Genetica **115:** 65–80, 2002.

Expansion of genome coding regions by acquisition of new genes

Esther Betrán & Manyuan Long
Department of Ecology and Evolution, 304 Zoology Building, The University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA (Phone: (773) 834 3567; Fax: (773) 702 9740; E-mails: mlong@midway.uchicago.edu, ebetran@midway.uchicago.edu)

Key words: new genes, G-value paradox, natural selection

Abstract

As it is the case for non-coding regions, the coding regions of organisms can be expanded or shrunk during evolutionary processes. However, the dynamics of coding regions are expected to be more correlated with functional complexity and diversity than are the dynamics of non-coding regions. Hence, it is interesting to investigate the increase of diversity in coding regions – the origin and evolution of new genes – because this provides a new component to the genetic variation underlying the diversity of living organisms. Here, we examine what is known about the mechanisms responsible for the increase in gene number. Every mechanism affects genomes in a distinct way and to a different extent and it appears that certain organisms favor particular mechanisms. The detail of some interesting gene acquisitions reveals the extreme dynamism of genomes. Finally, we discuss what is known about the fate of new genes and conclude that many of the acquisitions are likely to have been driven by natural selection; they increase functional complexity, diversity, and/or adaptation of species. Despite this, the correlation between complexity of life and gene number is low and closely related species (with very similar life histories) can have very different number of genes. We call this phenomenon the G-value paradox.

Introduction

A striking evolutionary feature of genomes is the difference in genome size among various organisms. Early efforts in the study of genome evolution were an attempt to understand the relationship between genome size and biological complexity, which gave rise to a generalized observation summarized as the C-value paradox. The term C-value paradox applies to the fact that genome size does not closely correlate with the biological complexity of organisms and thus with the phylogenetic relationships between species (Li, 1997). However, a genome can be partitioned into two portions: coding and non-coding regions. Does size of the coding region correlate with biological complexity? Or, asked in another way, does the paradox as observed at the level of the whole genome disappear for the coding region?

An analysis of ample genome sequence information from various genome projects rejects a positive answer to the above questions. The first impression is that the genomes of different organisms encode different number of genes (Table 1), which allows a further analysis of genome coding region evolution. When examining the relationship between gene number and species differences, we see two phenomena. First, closely related organisms with similar life histories can contain very different gene numbers, for example, the genomes of *Mycoplasma pneumonia* contain 50% more genes than *M. genitalium*. And second, gene number does not clearly correlates with complexity. For example, *Arabidopsis thaliana* and *Drosophila melanogaster* contain around 25,500 and 14,000 genes, respectively; the former encodes nearly twice as many genes as the latter, but it is not clear that it is twice as complex. As in the C-value paradox, these simple numerical analyses reveal similar phenomena: gene number does not tightly correlate with biological complexity. Thus, following the analogy with the C-value paradox, we can also simply call the phenomena described above as the G-(gene) value paradox.

The G-value paradox raises a question related to genomic evolution. Why does not the number of genes correlate highly with biological complexity? That gene numbers differ between organisms suggest

Table 1. Genome size and gene number in various organisms

	Genome size	Gene number
Homo sapiens	3.0×10^9	40,000
Fugu rubripes	4.0×10^8	60,000
Drosophila melanogaster	1.2×10^8	13,601
Caenorhabditis elegans	1.0×10^8	18,424
Ciona intestinalis	1.6×10^8	15,500
Arabidopsis thaliana	1.2×10^8	25,498
Saccharomyces cerevisiae	1.2×10^7	6,241
Plasmodium falciparum	3.0×10^7	6,630
Methanococcus jannaschi	1.7×10^6	1,738
Archaeoglobus fulgidus	2.2×10^6	2,436
Escherichia coli	4.6×10^6	4,288
Bacillus subtilis	4.2×10^6	4,100
Haemophilus influenzae	1.8×10^6	1,743
Treponema pallidum	1.1×10^6	1,041
Mycoplasma genitalium	5.8×10^5	470
Mycoplasma pneumoniae	8.1×10^5	716

a process of birth and death and, more interestingly, poses the general question: How do coding regions of genome expand by acquiring new genes and how is that related to complexity? A great deal of effort has been made to accomplish some understanding in this direction. One major accomplishment is the comprehension of various evolutionary mechanisms to create new genes. We will review various major mechanisms that increase the amount of coding region, their relative importance and finally we will discuss the general fate of new genes. A review of what is known will provide a basis for understanding how the expansion of coding regions might be related to complexity.

Duplication

Gene duplication was the first proposed mechanism for the increase of genetic content (Haldane, 1932; Muller, 1935). As early as the 1930's, the first duplication was identified; the Bar duplication. This duplication occurs as a consequence of unequal crossing-over involving section 16A of the X chromosome in *Drosophila melanogaster* (Bridges, 1936). Duplications can be classified according to their size (Ohno, 1970): domain duplication, gene duplication, segmental duplication, and genome duplication. The mechanisms that give rise to these events are different. Domain, gene or segment duplication are generated by means of unequal crossing-over. Unequal crossing-over during meiosis between misaligned sequences generates a chromosome with a duplicated region and a chromosome with a deleted region. Repetitive sequences can increase the chances of unequal crossing-over. Genome duplication takes place as a consequence of errors during meiosis or mitosis that end in the lack of reduction division within species or in hybrids (Haldane, 1932; Ohno, 1970).

Innumerable examples of domain, gene, and genome duplicates have been added to the literature (Ohno, 1970; Graur & Li, 2000; Otto & Whitton, 2000) but, in this era of post-genomics, the importance of duplication is even greater. During the year 2001, the number of completely sequenced eukaryotic genomes has increased to five: *Saccharomyces* (Goffeau et al., 1996), *C. elegans* (The C. elegans Sequencing Consortium, 1998), *Arabidopsis* (The Arabidopsis Genome Initiative, 2000), *Drosophila melanogaster* (Adams et al., 2000), and *H. sapiens* (Lander et al., 2001; Venter et al., 2001). Percentages of the genes that belong to identifiable families of duplicates in these genomes are 30, 48, 60, 40, and 38% of the genome, respectively (Blanc et al., 2000; Rubin et al., 2000; Ball & Cherry, 2001; Li et al., 2001). There are enormous numbers of duplicated coding regions in these genomes.

Genomic organization of these duplicated genes differs depending on the lineage. Yeast, worm, *Arabidopsis*, and human show clear blocks of duplicated genes (Friedman & Hughes, 2001; Lander et al., 2001; Venter et al., 2001). The yeast genome contains 39 duplicated blocks, all except one of which are inter-chromosomal. Twenty eight of these blocks are ancient. These data are consistent with a polyploidization event in the yeast lineage (Wolfe & Shields, 1997; Friedman & Hughes, 2001). The orientation of duplicated blocks with respect to their centromeres also supports this view (Wolfe & Shields, 1997). In the worm, only five duplicated pairs of segments (four of them recent) were found, all of them within the same chromosome (Friedman & Hughes, 2001), suggesting that they are a consequence of local duplication rather than genome polyploidization. The *Drosophila* sequence shows that most of the duplications are local (Friedman & Hughes, 2001). The human genome shows extensive duplication of segments; however, there is no adequate evidence for ancient polyploidization (Lander et al., 2001; Venter et al., 2001). The *Arabidopsis* genome is the most extensively duplicated and reorganized of the sequenced genomes (Blanc et al., 2000; Bancroft, 2001). The genome sequencing has provided clear evidence of

large regions that are duplicated. Additionally, transposition and/or translocation might have to be invoked to explain the organization of *Arabidopsis* genome today (Blanc et al., 2000).

All these genome data illustrate that, by far, the most general and important mechanism to generate new copies of genes is the duplication of genes and/or genomes. Duplications of genes and genomes are believed to result in the increase in complexity of organisms and an increase in rates of diversification (Ohno, 1970). It has, for example, been suggested that the ancestors of vertebrates had only ∼15,000 genes while vertebrates have ∼60,000 genes (Spring, 1997; Ohno, 1999). Two rounds of genome-wide duplication have been hypothesized to explain this and, hence, the diversity of forms generated during vertebrate evolution (Spring, 1997). Recently, this one-to-four rule has been extended to a one-to-eight rule in fish (Meyer & Schartl, 1999). The genome duplication and the diversification of fish seem to coincide in time, making fish the most successful group of vertebrates (Meyer & Schartl, 1999). The extent of the possible correlation between polyploidization and levels of species diversification remains to be shown. However, it is a suggestive working hypothesis.

Above, we summarized general patterns of gene duplication; the detail of the following examples should reveal the deeper understanding we have about how duplications lead to the evolution of new functions.

Adh-Adhr *in* Drosophila*: tandem duplication, close linkage and dicistronic expression*

Adh and *Adhr* genes of *Drosophila* are tandem duplicates, as evidenced by two observations: one is their significant sequence similarity (about 40% amino acid identity); the other their common intron/exon structure (Schaefer & Aquadro, 1987; Rat, Veuille & Lepesant, 1991). This duplication may be as old as 180 million years (My) (Russo, Takezaki & Nei, 1995). *Adh* functions to detoxify dietary alcohols (Chambers, 1988; Ashburner, 1998) but *Adhr* has unknown function, although it seems not to be a NAD-dependent dehydrogenase (Jeffs, Holmes & Ashburner, 1994). After duplication, *Adh* and *Adhr* were very close, separated by ∼300 bp (Figure 1). The expression of both genes has been studied in great detail in *D. melanogaster*. There are only two kinds of transcripts: one (∼1 kb) includes only *Adh* and the other (∼2 kb) is dicistronic including both genes (Brogna & Ahsburner, 1997), suggesting that dicistronic transcription is the only means by which ADHR protein is produced. Brogna and Ashburner (1997) detected ADHR with antibodies. Within the genus *Drosophila* the ADHR protein is evolving slightly slower than the ADH protein, supporting its functionality (Albalat Marfany & González-Duarte, 1994). The dicistronic transcript includes the intergenic region and it is rare, its abundance being roughly 1% or so that of the monocistronic *Adh* transcript. The relative level of the two transcripts is controlled by the efficiency of the termination and polyadenylation signals in the intergenic region (Brogna & Ashburner, 1997): the more transcriptional read-through, the more dicistronic transcript is produced.

Recent data on the expression of *Adhr* in *D. lebanonensis* and *D. buzzatii* (Betrán & Ashburner, 2000) suggest that the co-transcription of *Adh* and *Adhr* is the primitive state, and that *Adhr* acquired independent (monocistronic) transcription in the subgenus *Drosophila* (Figure 1). Given that these two genes originated from an adjacent duplication, most parsimoniously the two genes have been co-transcribed in many species (Figure 1) since the duplication event and have evolved different functions (see above) despite that. This is an example of how a duplicated gene can be expressed without a promoter if it is sufficiently close to the parental gene and, still, evolve a new function.

Origin of antifreeze protein genes

The origin of the *antifreeze glycoproteins* (*AFGPs*) genes in fishes is one of the best examples of how a similar function can originate independently under a common environmental pressure (Chen, Devries & Cheng, 1997a, b; Logsdon & Doolittle, 1997). These proteins function by preventing intracellular crystal growth (Knight, Cheng & Devries, 1991). Chen, Devries and Cheng (1997a) studied the *AFGP* gene family in Antarctic notothenioid fish and found certain homology of this gene with the trypsinogen gene. Sequence divergence between the homologous regions of the two genes was very small (4–7%). Chen, Devries and Cheng (1997a) estimated that the transformation of the *trypsinogen* gene (proteinase gene) into an ice-binding protein happened recently. The young ages of these antifreeze genes allowed authors to follow the whole detail of the process and to correlate their appearance with the freezing of the Antarctic Ocean 10–14 My ago.

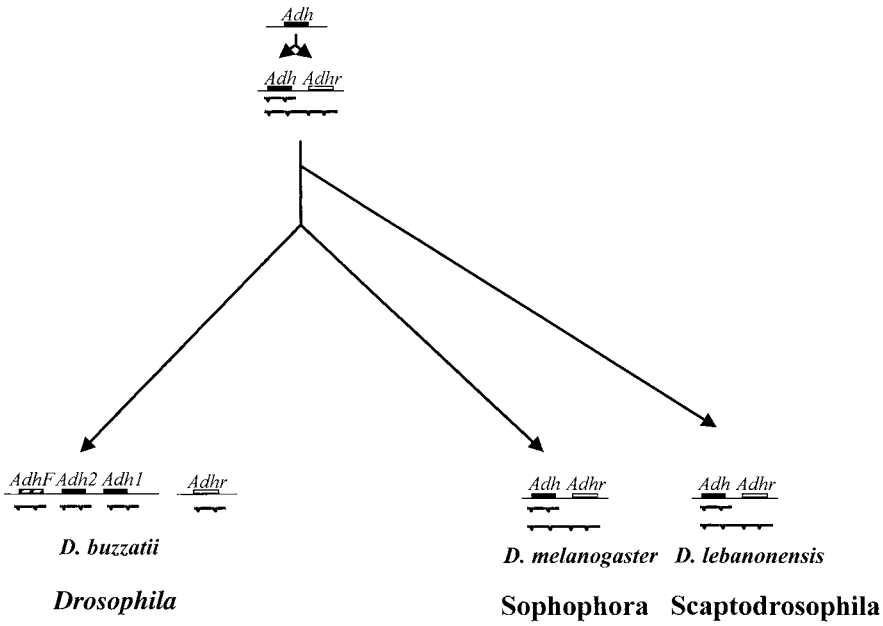

Figure 1. Distribution and expression of *Adh* and *Adhr* in the genus *Drosophila* (Russo et al., 1995; Brogna & Ashburner, 1997; Betrán & Ashburner, 2000). Lines indicate species and gene phylogenies. A double arrow indicates a duplication event. Expression is shown under the gene.

Figure 2 summarizes the mechanism in which these *antifreeze protein* genes originated. Given the common elements in the sequence, it seems that the *AFGP* gene in Antarctic notothenioid fish was born as a *trypsinogen* gene that underwent a big deletion and a series of tandem duplications of a nine nucleotide (Thr-Ala-Ala) motif. Thr is glycosylated and that allows the protein to bind ice. The spacers provide sites of postranscriptional proteolytic cleavage. A completely new function was born from pieces and an intron of the *trypsinogen* gene. This occurred under the appropriate selective pressure allowing a species to adapt to a new environment.

In addition, Chen, Devries and Cheng (1997b) studied the *AFGP* gene in the Arctic cod. They found that this gene is not related to *trypsinogen* and shows a different intron/exon structure despite the repetitive nature of both proteins (Figure 2). Antarctic notothenioid fish and Arctic cod are in different orders and although they produce nearly identical antifreeze proteins, the origin of these proteins is apparently unrelated.

This is a unique example because it is the only known case in which duplication of a small nonfunctional motif played a role in adaptation repeatedly in different lineages.

Processed genes

Processed copies of genes are copies of genes created by retroposition in which an mRNA of a gene is reverse-transcribed and inserted in the genome. Generation of processed copies of genes has been experimentally studied in human cells (Maestre et al., 1995). By using an assay for detection of reverse transcription and insertion events from a naive reporter gene, Maestre et al. (1995) demonstrated that genes indeed are reverse-transcribed and integrated in the genome. The integrated sequence has the length of the mRNA of the original gene, with a polyA tail and, often, direct flanking repeats (<15 bp) (Figure 3). Recently, Esnault, Maestre and Heidmann (2000) showed that LINE (long interspersed element) retrotransposon enzymatic machinery was involved in generating processed copies of genes in human cells. LINEs are endogenous mobile elements, which transpose via reverse transcription of their own transcript and have accumulated in $\sim$100,000 copies in mammals. Furthermore, it was shown that the presence of the LINE machinery increases 10-fold the rate at which processed copies of a reporter gene are produced in a cell culture assay, suggesting an important role of the LINE retrotransposons in the origin of processed copies of genes.

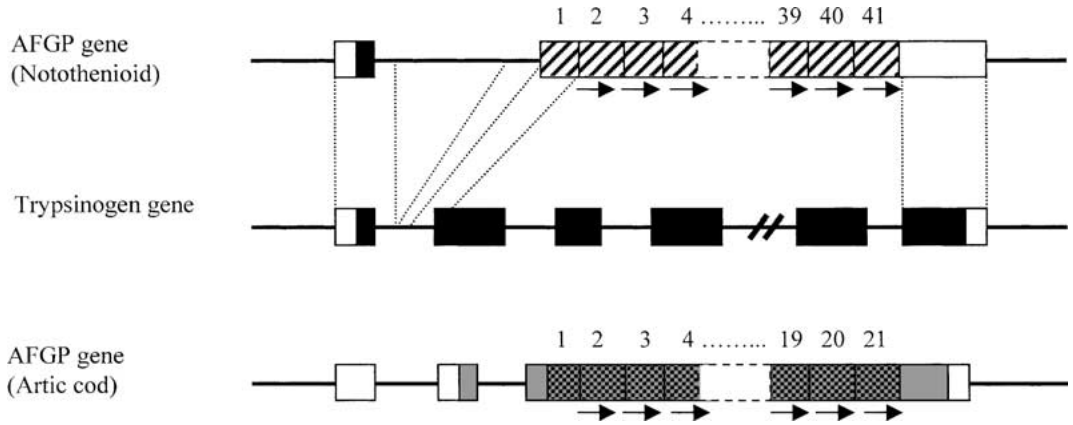

Figure 2. Gene structure of the AFGP genes in polar fishes. Regions of homology with other genes are shown when known. Arrows are used to describe tandem duplications.

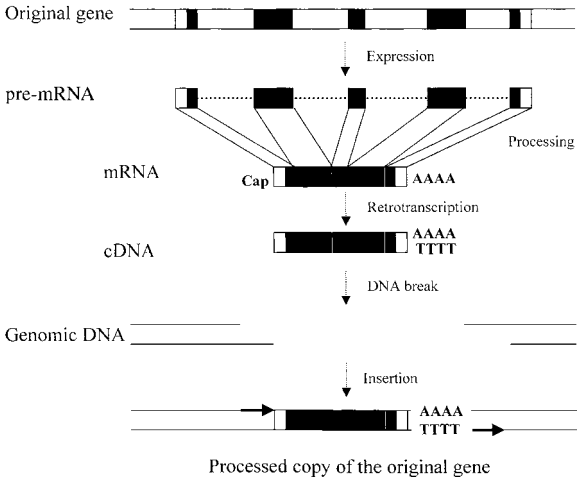

Figure 3. Generation of processed copies of genes in the genome.

Processed genes were first described in mammals because of their high copy number in mammalian genomes (Wagner, 1986). For example, in human chromosome 22 there are 134 pseudogenes. Out of these 134, 82% are sequences 'with homology with known genes but that lack intron/exon boundaries', a conspicuous feature of processed pseudogenes (Dunham et al., 1999). Now it is clear that they are also present in other organisms (Mighell et al., 2000). LINE-mediated retrotransposition of protein coding genes (Esnault, Maestre & Heidmann, 2000) could explain why processed copies of genes are more abundant in mammalian genomes than in the LINE-element-poor lineages as *Drosophila* (Petrov, Lozovskaya & Hartl, 1996) or *C. elegans* (Harrison, Echols, & Gerstein, 2001).

Processed copies of genes are often called pseudogenes because these elements do not carry promoter regions and insert randomly in the genome (Graur & Li, 2000). In addition, processed copies of genes often have traits (deletions and/or premature stop codons) that preclude their functionality. Because of that, retrotransposed paralogous copies of genes were first viewed as dead-end sequences. Although it is still accepted that many processed copies of genes are indeed pseudogenes (Mighell et al., 2000), an increasing number of examples have shown that this may be an important mechanism for generation of new genes (Brosius, 1991, 1999).

Two well-studied instances in humans are *Pyruvate dehydrogenase 2* (*Pdha2*) and *Phosphoglycerate kinase 2* (*Pgk-2*). Both of these genes are intronless autosomal copies of intron-containing X-linked genes (*Pdha1* and *Pgk-1*). The original genes, *Pdha1* and *Pgk-1*, have constitutive function, while the retroposed copies *Pdha2* and *Pgk-2* are expressed only in testes, suggesting newly evolved specific function. They share 86 and 87% protein sequence identities with the parental genes, respectively (McCarrey, 1987; Dahl et al., 1990; Fitzgerald et al., 1996; McCarrey et al., 1996). How did these new genes acquire their new promoters? The *Pdha2* promoter region derives from a recent insertion of *Pdha1* gene promoter into 5′ region of *Pdha2* (Datta et al., 1999), while the *Pgk-2* promoter region arose originally from an aberrant transcript that included a region transcribed from promoter region of *Pgk-1* (McCarrey, 1987).

We surveyed the literature for more cases of functional protein-coding processed genes in mammals

Table 2. Mammalian genes generated by retrotransposition

Retrotransposed gene name, function, and expression	Original gene name, function, and expression	Distribution	Hallmarks of retrosequences	Age	References
Pgk-2; testes, Cr 19	*Pgk-1*, ubiquitously, Cr X	Mammals	Intronless (abnormal 5′ long mature mRNA)	125 Mya	McCarrey, 1987, 1990; McCarrey et al., 1996
Pdha2; testes. Cr 4	*Pdha1*; ubiquitously, Cr X	Placentals	Intronless	∼70 Mya	Dahl et al., 1990; Datta et al., 1999
Cetn1, testes, Cr 18	*Cetn2*, Cr X	Mammals	Intronless and flanking direct repeats	>75 Mya	Hart et al., 1999
Cγ subunit of cAMP-dependent protein kinase, testes. Cr 9	Cα subunit of cAMP-dependent protein kinase; ubiquitously. Cr 19	Catarrhini primates	Intronless, poly A tail and flanking direct repeats	40 Mya	Reiton et al., 1998
GK, testes, Cr 4	*GK*; constitutively. Cr X	Human	Intronless	11 Mya	Sargent et al., 1994; Pan et al., 1999
GA733-1, placenta, Cr 1	*GA733-2*; placenta, Cr 4	Human	Intronless	300 Mya	Linnenbach et al., 1993;
OTF3C, pancreatic islet, Cr 8	*Oct3*, adult tissues, Cr 6	Mammals	Intronless	∼75 Mya	Takeda et al., 1992
XAP-5 like, testes, Cr 6	*XAP-5*, Cr X	Human and mouse	Intronless	>75 Mya	Sedlacek et al., 1999
αCP1, ubiquitous, Cr 2	*αCP2,* Cr 12	Human and mouse	Intronless	∼75–120 Mya	Makeyev et al., 1999
CDY, testes, Cr Y	CDYL, ubiquitous, Cr 13	Human–squirrel monkey	Intronless	30–40 Mya	Lahn and Page, 1999
hnRNP, testes, Cr 11	*hnRNP*, Cr X	Human	Intronless	_	Elliot et al., 2000
GLUD2, retina, testes and brain. Cr X	*GLUD1*, ubiquitous. Cr 10	Human but not mouse	Intronless	<70 Mya	Shashidharan et al., 1994
CaM like, epitelial cells	CaM	Human, rat and chicken	Intronless	≫75–120 Mya	Rhyner et al., 1992
SRp46, ubiquitous, Cr 11	PR264/SC35. Cr 17	Human and simians	Intronless	∼89 Mya	Soret et al., 1998
eIF4E2, placenta, Cr 20	*eIF4E1*, placenta, Cr 4	Human not mouse	Intronless, direct repeats	<70 Mya	Gao et al., 1998
ADAM20 and *ADAM21*, testes, Cr 14	*ADAM9*	Human but not macaque	Intronless	<20 Mya	Poindexter et al., 1999; Hooft van Huijsduijnen, 1998
SP-100-HMG, Cr 2	*HMG1*, Cr 13	Chimp and gorilla	Intronless-chimeric	35 Mya	Rogalla et al., 2000

Table 2. (*continued*)

Retrotransposed gene name, function, and expression	Original gene name, function, and expression	Distribution	Hallmarks of retrosequences	Age	References
MYCL2, testes, Cr X	*MYCL1*	Human	Intronless	_	Morton et al., 1989 Robertson et al., 1991
HPRT2, liver	*HPRT1*, ubiquitous, Cr X	Marsupials	Intronless	130–150 Mya	Noyce et al., 1997
Pabp2, testes	*Pabp1*, testes and somatic cells	Mouse only	Intronless	~80 Mya	Kleene et al., 1998; Kleene et al., 1999
AdoMetDC like, liver	*AdoMetDC*, ubiquitous	Mouse and possibly rat	Intronless	_	Persson et al.,1995
G6pd2, testes	*G6pd1*, ubiquitous	Mouse	Intronless	<30 Mya	Hendriksen et al., 1997
PMSE2b, transcribed	*PMSE2*	Mouse	Intronless and LINE1 sequence	At least 75–120 Mya	Zaiss and Kloetzel, 1999
Preproinsulin I, pancreas, Cr 1	*Preproinsulin II*, pancreas, Cr 1	Mouse and rat	Intronless (abnormal 5′ long mature mRNA)	~35 Mya	Soares et al., 1985
Zfa; testes. Cr 10	*Zfx*; ubiquitous. Cr X	Mouse	Intronless	~5 Mya	Ashworth et al., 1990; Louh and Page, 1994
U2af1-rs1, Cr 11	*U2af1-rs2* Cr X	Mouse	Intronless and inserted in an intron of another gene	<80 Mya	Nabetani et al., 1997
Ubc9-φ1 and *Ubc9-φ2*	*Ubc9*	Mouse	Intronless, poly-A tail and direct repeats	~80 Mya	Tsytsykova et al., 1998
Supt4h2, Cr 10	*Supt4h*, Cr 11	Mouse	Intronless	<70 Mya	Chiang et al., 1998
Pem2, epididymis, Cr 4	*Pem1*, testes, Cr X	Rat	Intronless	_	Nhim et al., 1997

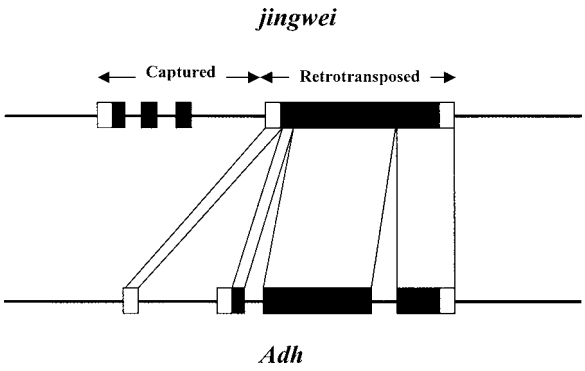

Figure 4. Structure of the chimerical gene: *jingwei*. Homology with the *Adh* gene of the retrotransposed region of the gene is shown.

and list the results in Table 2. These integrated sequences have the length of the mRNA of the original gene; sometimes polyA tracts still remain and direct repeats flank the inserted sequence. Interestingly, some cases show partially processed copies of genes or processed copies of abnormal transcripts that include 5′ promoter regions of the parental gene, for example, *Preproinsulin I* in mouse and *Pgk2* in human. In addition, there are also many functionally important RNA genes, which are not translated into proteins, that originated via retroposition, for example, BC1 RNA gene in rodents (Brosius & Gould, 1992) and BC200 RNA gene in primates (Martignetti & Brosius, 1993).

Origin of jingwei *a chimerical processed functional gene in* Drosophila

In *Drosophila*, there are also processed genes. One revealing example is a new chimerical gene called *jingwei* (*jgw*); (Long & Langley, 1993). This gene is located on chromosome 3 in *D. teissieri* and *D. yakuba* and it is not present in the closest relatives. The age of the gene has been estimated to be less than 2.5 My (Long, Wang & Zhang, 1999). A part of *jgw* was initially observed to hybridize to the *alcohol dehydrogenase* (*Adh*) probe in polytenic chromosomes (Langley, Montgomery, & Quattlebaum, 1982). Further analysis suggested that in the ancestor of these two species *Adh* (chromosome 2) generated a processed gene (Jeffs & Ashburner, 1991). The structure of this new gene is depicted in Figure 4. *Jgw* cDNA has been examined (Long & Langley, 1993). *Jgw* has an *Adh*-like 3′ end and a 5′ end (3 exons) recruited from another gene, *yellow emperor* (*ymp*) as shown in Figure 4 (Wang et al., 2000).

The interest in the origin of *jgw* includes not only the described initial molecular events but also the subsequent population dynamics. Population genetics analysis of the gene reveals that natural selection participated in the subsequent evolution of this gene (Long & Langley, 1993). The *Adh*-derived portion was sequenced from 10 *jgw* alleles of *D. teissieri* and 20 of *D. yakuba*. Most polymorphisms were silent as expected for genes under purifying selection. To explore the divergence of *jgw* in the period immediately after the *Adh* retrotransposition but before the split of *D. teissieri* and *D. yakuba*, the *Adh* region of *jgw* was compared with *Adh* of the other species in the *D. melanogaster* subgroup. Eight substitutions took place in that period; all of them replacements. This excess of replacement substitutions is consistent with *jgw* responding to positive selection and evolving a new function. A relative excess of (fixed) replacement substitutions over (fixed) silent substitutions between species (21:16) is apparent when compared to the proportion of replacement polymorphisms over silent polymorphisms within species (4:27). These results reveal that adaptive protein evolution remained important in the evolutionary history of *jgw* after the separation of the two species (McDonald & Kreitman, 1991).

Transposition

The movement of genetic material from one genome location to another is known as transposition. Usually, the ability to transpose is associated with mobile elements however, examples of coding regions that can generate new copies by transposition are also found (Saxena et al., 1996; Yi & Charlesworth, 2000). This phenomenon is different from retrotransposition since the origin of this newly originated copy does not involve mRNA retrotranscription but ectopic recombination. Convincing cases of transposition of genes have been described but its importance genome wide is difficult to quantify because duplication of a gene with subsequent reorganization leaves similar genome fingerprints.

A recent event of transposition of a gene in *Drosophila miranda* involves the ‘resurrection’ of a gene via this means. In *D. miranda*, the fusion between the Y chromosome and an autosome (Muller element C) has created a new sex determining system (Figure 6), (Powell, 1997). The chromosome attached to the Y chromosome is transmitted paternally and it

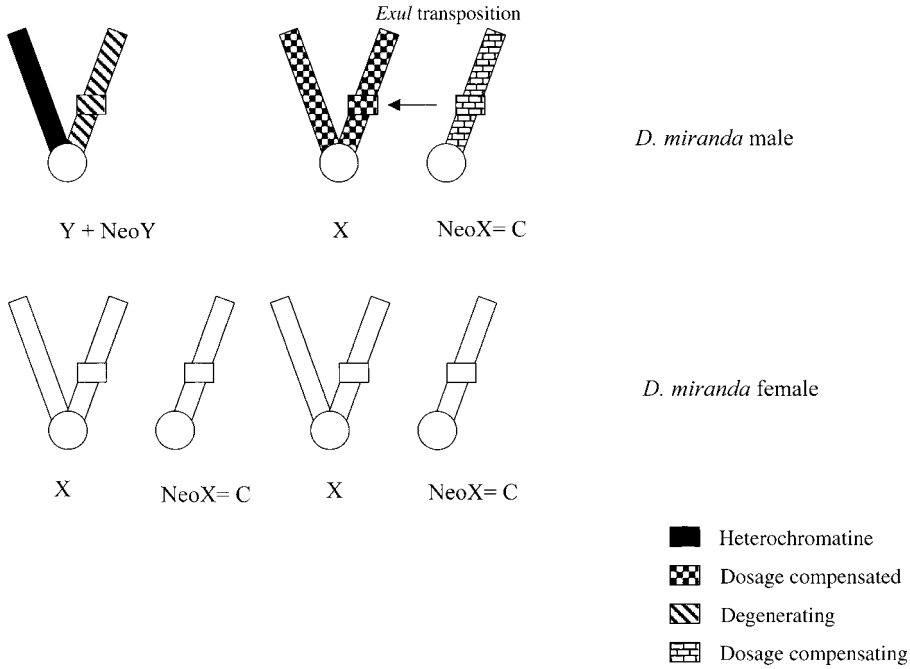

Figure 5. Sex chromosome composition of *D. miranda* male and female are shown. C refers to the Muller element (Powell, 1997). *Exu1* transposition is indicated with an arrow. Status of the chromosomal arm is indicated with a texture.

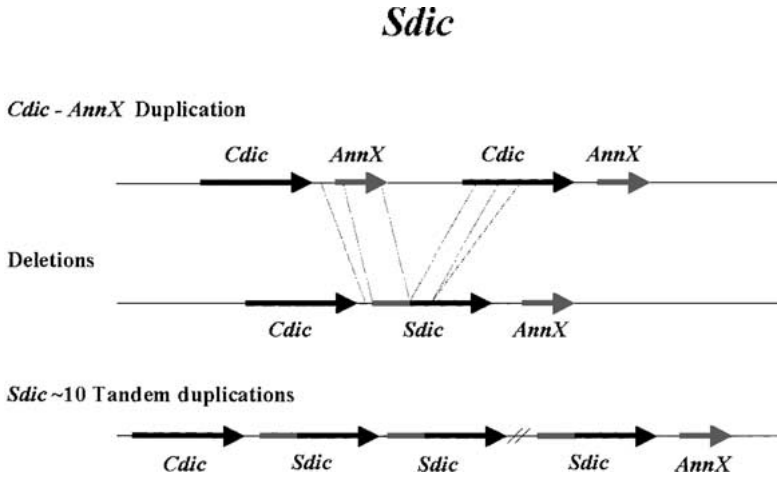

Figure 6. Sdic structure and mechanism of generation. One duplication and three deletions gave birth to a new chimerical gene. *Sdic* duplicated later on ~10 times.

is not undergoing crossing-over. This chromosome is called neo-Y and its non-fused homolog is called neo-X (Figure 5). In these developing sex chromosomes, two main processes are taking place: the neo-Y chromosome is degenerating and the neo-X chromosome is developing dosage compensation (Steinemann & Steinemann, 1998). Yi and Charlesworth (2000) found that a gene, *exuperantia 1* (*exu1*), (originally located in the neo-X chromosome) had recently transposed a copy into the ancestral X chromosome. The transposition fixed in the population in very short time: the transposed gene is not found in closely related species and its fixation swept variation away (Yi & Charlesworth, 2000). Its rapid fixation on the ancestral X implies very likely a selective advantage for its new location. *Exu1* plays a role in proper localization

of maternal mRNA and also in spermatogenesis. The degeneration of the neo-Y chromosome will promote deterioration of the neo-Y copy of *exu1*. That is expected to have a deleterious effect resulting in favorable selection for the transposed copy. Authors believe that the mechanism of transposition could have been ectopic exchange due to some homology between donor and target site; positive selection would have driven the copy to fixation.

Exon shuffling

Exon shuffling was first proposed in the late 70's. At that time the intron/exon structure of eukaryote genes was newly described and scientists recognized the presence of the same functional domains in different proteins. Hence, a new way of generating new proteins was suggested (Gilbert, 1978): introns (noncoding interspersed part of the gene) could help exons (coding interspersed part of the gene) to 'shuffle'. The word shuffle was used at that time to indicate that the process is similar to card shuffling in the sense that different sets of protein domains can combine to form different proteins. Exon shuffling occurs when exons duplicate or when they are inserted or deleted by ectopic recombination between introns, retrotransposition or unequal crossing over. In any case, intron phase, the intron position with reference to the translation open reading frame, should be compatible to maintain the frame of the protein.

An example of exon shuffling is *jingwei* (Long & Langley, 1993): the *Adh* part appears in the *Adh* gene and in the *jingwei* gene. *Sdic* could also be considered another example of exon shuffling. Two genes duplicated and in one of the copies the gene became chimerical after gene fusion (Nurminsky et al., 1998). Domains of *AnnX* and *Cdic* are present in *Sdic* (see below). The list of examples could be long but the *antifreeze glycoproteins* are another example. In this family of proteins, a 9-nucleotide (Thr-Ala-Ala) motif has been duplicated to generate a family with genes that vary in motif number from 4 to 55 copies (Chen, Devries & Cheng, 1997a). This motif has shuffled.

Genome wide analysis of exon shuffling has been centered on three topics: (1) study of the intron/exon phase distribution, (2) the correlation between protein domains and intron position, and (3) estimation of the extent of domain sharing between proteins. In all cases the aim is to determine the amount of exon shuffling.

If exon shuffling is very common and ancient, a biased intron phase distribution is expected (Gilbert, de Souza & Long, 1997). The intron/exon phase analysis reveals that there is an excess of symmetric introns and of introns with phase 0, which keeps codons intact and putative shuffling domains complete (Fedorov et al., 1998; Long & Deutsch, 1999). It has also been shown that intron position correlates with module boundaries in ancient proteins (de Souza et al., 1996; Long et al., 1998).

Domain sharing and shuffling have been investigated by Li et al. (2001) in human, fly, worm, and yeast complete genomes. Domains are functional units in a protein. Out of 1865, 1218, 1183, and 973 domains in human, fruitfly, worm, and yeast, respectively, they found 3433, 1702, 1248, and 470 distinct arrangements of two or more domains in these genomes. Many proteins showed extensive domain repetition. The largest total number of protein domains in a protein was 130. Lander et al. (2001) have also devoted time to the analysis of protein architecture from domains. They conclude that protein evolution has been centered not on domain innovation but on domain architecture (linear arrangement of domains in a protein). The human genome contains 1.8 times more protein architectures than fly or worm and 5.8 more than yeast. Shuffling creates different architectures; adding and deleting domains generating protein diversity. Morgensten and Atchley (1999) have recently studied 122 basic helix-loop-helix proteins in different species (bHLH). These proteins are classified together because they share the bHLH domain but the authors investigated the evolutionary history of the remaining parts of these proteins. Several lines of evidence support the modular history. First, the bHLH domain varies in position within these proteins. Second, there is a lack of sequence similarity in the regions flanking the bHLH domain. Third, there is a mosaic pattern for other very conserved domains: some bHLH proteins include these domains (leucine zipper domain or PAS dimerization domain) and others do not.

These genome-wide analyses reveal the high present and past importance of exon shuffling. Although exon shuffling has been studied mainly in animal species, plant species also show it (Long et al., 1996).

Horizontal transfer

Horizontal gene transfer is the transfer of genetic information from one species to another, by which

large number of 'foreign' genes can be introduced into genomes. In this way, species that are genetically separated entities can acquire alien coding DNA. Horizontal transfer can easily occur between different bacteria species by transformation, transduction and conjugation (Graur & Li, 2000). The sequencing of bacterial genomes has confirmed that lateral transfer is the major way by which bacterial genomes innovate (Ochman, Lawrence, & Groisman, 2000). Ochman, Lawrence and Groisman (2000) have recently reviewed the investigations on this phenomenon in bacteria. It was found that some bacteria have gained up to 16% of their genome by horizontal transfer. Horizontal transfer in unicellular eukaryotes can also occur via phagocytosis (de Koning et al., 2000). In multicellular eukaryotes, it has to be rare and most likely involves the action of retrovirus (Graur & Li 2000; de Koning et al., 2000). Surprisingly, a set of 223 genes has recently been reported to be horizontally transferred from bacteria to vertebrates as represented by humans (Lander et al., 2001). However, Salzberg et al. (2001) have compared these genes with all the sequenced genomes and have found that only 40 genes are exclusively shared by human and bacteria. Although these 40 genes are candidates for horizontal gene transfer, the fact the rest have been found in other lineages and are ancient genes lost in some species puts some doubt on their actual origin.

Recruitment of new regions and gene fusion

New coding sequences can also be generated by recruitment of new regions and gene fusion (Begun, 1997; Nurminsky et al., 1998; Thomson et al., 2000).

Adh-Finnegan (Begun, 1997) is an *Adh* duplicate that recruited a new 5′ untranslated region (UTR) and a new exon1 and used part of the old 5′ UTR to code for 60 new N-terminal amino acids in the *D. repleta* group. This gene does not have ADH activity. The protein is more basic than ADH and evolves faster.

Sdic is a tandem duplicated chimerical gene (Nurminsky et al., 1998). The repeating unit consists of a 5′ end from a gene playing a role in cell-adhesion (*annexin X*) and a 3′ end from a cytoplasmic dynein intermediate chain gene (*Cdic*). This chimera formed after an initial duplication of the *Cdic-AnnX* region followed by a set of deletions (Figure 7). In addition, the first exon of *Sdic* is derived from a portion of the third intron of *Cdic*. This new gene evolved testis-specific expression and is producing a protein that can be detected by GFP fusion in the tails of mature sperm. All these events happened in the 3 My old *D. melanogaster* lineage (Powell, 1997). This means that the events fixed in the population very rapidly and likely by positive selection (Nurminsky et al., 1998; but see Charlesworth & Charlesworth, 1999).

Birth and death: where do new genes go?

Whenever a new copy of a gene appears in a genome, two outcomes are possible. One is that the copy accumulates deleterious mutations and degenerates. Alternatively, the gene remains in the genome. If the gene is kept, it can gain new function or maintain previous function. Theoretically, given that most new mutations are deleterious and two functional genes are present after duplication, degeneration, and loss of the first gene to fix a deleterious mutation should happen more often than retention. Classical work was centered on the estimation of the time to silencing of one of the copies (Kimura & King, 1979, Watterson, 1983). Degeneration of one of the copies is supposed to occur within a few million generations. Walsh (1995) elaborated a population genetic model for the two fates of new gene copies. This model showed that in general, the probability that a new gene duplicate becomes a pseudogene instead of acquiring new function is high. However, the probability of a new gene acquiring a new function and staying in the population increases with population size if the ratio of advantageous to null mutations is high and selective advantage is high ($4N_e s \gg 1$).

Recently, Lynch and Conery (2000) conducted a genome-wide analysis of duplicates in the available genomes: human, mouse, *D. melanogaster, C. elegans*, *S. cerevisiae*, and *A. thaliana*, They observed the interesting phenomenon that all species under study show a high frequency of recent duplication, suggesting that some unknown proportions of older duplicates have been silenced. However, caution must be taken for their further interpretation of the results (Long & Thornton, 2001). For example in their work, age of the duplicates was estimated from synonymous changes. It is known that synonymous substitution rates vary among different genes by more than an order of magnitude with a flat distribution in the analyzed genes (Li, 1997; Zeng et al., 1998). Thus an adequate clock usable across different genes, critically required in their analysis, is not supported.

A recent review on polyploid evolution by Otto and Whitton (2000) cites revisions of the fraction of genes

retained after genome duplication. This fraction is 8% in yeast after 100 My, 72% in maize after 11 My, 77% in Xenopus after 30 My, 70% in salmonids after 25–100 My, 47% in catastomids after 50 My and 33% in vertebrates after 500 My. These numbers suggest that many duplicated copies have persisted longer in these genomes than expected. Several pertinent mechanisms have been suggested to explain these observations of long retention time of new gene duplicates (see below).

Concerted evolution

Redundant number of copies of one gene can be adaptive because it allows the organism to produce large quantity of a particular product (Ohno, 1970). If that is the case for a given gene product, after duplication we expect the copies to be kept and to evolve concertedly. Many families of genes arranged in tandem are kept and evolving in this way under unequal crossing-over and gene conversion (Graur & Li, 2000). Concerted evolution is not only achieved in closely linked duplicates. Fitzerald et al. (1996) studied *Pdha1/Pdha2* and *Pgk1/Pgk2*, and showed that gene conversion occurred between these copies although they are in different chromosomes. Duplicates can share information rarely revealing the existence of moderate concerted evolution (Nurminsky et al., 1996).

Additional effects of gene conversion

Resurrection of genes can take place by gene conversion. These resurrections can preserve copies that otherwise would be lost. One example of resurrection is the ribonuclease pseudogene in the bovine linage (Trabesinger-Ruef et al., 1996). The conversion event in this case removed a deletion and restored the amino acid sequence. Other examples are found in the amylase gene family in *D. pseudoobscura* in which gene conversion retarded pseudogene evolution (Popadic et al., 1996) or 18S rRNAs in *D. melanogaster* (Benevolenskaya et al., 1997). Recently, gene conversion has also been claimed to play a role in adaptive peak shift (Hansen, Carter & Chin, 2000). Hansen, Carter and Chin (2000) present a mathematical model in which mutations can accumulate in an unconstrained pseudogene of the family and convert a functional copy of the gene into a copy with new function. Chiu et al. (1997) showed that the changes accumulated in γ^2 globin were transferred by gene conversion to γ^1 globin. These changes increase the capture of oxygen from the mother's blood.

Population genetic factors

Purifying selection through continued expression have also been claimed to keep copies from degenerating. The production of abnormal products would be harmful and will be selected against (Hughes, 1994; Ohta, 1994). Some other families (immunoglobulins, T-cell receptors, and major histocompatibility complex genes) are maintained with efficient generation of diversity through mutation, gene conversion, balancing selection, and/or birth and death of copies (Gu & Nei, 1999; Sharon et al., 1999; Ohta, 2000). Given the existence of heterozygous advantage for a gene, pre-existing heterozygous benefit can be fixed for all individuals in a population after the duplication of the gene and fixation of a different allele in each gene (Spofford, 1969; Ohno, 1970).

Subfunctionalization

Recently, Force et al. (1999) and Lynch and Force (2000) have proposed a model in which the role of complex promoter regions is considered. If the duplication occurs in a gene with complex promoter region, complementary degenerative mutations can occur in both regulatory regions making both copies essential and preserving duplicates. The model is supported by the existence of developmental genes that show temporal partitioning of expression (Force et al., 1999). The role of subfunctionalization has been studied in mammalian developmental genes (Dermitzakis & Clark, 2001). Dermitzakis and Clark (2001) studied how different coding domains of developmental genes evolve after duplication. Positive selection was found to act in different regions of the gene in the different copies. Coding subfunctionalization was claimed to explain the results.

New functions

Novel functions arising in new genes are no doubt important to maintain many newly generated gene copies. Todd, Orengo and Thornton (2001) compared a large number of protein superfamilies for which structural information is available. They conclude that all superfamilies exhibit functional diversity generated by local sequence variation and domain shuffling, likely due to exon shuffling. If currently solved protein structures random, by sample the proteins in genomes, Todd et al.'s observation likely points to a high evolutionary rate of new gene functions, thus suggesting that evolution of new functions may be a factor

maintaining a substantial proportion of new genes. In addition, more and more particular examples are accumulating in which positive selection played an important role after duplication. There are examples of positive selection driving the gene to fixation, diversifying function, and/or changing the function of the gene (Long & Langley, 1993; Ohta, 1994; Nurminsky et al., 1998; Ting et al., 1998; Duda & Palumbi, 1999; Yi & Charlesworth, 2000). Additionally, it has been demonstrated that transition to a new function does not require many amino acid changes (Golding & Dean, 1998), simplifying the gain of a new function.

Concluding remarks

We have identified new gene acquisition as an important component in the evolution of coding regions in genomes. There are a number of mechanisms responsible for this process, as we summarized above, with ample evidence. An initial picture of the actual origin of new genes is emerging. It seems that by far the most general and important mechanism to generate new copies of genes is the tandem duplication of genes. The importance of other mechanisms has also been shown; especially retrotransposition and transposition, two mechanisms whose importance resides in allowing movement of genes between chromosomes. A genome wide study of exon shuffling also reveals that protein evolution has centered not on domain innovation but on domain architecture change. Horizontal gene transfer has been confirmed as a major way in which bacterial genomes innovate. However, what part of these acquisitions increases complexity or only redundancy remains unclear and a paradox. Recent progress in identifying and characterizing new gene functions and analyzing genomes has confirmed the tremendous importance of Darwinian selection in fixation and improvement of new gene structures. But as we discussed in the introduction there are cases of increase in size of coding regions without apparent increase in functional complexity. One explanation that does not conflict with other observations is that some of these extra genes are transient redundant copies whose fate is to be lost. The expansion of genome coding regions may be balanced by gene loss, a process that has been revealed by analysis of genome sequences in some organisms (e.g. Aravind et al., 2000). In addition, subfunctionalization would be a way of maintaining additional coding regions without increasing complexity. Much remains, both conceptually and technically, in the domain of mystery. We do not know how often new genes with novel functions originate in genomes; little solid data are available. While genome sequences provide a lot of information for computational analysis, they have not provided good indicators to measure age of a new gene, a key parameter in the study of new acquisitions. These challenges will stimulate further efforts to explore new gene evolution, the evolution of an important part of the genome.

Acknowledgements

We thank Janice Spofford and an anonymous reviewer for their suggestions to improve the manuscript.

References

Adams, M.D. et al., 2000. The genome sequence of *Drosophila melanogaster*. Science 287: 2185–2195.

Albalat, R., G. Marfany & R. González-Duarte, 1994. Analysis of nucleotide substitutions and amino acid conservation in the *Drosophila Adh* genomic region. Genetica 94: 27–36.

Aravind, L., H. Watanabe, D.J. Lipman & E.V. Koonin, 2000. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. Proc. Nat. Acad. Sci. USA 97: 11319–11324.

Ashburner, M., 1998. Speculations on the subject of alcohol dehydrogenase and its properties in *Drosophila* and other flies. BioEssays 20: 949–954.

Ashworth, A., B. Skene, S. Swift & Lovell-Badge, R., 1990. Zfa is an expressed retroposon derived from an alternative transcript of the Zfx gene. EMBO J. 9(5): 1529–1534.

Ball, C.A. & J.M. Cherry. 2001. Genome comparisons highlight similarity and diversity within the eukaryotic kingdoms. Curr. Opin. Chem. Biol. 5: 86–89.

Bancroft, I., 2001. Duplicate and diverge: the evolution of plant genome microstructure. Trends Genet. 17: 89–93.

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

Benevolenskaya, E.V., G.L. Kogan, A.V. Tulin, D. Philipp & V.A. Gvozdev, 1997. Segmented gene conversion as a mechanism of correction of 18S rRNA pseudogene located outside of rDNA cluster in *D. melanogaster*. J. Mol. Evol. 44: 646–651.

Betrán, E. & M. Ashburner, 2000. Duplication, dicistronic transcription, and subsequent evolution of the *Alcohol dehydrogenase* and *Alcohol dehydrogenase-related* genes in *Drosophila*. Mol. Biol. Evol. 17: 1344–1352.

Blanc, G., A. Barakat, R. Guyot, R. Cooke & M. Delseny, 2000. Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell 12: 1093–1101.

Bridges, C.B., 1936. The bar 'gene' a duplication. Science 83: 210–211.

Brogna, S. & M. Ashburner, 1997. The *Adh-related* gene of *Drosophila melanogaster* is expressed as a functional dicistronic messenger RNA: multigenic transcription in higher organisms. EMBO J. 16: 2023–2031.

Brosius, J., 1991. Retroposons – seeds of evolution. Science 251: 753.

Brosius, J., 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238: 115–113.

Brosius, J. & S.J. Gould, 1992. On 'genomenclature': a comprehensive (and respectful) taxonomy for pseudogenes and other 'junk DNA'. Proc. Natl. Acad. Sci. USA 89: 10706–10710.

Chambers, G.K., 1988. The *Drosophila alcohol dehydrogenase* gene-enzyme system. Adv. Genet. 25: 40–107.

Charlesworth, B. & D. Charlesworth, 1999. How was the *Sdic* gene fixed? Nature 400(6744): 519–520.

Chen, L., A.L. DeVries & C.H. Cheng, 1997a. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. Proc. Natl. Acad. Sci. USA 94: 3811–3816.

Chen, L., A.L. DeVries & C.H. Cheng, 1997b. Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. Proc. Natl. Acad. Sci. USA 94: 3817–3822.

Chiang, P.W., R. Zhang, L. Stubbs, L. Zhang, L. Zhu & D.M. Kurnit, 1998. Comparison of murine Supt4h and a nearly identical expressed, processed gene: evidence of sequence conservation through gene conversion extending into the untranslated regions. Nucl. Acids Res. 26(21): 4960–4964.

Chiu, C.H., H. Schneider, J.L. Slightom, D.L. Gumucio & M. Goodman, 1997. Dynamics of regulatory evolution in primate beta-globin gene clusters: *cis*-mediated acquisition of simian gamma fetal expression patterns. Gene 205: 47–57.

Dahl, H.H., R.M. Brown, W.M. Hutchison, C. Maragos & G.K. Brown, 1990. A testis-specific form of the human pyruvate dehydrogenase E1 alpha subunit is coded for by an intronless gene on chromosome 4. Genomics 8: 225–232.

Datta, U., I.D. Wexler, D.S. Kerr, I. Raz & M.S. Patel, 1999. Characterization of the regulatory region of the human testis-specific form of the pyruvate dehydrogenase alpha-subunit (PDHA-2) gene. Biochim. Biophys. Acta 1447: 236–243.

de Koning, A.P., F.S.L. Brinkman, S.J.M. Jonees & P.J. Keeling, 2000. Lateral gene transfer and metabolic adaptation in the human parasite *Trichomonas vaginalis*. 17(11): 1769–1773.

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.

Dermitzakis, E.T. & A.G. Clark, 2001. Differential selection after duplication in mammalian developmental genes. Mol. Biol. Evol. 18: 557–562.

Duda, T.F., Jr. & S.R. Palumbi, 1999. Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod Conus. Proc. Natl. Acad. Sci. USA 96: 6820–6823.

Dunham, I. et al., 1999. The DNA sequence of human chromosome 22 [see comments] [published erratum appears in Nature 2000 Apr 20; 404(6780): 904]. Nature 402: 489–495.

Elliott, D.J., J.P. Venables, C.S. Newton, D. Lawson, S. Boyle, I.C. Eperon, & H.J. Cooke, 2000. An evolutionarily conserved germ cell-specific hnRNP is encoded by a retrotransposed gene. Hum. Mol. Genet. 9: 2117–2124.

Esnault, C., J. Maestre & T. Heidmann, 2000. Human LINE retrotransposons generate processed pseudogenes. Nat. Genet. 24: 363–367.

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.

Fitzgerald, J., H.H. Dahl, I.B. Jakobsen & S. Easteal, 1996. Evolution of mammalian X-linked and autosomal Pgk and Pdh E1 alpha subunit genes. Mol. Biol. Evol. 13: 1023–1031.

Force, A., M. Lynch, F.B. Pickett, A. Amores, Y.L. Yan & J. Postlethwait, 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531–1545.

Friedman, R. & A.L. Hughes, 2001. Gene duplication and the structure of eukaryotic genomes. Genome Res. 11: 373–381.

Gao, M., W. Rychlik & R.E. Rhoads, 1998. Cloning and characterization of human eIF4E genes. J. Biol. Chem. 273: 4622–4628.

Gilbert, W., 1978. Why genes in pieces? Nature 271: 44.

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

Goffeau, A., B.G. Barrell, H. Bussey, R.W. Davis, B. Dujon, H. Feldmann, F. Galibert, J.D. Hoheisel, C. Jacq, M. Johnston, E.J. Louis, H.W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, & S.G. Oliver, 1996. Life with 6000 genes. Science 274: 546, 563–567.

Golding, G.B. & A.M. Dean, 1998. The structural basis of molecular adaptation. Mol. Biol. Evol. 15: 355–369.

Graur, D. & W.-H. Li, 2000. Fundamentals of Molecular Evolution. Sinauer Associates, Sunderland, MA.

Gu, X. & M. Nei, 1999. Locus specificity of polymorphic alleles and evolution by a birth-and-death process in mammalian MHC genes. Mol. Biol. Evol. 16: 147–156.

Haldane J.B.S., 1932. The Causes of Evolution. Longmans & Green, London.

Hansen, T.F., A.J. Carter & C.H. Chiu, 2000. Gene conversion may aid adaptive peak shifts. J. Theor. Biol. 207: 495–511.

Harrison, P.M., N. Echols & M.B. Gerstein, 2001. Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome. Nucl. Acids. Res. 29: 818–830.

Hart, P.E., J.N. Glantz, J.D. Orth, G.M. Poynter & J.L. Salisbury, 1999. Testis-specific murine centrin, Cetn1: genomic characterization and evidence for retroposition of a gene encoding a centrosome protein. Genomics 60: 111–120.

Hendriksen P.J., J.W. Hoogerbrugge, W.M. Baarends, P. de Boer, J.T. Vreeburg, E.A. Vos, T. van der Lende & J.A. Grootegoed, 1997. Testis-specific expression of a functional retroposon encoding glucose-6-phosphate dehydrogenase in the mouse. Genomics 41(3): 350–359.

Hooft van Huijsduijnen, R., 1998. ADAM 20 and 21; two novel human testis-specific membrane metalloproteases with similarity to fertilin-alpha. Gene 206: 273–282.

Hughes, A.L., 1994. The evolution of functionally novel proteins after gene duplication. Proc. R Soc. Lond. B Biol. Sci. 256: 119–124.

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

Jeffs, P.S., E.C. Holmes & M. Ashburner, 1994. The molecular evolution of the *Alcohol dehydrogenase* and *Alcohol dehydrogenase-related* genes in the *Drosophila melanogaster* species subgroup. Mol. Biol. Evol. 11: 287–304.

Kimura, M. & J.L. King. 1979. Fixation of a deleterious allele at one of two duplicate; loci by mutation pressure and random drift. Proc. Natl. Acad. Sci. USA 76: 2858–2861.

Kleene K.C., E. Mulligan, D. Steiger, K. Donohue & M.A. Mastrangelo, 1998. The mouse gene encoding the testis-specific isoform of Poly(A) binding protein (Pabp2) is an expressed retroposon: intimations that gene expression in spermatogenic cells facilitates the creation of new genes. J. Mol. Evol. 47(3): 275–281.

Kleene K.C. & M.A. Mastrangelo, 1999. The promoter of the Poly(A) binding protein 2 (Pabp2) retroposon is derived from the 5'-untranslated region of the Pabp1 progenitor gene. Genom. Genom. 61(2): 194–200.

Knight, C.A., C.C. Cheng & A.L. DeVries, 1991. Adsorption of alpha-helical antifreeze peptides on specific ice crystal surface planes. Biophys. J. 59: 409–418.

Lahn, B.T. & D.C. Page, 1999. Retroposition of autosomal mRNA yielded testis-specific gene family on human Y chromosome. Nat. Genet. 21: 429–433.

Lander, E.S. 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.

Li, W.H., 1997. Molecular Evolution. Sinauer Associates, Sunderland, MA.

Li, W.H., Z. Gu, H. Wang & A. Nekrutenko, 2001. Evolutionary analyses of the human genome. Nature 409: 847–849.

Linnenbach, A.J., B.A. Seng, S. Wu, S. Robbins, M. Scollon, J.J. Pyrc, T. Druck & K. Huebner. 1993. Retroposition in a family of carcinoma-associated antigen genes. Mol. Cell. Biol. 13: 1507–1515.

Logsdon, J.M., Jr. & W.F. Doolittle, 1997. Origin of antifreeze protein genes: a cool tale in molecular evolution. Proc. Natl. Acad. Sci. USA 94: 3485–3487.

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

Long, M. & Thorton, K. 2001 Evolution of gene duplication. *Science* 293: 1551a.

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.

Louh S.W. & D.C. Page, 1994. The structure of the Zfx gene on the mouse X chromosome. Genomics 19(2): 310–319.

Lynch, M. & J.S. Conery, 2000. The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155.

Lynch, M. & A. Force, 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154: 459–473.

Maestre, J., T. Tchenio, O. Dhellin & T. Heidmann, 1995. mRNA retroposition in human cells: processed pseudogene formation. EMBO J. 14: 6333–6338.

Makeyev, A.V., A.N. Chkheidze & S.A. Liebhaber, 1999. A set of highly conserved RNA-binding proteins, alphaCP-1 and alphaCP-2, implicated in mRNA stabilization, are coexpressed from an intronless gene and its intron-containing paralog. J. Biol. Chem. 274: 24849–24857.

Martignetti, J.A. & Brosius, J., 1993. BC200 RNA: a neural RNA polymerase III product encoded by a monomeric Alu element. Proc. Natl. Acad. Sci. USA 90: 11563–11567.

McCarrey, J.R., 1987. Nucleotide sequence of the promoter region of a tissue-specific human retroposon: comparison with its housekeeping progenitor. Gene 61: 291–298.

McCarrey, J.R., 1990. Molecular evolution of the human *Pgk-2* retroposon. Nucl. Acids Res. 18: 949–955.

McCarrey, J.R., M. Kumari, M.J. Aivaliotis, Z. Wang, P. Zhang, F. Marshall & J.L. Vandeberg, 1996. Analysis of the cDNA and encoded protein of the human testis-specific *Pgk-2* gene. Dev. Genet. 19: 321–332.

McDonald, J.H. & M. Kreitman, 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila.* Nature 351: 652–654.

Meyer, A. & M. Schartl, 1999. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. Curr. Opin. Cell Biol. 11: 699–704.

Mighell, A.J., N.R. Smith, P.A. Robinson & A.F. Markham, 2000. Vertebrate pseudogenes. FEBS Lett. 468: 109–114.

Morgenstern, B. & W.R. Atchley, 1999. Evolution of bHLH transcription factors: modular evolution by domain shuffling? Mol. Biol. Evol. 16: 1654–1663.

Morton, C.C., M.C. Nussenzweig, R. Sousa, G.D. Sorenson, O.S. Pettengill & T.B. Shows, 1989. Mapping and characterization of an X-linked processed gene related to MYCL1. Genomics 4: 367–375.

Muller H., 1935 The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. Genetics 17: 237–252.

Nabetani A., I. Hatada, H. Morisaki, M. Oshimura & T. Mukai, 1997. Mouse U2af1-rs1 is a neomorphic imprinted gene. Mol. Cell. Biol. 17(2): 789–778.

Nhim R.P., J.S. Lindsey & M.F. Wilkinson, 1997. A processed homeobox gene expressed in a stage-, tissue- and region-specific manner in epididymis. Gene 185(2): 271–276.

Noyce, L., J. Conaty, & A.A. Piper, 1997. Identification of a novel tissue-specific processed HPRT gene and comparison with X-linked gene transcription in the Australian marsupial *Macropus robustus*. Gene 186: 87–95.

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.

Nurminsky, D.Y., E.N. Moriyama, E.R. Lozovskaya & D.L. Hartl, 1996. Molecular phylogeny and genome evolution in *Drosophila virilis* species group: duplication of the *Alcohol dehydrogenase* gene. Mol. Biol. Evol. 13: 132–149.

Ochman, H., J.G. Lawrence & E.A. Groisman, 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405: 299–304.

Ohno, S., 1970. Evolution by Gene Duplication. Springer, Berlin.

Ohno, S., 1999. The one-to-four rule and paralogues of sex-determining genes. Cell. Mol. Life. Sci. 55: 824–830.

Ohta, T., 1994. Further examples of evolution by gene duplication revealed through DNA sequence comparisons. Genetics 138: 1331–1337.

Ohta, T., 2000. Evolution of gene families. Gene 259: 45–52.

Otto, S.P. & J. Whitton, 2000. Polyploid incidence and evolution. Annu. Rev. Genet. 34: 401–437.

Pan, Y., W.K. Decker, A.H. Huq & W.J. Craigen, 1999. Retrotransposition of glycerol kinase-related genes from the X chromosome to autosomes: functional and evolutionary aspects. Genomics 59: 282–290.

Persson K., I. Holm & O. Heby, 1995. Cloning and sequencing of an intronless mouse S-adenosylmethionine decarboxylase gene coding for a functional enzyme strongly expressed in the liver. J. Biol. Chem. 270(10): 5642–5648.

Petrov, D.A., E.R. Lozovskaya & D.L. Hartl, 1996. High intrinsic rate of DNA loss in *Drosophila*. Nature 384: 346–349.

Poindexter, K., N. Nelson, R.F. DuBose, R.A. Black & D.P. Cerretti, 1999. The identification of seven metalloproteinase-disintegrin (ADAM) genes from genomic libraries. Gene 237: 61–70.

Popadic, A., R.A. Norman, W.W. Doanet & W.W. Anderson, 1996. The evolutionary history of the amylase multigene family in *Drosophila pseudoobscura*. Mol. Biol. Evol. 13: 883–888.

Powell, J.R., 1997. Progress and Prospects in Evolutionary Biology: The Drosophila Model. Oxford University Press, New York.

Rat, L., M. Veuille &s J.A. Lepesant, 1991. *Drosophila Fat body protein P6* and *Alcohol dehydrogenase* are derived from a common ancestral protein. J. Mol. Evol. 33: 194–203.

Reinton, N., T.B. Haugen, S. Orstavik, B.S. Skalhegg, V. Hansson, T. Jahnsen & K. Tasken, 1998. The gene encoding the C gamma catalytic subunit of cAMP-dependent protein kinase is a transcribed retroposon. Genomics 49: 290–297.

Rhyner, J.A., M. Koller, I. Durussel-Gerber, J.A. Cox & E.E. Strehler, 1992. Characterization of the human calmodulin-like protein expressed in *Escherichia coli*. Biochemistry 31: 12826–12832.

Robertson, N.G., R.J. Pomponio, G.L. Mutter & C.C. Morton, 1991. Testis-specific expression of the human MYCL2 gene. Nucl. Acids Res. 19: 3129–3137.

Rogalla, P., Kazmierczak, B., Flohr A.M., Hauke, S. & Bullerdiek, J, 2000. Back to the roots of a new exon: the molecular archaeology of a *SP100* splice variant. Genomics 63: 117–122.

Rubin, G.M. et al., 2000. Comparative genomics of the eukaryotes. Science 287: 2204–2215.

Russo, C.A.M., N. Takezaki & M. Nei, 1995. Molecular phylogeny and divergence times of Drosophilid species. Mol. Biol. Evol. 13: 391–404.

Salzberg, S.L., O. White, J. Peterson & J.A. Eisen, 2001. Microbial genes in the human genome: lateral transfer or gene loss? Science 17: 17.

Sargent, C.A., C. Young, S. Marsh, M.A. Ferguson-Smith & N.A. Affara, 1994. The glycerol kinase gene family: structure of the Xp gene, and related intronless retroposons. Hum. Mol. Genet. 3: 1317–1324.

Saxena, R., L.G. Brown, T. Hawkins, R.K. Alagappan, H. Skaletsky, M.P. Reeve, R. Reijo, S. Rozen, M.B. Dinulos, C.M. Disteche, & D.C. Page, 1996. The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned. Nat. Genet. 14: 292–299.

Schaefer, S.W. & C.F. Aquadro, 1987. Nucleotide sequence of the *Adh* gene region of *Drosophila pseudoobscura*: evolutionary change and evidence for an ancient gene duplication. Genetics 117: 61–73.

Sedlacek, Z., E. Munstermann, S. Dhorne-Pollet, C. Otto, D. Bock, G. Schutz & A. Poustka. 1999. Human and mouse *XAP-5* and *XAP-5-like* (*X5L*) genes: identification of an ancient functional retroposon differentially expressed in testis. Genomics 61: 125–132.

Sharon, D., G. Glusman, Y. Pilpel, M. Khen, F. Gruetzner, T. Haaf & D. Lancet, 1999. Primate evolution of an olfactory receptor cluster: diversification by gene conversion and recent emergence of pseudogenes. Genomics 61: 24–36.

Shashidharan, P., T.M. Michaelidis, N.K. Robakis, A. Kresovali, J. Papamatheakis & A. Plaitakis, 1994. Novel human glutamate dehydrogenase expressed in neural and testicular tissues and encoded by an X-linked intronless gene. J. Biol. Chem. 269: 16971–16976.

Soares M.B., E. Schon, A. Henderson, S.K. Karathanasis, R. Cate, S. Zeitlin, J. Chirgwin & A. Efstratiadis, 1985. RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. Mol. Cell. Biol. 5(8): 2090–2103.

Soret, J., R. Gattoni, C. Guyon, A. Sureau, M. Popielarz, E. Le Rouzic, S. Dumon, F. Apiou, B. Dutrillaux, H. Voss, W. Ansorge, J. Stevenin & B. Perbal, 1998. Characterization of SRp46, a novel human SR splicing factor encoded by a PR264/SC35 retropseudogene. Mol. Cell. Biol. 18: 4924–4934.

Spofford, J., 1969. Heterosis and evolution of duplications. Amer. Natural. 103: 407–432.

Spring, J., 1997. Vertebrate evolution by interspecific hybridisation – are we polyploid? FEBS Lett 400: 2–8.

Steinemann, M. & S. Steinemann, 1998. Enigma of Y chromosome degeneration: neo-Y and neo-X chromosomes of *Drosophila miranda* a model for sex chromosome evolution. Genetica 103: 409–420.

Takeda, J., S. Seino & G.I. Bell, 1992. Human Oct3 gene family: cDNA sequences, alternative splicing, gene organization, chromosomal location, and expression at low levels in adult tissues. Nucl. Acids Res. 20: 4613–4620.

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

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

Thomson, T.M., J.J. Lozano, N. Loukili, R. Carrio, F. Serras, B. Cormand, M. Valeri, V.M. Diaz, J. Abril, M. Burset, J. Merino, A. Macaya, M. Corominas & R. Guigo, 2000. Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene. Genome Res. 10: 1743–1756.

Ting, C.T., S.C. Tsaur, M.L. Wu & C.I. Wu, 1998. A rapidly evolving homeobox at the site of a hybrid sterility gene. Science 282: 1501–1504.

Todd, A.E., C. A.S. Orengo & J.M. Thornton, 2001. Evolution of function in protein superfamilies, from a structural perspective. J. Mol. Biol. 307: 1113–1143.

Trabesinger-Ruef, N., T. Jermann, T. Zankel, B. Durrant, G. Frank & S.A. Benner, 1996. Pseudogenes in ribonuclease evolution: a source of new biomacromolecular function? FEBS Lett. 382: 319–322.

Tsytsykova A.V., E.N. Tsitsikov, D.A. Wright, B. Futcher & R.S. Geha, 1998. The mouse genome contains two expressed intronless retroposed pseudogenes for the sentrin/sumo-1/PIC1 conjugating enzyme Ubc9. Mol. Immunol. 35(16): 1057–1067.

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

Wagner, M., 1986. A consideration of the origin of processed pseudogenes (review). TIG: 134–137.

Walsh, J.B., 1995. How often do duplicated genes evolve new functions? Genetics 139: 421–428.

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.

Watterson, G.A., 1983. On the time for gene silencing at duplicate loci. Genetics 105: 745–766.

Wolfe, K.H. & D.C. Shields, 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387: 708–713.

Yi, S. & B. Charlesworth, 2000. A selective sweep associated with a recent gene transposition in *Drosophila miranda.* Genetics 156: 1753–1763.

Zaiss, D.M. & P.M. Kloetzel, 1999. A second gene encoding the mouse proteasome activator PA28beta subunit is part of a LINE1 element and is driven by a LINE1 promoter. J. Mol. Biol. 287 (5): 829–835.

Zeng L., J.M. Comeron, B. Chen & M. Kreitman, 1998. The molecular clock revisited: the rate of synonymous vs replacement change in *Drosophila*. Genetica 102/103: 369–382.