Genetica **118:** 171–182, 2003.

Origin of new genes: evidence from experimental and computational analyses

Manyuan Long, Michael Deutsch, Wen Wang, Esther Betrán, Frédéric G. Brunet & Jianming Zhang

Department of Ecology and Evolution, The University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA (Phone: +1-773-702-0557; Fax: +1-773-702-9740; E-mail: mlong@midway.uchicago.edu)

Key words: domain shuffling, exon–intron evolution, exon shuffling, evolutionary genomics, intron phases, new genes

Abstract

Exon shuffling is an essential molecular mechanism for the formation of new genes. Many cases of exon shuffling have been reported in vertebrate genes. These discoveries revealed the importance of exon shuffling in the origin of new genes. However, only a few cases of exon shuffling were reported from plants and invertebrates, which gave rise to the assertion that the intron-mediated recombination mechanism originated very recently. We focused on the origin of new genes by exon shuffling and retroposition. We will first summarize our experimental work, which revealed four new genes in *Drosophila*, plants, and humans. These genes are 10^6 to 10^8 million years old. The recency of these genes allows us to directly examine the origin and evolution of genes in detail. These observations show firstly the importance of exon shuffling and retroposition in the rapid creation of new gene structures. They also show that the resultant chimerical structures appearing as mosaic proteins or as retroposed coding structures with novel regulatory systems, often confer novel functions. Furthermore, these newly created genes appear to have been governed by positive Darwinian selection throughout their history, with rapid changes of amino acid sequence and gene structure in very short periods of evolution. We further analyzed the distribution of intron phases in three non-vertebrate species, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Arabidosis thaliana*, as inferred from their genome sequences. As in the case of vertebrate genes, we found that intron phases in these species are unevenly distributed with an excess of phase zero introns and a significant excess of symmetric exons. Both findings are consistent with the requirements for the molecular process of exon shuffling. Thus, these non-vertebrate genomes may have also been strongly impacted by exon shuffling in general.

Introduction

Origin of new genes with novel functions is a fundamental biological process in nature. Information about this process is essential for insight into genetic diversity in organisms. Successes in sequencing whole genomes of many organisms including humans in recent years show that organisms differ in the number and types of genes they possess (e.g., for bacterial genomes, see Fraser et al., 1995 and Himmelreich et al., 1996; for eukaryotic genomes, see Rubin et al., 2000; Lander et al., 2001; Venter et al., 2001; Betran & Long, 2002), revealing an evolutionary birth and death process for coding portions in genomes. New genes arise at various points in evolutionary time; they confer novel functions on organisms enabling them to confront the challenges of changing environments. Novel genes can be created by the juxtaposition of various preexisting exons in new combinations (Gilbert, 1978). Novel genes can also be created by mutational modification of duplicate genes (Ohno, 1970). An early sequence comparison revealed that there are around 250 modular protein families in eukaryotes (Patthy, 1995) that may have originated via exon shuffling; our further genomic analysis indicated that more than 20% of eukaryotic exons were created

by this mechanism (Long, Rosenberg & Gilbert, 1995; Long, de Souza & Gilbert, 1995).

Despite the progress in understanding the molecular and evolutionary mechanisms, several important questions remained. For example, what are major mechanisms for generating new gene structures? What evolutionary forces were involved in the fixation of the new genes across whole species? Are there any general patterns in new gene evolution? A general answer to these questions may not be available in the near future, because primary explorations have indicated that there might be multiple factors involved in each of the processes in new gene evolution. A serious challenge is that many of the early features of evolution of new genes have been lost in the long history of the genes comprising a most common data sources in molecular research, because most genes available for analysis are very old. For example, one consequence is that many of the introns have been lost, which might have been involved in exon shuffling in the primitive world of genes. Obviously, a solution to this problem is to directly observe the early stage in evolution of a young gene (Long, Wang & Zhang, 1999; Long, 2001). Unfortunately, identified young genes are rare in available databases of genes and genomes (e.g., Gu et al., 2002). Several possibilities may have contributed to this impasse. First, most efforts in molecular analysis are driven by medically related researches, which have been trying to identify those genes common to model organisms and humans. Molecular biology so far does not require analysis of young genes. Second, the power of gene annotation for genome sequences is limited (Attwood, 2000), due to the insufficient criteria for identifying functional genes and other technical challenges. Finally, the identification of a young gene is technically demanding, requiring understanding of molecular evolution, skill at handling sophisticated molecular technology, and bioinformatic analysis of genome sequences.

However, screening for the hallmarks of some of the most dynamic molecular processes for generating new gene structures, and comparing young and parental genes in one species with the orthologous parental genes in related species, we have built observation systems to study the new gene evolution. The molecular process for our screen includes exon shuffling and retroposition, which together can rapidly create new gene loci with chimerical gene structures – mosaic proteins or hybrid genes with new regulatory structures for the coding regions. Such structures often develop new protein functions or expression patterns distinct from their parental genes and thus provide a good system to investigate evolution of new gene functions. We will analyze the new genes we identified in the past decade. These new genes aged from 1 to 100 million years were found in *Drosophila*, higher plans, and primates including humans, thus represent a relatively broad range of higher eukaryotes. Our data from unambiguous young genes, together with computational analysis of genome sequences, have offered some insights. Finally, we will present a statistical analysis of intron phases in the genomes of *D. melanogaster*, *C. elegans*, and *A. thaliana*. In this analysis, we will show strong signals of exon shuffling in these plant and invertebrate genomes.

Case analysis of new genes

Exon shuffling, proposed by Gilbert (1978), has been demonstrated important in the origin of new genes (Long, 2001; Long, Rosenberg & Gilbert, 1995; Long, de Souza & Gilbert, 1995; Patthy, 1995). As a molecular evolutionary mechanism by which new genes are created by recombination of exons, exon shuffling may occur in a number of ways. The exon theory of genes (Gilbert, 1987) proposes that exon shuffling was common in the RNA world that preceded the advent of DNA as the carrier of genetic information – that introns, acting as ribozymes, catalyzed the shuffling of exons in the RNA progenote. In a DNA genome, exon shuffling may occur by nonhomologous recombination in introns. This includes illegitimate recombination at the genomic DNA level and retroposition at the RNA level (Brosius, 1999, 2003). Introns greatly extend the area of a chromosome in which recombination within a gene can occur, significantly elevating the probability of nonhomologous recombination between exons of different genes (Gilbert, 1987).

Jingwei

The chimeric *jingwei* gene in *Drosophila teissieri* and *D. yakuba*, for example, provides evidence for recent exon shuffling by retroposition (Long & Langley, 1993). *Jingwei* was created when a processed *alcohol dehydrogenase* messenger RNA was reverse-transcribed and inserted into the third intron of the *yellow emperor* (*ymp*) gene (Long, Wang & Zhang, 1999; Wang et al., 2000), as shown in Figure 1. The *Adh*-derived portion of *jingwei* was discovered in restriction analysis of the *Adh* gene in the *melanogaster* subgroup by Langley, Montgomery and Quattlebaum (1982) as a additional signal in two sibling species,

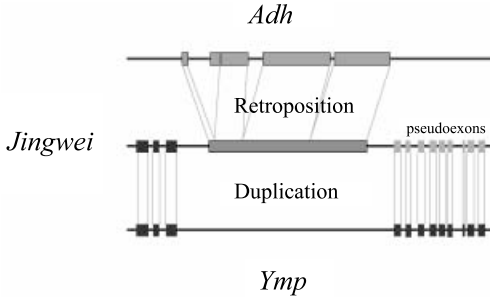

Figure 1. Origin of *jingwei* as a chimera from its two parents: *Adh* through retroposition and yellow emperor via duplication. Sequence analysis indicated that *jingwei* originated within 2.5 million years. The downstream exons become nontranscribed pseudoexons because the transcription terminates at the *Adh*-derived portion of *jingwei* (Long, unpublished data).

D. teissieri and *D. yakuba*, while all other species in the subgroup were found to contain only a single *Adh* copy. However, the structure and functionality of the new *Adh*-related locus were not investigated until Jeffs and Ashburner (1991) cloned and sequenced this new genetic element. Two remarkable observations were made: (i) the start codon is deleted in both species; (ii) all three introns are missing in both species. The nucleotide sequence of the new element is highly similar to that of the *Adh* gene (92.9% for *D. yakuba* and 95.6% for *D. teissieri*). These observations, with an additional assumption that the chance that a retrosequence jumps into a genomic region with an appropriate preexisting regulatory system is low, led to the conclusion that the second locus is an *Adh* processed pseudogene, that is, a functionless gene created by retroposition. However, the strong functional constraint at the protein sequence level was detected in a population genetic investigation (Long & Langley, 1993). It was found that among 21 polymorphic sites in a *D. yakuba* population (21 alleles), 19 changes were synonymous, while 8 out of 10 polymorphic sites in a *D. teissieri* population (10 alleles) had synonymous changes. This observation, in conjunction with the unique expression patterns in the two species, suggest that the second *Adh* locus has a newly evolved function. In contrast to the strong purifying selection in population genetic analysis, divergence between species was unusually rapid, bearing the signs of adaptive protein evolution. From the time point of the retroposition event to the speciation event in which *D. yakuba* and *D. teissieri* diverged, there were nine fixed nucleotide substitutions; amazingly, all these substitutions occurred in replacement sites and no synonymous change was fixed. Many of these changes in the three-dimensional *Adh* protein structure are scattered in one small area surrounding the substrate binding sites (Brunet, Zhang & Long, unpublished results). These observations strongly suggest that in its early stage *jingwei* was under strong selection for adaptive protein sequence changes, in contrast to the hypothesis that it was a functionless pseudogene in the very beginning (e.g. Brookfield & Sharp, 1994). After the speciation event, the JGW proteins continued to evolve rapidly under strong positive Darwinian selection (Long & Langley, 1993). Then, how did *jingwei* acquire start codons and an appropriate regulatory system? Molecular analyses showed that the retroposed *Adh* sequences recruited three nearby exons containing an in-frame start codon (Figure 1), yielding a chimerical protein. The regulatory sequence associated with the acquired exon was also recruited by *jingwei* as well, shown by the identical testis-specific expression pattern both in *D. teissieri jingwei* and its other parental gene, *yellow-emperor* (Wang et al., 2000; Long & Langley, 1993). Concerning the process of retroposition that created *jingwei*, Boeke and Pickeral (1999) speculated that the retroposition process was driven by the movement of some LINE-1-like retrotranspon in *Drosophila*.

Cytochrome c1

This finding is related to a classic problem in gene evolution. Many nuclear genes that encode organellar proteins may have been transferred from organellar genomes in a transposition process (Palmer, 1985; Nugent & Palmer, 1991). Because these genes are transcribed in the nucleus and translated in the cytoplasm, the derived proteins have to be moved to and specifically recognize the organelles (e.g., mitochondria). Such proteins contain an organellar targeting domain to carry out this function. However, because their ancestral genes that are located in the organellar genomes were unlikely to need such a function, the function must have originated after transfer of the genes into nuclear genomes. The mechanism conferring a target domain on these genes was assumed to be exon shuffling from other parts of the nuclear genomes (Nugent & Palmer, 1991). The cytochrome c1 in potato provides a clear piece of evidence for this hypothesis and is one of a few examples of exon shuffling in plants.

The cytochrome c1 in potato comprises nine exons of which the first three are responsible for mitochondrial targeting (Wegener & Schmitz, 1993). In an

exhaustive comparison of an exon database derived from GenBank (Long et al., 1996), we found that this portion of the gene had a high identity with the first three exons in the *gapdh* gene encoding glyceraldehyde-3-phosphate dehydrogenase (the identity at the level of amino acid sequences between the pea GapC, a member of GAPDH family, and the potato cytochrome c1 is 44% without gap in the sequence alignment, $P = 1.1 \times 10^{-6}$), suggesting that the first three exons in the cytochrome c1 gene had been created via exon shuffling from the *gapdh* gene (Figure 2). The shuffled peptides of both *gapdh* and the cytochrome c1 gene contain amphiphilic α-helices required for targeting activities in organelle-specific proteins (Roise et al., 1986; Schatz & Dobberstein, 1996). The shuffled region in cytochrome c1 contains introns, suggesting that the exon shuffling event may have taken place at the DNA level. The examination of phylogenetic distribution indicated that the shuffling occurred within the past 100 million years – a divergence time between *Solanaceae* and *Brassicaceae* and that the newly acquired peptide evolved much more rapidly than its counterpart in GAPDH.

Sphinx

Exon shuffling, as shown in *jingwei* in *Drosophila*, cytochrome c1 in plants and other cases, has been demonstrated to be a mechanism that can create new proteins rapidly. However, recent literature shows that eukaryotic genomes contain many noncoding but functionally important RNA (ncRNA) genes (Eddy, 2001). How does this type of genes originate? Does the mechanism of exon shuffling, which is known to play an important role in evolution of protein coding genes, also apply to the origin and evolution of ncRNA genes? Our recent finding of the new ncRNA gene, *sphinx*, in *D. melanogaster* revealed that exon shuffling mediated by retroposition, as for the young protein coding gene *jingwei*, is also an efficient mechanism for the origin of RNA genes (Wang et al., 2002a).

Sphinx originated within the last 2–3 million years in the *D. melanogaster* chromosome 4 as a chimerical ncRNA gene. Its first exon was recruited by its inserted exon 2 after retroposition from the gene encoding ATP synthase on chromosome 2, as summarized by Figure 3. The evidence that *sphinx* is probably a nonprotein-coding RNA (ncRNA) gene was provided by several observations. First, it appears that *sphinx* does not contain any ORFs with significant coding potential. The coding ability inherited from the parental gene has been rapidly eliminated by a series of sequence changes, including a change in the start codon, introduction of a stop codon in the original reading frame, and several deletions causing frameshift mutations. Second, it was found that several observed within-species polymorphisms are at nonsynonymous sites, suggesting no protein-coding constraint. However, it is not a pseudogene. It is expressed and displays sex- and development-specific alternative splicing, suggesting some biological functions associated with male and female individuals in various developmental stages; the expression is under tight control. Similar to *jingwei*, the site of insertion also allowed recruitment of a new regulatory region to the retroposed sequence, thus avoiding a fate of nonfunctionalization that many retrosequences face. Its sequence evolved with a significantly accelerated rate of substitutions: There are 18 fixed substitutions in 352 nucleotides of the retroposed region, compared to 2 changes in its parental gene. These observations suggest that the *sphinx* gene evolved toward new functions conferred by the novel chimerical RNA. In addition, the high level of nucleotide variation in *sphinx* negated a classic genetic conclusion that the fourth chromosome in *D. melanogaster* was nonrecombining and unvarying (Wang et al., 2002b).

PGAM3

PGAM3 is a new primate gene originated by retrotransposition (Figure 4) (Betran et al., 2002). Its parental gene encodes a phosphoglycerate mutase brain isoform (*PGAM1*) that catalyzes the transformation of 2-phosphoglycerate to 3-phosphoglycerate (Grisolia & Joyce 1959; Grisolia & Carreras 1975). Dierick, Mercer, and Glover (1997) investigated the structures and molecular evolution of the PGAM family. *PGAM1* conains 3 introns while *PGAM3* is intronless. The two genes are in different locations; *PGAM1* is on chromosome 10, at 10q25.3 and *PGAM3* is on the X chromosome region Xq13.3. *PGAM3* is located within the first intron of the Menkes disease gene (*MNK*). Flanked by 10-bp direct repeats, *PGAM3* has a poly-A tail of 16 bp after the polyadenylation signal at the end of the 3′UTR. *PGAM3* had so far been only found in humans and it was described as a pseudogene: *Phosphoglycerate mutase 1* processed pseudogene (ψ*PGAM1*). However, our data revealed that it is a functional processed gene present in humans and other species (Betrán et al., 2002).

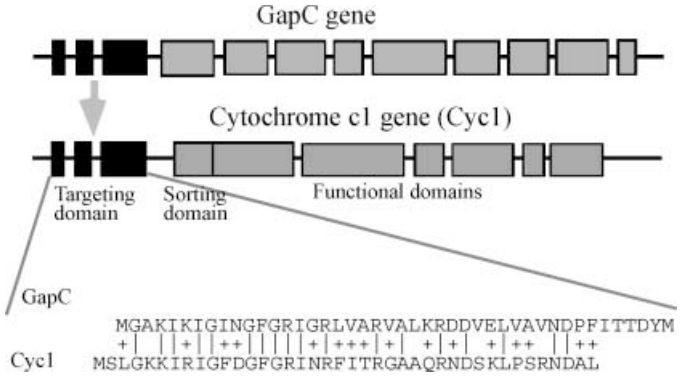

Figure 2. Cytochrome c1 in potato acquired its mitochondrial targeting domain as a consequence of exon shuffling from the GapC gene. In the aligned sequences, '|' indicates identical amino acid residues; '+' similar amino acid residues, measured according to scoring matrix BLOSUM62.

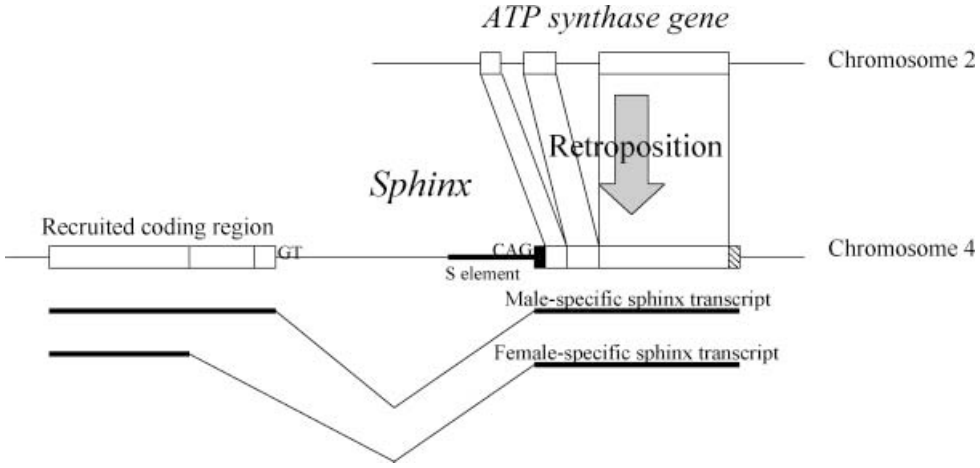

Figure 3. The ncRNA gene *sphinx* originated in a similar molecular process by which *jingwei* created. The ATP synthase gene in chromsome 2 retroposed and the retrosequences were inserted into chromosome 4 and recruited a nearby exon/intron to form a chimerical structure. The transposable S element provided 3′ splicing site for generating the novel intron. Alternative splicing created male-specific and female-specific isoform.

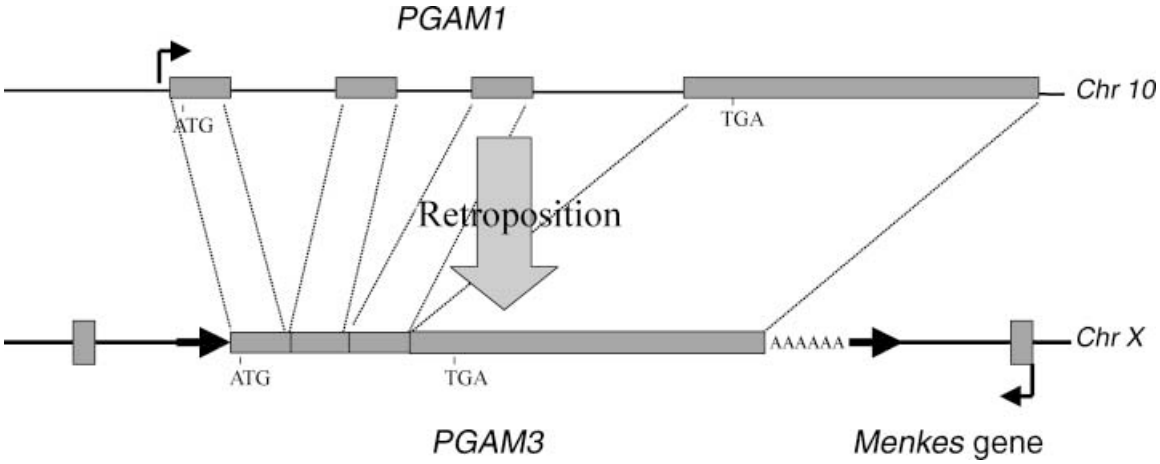

Figure 4. New primate gene *PGAM3* originated as an insertion of retrosequence of *PGAM1* gene into an intron in an unrelated gene *Menkes*. Reflecting the young age of *PGAM3*, the poly A tail of 16 bp is obvious in its 3′ end with two 10 bp direct repeats, as expected as additional hallmarks of retroposition, besides the three lost introns. The complete structure of the *Menkes* gene is not shown.

PGAM3 is present in humans, chimpanzees and macaques (Betrán et al., 2002). According to its phylogenetic distribution, it is more than 25 My old (Goodman et al., 1998). *PGAM3* consists of an intact open reading frame with many deletions scattered through the 3′UTR region. Perhaps it is more remarkable that comparing human *PGAM1* with human *PGAM3* and with chimpanzee *PGAM3* showed K_A bigger than K_S. K_A/K_S values were 2.625 and 5.000, respectively, with probabilities 0.052 and 0.008. The comparison of *PGAM1* and *PGAM3* with an outgroup sequence, *PGAM2*, provides the necessary information to determine changes in each lineage using a parsimony criterion. Most of the nonsynonymous changes between *PGAM1* and *PGAM3* happened in the *PGAM3* lineage. The number of amino acid changes in the *PGAM3* lineage was significantly higher than in the lineage leading to *PGAM1*: 9 versus 1 ($P = 0.0110$) in human and chimpanzee. Changes occurred preferentially in the first and second codon position. These two codon positions together explain the significant difference in amino acid rate and evolve at a different rate than the third position ($P = 0.0172$) in chimpanzee. These results reveal the action of positive selection. The excess of rare alleles in the population genetic analysis further reveals the role of selection. Because the amino acids at the enzyme active site are unchanged in human, chimpanzee and macaque *PGAM3*, we inferred that *PGAM3* might have PGAM function (Betrán et al., 2002).

These case analyses of new genes revealed two common features. First, *the introns can serve as the target sites for exon recombination and retrosequence insertion*. In the cases of *jingwei*, *cytochrome c1*, and *PGAM3*, clearly an intron in the recipient genes was involved in the exon recombination. In the case of *sphinx*, the recipient gene contributed 5′ splicing sites, CT|GTAAGT for the female transcript and AG|GTTTGT for the male transcript, which both are similar to the consensus for eukaryotic genes, AG|GTAAGT (Long & Deutsch, 1999). These splice sites may have previously existed for splicing out the intron into which the retroposed ATP synthase sequence inserted. Therefore, these cases, corroborated by results of the genomic analysis to be presented in the following section, support the concept, proposed by Gilbert (1978, 1987), that introns can facilitate exon recombination, leading to origin of new genes. Second, *all four cases showed accelerated substitution rates*, suggesting that the rapid changes driven by positive Darwinian selection, as revealed in the early study of the young gene *jingwei*, may be a general phenomenon. Indeed, rapid evolution was also observed in other new genes, for example, *Sdic* in *D. melanogaster* (Nurminsky et al., 1997 and see Ranz et al., this volume), *Adh-Fennigan* in the repleta group of *Drosophila* (Begun, 1997), and primate new genes (see Nahon & Brosius, 1999). The rapid evolution points to rapid acquisition of a novel function. In fact, the detailed study of the biochemical functions of the *jingwei* gene (Zhang, Dean & Long, unpublished data) unveiled a correlation between rapidly occurring substitutions and changes in substrates and enzymatic activities. Moreover, the rapid evolution suggests the imperfection of newly created genes and the evolutionary potential for a more improved function.

Exon shuffling in plants and invertebrates as revealed in intron phase analysis

A direct approach to detecting exon shuffling is sequence comparison between different genes. For example, in the case of *jingwei*, the sequence comparison among *jingwei*, *Adh*, and *yellow emperor* revealed significant sequence similarities between the recruited portion of *jingwei* and *yellow emperor* and between the retroposed region and *Adh*. In this case, because of the extremely recent age of *jingwei*, the similarities between the new gene and parental genes are higher than 90% at the DNA sequence level. Thus, *jingwei* is clearly established as a hybrid gene from a recombination of exon groups from two different gene families. Likewise, the cytochrome c1 gene in potato recruited a mitochondrial-targeting domain from the *gapdh* gene by exon shuffling 100 million years ago. The recruited targeting domain in the gene still keeps significant identity with the homologous region of the parental gene, *gapdh*, at the protein sequence level.

The direct comparative approach has revealed many cases of exon shuffling. For example, Patthy (1991, 1995) reported that numerous protein families were created by this mechanism. In these investigations, in addition to the sequence similarity criterion, other criteria have also been taken into considerations, such as the distribution of intron phases (Patthy, 1987), increasing the rigor of the comparisons. Dorit, Schoenbach and Gilbert (1991) also tried to systematically seek exon shuffling by

sequence comparison. A conspicuous feature of this early work based on sequence comparison is that most of the cases of exon shuffling identified were the genes from verteberates. There were far fewer cases from plants and invertebrates, although modular proteins were observed in all major groups of metazoa (Patthy, 1995). In fact, only several cases in plants were reported: potato cytochrome c1 as discussed above (Long et al., 1996), two anther-specific genes in the sunflower with unclear functions and evolutionary histories (Domon & Steinmetz, 1994), and an unfixed new homeobox-related locus in tomato (Chen et al., 1997). Similar to the cytochrome c1 gene, the sunflower genes acquired new signal peptides. Concerning intron origination, these observations led to the inference that exon shuffling was a recent process, probably associated with a recent origin of introns in vertebrates.

Complementary to the direct sequence comparison approach, other approaches have been developed to detect signs of past exon shuffling at the genomic level. These are based on the statistical analysis of the distribution of exon shuffling signals. The search for excess of certain types of introns in protein linker regions of protein structures (De Souza et al., 1996, 1998) is one such approach; the analysis of intron phase distribution provides another efficient tool (Long, Rosenberg & Gilbert, 1995; Long, de Souza & Gilbert, 1995). These new approaches differ from direct sequence comparison in that one no longer asks whether or not an individual gene is created by exon shuffling or if single intron was created by recent insertion or is a relic of ancient gene structure. Instead, one is looking for statistical signals of exon shuffling among a varied host of signals from other processes, including the gain and loss of introns. A test using one or a few genes is far too simple to account for the diverse array of contemporary gene structures. While direct sequence comparison reveals concrete cases of exon shuffling, the statistical analysis of molecular signals of various processes including exon shuffling has proven to be more efficient in detecting both ancient and recent events of exon shuffling (Long, de Souza & Gilbert, 1995; Long, 2001).

Is exon shuffling limited only to vertebrates? Are invertebrates and plants devoid of such mechanisms to create new genes? The current cases of exon shuffling identified by sequence comparison may not be able to provide a certain answer for several untested possibilities. First, it might be that the rates of exon shuffling are variable so that there might be more exon shuffling in the recent history of vertebrates while there might be ample exon shuffling in the early stages of non-vertebrates. Second, the shuffled exons in vertebrates might be more conserved than in non-vertebrates, so we would be more likely to observe cases in vertebrates. Finally, we cannot rule out the possibility that the data in the public databases we examined in the past are biased toward genes of humans or vertebrates because of the rapid progress of medically-related research. Fortunately, today, there has been great progress in the genomics of invertebrates and plants, which has provided genome sequences from *Drosophila melanogaster* (fruit flies), *Caenorhabditis elegans* (worms), and *Arabidosis thaliana* (cress). The annotated genes in these genomes can be used to detect exon shuffling. In the following analysis, we will answer the particular question whether the plant and invertebrate genomes are deficient in exon shuffling by analyzing intron phase distribution, as we designed previously. We will also review previous work to test the alternative hypotheses interpreting intron phase distributions.

Intron phases are defined by the position of an intron relative to the reading frame of the coded message. An intron can be located between two intact codons, or after the first, or after second nucleotide within a codon. The corresponding intron phases are phase 0, 1, or 2. Intron phase, thus simply defined, is one of the most conserved molecular features in eukaryotic genes. For example, the introns in TPI genes encoding triosephosphate isomerase at homologous positions in plants and animals are identical in phase (see the data of Cerff, 1995). This conservation is because changes of intron phase require demanding molecular processes, for example, the coupled deletion and insertion of nucleotides in the two flanks of an intron (for more hypothesized mechanisms involved in the process, see discussion of Stultzfus et al., 1997). Therefore, it is expected that this feature of introns may retain the signal of early events of intron evolution in addition to unambiguous recent changes of intron–exon structures of genes.

The first step in a statistical analysis is to construct a null hypothesis and derive a corresponding expectation for its test. This is called the method of hypothesis test. For the intron phase distribution, a valid test must be based on a null hypothesis of random distribution of intron phases, a feature typical of the predictions made by the introns-late theory. Because there are three intron phases, the simplest prediction of the random

insertion hypothesis is 1/3 for the frequency of each phase:

$$E(i) = \tfrac{1}{3},$$

where phase $i = 0$, 1, and 2. It should be noted, as discussed previously (Long, Rosenberg & Gilbert, 1995), that this is just the simplest but not the only possible prediction based on a random insertion hypothesis. Fortunately a modified random insertion hypothesis can be tested by testing its corresponding predictions. However, the test of the simplest prediction offered an unexpected insight in this case.

Any deviation from equiprobable distribution of intron phases may also reflect a more complicated insertion process. Therefore, a more conservative test is simply to assume that the biased phase distribution of single introns is a result of some unknown constraints on intron insertion, and then to test if the *associations* of different intron phases are random when a gene contains two or more introns. This approach was supported by the observations that splicing processes of two introns are not affected by their phases (Long & Deutsch, 1999), which thus seems to be consistent with the hypothesis of independent intron insertions. For any two introns in a gene, there are nine possible recombinations of phases: (0,0), (1,1), (2,2), (0,1), (0,2), (1,2), (1,0), (2,0), and (2,1), where (i, j) indicate phase i at the 5′ intron and j at the 3′ intron. The number of exons between the two introns under consideration can be 1, 2, 3, ... The (0,0), (1,1), and (2,2) are also called symmetric exons; the rest, asymmetric exons. Three predictions can be made about frequencies of these associations expected if they are random.

The first one is based on the equal-probability assumption

$$\begin{aligned} E_1(i, j) &= \tfrac{1}{3} \times \tfrac{1}{3} \\ &= \tfrac{1}{9}, \end{aligned}$$

where $i, j = 0$, 1, and 2. However, the three phases themselves are not evenly distributed. Hence we can make the conservative assumption that the observed frequency of intron phase $f(i)$ may reflect random intron insertions into nonrandomly distributed sites, while association of the phases of two introns is random. Thus, the second prediction can be made as

$$E_2(i, j) = f(i) \times f(j).$$

On the other hand, we can directly take observed frequencies of associations of intron phases, that is, the observed frequency of internal exons ($f(k, l)$), which were flanked by two introns with phase association (k, l) ($k, l = 0, 1,$ and 2) to compute expected frequencies of the nine associations, conditioning on the 3′ end of the 5′ exon and 5′ end of next internal exon that is identical in phase

$$\begin{aligned} E_3(i, j) &= \frac{A(i, j)}{S(k, l)}, \\ A(i, j) &= \textstyle\sum_{l_1} \sum_{l_2} \cdots \sum_{l_{n-1}} f(i, l_1) f(l_1, l_2) \cdots \\ &\quad \cdots f(l_{n-2}, l_{n-1}) f(l_{n-1}, j), \\ S(k, l) &= \textstyle\sum_{l_0} \sum_{l_1} \sum_{l_2} \cdots \sum_{l_{n-1}} \sum_{l_n} f(l_0, l_1) \cdot \\ &\quad \cdot f(l_1, l_2) \cdots f(l_{n-2}, l_{n-1}) \cdot \\ &\quad \cdot f(l_{n-1}, l_n), \end{aligned}$$

where n is number of internal exons within the two introns under consideration. Because the comparison between $E_2(i, j)$ and $E_3(i, j)$ in general databases showed that the difference between the two expectations was small (Long, unpublished data), we will only use $E_2(i, j)$ in the computation of expected number of intron association.

Flat files containing the genome sequences of *C. elegans* (CESC, 1998), *D. melanogaster* (Adams et al., 2000), and *A. thaliana* (chromosomes 2 and 4) (AGI, 2000) were downloaded from the GenBank database of the National Center for Biotechnology Information (Burks et al., 1990) to our Alpha WDPS 500au workstation (Digital) by anonymous FTP from ftp.ncbi.nlm.nih.gov. All analyses were based on the information in these files, and were performed separately for each species on our Alpha workstation. A computer program was developed to extract all intron-containing genes (identified as gene entries in the feature table that are followed by 'CDS..join' and 'CDS..complement(join' entries) from the GenBank flat files and construct a database containing gene identifiers, intron locations and phases, and protein sequences. The three species contain many families of paralogous genes. To remove false positives that would be caused by comparing exons from these paralogous genes, the database was purged of redundancy using the method of Long, Rosenberg and Gilbert (1995) with a criterion of 20% amino acid sequence similarity. When two genes show greater than 20% similarity scaled to the length of the shorter gene as measured by fasta3 (Pearson, 1994), the gene with more exons is kept and the other discarded to create the largest possible database of exons (Table 1).

First, we find that intron phase distributions of the three species in this study, as shown in Table 2, are similar to those previously reported from the analysis

Table 1. Numbers of genes and exons in each species database before and after purging[a]

	Unpurged		Purged (20%)	
	Genes	Exons	Genes	Exons
C. elegans	16,800	100,691	5,699	45,841
D. melanogaster	14,095	52,154	5,074	25,040
A. thaliana	7,858	39,677	2,587	18,815

[a] *C. elegans* and *Drosophila* data are from complete genomes.

Table 2. Observed intron phase distributions in the purged databases

	Phase			N[a]
	0	1	2	
C. elegans (%)	46.9	26.6	26.5	40,142
D. melanogaster (%)	42.3	34.1	26.6	19,966
A. thaliana (%)	57.1	21.5	21.4	16,228

[a] Number of introns in database.

Table 3. Observed distribution of intron associations

3′ phase	5′ phase		
	0	1	2
C. elegans			
0	8,010 (5.5)	3,930 (−9.4)	4,142 (−3.5)
1	4,050 (−6.1)	2,872 (15.0)	2,357 (−3.3)
2	4,126 (−3.9)	2,247 (−8.3)	2,743 (11.5)
D. melanogaster			
0	2,894 (4.8)	1,862 (−8.9)	1,765 (1.9)
1	1,932 (−4.9)	1,638 (9.0)	1,234 (−3.2)
2	1,687 (−2.6)	1,154 (−10.4)	1,233 (11.8)
A. thaliana			
0	4,792 (5.8)	1482 (−14.9)	1678 (−0.6)
1	1,596 (−6.7)	721 (10.9)	614 (−3.8)
2	1,721 (−1.9)	596 (−6.9)	648 (2.5)

The numbers in parentheses are excess over expectation (excess $= (O - E)/E \times 100$) (see methods). A chi-square test shows that the deviations from expectation are highly significant ($P \ll 10^{-10}$) in both species.

of a general database, GenBank (Fedorov et al., 1992; Long, Rosenberg & Gilbert, 1995; Long, de Souza & Gilbert, 1995; Long & Deutsch, 1999). The frequency of the zero phase intron is highest, from 42.3% for *D. melanogaster* to 46.9% for *C. elegans* to 57.1% for *A. thaliana*. Phase 2 introns are least frequent. In the cress and worm, phase 2 introns are slightly lower than the phase 1 intron by 0.1%. Thus, similar to the general database consisting mostly of vertebrate, especially human, genes, phase zero introns predominate in all three nonvertebrates, tending to keep codons intact. This distribution biased toward phase zero introns, while rejecting the equiprobable distribution predicted by the simplest model of intron insertion, was also tested using a biased proto-splice site model of intron insertion in a dicodon analysis. The observed distribution was not found to be consistent with the prediction of the biased proto-splice site model of the insertional theory (Long et al., 1998).

In the further conservative analysis of intron phase association, despite the negative result in the test under biased distribution of proto-splice sites, we assumed the observed biased intron phase distribution with dominant phase zero introns to be a consequence of some unknown insertion mechanism. Table 3 reveals a significant correlation between the introns: all symmetric exons are present in excess (observed numbers are larger than expected); asymmetric exons are less abundant than predicted. The deviations from random expectations in all three species are significant. These observations of highly significant excess of symmetric internal exons and correspondingly fewer asymmetric exons are in accordance with previous observations in general databases in which vertebrate genes predominated (Long, Rosenberg & Gilbert, 1995; Gilbert, de Souza & Long, 1997; Long & Rosenberg, 2000).

How were these patterns created in evolution? The simple model of random intron insertions was rejected by these observations, because both the predictions for single intron proportions (1/3) and for intron association within genes (1/9) are significantly different from the observed numbers. A more sophisticated model of intron insertion would be that introns randomly insert into nonrandomly distributed insertion sites. For example, there are conservative exon sites flanking each intron in many genes, whose sequence is, for example, AG|G, where '|' represents the intron. If AG|G is distributed randomly because of nonrandom factors, for example, amino acid compositions, then inserted introns would possibly create nonrandom distribution of intron phases. This hypothesis was tested in two directions. First, the dicodon usage (frequencies of 64×64 combinations) in six species (*Homo sapiens, D. melanogaster, C. elegans, Schizosaccharomyces pombe, Saccharomyces cerevisiae,* and *A. thaliana*) were computed using GenBank sequence data; each dicodon generated a phase prediction for various hypothetical insertion

sites including G|G, AG|G, (AG|GT), (C/A)AG|A/G ('C/A' means 'C or A'; A/G 'A or G'). It was found that the predicted frequencies from these hypothesized insertion sites were significantly different from the observed frequencies of intron phases (Long et al., 1998). Furthermore, a Monte Carlo simulation was conducted in which the observed introns were randomized into these hypothetic insertion sites in the same genes and then the association of intron phases were analyzed. It was found that these simulations failed to generate patterns similar to the observed excess of symmetric exons (Long & Rosenberg, 2000). These two observations negated the hypothesis of intron insertion into actually distributed proto-splice sites. These sites do not have equal-probable phases, but the patterns are significantly different from the observed ones.

The observed patterns of intron phase distribution are more likely created by exon shuffling. Patthy (1987) pointed out two requirements of exon shuffling: the length of a single inserted exon has to be 3*n* (*n*, a positive integer) to avoid frameshift mutations in the downstream exon(s) of recipient genes. Meanwhile, the phases of introns flanking inserted exons should be identical to the phase of recipient intron, otherwise the reading frame of the inserted exon would be changed even if the inserted exon is symmetric. Thus, if a large number of exon shuffling events take place in genome, then the correlation of intron phases would be amplified. Theoretical analysis also showed that, as a consequence of exon shuffling, one phase (zero phase intron) would predominate even if the phase distribution was equiprobable in a hypothetical beginning world of genes (Fedorov et al., 1998). Observed patterns are consistent with these predictions based on exon shuffling and do not support both simple and complex forms of insertional theory of introns. Therefore, genes of these three nonvertebrate species evolved likely via exon shuffling, supporting a ubiquitous role of this mechanism in evolution of exon–intron structures.

Lynch (2002) recently suggested that intron sliding through expansion/contraction (Stoltzfus et al., 1997) might involve 3*n* (*n* is an integer) nucleotides to avoid frameshift in the downstream exon and thus favor phase zero introns over phase 1 and 2 introns. The rationale is that a 3*n* nucleotide expansion would results in the loss of one or more codons for phase zero intron while for phase one and two introns, one or more codons would be lost and a codon would be changed into a stop codon or other deleterious non-synonymous codon. In this model, all different sequences were assumed to be equally likely contributors to splicing signals. For example, 3/64 of the codon splicing events was assumed to produce a stop codon. However, this assumption is incorrect. As ample evidence from previous analyses of the splicing process have shown, there are only a few types of conservative sites in both exon and intron sequences required for recognition and reactions of splicing processes in introns of all three phases (Horowitz & Krainer, 1974; Reed, 1996; Long et al., 1998). Under such restriction, the frequency of the induced stop codons is much lower than 3/64 and cannot interpret the observed big difference (∼20%) between phase-zero and phase-one/two introns (e.g. Long, Rosenberg & Gilbert, 1995) unless a high rate of sliding is assumed. Further, this model cannot interpret the significant correlation between module boundaries and intron positions observed by De Souza et al. (1996, 1998) and the correlation between module boundaries and symmetrical intron phase combinations found by Kaessmann et al. (2002).

Acknowledgements

We acknowledge an NSF grant and a Packard Fellowship in Science and Engineering to support part of the project of Manyuan Long's laboratory.

References

Adams, M.D., S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, R.A. Hoskins, R.F. Galle, et al., 2000. The genome sequence of *Drosophila melanogaster*. Science 287(5461): 2185–2195.

AGI (The Arabidopsis Genome Initiative), 2000. Analysis of the genome sequence of the flowering plant. Nature 408: 796–815.

Attwood, T.K., 2000. The Babel of bioinformatics. Science 290: 471–473.

Begun, D.J., 1997. Origin and evolution of a new gene descended from alcohol dehydrogenase in Drosophila. Genetics 145: 375–382.

Betrán, E. & M. Long, 2002. Expansion of genome coding regions by acquisition of new genes. Genetica 115: 65–80.

Betrán, E., W. Wang, L. Jin & M. Long, 2002. Evolution of the phosphoglycerate mutase processed gene in human and chimpanzee revealing the origin of a new primate gene. Mol. Biol. Evol. 19: 654–663.

Boeke, J.D. & O.K. Pickeral, 1999. Retroshuffling the genomic deck. Nature 398: 108–109, 111.

Brookfield, J.F. & P.M. Sharp, 1994. Neutralism and selectionism face up to DNA data. Trends Genet. 10: 109–111.

Brosius, J., 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238: 115–134.

Brosius, J., 2003. The contribution of RNAs and retroposition to evolutionary novelties. Genetica 118: 99–115.

Burks, C., M.J. Cinkosky, P. Gilna, J.E. Hayden, Y. Abe, E.J. Atencio, S. Barnhouse, D. Benton, C.A. Buenafe & K.E. Cumella, 1990. GenBank: current status and future directions. Meth. Enzymol. 183: 3–22.

Cerff, R., 1995. The chimeric nature of nuclear genoms and the antiquity of introns as demonstrated by the GAPDH gene system, pp. 205–228 in Tracing Biological Evolution in Protein and Gene Structures, edited by M. Go & P. Schimmel. Elsevier, Amsterdam.

CESC (The *C. elegans* Sequencing Consortium), 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282: 2012–2018.

Chen, J.J., B.J. Janssen, A. Williams & N. Sinha, 1997. A gene fusion at a homeobox locus: alterations in leaf shape and implications for morphological evolution. Plant Cell 9: 1289–1304.

De Souza, S.J., M. Long, R.J. Klein, S. Roy, S. Lin & 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. USA 95: 5094–5099.

De Souza, S.J., M. Long, L. Schoenbach, S.W. Roy & W. Gilbert, 1996. Intron positions correlate with module boundaries in ancient proteins. Proc. Natl. Acad. Sci. USA 93: 14632–14636.

Dierick, H.A., J.F.B. Mercer & T.W. Glover, 1997. A phosphoglycerate mutase brain isoform (*PGAM1*) pseudogene is localized within the human Menkes disease gene (*ATP7A*). Gene 198: 37–41.

Domon, C. & A. Steinmetz, 1994. Exon shuffling in anther-specific genes from sunflower. Mol. Gen. Genet. 244: 312–317.

Dorit, R.L., L. Schoenbach & W. Gilbert, 1991. How big is the universe of exons? Science 250: 1377–1382.

Eddy, S.R., 2001. Non-coding RNA genes and the modern RNA world. Nat. Rev. Genet. 2: 919–929.

Fedorov, A., L. Fedorova, V. Starshenko, V. Filatov & E. Grigor'ev, 1998. Influence of exon duplication on intron and exon phase distribution. J. Mol. Evol. 46: 263–271.

Fedorov, A., G. Suboch, M. Bujakov & L. Fedorova, 1992. Analysis of nonuniformity in intron phase distribution. Nucl. Acids Res. 20(10): 2553–2557.

Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann et al., 1995. The minimal gene complement of *Mycoplasma*. Science 270: 397–403.

Gilbert, W., 1978. Why gene in pieces? Nature 271(5645): 501.

Gilbert, W., 1987. The exon theory of genes. Cold Spring Harb. Symp. Quant. Biol. 52: 901–905.

Gilbert, W., S.J. de Souza & M. Long, 1997. Origin of genes. Proc. Natl. Acad. Sci. USA 94: 7698–7703.

Goodman, M., C.A. Porter, J. Czelusniak, S.L. Page, H. Schneider, J. Shoshani, G. Gunnell & C.P. Groves, 1998. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol. Phyl. Evol. 9: 585–598.

Grisolia, S. & B.K. Joyce, 1959. Distribution of two types of phosphoglyceric acid mutase, diphosphoglycerate mutase and D-2, 3-dipphosphoglyceric acid. J. Biol. Chem. 234, 6: 1335–1337.

Grisolia, S. & J. Carreras, 1975. Phosphoglycerate mutase from Yeast, chicken, breast muscle and kidney (2,3-PGA-dependent). Meth. Enzymol. 42: 435–450.

Gu, Z., A. Cavalcanti, F.C. Chen, P. Bouman & W.H. Li, 2002. Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. Mol. Biol. Evol. 19: 256–262.

Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B.C. Li & R. Herrmann, 1996. Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae Nucl. Acids Res. 24: 4420–4449.

Horowitz, D.S. & A.R. Krainer, 1994. Mechanisms for selecting 5′ splice sites in mammalian pre-mRNA splicing. Trends Genet. 10: 100–106.

Jeffs, P. & M. Ashburner, 1991. Processed pseudogenes in *Drosophila*. Proc. R. Soc. Lond. B. 244: 151–159.

Kaessmann, H., S. Zöllner, A. Nekrutenko & W.H. Li, 2002. Signatures of domain shuffling in the human genome. Genome Res. 12: 1642–1650.

Lander, et al., 2001. Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Langley, C.H., E. Montgomery & W.F. Quattlebaum, 1982. Restriction map variation in the Adh region of *Drosophila*. Proc. Natl. Acad. Sci. USA 79: 5631–5635.

Long, M., 2001. Evolution of novel genes. Curr. Opin. Genet. Dev. 11: 673–680.

Long, M. & M. Deutsch, 1999. Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns. Mol. Biol. Evol. 16: 1528–1534.

Long, M., W. Wang & J. Zhang, 1999. Origin of new genes and source for N-terminal domain of the chimerical gene, jingwei, in *Drosophila*. Gene 238: 135–141.

Long, M. & 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.

Long, M., C. Rosenberg & W. Gilbert, 1995. Intron phase correlations and the evolution of the intron/exon structure of genes. Proc. Natl. Acad. Sci. USA 92(26): 12495–12499.

Long, M., S.J. de Souza & W. Gilbert, 1995. Evolution of intron/exon structure of eukaryotic genes. Curr. Opin. Genet. Dev. 5: 774–778.

Long, M., S.J. de Souza, C. Rosenberg & W. Gilbert, 1996. Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor. Proc. Natl. Acad. Sci. USA 93: 7727–7731.

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

Long, M. & C.H. Langley, 1993. Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. Science 260: 91–95.

Lynch, M., 2002. Intron evolution as a population-genetic process. Proc. Natl. Acad. Sci. USA 99: 6118–6123.

Nugent, J.M. & J.D. Palmer, 1991. RNA-mediated transfer of the gene coxII from the mitochondrion to the nucleus during flowering plant evolution. Cell 66: 473–481.

Nurminsky, D.I., M.V. Nurminskaya, D. De Aguiar & D.L. Hartl, 1998. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. Nature 396: 572–575.

Ohno, S., 1970. Evolution by Gene Duplication. Springer, New York.

Palmer, J.D., 1985. Comparative organization of chloroplast genomes. Annu. Rev. Genet. 19: 325–354.

Patthy, L., 1987. Intron-dependent evolution: preferred types of exons and introns. FEBS Lett. 214: 1–7.

Patthy, L., 1991. Modular exchange principles in proteins. Curr. Opin. Struct. Biol. 1: 351–361.

Patthy, L., 1995. Protein Evolution by Exon-shuffling. Molecular biology intelligence unit, edited by R.G. Landes. Springer, Austin, TX.

Pearson, W.R., 1994. Using the FASTA program to search protein and DNA sequence databases. Meth. Mol. Biol. 24: 307–331.

Reed, R., 1996. Initial splice-site recognition and pairing during pre-mRNA splicing. Curr. Opin. Genet. Dev. 6: 215–220.

Roise, D., S.J. Horvath, J.M. Tomich, J.H. Richards & G. Schatz, 1986. A chemically synthesized pre-sequence of an imported mitochondrial protein can form an amphiphilic helix and perturb natural and artificial phospholipid bilayers. EMBO J. 5: 1327–1334.

Rubin, G.M., M.D. Yandell, J.R. Wortman, G.L. Gabor Miklos, C.R. Nelson, I.K. Hariharan et al., 2000. Comparative genomics of the eukaryotes. Science 287: 2204–2215.

Schatz, G. & B. Dobberstein, 1996. Common principles of protein translocation across membranes. Science 271: 1519–1526.

Stoltzfus, A., J.M. Logsdon Jr., J.D. Palmer & W.F. Doolittle, 1997. Intron “sliding” and the diversity of intron positions. Proc. Natl. Acad. Sci. USA 94: 10739–10744.

Venter, J.C. et al., 2001. The sequence of the human genome. Science 291: 1304–1351.

Wang, W., J. Zhang, C. Alvarez, A. Llopart & M. Long, 2000. The origin of the *Jingwei* gene and the complex modular structure of its parental gene, yellow emperor, in *Drosophila melanogaster*. Mol. Biol. Evol. 17: 1294–1301.

Wang, W., F.G. Brunet, E. Nevo & M. Long, 2002a. Origin of sphinx, a young chimeric RNA gene in *Drosophila melanogaster.* Proc. Natl. Acad. Sci. USA 99: 4448–4453.

Wang, W., K. Thornton, A. Berry & M. Long, 2002b. Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. Science 295: 134–137.

Wegener, S. & U.K. Schmitz, 1993. The presequence of cytochrome c1 from potato mitochondria is encoded on four exons. Curr. Genet. 24: 256–259.