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 with 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	419	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, whch 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). Ths 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

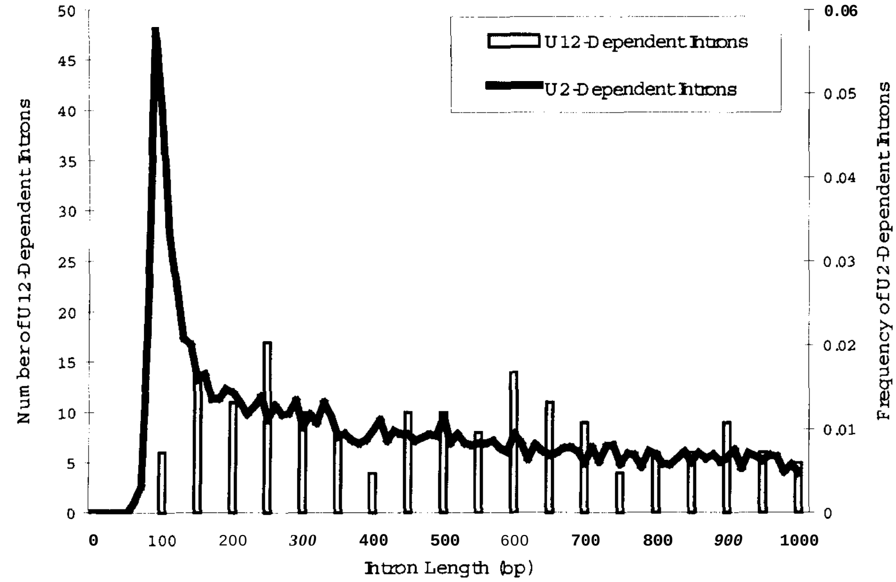

Figure 4.1 – The length of 168 U12-dependent introns and 11,402 RefScq-confirmed U2-dependent introns less than lkb in length are plotted. Grey bars represent the counts of U12-dependent introns grouped into 50-bp wide bins and the black line represents the frequency of U2-dependent introns grouped in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

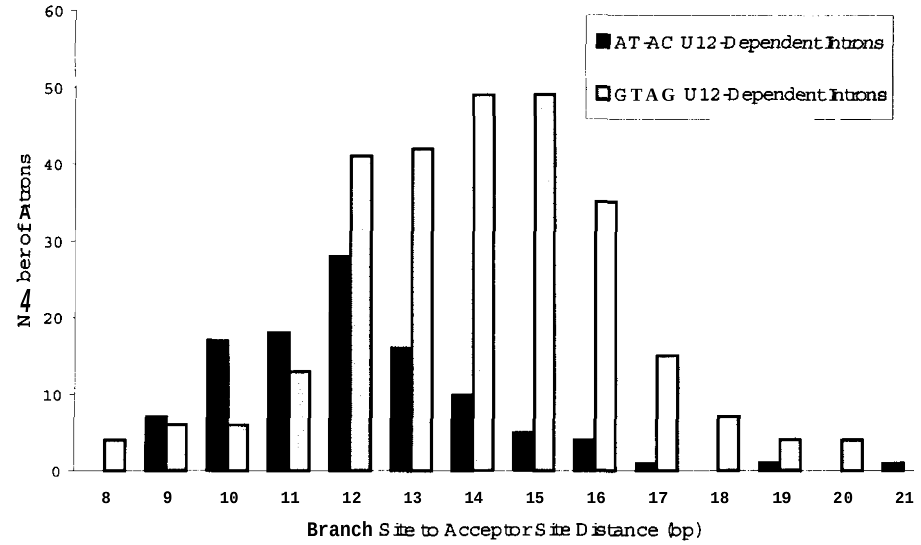

Figure 4.2 – The distance between the branch site and the acceptor site is plotted for 108 AT-AC U12-dependent introns (black bars) and 275 GT-AG U12-dependent introns (grey bars).

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 whch 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 sdar 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 with 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-sensitive L-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 hgher 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.