5 Snyder, M. and Gerstein, M. (2003) Genomics. Defining genes in the genomics era. *Science* 300, 258–260
6 Ciccarelli, F.D. *et al.* (2005) Complex genomic rearrangements lead to novel primate gene function. *Genome Res.* 15, 343–351
7 Zdobnov, E.M. *et al.* (2002) Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* 298, 149–159
8 Whiting, M.F. (2002) Phylogeny of the holometabolous insect orders based on 18S ribosomal DNA: when bad things happen to good data. *EXS* 92, 69–83
9 Raible, F. *et al.* (2005) Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*. *Science* 310, 1325–1326
10 Savard, J. *et al.* (2006) Genome-wide acceleration of protein evolution in flies (Diptera). *BMC Evol. Biol.* 6, 7
11 Savard, J. et al. (2006) Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. *Genome Res.* DOI:10.1101/gr.5204306 (www.genome.org)
12 Hillier, L.W. *et al.* (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432, 695–716
13 Nadeau, J.H. and Taylor, B.A. (1984) Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci. U. S. A.* 81, 814–818
14 Negre, B. *et al.* (2005) Conservation of regulatory sequences and gene expression patterns in the disintegrating *Drosophila* Hox gene complex. *Genome Res.* 15, 692–700
15 Miller, J.R. (2002) The Wnts. *Genome Biol* 3 REVIEWS3001
16 Xia, Q. *et al.* (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science* 306, 1937–1940
17 Holt, R.A. *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298, 129–149
18 Drysdale, R.A. and Crosby, M.A. (2005) FlyBase: genes and gene models. *Nucleic Acids Res.* 33, D390–D395
19 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
20 Salamov, A.A. and Solovyev, V.V. (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* 10, 516–522
21 Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797
22 Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552
23 Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704
24 Schmidt, H.A. *et al.* (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18, 502–504
25 Jones, D.T. *et al.* (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8, 275–282

0168-9525/$ – see front matter
doi:10.1016/j.tig.2006.10.004

Intron size in mammals: complexity comes to terms with economy

Uberto Pozzoli[1], Giorgia Menozzi[1], Giacomo P. Comi[2], Rachele Cagliani[1], Nereo Bresolin[1,2] and Manuela Sironi[1]

[1] Bioinformatic Laboratory, Scientific Institute IRCCS E. Medea, Via don L. Monza 20, 23842 Bosisio Parini (LC), Italy
[2] Dino Ferrari Centre, Department of Neurological Sciences, University of Milan, IRCCS Ospedale Maggiore Policlinico, Mangiagalli and Regina Elena Foundation, 20100 Milan, Italy

Different and contrasting models have been proposed to explain intron size evolution in mammals. Here, we demonstrate that intron and intergenic size *per se* has no adaptive role in gene expression regulation but reflects the need to preserve conserved intronic elements. Although the amount of non-coding functional elements explains the within-genome size variation of intergenic spacers, we show that an additional, additive pressure has been acting on highly expressed introns to reduce the cost of their transcription.

Introduction

Two main models have been proposed to explain within-genome variation in intron size in humans and other eukaryotes. The 'selection for economy' model envisages a situation whereby highly expressed genes are subjected to stronger pressure for intron shortening compared with genes expressed at low levels or in few tissues [1–5]. Conversely, the 'genomic design' model [6–8] hypothesizes that longer introns (and long intergenic regions) preferentially occur in tissue-specific genes because they allow chromatin-mediated gene suppression and complex regulation.

These two models have at least one thing in common: they postulate that gene expression patterns or levels have been conserved for an evolutionary time long enough for selection to leave a signature. Moreover, they are not mutually exclusive: low-expressed and tissue-specific genes might have greater regulatory needs and therefore longer intronic sequences, whereas widely or highly expressed genes, lacking strong intron constraints, might have been subjected to selective pressure to reduce intron length.

Recent studies (reviewed in Refs [9,10]) have shown that the non-coding regions of eukaryotic genomes are punctuated by numerous, relatively short sequences that have been conserved through long evolutionary times (multi-species conserved sequences or MCSs). Given that we have previously demonstrated [11] that fixation of MCSs influences intron size in humans, we wished to verify whether this finding might be integrated in a unifying model for intron size evolution.

Corresponding author: Sironi, M. (manuela.sironi@bp.lnf.it)
Available online 30 October 2006.

The presence of MCSs constrains intron size

To study the impact of the presence of MCSs on intron size, we aligned 27 008 pairs of introns that are orthologous between mouse and rat (see the Supplementary Data online for methods) and recorded insertions of transposable elements (TEs) and deletion events in the mouse sequences. Deletion events were inferred by retrieving mouse sequence gaps longer than 80 bp that are not accounted for by repeat insertions or microsatellite sequences in rat. Intron lengths before the mouse–rat divergence (original lengths) were estimated by adding deletions and subtracting TE insertions to present-day mouse intron lengths. The same procedures were applied to human–chimpanzee alignments (39 323 intron pairs). The data reported here refer to the rodent genome comparison; for the analysis of human deletion and insertion frequencies, a different approach based on simulations was used because of the paucity of events; these data confirm the observations in rodents and are available in the Supplementary Data online.

TE insertion and deletion frequencies were analyzed after dividing mouse introns into four length classes and, independently, into four groups depending on MCS density; Kruskall-Wallis tests were used to study the effect of MCS density within all size groups. Deletion frequency increases with intron size but, similarly to TE insertion frequency (Figure 1a), within each length class, frequencies diminish with increasing MCS content (for both TE insertions and deletions, Kruskall-Wallis $p < 0.01$ for differences within all size groups). Interestingly, the ratio of deletion frequency to insertion frequency also decreases with increasing MCS density (Figure 1b, Kruskall-Wallis $p < 0.01$ for the second and third length classes). We therefore wished to estimate the variation in mouse intron size relative to the common ancestor of the mouse and rat. An average shrinkage, stronger for longer introns, is evident in Figure 1c. However, it is equally clear that size contraction is progressively reduced, within each length class (except the smallest intron class), for increasing MCS densities (Kruskall-Wallis $p < 0.01$ for differences within all size groups), with introns showing very small size variations when extremely rich in conserved sequences.

These data suggest that the presence of MCSs selects against deletion events and, to a lesser extent, against repeat insertions (Figure 1b) so that the higher the intronic MCS density is, the less the size decreases (Figure 1c). As a consequence, size distributions are quite different when MCS-containing and MCS-lacking introns are analyzed (Figure 1d for mouse and Figure S1c in the Supplementary Data online for human). Indeed, introns that had no MCSs, being free of constraints, have probably undergone progressive size contraction over time so that the majority of them are nowadays relatively short. Figure 1e (Figure S1a in the Supplementary Data online for human) shows the variation of MCS density with size: consistent with these speculations, an increasing trend is observed for relatively long introns. Conversely, MCS density decreases with size for shorter introns (below $\sim 10^{3.5}$ bp); although this observation partially reflects the fact that the minimum MCS density for an intron is the reciprocal of its size, the presence of a minority of short introns with a high MCS density might be explained by considering that TE insertions and relatively large deletions are unlikely in short introns (Figure S2a in the Supplementary Data online), which are therefore only amenable to micro-insertions or deletions. These are skewed in favor of deletions in eukaryotes [12] and are expected to be selected against within MCSs; the result is a high MCS density in a proportion of short introns (Figure 1e and Figure S1a in the Supplementary Data online for mouse and human introns, respectively). It is worth noting that similar findings also apply to intergenic spacers (Figure 1f and Figures S1b, S1d and S2b in the Supplementary Data online).

Consistent with the above findings is the analysis of normalized size variation in human–mouse orthologous intron pairs [normalized length = (human length − mouse mouse length)/(human length + mouse length); (Figure 2g)]; lowess curves [13] indicate that the length of MCS-containing introns is extremely conserved irrespective of their size; conversely, the length of intronic regions lacking MCSs are remarkably conserved when shorter than 1 kb but diverge rapidly when longer. Stronger size conservation of MCS-containing introns than MCS-lacking introns also occurs over longer evolutionary periods, as shown by the analysis of human–chicken orthologous intron pairs (Figure 2h).

A role for gene expression in intron size evolution

These results indicate that MCSs have been imposing a strong constraint on intron size evolution. Nevertheless, <50% of human and mouse introns contain MCSs, and these findings do not rule out the possibility that factors other than MCSs, such as gene expression, might have had a role in shaping intron size. We therefore retrieved expression data for human and mouse genes; given that caution should be used when inferring expression rates from microarray data [14], we performed all analyses in parallel using both microarray and SAGE expression data. In line with previous findings [6], negative and significant, albeit extremely weak, correlations were identified between intron size and mean expression level (with microarray data, Spearman rho = −0.097 and −0.100 for human and mouse, respectively) or expression breadth (i.e. the number of tissues that express a given gene) (with the same data, Spearman rho = −0.0325 and −0.098 for human and mouse, respectively; the correlation coefficients obtained with SAGE data were very similar). Such weak correlations would usually deserve dismissal as not biologically significant; however, it has been previously suggested [3] that, especially in species with relatively small population sizes, the effects of selection might be detectable exclusively on highly expressed genes and therefore weak correlations might stem from the presence of a minority of cases accounting for the effect.

The lowess curves [13] in Figure 2a indicate that this might be the case: a decrease in human intron size is observed for expression levels above the 60th–80th percentile (depending on the experimental data used; in the Supplementary Data online see Figure S3 for SAGE and Figures S4 and S5 for mouse genes), whereas the curves are relatively flat for lower expression values. Conversely,

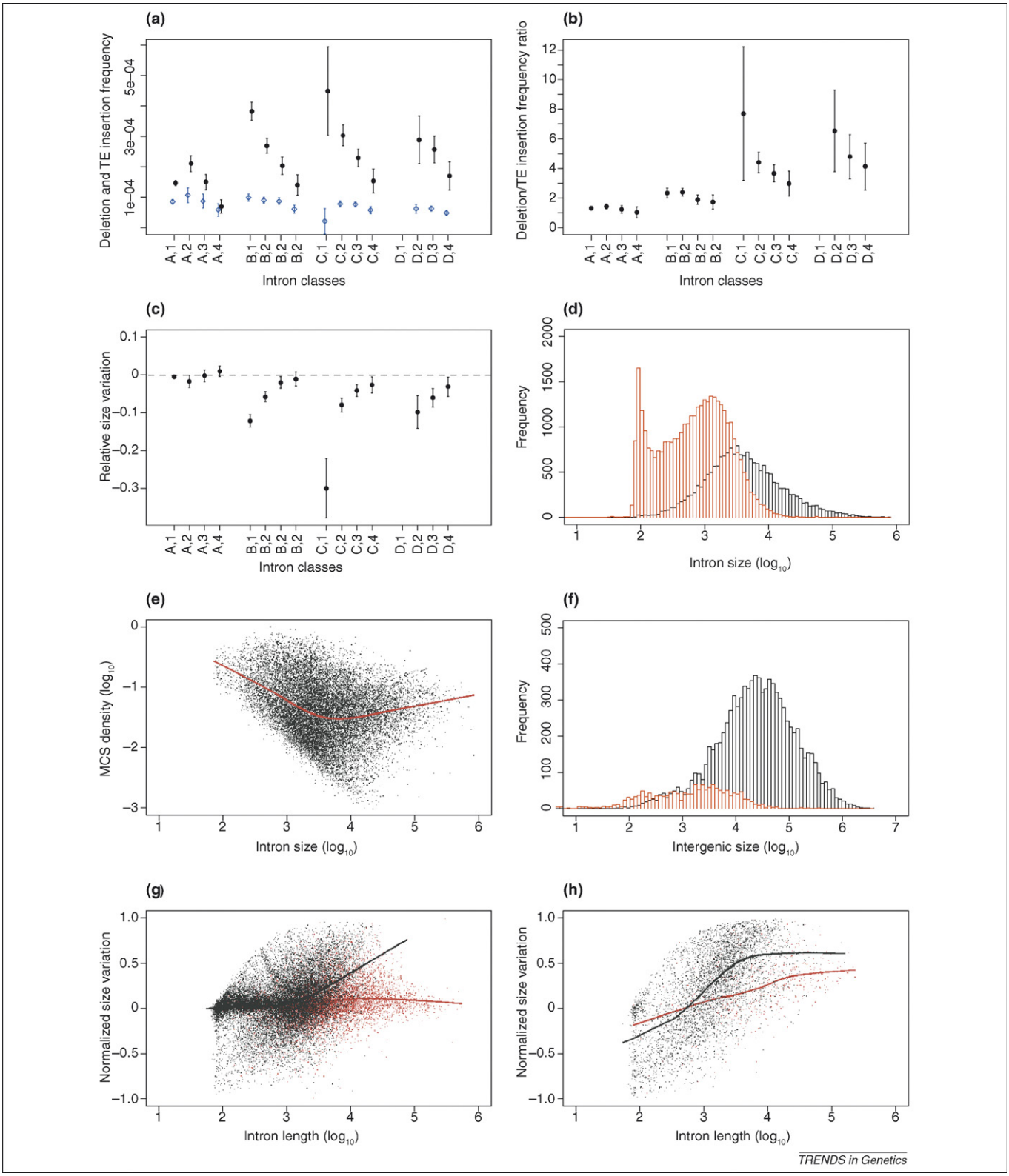

Figure 1. MCSs influence non-coding sequence length over time. **(a)** Mouse TE insertion (blue) and deletion frequencies (black) were calculated in four intron size and four MCS density classes (see methods in the Supplementary Data online). Class partition and MCS density calculation were performed using the inferred original intron size. Letters designate length classes (classes were separated at the following lengths in bp: 4507, 17 757 and 68 802) and numbers refer to MCS density classes (calculated as MCS length in bp divided by intron size; classes were separated at the following densities: 0.0008, 0.0239 and 0.0656). Frequencies (per intronic bp) are represented as mean values with 0.99 confidence intervals. Note that class D,1 (introns without MCSs and longer than 68 kb) contained only two introns and was therefore omitted. **(b)** Ratios between mouse deletion and insertion frequencies in introns of different length and MCS density classes. **(c)** Relative (to the mouse–rat common ancestor) mouse intron size variation in introns of different length and MCS density classes. The dashed line represents no size variation. **(d)** Histogram of mouse intron size (logarithmic scale) for MCS-lacking (red) and MCS-containing (black) introns. **(e)** Scatter plot and lowess smooth (red) of intronic MCS density against mouse intron size (logarithmic scales). **(f)** As (d) but for intergenic spacers. **(g)** The normalized size variation of 31 268 human–mouse orthologous intron pairs is plotted against human intron size (logarithmic scale). Lowess curves [10] were used to fit the scatter plots. For this analysis, only MCSs that are present in both human and mouse were considered (MCSs were identified as described in Ref. [7]). Red, MCS-containing introns; black, MCS-lacking introns. **(h)** Normalized size variation of MCS-containing (red) and MCS-lacking (black) human–chicken orthologous introns pairs ($n = 5055$). For this analysis, only MCSs that were fixed before the bird–mammalian split were considered [7].

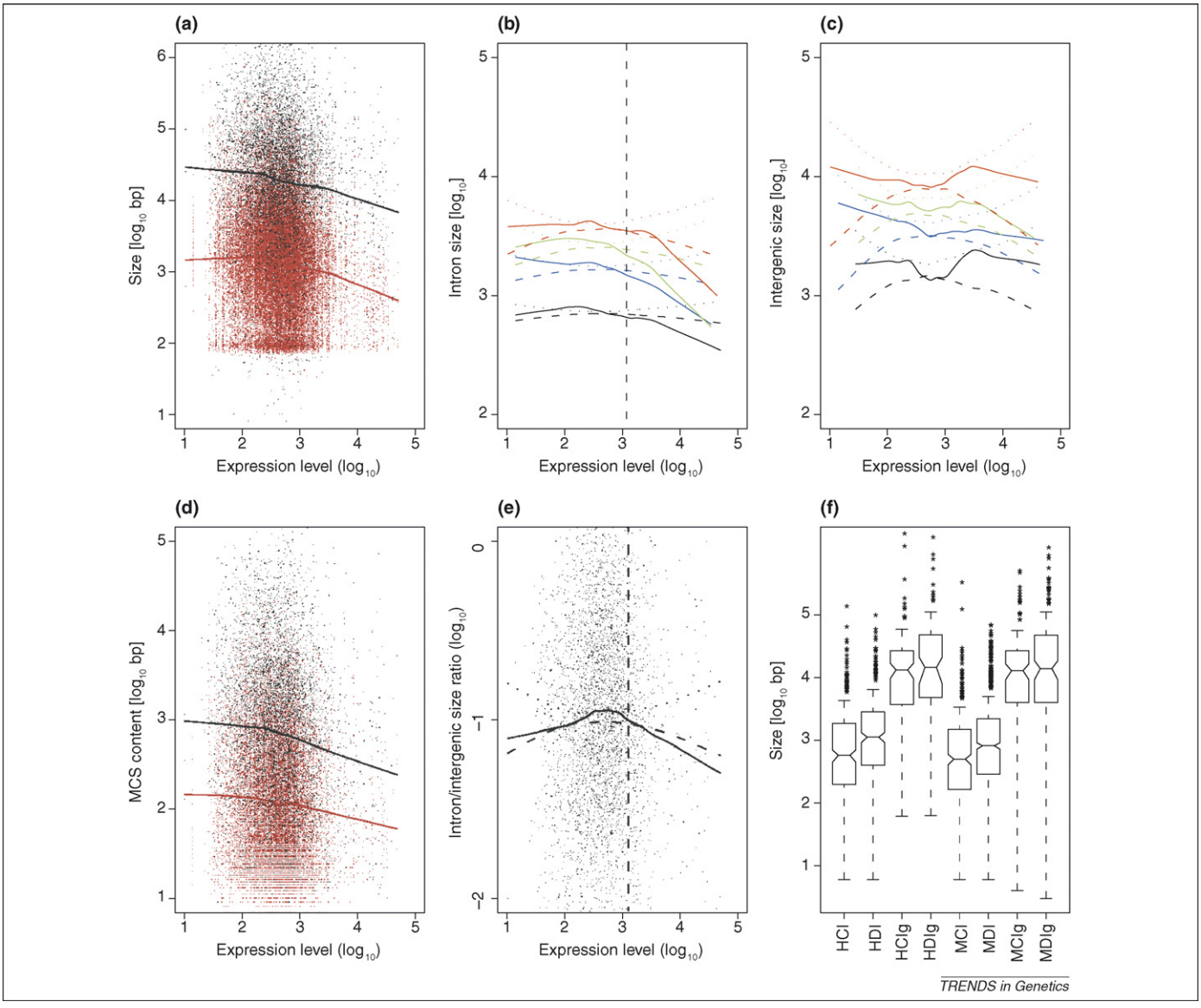

Figure 2. Gene-expression dependent variations of human intron and intergenic size. All expression data were derived from microarray experiments. **(a)** Scatter plot and lowess smooths of intron sequences (red) and intergenic sequences (black) versus gene expression. **(b)** Lowess curves [10] and relative probability intervals of intron length versus gene expression for introns containing no (black), one (blue), two (green) or three (red) MCSs. The 85thh expression percentile is marked by a vertical dashed line. **(c)** As (b) but for intergenic spacers. **(d)** Scatter plot and lowess fitting of intronic (red) and intergenic (black) conserved length (absolute number of MCS bases) versus gene expression. **(e)** Scatter smooth and lowess fitting of the ratio between intronic length and intergenic length versus gene expression. The 85th expression percentile is marked by a vertical dashed line. **(f)** Comparison of human and mouse introns and intergenic size from concordant and discordant genes. HCI, human concordant introns (n = 450); HCIg, human concordant intergenic spacers (n = 102); HDI, human discordant introns (n = 374); HDIg, human discordant intergenic spacers (n = 106); MCI, mouse concordant introns (n = 454); MCIg, mouse concordant intergenic spacers (n = 102); MDI, mouse discordant introns (n = 1016); MDIg, mouse discordant intergenic spacers (n = 226).

intergenic sequences show a steady size decrease with increasing expression (Figure 2a).

How can these findings be explained and reconciled with the demonstration of MCSs posing a constraint? To answer this question, we divided introns and intergenic sequences on the basis of their MCS content (the possibility of a sampling bias is addressed in the Supplementary Data online) and analyzed the variation of size with gene expression. The data are reported in Figures 2b, 2c and Figure S3 in the Supplementary Data online (Figures S4 and S5 for mouse genes) and indicate that, when MCS content is fixed, introns in highly expressed genes (expression levels higher than the 85th percentile) are significantly shorter (empirical $p = 0.01$) than their less expressed counterparts (see methods in the Supplementary Data online). The observed size decrease is not accounted for by highly expressed introns having, on average, shorter MCSs (Figure S6 in the Supplementary Data online). Conversely, curves corresponding to intergenic sequences show no expression-dependent variation if the MCS content is fixed (Figure 2c). As shown in Figure 2d, for both introns and intergenic spacers a marked decrease in MCSs (absolute number of conserved bases) is observed with increasing expression; this explains the average size decrease with increased expression and is the only effect acting on intergenic spacers. Conversely, data on intronic regions are consistent with the existence of an additional economizing pressure on highly expressed introns that combines with the other effect in an additive manner.

For further confirmation, we analyzed expression-dependent variations in intronic to intergenic size ratios. As previously noted [3], intronic and intergenic sequences can be conveniently compared in a pair-wise manner because their lengths and GC content are similar [7]. We also found the amount of sequence conservation (MCS number) to be highly correlated between the two (Spearman rho = 0.37, $p < 10^{-16}$). Moreover, comparison with intergenic spacers rules out the possibility that size differences are accounted for by local differences in mutation rates. In line with our hypothesis and with the data reported earlier, a significant (empirical $p = 0.01$) decrease in the ratio of intronic size to intergenic size is observed (Figure 2e) for highly expressed genes.

The parallel decrease of intron and intergenic sequence length with increased expression level had been one of the major criticisms of the selection-for-economy model [6]. All these analyses were also performed for mouse introns and intergenic sequences (see the figures in the Supplementary Data online), and gave similar results. Also, similar results were obtained when expression breadth or peak rate were used instead of mean expression level (not shown).

As we have hinted, postulating a role for gene expression in intron size evolution implies an assumption that expression patterns have been conserved for long enough for selection to act. The availability of extensive mouse and human data on gene expression enables insight into this issue. We assumed that orthologous genes that are highly expressed (a level higher than the 85th percentile with both microarrays and SAGE) in both human and mouse (concordant genes) have been experiencing at least 65–75 million years of expression-dependent selection [15,16]. According to our hypothesis, these genes should have shorter introns (but not shorter intergenic spacers) than discordant ones (genes that, from both microarray and SAGE data, are highly expressed in human but not in mouse or vice versa). Figure 2f indicates that this is the case: introns of concordant genes are significantly shorter than discordant introns in the respective species (Wilcoxon Rank sum test, $p < 10^{-6}$ and $p < 10^{-7}$ for human and mouse, respectively); conversely, no significant difference is observed between intergenic sequences.

Concluding remarks

The genomic-design model, as previously formulated [6–8], implies an adaptive interpretation of intron size increase in that it postulates that longer regions have a higher fraction of conserved sequence and mainly function to regulate and suppress complex tissue-specific genes. The data we report here support and extend this view. In particular, our model proposes that regulatory needs (accounted for by MCSs) shape intron size and tend to be stronger in genes that are not highly expressed. We also show that, when MCS content is fixed, no variation of size with expression level is observed for intergenic spacers and for introns in genes expressed at a medium–low level. We therefore propose that the fixation of functional conserved elements is the adaptive event underlying size increase. With respect to the more general, widely discussed issue of noncoding sequence proliferation in higher eukaryotes, our data challenge the neutralist view of genome expansion [17] and add insight into its adaptive interpretation [6–8]. However, noncoding DNA is composed of distinct functional categories and it is not surprising that additional selective forces act on introns, especially if they are actively transcribed.

We therefore propose a double-faceted model for intron size evolution in which the need to attain higher regulatory capacity (and therefore complexity) is balanced with energetic cost. The observed economizing pressure might also be accounted for by the merging of both energy and time economy (for genes that need rapid transcription), as recently suggested [18,19].

Acknowledgements

We thank R. Giorda for useful discussions about the article.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2006.10.003.

References

1 Hurst, L.D. *et al.* (1996) Imprinted genes have few and small introns. *Nat. Genet.* 12, 234–237
2 Castillo-Davis, C.I. *et al.* (2002) Selection for short introns in highly expressed genes. *Nat. Genet.* 31, 415–418
3 Urrutia, A.O. and Hurst, L.D. (2003) The signature of selection mediated by expression on human genes. *Genome Res.* 13, 2260–2264
4 Eisenberg, E. and Levanon, E.Y. (2003) Human housekeeping genes are compact. *Trends Genet.* 19, 362–365
5 Seoighe, C. *et al.* (2005) Gametophytic selection in *Arabidopsis thaliana* supports the selective model of intron length reduction. *PLoS Genet.* 1, e13
6 Vinogradov, A.E. (2004) Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.* 20, 248–253
7 Vinogradov, A.E. (2005) Noncoding DNA, isochores and gene expression: nucleosome formation potential. *Nucleic Acids Res.* 33, 559–563
8 Vinogradov, A.E. (2006) 'Genome design' model: evidence from conserved intronic sequence in human-mouse comparison. *Genome Res.* 16, 347–354
9 Dermitzakis, E.T. *et al.* (2005) Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat. Rev. Genet.* 6, 51–57
10 Boffelli, D. *et al.* (2004) Comparative genomics at the vertebrate extremes. *Nat. Rev. Genet.* 5, 456–465
11 Sironi, M. *et al.* (2005) Fixation of conserved sequences shapes human intron size and influences transposon-insertion dynamics. *Trends Genet.* 21, 484–488
12 Petrov, D.A. (2002) Mutational equilibrium model of genome size evolution. *Theor. Popul. Biol.* 61, 531–544
13 Cleveland, W. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* 74, 829–836
14 Draghici, S. *et al.* (2006) Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.* 22, 101–109
15 Madsen, O. *et al.* (2001) Parallel adaptive radiations in two major clades of placental mammals. *Nature* 409, 610–614
16 Murphy, W.J. *et al.* (2001) Molecular phylogenetics and the origins of placental mammals. *Nature* 409, 614–618
17 Lynch, M. and Conery, J.S. (2003) The origins of genome complexity. *Science* 302, 1401–1404
18 Chen, J. *et al.* (2005) Human antisense genes have unusually short introns: evidence for selection for rapid transcription. *Trends Genet.* 21, 203–207
19 Chen, J. *et al.* (2005) The small introns of antisense genes are better explained by selection for rapid transcription than by "genomic design". *Genetics* 171, 2151–2155

0168-9525/$ – see front matter
doi:10.1016/j.tig.2006.10.003