTOMA COT 25-NORMALIZED Homo sapiens cDNA clone CS0DC005YA21 5-PRIME, mRNA sequence |
| AL527437.5 | NEUROBLASTOMA COT 25-NORMALIZED Homo sapiens cDNA clone CS0DC021YL21 5-PRIME, mRNA sequence |
| AL832046.15 | cDNA DKFZp313I2410 (from clone DKFZp313I2410) |
| AL833938.2 | cDNA DKFZp586G2119 (from clone DKFZp586G2119) |
| AL833941.3 | cDNA DKFZp434D0316 (from clone DKFZp434D0316) |
| AY026895.4 | NREBP mRNA, complete cds |
| AY040554.4 | SLAM mRNA, complete cds |
| AY078985.4 | MTO1 protein isoform III mRNA, complete cds; nuclear gene for mitochondrial product |
| AY078986.4 | MTO1 protein isoform IV mRNA, complete cds; nuclear gene for mitochondrial product |
| AY147037.25 | MLL5 (MLL5) mRNA, complete cds; alternatively spliced |
| AY149920.21 | activated T-cell marker CD109 (CD109) mRNA, complete cds |
| AY189675.6 | dolichyl pyrophosphate phosphatase CWH8 mRNA, complete cds |
| AY238437.13 | inter-alpha trypsin inhibitor heavy chain precursor 5 (ITIH5) mRNA, complete cds |
| AY256823.10 | prestin isoform SLC26A5b (PRES) mRNA, complete cds |
| AY333285.22 | truncated transient receptor potential cation channel subfamily M member 6 variant a (TRPM6) mRNA, complete cds, alternatively spliced |
| AY532280.19 | G protein coupled receptor 133 (GPR133) mRNA, complete cds |
| AY557192.5 | growth hormone-releasing hormone receptor mRNA, complete cds |
| AY572488.14 | 200 kDa U5 snRNP-specific spliceosomal protein (BRR2) mRNA, complete cds |

| | |
|---|---|
| AY572806.5 | constitutive androstane receptor SV1 (NR1I3) mRNA, partial cds, alternatively spliced |
| AY572807.5 | constitutive androstane receptor SV2 (NR1I3) mRNA, complete cds, alternatively spliced |
| AY572819.5 | constitutive androstane receptor SV14 (NR1I3) mRNA, partial cds, alternatively spliced |
| AY601812.6 | natural killer cell immunoglobulin-like receptor (KIR2DL4) mRNA, complete cds |
| BC000133.12 | zinc finger protein 76 (expressed in testis), mRNA (cDNA clone MGC:5059 IMAGE:2900036), complete cds |
| BC000711.2 | translocase of inner mitochondrial membrane 8 homolog B (yeast), mRNA (cDNA clone MGC:1102 IMAGE:2823930), complete cds |
| BC001160.7 | LIM homeobox 3, mRNA (cDNA clone IMAGE:3354063), containing frame-shift errors |
| BC002549.12 | zinc finger protein 76 (expressed in testis), mRNA (cDNA clone MGC:1577 IMAGE:3139040), complete cds |
| BC005808.4 | mitochondrial translation optimization 1 homolog (S. cerevisiae), mRNA (cDNA clone IMAGE:3842973), partial cds |
| BC006112.2 | ATP-dependent glucokinase, mRNA (cDNA clone MGC:12975 IMAGE:3347312), complete cds |
| BC007363.10 | branched chain alpha-ketoacid dehydrogenase kinase, mRNA (cDNA clone MGC:16138 IMAGE:3630050), complete cds |
| BC009493.6 | dolichyl pyrophosphate phosphatase 1, mRNA (cDNA clone IMAGE:3938659), partial cds |
| BC011770.2 | FAST kinase, transcript variant 1, mRNA (cDNA clone MGC:19784 IMAGE:3831196), complete cds |
| BC012764.14 | RALBP1 associated Eps domain containing 1, mRNA (cDNA clone MGC:16228 IMAGE:3677286), complete cds |
| BC013831.4 | hypothetical protein MGC4189, mRNA (cDNA clone IMAGE:4149678) |
| BC013831.5 | hypothetical protein MGC4189, mRNA (cDNA clone IMAGE:4149678) |
| BC018771.5 | hypothetical protein PP1665, mRNA (cDNA clone MGC:32046 IMAGE:4845520), complete cds |
| BC021219.5 | dachshund homolog 1 (Drosophila), mRNA (cDNA clone IMAGE:4301720), partial cds |
| BC022553.6 | cDNA clone IMAGE:4819705, containing frame-shift errors |
| BC025304.3 | TcD37 homolog, mRNA (cDNA clone MGC:39274 IMAGE:5458046), complete cds |
| BC026080.7 | TUWD12, mRNA (cDNA clone MGC:26753 IMAGE:4827689), complete cds |
| BC027603.4 | KCCR13L, mRNA (cDNA clone MGC:26765 IMAGE:4830264), complete cds |

| | |
|---|---|
| BC033686.6 | dolichyl pyrophosphate phosphatase 1, mRNA (cDNA clone MGC:45389 IMAGE:5195127), complete cds |
| BC035373.17 | chloride channel Ka, mRNA (cDNA clone IMAGE:5172279), containing frame-shift errors |
| BC036349.12 | chromosome 9 open reading frame 68, mRNA (cDNA clone IMAGE:5168196), complete cds |
| BC037298.4 | FLJ44216 protein, mRNA (cDNA clone MGC:33586 IMAGE:4823991), complete cds |
| BC039815.13 | tumor protein p73-like, mRNA (cDNA clone MGC:48678 IMAGE:5552611), complete cds |
| BC040071.27 | alpha-2-macroglobulin, mRNA (cDNA clone MGC:47683 IMAGE:6056126), complete cds |
| BC040259.18 | erythrocyte membrane protein band 4.1-like 1, mRNA (cDNA clone MGC:34764 IMAGE:5195986), complete cds |
| BC041611.6 | killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 4, mRNA (cDNA clone MGC:52132 IMAGE:5485491), complete cds |
| BC043211.8 | Similar to G protein-coupled hepta-helical receptor Ig-Hepta, clone IMAGE:5295143, mRNA |
| BC044589.2 | ATP-dependent glucokinase, mRNA (cDNA clone IMAGE:5269974), partial cds |
| BC047296.3 | Similar to RIKEN cDNA 6720460I06 gene, clone IMAGE:4428383, mRNA |
| BC047421.5 | clone IMAGE:5311299, mRNA |
| BC050388.7 | similar to bK246H3.1 (immunoglobulin lambda-like polypeptide 1, pre-B-cell specific), mRNA (cDNA clone IMAGE:5728597), partial cds |
| BC051849.4 | hypothetical protein MGC4189, mRNA (cDNA clone MGC:60172 IMAGE:5493799), complete cds |
| BC051849.5 | hypothetical protein MGC4189, mRNA (cDNA clone MGC:60172 IMAGE:5493799), complete cds |
| BC052250.6 | solute carrier family 7 (cationic amino acid transporter, y+ system), member 8, transcript variant 1, mRNA (cDNA clone MGC:59794 IMAGE:6260264), complete cds |
| BC053320.4 | C-terminal binding protein 1, mRNA (cDNA clone MGC:59831 IMAGE:6461670), complete cds |
| BC054518.10 | hypothetical protein FLJ10211, mRNA (cDNA clone IMAGE:6501489), partial cds |
| BC056408.8 | dihydropyrimidinase-like 2, mRNA (cDNA clone MGC:64873 IMAGE:6044141), complete cds |
| BC060866.3 | hypothetical protein LOC163782, mRNA (cDNA clone MGC:71814 IMAGE:30344114), complete cds |
| BC063649.7 | chromosome 7 open reading frame 2, mRNA (cDNA clone IMAGE:4803998), partial cds |
| BC063825.3 | TRIAD3 protein, transcript variant 2, mRNA (cDNA clone MGC:74609 IMAGE:6504476), complete cds |

| | |
|---|---|
| BC064370.17 | PRP8 pre-mRNA processing factor 8 homolog (yeast), mRNA (cDNA clone MGC:74762 IMAGE:5587081), complete cds |
| BC065290.2 | cytochrome b reductase 1, mRNA (cDNA clone MGC:74716 IMAGE:6063145), complete cds |
| BC066592.16 | cut-like 1, CCAAT displacement protein (Drosophila), mRNA (cDNA clone MGC:75164 IMAGE:5740343), complete cds |
| BC067109.8 | dihydropyrimidinase-like 2, mRNA (cDNA clone MGC:70846 IMAGE:3870039), complete cds |
| BC068974.17 | splicing factor 3b, subunit 3, 130kDa, mRNA (cDNA clone MGC:74943 IMAGE:6168677), complete cds |
| BC069626.5 | nuclear receptor subfamily 1, group I, member 3, mRNA (cDNA clone MGC:97209 IMAGE:7262456), complete cds |
| BC070081.20 | cDNA clone IMAGE:30346915, containing frame-shift errors |
| BE207088.2 | ba10a10.y1 NIH_MGC_7 Homo sapiens cDNA clone IMAGE:2823930 5' similar to SW:DDP_HUMAN O60220 X-LINKED DEAFNESS DYSTONIA PROTEIN. [1] ;, mRNA sequence |
| BE251894.4 | 601107535F1 NIH_MGC_16 Homo sapiens cDNA clone IMAGE:3343856 5', mRNA sequence |
| BE251894.5 | 601107535F1 NIH_MGC_16 Homo sapiens cDNA clone IMAGE:3343856 5', mRNA sequence |
| BE257774.2 | 601116738F1 NIH_MGC_16 Homo sapiens cDNA clone IMAGE:3357440 5', mRNA sequence |
| BE259199.2 | 601109728F1 NIH_MGC_16 Homo sapiens cDNA clone IMAGE:3350455 5', mRNA sequence |
| BE385121.5 | 601275187F1 NIH_MGC_20 Homo sapiens cDNA clone IMAGE:3616317 5', mRNA sequence |
| BE736596.2 | 601306301F1 NIH_MGC_39 Homo sapiens cDNA clone IMAGE:3640585 5', mRNA sequence |
| BE795191.2 | 601592447F1 NIH_MGC_7 Homo sapiens cDNA clone IMAGE:3946571 5', mRNA sequence |
| BE797157.2 | 601587161F1 NIH_MGC_7 Homo sapiens cDNA clone IMAGE:3941463 5', mRNA sequence |
| BE891464.3 | 601435329F1 NIH_MGC_72 Homo sapiens cDNA clone IMAGE:3920358 5', mRNA sequence |
| BE904870.5 | 601496630F1 NIH_MGC_70 Homo sapiens cDNA clone IMAGE:3898878 5', mRNA sequence |
| BF794505.2 | 602255757F1 NIH_MGC_85 Homo sapiens cDNA clone IMAGE:4339094 5', mRNA sequence |
| BF981843.3 | 602309628F1 NIH_MGC_88 Homo sapiens cDNA clone IMAGE:4400694 5', mRNA sequence |
| BG475390.5 | 602491929F1 NIH_MGC_20 Homo sapiens cDNA clone IMAGE:4623078 5', mRNA sequence |
| BG493175.3 | 602541857F1 NIH_MGC_59 Homo sapiens cDNA clone IMAGE:4672790 5', mRNA sequence |

| | |
|---|---|
| BG696005.4 | 602658151F1 NCI_CGAP_Skn3 Homo sapiens cDNA clone IMAGE:4800780 5', mRNA sequence |
| BG704952.2 | 602688932F1 NIH_MGC_95 Homo sapiens cDNA clone IMAGE:4821366 5', mRNA sequence |
| BG705743.3 | 602668979F1 NIH_MGC_96 Homo sapiens cDNA clone IMAGE:4791762 5', mRNA sequence |
| BG705753.2 | 602668992F1 NIH_MGC_96 Homo sapiens cDNA clone IMAGE:4791622 5', mRNA sequence |
| BG722435.2 | 602693661F1 NIH_MGC_97 Homo sapiens cDNA clone IMAGE:4825839 5', mRNA sequence |
| BG825711.6 | 602747021F2 NIH_MGC_17 Homo sapiens cDNA clone IMAGE:4899941 5', mRNA sequence |
| BI560844.5 | 603254010F1 NIH_MGC_97 Homo sapiens cDNA clone IMAGE:5296442 5', mRNA sequence |
| BI753928.2 | 603022935F1 NIH_MGC_114 Homo sapiens cDNA clone IMAGE:5193584 5', mRNA sequence |
| BI758369.4 | 603025836F1 NIH_MGC_114 Homo sapiens cDNA clone IMAGE:5196321 5', mRNA sequence |
| BI772282.3 | 603055864F1 NIH_MGC_122 Homo sapiens cDNA clone IMAGE:5205374 5', mRNA sequence |
| BM451292.4 | AGENCOURT_6390834 NIH_MGC_67 Homo sapiens cDNA clone IMAGE:5493799 5', mRNA sequence |
| BM451292.5 | AGENCOURT_6390834 NIH_MGC_67 Homo sapiens cDNA clone IMAGE:5493799 5', mRNA sequence |
| BM545926.2 | AGENCOURT_6497810 NIH_MGC_125 Homo sapiens cDNA clone IMAGE:5588307 5', mRNA sequence |
| BM545926.3 | AGENCOURT_6497810 NIH_MGC_125 Homo sapiens cDNA clone IMAGE:5588307 5', mRNA sequence |
| BM684825.3 | UI-E-EJ1-ajj-l-01-0-UI.s1 UI-E-EJ1 Homo sapiens cDNA clone UI-E-EJ1-ajj-l-01-0-UI 3', mRNA sequence |
| BM747300.2 | K-EST0021829 S3SNU16s1 Homo sapiens cDNA clone S3SNU16s1-2-E11 5', mRNA sequence |
| BM839918.5 | K-EST0116904 S13KMS5 Homo sapiens cDNA clone S13KMS5-31-G08 5', mRNA sequence |
| BQ051323.2 | AGENCOURT_6763708 NIH_MGC_68 Homo sapiens cDNA clone IMAGE:5792711 5', mRNA sequence |
| BQ231901.2 | AGENCOURT_7554671 NIH_MGC_72 Homo sapiens cDNA clone IMAGE:6063145 5', mRNA sequence |
| BQ423777.2 | AGENCOURT_7805574 NIH_MGC_92 Homo sapiens cDNA clone IMAGE:6068214 5', mRNA sequence |
| BQ719021.2 | AGENCOURT_8109235 Lupski_sympathetic_trunk Homo sapiens cDNA clone IMAGE:6189375 5', mRNA sequence |
| BQ919271.4 | AGENCOURT_8794262 NIH_MGC_18 Homo sapiens cDNA clone IMAGE:6370682 5', mRNA sequence |
| BQ919271.5 | AGENCOURT_8794262 NIH_MGC_18 Homo sapiens cDNA clone IMAGE:6370682 5', mRNA sequence |
| BU158615.2 | AGENCOURT_7968610 NIH_MGC_68 Homo sapiens cDNA clone IMAGE:6011181 5', mRNA sequence |

| | |
|---|---|
| BU176606.2 | AGENCOURT_7825223 NIH_MGC_67 Homo sapiens cDNA clone IMAGE:6142965 5', mRNA sequence |
| BU500170.2 | AGENCOURT_7859851 NIH_MGC_64 Homo sapiens cDNA clone IMAGE:6108978 5', mRNA sequence |
| BU528662.3 | AGENCOURT_10183145 NIH_MGC_101 Homo sapiens cDNA clone IMAGE:6538626 5', mRNA sequence |
| BU541379.2 | AGENCOURT_10333359 NIH_MGC_40 Homo sapiens cDNA clone IMAGE:6573052 5', mRNA sequence |
| BU655927.5 | cl04b12.z1 Hembase; Erythroid Precursor Cells (LCB:cl library) Homo sapiens cDNA clone cl04b12 5', mRNA sequence |
| BU656904.3 | cl17f06.z1 Hembase; Erythroid Precursor Cells (LCB:cl library) Homo sapiens cDNA clone cl17f06 5', mRNA sequence |
| BU659919.5 | cl53f03.z1 Hembase; Erythroid Precursor Cells (LCB:cl library) Homo sapiens cDNA clone cl53f03 5', mRNA sequence |
| BU853397.5 | AGENCOURT_10417826 NIH_MGC_82 Homo sapiens cDNA clone IMAGE:6620233 5', mRNA sequence |
| BX095212.5 | BX095212 Soares fetal liver spleen 1NFLS Homo sapiens cDNA clone IMAGp998A19654, mRNA sequence |
| BX100703.4 | BX100703 Soares_parathyroid_tumor_NbHPA Homo sapiens cDNA clone IMAGp998D21684, mRNA sequence |
| BX330997.4 | BX330997 Homo sapiens HELA CELLS COT 25-NORMALIZED Homo sapiens cDNA clone CS0DK005YM04 3-PRIME, mRNA sequence |
| BX330997.5 | BX330997 Homo sapiens HELA CELLS COT 25-NORMALIZED Homo sapiens cDNA clone CS0DK005YM04 3-PRIME, mRNA sequence |
| BX335807.3 | BX335807 Homo sapiens PLACENTA COT 25-NORMALIZED Homo sapiens cDNA clone CS0DI020YE24 5-PRIME, mRNA sequence |
| BX335807.4 | BX335807 Homo sapiens PLACENTA COT 25-NORMALIZED Homo sapiens cDNA clone CS0DI020YE24 5-PRIME, mRNA sequence |
| BX353062.2 | BX353062 Homo sapiens NEUROBLASTOMA COT 10-NORMALIZED Homo sapiens cDNA clone CS0DB008YL06 5-PRIME, mRNA sequence |
| BX362900.4 | BX362900 Homo sapiens HELA CELLS COT 25-NORMALIZED Homo sapiens cDNA clone CS0DK005YM04 5-PRIME, mRNA sequence |
| BX362900.5 | BX362900 Homo sapiens HELA CELLS COT 25-NORMALIZED Homo sapiens cDNA clone CS0DK005YM04 5-PRIME, mRNA sequence |
| BX367559.5 | BX367559 Homo sapiens NEUROBLASTOMA COT 10-NORMALIZED Homo sapiens cDNA clone CS0DB001YH22 5-PRIME, mRNA sequence |

| | |
|---|---|
| BX374377.5 | BX374377 Homo sapiens NEUROBLASTOMA COT 10-NORMALIZED Homo sapiens cDNA clone CS0DB001YH22 5-PRIME, mRNA sequence |
| BX398157.3 | BX398157 Homo sapiens PLACENTA COT 25-NORMALIZED Homo sapiens cDNA clone CS0DI053YH10 5-PRIME, mRNA sequence |
| BX398157.4 | BX398157 Homo sapiens PLACENTA COT 25-NORMALIZED Homo sapiens cDNA clone CS0DI053YH10 5-PRIME, mRNA sequence |
| BX441333.1 | BX441333 Homo sapiens FETAL BRAIN Homo sapiens cDNA clone CS0DF016YN22 5-PRIME, mRNA sequence |
| BX460119.6 | BX460119 Homo sapiens FETAL BRAIN Homo sapiens cDNA clone CS0DF011YC02 5-PRIME, mRNA sequence |
| BX537389.19 | mRNA; cDNA DKFZp779G2251 (from clone DKFZp779G2251); complete cds |
| BX537406.3 | mRNA; cDNA DKFZp686B0852 (from clone DKFZp686B0852); complete cds |
| BX537590.26 | mRNA; cDNA DKFZp686I155 (from clone DKFZp686I155) |
| BX537590.28 | mRNA; cDNA DKFZp686I155 (from clone DKFZp686I155) |
| BX537590.41 | mRNA; cDNA DKFZp686I155 (from clone DKFZp686I155) |
| BX537784.19 | mRNA; cDNA DKFZp779O2152 (from clone DKFZp779O2152); complete cds |
| BX537849.19 | mRNA; cDNA DKFZp686E1588 (from clone DKFZp686E1588); complete cds |
| BX537849.25 | mRNA; cDNA DKFZp686E1588 (from clone DKFZp686E1588); complete cds |
| BX538166.19 | mRNA; cDNA DKFZp686E18109 (from clone DKFZp686E18109); complete cds |
| BX538166.25 | mRNA; cDNA DKFZp686E18109 (from clone DKFZp686E18109); complete cds |
| BX538181.7 | mRNA; cDNA DKFZp686K15188 (from clone DKFZp686K15188) |
| BX647115.8 | mRNA; cDNA DKFZp686B14144 (from clone DKFZp686B14144) |
| BX647366.19 | mRNA; cDNA DKFZp779N0852 (from clone DKFZp779N0852) |
| BX647367.19 | mRNA; cDNA DKFZp779A1459 (from clone DKFZp779A1459) |
| BX647691.1 | mRNA; cDNA DKFZp686E07118 (from clone DKFZp686E07118) |
| BX647873.15 | mRNA; cDNA DKFZp313O0321 (from clone DKFZp313O0321) |
| BX648035.26 | mRNA; cDNA DKFZp686P0286 (from clone DKFZp686P0286) |
| BX648544.19 | mRNA; cDNA DKFZp779D1555 (from clone DKFZp779D1555) |

| | |
|---|---|
| BX648732.15 | mRNA; cDNA DKFZp686E0628 (from clone DKFZp686E0628) |
| BX648909.2 | mRNA; cDNA DKFZp686G13268 (from clone DKFZp686G13268) |
| BX649072.19 | mRNA; cDNA DKFZp779D1050 (from clone DKFZp779D1050) |
| BX649159.2 | mRNA; cDNA DKFZp686F0129 (from clone DKFZp686F0129) |
| CA454130.2 | AGENCOURT_10738534 MAPcL Homo sapiens cDNA clone IMAGE:6718791 5', mRNA sequence |
| CA455101.2 | AGENCOURT_10735752 MAPcL Homo sapiens cDNA clone IMAGE:6722619 5', mRNA sequence |
| CB133422.3 | K-EST0184328 L12JSHC0s1 Homo sapiens cDNA clone L12JSHC0s1-1-F05 5', mRNA sequence |
| CD014040.13 | 90115366 Single gene library Homo sapiens cDNA, mRNA sequence |
| CD014044.13 | 90115482 Single gene library Homo sapiens cDNA, mRNA sequence |
| CD014122.13 | 90115350 Single gene library Homo sapiens cDNA, mRNA sequence |
| CD632310.2 | 56072143H1 FLPRSV Homo sapiens cDNA, mRNA sequence |
| CD694367.3 | EST10890 human nasopharynx Homo sapiens cDNA, mRNA sequence |
| CF129958.2 | UI-HF-ES0-avz-p-15-0-UI.r1 NIH_MGC_213 Homo sapiens cDNA clone IMAGE:30561686 5', mRNA sequence |
| CN262434.4 | 17000600176961 GRN_PRENEU Homo sapiens cDNA 5', mRNA sequence |
| CN262434.5 | 17000600176961 GRN_PRENEU Homo sapiens cDNA 5', mRNA sequence |
| CN289787.5 | 17000423762706 GRN_EB Homo sapiens cDNA 5', mRNA sequence |
| CN357165.2 | 17000424180130 GRN_ES Homo sapiens cDNA 5', mRNA sequence |
| CN367895.2 | 17000455125353 GRN_EB Homo sapiens cDNA 5', mRNA sequence |
| CN481392.2 | hw07b05.y1 Human primary human ocular pericytes. Unamplified (hw) Homo sapiens cDNA clone hw07b05 5', mRNA sequence |
| CN482999.3 | hw27c07.y1 Human primary human ocular pericytes. Unamplified (hw) Homo sapiens cDNA clone hw27c07 5', mRNA sequence |
| CR590430.4 | full-length cDNA clone CS0DI065YO02 of Placenta Cot 25-normalized of Homo sapiens (human) |
| CR590567.10 | full-length cDNA clone CS0DB003YG07 of Neuroblastoma Cot 10-normalized of Homo sapiens (human) |
| CR590845.2 | full-length cDNA clone CS0DK003YB10 of HeLa cells Cot 25-normalized of Homo sapiens (human) |

| | |
|---|---|
| CR593551.10 | full-length cDNA clone CS0DA004YC14 of Neuroblastoma of Homo sapiens (human) |
| CR595036.2 | full-length cDNA clone CS0DB002YB21 of Neuroblastoma Cot 10-normalized of Homo sapiens (human) |
| CR597108.3 | full-length cDNA clone CS0DI020YE24 of Placenta Cot 25-normalized of Homo sapiens (human) |
| CR597108.4 | full-length cDNA clone CS0DI020YE24 of Placenta Cot 25-normalized of Homo sapiens (human) |
| CR597531.2 | full-length cDNA clone CS0DI008YN22 of Placenta Cot 25-normalized of Homo sapiens (human) |
| CR598131.7 | full-length cDNA clone CS0DK010YC14 of HeLa cells Cot 25-normalized of Homo sapiens (human) |
| CR598553.10 | full-length cDNA clone CS0DA009YH03 of Neuroblastoma of Homo sapiens (human) |
| CR600084.10 | full-length cDNA clone CS0DI029YM17 of Placenta Cot 25-normalized of Homo sapiens (human) |
| CR601282.2 | full-length cDNA clone CS0DC015YO12 of Neuroblastoma Cot 25-normalized of Homo sapiens (human) |
| CR605756.4 | full-length cDNA clone CS0DK005YM04 of HeLa cells Cot 25-normalized of Homo sapiens (human) |
| CR605756.5 | full-length cDNA clone CS0DK005YM04 of HeLa cells Cot 25-normalized of Homo sapiens (human) |
| CR606698.3 | full-length cDNA clone CS0DF004YA19 of Fetal brain of Homo sapiens (human) |
| CR606750.5 | full-length cDNA clone CS0DJ014YK18 of T cells (Jurkat cell line) Cot 10-normalized of Homo sapiens (human) |
| CR607075.10 | full-length cDNA clone CS0DI034YK18 of Placenta Cot 25-normalized of Homo sapiens (human) |
| CR607185.6 | full-length cDNA clone CS0DK007YF19 of HeLa cells Cot 25-normalized of Homo sapiens (human |
| CR609253.5 | full-length cDNA clone CS0DC002YM14 of Neuroblastoma Cot 25-normalized of Homo sapiens (human) |
| CR610635.10 | full-length cDNA clone CS0DI085YK18 of Placenta Cot 25-normalized of Homo sapiens (human) |
| CR611357.7 | full-length cDNA clone CS0DL009YD13 of B cells (Ramos cell line) Cot 25-normalized of Homo sapiens (human) |
| CR612429.15 | full-length cDNA clone CS0DI042YI18 of Placenta Cot 25-normalized of Homo sapiens (human) |
| CR612524.10 | full-length cDNA clone CS0DI041YC02 of Placenta Cot 25-normalized of Homo sapiens (human) |
| CR613328.10 | full-length cDNA clone CS0DL005YO09 of B cells (Ramos cell line) Cot 25-normalized of Homo sapiens (human) |
| CR614544.6 | full-length cDNA clone CS0DL001YG03 of B cells (Ramos cell line) Cot 25-normalized of Homo sapiens (human) |
| CR619065.10 | full-length cDNA clone CS0DK002YM17 of HeLa cells Cot 25-normalized of Homo sapiens (human) |
| CR620199.2 | full-length cDNA clone CS0DN001YM15 of Adult brain of Homo sapiens (human) |

| | |
|---|---|
| CR621708.5 | full-length cDNA clone CS0DB001YH22 of Neuroblastoma Cot 10-normalized of Homo sapiens (human) |
| CR623465.2 | full-length cDNA clone CS0DM003YB16 of Fetal liver of Homo sapiens (human) |
| CR624926.7 | full-length cDNA clone CS0DK002YO07 of HeLa cells Cot 25-normalized of Homo sapiens (human) |
| CR626330.10 | full-length cDNA clone CS0DI015YK13 of Placenta Cot 25-normalized of Homo sapiens (human) |
| D28475.16 | KIAA0046 mRNA, partial cds |
| D78013.8 | mRNA for dihydropyrimidinase related protein-2, complete cds. |
| D87328.3 | mRNA for HCS, complete cds. |
| D87686.17 | mRNA for KIAA0017 protein, partial cds |
| D88308.4 | mRNA for very-long-chain acyl-CoA synthetase, complete cds. |
| D90228.5 | mRNA for mitochondrial acetoacetyl-CoA thiolase precursor, complete cds |
| L01406.5 | Human growth hormone-releasing hormone receptor mRNA, complete cds. |
| L05628.31 | Human multidrug resistance-associated protein (MRP) mRNA, complete cds |
| L09237.5 | growth hormone releasing hormone receptor, human, G-protein coupled receptor, secretin family |
| L11695.3 | Human activin receptor-like kinase (ALK-5) mRNA, complete cds. |
| L25876.5 | protein tyrosine phosphatase (CIP2)mRNA, complete cds. |
| L27711.5 | Human protein phosphatase (KAP1) mRNA, complete cds. |
| M11313.27 | Human alpha-2-macroglobulin mRNA, complete cds. |
| M18533.6 | Homo sapiens dystrophin (DMD) mRNA, complete cds. |
| M74099.16 | Human displacement protein (CCAAT) mRNA. |
| M77198.13 | Homo sapiens rac protein kinase-beta mRNA, complete cds. |
| M96652.3 | Human interleukin 5 receptor alpha-subunit (IL5R) mRNA |
| NM_000014.27 | alpha-2-macroglobulin (A2M), mRNA |
| NM_000019.5 | acetyl-Coenzyme A acetyltransferase 1 (acetoacetyl Coenzyme A thiolase) (ACAT1), nuclear gene encoding mitochondrial protein, mRNA |
| NM_000053.12 | ATPase, Cu++ transporting, beta polypeptide (Wilson disease) (ATP7B), mRNA |
| NM_000107.7 | damage-specific DNA binding protein 2, 48kDa (DDB2), mRNA |
| NM_000109.6 | dystrophin (muscular dystrophy, Duchenne and Becker types) (DMD), transcript variant Dp427c, mRNA |
| NM_000564.3 | interleukin 5 receptor, alpha (IL5RA), transcript variant 1, mRNA |

| | |
|---|---|
| NM_000610.9 | CD44 antigen (homing function and Indian blood group system) (CD44), transcript variant 1, mRNA |
| NM_000823.5 | growth hormone releasing hormone receptor (GHRHR), mRNA |
| NM_000856.7 | guanylate cyclase 1, soluble, alpha 3 (GUCY1A3), mRNA |
| NM_000856.8 | guanylate cyclase 1, soluble, alpha 3 (GUCY1A3), mRNA |
| NM_000950.2 | proline rich Gla (G-carboxyglutamic acid) 1 (PRRG1), mRNA |
| NM_001148.38 | ankyrin 2, neuronal (ANK2), transcript variant 1, mRNA |
| NM_001286.16 | chloride channel 6 (CLCN6), transcript variant ClC-6a, mRNA |
| NM_001386.8 | dihydropyrimidinase-like 2 (DPYSL2), mRNA |
| NM_001626.13 | v-akt murine thymoma viral oncogene homolog 2 (AKT2), mRNA |
| NM_002255.6 | killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 4 (KIR2DL4), mRNA |
| NM_002445.9 | macrophage scavenger receptor 1 (MSR1), transcript variant SR-AII, mRNA |
| NM_003037.4 | signaling lymphocytic activation molecule family member 1 (SLAMF1), mRNA |
| NM_003103.4 | SON DNA binding protein (SON), transcript variant g, mRNA |
| NM_003427.12 | zinc finger protein 76 (expressed in testis) (ZNF76), mRNA |
| NM_003494.54 | dysferlin, limb girdle muscular dystrophy 2B (autosomal recessive) (DYSF), mRNA |
| NM_003722.13 | tumor protein p73-like (TP73L), mRNA |
| NM_004006.6 | dystrophin (muscular dystrophy, Duchenne and Becker types) (DMD), transcript variant Dp427m, mRNA |
| NM_004009.6 | dystrophin (muscular dystrophy, Duchenne and Becker types) (DMD), transcript variant Dp427p1, mRNA |
| NM_004010.6 | dystrophin (muscular dystrophy, Duchenne and Becker types) (DMD), transcript variant Dp427p2, mRNA |
| NM_004043.7 | acetylserotonin O-methyltransferase (ASMT), mRNA |
| NM_004320.19 | ATPase, Ca++ transporting, cardiac muscle, fast twitch 1 (ATP2A1), transcript variant b, mRNA |
| NM_004385.7 | chondroitin sulfate proteoglycan 2 (versican) (CSPG2), mRNA |
| NM_004612.3 | transforming growth factor, beta receptor I (activin A receptor type II-like kinase, 53kDa) (TGFBR1), mRNA |
| NM_004977.2 | potassium voltage-gated channel, Shaw-related subfamily, member 3 (KCNC3), mRNA |
| NM_004996.31 | ATP-binding cassette, sub-family C (CFTR |
| NM_005106.11 | deleted in lung and esophageal cancer 1 (DLEC1), transcript variant DLEC1-N1, mRNA |

| | |
|---|---|
| NM_005122.5 | nuclear receptor subfamily 1, group I, member 3 (NR1I3), mRNA |
| NM_005192.5 | cyclin-dependent kinase inhibitor 3 (CDK2-associated dual specificity phosphatase) (CDKN3), mRNA |
| NM_005228.20 | epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian) (EGFR), transcript variant 1, mRNA |
| NM_005881.10 | branched chain alpha-ketoacid dehydrogenase kinase (BCKDK), mRNA |
| NM_006445.17 | PRP8 pre-mRNA processing factor 8 homolog (yeast) (PRPF8), mRNA |
| NM_006712.2 | FAST kinase (FASTK), transcript variant 1, mRNA |
| NM_006910.17 | retinoblastoma binding protein 6 (RBBP6), transcript variant 1, mRNA |
| NM_012244.6 | solute carrier family 7 (cationic amino acid transporter, y+ system), member 8 (SLC7A8), transcript variant 1, mRNA |
| NM_012307.13 | erythrocyte membrane protein band 4.1-like 3 (EPB41L3), mRNA |
| NM_012307.19 | erythrocyte membrane protein band 4.1-like 3 (EPB41L3), mRNA |
| NM_012426.17 | splicing factor 3b, subunit 3, 130kDa (SF3B3), mRNA |
| NM_013229.15 | apoptotic protease activating factor (APAF1), transcript variant 1, mRNA |
| NM_013440.14 | paired immunoglobin-like type 2 receptor beta (PILRB), transcript variant 1, mRNA |
| NM_014014.14 | U5 snRNP-specific protein, 200-KD (U5-200KD), mRNA |
| NM_015046.19 | senataxin (KIAA0625), mRNA |
| NM_015046.25 | senataxin (KIAA0625), mRNA |
| NM_015542.11 | UPF2 regulator of nonsense transcripts homolog (yeast) (UPF2), transcript variant 2, mRNA |
| NM_015678.30 | neurobeachin (NBEA), mRNA |
| NM_016210.2 | g20 protein (LOC51161), mRNA |
| NM_018125.13 | hypothetical protein FLJ10521 (FLJ10521), mRNA |
| NM_020438.6 | dolichyl pyrophosphate phosphatase 1 (DOLPP1), mRNA |
| NM_020859.7 | Shroom-related protein (ShrmL), mRNA |
| NM_020973.3 | glucosidase, beta, acid 3 (cytosolic) (GBA3), mRNA |
| NM_020987.37 | ankyrin 3, node of Ranvier (ankyrin G) (ANK3), transcript variant 1, mRNA |
| NM_021222.3 | TcD37 homolog (HTCD37), mRNA |
| NM_021907.18 | dystrobrevin, beta (DTNB), transcript variant 1, mRNA |
| NM_022648.28 | tensin (TNS), mRNA |
| NM_024091.2 | hypothetical protein MGC5297 (MGC5297), mRNA |
| NM_024345.5 | hypothetical protein MGC10765 (MGC10765), mRNA |

| | |
|---|---|
| NM_024505.10 | NADPH oxidase, EF hand calcium-binding domain 5 (NOX5), mRNA |
| NM_024611.15 | NMDA receptor-regulated gene 2 (NARG2), mRNA |
| NM_024744.5 | amyotrophic lateral sclerosis 2 (juvenile) chromosome region, candidate 8 (ALS2CR8), mRNA |
| NM_024843.2 | cytochrome b reductase 1 (CYBRD1), mRNA |
| NM_030569.13 | inter-alpha (globulin) inhibitor H5 (ITIH5), transcript variant 1, mRNA |
| NM_031284.2 | ATP-dependent glucokinase (ADP-GK), mRNA |
| NM_031846.7 | microtubule-associated protein 2 (MAP2), transcript variant 3, mRNA |
| NM_031922.14 | Homo sapiens RALBP1 associated Eps domain containing 1 (REPS1), mRNA |
| NM_032217.26 | ankyrin repeat domain 17 (ANKRD17), transcript variant 1, mRNA |
| NM_032523.19 | oxysterol binding protein-like 6 (OSBPL6), transcript variant 1, mRNA |
| NM_058183.3 | SON DNA binding protein (SON), transcript variant e, mRNA |
| NM_080599.11 | UPF2 regulator of nonsense transcripts homolog (yeast) (UPF2), transcript variant 1, mRNA |
| NM_080759.5 | dachshund homolog 1 (Drosophila) (DACH1), transcript variant 1, mRNA |
| NM_130854.12 | protein tyrosine phosphatase, receptor type, S (PTPRS), transcript variant 2, mRNA |
| NM_133338.5 | RAD17 homolog (S. pombe) (RAD17), transcript variant 1, mRNA |
| NM_133493.21 | CD109 antigen (Gov platelet alloantigens) (CD109), mRNA |
| NM_133645.4 | mitochondrial translation optimization 1 homolog (S. cerevisiae) (MTO1), mRNA |
| NM_138576.3 | B-cell CLL |
| NM_138925.4 | SON DNA binding protein (SON), transcript variant a, mRNA |
| NM_138926.3 | SON DNA binding protein (SON), transcript variant c, mRNA |
| NM_139179.4 | KCCR13L (LOC221955), mRNA |
| NM_144601.3 | chemokine-like factor super family 3 (CKLFSF3), transcript variant 1, mRNA |
| NM_144601.4 | chemokine-like factor super family 3 (CKLFSF3), transcript variant 1, mRNA |
| NM_144963.6 | hypothetical protein FLJ23790 (FLJ23790), mRNA |
| NM_145003.4 | hypothetical protein FLJ31164 (FLJ31164), mRNA |
| NM_145316.2 | chromosome 6 open reading frame 128 (C6orf128), mRNA |

| | |
|---|---|
| NM_145858.12 | crystallin, zeta (quinone reductase)-like 1 (CRYZL1), transcript variant 3, mRNA |
| NM_152788.3 | E2a-Pbx1-associated protein (EB-1), transcript variant 1, mRNA |
| NM_153208.7 | hypothetical protein MGC35048 (MGC35048), mRNA |
| NM_170674.1 | Meis1, myeloid ecotropic viral integration site 1 homolog 2 (mouse) (MEIS2), transcript variant b, mRNA |
| NM_170677.1 | Meis1, myeloid ecotropic viral integration site 1 homolog 2 (mouse) (MEIS2), transcript variant a, mRNA |
| NM_172240.7 | TUWD12 (TUWD12), mRNA |
| NM_174873.2 | purinergic receptor P2X, ligand-gated ion channel, 2 (P2RX2), transcript variant 2, mRNA |
| NM_177559.2 | casein kinase 2, alpha 1 polypeptide (CSNK2A1), transcript variant 1, mRNA |
| NM_181287.2 | chemokine-like factor super family 1 (CKLFSF1), transcript variant 8, mRNA |
| NM_181288.2 | chemokine-like factor super family 1 (CKLFSF1), transcript variant 9, mRNA |
| NM_181298.2 | chemokine-like factor super family 1 (CKLFSF1), transcript variant 20, mRNA |
| NM_181299.2 | chemokine-like factor super family 1 (CKLFSF1), transcript variant 21, mRNA |
| NM_181552.16 | cut-like 1, CCAAT displacement protein (Drosophila) (CUTL1), transcript variant 1, mRNA |
| NM_181554.3 | chemokine-like factor super family 3 (CKLFSF3), transcript variant 3, mRNA |
| NM_181554.4 | chemokine-like factor super family 3 (CKLFSF3), transcript variant 3, mRNA |
| NM_181712.3 | hypothetical protein LOC163782 (LOC163782), mRNA |
| NM_181861.15 | apoptotic protease activating factor (APAF1), transcript variant 3, mRNA |
| NM_182931.25 | myeloid |
| NM_182964.21 | Neuron navigator 2 (NAV2), transcript variant 1, mRNA |
| NM_198567.4 | FLJ44216 protein (FLJ44216), mRNA |
| NM_198827.19 | G protein-coupled receptor 133; G-protein coupled receptor GPR133 [Homo sapiens], mRNA sequence |
| NM_198999.10 | prestin (motor protein) (PRES), transcript variant a, mRNA |
| NM_206883.10 | prestin (motor protein) (PRES), transcript variant b, mRNA |
| NM_207111.3 | TRIAD3 protein (TRIAD3), transcript variant 1, mRNA |
| NM_207116.3 | TRIAD3 protein (TRIAD3), transcript variant 2, mRNA |
| NM_207414.2 | FLJ43860 protein (FLJ43860), mRNA |
| NM_212475.26 | fibronectin 1 (FN1), transcript variant 2, mRNA |
| NM_212475.28 | fibronectin 1 (FN1), transcript variant 2, mRNA |

| | |
|---|---|
| NM_212475.41 | fibronectin 1 (FN1), transcript variant 2, mRNA |
| NM_213653.3 | Homo sapiens hemochromatosis type 2 (juvenile) (HFE2), transcript variant a, mRNA |
| U02681.5 | cyclin-dependent kinase interactor 1; Cdi1 [Homo sapiens], mRNA sequence |
| U03464.12 | Human P-type ATPase ATP7B mRNA, complete cds |
| U11090.7 | Human hydroxyindole-O-methyltransferase promoter A-derived (HIOMT) mRNA, complete cds. |
| U11700.12 | Human copper transporting ATPase mRNA, complete cds. |
| U13616.37 | ankyrin G (ANK-3) mRNA, complete cds. |
| U16306.7 | Human chondroitin sulfate proteoglycan versican V0 splice-variant precursor peptide mRNA, complete cds. |
| U17279.8 | Human collapsin response mediator protein hCRMP-2 mRNA, complete cds. |
| U18300.7 | damage-specific DNA binding protein p48 subunit (DDB2) mRNA, complete cds. |
| U27327.2 | alpha (1,3 |
| U33017.4 | signaling lymphocytic activation molecule (SLAM) mRNA, complete cds. |
| U40317.12 | Human protein tyrosine phosphatase PTPsigma (PTPsigma) mRNA, complete cds. |
| U71199.6 | Human natural killer cell receptor (KIR) mRNA, complete cds |
| U94780.19 | Human meningioma-expressed antigen 6 (MEA6) mRNA, complete cds |
| X00588.20 | Human mRNA for precursor of epidermal growth factor receptor |
| X66534.7 | soluble guanylate cyclase large subunit mRNA |
| X66534.8 | soluble guanylate cyclase large subunit mRNA |
| X83378.16 | mRNA for putative chloride channe |
| X86779.2 | mRNA for FAST kinase |
| XM_035572.15 | chromosome 4 open reading frame 9 (C4orf9), mRNA |
| XM_045581.24 | likely ortholog of mouse 5-azacytidine induced gene 1 (AZI1), mRNA |
| XM_113947.5 | KIAA0565 gene product (KIAA0565), mRNA |
| XM_290820.10 | hypothetical protein FLJ10211 (FLJ10211), mRNA |
| XM_291015.25 | likely homolog of rat kinase D-interacting substance of 220 kDa (KIDINS220), mRNA |
| XM_291141.17 | KIAA0303 protein (KIAA0303), mRNA |
| XM_371114.11 | formin homology 2 domain containing 3 (FHOD3), mRNA |
| XM_371116.22 | myosin VB (MYO5B), mRNA |
| XM_371561.5 | similar to CDNA sequence BC043098 (LOC389039), mRNA |

| | |
|---|---|
| XM_371822.20 | chromosome 6 open reading frame 110 (C6orf110), mRNA |
| XM_372060.2 | similar to FLJ46354 protein (LOC389690), mRNA |
| XM_374388.9 | similar to KIAA1691 protein (LOC392619), mRNA |
| XM_376158.2 | hypothetical gene supported by AK126104; BX648733 (LOC401024), mRNA |
| XM_376593.10 | similar to RIKEN cDNA 9330128H10 gene (LOC401323), mRNA |
| XM_376593.5 | similar to RIKEN cDNA 9330128H10 gene (LOC401323), mRNA |
| XM_378044.14 | similar to hypothetical protein (LOC402354), mRNA |
| XM_379807.10 | similar to RIKEN cDNA 9330128H10 gene (LOC402479), mRNA |
| XM_379807.5 | similar to RIKEN cDNA 9330128H10 gene (LOC402479), mRNA |
| Y15723.7 | mRNA for soluble guanylyl cyclase |
| Y15723.8 | mRNA for soluble guanylyl cyclase |
| Y16961.13 | mRNA for KET protein |
| Y18483.6 | mRNA for SLC7A8 protein |
| Z26634.38 | mRNA for ankyrin B (440 kDa) |

# APPENDIX C

## LIST OF MOUSE TARGET GENES PREDICTED BY NTHunter (FROM CHAPTER 5)

| GenBank ID<br>Exon No | Annotation |
|---|---|
| AB036749.3 | Mporc-d mRNA for porcupine-D, complete cds |
| AB041547.10 | brain cDNA, clone MNCb-3763, similar to AC004410 fos39554_1 (Homo sapiens) |
| AB064265.8 | mRNA for RA175, complete cds |
| AF017112.16 | non-erythrocyte beta spectrin mRNA, complete cds |
| AF075436.13 | TA*p63 alpha mRNA, complete cds |
| AF098967.4 | interleukin enhancer binding factor 3 (Ilf3) mRNA, complete cds |
| AF116847.4 | bright and dead ringer gene product homologous protein Bdp mRNA, complete cds |
| AF144095.2 | unconventional myosin-15 mRNA, complete cds |
| AF167568.1 | mu opioid receptor variant F mRNA, complete cds |
| AF167568.3 | mu opioid receptor variant F mRNA, complete cds |
| AF226656.3 | peroxisomal N1-acetyl-spermine |
| AF247654.15 | betaPix-b mRNA, complete cds |
| AF323958.9 | prostaglandin transporter PGT mRNA, complete cds |
| AF333770.17 | cell recognition molecule CASPR4 mRNA, complete cds |
| AF340028.2 | Rab6-interacting protein 2 isoform A mRNA, complete cds |
| AF340029.2 | Rab6-interacting protein 2 isoform B mRNA, complete cds |
| AF402611.3 | actin-binding protein frabin-alpha mRNA, complete cds |
| AF532979.7 | putative RNA methylase mRNA, complete cds |
| AJ007938.7 | S6 kinase 2 |
| AJ007938.8 | S6 kinase 2 |
| AJ414734.12 | mRNA for 53BP1 protein |
| AK010722.7 | ES cells cDNA, RIKEN full-length enriched library, clone:2410075D05 product:hypothetical N-6 Adenine-specific DNA methylase containing protein, full insert sequence |
| AK028874.5 | 10 days neonate skin cDNA, RIKEN full-length enriched library, clone:4732465J15 product:weakly similar to N-ACETYLGLUCOSAMINE-6-SULFATASE [Coturnix coturnix], full insert sequence |

| | |
|---|---|
| AK029844.7 | adult male testis cDNA, RIKEN full-length enriched library, clone:4931410H15 product:hypothetical protein, full insert sequence |
| AK030664.3 | 6 days neonate head cDNA, RIKEN full-length enriched library, clone:5430412B19 product:hypothetical Amine oxidase containing protein, full insert sequence |
| AK034251.2 | adult male diencephalon cDNA, RIKEN full-length enriched library, clone:9330168O12 product:weakly similar to DJ846F13.1 (PHOSPHATIDIC ACID PHOSPHATASE TYPE 2C) (FRAGMENT) |
| AK034712.5 | 12 days embryo embryonic body between diaphragm region and neck cDNA, RIKEN full-length enriched library, clone:9430027K19 product:weakly similar to N-ACETYLGLUCOSAMINE-6-SULFATASE |
| AK036685.5 | adult male bone cDNA, RIKEN full-length enriched library, clone:9830162M14 product:weakly similar to N-ACETYLGLUCOSAMINE-6-SULFATASE |
| AK037822.7 | 16 days neonate thymus cDNA, RIKEN full-length enriched library, clone:A130052K22 product:immunoglobulin heavy chain 6 (heavy chain of IgM), full insert sequence |
| AK038882.6 | adult male hypothalamus cDNA, RIKEN full-length enriched library, clone:A230072B04 product:hypothetical protein, full insert sequence |
| AK039161.10 | adult male hypothalamus cDNA, RIKEN full-length enriched library, clone:A230103M05 product:calcitonin receptor, full insert sequence |
| AK041750.6 | 3 days neonate thymus cDNA, RIKEN full-length enriched library, clone:A630034L20 product:weakly similar to FC RECEPTOR-LIKE PROTEIN 1 |
| AK042401.11 | 3 days neonate thymus cDNA, RIKEN full-length enriched library, clone:A630088O16 product:similar to DACHSHUND-LIKE PROTEIN DACH2 |
| AK046272.3 | adult male corpora quadrigemina cDNA, RIKEN full-length enriched library, clone:B230363K08 product:hypothetical PapD-like structure containing protein |
| AK077689.9 | 8 days embryo whole body cDNA, RIKEN full-length enriched library, clone:5730523J24 product:PP1201 PROTEIN homolog |
| AK081643.5 | 16 days embryo head cDNA, RIKEN full-length enriched library, clone:C130058L07 product:weakly similar to N-ACETYLGLUCOSAMINE-6-SULFATASE |
| AK081865.3 | 16 days embryo head cDNA, RIKEN full-length enriched library, clone:C130083H12 product:hypothetical Phospholipid and glycerol acyltransferase (from 'motifs_6.msf') and EF-hand containing protein |
| AK087350.9 | 0 day neonate lung cDNA, RIKEN full-length enriched library, clone:E030047P06 product:solute carrier family 21 (prostaglandin transporter), member |

| | |
|---|---|
| AK087439.2 | 0 day neonate eyeball cDNA, RIKEN full-length enriched library, clone:E130118N02 product:weakly similar to DJ846F13.1 (PHOSPHATIDIC ACID PHOSPHATASE TYPE 2C) (FRAGMENT) |
| AK090093.7 | kidney CCL-142 RAG cDNA, RIKEN full-length enriched library, clone:G430107K15 product:ribosomal protein S6 kinase, 70kD, polypeptide 2 |
| AK090093.8 | kidney CCL-142 RAG cDNA, RIKEN full-length enriched library, clone:G430107K15 product:ribosomal protein S6 kinase, 70kD, polypeptide 2 |
| AK122318.29 | mRNA for mKIAA0587 protein |
| AK122318.30 | mRNA for mKIAA0587 protein |
| AK122322.9 | mRNA for mKIAA0595 protein |
| AK122470.7 | mRNA for mKIAA1219 protein |
| AK122478.22 | mRNA for mKIAA1250 protein |
| AK129316.5 | mRNA for mKIAA1247 protein |
| AY008297.17 | RPGR-interacting protein mRNA, complete cds |
| AY037807.17 | CCAAT displacement protein CDP (Cutl1) mRNA, complete cds |
| AY037807.22 | CCAAT displacement protein CDP (Cutl1) mRNA, complete cds |
| AY206701.2 | phosphatidylinositol-binding clathrin assembly protein (Picalm) mRNA, complete cds |
| AY220301.15 | betaPix-bL mRNA, complete cds |
| AY251601.5 | zona pellucida binding protein 2 (Zpbp2) mRNA, complete cds |
| AY331142.12 | potassium voltage-gated channel major isoform 1 (Kcnq1) mRNA, complete cds |
| BB610056.2 | BB610056 RIKEN full-length enriched, adult male liver Mus musculus cDNA clone 1300006B11 5', mRNA sequence |
| BC011470.2 | phosphatidylinositol binding clathrin assembly protein, mRNA (cDNA clone MGC:19382 IMAGE:2651109), complete cds |
| BC034628.17 | contactin associated protein 4, mRNA (cDNA clone MGC:36519 IMAGE:5369648), complete cds |
| BC045142.12 | potassium voltage-gated channel, subfamily Q, member 1, mRNA (cDNA clone MGC:54702 IMAGE:6314642), complete cds |
| BC046979.21 | ATPase, H+ transporting, lysosomal V0 subunit A isoform 4, mRNA (cDNA clone MGC:54824 IMAGE:6315485), complete cds |
| BC048410.7 | RIKEN cDNA 9330174J19 gene, mRNA (cDNA clone MGC:57080 IMAGE:6486835), complete cds |
| BC050879.2 | cDNA clone MGC:63264 IMAGE:5687835, complete cds |

| | |
|---|---|
| BC050986.5 | RIKEN cDNA 1700017D11 gene, mRNA (cDNA clone IMAGE:6772587) |
| BC051631.8 | ribosomal protein S6 kinase, polypeptide 2, mRNA (cDNA clone IMAGE:6331314), complete cds |
| BC051631.9 | ribosomal protein S6 kinase, polypeptide 2, mRNA (cDNA clone IMAGE:6331314), complete cds |
| BC052048.2 | cDNA clone MGC:62428 IMAGE:5687835, complete cds |
| BC052718.4 | SUMO |
| BC054544.3 | catenin src, mRNA (cDNA clone MGC:62385 IMAGE:6408956), complete cds |
| BC055033.3 | cDNA sequence BC006909, mRNA (cDNA clone MGC:62672 IMAGE:6310272), complete cds |
| BC055304.12 | potassium voltage-gated channel, subfamily Q, member 1, mRNA (cDNA clone MGC:62642 IMAGE:6397978), complete cds |
| BC055830.5 | cDNA clone MGC:67780 IMAGE:3599941, complete cds |
| BC057354.17 | oxoglutarate dehydrogenase (lipoamide), mRNA (cDNA clone MGC:67193 IMAGE:6827509), complete cds |
| BC058790.6 | cDNA clone MGC:67710 IMAGE:6395070, complete cds |
| BC061480.5 | SUMO |
| BC062654.2 | GNAS (guanine nucleotide binding protein, alpha stimulating) complex locus, mRNA (cDNA clone IMAGE:6822390), complete cds |
| BC062900.5 | sulfatase 2, mRNA (cDNA clone MGC:86096 IMAGE:6810085), complete cds |
| BC065159.3 | guanine nucleotide binding protein, alpha inhibiting 2, mRNA (cDNA clone MGC:90102 IMAGE:5717305), complete cds |
| BC066048.9 | peroxisome proliferative activated receptor, gamma, coactivator-related 1, mRNA (cDNA clone MGC:90133 IMAGE:6825158), complete cds |
| BC066095.8 | RIKEN cDNA E030049G20 gene, mRNA (cDNA clone MGC:90076 IMAGE:6856184), complete cds |
| BC066809.3 | cDNA sequence BC005662, mRNA (cDNA clone MGC:76519 IMAGE:30475544), complete cds |
| BC068128.4 | expressed sequence AI118078, mRNA (cDNA clone MGC:92972 IMAGE:6835202), complete cds |
| BC069846.10 | cDNA clone MGC:78014 IMAGE:4196478, complete cds |
| BG864492.2 | 602798514F1 NCI_CGAP_Mam4 Mus musculus cDNA clone IMAGE:4919571 5', mRNA sequence |
| BM453600.2 | AGENCOURT_6419456 NCI_CGAP_Ov44 Mus musculus cDNA clone IMAGE:5504484 5', mRNA sequence |
| CB273255.2 | mai76g07.y1 McCarrey Eddy spermatocytes Mus musculus cDNA clone IMAGE:6447637 5', mRNA sequence |

| | |
|---|---|
| CF736440.2 | UI-M-HD0-ckp-a-06-0-UI.r1 NIH_BMAP_HD0 Mus musculus cDNA clone IMAGE:30612005 5', mRNA sequence |
| CN704086.2 | E0483D06-5 NIA Mouse E11.5 whole embryo cDNA library (Long) Mus musculus cDNA clone NIA:E0483D06 IMAGE:30876521 5', mRNA sequence |
| CN704534.2 | E0489A01-5 NIA Mouse E11.5 whole embryo cDNA library (Long) Mus musculus cDNA clone NIA:E0489A01 IMAGE:30877056 5', mRNA sequence |
| D28599.7 | mRNA for proteoglycan, PG-M. |
| D28599.8 | mRNA for proteoglycan, PG-M. |
| D88187.14 | mRNA for Ftp-1, complete cds. |
| D88187.7 | mRNA for Ftp-1, complete cds. |
| L04275.8 | macrophage scavenger receptor type II mRNA, complete cds. |
| L26507.6 | Mouse myocyte nuclear factor (MNF) mRNA, complete cds. |
| NM_007588.10 | calcitonin receptor (Calcr), mRNA |
| NM_008138.3 | guanine nucleotide binding protein, alpha inhibiting 2 (Gnai2), mRNA |
| NM_008434.12 | potassium voltage-gated channel, subfamily Q, member 1 (Kcnq1), mRNA |
| NM_009260.16 | spectrin beta 2 (Spnb2), transcript variant 2, mRNA |
| NM_009986.17 | cut-like 1 (Drosophila) (Cutl1), transcript variant 1, mRNA |
| NM_009986.22 | cut-like 1 (Drosophila) (Cutl1), transcript variant 1, mRNA |
| NM_010561.4 | interleukin enhancer binding factor 3 (Ilf3), mRNA |
| NM_010703.7 | lymphoid enhancer binding factor 1 (Lef1), mRNA |
| NM_010862.2 | myosin XV (Myo15), transcript variant 1, mRNA |
| NM_011214.14 | protein tyrosine phosphatase, receptor type, U (Ptpru), mRNA |
| NM_011214.7 | protein tyrosine phosphatase, receptor type, U (Ptpru), mRNA |
| NM_011378.2 | transcriptional regulator, SIN3A (yeast) (Sin3a), mRNA |
| NM_013735.12 | transformation related protein 53 binding protein 1 (Trp53bp1), mRNA |
| NM_019689.4 | AT rich interactive domain 3B (Bright like) (Arid3b), mRNA |
| NM_021485.7 | ribosomal protein S6 kinase, polypeptide 2 (Rps6kb2), mRNA |
| NM_021485.8 | ribosomal protein S6 kinase, polypeptide 2 (Rps6kb2), mRNA |
| NM_023220.10 | RIKEN cDNA 2010106G01 gene (2010106G01Rik), mRNA |
| NM_023638.3 | porcupine homolog (Drosophila) (Porcn), transcript variant Mporc-d, mRNA |

| | |
|---|---|
| NM_023879.17 | retinitis pigmentosa GTPase regulator interacting protein 1 (Rpgrip1), mRNA |
| NM_027061.5 | RIKEN cDNA 1700017D11 gene (1700017D11Rik), transcript variant 1, mRNA |
| NM_027154.9 | RIKEN cDNA 2310061B02 gene (2310061B02Rik), mRNA |
| NM_028604.7 | RIKEN cDNA 2410075D05 gene (2410075D05Rik), mRNA |
| NM_031195.8 | macrophage scavenger receptor 1 (Msr1), mRNA |
| NM_033314.9 | solute carrier organic anion transporter family, member 2a1 (Slco2a1), mRNA |
| NM_053204.2 | Rab6 interacting protein 2 (Rab6ip2), mRNA |
| NM_130457.17 | contactin associated protein 4 (Cntnap4), mRNA |
| NM_139232.3 | FYVE, RhoGEF and PH domain containing 4 (Fgd4), mRNA |
| NM_145376.3 | cDNA sequence BC005662 (BC005662), mRNA |
| NM_145589.3 | cDNA sequence BC006909 (BC006909), mRNA |
| NM_146003.5 | SUMO |
| NM_146194.2 | phosphatidylinositol binding clathrin assembly protein (Picalm), mRNA |
| NM_153783.3 | polyamine oxidase (Paox), mRNA |
| NM_172613.7 | RIKEN cDNA 9330174J19 gene (9330174J19Rik), mRNA |
| NM_172848.3 | RIKEN cDNA B230363K08 gene (B230363K08Rik), mRNA |
| NM_199068.6 | forkhead box K1 (Foxk1), transcript variant 1, mRNA |
| NM_207675.8 | immunoglobulin superfamily, member 4A (Igsf4a), transcript variant 1, mRNA |
| U18542.10 | calcitonin receptor 1b mRNA, complete cds. |
| U20975.5 | kidney-specific Na-K-Cl cotransport protein splice isoform F (NKCC2) mRNA, complete cds. |
| U22394.2 | mSin3A (sin3A) mRNA, complete cds. |
| U55057.14 | receptor protein tyrosine phosphatase-lamda (ptp-lambda) mRNA, complete cds. |
| U55057.7 | receptor protein tyrosine phosphatase-lamda (ptp-lambda) mRNA, complete cds. |
| U70068.12 | potassium channel subunit (KvLQT1) mRNA, complete cds. |
| X58636.7 | Mouse LEF1 mRNA for lymphoid enhancer binding factor 1 |
| XM_127605.29 | discs, large homolog 5 (Drosophila) (Dlg5), mRNA |
| XM_129836.4 | hypothetical protein LOC213109 (LOC213109), mRNA |
| XM_130851.3 | RIKEN cDNA 4933421B21 gene (4933421B21Rik), mRNA |
| XM_138692.4 | hypothetical protein A530054K11 (A530054K11), mRNA |
| XM_139336.4 | similar to putative pheromone receptor (LOC223385), mRNA |

| | |
|---|---|
| XM_140451.11 | laminin, alpha 3 (Lama3), mRNA |
| XM_146954.16 | similar to Myosin Vc (Myosin 5C) (LOC208943), mRNA |
| XM_148343.3 | similar to Putative Rho |
| XM_284502.5 | SUMO-1-specific protease (Susp1-pending), mRNA |
| XM_358923.3 | similar to RIKEN cDNA 1810036I24 (LOC385719), mRNA |

# APPENDIX D

## LIST OF RAT TARGET GENES PREDICTED BY NTHunter (FROM CHAPTER 5)

| GenBank ID - Exon No | Annotation |
|---|---|
| AF239045.23 | KIDINS220 (Kidins220) mRNA, complete cds |
| AJ250425.2 | mRNA for collybistin I |
| AJ272428.3 | mRNA for cyclic nucleotide-gated channel 2b |
| AJ288898.15 | mRNA for GABA-A receptor interacting factor-1 (GRIF-1 gene), splice variants |
| AJ537441.5 | partial mRNA for MHC class Ib antigen (rt1-E2d gene), allele E2d(c) |
| AY325159.8 | Ab1-346 mRNA, complete cds |
| AY325245.11 | Cc2-5 mRNA, complete cds |
| AY348867.8 | rhotekin isoform 1 mRNA, complete cds; alternatively spliced |
| AY557199.8 | Kv4 potassium channel auxiliary subunit mRNA, complete cds |
| CB576499.2 | AMGNNUC:NRHY5-00391-B3-A W Rat hypothalamus (10471) Rattus norvegicus cDNA clone nrhy5-00391-b3 5', mRNA sequence |
| CF108157.2 | hultzomica01408 Rat lung airway and parenchyma cDNA libraries Rattus norvegicus cDNA clone Contig1074 5', mRNA sequence |
| CF978525.2 | F32G05_035.ab1.R Rat retinal ganglion cell Rattus norvegicus cDNA, mRNA sequence |
| CO390016.2 | AGENCOURT_26621744 NIH_MGC_253 Rattus norvegicus cDNA clone IMAGE:7302753 5', mRNA sequence |
| L14618.9 | Rattus norvegicus C1b receptor mRNA, complete cds |
| NM_019221.13 | transformation related protein 63 (Trp63), mRNA |
| NM_023957.2 | collybistin I (Arhgef9), mRNA |
| NM_030989.5 | tumor protein p53 (Tp53), mRNA |
| NM_031075.3 | purinergic receptor P2X ligand-gated ion channel, 3 (P2rx3), mRNA |
| NM_053795.23 | kinase D-interacting substance of 220 kDa (Kidins220), mRNA |
| NM_053816.9 | calcitonin receptor (Calcr), mRNA |

|  |  |
|---|---|
| NM_184046.8 | rhotekin (Rtkn), mRNA |
| U60562.5 | amelogenin mRNA, complete cds |
| X13058.5 | nuclear oncoprotein p53 |
| X90651.3 | mRNA for P2X3 receptor |
| XM_232212.16 | GRP1 binding protein GRSP1 (Grsp1), mRNA |
| XM_237523.14 | myomesin 1 (Myom1), mRNA |
| XM_238806.21 | similar to harmonin isoform b3 (LOC308596), mRNA |
| XM_342209.2 | E2F transcription factor 5 (E2f5), mRNA |
| XM_343385.8 | similar to RA175 (LOC363058), mRNA |
| XM_343402.27 | neogenin (Neo1), mRNA |
| XM_347229.8 | similar to cystic fibrosis transmembrane conductance regulator (LOC368064), mRNA |
| XM_347232.8 | cystic fibrosis transmembrane conductance regulator homolog (Cftr), mRNA |
| Y10258.13 | mRNA for TA2 KET alpha protein (p63 gene) |
| Y10473.6 | mRNA for P2X2 receptor, splice variant P2X2b |

# APPENDIX E

# AutoDB DATABASE BUILDING PROCESS (FROM CHAPTER 7)

## General information about the directory structure:

**Note:** All build steps assume you are using the Bash shell. STDERR and STDOUT redirect work differently for other shells, such as csh; please check to make sure you use the appropriate redirects for your shell.

The root directory for all input data and building tools for AutoDB is `/projects/AutoDB`.

The `/projects/AutoDB/InputData` directory contains data directories for each organism and database version. Create a new organism directory and database version directory for each new AutoDB project, if necessary, and then create the `ExpressedSequence` and `GenomicSequence` directories.

The `GenomicSequence` directory contains symbolic links to the genomic chromosomal sequence data files (in the directories `/seqdata/genomes/<organism>/<chromosome_directory>`), plus the `nibs` directory for the BLAT `.nib` files (see step 1-A below) and the `str` directory for the stripped data files (see step 1-B below).

## The 6 steps required for running AutoDB and building the web interface:

### Step 1: Preparing Genomic Data

**Note:** The Genome Sequence Administrator should perform the first two data preparation steps, since not everyone has permission to write to the `/seqdata` genome sequence repository.

The genome sequence files for the organism should be downloaded from the source and placed in:

`/seqdata/genomes/<organism>/<genome_version>`

The GenomicSequence directory of the new AutoDB project should be set up as a symbolic link to the sequence data directory. The genome data directory /seqdata/genomes/<organism>/<genome_version> should then be set up to contain the /nibs and /str directories detailed in the following steps.

**(A) .nib files** have to be generated for BLAT from the FASTA genome sequence files in the genome data directory. To create .nib files use BLAT's FaToNib tool, /usr/local/biotools/blat/FaToNib.

Method:

```
$ cd /projects/AutoDB/InputData/<organism>/<DB_NAME>/
GenomicSequence
$ mkdir nibs
$ /usr/local/biotools/blat/faToNib chr1.fa nibs/chr1.nib
```

Repeat for each FASTA chromosome sequence file.

**Note:** The directory GenomicSequence above is a symbolic link to /seqdata/genomes/<organism>/<genome_version>, which has limited write permission, so an authorized administrator must perform these steps.

**(B) .str files** need to be created for the genomic sequence data as follows. These are stripped data files, where the > identifier line and all whitespace are removed from the genomic FASTA file. From directory /projects/AutoDB/build/bin, run format-str-files.pl.

Method:

```
$ mkdir /seqdata/genomes/<organism>/<genome_version>/str
$ cd /projects/AutoDB/build/bin
$./format-str-files.pl -i
/projects/AutoDB/InputData/organism/DB_NAME/GenomicSequen
ce
-o
/projects/AutoDB/InputData/organism/DB_NAME/GenomicSequen
ce/str/ -p 'chr.*\.fa'
```

**Note:** Again, the directory GenomicSequence above is a symbolic link to /seqdata/genomes/<organism>/<genome_version>.

**(C)** For each organism, **.ini configuration files** should be checked. These files are named after the organisms, and are placed in `/projects/AutoDB/build/conf`. Before each run, these files need to be checked and updated. The fields in this file are as follows:

1. `Blat`: This specifies the path to the blat gfClient. Probably does not need to be changed.
2. `nibs_dir`**:** Path to the nibs files used for the particular run. This will likely need to be updated with each new version of the organism's database.
3. `hosts_ports`: This is a list of the hosts and port numbers on which gfServers for this organism are running. Each entry consists of a host name and port number, separated by a single colon. Multiple entries are separated by three colons.
   Example:
   `hosts_ports=HOST:17777` – This entry is for one host/port.
   `hosts_ports=HOST1:17777:::HOST1:17778:::HOST2:17777:::HOST2:17778` – This entry lists four gfServers across two different hosts.
4. `genome_data`: Full path to the `str` directory. Example: `genome_data=/projects/AutoDB/InputData/mouse/MouSDB5/GenomicSequence/str/`
5. `chr_ids`: single-colon separated list of chromosome ids. Must match the `str` and `nibs` file names, without their extensions. Example: `chr_ids=chr1:chr2:chr3:chr4`
6. `min_trans_per_clust`: Minimum transcripts per cluster. Default is 3, which eliminates singleton and doubleton clusters.
7. `percent_coverage`: Minimum percent ID that a transcript must have to the genome else it will be discarded. Default is 75.
8. `exon_percent_coverage`: Minimum percent ID of each exon to the genome, else the transcript is discarded. If the exon has percent ID less than this number, it may still be considered if it passes based on `mm_threshold`. Default is 95.
9. `mm_threshold`: Maximum amount of mismatches an exon can have if it fails exon_percent_coverage. This allows small exons which may have a low percent ID but also low number of mismatches to still be considered valid. Default is 5.
10. `min_exons`: minimum number of exons that a transcript must have to be included in the database. Since AutoDB analyzes splice variation, the default for this option is 2 (2 or more exons are required to have splice sites).
11. `min_exons_est` and `est_regex` can be disregarded as they are not currently used.

## Step 2: Running Blat

The file /projects/AutoDB/gfServers lists each organism followed by one or more host names and one or more port numbers. This is to help distribute hosts and ports across all organisms, to prevent running all organisms on one server, on one port. If your organism is not listed, please add it to the end of the file and take the next port number (i.e. if the highest port number in the file that has been used is 17789, please use 17790).

To start the gfServer(s), first ssh into the host you need to run it on, then navigate to the nibs file directory:

```
$ cd /projects/AutoDB/InputData/<organism>/<DB_NAME>/
GenomicSequence/nibs
```

While in the nibs directory, run:

```
$ /usr/local/biotools/blat/gfServer start <hostname>
<port_number> *.nib &
```

Then log out of the host.

**Important:** Before going on to the next step, it is vital that you double-check the availability of the gfServer(s) you started. To do so run the following command for each host/port combination.

```
$ /usr/local/biotools/blat/gfServer status <hostname>
<port_number>
```

If the server is available it will return with multiple lines of information. If not, it will state that it cannot connect to the server.

## Step 3: Starting AutoDB

From the directory /projects/AutoDB/build/bin, run generate-project-data.pl, providing transcript input data from ExpressedSequence directory. Following command shows starting AutoDB for a *Drosophila melanogaster* database.

```
$ ./generate-project-data.pl –p DmelSDB –o dmelanogaster
–data
/projects/AutoDB/InputData/dmelanogaster/DmelSDB/Expresse
dSequence/Dmelanogaster-20050315.gbank > DmelSDB.err 2>&1
&
```

**Note:** If you have multiple data files, please pass multiple –data arguments to `generate-project-data.pl`. The -data option is used for GenBank, UniGene, dbEST and Ensembl formats. The files must end `in.gbank, .ugn, .dbest` or `.embl` respectively.

### Step 4: Loading information into databases

After the first three steps are complete, information is loaded into databases with the following script. From the directory `/projects/AutoDB/build/bin`, run `project-step2.pl`. For each run, the database version number `(dbver)` should be updated. Following command shows loading information into a *Drosophila melanogaster* database.

```
$ ./project-step2.pl —p DmelSDB —o dmelanogaster —abbrev
Dm —dbver 1 > DmelSDB.step2.err 2>&1 &
```

### Step 5: Stopping Blat

After the runs are done, it is important to stop the gfServer with the following command. This frees the resources that the gfServer(s) were using.

```
$ /usr/local/biotools/blat/gfServer stop <hostname>
<port_number>
```

### Step 6: Creating the web interface

After the SDB is created and populated, run the following command:

```
$ /projects/AutoDB/www/createPhpInterface <NewSDB> <user>
<hostname> <schema>
```

Example:

```
$ /projects/AutoDB/www/createPhpInterface DmelSDB anovo
emmy 3
```

**Note:** Variable <schema> refers to the SDB schema, which has changed to its current state after HumanSDB2. All SDBs starting from HumanSDB3, MouSDB5, and RatSDB2 have schema version 3.

The script `createPhpInterface` sequentially launches the following scripts:

```
/projects/AutoDB/www/CREATE_TITLE_TABLE.pl <NewSDB>
<user> <host>

/projects/AutoDB/www/POPULATE_MESH_ANNOTATION.pl
<NewSDB> <user> <host>

/projects/AutoDB/www/CHECK_VAR_CLUSTERS.pl <NewSDB>
<user> <host>

/projects/AutoDB/www/CREATE_STATS.pl <NewSDB> <user>
<host> <schema>
```

It also adds <NewSDB> lines to the following files:

```
/projects/AutoDB/www/sdb.php
/projects/AutoDB/www/include/env.php
/projects/AutoDB/www/cluster_map.php
```

The list of AutoDB databases is accessible through the following URL:

http://sgc.ucsd.edu/autodb/sdb.php.

# REFERENCES

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter P. (2002) Molecular Biology of the Cell. 4th edition. Garland Science, New York, NY.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol.* **215**: 403-410.

Ast, G. (2004) How did alternative splicing evolve? *Nat. Rev. Genet.* **5(10)**: 773-782.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., Sonnhammer, E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.* **30**: 276-280.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C., Eddy, S.R. (2004) The Pfam protein families database. *Nucleic Acids Res.* **32:** Database issue: D138-41.

Berget, S.M. (1995) Exon recognition in vertebrate splicing. *J Biol Chem.* **270(6)**: 2411-2414.

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

Black, D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72**: 291-336.

Brett, D., Popisil, H., Valcarel, J., Reich, J., Bork, P. (2002) Alternative splicing and genome complexity. *Nature Genetics* **1**: 29-30.

Brinkman, B.M. (2004) Splice variants as cancer biomarkers. *Clin. Biochem.* **37(7)**: 584-594.

Brow, D.A. (2002) Allosteric cascade of spliceosome activation. *Annu. Rev. Genet.* **36**: 333-360.

Buckanovich, R.J., Yang, Y.L., Darnell, R.B. (1996) The onconeural antigen Nova-1 is a neuron-specific RNA-binding protein, the activity of which is inhibited by paraneoplastic antibodies. *The Journal of Neuroscience* **16**: 1114-1122.

Buckanovich, R.J. and Darnell, R.B. (1997) The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. *Molecular and Cellular Biology* **17**: 3194-3201.

Burset, M., Seledtsov, I.A., Solovyev, V.V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* **28(21)**: 4364-4375.

Burset, M., Seledtsov, I.A., Solovyev, V.V. (2001) SpliceDB: database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Res.* **29**: 255-259.

Byers, P.H., Duvic, M., Atkinson, M., Robinow, M., Smith, L.T., Krane, S.M., Greally, M.T., Ludman, M., Matalon, R., Pauker, S., Quanbeck, D., Schwarze, U. (1997) Ehlers-Danlos syndrome type VIIA and VIIB result from splice-junction mutations or genomic deletions that involve exon 6 in the COL1A1 and COL1A2 genes of type I collagen. *Am J Med Genet.* **72(1)**: 94-105.

Caceres, J.F., Kornblihtt, A.R. (2002) Alternative splicing; multiple control mechanisms and involvement in human disease. *Trends in Genet.* **18**: 186 - 193.

Carmel, I., Tal, S., Vig, I., Ast, G. (2004) Comparative analysis detects dependencies among the 5′ splice-site positions. *RNA* **10(5)**: 828–840.

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

Castle, J., Garrett-Engele, P., Armour, C.D., Duenwald, S.J., Loerch, P.M., Meyer, M.R., Schadt, E.E., Stoughton, R., Parrish, M.L., Shoemaker, D.D., Johnson, J.M. (2003) Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing. *Genome Biol.* **4(10)**: R66.

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

Celotto, A.M., Graveley, B.R. (2001) Alternative splicing of the Drosophila Dscam pre-mRNA is both temporally and spatially regulated. *Genetics* **159(2)**: 599-608.

Cline, M.S., Shigeta, R., Wheeler, R.L., Siani-Rose, M.A., Kulp, D., Loraine, A.E. (2004) The effects of alternative splicing on transmembrane proteins in the mouse genome. *Pac. Symp. Biocomput.* 17-28.

Cooper, T.A., Mattox, W. (1997) The regulation of splice-site selection, and its role in human disease. *Am. J. Hum. Genet.* **61**: 259-266.

Coward, E., Haas, S.A., Vingron, M. (2002) SpliceNest: visualizing gene structure and alternative splicing based on EST clusters. *Trends Genet.* **18**: 53-55.

Cramer, P., Pesce, C.G., Baralle, F.E., Kornblihtt, A.R. (1997) Functional association between promoter structure and transcript alternative splicing. *Proc Natl Acad Sci USA.* **94(21)**:11456-11460.

Cramer, P., Caceres, J.F., Cazalla, D., Kadener, S., Muro, A.F., Baralle, F.E., Kornblihtt, A.R. (1999) Coupling of transcription with alternative splicing: RNA pol II promoters modulate SF2/ASF and 9G8 effects on an exonic splicing enhancer. *Mol. Cell.* **4(2)**: 251-258.

Cramer, P., Srebrow, A., Kadener, S., Werbajh, S., de la Mata, M., Melen, G., Nogues, G., Kornblihtt, A.R. (2001) Coordination between transcription and pre-mRNA processing. *FEBS Letters* **498**: 179-182.

de la Mata, M., Alonso, C.R., Kadener, S., Fededa, J.P., Blaustein, M., Pelisch, F., Cramer, P., Bentley, D., Kornblihtt, A.R. (2003) A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell.* **12(2)**: 525-532.

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

Dredge, B.K., Darnell, R.B. (2003) Nova regulates GABA$_A$ receptor 2 alternative splicing via a distal downstream UCAU-rich intronic splicing enhancer. *Molecular and Cellular Biology* **23**: 4687-4700.

**Ensembl** (2005a) [http://www.ensembl.org/Fugu_rubripes/]

**Ensembl** (2005b) [http://www.ensembl.org/Danio_rerio/]

**Ensembl** (2005c) [http://www.ensembl.org/Mus_musculus/]

**Ensembl** (2005d) [http://www.ensembl.org/Rattus_norevegicus/]

**Ensembl** (2005e) [http://www.ensembl.org/Drosophila_melanogaster/]

**Ensembl** (2005f) [http://www.ensembl.org/Caenorhabditis_elegans/]

Feltes, C.M., Kudo, A., Blaschuk, O., Byers, S.W. (2002) An alternatively spliced cadherin-11 enhances human breast cancer cell invasion. *Cancer Res.* **62(22)**: 6688-6697.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., Miller, W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8**: 967-974.

Foulkes, N.S., Sassone-Corsi, P. (1992) More is better activators and repressors from the same gene. *Cell* **68:** 411-414.

Friedman, K.J., Kole, J., Cohn, J.A., Knowles, M.R., Silverman, L.M., Kole, R. (1999) Correction of aberrant splicing of the cystic fibrosis transmembrane conductance regulator (CFTR) gene by antisense oligonucleotides. *J Biol Chem.* **274(51)**: 36193-36199.

Grabowski, P., Black, D.L. (2001) Alternative RNA splicing in the nervous system. *Progress in Neurobiology* **65**: 289-308.

Grasso, C., Modrek, B., Xing, Y., Lee, C. (2004) Genome-wide detection of alternative splicing in expressed sequences using partial order multiple sequence alignment graphs.
*Pac. Symp. Biocomput.* 29-41.

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

Graveley, B.R. (2002) Sex, AGility, and the regulation of alternative splicing. *Cell* **109(4)**: 409-412.

Graveley, B.R., Kaur, A., Gunning, D., Zipursky, S.L., Rowen, L., Clemens, J.C. (2004) The organization and evolution of the dipteran and hymenopteran Down syndrome cell adhesion molecule (Dscam) genes. *RNA* **10(10)**: 1499-1506.

Gupta, S., Zink, D., Korn, B., Vingron, M., Haas, S.A. (2004) Strengths and weaknesses of EST-based prediction of tissue-specific alternative splicing. *BMC Genomics* **5**: 72.

Hu, G.K., Madore, S.J., Moldover, B., Jatkoe, T., Balaban, D., Thomas, J., Wang, Y. (2001) Predicting splice variant from DNA chip expression data. *Genome Res.* **11(7):** 1237-1245.

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

Hui, L., Zhang, X., Wu, X., Lin, Z., Wang, Q., Li, Y., Hu, G. (2004) Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. *Oncogene* **23(17)**: 3013-3023.

Iida, K., Seki, M., Sakurai, T., Satou, M., Akiyama, K., Toyoda, T., Konagaya, A., Shinozaki, K. (2004) Genome-wide analysis of alternative pre-mRNA splicing in Arabidopsis thaliana based on full-length cDNA sequences. *Nucleic Acids Res.* **32(17)**: 5096-5103.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature **431**: 931-945.

Jensen, K.B., Dredge, B.K., Stefani, G., Zhong, R., Buckanovich, R.J., Okano, H.J., Yang, Y.Y., Darnell, R.B. (2000a) Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron* **2**: 359-371.

Jensen, K.B., Musunuru, K., Lewis, H.A., Burley, S.K., Darnell, R.B. (2000b) The tetranucleotide UCAY directs the specific recognition of RNA by the Nova K-homology 3 domain. *Proc Natl Acad Sci* USA **97(11)**: 5740-5745.

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

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

Jurica, M.S., Morre, M.J. (2003) Pre-mRNA Splicing: Awash in a Sea of Proteins. *Molecular Cell* **12**: 5-14.

Kalnina, Z., Zayakin, P., Silina, K., Line, A. (2005) Alterations of pre-mRNA splicing in cancer. *Genes Chromosomes Cancer* **42(4)**: 342-357.

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

Kent, W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.* **12**: 656-664.

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(46)**: 42986-42993.

Kriventseva, E.V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M.S., Sunyaev, S. (2003) Increase of functional diversity by alternative splicing. *Trends Genet.* **19(3)**: 124-128.

Latchman, D.S. (2001) Transcription factors: bound to activate or repress. *Trends Biochem Sci.* **26(4)**: 211-213.

Le, K., Mitsouras, K., Roy, M., Wang, Q., Xu, Q., Nelson, S.F., Lee, C. (2004) Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. *Nuc Acids Res.* **32(22)**: e180.

Lee, C., Roy, M. (2004) Analysis of alternative splicing with microarrays: successes and challenges. *Genome Biol.* **5(7)**: 231.

Lee, C., Atanelov, L., Modrek, B., Xing, Y. (2003) ASAP: the Alternative Splicing Annotation Project. *Nucleic Acids Res.* **31**: 101-105.

Lehman, K., Schmidt, U. (2003) Group II introns: structure and catalytic versatility of large natural ribozymes. *Crit Rev Biochem Mol Biol.* **38(3)**: 249-303.

Letunic, I., Goodstadt, L., Dickens, N.J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R.R., Ponting, C.P., Bork, P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.* **30:** 242-244.

Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P., Bork, P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.* **32:** Database issue: D142-4.

Liu, S., Altman, R.B. (2003) Large scale study of protein domain distribution in the context of alternative splicing. *Nucleic Acids Res.* **31(16)**: 4828-4835.

Lopez, A.J. (1995) Developmental role of transcription factor isoforms generated by alternative splicing. *Dev Biol* **172**: 396–411.

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

Ma, X., He, F. (2003) Advances in the study of SR protein family. *Genomics Proteomics Bioinformatics.* **1(1)**: 2-8.

Mercatante, D.R., Bortner, C.D., Cidlowski, J.A., Kole, R. (2001a) Modification of alternative splicing of Bcl-x pre-mRNA in prostate and breast cancer cells. Analysis of apoptosis and cell death. *J Biol Chem.* **276(19)**: 16411-16417.

Mercatante, D.R., Sazani, P., Kole, R. (2001b) Modification of alternative splicing by antisense oligonucleotides as a potential chemotherapy for cancer and other diseases. *Curr Cancer Drug Targets* **1(3)**: 211-230.

Mercatante, D.R., Kole, R. (2002) Control of alternative splicing by antisense oligonucleotides as a potential chemotherapy: effects on gene expression. *Biochim Biophys Acta.* **1587**: 126-132.

Mironov, A.A., Fickett, J.W., Gelfand, M.S. (1999) Frequent alternative splicing of human genes. *Genome Res.* **9(12):** 1288-1293.

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

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

Modrek, B., Lee, C. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genet.* **34(2)**: 177-180.

Mouse Genome Sequencing Consortium. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520-562.

Musunuru, K. (2003) Cell-specific RNA-binding proteins in human disease. *Trends Cardiovasc Med.* **13(5):** 188-195.

Nogues, G., Kadener, S., Cramer, P., Bentley, D., Kornblihtt, A.R. (2002) Transcriptional activators differ in their abilities to control alternative splicing. *J Biol Chem.* **277(45)**: 43110-43114.

Nogues, G., Kadener, S., Cramer, P., de la Mata, M., Fededa, J.P., Blaustein, M., Srebrow, A., Kornblihtt, A.R. (2003) Control of alternative pre-mRNA splicing by RNA Pol II elongation: faster is not always better. *IUBMB Life* **55:** 235-241.

Nilsen, T.W. (2003) The spliceosome: the most complex macromolecular machine in the cell? *BioEssays* **25**: 1147-1149.

Nurtdinov, R.N., Artamonova, I.I., Mironov, A.A., Gelfand, M.S. (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum Mol Genet.* **12(11)**: 1313-1320.

Pacheco, T.R., Gomes, A.Q., Barbosa-Morais, N.L., Benes, V., Ansorge, W., Wollerton, M., Smith, C.W., Valcarcel, J., Carmo-Fonseca, M. (2004) Diversity of vertebrate splicing factor U2AF35: identification of alternatively spliced U2AF1 mRNAs. *J Biol Chem.* **279(26)**: 27039-27049.

Palm, K., Metsis, M., Timmusk, T. (1999) Neuron-specific splicing of zinc finger transcription factor REST/NRSF/XBR is frequent in neuroblastomas and conserved in human, mouse and rat. *Brain Res Mol Brain Res.* **72(1)**: 30-39.

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

Pasquinelli, A.E., Ruvkun, G. (2002) Control of developmental timing by microRNAs and their targets. *Annu. Rev. Cell. Dev. Bio.* **18**: 495–513.

Ponting, C.P., Schultz, J., Milpetz, F., Bork, P. (1999) SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res.* **27(1)**: 229-232.

Pospisil, H., Herrmann, A., Bortfeldt, R.H., Reich, J.G. (2004) EASED: Extended Alternatively Spliced EST Database. *Nucleic Acids Res.* **32**: D70-74.

Qi, M., Byers, P.H. (1998) Constitutive skipping of alternatively spliced exon 10 in the ATP7A gene abolishes Golgi localization of the menkes protein and produces the occipital horn syndrome. *Hum Mol Genet.* **7(3)**: 465-469.

Relogio, A., Ben-Dov, C., Baum, M., Ruggiu, M., Gemund, C., Benes, V., Darnell, R.B., Valcarcel, J. (2005) Alternative splicing microarrays reveal functional expression of neuron-specific regulators in Hodgkin lymphoma cells. *J Biol Chem.* **280(6)**: 4779-4784.

Resch, A., Xing, Y., Alekseyenko, A., Modrek, B., Lee, C. (2004a) Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Res.* **32(4)**: 1261-1269.

Resch, A., Xing, Y., Modrek, B., Gorlick, M., Riley, R., Lee, C. (2004b) Assessing the impact of alternative splicing on domain interactions in the human proteome. *J Proteome Res.* **3(1)**: 76-83.

Sakai, H., Maruyama, O. (2004) Extensive search for discriminative features of alternative splicing. *Pac Symp Biocomput.* 54-65.

Sazani, P., Kole, R. (2003) Therapeutic potential of antisense oligonucleotides as modulators of alternative splicing. *J Clin Invest.* **112(4)**: 481-486.

Schultz, J., Milpetz, F., Bork, P., Ponting, C.P. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci USA* **95(11)**: 5857-5864.

Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P., Bork, P. (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* 2000 **28(1)**: 231-234.

Schwarze, U., Goldstein, J.A., Byers, P.H. (1997) Splicing defects in the COL3A1 gene: marked preference for 5' (donor) spice-site mutations in patients with exon-skipping mutations and Ehlers-Danlos syndrome type IV. *Am J Hum Genet.* **61(6)**: 1276-1286.

Schwarze, U., Starman, B.J., Byers, P.H. (1999) Redefinition of exon 7 in the COL1A1 gene of type I collagen by an intron 8 splice-donor-site mutation in a form of osteogenesis imperfecta: influence of intron splice order on outcome of splice-site mutation. *Am. J. Hum. Genet.* **65**: 336-344.

Singh, R., Valcarcel, J., Green, M.R. (1995) Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. *Science* 1995 **268(5214)**: 1173-1176.

Singh, N., Preiser, P., Renia, L., Balu, B., Barnwell, J., Blair, P., Jarra, W., Voza, T., Landau, I., Adams, J.H. (2004) Conservation and developmental control of alternative splicing in maebl among malaria parasites. *J Mol Biol.* **343(3)**: 589-599.

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

Sorek, R., Ast, G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* **13(7)**: 1631-1617.

Sorek, R., Shamir, R., Ast, G. (2004) How prevalent is functional alternative splicing in the human genome? *Trends Genet.* **20(2)**: 68-71.

Stojdl, D.F., Bell, J.C. (1999) SR protein kinases: the splice of life. *Biochem Cell Biol.* **77(4)**: 293-298.

Sugnet, C.W., Kent, W.J., Ares, M. Jr, Haussler, D. (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac Symp Biocomput.* 66-77.

Takahara, K., Schwarze, U., Imamura, Y., Hoffman, G.G., Toriello, H., Smith, L.T., Byers, P.H., Greenspan, D.S. (2002) Order of intron removal influences multiple splice outcomes, including a two-exon skip, in a COL5A1 acceptor-site mutation that results in abnormal pro-alpha1(V) N-propeptides and Ehlers-Danlos syndrome type I. *Am J Hum Genet.* **71(3)**: 451-465.

Taneri, B., Snyder, B., Novoradovsky, A., Gaasterland, T. (2004) Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific. *Genome Biology* **5**: R75.

Taneri, B., Novoradovsky, A., Snyder, B., Gaasterland, T. (2005) Databases for comparative analysis of human-mouse orthologous alternative splicing. *Lecture Notes in Bioinformatics* **3388**: 123-131.

Thanaraj, T.A., Clark, F., Muilu, J. (2003) Conservation of human alternative splice events in mouse. *Nucleic Acids Res* **31**: 2544-2552.

Thanaraj, T.A., Stamm, S., Clark, F., Riethoven, J.J., Le Texier, V., Muilu, J. (2004) ASD: the Alternative Splicing Database. *Nucleic Acids Res.* **32**: D64-D69.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., Darnell, R.B. (2003) CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302**: 1212-1215.

Valcarcel, J., Gebauer, F. (1997) Post-transcriptional regulation: the dawn of PTB. *Curr Biol.* **7(11)**: R705-708.

Valenzuela, A., Talavera, D., Orozco, M., de la Cruz, X. (2004) Alternative splicing mechanisms for the modulation of protein function: conservation between human and other species. *J Mol Biol.* **335(2)**: 495-502.

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

Venables, J.P. (2004) Aberrant and alternative splicing in cancer. *Cancer Res.* **64(21)**: 7647-7654.

Wang, H., Hubbell, E., Hu, J., Mei, G., Cline, M., Lu, G., Clark, T., Sinai-Rose, M.A., Ares, M., Kulp, D.C., Haussler, D. (2003) Gene structure-based splice variant deconvolution using a microarray platform. *Bioinformatics* **19**: 315–322.

Wang, X. (2004) Prediction and functional analysis of *Arabidopsis thaliana* microRNAs. Doctoral Dissertation. The Rockefeller University.

Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., Burge, C.B. (2004) Systematic identification and analysis of exonic splicing silencers. *Cell* **119(6)**: 831-845.

Weir, M., Rice, M. (2004) Ordered partitioning reveals extended splice-site consensus information. *Genome Res.* **14(1)**: 67-78.

Will, C.L., Luhrmann, R. (1997) Protein functions in pre-mRNA splicing. *Curr Opin Cell Biol.* **9:** 320-328.

Woodley, L., Valcarcel, J. (2002) Regulation of alternative pre-mRNA splicing. *Brief. Funct. Genomic Proteomic.* **1(3)**: 266-277.

Xing, Y., Resch, A., Lee, C. (2004) The multiassembly problem: reconstructing multiple transcript isoforms from EST fragment mixtures. *Genome Res.* **14(3)**: 426-41.

Xu, Q., Modrek, B., Lee, C. (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.* **30(17):** 3754-3766.

Yang, Y.Y, Yin, G.L., Darnell, R.B. (1998) The neuronal RNA-binding protein Nova-2 is implicated as the autoantigen targeted in POMA patients with dementia. *Proc Natl Acad Sci* USA **95(22)**: 13254-13259.

Yeakley, J.M., Fan, J.B., Doucet, D., Luo, L., Wickham, E., Ye, Z., Chee, M., Fu, X.D. (2002) Profiling alternative splicing on fiber-optic arrays. *Nature Biotechnology* **20**: 353-358.

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

Yeo, G.W., Van Nostrand, E., Holste, D., Poggio, T., Burge, C.B. (2005) Identification and analysis of alternative splicing events conserved in human and mouse. *Proc Natl Acad Sci USA.* **102(8)**: 2850-2855.

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

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

Zeeberg, B.R., Feng, W., Wang, G. Wang, M.D., Fojo, A.T. *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology* **4**: R28.

Zhang, L., Liu, W., Grabowski, P.J. (1999) Coordinate repression of a trio of neuron-specific splicing events by the splicing regulator PTB. *RNA.* **5(1)**: 117-130.

Zhang, W., Liu, H., Han, K., Grabowski, P.J. (2002) Region-specific alternative splicing in the nervous system: implications for regulation by the RNA-binding protein NAPOR. *RNA* **8:** 671–685.

Zheng, C.L., Nair, T.M,. Gribskov, M., Kwon, Y.S., Li. H,R., Fu, X.D. (2004) A database designed to computationally aid an experimental approach to alternative splicing. *Pac. Symp. Biocomput.* 78-88.

Zhou, Z., Licklider, L.J., Gygi, S.P., Reed, R. (2002) Comprehensive proteomic analysis of the human spliceosome. *Nature* **419:** 182–185.

Zhuang, Y., Ma, F., Li-Ling, J., Xu, X., Li, Y. (2003) Comparative analysis of amino acid usage and protein length distribution between alternatively and non-alternatively spliced genes across six eukaryotic genomes. *Mol Biol Evol.* **20(12)**: 1978-1985.