

Investigating the Intron Recognition Mechanism in Eukaryotes

Lesley Collins and David Penny

Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand

Recent studies indicate that many introns, as well as the complex spliceosomal mechanism to remove them, were present early in eukaryotic evolution. This study examines intron and exon characteristics from annotations of whole genomes to investigate the intron recognition mechanism. Exon definition uses the exon as the unit of recognition, placing length constraints on the exon but not on the intron (allowing it a greater range of lengths). In contrast, intron definition uses the intron itself as the unit of recognition and thus removes constraints on internal exon length forced by the use of an exon definition mechanism. Thus, intron and exon lengths within a genome can reflect the constraints imposed by its splicing. This study shows that it is possible firstly to recover valid intron and exon information from genome annotation. We then compare internal intron and exon information from a range of eukaryotic genomes and investigate possible evolutionary length constraints on introns and exons and how they can impact on the intron recognition mechanism. Results indicate that exon definition-based mechanisms may predominate in vertebrates although the exact system in fish is expected to show some differences with the better characterized system from mammals. We also raise the possibility that the last common ancestor of plants and animals contained some type of exon definition and that this mechanism was replaced in some genes and lineages by intron definition, possibly as a result of intron loss and/or intron shortening.

Introduction

Introns, small nuclear (sn)RNAs, and spliceosomal components have now been characterized from a number of eukaryotic lineages, some of which were once thought to be early branching, consistent with the premise that both introns and the spliceosomal machinery to process them, were present in the last common ancestor of eukaryotes. Although it is now accepted that introns are ancestral gene features of eukaryotes (Fedorov, Merican, and Gilbert 2002; Rogozin et al. 2003, 2005; Collins and Penny 2005; Roy and Gilbert 2005a, 2005b), we need to know more about the likely characteristics of these ancient introns. In modern eukaryotes, intron and exon sizes are extremely variable, for example, the average intron size for humans is ~5 kb (Sakharkar, Chow, and Kanguene 2004), whereas the yeast *Schizosaccharomyces pombe* has an average intron length of only 107 nt (Kupfer et al. 2004). Examination of intron and exon characteristics can reveal the nature of the underlying mechanism that recognizes the intron in order for it to be spliced. We start with what has been learned about the spliceosome in ancestral eukaryotes by adding intron and splicing recognition information to increase our knowledge of splicing-related mechanisms in ancestral eukaryotes. We first examine intron and exon length constraints and how they relate to the splicing recognition system in extant eukaryotes and then see whether we can use this information to infer ancestral systems.

Comparative analysis of the spliceosomal complex (Collins and Penny 2005) indicates that not only was a spliceosome likely to be present in the eukaryotic ancestor but it also contained most of the key components (RNA and proteins) found in today's eukaryotes. It appears that the spliceosome had already formed links with other cellular

processes such as transcription, capping, and transport from the nucleus. One of the very first steps in the splicing process is the initial interaction between components of the spliceosome and the intron/exon boundaries (splice sites) of the unprocessed mRNA. Up to 15% of human genetic diseases are caused by point mutations that occur at or near the splice sites where they most likely result in aberrant splicing (Clark and Thanaraj 2002).

In 1990, Robberson et al. proposed the exon definition model in vertebrates where internal exons (exons between introns) are recognized as units before prespliceosomal assembly. The binding of the U1 small nuclear ribonucleoprotein (snRNP) at the downstream 5' splice site stabilizes U2AF binding across the exon at the upstream 3' splice site (Hoffman and Grabowski 1992; Cunningham, Hagan, and Grabowski 1995). The bridging of the exon is achieved by the binding of serine-/arginine-rich (SR) proteins to the proteins already bound at the 3' and 5' splice sites (Wu and Maniatis 1993; Graveley, Hertel, and Maniatis 1999; Furuyama and Bruzik 2002) (fig. 1). SR proteins contain an extensively phosphorylated RS domain, rich in arginine and serine residues that promotes protein-protein interactions and directs subcellular localization (Cazalla et al. 2002). This domain can also interact directly with the intronic branch point site (Ibrahim el et al. 2005). Under this mechanism, exonic splicing enhancer (ESE) sites, which are binding sites for SR proteins (Lam and Hertel 2002), act as barriers to prevent exon skipping. ESE sites also play a key role in ensuring the correct linear order of exons (Ibrahim el et al. 2005). The presence of ESEs, however, is not a sufficient indicator of exon definition because ESEs have also been found both in intronless genes (Pozzoli et al. 2004) and in introns (Zhang and Chasin 2004) and are also implicated in RNA export (Pozzoli et al. 2004). Similarly, SR proteins cannot be used as an indicator of exon definition as they are involved at multiple stages of the splicing reaction (Furuyama and Bruzik 2002) and have been found in species that show no evidence for exon definition. With exon definition, the most common consequence of

Key words: intron definition, exon definition, molecular evolution.

E-mail: lj.collins@massey.ac.nz.

Mol. Biol. Evol. 23(5):901–910, 2006

doi:10.1093/molbev/msj084

Advance Access publication December 21, 2005

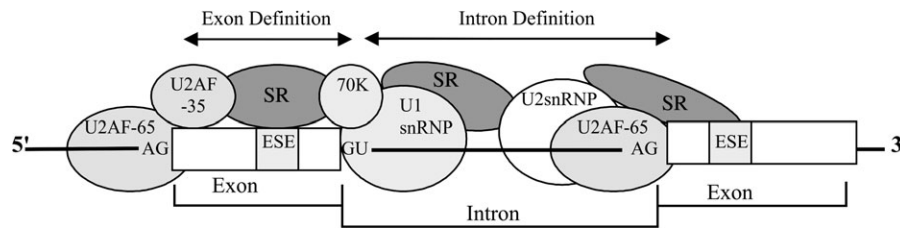


FIG. 1.—Exon and intron definition models for splice site recognition (based on Lorkovic et al. 2000; Romfo et al. 2000; Lam and Hertel 2002). The boundaries between the introns and exons are recognized through the binding of multiple proteins either across the exon (exon definition) or across the intron (intron definition).

mutations, in either the 5' splice site of the downstream intron or the 3' splice site of the upstream intron, is exon skipping where both the introns and the exon are omitted from the processed mRNA (Berget 1995; Black 1995; Simpson et al. 1999). Another consequence can be the activation of a cryptic splice site (Black 1995).

Exon definition in vertebrates appears to be optimally effective when the exon size is between 50 nt and 500 nt (Robberson, Cote, and Berget 1990; Berget 1995; Black 1995) because expanding or reducing the exon to outside these limits can either cause cryptic splice sites to be activated or the exon to be skipped (Berget 1995). The exon definition mechanism has been shown to be present in vertebrates (Berget 1995; Black 1995) and plants (McCullough, Baynton, and Schuler 1996; Simpson et al. 1999). With exon definition there appears to be little constraint on the length of the intron (Berget 1995), and it has been shown that intron size does not affect exon skipping (Ibrahim et al. 2005). Although long exons (>500 nt) and short exons (<50 nt—micro exons) are still present and efficiently spliced (Berget 1995; Niu, Hou, and Li 2005), these exons are more likely to be alternative exons or to have additional features that enhance their recognition (Bruce and Peterson 2001). A recent example of an alternative splicing strategy is the use of “recursive” splicing in *Drosophila melanogaster* where large exons are “subdivided” before intron splicing (Burnette et al. 2005). Small exons often use intronic splicing enhancers or intronic splicing silencers, small sequence motifs found within the intron to either promote or silence splicing, respectively (Simpson et al. 1999). Studies have also shown that increasing the length of the polypyrimidine (Py) tract, or perhaps the positioning of the Py tract, can prevent exon skipping of short exons (Dominski and Kole 1991).

However, in some species, short introns are defined directly without the need for exon definition (Romfo et al. 2000). This is termed intron definition (fig. 1), and here a mutation of a 5' splice site does not lead to exon skipping but instead the mutated intron is included in the final mRNA (intron inclusion). The yeast *S. pombe* appears to use the intron definition mechanism exclusively (Romfo et al. 2000). Other species such as *Caenorhabditis elegans*, *D. melanogaster* (Berget 1995; Romfo et al. 2000), and plants (Lorkovic et al. 2000), where there are small as well as large introns, appear to have both intron and exon definition mechanisms in operation. In vertebrates, it has been shown that large exons were incompatible for splicing only if they were flanked by large introns (Stern, Carlo, and

Berget 1996). When large exons were flanked by short introns, intron definition could then be used (Stern, Carlo, and Berget 1996). In *D. melanogaster*, it has been shown that it is possible for intron definition and exon definition to operate in a single pre-mRNA (Kennedy, Kramer, and Berget 1998).

Not all eukaryotic introns are removed by the same type of spliceosome. Minor introns (sometimes called AT-AC or U12-dependent introns) are removed using a spliceosome that contains U11 and U12 snRNPs instead of the U1 and U2 snRNPs of the major (or U2 dependent) spliceosome. Minor-spliced introns are found in vertebrates, plants, and some invertebrates (Burge, Padgett, and Sharp 1998; Patel and Steitz 2003). SR proteins, important for exon definition, appear to bind to the U11 snRNP in the same way as they do to the U1 snRNP (Lorkovic et al. 2005). In minor splicing, the U11 and U12 snRNPs exist as a stable di-snRNP complex (Lorkovic et al. 2005), unlike the separate complexes of U1 and U2 snRNPs found in major splicing. Therefore, it is possible that due to this, there are some differences in the precise intron recognition mechanism used for major and minor introns. However, minor introns make up only a small percentage of introns found in all species examined to date.

Trans-splicing, the joining together of two independently transcribed transcripts occurs naturally in many eukaryotic lineages (nematodes, insects, and trypanosomes) (Mayer and Floeter-Winter 2005). Intron boundaries are defined by the same elements observed in *cis*-splicing (major and minor splicing) although the 5' splice site is present on the spliced-leader-RNA molecule while the 3' splice site is present in the pre-mRNA. SR proteins and ESEs have been shown to interact in the same way in *trans*-spliced genes as for *cis*-spliced genes (reviewed in Mayer and Floeter-Winter 2005).

Alternative splicing, where one gene can result in multiple products, is also found in many eukaryotic species (reviewed in Graveley 2001; Sorek et al. 2004; Thanaraj et al. 2004). SR proteins and heterogeneous nuclear RNPs have important roles in alternative splicing and are involved in binding to exonic and intronic repressor sites (Maniatis and Tasic 2002). In mammalian cells, the Py tract-binding protein complex is a key splicing repressor of exon definition and causes the exon to be skipped (i.e., it is incorporated into the intron) (Wagner and Garcia-Blanco 2001). Alternative splicing is also found in *Plasmodium* species where introns from the *var* genes can act as transcriptional silencing elements that help control antigenic

variations (Gannoun-Zaki et al. 2005). Alternative splicing can operate using either the *cis*- (major and minor) or *trans*-splicing mechanism (Maniatis and Tasic 2002).

The *Introduction* thus far has only considered introns and exons away from the ends of the gene. It is useful to classify exons into three groups, first exons, internal exons, and final exons. Terminal exons (first and final) require different mechanisms for their recognition from internal exons (Berget 1995; Cooke, Hans, and Alwine 1999; Gornemann et al. 2005). First exons are connected to the 5'-capping system (Gornemann et al. 2005) while the last exon is connected to the 3'-polyadenylation system (Berget 1995; Cooke, Hans, and Alwine 1999), and hence their lengths are expected to fall into different ranges than internal exons. Introns are similarly classed into four groups, first introns, internal introns, and final introns and also a class from genes with only single introns. The first and final introns can be separated from the middle introns to exclude introns that lie within the 5'-untranslated region (UTR) and the 3' UTR of the pre-mRNA (Mignone et al. 2002) and which may be subjected to different length constraints. It has also been reported that first introns are often enriched in regulatory elements (Marais et al. 2005) and thus will be subjected to different evolutionary constraints than internal introns.

With both exon definition and intron definition, internal intron and exon lengths are expected to reflect constraints imposed by specific mechanisms to each species. However, only a few species have had their intron recognition system characterized experimentally. Therefore, we have collated intron and exon information from complete and nearly completely sequenced eukaryotic genomes to compare patterns of intron and exon lengths with species for which the intron recognition mechanisms are known. We first need to check the validity of intron/exon information from whole genome annotations in order to show that whole well-annotated genomes generate similar data to annotations of genomes of expressed sequence tag (EST) data sets. If these data sets produce similar results, then the annotations are valid vehicles for in-depth intron and exon analysis. Similarly, we then analyze information from constitutively and alternatively spliced genes to investigate whether internal introns and exon lengths show differences due to the presence of alternative splicing. Thirdly, we examine length information from many species showing how both intron and exon information, even from genomes that are not highly annotated, can provide length constraint information for internal introns and exons.

Our results indicate that exon definition-based mechanisms may predominate in vertebrates, although the exact system in fish is expected to show some differences from the better characterized system from mammals such as human and mouse. We also raise the possibility that the last common ancestor of plants and animals contained some type of exon definition and that this mechanism was replaced by intron recognition possibly as a result of intron loss and/or intron shortening in some genes and lineages.

Materials and Methods

The complete genomes of mouse (*Mus musculus* Build 33.1), rat (*Rattus norvegicus* Build 2.1), human

(*Homo sapiens* Build 35.2), *Saccharomyces cerevisiae* (NC_001133-48), *S. pombe* (NC_003421, NC_003423, NC_003424), *Encephalitozoon cuniculi* (Build 1.1), *C. elegans* (Build 1.1), *D. melanogaster* (Build 4.0), *Caenorhabditis briggsae* (CAAC00000000.1), *Cryptococcus neoformans* (NC_006670, NC_006679-87, NC006691-4), *Arabidopsis thaliana* (NC_003070.5, NC_003071.3, NC_003074.4, NC_003075.3, NC_003076.4), rice (*Oryza sativa* Build 2.1), *Plasmodium falciparum* (Build 1.1), *Entamoeba histolytica* (AAFB00000000.1), *Cryptosporidium parvum* (AAEE00000000), and *Dictyostelium discoideum* (AAFI00000000.1) were downloaded from the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov>). The complete genomes of chicken (*Gallus gallus* Build 29.1e), zebrafish (*Danio rerio* Build 29.4e), pufferfish (*Takifugu rubripes* Build 29.2e and *Tetraodon nigroviridis* Build 29.1b), sea squirt (*Ciona intestinalis* Build 31.195), mosquito (*Anopheles gambiae* Build 29.2e), and honeybee (*Apis mellifera* Build 29.1a) were downloaded from the Ensemble genome database (<http://www.ensembl.org/index.html>). EST-supported information for human, *C. elegans*, *C. neoformans*, and *A. thaliana* was extracted from their annotated genomes into separate data sets. The EST data set for *D. discoideum* was downloaded from Dictybase (<http://dictybase.org/>) and then linked to its annotated genome. Genome status (shown in table 3) was determined by the level of EST information contained within the genome. Genomes with predicted gene information only without using EST data linked to annotation were given the genome status of 0. Genomes with some experimental or EST data included in their annotation were given the status of 1. Genomes with EST data sets that could be linked to EST data sets using a transcript_id included in the genome annotation were given the status of 2. Genomes that contained EST and experimental evidence directly within the genome annotation were given the status of 3. Although EST data sets for honeybee, *C. intestinalis*, and *P. falciparum* were available, the linkage of the EST information to the genome annotation was not clear, so the genome status of these species was set at 1.

Gene, intron, and exon information was extracted from the annotated genomes using a perl script "intron_finder.pl" (available upon request). Intron_finder.pl reads an annotated genome file, and outputs database records of the introns found. The input file may be in any format which is understood by Bioperl (<http://www.bioperl.org>). It may contain several accessions and an accession name genes. A gene may have several splicing products with each product containing introns, exons, and notes. Other perl scripts (also available upon request) were used to sort and collate internal intron and exon information from the output of intron_finder.pl.

For the purpose of this study, a minimum intron length was determined from the literature and set at 12 nt for animals, plants, and fungi and 8 nt for protists. Major and minor splicing information was not distinguished during this study as the number of introns and exons represented by minor splicing is expected to be small (estimated as 26/17,408 in primates and 11/19,553 in *A. thaliana* [Burge, Padgett, and Sharp 1998]) and underestimated in a number

Table 1
Comparison of Whole Genome and EST-Based Data Sets

		Internal Exons				Internal Introns			
		Number	5%	50%	95%	Number	5%	50%	95%
<i>Caenorhabditis elegans</i>	Genome	92,169	66	154	574	73,577	44	63	1,135
	EST	53,175	68	157	585	43,297	44	64	1,184
<i>Arabidopsis thaliana</i>	Genome	126,192	49	115	448	103,820	85	97	371
	EST	12,606	50	114	444	14,318	75	97	365
Human (<i>Homo sapiens</i>)	Genome	283,216	51	124	289	252,375	98	1422	18,062
	EST	39,853	48	124	332	35,458	97	1464	19,738
<i>Cryptococcus neoformans</i>	Genome	32,443	30	142	754	25,874	46	55	113
	EST	14,394	31	150	805	18,431	46	54	95
<i>Dictyostelium discoideum</i>	Genome	8,208	36	246	1,391	3,701	53	99	301
	EST	7,124	35	249	1,387	3,228	50	99	304

of genomes. Statistics were done using the *R* statistics package (version 2.1.0; <http://www.r-project.org/>).

Results

The first step was to assess the accuracy of the intron and exon length data obtained from annotated genomes. We did this by comparing data sets from whole genome annotations to data sets containing information from EST-confirmed genes. The species used here (human, *C. elegans*, *A. thaliana*, *C. neoformans*, and *D. discoideum*) represented a cross section of the levels of genomic information presently available. These species vary from well-annotated genomes (*C. elegans* and *A. thaliana*), to recently sequenced genomes with EST information included in the annotation (*C. neoformans*), to sequenced genomes with EST data sets available separately (*D. discoideum*), and to genomes where computer predicted information is linked to supporting EST information (human) within the genome annotation. *Dictyostelium discoideum* is often placed as a basal eukaryote on eukaryotic trees, but recently its placement is under debate and is now suggested to be placed between the fungi + animal and the plant lineages (Williams, Noegel, and Eichinger 2005).

In every genome there will always be special cases that result in very long or very short introns and/or exons as well as occasional misannotations. We wished in this study to use typical events, and therefore internal exon and internal lengths were compared at the 5%, 50% (median), and 95% "quartiles" to remove biases that result from occasional very long or short lengths.

Results (table 1) showed that for internal exon lengths there is little difference between whole genome and EST-based data sets, thus giving confidence that the annotated genomes were acceptable for further analysis. The 5% and median quartiles differed normally by only a few nucleotides (except in *C. neoformans* where the difference was 8 nt at the median). Larger differences were seen at the 95% quartiles with the largest difference again with *C. neoformans* (51 nt). Similarly, internal introns showed only a few nucleotides different at the 5% and median quartiles and larger differences at the 95% quartiles (the human data sets produced the largest intron differences). From a statistical point of view, the differences are seen as significant. Statistical analysis (e.g., *t*-tests, Wilcoxon, Kolmogorov-Smirnov tests) between the whole genome and EST data sets tended

to highlight even a single nucleotide difference as being highly significant due to the large amount of data in the data sets and the skewed distribution of lengths. However, from a biochemical viewpoint the differences are well within variation that can be handled by the biological mechanism.

The results between whole genome and EST data sets are also a reflection on the accuracy of information from predicted and experimentally confirmed annotation. It should be remembered that predicted annotation, especially of splice sites, is always trained using experimentally confirmed data. Early annotation is expected to underestimate the amount of splicing as unusual (e.g., minor) or species-specific splicing may not be accurately predicted using computer prediction. On the other hand, another reason for discrepancies between whole genome and EST databases is that EST transcripts may not be available from all tissues, developmental stage, and cell types associated with a particular species and represent only a slice of genomic information. Although there are now a number of completely sequenced eukaryotic genomes available, they vary in the amount of experimental information linked to the annotation. We concluded from our results in table 1 that we could use whole genomic annotations (predicted genes as well as EST-confirmed annotation) to analyze intron and exon lengths, so long as the genome "status" was taken into consideration when drawing conclusions from the information. In this way we could compare intron and exon information from a number of different eukaryotic species.

We next examined information from a number of species that are known to use alternative splicing (human, *A. thaliana*, *D. melanogaster*, *C. neoformans*, and the honeybee *A. mellifera*). Alternatively spliced products are often detected using EST information (Hiller et al. 2005). However, EST databases can under represent the amount of alternative splicing because a product must be expressed sufficiently highly to be detected in the tissue or cell type being sampled and also at the correct time of development (Hiller et al. 2005). Thus, annotated genomes being at different levels of annotation may vary in their information about alternatively spliced products. Our aim was to see whether genes that are annotated as producing a single product or multiple products resulted in a similar range of intron and exon lengths. This can be important when comparing genomes as a gene that is alternatively spliced in one species may be constitutively spliced in another (Pan et al. 2005). In alternatively spliced data sets, we could expect to see a rise

Table 2
Comparison of Single, Multiple, and Both Product Data Sets

		Internal Exons				Internal Introns			
		Number	5%	50%	95%	Number	5%	50%	95%
Human (<i>Homo sapiens</i>)	S	87,729	52	123	275	77,405	95	1,389	17,144
	M	196,087	50	124	297	174,970	99	1,436	18,471
	B	283,816	51	124	289	252,375	98	1,422	18,062
	DiffSB		1	1	14		3	33	918
	DiffSM		2	1	22		4	47	1327
<i>Cryptococcus neoformans</i>	S	23,446	31	144	739	18,551	46	55	109
	M	8,997	28	137	801	7,323	46	55	121
	B	32,443	30	142	754	25,874	46	55	113
	DiffSB		1	2	15		0	0	4
	DiffSM		3	7	62		0	0	12
<i>Caenorhabditis elegans</i>	S	70,536	56	107	563	55,210	42	59	1,046
	M	21,633	66	155	614	18,367	44	85	1,472
	B	92,169	66	154	574	73,577	44	63	1,135
	DiffSB		10	47	11		2	4	89
	DiffSM		10	48	51		2	26	426
<i>Arabidopsis thaliana</i>	S	69,870	50	118	444	56,781	73	96	375
	M	56,322	48	111	455	47,039	75	97	365
	B	126,192	49	115	448	103,820	85	97	371
	DiffSB		1	3	4		12	1	4
	DiffSM		2	7	11		2	1	10
<i>Drosophila melanogaster</i>	S	14,615	82	225	1,203	9,810	54	66	1,801
	M	54,243	70	198	1,128	43,396	56	90	5,481
	B	68,858	72	204	1,143	53,206	55	78	4,830
	DiffSB		10	21	60		1	12	3,029
	DiffSM		12	88	75		2	24	3,680
Honeybee (<i>Apis mellifera</i>)	S	14,676	16	100	333	10,473	38	185	7,071
	M	106,224	19	154	455	87,184	36	117	3,450
	B	120,900	18	148	441	97,657	36	122	3,937
	DiffSB		2	48	108		2	63	3,134
	DiffSM		3	54	122		2	68	3,621

NOTE.—S, single product genes only; M, multiple product genes only; B, all genes; DiffSB, difference between the single product genes and both; and DiffSM, difference between single and multiple product genes.

in mean and 95% quartile in the length of internal introns due to inclusion of the skipped exon in the intron length.

Results (table 2) indicate that data sets containing single product genes, multiple product genes, and all genes produced reasonably similar length characteristics. In the human multiple product data set, the internal intron median and 95% quartile were increased compared to the single product data set and a data set containing both types of genes. However, the internal exon lengths stayed almost the same. These results suggest that constraint on the internal exon length is being maintained in both constitutive and alternatively spliced genes in humans. In contrast, *C. elegans*, *D. melanogaster*, and honeybee show large rises in intron lengths between their single and multiple product data sets. *Caenorhabditis elegans* and *D. melanogaster* contain both exon definition and intron definition, but the intron recognition system in honeybee is not yet known. A common form of alternative splicing uses exon skipping resulting in the exon being included in the intron (reviewed in Graveley 2001). It is possible that although both exon definition and intron definition are used in these species, alternatively spliced genes may predominantly use the exon definition mechanism for an exon that has a possibility to be skipped; hence, it is possible for introns that are not constrained by intron definition to become longer.

Arabidopsis thaliana was once thought to use intron definition because its introns were sufficiently short for this

to be possible; however, further investigation demonstrated that exon definition is also found in *A. thaliana* and other plants (Lorkovic et al. 2000; Simpson et al. 2004). The similarities between the single and multiple data sets are interesting in that there appears to be constraints on both the internal introns and exons for both constitutive and alternatively spliced genes.

Although the mechanism of intron recognition is not yet characterized in *C. neoformans*, we can see that generally internal intron length has a more restricted range than exon length which is suggestive of intron definition. Comparing the single and multiple product data sets from *C. neoformans*, we see a small rise in internal exon length but a smaller rise in intron length. Results suggest that for both constitutive and alternatively spliced genes, internal introns are being highly constrained in *C. neoformans*, but internal exons are also under some constraint. There is evidence for a number of forms of alternative splicing in *C. neoformans* including exon skipping, truncation, and extension at both 5' and 3' ends (Loftus et al. 2005), but we do not yet know if exon definition is present in *C. neoformans* and is involved in the exon skipping mechanism found in at least some of its alternatively spliced genes.

Although for some species there are differences between constitutively and alternatively spliced gene data sets, when we combine the data sets we find that the results are intermediate. This gives us an overview of internal introns

Table 3
Internal Exon and Intron Lengths from Whole Genome Data Sets

Species	Status	Splicing Types	Internal Exon Data				Internal Intron Data					
			Number	5%	50%	95%	Range	Number	5%	50%	95%	Range
Human	2	M, m, A	283,216	51	124	289	238	252,375	98	1,422	18,062	17,694
Mouse	2	M, m, A	228,593	51	124	288	237	201,385	94	1,244	14,042	13,948
Rat	2	M, m, A	169,149	54	126	287	233	149,620	93	1,208	12,584	12,491
Chicken	1	M, A	233,360	56	127	270	214	206,945	99	846	7,872	7,773
Zebrafish	1	M, A	197,442	57	130	300	243	171,411	80	892	8,248	8,168
<i>Takifugu rubripes</i>	1	M, A	273,209	57	151	283	226	243,624	71	142	2,206	2,135
<i>Tetraodon nigroviridis</i>	1	M, A	160,037	50	127	353	303	138,208	68	131	1,992	1,924
Sea squirt	1	M, T, A	195,267	29	129	233	204	174,261	53	318	1,221	1,168
<i>Drosophila melanogaster</i>	2	M, m, A	68,858	72	204	1,143	1,071	53,206	55	78	4,830	4,775
Mosquito	2	M, A	31,828	77	200	968	891	20,726	62	92	4,496	4,434
Honeybee	1	M, A	120,900	18	148	441	423	97,657	36	122	3,621	3,585
<i>Caenorhabditis elegans</i>	3	M, T, A	92,169	66	154	574	508	73,577	44	63	1,135	1,091
<i>Caenorhabditis briggsae</i>	1	M, T, A	62,114	72	166	632	560	48,615	43	53	1,339	1,296
<i>Schizosaccharomyces pombe</i>	3	M, A	2,446	33	138	891	858	1,202	38	52	164	126
<i>Saccharomyces cerevisiae</i> ^a	3	M	527	4	329	1,370	1,366	268	69	148	525	456
<i>Cryptococcus neoformans</i>	3	M, A	32,443	30	142	754	724	25,874	46	55	113	67
<i>Encephalitozoon cuniculi</i> ^a	0	M	29	3	156	547	544	15	24	33	47	23
<i>Arabidopsis thaliana</i>	3	M, m, A	126,192	49	115	448	399	103,820	85	97	371	286
<i>Oryza sativa</i>	1	M, A	84,892	48	120	571	523	65,650	73	145	1,175	1,102
<i>Plasmodium falciparum</i>	1	M, A	5,851	41	109	1,222	1,181	3,923	80	131	274	194
<i>Entamoeba histolytica</i>	0	M	693	84	258	1,309	1,225	137	36	55	240	204
<i>Dictyostelium discoideum</i>	2	M, A	8,208	36	246	1,391	1,355	3,701	53	99	301	248
<i>Cryptosporidium parvum</i>	0	M	10	82	165	702	620	7	47	64	88	41

NOTE.—Status: 0, genes predicted only, no EST data linked to annotation; 1, some experimental or EST data included in annotation; 2, genome-wide EST data in separate files but able to be easily linked to EST data using a transcript_id included in the genome annotation; and 3, genome-wide EST data included in annotation. Splicing types: M, major splicing; m, minor splicing; T, *trans*-splicing; and A, alternative splicing.

^a *Saccharomyces cerevisiae* uses total introns (as there are no middle introns) and middle exons; *E. cuniculi* uses total introns and total exons as there are no middle introns and only a single middle exon.

and internal exons within a species regardless of the number of products its genes produces. We suggest that in humans, the exon definition system may be limiting any differences in internal exon lengths between constitutive and alternatively spliced genes. However, there are no similar intron constraints allowing intron lengths to become larger as they can include differentially spliced exons. This is in direct contrast to the situation in *C. neoformans* which appears to have tight constraints on intron size but shows larger length differences in their internal exons. Results from species that contain both intron and exon definition may reflect constraints imposed on some exons (those that are affected by exon definition) while others may be not be as constrained. Further study is required to reveal more about the effect of alternative splicing on intron and exon lengths.

Information from a larger set of organisms (animals, plants, fungi, and unicellular eukaryotes) was then extracted. This meant including genomes with little or no annotation-linked EST information, little experimental evidence linkage, and little splicing information. Species such as *S. cerevisiae* and *E. cuniculi* were included because they are examples of extreme intron loss. In these species, single intron genes are more common than multiple intron genes and appear to result from preferential intron loss from the 3' end of the gene (Mourier and Jeffares 2003). This process results in a loss of most internal intron and internal exon information. For this reason, calculations for *S. cerevisiae* used total introns and internal exons (as there are no middle introns); and *E. cuniculi* used total introns and total exons (as there are no internal introns and only a single internal exon).

For this set of comparisons, we used whole genome data sets incorporating both single and multiple product information and recorded each genome status according to the level of EST information contained within each species annotation and the presence of alternative splicing (from general literature). The length range (95%–5%) for internal exons and introns was also calculated. The results of these comparisons are shown in table 3.

Vertebrates contain very high numbers of introns and exons, but the range of their internal exon lengths are extremely similar (e.g., only a 5-nt difference in ranges between human, mouse, and rat). In contrast, intron lengths have a very high range difference with most of the difference coming from the 95% quartile. It is interesting that the two pufferfish (*T. rubripes* and *T. nigroviridis*) have rather different internal exon ranges but reasonably similar internal intron ranges. These genomes are based on some experimental evidence, but there is still much annotation work required on these genomes, therefore we cannot rule out annotation “effects” as a cause of these differences. However, fundamental biological differences between these two species (*T. nigroviridis* lives in tropical fresh water while *T. rubripes* lives in the Sea of Japan and has a 15° lower body temperature) have been recently correlated with genomic differences such as the formation of GC-rich regions (Jabbari and Bernardi 2004). There is thus a possibility that the differences seen in exon lengths are real and reflect other as yet unknown genomic consequences of a major environmental change. *Takifugu rubripes* has a genome ~7.5 times smaller than that of human with compaction occurring

within both introns and intergenic regions (Miles et al. 2003). Internal exon length constraints suggest that exon definition in fish may be predominant; however, is likely that the precise motifs and mechanism involved will be different from that observed in mammals.

Results from the genome of the sea squirt, *Ciona intestinalis* (a nonvertebrate chordate) are interesting in that although they show generally shorter internal exons and introns length ranges than the other chordates. Internal exons have a much lower 5% quartile and slightly lower 95% quartile which accounts for their lower range, but their median quartile is comparable to the vertebrates. Internal introns also show lower 5% and 95% quartiles, but its median is higher than that of the pufferfish. The sea squirt genome was given a lower status due to our inability to relate the available EST data to the coordinates in the genome annotation. However, the sea squirt has a highly modified genome that has undergone massive deletions of genes and gene families that were present in the chordate ancestor (Hughes and Friedman 2005), thus it is likely that intron and exon characteristics may have changed.

In the insects, nematodes, and plants we see that the exon range is larger than that of the vertebrates whereas there is still a large amount of internal length variation. Because both types of intron recognition are thought to be present in these species, on a genome-wide basis, we would expect to see exon lengths being constrained where exon definition was being used but becoming larger in areas where it was not being used. The three insect species used in this study show interesting differences in their internal exon lengths; honeybee showing a much lower 5% and 95% range than the two dipteran species. The honeybee genome, like that of the sea squirt, is still not fully annotated, and it cannot be excluded that automated annotation is artificially creating shorter exons than are seen when a larger amount of EST and experimental information is available for the genome.

In contrast to what is seen in animals, the fungi *S. pombe*, with no evidence of exon definition (Romfo et al. 2000) and *C. neoformans* (intron recognition mechanism not yet characterized) have narrow intron length ranges with much larger exon length ranges. *Schizosaccharomyces pombe* has lost many of its introns (Sverdlov et al. 2004), but *C. neoformans* not only contains a large number of introns but, as mentioned above, also has alternative splicing (Loftus et al. 2005). These results suggest that the introns in these species are under some sort of constraint, but we cannot determine if this is because of intron definition or other constraints which are discussed later.

The other yeast *S. cerevisiae* gives different results, but there are only a small number of multiple exons in this species. *Saccharomyces cerevisiae* does not appear to use exon definition but instead uses a type of intron definition where distinct RNA-RNA complementary sequences within the intron form RNA secondary structure (Howe and Ares 1997); or by interaction with the transcription elongation system (Howe, Kane, and Ares 2003). However, for one intron in *S. cerevisiae*, mutation of the 5' splice site causes intron retention rather than exon skipping, indicating that some type of intron definition may be present (Howe and Ares 1997).

The microsporidian *E. cuniculi* has an extremely reduced genome and has retained only a few introns (23 in total), mostly in 5' positions (Mourier and Jeffares 2003) and as expected shows an extremely narrow intron length range. It is expected that with such small lengths the introns could be defined directly without the need for an exon definition system. In addition to splicing, other pre-mRNA processing events include 5' end capping with 7-methyl guanosine, and once formed, the cap structure can enhance recognition of the 5'-most intron (Le Hir, Nott, and Moore 2003). Because most introns in highly reduced genomes such as *E. cuniculi* show a 5'-positional bias (Mourier and Jeffares 2003), it is possible that enhancement by the capping process is being used for intron recognition, bypassing the requirement for both intron or exon definitions.

The four protist genomes that contained enough introns for analysis (*P. falciparum*, *E. histolytica*, *D. discoideum*, and *C. parvum*) show a narrow intron range and a wider exon length range. More than 50% of *P. falciparum* genes contain introns (Gardner et al. 2002), but less than 10% of genes contain introns in the other apicomplexan genome *C. parvum* (Templeton et al. 2004). Although *C. parvum* is intron poor, its genome does not appear to be otherwise reduced (Jeffares, Mourier, and Penny 2006), suggesting that introns may have been selectively lost. *Dictyostelium discoideum* contains many introns and also has alternative splicing (Escalante, Moreno, and Sastre 2003). Internal intron and exon lengths from these protists suggest that internal exons are not highly constrained, whereas some constraint is being placed on internal introns.

Discussion

Our primary conclusion is that annotated genomes can give good information about the processing mechanisms for mRNA splicing. This conclusion is based on several tests described in the *Results*. Whole genome annotations are not mere collections of computer predictions; they form an organism-wide vehicle linking genomic and biological information. EST databases provide valuable information for genomic analysis, but the coverage of the transcriptome can be limited especially for genes that are expressed at low levels or under limited conditions (Ohler, Shomron, and Burge 2005). Precise and species-specific computer prediction followed by experimental confirmation (e.g., Zhu and Brendel 2003; Ohler, Shomron, and Burge 2005) can detect transcripts that may be otherwise undetected. For this reason, genome annotations that combine experimental information (EST and otherwise) with computer predictions are valid tools in complete genome analysis. However, when analyzing the results, one should always bear in mind the status of the genome (its completeness, state of assembly, linkage to EST, and other experimental data).

From our investigation of data sets from constitutive and alternatively spliced genes, we can query the extent that the exon definition mechanism contributes to the exon skipping form of alternative splicing. Exon skipping is the main mechanism by which genes are alternatively spliced in mammals (Sorek et al. 2004). However, alternative splicing is also found in *S. pombe* (Tang, Kaufer, and Lin 2002) which does not use exon definition, indicating that the "exon

skipping” seen in alternative splicing may be the result of a different process from that seen with splice site mutation. The characterization of the intron recognition mechanisms in such species as *C. neoformans*, *P. falciparum*, and *D. discoideum* that use alternative splicing but by having short introns may not require exon definition, would aid in determining any link between the alternative splicing and exon definition.

Conserved minor alternatively spliced exons in mammals show a higher level of RNA sequence conservation than major alternatively spliced exons (Xing and Lee 2005), suggesting that minor exons require more regulatory signals and that their splicing may be highly regulated. Our study did not distinguish minor-spliced introns and their associated exons from the large majority of major introns and exons. Further study is required to determine if length constraints are acting in the same way between those introns and exons affected by major or minor splicing.

Introns incur an extra cost both in terms of energy and time for replication, transcription, and processing. Genes expressed at higher levels tend to have shorter introns in both humans and *C. elegans*, possibly to reduce the energetic cost of transcription and processing (Castillo-Davis et al. 2002; Llopert et al. 2002). Therefore, it could be expected that the loss of an intron would be beneficial. However, due to the complex interaction between splicing and other RNA processing interactions, an intron-containing gene (at least in humans) is expressed more than are intronless versions of the same gene (Wiegand, Lu, and Cullen 2003). Reduction in intron size may be the result of selection to reduce the transcriptional costs (Comeron 2004; Wagner 2005). Short introns are favored in human genes requiring a minimal response time (so-called “nimble” genes), and some short introns may be involved in antisense regulation (Chen et al. 2005). Intron density correlates with generation time; species that reproduce rapidly have fewer introns per gene than species that have longer life cycles (Jeffares, Mourier, and Penny 2006). Thus, it is likely that species with short generation times could constrain shorter introns if they are unable to “lose” their introns altogether.

Highly reduced genomes usually retain few introns and multiple introns within their genes are rare. An example is *S. cerevisiae* which can use RNA secondary structure within the intron (intron self-complementation) for intron boundary recognition (Howe and Ares 1997). The splicing mechanism in *S. cerevisiae* is also highly coordinated with the transcription elongation system which enhances exon fidelity (Howe, Kane, and Ares 2003). *Saccharomyces cerevisiae* is incapable of removing introns correctly from genes from a number of other eukaryotes, including *D. melanogaster* (Langford et al. 1983), an indication of differences in the yeast splicing system. Another yeast in this study, *S. pombe*, however, is able to splice plant introns although some aberrant splicing occurs (Sarmah et al. 2002), indicating that there is significant similarity between the intron recognition mechanisms found in *S. pombe* and plants.

Is exon definition a recent system that evolved when intron length expanded for some reason (e.g., inclusion of functional ncRNA), or is it an “ancestral” system that has been lost in some lineages; this loss being connected in

some way to widespread genomic intron loss. To answer this question we need to consider intron loss throughout eukaryotic lineages. Some introns appear old, whereas others with narrow phylogenetic distributions appear to be gained more recently (Roy and Gilbert 2005a). Rates of intron loss and gain differ significantly between eukaryotic genomes (Jeffares, Mourier, and Penny 2006). Intron loss and gain is very slow in vertebrates, but intron loss has been more pronounced in *D. melanogaster* and *C. elegans*, whereas the plant *A. thaliana* appears to have retained most of its ancestral introns (Rogozin et al. 2003; Roy, Fedorov, and Gilbert 2003; Roy and Gilbert 2005a). The last common ancestor of plants, animals, and fungi is inferred to have had a large number of introns, suggesting that intron-rich gene structures are ancestral features and that genomes containing small numbers of introns have been subjected to mass intron loss (Roy and Gilbert 2005a).

Intron loss (especially mRNA-mediated losses) can produce the fusion of several exons to produce what is known as extraordinarily large exons, which can be found throughout multicellular and unicellular eukaryotes (Niu, Hou, and Li 2005). Such exons (if internal) subjected to pressure from an exon definition type of splicing recognition system have a number of options. They may activate cryptic splice sites to create a new intron and reduce exon size but in the process lose coding information, or they could use other means to attract the spliceosome to an adjacent intron. Intron definition, attraction by the 5' cap, 3' polyadenylation site attraction, and the RNA secondary structure mechanism found in *S. cerevisiae* are four alternatives. Because detailed examination of intron recognition mechanisms has not yet progressed further than model eukaryotes, it is possible that in the future additional alternatives to exon recognition may be revealed. We therefore raise the possibility that the last common ancestor of plants and animals contained some type of exon definition and that this mechanism was replaced by intron recognition in some genes and lineages, possibly as a result of intron loss and/or intron shortening.

However, we cannot as yet understand the types of intron recognition systems that may have been present in the last common ancestor of all eukaryotes. Because it is now likely that the eukaryotic ancestor contained both many introns (Rogozin et al. 2005; Roy and Gilbert 2005b) and a more complex genomic organization than previously thought, we conclude that such an organism must have contained some mechanism of intron recognition. From the initial analysis carried out in this study, short introns and a slightly relaxed constraint on exon length are predominantly found throughout *P. falciparum*, *E. histolytica*, and *C. parvum* consistent with intron definition being present. However, these parasitic eukaryotes could have modified their genomes in some way and evolved their intron recognition systems secondarily. The improved linkage of experimental and biological information to sequenced genomes, as well as the sequencing of free-living unicellular eukaryotes, will help understand this question. In the meantime, it is now clear that information about splicing mechanisms can be gained from well-annotated genomes, opening the way to understand more about molecular evolution in eukaryotes.

Acknowledgments

Many thanks to Michael Woodhams for intron_finder.pl and Klaus Schliep for statistical analysis. Thanks to Dan Jeffares for advice and genome analysis during the early stages of this study. Thanks also to the Helix Parallel Processing Cluster for access to their facilities. Finally, we would like to give many thanks to the organizers of the Society for Molecular Biology and Evolution Tri-National Young Investigators' Workshop for the opportunity to present this research. This work was supported by the New Zealand Marsden Fund and the Allan Wilson Centre for Molecular Ecology and Evolution.

Literature Cited

- Berget, S. M. 1995. Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270**:2411–2414.
- Black, D. L. 1995. Finding splice sites within a wilderness of RNA. *RNA* **1**:763–771.
- Bruce, S. R., and M. L. Peterson. 2001. Multiple features contribute to efficient constitutive splicing of an unusually large exon. *Nucleic Acids Res.* **29**:2292–2302.
- Burge, C. B., R. A. Padgett, and P. A. Sharp. 1998. Evolutionary fates and origins of U12-type introns. *Mol. Cell* **2**:773–785.
- Burnette, J. M., E. Miyamoto-Sato, M. A. Schaub, J. Conklin, and A. J. Lopez. 2005. Subdivision of large introns in *Drosophila* by recursive splicing at non-exonic elements. *Genetics* **170**:661–674.
- Castillo-Davis, C. I., S. L. Mekhedov, D. L. Hartl, E. V. Koonin, and F. A. Kondrashov. 2002. Selection for short introns in highly expressed genes. *Nat. Genet.* **31**:415–418.
- Cazalla, D., J. Zhu, L. Manche, E. Huber, A. R. Krainer, and J. F. Caceres. 2002. Nuclear export and retention signals in the RS domain of SR proteins. *Mol. Cell. Biol.* **22**:6871–6882.
- Chen, J., M. Sun, L. D. Hurst, G. G. Carmichael, and J. D. Rowley. 2005. Human antisense genes have unusually short introns: evidence for selection for rapid transcription. *Trends Genet.* **21**:203–207.
- Clark, F., and T. A. Thanaraj. 2002. Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.* **11**:451–464.
- Collins, L., and D. Penny. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.* **22**:1053–1066.
- Comeron, J. M. 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics* **167**:1293–1304.
- Cooke, C., H. Hans, and J. C. Alwine. 1999. Utilization of splicing elements and polyadenylation signal elements in the coupling of polyadenylation and last-intron removal. *Mol. Cell. Biol.* **19**:4971–4979.
- Cunningham, T. P., J. P. Hagan, and P. J. Grabowski. 1995. Reconstitution of exon-bridging activity with purified U2AF and U1 snRNP components. *Nucleic Acids Symp. Ser.* **33**:218–219.
- Dominski, Z., and R. Kole. 1991. Selection of splice sites in pre-mRNAs with short internal exons. *Mol. Cell. Biol.* **11**:6075–6083.
- Escalante, R., N. Moreno, and L. Sastre. 2003. *Dictyostelium discoideum* developmentally regulated genes whose expression is dependent on MADS box transcription factor SrfA. *Eukaryot. Cell* **2**:1327–1335.
- Fedorov, A., A. F. Merican, and W. Gilbert. 2002. Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc. Natl. Acad. Sci. USA* **99**:16128–16133.
- Furuyama, S., and J. P. Bruzik. 2002. Multiple roles for SR proteins in trans splicing. *Mol. Cell. Biol.* **22**:5337–5346.
- Gannoun-Zaki, L., A. Jost, J. Mu, K. W. Deitsch, and T. E. Wellems. 2005. A silenced *Plasmodium falciparum* var promoter can be activated in vivo through spontaneous deletion of a silencing element in the intron. *Eukaryot. Cell* **4**:490–492.
- Gardner, M. J., N. Hall, E. Fung et al. (36 co-authors). 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**:498–511.
- Gornemann, J., K. M. Kotovic, K. Hujer, and K. M. Neugebauer. 2005. Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex. *Mol. Cell* **19**:53–63.
- Graveley, B. R. 2001. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* **17**:100–107.
- Graveley, B. R., K. J. Hertel, and T. Maniatis. 1999. SR proteins are 'locators' of the RNA splicing machinery. *Curr. Biol.* **9**:R6–R7.
- Hiller, M., K. Huse, M. Platzer, and R. Backofen. 2005. Non-EST based prediction of exon skipping and intron retention events using Pfam information. *Nucleic Acids Res.* **33**:5611–5621.
- Hoffman, B. E., and P. J. Grabowski. 1992. U1 snRNP targets an essential splicing factor, U2AF65, to the 3' splice site by a network of interactions spanning the exon. *Genes Dev.* **6**:2554–2568.
- Howe, K. J., and M. Ares Jr. 1997. Intron self-complementarity enforces exon inclusion in a yeast pre-mRNA. *Proc. Natl. Acad. Sci. USA* **94**:12467–12472.
- Howe, K. J., C. M. Kane, and M. Ares Jr. 2003. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in *Saccharomyces cerevisiae*. *RNA* **9**:993–1006.
- Hughes, A. L., and R. Friedman. 2005. Loss of ancestral genes in the genomic evolution of *Ciona intestinalis*. *Evol. Dev.* **7**:196–200.
- Ibrahim el, C., T. D. Schaal, K. J. Hertel, R. Reed, and T. Maniatis. 2005. Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. *Proc. Natl. Acad. Sci. USA* **102**:5002–5007.
- Jabbari, K., and G. Bernardi. 2004. Body temperature and evolutionary genomics of vertebrates: a lesson from the genomes of *Takifugu rubripes* and *Tetraodon nigroviridis*. *Gene* **333**:179–181.
- Jeffares, D. C., T. Mourier, and D. Penny. 2006. The biology of intron gain and loss. *Trends Genet.* **22**:16–22.
- Kennedy, C. F., A. Kramer, and S. M. Berget. 1998. A role for SRp54 during intron bridging of small introns with pyrimidine tracts upstream of the branch point. *Mol. Cell. Biol.* **18**:5425–5434.
- Kupfer, D. M., S. D. Drabenstot, K. L. Buchanan, H. Lai, H. Zhu, D. W. Dyer, B. A. Roe, and J. W. Murphy. 2004. Introns and splicing elements of five diverse fungi. *Eukaryot. Cell* **3**:1088–1100.
- Lam, B. J., and K. J. Hertel. 2002. A general role for splicing enhancers in exon definition. *RNA* **8**:1233–1241.
- Langford, C., W. Nellen, J. Niessing, and D. Gallwitz. 1983. Yeast is unable to excise foreign intervening sequences from hybrid gene transcripts. *Proc. Natl. Acad. Sci. USA* **80**:1496–1500.
- Le Hir, H., A. Nott, and M. J. Moore. 2003. How introns influence and enhance eukaryotic gene expression. *Trends Biochem. Sci.* **28**:215–220.
- Llopert, A., J. M. Comeron, F. G. Brunet, D. Lachaise, and M. Long. 2002. Intron presence-absence polymorphism in *Drosophila* driven by positive Darwinian selection. *Proc. Natl. Acad. Sci. USA* **99**:8121–8126.
- Loftus, B. J., E. Fung, P. Roncaglia et al. (50 co-authors). 2005. The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. *Science* **307**:1321–1324.

- Lorkovic, Z. J., R. Lehner, C. Forstner, and A. Barta. 2005. Evolutionary conservation of minor U12-type spliceosome between plants and humans. *RNA* **11**:1095–1107.
- Lorkovic, Z. J., D. A. Wieczorek Kirk, M. H. Lambermon, and W. Filipowicz. 2000. Pre-mRNA splicing in higher plants. *Trends Plant Sci.* **5**:160–167.
- Maniatis, T., and B. Tasic. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418**:236–243.
- Marais, G., P. Nouvellet, P. D. Keightley, and B. Charlesworth. 2005. Intron size and exon evolution in *Drosophila*. *Genetics* **170**:481–485.
- Mayer, M. G., and L. M. Floeter-Winter. 2005. Pre-mRNA trans-splicing: from kinetoplasts to mammals, an easy language for life diversity. *Mem. Inst. Oswaldo Cruz* **100**:501–513.
- McCullough, A. J., C. E. Baynton, and M. A. Schuler. 1996. Interactions across exons can influence splice site recognition in plant nuclei. *Plant Cell* **8**:2295–2307.
- Mignone, F., C. Gissi, S. Liuni, and G. Pesole. 2002. Untranslated regions of mRNAs. *Genome Biol.* **3**: 0004.1–0004.10.
- Miles, C. G., L. Rankin, S. I. Smith, M. Niksic, G. Elgar, and N. D. Hastie. 2003. Faithful expression of a tagged Fugu WT1 protein from a genomic transgene in zebrafish: efficient splicing of pufferfish genes in zebrafish but not mice. *Nucleic Acids Res.* **31**:2795–2802.
- Mourier, T., and D. C. Jeffares. 2003. Eukaryotic intron loss. *Science* **300**:1393.
- Niu, D. K., W. R. Hou, and S. W. Li. 2005. mRNA-mediated intron losses: evidence from extraordinarily large exons. *Mol. Biol. Evol.* **22**:1475–1481.
- Ohler, U., N. Shomron, and C. B. Burge. 2005. Recognition of unknown conserved alternatively spliced exons. *PLoS Comput. Biol.* **1**:113–122.
- Pan, Q., M. A. Bakowski, Q. Morris, W. Zhang, B. J. Frey, T. R. Hughes, and B. J. Blencowe. 2005. Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends Genet.* **21**:73–77.
- Patel, A. A., and J. A. Steitz. 2003. Splicing double: insights from the second spliceosome. *Nat. Rev. Mol. Cell. Biol.* **4**: 960–970.
- Pozzoli, U., L. Riva, G. Menozzi, R. Cagliani, G. P. Comi, N. Bresolin, R. Giorda, and M. Sironi. 2004. Over-representation of exonic splicing enhancers in human intronless genes suggests multiple functions in mRNA processing. *Biochem. Biophys. Res. Commun.* **322**:470–476.
- Robberson, B. L., G. J. Cote, and S. M. Berget. 1990. Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol. Cell. Biol.* **10**:84–94.
- Rogozin, I. B., A. V. Sverdlov, V. N. Babenko, and E. V. Koonin. 2005. Analysis of evolution of exon-intron structure of eukaryotic genes. *Brief. Bioinform.* **6**:118–134.
- Rogozin, I. B., Y. I. Wolf, A. V. Sorokin, B. G. Mirkin, and E. V. Koonin. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.* **13**:1512–1517.
- Romfo, C. M., C. J. Alvarez, W. J. van Heeckeren, C. J. Webb, and J. A. Wise. 2000. Evidence for splice site pairing via intron definition in *Schizosaccharomyces pombe*. *Mol. Cell. Biol.* **20**:7955–7970.
- Roy, S. W., A. Fedorov, and W. Gilbert. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl. Acad. Sci. USA* **100**: 7158–7162.
- Roy, S. W., and W. Gilbert. 2005a. Complex early genes. *Proc. Natl. Acad. Sci. USA* **102**:1986–1991.
- . 2005b. Rates of intron loss and gain: implications for early eukaryotic evolution. *Proc. Natl. Acad. Sci. USA* **102**: 5773–5778.
- Sakharkar, M. K., V. T. Chow, and P. Kanguane. 2004. Distributions of exons and introns in the human genome. *In Silico Biol.* **4**:387–393.
- Sarmah, B., N. Chakraborty, S. Chakraborty, and A. Datta. 2002. Plant pre-mRNA splicing in fission yeast, *Schizosaccharomyces pombe*. *Biochem. Biophys. Res. Commun.* **293**:1209–1216.
- Simpson, C. G., G. P. Clark, J. M. Lyon, J. Watters, C. McQuade, and J. W. S. Brown. 1999. Interactions between introns via exon definition in plant pre-mRNA splicing. *Plant J.* **18**:293–302.
- Simpson, C. G., S. N. Jennings, G. P. Clark, G. Thow, and J. W. Brown. 2004. Dual functionality of a plant U-rich intronic sequence element. *Plant J.* **37**:82–91.
- Sorek, R., R. Shemesh, Y. Cohen, O. Basechess, G. Ast, and R. Shamir. 2004. A non-EST-based method for exon-skipping prediction. *Genome Res.* **14**:1617–1623.
- Sterner, D. A., T. Carlo, and S. M. Berget. 1996. Architectural limits on split genes. *Proc. Natl. Acad. Sci. USA* **93**:15081–15085.
- Sverdlov, A. V., V. N. Babenko, I. B. Rogozin, and E. V. Koonin. 2004. Preferential loss and gain of introns in 3' portions of genes suggests a reverse-transcription mechanism of intron insertion. *Gene* **338**:85–91.
- Tang, Z., N. F. Kaufer, and R. J. Lin. 2002. Interactions between two fission yeast serine/arginine-rich proteins and their modulation by phosphorylation. *Biochem. J.* **368**:527–534.
- Templeton, T. J., L. M. Iyer, V. Anantharaman, S. Enomoto, J. E. Abrahamte, G. M. Subramanian, S. L. Hoffman, M. S. Abrahamson, and L. Aravind. 2004. Comparative analysis of apicomplexa and genomic diversity in eukaryotes. *Genome Res.* **14**:1686–1695.
- Thanaraj, T. A., S. Stamm, F. Clark, J. J. Riethoven, V. Le Texier, and J. Muilu. 2004. ASD: the alternative splicing database. *Nucleic Acids Res.* **32**:D64–D69.
- Wagner, A. 2005. Energy constraints on the evolution of gene expression. *Mol. Biol. Evol.* **22**:1365–1374.
- Wagner, E. J., and M. A. Garcia-Blanco. 2001. Polypyrimidine tract binding protein antagonizes exon definition. *Mol. Cell. Biol.* **21**:3281–3288.
- Wiegand, H. L., S. Lu, and B. R. Cullen. 2003. Exon junction complexes mediate the enhancing effect of splicing on mRNA expression. *Proc. Natl. Acad. Sci. USA* **100**:11327–11332.
- Williams, J. G., A. A. Noegel, and L. Eichinger. 2005. Manifestations of multicellularity: Dictyostelium reports in. *Trends Genet.* **21**:392–398.
- Wu, J. Y., and T. Maniatis. 1993. Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell* **75**:1061–1070.
- Xing, Y., and C. Lee. 2005. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc. Natl. Acad. Sci. USA* **102**: 13526–13531.
- Zhang, X. H., and L. A. Chasin. 2004. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.* **18**:1241–1250.
- Zhu, W., and V. Brendel. 2003. Identification, characterization and molecular phylogeny of U12-dependent introns in the *Arabidopsis thaliana* genome. *Nucleic Acids Res.* **31**:4561–4572.

Billie Swalla, Associate Editor

Accepted December 14, 2005