Title

Models of spliceosomal intron proliferation in the face of widespread ectopic expression

Permalink

https://escholarship.org/uc/item/0s4817js

Journal

Gene, 366(2)

ISSN

0378-1119

Authors

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

Publication Date

2006-02-01

DOI

10.1016/j.gene.2005.09.004

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Gene 366 (2006) 201 – 208

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

Review

Models of spliceosomal intron proliferation in the face of widespread ectopic expression

Francisco Rodríguez-Trelles [a,b,*], Rosa Tarrío [a,b], Francisco J. Ayala [a]

[a] *Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92697-2525, USA*
[b] *Grupo de Xenética Evolutiva, Departamento de Anatomía Patolóxica e Ciencias Forenses, Universidade de Santiago de Compostela, 15706-Santiago, Spain*

Received 24 June 2005; received in revised form 4 August 2005; accepted 2 September 2005
Available online 8 Novemer 2005
Received by G. Bernardi

Abstract

It is now certain that today living organisms can acquire new spliceosomal introns in their genes. The proposed sources of spliceosomal introns are exons, transposons, and other introns, including spliceosomal and group II self-splicing introns. Spliceosomal introns are thought to be the most likely source, because the inserted sequence would immediately be endowed with the essential set of intron recognition sequences, thereby preventing the deleterious effects associated with incorrect splicing. The most obvious spliceosomal intron duplication pathways involve an RNA transcript intermediate step. Therefore, for a spliceosomal intron to be originated by duplication, either the source gene from which the novel intron is derived, or that gene and the recipient gene, which contains the novel intron, would need to be expressed in the germ line. Intron proliferation surveys indicate that putative intron duplicate-containing genes do not always match detectable expression in the germ line, which casts doubt on the generality of the duplication model. However, judging mechanisms of intron gain (or loss) from present-day gene expression profiles could be erroneous, if expression patterns were different at the time the introns arose. In fact, this may likely be so in most cases. Ectopic expression, i.e., the expression of genes at times and locations where the target gene is not known to have a function, is a much more common phenomenon than previously realized. We conclude with a speculation on a possible interplay between spliceosomal introns and ectopic expression at the origin of multicellularity.

Keywords: Introns early theory; Introns late theory; Duplication; Intron evolution; Intron function; Origin of multicellularity

1. 'Introns early' and the late proliferation of spliceosomal introns

The debate on the origins and evolution of spliceosomal introns calls for two distinctions. The first distinction revolves around the two uses of the term 'intron'; specifically, introns as a theoretical construct, which should be distinguished from introns such as they become eventually instantiated into the specific types of intervening sequences which have hitherto been discovered (e.g., spliceosomal introns, group I and II introns, tRNA introns). As a theoretical construct–i.e., the notion of some sort of intragene non-coding sequence which is spliced out from the RNA before translation into amino acid sequence–introns are an invocation of the 'introns early' theory. The theory pushed the origin of the presently observed 'genes-in-pieces' structure of eukaryotic genes back to the RNA world (Doolittle, 1978). This claim was justified on several assumptions. One is that the first self-replicating coding nucleic acid sequences–the so-called 'minigenes' and/or the larger molecules that would have resulted from their accretion–would have necessarily included some stretches with and some without coding information (Darnell, 1981; Doolittle, 1981; Darnell and Doolittle, 1986). A second assumption is that RNA–RNA processing already existed in the earliest RNA world (Darnell, 1981; Darnell and Doolittle, 1986). This notion has recently been strengthened after the discovery that mRNA splicing by the spliceosome is, in fact, an RNA catalyzed reaction (Valadkhan, 2005), which confirms the previous conjecture based on the apparent similarities between the

[*] Corresponding author. Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92697-2525, USA. Tel.: +1 949 824 8293; fax: +1 949 824 2474.
E-mail address: ftrelles@usc.es (F. Rodríguez-Trelles).

0378-1119/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2005.09.004

splicing mechanisms of group II self-splicing introns and spliceosomal introns (Cavalier-Smith, 1991; Sharp, 2005). A third assumption, borrowed from the 'exon theory' of genes (Blake, 1978; Gilbert, 1978, 1987; Tonegawa et al., 1978), is that, once they originated, introns became co-opted into spacers, which increased the chances of illegitimate recombination between existing coding units. The earliest introns would have been retained because they facilitated the generation of novel proteins from pre-existing functional modules (in a similar way as exon shuffling fostered the assemblage of mosaic proteins at the origin of the metazoans; Doolittle, 1978; Patthy, 1999; Liu and Grigoriev, 2004; Cohen-Gihon et al., 2005). One corollary of the 'introns early' theory is that today's intron-lacking–or intron-sparse–genomes have resulted from extensive intron loss (which implies that the eukaryotic mode of gene organization antedates the origin of prokaryotes; Darnell and Doolittle, 1986).

The 'introns early' theory was inspired by the discovery of the 'split-gene' structure (Berget et al., 1977; Chow et al., 1977). Efforts to falsify the theory focused, accordingly, on the empirical evidence gathered from spliceosomal introns. These efforts, which gave body to the 'introns late' theory (Cavalier-Smith, 1978; Palmer and Logsdon, 1991; Logsdon, 1998), failed to demonstrate a correspondence between the position of introns and the boundaries of coding modules–identified following a variety of alternative criteria–for ancient proteins (i.e., defined as those whose origin preceded the eukaryote–prokaryote split; Logsdon, 1998). Moreover, neither spliceosomal introns, nor traces of the splicing machinery, have been found in the more than one hundred bacterial and archaeal genomes that have been completely sequenced (Lynch and Richardson, 2002).

The controversy between "introns early" and "introns late" has not been conclusively settled, to a large extent because of the uncertainty associated with the reconstruction of the historical pathway of extremely old events (Collins and Penny, 2005). For example, a study of mosaic genes assembled by exon shuffling from symmetrical modules of class 1–1 (i.e., modules flanked by introns placed between the first and second codon letters) at the origin of the metazoans–a relatively recent event compared with the origin of genes–has found that much of the inferred original intron-module structure has disappeared in flies and worms (Bányai and Patthy, 2004). Moreover, there is not conclusive demonstration of why natural selection for a streamlined genome, such as that typically exhibited by prokaryotes, would not result into wholesale intron elimination (Mourier and Jeffares, 2003; Charlesworth and Barton, 2004). Extensive intron losses have accompanied genome compaction in arthropods and nematodes, and apparently in many extant lineages of unicellular eukaryotes, as shown by phylogenetic (Rogozin et al., 2003; Roy and Gilbert, 2005a) and phylogeny-independent (Archibald et al., 2002; Simpson et al., 2002; Anantharaman et al., 2002; Bányai and Patthy, 2004; Collins and Penny, 2005) criteria.

Theories of nuclear genome size variation postulate specific evolutionary forces for wholesale intron elimination. The 'near-optimal DNA' theory (Cavalier-Smith, 1978) sets off the structural role of non-genic DNA in providing the nuclear skeleton ("nucleoskeletal" DNA), such that genome sizes (plus their tightness of packing and degree of unfolding) causally determine nuclear volumes. The ratio of the volume of the nucleus to that of the cytoplasm is invariant with cell volume owing to metabolic and steric constraints. Cell volume is a highly adaptive feature under cell-cycle genes control. Cell size decreases caused by mutations in cell-cycle genes will, accordingly, effect positive selection for a corresponding decrease in nuclear volume, which would be optimally achieved by decreasing the amount of nuclear DNA, including introns (Cavalier-Smith, 2005).

Plausible molecular mechanisms that could yield extensive intron loss have been identified (Mourier and Jeffares, 2003). If the ancestors at the divergences between major eukaryotic kingdoms (including protists and multicellular eukaryotes) already exhibited an intron-dense genome structure (Roy and Gilbert, 2005a), spliceosomal introns would have arisen much earlier. This conclusion is consistent with recent comparative analyses of many basal eukaryotic lineages showing that the last common ancestor of extant eukaryotes already contained a spliceosome that was similar in complexity to the spliceosome present in today's eukaryotes (Anantharaman et al., 2002; Collins and Penny, 2005).

Discovery of spliceosomal introns of late origin (Giroux et al., 1994; Logsdon et al., 1995; Hankeln et al., 1997; Frugoli et al., 1998; Tarrío et al., 2003; Coghlan and Wolfe, 2004) does not disprove the 'introns early' theory, but only indicates that new spliceosomal introns continue to be born in evolution. Moreover, inferred contemporary rates of intron birth are too slow to have been able to accrue present-day intron numbers (Roy and Gilbert, 2005b). Likewise, corroboration of the widespread notion that spliceosomal introns–together with the ribonucleic acid components of the spliceosomal apparatus–are descendants of the pieces of an hypothetical fragmentation underwent by type II self-splicing introns after they invaded the nucleus from the mitochondria (Sharp, 1991), would only imply that spliceosomal introns are not the right objects to falsify Doolittle's (1978) theoretical constructs.

It could conversely be entertained that the current organization of the ribonucleotide fraction of the spliceosome into pieces, including the five small nuclear RNAs in addition to the spliceosomal intron itself, reflect the ancestral state from which contemporary one-piece self-splicing introns II derived. Recent findings in the hyperthermophile parasite *Nanoarchaeum equitans* strongly suggest that this might have been the case for modern tRNA introns (Randau et al., 2005). This organism builds its functional tRNA genes from pieces. Each piece consists of two halves including a tRNA module and a terminal sequence. The two pieces join via reverse-complementation of their terminal sequences, which gives place to intervening duplex sequences. Taking into account the basal position of *Nanoarchaeum* within the archaea, it seems plausible that these intervening duplexes represent the ancestors from which modern one-piece tRNA introns evolved (Randau et al., 2005). Moreover, fusion events are estimated to be at least four times more common than fission events in the evolution of multi-domain protein genes (Kummerfeld and Teichmann, 2004).

The arguments we have advanced suggest that the current mechanisms by which spliceosomal introns proliferate may not be the same as the mechanisms that first caused spliceosomal introns to become integral parts of genes (Roy and Gilbert, 2005b). In this review, we focus on the current proliferation of introns. In particular, we explore some potentially important implications that the discovery that ectopic expression–i.e., the expression of genes at times and locations where the target gene is not known to have a function–is a widespread phenomenon (Khaitovich et al., 2004; Yanai et al., 2004; Rodríguez-Trelles, 2004; Rodríguez-Trelles et al., 2005) has for evaluating mechanisms of intron proliferation.

2. Intricacies of spliceosomal intron proliferation

It is now certain that living organisms can gain new spliceosomal introns in their genes. However, little is known about the frequencies and rates at which this happens. Do new introns proliferate steadily or episodically? Which, if any, are the correlations between intron gain and intron loss? Intron proliferation (like many other features of genomes, such as GC content, deletion/insertion rate ratios, and others) is a non-homogeneous non-stationary process, which varies both between lineages and within a lineage's evolution. Comparative genomics studies have evinced large differences in intron numbers that do not reflect the phylogeny of the species (e.g., Bhattacharya et al., 2000; Wada et al., 2002; Rogozin et al., 2003; Castillo-Davis et al., 2004; Cho et al., 2004; Edvardsen et al., 2004; Nielsen et al., 2004; Roy and Gilbert, 2005b). The causes underlying this variation are poorly understood, but doubtlessly involve a complex intertwining between natural selection and internal, bias-at-the-origin factors, none of which has been satisfactorily accounted for.

Natural selection influences intron proliferation if only because the greater the number of introns the most likely it will be that mutations will occur that impact essential intron recognition sequences and yield non-functional alleles (Lynch, 2002; Sharp, 2005). Also, transcription of long introns represents a metabolic cost to the cell (Castillo-Davis et al., 2002). According to the near-optimal DNA theory (Cavalier-Smith, 1978; see above), intronic DNA is selected against when nucleoskeletal DNA–and thus the volume of the nucleus–decreases in response to reductions in cell size (Cavalier-Smith, 2005). Yet, a growing body of data indicates that introns might also be favored by natural selection. Introns are no longer regarded as 'junk' DNA (because of their condition as rapidly evolving untranslated sequences at a time dominated by the 'central dogma'); rather, they are increasingly considered as epitome of functionality (Mattick, 1994, 2003; Mattick and Gagen, 2001; Le Hir et al., 2003; Lynch and Kewalramani, 2003; Rodríguez-Trelles et al., 2003; Cavalier-Smith, 2005; Bompfünewerer et al., 2005).

This change of perception is favored by three recent considerations: (i) proteins are not the end product of a serial processing, but a branch in a parallel information network where untranslated sequences might turn out to be central players (Mattick, 2003); (ii) in addition to its genic role, DNA performs regulatory and structural functions (e.g., nucleoskeletal role; Cavalier-Smith, 2005); and (iii) unlike protein-encoded messages, non-coding information, such as it is effected in the folding and interactions of nucleic acids, often does not impose a high degree of primary sequence definition (Bergman and Kreitman, 2001; Rodríguez-Trelles et al., 2003).

There is an increasing sense that introns are highly plastic, dynamically co-optable entities, whose role can change repeatedly during the course of their existence, in addition to frequently being multi-functional. Besides the constraints imposed by their own splicing, and their role in exon shuffling (Gilbert, 1978; Patthy, 1999), introns are intimately associated with the regulation of gene expression through the increasing amount of couplings between splicing and transcription/translation recently uncovered (Kornblihtt et al., 2004; Maquat, 2004); carry cis-regulatory information (Le Hir et al., 2003; Rodríguez-Trelles et al., 2003; Pagani and Baralle, 2004); are suspected to be integral components of a trans-acting regulatory network that would act in parallel with proteins in the development of complex organisms (Mattick, 1994, 2003; Mattick and Gagen, 2001); and play a role in RNA editing (Herbert, 1996). In addition, introns promote transcriptome variation by means of alternative splicing (Boue et al., 2003; Kondrashov and Koonin, 2003; Modrek and Lee, 2003; Ast, 2004); provide coordinates for the identification of premature termination codons in nonsense-mediated decay (Lynch and Kewalramani, 2003); may be an important factor for the spatial distribution of nucleosomes (Csordas, 1989; Lauderdale and Stein, 1992; Baldi et al., 1996; Denisov et al., 1997; Levitsky et al., 2001; Vinogradov, 2005); and likely are significant components of the nucleoskeletal DNA (Cavalier-Smith, 1978, 2005).

Some of these functions rest upon intron sequence features, whereas some others, like their role in non-sense mediated decay, appear to depend only on the positional information that is left at the exon–exon junctions after the introns are spliced out from the primary transcript. These and possibly other as yet undiscovered functions may have contributed to intron proliferation at different times and with varying intensities for different introns and lineages. Thus, cell size increases may create propitious conditions for intron proliferation because of the concomitant necessity of increasing nucleoskeletal DNA (Cavalier-Smith, 2005). However, as noted above, the number of introns per gene is expected to have an upper threshold above which additional insertions are disfavored by natural selection (Lynch, 2002; Lynch and Kewalramani, 2003).

Natural selection operates on the variation originated by mutation. Albeit mutations are random with respect to their effects on fitness, they are non-random in the sense that not all types of mutations are mechanistically equally probable (e.g., Rodríguez-Trelles et al., 1999, 2000a). Mutation biases can thus imprint orientation on intron proliferation. Intron gain is predominantly viewed as a complex mutational process which involves features of the recipient sequence–the so-called proto-splice site, and possibly various other recognition sequences, such as those that modulate exon splicing (Dibb and Newman, 1989; Sadusky et al., 2004)–as well as intron sources. The spatial distribution of proto-splice sites is strongly correlated with codon-usage biases, which could account in part for the excess of phase 0 introns (i.e., introns between codons) over phases 1 and 2 introns (Ruvinsky et al., 2005). Introns are rarely found very near one another within genes, which may be due to spatial constraints associated with a need of room for splicing and/or other restraints. The probability of an intron insertion at any particular target site will, therefore, be impacted by its distance to the nearest intron, which may change over evolutionary time. It remains unknown the extent to which such mechanical constraints, in conjunction with variable combinations of the selective forces discussed in the previous paragraph, can result in convergent intron distributions across independent lineages (Fedorov et al., 2002; Rogozin et al., 2003; Tarrío et al., 2003; Sverdlov et al., 2005).

Spliceosomal introns can arise from exons (Rogers, 1989), from transposons (Crick, 1979), or other introns including spliceosomal (Sharp, 1985) and group II self-splicing introns (Rogers, 1989). Each different intron source might result into a distinctive mode of intron proliferation. Because of their invasive properties, transposons may be favored as natural agents for bursts of intron spread, and this may account for the observed disparate phylogenetic patterns of intron density (Purugganan, 1993; Rzhetsky and Ayala, 1999; Fedorov et al., 2003). But the heterogeneity of phylogenetic patterns could be associated with other sources of introns, such as spliceosomal introns if the rates of duplication depend on the effectiveness of the biochemical processes involved (e.g., mutations altering the kinetics of reverse-splicing and/or reverse-transcription—see below). Ascertaining which, if any, of the aforementioned intron sources is most important, or if, alternatively, their relative significance varies during the course of evolution, will have a profound impact on our current understanding of the evolution of gene structure.

3. Germ line gene expression in models of spliceosomal intron duplication

Spliceosomal introns themselves have long been favored as the most likely source of new spliceosomal introns (Sharp, 1985; Hankeln et al., 1997; Logsdon, 1998; Tarrío et al., 1998; Coghlan and Wolfe, 2004). This “intron duplication” or “intron-transposition” model is appealing because it ensures that the inserted sequence is immediately endowed with the essential recognition sequences, which would prevent deleterious effects due to incorrect splicing (Sharp, 1985; Palmer and Logsdon, 1991; Lynch and Richardson, 2002). Three different mechanisms by which spliceosomal introns could beget new introns have been proposed (Fig. 1): (i) reverse-splicing of spliced introns into a new site in the same or a different mRNA, which is subsequently reverse-transcribed to a cDNA that recombines with the genome (Sharp, 1985); (ii) reverse-transcription of spliced introns which reinsert themselves into the nuclear genome (Lynch, 2002); and (iii) conversion of intron-less genes by reverse-transcription of unspliced transcripts of their intron-containing paralogs (Hankeln et al., 1997).

Of these three mechanisms, only the first two can originate novel intron positions; the third mechanism can result only in a new intron at the same position as the source intron (Coghlan and Wolfe, 2004). Moreover, the first pathway entails the extra benefit of providing a guidance mechanism by which introns

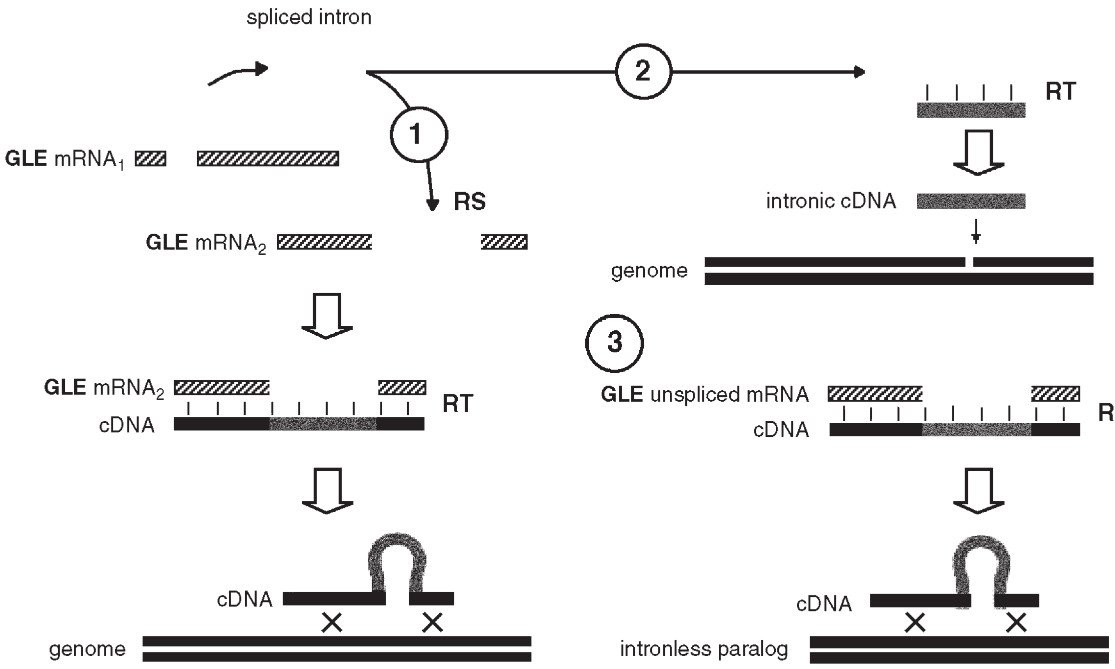

Fig. 1. Mechanisms of spliceosomal intron duplication. (1) Reverse-splicing (RS) of spliced introns into a new site in the same (mRNA$_1$) or a different (mRNA$_2$) germ line-expressed (GLE) mRNA, which is subsequently reverse-transcribed (RS) to a cDNA that recombines with the genome (Sharp, 1985); (2) reverse-transcription (RT) of spliced introns which reinsert themselves into the nuclear genome (Lynch, 2002); and (3) conversion of intron-less genes by reverse-transcription of unspliced transcripts of their intron-containing paralogs (Hankeln et al., 1997).

can be preferentially inserted into regions containing exon splicing enhancers (Lynch and Richardson, 2002). In order to be vertically transmitted, an intron gain must occur in a germ line cell (Logsdon et al., 1998; Coghlan and Wolfe, 2004; Roy, 2004). All three intron duplication pathways involve necessarily an RNA transcript intermediary. Therefore, for a spliceosomal intron to be originated by duplication, either the source gene from which the intron is derived, or–if the new intron originates via a reverse-splicing mechanism–that gene and the recipient gene containing the novel intron need to be expressed in the germ line (Logsdon, 1998; Logsdon et al., 1998; Coghlan and Wolfe, 2004; Roy, 2004).

Detection of intron duplication events usually starts with identification of newly gained introns on the basis of their restricted phylogenetic distribution. The sequence of each new intron is then matched to the remaining available intron sequences of the same genome in search for similarity. Only three studies have obtained positive results (Hankeln et al., 1997; Tarrío et al., 1998; Coghlan and Wolfe, 2004). Only the most recent study has examined expression data (Coghlan and Wolfe, 2004). The study identified 122 recent gains, defined as introns present in either *Caenorhabditis elegans* or *C. briggsae* that are absent from two independent distantly related outgroups (the parasitic nematode *Brugia malayi* and the arthropod-vertebrate clade consisting of fly, mosquito, human, and mouse). This definition of recent intron gain hinges on the assumption that one intron gain is more likely than three independent intron losses. This may be too liberal an assumption, particularly if we take into account that rates of intron loss appear to be much greater, perhaps one order of magnitude greater, than rates of intron gain (Kryzwinski and Besansky, 2002; Lynch, 2002; Wada et al., 2002; Cho et al., 2004; Kiontke et al., 2004; Roy and Gilbert, 2005b). Putative novel introns are on average longer than control introns (Coghlan and Wolfe, 2004); this, however, may be because nematodes preferentially lose shorter introns (Cho et al., 2004). Moreover, one intron gain considered certain by Coghlan and Wolfe (2004) may represent an insertion of a palindromic element into a pre-existing intron (Roy, 2004). Be that as it may, out of the 122 presumptive novel introns, 28 exhibit significant sequence identity to other introns of the same genome, which strongly suggests that they arose by duplication of older introns (Coghlan and Wolfe, 2004).

Both, the novel introns and the set of older introns that they match are preferentially located in genes with detectable germ line expression (Coghlan and Wolfe, 2004); although the correlation is not perfect, which casts doubt on the intron duplication model (Logsdon, 1998; Logsdon et al., 1998; Coghlan and Wolfe, 2004; Roy, 2004). Should a perfect correlation be expected? In our view, not necessarily, if the following two conditions obtain. (i) There is a stochastic component to gene expression. The conventional notion of the 'expression level of a gene' as a cell type- or tissue-feature is an artefact that was prompted by the study of gene expression with methods that require large cell populations (Paldi, 2003; Kurakin, 2005). Modern single-cell based approaches indicate that the activity of a gene can vary significantly between cells of the same tissue, simply because of the fact that gene expression intrinsically relies upon random encounters between finite, and often small, numbers of diffusible molecules (Sternberg and Félix, 1997; Paldi, 2003; Kurakin, 2005; Kaern et al., 2005; Theise, 2005). The probabilistic character of gene activation makes it conceivable that RNA-mediated intron duplication could occur via stochastically produced transcripts in cells from germ line tissues where the average steady state of expression of the encoding genes may pass undetected. (ii) There is gene expression turnover, which in the long term is likely to be of greater consequence than (i). Evolutionary transcriptomics studies (Khaitovich et al., 2004; Yanai et al., 2004; reviewed in Rodríguez-Trelles et al., 2005) suggest that present-day gene expression profiles may carry only limited information about the expression profiles of the recent past.

4. Widespread ectopic expression and the proliferation of *Xdh* introns

The requirement of germ line gene expression in models of intron duplication emanates from a long-standing regulatory paradigm, which claims that gene expression profiles are controlled down to the last detail (Carroll et al., 2001; Davidson, 2001; Wilkins, 2002). Under this scheme, ectopic expression, i.e., the expression of genes at times and locations where the target gene is not known to have a function, would be mostly deleterious. This paradigm has been challenged by molecular geneticists who have shown that any gene may be transcribed in any cell type (Humphries et al., 1976; Weintraub and Groudine, 1976; Chelly et al., 1989), and evolutionists who have shown that enzymatic-protein expression profiles are greatly variable, even among closely related species (Dickinson, 1980; see Rodríguez-Trelles et al., 2005).

Evolutionary transcriptomics studies have shown that (i) a substantial fraction of gene expression differences between species is adaptively neutral or nearly neutral (Khaitovich et al., 2004), and (ii) that for any given species and tissue, it is frequently not possible to anticipate whether a gene will be transcriptionally active or not on the basis of its expression status in the same tissue in related species (Yanai et al., 2004). These findings indicate that ectopic expression is widespread, an interpretation consistent with (i) the properties of cis-regulatory sequences, which are typically short and thus can easily arise–or be dismantled–by mutation randomly throughout the genomes, and (ii) gene cross-talking conflicts arising because unrelated promoters often carry cis-regulatory sequences for the same transcription factor (see Rodríguez-Trelles et al., 2005). Apparently, many genes can change their transcriptional status erratically during the course of evolution without major functional impingement. A corollary of this conclusion is that present-day germ line expression status ('on' or 'off') of newborn-intron-containing genes might be irrelevant for evaluating intron duplication models.

As an example, consider the case of the *Xdh* (xanthine dehydrogenase) gene. In *D. sucinea* and *D. capricorni*, two closely related species of the *Drosophila willistoni* group, the *Xdh* gene carries two short introns referred to as introns A and B

(Tarrío et al., 1998). Introns A and B are most likely novel introns (40 My old, the approximate age of the *sucinea–capricorni* lineage, or less) because they are absent from all 16 increasingly distantly related lineages of animals and fungi (those listed in Tarrío et al., 2003, plus *Apis mellifera* and *Tribolium castaneum*). The two introns exhibit significant sequence similarity to an older intron located nearby within the *Xdh* gene, which indicates their origin by duplication (Tarrío et al., 1998). Retention of similarity between the old and new introns might have been facilitated because the *Xdh* region has evolved more slowly in *D. sucinea* and *D. capricorni* than in other *willistoni* species (Rodríguez-Trelles et al., 2000b). The expression status of *Xdh* in the germ line of these species is unknown. However, Dickinson (1980) detected *Xdh* activity in the ovaries of *D. adiastola* and *D. ornata*, two members of the Hawaiian *Drosophila adiastola* subgroup, using protein electrophoresis. The fact that he was not able to detect *Xdh* expression in the remaining 25 Hawaiian species of his survey–including additional *adiastola* subgroup representatives–was taken as an indication that the referred *Xdh* activities were ectopic. But we want to point out that even if *Xdh* is currently inactive in the germ lines of *D. sucinea* and *D. capricorni*, it could well have been active at the time introns A and B arose. Widespread ectopic expression might thus account for the imperfect matching between germ line expressed genes and genes that carry new introns that still resemble their parental introns (Coghlan and Wolfe, 2004). This is, of course, the case for any model of intron gain–or loss–that requires an RNA transcript intermediary.

Studies seeking to evaluate the mechanisms of intron origins should show that the implicated genes were transcriptionally active in the germ line at the relevant times. There are at least two modes of tackling this issue. One is reconstructing ancestral gene expression states from appropriate phylogenetic sampling. A second way is by circumscribing the analyses to genes which are known to be performing essential functions in the germ line (hence being realistically expected to have remained stably expressed). In cases of ancient intron gain events, it may be all but impossible to ascertain that the target genes were expressed in the germ line, but ancient intron duplications may be difficult to establish because their sequence similarity would have largely decayed.

5. Spliceosomal introns and ectopic expression in the evolution of multicellularity

The evolution of multicellularity represents a major transition in the history of life, which may have independently occurred several times (Kirk, 2005). Multicellular organisms develop from a single cell that replicates to give rise to a spatially structured individual with a number of differentiated cell types. The starting condition for the evolution of multicellularity is assumed to be a colony of identical cells derived from the clonal expansion of a single cell (Aravind and Subramanian, 1999; Maynard-Smith and Szathmáry, 1999; Kirk, 2005). Subsequently, certain changes would have assorted the expression of ancestral regulatory effectors among distinct subsets of cells in the colony, thus triggering spatial differentiation for the evolution of functional differentiation among distinct cell types (Aravind and Subramanian, 1999; Maynard-Smith and Szathmáry, 1999). In unicellular organisms, differentiation consists of a succession of gene expression states in response to environmental conditions; ectopic expression can only be displayed on a temporal dimension. Changes in the regulatory transcriptional profiles of different cell subsets of early cell aggregates could have represented the earliest evolutionary manifestation of ectopic expression on a spatial scale. Novel expression profiles could become heritable traits by concomitant changes in DNA methylation patterns or chromatin marks. Insofar as ectopic expression is a reflection of architectural constraints of the regulatory system (Rodríguez-Trelles et al., 2005), the origin of multicellularity might be contemplated as a natural outcome of cell aggregation.

The unfolding of ectopic expression along the spatial axis would represent an explosion in the number of cellular environments to which gene products were exposed, increasingly so as the cell types became more and more specialized. Environmental diversification would have further expanded the range of potential interactions, thus opening new avenues for the recruitment of genetic variation. Spliceosomal introns may have been primary players in this scenario, first by allowing the generation of novel combinations of exons by exon shuffling (Patthy, 1999; Cohen-Gihon et al., 2005), and second, because they became readily co-opted for alternative splicing (Boue et al., 2003; Ast, 2004). The origin of multicellularity might thus have left its own imprint in the subsequent proliferation of spliceosomal introns.

Acknowledgements

F. R-T and R. T. have received support from contracts Ramón y Cajal and F.I.S. G03/184, from the Ministerio de Ciencia y Tecnología and Ministerio de Sanidad (Spain), respectively.

References

Anantharaman, V., Koonin, E.V., Aravind, L., 2002. Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res. 30, 1427–1464.

Aravind, L., Subramanian, G., 1999. Origin of multicellular eukaryotes—insights from proteome comparisons. Curr. Opin. Genet. Dev. 9, 688–694.

Archibald, J.M., O'Kelly, C.J., Doolittle, W.F., 2002. The chaperonin genes of jakobid and jakobid-like flagellates: implications for eukaryotic evolution. Mol. Biol. Evol. 19, 422–431.

Ast, G., 2004. How did alternative splicing evolve? Nat. Rev. Genet. 5, 773–782.

Baldi, P., Brunak, S., Chauvin, Y., Krogh, A., 1996. Naturally occurring nucleosome positioning signals in human exons and introns. J. Mol. Biol. 263, 503–510.

Bányai, L., Patthy, L., 2004. Evidence that human genes of modular proteins have retained significantly more ancestral introns than their fly or worm orthologues. FEBS Lett. 565, 127–132.

Berget, S.M., Moore, C., Sharp, P.A., 1977. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. Proc. Natl. Acad. Sci. U. S. A. 74, 3171–3175.

Bergman, C.M., Kreitman, M., 2001. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. Genome Res. 11, 1335–1345.

Bhattacharya, D., Lutzoni, F., Reeb, V., Simon, D., Nason, J., Fernández, F., 2000. Widespread occurrence of spliceosomal introns in the rDNA genes of ascomycetes. Mol. Biol. Evol. 17, 1971–1984.

Blake, C.C.F., 1978. Do genes-in-pieces imply protein-in-pieces? Nature 273, 267.

Bompfünewerer, A.F., et al., 2005. Evolutionary patterns of non-coding RNAs. Theory Biosci. 123, 301–369.

Boue, S., Letunic, I., Bork, P., 2003. Alternative splicing and evolution. BioEssays 25, 1031–1034.

Carroll, S.B., Grenier, J.K., Weatherbee, S.D., 2001. From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design. Blackwell Scientific, Malden, MA, USA.

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

Castillo-Davis, C.I., Bedford, T.B.C., Hartl, D.L., 2004. Accelerated rates of intron gain/loss and protein evolution in duplicate genes in human and mouse malaria parasites. Mol. Biol. Evol. 21, 1422–1427.

Cavalier-Smith, T., 1978. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. J. Cell Sci. 34, 247–278.

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

Cavalier-Smith, T., 2005. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. Ann. Bot. 95, 147–175.

Charlesworth, B., Barton, N., 2004. Does bigger mean worse? Curr. Biol. 14, R233–R235.

Chelly, J., Concordet, J.P., Kaplan, J.C., Kahn, A., 1989. Illegitimate transcription: transcription of any gene in any cell type. Proc. Natl. Acad. Sci. U. S. A. 86, 2617–2621.

Cho, S., Jin, S.W., Cohen, A., Ellis, R.E., 2004. A phylogeny of *Caenorhabditis* reveals frequent loss of introns during nematode evolution. Genome Res. 14, 1207–1220.

Chow, L.T., Gelinas, R.E., Broker, T.R., Roberts, R.T., 1977. An amazing sequence arrangement at the 5′ ends of adenovirus-2 messenger RNA. Cell 12, 1–8.

Coghlan, A., Wolfe, K.H., 2004. Origins of recently gained introns in *Caenorhabditis*. Proc. Natl. Acad. Sci. U. S. A. 101, 11362–11367.

Cohen-Gihon, I., Lancet, D., Yanai, I., 2005. Modular genes with metazoan-specific domains have increased tissue specificity. Trends Genet. 21, 210–213.

Collins, L., Penny, D., 2005. Complex spliceosomal organization ancestral to extant eukaryotes. Mol. Biol. Evol. 22, 1053–1066.

Crick, F., 1979. Split genes and RNA splicing. Science 204, 264–271.

Csordas, A., 1989. A proposal for a possible role of nucleosome positioning in the evolutionary adjustment of introns. Int. J. Biochem. 21, 455–461.

Darnell, J.E., 1981. Do features of present-day eukaryotic organisms reflect ancient sequence arrangements? In: Scudder, G.G.E., Reveal, J.L. (Eds.), Evolution Today. Proc. 2nd Int. Cong. Syst. Evol. Biol. Hunt. Inst. Bot. Document. Carnegie-Mellon Univ., Pittsburgh, pp. 207–213.

Darnell, J.E., Doolittle, W.F., 1986. Speculations on the early course of evolution. Proc. Natl. Acad. Sci. U. S. A. 83, 1271–1275.

Davidson, E.H., 2001. Genomic Regulatory Systems—Development and Evolution. Academic Press, San Diego.

Denisov, D.A., Shpigelman, E.S., Trifonov, E.N., 1997. Protective nucleosome centering at splice sites as suggested by sequence-directed mapping of the nucleosomes. Gene 205, 145–149.

Dibb, N.J., Newman, A.J., 1989. Evidence that introns arose at proto-splice sites. EMBO J. 8, 2015–2021.

Dickinson, W.J., 1980. Evolution of patterns of gene expression in Hawaiian Picture-winged *Drosophila*. J. Mol. Evol. 16, 73–94.

Doolittle, W.F., 1978. Genes in pieces: were they ever together? Nature 272, 581–582.

Doolittle, W.F., 1981. Prejudices and preconceptions about genome evolution. In: Scudder, G.G.E., Reveal, J.L. (Eds.), Evolution Today. Proc. 2nd Int. Cong. Syst. Evol. Biol. Hunt. Inst. Bot. Document. Carnegie-Mellon Univ., Pittsburgh, pp. 197–205.

Edvardsen, R.B., et al., 2004. Hypervariable and highly divergent intron–exon organizations in the chordate *Oikopleura dioica*. J. Mol. Evol. 59, 448–457.

Fedorov, A., Merican, A.F., Gilbert, W., 2002. Large-scale comparison of intron positions among animal, plant, and fungal genes. Proc. Natl. Acad. Sci. U. S. A. 99, 16128–16133.

Fedorov, A., Roy, S., Fedorova, L., Gilbert, W., 2003. Mystery of intron gain. Genome Res. 13, 2236–2241.

Frugoli, J.A., McPeek, M.A., Thomas, L.T., McClung, C., 1998. Intron loss and gain during evolution of the catalase gene family in angiosperms. Genetics 149, 355–365.

Gilbert, W., 1978. Why genes in pieces? Nature 271, 501.

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

Giroux, M.J., Clancy, M., Baier, J., Ingham, L., McCarty, D., Hannah, L.C., 1994. De novo synthesis of an intron by the maize transposable element *Dissociation*. Proc. Natl. Acad. Sci. U. S. A. 91, 12150–12154.

Hankeln, T., Friedl, H., Ebersberger, I., Martin, J., Schmidt, E.R., 1997. A variable intron distribution in globin genes of *Chironomus*: evidence for recent intron gain. Gene 205, 151–160.

Herbert, A., 1996. RNA editing, introns and evolution. Trends Genet. 12, 6–9.

Humphries, S., Windass, J., Williamson, R., 1976. Mouse globin gene expression in erythroid and non-erythroid tissues. Cell 7, 267–277.

Kaern, M., Elston, T.C., Blake, W.J., Collins, J.J., 2005. Stochasticity in gene expression: from theories to phenotypes. Nat. Rev. Genet. 6, 451–464.

Khaitovich, P., et al., 2004. A neutral model of transcriptome evolution. PloS Biol. 2, 682–689.

Kiontke, K., Gavin, N.P., Raynes, Y., Roehrig, C., Piano, F., Fitch, D.H., 2004. *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. Proc. Natl. Acad. Sci. U. S. A. 101, 9003–9008.

Kirk, D.L., 2005. A twelve-step program for evolving multicellularity and a division of labor. BioEssays 27, 299–310.

Kondrashov, F.A., Koonin, E.V., 2003. Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from introns sequences. Trends Genet. 19, 115–119.

Kornblihtt, A.R., De la Mata, M., Fededa, J.P., Muñoz, M.J., Nogués, G., 2004. Multiple links between transcription and splicing. RNA 10, 1489–1498.

Kryzwinski, J., Besansky, N.J., 2002. Frequent intron loss in the *White* gene: a cautionary tale for phylogeneticists. Mol. Biol. Evol. 19, 362–366.

Kummerfeld, S.K., Teichmann, S.A., 2004. Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet. 21, 25–30.

Kurakin, A., 2005. Self-organization vs. watchmaker: stochastic gene expression and cell differentiation. Dev. Genes Evol. 215, 46–52.

Lauderdale, J.D., Stein, A., 1992. Introns in the chicken ovalbumin gene promote nucleosome alignment in vitro. Nucleic Acids Res. 20, 6589–6596.

Le Hir, H., Nott, A., Moore, M.J., 2003. How introns influence and enhance eukaryotic gene expression. Trends Biomed. Sci. 28, 215–220.

Levitsky, V.G., Podkolodnaya, O.A., Kolchanov, N.A., Podkolodny, N.L., 2001. Nucleosome formation potential of exons, introns, and *Alu* repeats. Bioinformatics 17, 1062–1064.

Liu, M.Y., Grigoriev, M.A., 2004. Protein domains correlate strongly with exons in multiple eukaryotic genomes—evidence of exon shuffling? Trends Genet. 20, 399–403.

Logsdon, J.M., 1998. The recent origins of spliceosomal introns revisited. Curr. Opin. Genet. Dev. 8, 637–648.

Logsdon, J.M., Tyshenko, M.G., Dixon, C., Jafari, J.D., Walker, V.K., Palmer, J. D., 1995. Seven newly discovered intron positions in the triose-phosphate-isomerase gene: evidence for the introns-late theory. Proc. Natl. Acad. Sci. U. S. A. 92, 8507–8511.

Logsdon, J.M., Stoltzfus, A., Doolittle, W.F., 1998. Molecular evolution: recent cases of spliceosomal intron gain? Curr. Biol. 8, R560–R563.

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

Lynch, M., Kewalramani, A., 2003. Messenger RNA surveillance and the evolutionary proliferation of introns. Mol. Biol. Evol. 20, 563–571.

Lynch, M., Richardson, A.O., 2002. The evolution of spliceosomal introns. Curr. Opin. Genet. Dev. 12, 701–710.

Maquat, L.E., 2004. Non-sense mediated RNA decay: splicing, translation and mRNP dynamics. Nat. Rev. Mol. Cell Biol. 5, 89–99.

Mattick, J.S., 1994. Introns evolution and function. Curr. Opin. Genet. Dev. 4, 823–831.

Mattick, J.S., 2003. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. BioEssays 25, 930–939.

Mattick, J.S., 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.

Maynard-Smith, J., Szathmáry, E., 1999. The Origins of Life. From the Birth of Life to the Origin of Language. Oxford University Press, Oxford.

Modrek, B., Lee, C.J., 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat. Genet. 34, 177–180.

Mourier, T., Jeffares, D.C., 2003. Eukaryotic intron loss. Science 300, 1393.

Nielsen, C.B., Friedman, B., Birren, B., Burge, C.B., Galagan, J.E., 2004. Patterns of intron gain and loss in fungi. PLoS Biol. 2, e422.

Pagani, F., Baralle, F.E., 2004. Genomic variants in exons and introns: identifying the splicing spoilers. Nat. Rev. Genet. 5, 389–396.

Paldi, A., 2003. Stochastic gene expression during cell differentiation: order from disorder? Cell. Mol. Life Sci. 60, 1775–1778.

Palmer, J.D., Logsdon, J.M., 1991. The recent origin of introns. Curr. Opin. Genet. Dev. 1, 470–477.

Patthy, L., 1999. Genome evolution and the evolution of exon-shuffling—a review. Gene 238, 103–114.

Purugganan, M.D., 1993. Transposable elements as introns: evolutionary connections. Trends Genet. 8, 239–243.

Randau, L., Münch, R., Hohn, M.J., Jahn, D., Söll, D., 2005. Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5′- and 3′-halves. Nature 433, 537–541.

Rodríguez-Trelles, F., 2004. Transcriptome evolution: much ado about nothing? Heredity 93, 405–406.

Rodríguez-Trelles, F., Tarrío, R., 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., Ayala, F.J., 2000a. Fluctuating mutation bias and the evolution of base composition in *Drosophila*. J. Mol. Evol. 50, 1–10.

Rodríguez-Trelles, F., Tarrío, R., Ayala, F.J., 2000b. Disparate evolution of paralogous introns in the *Xdh* gene of *Drosophila*. J. Mol. Evol. 50, 1–10.

Rodríguez-Trelles, F., Tarrío, R., Ayala, F.J., 2003. Evolution of cis-regulatory regions versus codifying regions. Int. J. Dev. Biol. 47, 665–673.

Rodríguez-Trelles, F., Tarrío, R., Ayala, F.J., 2005. Is ectopic expression caused by deregulatory mutations or due to gene-regulation leaks with evolutionary potential? BioEssays 27, 592–601.

Rogers, J.H., 1989. How were introns inserted into nuclear genes? Trends Genet. 5, 213–216.

Rogozin, I.B., Wolf, Y.I., Sorokin, A.V., Mirkin, B.G., Koonin, E.V., 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr. Biol. 13, 1512–1517.

Roy, S.W., 2004. The origin of recent introns: transposons? Genome Biol. 5, 251.

Roy, S.W., Gilbert, W., 2005a. Complex early genes. Proc. Natl. Acad. Sci. U. S. A. 102, 1986–1991.

Roy, S.W., Gilbert, W., 2005b. Rates of intron loss and gain: implications for early eukaryotic evolution. Proc. Natl. Acad. Sci. U. S. A. 102, 5773–5778.

Ruvinsky, A., Eskesen, S.T., Eskesen, F.N., Hurst, L.D., 2005. Can codon usage bias explain intron phase distributions and exon symmetry? J. Mol. Evol. 60, 99–104.

Rzhetsky, A., Ayala, F.J., 1999. The enigma of intron origins. Cell. Mol. Life Sci. 55, 3–6.

Sadusky, T., Newman, A.J., Dibb, N.J., 2004. Exon junction sequences as cryptic splice sites: implications for intron origins. Curr. Biol. 14, 505–509.

Sharp, P.A., 1985. On the origin of RNA splicing and introns. Cell 42, 397–400.

Sharp, P.A., 1991. Five easy pieces. Science 254, 263.

Sharp, P.A., 2005. The discovery of split genes and RNA splicing. Trends Biochem. Sci. 30, 279–281.

Simpson, A.G.B., MacQuarrie, E.K., Roger, A.J., 2002. Early origin of canonical introns. Nature 419, 270.

Sternberg, P.W., Félix, M.-A., 1997. Evolution of cell lineage. Curr. Opin. Genet. Dev. 7, 543–550.

Sverdlov, A.V., Rogozin, I.B., Babenko, V.N., Koonin, E.V., 2005. Conservation versus parallel gains in intron evolution. Nucleic Acids Res. 33, 1741–1748.

Tarrío, R., Rodríguez-Trelles, F., Ayala, F.J., 1998. New *Drosophila* introns originate by duplication. Proc. Natl. Acad. Sci. U. S. A. 95, 1658–1662.

Tarrío, R., Rodríguez-Trelles, F., Ayala, F.J., 2003. A new Drosophila spliceosomal intron position is common in plants. Proc. Natl. Acad. Sci. U. S. A. 100, 6580–6583.

Theise, N.D., 2005. Now you see it, now you don’t. Nature 435, 1165.

Tonegawa, S., Maxam, A.M., Tizard, R., Bernard, O., Gilbert, W., 1978. Sequence of a mouse germ-line gene for a variable region of an immunoglobulin light chain. Proc. Natl. Acad. Sci. U. S. A. 75, 1485–1489.

Valadkhan, S., 2005. Construction of a minimal protein-free spliceosome. Science 307, 863–864.

Vinogradov, A.E., 2005. Noncoding DNA, isochores and gene expression: nucleosome formation potential. Nucleic Acids Res. 33, 559–563.

Wada, H., Kobayashi, M., Sato, R., Satoh, N., Miyasaka, H., Shirayama, Y., 2002. Dynamic insertion-deletion of introns in deuterostome EF-1α genes. J. Mol. Evol. 54, 118–128.

Weintraub, H., Groudine, M., 1976. Chromosomal subunits in active genes have an altered conformation. Science 193, 848–856.

Wilkins, A.S., 2002. The Evolution of Developmental Pathways. Sinauer Associates, Sunderland, MA, USA.

Yanai, I., Graur, D., Ophir, R., 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. OMICS 8, 15–24.