Bioinformatics Approaches to RNA Splicing

Aaron Levine

Thesis presented for the degree of

Master of Philosophy

University of Cambridge

and

The Sanger Centre
Wellcome Trust Genome Campus
Hinxton, Cambridgeshire

August 8, 2001

Summary

With the completion or near-completion of many large genome sequencing projects, automated annotation of vertebrate genomes has become an important research priority. Yet the complex intron/exon structure and the prevalence of alternative splicing of genes in higher organisms have rendered accurate gene prediction a difficult and still unsolved problem.

In order to aid in the ongoing genome annotation projects, I have developed a splice site prediction program, StrataSplice, which predicts both canonical (GT-AG) and minor variant (GC-AG) splice sites and is designed to integrate easily into a variety of gene prediction and annotation systems. StrataSplice utilises a new splice site prediction model that combines local GC content with a standard probabilistic pattern recognition technique and shows modest but significant improvement over standard splice site prediction models. Much of this improvement occurs in gene-rich high GC regions in which previous models perform more poorly.

U12-dependent introns are a distinct class of introns found in small numbers in the genomes of most higher eukaryotes, yet they have been largely ignored in genome annotation efforts. I have conducted a computational scan for these rare introns in the draft human genome sequence and generated a new reference set of 404 U12-dependent introns, an increase of 6-fold over the number previously available in all genomes. Analysis of these introns suggested that there is a significant error rate (>0.25 percent) at the acceptor site in U12-dependent splicing and that, in contrast to U2-dependent introns, U12-dependent introns may be recognised in an exclusively exon-dependent manner.

Chromosome 22 was the first human chromosome to be sequenced and has been subject to extensive experimental and computational annotation. Combining the predictions generated by StrataSplice with expressed sequence evidence, I have generated a set of 3,199 expressed sequence confirmed introns on chromosome 22. Nearly 80 percent of these introns were previously annotated, but the remaining 20 percent (671 introns) may help identify either alternatively spliced forms of known genes or previously unidentified genes.

Acknowledgements

I feel fortunate to have spent a year working in the Informatics group at the Sanger Centre and living at Churchill College and many members of both of these communities have helped me immensely during my time in Cambridge.

I am particularly grateful for the support and guidance of my supervisor, Richard Durbin, who introduced me to the field of probabilistic sequence analysis and helped to shape my varied and only semi-connected ideas into a coherent and feasible project

I am also grateful for the useful discussions I had with and the technical assistance I received from the other members of Richard's group, namely Irmtraud Meyer, Kevin Howe and Marc Sohrmann. I would like to thank Marc in particular for the generous loan of his laptop, which was of great assistance in the preparation of this thesis.

Alex Bateman and Sam Griffiths-Jones from the pfam group have been generally helpful, as has my University supervisor, Nick Goldman from the department of Zoology. Additionally, I am particularly indebted to Matt Pocock and Thomas Down, who have patiently answered numerous Java questions, James Gilbert, who has solved several frustrating Perl and postscript problems and Michele Clamp, who has provided invaluable assistance extracting data from Ensembl.

I would also like to thank the Café Churchill crowd at Churchill College for the good times, foosball, inspiration, and many helpful discussions on the true importance of U12 introns.

I was supported by a studentship from the Winston Churchill Foundation and the Sanger Centre is supported by the Wellcome Trust.

Preface

This thesis is the result of my own work and not the product of any collaboration. As with almost any scientific endeavour the work described herein builds on the research of countless others, and they are referenced throughout the body of this thesis and in a bibliography.

This thesis is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other University nor has any part of this thesis been submitted for any such degree, diploma or other qualification.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The discovery of the interrupted nature of many eukaryotic genes has to rank as one of the most startling of the era of molecular biology (Berget *et al.,* 1977; Chow *et al.,* 1977). In the last 25 years, remarkable progress has been made understanding the structure of eukaryotic genes and the complex intracellular machine, known as a spliceosome, which processes these interrupted genes to yield final, translatable mRNA products. However, recently the RNA splicing story has been growing more complex rather than less and numerous crucial questions remain unanswered.

For many years, simplicity prevailed and *all* eukaryotic mRNA introns were believed to be processed by the same mechanism. However, this picture of a single spliceosome recognising a simple and conserved set of sequence signals has broken down over the last decade. In its place, we are now aware of two distinct spliceosomes, each processing a disjoint subset of introns, defined not by clear and conserved signals, but by a variety of semi-conserved signals that combine to direct splicing in the cell.

Now as the field of genomics continues to grow, and large-scale analysis becomes an important, if not dominant, research paradigm, the need to improve understanding of RNA splicing becomes even more acute. For the recently completed and ongoing genome projects of higher organisms to reach their full potential, it is critical that researchers be able to accurately identify the protein coding regions of the genome and thus create from the genome sequence a catalogue first of all the genes and then eventually of all the proteins in an organism. And a prerequisite to identifying protein coding regions is an understanding of how the cell demarcates coding and non-coding regions at RNA splice sites.

Additionally, the recent revelations regarding the surprisingly small number of genes in the human genome (International Human Genome Sequencing Consortium, 2001; Venter *et al.,* 2001) have thrust the intron/exon structure of eukaryotic genes into the limelight as a potential generator of protein diversity (Graveley, 2001). In particular, the process of alternative splicing, due to the potentially combinatorial increase in protein diversity that can result from it, has been hypothesised to account for much of the complexity apparently missing from the relatively gene-poor genome sequence (Modrek *et al.,* 2001).

In short, although RNA splicing has been extensively studied both from an experimental and an informatics perspective over the last 20 or so years, much remains to be learned. Thts thesis addresses a variety of RNA splicing and intron-based analyses,

covering both major and minor class introns and hopefully will prove useful to the ongoing project of annotating the human and other vertebrate genome sequences.

Chapter 2 of this thesis is a brief review, both of the mechanism of RNA splicing and of computational approaches to identifying splice sites. It is intended as an introduction to the non-specialist and should make chapters 3 though 5 more accessible to readers not familiar with either the biology of RNA splicing or basic principles of pattern recognition in DNA sequences.

Chapter 3 discusses the identification of RNA splice sites from genomic DNA sequences. Two new splice site prediction models are introduced, one utilising hgher-order dependencies within the splice site signal and one that considers local GC content in its predictions. This latter model forms the basis of a splice site prediction program called StrataSplice that is discussed at the end of the chapter.

Chapter 4 discusses a scan for members of the rare subclass of U12-dependent introns in the draft human genome sequence. Many new U12-dependent introns were identified and analysis of their properties led to several interesting observations.

Chapter 5 briefly discusses a scan for introns on human chromosome 22. Over three thousand introns were identified of which 20 percent were not part of current gene models and these should prove helpful in the ongoing project to annotate chromosome 22.

Finally chapter 6 concludes the thesis by discussing briefly new directions that could be taken, should this research be continued.

Chapter 2

Introns and RNA Splicing: a short review

2.1 An overview of human gene structure

The complexity of gene structure in hgher organisms is one reason that automated annotation of the human and other large genomes has proven difficult (International Human Genome Sequencing Consortium, 2001; Zhang, 1998; Guigo *et al.*, 2000). The average human gene covers nearly 30 kb of genomic sequence and consists of several promoter signals and numerous splice sites as well as at least one transcription start, translation start, translation stop and polyadenylation site. These features determine how mature messenger RNAs (mRNAs) are produced from the genomic DNA sequence and play important roles in the processing of the gene into a final protein product (see Figure 2.1).

In lower eukaryotes, such as *S. cerevisiae*, most genes produce a single mRNA containing a continuous coding sequence flanked by short untranslated regions (UTRs). In contrast, many, if not most, human genes produce multiple messages, typically with long UTRs and interrupted by intervening sequences called introns that are spliced out during mRNA processing. The mRNA sequences that are spliced together when the introns are excised are known as exons and these form the final mRNA transcript. Internal exons tend to be short, with most less than 300 bp in length, whde introns, in contrast, vary greatly in size but are generally longer, with a mean size of nearly 3,400 bp (see Table 2.1). Terminal exons, at the beginning and end of mRNA transcripts, can be significantly longer than internal exons, and these long exons are found quite frequently in the *3'* UTR.

	Median	
Internal exon length	122 bp	145 bp
Number of exons	7	
Intron length	1,023 bp	3,365 bp
Coding sequence length	1,100 bp	1,340 bp
Genomic extent	14 kb	M

Table 2.1 - Characteristics of human genes. The median and mean values for a number of charactensucs of human genes are provided. Data from (International Human Genome Sequencing Consortium, 2001).

Human gene structure varies within the genome as well. For instance, while exon length remains relatively constant across a variety of GC content levels, intron length decreases dramatically in regions of high GC content (International Human Genome Sequencing consortium, 2001).

Exon/intron boundaries are known as splice sites, and as more and more genes

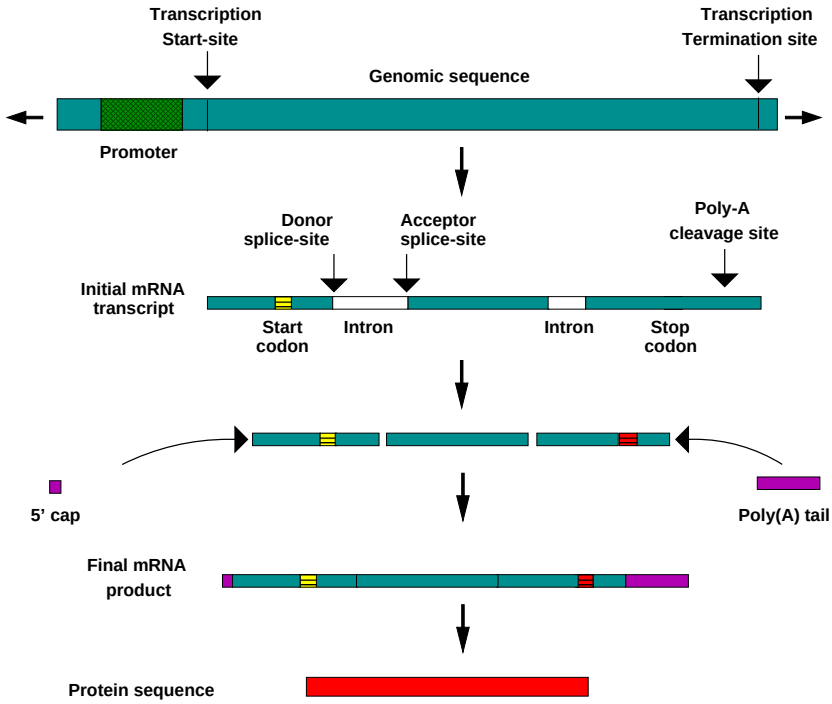


Figure 2.1 - The major components of a typical protein-coding eukaryotic gene and the role these components play in the processing of a nascent transcript into a mature mRNA are shown. (Figure courtesy of K. Howe).

were sequenced, it was quickly noticed that most splice sites had similar sequences (Mount, 1982). In particular, almost all introns started with a GT dinucleotide and ended with an AG dinucleotide. These bases became known as the canonical dinucleotides and formed the basis of the GT-AG rule for introns. Additionally several other bases near splice site junctions show varying degrees of conservation and these form the consensus sequences for donor and acceptor splice sites (see Figure 2.2).

Although more than 99 percent of introns obey the GT-AG rule, both GC-AG and AT-AC are known to be valid intron boundaries as well (Jackson, 1991, Burset *et al.*, 2000). In addition to the splice site consensus sequences, a typical intron has two other semi-conserved sequences, a polypyrimidine tract upstream of the acceptor signal and a branch site region upstream of the polypyrimidine tract.

2.2 The RNA splicing process

Introns are excised from pre-mRNA by a large complex consisting of five snRNAs and 50-100 proteins. A comprehensive review of RNA splicing and spliceosome formation is beyond the scope of this chapter, but I will provide a brief overview of the splicing process and make note of some interesting recent results. For a readable

introduction to splicing, I direct the reader to chapter 22 of Lewin's *Genes VII* (2000), whilst for more detailed coverage, I refer the reader to a number of recent reviews below.

In brief, the excision of a single intron from a nascent pre-mRNA transcript is a two-step process requiring two distinct transesterification reactions. Initially cleavage occurs at the 5' splice site and the first base of the intron forms a lariat by binding to an adenosine nucleotide at the branch site, upstream of the 3' splice site. Next a new phosphodiester bond is formed between the last base of the upstream exon and the first base of the downstream exon and the intron is released as a product of this reaction (reviewed in Burge *et al.,* 1999).

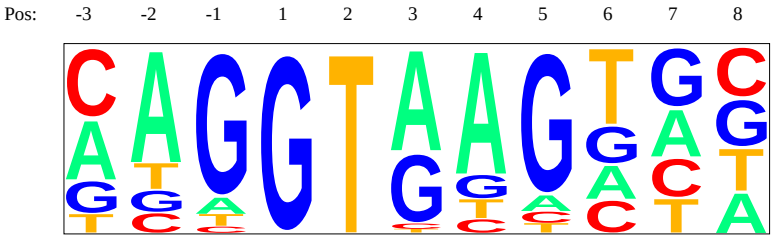
The reactions described above occur within the spliceosome complex, which is responsible for the crucial tasks of recognising the appropriate splice sites and catalysing the splicing reactions. The spliceosome consists largely of five RNA-protein complexes known as small nuclear ribonucleoprotein particles (snRNPs).

The first step in splicing is typically the ATP-independent recognition of the 5' splice site by the U1 snRNA and the association of the U1 snRNP with this region, which results in the formation of the commitment (E) complex. This interaction, while thought to occur at the vast majority of introns, is not strictly required as some introns have been identified in which splicing proceeds efficiently in the absence of the U1 snRNP *in vitro* (Crispino *et al.*, 1994, 1996; Tarn and Steitz, 1994).

A key role of the U1 snRNP is to promote the association of the U2 snRNP with the branch point region of the intron. U2 snRNP association depends on two key interactions: reception of the polypyrimidine tract region by the protein $U2AF^{65}$ and the recently discovered interaction between the protein $U2AF^{35}$ and the intron's terminal AG dinucleotide (Zorio and Blumenthal, 1999; Wu *et al.,* 1999; reviewed in Reed, 2000; Moore, 2000). The association of the U2 snRNP with the branch point region is an ATP-dependent process in which at least six proteins, components of the essential splicing factors SF3a and SF3b, bind either upstream or downstream of the branch point region (Kramer *et al.,* 1999; reviewed in Reed, 2000). The association of both the U1 and U2 snRNPs defines complex A (the pre-spliceosome).

Association of the tri-snRNP complex containing the U4, U5 and U6 snRNPs with the pre-spliceosome is required to form the B complex. Recently, the splicing factor SPF30 has been shown to play a key role in the integration of the tri-snRNP complex into the pre-spliceosome although this transition remains poorly defined (Rappsilber *et*

a



b

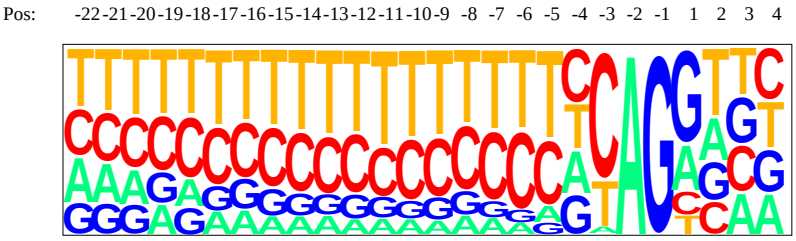


Figure 2.2 - Consensus sequences at canonical donor (a) and acceptor (b) splice sites. The size of each base represents the relative frequency at the specified position. Position one represents the first base of the intron for donor sites and the first base of the exon for acceptor sites. Graphical representations of frequencies were generated using Pictogram (Burge,C., unpublished, available at http://genes.mit.edu/pictogram.html).

al., 2001; Meister *et al.*, 2001). Although the full role of the tri-snRNP complex is not understood, it interacts with both the 5' and 3' ends of the intron and contains the highly conserved splicing factor Prp8, which is thought to contribute to the catalytic core of the spliceosome (reviewed in Collins and Guthrie, 2000). The exact nature of the catalytic core remains an open question.

Tri-snRNP addition initiates a series of RNA-RNA rearrangements, most notably the displacement of the U1 snRNA from the 5' splice site by the U6 snRNA, which create the catalytically competent C complex (reviewed in Burge *et al.*, 1999; Konarska, 1998). Recent results have found that many of these rearrangements are directed by RNA helicases in the DExD/H box family (Staley and Guthrie, 1999; reviewed in Schwer, 2001).

In yeast, the organism in which most of the interactions detailed above were worked out, introns are generally short and often interrupt larger exons. In this situation, the spliceosome is thought to form directly around the intron, in a process known as intron recognition (Talerico and Berget, 1994). In contrast, most human genes have short exons interrupted by long introns, potentially causing problems for an intron recognition

based splicing system. However, an alternate model, known as exon recognition is thought to function in the splicing of longer introns; in this model, the spliceosome assembles initially around the shorter exon sequence as opposed to around the intron (Berget, 1995). Recognition in both models involves splicing-associated SR proteins, which are believed to play an important role in bridging the sequence between neighbouring splice sites and bringing spliceosome components together (reviewed in Gravely, 2000).

2.3 U12-dependent introns and the U12 spliceosome

Careful analysis of splice junctions in the early 1990's revealed a small number of introns with highly unusual donor and acceptor sites containing AT and AC in place of the typical GT and AG (Jackson, 1991; Hall and Padgett, 1994). Experimental work quickly verified suggestions that this subset of introns was excised by a novel spliceosome and characterisation of the so-called U12 spliceosome, whch contains the U11, U12, U4atac, U6atac and U5 snRNPs, began (Tarn and Steitz, 1996a, 1996b).

Many genes contain both U2- and U12-dependent introns but little is known about how the two spliceosomes cooperate to identify and splice the correct introns *in vivo.* Distinct differences are observed, however, between the splice site signals associated with the two types of introns. U12-dependent introns exhibit strongly conserved and informative donor and branch signals, whereas U2-dependent introns exhibit only moderately informative signals at the donor and acceptor sites and a highly degenerate branch site signal. Additionally the polypyrimidine tract seen between the branch site and acceptor site of U2-dependent introns is lacking, or is at least significantly weaker (see Chapter 4), in U12-dependent introns.

The evolutionary history of these two classes of introns and their respective spliceosomes remains unclear. Burge *et al.* (1998) have reported both intron subtype switching (e.g. conversion from AT-AC to GT-AG termini among U12-dependent introns) and U12- to U2-dependent intron conversion and concluded that U12-dependent introns tend to convert to U2-dependent over evolutionary time. They also reported a biased distribution of U12-dependent introns within a variety of genomes, a result they found suggestive of a fission-fusion model of spliceosome evolution in whch the U2 and U12 systems diverged in separate lineages and were later united through a merging of genetic material in a progenitor of hgher eukaryotes (Burge *et al.,* 1998).

Recent results have found a strikingly high degree of overlap between the proteins and non-coding RNAs involved in U2- and U12-dependent splicing. In addition to the U5 snRNA (Tarn and Steitz, 1996a), all 8 snRNP Sm proteins (Will *et al.,* 1999), the 4 proteins that constitute the splicing factor SF3b (Will *et al.,* 1999), and the splicing-associated protein Prp8 (Luo *et al.,* 1999) have been found in both the U2 and U12 spliceosomes. Recent evidence has indicated that splicing-associated SR proteins, long known to function in the major spliceosome, play functional roles in U12-dependent splicing as well (Hastings and Krainer, 2001). Extensive similarity in secondary structures and interactions between the set of non-coding RNAs U11, U12, U4atac and U6atac involved in the U12 spliceosome and the set U1, U2, U4 and U6 involved in the U2 spliceosome argue for homology of the two systems as well (Burge *et al.*, 1998) as do recent results that have found the stem-loop structures of U6 and U6atac to be functionally analogous (Shukla and Padgett, 2001). Although the evolutionary implications of this high degree of overlap are not entirely clear, these findings may indicate that the U12 spliceosome evolved in the presence of the U2 spliceosome rather than in a different lineage as the fission-fusion model suggests (Will *et al.*, 1999).

As more genomes have been sequenced, U12-dependent introns have been identified in a variety of hgher organisms, including human, mouse, fly and arabidopsis. Interestingly, U12-dependent introns seem to be entirely lacking from the model organisms *S. cerevisiae* and *C. elegans.*

24 Computational analysis of RNA splicing

Ever since the recognition of consensus signals for RNA splice sites, research into computational identification of splice sites and, thus, toward the determination of gene structure has been quite active. Originally, splice sites were identified simply by "eyeballing" a DNA sequence and looking for matches to the consensus splice site sequences. However, it quickly became apparent that many functional splice sites shared only a few bases of similarity and more sophisticated computer models were required.

Simple independent weight matrices, or frequency tables, whch yield a probabilistic log-odds score for each base at each position in a sequence, were one of the first methods developed and still prove useful today (see Figure 2.2, Staden, 1984; Harr *et al.,* 1983). Weight matrices and the many derivatives of this method require a training set of true sites to generate the frequency table and then score potential sites by summing the scores of individual bases in a pre-defined window. The incorporation of

dependencies between neighbouring bases (first-order dependencies) into the weight matrix framework represents one of the most significant advancements on this simple predictive framework (Zhang and Marr, 1993).

Although several new approaches, such as finite state automata (Kudo *et al.*, 1987) and neural networks (Brunak and Engelbrecht, 1991), were developed to identify splice sites in the 1980's and early 1990's, the next models to gain widespread use were not introduced until 1997 as components of the hghly successful GENSCAN gene prediction system (Burge and Karlin, 1997). Maximal dependence decomposition, which GENSCAN uses to identify donor splice sites, is a tree-based decomposition approach that breaks down donor sites into a set of classes, based on dependencies between bases in the splice site signal, and uses a simple weight matrix to model each class individually (Burge, 1998). The GENSCAN system uses a new model for acceptor sites as well, termed a windowed weight array method, whch models the branch point region using a modification of the first-order dependencies approach that groups sets of neighbouring bases together in order to avoid problems caused by limited data (Burge, 1998).

More recent approaches have tended to integrate multiple signals into the prediction process. For instance, GeneSplicer (Pertea *et al.*, 2001) combines a traditional log-odds score based on a slight variant of maximal dependence decomposition, a measure of local coding potential and a local optimality requirement. Although combining these signals does yield improvements in splice site identification, the utility of this approach for gene prediction is more questionable, as many gene prediction systems already consider the additional signals.

Progress has been made recently as well on the problem of identifying the precise splice site from among a number of nearby, or proximal, false positives, a problem that has significant implications for automated genome analysis. One promising approach uses decision trees to discriminate between true sites and proximal false sites and may prove useful for annotation efforts (Thanaraj and Robinson, 2000). However, the current position is that all these methods generate very large numbers of false positive predictions. Typical behaviour for the analysis of genomic sequences is roughly 12 false positives per kb if thresholds are set to include 99 percent of true sites and 6 false positives per kb if thresholds are set to include 95 percent of true sites.

Finally, the large expressed sequence datasets that have been generated in the last few years have permitted the compilation of EST-confirmed splice sites on a large scale

and facilitated analysis of both canonical and non-canonical introns (Burset *et al.*, 2000; International Human Genome Sequencing Consortium, 2001).

Chapter 3

Identifying Splice Sites in Human Genomic DNA Sequences

Summary

The presence of conserved sequences at splice sites has been well documented over the last 25 years. However, these sequences are not sufficiently informative to permit unambiguous identification of gene structure. Gene prediction programs, such as GENSCAN (Burge and Karlin, 1997), combine splice site predictions with other information to predict complete gene structures. This chapter describes two novel models for the identification of canonical splice sites (sections 3.1 and 3.2) and one model, which applies standard methodology to identify the most frequent non-canonical splice site (section 3.3). The chapter concludes with a discussion of a human splice site predictor, Stratasplice, which incorporates the best of these models and should prove useful for genome annotation. This analysis led to the observation that splice sites in GC-rich regions of the genome are slightly different from, and harder to predict than, splice sites in GC-poor regons.

3.1 A block dependence model for donor site identification

Introduction

Probabilistic signal recognition relies on the detection of differences between a training set of confirmed signals and a control set. Simple models, which detect, for instance, only the order of individual nucleotides, require relatively small training sets, while more complex models, which may consider overlapping pairs or groups of nucleotides, necessitate much larger sets of training data. Traditionally, a major stumbling block in the development of splice site detectors has been the shortage of reliable training data. However, the recent publication of SpliceDB (Burset *et al.*, 2001), which contains more than 15,000 confirmed human splice site pairs has largely alleviated this concern.

Previous reports have suggested that, in addition to dependencies between neighbouring bases, the donor splice site contains longer range dependencies, perhaps relating to the binding of the U1 snRNA to the donor site (Burge and Karlin, 1997). This analysis attempts to quantify these longer-range interactions and take advantage of the information they provide to improve *ab initio* splice site identification.

Materials & Methods

Test sets

Training and evaluation sets were generated from the 15,263 confirmed canonical human splice site pairs in SpliceDB (Burset *et al.*, 2001). 786 donor sites and 1,295 acceptor sites with poor or incomplete sequence data were removed from this set, yielding a total of 14,477 confirmed 5' splice sites and 13,968 confirmed 3' splice sites. A control set of genomic DNA used to calculate null model frequencies was extracted from the first 10 kb of repeat-masked DNA chosen from 100 randomly selected Ensembl clones (International Human Genome Sequencing Consortium, 2001). Sets of "false" splice sites were generated by extracting sequences around GT, GC or AG dinucleotides in this random set of genomic DNA. (Some small fraction of these sites will in fact be true).

Independence and first-order dependence models

Two classic pattern recognition techniques, independent weight matrices (Staden, 1984) and first-order dependent weight matrices (Zhang and Marr, 1993) were re-implemented for comparative purposes. These two models yield log-likelihood scores for each potential splice site by comparing the frequency of either individual nucleotides (independent model) or dinucleotides (first-order model) at each position in the splice site window with background genomic frequencies. Given a sequence $X = \{x_1, x_2, \ldots, x_n\}$, scores were derived from each model as follows:

$S(X) = \sum_i \log_2 \frac{f^i_{x_i}}{q_{x_i}}$	Independence Model
$S(X) = \sum_i \log_2 \frac{f^i_{x_i \mid x_{i-1}}}{q_{x_i \mid x_{i-1}}}$	First-order dependence Model

where $f^i_{x_i}$ is the frequency of base x_i at position i in the training set, $f^i_{x_i|x_{i-1}}$ is the frequency of base x_i at position i following base x_j at position $i-1$ and q_{x_i} and $q_{x_i|x_{i-1}}$ are genomic nucleotide and dinucleotide frequencies, respectively.

Detection rates

Detection rate curves (see Figure 3.3, for example), which illustrate a model's

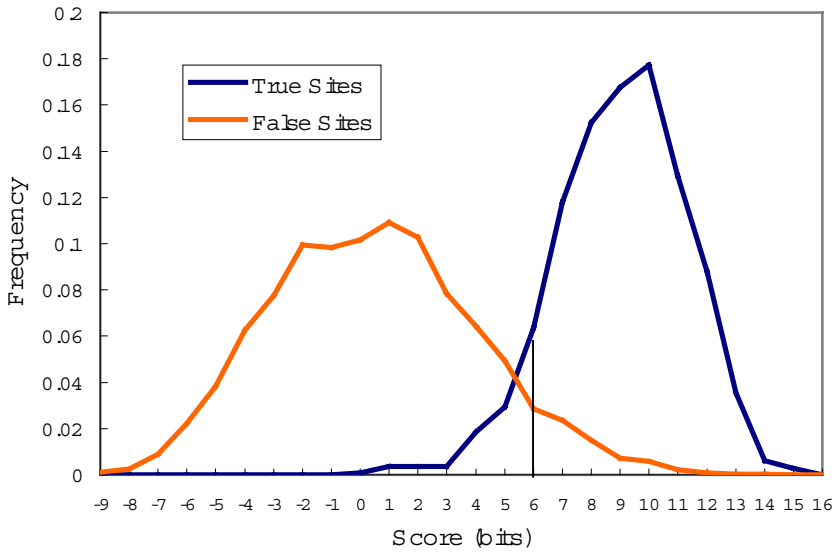


Figure 3.1 - Typical log-odds score distributions for evaluation and control sets. The vertical line, drawn here at an arbitrary threshold of 6 bits, divides the true distribution (blue line) into true positives and false negatives and the false distribution (red line) into true negatives and false positives. Distributions such as this one were used to generate the detection rate curves as described in Materials and Methods.

performance at a variety of threshold values, were used to compare the performance of the various models. The fraction of true sites included and the false positive rate are calculated directly from the evaluation and false sets, respectively. Using 6 bits as the threshold value (illustrated by the vertical bar in Figure 3.1), a point (*x*,*y*) on the detection rate curve would be calculated as follows:

$$x = C(\mathit{false} > 6) * (10{,}000 / G)$$
$$y = C(\mathit{true} > 6) / C(\mathit{true})$$

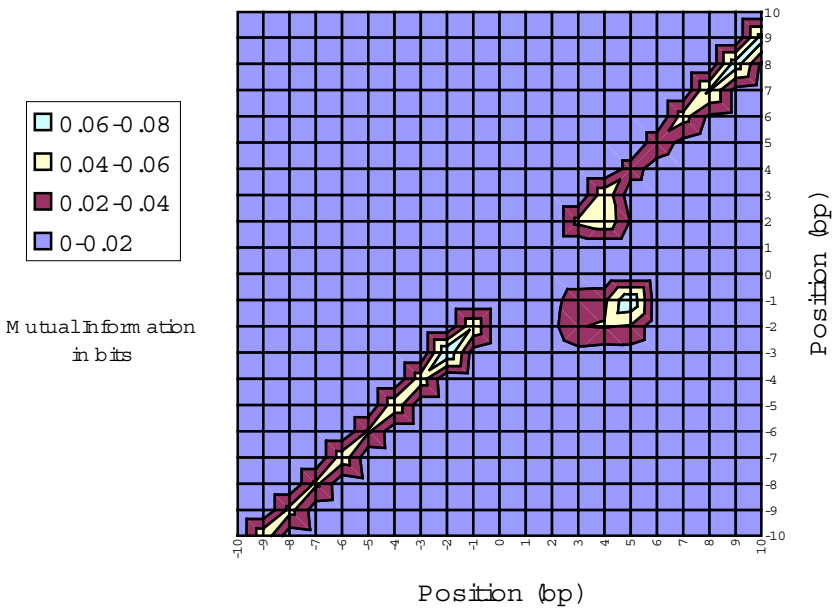
where C(condition) represents a simple conditional count and *G* is the amount of genomic DNA from which the relevant set of false sites was extracted.

Mutual information analysis

To identify dependencies between non-neighbouring bases, the mutual information was calculated between all pairwise combinations of bases in donor and acceptor splice sites using the SpliceDB dataset (Burset *et al*, 2000). The mutual information

$$M(i,j) = \sum_{a,b} f(x_i = a, x_j = b) \log_2 \frac{f(x_i = a, x_j = b)}{f(x_i = a) f(x_j = b)}$$

a



b

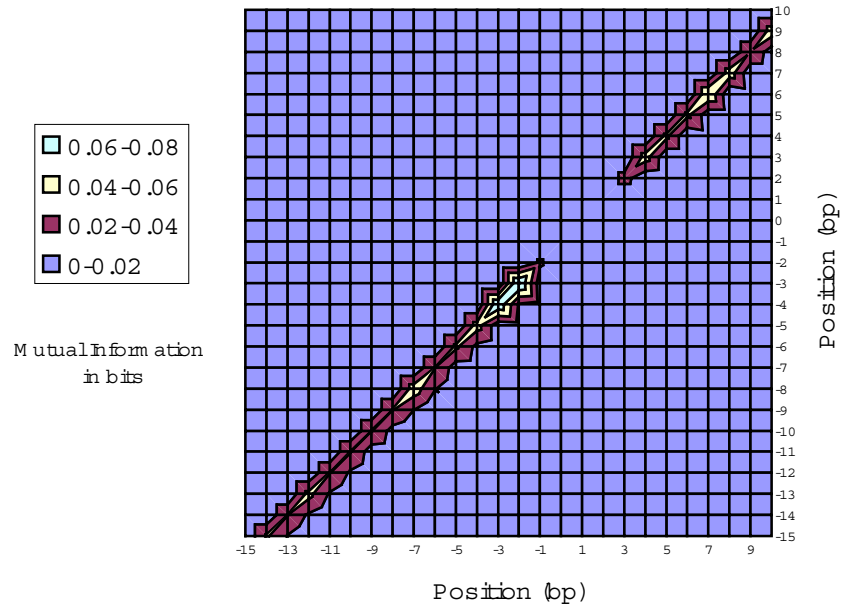
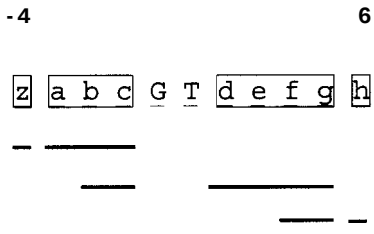


Figure 3.2 - Mutual information around donor (a) and acceptor (b) splice sites. Positions in the donor site window [-10, +10] and acceptor site window [-15, +10] are shown along both axes. The canonical G is at position 0 in (a) and position -1 in (b). The diagonal line present in both (a) and (b) represents the mutual information between neighbouring pairs of bases.

between two positions yields a value in bits indicating the degree of dependence between positions i and j (Durbin *et al.,* 1998).

Score calculations

Based on the mutual information results (see Figure 3.2a) a model was derived whtch divided the region around the donor splice site into blocks as shown below:



Thts model was scored using log-likelihood scoring considering the conditional probabhties of the blocks above (dependencies indicated by the horizontal black lines) and using genomic dinucleotide frequencies for the null model. 'Thus, the score of a sequence X in bits is

$$S(X) = \log_2 \frac{f(abc \mid z) * f(defg \mid bc) * f(h \mid fg)}{q(z)q(a \mid z)...q(c \mid b) * q(d)q(e \mid d)...q(h \mid g)}$$

Frequency values for each possible base combination of each block given its dependencies in the model were calculated by adding pseudocounts based on genomic dinucleotide frequencies to the observed counts. 'Thus, for example,

$$f(abc \mid z) = \frac{C(x_{-4} = z, x_{-3} = a, x_{-2} = b, x_{-1} = c) + 4^3 q(a \mid z)q(b \mid a)q(c \mid b)}{C(x_{-4} = z) + 4^3}$$

Results

Mutual information analysis (see Figure 3.2a) revealed a fair amount of information (> 0.3 bits) between non-neighbouring bases in donor splice sites and a novel block dependence model of donor splice sites was developed to take advantage of this information. 'This model showed moderate improvement over first-order dependence and independent weight matrix models for prediction of canonical donor splice sites (see Figure 3.3).

Mutual information analysis was also performed on the acceptor splice site dataset, but no significant information was found between non-neighbouring bases (see Figure 3.2b).

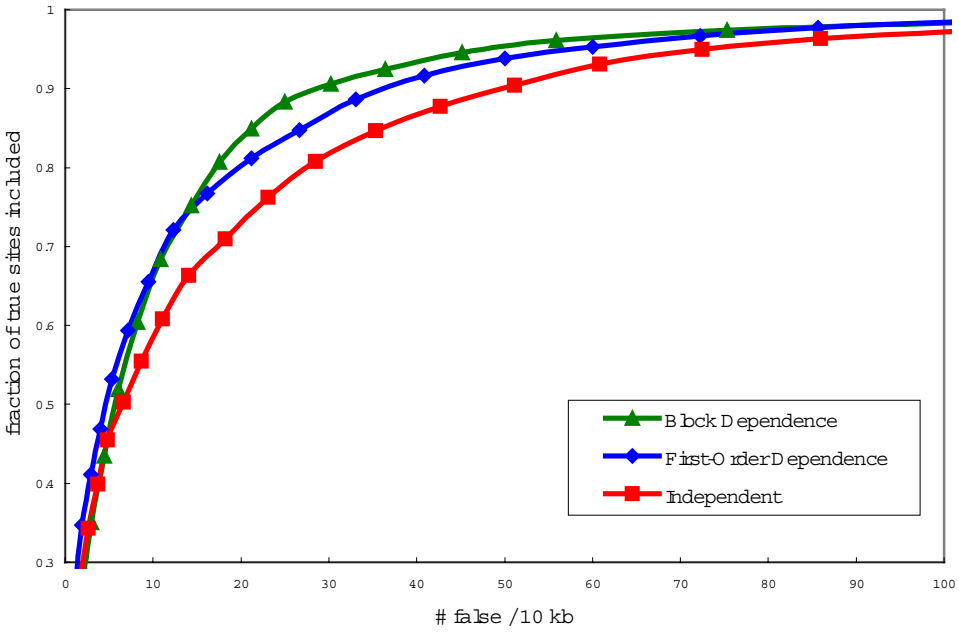


Figure 3.3 - Comparison between independent, first-order dependence and block dependence models for donor splice site identification. The block dependence model predicted fewer false positives at most sensitivity levels than did the other two models.

Discussion

The block dependence model described here shows that splice site signal recognition can be improved by considering higher order and long-range interactions. However, this improvement is quite modest when compared to the first order dependence model. Other splice site identification models including maximal dependence decomposition, the model used by the GENSCAN gene prediction program (Burge and Karlin, 1997), also show modest improvements over a first-order dependence model and the block dependence model presented here is unlikely to represent a substantial improvement over these models.

3.2 Improved splice site identification by considering local GC content

Introduction

Analysis of the draft sequence of the human genome has confirmed the observation that large regions of the genome deviate from the genome-wide average GC level of 41 percent (International Human Genome Sequencing Consortium, 2001). Huge regions (>10 Mb) were observed which differed significantly from the average, while smaller regions (20 kb) showed more variation with GC content levels ranging from 30

to 65 percent. 'These results suggest that traditional probabilistic signal recognition techniques, such as those used to identify RNA splice sites, whch identify differences between a positive model and a genomic null model, are likely to suffer from substandard performance in regions where the actual genome sequence differs greatly from the genome average. Conversely, more accurate modelling of background DNA composition should allow for more accurate discrimination of true splice signals.

The most successful gene prediction programs, such as GENSCAN (Burge and Karlin, 1997) fit different coding models and length distributions for regions of different GC content, but I am not aware of previous stratification of splice site models. I describe here an approach to splice site identification, whch extends the first-order dependence weight matrix technique (Zhang and Marr, 1993) by stratifying the prediction process according to local GC content. This yields significantly improved performance, particularly in GC-rich and, thus, gene-rich areas (Zoubak *et al.,* 1996).

Materials & Methods

Test sets were derived and detection rate curves were generated as described in section 3.1.

GC stratification

Canonical donor (GT) and acceptor (AG) splice sites were stratified by local GC content according to the base composition in the total surrounding sequence (generally 80 bases, excluding 8 bases immediately around the splice junction) included in SpliceDB. The control set was stratified accordmg to the GC content in 300 base chunks. During sequence scans (and during derivation of the false set from the genomic set) potential splice sites were stratified according the base composition in the 75 bases preceding and following an eight-base window centred on the splice site itself.

Splice site windows

I chose splice site windows that included all positions significantly deviating from random background frequencies on the basis of relative entropy calculations (Durbin *et al.,* 1998) and were expanded to convenient sizes.

$$Entropy(i) = f(x_i)\log_2\frac{f(x_i)}{q(x_i)}$$

This yielded windows from –10 to +10 around GT donor sites (canonical G at position 0) and –25 to +5 around AG acceptor sites (canonical A at position 0).

Score calculations

Log-likelihood scoring was used to generate a log-odds score for each potential splice site (Durbin *et al.,* 1998). Conditional frequency values for each dinucleotide pair at each position in the splice site window ($f^i_{a|b}$) were determined by adding pseudocounts to the observed values as follows:

$$f^i_{a|b} = \frac{C^i_{a|b} + 4q_{a|b}}{\sum_b C^i_{a|b} + 4}$$

where $C^i_{a|b}$ is the observed count of base *a* occurring at position *i* following base *b* at position *i*-1 and $q_{a|b}$ is the observed conditional frequency for the appropriate control set. Observed conditional frequencies in the appropriate stratified control set were used for the null model. One model was trained (*e.g.* calculation of both *f* and *q* values) for each stratum of each splice signal. Score values in bits for a sequence $X = \{x_1, x_2, ..., x_n\}$ were derived from the appropriate frequency data as follows:

$$S(X) = \sum_i \log_2 \frac{f^i_{x_i|x_{i-1}}}{q_{x_i|x_{i-1}}}.$$

Prior probability estimation

'The prior probability that a given GT or AG dinucleotide defined a true splice site was calculated using estimates of the total number of GT dinucleotides, AG dinucleotides and introns in the genome. The estimates of the total number of each dinucleotide were generated by counting dinucleotides on one strand of 2 MB of random genomic sequences (10 kb chunks from 200 randomly selected clones) and scaling this value to fit the 3000 MB genome. An estimate of 400,000 introns in the genome was generated by considering a genome consisting of 40,000 genes where each gene had an average of 10 introns. Limiting the analysis to one strand and scaling this number by the overall frequencies of each of the various types of splices sites (e.g. 99.24% GT-AG, 0.69% GC-AG, etc) reported in SpliceDB (Burset *et al.,* 2000) allowed the calculation of "per strand estimates" for each splice site type. Dividing by the corresponding total number of the relevant dinucleotide estimated per strand yielded the final priors ($P(T)$, see 'Table 3.1). As gene densities vary with GC content (Zoubak *et al.*, 1996), the prior probabilities for GT and AG dinucleotides were scaled accordmg to the frequencies of true ($f(T|\mathrm{GC})$) and false splice sites ($f(F|\mathrm{GC})$) at each GC level as follows:

$P(T \mid \mathrm{GC}) = \frac{f(T \mid GC)}{f(F \mid GC)} P(T)$. The necessary GC-level dependent frequency values were derived from the stratification of the true and false splice site sets (see Figure **3.4).**

	GT Donor Sites		AG Acceptor Sites	
Prior		1.37e-3		9.94e-4
Posterior Threshold	Sensitivity	Specificity	Sensitivity	Specificity
-∞	100	0.1	100	0.1
1e-6	99.8	0.3	99.8	0.3
le-5	99.6	0.4	99.7	0.4
le-4	99.1	0.7	99.0	0.6
lc-3	96.6	1.5	96.2	1.1
1e-2	84.8	3.5	70.4	3.6
5e-2	58.4	7.4	16.3	10.1
le-1	41.5	10.5	0.0	**N/A**

Table 3.1 - Performance of the stratified splice model on genomic sequences. Prior probabilities and sensitivity and specificity values of the stratified splice model at various posterior probability threshold values are indicated. Sensitivity and specificity values are provided as percentages and are calculated assuming an intron density of 67/MB (see Materials and Methods).

Posterior probability calculations

Posterior probability values, which incorporate prior biological information into a statistical framework (Durbin *et al.*, 1998), were used to generate probability values that combined the log-odds scores and the estimated prior probabhties for each potential splice site. Bayes' theorem was used:

$$P(T \mid S(X) = s) = \frac{P(S(X) = s \mid T)P(T)}{P(S(X) = s \mid T)P(T) + P(S(X) = s \mid F)(1 - P(T))}$$

where $P(S(X) = s \mid T)$ reads the probability that the score of sequence X is s, given the knowledge that the sequence is a true splice site and $P(T)$ is the scaled prior probability described above.

The conditional probability values used in the posterior calculation were calculated from the strata-specific distributions of true and false splice sites (see Figure 3.1, for example), assuming that these distributions were Gaussian. In brief, the mean and standard deviation were estimated for each distribution using standard formulas and the conditional frequency values were taken from the hypothetical Gaussian distribution that these two values defined. This approach led to more robust estimation of values in the tails of the distribution than simply using the observed values due to the small number of data points in the tails.

Model evaluation

The choice of GC-level boundaries for the stratification process was evaluated by a modified version of the equivalence number statistic (EN), which summarises the selectively and specificity of a given model by comparing the number of false positives and true negatives (Pearson, 1995). In this situation, I use probability distributions rather than raw numbers, and define the equivalence number as the frequency of false positives when the log-odds bit threshold is set to equalise the frequency of false positives and true negatives. As my model seeks to minimise both false positives and true negatives, the lower the EN value, the better the model. In order to take into account the effects of stratifying the prediction process, the final EN value was a weighted average of the EN values of each individual model. Weighting was done accordmg to the frequency of true splice sites within each stratum. Given a stratified splice prediction model with n strata (e.g. $M = \{m_1, m_2, ... m_n\}$) the final EN value would be

$$EN(M) = \sum_i EN(m_i) * f(T \mid m_i).$$

In order to maximise use of the available data, I used a jack-knife procedure in which the avadable data was divided into four sets. Pour training and evaluation cycles were performed holding out each set for evaluation in turn and using the other three sets for training. The results of these four cycles were averaged to produce the final value.

Sensitivity and specificity

Sensitivity and specificity values were determined using the posterior values generated when the model was trained and evaluated using disjoint subsets of the set of all true sites in SpliceDB and on all false sites in the genomic control set. A jack-knife procedure identical to the one described above was used and final values are the average of four different training and evaluation cycles. Sensitivity was calculated as the ratio of true positives to all true sites. Specificity calculations depended on an estimate of the density of introns in the genome. Two values were used: 67 introns/MB (consistent with the intron density estimates for the prior probability calculations) and 563 introns/MB (the intron density of GENSCAN's evaluation set). Using these estimates I scaled the total number of observed true positives to the expected number in a set the size of the control set and then calculated the specificity as the ratio of true positives to true positives plus false positives.

Results

Previous studies have indicated that GC-rich regions of the human genome are also gene-rich (Zoubak *et al.,* 1996). This is reflected in the distribution of GC content levels near intron splice sites (see Figure 3.4). As expected the GC distribution around GT dinucleotides in a random genomic sample (my control set) was approximately normally distributed with a peak near 40 percent GC. Only 10 percent of background GT dinucleotides are in areas of 60 percent GC or greater. In contrast, 27 percent of true donor splice sites are located in areas of 60 percent GC or greater. Similar results were seen for AG acceptor sites (data not shown).

To determine whether splice site signals had the same composition across the full range of GC content levels, all true splice sites from SpliceDB (Burset *et al.,* 2001) were divided into 3 groups based on the surrounding sequences (excluding 8 bp around the actual splice junction) and simple frequency tables were derived around the splice sites (see Tables 3.2a,b). Interestingly the donor site signal is largely conserved across all GC levels except for the thud base in the intron, which changes from 71 percent **A** and 23 percent C in the low GC content group to 33 percent A and 62 percent G in the high GC content group. **A** similar though less dramatic change involving C and T nucleotides is seen for the thud base of the intron (just before the AG) at acceptor sites as well. The polypyrimidine (C|T) tract found upstream of acceptor splice sites is biased toward C in high GC content regions and toward T in low GC content regions (data not shown).

To explore whether splice site identification could be facilitated by considering local GC content, I developed a splice site identification model based on the first-order dependence weight matrix approach (Zhang and Marr, 1993), which stratifies both the training data and the null model data according to local GC content. Figures 3.5a and 3.6a use detection rate curves (as described in section 3.1) to compare the performance of two standard weight matrix models and the new stratified model. Strikingly, at GT donor sites, the new stratified model outperformed the non-stratified first-order dependence model as least as dramatically as this model outperformed the independent weight matrix model. Less dramatic, but still useful, improvements were seen for the acceptor site model.

To explore the reasons behind these improvements, I compared the performance of the stratified and the non-stratified first order model on stratified test sets (see Figures 3.5b, 3.6b). These graphs indicate the relative performance of each predictor on splice

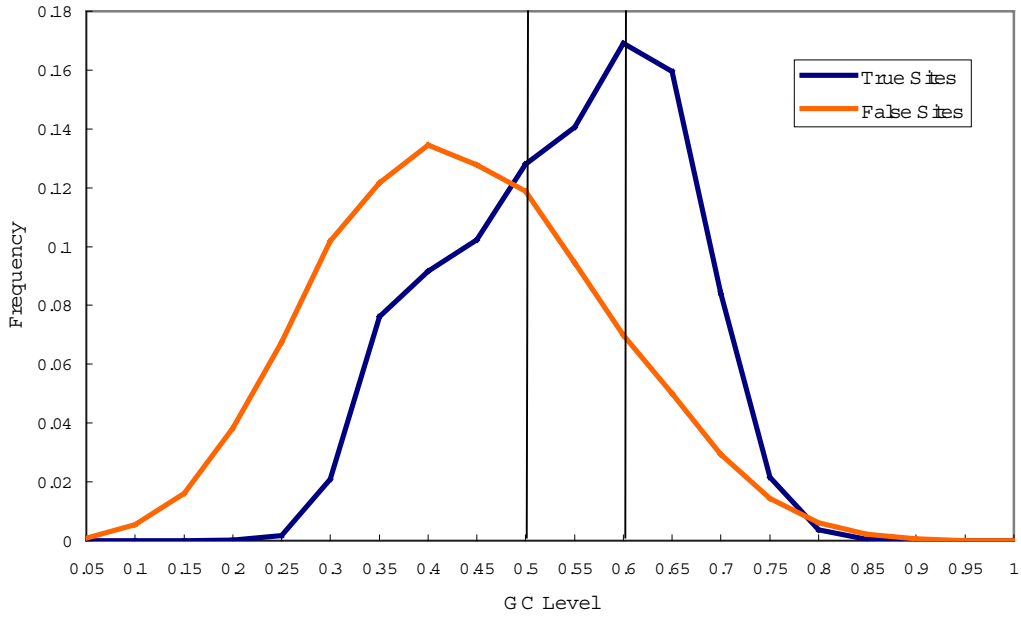


Figure 3.4 - Local GC content near true and false splice sites. The red line indicates the frequency distribution of local GC content near GT sites in the control set, while the blue line indicates this same distribution for true GT donor sites in SpliceDB. The black vertical bars indicate the final stratification boundaries used in the model.

sites in each GC stratum. Interestingly, for the non-stratified donor site predictor, I found significant differences in my ability to accurately identify splice sites according to the stratum. In particular, splice site prediction was easiest in low GC content regions, slightly harder in medium GC content regions, and significantly harder in high GC content regions (Figure 3.5b, compare dotted lines). Using the stratified predictor, I observed only minor improvements in the low and medium strata but found a striking improvement in performance in the high GC stratum.

Breaking down the performance by strata at acceptor sites led to slightly different results (see Figure 3.6b). For the non-stratified model, both the low and medium strata showed similar performance profiles, while the high GC stratum showed significantly worse performance (Figure 3.6b, compare dotted lines). The stratified model yielded improvements across all three strata with the high GC stratum showing the most dramatic improvement. However, improvement for this stratum was not as dramatic as it was for the donor site predictor.

To further quantify the performance of this new splice site identification model, sensitivity and specificity values were calculated at a variety of posterior probability thresholds. Results are shown for two intron density levels. Table 3.1, which uses 67

a

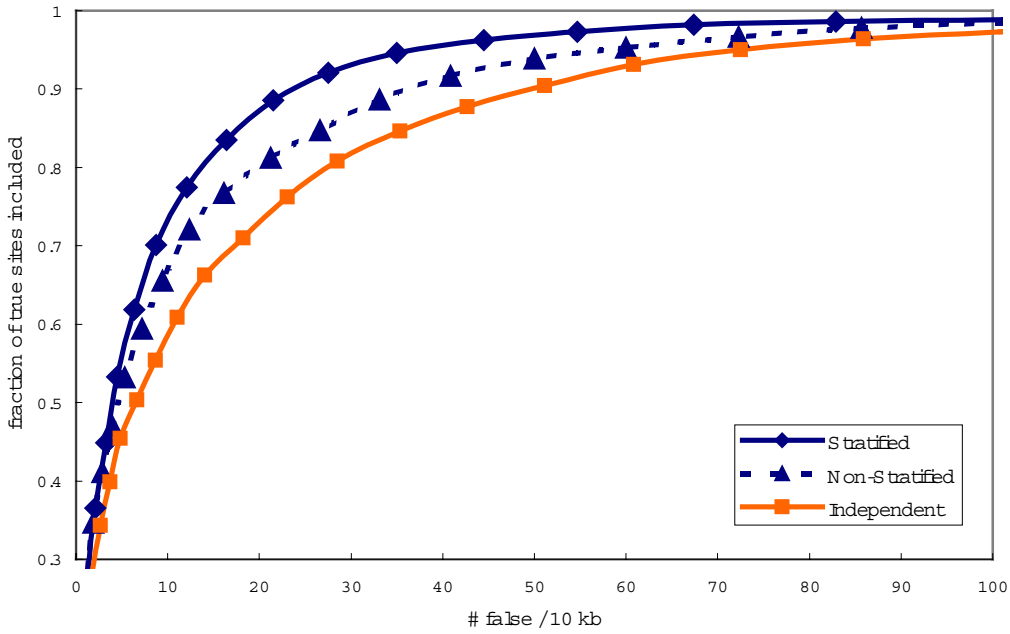
	Low GC Content							
A	0.64	0.11	0.00	0.00	0.71	0.72	0.10	0.20
C	0.10	0.03	0.00	0.00	0.02	0.05	0.05	0.11
G	0.10	0.78	**1.00**	0.00	0.23	0.09	0.76	0.16
T	0.16	0.08	0.00	**1.00**	0.04	0.14	0.09	0.54
	Medium GC Content							
A	0.61	0.08	0.00	0.00	0.44	0.71	0.06	0.15
C	0.13	0.04	0.00	0.00	0.04	0.08	0.06	0.18
G	0.13	0.81	**1.00**	0.00	0.50	0.13	0.84	0.26
T	0.13	0.07	0.00	**1.00**	0.02	0.08	0.04	0.41
	High GC Content							
A	0.55	0.08	0.00	0.00	0.33	0.69	0.03	0.10
C	0.18	0.03	0.00	0.00	0.03	0.11	0.06	0.24
G	0.15	0.83	**1.00**	0.00	0.62	0.15	0.88	0.26
T	0.12	0.06	0.00	**1.00**	0.01	0.05	0.03	0.39
	A	G	**G**	**T**	A|G	A	G	T

b

	Low GC Content						
A	0.10	0.27	0.07	**1.00**	0.00	0.27	0.26
C	0.22	0.22	0.56	0.00	0.00	0.12	0.16
G	0.05	0.17	0.00	0.00	**1.00**	0.50	0.18
T	0.62	0.34	0.37	0.00	0.00	0.11	0.40
	Medium GC Content						
A	0.07	0.22	0.03	**1.00**	0.00	0.22	0.21
C	0.41	0.36	0.78	0.00	0.00	0.15	0.21
G	0.06	0.22	0.00	0.00	**1.00**	0.52	0.22
T	0.46	0.20	0.19	0.00	0.00	0.11	0.35
	High GC Content						
A	0.04	0.17	0.02	**1.00**	0.00	0.21	0.18
C	0.49	0.39	0.87	0.00	0.00	0.14	0.24
G	0.09	0.30	0.00	0.00	**1.00**	0.57	0.28
T	0.38	0.14	0.11	0.00	0.00	0.08	0.30
	C|T		C|T	**A**	**G**	G	

Table 3.2 –Nucleotide frequencies at stratified donor (a) and acceptor (b) splice sites. Splice sites from SpliceDB (Burset *et al.*, 2001) were divided into three groups, less than 50 percent GC, 50-60 percent GC and greater than 60 percent GC, based on local GC content in an 80 bp window around the splice site, and the frequencies of bases at positions surrounding the splice junction are shown. The splice site consensus sequence is shown in the bottom row. Bold text indicates the canonical dinucleotide. Red text indicates bases that show large changes in frequency between strata.

a



b

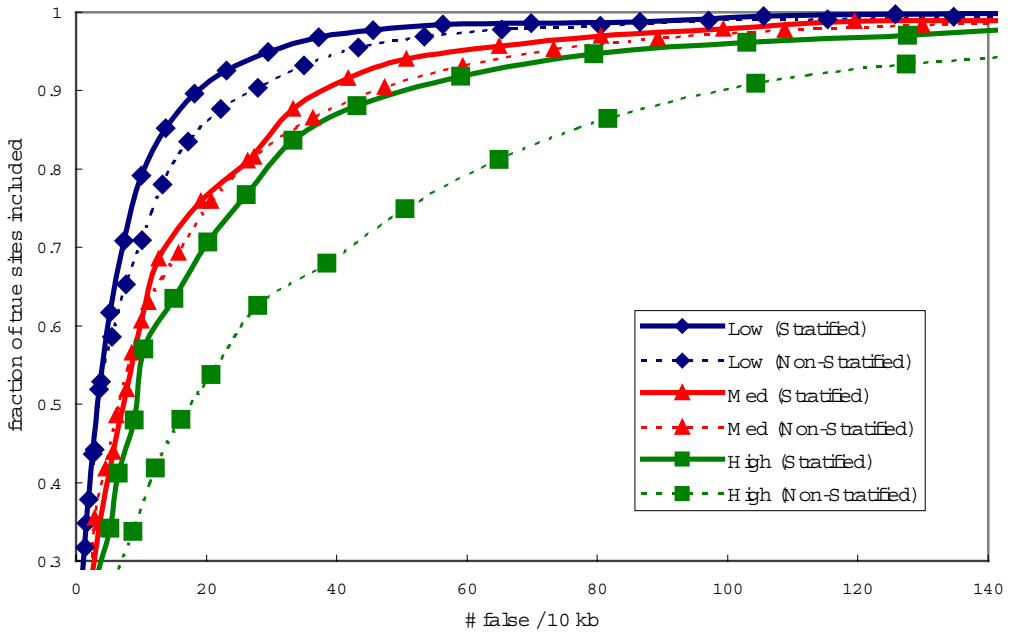
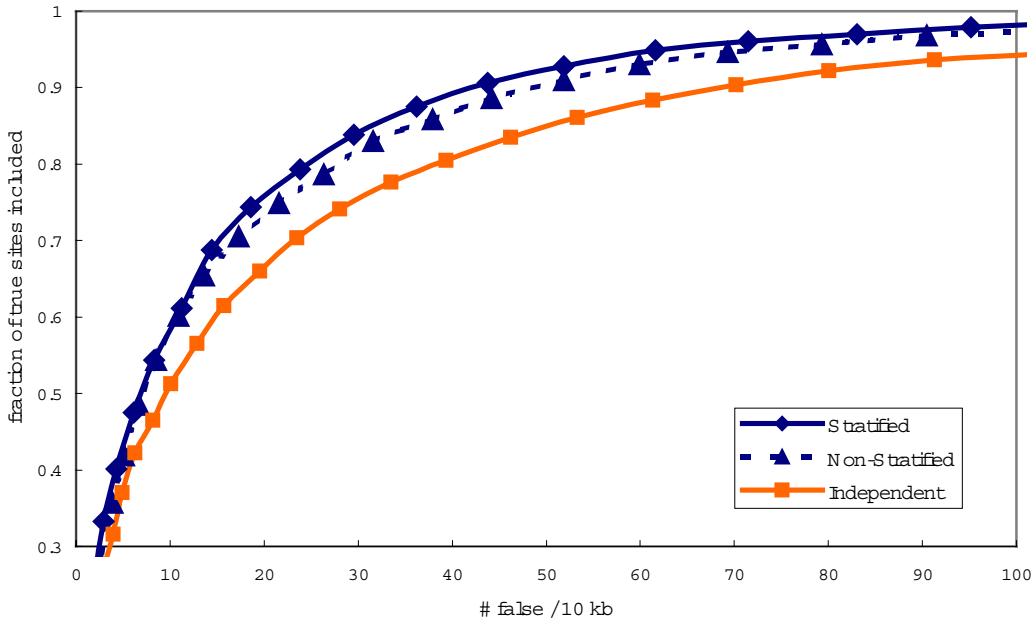


Figure 3.5 - Performance of the stratified splice predictor at GT donor sites. (a) Performance comparison with first-order dependence model (non-stratified) and independent model. (b) Performance comparison with first-order dependence model (non-stratified) on stratified test sets.

a



b

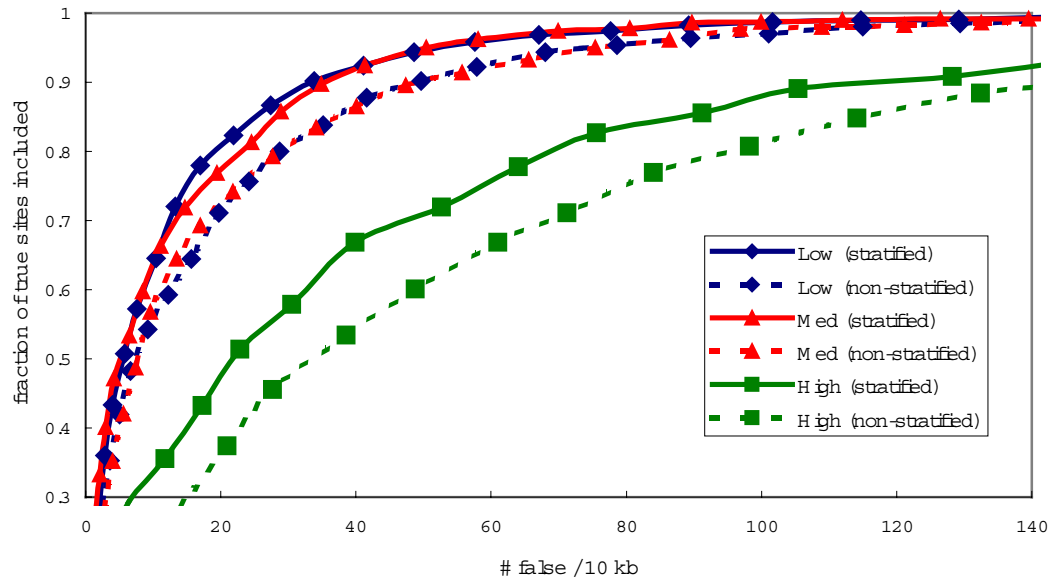


Figure 3.6 – Performance of the stratified splice predictor at AG acceptor sites. (A) Performance comparison with first-order dependence model (non-stratified) and independent model. (B) Performance comparison with first-order dependence model (non-stratified) on stratified test sets.

introns/MB, represents the performance expected on typical genomic DNA sequences whde Table 3.3 uses the much higher density of 563 introns/MB for comparative purposes with the GENSCAN splice site predictors. As Table 3.3 indicates, the stratified splice predictor described here generally gives higher specificity values for a given sensitivity level at both donor and acceptor splice sites than the GENSCRN splice site predictors.

	GT Donor Sites		AG Acceptor Sites	
Sensitivity	Genscan	Stratified	Genscan	Stratified
99	N/A	5.9	N/A	4.6
95	8.7	13.1	5.6	9.5
90	13.4	18.4	8.8	13.7
50	**36.0**	**43.4**	33.8	30.4

Table 3.3 – Comparison between GENSCAN's splice site predictors and my new stratified splice model. The table indicates specificity values as percentages at **the** indicated sensitivity level. GENSCAN values are taken from (Burge, 1998). Stratified splice model values are calculated assuming the same intron density (563 introns / Mb) as GENSCAN's evaluation set (Burge, 1997). Specificity values for GENSCAN at 99 percent sensitivity have not been published.

In theory, the stratification process can divide the data into any number of strata, but in practice limited data means only four or five models can be reliably trained. A variety of different strata boundaries were explored and evaluated using a modified version of the equivalence number metric (described in Materials and Methods), whch indicates the frequency of false positives when a threshold is selected to balance the frequency of false positives and true negatives. Switching from a non-stratified model to a three-stratum (< 50 percent, 50-60 percent, > 60 percent) model decreased the equivalence number from 12.7 percent to 10.3 percent for the GT model and from 12.0 percent to 10.7 percent for the AG model. Similar results were seen for other three-stratum models and for models with four or five strata. Thus, a three-stratum model with the boundaries at 50 and 60 percent GC was selected.

Discussion

'The simple approach presented here of stratifying the data and training a first-order model for each stratum outperforms higher order models, such as those used in GENSCAN and section 3.1 and shows that stratification by local GC content levels is a powerful technique for improving genomic signal recognition. Although some differences were observed among the consensus sequences of splice sites after stratifying by GC content, much of the improvement seems to be due to the improved null model, whch was generated by stratifying the control set. This observation suggests that similar

stratification approaches may yield significant improvements in other signal detection problems, such as promoter motif detection.

The observation that splice site identification is more difficult in GC-rich regions of the genome than it is in GC-poor regions is quite intriguing, particularly considering the correlation between GC content and intron length seen in the human genome (International Human Genome Consortium, 2001). Although my results were derived *in silico*, they seem likely to indicate a biological reality; splice site consensus signals provide less information in GC-rich regions of the genome than they do in others. These observations lead to the enticing hypothesis that splice site recognition by the spliceosome may be a significant constraint on intron evolution, particular in GC-rich regions. Short introns are not associated with GC-rich regions in all vertebrate genomes (Hurst *et al.,* 1999), however, and it would be interesting to consider other hgher organisms to see if these support the observed association.

It is worth observing as well that the problem under consideration here, namely the identification of RNA splice sites from genomic sequence is rather more difficult than that whch the cell performs *in vivo.* Whereas I must attempt to identify splices sites from raw genomic sequence, the cell must only accurately identify splice sites on pre-mRNA. If roughly a quarter of the genome is transcribed (Venter *et al.,* 2001), the splicing machinery in the cell has a search space reduced 8 fold in size (the extra factor of *two* comes because mRNA is single stranded). This partially explains the low specificity scores seen for the genomic analysis (see Table 3.1) and emphasises the importance of considering as much evidence as possible when predicting genes. Systems such as GAZE (Howe and Durbin, unpublished) whch can integrate splice site predictions from one source with promoter predictions from another as well as homology and comparative information seem likely to be the way forward in automated gene prediction.

3.3 − Identifying non-canonical GC donor sites

Introduction

Recent analyses have indicated that roughly 0.7 percent of human introns start with the non-canonical dinucleotide GC in place of the much more common GT (International Human Genome Sequencing Consortium, 2001; Burset *et al.,* 2001). However most gene prediction packages do not consider GC as a potential donor site and miss several thousand introns for this reason. Additionally, automated analysis pipelines, such as the Ensembl project, whch are becoming increasingly important

gateways to the human genome sequence, have few of these non-canonical, but still relatively common, introns annotated correctly (M. Clamp, personal communication). Thts chapter describes the development and performance of a simple GC donor site model, whtch should prove useful for genome annotation.

Materials & Methods

Test sets

A training set of 122 true GC donor sites was derived from the set of 270 EST-confirmed and verified non-canonical introns included in SpliceDB (Burset *et al.,* 2001). Control and false sets were generated as described in section 3.1.

First-order dependence model

A first order dependence weight matrix splice site predictor (Zhang and Marr, 1993) was implemented as described in section 3.1 and trained using the training set of GC donor sites.

Prior and posterior probabilities

A prior probability for GC dinucleotides was derived as in section 3.2 except that no corrections were made for local GC content. Posterior probabhties were calculated as in section 3.2.

Results

A first-order dependence weight matrix was built from the training set and used to score sets of true and false GC donor sites. Although GC donor sites are roughly 100-fold less common than GT donor sites, and the prior probability is therefore roughly 100-fold less, performance (see Table 3.4) is only marginally worse at GC sites when compared to GT sites. The GC model is difficult to evaluate accurately due to limited data, but specificity of GC donor predlctions tends to be roughly 15-25 fold worse than for GT donor predictions at a given sensitivity level. For instance at a threshold that includes roughly 96 percent of all true sites, 150 out of 10,000 predicted GT donor sites will be true whde 7 out of 10,000 predicted GC donor sites would be true.

Although the GT and GC donor consensus sequences are similar, the GC donor consensus is more highly conserved and contains nearly 13 bits of information, as opposed to approximately 8 bits of information in the GT donor consensus. For this

	GT Donor Sites		GC Donor Sites	
Prior		1.37e-3		1.12e-5
Posterior Threshold	Sensitivity	Specificity	Sensitivity	Specificity
-∞	100	0.1	100	0.001
le-6	99.8	0.3	100	0.008
le-5	99.6	0.4	100	0.02
le-4	99.1	0.7	95.9	0.07
le-3	96.6	1.5	76.2	0.3
1e-2	84.8	3.5	48.4	1.7
5e-2	58.4	7.4	27.9	6.9
le-1	41.5	10.5	17.2	76.8

reason, the GC model developed here should be expected to significantly outperform the simple approach of identifying GC donor sites by simply replacing the 'T' with a C in a standard GT donor site model.

Discussion

The model described here for identifying GC donor sites is interesting not because it is novel or complex, but because it is immediately useful. Many of the 2000 or so genes with GC introns may have been incorrectly annotated during the early stages of automated genome analysis. Yet, if the goal of delineating the full collection of human genes is to be achieved, these genes, whch contain non-canonical introns, must be included.

Although the number of false positives is high for GC donor site identification, this result is not unexpected, nor is it a major concern. Many gene prediction systems are tailored to work with a large set of predlctions and can combine a variety of types of evidence to separate true and false signals.

3.4 – Stratasplice: A human splice site predictor

Introduction

Stratasplice is a stand-alone splice site predictor designed for use on human genomic sequences. It utilises the stratified splice site identification model described in section 3.2 to identify canonical GT and AG splice sites and the model described in

section *3.3* to identify non-canonical GC donor sites. Stratasplice utilises a Bayesian probabilistic framework and reports both log-odds bit scores and posterior probabhties for all of its predlctions. For easy integration into gene prediction systems such as DOUBLESCAN (Meyer and Durbin, unpublished) or GAZE (Howe and Durbin, unpublished), Stratasplice accepts fasta files as input and outputs its predlctions in GFF format (see http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml). Stratasplice is avadable free of charge from the Sanger Centre website at http://www.sanger.ac.uk/software/analysis/stratasplice/).

Program usage

StrataSplice is a command-line program written in Java 1.2 (available from http://java.sun.com). It has been extensively tested on a variety of Unix platforms but should run on Windows and other environments that support Java as well. Stratasplice is provided as a Java Archive file (**.jar**) and is run in its default mode as follows:

```
java -fast -jar Stratasplice.jar filename
```

In addition to the filename, whch should be the full path to any valid fasta file containing one or more sequences, a number of parameters (see Table 3.5) may be used to customise StrataSplice's performance. The order of the parameters is not important as long as each parameter is provided at most one time and all parameters precede the file name.

Flag	Type	Description	Default
s	Numeric	Log-odds bit threshold	Negative infinity
p	Numeric	Posterior probability score threshold	*0*
g	String	Genomic data file	–1 Mb taken in 10 kb chunk from 100 random genomic clones
a	String	AG true file	training file derived from SpliceDB
b	String	AG false file	training file derived from genomic data set
c	Numeric	AG prior probability	9.94e-4
d	String	GT true filc	training file derived from SpliceDB
e	String	GT false file	training file derived from genomic data set
f	Numeric	GT prior probability	1.37e-3
h	String	GC true file	training file derived from SpliceDB
i	String	GC false file	training file derived from genomic data set
j	Numeric	GC prior probabhty	1.12e-5

Chapter 4

A Computational Scan for U12-Dependent Introns in the Human Genome Sequence

Summary

U12-dependent introns are found in small numbers in most eukaryotic genomes, but their scarcity makes accurate characterisation of their properties challengmg. A computational search for U12-dependent introns was performed using the draft version of the human genome sequence. Human expressed sequences confirmed 404 U12-dependent introns w i t h the human genome, a 6-fold increase over the total number of non-redundant U12-dependent introns previously identified in all genomes. Although most of these introns had AT-AC or GT-AG terminal dinucleotides, small numbers of introns with a surprising diversity of termini were found, suggesting that many of the non-canonical introns found in the human genome may be variants of U12-dependent introns and, thus, spliced by the minor spliceosome. Comparisons with U2-dependent introns revealed that the U12-dependent intron set lacks the "short intron" peak characteristic of U2-dependent introns, suggesting that U12-dependent introns may be recognised exclusively in an exon dependent manner. Analysis of this U12-dependent intron set confirmed reports of a biased distribution of U12-dependent introns in the genome and allowed the identification of several alternative splicing events as well as a surprising number of apparent splicing errors. This new larger reference set of U12-dependent introns will serve as a resource for future studies of both the properties and evolution of the U12 spliceosome.

Introduction

Two distinct types of pre-mRNA introns, termed U2- and U12-dependent based on the spliceosome complexes that excise them during RNA processing, are found in most higher organisms (reviewed in Burge *et al.*, 1999). While the roughly 99.9% of introns spliced by the major (U2-dependent) spliceosome have been extensively characterised (International Human Genome Sequencing Consortium, 2001; Zhang, 1998), less is known regarding the remaining 0.1% of introns, whch fall into the U12-dependent class. This minor class of introns was originally identified due to its unusual conserved donor and branch signals and highly atypical AT-AC terminal dinucleotides (Jackson, 1991; Hall and Padgett, 1994). More recently, analyses have found that AT-AC termini are not strictly required and identified many U12-dependent introns with GT-AG terminal dinucleotides as well as a few with other termini (Sharp and Burge, 1997; Burge *et al.,* 1998; Wu and Krainer, 1999). Additionally, a small number of U2-dependent introns with U12-like AT-AC terminal dinucleotides have been identified, confirming

that analysis of the entire splice site signal and not just the terminal dmucleotides is required for accurate classification (Dietrich *et al.,* 1997).

Although U12-dependent introns have been identified previously through homology searches and by analysing annotated intron junctions (Burge *et al.,* 1998), the limited number of U12-dependent introns available to researchers remains a major factor hindering understanding of this rare class of introns. The analysis presented here represents the first large-scale search for U12-dependent introns in the recently completed human genome sequence. A greater than expected diversity in the terminal dmucleotides of U12-dependent introns was observed, giving further evidence to the idea that flexibility in these positions has played an important role in intron evolution (Burge *et al.,* 1998; Deitrich *et al.*, 1997). This analysis generated a new reference set of human U12-dependent introns eight-fold larger than the previously available set and allows a more extensive characterisation of these introns to be carried out.

Materials and Methods

Human U12-dependent introns were identified using a two-step procedure. First potential donor and branch site signals were identified based on statistical pattern recognition techniques. Low threshold values that detected almost all known sites whde accepting a large number of false positives were used. Prom these signals, potential introns (donor/acceptor pairs) were generated and expressed sequence evidence was used to identify a subset of these potential introns as valid. All genomic scans used the 9 January 2001 assembly of the 7 October 2000 freeze of the human genome draft sequence (International Human Genome Sequencing Consortium, 2001; available from `http://genome.cse.ucsc.edu/`).

Candidate U12-dependent intron donor and branch sites were identified using a standard weight matrix approach (Staden, 1984). The weight matrix models were trained using a previously described non-redundant set of 48 U12-dependent introns from a variety of species (Sharp and Burge, 1997). Simple pseudocounts based on genomic nucleotide frequencies (the null model) were added during the training process to avoid overfitting the model to the training data. Any sequences whose log-odds scores from the donor signal weight matrix exceeded an empirically derived bit threshold were considered as potential U12-dependent intron donor sites. Potential U12-dependent acceptor sites were identified by considering all high scoring branch signals (again using an empirically derived threshold) and includmg only those that had a putative acceptor site (an AC dinucleotide, for instance) within a certain distance range from the putative

branch site. The traditional consensus branch site for U12-dependent introns is TTCCTTAA, although my search pattern extended slightly beyond thts consensus and none of the bases were strictly required in my analysis. All pairs of potential donors and acceptors that met the above criteria and were within a certain distance of each other were considered to define potential U12-dependent introns. For each of these cases, 64 bp of potential exon sequence, 32 bp from before and 32 bp from after the hypothetical intron were extracted and saved for later analysis.

The analysis described above involved five parameters: a donor site score threshold (9 bits), a branch site score threshold (6 bits), both a minimum and a maximum branch site to acceptor site &stance (8 bp, 21 bp), and a maximum intron size (20 kb). The first four of these were selected to be as inclusive as possible (based on the training data) while still minimising time required for computation, while the final parameter, maximum intron size, had to be limited to relatively small values to render the analysis computationally tractable. The analysis did, therefore, overlook some longer U12-dependent introns (see Discussion). After confirmation of introns, the distributions of donor scores, branch scores and the branch to acceptor distance were plotted and showed approximately normal distributions with the thresholds well separated from the peaks (see Figure 2 and data not shown), suggesting that the empirical thresholds did not eliminate a large number of valid results. Parameter values for the GT-AG and AT-AC scans are provided above; parameter values for all scans are provided in the legend to Table 4.1.

Expressed sequence data were used to confirm a small portion of the large set of potential U12-dependent introns as true introns. For this purpose a specialised human expressed sequence database was developed whtch contained 54,484 human mRNA sequences from EMBL release 65 (Baker *et al.,* 2000) and 3,268,161 human ESTs from dbEST downloaded from the NCBI on 28 February 2001 (Boguski *et al.*, 1993; available from **ftp://ncbi.nlm.nih.gov/genbank/).**

High-speed SSAHA similarity searches (Ning, Z., Cox, A.J. and Mullikin,J.C., in press) were performed looking for matches between each potential U12-dependent intron and a repeat-masked version of the database described above. Repeat masking was performed using DUST (Tatusov and Lipman, unpublished). SSAHA (version 1.1) was used with the following options: wordlength, 13; minprint, 39; maxstore, 50000; reportmode, replaceC. The results of this search were parsed to include only those expressed sequence matches that extended at least 15 bp on both sides of the

hypothetical splice junction. Two potential introns were considered duplicate if they showed identical sequences along the full 64 bp of potential exon regions. Although such a situation could potentially result from gene duplication events and represent a valid intron, redundancy in the draft sequence assembly presents an equally plausible explanation. Accordingly only one copy of each potentially duplicate intron was saved for further analysis. Introns supported by a variety of SSAHA matches extending from at least position 3 to position 61 were considered verified at this point. As SSAHA functions in a phased manner and does not necessarily report the full length of the sequence match, introns whch showed support but did not meet this stringent SSAHA criterion were analysed using BLAST (Altschul *et al.*, 1997, version 2.0.6, installed locally). Introns supported by a perfect BLAST match over all 64 bp were considered as verified. The remaining set of candldate introns, whch showed some support but met neither the SSAHA nor the BLAST criteria were examined and classified manually.

Scans were performed for standard U12-dependent introns with AT-AC and GT-AG terminal &nucleotides as well as a variety of non-standard introns (see Table 4.1). Non-standard donor signals were identified using modified training sets, whch had, for instance, each GT dinucleotide at the donor position replaced with a GC &nucleotide. Non-standard acceptors were identified by using the original branch site training set but scanning the downstream region after high scoring branch sites for the non-standard dinucleotide of interest.

Non-standard splice junctions were checked for possible ambiguities in the form of cases where a single expressed sequence could support a variety of splice junctions, as previously described (Burset *et al.*, 2000). No such cases were found.

Distributions of U12-dependent introns in the genome were modelled using binomial distributions as previously described (Burge *et al.*, 1998).

Results

Characteristics of human U12-dependent introns

Scans of the human genome draft sequence were performed to identify both typical AT-AC and GT-AG U12-dependent introns and atypical U12-dependent introns with a variety of other splice junctions (see Table 4.1). The searches for AT-AC and GT-AG introns examined all candldate introns up to 20 kb in length whde the other searches only examined potential introns of up to 2 kb in length. Accordingly atypical introns are likely to be somewhat underrepresented in my results. Unlike the only previous large

Intron Termini	Reported in (Burge *et al.*, 1998)	Total Found	Putative Splicing Errors	Total Confirmed
GT-AG	**34**	279	**4**	275
AT-AC	12	109	1	108
AT-AG	1	8	1	7
GT-AT	0	5	1	**4**
AT-AT	0	**4**	0	**4**
GT-GG	0	7	**4**	**3**
AT-AA	1	5	**3**	2
GT-AA	0	1	0	1
GT-CA	0	1	1	0
GC-AG	1	0	0	0
Totals	**49**	**41**9	15	**404**

scale U12-dependent intron search, these scans analysed unannotated genome sequence data and were neither biased nor aided by previous annotation (Burge *et al.*, 1998).

The search for AT-AC and GT-AG introns examined approximately 20 million candidate introns found by pairing high-scoring U12-dependent donor and branch site signals. **388** of these candidates were confirmed by expressed sequence data using the stringent criteria described above. Five out of these **388** were classified as likely splicing errors and removed from further analysis. The **383** AT-AC and GT-AG human U12-dependent introns reported here represent an increase of **337** (more than 8-fold) over the introns reported in the only similar study (Burge *et al.*, 1998).

In total, scans for U12-dependent introns with 16 different combinations of terminal dinucleotides were performed (see Table **4.1).** 419 introns, including the **388** AT-AC and GT-AG introns discussed above, met the confirmation criteria. Of the additional **31** introns, 10 were classified as likely splicing errors, leaving a total of 21 non AT-AC or GT-AG human U12-dependent introns, distributed among 6 classes, including the previously documented AT-AG and AT-AA (Burge *et al.*, 1998) as well as several previously undocumented classes. Examination of the donor and acceptor signals of the atypical U12-dependent introns reveals almost perfect conservation of both the donor and branch sites with the U12-dependent intron consensus sequences. Detailed

information for all 404 confirmed U12-dependent introns is available as supplementary information in Appendix A or from **http://www.sanger.ac.uk/Users/rd/U12/.**

Despite searches for introns starting with GC or GG, all confirmed introns showed standard AT or GT dinucleotides at the donor position, suggesting that these bases may be almost universally required for successful splicing. One GC-AG U12-dependent intron, which was missed during my analysis due to its atypical and low-scoring donor site, has been reported previously indicating that an AT or GT &nucleotide is not an absolute requirement (Burge *et al.,* 1998). In contrast, a variety of terminal dinucleotides (including AG, AC, AT, AA, and GG) were observed at the acceptor position. The diversity of terminal dinucleotides observed at the acceptor site of human U12-dependent introns confirmed recent experimental work, which indicated that a variety of dinucleotides can serve as functional U12-dependent acceptor sites *in vitro* (Dietrich *et al.,* 2001). This flexibility fits well with the idea that the branch site serves as the primary recognition point for the 3' end of U12-dependent introns and suggests that the mechanism of 3' site identification may be only loosely constrained.

282 confirmed GT donor sites were also scored as U2-dependent donor sites, using the stratified splice predictor described in section 3.2. The vast majority of these sites scored poorly as U2 sites. Only 7 out of 282 (2.5%) received a log-odds score greater than 5 bits and even these scores were generally well below the mean score (mean: 8.66, SD: 2.31) for a set of 3,620 true sites scored with the U2 model.

Estimating the frequency of U12-dependent introns within the genome is a difficult problem and, due to the lack of comparable data for U2-dependent introns, my results do not lead to an easy solution. However, comparing the small sample of 11 U12-dependent introns I identified on chromosome 22 with the 3,199 U2-dependent introns identified in a similar search for U2-dependent introns on chromosome 22 (see Chapter 5) suggests that as many as **0.34** percent of human introns are spliced by the U12 spliceosome. This number is larger than earlier estimates that suggested roughly 0.15 percent of human introns were likely to be U12-dependent (Burge *et al.,* 1998), but, due to the small sample size, must be taken as only a rough estimate.

Access to this large set of confirmed U12-dependent introns allowed me to analyse several characteristics of this rare class of introns. Figure 4.1 compares the length distribution of 168 confirmed AT-AC and GT-AG U12-dependent introns with 11,402 RefSeq-confirmed U2-dependent introns (length < 1 kb) from version 1.0 of Ensembl

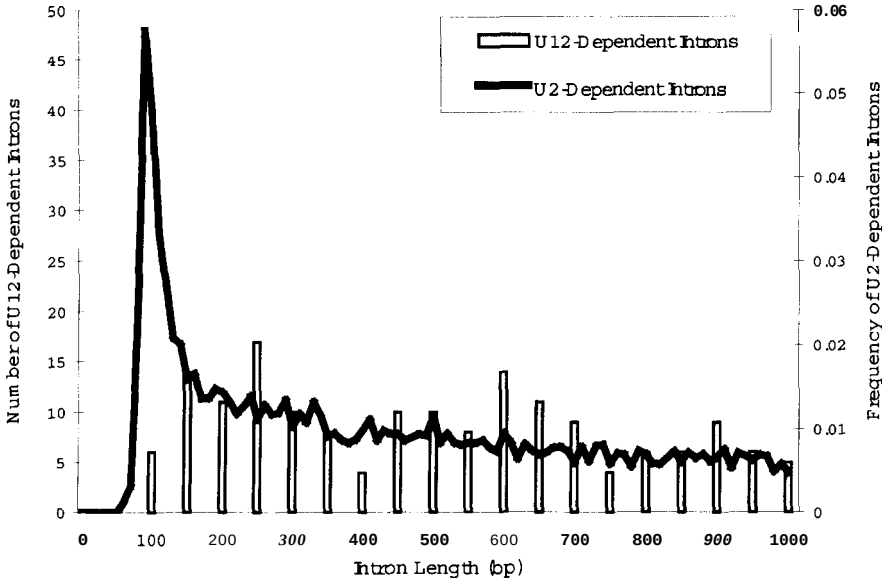


Figure 4.1 – The length of 168 U12-dependent introns and 11,402 RefSeq-confirmed U2-dependent introns less than 1kb in length are plotted. Grey bars represent the counts of U12-dependent introns grouped into 50-bp wide bins and the black line represents the frequency of U2-dependent introns grouped in 10-bp wide bins.

(International Human Genome Sequencing Consortium, 2001). U2-dependent introns have a two-component distribution, with a peak at approximately 90 bp and an exponential-like component for longer lengths. U12-dependent introns seem to be lacking the short component of the U2-dependent intron length distribution. In contrast, U12-dependent introns show a gradual peak between 200 and 250 bp, then a slow decay. The distributions are s d a r for larger introns between 1 and 20 kb (U2: mean: 4,130 bp, SD: 3,720 bp. U12: mean: 3,600 bp, SD: 3,300 bp, and data not shown), showing that the exponential components are similar.

The distribution of the distance between the branch site and the acceptor site for both AT-AC and GT-AG U12-dependent introns is illustrated in Figure 4.2. These results confirm earlier findings that this distance is much more sharply restricted for U12-dependent introns than it is for U2-dependent introns and verify suggestions (Dietrich *et al.*, 2001) that AT-RC and GT-AG U12-dependent introns show different distributions for this distance (Chi-square test: $P < 0.001$). No functional relevance for this difference has been identified.

Table 4.2 compares the phase of 284 of the U12-dependent introns found in this study with 11,117 predominately U2-dependent introns previously analysed (Long *et al.*, 1995). The two distributions differ significantly (Chi-square test: $p < 0.001$) with the

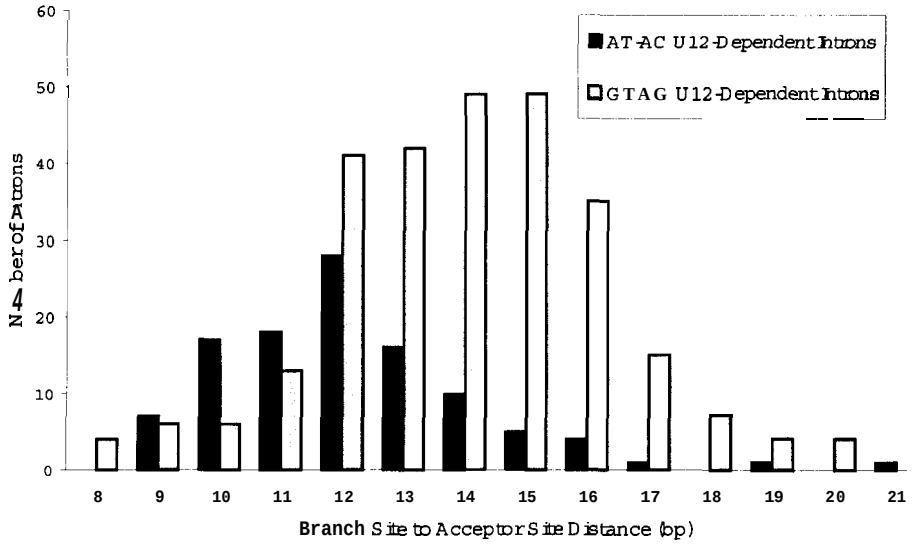


Figure 4.2 – The distance between the branch site and the acceptor site is plotted for 108 AT-AC U12-dependent introns (black bars) and 275 GT-AG U12-dependent introns (grey bars).

most striking difference being the bias against phase 0 introns in the U12-dependent intron data, compared to the bias toward phase 0 introns in the U2-dependent intron data. These results generally agree with previously analysed intron phase data from a smaller dataset (Burge *et al.*, 1998).

	U2-dependent Introns		U12-dependent Introns	
	Number	Percentage	Number	Percentage
Phase 0	5,263	47.4	59	20.8
Phase 1	3,372	30.3	118	41.5
[illegible]	[illegible]	[illegible]	107	[illegible]
Total	11,117	100	284	100

Analysis of the base composition between the branch site and the acceptor site of U12-dependent introns reveals a slight pyrimidine bias in this region. 66 percent of a sample of 2,191 nucleotides from between the branch and acceptor consensus sequences were pyrimidines, while only 54 percent of a control set of 3,060 nucleotides from upstream of the branch site consensus sequence were pyrimidines. Although extracting a comparable set of data for U2-dependent introns is difficult, pyrimidmes make up nearly

Confirmed Intron			Putative Splicing Error		3' Difference
ID	Termini	Evidence	Termini	Evidence	
2	GT-AG	8	GT-GG	1	-4
14	GT-AG	94	GT-CA	1	-1
45	GT-AG	7	GT-AG	1	-3
92	GT-AT	2	GT-GG	1	+3
97	AT-AC	7	AT-AC	1	+4
122	GT-AG	60	GT-AG	1	+2
124	GT-AG	266	GT-GG	7	+1
127	GT-AG	12	GT-AT	1	-2
145	AT-AC	12	AT-AA	1	+5
216	AT-AC	16	AT-AA	1	-3
226	AT-AC	3	AT-AA	1	+6
236	GT-AG	15	GT-GG	1	-3
251	GT-AG	7	GT-AG	1	-4
290	AT-AC	13	AT-AG	1	+2
393	GT-AG	24	GT-AG	1	+2

80 percent of the nucleotides in the 9 bp upstream of the acceptor site consensus (CAG), suggesting that the pyrimidine bias at U12-dependent introns is not as strong as it is at U2-dependent introns.

High error rates at the acceptor site in U12-dependent splicing

A surprisingly high number of introns were identified which met all confirmation criteria, yet seemed unlikely to represent real introns. In general, these introns shared donor sites with other confirmed introns yet differed slightly (1-6 bp) in acceptor site positions. In most cases one member of these pairs of introns had typical terminal &nucleotides and was strongly supported by a large number of expressed sequences while the second exhibited atypical dmucleotidcs and was weakly supported. In many cases the second intron led to the subsequent exon being out of frame and thus is unlikely to represent a true alternatively spliced variant of the gene. 15 introns exhibited these criteria and were classified as likely splicing errors (see Table **4.3).** Although a few of these so-called splicing errors may represent errors in EST sequencing, most seem likely to represent mistakes made by the U12 spliceosome.

In total 21 ESTs were observed confirming likely splicing errors and 5,864 ESTs were observed confirming accepted introns. These numbers suggest that splicing mistakes at the *3'* end of U12-dependent introns occur at a rate of approximately 1 error

in every 280 splices. This value likely underestimates the true error rate in U12 acceptor site selection as only a small subset of terminal dinucleotides was considered in this study. Similar genomic scans with other pairs of bases at the acceptor position could potentially uncover even more evidence of errors during U12-dependent splicing.

Alternative splicing of U12 introns

The approach to intron identification used for these analyses allowed me to identify alternative splicing situations in whch one splice site was used in two or more confirmed introns. Among the 404 U12-dependent introns, 13 such pairs of alternatively spliced introns were observed (see Table 4.4). Eleven cases were identified where the same donor site was used with a different acceptor site and two cases were found in whch different donor sites were paired with the same acceptor site. Interestingly, three of these alternative splicing events involved introns with different pairs of terminal &nucleotides, the first time, to the best of my knowledge, this has been observed. For instance 14 expressed sequences supported an AT-AT intron of length 620 bp in a hypothetical human protein (genbank accession NM_024549) while two expressed sequences supported an AT-AC intron with the same donor site but a different acceptor site 3,344 bp downstream of the donor site.

These results suggest that, at a minimum, 13 out of 391, or roughly 3.3 percent, of human U12-dependent introns have an associated intron truncation/extension type alternatively spliced form. A bias (11 out of 13) towards alterations at the acceptor site was also observed, although the numbers are too small to draw any strong conclusions in this regard. A s d a r analysis of approximately 3,200 expressed sequence confirmed U2-dependent introns (of length < 20 kb) on human chromosome 22 found truncation/extension alternative splicing events to occur at roughly 12 percent of introns and only negligible differences between the frequency of events involving donor and acceptor sites (see Chapter 5).

Non-random distribution of U12-dependent introns in the genome

The distribution of U12-dependent introns w i t h the human genome has important implications for understanding the evolutionary history of the major and

Gene	ID	Termini	Length	Evidence	Accession
Porphobilinogen deaminase (PBG-D) mRNA	3	GT-AG	1145	26	X04217
	4	GT-AG	1593	2	R06263
Quinone oxidoreductase homolog-1 mRNA	343	AT-AC	4501	3	AA370151
	342	AT-AC	4522	11	AF029689
Von Hippel-Lindau binding protein (VBP-1) mRNA	132	GT-AG	2403	35	U96759
	133	GT-AG	3187	1	BF667071
Calcium channel, alpha 2/delta subunit 2 CACNA2D2 gene mRNA	247	GT-AG	103	2	A1251367
	248	GT-AT	97	4	AF042972
Unknown	105	GT-AG	2951	2	AV725561
	106	GT-AG	5038	1	A1917412
Unknown	304	GT-AG	13385	1	BF373273
	303	GT-AG	13423	2	BE887649
Unknown	158	AT-AC	3344	2	AK024780
	157	AT-AT	620	14	BE275895
Unknown	287	GT-AG	1471	39	AK001916
	288	GT-AG	2747	1	BE263460
Unknown	67	GT-AG	605	17	T50022
	68	GT-AG	2677	1	AL523899
Cullin 4a (CUL4A)	257	AT-AC	8926	21	AF077188
	258	AT-AC	277	1	AL560997
Unknown	386	GT-AG	12503	2	AK000443
	387	GT-AG	14540	4	AK022732
JNK1 protein kinase	367	GT-AG	1727	2	L26318
	368	GT-AG	1301	3	L35004
Unknown	106	GT-AG	5038	1	A1917412
	107	AT-AG	1984	5	A1023856

Table 4.4 – Alternatively spliced U12-dependent introns. 13 examples of alternatively spliced U12-dependent introns are shown. For each splicing variant, the ID matching the supplementary intron table, the intron terminal dinucleotides, the intron length, the total evidence supporting the intron and an accession number of a confirming expressed sequence are presented.

minor spliceosomes. Among the 404 U12-dependent introns identified in this analysis, 16 cases were identified where the same expressed sequence confirmed two or more U12-dependent introns, indicating that the two introns occurred within a single gene (see Table **4.5).** One of these cases *(Homo sapiens* NHE-6, genbank accession AF030409) had 3 U12-dependent introns (1 AT-AC, 2 GT-AG) supported by a single expressed sequence.

Assuming that U12-dependent introns are randomly distributed throughout the genome, the probability of identifying 16 or more genes with multiple U12-dependent introns among 388 genes with at least one Ul2-dependent intron is $P < 0.009$. This strongly confirms earlier reports that suggested U12-dependent introns were distributed non-randomly within genomes (Burge *et al.,* 1998). It is worth noting that the strict requirement for multiple introns to be supported by a single expressed sequence almost

Gene	U12-dependent Introns	Accession
Smg GDS-associated protein (SMAP) mRNA	GT-AG (84) AT-AC (85)	U59919
Transcription elongation factor TFIIS.h	AT-AC (239) GT-AG (240)	AJ223473
Inositol polyphosphate 5-phosphatase (5ptase) mRNA	GT-AG (321) AT-AG (322)	M74161
WDR10p-L (WDR10) mRNA	GT-AG (235) GT-AG (236)	AF244931
Diaphanous 1 (HDIA1) mRNA	AT-AC (81) AT-AC (82)	AF051782
Erythroid K:Cl cotransporter (KCC1) mRNA	GT-AG (243) GT-AG (244)	AF047338
Hypothetical transmembrane protein SBBI53 mRNA	GT-AG (381) GT-AG (382)	AF242523
Spermidine aminopropyltransferase mRNA	AT-AC (312) GT-AG (313)	AD001528
Dihydropyridine-sensitiveL-type calcium channel alpha-1 subunit CACNL1A3 (CACNA1S) mRNA	GT-AG (9) GT-AG (10)	L33798
Hypothetical protein FLJ22028	GT-AG (105) AT-AG (107)	AV725561
Autoantigen mRNA	GT-AG (245) GT-AG (246)	L26339
KIAA0136 gene mRNA	AT-AC (344) GT-AG (345)	D50926
Histidase mRNA	GT-AG (98) GT-AG (99)	D16626
ERCC5 excision repair protein (XPG) mRNA	GT-AG (212) AT-AT (213)	L20046
KIAA1176 gene mRNA	GT-AG (188) GT-AG (189)	AB033002
Sodium-hydrogen exchanger 6 (NHE-6) mRNA	AT-AC (401) GT-AG (302) GT-AG (303)	AF030409

Table 4.5 – Genes with multiple U12-dependent introns. 16 Genes with at least two U12-dependent introns are shown. For each U12-dependent intron in the specified gene, the terminal dinucleotides and ID (matching the complete intron list provided as supplementary information) are provided. The accession number of a confirming expressed sequence is provided for each gene.

certainly leads to an underestimate of the true number of genes with multiple U12-dependent introns and, thus, an overestimate of the hkelihood of thls distribution occurring by chance. This underestimation occurs due to the short length of most ESTs and the correspondingly small chance that a single EST would support multiple introns. Furthermore, in this analysis duplicate U12-dependent introns, whch arose from gene duplications during evolution, are counted as distinct introns. If each group of duplicate introns was counted as a single intron, the hkelihood of seeing this distribution arising randomly would be reduced.

Relative utility of the mRNA and EST datasets

'The use of expressed sequences to confirm introns in this analysis provides an opportunity to compare the coverage and utility of the two major expressed sequence datasets: the set of mRNA and the set of ESTs (see 'Table 4.6). Of the 404 introns identified in this analysis, 267 (66 percent) were supported by both mRNA and EST sequences, 101 (25 percent) were supported only by EST sequences and 36 (9 percent) were supported only by mRNA sequences. As expected, there was much higher redundancy in the EST set, as the median number of EST sequences supporting an intron was four and the median number of mRNA sequences supporting an intron was one.

	mRNA	EST
Total Introns	303 (75%)	368 (91%)
Exclusive Introns	36 (9%)	101 (25%)
Mean	1.5	14.5
Median	1	4
Maximum	14	284

Table 4.6 – Summary of expressed sequences supporting U12-dependent introns. The total number of introns and the number of introns exclusively supported by each type of expressed sequence are shown in the top two rows. The average and median numbers of expressed sequences supporting each U12-dependent intron are shown as well as the maximum number of expressed sequences supporting a single intron.

These results suggest that the coverage of both datasets is good but far from complete. Furthermore the differing characteristics of the two sets render them useful for different types of analyses. In some cases the higher coverage of the EST set may be required, while in others (such as the analysis of the distribution of U12-dependent introns above) the longer length of sequences in the mRNA set may make this collection of sequences more valuable.

Discussion

The analysis presented here greatly increases both the number of U12-dependent introns identified and the diversity of these introns. The observation that a significant number of U12-dependent introns exhibit atypical terminal &nucleotides suggests that a good number of the so-called non-canonical introns identified in a variety of genomes (Burset *et al.*, 2001), may represent variants of U12-dependent introns. Furthermore, due to the different parameters used in the searches for typical and atypical U12-dependent introns, the results presented here most likely reflect an under-representation of atypical U12-dependent introns. For instance, only 76 percent of AT-AC and GT-AG U12-

dependent introns have lengths under 2 kb. If this ratio holds for atypical U12-dependent introns as well, the 21 examples reported here should increase to 27 or 28. Furthermore, scans for introns with pairs of terminal dmucleotides not considered in this study may identify additional atypical U12-dependent introns.

The 404 U12-dependent introns identified here represent a lower bound on the genome's full complement of these introns for a variety of reasons. Firstly, as noted previously, the arbitrary limit of 20 kb as the maximum intron length for AT-AC and GT-AG U12-dependent introns almost certainly excluded a significant number of true introns from my analysis. For comparison roughly five percent of Ensembl U2-dependent introns confirmed by RefSeq entries are greater than 20 kb in length (International Human Genome Sequencing Consortium, 2001). In addition the threshold values used for donor and branch site scores, whde chosen to be inclusive, likely excluded a small number of valid introns from the analysis.

Furthermore, the incomplete nature of the EST and mRNA sets used to confirm introns means that some number of true introns, whch were identified as potential introns in the first stage of this analysis, failed to meet the confirmation criteria and were not included in the final counts. EST datasets in particular are biased towards the 5' and 3' ends of genes and are less likely to provide evidence for introns near the middle of larger genes.

A large majority of the human U12-dependent introns reported previously were identified in this large-scale genomic analysis. However a few were missed. For instance, intron 5 of FHIT (human fragile histidine triad gene), and intron 16 of HPS (human Hermansky-Pudlak syndrome gene) previously noted to be U12-dependent introns (Burge *et al.*, 1998), were both missed by my analysis. Careful examination of these particular introns reveals that the FHIT intron was missed due to its exceptionally long length whde the HPS intron was missed to due to its atypical and low scoring donor and branch sites.

The large set of U12-dependent introns presented here should prove helpful for future studies regarding the evolution of the two-spliceosome system. Comparisons with the nearly complete mouse genome should prove useful in analysing the frequency of subtype switching, as well as intron conversion and loss.

The differences observed between the length distribution of U12- and U2-dependent introns raise interesting questions about the two splicing mechanisms. In particular the accurate pairing of donor and acceptor sites is thought to occur by two

different models in hgher eukaryotes, an intron definition model, whch functions in the excision of small introns (Talerico and Berget, 1994), and an exon definition model, whch functions in the excision of larger introns (Berget, 1995). U12-dependent introns have been shown to participate to some degree in exon definition interactions (Wu and Krainer, 1996) and one possible explanation for the relative dearth of short U12-dependent introns may be that they are recognised exclusively in an exon-dependent fashion, eliminating any selective benefit potentially associated with the short length of many U2-dependent introns.

A number of the U12-dependent introns found in the human genome occur within larger gene families, suggesting that the intron arose originally in a single ancestral gene and was duplicated along with the rest of the gene as the families grew. The presence of U12-dependent introns in some gene families, including the calcium and sodium voltage-gated cation channels (Wu and Krainer, 1999), the matrilin family (Muratoglu *et al.*, 2000), the protein kinase superfamily (Burge *et al.*, 1998) and the E2F transcription factory family, has been well studied. This analysis found conservation of U12-dependent introns in the phospholipase C famdy, the transportin family, the diaphanous family and the CAMP-binding guanine nucleotide exchange factor family (see supplementary information) in addition to these previously identified gene families. Additionally, U12-dependent intron containing genes seem to be over-represented in the ras-raf signal transduction pathway, although further work is required to determine the significance of this observation.

The observation of alternative splicing of U12-dependent introns poses interesting evolutionary questions as well. If U12-dependent introns convert to U2-dependent over evolutionary time by accumulation of mutations at the splicing junctions as previously hypothesised (Burge *et al.*, 1998; Dietrich *et al.*, 1997), how would this work for alternatively spliced introns. In the case of an intron truncation event where two different acceptors could pair with a single donor, the intron conversion process might necessitate either the seemingly unlikely simultaneous conversion of multiple intron junctions or the loss of one of the splicing alternatives. This scenario suggests that alternatively spliced U12-dependent introns would be preferentially preserved, but is in conflict with the observation that alternative splicing is rarer at U12-dependent introns. A possible explanation may be that the U12 spliceosome is less amenable to the complex regulation patterns that alternative splicing requires and that alternative splicing, therefore, arises less frequently at U12-dependent introns.

Although little is known about error rates of U2-dependent splicing, the calculation of a preliminary error rate for U12-dependent splicing presents some interesting possibilities. In particular, if errors occur with a significantly higher frequency at U12-dependent introns than at U2-dependent introns, this may point to a reason that U12-dependent introns seem to be selected against during evolution and even are found to be lacking entirely from some eukaryotes, such as *C. elegans*.

In addition to the observations made here, I hope the set of U12-dependent introns generated by this analysis will provide a useful resource for future examinations of the minor spliceosome and its evolution.

Chapter 5

A Computational Scan for U2-Dependent Introns on Human Chromosome 22

Summary

A computational scan for U2-dependent introns on human chromosome 22 identified 3,199 introns strongly supported by expressed sequences. Of these, 0.7 percent were non-canonical GC-AG introns and the remaining 99.3 percent were canonical GT-AG introns. Approximately 12 percent of introns were involved in an intron truncation/extension alternative splicing event, with roughly equal numbers of these occurring at the donor and acceptor ends of the intron. This large set of confirmed introns should prove to be a useful resource for continued detailed gene annotation.

Introduction

Chromosome 22 was the first human chromosome to be essentially completely sequenced (Dunham *et al.,* 1999) and has been extensively annotated using both computational and experimental approaches. Because of this annotation, chromosome 22 has become the de facto test sequence for a large variety of novel informatics analyses. However, annotation is still not complete. In order to assist in the ongoing annotation of chromosome 22, I have developed an expressed sequence based intron identifier, which pairs strong donor and acceptor signals in an attempt to identify as many U2-dependent introns as possible. Results of a preliminary analysis on chromosome 22 and prospects for performing such an analysis on a genomic scale are discussed.

Materials and Methods

Human U2-dependent introns were identified using a two-step procedure, similar to that used in Chapter 4 to identify U12-dependent introns. Potential donor and branch site signals were identified based on statistical pattern recognition techniques, as implemented in StrataSplice (see section 3.4). A posterior probability threshold value of 1e-3 was selected to balance sensitivity with computational demands. Earlier analysis (see sections 3.2, 3.3) suggests that this included roughly 96 percent of GT donor sites, 76 percent of GC donor sites and 96 percent of AG acceptor sites. From these signals, potential introns (donor/acceptor pairs) of less than 20 kb were generated and expressed sequence evidence was used, as described in detad in Chapter 4, to identify a subset of these potential introns as valid. All scans used the 19 May 2000 'Release 2' build of the chromosome 22 sequence (Dunham *et al.,* unpublished, avadable from **http://www.sanger.ac.uk/HGP/Chr22/).**

Confirmed introns were compared to a collection of known repeats on chromosome 22 generated using RepeatMasker (Smit, A.P.R. & Green, P., unpublished) and introns with at least one splice site within a known repeat were discarded and ignored in later analyses.

Results

The computational scan for U2-dependent introns on chromosome 22 generated roughly 72 million potential introns less than 20 kb in length. 4,719 of these potential introns were strongly confirmed by either mRNA or EST sequences. 1,520 of these were found to have at least one splice site (and usually both) within an annotated repeat element. Removing these 1,520 introns left a total of 3,199 expressed sequence confirmed introns on chromosome 22. Nearly 80 percent of these introns agree precisely with chromosome 22 annotations, and this accounts for approximately 70 percent of previously annotated introns (see Table 5.1). 671 of the identified introns were missing from the chromosome 22 annotations and some of these may represent either alternatively spliced forms of known genes or previously unidentified genes.

	Number of Introns
Chromosome 22 Annotation	3,584
U2-Dependent Intron Finder	3,199
Both Sets	2,528
Annotation only	1,056
Intron Finder only	671

Table 5.1 – Comparison between introns identified in this study and annotation of chromosome 22. The total number of introns identified by the chromosome 22 annotation team and the intron finder described in this chapter are provided. The number of introns found in both sets or exclusively in one set is shown as well. Annotation data (Release 2.3, 6 March 2001) were produced by the Chromosome 22 Gene Annotation Group at the Sanger Centre and were obtained from the World Wide Web at **http://www.sanger.ac.uk/HGP/Chr22** (Dunham *et al.*, unpublished).

3,178, or 99.3 percent, of the introns identified in this study were canonical GT-AG introns and the remaining 0.7 percent were non-canonical GC-AG introns. Although the thresholds used in this analysis were biased slightly toward the inclusion of GT-AG introns, these percentages compare well with previously determined estimates of GC-AG intron frequency (International Human Genome Sequencing Consortium, 2001; Burset *et al.*, 2000).

The 3,199 introns identified on chromosome 22 were searched for intron truncation/extension type alternative splicing events in which one splice site remained the same. Of the 3,001 unique donor sites found in the set of 3,199 introns, 175, or 5.8

percent, were associated with two or more acceptor sites. Similarly, 187 or 6.2 percent, of the 2,996 unique acceptor sites were associated with two or more donor sites. Combining these numbers for individual splices sites suggests that roughly 12 percent of U2-dependent introns on chromosome 22 showed intron truncation/extension type alternative splicing. Most alternatively spliced introns only had one alternative form, but a small number of donor and acceptor sites were found that were used in three or four introns (see Table 5.2).

	Donor Sites		Acceptor Sites	
	Number	Percent	Number	Percent
1 Intron	2,826	94.17	2,809	93.76
2 Introns	154	5.13	173	5.77
3 Introns	19	0.63	12	0.40
4 Introns	2	0.07	2	0.07
Total	3,001		2,996	

Discussion

The study described here shows the utility of large-scale intron identification projects for genome annotation and analysis. Preliminary examination of this data by the chromosome 22 annotation team at the Sanger Centre suggests that this approach has identified a variety of potentially novel introns and has provided additional evidence for a large number of previously annotated gene structures.

However the analysis is computationally quite intensive. For instance the analysis of the roughly 35 Mb chromosome 22 sequence involved the generation of approximately 72 million potential introns, which required roughly 9 GB of storage space. Additionally, the search time, even using the high speed SSAHA search algorithm (Ning, Z., Cox, A.J. and Mullikin,J.C., in press) was significant. Searching the 72 million potential introns took 5.5 days of processor time on a 16 GB machme. Furthermore, such an analysis would ideally be performed using more inclusive thresholds, but this would lead to even more significant requirements in terms of disk space and processor time. Because the disk space is needed only transiently, the search time is the key factor determining whether or not such an analysis could be performed on a genomic scale. As analysis of the complete human genome using the same search parameters would take over a year on a high memory machme, the answer at the current time is, realistically, no, at least not as done here.

This approach is much more efficient, however, **if** the expressed sequences are localised within the genome first, allowing the searches to be performed against much smaller databases, and this approach could make such an analysis feasible on a genomic scale. However, it would arguably make more sense to work in the opposite direction and align full ESTs to the genome as done at Ensembl/UCSC and use these sequences to confirm introns. Comparing the results of Ensembl's alignment of ESTs to chromosome 22 with the results reported here would be an interesting project, but time to complete this work was not available.

Chapter 6

Conclusion

The process of annotating the human genome has now begun in earnest with the completion of the draft sequence and will continue for many years as the sequence is finished and as understanding of the genome's complex structure improves. The work described in this thesis aimed to assist this ongoing process by using a variety of computational approaches to help identify introns, and in the process also add to our knowledge about the mechanisms of RNA splicing. Below I outline briefly a few of the many ways these investigations could be taken further.

One step toward this goal was taken with the development of a new model for splice site identification that utilises local GC content to generate improved predictions compared to standard non-stratified models. This model identified differences in intron recognition signals that varied with GC content. These lead to interesting questions regardmg the role of splice site recognition in genome evolution. It is worth exploring whether splice site recognition has played a role in genome evolution, by, for instance, restricting most introns in GC-rich regions to short lengths. Another possibility that merits further examination is that small and large introns may exhibit slightly different splice site signals. If this is the case, my stratified model may capture this signal indirectly as a by-product of most small introns occurring in GC-rich regions of the genome.

Splice site signals are insufficiently informative by themselves and the chance of pinpointing splice sites in genomic sequences based on their signals alone seems slim. However, improvements in splice site identification lead to improvements in gene prediction and are worth exploring. In this regard, it seems that future work will further blur the distinction between splice site prediction and gene prediction with the emphasis falling on programs that take advantage of a variety of signals to accurately identify either introns or exons. An improved understanding of the splicing process *in vivo* will play a key role in this progression as utilising the information provided by intronic and exonic splicing enhancer sites (reviewed in Blencowe, 2000) is likely to be a difficult but important step in this process.

The advent of large expressed sequence libraries is having a dramatic impact on genome annotation. In regard to splicing, alignments of expressed sequences to the genome can identify splice sites with high confidence and are a powerful tool for determining gene structure. Expressed sequence datasets, however, are generally both incomplete and biased toward the start and end of transcripts and, whde a valuable annotation tool, complement rather than replace other annotation efforts.

Due to their scarcity, U12-dependent introns have traditionally been ignored in large-scale annotation efforts, hindering the accurate annotation of several hundred human genes. My analysis of these rare introns will help define these gene structures and work is ongoing to add these introns into Ensembl (International Human Genome Sequencing Consortium, 2001).

I hope that this research will also have implications for understanding the origin of the two spliceosome system found in many eukaryotes. In particular, comparative analysis of the processes of gain and loss of U12-dependent introns and conversion of U12- to U2-dependent introns (and perhaps vice versa) will increase our understanding of both splicing systems, and perhaps offer insights into eukaryotic evolution more generally. Scanning the mouse genome, once its sequencing reaches a suitable stage, would be the logical next step, as many genes and even gene structures should be conserved between mouse and human.

Although my analysis has led to a number of new observations regarding U12-dependent introns, the calculation of a preliminary error rate for U12-dependent splicing is particularly interesting. To the best of my knowledge, no previous estimates for splicing error rates by either spliceosome exist, and whde my estimate is only preliminary, it points the way toward an effective use of expressed sequence data to more accurately estimate this value and provide a comparable estimate for U2 splicing. These estimates could be obtained as by-products of the large-scale EST to genome alignment projects ongoing at the Sanger Centre and elsewhere.

Developing an accurate understanding of the frequency and mechanisms of alternative splicing looms as the next big goal in the ongoing effort to understand eukaryotic gene structure and hopefully a combination of expressed sequence based analysis and *ab initio* predictions can yield significant progress on this front. Although my work has not focused on this problem, I have identified a variety of alternatively spliced U12-dependent introns and hope that my stratified splice predictor will help in the determination of potential alternative U2-dependent splice sites.

Nearly 25 years of research into RNA splicing has yielded enormous progress in terms of understanding, but many key questions remain. Although *in vivo* work is crucial to an eventual understanding of RNA splicing, the wealth of data becoming avadable through genome projects has enabled informatics approaches to also make a significant contribution. As sequence data continues to accumulate, I expect this trend to continue

and imagine that computational approaches will play an important role in answering many of the still unresolved questions regardmg RNA splicing.

Chapter 7

References

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J, Zhang,Z., Miller,W., and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25,** 3389-3402

Baker,W., van den Broek,A., Camon,E., Hingamp,P., Sterk,P., Stoesser,G., and Tuli,M.A. (2000) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.,* **28**, 19-23.

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

Berget,S.M., Moore,C. and Sharp,P.A. (1977) Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. USA*, **74**, 3171-3175.

Blencowe, B.J. (2000) Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci.*, **25**, 106-110.

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

Brunak,S. and Engelbrecht,J. (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.*, **220**, 49-65.

Burge,C. (1997) Identification of complete gene structures in human genomic DNA. PhD thesis. Stanford University, Stanford, CA.

Burge, C. (1998) Modeling dependencies in pre-mRNA splicing signals. In Salzberg,S.L., Searls,D.B., and Kasif,S. (eds.), *Computational Methods in Molecular Biology*. Elsevier Science. Amsterdam, Netherlands.

Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78-94.

Burge,C.B., Padgett,R.A, and Sharp,P.A. (1998) Evolutionary fates and origins of U12-type introns. *Mol. Cell*, **2**, 773-785.

Burge,C.B., Tuschl,T., and Sharp,P.A. (1999) Splicing of precursors to mRNAs by the spliceosomes. In Gesteland,R.F. and Atkins,J.F. (eds.), *The RNA World.* (2nd edition) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

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

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

Chow,L.T., Gelinas,R.E., Broker,T.R., and Roberts,R.J. (1977) An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, **12**, 1-8.

Collins,C.A., and Guthrie,C. (2000) The question remains: is the spliceosome a ribozyme. *Nat. Struct. Biol.*, **7**, 850-854.

Crispino,J., Blencowe,B.J. and Sharp,P.A. (1994) Complementation by SR proteins of pre-mRNA splicing reactions depleted of U1 snRNP. *Science*, **265**, 1860-1869.

Crispino,J.D., Mermoud,J., Lamond,A., and Sharp,P.A. (1996) *Cis*-acting elements from the 5' splice site promote U1-independent pre-mRNA splicing. *RNA*, **2**, 664-673.

Dietrich,R.C., Incorvaia,R. and Padgett,R.A. (1997) Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. *Mol. Cell*, **1**, 151-160.

Dietrich,R.C., Peris,M.J.,Seyboldt,A.S., and Padgett,R.A. (2001) Role of the 3' splice site in U12-dependent intron splicing. *Mol. Cell. Biol.*, **21**, 1942-1952.

Dunham,I. *et al.*, (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, 489-495.

Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press. Cambridge, UK.

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

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

Hall,S.L. and Padgett,R.A. (1994) Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. *J. Mol. Biol.* **239**, 357-365.

Harr,R., Haggstrom,M. and Gustafsson,P. (1983) Search algorithm for pattern match analysis of nucleic acid sequences. *Nucleic Acids Res.*, **11**, 2943-2957.

Hastings,M.L. and Krainer,A.R. (2001) Functions of SR proteins in the U12-dependent AT-AC pre-mRNA splicing pathway. *RNA*, **7**, 471-482.

Hurst,L.D., Brunton,C.F.A, and Smith, N.G.C. (1999) Small introns tend to occur in GC-rich regions in some but not all vertebrates. *Trends Genet.*, **15**, 437-439.

International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.

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

Konarksa,M.M. (1998) Recognition of the 5' splice site by the spliceosome. *Acta Biochim Pol*, **45**, 869-81.

Kramer,A., Gruter,P., Groning,K. and Kastner,B. (1999) Combined biochemical and electron microscopic analyses reveal the architecture of the mammalian U2 snRNP. *J. Cell. Biol.*, **145**, 1355-1368.

Kudo,M., Lida,Y. and Shimbo,M. (1987) Syntactic pattern analysis of 5' splice site sequences of mRNA precursors in higher eucaryote genes. *Comput. Appl. Biosci.*, **3**, 319-324.

Lewin,B. (2000) *Genes VII*. Oxford University Press. Oxford, UK.

Long,M., Rosenberg,C., and Gilbert,W. (1995) Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc. Natl. Acad. Sci. USA*, **92**, 12495-12499.

Luo,H.R., Moreau,G.A, Levin,N. and Moore,M.J. (1999) The human Prp8 protein is a component of both U2- and U12-dependent spliceosomes. *RNA*, **5**, 893-908.

Meister,G., Hannus,S., Plottner,O., Baars,T., Hartmann,E., Fakan,S., Laggerbauer,B. and Fischer,U. (2001) SMNrp is an essential pre-mRNA splicing factor required for the formation of the mature spliceosome. *Embo J.*, **20**, 2304-2314.

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

Moore,M.J. (2000) Intron recognition comes of AGe. *Nat. Struct. Biol.*, **7**, 14-16.

Mount,S.M. (1982) A catalogue of splice junction sequences. *Nucleic Acids Res.*, **10**, 459-472.

Muratoglu,S., Krysan,K., Baláza,M., Sheng,H., Zákány,R., Módis,L., Kiss,I., and Deák,F. (2000) Primary structure of human matrilin-2, chromosome location of the MATN2 gene and conservation of an AT-AC intron in matrilin genes. *Cytogenet. Cell. Genet.*, **90**, 323-327.

Pearson, W.R. (1995) Comparison of methods for searching protein sequence databases. *Protein Sci.*, **4**, 1145-1160.

Rappsilber,J., Ajuh,P.M., Lamond,A.I. and Mann,M. (2001) SPF30, an essential human splice factor required for assembly of U4/U5/U6 tri-snRNP into the spliceosome. *J. Biol. Chem.*, in press.

Reed,R. (2000) Mechanisms of fidelity in pre-mRNA splicing. *Curr. Opin. Cell. Biol.*, **12**, 340-345.

Schwer,B. (2001) A new twist on RNA helicases: DExH/D box proteins as RNPases. *Nat. Struct. Biol.*, **8**, 113-116.

Sharp,P.A. and Burge,C.B. (1997) Classification of introns: U2-type or U12-type. *Cell*, **91**, 875-879.

Shukla,G.C. and Padgett,R.A. (2001) The intramolecular stem-loop structure of U6 snRNA can functionally replace the U6atac snRNA stem-loop. *RNA*, **7**, 94-105.

Staden,R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.*, **12**, 505-519.

Staley,J.P. and Guthrie,C. (1999) An RNA switch at the 5' splice site requires ATP and the DEAD box protein Prp28p. *Mol. Cell.*, **3**, 55-64.

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

Tarn,W.Y. and Steitz,J.A. (1994) SR proteins can compensate for loss of U1 snRNP functions in vitro. *Genes Dev.*, **9**, 2704-2717.

Tarn,W.Y. and Steitz,J.A. (1996a) A novel spliceosome containing the U11, U12 and U5 snRNPs excises a minor class (AT-AC) intron *in vitro*. *Cell*, 804, 801-811.

Tarn,W.Y. and Steitz,J.A. (1996b) Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. *Science*, **27**, 1824-1832.

Venter,J.C. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304-1351.

Will, C.L., Schneider,C., Reed,R. and Lührmann,R. (1999) Identification of both shared and distinct proteins in the major and minor spliceosomes. *Science*, *284*, 2003-2005.

Wu,Q. and Krainer,A.R. (1996) U1-mediated exon definition interactions between AT-AC and GT-AG introns. *Science*, **274**, 1005-1008.

Wu,Q. and Krainer,A.R. (1999) AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes. *Mol. Cell. Biol.*, **19**, 3225-3236.

Wu,S., Romfo,C.M., Nilsen,T.W. and Green,M.R. (1999) Functional recognition of the 3' splice site AG by the splicing factor U2AF^{35}. *Nature*, **402**, 832-835.

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

Zhang,M.Q. and Marr,T.G. (1993) A weight array method for splicing signal analysis. *Comp. Appl. Biol. Sci.*, **9**, 499-509.

Zorio,D.A.R. and Blumenthal,T. (1999) Both subunits of U2AF recognize the 3' splice site in Caenorhabditis elegans. *Nature*, **402**, 835-838.

Zoubak,S., Clay,O., and Bernardi,G. (1996) The gene distribution of the human genome. *Gene*, **174**, 95-102.

Appendix A

Complete List of U12-Dependent Introns Identified in Chapter 4

Appendix A

Complete List of U12-Dependent Introns Identified in Chapter 4

Description of the table of U12-dependent introns

The table below contains one line for each of the 404 U12-dependent introns identified in the computational scan for U12-dependent introns described in Chapter 4. Each intron has a unique ID, which matches the ID field on Tables 4.3, 4.4, 4.5 and 4.6. Additionally the golden path contig on which each intron was found is indicated, as are the donor and acceptor positions of the intron in the specified contig. The terminal dinucleotides (termini), the length, the distance from the branch site to the acceptor site (b_to_a), the number of mRNA and EST sequences supporting the intron and the name of the gene in which the intron was found (if known) are also provided.

ID	contig	donor	acceptor	termini	length	b_to_a	rna	est	gene (if known)
1	ctg_na	8406675	8405853	atac	822	9	3	8	serine threonine kinase (STK11/LKB1)
2	ctg12269	12008161	12008008	gtag	153	13	3	5	unassigned RNA
3	ctg12269	14606903	14608048	gtag	1145	14	3	23	porphobilinogen deaminase (PBG-D, EC 4.3.1.8)
4	ctg12269	14606903	14608496	gtag	1593	16	0	2	no RNA
5	ctg12269	15088391	15088226	gtag	165	17	1	3	ubiquitin-specific protease UBP41
6	ctg12316	1504800	1504510	gtag	290	13	0	1	no RNA
7	ctg12323	2773877	2775625	gtag	1748	14	4	33	Homo sapiens BAF53a (BAF53a) mRNA, complete cds.
8	ctg12364	630349	638172	gtag	7823	14	1	2	unassigned RNA
9	ctg12369	5058034	5057161	gtag	873	15	1	0	Human dihydropyridine-sensitive L-type calcium channel alpha-1 subunit CACNL1A3
10	ctg12369	5106119	5104281	gtag	1838	13	1	0	Human dihydropyridine-sensitive L-type calcium channel alpha-1 subunit CACNL1A3
11	ctg12387	11807771	11808673	gtag	902	10	2	115	Ku autoimmune antigen gene
12	ctg12393	4271894	4270764	gtag	1130	17	1	3	unassigned RNA
13	ctg12393	9130298	9147203	atac	16905	14	1	0	unassigned RNA
14	ctg12438	1007607	1008983	gtag	1376	16	1	93	RNA polymerase II subunit
15	ctg12475	783919	784561	gtag	642	12	2	12	unassigned RNA
16	ctg12475	2803637	2805326	gtag	1689	13	5	15	GTP-binding protein RAB6
17	ctg12475	2924257	2919307	gtag	4950	13	1	2	unassigned RNA
18	ctg12475	3207262	3207792	atac	530	11	3	22	phosphatase methylesterase-1 (PME-1)
19	ctg12475	4000723	4004365	atac	3642	13	1	61	unassigned RNA
20	ctg12475	7854989	7853175	gtag	1814	12	1	15	no RNA
21	ctg12475	16986837	16988015	gtag	1178	17	5	12	WAIT-1
22	ctg12482	13320567	13319384	gtag	1183	14	5	17	origin recognition complex subunit 3 (ORC3)
23	ctg12483	796536	797214	atac	678	12	0	2	no RNA
24	ctg12483	1359141	1356780	atac	2361	11	2	75	Arp2/3 complex 16kDa subunit (ARC16)
25	ctg12483	2769484	2767362	gtag	2122	15	1	9	unassigned RNA
26	ctg12559	9058733	9059659	atac	926	11	3	5	mediator (Sur2)
27	ctg12559	16869554	16873925	gtag	4371	14	0	10	no RNA
28	ctg12559	21752756	21751538	gtat	1218	12	0	5	no RNA
29	ctg12616	437115	438183	gtag	1068	15	1	11	tRNA-guanine transglycosylase
30	ctg12691	2366591	2366385	gtag	206	12	0	3	no RNA
31	ctg12719	2624577	2627566	gtag	2989	14	1	2	unassigned RNA
32	ctg12723	1314122	1325072	atac	10950	11	0	57	no RNA
33	ctg12729	2263390	2260721	atac	2669	13	1	4	unassigned RNA
34	ctg12743	1633901	1637859	gtag	3958	14	2	3	unassigned RNA

35	ctg12749	1119576	1120193	gtag	617	13	0	6	no RNA
36	ctg12749	7826157	7821817	gtag	4340	15	1	2	unassigned RNA
37	ctg12766	4129283	4128597	gtag	686	12	3	3	TATA binding protein associated factor (TAFII150)
38	ctg12770	448596	448803	gtag	207	16	1	0	neuronal apoptosis inhibitory protein
39	ctg12772	2626497	2623661	atac	2836	10	0	1	no RNA
40	ctg12824	4885849	4879798	atac	6051	12	0	2	no RNA
41	ctg12824	5036385	5035593	atac	792	10	1	1	unassigned RNA
42	ctg12824	24133076	24133856	gtag	780	11	0	1	no RNA
43	ctg12850	3847805	3846674	atac	1131	12	3	1	matrilin-3
44	ctg12913	3545077	3547144	gtag	2067	16	0	2	no RNA - AOX1
45	ctg13001	11511077	11514110	gtag	3033	15	1	6	unassigned RNA
46	ctg13010	10127356	10125651	gtag	1705	11	0	1	no RNA
47	ctg13023	6097907	6101881	gtag	3974	12	1	23	unassigned RNA
48	ctg13038	666224	667824	gtag	1600	9	2	23	cap-binding protein (CBP20)
49	ctg13067	2493412	2494448	gtag	1036	18	1	63	translocon-associated protein gamma subunit
50	ctg13079	905421	905917	gtag	496	15	2	2	digestive tract-specific calpain (nCL-4)
51	ctg13079	3127622	3122059	gtag	5563	15	0	2	no RNA
52	ctg13103	593749	592779	gtag	970	14	2	4	cap-binding protein (CBP80)
53	ctg13103	28344485	28348736	atac	4251	16	0	20	no RNA
54	ctg13116	5048234	5044745	gtag	3489	11	3	2	Mcd4p homolog (ER membrane protein)
55	ctg13155	344147	342208	atac	1939	12	0	2	no RNA
56	ctg13155	789323	789693	gtag	370	9	1	2	unassigned RNA
57	ctg13206	5570875	5571996	gtag	1121	16	1	1	unassigned RNA
58	ctg13206	7466044	7470648	gtag	4604	14	1	3	unassigned RNA
59	ctg13255	3171119	3189402	gtag	18283	16	1	9	unassigned RNA
60	ctg13284	4336888	4336361	gtag	527	14	0	1	no RNA
61	ctg13284	4338164	4337189	atac	975	14	0	5	no RNA
62	ctg13284	5500810	5501841	gtag	1031	13	2	2	TBP-associated factor 172 (TAF-172)
63	ctg13284	7392474	7395671	gtag	3197	16	0	7	no RNA
64	ctg13284	9369109	9368784	gtgg	325	21	1	6	unassigned RNA
65	ctg13284	11466790	11466375	gtat	415	11	0	1	no RNA
66	ctg13284	11480067	11466367	gtag	13700	19	1	46	CGI-108 protein
67	ctg13284	11495291	11495896	gtag	605	16	0	17	no RNA
68	ctg13284	11495291	11497968	gtag	2677	12	0	1	no RNA
69	ctg13284	11901924	11903243	gtag	1319	14	0	1	no RNA
70	ctg13284	36675973	36671828	atac	4145	14	2	9	unassigned RNA
71	ctg13284	38171842	38173541	atac	1699	12	5	10	unassigned RNA
72	ctg13286	5835001	5839465	gtag	4464	15	1	2	unassigned RNA
73	ctg13286	6858793	6865696	atac	6903	16	1	0	homeodomain protein (Prox 1)
74	ctg13286	7208221	7207276	atac	945	12	1	16	HSKM-B
75	ctg13290	406895	409087	gtag	2192	14	1	33	WS-3
76	ctg13321	2691917	2696822	gtag	4905	16	1	0	unassigned RNA
77	ctg13361	5083938	5076925	gtag	7013	14	4	16	RNA 3'-terminal phosphate cyclase-like protein (rcl1) - RNA processing
78	ctg13361	29016571	29018536	gtag	1965	14	2	22	unassigned RNA
79	ctg13420	1202058	1202336	atat	278	11	4	45	Homo sapiens mRNA for Prer protein
80	ctg13420	1229175	1230401	gtag	1226	15	1	14	histidyl-tRNA synthetase homolog (HO3)
81	ctg13420	2217557	2216441	atac	1116	15	1	8	diaphanous 1 (HDIA1) -cell motility
82	ctg13420	2268292	2267660	atac	632	19	2	2	diaphanous 1 (HDIA1)
83	ctg13446	3682407	3686111	atac	3704	10	0	1	no RNA

84	ctg13446	3911600	3906115	gtag	5485	15	2	17	Smg GDS-associated protein SMAP
85	ctg13446	3963637	3961251	atac	2386	13	2	3	Smg GDS-associated protein SMAP
86	ctg13458	146979	147679	gtag	700	15	5	4	exonuclease 1 (EXO1/HEX1)
87	ctg13464	1066836	1066653	gtag	183	11	1	0	unassigned RNA
88	ctg13513	10699073	10697038	gtag	2035	17	0	2	no RNA
89	ctg13516	940311	939774	gtag	537	13	0	25	no RNA
90	ctg13531	420977	420720	gtag	257	13	0	1	no RNA
91	ctg13617	3757596	3754663	gtag	2933	12	0	3	no RNA
92	ctg13617	8399283	8400103	gtat	820	14	0	2	no RNA
93	ctg13617	9191549	9192965	gtag	1416	8	0	1	no RNA
94	ctg13685	819711	818448	atac	1263	12	4	29	WDR1
95	ctg13698	1854642	1854357	atac	285	16	1	3	unassigned RNA
96	ctg13715	10611722	10615286	gtag	3564	13	0	1	no RNA
97	ctg13915	3025345	3020796	atac	4549	12	3	4	unassigned RNA
98	ctg13915	3748815	3745770	gtag	3045	14	1	2	histidase
99	ctg13915	3759837	3759731	gtag	106	12	1	0	histidase
100	ctg13934	572024	570979	gtag	1045	17	2	35	unassigned RNA
101	ctg13980	739832	742971	gtag	3139	15	5	0	chloride channel ClC-6
102	ctg14007	5421950	5419413	gtag	2537	13	3	13	phosphatidylcholine transfer protein (PC-TP)
103	ctg14015	1249901	1252645	gtag	2744	17	3	3	transportin (TRN)
104	ctg14055	3072440	3070650	gtag	1790	12	2	9	c-raf
105	ctg14071	1105167	1108118	gtag	2951	14	0	2	no RNA
106	ctg14071	1105167	1110205	gtag	5038	12	0	1	no RNA
107	ctg14071	1108221	1110205	atag	1984	12	0	5	no RNA
108	ctg14101	1195665	1193266	gtag	2399	17	1	6	succinyl CoA:3-oxoacid CoA transferase precursor (OXCT)
109	ctg14101	2623729	2622933	gtag	796	14	0	5	no RNA
110	ctg14246	821462	817648	gtag	3814	12	3	22	Homo sapiens UBA3 (UBA3) mRNA, complete cds.
111	ctg14250	3519653	3517060	atac	2593	14	1	12	syntaxin 6 - intracellular transport
112	ctg14250	4973363	4970770	gtag	2593	14	3	0	voltage-operated calcium channel, alpha-1 subunit (CACNA1E)
113	ctg14294	495670	496370	gtag	700	17	1	22	unassigned RNA
114	ctg14357	629389	630300	gtag	911	14	1	37	unassigned RNA
115	ctg14374	382695	387822	gtag	5127	17	2	12	unassigned RNA
116	ctg14391	3791571	3790979	gtag	592	13	2	6	unassigned RNA
117	ctg14391	16583662	16583551	gtag	111	16	2	13	peroxisomal biogenesis factor 16 (PEX16)
118	ctg14409	807680	808697	gtag	1017	15	2	4	unassigned RNA
119	ctg14422	1769401	1770614	gtag	1213	20	2	2	actin-related protein
120	ctg14468	171935	166225	gtag	5710	13	0	1	no RNA
121	ctg14473	3901018	3900657	gtag	361	15	1	0	voltage-dependent, calcium channel alpha-2b subunit (CACNB2)
122	ctg14473	9976267	9977017	gtag	750	12	2	58	sorcin (SRI) - interacts with alpha subunit of calcium channels
123	ctg14489	1150032	1149419	atac	613	13	1	2	Matrilin-1 (MATN1/CMP)
124	ctg14489	3061271	3061558	gtag	287	12	2	284	translation initiation factor eIF3 p36 subunit
125	ctg14493	4961242	4956440	gtag	4802	12	1	3	phospholipase c delta 1
126	ctg14502	123828	124496	gtag	668	15	0	19	no RNA
127	ctg14583	4420669	4420129	gtag	540	16	0	12	no RNA
128	ctg14637	2733434	2729462	gtag	3972	12	3	19	unassigned RNA
129	ctg14715	1891368	1891157	gtag	211	12	0	2	no RNA
130	ctg14716	9916	9061	atac	855	12	5	8	unassigned RNA
131	ctg14716	561957	561336	gtag	621	16	2	1	Lowe oculocerebrorenal syndrome (OCRL)
132	ctg14748	308334	310737	gtag	2403	13	3	32	VHL binding protein-1 (VBP-1)

133	ctg14748	308334	311521	gtag	3187	12	0	1	no RNA
134	ctg14797	1949969	1962548	gtag	12579	15	0	4	no RNA
135	ctg14824	1647813	1648294	gtag	481	15	4	83	small nuclear RNA protein E
136	ctg14927	479891	478533	gtag	1358	15	0	5	no RNA
137	ctg14947	462503	463948	gtag	1445	14	1	1	unassigned RNA
138	ctg14950	244053	242815	atac	1238	10	1	1	unassigned RNA
139	ctg14950	1247199	1234330	atac	12869	11	1	1	unassigned RNA
140	ctg14977	1979429	1986609	atac	7180	11	14	0	T-type calcium channel alpha 1 subunit G (CACNA1G)
141	ctg15054	12070488	12075387	gtag	4899	14	1	11	unassigned RNA
142	ctg15054	17805347	17802943	gtag	2404	14	1	9	guanylate binding protein isoform II (GBP2) - GTPase
143	ctg15054	17878464	17878263	gtag	201	13	0	1	no RNA
144	ctg15054	17952420	17952158	gtag	262	12	0	1	no RNA
145	ctg15058	958544	959953	atac	1409	9	0	12	no RNA
146	ctg15064	13222495	13223157	gtag	662	15	3	7	tetraspan NET-6
147	ctg15064	19066287	19061872	atac	4415	11	1	11	unassigned RNA
148	ctg15064	27659180	27659995	atac	815	12	3	42	glycyl-tRNA synthetase
149	ctg15064	29526923	29526328	gtag	595	14	4	60	U6 snRNA-associated Sm-like protein LSm5
150	ctg15071	5016381	5016550	gtag	169	12	1	77	dynactin 3 subunit (p22) (DCTN3) intracellular movement
151	ctg15071	5726108	5725550	gtag	558	8	1	1	unassigned RNA
152	ctg15082	2783499	2793213	gtag	9714	20	0	1	no RNA
153	ctg15082	2931568	2930258	gtag	1310	15	0	5	no RNA
154	ctg15082	3913049	3911462	atac	1587	11	2	7	unassigned RNA
155	ctg15105	135358	135913	atac	555	15	0	9	no RNA
156	ctg15131	21411977	21413577	atac	1600	11	1	2	unassigned RNA
157	ctg15140	1730299	1730919	atat	620	13	0	14	no RNA
158	ctg15140	1730299	1733643	atac	3344	12	1	1	unassigned RNA
159	ctg15140	4494303	4494812	gtag	509	14	0	52	no RNA
160	ctg15174	959704	961245	gtag	1541	12	1	6	transportin2 (TRN2) - mediates import into nucleus
161	ctg15174	2527802	2527650	gtag	152	16	5	1	voltage dependent, calcium channel alpha 1a subunit (CACNA1A)
162	ctg15247	179006	179162	atac	156	12	3	4	Ran binding protein 11
163	ctg15256	1289716	1290291	gtag	575	9	2	18	unassigned RNA
164	ctg15279	8355332	8353611	gtag	1721	13	2	12	peroxisomal phytanoyl-CoA alpha-hydroxylase (PAHX)
165	ctg15285	4771601	4773174	gtag	1573	15	1	226	c-myc binding protein
166	ctg15285	7447913	7447436	gtgg	477	12	0	1	no RNA
167	ctg15285	8374087	8372939	atac	1148	12	1	2	primase subunit p48 (PRIM1) - DNA replication
168	ctg15296	1320787	1319254	gtag	1533	12	0	4	no RNA
169	ctg15361	11723280	11722879	atag	401	14	0	2	no RNA
170	ctg15380	220141	218801	gtag	1340	12	0	1	no RNA
171	ctg15422	678097	677654	atac	443	10	0	10	no RNA
172	ctg15424	16078856	16089564	atac	10708	12	1	1	voltage-gated sodium channel, alpha subunit (SCN2A)
173	ctg15424	16220741	16222344	atac	1603	12	3	0	voltage-gated sodium channel, alpha subunit (SCN3A)
174	ctg15424	17632082	17632418	atac	336	12	1	0	voltage-gated sodium channel, alpha subunit (SCN6A)
175	ctg15467	4828879	4828262	gtag	617	14	1	3	unassigned RNA
176	ctg15478	200862	208439	gtag	7577	14	1	9	tubulin-folding cofactor E
177	ctg15485	396014	395568	atac	446	10	0	1	no RNA
178	ctg15493	2991532	2992090	atac	558	10	1	195	heat shock factor binding protein 1 (HSBP1)
179	ctg15540	2379952	2380862	gtag	910	13	1	13	GAR1
180	ctg15556	280785	278872	atac	1913	9	2	1	diaphanous 2
181	ctg15583	362230	360031	atac	2199	10	1	11	CAGF28

182	ctg15590	223017	212044	gtag	10973	14	3	8	transcription factor WSTF
183	ctg1564	202394	203654	atac	1260	21	2	2	Ran binding protein 16
184	ctg15649	24839	24233	gtag	606	11	2	0	testicular protein (TSPY) mRNA
185	ctg15649	1517488	1518323	atac	835	10	0	4	no RNA
186	ctg15649	6183029	6180918	gtag	2111	15	3	5	unassigned RNA
187	ctg15665	7010226	7010145	atac	81	9	1	3	matrilin-4
188	ctg15665	7804915	7805058	gtag	143	14	1	2	unassigned RNA
189	ctg15665	7818487	7818696	gtag	209	12	1	0	unassigned RNA
190	ctg15665	14192682	14194352	atac	1670	15	1	1	unassigned RNA
191	ctg15665	16701597	16701116	atag	481	12	0	1	no RNA
192	ctg15665	20538455	20538796	gtag	341	9	2	0	SPO11
193	ctg15684	235800	235557	gtag	243	14	1	3	unassigned RNA
194	ctg15751	632335	633044	gtaa	709	9	0	1	no RNA
195	ctg15811	1354967	1359317	gtag	4350	14	2	15	unassigned RNA
196	ctg15855	608204	608132	gtag	72	12	1	0	unassigned RNA
197	ctg15907	22033710	22037477	atac	3767	11	1	2	E2F3 transcription factor
198	ctg15907	38170325	38170611	gtgg	286	12	0	9	no RNA
199	ctg15907	41619758	41621171	gtag	1413	13	4	3	stress-activated protein kinase 4 (SAPK4)
200	ctg15907	45248089	45243757	gtag	4332	13	0	4	no RNA
201	ctg15907	49142084	49140321	atac	1763	14	4	10	unassigned RNA
202	ctg15944	3676837	3676967	gtag	130	15	1	4	unassigned RNA
203	ctg15944	7438132	7440203	atac	2071	14	4	45	damage-specific DNA binding protein 1 (DDB1)
204	ctg15944	8059651	8059081	gtag	570	14	4	88	unassigned RNA
205	ctg15951	37181	38072	gtag	891	16	4	3	K-Cl cotransporter KCC4
206	ctg15951	217765	219484	gtag	1719	9	1	1	unassigned RNA
207	ctg15966	141492	142094	gtag	602	12	3	8	unassigned RNA
208	ctg15988	7782645	7781130	atac	1515	12	2	0	unassigned RNA
209	ctg15988	11926421	11924866	gtag	1555	15	1	68	unassigned RNA
210	ctg15988	13393246	13396290	gtag	3044	14	1	6	vacuolar ATPase
211	ctg15995	608008	608452	gtag	444	8	0	1	no RNA
212	ctg16008	18497646	18503408	gtag	5762	13	3	8	excision repair protein (ERCC5/XPG)
213	ctg16008	18525768	18526627	atat	859	11	2	17	excision repair protein (ERCC5/XPG)
214	ctg16106	815887	815369	atag	518	13	1	1	unassigned RNA
215	ctg16106	900543	903194	gtag	2651	15	1	70	MCT-1
216	ctg16166	5418471	5418020	atac	451	13	2	14	glia maturation factor beta
217	ctg16166	5467290	5467965	gtag	675	14	1	5	CGR19
218	ctg16170	732939	733805	gtag	866	13	3	5	huntingtin
219	ctg16235	274250	272440	gtag	1810	15	2	4	nummuscle myosin heavy chain-B(MYH10)
220	ctg16282	693960	694394	gtag	434	16	2	67	ionizing radiation resistance conferring protein (DAP-3)
221	ctg16307	2640943	2641913	gtag	970	13	0	2	no RNA
222	ctg16307	3656542	3658072	gtag	1530	16	1	9	protein phosphatase 2A alpha subunit mRNA,
223	ctg16335	751972	753283	atac	1311	13	0	2	no RNA
224	ctg16335	2141893	2143100	gtag	1207	12	3	11	RAB-like protein 2a or 2b ?
225	ctg16335	6765525	6768939	gtag	3414	13	2	5	insulin induced protein
226	ctg16391	891630	891155	atac	475	13	0	3	no RNA
227	ctg16408	1261706	1260537	atac	1169	14	0	41	no RNA
228	ctg16453	1353893	1353656	gtag	237	10	0	6	no RNA
229	ctg16478	747742	748168	gtag	426	16	3	56	poly(ADP-ribose) polymerase
230	ctg16537	2759032	2759131	gtag	99	11	0	4	no RNA

231	ctg16537	9860000	9860338	gtag	338	11	2	4	phospholipase c beta 4
232	ctg16537	13762255	13761061	gtag	1194	16	1	2	unassigned RNA
233	ctg16537	21773758	21776953	gtag	3195	12	1	38	N-terminal acetyltransferase complex ard1 subunit
234	ctg16537	21799041	21796764	atac	2277	12	4	9	unassigned RNA
235	ctg16551	142124	140613	gtag	1511	15	9	16	WDR10p/SPG
236	ctg16551	147305	145331	gtag	1974	16	6	9	WDR10p/SPG
237	ctg16574	125108	127679	gtag	2571	15	0	1	no RNA
238	ctg16574	356281	358062	atac	1781	15	1	0	E2F2 transcription factor
239	ctg16574	657348	653838	atac	3510	10	2	8	transcription elongation factor TFIIS.h
240	ctg16574	699775	691254	gtag	8521	18	2	1	transcription elongation factor TFIIS.h
241	ctg16586	2170190	2170756	atag	566	14	0	1	no RNA
242	ctg16586	2475706	2475047	atac	659	11	2	3	E2F1 transcription factor
243	ctg16704	2269288	2269537	gtag	249	12	6	7	erythroid K:Cl cotransporter (KCC1)
244	ctg16704	2275163	2275317	gtag	154	16	4	7	erythroid K:Cl cotransporter (KCC1)
245	ctg16704	2338461	2338354	gtag	107	12	2	4	autoantigen
246	ctg16704	2342900	2342799	gtag	101	13	1	3	autoantigen
247	ctg16733	591947	591844	gtag	103	19	2	0	calcium channel, alpha2/delta 2 (CACNA2D2)
248	ctg16733	591947	591850	gtat	97	13	4	0	calcium channel, alpha2/delta 2 (CACNA2D2)
249	ctg16733	1400085	1401149	gtag	1064	13	1	0	unassigned RNA
250	ctg16733	1553239	1552660	gtag	579	16	4	5	HIV-1 Vpr binding protein (VprBP)
251	ctg16734	3065528	3062694	gtag	2834	12	1	6	unassigned RNA
252	ctg16736	59874	59737	gtag	137	11	1	5	DNA binding regulatory factor (RFX5)
253	ctg16751	982697	982896	gtag	199	13	2	24	unassigned RNA
254	ctg16827	93452	102023	atac	8571	14	0	1	no RNA
255	ctg16842	2089152	2085000	gtag	4152	14	2	0	voltage-dependent, calcium channel alpha-1D subunit (CACNA1D)
256	ctg16842	3242482	3242700	atac	218	12	2	4	calcium channel alpha2-delta3 subunit (CACNA2D3)
257	ctg16861	362833	353907	atac	8926	12	1	20	cullin 4A (CUL4A)
258	ctg16861	362833	362556	atac	277	12	0	1	unassigned RNA
259	ctg16861	384494	385448	atac	954	9	4	25	unassigned RNA
260	ctg16920	87556	84880	gtag	2676	15	2	144	calcium dependent protease
261	ctg16938	622316	618364	gtag	3952	18	0	1	no RNA
262	ctg17028	225680	224089	gtag	1591	9	4	12	inositol hexakisphosphate kinase 2 (IP6K2)
263	ctg17036	610045	611041	atag	996	13	2	3	cAMP-binding guanine nucleotide exchange factor III (cAMP-GEFIII)
264	ctg17042	511298	511012	gtag	286	14	2	1	phosphlipase C beta 3
265	ctg17042	744623	740530	gtag	4093	16	1	25	ADP-ribosylation factor-like protein 2 (ARL2)
266	ctg17042	1179863	1179356	atac	507	14	5	7	calcium- and diacylglycerol-regulated guanine nucleotide exchange factor I (CalDAG-GEFI)
267	ctg17042	1585240	1585700	gtag	460	12	2	35	calcium activated neutral protease large subunit
268	ctg17042	1643482	1644584	gtag	1102	14	2	11	DNA polymerase alpha
269	ctg17042	2319296	2319390	gtag	94	11	2	17	Dr1-associated corepressor (DRAP1)
270	ctg17042	2862671	2861338	gtag	1333	13	3	18	unassigned RNA
271	ctg17057	150869	148794	atac	2075	15	0	10	no RNA
272	ctg17064	105526	105725	gtag	199	15	1	48	unassigned RNA
273	ctg17077	14755	14975	gtag	220	19	0	4	no RNA
274	ctg17082	2828242	2829125	gtag	883	17	0	5	no RNA
275	ctg17129	1016435	1017201	gtag	766	12	0	47	no RNA
276	ctg17189	530732	537766	gtag	7034	16	0	1	no RNA
277	ctg17207	20648	21739	atac	1091	11	2	5	unassigned RNA
278	ctg17207	407037	404994	gtag	2043	14	0	17	no RNA
279	ctg17326	1011062	1010964	atac	98	10	2	9	proliferating-cell nucleolar protein P120

280	ctg17326	1025477	1025367	gtag	110	15	1	22	Mi-2 putative helicase
281	ctg17326	1455072	1454646	atac	426	11	4	98	B-cell receptor associated protein (BAP/REA)
282	ctg17328	3847091	3848228	gtag	1137	14	0	2	no RNA
283	ctg17328	3850251	3852930	gtag	2679	16	1	0	unassigned RNA
284	ctg17403	1436192	1435940	atac	252	9	2	10	protein kinase Njmu-R1
285	ctg17403	1448312	1449770	gtag	1458	14	1	60	zinc finger transcription factor (ZNF207)
286	ctg17436	62912	59269	gtag	3643	16	1	113	3-phosphoinositide dependent protein kinase-1 (PDK1)
287	ctg17441	554593	556064	gtag	1471	14	2	37	unassigned RNA
288	ctg17441	554593	557340	gtag	2747	10	0	1	no RNA
289	ctg17441	1509924	1507805	atac	2119	13	1	11	ras guanine nucleotide releasing factor (C3G/GNRP)
290	ctg17450	79338	78013	atac	1325	10	2	11	c6.1A protein
291	ctg17450	632419	632981	gtag	562	13	2	1	muscle-specific serine kinase 1 (MSSK1)
292	ctg17450	1217559	1217245	gtag	314	16	1	12	no RNA
293	ctg17490	3218237	3217909	gtag	328	18	1	1	unassigned RNA
294	ctg17492	287351	287665	gtag	314	17	0	10	no RNA
295	ctg17565	132197	132405	ataa	208	13	1	0	arginine methyltransferase
296	ctg17565	1280410	1276853	gtag	3557	15	0	1	no RNA
297	ctg17568	404113	407685	gtag	3572	8	0	1	no RNA
298	ctg17689	543157	540679	atac	2478	11	2	7	unassigned RNA
299	ctg17714	1177567	1185148	gtag	7581	15	0	1	no RNA
300	ctg17737	961395	962640	gtag	1245	15	1	19	unassigned RNA
301	ctg17748	4534286	4534924	gtag	638	13	1	14	Homo sapiens cyclin K (CPR4) mRNA, complete cds.
302	ctg17748	4936163	4938548	gtag	2385	12	1	1	echinoderm microtubule-associated protein homolog HuEMAP
303	ctg17777	5860664	5847241	gtag	13423	19	0	2	no RNA
304	ctg17777	5860664	5847279	gtag	13385	16	0	1	no RNA
305	ctg17782	10920	9461	gtag	1459	16	1	2	S2P
306	ctg17872	240784	241252	atac	468	12	1	0	unassigned RNA
307	ctg17887	5412	21221	gtag	15809	13	2	2	TFIIB related factor hBRF (HBRF)
308	ctg17936	210218	210813	atac	595	12	1	21	Not56-like protein
309	ctg18037	948689	947897	gtag	792	16	2	3	B-raf
310	ctg18037	11933911	11934471	atac	560	13	3	12	serine threonine protein kinase (PSSALRE)
311	ctg18037	11938328	11938180	gtag	148	17	4	0	proton-gated cation channel ASIC3 mRNA,
312	ctg18073	67763	68524	atac	761	13	2	51	spermidine aminopropyltransferase
313	ctg18073	82361	83959	gtag	1598	13	2	34	spermidine aminopropyltransferase
314	ctg18126	322495	321974	atac	521	11	0	7	no RNA
315	ctg18147	809169	810258	gtag	1089	15	0	161	no RNA
316	ctg18251	465553	466711	gtag	1158	16	1	1	unassigned RNA
317	ctg18306	1072422	1071862	atac	560	13	1	2	unassigned RNA
318	ctg18322	250426	251296	gtag	870	15	1	100	neutrophil cytochrome b light chain p22 phagocyte b-cytochrome
319	ctg18348	4421	5521	atac	1100	12	1	6	H.sapiens mRNA for stress activated protein kinase-3
320	ctg18433	380147	382873	gtag	2726	18	0	18	no RNA
321	ctg18539	96086	95515	gtag	571	16	2	1	inositol polyphosphate 5-phosphatase (5ptase)
322	ctg18539	141393	140991	atag	402	15	1	0	inositol polyphosphate 5-phosphatase (5ptase)
323	ctg18591	458333	458466	gtag	133	15	1	2	DOCK180 - cytokinesis related
324	ctg18647	69156	69368	gtag	212	12	1	16	phosphoethanolamine cytidylyltransferase
325	ctg18730	733964	735578	gtag	1614	13	2	36	unassigned RNA
326	ctg18743	128944	127430	gtag	1514	13	2	4	unassigned RNA
327	ctg18743	2159516	2160690	atac	1174	11	1	3	unassigned RNA
328	ctg18846	7470681	7473601	atac	2920	16	4	25	E2F5 transcription factor

329	ctg18867	4302903	4290778	gtag	12125	15	1	4	phosphatase 2A beta subunit
330	ctg18891	175534	173705	gtag	1829	17	5	171	unassigned RNA - muscle specific
331	ctg18961	67022	66819	ataa	203	12	0	1	no RNA
332	ctg19078	317626	317854	gtag	228	13	1	44	FX protein
333	ctg19085	380214	379830	atac	384	10	0	17	no RNA
334	ctg19153	800390	798728	gtag	1662	13	0	1	no RNA
335	ctg19153	1479008	1476247	gtag	2761	16	3	0	Human xanthine dehydrogenase/oxidase mRNA,
336	ctg19153	9470504	9468833	atat	1671	11	1	10	unassigned RNA
337	ctg19153	13443567	13445415	gtag	1848	17	1	3	no RNA
338	ctg19247	6229884	6223216	atac	6668	12	1	1	huntingtin interacting protein (HIP1)
339	ctg19247	6512938	6514398	gtag	1460	15	1	30	unassigned RNA
340	ctg19262	1221484	1221690	gtag	206	11	3	84	26S proteasome ATPase subunit (PSMC4/MIP224)
341	ctg21fin2	20530904	20530455	gtag	449	14	4	9	B17
342	ctg21fin2	20544684	20540162	atac	4522	17	2	9	quinone oxidoreductase homolog-1
343	ctg21fin2	20544684	20540183	atac	4501	10	0	3	no RNA
344	ctg21fin2	23286697	23287523	atac	826	13	1	3	unassigned RNA
345	ctg21fin2	23293458	23293661	gtag	203	10	1	0	unassigned RNA
346	ctg21fin5	812324	812633	gtag	309	14	2	3	GT334 protein - sodium transporter?
347	ctg21fin5	875982	877440	atac	1458	11	4	33	GT335/HES1
348	ctg22fin10	269229	269115	gtag	114	12	3	2	ERK6 mRNA for extracellular signal regulated kinase
349	ctg22fin3	2640958	2642861	gtag	1903	13	4	7	CDC45L - cell cycle regulator
350	ctg22fin3	3034523	3033606	gtag	917	15	2	2	thioredoxin reductase II
351	ctg22fin4	1460899	1459276	gtag	1623	14	3	30	mitogen-activated protein kinase (MAPK1)
352	ctg22fin4	3476936	3476440	gtag	496	12	1	20	unassigned RNA
353	ctg22fin4	9498501	9497117	gtag	1384	14	2	7	unassigned RNA
354	ctg22fin4	10762104	10761734	gtag	370	13	0	22	no RNA
355	ctg22fin4	15925520	15924660	gtag	860	12	2	4	nonmuscle myosin heavy chain-A (MYH9)
356	ctg22fin4	18174711	18172305	atac	2406	14	0	1	no RNA
357	ctg22fin4	21107040	21105836	gtag	1204	10	1	5	unassigned RNA
358	ctg22fin4	21383941	21384825	gtag	884	16	1	3	unassigned RNA
359	ctg24	7455313	7460459	gtag	5146	15	2	4	unassigned RNA
360	ctg25099	1737750	1730988	gtag	6762	12	1	35	serine kinase SRPK2
361	ctg25099	16016139	16017537	gtag	1398	13	1	9	U6 snRNA-associated Sm-like protein LSm8 mRNA,
362	ctg25099	28277769	28269840	atac	7929	10	2	29	unassigned RNA
363	ctg25099	28574133	28573075	gtag	1058	17	2	3	unassigned RNA
364	ctg25122	37541	35997	atac	1544	11	1	18	RNP L protein
365	ctg25256	3760070	3773188	atac	13118	12	1	0	cAMP-regulated guanine nucleotide exchange factor II (cAMP-GEFII)
366	ctg25448	1509029	1509768	gtag	739	20	1	5	Homo sapiens gamma SNAP mRNA, complete cds.
367	ctg25478	890414	892141	gtag	1727	15	2	0	protein kinase JNK1alpha
368	ctg25478	890840	892141	gtag	1301	15	2	1	protein kinase JNK1beta
369	ctg2712	1031063	1037862	gtag	6799	15	0	3	no RNA
370	ctg2712	1088091	1087917	gtag	174	11	0	2	no RNA
371	ctg2712	2077364	2077225	gtag	139	12	1	1	phosphlipase C beta 2
372	ctg2712	4131796	4129425	atac	2371	12	1	10	unassigned RNA
373	ctg2712	6749294	6751075	gtag	1781	14	0	1	no RNA
374	ctg2712	13478834	13480499	atac	1665	13	2	0	adaptor-related protein complex AP-4 epsilon subunit
375	ctg28041	686879	689242	gtag	2363	14	1	8	unassigned RNA
376	ctg28078	495770	493168	gtag	2602	16	1	40	unassigned RNA
377	ctg28098	2577564	2571847	gtag	5717	16	1	2	small GTPase RAB6B

378	ctg2810	968984	971060	atac	2076	9	1	9	unassigned RNA
379	ctg30153	8866080	8868016	atac	1936	10	1	26	no RNA
380	ctg3235	1449702	1453964	atac	4262	13	2	0	unassigned RNA
381	ctg3454	1046368	1045656	gtag	712	14	2	8	hypothetical transmembrane protein SBBI53
382	ctg3454	1047881	1046463	gtag	1418	14	2	18	hypothetical transmembrane protein SBBI53
383	ctg3461	2469486	2466933	gtag	2553	11	3	3	CSNK1G1 mRNA for casein kinase 1 gamma 1
384	ctg36	8022436	8022646	atac	210	12	2	1	cAMP-regulated guanine nucleotide exchange factor I cAMP-GEFI
385	ctg36	8972276	8974265	gtag	1989	15	2	0	cyclin T1
386	ctg36	9031169	9043672	gtag	12503	14	1	1	unassigned RNA
387	ctg36	9031169	9045709	gtag	14540	15	1	3	unassigned RNA
388	ctg3678	130684	134088	gtag	3404	11	2	18	Ca2-activated neutral protease large subunit (CANP)
389	ctg38	42393	42552	gtag	159	13	2	27	unassigned RNA
390	ctg44	2600507	2598508	atac	1999	13	2	17	matrilin-2
391	ctg45	15904279	15917906	gtag	13627	13	1	28	N33 - putative membrane protein
392	ctg46	63383	60928	gtag	2455	15	2	1	LLGL - putative cytoskeletal protein
393	ctg46	169621	168789	gtag	832	15	1	23	GTP-binding protein (DRG2)
394	ctg53	3106670	3095974	gtag	10696	18	1	2	unassigned RNA
395	ctg53	4278471	4278361	gtag	110	12	1	3	unassigned RNA
396	ctg53	4363072	4362407	gtag	665	15	1	38	rab geranylgeranyl transferase, alpha-subunit
397	ctg53	25845384	25843610	gtag	1774	13	3	47	rapamycin binding protein (FKBP3)
398	ctg595	7436975	7443108	atac	6133	13	0	2	no RNA
399	ctg595	11831740	11831262	gtag	478	13	1	10	nuclear pore complex protein (NUP107)
400	ctg7013	207167	209191	gtag	2024	12	1	3	H.sapiens mRNA for human giant larvae homolog
401	ctg78	98515	100884	atac	2369	10	2	2	sodium-hydrogen exchanger 6 (NHE-6)
402	ctg78	104694	110650	gtag	5956	18	2	1	sodium-hydrogen exchanger 6 (NHE-6)
403	ctg78	121558	128138	gtag	6580	10	2	2	sodium-hydrogen exchanger 6 (NHE-6)
404	ctg9872	61949	64475	gtag	2526	20	0	15	no RNA