# The Role of Reverse Transcriptase in Intron Gain and Loss Mechanisms

Noa E. Cohen,[1,2] Roy Shen,[1] and Liran Carmel*[,1]

[1]Department of Genetics, The Alexander Silberman Institute of Life Sciences, Faculty of Science, The Hebrew University of Jerusalem, Jerusalem, Israel

[2]School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel

***Corresponding author:** E-mail: carmell@cc.huji.ac.il.

**Associate editor:** Aoife McLysaght

## Abstract

Intron density is highly variable across eukaryotic species. It seems that different lineages have experienced considerably different levels of intron gain and loss events, but the reasons for this are not well known. A large number of mechanisms for intron loss and gain have been suggested, and most of them have at least some level of indirect support. We therefore figured out that the variability in intron density can be a reflection of the fact that different mechanisms are active in different lineages. Quite a number of these putative mechanisms, both for intron loss and for intron gain, postulate that the enzyme reverse transcriptase (RT) has a key role in the process. In this paper, we lay out three predictions whose approval or falsification gives indication for the involvement of RT in intron gain and loss processes. Testing these predictions requires data on the intron gain and loss rates of individual genes along different branches of the eukaryotic phylogenetic tree. So far, such rates could not be computed, and hence, these predictions could not be rigorously evaluated. Here, we use a maximum likelihood algorithm that we have devised in the past, Evolutionary Reconstruction by Expectation Maximization, which allows the estimation of such rates. Using this algorithm, we computed the intron loss and gain rates of more than 300 genes in each branch of the phylogenetic tree of 19 eukaryotic species. Based on that we found only little support for RT activity in intron gain. In contrast, we suggest that RT-mediated intron loss is a mechanism that is very efficient in removing introns, and thus, its levels of activity may be a major determinant of intron number. Moreover, we found that intron gain and loss rates are negatively correlated in intron-poor species but are positively correlated for intron-rich species. One explanation to this is that intron gain and loss mechanisms in intron-rich species (like metazoans) share a common mechanistic component, albeit not a RT.

**Key words:** eukaryotic gene structure, intron evolution, intron gain rate, intron gain mechanism, intron loss rate, intron loss mechanism, reverse transcriptase, intron positional bias.

## Introduction

Eukaryotic genomes vary considerably in their intron occupancy. Humans, for example, have on average around eight introns per gene (Sakharkar et al. 2004). Some excavates, as a counter example, have only a few in their entire genome (Nixon et al. 2002; Simpson et al. 2002). This broad spectrum reflects a varied evolutionary history wherein different lineages experienced different forces that shaped the intron–exon structure of their genes. This evolutionary history is customarily described by looking at the rates by which introns are inserted (gained) or removed (lost) from genes (Stoltzfus et al. 1997; Rogozin et al. 2003).

Indeed, a number of comparative studies have shown that eukaryotes differ significantly in their intron gain and loss rates (Rogozin et al. 2003; Nguyen et al. 2005; Roy and Gilbert 2005; Carmel, Wolf, et al. 2007; Csuros et al. 2008). Some taxonomic groups exhibit almost stagnant gene architecture, whereas others underwent frequent changes. For example, although the gene structure of vertebrates and apicomplexans has changed very little during the last 100 million years (Babenko et al. 2004; Roy and Penny 2006), it has changed rapidly in nematodes (Banyai and Patthy 2004; Cho et al. 2004; Coghlan and Wolfe 2004) and in arthropods (Banyai and Patthy 2004; Li et al. 2009; Colbourne et al. 2011). In plants, high number of intron gain and loss events have been detected, but it remains to be seen at what evolutionary times they had occurred (Knowles and McLysaght 2006; Carmel, Wolf, et al. 2007; Roy and Penny 2007).

Revealing the biological mechanisms that underlie intron gain and loss may explain the varied evolution of gene structure. However, these mechanisms prove hard to pin down, mainly because intron sequences evolve fast, thus rapidly erase traces of their origin. This has led to a series of attempts to identify recent intron gain and loss events. Although some success have been recorded with detecting recent intron losses, very few unequivocal intron gain events were found (Babenko et al. 2004; Coghlan and Wolfe 2004; Knowles and McLysaght 2006; Coulombe-Huntington and Majewski 2007; Roy and Penny 2007; Li et al. 2009). The task is even more complicated as it is gradually recognized that intron gain and loss are driven by multiple mechanisms. So far, no single mechanism gained comprehensive support, but many have gained some support. Moreover, some mechanisms may be specific to certain evolutionary times, others may be specific to certain taxonomic groups, and others may act simultaneously within the same organism.

Three main mechanisms have been suggested to explain intron loss. Arguably, all have gained some indirect support. The reverse transcriptase (RT)–mediated intron loss model, probably the most popular of the three, suggests that processed, or semiprocessed, mRNA is reverse transcribed by RT, and the resulting cDNA integrates into the genome by homologous recombination (Fink 1987; Derr and Strathern 1993; Mourier and Jeffares 2003). The simple intron deletion model puts forward the idea that introns are lost by direct genomic deletion. An exact intronic deletion may be assured, for example, by the presence of short direct repeats at the intronic ends (Cho et al. 2004; Rodriguez-Trelles et al. 2006b). It is also conceivable that such exact (and also not exact) intron loss may be a result of nonhomologous end joining (NHEJ) repair of DNA double-strand breaks (Farlow et al. 2010). According to the exonization model, splice signal mutations impede intron recognition, leading to intron retention (Parma et al. 1987; Catania and Lynch 2008).

Intron gain mechanism has proved even more elusive than intron loss due to the scarcity of definite recent gains (Coulombe-Huntington and Majewski 2007). Consequently, it is more difficult to test models, and many have been so far suggested. The self-splicing intron origin model suggests that mobile group II self-splicing introns accumulate point mutations that turn them into spliceosomal introns (Rogers 1989; Cavalier-Smith 1991; Sharp 1991; Koonin 2006; Martin and Koonin 2006). The transposon model assumes that transposable elements can become introns and lose their mobility, either they are already equipped with the required splicing signals or they rapidly acquire them upon integration (Crick 1979; Purugganan and Wessler 1992; Roy 2004). According to the intronization model, mutations along the coding sequence create new splice sites, turning part of an exon into an intron (Sela et al. 2007; Catania and Lynch 2008; Irimia et al. 2008). The tandem duplication model argues that when an exonic region that contains AGGT or a similar motif is duplicated, 5′ and 3′ splice sites are automatically generated, forming a new intron (Rogers 1989; Lynch and Richardson 2002). The NHEJ-mediated intron gain model asserts that short DNA sequences of various origins are integrated as introns into nuclear DNA following NHEJ repair of double-strand breaks (Li et al. 2009; Farlow et al. 2010; Ragg 2011). The intron transposition model, perhaps the most widely accepted, states that excised introns may be integrated into nearby mRNAs by reverse splicing and then incorporate into the genome by reverse transcription followed by homologous recombination (Sharp 1985; Rodriguez-Trelles et al. 2006a; Roy and Irimia 2009). Interestingly, the intron transposition model requires RT—the same RT that is also believed to be associated with intron loss. The spliceosomal retrohoming (Roy and Irimia 2009), template switching (Roy and Irimia 2009), and RT-slippage (Sverdlov et al. 2004) models propose alternative mechanisms in which RT is involved in intron gain.

From the above survey, it is apparent that RT is conjectured to fulfill a central role in intron loss and might also affect intron gain. In order to test this, we formulated three predictions, some of which have already been suggested in the literature:

- In RT-mediated intron loss, the loss rate is expected to be higher near the 3′ end of the gene than near its 5′ end. This is due to the fact that the process of reverse transcription initiates at the 3′ end and because RT frequently disassociates from the template prematurely. Hence, the effect of RT is expected to decrease with the distance from the 3′ end of the RNA.
- In RT-mediated intron gain, the gain rate is expected to be higher near the 3′ end of the gene than near its 5′ end. This is for the same reason as in the previous item.
- If RT is involved in both intron loss and gain, a positive correlation between the respective rates is expected. If these rates are determined by selective forces but are otherwise unrelated, a negative correlation is expected (Roy and Irimia 2009). In contrast, a positive correlation is expected if the gain and loss processes share a common mechanistic component (Carmel, Wolf, et al. 2007; Roy and Irimia 2009).

In this paper we test the above hypotheses by reconstructing the evolutionary history of gene architecture and mapping intron gain and loss events on the eukaryotic phylogenetic tree. To this end, we use the Evolutionary Reconstruction by Expectation Maximization (EREM) maximum likelihood algorithm that we had previously developed (Carmel et al. 2005, 2007b, 2010) and apply it to 391 genes with full set of orthologs in 19 eukaryotic species (see "Methods").

The first hypothesis was already laid out by Fink (1987) to explain the tendency of the yeast introns to reside in the 5′ end of the genes. Since his work, a number of studies have analyzed intron positions along the gene, reporting similar 5′ positional bias in several cases, most notably in intron-poor species (Sakurai et al. 2002; Mourier and Jeffares 2003; Lin and Zhang 2005). Such 5′ positional bias can be explained either by an elevated intron loss rate at the 3′ end of the gene, as Fink suggested, or by an elevated intron gain rate at the 5′ end. Previous analyses did not take evolutionary perspective but rather analyzed each species separately, thus were unable to decide between the two explanations. Here, we use EREM to compute intron gain and loss rates at the opposing ends of each gene, thus computing—for each organism—whether there is a consistent difference in these rates between the 3′ and the 5′ ends of its genes. We find that the 5′ positional bias of intron-poor species is due to elevated loss rate at the 3′ end, whereas intron gain rates were not significantly different between the two ends. Intron-rich species do not exhibit similar positional bias, and correspondingly, we detect no difference in intron gain and loss rates between the gene ends.

To further test whether RT plays a role in both intron gain and loss, we computed for each species the correlation between the intron gain and loss rates of its genes. Strikingly, these correlations depend linearly on the average intron number per gene. In species with a low intron density (roughly, less than three introns per gene), intron gain and loss rates are negatively correlated, whereas in species with a higher intron density, a positive correlation is observed.

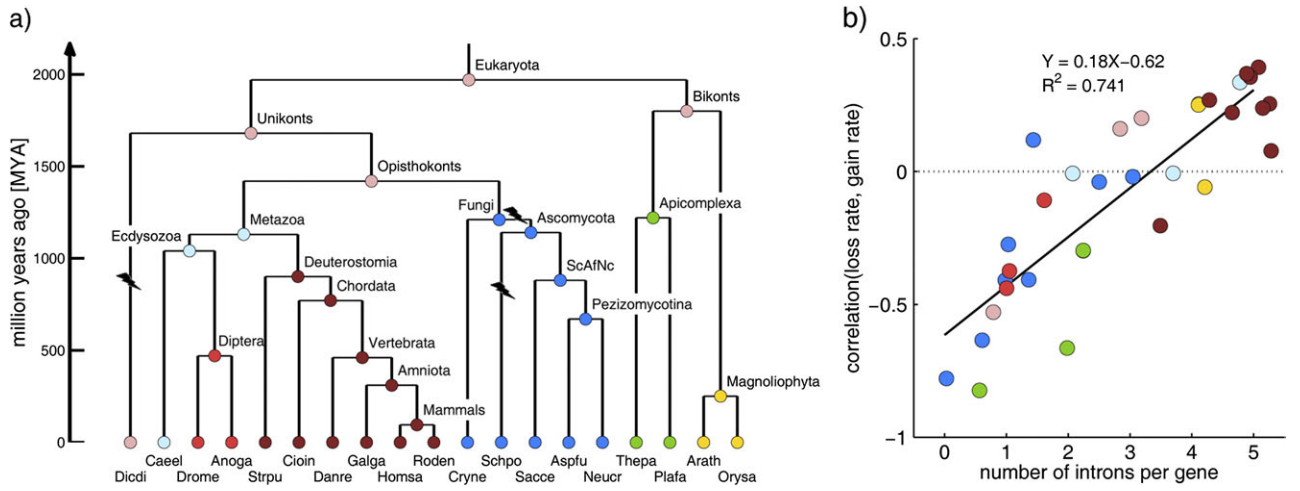


**Fig. 1.** Node color indicates phylogenetic association: fungi (blue), deuterostomia (brown), plants (yellow), diptera (red), apicomplexa (green), deep metazoan (cyan), and deep eukaryotes (pink). (*a*) The phylogenetic tree of eukaryotes used throughout this paper. Black lightning symbol indicates significantly higher intron loss rate at the 3′ end of the genes. Species and lineage abbreviations: (Caeel) *C. elegans*, (Strpu) *S. purpuratus*, (Cioin) *C. intestinalis*, (Danre) *D. rerio*, (Galga) *G. gallus*, (Homsa) *H. sapiens*, (Roden) *M. musculus* and *R. norvegicus* combined, (Drome) *D. melanogaster*, (Anoga) *A. gambiae*, (Cryne) *C. neoformans*, (Schpo) *S. pombe*, (Sacce) *S. cerevisiae*, (Aspfu) *A. fumigatus*, (Neucr) *N. crassa*, (Arath) *A. thaliana*, (Orysa) *O. sativa*, (Thepa) *T. parva*, (Plafa) *P. falciparum*, and (Dicdi) *D. discoideum*. (*b*) Spearman correlation between intron gain rate and intron loss rate as a function of the average number of introns per gene. Lineages for which $\rho(s)$ was not available were excluded (Eukaryota, Unikonts, Amniota, ScAfNc, and Apicomplexa).

## Methods

### Intron Presence/Absence Data

We used a data set that was compiled by us in a previous work (Carmel, Wolf, et al. 2007). It comprises 391 genes that have orthologs in 19 eukaryotic species: 9 metazoans (*Caenorhabditis elegans*, *Strongylocentrotus purpuratus*, *Ciona intestinalis*, *Danio rerio*, *Gallus gallus*, *Homo sapiens*, rodents [*Mus musculus* and *Rattus norvegicus* combined], *Drosophila melanogaster*, *Anopheles gambiae*), 5 fungi (*Cryptococcus neoformans*, *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, *Aspergillus fumigatus*, *Neurospora crassa*), 2 plants (*Arabidopsis thaliana*, *Oryza sativa*), 2 apicomplexans (*Theileria parva*, *Plasmodium falciparum*), and the protist *Dictyostelium discoideum*.

The preparation of the data is described in detail in (Carmel, Wolf, et al. 2007). In short, the protein-coding sequences of the orthologs of each gene were aligned, and the intron positions were mapped on the alignment. The data were then converted into binary data—bases immediately followed by an intron are represented by one, whereas zero stands for a base that is followed by another coding base. Only reliable gap-less portions of the multiple alignment were used for further analysis. Untranslated regions (UTRs) were excluded from the analysis in order to avoid possible biases (3′UTR is depleted with introns).

### Phylogenetic Tree

Some branching patterns within the eukaryotic tree are still debated. Here, we used the widely accepted unikonts–bikonts split (Stechmann and Cavalier-Smith 2002) and the ecdysozoan topology (Aguinaldo et al. 1997) and took the divergence times from Carmel, Wolf, et al. (2007), see figure 1*a*. Repeating the analysis using alternative topologies yielded very similar results (see "Results").

### Intron Gain and Loss Rates of Individual Genes

The binary alignments served as input for the EREM software, which uses an expectation-maximization algorithm to reconstruct the evolutionary history of the gene architecture (Carmel et al. 2005, 2007b, 2010). EREM is unique among similar software in that it can estimate the intron gain and loss rates of individual genes (Carmel et al. 2007a). Its main output consists of three matrices, $P(g,s)$, $L(g,s)$, and $G(g,s)$. $P(g,s)$ is the expected number of (known or inferred) introns present in gene $g$ at node $s$ of the phylogenetic tree. Likewise, $L(g,s)$ and $G(g,s)$ count the expected number of intron loss and gain events, respectively, in gene $g$ along the branch leading to node $s$. Using these matrices, we compute a fourth matrix that measures the number of introns that were not lost (but rather retained) in gene $g$ along branch $s$, $R(g,s)=P(g,\mathrm{Pa}(s))-L(g,s)$, where $\mathrm{Pa}(s)$ is the parent node of $s$.

Based on these matrices, we calculate the intron loss rate of gene $g$ along branch $s$ as the expected number of loss events per intronic site, $r_L(g,s)=L(g,s)/P(g,\mathrm{Pa}(s))=L(g,s)/(L(g,s)+R(g,s))$. For genes with zero introns in $\mathrm{Pa}(s)$, $L(g,s)=R(g,s)=0$, and the rate is designated as unknown.

Similarly, the intron gain rate of gene $g$ along branch $s$ was defined as the expected number of gain events per intron-free site, $r_G(g,s)=G(g,s)/(L_g-R(g,s))$, where $L_g$ is the length (in nucleotides) of gene $g$.

### Intron Gain and Loss Rates At the Opposing Ends of the Genes

In order to compute the rates at the opposing ends of the genes, we split each gene into two halves (the middle nucleotide was omitted from both halves in genes of odd length). In some genes, the reliable portion of the alignment of one of the halves was too short (less than 15 bases)

to be included in the analysis. In this case, the other half was excluded as well. Overall, 305 genes had a long-enough reliable alignment for both halves (supplementary table S1, Supplementary Material online). For these genes, we used EREM to compute the intron loss and gain rates in each half gene, $r_L^{5'}(g,s)$, $r_L^{3'}(g,s)$, $r_G^{5'}(g,s)$, and $r_G^{3'}(g,s)$.

### Positional Bias

To measure inhomogeneity in intron density along the genes, we counted, for each organism s, the number of genes in which the intron density at the 5′ half is greater than their density in the 3′ half, $B^{5'}(s)$. Similarly, we counted the number of genes in which the intron density at the 3′ half is greater than their density at the 5′ half, $B^{3'}(s)$. The positional bias of organism s is defined as $B(s)=B^{5'}(s)/(B^{5'}(s)+B^{3'}(s))$. If introns are equally likely to be in each side of the gene, we expect to obtain $B \approx 1/2$. Values higher than 1/2 indicate that introns tend to reside at the 5′ half, whereas values lower than 1/2 indicate the opposite trend.

## Results

### Footprints of RT-Mediated Intron Loss in Intron-Poor Species

EREM was used for computing the intron loss rate in the 5′ half and in the 3′ half of each of the 305 genes in each lineage (both living species and ancestral forms). In every lineage, we excluded from the analysis genes with an unknown rate in at least one of their halves and genes with zero loss rate at both halves. This procedure resulted in a variable number of genes suitable for analysis in each lineage (see "Methods," supplementary table S2, Supplementary Material online). For each lineage, we used the two-sided paired Wilcoxon sign test to compare the loss rates at the opposing ends of the genes. The *P* values were corrected for multiple comparisons using the false discovery rate (FDR) procedure at a level of 0.05 (Benjamini and Hochberg 1995). We found that the loss rate is significantly elevated at the 3′ end in Ascomycota, *D. discoideum*, and *S. pombe* (fig. 1*a*, Table 1). In general, these organisms tend to have low number of introns per gene (fig. 2). A notable exception is *S. cerevisiae* for which no loss rate difference was detected as both ends of the genes experienced many loss events.

To better understand the relation between the loss rate difference and the number of introns in each side of the gene, we computed for each organism *s* the positional bias $B(s)$ that measures the tendency of the introns to reside in the 5′ half of its genes (see "Methods," Table 1). We observed a highly significant negative correlation between $B(s)$ and the average number of introns per gene (Spearman correlation $-0.573$, $P = 2.6 \times 10^{-4}$; fig. 2).

Previous analyses measured positional bias by counting introns at opposite ends of the genes and therefore were unable to tell whether the bias is due to an increased loss rate at the 3′ end or an increased gain rate at the 5′ end. Here, by analyzing rates, we show that the positional bias is positively associated with elevated loss rate at the 3′ end. Taking this rate bias as an indication of RT activity (more about this assumption in the "Discussion"), we conclude that RT-mediated decay is a major mechanism of intron loss in intron-poor organisms.

In order to provide further credence to our results, we repeated the analysis with modifications to our definitions of the 3′ and 5′ ends by splitting each gene into four quarters. First, we defined the 3′ end of the gene as the 3′ most quarter, whereas the 5′-end was taken as the 5′ most quarter (supplementary fig. S1, Supplementary Material online). Second, we defined the 3′ end of the gene as the 3′ most quarter, whereas the 5′ end was taken as the rest of the gene (supplementary fig. S2, Supplementary Material online). In both cases, the results remained essentially unchanged.

### No Strong Evidence For RT Activity in Intron Gain

We have repeated the above analysis for intron gain rates. Genes with zero gain rate at both halves were excluded from the analysis (supplementary table S2, Supplementary Material online). No lineage showed significantly elevated intron gain rate at either end, but *O. sativa* (rice), magnoliophyta (flowering plants), and *D. discoideum* showed marginal *P* values, suggesting very weak elevated rate at the 3′ end (Table 1). Yet, gain rates in plants were previously shown to have a reduced accuracy (Carmel, Wolf, et al. 2007). Notably, *D. discoideum* is also characterized by a significant loss rate difference, with a net effect of many more introns at the 5′ end (fig. 2).

There are only five genes in *S. cerevisiae* that show nonzero intron gain rate (KOG0400, KOG0407, KOG1753, KOG1790, and KOG3430; supplementary table S2, Supplementary Material online). In all of them, the gain rate is zero at the 3′ end and is positive at the 5′ end. In addition, whenever loss rate is available, these genes show intron loss rate that is higher at the 3′ end. This observation suggests that these genes are under selection to host introns in their 5′ end, possibly due to intronic regulatory functions. Despite of the fact that many yeast introns carry snoRNAs, no known snoRNAs reside in these five genes. Interestingly, four of the five genes are ribosomal proteins (KOG1753 [Rps16ap], KOG1790 [Rpl34bp], KOG0407 [Rps14ap], KOG0400 [Rps13p]), which is a statistically significant enrichment ($P = 4 \times 10^{-5}$, hypergeometric test).

### Intron Gain and Loss Rates Are Positively Correlated for Intron-Rich Species and Are Negatively Correlated for Intron-Poor Species

The analysis conducted so far (based on data of rate bias) suggests that RT has a role in intron loss but not in intron gain. To further test this observation, we calculated, for each lineage, the correlation between the intron loss and gain rates of the genes (this time, the rates were measured along the entire length of the genes). To this end, we have used all the genes of a lineage *s* for which the loss rate could be inferred (see "Methods") and computed the Spearman correlation $\rho(s)=\mathrm{corr}_g(r_L(g,s), r_G(g,s))$.

In general, it is unclear what value $\rho(s)$ should have. When the processes of intron gain and loss share a mechanistic

**Table 1.** Intron-Related Features Measured for All Lineages (both living species and ancestral forms).

| Lineage | Introns Per Gene | Gain–Loss Correlation[a] | Loss Bias[b] | Gain Bias[b] | Positional Bias, *B* |
|---|---|---|---|---|---|
| Unikonts | 2.84 | 0.16 | 0.08 (3′) | 0.77 (5′) | 0.55 |
| Opisthokonts | 3.18 | 0.20 | 0.69 (3′) | 0.86 (3′) | 0.55 |
| Metazoa | 4.78 | 0.34 | 0.51 (5′) | 0.36 (3′) | 0.54 |
| Ecdysozoa | 3.70 | −0.01 | 0.30 (3′) | 0.49 (5′) | 0.53 |
| Deuterostomia | 5.28 | 0.08 | 0.95 (3′) | 0.64 (3′) | 0.52 |
| Chordata | 5.26 | 0.25 | 0.49 (5′) | 1.00 (3′) | 0.51 |
| Vertebrata | 5.15 | 0.24 | 1.00 (3′) | 0.45 (5′) | 0.52 |
| Amniota | 5.07 | N\A | 0.06 (5′) | N\A | 0.52 |
| Diptera | 1.61 | −0.11 | 0.20 (3′) | 0.30 (5′) | 0.52 |
| Fungi | 2.50 | −0.04 | 0.11 (3′) | 0.08 (5′) | 0.59 |
| Ascomycota | 2.24 | −0.30 | *$7.9 \times 10^{-5}$ (3′) | 0.02 (5′) | 0.62 |
| ScAfNc | 0.98 | −0.41 | 0.01 (3′) | 0.86 (5′) | 0.58 |
| Pezizomycotina | 1.43 | 0.12 | 0.92 (3′) | 0.39 (5′) | 0.58 |
| Bikonts | 2.50 | N\A | 0.41 (5′) | N\A | 0.51 |
| Magnoliophyta | 4.21 | −0.06 | 0.90 (3′) | 0.02 (3′) | 0.47 |
| Apicomplexa | 2.24 | −0.3 | 0.86 (3′) | 0.91 (5′) | 0.53 |
| Mammals | 5.02 | N\A | 0.47 (3′) | N\A | 0.52 |
| Dicdi | 0.79 | −0.53 | *$1.4 \times 10^{-6}$ (3′) | 0.00 (5′) | 0.69 |
| Caeel | 2.07 | −0.01 | 0.90 (3′) | 0.37 (3′) | 0.53 |
| Strpu | 4.65 | 0.22 | 0.04 (5′) | 0.87 (5′) | 0.54 |
| Cioin | 3.49 | −0.20 | 0.52 (5′) | 0.82 (3′) | 0.46 |
| Danre | 5.08 | 0.39 | 0.33 (5′) | 0.02 (3′) | 0.53 |
| Galga | 4.94 | 0.36 | 0.76 (3′) | 0.03 (3′) | 0.53 |
| Homsa | 4.89 | 0.37 | 0.26 (3′) | 0.81 (3′) | 0.54 |
| Roden | 4.28 | 0.27 | 0.30 (3′) | 0.21 (3′) | 0.55 |
| Drome | 1.05 | −0.37 | 0.21 (3′) | 0.76 (3′) | 0.56 |
| Anoga | 1.00 | −0.44 | 0.69 (3′) | 1.00 (3′) | 0.51 |
| Cryne | 3.05 | −0.02 | 1.00 (3′) | 0.57 (3′) | 0.53 |
| Schpo | 0.61 | −0.63 | *$2.6 \times 10^{-5}$ (3′) | 0.05 (3′) | 0.73 |
| Sacce | 0.03 | −0.78 | 0.38 (3′) | 0.06 (3′) | 0.88 |
| Aspfu | 1.36 | −0.41 | 0.10 (3′) | 0.73 (5′) | 0.60 |
| Neucr | 1.03 | −0.27 | 0.08 (3′) | 0.19 (5′) | 0.62 |
| Arath | 4.11 | 0.25 | 0.93 (3′) | 0.89 (3′) | 0.47 |
| Orysa | 4.11 | 0.25 | 0.46 (5′) | 0.01 (3′) | 0.45 |
| Thepa | 1.98 | −0.66 | 0.51 (5′) | 0.39 (3′) | 0.54 |
| Plafa | 0.56 | −0.82 | 0.06 (3′) | 1.00 (3′) | 0.58 |

[a] NOTE.—Spearman correlation between the intron gain and loss rates.
[b] *P* value by two-sided paired Wilcoxon sign test. In parenthesis, the side with the higher rate. Asterisk stands for values that are significant after FDR correction at a level of 0.05.

component, a positive correlation is expected. But, if both processes are driven by selection but are mechanistically unrelated, a negative correlation is expected. We observed $\rho(s)$ values ranging from −0.82 to 0.39 (Table 1). Unexpectedly, $\rho(s)$ was found to have significant linear dependence upon the average number of introns per gene (Spearman correlation 0.868, $P = 1.3 \times 10^{-7}$; fig. 1*b*).

It seems that intron gain and loss processes are uncoupled in intron-poor organisms. This is consistent with the fact that we identified traces of RT activity affecting intron loss in these organisms but no signs for RT activity affecting intron gain. In contrast, intron-rich organisms are characterized by a positive correlation between intron loss and gain rates but show no traces of RT activity. This suggests that the intron gain and loss mechanisms are coupled but are probably not dependent on RT.

## The Results Are Insensitive to the Tree Topology

Some topological aspects of the eukaryotic tree are controversial, and we have therefore repeated the analysis for an alternative tree topology. Instead of the unikonts–bikonts split, we have assumed the crown group topology (plants group with unikonts to the exclusion of apicomplexans), and instead of ecdysozoan topology, we have assumed the ceolomate topology (insects group with vertebrates to the exclusion of nematodes). The analysis gave qualitatively very similar results, with some minor changes. The most notable change is that more lineages showed significant rate bias following the FDR procedure. More specifically, an elevated intron loss at the 3′ side was significant also in unikonts, fungi, and pezizomycotina (supplementary fig. S3*a*, Supplementary Material online). Magnoliophyta and *O. sativa* showed significant elevation of intron gain rate at the 3′ end. Interestingly, the intron gain rate in *D. discoideum* was found to be higher at the 5′ end, whereas previously, it was higher at the 3′ end. This might be a reflection of the very long branch connecting *D. discoideum* to the rest of the tree. Nevertheless, essentially the same relationships as before were obtained between the average number of introns per gene and $\rho(s)$ (supplementary fig. S3*b*, Supplementary Material online) and between the average number of introns per gene and the positional bias (supplementary fig. S4, Supplementary Material online).

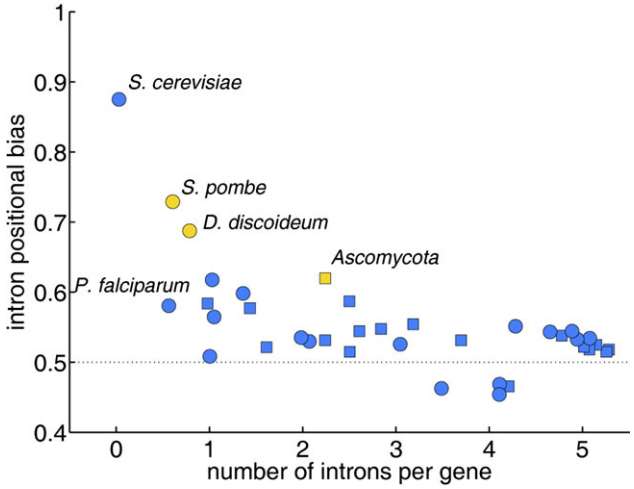


FIG. 2. Intron positional bias versus the average number of introns per gene for all organisms, both living species (circles) and extinct ancestral forms (squares). Yellow indicates significantly higher intron loss rate at the 3′ end of the genes.

## Discussion

The presented results indicate that intron gain and loss mechanisms are probably very different among different taxonomic groups. We see elevated intron loss at the 3′ end of the gene for intron-poor organisms, mostly within fungi. Intron-rich species, in particular vertebrates and plants, seem to rely on different mechanisms.

It is an open question why intron-poor species show loss rate bias, resulting in a high density of introns near the 5′ end. Our results put forward the possibility that RT-mediated intron loss, when active, is very efficient in removing introns. Therefore, lineages in which this mechanism becomes active rapidly lose many of their introns and are characterized by low intron numbers. This is consistent with the growing recognition that ancient eukaryotic ancestors carried many introns (Carmel, Wolf, et al. 2007), and all intron-poor organisms have experienced many intron loss events.

Unexpectedly, we revealed a clear linear relationship between the intron gain–loss rate correlation in the branch leading to a species and the average number of introns in that species (fig. 1*b*). Such a relationship, however, is consistent with the hypothesis above as when RT-mediated intron loss becomes more dominant, the more negative the correlation between the intron gain and loss rates would be. One of the steps in RT-mediated intron loss is homologous recombination of the cDNA with the genomic copy of the gene. Three eukaryotic groups are characterized by a negative correlation between the intron gain and loss rates—fungi, insects, and plants (fig. 1*b*). Consistent with our observation, many species of fungi predominantly use the homologous recombination pathway for DNA repair (Farlow et al. 2010; Zhang et al. 2010), leading to increased efficiency of RT-mediated intron loss. Drosophila shows a slight preference to the NHEJ pathway, but the level of homologous recombination is still comparable to that of NHEJ (Johnson-Schlitz et al. 2007). In contrast, plants are thought to predominantly use the NHEJ DNA repair pathway. However, the low levels of homologous recombination in plants may be compensated by their tendency to polyploidity (that increases the number of homologous sequences) and by high levels of retroelements activity (Bennetzen 1996) (that increases the level of their cellular RT enzyme).

Throughout this paper, we interpret high intron loss rate at the 3′ end as a hallmark of RT activity. Yet, one should note that there is another option. Elevated loss rate at the 3′ end may also be a result of preferential retention of introns at the 5′ end. Indeed, 5′ most introns are known to participate in transcription regulation; thus, there may be a selection against their removal (Bradnam and Korf 2008). Further work is required to tell whether the bias is due to increased loss rate at the 3′ end or due to increased selection against intron loss at the 5′ end. Moreover, it had been suggested that priming of the poly(A) terminus on a T-rich fragment of the mRNA can initiate the reverse transcription at different locations along the transcript (Feiber et al. 2002; Niu et al. 2005). In such cases, the activity of RT will not result in a 3′ bias.

In vertebrates, intron gain and loss may be coupled, although independent on RT. A recent interesting suggestion for a mechanism that may contribute to both intron loss and intron gain is the NHEJ DNA repair pathway. The discovery of new introns in *Daphnia pulex* with short directed repeats at their ends triggered the suggestion that these introns were gained through NHEJ DNA repair (Li et al. 2009). This possibility has gained further support from observations in other species (Zhang et al. 2010; Ragg 2011). Recently, it had been suggested that during NHEJ, not only that exogenous DNA can be integrated into the genome but also that genomic fragments, including whole introns, can be deleted (Farlow et al. 2010).

## Supplementary Material

Supplementary figures S1–S4 and supplementary tables T1–T2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

## Acknowledgments

This work was supported by the European Union Marie Curie International Reintegration Grant (PIRG05-GA-2009-248639) and by the Lohenstein August and Elza Foundation. NEC is partially supported by the Sudarsky Center for Computational Biology at the Hebrew University of Jerusalem.

## References

Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* 387:489–493.

Babenko VN, Rogozin IB, Mekhedov SL, Koonin EV. 2004. Prevalence of intron gain over intron loss in the evolution of paralogous gene families. *Nucleic Acids Res.* 32:3724–3733.

Banyai L, Patthy L. 2004. Evidence that human genes of modular proteins have retained significantly more ancestral introns than their fly or worm orthologues. *FEBS Lett.* 565:127–132.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J R Stat Soc Series B Methodol.* 57:289–300.

Bennetzen JL. 1996. The contributions of retroelements to plant genome organization, function and evolution. *Trends Microbiol.* 4:347–353.

Bradnam KR, Korf I. 2008. Longer first introns are a general property of eukaryotic gene structure. *PLoS One.* 3:e3093.

Carmel L, Rogozin IB, Wolf YI, Koonin EV. 2005. An expectation-maximization algorithm for analysis of evolution of exon-intron structure of eukaryotic genes. In: McLysaght A, Huson DH, editors. RECOMB 2005 Comparative Genomics International Workshop (RCG 2005). Lecture notes in bioinformatics. Berlin, (Germany): Springer. p. 35–46.

Carmel L, Rogozin IB, Wolf YI, Koonin EV. 2007a. Evolutionarily conserved genes preferentially accumulate introns. *Genome Res.* 17:1045–1050.

Carmel L, Rogozin IB, Wolf YI, Koonin EV. 2007b. Patterns of intron gain and conservation in eukaryotic genes. *BMC Evol Biol.* 7:192.

Carmel L, Wolf YI, Rogozin IB, Koonin EV. 2007. Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res.* 17:1034–1044.

Carmel L, Wolf YI, Rogozin IB, Koonin EV. 2010. EREM: parameter estimation and ancestral reconstruction by expectation-maximization algorithm for a probabilistic model of genomic binary characters evolution. *Adv Bioinformatics.* 2010:167408.

Catania F, Lynch M. 2008. Where do introns come from? *PLoS Biol.* 6:e283.

Cavalier-Smith T. 1991. Intron phylogeny: a new hypothesis. *Trends Genet.* 7:145–148.

Cho S, Jin SW, Cohen A, Ellis RE. 2004. A phylogeny of caenorhabditis reveals frequent loss of introns during nematode evolution. *Genome Res.* 14:1207–1220.

Coghlan A, Wolfe KH. 2004. Origins of recently gained introns in Caenorhabditis. *Proc Natl Acad Sci U S A.* 101:11362–11367.

Colbourne JK, Pfrender ME, Gilbert D, et al. (69 co-authors). 201. The ecoresponsive genome of Daphnia pulex. *Science* 331:555–561.

Coulombe-Huntington J, Majewski J. 2007. Characterization of intron loss events in mammals. *Genome Res.* 17:23–32.

Crick F. 1979. Split genes and RNA splicing. *Science* 204:264–271.

Csuros M, Rogozin IB, Koonin EV. 2008. Extremely Intron-rich genes in the alveolate ancestors inferred with a flexible maximum-likelihood approach. *Mol Biol Evol.* 25:903–911.

Derr LK, Strathern JN. 1993. A role for reverse transcripts in gene conversion. *Nature* 361:170–173.

Farlow A, Meduri E, Schlotterer C. 2010. DNA double-strand break repair and the evolution of intron density. *Trends Genet.* 27:1–6.

Feiber AL, Rangarajan J, Vaughn JC. 2002. The evolution of single-copy Drosophila nuclear 4f-rnp genes: spliceosomal intron losses create polymorphic alleles. *J Mol Evol.* 55:401–413.

Fink GR. 1987. Pseudogenes in yeast? *Cell* 49:5–6.

Irimia M, Rukov JL, Penny D, Vinther J, Garcia-Fernandez J, Roy SW. 2008. Origin of introns by 'intronization' of exonic sequences. *Trends Genet.* 24:378–381.

Johnson-Schlitz DM, Flores C, Engels WR. 2007. Multiple-pathway analysis of double-strand break repair mutations in Drosophila. *PLoS Genet.* 3:e50.

Knowles DG, McLysaght A. 2006. High rate of recent intron gain and loss in simultaneously duplicated Arabidopsis genes. *Mol Biol Evol.* 23:1548–1557.

Koonin EV. 2006. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol Direct.* 1:22.

Li W, Tucker AE, Sung W, Thomas WK, Lynch M. 2009. Extensive, recent intron gains in Daphnia populations. *Science.* 326:1260–1262.

Lin K, Zhang DY. 2005. The excess of 5' introns in eukaryotic genomes. *Nucleic Acids Res.* 33:6522–6527.

Lynch M, Richardson AO. 2002. The evolution of spliceosomal introns. *Curr Opin Genet Dev.* 12:701–710.

Martin W, Koonin EV. 2006. Introns and the origin of nucleus-cytosol compartmentalization. *Nature.* 440:41–45.

Mourier T, Jeffares DC. 2003. Eukaryotic intron loss. *Science* 300:1393.

Nguyen HD, Yoshihama M, Kenmochi N. 2005. New maximum likelihood estimators for eukaryotic intron evolution. *PLoS Comput Biol.* 1:e79.

Niu DK, Hou WR, Li SW. 2005. mRNA-mediated intron losses: evidence from extraordinarily large exons. *Mol Biol Evol.* 22:1475–1481.

Nixon JE, Wang A, Morrison HG, McArthur AG, Sogin ML, Loftus BJ, Samuelson J. 2002. A spliceosomal intron in Giardia lamblia. *Proc Natl Acad Sci U S A.* 99:3701–3705.

Parma J, Christophe D, Pohl V, Vassart G. 1987. Structural organization of the 5' region of the thyroglobulin gene. Evidence for intron loss and "exonization" during evolution. *J Mol Biol.* 196:769–779.

Purugganan M, Wessler S. 1992. The splicing of transposable elements and its role in intron evolution. *Genetica* 86:295–303.

Ragg H. 2011. Intron creation and DNA repair. *Cell Mol Life Sci.* 68:235–242.

Rodriguez-Trelles F, Tarrio R, Ayala FJ. 2006a. Models of spliceosomal intron proliferation in the face of widespread ectopic expression. *Gene* 366:201–208.

Rodriguez-Trelles F, Tarrio R, Ayala FJ. 2006b. Origins and Evolution of Spliceosomal Introns. *Ann Rev Genet.* 40:47–76.

Rogers JH. 1989. How were introns inserted into nuclear genes? *Trends Genet.* 5:213–216.

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol.* 13:1512–1517.

Roy SW. 2004. The origin of recent introns: transposons? *Genome Biol.* 5:251.

Roy SW, Gilbert W. 2005. Complex early genes. *PNAS.* 102:1986–1991.

Roy SW, Irimia M. 2009. Mystery of intron gain: new data and new models. *Trends Genet.* 25:67–73.

Roy SW, 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.

Roy SW, 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.

Sakharkar MK, Chow VT, Kangueane P. 2004. Distributions of exons and introns in the human genome. *In Silico Biol.* 4:387–393.

Sakurai A, Fujimori S, Kochiwa H, Kitamura-Abe S, Washio T, Saito R, Carninci P, Hayashizaki Y, Tomita M. 2002. On biased distribution of introns in various eukaryotes. *Gene* 300:89–95.

Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, Ast G. 2007. Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. *Genome Biol.* 8:R127.

Sharp PA. 1985. On the origin of RNA splicing and introns. *Cell* 42:397–400.

Sharp PA. 1991. "Five easy pieces". *Science* 254:663.

Simpson AG, MacQuarrie EK, Roger AJ. 2002. Eukaryotic evolution: early origin of canonical introns. *Nature.* 419:270.

Stechmann A, Cavalier-Smith T. 2002. Rooting the eukaryote tree by using a derived gene fusion. *Science* 297:89–91.

Stoltzfus A, Logsdon JM Jr., Palmer JD, Doolittle WF. 1997. Intron "sliding" and the diversity of intron positions. *Proc Natl Acad Sci U S A.* 94:10739–10744.

Sverdlov AV, Babenko VN, Rogozin IB, Koonin EV. 2004. Preferential loss and gain of introns in 3′ portions of genes suggests a reverse-transcription mechanism of intron insertion. *Gene* 338:85–91.

Zhang LY, Yang YF, Niu DK. 2010. Evaluation of models of the mechanisms underlying intron loss and gain in Aspergillus fungi. *J Mol Evol*. 71:364–373.