UC Irvine
UC Irvine Previously Published Works

Title
Evolution of cis-regulatory regions versus codifying regions.

Permalink
https://escholarship.org/uc/item/6qs6t4sf

Journal
The International Journal of Developmental Biology, 47(7-8)

ISSN
0214-6282

Authors
Rodríguez-Trelles, Francisco
Tarrío, Rosa
Ayala, Francisco J

Publication Date
2003

DOI
10.1387/ijdb.14756342

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Int. J. Dev. Biol. 47: 665-673 (2003)

Evolution of cis-regulatory regions versus codifying regions

FRANCISCO RODRÍGUEZ-TRELLES*,[1], ROSA TARRÍO[1] and FRANCISCO J. AYALA

Department of Ecology and Evolutionary Biology, University of California, Irvine, California, U.S.A. and [1]Unidad de Medicina Molecular INGO, Hospital Clínico Universitario, Universidad de Santiago de Compostela, Santiago, Spain

ABSTRACT Efforts to understand the genetic basis of evolutionary change have concentrated on proteins and their encoding DNA sequences. These studies have brought to light patterns and processes at the nucleotide level, yet the complex functional relationships between genetic variants and phenotypes remain poorly known. The realization that even a complete description of proteins and the effects of their activity will not suffice to understand the conditions under which they are time- and tissue-specifically expressed or repressed during development has refocused attention on cis-regulatory regions. In particular, promoter sequences are thought to hold the key for understanding the evolution of phenotypic differences between species. This is because of their complex organization into independent modules such that, unlike coding sequences in which mutations affect protein function every time the protein is expressed, mutations in cis-regulatory sequences may have minor or no pleiotropic effects. Complex information-encoding makes cis-regulatory regions poorly amenable to comparative methods designed for coding sequences. Some general conclusions are emerging as to how genetic variation is distributed across regulatory networks and the processes modulating the structure of this variation. We bring into this emerging scenario several recent findings pointing to different ways in which spliceosomal introns, pseudogenes and patterns of point mutation can be active players for the evolution of novel transcriptional profiles.

KEY WORDS: *cis-regulation, intron regulation, pseudogene, phenotypic evolution*

Introduction

The one-gene, one enzyme hypothesis and the central dogma configured a view of the genetic systems as protein-coding entities. The phenotype space was reduced almost exclusively to protein features as the organisms' structural and functional building blocks (see Mattick and Gagen, 2001). This abstraction had a profound influence on the research agenda in evolutionary biology, as epitomized by the protagonism achieved by protein-gel electrophoresis immediately after its introduction as a tool for investigating the genetics of evolutionary change (Lewontin, 1974). This influence became passively fostered with the surge of DNA sequencing methods, to a large extent because easily-discernible patterns of conservation exhibited by triplet-code-structured sequences provide a more amenable ground for positional homology inferences than untranslated sequences (see below). The vast majority of the available information on genetic variation is, indeed, from protein-coding regions. This information has revealed patterns and processes of evolution at the nucleotide level (reviewed in Kreitman and Comeron, 1999; Yang and Bielawski, 2000). But much less effort has been dedicated to the functional relationships between genetic variants and phenotypes, even at primary levels of protein conformation and function (see Golding and Dean, 1998; Patthy, 1999; Yang and Bielawski, 2000; Lewontin, 2002).

It has become increasingly clear that even a complete description of the proteins and their effects will not suffice to unravel the conditions under which encoding sequences are time- and tissue-specific expressed or repressed during development (Jacob and Monod, 1961; Zuckerkandl, 1963; Britten and Davidson, 1969; Wilson, 1975; Raff and Kaufman, 1983). The observation that chimps and humans exhibit nearly identical protein sequences, while being so different behaviorally and morphologically, made it obvious that the key to the making of a complex eukaryote resides on some kind of regulatory apparatus (King and Wilson, 1975). This inference has reappeared with renewed strength in the post-genome era, triggered by the observation that organismal complexity correlates only barely with gene number; e.g., the fruitfly, *Drosophila melanogaster*, contains less than 14,000 genes (Adams, *et al.,* 2000), whereas the considerably less complex nematode worm *Caenorhabditis elegans* contains approximately 20,000 (Waterston and Sulston, 2000; see Hahn and Wray,

***Address correspondence to:** Dr. Francisco Rodríguez-Trelles, c/o Dr. Francisco J. Ayala. Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92697-2525, USA. Fax: +34-981-951-491. e-mail: ftrelles@iiag.cesga.es

0214-6282/2003/$25.00
© UBC Press
Printed in Spain
www.ijdb.ehu.es

2002). In particular, cis-regulatory regions hold the promise of a synthesis between developmental biology and evolutionary genetics (Akam, 1998; Stern, 2000).

The Evo-Devo divide

Developmental geneticists and evolutionists have long been reluctant to shake each other hands, recently because the dramatic laboratory homeotic transformations set forth by the former to explain dramatic morphological transitions reported by paleontologists (e.g., the Cambrian explosion) are largely evolutionarily unviable. One way out of this riddle is to claim that the drastic anatomical transitions seen in the strata are artifactual, owed to imperfections of the fossil record. Experimental support for this claim is indirect, for the most part coming from molecular clock approximations that move substantially backwards fossil dates for the origin of many taxa (Wray, 2001; but see Rodríguez-Trelles *et al.*, 2002; Benton and Ayala, 2003). Some discrepancy between fossil-based and sequence-based timing estimates is expected, among other reasons because sequence differences reflect the time since two taxa last shared a common ancestor (or earlier, for gene polymorphisms present in the ancestor; Benton and Ayala, 2003), whereas fossils reflect the oldest known record of anatomical structures that define a specific group (reviewed in Wray, 2001; Benton and Ayala, 2003). But molecular clock estimates have been shown unreliable, with a statistical bias towards inflated dates (Rodríguez-Trelles *et al.*, 2002; Benton and Ayala, 2003; Glazko and Nei, 2003), and unacceptably large individual variances (Gillespie, 1991; Ayala, 1997; Li, 1997; Rodríguez-Trelles *et al.*, 2001a).

Alternatively, the picture reflected by the fossil record might be essentially correct, but novel morphologies would result from the accumulation of small, tolerable steps, instead of gross, unlikely-to-be-viable homeotic mutations. This account has long been clouded by the standing model of developmental genetic control as a cascade of key master genes (e.g., the *Hox* genes) operating on an 'all-or-nothing basis' between developmental states (i.e., the 'gene selector' model; García-Bellido, 1975; Morata and Lawrence, 1977; Lewis, 1978). As Akam (1998) has posed it, "if selector genes work as stable binary switches, their role cannot change in small steps". This model, inspired in perceptions of the quantum nature of some laboratory aberrations (the so-called 'hopeful monsters'), has now been challenged, particularly after the observation that (i) quantitative variations at the level of the products of some conserved sets of selector genes can alter morphology and behavior in subtle ways (Carroll, 1995; Gibson and Hogness, 1996; Akam, 1998; Stern, 1998, 2000; Skaer and Simpson, 2000; Sucena and Stern, 2000; Mann and Carroll, 2002); (ii) wild populations exhibit substantial heritable variation in gene expression (Rockman and Wray, 2002; Wray *et al.*, 2003); and (iii) promoter sequence variation is influenced by natural selection (e.g., Crawford *et al.*, 1999; Segal *et al.*, 1999; Daborn *et al.*, 2002; Lerman *et al.*, 2003). Consequently, attention has shifted towards the regulation of gene expression.

Promoters and the evolution of pattern

The chief control point in gene expression is translational initiation (Wray *et al.*, 2003). The initiation of transcription is regulated by the direct interaction, in a sequence-specific manner, of transcription-factor proteins with short stretches of DNA surrounding the target gene, called cis-regulatory DNA (reviewed in Davidson, 2001). Gene expression can be altered by changing either the spatial distribution or concentration of trans-acting transcription factors, or the sequence of cis-regulatory DNA (here we use the term promoter as synonymous of the entire cis-regulatory apparatus of a gene). Because a given transcription factor usually interacts with many promoters, a change in its specificity can cause changes at many loci. It seems more likely that regulatory novelties evolve by changes in promoter sequences than by changes in transcription factor specificities (Tautz, 2000**).** Promoters hold the key to understand the evolution of phenotypic differences between species, specifically because of their complex organization into independent, often redundant, and able to interact epigenetically, modules. This module organization entails an ability to generate elaborate responses (in the form of concerted changes across functionally related loci in the rate of transcriptional initiation) that have minor or not pleiotropic consequences, i.e., the disturbance of a module need not disrupt the development of the entire organism (Akam, 1998; Raff, 2000; Stern, 2000).

Typical eukaryotic promoter sequences range from a few hundred bp to more than 100 Kb. The core promoter, a region lacking significant regulatory function that provides the docking site for the assembly of the transcription complex and a position for the start of transcription (reviewed in Wray *et al.*, 2003), represents only a minor fraction (~100 bp) of this length. Dispersed across the remainder promoter sequence (i.e., the vast majority of it) there are a number of regulatory transcription-factor binding sites (comprising around 10-20% of the total promoter nucleotides) which confer the specificity of transcription. Either singly or as small but distinct clusters (typically 6 to 10 bp-long; Fairall and Schwabe, 2001; often termed enhancers, but see Wray *et al.*, 2003), these binding sites operate as remarkably independent units or modules (Travers, 1993; Latchman, 1995; Arnone and Davidson, 1997), even retaining their function when transposed to a new location in the genome (Gray and Levine, 1996; Kirchhamer *et al.*, 1996). Individual modules direct or repress transcription in specific cell types and at particular times in development, each influencing just a discrete aspect of the overall transcription profile. This modular cis-regulatory organization would, therefore, allow explaining dramatic regulatory changes between distantly related taxa as due to the accumulation of subtle, tolerable changes in the promoter regions, avoiding the difficulties ingrained in the notion of hopeful monsters (Akam, 1998; Stern, 2000). This model of gene function would explain why organismal complexity does not correlate with gene number, because complexity is a manifestation of gene expression profiles produced during development, rather than being simply due to the number of genes (Markstein and Levine, 2002; Hahn and Wray, 2002).

Features of cis-regulatory sequences

To understand the evolution of cis-regulatory DNA, it is necessary to know how genetic variation is distributed across regulatory networks, and the processes influencing the structure of this variation (Stern, 2000). In the case of protein coding regions this knowledge has chiefly been acquired by means of comparative sequence analyses. Nucleotide substitution is a slow process observed by comparing two (or multiple) sequences descended from a common ancestral sequence (see Graur and Li, 2000). Comparative approaches are a powerful tool for investigating molecular evolution because functionally constrained sequences are evolutionarily conserved (Bergman and Kreitman, 2001). The

connection between conservation and function makes possible positional homology inferences, even for deep evolutionary divergences. A second feature of coding sequences that makes them particularly amenable to comparative analysis derives from the genetic code, which provides an a priori conceptual frame for the statistical interpretation of the amount and pattern of nucleotide variation (Bergman and Kreitman, 2001). Comparative approaches have successfully exploited these features, not just to identify genes and delimit their intron-exon boundaries, but for evidencing molecular adaptation as well (e.g., inferences derived from an excess of nonsynonymous substitutions relative to synonymous substitutions; see Yang, 1998; Nielsen and Yang, 1998; Kreitman and Comeron, 1999; Yang and Bielawski, 2000; Yang and Nielsen, 2002).

On the contrary, the complex information-encoding that yields cis-regulatory sequences so appealing from a theoretical standpoint makes them difficult for comparative analysis. Specifically:
(i) regulatory sequences do not have properties directly comparable to open reading frames and codons in coding sequences, which hampers any effort seeking to define the position, amount, and strength of selective constraints in functional regulatory elements, based only on the inspection of sequence data (Bergman and Kreitman, 2001; Dermitzakis *et al.*, 2003);
(ii) models of transcriptional regulation do not simply involve activation or suppression by transcription factors, but also include competitive binding of proteins (Small *et al.*, 1991), cooperative binding (Burz *et al.*, 1998), chromatin bending (Bell *et al.*, 2001; Xin *et al.*, 2003), and other complex, non-linear, often strongly context-dependent molecular interactions (see Wray *et al.*, 2003); and
(iii) the structural and functional properties of cis-regulatory elements are not always reflected in the nucleotide sequence (see Dermitzakis *et al.*, 2003). For example, regulatory sequences can maintain regulatory function despite structural reorganization as a result of species-specific loss and gain of transcription-factor binding sites (Piano *et al.*, 1999; Ludwig *et al.*, 2000; Cuadrado *et al.*, 2001). Analogously, normal assembly of the protein transcription complex can be altered by artificially lengthening the DNA stretches spanning between binding sites, which can lead to unpredictable deregulatory effects (Bonifer, 2000). The sequence space of those stretches can, nonetheless, be constrained by natural selection to avoid specific motifs matching binding sites or other functional signals, whose fortuitous appearance at unspecific positions could disrupt transcription (Hahn *et al.*, 2003).

Because of all these features, alignment methods designed for coding sequences often perform poorly on cis-regulatory sequences. In particular, comparative analyses of distantly related cis-regulatory sequences invariably miss multiple hits, including the individual mutations altering developmental processes that were initially exposed to natural selection (Sucena and Stern, 2000). These changes can be better evidenced from comparisons across closely related sequences, but the alignments do not allow distinguishing functional from passive conservation, often leading to many false positives. Because of these intricacies, the evolution of cis-regulatory sequences is only beginning to be placed in a quantitative analytical framework (Bergman and Kreitman, 2001).

To a large extent, progress in understanding the evolution of cis-regulatory sequences is still tightly linked to methodological advances. On the bioinformatics side, these have concentrated in enhancing the sensitivity of comparative sequence strategies along two mutually nonexclusive avenues. First, by lengthening the trees in comparisons of phylogenetically closely related sequences by including as many as possible least-related taxa for a given depth of phylogenetic coverage. This approach, termed 'phylogenetic shadowing' (Bofelli *et al.*, 2003), minimizes ambiguity in the computation of the multiple alignment, while at the same time reducing the likelihood of passive conservation. It has enabled the discovery of previously undetected (by classic phylogenetic footprinting methods) primate-specific gene regulatory elements (Bofelli *et al.*, 2003). A second strategy consists of developing improved probability weight matrices for accurate 'in silico' binding site prediction (reviewed in Dermitzakis *et al.*, 2003). Setting as the null either the background sequence, or a consensus from already functionally characterized binding sites, these methods are proving useful for investigating binding site turnover (e.g., Stone and Wray, 2001; Dermitzakis *et al.*, 2003).

One obvious limitation of these approaches is that they rest on the assumption that the binding site is the fundamental unit of regulatory evolution (Dermitzakis *et al.*, 2003). However, not all binding sites are functional; in fact, they can appear by chance quite easily, because they comprise few (generally six to ten) nucleotides and there are many possible binding matrices to match. Identifying the potential binding sites that actually bind protein requires biochemical data and *in vivo* functional assays. Even so, it is difficult to rule out the possibility that a supposed non-binding site nucleotide might in fact be part of an unrecognized binding site (Wray *et al.*, 2003). An additional problem is that nucleotides not directly involved in transcription-factor binding specificity (eventually 80-90% of all nucleotides in the promoter) can influence transcription-factor binding in ways not yet well understood; for example through changes in the local conformation of DNA, or in the spacing between binding sites. Indeed, conserved sequence blocks in promoter alignments often do not coincide with functionally characterized binding sites (Bergman and Kreitman, 2001). Ultimately, non-coding conservation might be not a reflection of functional constraint, but the result of a local reduction in the rate of mutation (Clark, 2001). Local heterogeneity in mutation rates has been demonstrated for two short (~55bp), constitutively spliced *Drosophila* introns (introns 2 and B) without known cis-regulatory function. Despite being paralogs (intron B has recently been acquired by duplication of intron 2; Tarrío *et al.*, 1998) located near each other (~600bp apart) within the *xanthine dehydrogenase* (*Xdh*) gene, they exhibit disparate evolutionary rates and nucleotide base compositions, which could not be ascribed to natural selection (Rodríguez-Trelles *et al.*, 2000a). Besides these limitations it is the relatively low number of modules so far characterized (~100 in all animals combined, although rapidly increasing; Markstein and Levine, 2002), the majority of them in *Drosophila* and mammals. Despite conceptual and methodological limitations, the available information allows to outline some general features of cis-regulatory sequence evolution.

Evolution of cis-regulatory sequences

Empirically, the pattern of cis-regulatory sequence evolution has qualitatively been described by conserved blocks of DNA separated by unalignable gaps (Bergman and Kreitman, 2001). The average cis-regulatory nucleotide site evolves very fast, but there can be large substitution rate differences between promoters, and between sites of a given promoter. A rough comparison of the rates of substitution in different gene regions between mouse

and human (assuming they diverged 80 million years ago; Li, 1997) indicates that the average promoter nucleotide site (meaning by "promoter" either the 5' flanking region or the 5' untranslated region) evolves approximately as fast as the average two-fold degenerate protein coding site (i.e., ~2.2 x 10^{-10} substitutions per year), faster than nondegenerate sites (~0.8 x 10^{-10} substitutions per year) but distinctly more slowly that four-fold degenerate sites and introns (~3.6 x 10^{-10}). However, in a recent comparison between *Drosophila melanogaster* and *Drosophila virilis* (assuming they diverged 40 million years ago, more than enough time to discern functional constraint in non-coding sequences), intergenic regions and introns, known a priori to contain cis-regulatory activity, exhibit essentially identical fractions of conserved DNA; and the rate and pattern of point substitutions and indels within conserved DNA blocks is the same across the two types of regions. This result seems to indicate that the evolutionary dynamics of the *average* cis-regulatory nucleotide site is the same whether the site represent an intergenic region or an intron, at least in *Drosophila* (Bergman and Kreitman, 2001).

There can be large substitution rate differences between promoters. For example, comparative analysis of the *Dlx5/Dlx6* intergenic region across zebrafish and mammals unveiled a highly conserved segment (i.e., 80% identity over a 660 bp-long alignment), shown by functional assay to be the site of cross-regulatory interactions between *Dlx* genes in the embryonic forebrain (Zerucha *et al.*, 2000; Müller *et al.*, 2002). This means that cis-regulatory sequences can be conserved for time periods as long as 5 x 10^{-8} years. Examples of cis-regulatory sequence conservation across zebrafish and mammals include two intron regions containing the *Ar-A* and *Ar-C* transcription-factor binding modules, which drive expression of the *sonic hedgehog* (*shh*) gene in the ventral neural tube and notochord of the developing embryo. In contrast, neither the brain-specific upstream *Ar-D* nor the floor-plate *Ar-B* binding modules retain any trace of homology detectable by comparative analysis. Hence, even neighboring cis-regulatory elements involved in the same regulatory network can evolve at disparate rates. Interestingly, however, all four binding modules have retained their regulatory function (Müller *et al.*, 2002), illustrating the principle that there is not a linear relationship between functional differentiation and amount of cis-regulatory sequence evolution. Sequence divergence with retention of regulatory function has also been shown for the enhancer-driving *even skipped* (*eve*) expression in stripe number 2 in closely related species of *Drosophila* (Ludwig *et al.*, 2000). The model inferred from these seemingly paradoxical observations calls for stabilizing selection acting on gene expression, while allowing for variation in the composition of cis-regulatory sequences (Ludwig *et al.*, 2000; Bergman and Kreitman, 2001).

Rate variation among sites in cis-regulatory regions is largely accounted for by substitution rate differences between transcription-factor binding sites and the nonbinding sites located between them, although conserved cis-regulatory blocks do not always coincide with functionally identified binding sites (Bergman and Kreitman, 2001). Variation in substitution rates among sites is most often accommodated in models of protein coding sequence evolution using the discrete gamma approximation (Yang, 1996). The extent of among-site rate variation is inversely proportional to the value of the shape parameter (α) of the gamma distribution. Lowest values of α are obtained for genes combining a large fraction of invariable sites with a few rapidly evolving sites. Therefore, in absence of positive selection, the value of α can be interpreted as a measure of the functional constraint of a gene (Zhang and Gu, 1998). Small values of α (<0.5) indicate strong constraint, and large values (>1) indicate weak constraint. Hence, it is expected that cis-regulatory regions, which typically exhibit low density of functionally important sites, will yield larger α values than protein coding regions. Analogously, it is expected that α will vary among different types of promoters depending on their number of transcription factor binding sites and organizational complexity. Thus, α provides a summary statistic for the overall degree of constraint of a promoter, useful for comparative purposes.

Attaining accurate estimates of α for cis-regulatory regions is problematic, because promoters can reliably be aligned only when they are low diverged. But in such cases there has not been enough time for substitutions to occur at potentially variable sites, which are then regarded as invariant sites, and α is underestimated (see Zhang and Gu, 1998). A further complication might be that promoter among-site rate variation appears to be extremely nonstationary, i.e., single transcription-factor binding sites can appear and disappear among relatively closely related species (e.g., Wu and Brennan, 1993; Takahasi *et al.*, 1999; Liu *et al.*, 2000) and even within populations (e.g., Stone and Wray, 2001; Rockman and Wray, 2002). A given promoter nucleotide site can switch function (i.e., from binding site to other roles or non functionality) quite frequently in evolution. This feature of promoters raises additional difficulties, for standard tests of molecular evolution are based on straightforward, e.g., synonymous-nonsynonymous, categorizations (Wray *et al.*, 2003). In general, cis-regulatory sequences have limited utility for phylogenetic reconstruction, being most useful to resolve branching patterns below the species group level.

Qualitatively, cis-regulatory sequences undergo basically the same types of mutations as protein coding sequences. These range from small changes, such as point mutations and small indels, through variation in short repeat structure, to large restructurations, such as the acquisition of new regulatory sequences by transposition, evolution of new promoters in association with promoter-less genes created by retroposition, promoter fragmentation or recombination associated to gene duplication. Available information on these and other changes, their specificities compared to coding regions (e.g., in absence of the requirements imposed by the genetic code, indels are not constrained to be multiples of three in cis-regulatory regions), and their potential effects on gene expression have been extensively discussed by Wray *et al.* (2003).

Recently, it has become apparent that a significant source of cis-regulatory sequences emerges from duplicated pseudogenes. The duplicates may have, first, become pseudogenes as a consequence of premature stop codons or other disabling mutations. But it has been shown in a number of cases that some "pseudogenes" have acquired regulatory functions, typically relative to the original genes source of the duplication (Balakirev and Ayala, 2003). Korneev *et al.* (1999) have shown that a *nitric oxide synthase* pseudogene (*pseudo-NOS*) and its paralogous functional gene (*nNOS*) are co-expressed in identifiable neurons of the mollusk *Lymnaea stagnalis*. The *pseudo-NOS* transcript includes a region with significant antisense homology to the *nNOS* mRNA. The antisense region of the *pseudo-NOS* RNA specifically suppresses

the synthesis of the *nNOS* protein. Thus the *pseudo-NOS* transcript acts as an antisense regulator of *nNOS* protein synthesis. Healy *et al.* (1996) have shown that 3' sequences that lie within the ψ*Est-6* pseudogene transcription unit of *D. melanogaster* contain elements that modulate the expression of *Est-6*, which obviously implies some regulatory function for ψ*Est-6*.

Troyanovsky and Leube (1994) have described an interesting example of gene/pseudogene cooperation in human *cytokeratin 17* expression. A detailed examination of *cytokeratin* transcription regulation using gene/pseudogene chimeric constructs has identified specific promoter/enhancer elements that are inactive by themselves but can interact to induce strong transcriptional activity of reporter genes. The process includes the interaction between the proximal region of the inactive *cytokeratin* pseudogene promoter and the distal upstream region of the actively transcribed *cytokeratin* gene. Troyanovsky and Leube (1994) conclude that cis elements in the proximal 5'-upstream region of the pseudogene promoter can cooperate with distal enhancer elements of the functional gene to induce strong transcriptional activity in transfected HeLa cells. In mice, the *Makorin1-p1* pseudogene regulates the messenger-RNA stability of its homologous coding gene (*Makorin1*) by competitive interaction, either at the RNA or DNA level (Hirotsune *et al.*, 2003). Similar mechanisms had been previously suggested for other instances of regulatory gene-pseudogene interaction (Livak, 1990; Kalmykova *et al.*, 1998; see also Balakirev and Ayala, 2003).

The role of introns

It is almost certain that modern nuclear introns are not ancient remnants of the prebiotic assembly of genes, but evolutionary descendants of type II self-splicing introns which would have populated the eukaryotic lineage late in evolution (Logsdon, 1998; Lynch, 2002; Lynch and Richardson, 2002; Tarrío *et al.*, 2003). Once released from the constraints of self-splicing, spliceosomal introns became free to evolve and explore new evolutionary space. Spliceosomal introns are increasingly viewed as genomic parasites that have been co-opted into many essential functions such that few, if any, eukaryotes could survive without them (Lynch and Richardson, 2002; Le Hir *et al.*, 2003). Many intron sequences are known to contain transcription- factor binding sites [e.g., the *Sog* intron in *Drosophila* (Francois *et al.*, 1994); immunoglobulin μ and κ intronic enhancers (Sleckman, 1996); CCR5 in humans (Bamshad *et al.*, 2002); *Ar-A* and *Ar-C* in the vertebrate *shh* gene (Müller *et al.*, 2002); *Otx* in the sea urchin *Strongylocentrotus purpuratus* (Yuh *et al.*, 2002); see also Bergman and Kreitman, 2001]. In addition, introns can regulate transcription by controlling DNA accessibility through modulation of nucleosome position (Liu *et al.*, 1995; Le Hir *et al.*, 2003), and the splicing signals of transcribed introns can enhance the activity of RNA polymerase II (see Le Hir *et al.*, 2003). Apparently, some introns have become an integral part of the cis-regulatory apparatus. As yet not sufficiently acknowledged, the possibility of intron proliferation by reverse splicing and retropositioning endows intervening sequences with the potential to be significant players in the evolution of cis-regulatory networks.

The notion that introns can spread by duplication has been theoretically entertained, first, because intronic RNAs are produced in large numbers, which is not surprising since they are processed in parallel with gene expression (see Mattick and Gagen, 2001); second, some translation occurs within the nucleus, which can make reverse transcriptase available within the nuclear domain (Iborra *et al.*, 2001); and, third, after excision, intron lariats remain attached to the splicing machinery long enough for the intron to be reverse-transcribed and reverse-spliced into an ectopic site (Clement *et al.*, 2001; Lynch and Richardson, 2002). (Multiple output in parallel with gene expression, large numbers, and the potential for specifically targeted interactions as a function of their sequence complexity has been advanced to highlight intronic RNAs as 'excellent' candidates for trans-acting factors enabling dynamical gene-gene communication, genetic multitasking, and system integration; Mattick and Gagen, 2001.)

Intron insertion is now widely acknowledged as a frequent phenomenon (Logsdon, 1998; Lynch, 2001; Tarrío *et al.*, 2003). So far, however, only a single study has provided evidence for the origin of a nuclear intron, by intragenomic duplication of a preexisting intron (Tarrío *et al.*, 1998). The study detected three newly inserted introns in the *Xdh* gene of *Drosophila* (introns A and B), and the medfly *Ceratitis capitata* (intron C). On the basis of significant sequence similarity, all three introns are likely transposed copies of a preexisting *Xdh* intron (intron 2), which is pervasive in *Drosophila* and other dipterans (and has a homologous position as an intron found in humans and other diverse organisms). Even though none of these introns was found to contain cis-regulatory elements, the finding raises the possibility that introns carrying cis-regulatory modules could transpose to other genomic locations where they could exert their activity (as noted above, introns devoid of regulatory motifs can affect transcription in other ways). Interestingly, the site of intron A, the most circumscribed phylogenetically (thus, the most recent) of the three introns, has been occupied independently by another intron in plants (Tarrío *et al.*, 2003), a quite unlikely coincidence considering that the *Xdh* locus comprises ~4000 coding sites for intron insertion. This observation strongly suggests both that there are constraints on the spatial distribution of sites potentially settled by spliceosomal introns in protein coding sequences; and that there exists a certain pressure for intron insertion. This pressure should be better tolerated in noncoding regions, owed to their overall greater sequence malleability. Introns carrying transcription-factor binding sites can thus contribute new regulatory modules to other cis-regulatory regions. In those cases where they lack regulatory function, intron insertion could alter the spacing between preexisting binding sites or provide raw material for the evolution of new regulatory signals. Consequently, introns should be encountered in promoter regions, especially in those that are transcribed and, thus, remain recognizable (see Mattick and Gagen, 2001). Also, note that these introns should be recent (i.e., not ancient remains of the prebiotic assembly of genes) in all accounts.

Macroevolutionary potential of fluctuating mutation bias

For any given genetic character, the incidence of alternative character states is often not uniform. For instance, some nucleotide sites vary more rapidly than others, indels are more frequently observed in introns and intergenic regions than in coding regions, some genomes and genome regions contain preferably G and C nucleotides whereas others exhibit extremely low GC content, and so on. Two disparate views that seek to account for these asymmetries are:

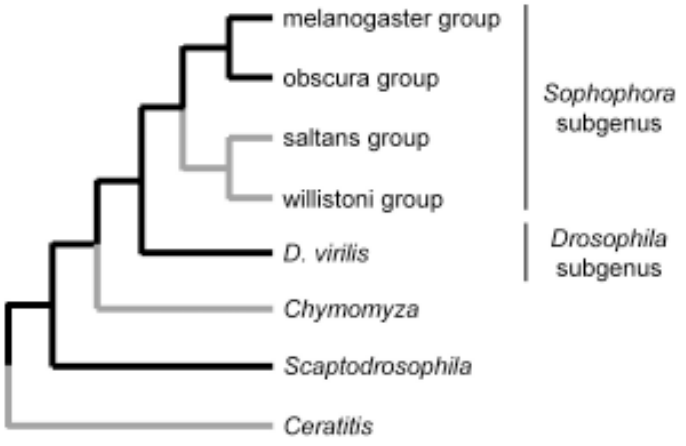

Fig. 1. Cladogram of the phylogenetic relationships among the species discussed in the text (according to TARRÍO *et al.*, 2001). *Black and grey branches represent high and low-GC lineages, respectively.*

(i) The asymmetries reflect 'internal' biases in the spontaneous production of variation, or mutational biases, which may arise from specific aspects of the machinery of DNA replication and repair. The potential of mutational biases to imprint direction to evolution is obvious in strictly neutral models, in which the rate of substitution is a function solely of the neutral rate of interconversion between allelic states. But it has also been demonstrated in models of selection under very general conditions (Shields, 1990; Yampolsky and Stoltzfus, 2001).

(ii) Alternatively, the asymmetries arise subsequently to the origin of variation, as a result of 'external' factors, specifically the culling effects of Darwinian natural selection on finite genetic pools containing uniform representations of all relevant genetic variation.

These alternative views have emerged, for example, in attempts to settle the causes of regional GC content differences within the vertebrates' genome (the so called "isochors") (see Sueoka, 1962; Bernardi *et al.*, 1985; Gillespie, 1991; Eyre-Walker, 1999; Montoya-Burgos *et al.*, 2003).

In the case of *Drosophila*, it has long been thought that the pattern of point mutation has been a negligible source of heritable variation during the diversification of the genus (Shields *et al.*, 1988; Moriyama and Hartl, 1993; Kliman and Hey, 1994; Petrov and Hartl, 1999), initiated around 60 My ago (Fitch and Ayala, 1994; Powell and DeSalle, 1995). This notion has recently been challenged after the observation that the *Drosophila saltans* and *D. willistoni* species groups exhibit patterns of GC content markedly different from those previously known in *Drosophila* (Rodríguez-Trelles *et al.*, 1999; 2000b; 2000c). Specifically, the GC content in synonymous codon sites and, to a lesser extent, in nonsynonymous sites and non-coding regions is higher in the *D. melanogaster* and *D. obscura* groups, and in *D. virilis,* which belongs to a different subgenus (see Fig. 1), than in the *D. saltans* and *D. willistoni* groups. In addition, the *saltans* and *willistoni* groups exhibit an increased rate of amino acid substitution in less functionally constrained regions, with the new replacements occurring preferably by amino acids encoded by low-GC content codons. These findings have subsequently been corroborated by more extensive surveys of the *D. willistoni* genome (Bergman and Kreitman, 2001; Bergman *et al.*, 2002; see also Begun and Whitley, 2002).

Large GC content differences among lineages are responsible for the tendency of molecular phylogenies inferred with conventional homogenous stationary models (review in Powell 1997) to place the *D. saltans* and *D. willistoni* groups outside their genus (Tarrío *et al.*, 2001), strongly advising the use of more realistic, nonhomogeneous nonstationary representations for modelling the molecular evolution of *Drosophila*. (Rodríguez-Trelles *et al.*, 2000c; Tarrío *et al.,* 2001; Bergman and Kreitman, 2001). The reconstruction of ancestral GC content by these methods has shown that most of the change in nucleotide base composition occurred along the branch ancestral to the fast-evolving *saltans-willistoni* lineages (Rodríguez-Trelles *et al.*, 2000c). These observations are best explained by a shift in the pattern of point mutation that occurred in the ancestor of the *saltans-willistoni* offshoot, after it split from the lineage that gave rise to the *melanogaster* and *obscura* groups, possibly reinforced by synergistic effects of reduced population numbers (Rodríguez-Trelles *et al.*, 1999). Similar compositional changes observed in other fruit fly species (e.g., *Scaptodrosophila* has an extremely high GC content compared to *Ceratitis*; and *Chymomyza* has a very low GC content compared to *Drosophila*; see Fig. 1) suggest that changeability of the pattern of point mutation might be a distinctive characteristic not only of a limited group of species within the subgenus *Sophophora*, but a general feature of the drosophilid genome (Rodríguez-Trelles *et al.*, 2000b).

Nucleotide-base compositional changes have dramatically impacted protein amino acid composition in dipterans. For example, XDH contains ~3-5% more amino acids encoded by high-GC codons in *obscura*-group species than in species of the *saltans* group, *Chymomyza* or *Ceratitis* (Rodríguez-Trelles *et al.*, 2001b). Despite many amino acid replacements, XDH seems to function equally in different species; i.e., XDH function can be achieved by a large array of amino acid compositions (Rodríguez-Trelles *et al.*, 2001b). Homologous proteins can conserve their tri-dimensional structure even after having lost all trace of homology in their primary sequences (e.g., Patthy, 1999; Torrents *et al.*, 2002). In this respect, protein function resembles cis-regulatory function, except that the later exhibits notably greater flexibility at the sequence level. Cis-regulatory function is substantially less constrained than protein function, presumably because of its physical organization into independent modules so that mutations in promoter sequences may have minor or no pleiotropic effects, unlike coding regions in which mutations affect protein function every time the protein is expressed (Stern, 2002). For instance, new binding sites are continuously appearing by mutation in many places in a genome, an important way in which new transcription patterns can evolve (Stone and Wray, 2001). Yet non-functional binding of transcription factors to suitable motifs, but in inappropriate genome locations, may introduce noise into the efficient functioning of the cell.

Spurious binding sites are underrepresented in bacterial genomes (Hahn *et al.*, 2003). However, the strength of natural selection against spurious binding sites is weak ($N_e s \sim 0.09$, averaged across Eubacteria and Archaea), similar to that of codon bias (Hahn *et al.*, 2003). Hence, if the shift in the pattern of point mutation has been powerful enough as to switch the pattern of codon usage in *D. saltans* and *D. willistoni*, even in the highly

expressed, putatively more constrained genes *Adh* and *Sod* (Rodríguez-Trelles *et al.*, 1999), it should also be expected to have overcome selection against spurious transcription-factor binding sites. Since the new mutation bias has increased AT content in these species, their genomes should, therefore, be enriched in low-GC binding sites (and impoverished in GC-rich binding sites) compared to *D. melanogaster* and *D. pseudoobscura*. If there is a connection between the GC content of a binding site and its function (e.g., GC content might vary between different types of enhancers; or between enhancers, silencers, insulators, and other classes of regulatory elements), fluctuating mutation bias might have ultimately been a major trigger of the phenotypic differences between these groups of fruit flies. Because the species of the *saltans* group are more diverged compositionally than *D. willistoni* from *D. melanogaster* and *D. pseudoobscura* (Rodríguez-Trelles *et al.*, 1999; 2000c), they are most appropriate for investigating this issue. The *saltans* species might be more useful than *D. willistoni* also for dissecting regions of the *Drosophila* genome under different levels of functional constraint (see Bergman *et al.*, 2002). Therefore, we propose that (in addition to the already selected *D. melanogaster* and *D. pseudoobscura*) a *saltans*-group species would be a most suitable candidate as the third *Drosophila* species for complete genome sequencing.

Acknowledgements

F. R-T and R. T. have received support from contracts Ramón y Cajal and Doctor I3P, respectively, from the Ministerio de Ciencia y Tecnología (Spain). Research supported by NIH Grant GM42397 to F.J.A.

References

ADAMS, M.D., CELNIKER, S.E., HOLT, R.A., EVANS, C.A., GOCAYNE, J.D., AMANATIDES, P.G., SCHERER, S.E., LI, P.W., HOSKINS, R.A., GALLE, R.F., *et al.* (2000). The genome sequence of *Drosophila melanogaster*. *Science* 287: 2185-2195.

AKAM, M. (1998). *Hox* genes, homeosis and the evolution of segment identity: no need for hopeless monsters. *Int. J. Dev. Biol.* 42: 445-451.

ARNONE, M.I. and DAVIDSON, E.H. (1997). The hardwiring of development: Organization and function of genomic regulatory systems. *Development* 124: 1851-1864.

AYALA, F.J. (1997). Vagaries of the molecular clock. *Proc. Natl. Acad. Sci. USA*. 94: 7776-7783.

BALAKIREV, E. S. and AYALA, F. J. (2003). Pseudogenes: Are They "Junk" or Functional DNA? *Annu. Rev. Genet.* 37: 123-151.

BAMSHAD, M.J., MUMMIDI, S., GONZÁLEZ, E., AHUJA, S.S., DUNN, D.M., WATKINS, W.S., WOODING, S., STONE, A.C., JORDE, L.B., WEISS, R.B., *et al.* (and 1 co-author). (2002). A strong signature of balancing selection in the 5' cis-regulatory region of *CCR5*. *Proc. Natl. Acad. Sci. USA.* 99: 10539-10544.

BEGUN, D.J. and WHITLEY, P. (2002). Molecular population genetics of *Xdh* and the evolution of base composition in *Drosophila*. *Genetics* 162: 1725-1735.

BELL, A.C., WEST, A.G. and FELSENFELD, G. (2001). Insulators and boundaries: versatile regulatory elements in the eukaryotic genome. *Science* 291: 447-450.

BENTON, M. J. and AYALA, F. J. (2003). Dating the Tree of Life. *Science* 300: 1698-1700.

BERGMAN. C.M. and KREITMAN, M. (2001). Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res.* 11: 1335-1345.

BERGMAN, C.M., PFEIFFER, B.D., RINCÓN-LIMAS, D.E., HOSKINS, R.A., GNIRKE, A., MUNGALL, C.J., WANG, A.M., KRONMILLER, B., PACLEB, J., PARK, S., *et al.* (2002). Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biol.* 3: research0086.1-0086.20.

BERNARDI, G., OLOFFSSON, B., FILIPSKI, J. *et al.* (1985). The mosaic genome of warm-blooded vertebrates. *Science* 228: 953-958.

BOFFELLI, D., MCAULIFFE, J., OVCHARENKO, D., LEWIS, K.D., OVCHARENKO, I., PACHTER, L. and RUBIN, E.M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* 299: 1391-1394.

BONIFER, C. (2000). Developmental regulation of eukaryotic gene loci: Which cis-regulatory information is required? *Trends Genet.* 16: 310-315.

BRITTEN, R.J. and DAVIDSON, E.H. (1969). Gene regulation for higher cells: A theory. *Science*. 165: 349-357.

BURZ, D.S., RIVERA-POMAR, R., JACKLE, H. and HANES, S.D. (1998). Cooperative DNA-binding by Bicoid provides a mechanism for threshold-dependent gene activation in the *Drosophila* embryo. *EMBO J.* 17: 5998-6009.

CARROLL, S.B. (1995). Homeotic genes and the evolution of arthropods and chordates. *Nature* 376: 479-485.

CLARK, A.G. (2001). The search for meaning in non-coding DNA. *Genome Res.* 11: 1319-1320.

CLEMENT, J., MAITI, S. and WILKINSON, M.F. (2001). Localization and stability of introns spliced from the *Pem* homeobox gene. *J. Biol. Chem.* 20: 16919-16930.

CRAWFORD, D.L., SEGAL, J.A. and BARNETT, J.L. (1999). Evolutionary analysis of TATAless proximal promoter function. *Mol. Biol. Evol.* 16: 194-207.

CUADRADO, M., SACRISTÁN, M. and ANTEQUERA, F. (2001). Species specific organization of CpG island promoters at mammalian homologus genes. *EMBO Rep.* 2: 586-592.

DABORN, P.J., YEN, J.L., BOGWITZ, M.R., GOFF, G.L., FEIL, E., JEFFERS, S. TIJET, N., PERRY, T., HECKEL, D., BATTERHAM, P. *et al.* (2002). A single P450 allele associated with insecticide resistance in *Drosophila*. *Science* 297: 2253-2225.

DAVIDSON, E.H. (2001). *Genomic Regulatory Systems: Development and Evolution*. Academic Press, San Diego.

DERMITZAKIS, E.T., BERGMAN, C.M. and CLARK, A.G. (2003). Tracing the evolutionary history of Drosophila regulatory regions with models that identify transcription factor binding sites. *Mol. Biol. Evol.* 20: 703-714.

EYRE-WALKER, A. (1999). Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics* 152: 675-683.

FAIRALL, L. and SCHWABE, J.W.R. (2001). DNA binding by transcription factors. In *Transcription Factors*. (ed. J. LOCKER). Academic Press, Inc. San Diego. Pp. 65-84.

FITCH, W.M. and AYALA, F.J. (1994). The superoxide dismutase molecular clock revisited. *Proc. Natl. Acad. Sci. USA*. 91: 6802-6807.

FRANCOIS, V., SOLLOWAY, M., O'NEILL, J.W., EMERY, J. and BIER, E. (1994). Dorsal-ventral patterning of the *Drosophila* embryo depends on a putative negative growth factor encoded by the short gastrulation gene. *Genes Dev.* 8: 2602-2616.

GARCÍA-BELLIDO, A. (1975). Genetic control of wing disc development in *Drosophila*. In *Cell Patterning. Ciba Found. Symp.* 29: 161-178.

GIBSON, G. and HOGNESS, D.S. (1996). Effect of polymorphism in the *Drosophila* regulatory gene *Ultrabithorax* on homeotic stability. *Science* 271: 200-203.

GILLESPIE, J.H. (1991). *The Causes of Molecular Evolution*. Oxford University Press. New York.

GLAZKO, G.V. and NEI. M. (2003). Estimation of divergence times for major lineages of primate species. *Mol. Biol. Evol.* 20: 424-434.

GOLDING, G.B. and DEAN, A.M. (1998). The structural basis of molecular adaptation. *Mol. Biol. Evol.* 15: 355-369.

GONZÁLEZ, P., RAO, P.V., NUÑEZ, S.B. and ZIGLER, J.S. Jr. (1995). Evidence for independent recruitment of zeta-crystallin/quinone reductase (CRYZ) as a crystalline in camelids and hystricomorph rodents. *Mol Biol Evol* 12: 773-781.

GRAY, S. and LEVINE, M. (1996). Transcriptional repression in development. *Curr. Opin. Cell. Biol.* 8: 358-364.

GRAUR, D. and LI, W-H. (2000). *Fundamentals of molecular evolution*. Sinauer Associates Inc. Sunderland, Massachusetts.

HAHN, M.W. and WRAY, G.A. (2002). The g-value paradox. *Evolution and Development* 4: 73-75.

HAHN, M.W., STAJICH, J.E. and WRAY, G.A. (2003). The effects of selection against spurious transcription factor binding sites. *Mol. Biol. Evol.* 20: 901-906.

HEALY, M.J., DUMANCIC, M.M., CAO, A., and OAKESHOTT, J.G. (1996). Localization of sequences regulating ancestral and acquired sites of esterase 6 activity in *Drosophila melanogaster*. *Mol. Biol. Evol.* 13: 784-797.

HIROTSUNE, S., YOSHIDA, N., CHEN, A., GARRETT, L., SUGIYAMA, F., TAKAHASHI, S., YAGAMI, K., WYNSHAW-BORIS, A. and YOSHIKI, A. (2003). An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* 423: 91-96.

IBORRA, F.J., JACKSON, D.A. and COOK, P.R. (2001). Coupled transcription and translation within nuclei of mammalian cells. *Science* 293: 1139-1142.

JACOB, F. and MONOD, J. (1961). On the regulation of gene activity. *Cold Spring Harbor Symp. Quant. Biol.* 26: 193-211.

KALMYKOVA, A.I., DOBRITSA, A.A. and GVOZDEV, V.A. (1998). *Su*(*Ste*) diverged tandem repeats in a *Y* chromosome of *Drosophila melanogaster* are transcribed and variously processed. *Genetics* 148: 243-249.

KING, M.C. and WILSON, A.C. (1975). Evolution at two levels in humans and chimpanzees. *Science* 188: 107-116.

KIRCHHAMER, C.V., BOGARAD, L.D. and DAVIDSON, E.H. (1996). Developmental expression of synthetic cis-regulatory systems composed of spatial control elements from two different genes. *Proc. Natl. Acad. Sci. USA* 93: 13849-13854.

KIMURA, M. (1983). *The neutral theory of molecular evolution*. Cambrigde University Press. Cambrigde, U.K.

KLIMAN, R.M. and HEY, J. (1994). The effects of mutation and natural selection on codon bias in the genes of *Drosophila*. *Genetics* 137: 1049-1056.

KORNEEV, S.A., PARK, J.-H. and O'SHEA, M. (1999). Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *J. Neurosci.* 19: 7711-7720.

KREITMAN, M. and COMERON, J.M. (1999). Coding sequence evolution. *Curr. Opin. Genet. Dev.* 9: 637-641.

LATCHMAN, D.S. (1998). Eukaryotic *transcription factors*. Academic Press, San Diego.

LE HIR, H., NOTT, A. and MOORE, M. J. (2003). How introns influence and enhance eukaryotic gene expression. *TIBS* 28: 215-220.

LERMAN, D.N., MICHALAK, P., HELIN, A.B., BETTENCOURT, B.R. and FEDER, M.E. (2003). Modification of heat-shock gene expression in *Drosophila melanogaster* populations via transposable elements. *Mol. Biol. Evol.* 20: 135-144.

LEWONTIN, R.C. (1974). *The Genetics Basis of Evolutionary Change*. Columbia University Press. New York.

LEWONTIN, R.C. (2002). Directions in evolutionary biology. *Annu. Rev. Genet.* 36: 1-18.

LEWIS, E.B. (1978). A gene complex controlling segmentation in Drosophila. *Nature* 276: 565-570.

LI, W-H. (1997). *Molecular Evolution*. Sinauer Associates, Inc. Sunderland, Massachusetts.

LIU, K., SANDGREN, E.P., PALMITER, R.D. and STEIN, A. (1995). Rat growth hormone gene introns stimulate nucleosome alignment in vitro and in transgenic mice. *Proc. Natl. Acad. Sci. USA* 92: 7724-7728.

LIU, T., WU, J. and HE, F. (2000). Evolution of *cis*-acting elements in 5' flanking regions of vertebrate acting genes. *J. Mol. Evol.* 50: 22-30.

LIVAK, K.J. (1990). Detailed structure of the *Drosophila melanogaster Stellate* genes and their transcripts. *Genetics* 124: 303-316.

LOGSDON, J.M. Jr. (1998). The recent origins of spliceosomal introns revisited. *Curr. Opin. Genet. Dev.* 8: 637-648.

LUDWIG, M.Z., BERGMAN, C.M., PATEL, N.H. and KREITMAN, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* 403: 564-567.

LYNCH, M. (2002). Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci. USA.* 99: 6118-6123.

LYNCH, M. and RICHARDSON, A.O. (2002). The evolution of spliceosomal introns. *Curr. Opin. Genet. Dev.* 12: 701-710.

MANN, R.S. and CARROLL, S.B. (2002). Molecular mechanisms of selector gene function and evolution. *Curr. Opin. Genet. Dev.* 12: 592-600.

MARKSTEIN, M. and LEVINE, M. (2002). Decoding cis-regulatory DNAs in the *Drosophila* genome. *Curr. Opin. Genet. Dev.* 12: 601-606.

MATTICK, J.S. and GAGEN, M.J. (2001). The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Mol. Biol. Evol.* 18: 1611-1630.

MONTOYA-BURGOS, J.I., BOURSOT, P. and GALTIER, N. (2003). Recombination explains isochores in mammalian genomes. *Trends Genet.* 19: 128-130.

MORATA, G. and LAWRENCE, P.A. (1977). Homeotic genes, compartments and cell determination in *Drosophila*. *Nature* 265: 211-216.

MORIYAMA, E.N. and HARTL, D.L. (1993). Codon usage bias and base composition of nuclear genes in *Drosophila*. *Genetics* 134: 847-858.

MÜLLER, F., BLADRE, P. and STRÄHLE, U. (2002). Search for enhancers: teleost models in comparative genomic and transgenic analysis of *cis* regulatory elements. *BioEssays* 24: 564-572.

NIELSEN, R. and YANG, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148: 1271-1276.

PATTHY, L. (1999). *Protein Evolution*. Blackwell Science. Oxford.

PETROV, D.A. and HARTL, D.L. (1999). Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. *Proc. Natl. Acad. Sci. USA*. 96: 1475-1479.

PIANO, F., PARISI, M.J., KARESS, R. and KAMBYSELLIS, M.P. (1999). Evidence for redundancy but not trans factor-cis element coevolution in the regulation of *Drosophila* Yp genes. *Genetics* 152: 605-616.

POWELL, J.R. (1997). *Progress and prospects in evolutionary biology: the Drosophila model*. Oxford University Press. New York.

POWELL, J.R. and DESALLE, R. (1995). Drosophila molecular phylogenies and their uses. *Evol. Biol.* 28: 87-138.

RAFF, R.A. (2000). Evo-devo: the evolution of a new discipline. *Nature Rev.* 1: 74-79.

RAFF, R.A. and KAUFMAN, T.C. (1983). *Embryos, Genes, and Evolution: The Developmental-Genetic Basis of Evolutionary Change*. Macmillan Publishing Co. Inc. New York.

ROCKMAN, M.V. and WRAY, G.A. (2002). Abundant raw material for cis-regulatory evolution in humans. *Mol. Biol. Evol.* 19: 1991-2004.

RODRÍGUEZ-TRELLES, F., TARRÍO, R. and AYALA, F.J. (1999). Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. *Genetics* 153: 339-350.

RODRÍGUEZ-TRELLES, F., TARRÍO, R. and AYALA, F.J. (2000a). Disparate evolution of paralogous introns in the *Xdh* gene of Drosophila. *J. Mol. Evol.* 50: 123-130.

RODRÍGUEZ-TRELLES, F., TARRÍO, R. and AYALA, F.J. (2000b). Fluctuating mutation bias and the evolution of base composition in *Drosophila*. *J. Mol. Evol.* 50: 1-10.

RODRÍGUEZ-TRELLES, F., TARRÍO, R. and AYALA, F.J. (2000c). Evidence for a high ancestral GC content in *Drosophila*. *Mol. Biol. Evol.* 17: 1710-1717.

RODRÍGUEZ-TRELLES, F., TARRÍO, R. and AYALA, F.J. (2001a). Erratic overdispersion of three molecular clocks: GPDH, SOD, and XDH. *Proc. Natl. Acad. Sci. USA*. 98: 11405-11410.

RODRÍGUEZ-TRELLES, F., TARRÍO, R. and AYALA, F.J. (2001b). Xanthine dehydrogenase (XDH): episodic evolution of a "neutral" protein. *J. Mol. Evol.* 53: 485-495.

RODRÍGUEZ-TRELLES, F., TARRÍO, R. and AYALA, F.J. (2002). A methodological bias towards overestimation of molecular evolutionary time-scales. *Proc. Natl. Acad. Sci. USA.* 99: 8112-8115.

SEGAL, J.A., BARNETT, J.L. and CRAWFORD, D.L. (1999). Functional analysis of natural variation in Sp1 binding sites of a TATA-less promoter. *J. Mol. Evol.* 49: 736-749.

SHIELDS, D.C. (1990). Switches in species-specific codon preferences: the influence of mutation biases. *J. Mol. Evol.* 31: 71-80.

SHIELDS, D.C., SHARP, P.M., HIGGINS, D.G. and WRIGHT, F. (1988). "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Mol. Biol. Evol.* 5: 704-716.

SKAER, N. and SIMPSON, P. (2000). Genetic analysis of bristle loss in hybrids between *Drosophila melanogaster* and *D. simulans* provides evidence for divergence of *cis*regulatory sequences in the *achaete-scute* gene complex. *Dev. Biol.* 221: 148-167.

SLECKMAN, B.P., GORMAN, J.R. and ALT, F.W. (1996). Accessibility control of antigen-receptor variable-region gene assembly: role of cis-acting elements. *Annu. Rev. Immunol.* 14: 459-481.

SMALL, S., KRAUT, R., HOEY, T., WARRIOR, R. and LEVINE, M. (1991). Transcriptional regulation of a pair-rule stripe in Drosophila. *Genes Dev.* 5: 827-839.

STERN, D.L. (1998). A role of Ultrabithorax in morphological differences between *Drosophila* species. *Nature* 396: 463-466.

STERN, D.L. (2000). Evolutionary developmental biology and the problem of variation. *Evolution.* 54: 1079-1091.

STONE, J.R. and WRAY, G.A. (2001). Rapid evolution of cis-regulatory sequences via local point mutations. *Mol. Biol. Evol.* 18: 1764-1770.

SUCENA, E. and STERN, D.L. (2000). Divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by cis-regulatory evolution of *ovo/shaven-baby. Proc. Natl. Acad. Sci. USA.* 97: 4530-4534.

SUEOKA, N. (1988). Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. USA.* 85: 2653-2657.

TAKAHASI, H., MITANI, Y., SATOH, G. and SATOH, N. (1999). Evolutionary alterations of the minimal promoter for notochord-specific *Brachyury* expression in ascidian embryos. *Development* 126: 3725-3734.

TARRÍO, R., RODRÍGUEZ-TRELLES, F. and AYALA, F.J. (1998). New *Drosophila* introns originate by duplication. *Proc. Natl. Acad. Sci. USA.* 95: 1658-1662.

TARRÍO, R., RODRÍGUEZ-TRELLES, F. and AYALA, F.J. (2001). Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the *Drosophilidae*. *Mol. Biol. Evol.* 18: 1464-1473.

TARRÍO, R., RODRÍGUEZ-TRELLES, F. and AYALA, F.J. (2003). A new *Drosophila* spliceosomal intron position is common in plants. *Proc. Natl. Acad. Sci. USA.* 100: 6580-6583.

TAUTZ, D. (2000). Evolution of transcriptional regulation. *Curr. Opin. Genet. Dev.* 10: 575-579.

TORRENTS, E., ALOY, P., GIBERT, I. and RODRIGUEZ-TRELLES, F. (2002). Ribonucleotide reductases: divergent evolution of an ancient enzyme. *J. Mol. Evol.* 55: 138-152.

TRAVERS, A. (1993). *DNA-protein Interactions*. Chapman and Hall. London.

TROYANOVSKY, S.M. and LEUBE, R.E. (1994). Activation of the silent human cytokeratin 17 pseudogene-promoter region by cryptic enhancer elements of the cytokeratin 17 gene. *Eur. J. Biochem.* 223: 61-69.

WATERSTON, R. and SULSTON, J. (1995) The genome of Caenorhabditis elegans. *Proc. Natl. Acad. Sci. USA.* 92: 10836-10840.

WILSON, A.C. (1975). Evolutionary importance of gene regulation. *Stadler Symp* 7: 117-134.

WRAY, G.A. (2001). Dating branches on the tree of life using DNA. *Genome Biol.* 3: reviews0001.1-0001.7.

WRAY, G.A., HAHN, M.W., ABOUHEIF, E., BALHOFF, J.P., PIZER, M., ROCKMAN, M.V. and ROMANO, L. (2003). The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* 10.1093/molbev/msg140.

WU, C.Y. and BRENNAN, M.D. (1993). Similar tissue-specific expression of the *Adh* genes from different *Drosophila* species is mediated by distinct arrangements of *cis*-acting sequences. *Mol. Gen. Genet.* 240: 58-64.

XIN, L., LIU, D-P. and LING, C-C. (2003). A hypothesis for chromatin domain opening. *BioEssays* 25: 507-514.

YANG, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* 9: 367-372.

YANG, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15: 568-573.

YANG, Z. and BIELAWSKI, J.P. (2000). Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15: 496-503.

YANG, Z. and NIELSEN, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19: 908-917.

YAMPOLSKY, L.Y. and STOLTZFUS, A. (2001). Bias in the introduction of variation as an orienting factor in evolution. *Evol. Dev.* 3: 73-83.

YUH, C-H., BROWN, C.T., LIVI, C.B., ROWEN, L. CLARKE, P.J.C. and DAVIDSON, E.H. (2002). Patchy interspecific sequence similarities efficiently identify positive *cis*-regulatory elements in the sea urchin. *Dev. Biol.* 246: 148-161.

ZHANG, J. and GU. X. (1998). Correlation between the substitution rate and rate variation among sites in protein evolution. *Genetics* 149: 1615-1625.

ZERUCHA, T., STUHMER, T., HATCH, G., PARK, B.K., LONG, Q., YU, G., GAMBAROTTA, A. *et al.* (and 3 co-authors). (2000). A highly conserved enhancer in the Dlx5/Dlx6 region is the site of cross-regulatory interactions between Dlx genes in the embryonic forebrain. *J Neuro* 20: 709-721.

ZUCKERKANDL, E. (1963). Perspectives in molecular anthropology. In *Structural Chemistry and Molecular Biology* (eds. A. Rich and N. Davidson). Pp. 256-274. W. H. Freeman, San Francisco.