Published online 7 February 2008 *Nucleic Acids Research, 2008, Vol. 36, No. 5* ***1703–1712***
doi:10.1093/nar/gkn012

Spliceosomal introns as tools for genomic and evolutionary analysis

Manuel Irimia[1] **and Scott William Roy**[2,*]

[1]Departament de Genètica, Universitat de Barcelona, Barcelona, Spain and [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Received December 1, 2007; Revised January 7, 2008; Accepted January 8, 2008

ABSTRACT

Over the past 5 years, the availability of dozens of whole genomic sequences from a wide variety of eukaryotic lineages has revealed a very large amount of information about the dynamics of intron loss and gain through eukaryotic history, as well as the evolution of intron sequences. Implicit in these advances is a great deal of information about the structure and evolution of surrounding sequences. Here, we review the wealth of ways in which structures of spliceosomal introns as well as their conservation and change through evolution may be harnessed for evolutionary and genomic analysis. First, we discuss uses of intron length distributions and positions in sequence assembly and annotation, and for improving alignment of homologous regions. Second, we review uses of introns in evolutionary studies, including the utility of introns as indicators of rates of sequence evolution, for inferences about molecular evolution, as signatures of orthology and paralogy, and for estimating rates of nucleotide substitution. We conclude with a discussion of phylogenetic methods utilizing intron sequences and positions.

INTRODUCTION

Patterns and evolution of intron–exon structures

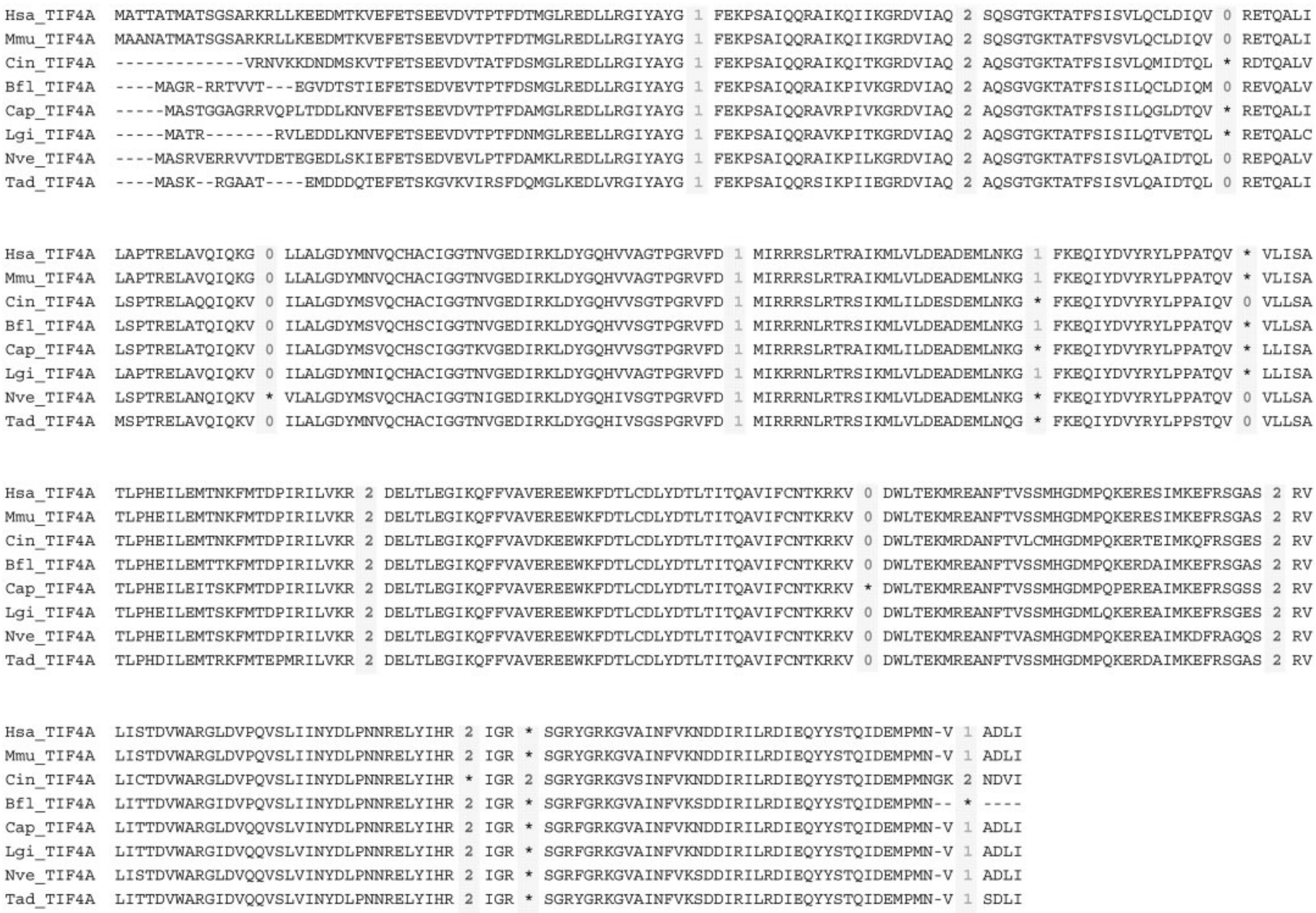
Spliceosomal introns are sequences that interrupt eukaryotic genes and are removed from RNA transcripts by the spliceosome, a complex cellular RNA–protein machine incorporating five RNAs and hundreds of proteins (1). Our understanding of the evolution of spliceosomal introns has increased exponentially over the past few years due to the release of many genome sequences from most major eukaryotic lineages, both about intron loss and gain dynamics (2–8), as well as the evolution of intron sequences and splicing (9–13). [Several reviews have recently tackled the question of the evolution of introns in eukaryotes (14–19)]. During the first 25 years after their discovery in 1977 (20), much of the study of spliceosomal introns focused on a debate about the timing of origin of the first introns (21), whether before the divergence of eukaryotes from prokaryotes (which lack spliceosomal introns) (22–25) or within the evolutionary history of eukaryotes (26–31). Although this debate continues, the momentum has clearly tipped towards the perspective that introns appeared once in early (or pre-) eukaryotic evolution by the proliferation and transformation of type II self-splicing introns (26,27,32,33), possibly transferred from the mitochondrion.

Over the past 5 years, the focus has shifted away from the question of the ultimate origin of introns to attempts to track the history of intron loss/gain and intron sequence evolution during eukaryotic history. We can now be confident that large numbers of introns were present by early eukaryotic history (14,34–40) and that many or even most modern introns date to the times of early eukaryotic ancestors. Over at least recent eukaryotic evolution (say, the last ~100 My), intron gain has been a very rare event, with most lineages experiencing rates of gain corresponding to <0.0002 gains per gene per million years (7,8,41–46). Rates of intron loss have been more variable: in some lineages, rates of loss are perhaps 10% per 100 My, whereas other lineages have experienced almost no intron loss over tens or hundreds of millions of years (2,6,7,42,44–48). Figure 1 shows the example of metazoans, where the majority of intron positions have been retained between vertebrates and basal animals.

Moreover, intron positions have been shown to be very constant over time—i.e. that intron 'sliding', in which an intron would migrate a few base pairs along a gene, is a very rare occurrence (4,49,50). Several studies have also documented the mechanisms of intron loss: patterns of intron loss including 3′-biased intron loss (7,51–53), exact

*To whom correspondence should be addressed. Tel: 301 496 8941; Fax: 301 480 2288; Email: scottwroy@gmail.com

The authors wish it to be known that, in their opinion, the authors should be regarded as joint First Authors.



```
Hsa_TIF4A  MATTATMATSGSARKRLLKEEDMTKVEFETSEEVDVTPTFDTMGLREDLLRGIYAYG 1 FEKPSAIQQRAIKQIIKGRDVIAQ 2 SQSGTGKTATFSISVLQCLDIQV 0 RETQALI
Mmu_TIF4A  MAANATMATSGSARKRLLKEEDMTKVEFETSEEVDVTPTFDTMGLREDLLRGIYAYG 1 FEKPSAIQQRAIKQIIKGRDVIAQ 2 SQSGTGKTATFSVSVLQCLDIQV 0 RETQALI
Cin_TIF4A  -------------VRNVKKDNDMSKVTFETSEEVDVTATFDSMGLREDLLRGIYAYG 1 FEKPSAIQQRAIKQITKGRDVIAQ 2 AQSGTGKTATFSISVLQMIDTQL * RDTQALV
Bfl_TIF4A  ----MAGR-RRTVVT---EGVDTSTIEFETSEDVEVTPTFDSMGLREDLLRGIYAYG 1 FEKPSAIQQRAIKPIVKGRDVIAQ 2 AQSGVGKTATFSISILQCLDIQM 0 REVQALV
Cap_TIF4A  -----MASTGGAGRRVQPLTDDLKNVEFETSEEVDVTPTFDAMGLREDLLRGIYAYG 1 FEKPSAIQQRAVRPIVKGRDVIAQ 2 AQSGTGKTATFSISILQGLDTQV * RETQALI
Lgi_TIF4A  -----MATR-------RVLEDDLKNVEFETSEEVDVTPTFDNMGLREELLRGIYAYG 1 FEKPSAIQQRAVKPITKGRDVIAQ 2 AQSGTGKTATFSISILQTVETQL * RETQALC
Nve_TIF4A  ----MASRVERRVVTDETEGEDLSKIEFETSEDVEVLPTFDAMKLREDLLRGIYAYG 1 FEKPSAIQQRAIKPILKGRDVIAQ 2 AQSGTGKTATFSISVLQAIDTQL 0 REPQALV
Tad_TIF4A  ----MASK--RGAAT----EMDDDQTEFETSKGVKVIRSFDQMGLKEDLVRGIYAYG 1 FEKPSAIQQRSIKPIIEGRDVIAQ 2 AQSGTGKTATFSISVLQAIDTQL 0 RETQALI


Hsa_TIF4A  LAPTRELAVQIQKG 0 LLALGDYMNVQCHACIGGTNVGEDIRKLDYGQHVVAGTPGRVFD 1 MIRRRSLRTRAIKMLVLDEADEMLNKG 1 FKEQIYDVYRYLPPATQV * VLISA
Mmu_TIF4A  LAPTRELAVQIQKG 0 LLALGDYMNVQCHACIGGTNVGEDIRKLDYGQHVVAGTPGRVFD 1 MIRRRSLRTRAIKMLVLDEADEMLNKG 1 FKEQIYDVYRYLPPATQV * VLISA
Cin_TIF4A  LSPTRELAQQIQKV 0 ILALGDYMSVQCHACIGGTNVGEDIRKLDYGQHVVSGTPGRVFD 1 MIRRRSLRTRSIKMLILDESDEMLNKG * FKEQIYDVYRYLPPAIQV 0 VLLSA
Bfl_TIF4A  LSPTRELATQIQKV 0 ILALGDYMSVQCHSCIGGTNVGEDIRKLDYGQHVVSGTPGRVFD 1 MIRRRNLRTRSIKMLVLDEADEMLNKG 1 FKEQIYDVYRYLPPATQV * VLLSA
Cap_TIF4A  LSPTRELATQIQKV 0 ILALGDYMSVQCHSCIGGTKVGEDIRKLDYGQHVVSGTPGRVFD 1 MIRRRSLRTRAIKMLILDEADEMLNKG * FKEQIYDVYRYLPPATQV * LLISA
Lgi_TIF4A  LAPTRELAVQIQKV 0 ILALGDYMNIQCHACIGGTNVGEDIRKLDYGQHVVAGTPGRVFD 1 MIKRRNLRTRSIKMLVLDEADEMLNKG 1 FKEQIYDVYRYLPPATQV * LLISA
Nve_TIF4A  LSPTRELANQIQKV * VLALGDYMSVQCHACIGGTNIGEDIRKLDYGQHIVSGTPGRVFD 1 MIRRRNLRTRSIKMLVLDEADEMLNKG * FKEQIYDVYRYLPPATQV 0 VLLSA
Tad_TIF4A  MSPTRELAVQIQKV 0 ILALGDYMNVQCHACIGGTNVGEDIRKLDYGQHIVSGSPGRVFD 1 MIRRRNLRTRSIKMLVLDEADEMLNQG * FKEQIYDVYRYLPPSTQV 0 VLLSA


Hsa_TIF4A  TLPHEILEMTNKFMTDPIRILVKR 2 DELTLEGIKQFFVAVEREEWKFDTLCDLYDTLTITQAVIFCNTKRKV 0 DWLTEKMREANFTVSSMHGDMPQKERESIMKEFRSGAS 2 RV
Mmu_TIF4A  TLPHEILEMTNKFMTDPIRILVKR 2 DELTLEGIKQFFVAVEREEWKFDTLCDLYDTLTITQAVIFCNTKRKV 0 DWLTEKMREANFTVSSMHGDMPQKERESIMKEFRSGAS 2 RV
Cin_TIF4A  TLPHEILEMTNKFMTDPIRILVKR 2 DELTLEGIKQFFVAVDKEEWKFDTLCDLYDTLTITQAVIFCNTKRKV 0 DWLTEKMRDANFTVLCMHGDMPQKERTEIMKQFRSGES 2 RV
Bfl_TIF4A  TLPHEILEMTTKFMTDPIRILVKR 2 DELTLEGIKQFFVAVEREEWKFDTLCDLYDTLTITQAVIFCNTKRKV 0 DWLTEKMREANFTVSSMHGDMPQKERDAIMKEFRSGAS 2 RV
Cap_TIF4A  TLPHEILEITSKFMTDPIRILVKR 2 DELTLEGIKQFFVAVEREEWKFDTLCDLYDTLTITQAVIFCNTKRKV * DWLTEKMREANFTVSSMHGDMPQPEREAIMKEFRSGSS 2 RV
Lgi_TIF4A  TLPHEILEMTSKFMTDPIRILVKR 2 DELTLEGIKQFFVAVEREEWKFDTLCDLYDTLTITQAVIFCNTKRKV 0 DWLTEKMREANFTVSSMHGDMLQKEREAIMKEFRSGES 2 RV
Nve_TIF4A  TLPHEILEMTSKFMTDPIRILVKR 2 DELTLEGIKQFFVAVEREEWKFDTLCDLYDTLTITQAVIFCNTKRKV 0 DWLTEKMREANFTVASMHGDMPQKEREAIMKDFRAGQS 2 RV
Tad_TIF4A  TLPHDILEMTRKFMTEPMRILVKR 2 DELTLEGIKQFFVAVEREEWKFDTLCDLYDTLTITQAVIFCNTKRKV 0 DWLTEKMREANFTVSSMHGDMPQKERDAIMKEFRSGAS 2 RV


Hsa_TIF4A  LISTDVWARGLDVPQVSLIINYDLPNNRELYIHR 2 IGR * SGRYGRKGVAINFVKNDDIRILRDIEQYYSTQIDEMPMN-V 1 ADLI
Mmu_TIF4A  LISTDVWARGLDVPQVSLIINYDLPNNRELYIHR 2 IGR * SGRYGRKGVAINFVKNDDIRILRDIEQYYSTQIDEMPMN-V 1 ADLI
Cin_TIF4A  LICTDVWARGLDVPQVSLIINYDLPNNRELYIHR * IGR 2 SGRYGRKGVSINFVKNDDIRILRDIEQYYSTQIDEMPMNGK 2 NDVI
Bfl_TIF4A  LITTDVWARGIDVPQVSLIINYDLPNNRELYIHR 2 IGR * SGRFGRKGVAINFVKSDDIRILRDIEQYYSTQIDEMPMN-- * ----
Cap_TIF4A  LITTDVWARGLDVQQVSLVINYDLPNNRELYIHR 2 IGR * SGRFGRKGVAINFVKNDDIRILRDIEQYYSTQIDEMPMN-V 1 ADLI
Lgi_TIF4A  LITTDVWARGIDVQQVSLVINYDLPNNRELYIHR 2 IGR * SGRFGRKGVAINFVKNDDIRILRDIEQYYSTQIDEMPMN-V 1 ADLI
Nve_TIF4A  LISTDVWARGLDVQQVSLVINYDLPNNRELYIHR 2 IGR * SGRFGRKGVAINFVKSDDIRILRDIEQYYSTQIDEMPMN-V 1 ADLI
Tad_TIF4A  LITTDVWARGIDVQQVSLVINYDLPNNRELYIHR 2 IGR * SGRYGRKGVAINFVKSDDIRILRDIEQYYSTQIDEMPMN-V 1 SDLI
```

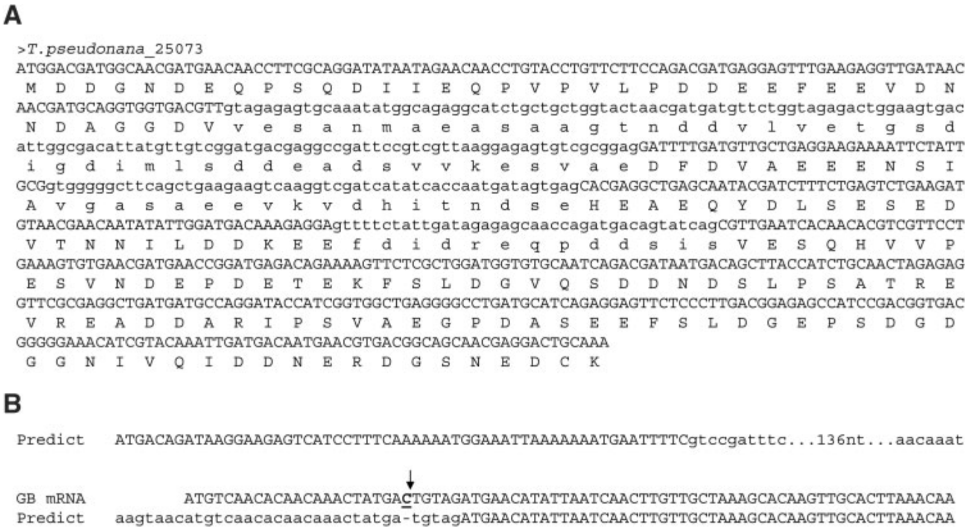
Figure 1. Intron positions are often conserved over long evolutionary times. Protein-level alignments of the translation initiation factor 4A gene (*TIF4A*) from a variety of metazoan species are shown. Intron positions are indicated by digits corresponding to the phase of the intron relative to the surrounding codons (phases 0, 1 and 2 introns fall before the first, second and third bases of a codon, respectively). Most intron positions are conserved at the exact homologous position and phase over all of animal history within these species, all the way from the placozoan *Trichoplax adhaerens* to chordates. Abbreviations: Hsa, *Homo sapiens*; Mmu, *Mus musculus*; Cin, *Ciona intestinalis*; Bfl, *Branchiostoma floridae*; Cap, *Capitella sp*; Lgi, *Lottia gigantea*; Nve, *Nematostella vectensis*; Tad, *Trichoplax adhaerens*.

removal of intron sequences (2), apparently coincident loss of adjacent introns (7,47,54) and possibly germline-biased intron loss (unpublished data), all seem to indicate that intron loss proceeds via reverse transcription of RNA intermediates (55,56). In addition, comparative analyses have uncovered an apparent (though incomplete) correspondence between rates of intron loss and rates of sequence evolution—degree of loss of ancestral introns appears directly correlated with degree of sequence change [(57) and unpublished data].

Introns as the repository of information about gene structure

The focus of this article is not on these patterns of evolution themselves, but on their implications for analysis of genome structures and eukaryotic evolution in general. Although introns have largely been regarded as a hindrance for genome analysis given the difficulties associated with gene annotation in the presence of introns, intron positions and sequences are potentially very useful in addressing a wide variety of important genomic and evolutionary problems. In particular, intron loss/gain has been shown to be a very slow process in many lineages relative to other genetic characters (sequence evolution of proteins, genes and non-coding DNA, insertion and deletion of transposable elements and even genome rearrangement), thus intron positions contain (and retain) a large amount of information about genome structure and deep evolutionary history.

Over the past few years, a variety of researchers from disparate fields have developed methods that harness this information for purposes ranging from reconstructing evolutionary phylogenies to improving gene prediction, from alignment of homologous protein sequences to assignment of orthology in large protein families. Many of these methods are already quite powerful, though often known primarily to those working on introns themselves, rather than those working on the problems addressed by the methods. Other methods are still largely undeveloped, and represent promising future lines of work. Here, we review these approaches, and delineate possible uses in genomic and evolutionary study.



```
A

>T.pseudonana_25073
ATGGACGATGGCAACGATGAACAACCTTCGCAGGATATAATAGAACAACCTGTACCTGTTCTTCCAGACGATGAGGAGTTTGAAGAGGTTGATAAC
 M  D  D  G  N  D  E  Q  P  S  Q  D  I  I  E  Q  P  V  P  V  L  P  D  D  E  E  F  E  E  V  D  N
AACGATGCAGGTGGTGACGTTgtagagagtgcaaatatggcagaggcatctgctgctggtactaacgatgatgttctggtagagactggaagtgac
 N  D  A  G  G  D  V  v  e  s  a  n  m  a  e  a  s  a  a  g  t  n  d  d  v  l  v  e  t  g  s  d
attggcgacattatgttgtcggatgacgaggccgattccgtcgttaaggagagtgtcgcggagGATTTTGATGTTGCTGAGGAAGAAAATTCTATT
 i  g  d  i  m  l  s  d  d  e  a  d  s  v  v  k  e  s  v  a  e  D  F  D  V  A  E  E  E  N  S  I
GCGgtgggggcttcagctgaagaagtcaaggtcgatcatatcaccaatgatagtgagCACGAGGCTGAGCAATACGATCTTTCTGAGTCTGAAGAT
 A  v  g  a  s  a  e  e  v  k  v  d  h  i  t  n  d  s  e  H  E  A  E  Q  Y  D  L  S  E  S  E  D
GTAACGAACAATATATTGGATGACAAAGAGGAgttttctattgatagagagcaaccagatgacagtatcagCGTTGAATCACAACACGTCGTTCCT
 V  T  N  N  I  L  D  D  K  E  E  f  d  i  d  r  e  q  p  d  d  s  i  s  V  E  S  Q  H  V  V  P
GAAAGTGTGAACGATGAACCGGATGAGACAGAAAAGTTCTCGCTGGATGGTGTGCAATCAGACGATAATGACAGCTTACCATCTGCAACTAGAGAG
 E  S  V  N  D  E  P  D  E  T  E  K  F  S  L  D  G  V  Q  S  D  D  N  D  S  L  P  S  A  T  R  E
GTTCGCGAGGCTGATGATGCCAGGATACCATCGGTGGCTGAGGGGCCTGATGCATCAGAGGAGTTCTCCCTTGACGGAGAGCCATCCGACGGTGAC
 V  R  E  A  D  D  A  R  I  P  S  V  A  E  G  P  D  A  S  E  E  F  S  L  D  G  E  P  S  D  G  D
GGGGGAAACATCGTACAAATTGATGACAATGAACGTGACGGCAGCAACGAGGACTGCAAA
 G  G  N  I  V  Q  I  D  D  N  E  R  D  G  S  N  E  D  C  K

B

Predict   ATGACAGATAAGGAAGAGTCATCCTTTCAAAAAATGGAAATTAAAAAAATGAATTTTCgtccgatttc...136nt...aacaaat

                                      ↓
GB mRNA          ATGTCAACACAACAAACTATGACTGTAGATGAACATATTAATCAACTTGTTGCTAAAGCACAAGTTGCACTTAAACAA
Predict   aagtaacatgtcaacacaacaaactatga-tgtagATGAACATATTAATCAACTTGTTGCTAAAGCACAAGTTGCACTTAAACAA
```

Figure 2. Intron length distributions and gene prediction. Since introns are removed from transcripts, they are not expected to respect coding frame, predicting roughly equal numbers of introns with lengths of a multiple of three nucleotides (3n), 3n + 1 and 3n + 2. (**A**) The majority of predicted introns in the genome of the diatom *Thalassiosira pseudonana* preserve reading frame (i.e. they are 3n nucleotides and lack in frame stop codons), suggesting that many predicted introns are instead coding sequence. All three introns of this predicted gene (JGI ID25063) preserve reading frame (56). (**B**) Predicted introns in the *E. histolytica* genomes are disproportionately 3n + 2. This excess appears to reflect single bases left out of the assembly (i.e. indels), with introns predicted in order to restore reading frame. Here, a predicted intronic sequence from the genome annotation is present in an mRNA from GenBank, and is missing a single cytosine nucleotide (7).

USES OF SPLICEOSOMAL INTRONS FOR GENOME SEQUENCING, ANNOTATION AND ALIGNMENT

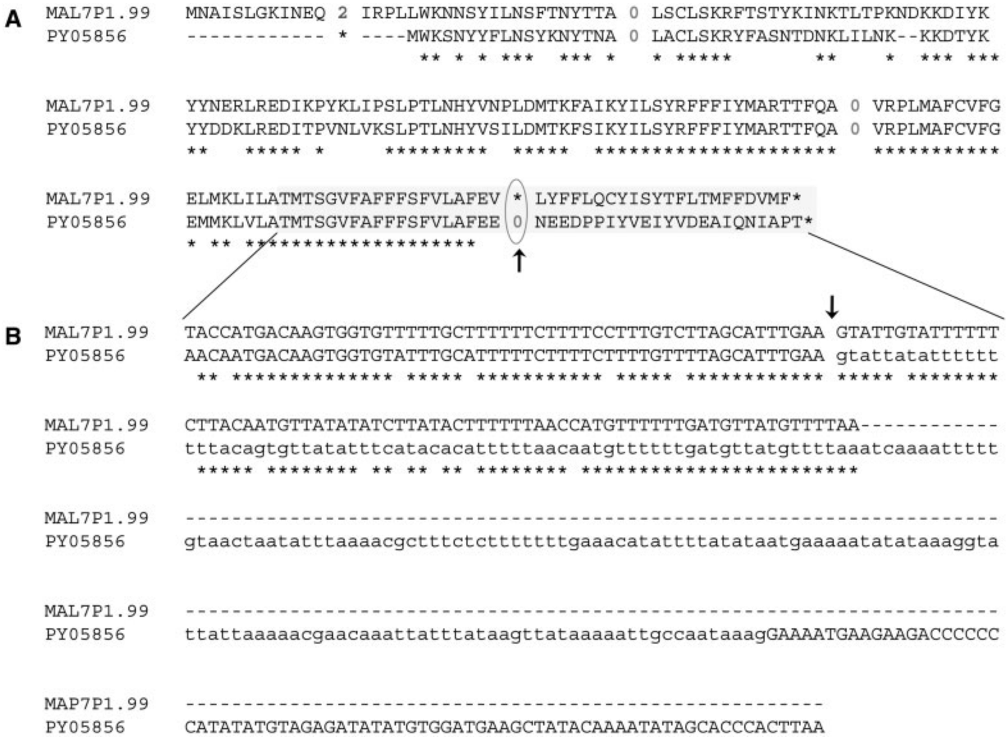
Intron length distributions and genome assembly and annotation

A potential utility of introns for large-scale sequencing efforts involves the distribution of intron lengths. Since introns are removed from protein-coding transcripts, intron lengths are not expected to respect coding frame: across the genome, we expect roughly equal proportions of introns that are multiple of three bases ('3n' introns), one more than a multiple of three bases (3n + 1) and two more (3n + 2). However, one of us and David Penny (58) recently reported a survey of predicted genes from genome annotations across 29 different species in which we found common deviations from this expectation. In some cases, the number of 3n introns was much larger than the numbers of 3n + 1 or 3n + 2 introns (Figure 2A), in other cases less. Further investigation indicated that many such cases seemed to be due to genome-wide problems in annotation. Such an internal check for genome annotations could constitute an important step in improving genome annotations before their public release. [While this paper was in press, a new report by Jaillon et al. (Nature 2008 451:359–62) showed a real biological deficit of 3n introns owing to selection for nonsense mediated decay. This result cannot however explain predicted proteomes with other types of skewed intron length distribution].

Our previous study (8) also showed a case in which analysis of gene and intron annotations was able to identify a previously unnoticed large number of indels in a genome assembly (example in Figure 2B). In the case of *Entamoeba histolytica*, the publicly available annotation showed a pronounced excess of 3n + 2 introns. Genome-wide computational and manual inspection of predicted introns indicated that the majority of these excess 3n + 2 introns were associated with a single missing base in the assembled genome sequence—many of these cases appear to be actual coding sequence which had been disrupted by the lack of a base, leading to false prediction of an intron in order to keep the predicted gene sequence in frame. Thus, in some cases, scrutiny of genome-wide intron length distributions from preliminary gene predictions could indicate otherwise undetected errors in genome assembly.

Intron position conservation and improved gene annotation

Intron positions are very often conserved over very long evolutionary distances (Figure 1). In some lineages, this reaches extremes. In *Theileria* apicomplexans, 99.7% of intron positions are conserved between *T. parva* and *T. annulata*, diverged roughly 82 Mya (2). In mammals, 99.9% of intron positions are conserved between human and dog, diverged around 100 Mya (42). Lineages with significant numbers of introns are particularly difficult to annotate in the absence of exhaustive transcript sequence information. Here, comparison with species which are known to have very similar intron–exon structures in orthologous regions could vastly improve uncertain annotations (59). When predicted intron positions are mapped onto protein-level alignments of predicted orthologs, the results are often very clear—protein-level sequence similarity will cease abruptly at the boundary of a species-specific intron position [(2,60,61) and unpublished observations]. While this could in fact reflect



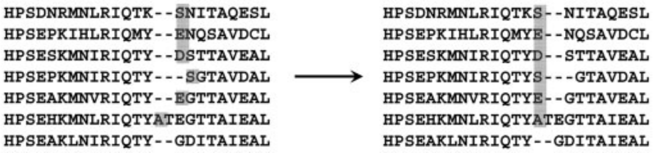
```
MAL7P1.99    MNAISLGKINEQ 2 IRPLLWKNNSYILNSFTNYTTA 0 LSCLSKRFTSTYKINKTLTPKNDKKDIYK
PY05856      ------------ * ----MWKSNYYFLNSYKNYTNA 0 LACLSKRYFASNTDNKLILNK--KKDTYK
                                 ** * * ***  *** *   * *****       **    *  *** ***

MAL7P1.99    YYNERLREDIKPYKLIPSLPTLNHYVNPLDMTKFAIKYILSYRFFFIYMARTTFQA 0 VRPLMAFCVFG
PY05856      YYDDKLREDITPVNLVKSLPTLNHYVSILDMTKFSIKYILSYRFFFIYMARTTFQA 0 VRPLMAFCVFG
             **   ***** *     *********   *****  *********************   ***********

MAL7P1.99    ELMKLILATMTSGVFAFFFSFVLAFEV * LYFFLQCYISYTFLTMFFDVMF*
PY05856      EMMKLVLATMTSGVFAFFFSFVLAFEE 0 NEEDPPIYVEIYVDEAIQNIAPT*
             * ** ********************

MAL7P1.99    TACCATGACAAGTGGTGTTTTTGCTTTTTTCTTTTCCTTTGTCTTAGCATTTGAA GTATTGTATTTTTT
PY05856      AACAATGACAAGTGGTGTATTTGCATTTTTCTTTTCTTTTGTTTTAGCATTTGAA gtattatattttttt
              ** ************** ***** *********** ***** ************ ***** ********

MAL7P1.99    CTTACAATGTTATATATCTTATACTTTTTTAACCATGTTTTTTGATGTTATGTTTTAA-----------
PY05856      tttacagtgttatatttcatacacatttttaacaatgtttttgatgttatgttttaaatcaaaatttt
              ***** ******** ** ** ** ******** ************************

MAL7P1.99    ---------------------------------------------------------------------
PY05856      gtaactaatatttaaaacgctttctctttttttgaaacatattttatataatgaaaaatatataaaggta

MAL7P1.99    ---------------------------------------------------------------------
PY05856      ttattaaaaacgaacaaattatttataagttataaaaattgccaataaagGAAAATGAAGAAGACCCCCC

MAP7P1.99    -------------------------------------------------------
PY05856      CATATATGTAGAGATATATGTGGATGAAGCTATACAAAATATAGCACCCACTTAA
```

Figure 3. Reconciling orthologous intron–exon structures to improve gene predictions. (**A**) Clear protein similarity between the *MAL7P1.99* and *PY05856* genes of *P. falciparum* and *P. yoelii* continues through two conserved intron positions, and then ends abruptly at a predicted *P. yoelii*-specific intron. (**B**) Alignment of the genes at the DNA level shows strong sequence similarity between the predicted *P. yoelii* intron and the predicted *P. falciparum* coding sequence, strongly suggesting that the predicted *P. yoelli* intron instead is coding.

biological reality, it seems very likely that this often reflects misprediction of the intron in one species—either there is an intron present in both species, which has gone unpredicted in one, or there is no intron at that position in either species, and truly exonic sequence in one species has been predicted as an intron.

Figure 3 shows a pair of orthologs from *Plasmodium falciparum* and *P. yoelii*. Clear protein sequence similarity continues through conserved intron positions, and then ends abruptly at a *P. yoelii*-specific predicted intron. Alignment of the regions at the DNA level clearly shows that the sequence of the *P. yoelii*-specific predicted intron is highly similar to the *P. falciparum* sequence. This pattern strongly suggests that the predicted intron sequence is instead coding sequence homologous to the *P. falciparum* coding sequence.

Reconciliation between protein models in species pairs or clusters could greatly improve gene predictions, particularly in species where highly skewed sequence composition or frequent repetitive sequence renders accurate gene prediction most difficult. For example, Coghlan and Durbin (62) recently presented a new method to combine predictions from different gene finders used for one species by comparing intron/exon structures of the different predictions and between gene models of closely related species, building gene structures based on the most conserved exons. They applied this methodology to the nematodes *Caenorhabditis briggsae* and *C. remanei*, obtaining increases of >10% in exon-level specificity and almost 3% in sensitivity, compared to the best outputs from previous gene finders.



```
HPSDNRMNLRIQTK--SNITAQESL         HPSDNRMNLRIQTKS--NITAQESL
HPSEPKIHLRIQMY--ENQSAVDCL         HPSEPKIHLRIQMYE--NQSAVDCL
HPSESKMNIRIQTY--DSTTAVEAL         HPSESKMNIRIQTYD--STTAVEAL
HPSEPKMNIRIQTY---SGTAVDAL  ---->  HPSEPKMNIRIQTYS---GTAVDAL
HPSEAKMNVRIQTY--EGTTAVEAL         HPSEAKMNVRIQTYE--GTTAVEAL
HPSEHKMNLRIQTYATEGTTAIEAL         HPSEHKMNLRIQTYATEGTTAIEAL
HPSEAKLNIRIQTY--GDITAIEAL         HPSEAKLNIRIQTY--GDITAIEAL
```

Figure 4. Using intron positions to improve protein sequence alignments. Protein sequence alignment in regions with significant change is often ambiguous (left). Alignment of intron positions indicates the likely true alignment (from 61).

Improved protein sequence alignments

Since introns often maintain their positions over very long evolutionary timescales, intron positions can often retain information about gene homology after protein sequences have experienced enough change as to render alignment difficult. As such, intron positions can be used as check points in protein alignments, improving their quality. Recently, Csuros and coauthors (63) developed a method to use intron positions to improve protein-level alignments in regions of questionable alignment (Figure 4). They introduced alignment penalties and rewards for intron positions into the alignment matrixes, considering intron position matches/mismatches scoring alternative alignments. The use of these algorithms significantly improved the quality of some gapped alignments.

INTRONS AS TOOLS TO STUDY MOLECULAR EVOLUTION

Intron density as a surrogate for rates of sequence evolution

A central goal of full genome sequencing is understanding the evolutionary history of ourselves and other organisms. Identification of slow evolving lineages (the so-called 'living fossils') is thus of central interest, both for what they can reveal about organismal complexity and genome structure evolution. In this context, intron density may serve as an important indicator for branch length. Since rates of intron gain across a wide variety of lineages have been very low (2,6,7,42,44–47), intron numbers in modern species often largely reflect the extent of intron loss since intron-rich ancestors. Intron number is therefore inversely correlated to 'branch length', in terms of intron loss.

Interestingly, and somewhat surprisingly given the very different sets of mutation and presumed evolutionary forces controlling intron loss and sequence evolution, a correspondence between degree of intron loss and degree of sequence change has been found in some eukaryotic lineages. The most straightforward case involves a genomic survey across metazoan genomes (57). The genomes which had experienced the least intron loss since the metazoan ancestor (vertebrates and the marine annelid *Platynereis*) also had experienced less sequence change since the ancestor than other studied lineages (from among insects, urochordates and nematodes). A second example concerns the multitude of nearly intronless protists that also exhibit very high degrees of sequence change (29,64). If in fact a relationship between intron number and branch lengths holds generally, intron density as estimated from small-scale genomic sequencing efforts could be useful in identifying short-branch taxa.

Inferences about molecular evolution

Patterns of intron distribution and evolution can also provide insights into other aspects of molecular evolution. First, intron presence/absence is useful in inferring the mechanism of gene duplication, since intron absence is a hallmark of gene duplication by retroposition (65). Thousands of intron-less copies of the so-called 'processed pseudogenes' are present in many eukaryotic genomes (66,67), originated by retrotranscription of processed mRNAs and subsequent insertion into the genome (65). This mechanism can be easily distinguished from segmental genome duplication by the absence of introns (65,68). Second, the correspondence of intron positions with the boundaries of domains whose reshuffling contributed to the origin of new proteins in metazoan lineages allows the reconstruction of the mechanism of origin of multi-domain protein-encoding genes (69–73).

Given the apparent dependency of intron loss on reverse transcriptase, rates of intron loss could also provide information about the presence and activity of retroelements through evolutionary history. For instance, we recently showed that rates of intron loss have varied by orders of magnitude in the history of apicomplexan evolution (74). Given the apparent dependence of intron loss on retroelement activity (52), the lack of known active retroelements in modern *Plasmodium* and *Theileria* species is consistent with the lack of recent intron loss (74). If so, the much more extensive loss in both the *Plasmodium* and *Theileria* ancestors since the genera's divergence suggests retroelement activity. Thus the pattern of intron loss through time may provide information about the activity of retroelements over evolutionary depths where the actual retroelement insertion history has been erased by subsequent mutation.

Intron positions as signatures of orthology and paralogy

Since the rate of intron loss in modern organisms is often very low, the pattern of intron positions can be used as an indication for orthology among paralogous groups (75–78). These studies usually complement others such as classical phylogenetic analysis of gene families or synteny comparisons, and may be especially useful in the annotation of newly sequenced genomes in assigning orthology among large gene families with very similar domains (such as kinases, TGF-βs, immunoglobulins, etc.) that are hard to distinguish by traditional phylogenetic methods. Some evidence suggests that patterns of intron gain and loss might be different among paralogous groups (79) [although see (80)]. If so, orthology inferences could be hampered by the higher rate of intron change. On the other hand, increased rates of loss and gain could increase intron positions' usefulness in identifying orthology even over short evolutionary times (with little sequence differentiation).

Estimation of neutral rates of nucleotide substitution

Intron sequences themselves appear to tolerate sequence changes quite easily. Putatively neutrally evolving (portions of) intron sequences are thus a key tool in estimating neutral rates of mutation. Hoffman and Birney (81) recently published a new method to estimate neutral rates of nucleotide substitution based on the study of the substitutions occurring on the alignable introns sequences. They compared their method to a previous method also based on intron sequences (82) and more classic methods based on substitutions in synonymous coding sites, finding a strong correlation between estimates from the two types of methods on different species comparisons. Interestingly, synonymous sites have been shown to be under purifying selection [reviewed in (83,84)], or even under positive selection (85–87). However, introns are also known to contain different types of functional elements (88–90) and thus selection of regions of estimation of neutral rates requires caution.

INTRONS IN PHYLOGENETIC ANALYSIS

Intron sequences and positions contain a record of the evolutionary history of a species or group of species, which presumably contains valuable phylogenetic information. Two phylogenetic strategies have utilized introns as phylogenetic markers at two very different evolutionary depths. Intron sequences, which are relatively fast evolving, have been commonly used to resolve relationships between closely related species. At the other end, evolution of gene structures by intron loss and gain, which can be very slow in some lineages, has been used in order to resolve deeper nodes over which our confidence in traditional sequence methods can be reduced by the large amount of change over the studied species.

Intron sequences as phylogenetic markers

The use of intron sequences for resolution of relationships between closely related species was established more than 15 years ago (91,92), and it is so common as to barely require comment (93–98). The appeal of intronic sequences for phylogenetics of recent divergences owes to their plausibly being both more rapidly and more neutrally evolving than protein coding and other clearly functional sequences (Figure 5A). As such, relatively simple phylogenetic methods are thought to be of use in utilizing intronic sequences over depths where multiple mutation is unlikely, and over which protein sequences may have experienced too little change to yield sufficient signal. Moreover, obtaining informative intron sequence data sets from non-model organisms is relatively easy and fast using PCR-based methods (91), and the global lack of functional constraints and high potential phylogenetic information content make introns a good complement to mtDNA-based phylogenies for poorly studied groups (99–101).

Intron positions as phylogenetic markers

As mentioned above, intron presence/absence is a relatively very slowly evolving character in most lineages studied to date. Whereas the average number of changes per nucleotide site in putatively unconstrained nucleotide sequence between mouse and human is estimated to be $K_s = 0.6$ (102), and the degree of protein sequence change around 21.5% (102), a survey of more than 150 000 intron positions between the species found only 120 changes (0.08%), a degree of change three orders of magnitude lower (42). Stajich and Dietrich (47) found <1% intron loss/gain change across a clade of four *Cryptococcus* species, compared to 35% change in nucleotide sites across the same species (47). Median dS is estimated to be around 0.49 for the *Plasmodium* parasites *P. falciparum* and *P. yoelii* (103), but fewer than 1.5% of intron sites have experienced a loss/gain event (45).

Relative to sequence-based studies, phylogenetics using intron presence/absence is truly in its infancy. Among the very few published studies, there are essentially three strategies. First, some authors have used one or a few intron loss/gain patterns to group species into a clade (78,104,105) (Figure 5B), since as rare genome change (RGC) losses/gains could be theoretically highly parsimonious (106). As RGCs, introns have the advantage of having a very wide taxonomic resolution, potentially low homoplasy and applicability to a broad range of eukaryotic groups (106). Here, however, significant caution is necessary. First, while such 'magic bullet' approaches may work well for groups with very few intron changes, the possibility of homoplasy (in particular, of multiple loss of the same intron) is likely in cases where there is more change, for instance, nematodes and dipterans (107); second, individual cases attest to the recurrent loss of the same intron, even while flanking introns remain intact (107,108). Until the degree of such variation across sites is better known, individual cases of intron loss/gain as phylogenetic markers should in our opinion be viewed as non-conclusive in many cases.

A second strategy involves using explicit phylogenetic models to analyze large collections of intron presence/absence data across species. One of us and Walter Gilbert (109) used intron presence/absence data for 684 sets of eukaryotic orthologs across eight animal, fungus, plant, and apicomplexan species (4) to develop a method for resolving deep nodes in metazoan phylogeny. Nguyen and coauthors (38) then developed a more general phylogenetic method for analysis of the same kind of data. These data were used to address the relationship between arthropods, nematodes and deuterostomes, with both analyses placing deuterostomes as the outgroup.

Again, there is reason for caution here. Both analyses assumed constancy of intron loss rates across intron sites, an assumption which is unlikely to hold generally, and which might bias methods towards long branch attraction and other recurrent problems in phylogenetics, especially important in this particular phylogeny (110–114). Moreover, the small absolute number of characters obtainable from even such an exhaustive comparative genomic effort may make it difficult to obtain sufficient knowledge of the shape of the distribution of rates across sites, which may make these shortcomings difficult to correct for. On the other hand, the availability of representatives of the groups in question that have experienced less intron loss will presumably allow for more confident reconstruction.

The third and perhaps most promising (as well as intriguing) strategy was developed by Krauss and coauthors (115), and aims to overcome these problems by restricting the analysis to cases in which ancestral and derived states can be more confidently inferred. Very short exons (e.g. shorter than 50 bp) are very rare across most characterized species, likely due to problems associated with accurate splicing of such regions. Therefore, introns found at nearby positions in orthologous coding regions are unlikely to have coexisted. Assuming that multiple insertions into the same site are very rare, one can then confidently infer that there is an edge on the phylogenetic tree (along which both intron loss and gain has occurred) separating those species that share one of the positions from those that share the other (Figure 5C). In cases in which a known outgroup shares one of the two positions, one can furthermore infer the directionality of this change, and therefore place all derived species into a clade. However, it remains to be seen whether (and in what cases) sufficient numbers of such intron pairs will be obtainable, thus it is not yet clear to what extent this model will be generalizable.

A fourth possibility that has not to our knowledge been explored would utilize shifts of intron boundaries (i.e. shortening or lengthening of adjacent coding sequence) (Figure 5D). Analysis of large numbers of alignments in various lineages shows that such changes are very rare indeed, and that such changes might represent useful

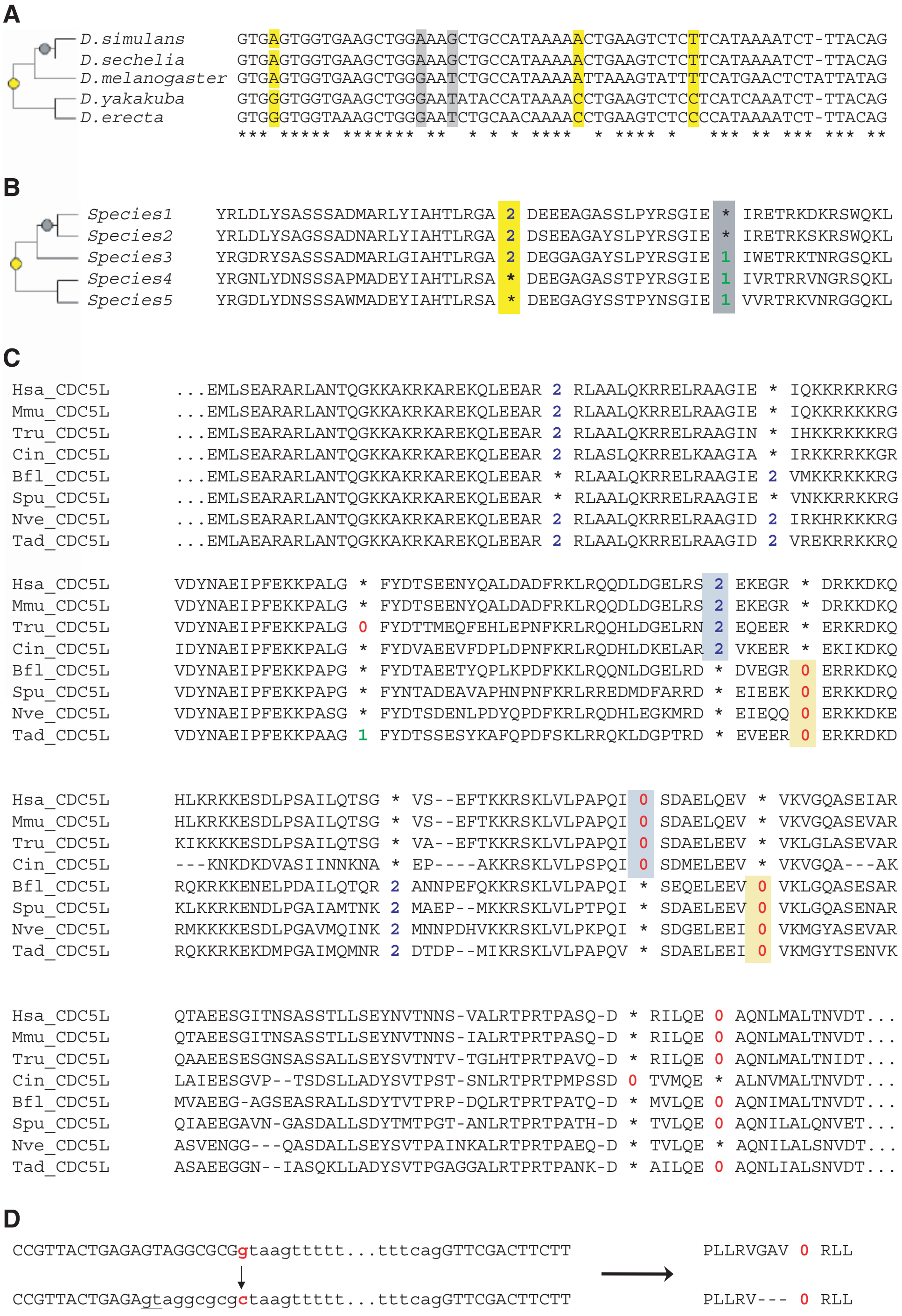

A
D.simulans GTGAGTGGTGAAGCTGGAAAGCTGCCATAAAAACTGAAGTCTCTTCATAAAATCT-TTACAG
D.sechelia GTGAGTGGTGAAGCTGGAAAGCTGCCATAAAAACTGAAGTCTCTTCATAAAATCT-TTACAG
D.melanogaster GTGAGTGGTGAAGCTGGGAATCTGCCATAAAAATTAAAGTATTTTCATGAACTCTATTATAG
D.yakakuba GTGGGTGGTGAAGCTGGGAATATACCATAAAACCTGAAGTCTCCTCATCAAATCT-TTACAG
D.erecta GTGGGTGGTAAAGCTGGGAATCTGCAACAAAACCTGAAGTCTCCCCATAAAATCT-TTACAG
*** ***** ******* ** * * * **** * **** * *** ** *** *** **
B
Species1 YRLDLYSASSSADMARLYIAHTLRGA 2 DEEEAGASSLPYRSGIE * IRETRKDKRSWQKL
Species2 YRLDLYSAGSSADNARLYIAHTLRGA 2 DSEEAGAYSLPYRSGIE * IRETRKSKRSWQKL
Species3 YRGDRYSASSSADMARLGIAHTLRGA 2 DEGGAGAYSLPYRSGIE 1 IWETRKTNRGSQKL
Species4 YRGNLYDNSSSAPMADEYIAHTLRSA * DEEGAGASSTPYRSGIE 1 IVRTRRVNGRSQKL
Species5 YRGDLYDNSSSAWMADEYIAHTLRSA * DEEGAGYSSTPYNSGIE 1 VVRTRKVNRGGQKL
C
Hsa_CDC5L ...EMLSEARARLANTQGKKAKRKAREKQLEEAR 2 RLAALQKRRELRAAGIE * IQKKRKRKRG
Mmu_CDC5L ...EMLSEARARLANTQGKKAKRKAREKQLEEAR 2 RLAALQKRRELRAAGIE * IQKKRKKKRG
Tru_CDC5L ...EMLSEARARLANTQGKKAKRKAREKQLEEAR 2 RLAALQKRRELRAAGIN * IHKKRKKKRG
Cin_CDC5L ...EMLSEARARLANTQGKKAKRKAREKQLEEAR 2 RLASLQKRRELKAAGIA * IRKKRRKKGR
Bfl_CDC5L ...EMLSEARARLANTQGKKAKRKAREKQLEEAR * RLAALQKRRELRAAGIE 2 VMKKRKKKRG
Spu_CDC5L ...EMLSEARARLANTQGKKAKRKAREKQLEEAR * RLAALQKRRELRAAGIE * VNKKRRKKRG
Nve_CDC5L ...EMLSEARARLANTQGKKAKRKAREKQLEEAR 2 RLAALQKRRELRAAGID 2 IRKHRKKKRG
Tad_CDC5L ...EMLAEARARLANTQGKKAKRKAREKQLEEAR 2 RLAALQKRRELRAAGID 2 VREKRRKKRQ
Hsa_CDC5L VDYNAEIPFEKKPALG * FYDTSEENYQALDADFRKLRQQDLDGELRS 2 EKEGR * DRKKDKQ
Mmu_CDC5L VDYNAEIPFEKKPALG * FYDTSEENYQALDADFRKLRQQDLDGELRS 2 EKEGR * DRKKDKQ
Tru_CDC5L VDYNAEIPFEKKPALG 0 FYDTTMEQFEHLEPNFKRLRQQHLDGELRN 2 EQEER * ERKRDKQ
Cin_CDC5L IDYNAEIPFEKKPALG * FYDVAEEVFDPLDPNFKRLRQDHLDKELAR 2 VKEER * EKIKDKQ
Bfl_CDC5L VDYNAEIPFEKKPAPG * FYDTAEETYQPLKPDFKKLRQQNLDGELRD * DVEGR 0 ERRKDKQ
Spu_CDC5L VDYNAEIPFEKKPAPG * FYNTADEAVAPHNPNFKRLRREDMDFARRD * EIEEK 0 ERKKDRQ
Nve_CDC5L VDYNAEIPFEKKPASG * FYDTSDENLPDYQPDFKRLRQDHLEGKMRD * EIEQQ 0 ERKKDKE
Tad_CDC5L VDYNAEIPFEKKPAAG 1 FYDTSSESYKAFQPDFSKLRRQKLDGPTRD * EVEER 0 ERKRDKD
Hsa_CDC5L HLKRKKESDLPSAILQTSG * VS--EFTKKRSKLVLPAPQI 0 SDAELQEV * VKVGQASEIAR
Mmu_CDC5L HLKRKKESDLPSAILQTSG * VS--EFTKKRSKLVLPAPQI 0 SDAELQEV * VKVGQASEVAR
Tru_CDC5L KIKKKKESDLPSAILQTSG * VA--EFTKKRSKLVLPAPQI 0 SDAELEEV * VKLGLASEVAR
Cin_CDC5L ---KNKDKDVASIINNKNA * EP----AKKRSKLVLPSPQI 0 SDMELEEV * VKVGQA---AK
Bfl_CDC5L RQKRKKENELPDAILQTQR 2 ANNPEFQKKRSKLVLPAPQI * SEQELEEV 0 VKLGQASESAR
Spu_CDC5L KLKKRKENDLPGAIAMTNK 2 MAEP--MKKRSKLVLPTPQI * SDAELEEV 0 VKLGQASENAR
Nve_CDC5L RMKKKKESDLPGAVMQINK 2 MNNPDHVKKRSKLVLPKPQI * SDGELEEI 0 VKMGYASEVAR
Tad_CDC5L RQKKRKEKDMPGAIMQMNR 2 DTDP--MIKRSKLVLPAPQV * SDAELEEI 0 VKMGYTSENVK
Hsa_CDC5L QTAEESGITNSASSTLLSEYNVTNNS-VALRTPRTPASQ-D * RILQE 0 AQNLMALTNVDT...
Mmu_CDC5L QTAEESGITNSASSTLLSEYNVTNNS-IALRTPRTPASQ-D * RILQE 0 AQNLMALTNVDT...
Tru_CDC5L QAAEESESGNSASSALLSEYSVTNTV-TGLHTPRTPAVQ-D * RILQE 0 AQNLMALTNIDT...
Cin_CDC5L LAIEESGVP--TSDSLLADYSVTPST-SNLRTPRTPMPSSD 0 TVMQE * ALNVMALTNVDT...
Bfl_CDC5L MVAEEG-AGSEASRALLSDYTVTPRP-DQLRTPRTPATQ-D * MVLQE 0 AQNIMALTNVDT...
Spu_CDC5L QIAEEGAVN-GASDALLSDYTMTPGT-ANLRTPRTPATH-D * TVLQE 0 AQNILALQNVET...
Nve_CDC5L ASVENGG---QASDALLSEYSVTPAINKALRTPRTPAEQ-D * TVLQE * AQNILALSNVDT...
Tad_CDC5L ASAEEGGN--IASQKLLADYSVTPGAGGALRTPRTPANK-D * AILQE 0 AQNLIALSNVDT...
D
CCGTTACTGAGAGTAGGCGCGgtaagttttt...tttcagGTTCGACTTCTT
CCGTTACTGAGAgtaggcgcgctaagttttt...tttcagGTTCGACTTCTT
PLLRVGAV 0 RLL
PLLRV--- 0 RLL

Figure 5. Introns as tools for phylogenetics. (**A**) Intron sequences themselves are often used to resolve phylogenetics relationships between closely related species. Here, four phylogenetically informative sites in an alignment of sequences of orthologous introns from five *Drosophila* species (the 1st intron from the *CG10050* locus) are consistent with the accepted consensus relationship of the species (left). (**B**) Intron loss/gain and position conservation as phylogenetic characters. Intron position conservation across subsets of species suggests phylogenetic groupings. Under many circumstances, the high degree of intron loss/gain change requires correction for the possibility of multiple changes at a site (3). (**C**) Nearby pairs of intron positions as phylogenetic characters. Krauss *et al.* (115) suggest a novel class of intron change in order to diminish the possibility of homoplasy. Since short exons are rare across various genomes, introns at nearby positions in orthologous genes are unlikely to have coexisted, suggesting multiple rare changes (an intron loss followed by a nearby gain) between putatively ancestral and derived gene structures. Here, two pairs of nearby intron positions in the *CDC5-like* gene of metazoans support grouping the three vertebrates (top three lines) with *Ciona intestinalis* (Cin_CDC5L). (**D**) Intron boundary sliding as a possible phylogenetic character. Rarely, an intron boundary may shift, leading to conversion of exonic sequence to intron (or vice versa). In this hypothetical case, mutation of the 5′ splice boundary (gt→ct) leads to use of an upstream previously exonic GT site. To our knowledge, no study has yet employed such changes for phylogenetic analysis. Abbreviations: Hsa, *Homo sapiens*; Mmu, *Mus musculus*; Tru, *Takifugu rubripes*; Cin, *Ciona intestinalis*; Bfl, *Branchiostoma floridae*; Spu, *Strongylocentrotus purpuratus*; Nve, *Nematostella vectensis*; Tad, *Trichoplax adhaerens*.

genomic changes. A major concern in developing methods using these changes will be exclusion of incorrect genome annotation as an explanation.

An important caveat to the seeming usefulness of intron loss/gain across species is the possibility of highly skewed distributions of rates across sites (alluded to above). In a few cases, careful study of closely related species with known phylogenies has indicated that a single intron position has been subject to striking recurrent intron loss while other intron positions have remained intact (107,108). In such cases, the observed phylogenetic position will give support to inaccurate phylogenetic groups. It is not currently known how general this pattern of recurrent loss of the same introns is, and so it is not yet clear how much of a problem this may constitute in large-scale or genome-level comparisons. Strong confidence in use of intron losses/gains as phylogenetic characters awaits a better understanding of the causes and generality of large differences in loss rates across sites.

CONCLUSION

Problems associated with signal-to-noise ratios are ubiquitous in bioinformatic, genomic and evolutionary analyses. As slowly evolving characters, intron positions provide useful and otherwise scarce information. We have reviewed well-developed, early-stage and potential uses of introns as tools for addressing a wide range of problems. We look forward to development of further methods.

ACKNOWLEDGEMENTS

M.I. was funded by the Spanish Ministerio of Educación y Ciencia, through the FPI grant (BFU2005-00252), and S.W.R. by the Intramural Research Program of the National Library of Medicine at National Institutes of Health/DHHS. Funding to pay the Open Access publication charges for this article has been provided by the National Institutes of Health. We thank Eugene Koonin and Jordi Garcia-Fernandez and their groups for intellectual support and stimulation, for financial support, and for fostering environments of open intellectual exploration in their respective groups. SWR thanks Rogério da Silva and all of Marcelo Ferreira's group for heroics of hospitality and friendship during the course of this project.

Conflict of interest statement. None declared.

REFERENCES

1. Nilsen,T.W. (2003) The spliceosome: the most complex macromolecular machine in the cell? *Bioessays*, **25**, 1147–1149.
2. Roy,S.W. and Penny,D. (2006) Large-scale intron conservation and order-of-magnitude variation in intron loss/gain rates in apicomplexan evolution. *Genome Res.*, **16**, 1270–1275.
3. Roy,S.W. and Penny,D. (2007) A very high fraction of unique intron positions in the intron-rich diatom *Thalassiosira pseudonana* indicates widespread intron gain. *Mol. Biol. Evol.*, **24**, 1447–1457.
4. Rogozin,I.B., Wolf,Y.I., Sorokin,A.V., Mirkin,B.G. and Koonin,E.V. (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.*, **13**, 1512–1517.
5. Carmel,L., Wolf,Y.I., Rogozin,I.B. and Koonin,E.V. (2007) Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res.*, **17**, 1034–1044.
6. Nielsen,C., Friedman,B., Birren,B., Burge,C. and Galagan,J. (2004) Patterns of intron gain and loss in fungi. *PLoS Biol.*, **2**, e422.
7. Roy,S.W. and Gilbert,W. (2005) The pattern of intron loss. *Proc. Natl Acad. Sci. USA*, **102**, 713–718.
8. Roy,S.W., Irimia,M. and Penny,D. (2006) Very little intron gain in *Entamoeba histolytica* genes laterally transferred from prokaryotes. *Mol. Biol. Evol.*, **23**, 1824–1827.
9. Kupfer,D.M., Drabenstot,S.D., Buchanan,K.L., Lai,H., Zhu,H., Dyer,D.W., Roe,D.A. and Murphy,J.W. (2004) Introns and splicing elements of five diverse fungi. *Eukaryotic Cell*, **3**, 1088–1100.
10. Bon,E., Casaregola,S., Blandin,G., Llorente,B., Neuveglise,C., Munsterkotter,M., Guldener,U., Mewes,H.-W., Helden,J.V., Dujon,B. *et al.* (2003) Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. *Nucleic Acids Res.*, **31**, 1121–1135.
11. Irimia,M., Penny,D. and Roy,S.W. (2007) Coevolution of genomic intron number and splice sites. *Trends Genet.*, **23**, 321–325.
12. Irimia,M., Rukov,J.L., Penny,D. and Roy,S.W. (2007) Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing. *BMC Evol. Biol.*, **7**, 188.
13. Schwartz,S., Silva,J., Burstein,D., Pupko,T., Eyras,E. and Ast,G. (2008) Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res.*, **18**, 88–103.
14. Rogozin,I., Sverdlov,A., Babenko,V. and Koonin,E. (2005) Analysis of evolution of exon-intron structure of eukaryotic genes. *Brief Bioinform.*, **6**, 118–134.
15. Roy,S.W. and Gilbert,W. (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat. Rev. Genet.*, **7**, 211–221.
16. Rodriguez-Trelles,F., Tarrio,R. and Ayala,F.J. (2006) Origins and evolution of spliceosomal introns. *Annu Rev. Genet.*, **40**, 47–76.
17. Fedorova,L. and Fedorov,A. (2003) Introns in gene evolution. *Genetica*, **118**, 123–131.
18. Lynch,M. and Richardson,A.O. (2002) The evolution of spliceosomal introns. *Curr. Opin. Genet. Dev.*, **12**, 701–710.
19. Zhaxybayeva,O. and Gogarten,J.P. (2003) Spliceosomal introns: New insights into their evolution. *Curr. Biol.*, **13**, R764–R766.
20. Berget,S.M. and Sharp,P.A. (1977) A spliced sequence at the 5′-terminus of adenovirus late mRNA. *Brookhaven Symp. Biol.*, **12–20**, 332–344.
21. Doolittle,W.F. (1978) Genes in pieces: were they ever together? *Nature*, **272**, 581–582.
22. Gilbert,W. (1987) The exon theory of genes. *Cold Spring Harb. Sym.*, **52**, 901–905.
23. de Souza,S.J., Long,M., Schoenbach,L., Roy,S.W. and Gilbert,W. (1996) Intron positions correlate with module boundaries in ancient proteins. *Proc. Natl Acad. Sci. USA*, **93**, 14632–14636.
24. Long,M., de Souza,S.J. and Gilbert,W. (1995) Evolution of the intron-exon structure of eukaryotic genes. *Curr. Opin. Genet. Dev.*, **5**, 774–778.
25. Darnell,J. Jr. (1978) Implications of RNA-RNA splicing in evolution of eukaryotic cells. *Science*, **202**, 1257–1260.
26. Cavalier-Smith,T. (1985) Selfish DNA and the origin of introns. *Nature*, **315**, 283–284.
27. Cavalier-Smith,T. (1991) Intron phylogeny: a new hypothesis. *Trends Genet.*, **7**, 145–148.
28. Stoltzfus,A. (1994) Origin of introns-early or late? *Nature*, **369**, 526–527.
29. Logsdon,J. (1998) The recent origins of spliceosomal introns revisited. *Curr. Opin. Genet. Dev.*, **8**, 637–648.
30. Logsdon,J. Jr, Tyshenko,M., Dixon,C., D.-Jafari,J., Walker,V. and Palmer,J. (1995) Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. *Proc. Natl Acad. Sci. USA*, **92**, 8507–8511.
31. Dibb,N.J. and Newman,A.J. (1989) Evidence that introns arose at proto-splice sites. *EMBO J.*, **8**, 2015–2021.
32. Cech,T.R. (1986) The generality of self-splicing RNA: Relationship to nuclear mRNA splicing. *Cell*, **44**, 207–210.

33. Sharp,P.A. (1985) On the origin of RNA splicing and introns. *Cell*, **42**, 397–400.
34. Fedorov,A., Merican,A. and Gilbert,W. (2002) Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc. Natl Acad. Sci. USA*, **99**, 16128–16133.
35. Archibald,J., O'Kelly,C. and Doolittle,W. (2002) The chaperonin genes of jakobid and jakobid-like flagellates: implications for eukaryotic evolution. *Mol. Biol. Evol.*, **19**, 422–431.
36. Roy,S.W. and Gilbert,W. (2005) Complex early genes. *Proc. Natl Acad. Sci. USA*, **102**, 1986–1991.
37. Sverdlov,A., Rogozin,I., Babenko,V. and Koonin,E. (2005) Conservation versus parallel gains in intron evolution. *Nucleic Acids Res.*, **33**, 1741–1748.
38. Nguyen,H., Yoshihama,M. and Kenmochi,N. (2005) New maximum likelihood estimators for eukaryotic intron evolution. *PLoS Comput. Biol.*, **1**, e79.
39. Yoshihama,M., Nakao,A., Nguyen,H.D. and Kenmochi,N. (2006) Analysis of ribosomal protein gene structures: implications for intron evolution. *PLoS Genet.*, **2**, e25.
40. Slamovits,C.H. and Keeling,P.J. (2006) A high density of ancient spliceosomal introns in oxymonad excavates. *BMC Evol. Biol.*, **6**, 34.
41. Csurös,M. (2005), *Third RECOMB Satellite Workshop on Comparative Genomics*. Springer LNCS 3678, pp. 47–60.
42. Coulombe-Huntington,J. and Majewski,J. (2007) Characterization of intron loss events in mammals. *Genome Res.*, **17**, 23–32.
43. Fedorov,A., Roy,S., Fedorova,L. and Gilbert,W. (2003) Mystery of intron gain. *Genome Res.*, **13**, 2236–2241.
44. Roy,S.W. and Gilbert,W. (2005) Rates of intron loss and gain: implications for early eukaryotic evolution. *Proc. Natl Acad. Sci. USA*, **102**, 5773–5778.
45. Roy,S.W. and Hartl,D.L. (2006) Very little intron loss/gain in *Plasmodium*: intron loss/gain mutation rates and intron number. *Genome Res.*, **16**, 750–756.
46. Roy,S.W. and Penny,D. (2007) Patterns of intron loss and gain in plants: intron loss-dominated evolution and genome-wide comparison of *O. sativa* and *A. thaliana*. *Mol. Biol. Evol.*, **24**, 171–181.
47. Stajich,J.E. and Dietrich,F.S. (2006) Evidence of mRNA-mediated intron loss in the human-pathogenic fungus *Cryptococcus neoformans*. *Eukaryotic Cell*, **5**, 789–793.
48. Carmel,L., Rogozin,I.B., Wolf,Y.I. and Koonin,E.V. (2007) Evolutionarily conserved genes preferentially accumulate introns. *Genome Res.*, **17**, 1045–1050.
49. Rogozin,I.B., Lyons-Weiler,J. and Koonin,E.V. (2000) Intron sliding in conserved gene families. *Trends Genet.*, **16**, 430–432.
50. Sakharkar,M.K., Tan,T.W. and de Souza,S.J. (2001) Generation of a database containing discordant intron positions in eukaryotic genes (MIDB). *Bioinformatics*, **17**, 671–675.
51. Sverdlov,A., Babenko,V., Rogozin,I. and Koonin,E. (2004) Preferential loss and gain of introns in 3′ portions of genes suggests a reverse-transcription mechanism of intron insertion. *Gene*, **338**, 85–91.
52. Mourier,T. and Jeffares,D.C. (2003) Eukaryotic intron loss. *Science*, **300**, 1393.
53. Lin,K. and Zhang,D.-Y. (2005) The excess of 5′ introns in eukaryotic genomes. *Nucleic Acids Res.*, **33**, 6522–6527.
54. Niu,D.-K., Hou,W.-R. and Li,S.-W. (2005) mRNA-Mediated intron losses: evidence from extraordinarily large exons. *Mol. Biol. Evol.*, **22**, 1475–1481.
55. Boeke,J.D., Garfinkel,D.J., Styles,C.A. and Fink,G.R. (1985) Ty elements transpose through an RNA intermediate. *Cell*, **40**, 491–500.
56. Fink,G. (1987) Pseudogenes in yeast? *Cell*, **49**, 5–6.
57. Raible,F., Tessmar-Raible,K., Osoegawa,K., Wincker,P., Jubin,C., Balavoine,G., Ferrier,D., Benes,V., de Jong,P., Weissenbach,J. *et al.* (2005) Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*. *Science*, **310**, 1325–1326.
58. Roy,S.W. and Penny,D. (2007) Intron length distributions and gene prediction. *Nucleic Acids Res.*, **35**, 4737–4742.
59. Siegel,N., Hoegg,S., Salzburger,W., Braasch,I. and Meyer,A. (2007) Comparative genomics of ParaHox clusters of teleost fishes: gene cluster breakup and the retention of gene sets following whole genome duplications. *BMC Genomics*, **8**, 312.
60. Louis,A., Ollivier,E., Aude,J.-C. and Risler,J.-L. (2001) Massive sequence comparisons as a help in annotating genomic sequences. *Genome Res.*, **11**, 1296–1303.
61. Roy,S., Fedorov,A. and Gilbert,W. (2003) Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl Acad. Sci. USA*, **100**, 7158–7162.
62. Coghlan,A. and Durbin,R. (2007) Genomix: a method for combining gene-finders' predictions, which uses evolutionary conservation of sequence and intron-exon structure. *Bioinformatics*, **23**, 1468–1475.
63. Csuros,M., Holey,J.A. and Rogozin,I.B. (2007) In search of lost introns. *Bioinformatics*, **23**, i87–i96.
64. Dacks,J.B., Marinets,A., Ford Doolittle,W., Cavalier-Smith,T. and Logsdon,J.M. Jr. (2002) Analyses of RNA Polymerase II genes from free-living protists: phylogeny, long branch attraction, and the eukaryotic big bang. *Mol. Biol. Evol.*, **19**, 830–840.
65. Vanin,E.F. (1985) Processed pseudogenes: characteristics and evolution. *Annual Rev. Genet.*, **19**, 253–272.
66. Harrison,P.M. and Gerstein,M. (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J. Mol. Biol.*, **318**, 1155–1174.
67. Zhang,Z., Carriero,N. and Gerstein,M. (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet.*, **20**, 62–67.
68. D'Errico,I., Gadaleta,G. and Saccone,C. (2004) Pseudogenes in metazoa: origin and features. *Brief. Funct. Genom. Proteom.*, **3**, 157–167.
69. Benito-Gutierrez,E., Nake,C., Llovera,M., Comella,J.X. and Garcia-Fernandez,J. (2005) The single AmphiTrk receptor highlights increased complexity of neurotrophin signalling in vertebrates and suggests an early role in developing sensory neuroepidermal cells. *Development*, **132**, 2191–2202.
70. Kaessmann,H., Zollner,S., Nekrutenko,A. and Li,W.-H. (2002) Signatures of domain shuffling in the human genome. *Genome Res.*, **12**, 1642–1650.
71. Vibranovski,M., Sakabe,N., Oliveira,R. and Souza,S. (2005) Signs of ancient and modern exon-shuffling are correlated to the distribution of ancient and modern domains along proteins. *J. Mol. Evol.*, **61**, 341–350.
72. Patthy,L. (1999) Genome evolution and the evolution of exon-shuffling — a review. *Gene*, **238**, 103–114.
73. Patthy,L. (2003) Modular assembly of genes and the evolution of new functions. *Genetica*, **118**, 217–231.
74. Roy,S.W. and Penny,D. (2007) Widespread intron loss suggests retrotransposon activity in ancient apicomplexans. *Mol. Biol. Evol.*, **24**, 1926–1933.
75. Ferrier,D.E.K., Minguillon,C., Holland,P.W.H. and Garcia-Fernandez,J. (2000) The amphioxus Hox cluster: deuterostome posterior flexibility and Hox14. *Evol. Dev.*, **2**, 284–293.
76. Endo,Y., Liu,Y., Kanno,K., Takahashi,M., Matsushita,M. and Fujita,T. (2004) Identification of the mouse H-ficolin gene as a pseudogene and orthology between mouse ficolins A/B and human L-/M-ficolins. *Genomics*, **84**, 737–744.
77. Franck,E., Madsen,O., van Rheede,T., Ricard,G.N., Huynen,M.A. and de Jong,W.W. (2004) Evolutionary diversity of vertebrate small heat shock proteins. *J. Mol. Evol.*, **59**, 792–805.
78. Jordal,B.H. (2002) Elongation factor 1 alpha resolves the monophyly of the haplodiploid ambrosia beetles *Xyleborini* (Coleoptera: Curculionidae). *Insect Mol. Biol.*, **11**, 453–465.
79. Babenko,V., Rogozin,I., Mekhedov,S. and Koonin,E. (2004) Prevalence of intron gain over intron loss in the evolution of paralogous gene families. *Nucleic Acids Res.*, **32**, 3724–3733.
80. Roy,S.W. and Penny,D. (2007) On the incidence of intron loss and gain in paralogous gene families. *Mol. Biol. Evol.*, **24**, 1579–1581.
81. Hoffman,M.M. and Birney,E. (2007) Estimating the neutral rate of nucleotide substitution using introns. *Mol. Biol. Evol.*, **24**, 522–531.
82. Castresana,J. (2002) Estimation of genetic distances from human and mouse introns. *Genome Biol.*, **3**, research0028.0021–research 0028.0027.
83. Chamary,J.V., Parmley,J.L. and Hurst,L.D. (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.*, **7**, 98–108.

84. Xing,Y. and Lee,C. (2005) Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc. Natl Acad. Sci. USA*, **102**, 13526–13531.
85. Resch,A.M., Carmel,L., Marino-Ramirez,L., Ogurtsov,A.Y., Shabalina,S.A., Rogozin,I.B. and Koonin,E.V. (2007) Widespread positive selection in synonymous sites of mammalian genes. *Mol. Biol. Evol.*, **24**, 1821–1831.
86. Nielsen,R., Bauer DuMont,V.L., Hubisz,M.J. and Aquadro,C.F. (2007) Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. *Mol. Biol. Evol.*, **24**, 228–235.
87. Neafsey,D. and Galagan,J. (2007) Positive selection for unpreferred codon usage in eukaryotic genomes. *BMC Evol. Biol.*, **7**, 119.
88. Yeo,G.W., Nostrand,E.L.V. and Liang,T.Y. (2007) Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS Genet.*, **3**, e85.
89. Epstein,D.J., McMahon,A.P. and Joyner,A.L. (1999) Regionalization of sonic hedgehog transcription along the anteroposterior axis of the mouse central nervous system is regulated by Hnf3-dependent and -independent mechanisms. *Development*, **126**, 281–292.
90. Kabat,J.L., Barberan-Soler,S., McKenna,P., Clawson,H., Farrer,T. and Zahler,A.M. (2006) Intronic alternative splicing regulators identified by comparative genomics in nematodes. *PLoS Comp. Biol.*, **2**, e86.
91. Lessa,E. (1992) Rapid surveying of DNA sequence variation in natural populations. *Mol. Biol. Evol.*, **9**, 323–330.
92. Slade,R.W., Moritz,C., Heideman,A. and Hale,P.T. (1993) Rapid assessment of single-copy nuclear DNA variation in diverse species. *Mol. Ecol.*, **2**, 359–373.
93. Pecon-Slattery,J., Pearks Wilkerson,A.J., Murphy,W.J. and O'Brien,S.J. (2004) Phylogenetic assessment of introns and SINEs within the Y chromosome using the cat family Felidae as a species tree. *Mol. Biol. Evol.*, **21**, 2299–2309.
94. Eick,G.N., Jacobs,D.S. and Matthee,C.A. (2005) A nuclear DNA phylogenetic perspective on the evolution of echolocation and historical biogeography of extant bats (Chiroptera). *Mol. Biol. Evol.*, **22**, 1869–1886.
95. Willows-Munro,S., Robinson,T.J. and Matthee,C.A. (2005) Utility of nuclear DNA intron markers at lower taxonomic levels: phylogenetic resolution among nine *Tragelaphus* spp. *Mol. Phylogenet. Evol.*, **35**, 624–636.
96. Creer,S., Pook,C.E., Malhotra,A. and Thorpe,R.S. (2006) Optimal intron analyses in the *Trimeresurus* radiation of asian pitvipers. *Syst. Biol.*, **55**, 57–72.
97. Matthee,C.A., Eick,G., Willows-Munro,S., Montgelard,C., Pardini,A.T. and Robinson,T.J. (2007) Indel evolution of mammalian introns and the utility of non-coding nuclear markers in eutherian phylogenetics. *Mol. Phylogenet. Evol.*, **42**, 827–837.
98. Matocq,M.D., Shurtliff,Q.R. and Feldman,C.R. (2007) Phylogenetics of the woodrat genus *Neotoma* (Rodentia: Muridae): implications for the evolution of phenotypic variation in male external genitalia. *Mol. Phylogenet. Evol.*, **42**, 637–652.
99. Slade,R., Moritz,C. and Heideman,A. (1994) Multiple nuclear-gene phylogenies: application to pinnipeds and comparison with a mitochondrial DNA gene phylogeny. *Mol. Biol. Evol.*, **11**, 341–356.
100. Prychitko,T.M. and Moore,W.S. (2000) Comparative evolution of the mitochondrial cytochrome b gene and nuclear {beta}-fibrinogen intron 7 in woodpeckers. *Mol. Biol. Evol.*, **17**, 1101–1111.
101. Prychitko,T.M. and Moore,W.S. (2003) Alignment and phylogenetic analysis of {beta}-fibrinogen intron 7 sequences among avian orders reveal conserved regions within the intron. *Mol. Biol. Evol.*, **20**, 762–771.
102. Waterston,R., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M. and An,P. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
103. Hall,N., Karras,M., Raine,J.D., Carlton,J.M., Kooij,T.W.A., Berriman,M., Florens,L., Janssen,C.S., Pain,A., Christophides,G.K. *et al.* (2005) A Comprehensive survey of the plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. *Science*, **307**, 82–86.
104. Venkatesh,B., Ning,Y. and Brenner,S. (1999) Late changes in spliceosomal introns define clades in vertebrate evolution. *Proc. Natl Acad. Sci. USA*, **96**, 10267–10271.
105. Rokas,A., Kathirithamby,J. and Holland,P.W.H. (1999) Intron insertion as a phylogenetic character: the engrailed homeobox of Strepsiptera does not indicate affinity with Diptera. *Insect Mol. Biol.*, **8**, 527–530.
106. Rokas,A. and Holland,P.W.H. (2000) Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol.*, **15**, 454–459.
107. Kiontke,K., Gavin,N.P., Raynes,Y., Roehrig,C., Piano,F. and Fitch,D.H.A. (2004) *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proc. Natl Acad. Sci. USA*, **101**, 9003–9008.
108. Krzywinski,J. and Besansky,N.J. (2002) Frequent intron loss in the white gene: a cautionary tale for phylogeneticists. *Mol. Biol. Evol.*, **19**, 362–366.
109. Roy,S.W. and Gilbert,W. (2005) Resolution of a deep animal divergence by the pattern of intron conservation. *Proc. Natl Acad. Sci. USA*, **102**, 4403–4408.
110. Aguinaldo,A.M.A., Turbeville,J.M., Linford,L.S., Rivera,M.C., Garey,J.R., Raff,R.A. and Lake,J.A. (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, **387**, 489–493.
111. Dopazo,H. and Dopazo,J. (2005) Genome-scale evidence of the nematode-arthropod clade. *Genome Biol.*, **6**, R41.
112. Philippe,H., Lartillot,N. and Brinkmann,H. (2005) Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. *Mol. Biol. Evol.*, **22**, 1246–1253.
113. Delsuc,F., Brinkmann,H., Chourrout,D. and Philippe,H. (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, **439**, 965–968.
114. Irimia,M., Maeso,I., Penny,D., Garcia-Fernandez,J. and Roy,S.W. (2007) Rare coding sequence changes are consistent with ecdysozoa, not coelomata. *Mol. Biol. Evol.*, **24**, 1604–1607.
115. Krauss,V., Pecyna,M., Kurz,K. and Sass,H. (2005) Phylogenetic mapping of intron positions: a case study of translation initiation factor eIF2-gamma. *Mol. Biol. Evol.*, **22**, 74–84.