Eukaryotic Intron Loss

Tobias Mourier* and Daniel C. Jeffares

Recently, attention has been drawn to eukaryotic genomes with very few introns (*1*, *2*) and to the biased position of introns within genes (*3*). We show here that intron-poor eukaryotes for which genome data is available have a 5′ bias in the position of their introns within genes. This decrease in introns toward the 5′ end of the gene is more pronounced with increasing intron paucity. We argue that this asymmetry is more consistent with models of intron loss from intron-poor organisms.

We analyzed intron positions from 18 relatively well-annotated eukaryotic genomes, which differ by three orders of magnitude in the average number of introns per gene. Because all eukaryotes have evolved from a common ancestor (*4*), there must have been considerable gain of introns in some lineages, intron loss in others, or a combination of both processes.

The genomes display marked differences in patterns; the introns in multicellular genomes are evenly distributed throughout genes, whereas those of unicellular organisms are biased toward the 5′ ends (Fig. 1A). The observed bias in intron position is correlated with intron paucity. Only *Plasmodium* genomes show a different trend. We have no clear explanation for this deviation, but it may be an artifact of gene prediction, which is particularly challenging in *Plasmodium* and would be expected to detect 3′ introns more reliably where expressed sequence tag (EST) coverage is most complete (*5*).

The most common mechanism cited to explain intron gain is the insertion of mobile genetic elements (*6*) or "reverse splicing" (*7*). Either a preference for intron accumulation in the 5′ end of genes or selection for 5′ position is required for intron gain to explain the pattern we observe. One possibility for selection of 5′ introns is that they could be preferentially maintained because they contain the majority of intronic enhancers (*8*). However, if introns were preferentially accumulated in the 5′ end of genes, this tendency would depend on the number of introns already present in the genome because there is no 5′ bias in intron-rich genomes.

The best-characterized model for intron loss is homologous recombination between the genomic copy of a gene and an intron-less cDNA produced by reverse transcription of the corresponding mRNA (*9*) (Fig. 1B). Because retrotransposons can reverse transcribe mRNAs other than their own transcripts (*10*), cDNA templates are expected to be present in eukaryotic cells. Further, Derr (*11*) showed that cDNAs could recombine with the corresponding gene, resulting in intron loss.

The pattern we observe could be derived from intron loss by this mechanism if reverse transcriptases begin from the 3′ end of RNA molecules and dissociate in a length-dependant manner, as is the case for pseudogenes derived from cDNAs (*12*). Most cDNAs would therefore be truncated at the 5′ end; thus, templates that could replace introns at the 5′ ends of genes would be expected to be less abundant. This mechanism would therefore be expected to predominantly remove 3′ introns.

Considering the known mechanisms by which introns are lost or gained, we believe the simplest explanation for the 5′ bias is that introns have been lost from these intron-poor eukaryotes by homologous recombination of cDNAs. Clearly, this does not prohibit intron gain in some cases, but the high asymmetry of intron position in intron-poor organisms appears to be secondarily derived.

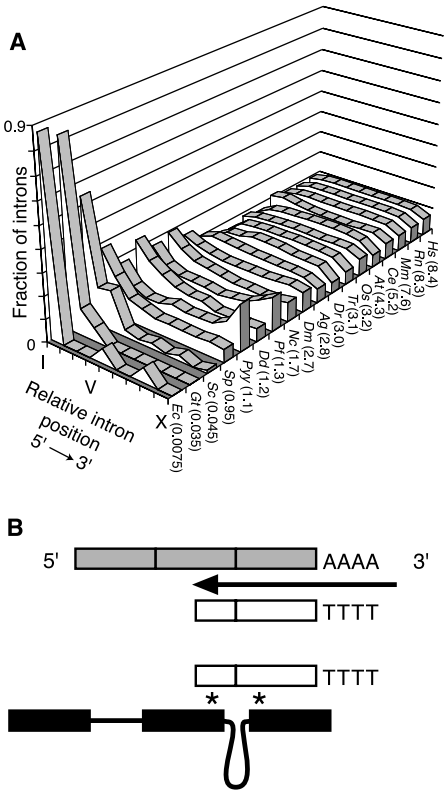

Fig. 1. The intron position bias of 18 eukaryotes and a model to explain the pattern. (**A**) The relative position of each intron within its host gene was calculated as the sequence length of the open reading frame (ORF) upstream of the intron divided by the full length of the ORF. For each organism, relative intron position data were pooled into 10 categories, where each category is one-tenth of the ORF length (denoted I to X progressively from 5′ to 3′). The fraction of introns each genome contains in each category was plotted on the *y* axis. Finer scale of up to 20 categories did not produce an essentially different pattern (data not shown). Species are sorted according to their intron-to-gene ratios (in parentheses). A significant correlation was found between average relative intron position and intron-to-gene ratio (Spearman rank, $r_s = 0.59$, $P = 9.4 \times 10^{-3}$; excluding *Plasmodium* data: $r_s = 0.85$, $P = 3 \times 10^{-5}$). Species abbreviations (and number of introns studied): *Hs*, *Homo sapiens* (139,418); *Rn*, *Rattus norvegicus* (33,083); *Mm*, *Mus musculus* (65,418); *Ce*, *Caenorhabditis elegans* (100,569); *At*, *Arabidopsis thaliana* (107,552); *Os*, *Oryza sativa japonica* (234,084); *Tr*, *Takifugu rubripes* (102,709); *Dr*, *Danio rerio* (3,895); *Ag*, *Anopheles gambiae* (36,503); *Dm*, *Drosophila melanogaster* (37,755); *Nc*, *Neurospora crassa* (17,215); *Pf*, *Plasmodium falciparum* (7,040); *Dd*, *Dictyostelium discoideum* (1,964); *Pyy*, *Plasmodium yoelii yoelii* (6,676); *Sp*, *Schizosaccharomyces pombe* (4,753); *Sc*, *Saccharomyces cerevisiae* (286); *Gt*, *Guillardia theta* nucleomorph (17); *Ec*, *Encephalitozoon cuniculi* (15). Sources for annotated genome data: *H. sapiens*, *R. norvegicus*, and *M. musculus*, The Genome Browser, assembly hg12 (*13*); *S. pombe* and *C. elegans*, The Sanger Institute (www.sanger.ac.uk); *N. crassa*, Whitehead Institute/MIT Center for Genome Research (*14*); *O. sativa japonica*, The Institute of Genomic Research (www.tigr.org); all other annotations from GenBank at NCBI (www.ncbi.nlm.nih.gov). The *D. discoideum* data set was an incomplete chromosome 2 annotation from GenBank (*15*). (**B**) The cDNA recombination model for intron loss. Gray boxes, exons of an mRNA; open boxes, exons of a cDNA produced by reverse transcription (arrow); black boxes, the corresponding genomic DNA showing the site of recombination (asterisks).

References

1. J. E. J. Nixon *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 3701 (2002).
2. A. G. B. Simpson, E. K. MacQuarrie, A. J. Roger, *Nature* **419**, 270 (2002).
3. A. Sakurai *et al.*, *Gene* **300**, 89 (2002).
4. J. R. Brown, W. F. Doolittle, *Microbiol. Mol. Biol. Rev.* **61**, 456 (1997).
5. M. J. Gardner *et al.*, *Nature* **419**, 531 (2002).
6. M. J. Giroux *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 12150 (1994).
7. L. Bonen, J. Vogel, *Trends Genet.* **17**, 322 (2001).
8. S. Schandorff, thesis, University of Copenhagen (2000).
9. G. R. Fink, *Cell* **49**, 5 (1987).
10. C. Esnault, J. Maestre, T. Heidmann, *Nature Genet.* **24**, 363 (2000).
11. L. K. Derr, *Genetics* **148**, 937 (1998).
12. A. M. Weiner, P. L. Deininger, A. Efstratiadis, *Annu. Rev. Biochem.* **55**, 631 (1986).
13. W. J. Kent *et al.*, *Genome Res.* **12**, 996 (2002).
14. Neurospora Sequencing Project, Release 3. Whitehead Institute/MIT Center for Genome Research (www.genome.wi.mit.edu).
15. G. Glöckner, personal communication.

Department of Evolutionary Biology, Zoological Institute, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen Ø, Denmark.

*To whom correspondence should be addressed. E-mail: tmourier@zi.ku.dk