Intron Loss and Gain in *Drosophila*

Jasmin Coulombe-Huntington and Jacek Majewski

Department of Human Genetics, McGill University, Montreal, Quebec, Canada

Although introns were first discovered almost 30 years ago, their evolutionary origin remains elusive. In this work, we used multispecies whole-genome alignments to map *Drosophila melanogaster* introns onto 10 other fully sequenced Drosophila genomes. We were able to find 1,944 sites where an intron was missing in one or more species. We show that for most (>80%) of these cases, there is no leftover intronic sequence or any missing exonic sequence, indicating exact intron loss or gain events. We used parsimony to classify these differences as 1,754 intron loss events and 213 gain events. We show that lost and gained introns are significantly shorter than average and flanked by longer than average exons. They also display quite distinct phase distributions and show greater than average similarity between the 5′ splice site and its 3′ partner splice site. Introns that have been lost in one or more species evolve faster than other introns, occur in slowly evolving genes, and are found adjacent to each other more often than would be expected for independent single losses. Our results support the cDNA recombination mechanism of intron loss, suggest that selective pressures affect site-specific loss rates, and show conclusively that intron gain has occurred within the Drosophila lineage, solidifying the "introns-middle" hypothesis and providing some hints about the gain mechanism.

Introduction

The evolutionary history of spliceosomal introns remains to this day an unresolved issue in many respects. It is still unclear whether their expansion occurred strictly in early eukaryotic or even preeukaryotic ancestors, often referred to as the "introns-early" hypothesis, or whether new introns still appear today, the "introns-late" or "introns-middle" hypothesis. Introns could have expanded in one or more bursts, in a fashion similar to transposable elements, they could be the result of tandem duplications within exons that accidentally code for a pair of functional splice sites (Zhuo et al. 2007), and it has also been suggested that they could appear through a reverse-splicing mechanism catalyzed by the splicing machinery itself (Coghlan and Wolfe 2004). Introns are generally considered to be nonfunctional, except for their passive role in some alternative splicing (AS) events, some transcriptional regulatory functions, and their suggested evolutionary role in facilitating exon shuffling. However, their level of interspecies conservation varies between and within genes, suggesting there are some selective pressures related to yet unknown biological functions (Majewski and Ott 2002; Gaffney and Keightley 2006). Although as much as 80% of intron positions are conserved across some very distant eukaryotic species, like humans and sea anemones (Putnam et al. 2007), many introns appear to be completely missing in some species. Either these differences are the result of novel intron insertions or of introns being completely and precisely deleted. We can obtain valuable insights into intron evolutionary dynamics and gain further understanding of the origin of spliceosomal introns by studying these loss or gain events.

So far, studies of intron dynamics have been mostly limited to comparing intron positions across highly conserved, orthologous or paralogous genes from often very distant species (Rogozin et al. 2005; Yoshihama et al. 2007). The problem is, the fewer the species included in these studies, and the more evolutionarily distant they are, the easier it is to mistake parallel losses for gains or vice versa. A dramatic example of this was how information from a single recently sequenced species, the sea anemone, changed the estimated proportion of human introns that are at least 500 Myr old from roughly 25% to an astounding 80% or more (Putnam et al. 2007). Mainly for this reason, reported cases of intron gain events have been criticized (Logsdon et al. 1998; Roy and Penny 2006). Technical issues aside, it appears losses, although rare, occur at a measurable rate in most eukaryotes, whereas intron gains, at least in the last 100 Myrs, seem to be restricted to specific clades. We have recently shown that over a hundred loss events, and not a single gain, could be detected across humans, dogs, and rodents (Coulombe-Huntington and Majewski 2007). Overall, many more loss events have been inferred and documented than gain events. Intron loss is by now a pretty well-established phenomenon. The prevailing theory for the biological mechanism, as portrayed in figure 1, is that a processed (intronless) mRNA expressed in the germ line is reverse transcribed to cDNA which then recombines with the genomic version of the gene, thereby precisely deleting the unmatched intronic sequence. This mechanism has been demonstrated experimentally in yeast (Derr et al. 1991). Many studies have demonstrated that lost introns display characteristics that support this mechanistic model (Mourier and Jeffares 2003; Sverdlov et al. 2004; Roy and Gilbert 2005; Roy and Hartl 2006; Coulombe-Huntington and Majewski 2007), such as small size, 3′ positional bias, and enrichment in highly expressed genes.

With the increasing number of sequenced eukaryotic genomes becoming available, we can zoom in and explore more recent intron evolutionary dynamics. As we increase the phylogenetic resolution, it should be easier to distinguish true gains from parallel losses. The advantages of studying intron dynamics in fruit flies stem from the fact that we have 12 fully sequenced fruit fly genomes (Adams et al. 2000; Myers et al. 2000; Richards et al. 2005; Clark et al. 2007), we know the position of nearly every gene in the model species *Drosophila melanogaster* and the fruit flies' short generation time makes them likely to experience many intron loss or gain events. Additionally, the Drosophila genome is a good place to look for selective pressures acting on intron loss or gain events due to their large effective population size and their tendency to preserve

Key words: intron loss, intron gain, intron dynamics, drosophila, comparative genomics.

E-mail: jacek.majewski@mcgill.ca.

Mol. Biol. Evol. 24(12):2842–2850. 2007
doi:10.1093/molbev/msm235
Advance Access publication October 27, 2007

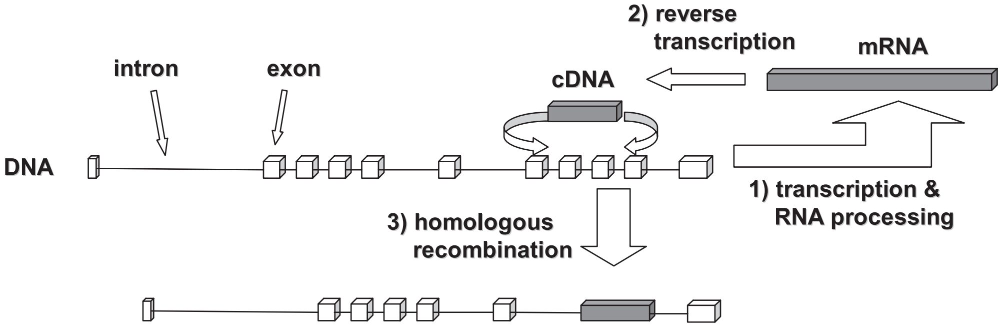

FIG. 1.—The most widely accepted model of intron loss, which is supported by our data, whereby an intronless cDNA recombines with the genomic version of the gene, deleting one or more introns.

a compact genome. Using *D. melanogaster* gene annotations to map the introns in the other flies, we can consider our study to be practically genome wide, providing enormous statistical power to detect the distinctive characteristics of lost or gained introns. These characteristics provide useful insights into the molecular mechanism of such events and about the selective pressures acting on them.

Methods

Reference-Based Intron Mapping

We used a previously described approach (Coulombe-Huntington and Majewski 2007) based on the latest MultiZ whole-genome alignments (Blanchette et al. 2004) and *D. melanogaster* RefSeq gene annotations, downloaded from the University of California at Santa Cruz Genome Browser database (Karolchik et al. 2003), to map the position of each *D. melanogaster* splice site directly onto the genomes of 10 other Drosophila species, *Drosophila sechellia*, *Drosophila erecta*, *Drosophila yakuba*, *Drosophila ananassae*, *Drosophila pseudoobscura*, *Drosophila persimilis*, *Drosophila willistoni*, *Drosophila virilis*, *Drosophila mojavensis*, and *Drosophila grimshawi*. We discarded *Drosophila simulans* from the analysis due to its poor genome assembly (Clark et al. 2007). In order to infer an intron in the target species, the start and end of the intron must be successfully aligned with the reference species and be separated by more than 30 bp in the target genome. A missing intron was inferred when both edges of the reference intron were mapped but were less than 20 bp apart. For cases where there was an alignment gap where the reference splice site mapped, the closest aligned base, up to 5 bp to each side, was used as the intron edge instead. We were able to map both edges of 28,933 *D. melanogaster* introns in every other species (see supplementary table 1, Supplementary Material online). Our predictions agree well with GeneMapper, another reference-based gene annotation system (Chatterji and Pachter 2006). A total of 84% of GeneMapper splice sites mapped to within 5 bp of one of our inferred intron edges. A total of 1,944 predicted introns were shorter than 20 bp in one or more species, 84% of which were shorter than 10 bp and 48% of zero size (see supplementary table 1, Supplementary Material online). We assumed these introns were missing due to the loss of the intron along the target species lineage or gain of the intron in the *D. melanogaster* lineage. As introns smaller than 20 bp are assumed to be unspliceable based on experimental studies in other organisms (Russell et al. 1994; Slaven et al. 2006), we presume that the majority of the remaining size to these tiny predicted introns is actually due to minor misalignments around the gap.

The presumed missing introns, predicted introns that were shorter than 20 bp, along with 200 bp of flanking exon sequence, were realigned to the homologous *D. melanogaster* sequence using ClustalW (Thompson et al. 1994), with high (80) gap opening penalty and low (0) gap extension penalty. Then, if possible, we made some minor adjustments on the ClustalW alignments to show that the missing intron was completely missing in the target species. We expected the gap in the target species would be delimited by a consensus GT/AG splice site signal sequence in the reference species. We could show that 82.3%, rather than 48%, of missing introns appeared after realignment to be of zero size, leaving the bordering exons intact.

It should be noted that our approach allows us to predict only events affecting introns present in the reference, annotated genome of *D. melanogaster*. We did not attempt to infer the positions of introns that are absent in the reference species but may be present in one or more of the remaining genomes. The latter is a much more difficult problem, which involves first inferring the coding sequence of orthologous genes in the nonannotated species, followed by determining whether the inferred coding sequence is interrupted by a legitimate, spliceable intron. We found such approaches to be considerably more error prone (data not shown). Although this bias in ascertainment prevents us from directly comparing the rates of intron gains and losses on the same branch of the tree, we believe it is necessary in order to maintain low false discovery rates of gene structure changes.

Inferring Intron Loss and Gain

Introns missing in one or more target species can be explained either by the loss of the intron somewhere along the target species' lineage or by the gain of the intron in the lineage of the reference species, *D. melanogaster*. We inferred losses and gains using Dollo parsimony, whereby independent parallel loss of the same intron is allowed but parallel gain is not. Dollo parsimony has been utilized before for the study of intron loss and gain (Rogozin et al. 2003). As a result of inferring introns from *D. melanogaster* annotations, we can only infer gains along the melanogaster lineage and losses on other branches. We did not infer a loss or gain for events that occurred on one of the 2 oldest branches of the tree because loss and gain are equally likely.

Correlating Transposon Numbers with the Species-Specific Loss Rates

We used standalone Blast 2.2.14 (Altschul et al. 1997) to look for 3 *D. melanogaster* transposase sequences and one reverse transcriptase sequence (National Center for Biotechnology Information accession numbers S60466, ANN39288, Q7M3K2, and AAB50148) in the genome of each species and counted the number of hits (E value < 10). We then calculated the Pearson correlation coefficient between the numbers of hits in each genome and the species-specific loss rate, defined as the total number of intron losses detected in a species over the divergence time from *D. melanogaster*.

Measuring Overrepresentation of Adjacent Losses

To determine whether there is significant clustering of lost introns within genes, we used all genes for which there were exactly 2 inferred intron losses in *D. pseudoobscura* and at least 3 introns in the reference species, *D. melanogaster*. We calculated the expected probability that independently occurring losses be adjacent as $P = 2/n$, where n is the total number of introns (lost and not lost) in the gene. We calculated the expected number of adjacent losses over all eligible genes as the sum of the individual expectations. We compared the expected number with the observed number of adjacent losses using a chi-square test to assess whether there is significant overrepresentation of adjacent losses.

Simulating the Evolution of Intron Phase Distribution

In order to assess whether the phase distribution of gained introns combined with the phase preference of intron loss could explain the current phase distribution in *D. melanogaster*, we used a simple simulation. We started the simulation with 10,000 introns, distributed according to the same ratios as gained introns with respect to phase, 65% phase 0, 20% phase 1, and 15% phase 2. Introns were then removed one by one, according to the phase preference of intron loss, which was obtained by dividing the phase ratios of lost introns by the phase ratios of all introns. The result is that phase 0 introns are 33% more likely to be lost than phase 1 or 2 introns. In order to choose the phase of the removed intron at a given round, we calculate the probabilities of choosing each phase by multiplying the number of remaining introns in each phase by the phase preference rates of losses and normalize to make the sum of probabilities equal to one. The question was whether or not the phase distribution in the simulation could come to within 1% root mean squared deviation (RMSD) of the current phase distribution in *D. melanogaster*. This was achieved after 93% of the original 10,000 introns were lost.

Measuring Relative Conservation

We measured the pairwise conservation between *D. melanogaster* and a target species as the ratio of matching amino acids over the total number of *melanogaster* residues, based on the translated MultiZ alignment. We performed an unpaired t-test, comparing the average pairwise conservation of genes with losses to the average pairwise conservation of other genes. Then, to assign a "relative conservation" to a given gene, we used the ratio of the gene's average pairwise conservation over the average conservation for all genes over all species. A relative conservation of 1 would mean a perfectly average conservation for that particular gene.

To compare the level of conservation of average introns with that of lost introns, which are missing in some species, we computed the average intronic pairwise conservation for all combinations of species required. We simply measure, for each species where the intron is still present, the ratio of bases that match *D. melanogaster*'s sequence, based on the MultiZ alignment. Then, we performed a paired student's t-test by matching each lost intron's average pairwise conservation with the average intron conservation for the same set of species. This means that the level of conservation used as control was always based solely on the species where the lost intron was still present. For the purpose of the multiple regression (see Results), we define the relative conservation as the ratio of an intron's average conservation to the average conservation of all introns over the same set of species.

Results

Mapping Introns

We were able to map 28,933 *D. melanogaster* introns onto every other species (see Methods). A total of 1,944 of these introns were missing from one or more species, assumed to be the result of a loss or gain event somewhere along the tree. A total of 82.3% of these were shown to be completely missing, leaving the exonic sequence intact (see Methods). Based on Dollo parsimony, allowing for parallel losses but no parallel gains, we infer 1,754 loss events and 213 gain events. Figure 2 shows the number of gains or losses inferred on each branch. As a direct result of the gene mapping approach, all studied introns have to be present in the reference species *D. melanogaster*. Therefore, we can only detect gain events on branches that are ancestral to *D. melanogaster* (dashed lines on fig. 2) and loss events on other branches. For events that happened on 1 of the 2 oldest branches, lacking further rooting of the tree, we cannot distinguish between gains and losses. There are 220 such differences, as shown in figure 2.

Varying Loss Rates

As shown in figure 2, the number of losses per branch does not follow a predictable, clock-like, pattern. Some clades, like the *willistoni* group, seem to undergo many more losses per million years than others. Many factors could produce diversity in the loss rate, such as generation time or

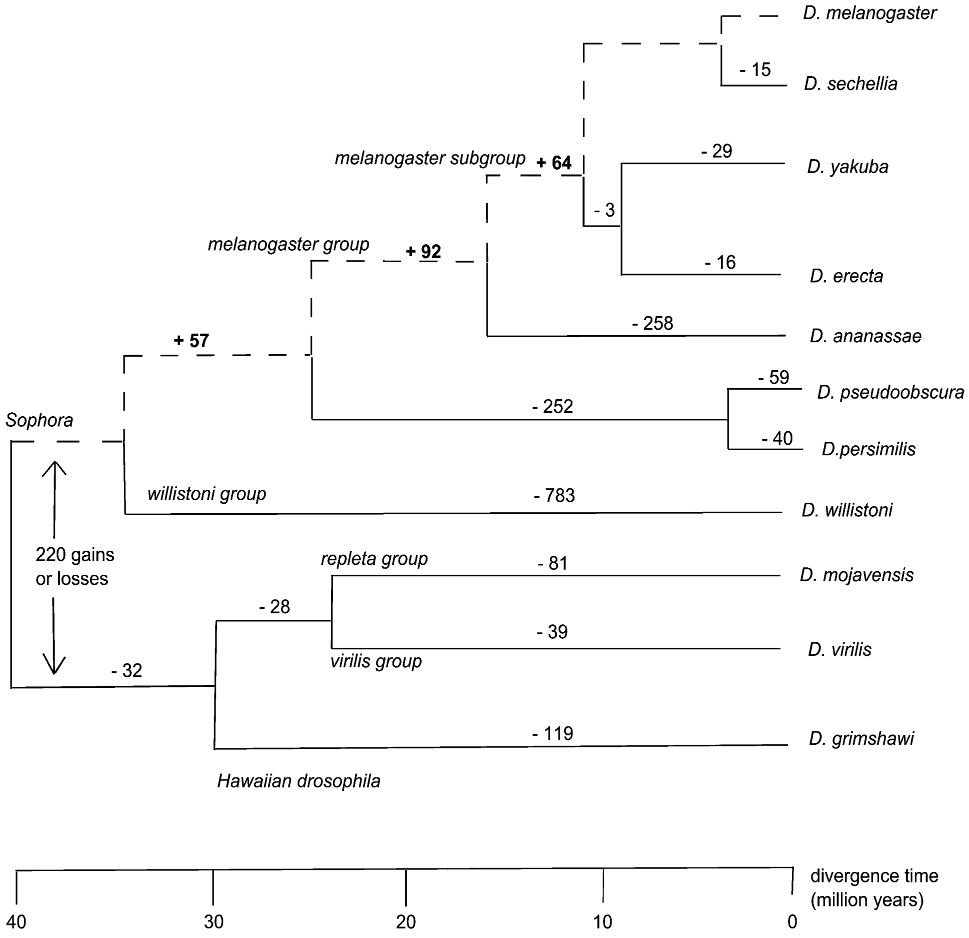

FIG. 2.—Number of predicted intron gains (in bold) and losses, as inferred using Dollo parsimony. The dashed lines indicate the branches where gains could be inferred and losses could not. The tree is based on an image from the assembly/alignment/annotation of 12 related Drosophila species (Clark et al. 2007).

effective population size. We find there is a fairly good correlation with genome size ($R^2 = 0.75, n = 10, P = 0.012$), which might suggest that the level of activity of some transposable elements, known to increase genome size (Kidwell 2002), might affect the branch-specific rate. This theory is consistent with the recombination model of intron loss because reverse transcriptase is involved in the mechanism. We attempted to correlate the loss rate of each species with the number of transposon sequences in the species' genome for different transposon sequences (see Methods). The only query sequence to yield a significant correlation with the species-specific rate was a *P* element ($R^2 = 0.69$, $n = 10$, $P = 0.027$). Although *D. willistoni* is known to contain many *P* elements, this gene codes for transposase, which acts through a "cut-and-paste" mechanism, unlike reverse transcriptase. It is possible that *P* elements are involved in intron loss via a currently unknown mechanism or that activity of yet other transposable elements has been involved with intron deletions. Assuming the cDNA recombination model is correct, we may not be able to detect the presence of reverse transcriptase-coding retroposons because they may have mostly degenerated since the loss events took place, which is likely given the high rate of sequence evolution in flies. Additionally, the branch lengths are very approximate (Pollard et al. 2006), which might not allow appropriate correlations to be drawn.

Characterizing Lost and Gained Introns

More than 80% of missing introns were shown to be exactly cut out, leaving a functional looking splice site pair (see Methods). Introns with missing orthologues, with a median length of 97 bp, are significantly shorter than average introns ($t = 22.4$, degrees of freedom [df] = 5783, *P* value $< 10^{-10}$). The average length of gained introns is similar to that of losses ($t = 1.2$, df = 322, *P* value = 0.23). We show that losses appear skewed to the 3′ of genes ($t = 3.6$, df = 2864, *P* value = 3.3×10^{-4}). Flanking coding exons of lost and gained introns are longer than average ($t = 9.4$, df = 4150, *P* value $< 10^{-10}$). Genes with loss or gain have significantly more introns than average genes ($t = 27.5$, df = 3202, *P* value $< 10^{-10}$), and genes with losses have significantly more introns than genes with gains ($t = 4.2$, df = 359.6, $P = 4 \times 10^{-5}$). We also demonstrate that pairs of adjacent losses occur more often than expected by chance ($\chi^2 = 52.6$, df = 1, *P* value 4×10^{-13}; see Methods for details). These characteristics of lost introns each provide support to the recombination model of intron

Table 1
Intron Phase Distributions in *Drosophila melanogaster*

Phase	All (%)	Lost (%)	Gained (%)	AGGT[a] (%)
0	41.2	49	65	72.5
1	32.6	28	20	19.3
2	26.2	23	15	8.2

[a] Hypothetical introns inserted in exonic AGGT sites.

loss. First of all, the recombination model accounts neatly for the precise "splicing out" of introns, which leaves only a fused exon. From the model, we also expect the occurrence of 2 or more neighboring introns disappearing in a single recombination event, which explains why we find more adjacent pairs of losses than expected. Furthermore, regions with short introns and long exons would have the greatest ratio of homologous sequence with the intronless cDNA, which should favor recombination. As the cDNA is created from 3′ to 5′, with respect to the gene's orientation, partial cDNAs should be enriched in 3′ exonic sequences. However, the fact that 3′ introns are generally shorter than 5′ introns ($R^2 = 0.068$, P value $< 2 \times 10^{-16}$) is also a potential source of this bias. The fact that losses occur preferentially in intron dense genes is expected regardless of the mechanism, but it supports the fact that we are looking at actual losses rather than misclassified gains or artifacts caused by genome assembly errors.

Intron Phases

Fruit fly introns, like human introns, are not evenly distributed over each of the 3 possible phases, phase 0 indicating an intron found between 2 codons, phase 1, between the first and second base of a codon, and phase 2, between the second and third. Most eukaryotes have significantly more than one-third phase 0 introns and usually more phase 1 than phase 2 introns. Based on RefSeq *D. melanogaster* gene annotations, we computed the percentage of introns in each phase for the whole genome. Distributions are displayed in table 1. Assuming that losses and gains occur with no preference with respect to phase, we expected to find similar ratios. Instead we found that, although the direction of the skew was the same, the ratios were significantly different for both losses ($\chi^2 = 38.2$, df $= 2$, P value $= 5 \times 10^{-9}$) and gains ($\chi^2 = 50.9$, df $= 2$, P value $= 8.7 \times 10^{-12}$). We observe a neat consistency in the phase preference of losses by demonstrating that the phase bias of the 143 introns that have been lost independently in 2 different branches follows the expected, more pronounced, distributional skew ($\chi^2 = 1.56$, df $= 2$, P value $= 0.46$), and the same goes for the 30 introns that have been lost in 3 independent lineages ($\chi^2 = 0.46$, df $= 2$, P value $= 0.79$). We simulated an evolutionary scenario that assumes the gains' ratios as the ancestral distribution and takes into account the phase preference of losses, whereby phase 0 introns are one-third more likely to be lost than phase 1 or 2 introns (see Methods). After 93% of the ancestral introns were lost, the phase ratios in the simulation were equal to *D. melanogaster*'s current ratios (within 1% RMSD). In fact, according to recent estimates, it seems probable that Drosophila did lose in excess of 85% of its ancestral introns (Putnam et al. 2007). Assuming that all introns originally appeared in the same ratios as our detected gains, we can explain the difference between the original ratios and the current ratios as caused by the inherent phase preference of intron loss events.

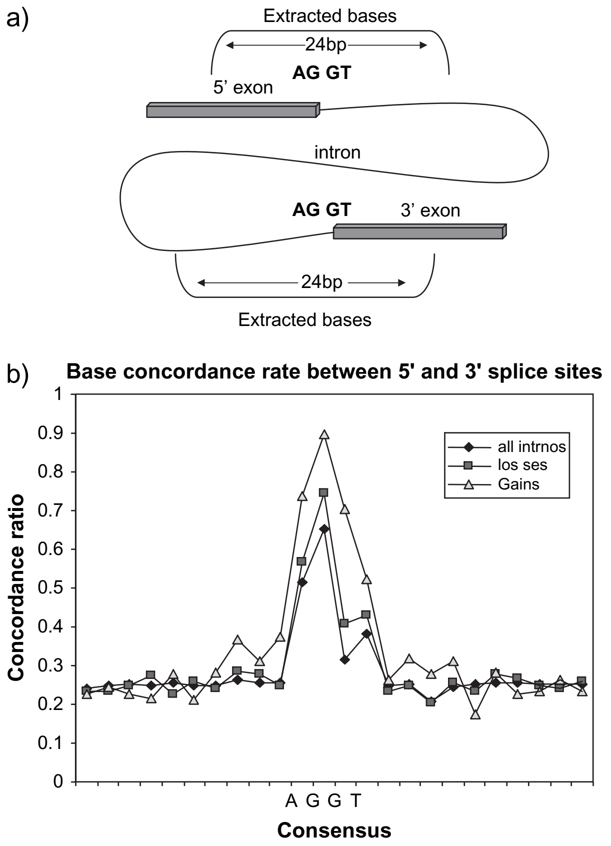

FIG. 3.—(*a*) We aligned 24 bp centered on the 5′ splice site with the 24 bp centered on the 3′ partner splice site in the direct orientation. AGGT is the consensus motif of the 4 innermost bases at both splice sites. (*b*) Concordance ratio between each base surrounding the 5′ splice site and the corresponding base, equally distanced from the 3′ splice site. The similarity of the 2 splice sites is significantly higher for the lost and gained categories than for the genome-wide average.

Characteristics of Splice Sites

We extracted 24 base sequences surrounding each splice site in *D. melanogaster* and aligned the sequence at the start of each intron with the sequence at the end, as displayed in figure 3*a*. Comparing the 4 bases centered on each 5′ splice site of lost introns with the 4 bases surrounding the 3′ splice site revealed that these sequence pairs were significantly more similar than they are for average introns ($\chi^2 = 106$, df $= 4$, $P = 4.8 \times 10^{-22}$). This difference was much more pronounced for gained introns ($\chi^2 = 337$, df $= 4$, $P = 8.9 \times 10^{-78}$). Figure 3*b* plots the average rate of concordance between each base surrounding the 5′ splice site and the base equally distanced from the 3′ partner splice site, over lost introns, gained introns, and all introns. The most common sequence we see at

Table 2
Intron Characteristics Multivariate Correlations to Loss and Gain

	Intron Size (logged)	5′ Exon Size (logged)	3′ Exon Size (logged)	Relative Position	Splice Site Similarity[a]	Intron Conservation	Gene Conservation
Losses	− $P < 2 \times 10^{-16}$	+ $P = 5.2 \times 10^{-9}$	+ $P = 1.3 \times 10^{-10}$	+ $P = 0.79$	+ $P < 2 \times 10^{-16}$	− $P = 1.6 \times 10^{-12}$	+ $P = 0.0094$
Gains	− $P < 2 \times 10^{-16}$	+ $P = 1.4 \times 10^{-13}$	+ $P = 8.6 \times 10^{-6}$	− $P = 9.8 \times 10^{-4}$	+ $P < 2 \times 10^{-16}$	− $P = 0.18$	− $P = 0.88$

[a] Number of matching bases out of 4 (see fig. 3).

these 4 bases is AGGT. If the middle of this sequence where to serve as a "proto-splice-site," a predetermined insertion site for new introns, it could explain this characteristic of gains. One possible mechanism would be the tandem duplication of an exonic sequence that contains an AGGT motif (Zhuo et al. 2007). Such a mechanism could potentially create a spliceable intron in a single event, without altering the coding sequence. The theory whereby introns are inserted through a reverse-splicing mechanism also favors proto-splice-sites. To assess whether the proto-splice-site hypothesis could explain the highly skewed phase distribution of gains, we calculated the expected phase distribution over all exonic AGGT motifs in *D. melanogaster*, assuming an intron would insert itself between the 2 guanines. The ratios are displayed in table 1. As it comes pretty close to the distribution of gained introns, this becomes a possible explanation, bearing in mind that introns might occasionally insert themselves in slightly different motifs and that the distribution obtained from *D. melanogaster* might be somewhat different than it was at the time the gains occurred. For example, the same motif in Human yields a very different phase distribution of 56% phase 0, 28% phase 1, and 16% phase 2. Therefore, if the proto-splice-site hypothesis is true, it may be difficult to deduce the phase distribution of proto-splice-sites present in the distant ancestors where most introns appeared.

Intron loss, Alternative Splicing, and Selection

Most introns are thought to be useless pieces of genes that merely slow down the transcription process. Some, however, are required for their role in AS events, and others are thought to play a role in transcriptional regulation. Based on *D. melanogaster* RefSeq annotations, we created a list of 4,718 exons that underwent a putative AS event that required the presence of either one or both flanking introns. The number of lost introns that seemed to disrupt an AS event in *D. melanogaster* was significantly lower than the null expectation ($\chi^2 = 66.7$, df = 1, $P = 3.1 \times 10^{-16}$). We only found 14 cases of a loss disrupting a putative AS event in *D. melanogaster*, whereas the expected was 88.6. Seven of these losses disrupted cryptic splice site usage events and 7 disrupted alternative usage of a cassette exon. The expected number of disruptions of each type was, respectively, 19.9 and 68.4. Therefore, the underrepresentation of disruptions of cassette exon AS events is much more significant. We also found that lost introns are less conserved than the average, considering the species where the intron is still present (t[paired] = 9.5, df = 2752, $P < 2 \times 10^{-16}$; see Methods for details), suggesting there are selective forces influencing the probability of fixation of loss events. Genes with missing introns on the other hand are more conserved at the protein sequence level than average genes, across all 11 species ($t = 19.4$, df = 3250, $P < 2.2 \times 10^{-16}$). The fact that the genes are more conserved than the average serves to show that the relatively lower conservation of introns is not an artifact of finding losses in poorly assembled regions of the genomes.

The Bigger Picture

We performed a linear multiple regression analysis to assess the combined power of studied intron characteristics to predict loss or gain and to control for the covariation of variables, such as size and intron position or size and sequence conservation. It has also been documented that large exons tend to be surrounded by short introns because they are otherwise undetected by the spliceosome (Sterner et al. 1996); therefore, this analysis should also allow to assess whether the observation of long flanking exons is independent of the small size of the lost and gained introns. We correlated the number of times an intron was lost or gained across the tree with the log of intron size, the log of flanking exon sizes, splice site partner similarity, relative position within the gene, the intron's relative conservation, and the gene's relative conservation (see Methods). The direction of inferred slopes and P values are displayed in table 2. Sizes, splice site similarity, and position are based on *D. melanogaster* gene annotations and sequence. The multiple R^2 for losses was 0.022 and for gains, 0.01, meaning the combined predictive power is not very high. However, this approach should allow us to discard presumed relationships between loss or gain probability and some covarying but not directly correlated variables. Most results were consistent with our earlier analysis. We could confirm with greater certainty that intron size, flanking exon size, and similarity between the 5′ and 3′ splice sites are significantly linked to the probability of both gain and loss. Intron conservation and gene conservation are directly linked to loss probability, rather than through the mediating effect of intron size or position. The position of introns within the gene, however, does not show any correlation with loss events within this analysis. This could mean that the finding was merely an artifact caused by the relationship between intron position and size. On the other hand, the regression reveals a possible 5′ positional bias for gained introns. We also notice that gains, unlike losses, show greater correlation with 5′ exon length than with 3′ exon length. These differences could

Table 3
Multivariate Correlations in Highly Conserved Genes

	Intron Size (logged)	5′ Exon Size (logged)	3′ Exon Size (logged)	Relative Position	Splice Site Similarity[a]	Intron Conservation	Gene Conservation
Losses	− $P < 2 \times 10^{-16}$	+ $P = 1.9 \times 10^{-7}$	+ $P = 2.4 \times 10^{-7}$	− $P = 0.56$	+ $P < 2 \times 10^{-16}$	− $P = 0.0026$	− $P = 8.7 \times 10^{-12}$
Gains	− $P = 0.0014$	+ $P = 7.1 \times 10^{-5}$	+ $P = 0.57$	− $P = 0.18$	+ $P = 8.7 \times 10^{-16}$	+ $P = 0.12$	− $P < 2 \times 10^{-16}$

[a] Number of matching bases out of 4 (see fig. 3).

reflect the directionality, with respect to the gene, of the mechanisms involved in intron loss and gain. As the creation of cDNAs occurs from 3′ to 5′, it makes sense that the length of the 3′ exon has a greater effect on the probability of loss than the length of the 5′ exon. As for gains, the relatively greater effect of 5′ exon length over 3′ exon length, as well as the 5′ positional bias, could suggest that the gain mechanism occurs in the same orientation as transcription.

Qualitative Conclusions and Quality of Data

It is unreasonable to expect the assembly of each fly genome or the alignments of these genomes to be absolutely perfect. Some of the fly genomes are still at the stage of their first assembly, as opposed to the Human genome that, at the time of writing of this article, has already reached its 18th assembly. In addition, whole-genome alignments introduce another source of error. As we expect a reasonable false discovery rate, we decided to use the regression analysis to compare our qualitative results with those of a highly confident subset of genes. We performed the same regression as described above on the subset of genes for which the pairwise protein sequence identity between the *D. melanogaster* gene and the predicted orthologues in every other species was beyond 80%. Slope orientations and *P* values are shown in table 3. We expect the alignment itself to be of much better quality within and around these genes as the highly conserved coding sequence severely restricts the number of ways to create the most probable alignment. The analysis revealed that the direction of every significant correlation was conserved in the subset, except for relative gene conservation. In the subset, it seems that both gains and losses correlate inversely with gene conservation, which could mean that introns in highly conserved genes execute some crucial functions, as exemplified by the relatively high frequency of introns in *Saccharomyces cerevisiae* ribosomal protein genes (Nakao et al. 2004) or, alternatively, that loss and gain mechanisms are inherently too error prone to evolve and reach fixation in such essential genes. In any case, we believe this example demonstrates the potential pitfalls of studies based on nonrandom subsets of genes, as are most analyses on intron dynamics to date.

Discussion

This study constitutes the largest scale investigation of intron dynamics in Drosophila to date. The fact that it is genome wide makes it possible to define the characteristics of lost and gained introns without bias. Many of the characteristics of lost introns support the cDNA-mediated recombination model of intron loss, such as short length, long flanking exons, and clustering of lost introns within genes. The statistical power obtained by using the entire genome as control also allowed us to show that lost introns have lower sequence conservation than average. The elevated sequence-level conservation of introns that are less likely to be lost suggests that introns do exert some biological functions. Binding to specific transcription factors would justify conservation of sequence motifs, as would roles in AS. It is probable that the significantly lower proportion of lost introns used for AS affects the average sequence conservation. The fact that the genes that lose introns are more conserved than average is interesting for a few reasons. First, it proves that the relatively low sequence conservation of lost introns is not an artifact caused by poor assembly or alignment quality. Second, the recombination model requires genes to be highly expressed in the germ line in order to undergo intron loss, and such genes are likely well conserved on average (Arango et al. 2006). There remains the interesting variation of the loss rate across different clades. Although we found an association with genome size and the number of *P* elements, we fail to find a clear, consistent explanation for this variation. It should be noted that the estimated divergence times of species are very approximate (Pollard et al. 2006). This uncertainty makes it hard to find the cause of the rate variation or indeed if there is significant variation.

As mentioned in the introduction, reported cases of intron gains so far have been criticized, mostly for being confounded by multiple parallel losses. How does our evidence of intron gain hold up in this respect? First, it should be recalled that our gains were inferred based solely on maximizing parsimony over the tree. After losses and gains were properly classified, we discovered significant differences between the 2 classes. One key difference is that there are significantly more introns in genes with losses than in genes with gains, as would be expected under the probabilistic model where the probability of a gene experiencing a loss (but not a gain) increases with the number of introns present. Second, gains are significantly different from losses in their level of similarity between pairs of splice sites and in their highly skewed phase distribution. Additionally, gains show a 5′ positional bias whereas losses do not, and the 2 groups have a different relationship to 3′ and 5′ exon length. The strongest evidence against the parallel loss theory is that inferring a false positive gain on the *melanogaster* group branch (see fig. 2) would require the same intron to be lost 3 times independently and for a false positive gain on the *melanogaster* subgroup branch, 4 times. Assuming that all differences were actually caused by losses and that intron loss generally affects random introns, as suggested by the low R^2 of the multiple regression analysis, we would only expect 0.23 false positive gains on the branch of the *melanogaster* group and 0.003 on the branch of the *melanogaster* subgroup. Thus, although Dollo parsimony may overestimate the number of gain events (Rogozin et al. 2005), and we cannot fully rule out possibility that some of our gains may, in fact, be misclassified recurrent losses, it is extremely unlikely that all or even a substantial proportion of our reported intron gains are false.

Our study sheds some much needed light on the ongoing introns-early versus introns-late debate. Like many in the field have concluded, the answer is the introns-middle hypothesis (de Souza 2003; Koonin 2006), whereby most introns must have been gained very early in eukaryotic evolution, if not in preeukaryotic ancestors, whereas some introns are still gained today, at least in some species. Although, because of our ascertainment method, we detected many more losses than gains, the average overall rates are comparable: 6.1 gains per million years versus 8.9 losses per million years. The fact that the loss rate is higher than the gain rate agrees with the fact that fruit flies are believed to have lost most of their introns, having many fold fewer introns than their eukaryotic ancestors (Putnam et al. 2007).

The question as to how introns originally appeared is a complicated one. In the introduction, we presented 3 proposed mechanisms of intron gain. All 3 are examples of the broader, insertional theory of intron gains. The alternative is the formative theory, whereby introns were created as a by-product of exon shuffling. In the formative theory, introns are simply pieces of DNA that got caught between newly inserted exons. Our study brings support to the insertional model, as a process that has been occurring in relatively recent history (<40 MYA). The main reason is that we find many examples of intron gains in between exons that have not been shuffled, being found in the same position relative to the rest of the gene across all 11 species. The fact that newly gained introns show exceptional similarity between splice site pairs also supports the insertional model. The similarity could be the result of a tandem duplication, a by-product of transposition or a characteristic of ideal insertion sites for a reverse-splicing mechanism. Some researchers have attempted to explain the uneven phase distribution of introns in most species as a result of proto-splice-site insertion motifs, without success (Long et al. 1998; Long and Rosenberg 2000). Others have suggested that exon shuffling and the formative theory of introns explain the phase distribution (Vibranovski et al. 2006). One problem is that basing the studies on extant species might give an inaccurate picture of the phase distribution of proto-splice-sites in the ancestral species, as exemplified by the considerable difference in the phase distributions predicted using the same motif (AGGT) in flies and Human. Another factor these studies did not consider is the inherent preference of intron loss for phase 0 introns, a relation which had not been documented prior to this study. We have shown that limiting intron insertion/creation to AGGT motifs can explain the phase distribution skew of newly gained introns fairly well. We have also shown, using simulation, that the inherent phase preference of intron loss can explain the difference between the phase distribution of gained introns and that of all current introns. Because our results favor the insertional theory of introns, we have at least a few models to test. To assess whether introns result from insertions of transposon-like elements, we Blasted *D. melanogaster* introns that have or have not been gained onto every other intron. We expected the recently gained introns to bear higher similarity to each other and to other introns, thus yielding a greater number of Blast hits, but we did not detect such an effect (data not shown). Assuming the tandem duplication model was correct, we expected to detect remnants of direct repeats that would be longer than the 4 bases around the splice sites. We also failed to find such long repeats using Blast (data not shown). It should be noted that our analysis did not uncover any truly recent intron gains, the latest events occurring around 10 MYA. This relatively long interval, combined with the fast rate of evolution in Drosophila, may have led to the decay of the original intronic sequences and prevented our approach to detect any similarity.

The reverse-splicing mechanism would leave virtually no trace, making it almost impossible to disprove. Furthermore, as reverse splicing is dependent on the splicing machinery, it seems almost impossible that the first introns would have been gained through such a mechanism. There are yet other theories of intron gain, including recombination between homologous copies of genes (Venkatesh et al. 1999) and the creation of new splice sites within exons via single nucleotide mutations (Wang et al. 2004), but these mechanisms are also very difficult to prove or disprove. It is also possible that recent gains occur through an entirely different mechanism than did ancestral introns, and thus, understanding the mechanism of recent gains might not be the ultimate answer to understanding the origin of most spliceosomal introns.

This genome-wide analysis in flies allowed us to confirm previous findings about the nature of lost introns. It further supports the cDNA recombination model of intron loss and provides a good estimate of how common this event is in Drosophila species and how this rate can vary wildly between different intronic sites and different species. Perhaps more surprisingly, our data strongly support that intron gain did occur in the Drosophila lineage, as recently as 8–10 MYA, and provide some hints concerning the nature of the gain mechanism.

Supplementary Material

Supplementary table is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to thank Dr. Daniel Gaffney for critical reading of the manuscript. This research was supported by funds from the Canadian Institute of Health Research and the Canada Research Chairs program.

Literature Cited

Adams MD, Celniker SE, Holt RA, et al. (194 co-authors). 2000. The genome sequence of Drosophila melanogaster. Science. 287:2185–2195.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Arango NA, Huang TT, Fujino A, Pieretti-Vanmarcke R, Donahoe PK. 2006. Expression analysis and evolutionary conservation of the mouse germ cell-specific D6Mm5e gene. Dev Dyn. 235:2613–2619.

Blanchette M, Kent WJ, Riemer C, et al. (12 co-authors). 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 14:708–715.

Chatterji S, Pachter L. 2006. Reference based annotation with GeneMapper. Genome Biol. 7:R29.

Clark AG, Eisen MB, Smith DR, et al. (416 co-authors). 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 450:203–218.

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

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

de Souza SJ. 2003. The emergence of a synthetic theory of intron evolution. Genetica. 118:117–121.

Derr LK, Strathern JN, Garfinkel DJ. 1991. RNA-mediated recombination in S. cerevisiae. Cell. 67:355–364.

Gaffney DJ, Keightley PD. 2006. Genomic selective constraints in murid noncoding DNA. PLoS Genet. 2:e204.

Karolchik D, Baertsch R, Diekhans M, et al. (13 co-authors). 2003. The UCSC Genome Browser Database. Nucleic Acids Res. 31:51–54.

Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica. 115:49–63.

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.

Logsdon JM Jr, Stoltzfus A, Doolittle WF. 1998. Molecular evolution: recent cases of spliceosomal intron gain? Curr Biol. 8:R560–R563.

Long M, de Souza SJ, Rosenberg C, Gilbert W. 1998. Relationship between “proto-splice sites” and intron phases: evidence from dicodon analysis. Proc Natl Acad Sci USA. 95:219–223.

Long M, Rosenberg C. 2000. Testing the “proto-splice sites” model of intron origin: evidence from analysis of intron phase correlations. Mol Biol Evol. 17:1789–1796.

Majewski J, Ott J. 2002. Distribution and characterization of regulatory elements in the human genome. Genome Res. 12:1827–1836.

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

Myers EW, Sutton GG, Delcher AL, et al. (28 co-authors). 2000. A whole-genome assembly of Drosophila. Science. 287:2196–2204.

Nakao A, Yoshihama M, Kenmochi N. 2004. RPG: the Ribosomal Protein Gene database. Nucleic Acids Res. 32:D168–D170.

Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. PLoS Genet. 2:e173.

Putnam NH, Srivastava M, Hellsten U, et al. (19 co-authors). 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science. 317: 86–94.

Richards S, Liu Y, Bettencourt BR, et al. (51 co-authors). 2005. Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. Genome Res. 15:1–18.

Rogozin IB, Sverdlov AV, Babenko VN, Koonin EV. 2005. Analysis of evolution of exon-intron structure of eukaryotic genes. Brief Bioinform. 6:118–134.

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, Gilbert W. 2005. The pattern of intron loss. Proc Natl Acad Sci USA. 102:713–718.

Roy SW, Hartl DL. 2006. Very little intron loss/gain in Plasmodium: intron loss/gain mutation rates and intron number. Genome Res. 16:750–756.

Roy SW, Penny D. 2006. Smoke without fire: most reported cases of intron gain in nematodes instead reflect intron losses. Mol Biol Evol. 23:2259–2262.

Russell CB, Fraga D, Hinrichsen RD. 1994. Extremely short 20-33 nucleotide introns are the standard length in Paramecium tetraurelia. Nucleic Acids Res. 22:1221–1225.

Slaven BE, Porollo A, Sesterhenn T, Smulian AG, Cushion MT, Meller J. 2006. Large-scale characterization of introns in the Pneumocystis carinii genome. J Eukaryot Microbiol. 53(Suppl 1):S151–S153.

Sterner DA, Carlo T, Berget SM. 1996. Architectural limits on split genes. Proc Natl Acad Sci USA. 93:15081–15085.

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.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Venkatesh B, Ning Y, Brenner S. 1999. Late changes in spliceosomal introns define clades in vertebrate evolution. Proc Natl Acad Sci USA. 96:10267–10271.

Vibranovski MD, Sakabe NJ, de Souza SJ. 2006. A possible role of exon-shuffling in the evolution of signal peptides of human proteins. FEBS Lett. 580:1621–1624.

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.

Yoshihama M, Nguyen HD, Kenmochi N. 2007. Intron dynamics in ribosomal protein genes. PLoS ONE. 2:e141.

Zhuo D, Madden R, Elela SA, Chabot B. 2007. Modern origin of numerous alternatively spliced human introns from tandem arrays. Proc Natl Acad Sci USA. 104:882–886.

Aoife McLysaght, Associate Editor

Accepted October 22, 2007