Intron Loss and Gain in Eukaryotes

by

Jasmin Coulombe-Huntington

Department of Human Genetics,
McGill University, Montreal

Thesis submitted to McGill University in partial fulfilment
of the requirements for the degree of

Master of Science
Human Genetics

Submitted in January 2008

Table of contents

Preface

The following thesis entitled Intron Loss and Gain in Eukaryotes is based on two manuscripts: "Characterizing Intron Loss Events in Mammals" (Coulombe-Huntington and Majewski 2007), published in the January 2007 issue of Genome Research and "Intron Loss and Gain in *Drosophila*", which was recently published in the December 2007 issue of Molecular Biology and Evolution. I would like to thank first and foremost my supervisor and co-author on both papers, Prof. Jacek Majewski. I would also like to thank the other members of my supervisory committee, Prof. Paul Harrison and Prof. Mathieu Blanchette for their useful guidance and criticisms as well as Dr. Daniel Gaffney for critical reading of the manuscript Intron Loss and Gain in *Drosophila*. This research was supported by funds from the Canadian Institute of Health Research and the Canada Research Chairs program.

Abstract

Although introns were first discovered almost 30 years ago, their evolutionary origin and function remains elusive. In this thesis, I describe a referenced-based intron mapping method based on multi-species whole-genome alignments. We applied this method in two distinct studies. First we studied intron loss and gain dynamics in mammals and subsequently in *Drosophila*. We mapped known human introns onto the mouse, rat and dog genomes, mouse introns onto the human genome and *Drosophila melanogaster* introns onto 10 other fully sequenced *Drosophila* genomes. This genome-wide approach allowed us to assess the presence or absence of over 150,000 known human introns across four mammalian species and more than 35,000 *D. melanogaster* introns across 11 fruit fly species. We inferred 122 intron loss events in mammals and no intron gain events. In flies, we were able to identify 1754 intron loss events and 213 gain events. In both studies we found that lost introns tend to be extremely short and show higher than average similarity between their 5' splice-site sequence and the 3' partner splice-site sequence. We also demonstrate that losses in mammals occur preferentially in highly expressed house-keeping genes, while in *Drosophila* we show that lost and gained introns are flanked by longer than average exons, display quite distinct phase distributions and losses demonstrate significant clustering within genes. Across flies, it appears introns that have been lost evolve faster than other introns while they occur in slowly evolving genes. Our results in both studies strongly support the cDNA recombination mechanism of intron loss. The results in flies also suggest that selective pressures affect site-specific loss rates and show that intron gain has occurred within the *Drosophila* lineage, solidifying the "introns-middle" hypothesis and providing some hints about the gain mechanism and origin of introns.

Résumé

Malgré le fait que les introns furent découverts il y a près de 30 ans, leur origine et leur fonction nous échappent encore. Au cours de cette thèse, je décrirais une méthode qui permet de projeter des introns d'une espèce de référence sur d'autres génomes, basée sur des alignements de génomes complets à plusieurs espèces. Nous avons appliqué cette méthode dans le cadre de deux études distinctes. Premièrement, nous avons étudié les pertes et les gains d'introns chez les mammifères et ensuite chez les Drosophiles. Nous avons projeté les introns humains sur le génome de la souris, du rat et du chien, les introns de la souris sur le génome humain et les introns de la *Drosophile melanogaster* sur les génomes de 10 autres espèces de Drosophiles complètement séquencées. Cette approche d'ordre génomique nous a permis de comparer la présence ou l'absence de plus de 150,000 introns humains dans quatre espèces de mammifères et plus de 35,000 introns de *D. melanogaster* dans 11 espèces de drosophiles. Nous avons détecté 122 pertes d'introns chez les mammifères mais aucun gain d'intron. Chez les mouches à fruits, nous avons identifié 1754 pertes d'introns et 213 gains d'introns. Dans les deux études, nous démontrons que les introns perdus sont extrêmement courts et démontrent une similarité relativement élevée entre le site d'épissage au début de l'intron et le site d'épissage à la fin de l'intron. Nous démontrons chez les mammifères les pertes d'introns se produisent de préférence dans des gènes hautement exprimés et de fonctions cruciales à la cellule. Chez les drosophiles nous démontrons que les introns perdus ou gagnés sont délimités par des exons plus longs que la moyenne, ont une distribution de phase plutôt distincte et les pertes démontrent une tendance à se retrouver en groupe à l'intérieur des gènes. Chez les mouches à fruits, il semble que les introns perdus évoluent plus rapidement que la moyenne, tout en se concentrant dans des gènes qui évoluent plus lentement que la moyenne. Nos résultats des deux études supportent fortement le mécanisme de perte d'intron par recombinaison avec une molécule de cDNA. Nos résultats chez les drosophiles suggèrent que la probabilité de perte d'intron varierait d'un site à l'autre dû à des pressions sélectives et démontrent aussi que des nouveaux introns ont apparus au cours de l'évolution des mouches à fruits, ce qui solidifie la théorie des "introns-milieu" et procure des indices quant au mécanisme d'apparition des introns et leur origine.

Introduction

Exons, which code for both the amino acid sequence of the protein and the untranslated region of the processed mRNA, make up only a relatively small part of eukaryotic genes. The greater part of genes is composed of introns, which intersect the exons and code neither for any part of the mRNA or the protein product. After transcription, the introns are entirely removed by the splicing machinery. Introns account for approximately 24% of the human genome (Venter et al. 2001) and, almost 30 years after their discovery, both their function and their origin remain poorly understood. Some introns are believed to play a role in transcriptional regulation and many alternative splicing events would be impossible without introns. However, these functions justify very little of the >150,000 introns present in the human genome, most spanning over thousands of bases. Additionally, their level of interspecies conservation varies between and within genes, suggesting there are selective pressures related to yet unknown biological functions (Gaffney and Keightley 2006; Majewski and Ott 2002).

As to where and how introns first appeared, there is a plethora of different theories. It is still unclear whether their expansion occurred strictly in early eukaryotic or even pre-eukaryotic 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 which 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). All of these theories fall into the insertional theory of introns (Cavalier-Smith 1991; Palmer and Logsdon 1991), as opposed to the formative theory, whereby introns were created simply as a byproduct of exon shuffling (Gilbert 1987).

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 insight 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, Nguyen and Kenmochi 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 million years 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, Stoltzfus and Doolittle 1998; Roy and Penny 2006). Technical issues aside, it appears losses, although rare, occur at a measurable rate in most eukaryotes, while intron gains, at least in the last 100 Myrs, seem to be restricted to specific clades. 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 germline 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, Strathern and Garfinkel 1991). Many studies have demonstrated that lost introns display characteristics that support this mechanistic model (Mourier and Jeffares 2003; Roy and Gilbert 2004; Sverdlov et al. 2005; Roy and Hartl 2006), such as small size, 3’ positional bias and enrichment in highly expressed genes.

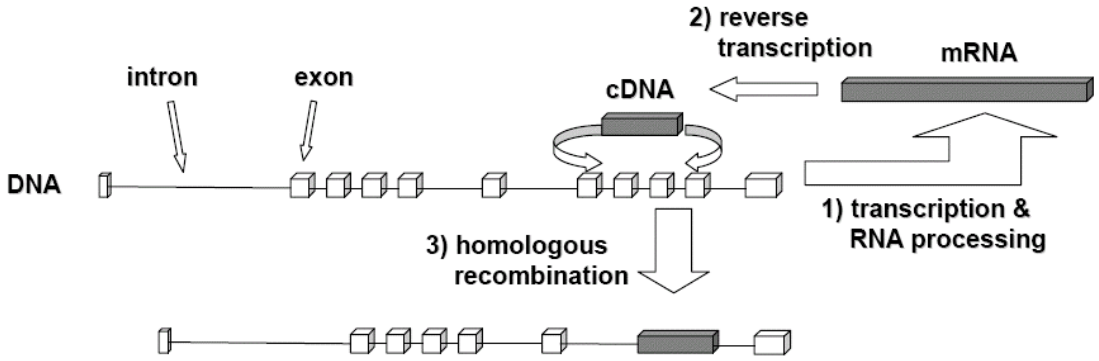

Figure 1: The most widely accepted model of intron loss, whereby an intronless cDNA recombines with the genomic version of the gene, deleting one or more introns

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. Also, as fully sequenced genomes become increasingly well annotated with gene positional information, we can begin to look at intron dynamics on a genome-wide scale.

This thesis presents data from two distinct studies on intron dynamics, divided into chapters 1 and 2. In the first chapter, "Characterizing Intron Loss Events in Mammals", we compared introns across human, mouse, rat and dog for more than 17,000 genes. We found and characterized 122 intron loss events but no gain events. Chapter 2, entitled "Intron Loss and Gain in *Drosophila*", looks at intron loss and gain events across 11 fruit fly species at >37,000 intronic sites. In this second study, we found 1754 intron losses and 213 gains. In both studies, we used MultiZ multi-species whole-genome alignments (Blanchette et al. 2004) to map introns from a reference species to other species, allowing us to study almost every intron in the reference species genome. If we depended on the prior annotation of genes in every species, like previous intron dynamics studies, the number of genes in the analyses would be severely reduced by the lack of annotation in most species. The two studies differ slightly in their approach, mostly due to the considerable difference in the number of loss and gain events studied and differences in

the types and quality of available datasets. The fact that these studies are genome-wide provides 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.

Chapter 1: Characterization of intron loss events in mammals

INTRODUCTION

Here we make use of the complete, high quality genomic sequences of four mammalian species (human, mouse, rat, and dog) to investigate intron gain and loss dynamics in mammals. We utilize a gene mapping technique to map annotated reference human genes onto genome-wide, multi-species sequence alignments, allowing us to investigate the predicted intron-exon boundaries of 152,146 introns within 17,242 autosomal genes. A recent study which considered a much smaller number of mammalian genes (Roy et al. 2003) uncovered 6 differences in intron positions between human and rodents, and suggested that there is no evidence for intron gain, and a very slow rate of intron loss in mammals. We detect over 100 cases of intron loss and still no evidence for any intron gain during mammalian evolution. Our large sample size allows us to determine the relative rates of intron losses in mammalian lineages and characterize the types of introns and genes that appear susceptible to loss, providing us with new insight regarding the mechanisms of intron deletion.

RESULTS

We used the mapping of annotated human exon-intron boundaries onto the mouse, rat and dog genomes to detect changes in gene architecture that occurred during the evolution of the four mammalian species. This approach makes use of the highest quality gene annotation (17242 human genes), but it allows us only to detect either intron loss events that occurred in rodent and dog, or intron gain events that occurred in the human lineage. Thus, we also employed the reverse approach: mapping known mouse genes onto mouse/human whole genome alignments. The latter strategy results in a slightly smaller dataset (16068 mouse genes) but allows us to detect intron losses in the human, and intron gains in the mouse genome.

We were able to uncover a total of 120 isolated changes: 4 occurring in human, 29 in mouse, 46 in rat, 34 in the rodent lineage prior to the mouse/rat divergence, and 7 in dog. Remarkably, all of the changes were consistent with a loss, rather than a gain of an intron; that is, for each case of a deletion of an intron relative to the reference gene structure (either mouse or human), the annotated intron was present in an earlier diverged organism. The loss of each intron was verified by using dog as the outgroup for changes occurring in human, mouse or rat, and using chicken as the outgroup for changes occurring in dog.

Figure 2 shows an example of an intron deletion event occurring in mouse displayed in the vertebrate MultiZ alignment track of the UCSC genome browser. This case illustrates common misalignments close to the splice sites, which is the reason we allowed for a 25 bp margin of error in the distance between exon edges in the target species during the search for intron loss and gain (see Methods). To confirm each gain/loss event, we extracted the original genomic sequences from the assemblies and used ClustalW to re-align the reference species intron and 100 bp of flanking upstream and downstream sequences with the homologous target species region. Our analysis shows that at least 117 of the detected intron losses are exact. The remaining 3 cases are also likely to be exact losses but fall into regions of relatively poor quality genomic sequence and require single base insertion/deletion events in the alignments.

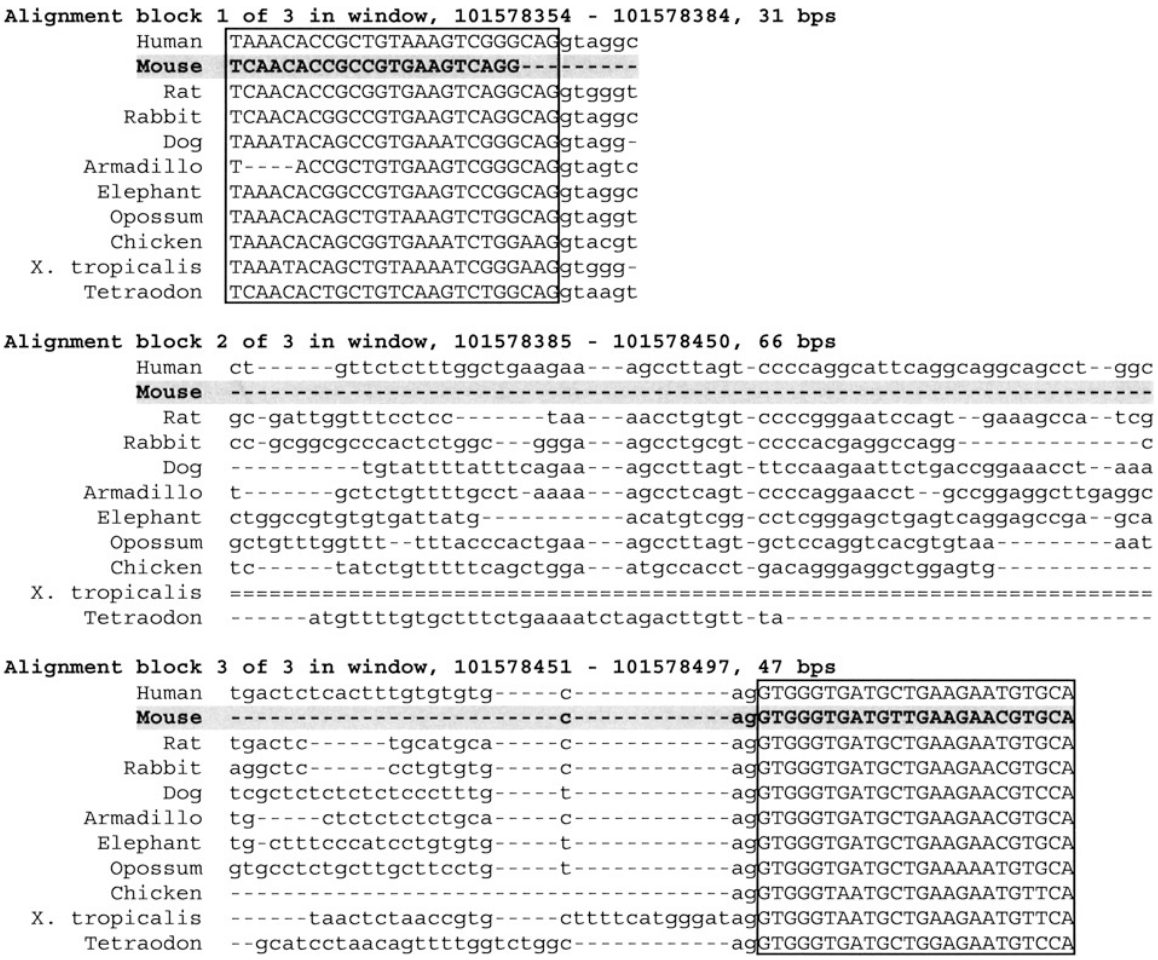

Figure 2: An example of intron loss in the DYNC1H1 gene in mouse visualized in the UCSC Genome Browser display of multi-species alignments. Uppercase, boxed sequences correspond to exons. Note that the alignment is inexact at the splice sites, resulting in an artefactual 3 bp intron length in mouse, which necessitates an approximate search strategy (described in the Methods), and realignment of sequences using an appropriate parameter choice in order to confirm all candidate intron deletions.

Rates of intron loss/gain

We find a very low rate of intron loss throughout the mammalian evolution and no evidence for intron gain. Based on the total number of donor/acceptor splice site pairs identified in the alignments (146,964 for mouse; 141,942 for rat; 146,727 for dog; and 124,474 for human) we determined the rates for intron loss per million years per intron

as: $5.32 \cdot 10^{-6}$ for the mouse-rat common ancestor, $6.58 \cdot 10^{-6}$ for mouse, $1.08 \cdot 10^{-5}$ for rat, $5.30 \cdot 10^{-7}$ for dog and $4.28 \cdot 10^{-7}$ for human. These estimates assume that human and dog lineages diverged 95 MYA, human and rodent 75 MYA (Waterston et al. 2002), and mouse and rat 30 MYA (Nei et al. 2001; Springer et al. 2003). In order to assess whether the rates are proportional to generation time, we multiplied these rates by the age of sexual maturity of each organism (1/6, 1/3, 3, and 12 years for mouse, rat, dog, and human), and normalized the resulting figures, so that the rate for human is equal to 1. (Note that we are making a somewhat simplistic assumption that the ages behaved proportionally during the evolution of each lineage). From the human ratio of 1, we obtain ratios of 0.21 for mouse, 0.70 for rat and 0.31 for dog, indicating that generation time is not the only factor affecting the rate of intron loss. As we expect the loss rate to be proportional to reverse transcriptase activity and the two known sources of endogenous non-telomeric reverse transcriptase activity are LINEs (long interspersed nuclear elements) and retroviruses, we attempted to correlate the loss rate to the number of LINE elements (Smith-Waterman conservation score >25,000) identified in the newest Repeat Masker annotations of each species, downloaded from the UCSC Genome Browser database. We see an almost perfect correlation coefficient of 0.98, but four species are not enough to assess the significance of this result. Other possible factors may include the effective population size of each lineage, which could lead to differences in the probabilities of fixation of the changes.

Sizes of deleted introns

One of the most striking characteristics distinguishing the deleted introns is their extremely small size. The mean size of a human intron is 6259 bp, while the deleted cases were on average 355 bases long in human. Figure 3 illustrates the difference in the size distribution of deleted introns and that of all introns. The difference in the distributions is highly statistically significant ($t = -57.3$, $df = 208$, $p < 10^{-10}$). Most of the deleted introns (81 out of 120) are smaller than 150 bases. We further investigated five cases of unusual intron deletions that exceeded 1000 bp in length (5968, 7100, 1030, 1158, and 4380 nucleotides in genes *FAM54A*, *TAF5L*, *FLJ11806*, *EEF2*, *CLTCL1*, respectively). Four of those cases occurred in the rat lineage and one in mouse. We identified the corresponding

intron in the closest relative (mouse, in case the loss occurred in rat, and *vice versa*) and observed that the introns in the closest relative were actually considerably shorter than in human (984, 4902, 224, 134, and 79 bases respectively) and hence were likely to be short at the time of their deletion. This suggests that the size of the intron must be an important factor affecting the underlying molecular mechanism of deletion.

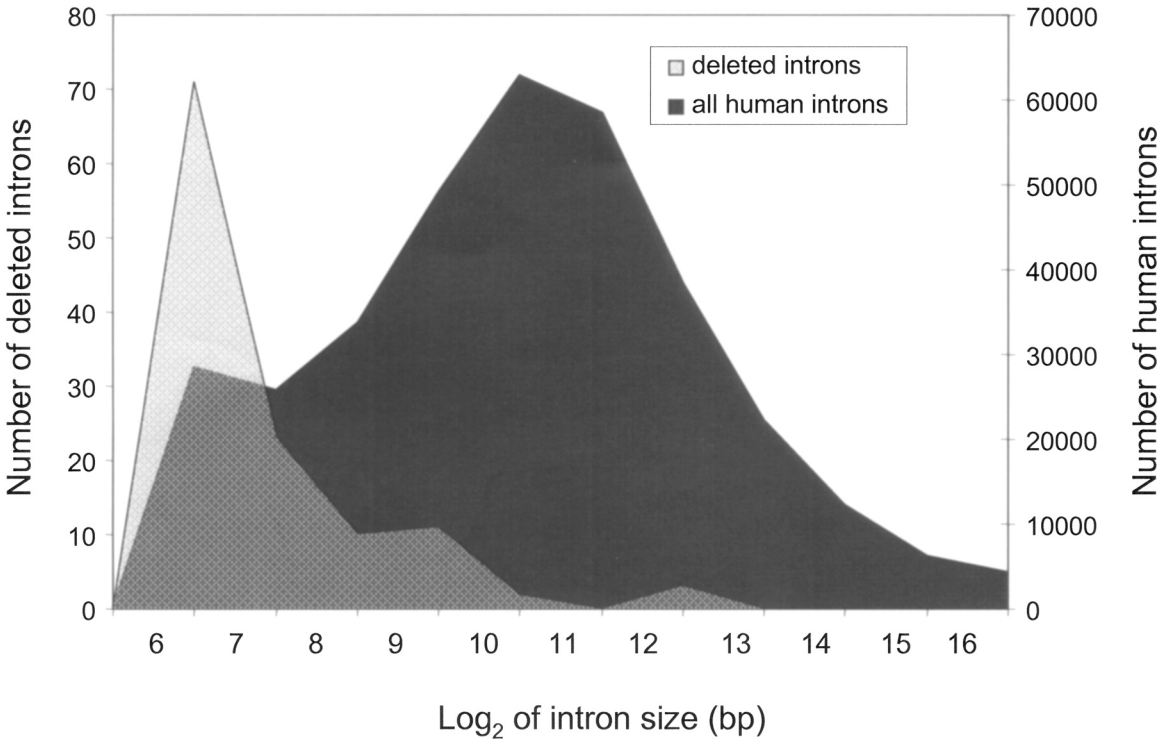

Figure 3: Log- size distribution of all introns (black) versus deleted introns (grey). The deleted introns are unusually short and much shorter than the human genome average.

Intron phases

Introns can be classified as phase 0 (inserted between two codons), phase 1 (after the first base of a codon), or phase 2 (after the second base). We examined the phase distribution of the 116 deleted introns from Table 1 and compared it to the phase distribution of all introns from the RefSeq dataset. The proportions for the deleted introns were 0.52, 0.26 and 0.22 for phases 0, 1 and 2 respectively, while the ratios for the genome average were 0.46, 0.32 and 0.22. The distribution of phases of deleted introns did not differ

significantly from the expected (χ^2 = 2.24, df = 2, p = 0.33). Intron deletions in mammals appear to occur randomly with respect to their phase.

Positions of deleted introns within genes

We defined the relative position of each intron within a gene as its ordinal number from 5' to 3', *n*, divided by the total number of CDS introns, *N*. In order to obtain a symmetrical distribution centered about ½, we subtracted ½ from the numerator. Hence, the adjusted relative position, $(n - ½)/N$ has a range between $1/(2N)$ and $1 - 1/(2N)$, and an expectation of ½. We used a χ^2 test to compare the proportion of deletions in the 5' half of the gene versus the 3' half. We found that the positions of deleted introns were significantly skewed toward the 3' half of each gene (χ^2 = 7.76, df = 1, p= 0.0053). 73 of the intron losses occurred closer to the 3' end of genes, as compared to 40 which were closer to the 5' end. As 5' introns are known to be generally longer than 3' introns we performed a multi-variate correlation between intron position, length and whether the intron had been lost or not (data not shown). Although intron size did decrease from 5' to 3', the relation between position and probability of loss was still significant.

Splice site characteristics

We examined the distributions of bases around both splice sites and compared it to the distributions for all introns. We found that the consensus at the 5' splice site was not significantly different from the control. However, at the 3' splice site, the two positions following the acceptor AG dinucleotide had a significantly greater frequency of the bases G (χ^2 = 3.82, df = 1, p = 0.05) and T (χ^2 = 4.93, df = 1, p = 0.03), respectively. Since the GT at the 3' splice site is very often preceded by an AG, this stronger consensus sequence may have served to promote a recombination event occurring between the two splice sites, leading to the deletion of the intron.

Expression patterns and ontology of genes undergoing intron loss

We used the EASE (Hosack et al. 2003) interface to classify our genes into GO categories (biological process) and characterize the types of genes that undergo intron deletion events. EASE calculates over-representation statistics for each GO category

using an EASE score, which approximates a p-value by using the upper bound of the distribution of Jackknife Fisher exact probabilities. In Table 1, we list the most over-represented biological processes (EASE score < 0.05). We note that most of the genes with intron deletions are involved in biosynthesis, metabolism, translation, transcription, and RNA processing. All of the over-represented categories correspond to ubiquitous housekeeping functions, suggesting that intron deletion events occur predominantly in genes that are both highly expressed and expressed in the germline. In order to further confirm this hypothesis, we utilized microarray expression data available from SymAtlas (Su et al. 2002) to determine the expression intensities and breadths of the candidate genes. Since germline gene expression levels are not known, we used averaged gcRMA (Robust Multichip Average with GC correction) expression over all tissues as a proxy of germline expression (Majewski 2003), and compared the averages of the intron-deleted sample to all genes. The average gcRMA expression level was of 952 overall, and significantly higher, 9560, for the genes with intron deletions ($t = -4.3$, $df = 108$, $p = 3.58 \cdot 10^{-5}$). In order to study the breadth of expression, we used MASS 5.0 present/absent calls from more than 300 tissues and cell lines and determined the fraction of tissues where expression was positively detected (present). Again, we compared the expression breadth for genes with intron deletions (0.54) to that of all genes (0.26) and found a highly significant difference ($t = -6.9$, $df = 92$, $p = 7.15 \cdot 10^{-10}$). Thus intron deletions occur preferentially in genes with housekeeping functions, which have experimentally been determined to be both highly and broadly expressed.

Table 1: Over-represented GO Biological Processes

GO Biological Process	List Hits	Population Hits	EASE score [1]
protein biosynthesis	19	650	$6.2*10^{-7}$
biosynthesis	26	1199	$6.4*10^{-7}$
macromolecule biosynthesis	22	1002	$5.7*10^{-6}$
metabolism	75	7637	$2.7*10^{-5}$
translation	9	236	$2.7*10^{-4}$
Pol II promoter transcription	12	477	$5.6*10^{-4}$

nucleic acid metabolism	38	3429	0.003
RNA processing	10	430	0.0035
RNA metabolism	10	460	0.0054
protein metabolism	31	2696	0.0057
nucleocytoplasmic transport	5	108	0.0073
spermine biosynthesis	2	2	0.014
translational elongation	3	27	0.016
spermine metabolism	2	3	0.021
spermidine metabolism	2	3	0.021
spermidine biosynthesis	2	3	0.021
intracellular transport	10	613	0.03
transcription	26	2426	0.03
polyamine biosynthesis	2	7	0.049

[1] A p-value approximation, uncorrected for multiple testing, based on the number of hits within a category for our list of 99 genes (which could be identified from our dataset by their Locus Link ID), as compared to total hits within a population of 13802 genes (null expectation).

Genes experiencing frequent intron loss: *GAPDH*

We performed a detailed analysis of the *GAPDH* gene, where we found evidence of multiple, independent intron losses occurring in mouse, human, and rat. *GAPDH* is a known, very highly expressed housekeeping gene, which supports the hypothesis that expression in the germline is essential for intron loss. We used genomic DNA and mRNA *GAPDH* sequences for 15 vertebrate species, and performed multiple sequence alignments to reconstruct the intron/exon structure of the gene in each species (Figure 3). A Dollo parsimony approach (assuming a single appearance of the derived character – intron) reveals no gain events throughout vertebrates, but numerous losses, including several independent losses of the same intron (intron 9 of the ancestral gene). The result also suggests that the phenomenon of intron loss in vertebrates (at least within this gene) may be accelerated in the mammalian branch.

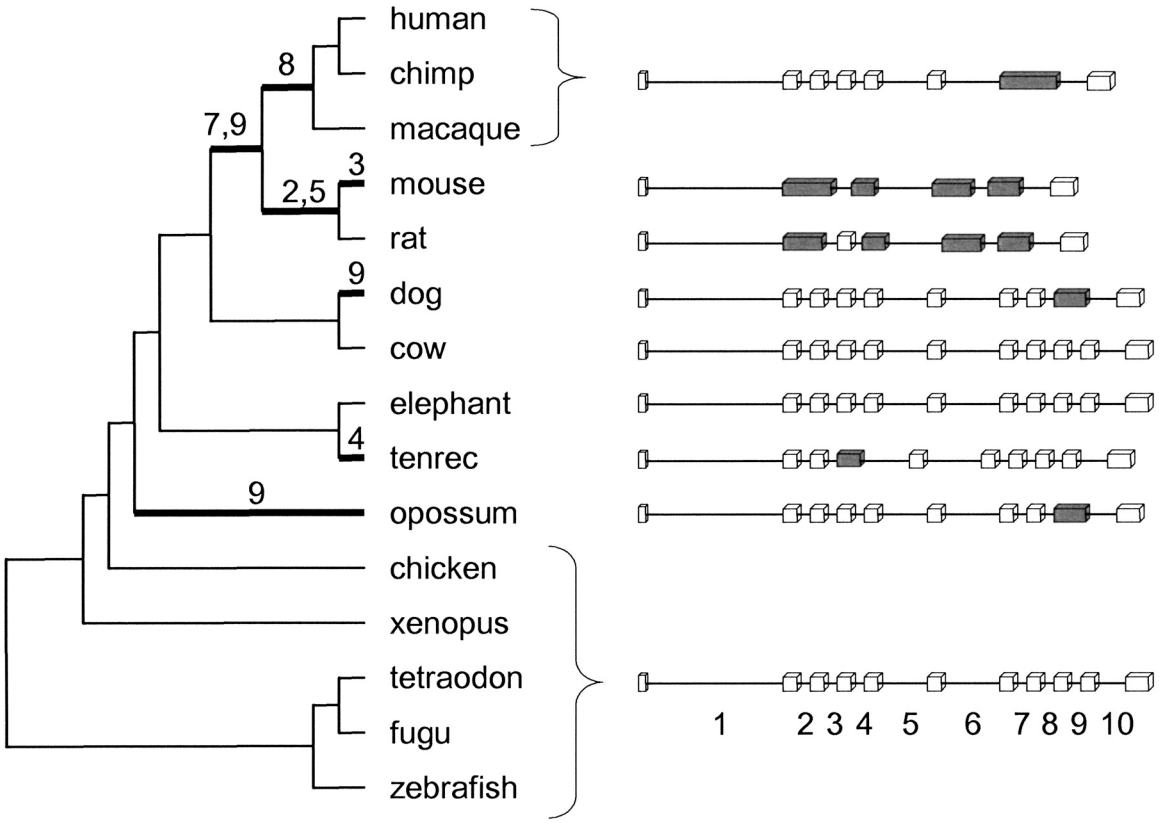

Figure 4: The evolution of the intron-exon structure of the *GAPDH* gene throughout the vertebrate phylogeny. The numbers on the branches indicate the inferred deletion events. The introns are numbered according to their position within the coding sequence of the ancestral gene.

Does intron loss disrupt alternative splicing?

We identified 2 cases of intron losses disrupting known (RefSeq-confirmed) alternative splicing (AS) events which alter the predicted amino acid sequence of the gene for details). We also detected 20 losses which disrupt predicted (EST-based) AS events. If intron losses occurred randomly, without any regard to preserving AS, the expected number, based on all the RefSeq introns used in this analysis is a disruption of 4 known events and 17 predicted events. It is unexpected that the number of observed losses that disrupt predicted AS events is actually slightly greater than the null expectation. However, since our ability to predict AS events is highly dependent on the availability of mRNAs and ESTs, and the set of genes undergoing intron losses is extremely highly and

broadly expressed, there is likely to be a bias in the annotation of the deleted sample. That is, because of their high expression levels, genes experiencing intron loss have deeper EST coverage and are better annotated with respect to AS than the genome average.

In view of the annotation bias, it is difficult to conclude whether the disrupted predicted AS events are truly functional or constitute an artefact of deep EST coverage and the presence of inadvertent splicing errors. It is also possible that, since AS may be only weakly conserved across species (Pan et al. 2005), a predicted disruption of AS in humans may have no effect on AS in the species where the deletion occurred.

DISCUSSION

We identify over 100 cases of intron loss in the four examined mammalian species. Our approach, based on mapping of known human genes to whole-genome sequence alignments of multiple species, allows us to utilize the annotation information from well studied model species, such as human and mouse, and predict gene structure in other, relatively poorly annotated species. Using our method, we recover all 6 intron deletion events detected in a smaller scale study (Roy et al. 2003), and extend previous conclusions regarding the patterns of intron loss in mammals. There are several remarkable characteristics of our dataset: 1) losses appear to occur almost exclusively for small introns; 2) essentially all of our examples of loss are consistent with an exact deletion event; 3) the loss events are biased towards the 3' ends of genes, but can be found at all positions; 4) genes that are associated with intron loss events are generally highly expressed and have housekeeping functions; 5) the rate of intron loss is related to the generation time and the number of conserved LINE elements in each species; 6) all of the differences in gene structure are consistent with intron loss events – no detectable intron insertions have occurred in human or mouse since the divergence of their lineages. Below, we discuss some implications of these findings.

Mechanism of intron loss

It has been suggested that introns loss may be mediated either by genomic deletion events or recombination of the genomic locus with a reverse-transcribed, processed mRNA molecule of the gene (Logsdon et al. 1998). Our analysis suggests that at least 98% (and possibly all) of the observed deletions are exact. In addition, we do not find any evidence for inexact deletions, which would retain a small part of the intron or remove parts of neighbouring exons. It has been argued (Roy et al. 2003) that random genomic deletion events would be unlikely to always result in exact intron losses. This is even more evident in our large dataset. It would be extremely unlikely that, if intron loss were generally mediated by random deletions, we would fail to recover any cases of inexact losses. Even in the presence of purifying selection against such potentially deleterious events, it seems plausible that some minor insertion/deletions of the boundary sequence, particularly ones that do not alter the reading frame, would be evolutionarily neutral. Thus the exact character of the detected intron loss events supports the latter model, namely recombination with an intronless cDNA of the gene.

The small size of the introns provides another insight into the mechanism of loss. It is well documented that genetic recombination events occur less frequently in the presence of mismatches, insertions, or deletions within the recombining substrates (Majewski and Cohan 1999). We propose that in the cases of intron loss, recombination with cDNA is much more likely if the introns are small, resulting in a high relative effective proportion of sequence identity.

We also find that genes susceptible to intron loss tend to be involved in housekeeping functions and expressed at relatively high levels. Again, high level of expression most likely results in relatively high levels of reverse-transcribed copies of the gene, leading to an increased probability of recombination. A similar effect has been demonstrated for the frequency of processed pseudogenes (Zhang et al. 2003). Furthermore, in order for the recombination events to result in intron losses that are transmitted to the next generation and have a chance to increase in frequency in the population, the loss events must occur in the germline, and not be restricted to somatic cells. Thus, germline expression of the gene would be an essential condition for intron loss. In accordance with this prediction, we find that our intron-deleted dataset is highly

enriched in housekeeping (ubiquitously expressed) genes. Thus both the expression levels and the expression patterns of the genes support the recombination-mediated model of intron loss.

Finally, we find that the position of the lost introns is significantly biased towards the 3' ends of the genes. This is in accordance with recent studies of lower eukaryotes (Roy and Gilbert 2005b; Sverdlov et al. 2004) and again supports intron loss being mediated by recombination, since reverse transcription of the mRNA is believed to occur preferentially from the 3' end (Weiner et al. 1986). However, this result may also reflect the bias in distribution of intron sizes (first introns are generally longer and more difficult to remove by recombination) and selective pressures against deleting potentially regulatory regions, which may be present close to the 5' termini of genes (Majewski and Ott 2002).

Selection favouring intron loss?

The preferential intron loss in highly expressed housekeeping genes is also consistent with selection for transcription efficiency favouring the resulting short transcript (Castillo-Davis et al. 2002). While the selection pressure and the increased likelihood of recombination in highly transcribed genes are not mutually exclusive and, in theory, may both contribute to the association of intron loss and expression levels, it seems unlikely that this type of selection is a major force responsible for the observed intron losses. Most of the deleted introns are extremely short (~100 bp) while much longer introns are present in the corresponding genes and had not been deleted. Selection alone would favour the loss of longer introns. In the example of the *GAPDH* gene, a loss of an 82 bp intron from a 3783 bp transcript would result in only a very modest 2% decrease in the time of transcription. In comparison, loss of the first intron fully contained within the CDS (~1700 bp) could result in a 45 % reduction. We propose that the availability of cDNA and the length of unmatched intronic sequences in the recombining strands are the primary limiting factors in the process of intron loss. Once the gene conversion event occurs, selection may be an additional force increasing the probability of fixation of such events.

Rates of intron loss and gain

The rate of intron loss in mammals appears extremely slow. The fastest genome-wide rate, in the rat lineage is approximately 1 intron loss per 1.53 million years. We note that the rates are not clocklike and appear to be dependent on the generation time of each lineage: rodent > dog > human, but they are also likely to be affected by other factors, such as the effective population size. At the current rate, it would take more than 10^{12} years for the human genome to shed half of its introns. Hence, intron loss/gain does not appear to be a major factor in mammalian evolution.

Since we have not detected any cases of intron gain, we estimate the process to proceed considerably slower than intron loss. Our approach would allow us to detect intron gain events occurring in the mouse/rodent lineage, after its divergence from the human lineage. Since no cases of intron gain (and 63 losses) were found the genome-wide rate of gain is at least 60-fold slower than the rate of loss. Our observations support the premise that modern mammalian introns are evolutionarily inert and, having expanded through early eukaryotic genomes (or being inherited through even earlier ancestors), have been gradually, albeit very slowly, disappearing within mammalian lineages over at least the past 100 million years.

METHODS

DNA sequences and interspecies alignments

We used the RefSeq annotation of the human genomic sequence to extract coding sequences of human genes (Hinrichs et al. 2006). Only the sequences which could be *in silico* translated into their predicted protein were retained. This strategy resulted in a high confidence, non-redundant dataset of 17,242 human autosomal genes, containing 152,146 distinct introns within their coding sequences. The rationale for discarding the sex chromosomes is that the X chromosome is significantly enriched in processed pseudogenes (Drouin 2006) and the Y chromosome, containing only a few dozen genes, is the most difficult to sequence and map due to the lack of meiotic recombination over

most of its length (Bachtrog and Charlesworth 2001). We based our analysis on the four available highest-quality mammalian genome assemblies: human (hg17), mouse (mm7), rat (rn3), and dog (canFam2). We mapped the well annotated human genes onto the genome-wide alignments present within the 17-way MultiZ (Blanchette et al. 2004) alignment tracks in order to determine the intron-exon structures in the target species. We considered only introns that were flanked by coding, or partially coding, exons, since non-coding UTR sequences are poorly conserved (and often not conserved) among species, and provide poor anchors for detecting splice sites within alignments. We also performed the reverse analysis by mapping a set of 16,068 mouse RefSeq genes (129,336 CDS introns) onto the mouse vs. human genomic sequence alignments.

We used the following criteria to detect intron loss events in the target sequence (or gain in the reference sequence): 1) for each reference species intron we identified the positions of both the donor and acceptor splice sites within the MultiZ alignment; 2) within the target species, we flagged an intron as potentially lost if the distance between the donor and acceptor sites was lower than a predetermined cutoff of 25 bp. The latter condition was necessary since alignments are often imperfect at the exon-intron boundary (Figure 1). In particular, especially in the case of intron loss events, the last two base pairs of an exon, which have an AG consensus, tend to align with the downstream intronic acceptor site (also AG), but more serious misalignments are also common. Nevertheless, allowing a margin of 25 bp did not introduce any false positive results (as manually verified in the final curated results), since sequences shorter than 25 bp cannot be efficiently spliced in mammals (Lim and Burge 2001) and correspond to imperfect alignments, rather than actual introns. Using the above first pass search criteria, we identified 623 cases of potential intron loss/gain.

Since the genome assemblies and the resulting alignment contain numerous sequencing, assembly, and alignment artefacts, all potential intron loss events were further filtered based on the quality of the underlying alignment. In the process of constructing the BlastZ alignments, gaps in the sequences may be filled in using secondary (non-syntenic) sequences. This significantly increases the proportion of aligned sequences but also results in an increased probability of introducing alignment errors. Thus, only potential intron loss cases which mapped to the highest confidence,

top, syntenic, long (encompassing at least two neighbouring genes) alignment nets (Kent et al. 2003) were retained for further analysis. Cases occurring in genes which were aligned to multiple or non-syntenic portions of target genomes, which could potentially constitute alignments to duplicate genes or pseudogenes, were rejected. This strategy resulted in 157 cases of intron loss/gain, of which 35 occurred in both rat and mouse, for a total of 122 events.

For all the candidates, we extracted the sequence of 100 bp flanking the intronic site from the genomic sequence assembly and used ClustalW (Thompson et al. 1994) with high gap opening penalty (80) and low gap extension penalty (0) to align it to the human intron-containing sequence, and visualize the detailed evidence for intron loss. After performing some minor supervised adjustments, mainly correcting the misalignment of the terminal AG of an upstream exon with the downstream acceptor site (see above), this allowed us to confirm the deletion events and demonstrate that essentially all of the events are cases of exact deletion, with no alteration to the coding sequence. All of the 120 isolated intron loss/gain events were successfully validated using the sequence alignments (see the Supplementary Data for the manuscript Characterizing Intron Losses in Mammals [Coulombe-Huntington and Majewski 2007]).

Characterization of genes involved in loss events

In order to functionally classify the genes involved in intron loss events, we used the EASE (Hosack et al. 2003) interface to the Gene Ontology annotation. We identified the GO categories with the highest support – lowest EASE score – for over-representation by the genes within our list, as compared to all known genes.

In order to approximate expression levels and expression breadth of the genes, we used microarray expression data from SymAtlas (Su et al. 2002). Although the relevant variable is the expression level in the germline, this information is currently not available. As a proxy for gene expression levels, we used the mean values of gcRMA summaries across all tissues studied. Note that, because of developmental history of germ cells, testes and ovary-specific expression levels may not be the appropriate indicator of germline expression, and a global average expression may provide a better estimate (Majewski 2003), particularly in the case of housekeeping genes. As an estimate of

expression breadth, we used the present/absent calls from the MASS 5.0 summaries and, for each gene, calculated the percentage of tissues where expression was detected.

Intron loss and alternative splicing

In order to study the relationship of intron loss and alternative splicing, we cross-referenced the set of lost intron positions with alternative gene isoforms present in the RefSeq dataset. We identified all the introns where the deletion in the target species disrupts a known alternative splicing event in human. For example, deletion of an intron would prevent alternative usage of the adjacent exons (cassette events), as well as alternative (cryptic) splice site usage of the adjacent splice sites. We further limited the alternative splicing events of interest to only those that altered the predicted amino acid sequence of the gene. In order to obtain a background genome-wide estimate for the probability of any intron loss disrupting alternative splicing, we also determined how many introns from our entire input dataset border alternatively spliced exons that would be disrupted by a deletion.

While the RefSeq set of genes is manually curated and highly accurate, it contains relatively few alternatively spliced isoforms. Hence we also analyzed predicted alternative splicing events from the ExonWalk annotation of the UCSC database. Briefly, the ExonWalk program merges EST and cDNA evidence together to predict full length isoforms, including alternative transcripts. To predict transcripts that are biologically functional, rather than the result of technical or biological noise, ExonWalk requires that every intron and exon be either: 1) Present in cDNA libraries of another organism (i.e. also present in mouse), 2) Have three separate cDNA GenBank entries supporting it, or 3) Be evolving like a coding exon as determined by the Exoniphy program (Siepel and Haussler 2004). Once the transcripts are predicted an open reading frame finder is used to find the best open reading frame. Transcripts that are targets for nonsense mediated decay are filtered. We further filtered out all predicted transcripts that did not begin with an ATG and did not end with a stop codon.

Chapter 2: Intron Loss and Gain in *Drosophila*

INTRODUCTION

As our technique was quite successful in a mammalian study, we decided to extend it to *Drosophila*. 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 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.

RESULTS

Mapping introns

We were able to map 28,933 *D. melanogaster* introns onto every other species (see Methods). 1944 of these introns were missing from one or more species, assumed to be the result of a loss or gain event somewhere along the phylogenetic tree. 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 1754 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 which are ancestral to *D. melanogaster* (dashed lines

on Figure 5) and loss events on other branches. For events which happened on one of the two oldest branches, lacking an appropriate reference outgroup, we cannot distinguish between gains and losses. There are 220 such differences, as shown in Figure 5.

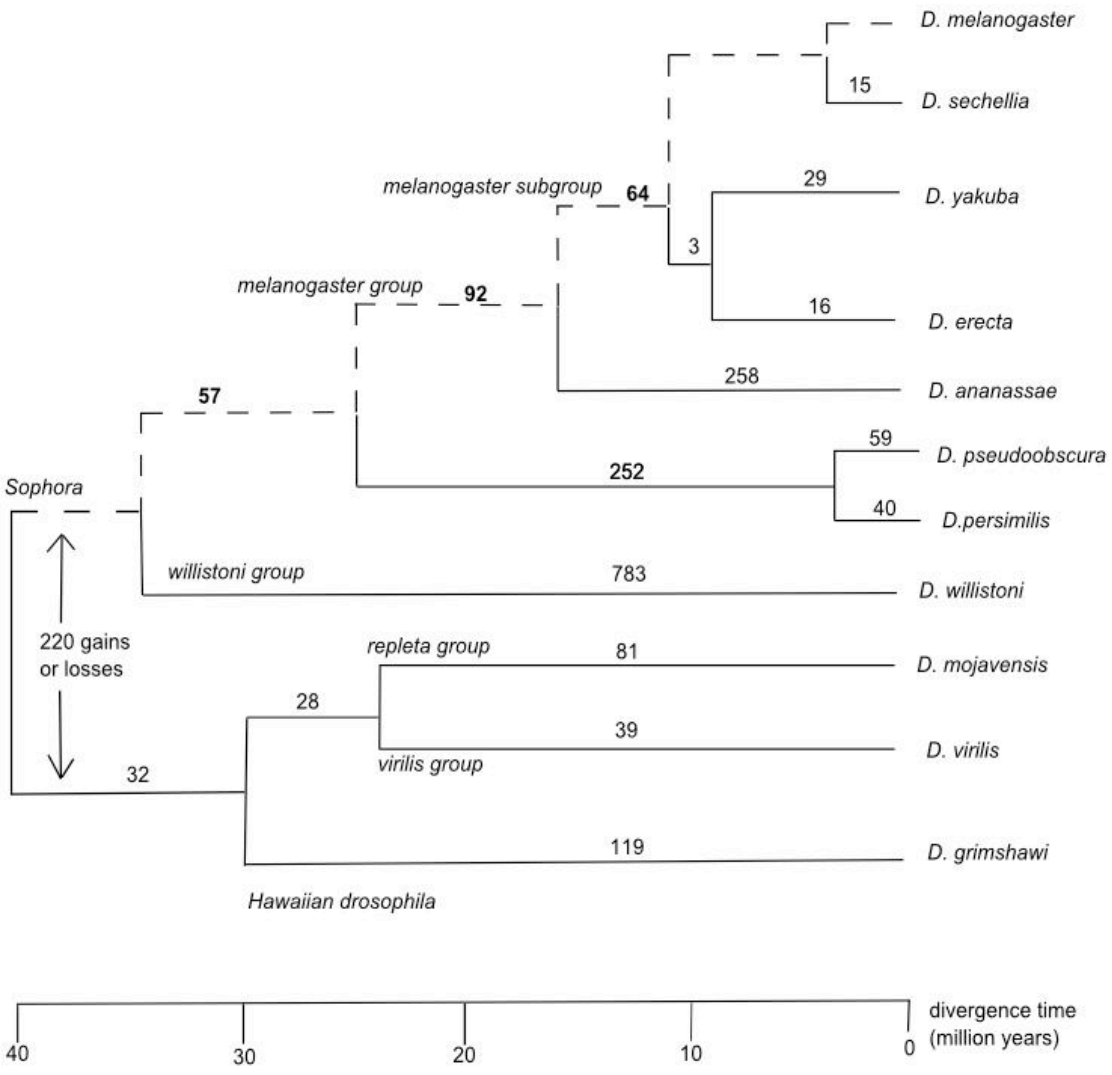

Figure 5: Number of predicted intron gains (in bold) and losses, as inferred using Dollo parsimony. The dashed lines indicate the branches where gains could be inferred and losses could not. The tree is based on an image from the Assembly/Alignment/Annotation of 12 related *Drosophila* species (Clark et al. 2007)

Varying loss rates

As shown in Figure 5, 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 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 since reverse transcriptase is involved in the mechanism and retrotransposons, as mentioned in Chapter 1, are one of the two known sources of endogenous reverse transcriptase activity. 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 or viruses has been involved with intron deletions.

Characterizing lost and gained introns

More than 80% of missing introns were shown to be precisely excised, leaving a functional looking splice-site pair (see Methods). Introns with missing orthologues, with a median length of 97 base pairs, are significantly shorter than average introns (t=22.4, 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=$4*10^{-5}$). We also demonstrate that pairs of adjacent losses occur more often than

expected by chance (χ^2=52.6, df=1, p-value-4*10^{-13}, see Methods for details). These characteristics of intron loss each provide support to the recombination model of intron 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 two 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 cDNA's 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*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. Unlike human genes, there was no information for *Drosophila* in the EASE database. Therefore, we could not look for overrepresentation of house-keeping functions. Although there is some publicly available microarray expression data for *D. melanogaster*, different datasets yielded contradictory results; some suggested that genes with losses were significantly more highly expressed while others suggested the opposite (data not shown). This is probably due to undocumented biases in the expression datasets, either at the experimental or analysis level.

Intron phases

Fruit fly introns, like human introns, are not evenly distributed over each of the three possible phases. 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 2. 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 (χ^2=38.2, df=2, p-value=5*10^{-9}) and gains (χ^2=50.9, df=2, p-

value=$8.7*10^{-12}$). We observe consistency in the phase preference of losses by demonstrating that the phase bias of the 143 introns that have been lost independently in two different branches follows the expected, more pronounced, distributional skew (χ^2=1.56, df=2, p-value=0.46) and the same goes for the 30 introns that have been lost in three independent lineages (χ^2=0.46, df=2, p-value=0.79). We simulated an evolutionary scenario which 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.

Table 2: Intron phase distributions in *D. melanogaster*

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

1: hypothetical introns inserted in exonic AGGT sites

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 6-a. Comparing the four bases centered on each 5' splice-site of lost introns to the four bases surrounding the 3' splice-site revealed that these sequence pairs were significantly more similar than they are for average introns (χ^2=106, df=4, p=$4.8*10^{-22}$). This difference was much more pronounced for gained introns (χ^2=337, df=4, p=$8.9*10^{-78}$). Figure 6-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 these four bases is AGGT. This characteristic is related to the A and G base preferences of lost introns demonstrated in the mammalian study. In fact, the same test applied to lost introns in mammals shows the same splice-site similarity (χ^2=8.5, df=4, p=3.5*10^{-3}), although less pronounced. 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 which 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 two guanines. The ratios are displayed in Table 2. As it comes fairly 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.

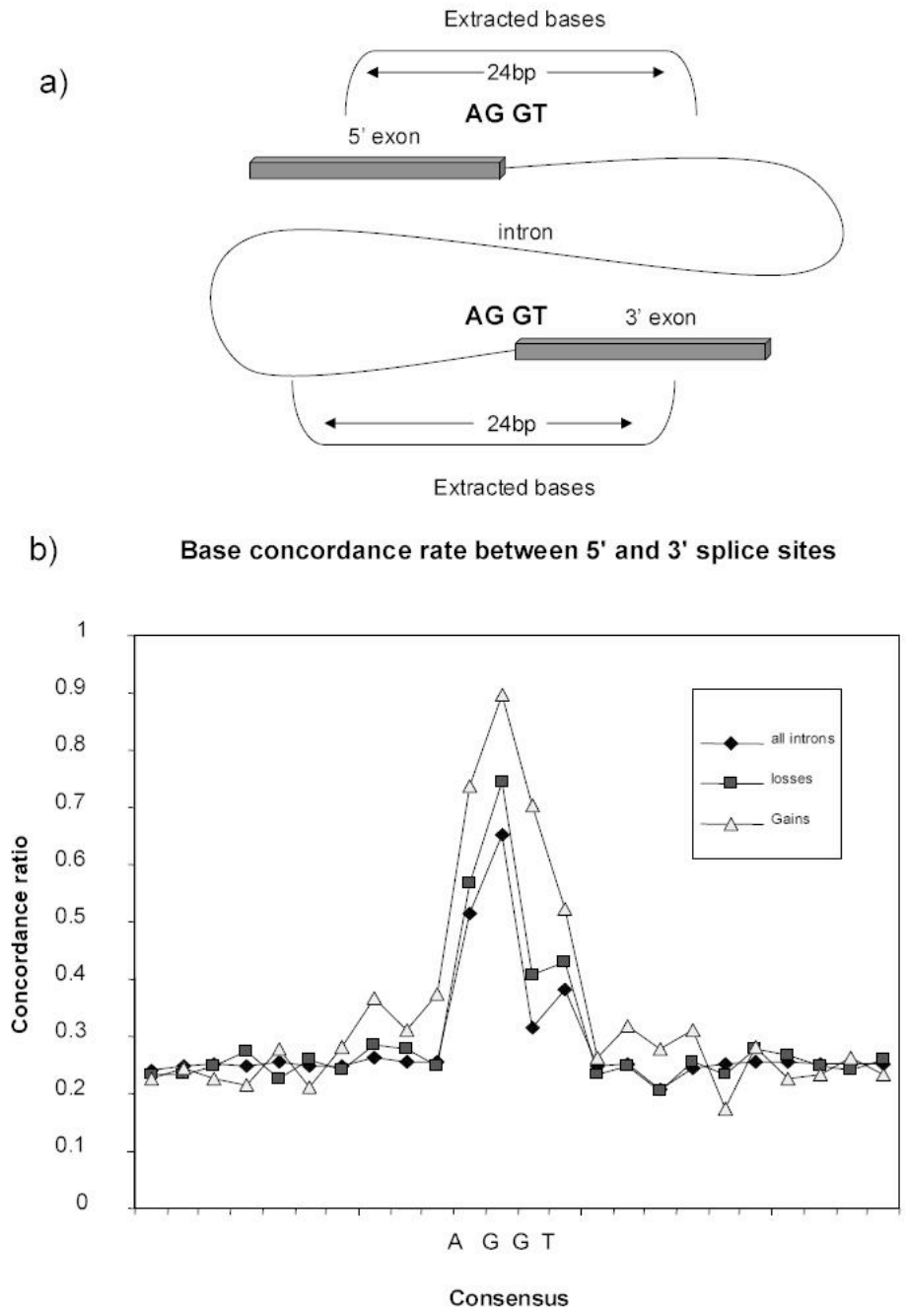

Figure 6: a) We aligned 24 bp centered on the 5' splice-site with the 24 bp centered on the 3' partner splice-site in the direct orientation. AGGT is the consensus motif of the four innermost bases at both splice-sites. **b)** Concordance ratio between each base surrounding the 5' splice site and the corresponding base, equally distanced from the 3'

splice site. The similarity of the two splice sites is significantly higher for the lost and gained categories than for the genome-wide average.

Intron loss, alternative splicing and selection

Most introns have no known function and appear to be useless pieces of genes which merely slow down the transcription process. Some however are required for their role in alternative splicing (AS) events and others are thought to play a role in transcriptional regulation. Based on *D. melanogaster* RefSeq annotations, we created a list of 4718 exons which underwent a putative AS event which 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 (χ^2=66.7, df=1, p=3.1*10^{-16}). We only found 14 cases of a loss disrupting a putative alternative splicing 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*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*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 co-variation 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., PNAS, 1996). This analysis

should allow us to asses 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 3. 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 co-varying 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 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.

Table 3: Intron Characteristics Multivariate Correlations to Loss and Gain

	Intron size (logged)	5' exon size (logged)	3' exon size (logged)	relative position	Splice-site similarity[1]	Intron conservation	Gene conservation
losses	-	+	+	+	+	-	+
	$p<2*10^{-16}$	$p=5.2*10^{-9}$	$p=1.3*10^{-10}$	p=0.79	$p<2*10^{-16}$	$p=1.6*10^{-12}$	p =0.0094
gains	-	+	+	-	+	-	-
	$p<2*10^{-16}$	$p=1.4*10^{-13}$	$p=8.6*10^{-6}$	$p=9.8*10^{-4}$	$p<2*10^{-16}$	p = 0.18	P = 0.88

1: number of matching bases out of four (see Figure 6)

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 which, at the time of writing of this thesis, has already reached its eighteenth 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 pair-wise protein sequence identity between the *D. melanogaster* gene and the predicted orthologues in every other species was >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 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, Yoshihama and Kenmochi 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 non-random subsets of genes, as are most analyses on intron dynamics to date.

Table 4: Multivariate Correlations in Highly Conserved Genes

	Intron size (logged)	5' exon size (logged)	3' exon size (logged)	relative position	Splice-site similarity[1]	Intron conservation	Gene conservation
losses	-	+	+	-	+	-	-
	$p<2*10^{-16}$	$p=1.9*10^{-7}$	$p=2.4*10^{-7}$	p=0.56	$p<2*10^{-16}$	p=0.0026	$p=8.7*10^{-12}$
gains	-	+	+	-	+	+	-
	p=0.0014	$p=7.1*10^{-5}$	p=0.57	p=0.18	$p=8.7*10^{-16}$	p = 0.12	$P<2*10^{-16}$

1: number of matching bases out of four (see Figure 6)

DISCUSSION

There are major differences between this study and our first study in mammals. First, we didn't verify every loss or gain event manually as we did in the mammalian study simply due to the sheer quantity of events, making such manual validation impractical. Secondly, there are considerable differences in the quality of the genomic assemblies, as most *Drosophila* are sequenced at much lower depth than the human, mouse, rat or dog genomes, and we expect the quality of the whole-genome alignments to suffer as a consequence. However, we believe that the first study has proven the reliability of our method and for that reason we expected a low false discovery rate despite these weaknesses. There is also less publicly available information about gene expression and function, which affects the types of analyses we can perform. Another difference with the first study is that we do not filter events based on alignment quality. The reasons for this are the lower quality of the alignments and the fact that there have been many more genomic rearrangements between species of *Drosophila* than between mammalian species. We attempted to filter events based on the length of alignments blocks and found that the ratio of exact events did not increase as a consequence. This is

why we deemed it preferable to concentrate on the overall characteristics, allowing for an acceptable false discovery rate.

Loss and Gain Characteristics

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 alternative splicing. It is probable that the significantly lower proportion of lost introns used for alternative splicing affects the average sequence conservation. The fact that the genes which 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. Secondly, the recombination model requires genes to be highly expressed in the germline 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.

Real intron gains?

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 two 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. Secondly, 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 two 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 Figure 2) would require the same intron to be lost three times independently and for a false positive gain on the *melanogaster subgroup* branch, four 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, the overall probability that all of our inferred gains could be false positives caused by parallel losses is very small.

Loss rate vs. gain rate

Our study sheds some much needed light on the ongoing introns-early vs. 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 pre-eukaryotic ancestors, while some introns are still gained today, at least in some species. Although we detected many more losses than gains because of our ascertainment method, 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 origin of introns

The question as to how introns originally appeared is a complicated one. In the general introduction, we presented three proposed mechanisms of intron gain, transposition, tandem duplication and reverse splicing. All three are examples of the broader, insertional theory of intron gains. The alternative is the formative theory, whereby introns were created as a byproduct 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. 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 byproduct 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 eukaryotes 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 explains the phase distribution (Vibranovski, Sakabe and de Souza 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 relationship 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. Since 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 performed a Blast search, aligning *D. melanogaster* introns that have or have not been gained onto every other intron. The transposon-like model implies that new introns are copies of existing introns, therefore 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 which would be longer than the four bases around the splice sites. We also failed to find such long repeats using Blast (data not shown) and Figure 6 shows that the similarity stops abruptly after the four bases centered on the splice-sites, displaying a concordance ratio of around 0.25, which is the null expectation. 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 dependant 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, Ning ang Brenner 1999) and the creation of new splice-sites within exons via single nucleotide mutations (Wang, Yu and Long 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.

METHODS

Reference-based intron mapping

We used the approach described in the Methods section of Chapter 1, with slight modifications, based on the latest MultiZ whole-genome alignments (Blanchette et al. 2004) and *D. melanogaster* RefSeq gene annotations downloaded from the UCSC Genome Browser database (Karolchik et al. 2003) to map the position of each *D. melanogaster* splice-site within a coding region directly onto the genomes of 10 other *Drosophila* species, *D. sechellia*, *D. erecta*, *D. yakuba*, *D. annanasae*, *D. pseudoobscura*,

D. persimilis, *D. willistoni*, *D. virilis*, *D. mojavensis*, *D. grimshawi*. We discarded *D. 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. This criteria is stricter than the 25 bp cut-off used in mammals because less in known about the fruit fly's capability of splicing tiny introns. 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. Our predictions agree well with GeneMapper, another reference-based gene annotation system (Chatterji and Pachter 2006). 84% of GeneMapper splice-sites mapped to within 5 bp of one of our inferred intron edges. 1944 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. 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, Fraga and Hinrichsen 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.

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 which 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 non-annotated 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 a low false discovery rate of gene structure changes.

Realignment of sequence around missing introns

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, Higgins and Gibson 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.

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 displayed in Figure 5, 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 which occurred on one of the two oldest branches of the tree since 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 three *D. melanogaster* transposase sequences and one reverse transcriptase sequence (NCBI 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 two inferred intron losses in *D. pseudoobscura* and at least three 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 calculate 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 squared 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 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 pair-wise 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 pair-wise conservation of genes with losses to the average pair-wise conservation of other genes. Then, to assign a "relative conservation" to a given gene, we used the ratio of the gene's average pair-wise 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 to that of lost introns, which are missing in some species, we computed the average intronic pair-wise conservation for all combinations of species required. Then, we performed a paired student's t-test by matching each lost intron's average pair-wise 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.

General Conclusion

In this thesis I have presented a novel reference-based intron mapping technique to study intron dynamics on a genome-wide scale. I have demonstrated the efficiency of the method as well as the importance of conducting intron dynamics studies on complete genomes, which currently, due to lack of sufficient gene annotation in most sequenced species, can only be achieved through reference-based mapping. We have shown that intron loss has occurred in mammals and that characteristics of lost introns support the cDNA recombination model, while intron gain appears to have completely stopped in mammals over 100 Myrs ago. In Chapter 2, I have shown that both intron loss and gain has occurred in *Drosophila*, further supporting the cDNA model of intron loss and putting an end to the strict interpretation of the introns-early hypothesis. Although the exact mechanism responsible for the origin of introns remains a mystery, our results have demonstrated that recently gained introns appeared through an insertional mechanism rather than a formative mechanism and we have revived the proto-splice-site theory as a viable explanation for the phase bias of intron gain. To be able to distinguish between different insertional models of gain, such as transposition, tandem duplication or reverse splicing, more gain events would have to be discovered, which is highly feasible with the steadily increasing number of fully sequenced eukaryotic genomes becoming available. As for intron loss, we can conclude that it occurs through a similar mechanism across very distant eukaryotes, as results in both studies seem to point to the same, widely accepted, cDNA recombination mechanism, and although it is a relatively rare phenomenon, it is probably the cause of most differences in intron positions across distant eukaryotes. We have shown how selection seems to play a role in the fixation of intron loss events, which is consistent with the belief that some introns exert important biological functions that require sequence conservation, such as transcriptional regulation or alternative splicing regulation. More specifically, we have demonstrated that introns essential to alternative splicing events in *D. melanogaster* are much less likely to be lost in other *Drosophila* species, suggesting that natural selection restricts such mutations from spreading through fruit fly populations. Although we may not know exactly how

introns first came into being, we know that they will be with us for a very long time and that some introns are probably very useful.

Bibliography

Adams, M. D.S. E. CelnikerR. A. Holt et al. 2000. The genome sequence of Drosophila melanogaster. Science 287:2185-2195.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Arango, N. A., T. T. Huang, A. Fujino, R. Pieretti-Vanmarcke, and P. K. Donahoe. 2006. Expression analysis and evolutionary conservation of the mouse germ cell-specific D6Mm5e gene. Dev Dyn 235:2613-2619.

Babenko, V. N., I. B. Rogozin, S. L. Mekhedov, and E. V. Koonin. 2004. Prevalence of intron gain over intron loss in the evolution of paralogous gene families. Nucleic Acids Res 32:3724-3733.

Bachtrog, D., and B. Charlesworth. 2001. Towards a complete sequence of the human Y chromosome. Genome Biol **2**:REVIEWS1016.

Blanchette, M., W. J. Kent, C. Riemer et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14:708-715.

Castillo-Davis, C. I., S. L. Mekhedov, D. L. Hartl, E. V. Koonin, and F. A. Kondrashov. 2002. Selection for short introns in highly expressed genes. Nat Genet 31:415-418.

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

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

Cho, S., S. W. Jin, A. Cohen, and R. E. Ellis. 2004. A phylogeny of caenorhabditis reveals frequent loss of introns during nematode evolution. Genome Res 14:1207-1220.

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., and K. H. Wolfe. 2004. Origins of recently gained introns in Caenorhabditis. Proc Natl Acad Sci U S A 101:11362-11367.

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

Csuros, M. 2005. Likely scenarios of intron evolution. Comparative Genomics 3678:47-60.

de Souza, S. J. 2003. The emergence of a synthetic theory of intron evolution. Genetica 118:117-121.

de Souza, S. J., M. Long, R. J. Klein, S. Roy, S. Lin, and W. Gilbert. 1998. Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. Proc Natl Acad Sci U S A 95:5094-5099.

Derr, L. K., J. N. Strathern, and D. J. Garfinkel. 1991. RNA-mediated recombination in S. cerevisiae. Cell 67:355-364.

Drouin, G. 2006. Processed pseudogenes are more abundant in human and mouse X chromosomes than in autosomes. Mol Biol Evol **23**:1652-1655.

Duret, L. 2001. Why do genes have introns? Recombination might add a new piece to the puzzle. Trends Genet 17:172-175.

Fedorov, A., A. F. Merican, and W. Gilbert. 2002. Large-scale comparison of intron positions among animal, plant, and fungal genes. Proc Natl Acad Sci U S A 99:16128-16133.
Gaffney, D. J., and P. D. Keightley. 2006. Genomic selective constraints in murid noncoding DNA. PLoS Genet 2:e204.
Gilbert W. 1987. The exon theory of genes. Cold Spring Harb Symp Quant Biol 52:901–905.
Hinrichs, A. S., D. Karolchik, R. Baertsch et al. 2006. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res 34:D590-598.
Hosack, D. A., G. Dennis, Jr., B. T. Sherman, H. C. Lane, and R. A. Lempicki. 2003. Identifying biological themes within lists of genes with EASE. Genome Biol 4:R70.
Karolchik, D., R. Baertsch, M. Diekhans et al. 2003. The UCSC Genome Browser Database. Nucleic Acids Res 31:51-54.
Kent, W. J., R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A 100:11484-11489.
Kidwell, M. G. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica 115:49-63.
Kim, H., R. Klein, J. Majewski, and J. Ott. 2004. Estimating rates of alternative splicing in mammals and invertebrates. Nat Genet 36:915-916; author reply 916-917.
Koonin, E. V. 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.
Krzywinski, J., and N. J. Besansky. 2002. Frequent intron loss in the white gene: a cautionary tale for phylogeneticists. Mol Biol Evol 19:362-366.
Lim, L. P., and C. B. Burge. 2001. A computational analysis of sequence features involved in recognition of short introns. Proc Natl Acad Sci U S A 98:11193-11198.
Logsdon, J. M., Jr., A. Stoltzfus, and W. F. Doolittle. 1998. Molecular evolution: recent cases of spliceosomal intron gain? Curr Biol 8:R560-563.
Long, M., S. J. de Souza, C. Rosenberg, and W. Gilbert. 1998. Relationship between "proto-splice sites" and intron phases: evidence from dicodon analysis. Proc Natl Acad Sci U S A 95:219-223.
Long, M., and C. Rosenberg. 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. 2003. Dependence of mutational asymmetry on gene-expression levels in the human genome. Am J Hum Genet 73:688-692.
Majewski, J., and F. M. Cohan. 1999. DNA sequence similarity requirements for interspecific recombination in Bacillus. Genetics 153:1525-1533.
Majewski, J., and J. Ott. 2002. Distribution and characterization of regulatory elements in the human genome. Genome Res 12:1827-1836.
Mourier, T., and D. C. Jeffares. 2003. Eukaryotic intron loss. Science 300:1393.
Myers, E. W., G. G. Sutton, A. L. Delcher et al. 2000. A whole-genome assembly of Drosophila. Science 287:2196-2204.
Nakao, A., M. Yoshihama, and N. Kenmochi. 2004. RPG: the Ribosomal Protein Gene

database. Nucleic Acids Res 32:D168-170.
Nei, M., P. Xu, and G. Glazko. 2001. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. Proc Natl Acad Sci U S A 98:2497-2502.
Nguyen, H. D., M. Yoshihama, and N. Kenmochi. 2005. New maximum likelihood estimators for eukaryotic intron evolution. PLoS Comput Biol 1:e79.
Nielsen, C. B., B. Friedman, B. Birren, C. B. Burge, and J. E. Galagan. 2004. Patterns of intron gain and loss in fungi. PLoS Biol 2:e422.
Nurtdinov, R. N., Artamonova, II, A. A. Mironov, and M. S. Gelfand. 2003. Low conservation of alternative splicing patterns in the human and mouse genomes. Hum Mol Genet 12:1313-1320.
Palmer, J. D., and J. M. Logsdon, Jr. 1991. The recent origins of introns. Curr Opin Genet Dev **1**:470-477.
Pan, Q., M. A. Bakowski, Q. Morris, W. Zhang, B. J. Frey, T. R. Hughes, and B. J. Blencowe. 2005. Alternative splicing of conserved exons is frequently species-specific in human and mouse. Trends Genet 21:73-77.
Pollard, D. A., V. N. Iyer, A. M. Moses, and M. B. Eisen. 2006. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. PLoS Genet 2:e173.
Putnam, N. H., M. Srivastava, U. Hellsten et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86-94.
Richards, S., Y. Liu, B. R. Bettencourt et al. 2005. Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. Genome Res 15:1-18.
Rogozin, I. B., A. V. Sverdlov, V. N. Babenko, and E. V. Koonin. 2005. Analysis of evolution of exon-intron structure of eukaryotic genes. Brief Bioinform 6:118-134.
Rogozin, I. B., Y. I. Wolf, A. V. Sorokin, B. G. Mirkin, and E. V. Koonin. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr Biol 13:1512-1517.
Roy, S. W., A. Fedorov, and W. Gilbert. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. Proc Natl Acad Sci U S A 100:7158-7162.
Roy, S. W., and W. Gilbert. 2005. The pattern of intron loss. Proc Natl Acad Sci U S A 102:713-718.
Roy, S. W., and W. Gilbert. 2005. Complex early genes. Proc Natl Acad Sci U S A 102:1986-1991.
Roy, S. W., and W. Gilbert. 2005. The pattern of intron loss. Proc Natl Acad Sci U S A 102:713-718.
Roy, S. W., and W. Gilbert. 2005. Rates of intron loss and gain: implications for early eukaryotic evolution. Proc Natl Acad Sci U S A 102:5773-5778.
Roy, S. W., and W. Gilbert. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet 7:211-221.
Roy, S. W., and D. L. Hartl. 2006. Very little intron loss/gain in Plasmodium: intron loss/gain mutation rates and intron number. Genome Res 16:750-756.

Roy, S. W., and D. Penny. 2006. Smoke without fire: most reported cases of intron gain in nematodes instead reflect intron losses. Mol Biol Evol 23:2259-2262.
Russell, C. B., D. Fraga, and R. D. Hinrichsen. 1994. Extremely short 20-33 nucleotide introns are the standard length in Paramecium tetraurelia. Nucleic Acids Res 22:1221-1225.
Siepel, A., and D. Haussler. 2004. Combining phylogenetic and hidden Markov models in biosequence analysis. J Comput Biol 11:413-428.
Slaven, B. E., A. Porollo, T. Sesterhenn, A. G. Smulian, M. T. Cushion, and J. Meller. 2006. Large-scale characterization of introns in the Pneumocystis carinii genome. J Eukaryot Microbiol 53 Suppl 1:S151-153.
Springer, M. S., W. J. Murphy, E. Eizirik, and S. J. O'Brien. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc Natl Acad Sci U S A 100:1056-1061.
Sterner, D. A., T. Carlo, and S. M. Berget. 1996. Architectural limits on split genes. Proc Natl Acad Sci U S A 93:15081-15085.
Stoltzfus, A. 1994. Origin of introns--early or late. Nature 369:526-527; author reply 527-528.
Su, A. I., M. P. Cooke, K. A. Ching et al. 2002. Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A 99:4465-4470.
Sverdlov, A. V., V. N. Babenko, I. B. Rogozin, and E. V. Koonin. 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, J. D., D. G. Higgins, and T. J. Gibson. 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., Y. Ning, and S. Brenner. 1999. Late changes in spliceosomal introns define clades in vertebrate evolution. Proc Natl Acad Sci U S A 96:10267-10271.
Venter, J. C.M. D. AdamsE. W. Myers et al. 2001. The sequence of the human genome. Science 291:1304-1351.
Vibranovski, M. D., N. J. Sakabe, and S. J. de Souza. 2006. A possible role of exon-shuffling in the evolution of signal peptides of human proteins. FEBS Lett 580:1621-1624.
Wang, W., H. Yu, and M. Long. 2004. Duplication-degeneration as a mechanism of gene fission and the origin of new genes in Drosophila species. Nat Genet 36:523-527.
Waterston, R. H.K. Lindblad-TohE. Birney et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-562.
Weiner, A. M., P. L. Deininger, and A. Efstratiadis. 1986. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu Rev Biochem 55:631-661.
Yoshihama, M., A. Nakao, H. D. Nguyen, and N. Kenmochi. 2006. Analysis of ribosomal protein gene structures: implications for intron evolution. PLoS Genet 2:e25.
Yoshihama, M., H. D. Nguyen, and N. Kenmochi. 2007. Intron dynamics in ribosomal protein genes. PLoS ONE 2:e141.
Zhang, Z., P. M. Harrison, Y. Liu, and M. Gerstein. 2003. Millions of years of evolution

preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res 13:2541-2558.
Zhuo, D., R. Madden, S. A. Elela, and B. Chabot. 2007. Modern origin of numerous alternatively spliced human introns from tandem arrays. Proc Natl Acad Sci U S A 104:882-886.