

LETTERS

On the Incidence of Intron Loss and Gain in Paralogous Gene Families

Scott William Roy and David Penny

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

Understanding gene duplication and gene structure evolution are fundamental goals of molecular evolutionary biology. A previous study by Babenko et al. (2004. Prevalence of intron gain over intron loss in the evolution of paralogous gene families. *Nucleic Acids Res.* 32:3724–3733) employed Dollo parsimony to infer spliceosomal intron losses and gains in paralogous gene families and concluded that there was a general excess of gains over losses. This result contrasts with patterns in orthologous genes, in which most lineages show an excess of intron losses over gains, suggesting the possibility of fundamentally different modes of intron evolution between orthologous and paralogous genes. We further studied the data and found a low level of intron position conservation with outgroups, and this led to problems with using Dollo parsimony to analyze the data. Statistical reanalysis of the data suggests, instead, that intron losses have outnumbered intron gains in paralogous gene families.

Introduction

Two ongoing mechanisms of genomic and functional diversification in eukaryotes are duplication of genes and the gain and loss of spliceosomal introns, genomic sequences that are excised from RNA transcripts. A wealth of recent work has elucidated the evolution of spliceosomal introns through eukaryotic evolution; however the ultimate causes and timing of intron origin remain the subject of debate (Fedorov et al. 2003; Fedorov and Fedorova 2004, 2006; Collins and Penny 2005; Martin and Koonin 2006; Wang, Feng, and Niu 2007). The numbers and positions of introns vary across homologous genes, implying the occurrence of intron loss and/or gain through evolution (Perler et al. 1980; Federov, Merican, and Gilbert 2002; Rogozin et al. 2003). Over the past few years, a growing body of work has shown that intron loss is a major mechanism of genome evolution, and that intron losses tend to outnumber intron gains in orthologous genes, a pattern observed across a wide array of diverse lineages (e.g., Robertson 1998; Roy, Fedorov, and Gilbert 2003; Mourier and Jeffares 2003; Niu et al. 2004; Lin and Zhang 2005; Lin et al. 2006; Roy and Gilbert 2005a; Stajich and Dietrich 2006; Roy and Hartl 2006; Roy and Penny 2007).

Interestingly, then, the largest study to date of intron evolution in gene duplicates came to the opposite conclusion. Babenko et al. (2004) studied intron evolution in lineage-specific expansions (LSEs; genes that have undergone one or more duplications following divergence from the other studied species; see example in figure 1) in six eukaryotic lineages. For each LSE they aligned paralogous genes with orthologous genes from available outgroups and mapped intron positions onto the alignment to identify shared intron positions. Using Dollo parsimony to infer intron loss/gain for each intron position (fig. 1), they reported an excess of gains over losses. This result stands as the most important large-scale analysis to find considerably more gain than loss. Intriguingly, this pattern suggests that the pronounced differences between the evolution of orthologous and paralogous genes (e.g., Katju and Lynch 2003; Kopelman et al.

2005; Wang, Yu, and Long 2004) include intron–exon structure evolution.

However, accumulating evidence for frequent intron loss suggests that Dollo parsimony may have trouble distinguishing intron loss and gain (Roy and Gilbert 2005b; Roy and Penny 2006). The data from the original paper (ftp://ftp.ncbi.nih.gov/koonin/intron_evolution/LSEs/) are summarized in table 1, and compare with the illustrative example given in figure 1. Dollo parsimony will only be accurate if intron presence/absence in available outgroups reflects presence/absence at the time of gene duplication. Intron positions shared between both descendent lineages of a gene duplication (e.g., introns A and B in figure 1) very likely represent introns present at the time of duplication. The presence of such introns in outgroups thus represents an important test of the utility of Dollo parsimony. Of a total 3,809 such shared intron positions across 1,421 gene duplications in the data set, nearly half (45.9%) are not represented in any studied outgroups (available outgroup species as well as possible additional gene copies from the LSE; intron B in figure 1 is an example of such an intron). Thus Dollo parsimony frequently fails to infer correctly intron presence/absence at the time of duplication. This failure of outgroups is expected to lead to many actual intron losses' being misidentified as intron gains: indeed, for the 45.5% gene duplications for which there are no intron positions shared between both descendent lineages as well as outgroup(s), parsimony infers 4.1 gains per loss (101/248), compared to 2.4 (176/74) for duplications with one such shared intron and 1.2 (142/121) for duplications with more than one shared intron.

How can we correct for this? Consider a pair of duplicates (which may either represent terminal branches or internal branches that themselves bifurcate at later duplication events; figure 1). For each duplication event, we first estimate P , the probability that an intron present at the time of duplication is represented in an available outgroup. If there are l_2 total intron positions that are shared between (descendants of) both duplicates, of which l_{2A} are represented in an outgroup sequence, this suggests that P is around $\hat{p} = l_{2A}/l_2$. Now, if there are L introns that were present at the duplication event and subsequently lost, we likewise expect that only roughly a fraction p would be present in outgroups. If there are l_{1A} introns shared between exactly one descendent branch and outgroup(s) (i.e., losses inferred

Key words: gene duplication, gene families, genome evolution, parsimony, statistical inference.

E-mail: scottwroy@gmail.com.

Mol. Biol. Evol. 24(8):1579–1581. 2007

doi:10.1093/molbev/msm082

Advance Access publication April 29, 2007

	INTRON	A	B	C	D	E	F
	LSE gene 1	+	+	+	+	-	+
	LSE gene 2	+	-	+	-	-	+
	LSE gene 3	+	+	-	-	+	-
	LSE gene 4	+	-	+	-	-	-
	OUTG SP 1	+	-	-	+	-	-
	OUTG SP 2	-	-	-	+	-	-
	OUTG SP 3	-	-	-	-	-	-

FIG. 1.—Hypothetical example of a lineage-specific expansion (LSE). The LSE consists of four paralogs in one species (LSE genes 1–4) due to three gene duplications since the divergence from outgroup species (OUTG SP 1–3). There are six observed intron positions (introns A–F), with intron presence/absence data (+/–) for each homolog. We consider the second gene duplication (oval), and estimate numbers of intron losses and gains in the directly descendent branches (bold lines). Dollo parsimony infers one intron loss (intron C) and two gains (E and F). Intron D represent an event that occurred on a later branch (after the divergence of LSE genes 1 and 2); intron B represents at least two events, one before the duplication in question and one after the divergence of LSE genes 1 and 2. Among the two intron positions shared between both descendent lineages and thus likely present at the time of duplication (A and B), only one is present in an outgroup (A). This raises the possibility that apparent gains (E and F) may also have been present at the gene duplication and instead represent real intron losses. Employing the method described in the text, we have $l_2 = 2$ (introns A and B), $l_{2A} = 1$ (A), $l_1 = 3$ (C, E and F), and $l_{1A} = 1$ (C). Thus we estimate that there are roughly two losses ($=l_{1A}l_2/l_{2A}$) and one gain ($l_1 - l_{1A}l_2/l_{2A}$).

by parsimony), we have $\hat{p}\hat{L} = l_{1A}$; thus we estimate $\hat{L} = l_{1A}/\hat{p} = l_{1A}l_2/l_{2A}$. If there are l_1 total introns that are present in descendants of only one duplicate (both those present and those absent in outgroups), the total estimated number of gains is thus l_1 minus the estimated number of losses: $l_1 - l_{1A}l_2/l_{2A}$.

We applied this method to the data (Table 2). Contrary to the previous finding, we estimated that there are more total intron losses than gains (280 versus 233 in the entire data set; table 1). Losses are estimated to have outnumbered gains in three of five lineages, and to have been roughly equal to gains in a fourth.

However, we think that this measure, too, is likely to be biased toward intron gain. Consider a pair of gene duplicates with two shared introns, each with a p probability of being represented in outgroups. According to the binomial distribution, the probability that 0, 1, or 2 introns will be represented in outgroups is $(1-p)^2$, $2p(1-p)$, and p^2 , respectively. If 0 are represented in outgroups, the gene will be excluded as uninformative. Thus the probabilities of 1 or 2 introns being present in the outgroup, given that at least

Table 1
Summary of the Data

Species	Studied Gene Families	Duplication Events Per Family	Intron Positions
<i>H. sapiens</i>	278	1.18	1389
<i>C. elegans</i>	228	1.23	955
<i>D. melanogaster</i>	93	1.17	245
<i>S. pombe</i>	29	1.10	70
<i>S. cerevisiae</i>	17	1.00	17
<i>A. thaliana</i>	419	1.56	1854
All	1064	1.33	4530

Table 2
Estimated Intron Losses and Gains in Lineage Specific

Species	Estimated Ratio of Gains-to-Losses	
	Parsimony	Current
<i>H. sapiens</i>	3.4 (274/80)	3.8 (87.3/22.7)
<i>C. elegans</i>	3.3 (517/157)	1.1 (116.7/109.3)
<i>D. melanogaster</i>	2.4 (127/52)	0.4 (7.5/17.5)
<i>S. pombe</i>	2.4 (34/14)	0.0 (0/3)
<i>S. cerevisiae</i>	1.0 (1/1)	N/A (0/0)
<i>A. thaliana</i>	3.1 (434/139)	0.2 (23/127)
All	3.1 (1387/443)	0.8 (233.4/279.6)

NOTE.—Estimated intron losses and gains in lineage specific expansions by Dollo parsimony (from Babenko et al. 2004) and as described in the text (“Current”). Total numbers for the current estimates are lower because of the exclusion of cases with uninformative outgroups (i.e., for which $l_{2A} = 0$).

one is, are $2(1-p)/(2-p)$ and $p/(2-p)$, respectively. If one or two introns are present in outgroups, we estimate \hat{p} is 0.5 or 1, respectively. Thus on average we estimate $\hat{p} = 0.5 \times 2(1-p)/(2-p) + 1 \times p/(2-p) = 1/(2-p)$, which is larger than p since $1/(2-p) - p = (1-p)^2/(2-p) > 0$. In general for a gene with n introns, we will overestimate \hat{p} by on average $p(1-p)^n/[1-(1-p)^n]$. This overestimate of p will cause us to underestimate the number of intron losses that are incorrectly inferred by parsimony to be intron gains, and thus it will lead us to under/overestimate intron losses/gains.

The case is worst for genes with a single intron shared between duplicates: either the intron is absent in outgroups (in which case the duplication is not considered) or it is present (in which case we estimate $\hat{p} = 1$, leading to an overall estimate of $\hat{p} = 1$). The bias toward inference of intron gain was confirmed by simulations (data not shown). Correspondingly, the overall estimated ratio of intron losses to gains jumps from 1.2 (280/233) to 1.7 (228/134) when duplications with only a single shared intron (i.e., $l_{2A} = 1$) are excluded, and to 2.0 (174/87) among cases with $l_{2A} > 2$.

These results suggest that paralogous genes, like orthologous genes, have experienced an excess of intron loss over intron gain over most lineages. Further studies employing more (and more closely related) species, and accounting for the possibility of homoplastic intron insertion, will be necessary to finally resolve the issue. These results provide a further cautionary example in using parsimony in directionalizing intron loss/gain events, and underscore the importance of using more sophisticated statistical methods and/or more closely related species for accurate inferences about genome evolution.

Literature Cited

- Babenko VN, Rogozin IB, Mekhedov SL, Koonin EV. 2004. Prevalance of intron gain over intron loss in the evolution of paralogous gene families. *Nucleic Acids Res.* 32:3724–3733.
- Csurös M. 2005. Likely scenarios of intron evolution. Third RECOMB Satellite Workshop on Comparative Genomics. p. 47–60 Springer LNCS 3678.
- Fedorov A, Fedorova L. 2004. Introns: mighty elements from the RNA world. *J Mol Evol.* 59:718–721.
- Fedorov A, Fedorova L. 2006. Where is the difference between the genomes of humans and annelids? *Genome Biol.* 7:203.
- Fedorov A, Merican AF, Gilbert W. 2002. Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc Natl Acad Sci USA.* 99:16128–16133.

- Fedorov A, Roy S, Fedorova L, Gilbert W. 2003. Mystery of intron gain. *Genome Res.* 13:2236–2241.
- Katju V, Lynch M. 2003. The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. *Genetics.* 165:1793–1803.
- Kopelman NM, Lancet D, Yanai I. 2005. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nat Genet.* 37:588–589.
- Lin H, Zhu W, Silva J, Gu X, Buell CR. 2006. Intron gain and loss in segmentally duplicated genes in rice. *Genome Biol.* 7:R41.
- Lin K, Zhang D-Y. 2005. The excess of 5' introns in eukaryotic genomes. *Nucl Acids Res.* 33:6522–6527.
- 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 H, 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.
- Perler F, Efstratiadis A, Lomedico P, Gilbert W, Kolodner R, Dodgson J. 1980. The evolution of genes: the chicken preproinsulin gene. *Cell.* 20:555–566.
- Raible F, Tessmar-Raible K, Osogawa K, et al. 2005. Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*. *Science.* 310:1325–1326.
- Robertson HM. 1998. Two large families of chemoreceptor genes in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement, and intron loss. *Genome Res.* 8:449–463.
- 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, Fedorov A, 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.
- Roy SW, Gilbert W. 2005a. The pattern of intron loss. *Proc Natl Acad Sci USA.* 102:713–718.
- Roy SW, Gilbert W. 2005b. Complex early genes. *Proc Natl Acad Sci USA.* 102:1986–1991.
- Roy SW, Hartl DL. 2006. Very little intron loss/gain in Plasmodium: intron loss/gain mutation rates and intron number. *Genome Res.* gr.4845406
- Roy SW, Penny D. 2006. Smoke without fire: most reported cases of intron gain in nematodes instead reflect intron losses. *Mol Biol Evol.* 23:229–2262.
- 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.
- Stajich JE, Dietrich FS. 2006. Evidence of mRNA-mediated intron loss in the human-pathogenic fungus *Cryptococcus neoformans*. *Eukaryotic Cell.* 5:789–793.
- Wang HF, Feng L, Niu D-K. 2007. Relationship between mRNA stability and intron presence. *Biochem Biophys Res Commun.* [epub]
- Wang W, Yu H, Long M. 2004. Duplication-degeneration as a mechanism of gene fission and the origin of new genes in *Drosophila* species. *Nat Genet.* 36:523–527.

Kenneth Wolfe, Associate Editor

Accepted April 10, 2007