Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain

Scott W. Roy*†, Alexei Fedorov†‡, and Walter Gilbert*§

*Biological Laboratories, 16 Divinity Avenue, Cambridge, MA 02138; and ‡Department of Medicine, Medical College of Ohio, 3120 Glendale Avenue, Toledo, OH 43614-5809

Contributed by Walter Gilbert, April 17, 2003

We compared intron–exon structures in 1,560 human–mouse orthologs and 360 mouse–rat orthologs. The origin of differences in intron positions between species was inferred by comparison with an outgroup, *Fugu* for human–mouse and human for mouse–rat. Among 10,020 intron positions in the human–mouse comparison, we found unequivocal evidence for five independent intron losses in the mouse lineage but no evidence for intron loss in humans or for intron gain in either lineage. Among 1,459 positions in rat–mouse comparisons, we found evidence for one loss in rat but neither loss in mouse nor gain in either lineage. In each case, the intron losses were exact, without change in the surrounding coding sequence, and involved introns that are extremely short, with an average of 200 bp, an order of magnitude shorter than the mammalian average. These results favor a model whereby introns are lost through gene conversion with intronless copies of the gene. In addition, the finding of widespread conservation of intron–exon structure, even over large evolutionary distances, suggests that comparative methods employing information about gene structures should be very successful in correctly predicting exon boundaries in genomic sequences.

When it was discovered 25 years ago that eukaryotes, unlike prokaryotes, had split gene structures, it came as quite a shock. Where did these so-called introns come from? What uses did they have? How did they propagate? How labile were their positions and sequences? These are questions that have proved very difficult to solve despite occupying the minds and laboratories of a generation of biologists.

As the exon–intron structure of genes reaches its silver anniversary, much about this structure is still a mystery and the subject of intense research. In this postgenomic world, introns have proven themselves a singular pest in our attempts to predict gene structures from raw genome sequence. They have almost no consensus sequence over their lengths; they can be absurdly long; and they are the substrates of a bewildering array of different alternative splicing patterns. One promising avenue of improvement of the algorithms would further exploit comparisons with other genomic sequences. If intron–exon structures are highly conserved between two species, the viability of an exon prediction should be able to be evaluated by comparison with the orthologous gene copy in another organism.

However, before such comparisons can be used, it is important to know more about the degree of conservation of gene structure between related species. Two ill understood processes that may change the intron–exon structure of genes are intron loss, in which the intervening noncoding sequence between two exons is jettisoned, and intron gain, in which an intron appears *de novo*. Apparent instances of both have been described. The first such event was characterized by Perler *et al.* (1) in 1980. Rats have two insulin genes, one with a two-exon-one-intron structure and the other with a three-exon-two-intron structure (in which one intron matches the single intron of the other gene copy). The genes are paralogs as a result of a recent duplication. To determine which structure was ancestral, the authors screened a genomic chicken library to find insulin genes in chicken. They found a sole copy, with a three-exon-two-intron structure and thus inferred that the other copy had lost one of its introns.

Compelling evidence for intron gain was slower in coming. In 1995, Logsdon *et al.* (2) sequenced the TPI genes for a host of metazoans and found that the gene took on many different structures in different species. In some cases, an intron position in one species was not shared with any other species, such that its phylogenetic distribution could be explained by either a single insertion or by up to 15 losses. Thus it was generally accepted that intron gain also occurs.

Since that time, many instances of intron gain and loss have been described (refs. 3–5; see refs. 6 and 7 for thorough reviews). However, such studies have been for the most part case studies, uncovering one or two instances in a single gene. Thus, it is hard to infer from the literature the relative importance of the two processes and thus to develop a qualitative sense of the shaping processes of intron evolution and of the general conservation of intron–exon structures between orthologs.

To begin to fill this void, we undertook two global studies of the differences in intron–exon structure, one between mouse and rat and the other between mouse and human. Comparing discordant intron structures to those of an outgroup (human for the mouse–rat comparison; *Fugu* for the human–mouse comparison), we were able to infer whether a given difference was caused by a gain in the intron-containing lineage or by a loss in the other.

Comparing 10,020 introns in human–mouse orthologs and 1,459 in mouse–rat, we found convincing evidence for five intron losses in the mouse lineage since its divergence from humans and one intron loss in the rat lineage since its divergence from mouse. In each of the characterized losses, the intron was exacted precisely without change to the flanking coding region. In each case, the corresponding intron in the intron-containing gene copy is very short, suggesting a bias for loss of short introns. We found no instances that resembled intron gain, suggesting that the mechanisms of intron gain are nonfunctional in mammals. The fraction of introns that have been lost since each evolutionary divergence is thus tiny, on the order of 0.06%.

Methods

Exon–Intron Databases. All 4,310 known mouse and 1,800 known rat genes with characterized intron–exon structures were obtained from the latest release (132) of GenBank. All human genes, both confirmed and predicted, were obtained from the human genome annotation available on the National Center for Biotechnology Information web site. We used the EID programs of Saxonov *et al.* (9) to generate databases of the intron–exon structures and sequences of all genes for each organism.

Gene Pairs. We did reciprocal BLASTP searches between all rat and all mouse genes and between all mouse and all human genes. For the rat–mouse comparison, this yielded 360 unique gene pairs

†S.W.R. and A.F. contributed equally to this work.

§To whom correspondence should be addressed. E-mail: gilbert@nucleus.harvard.edu.

 www.pnas.org/cgi/doi/10.1073/pnas.1232297100

Table 1. Apparent (but not actual) cases of intron discordance

Compared species	No. of analyzed introns							
	GenBank mRNA	Sliding	Alignment boundary	Alternative splicing	Boundary sliding	Low homology	GenBank errors	No *Fugu* homolog
Mouse–rat	108	79	17	4	6	11	14	—
Human–mouse	277	89	58	43	41	16	14	9

GenBank mRNA, apparent species-specific intron position due to GenBank record that is a mosaic of genomic and mRNA sequences; sliding, introns in one species at a position near to an intron position in the other species; alignment boundary, intron position is found at the boundary of the BLAST alignment; alternative splicing, an extra exon in one species leads to an extra, and thus unmatched, intron position; boundary sliding, coding sequence appears to have expanded or contracted at the boundary of the intron position; low homology, introns found in areas of low sequence homology; GenBank errors, unlikely exon–intron structures in GenBank records (see text); no *Fugu* homolog, human–mouse discordances whose ancestral states cannot be inferred because of the lack of a convincing *Fugu* ortholog.

with 1,459 intron positions. For the mouse-human comparison, there were 10,020 intron positions in 1,576 gene pairs.

Automated Intron Comparisons. For each pair of genes, we used our CIP program (10) to map the intron positions of each gene onto the corresponding BLAST protein alignment. All alignments in which there was at least one unique intron position in one of the species were inspected by eye. Those instances in which there was an intron present at a position in one species with no corresponding intron in the other in a region of good alignment were marked as instances of discordant intron–exon structure.

For each instance of discordant intron–exon structure, an ortholog from an outgroup was sought. For the rat–mouse comparison, we used human; for mouse–human, we used *Fugu* (individual gene structures were obtained from www.jgi.doe.gov). The three sequences were aligned by using the default options of CLUSTALW. There were several cases in which CLUSTALW placed one of the two genes from the more closely related species (human or mouse for human–mouse–*Fugu*; mouse or rat for rat–mouse–human) as the outgroup. We eliminated these as probable cases of paralogs. We then again used our CIP program to align the mouse gene to that of the outgroup for intron position comparison. If the outgroup had an intron at the discordant position, the difference was attributed to intron loss. If not, the difference was attributed to gain.

Manual Confirmation. Initially, our program identified 543 cases of mouse–human and 259 cases of rat–mouse intron discordance. However, on detailed inspection of the alignments, we found that all but a few cases were explained by several recurrent patterns (Table 1).

The most common case involved long stretches of alignment in which one organism had relatively large numbers of introns and the other had none. We checked many of these cases and found that all were due to GenBank records that joined genomic data to mRNA or cDNA data, with intron positions marked only in the genomic regions. Such cases accounted for more than half of the discordant intron positions (277 human–mouse; 108 mouse–rat).

The next most common pattern appeared as an intron position that had moved a few or several bases with respect to the coding sequences. Fifty-four human–mouse (71 mouse–rat) of these were identified by CIP as ''sliding'' based on a difference in positions of two codons or less. A further 35 human–mouse (8 mouse–rat) were identified by eye at distances of up to six codons. Although there is still a lively debate about the frequency of such sliding events in nature, in this instance we take most of these cases to be caused by database errors and/or misprediction of intron–exon boundaries by the various prediction programs.

A third pattern showed an intron at the boundary of an alignment, just as would be expected in the case of gene truncation or elongation by an intron-mediated process. This also does not fit the pattern of simple intron loss or gain because the sequence on one side of the intron is not homologous to the sequence from the orthologous gene. There were 58 such human–mouse (17 mouse–rat) cases.

Many other cases were apparent instances of alternative splicing, where a matched intron position flanked a region of alignment gap that extended exactly to a second unmatched intron position, after which the ungapped alignment between the two protein sequences was restored. This is exactly the pattern expected in the case of a comparison between different isoforms of the same transcript in different species. Forty-three human–mouse (4 mouse–rat) cases showed this pattern.

Equally important were apparent instances of ''boundary sliding,'' where one (but not both) boundaries of an intron appeared to have moved by an integral number of codons, converting some intron sequence to coding sequence (in the case of an intron contraction) or vice versa (in the case of an intron expansion). Without further investigation, it is impossible to determine which of these cases represent actual instances of intron boundary sliding and which are attributable to various prediction or annotation errors. Thus, we simply note that such a computational error would be easy to make, and thus we expect at least some of these to be mispredictions. There were 41 human–mouse (6 mouse–rat) cases.

A further 16 unmatched human–mouse (11 mouse–rat) introns fell in regions of very low amino acid sequence conservation, a pattern that is not expected from a simple gain or loss. Therefore, we discarded these cases as probable gene fusions or results of other genome dynamics. Finally, nine mismatched human–mouse intron positions had no corresponding orthologous region in *Fugu*, rendering their further investigation impossible.

This manual inspection of the database eliminated all but 18 human–mouse and 14 mouse–rat cases of discordant intron positions, which we then analyzed further.

Confirmation of GenBank Records. For all instances of discordant introns, we checked the GenBank records for the genes involved to ensure that the record seemed to contain a feasible gene structure. We thus identified several GenBank entries that contained ''introns'' of lengths 3, 2, 1, and even 0 bp, which we discarded. In addition, we were able to eliminate as unlikely a few instances in the human genome where the entire intron is a series of N′s, instances that we attribute to misapprehension of the relationship between adjacent contigs in the assembly. This analysis yielded 12 human–mouse and 5 mouse–rat cases of apparent intron loss and no cases of gain.

We then did an online BLAST search of the genome sequence for each of the 17 genes that appeared to have lost an intron. In seven mouse–human and four mouse–rat cases, we found that

Table 2. Results of intron comparison

Compared species	No. of gene pairs	Total introns	Results	
			Gain	Loss
Mouse–rat	360	1,459	0	1 (Rn)
Human–mouse	1,560	10,020	0	5 (Mm)

the genomic copy of the gene harbored an intron at exactly the same position as the orthologous gene. Thus, these initial identifications seem to be caused by GenBank errors in which intron positions were left out of the annotation.

This final analysis yielded six cases of well supported intron loss, five in mouse since divergence with human and one in rat since divergence with mouse, in which the GenBank record and the genomic copy agree as to the absence of an intron position (Table 2).

Results and Discussion

The first striking result of this analysis was the number of genes for which all intron positions matched exactly. Of 1,590 orthologous human–mouse pairs, 1,410 showed no deviations at all in intron alignment. Of 360 mouse–rat pairs, 307 were identical. Thus, even taking into account the multitude of problems with such a large-scale analysis (see below), 85% of mouse–rat orthologs have unambiguously maintained the entirety of their genomic structures for at least 30 million years (My) and 90% of mouse–human orthologs have done the same for 75 My.

Furthermore, of the remaining 543 discordant positions in 180 human–mouse orthologs (239 in 53 for mouse–rat), half were easily explained by one basic problem with many GenBank records. In each of these cases, there appeared to have been massive intron loss in one of the two organisms. However, upon closer inspection, we found that the apparent absence of introns was caused by an erroneous gene structure that was a mosaic of mRNA and genomic DNA. Thus, in this case, although some intron positions were included in the GenBank record, most were not, due to the sequence in the record having been partially derived from an mRNA.

However, this does not exclude the possibility that some of these cases are indeed examples of massive intron loss. Such instances have been well documented (11, 12). However, because with our protocol it is hard to distinguish such events of massive loss from GenBank errors or from retrotransposed pseudogenes, we have discarded these examples from the analysis. Our analysis is thus only sensitive to examples of simple intron loss or gain, events in which up- and downstream intron positions are unchanged.

This leaves only 264 human–mouse (131 mouse–rat) discordant positions of any possible biological interest (Table 1). Among these, 89 human–mouse (79 mouse–rat) occur at sites where the intron appears to have shifted over a few bases in one organism relative to the other, putative cases of so-called intron sliding. Although a few of these examples may actually represent true changes in the gene structure, it is hard to reconcile such an apparently high rate of intron movement, the most plausible models of which invoke intron excision and subsequent reinsertion followed by reverse transcription, with such low rates of the two separate processes loss and gain. As such, we think it much more likely that most of these cases are attributable to various computer and human errors in which intron positions have been mismarked. This notion is supported by the marked reduction in sequence similarity between orthologs in the region between the intron positions. It appears that parts of intron have been called coding and vice versa in one of the orthologs.

There are 41 human–mouse (6 mouse–rat) instances of apparent contraction or expansion of an intron, in which there is extra coding sequence directly to one side of the intron. This phenomenon is referred to as "boundary sliding." Again, without further investigation, it is impossible to detect whether such instances reflect true differences or failures of the gene prediction programs.

A further 43 human–mouse (4 mouse–rat) positions appear to be alignments between different spliceosomal isoforms of the same transcript. In these cases, the inclusion of an extra exon in one organism has led to an apparent intron discordance. Such instances could either reflect a true species-specific exon or, more likely, a failure of BLAST to find the proper orthologous transcript or a lack of the proper transcript in the database. Lacking further confirmation, it is hard to say whether such discordances actually reflect biological differences, but they are clearly not intron losses or gains.

In 58 human–mouse (17 mouse–rat) cases, the discordant intron lies within three codons of the end of the alignment. In these cases, the intron marks the boundary of the region of homology, exactly as one would expect in an intron-mediated gene fusion or truncation. However, such a pattern would also be expected if true coding sequence has been marked as an intron or vice versa, thus disrupting the sequence homology. Again, without further investigation, such instances are hard to interpret, though we are more optimistic that a substantial fraction of these instances may in fact represent interesting biological changes.

Finally, 16 human–mouse (11 mouse–rat) discordant introns were found in regions of bad mouse–human (mouse–rat) alignment, and a further 9 human–mouse cases were found in regions of human–mouse alignment that had no ortholog in *Fugu*. The former are worthy of further investigation, though they certainly do not resemble the simple intron gain or loss that we sought. Some fraction of the latter may represent real cases of intron movement, though at present it is impossible to tell because of the lack of an outgroup.

This left 18 cases of human–mouse intron discordance and 14 cases of mouse–rat intron discordance. In each case, we examined the GenBank files for the genes. In every instance where comparison with an outgroup appeared to show an intron insertion, the given GenBank file was very suspect. We found multiple instances of introns with lengths from 0 to 3 bp, all of which we discarded as extremely unlikely. We also found cases in the human genome assembly where the entire intron was populated by N′s, suggesting that this was caused by a failure of the assembler to recognize that the two flanking regions were in fact adjacent in the genome. In no case did we find a viable example of intron insertion. This is interesting in light of the finding of the newly inserted intron found in the SRY gene of marsupials and suggests that either this insertion is an extremely rare case or that some silencing of the intron insertion machinery occurred since the divergence of placental and marsupial mammals (13).

However, we did find several cases of loss (Fig. 1). In each of these cases, there is an intron at a given position in exactly one of the aligned species, and that intron is found in the genomic copy of the outgroup as well. Interestingly, as Table 3 shows, each of the corresponding introns is very short, with an overall mean of 205 bp, compared with an overall human mean of >2,500 bp and median of 1,820 bp (calculated from our intron–exon database). As Fig. 1 shows, all six protein sequence alignments show extremely strong sequence conservation. In addition, the regions directly flanking the discordant intron positions show equally strong conservation. There are no gaps in the alignments, suggesting that the loss was exact.

It is important to note that these instances are not cases of massive intron loss, for instance, through the genomic incorporation of the product of a retroposition. This is evident in Fig. 1, where it can be seen that neighboring intron positions are conserved and that the intron loss event is indeed specific to a single internal intron in each case.

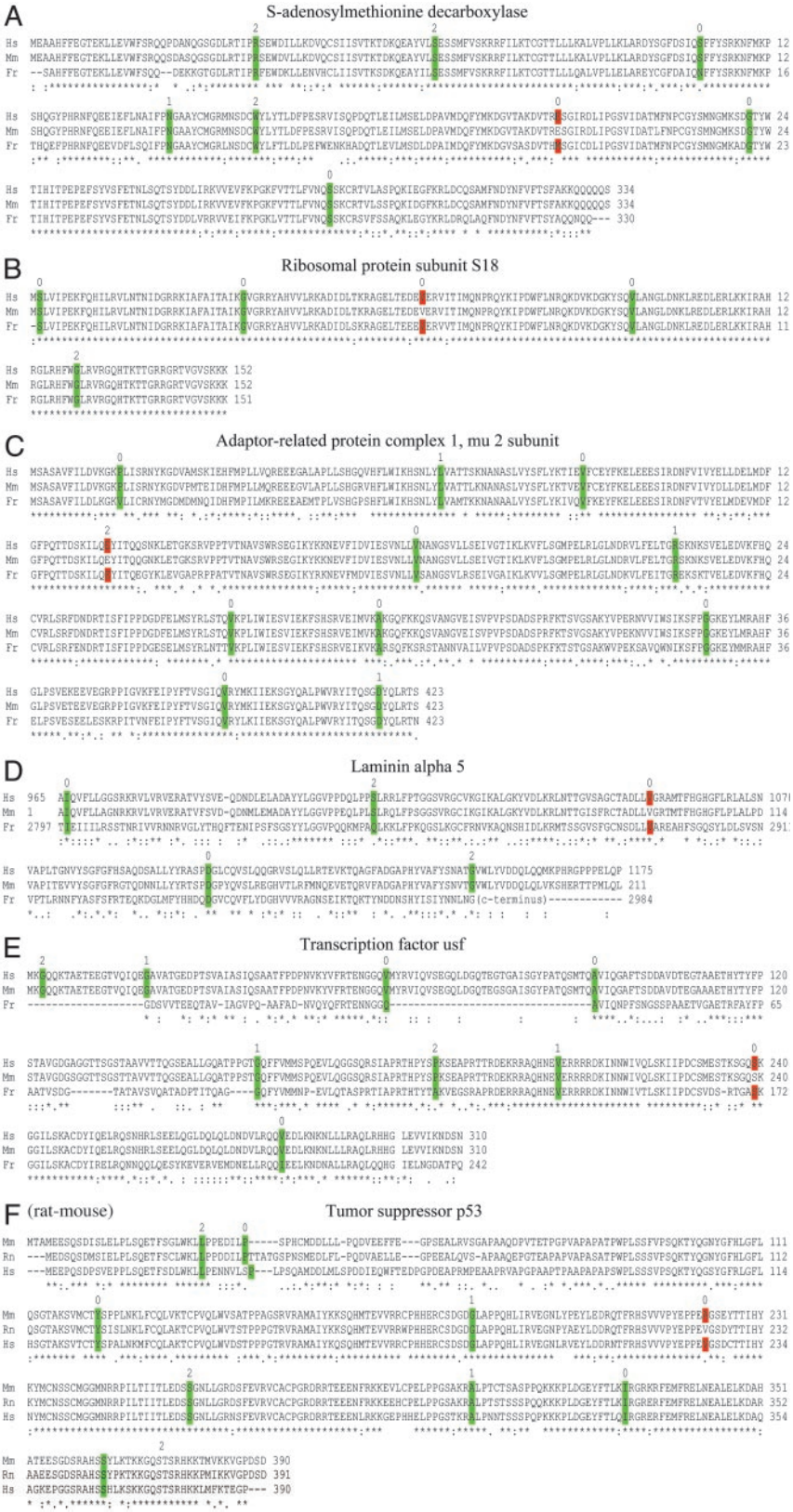

Fig. 1. Alignments with lost intron positions. Green boxes indicate introns shared among all three orthologs. Red boxes indicate introns that have been lost in one species. The phase of the intron is shown above the alignment.

It is striking that, in each case, the intron loss has not been accompanied by the creation of a gap in the alignment and that sequences flanking the lost intron have retained extremely high sequence similarly. The most straightforward model explaining this pattern invokes recombination between the genomic copy of a gene and a product of reverse transcription of a processed

Table 3. Characterization of deleted introns

Species	Gene identifier (EID release 132)	Gene name	Fig. 1 panel	Length of corresponding intron, nt
Mouse	128985_AB025024	*S*-adenosylmethionine decarboxylase	*A*	291 (Human); no (Rat)
Mouse	129712_AF100956	Ribosomal protein subunit S18	*B*	81 (Human); no (Rat)
Mouse	131047_F139406S11	Adaptor-related protein complex 1, mu 2 subunit	*C*	113 (Human); ? (Rat)
Mouse	131297_MMAJ6993	Laminin alpha 5	*D*	107 (Human); no (Rat)
Mouse	132494_MMU41741	Transcription factor usf	*E*	245 (Human); 223 (Rat)
Rat	133770_RATP53TS08	Tumor suppressor p53	*F*	393 (Mouse)

mRNA copy of the gene. The predictions of this model fit our data in two important ways. First, such a process is expected to cause exact deletions, because the intron will have been precisely excised in the mRNA, leaving the coding region intact. Second, such a process would be expected to favor the deletion of shorter introns, because the matching of sequence required to permit such double crossover events would be much easier. Although such a short-intron bias might also be expected from a bias in the length of spontaneous genomic deletions, such deletions would not be expected to be exact, as are the instances here.

These results are an important step in the debate over the relative roles of various processes in the shaping of the modern intron–exon structures of genes. First, they show that introns can be lost exactly, without alteration to nearby coding sequences. Second, they show that intron–exon structures can change in mammals. Third, although not statistically significant, they suggest that the rate of intron loss in rodents may be higher than in humans ($P = 0.07$, binomial distribution), and the fact that no case of gain or loss was found in humans suggests that human introns may have remained static for up to 75 million years. One explanation for a higher rate of loss in rodents could be rodents' shorter generation time. An analogous effect on the rate of synonymous site distribution has been demonstrated by Gillespie (14). Also intriguing is the observation that the intron corresponding to each loss in rodents is shorter than would be expected, suggesting that the mechanism of intron loss may favor short introns.

These results also cast doubt on two previously touted models. The most popular model of intron gain postulates retrotransposon properties of introns as responsible for their propagation (15, 16). However, if retrotransposon activity were a central mechanism for intron spread, one would expect to see many new introns in a group such as mammals, where so many transposable elements are so active. That we see no gains whatsoever suggests that introns do not move by simple retrotransposition. A second model is that of Rogers (17) and of Brenner and colleagues (3). These authors suggest that tandem duplications of exons could lead to the use of cryptic splice signals within the exons, thus creating an extra intron in the middle of a previously intact exon. Again, this model is difficult to reconcile with the lack of new introns in mammals, whose genomes exhibit thousands of tandem duplications.

Lastly, it is important to note that the intron losses observed here are not completely akin to the polymorphic loss recently found in the jingwei gene by Llopart *et al.* (18). Whereas the losses characterized here are exact, with no change in coding sequence, the post-intron-loss jingwei allele has an extra four codons relative to the ancestral allele as a result of an incomplete intron deletion. Thus, the jingwei example is unlikely to have been generated by recombination with a cDNA; the two different types of losses may be completely disparate in their mechanisms. In addition, the polymorphic state of the jingwei intron loss allowed the authors to demonstrate that the new allele shows a signature of positive selection, whereas it is impossible in the examples found here to infer the possible forces leading to the fixation of the post-intron-loss alleles.

The finding of six separate exact intron deletions in rodent lineages in the absence of any additions suggests that the process of intron loss may have a higher rate than that of intron gain in mammals. The pattern observed favors a model in which introns are lost through gene conversion with products generated by reverse transcription of mRNA copies of a gene, creating, after a double recombination, genomic gene copies in which the intron is exactly deleted and a pattern by which the loss of shorter introns is favored.

These results are also important in informing comparative gene prediction. If orthologs between human and mouse have virtually identical intron–exon structures, then cases of ambiguous assignment of intron boundaries should be resolvable by comparison with the other species. Although the intron–exon structure is nearly identical, silent positions have saturated between the genomes. Thus, when choosing, for instance, between multiple GT motifs as the beginning of an intron, comparison with the orthologous gene sequence will often strongly favor one as conserved between species. The inclusion of more genomic sequences from other species will only further strengthen such methods.

A.F. was supported by startup funds from the Bioinformatics Laboratory at the Medical College of Ohio.

1. Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. (1980) *Cell* **20,** 555–566.
2. Logsdon, J. M., Jr., Tyshenko, M. G., Dixon, C., D.-Jafari, J., Walker, V. K. & Palmer, J. D. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 8507–8511.
3. Venkatesh, B., Ning, Y. & Brenner, S. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 10267–10271.
4. Gotoh, O. (1998) *Mol. Biol. Evol.* **15,** 1447–1459.
5. Wada, H., Kobayashi, M., Satoh, R., Miyasaka, H. & Shirayama, Y. (2002) *J. Mol. Evol.* **54,** 118–128.
6. Logsdon, J. M., Stoltzfus, A. & Doolittle, W. F. (1998) *Curr. Biol.* **8,** R560–R563.
7. Logsdon, J. M., Jr. (1998) *Curr. Opin. Genet. Dev.* **8,** 637–648.
8. Paquette, S. M., Bak, S. & Feyereisen, R. (2000) *DNA Cell Biol.* **19,** 307–317.
9. Saxonov, S., Daizadeh, I., Fedorov, A. & Gilbert, W. (2000) *Nucleic Acids Res.* **28,** 185–190.
10. Fedorov, A., Merican, A. F. & Gilbert, W. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 16128–16133.
11. McCarrey, J. R. & Thomas, K. (1987) *Nature* **326,** 501–505.
12. Hendriksen, P. J., Hoogerbrugge, J. W., Baarends, W. M., de Boer, P., Vreeburg, J. T., Vos, E. A., van der Lende, T. & Grootegoed, J. A. (1997) *Genomics* **41,** 350–359.
13. O'Neill, R. J., Brennan, F. E., Delbridge, M. L., Crozier, R. H. & Graves, J. A. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 1653–1657.
14. Gillespie, J. H. (1991) *The Cause of Molecular Evolution* (Oxford Univ. Press, Oxford).
15. Cavalier-Smith, T. (1985) *Nature* **315,** 283–284.
16. Palmer, J. D. & Logsdon, J. M., Jr. (1991) *Curr. Opin. Genet. Dev.* **1,** 470–477.
17. Rogers, J. H. (1990) *FEBS Lett.* **268,** 339–343.
18. Llopart, A., Comeron, J. M., Brunet, F. G., Lachaise, D. & Long, M. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 8121–8126.