Opinion TRENDS in Genetics Vol.20 No.5 May 2004

Compactness of human housekeeping genes: selection for economy or genomic design?

Alexander E. Vinogradov

Institute of Cytology, Russian Academy of Sciences, Tikhoretsky Ave 4, St Petersburg 194064, Russia

The short length of human housekeeping genes, compared with tissue-specific genes, has been attributed to selection for economy in transcription and translation. In this article, I show that the non-transcribed intergenic spacer around housekeeping genes is also short, which suggests that short housekeeping genes are a result of local genome organization. Tissue-specific genes are longer than housekeeping genes because they have more functional domains, which is an indication of their more-complex protein architecture.

Recently, several papers have reported that highly and broadly expressed (i.e. expressed in many tissues) human genes are shorter, both in their intronic and coding sequences, than genes that are expressed in a tissue-specific fashion [1–3]. Because transcription and (particularly) translation are energetically costly, this shortness was interpreted as a result of selection for economy [1–3]. However, in *Saccharomyces cerevisiae* (and probably some other unicellular organisms) the highly expressed genes have longer introns compared with genes that have a lower expression level [4], contradicting this interpretation. In addition, a negative correlation between the gene expression level and the length of coding sequence (CDS) was not observed in *Arabidopsis thaliana* and *Drosophila melanogaster,* whereas positive correlation was found in *Caenorhabditis elegans* [5]. In the genomes of warm-blooded vertebrates, genes and intergenic sequences located in the GC-rich regions (called heavy isochores) are shorter than sequences located in the GC-poor regions (light isochores) [6–8]. Genes with a high level and breadth of expression (housekeeping genes) are on average more GC-rich [9–11]. The GC-rich sequences tend to be located in the central, open chromatin of the interphase nuclei of warm-blooded vertebrates, whereas the GC-poor sequences are located in the peripheral, more compact chromatin [12].

All these facts suggest another interpretation of the correlation between gene length and the level and breadth of expression in the human genome. They indicate that the length and GC content of intronic and intergenic sequences might be optimized for chromatin-mediated suppression and more-complex regulation of tissue-specific genes compared with housekeeping genes, resulting in a special epigenetic organization of the genome. (The suppression of tissue-specific genes is obviously not pertinent to yeast and other unicellular organisms.) This problem is also directly relevant to the long-standing 'C-value (genome size) paradox'. This paradox, after discovery that a significant part of eukaryotic genome consists of non-coding DNA (constituting intronic and intergenic sequences), can be redefined as whether a part of this non-coding DNA is related to organismal complexity. The greater length of CDS in the tissue-specific genes might not be a result of a weaker selection for economy but a consequence of the functional and regulatory complexity of tissue-specific proteins. For instance, when the proteome of multicellular eukaryotes (and particularly vertebrates, judging by the human genome) were compared with unicellular organisms, there was an expansion of families of multidomain proteins, which are generally longer and might be involved in multicellular organization (i.e. they are proteins involved in extracellular interactions) [8,13]. It is reasonable to suggest that these longer, multidomain proteins are mostly tissue-specific proteins. In this article, I build a case for the 'genomic design' interpretation.

Data acquisition

The data on the expression levels of human genes in normal tissues that were obtained from oligonucleotide microarray experiments (Affymetrix U95A; http://www.affymetrix.com/) were extracted from the Gene Expression Atlas (http://expression.gnf.org/cgi-bin/index.cgi) [14]. Only probes that represented characterized genes with links to the RefSeq database were used (http://www.ncbi.nlm.nih.gov/RefSeq/index.html), and signals from duplicate probes on the same chip and replicates representing the same tissue were averaged. This gave a total of 7708 genes and 32 tissues. A gene was regarded as expressed if its signal level exceeded a conservative threshold of 200 arbitrary units [14]. I found that there is a strong correlation (Spearman $r = 0.89$; $P < 10^{-8}$) between the expression breadth (number of tissues where a given gene is expressed) and the expression level (averaged over all tissues studied). The intronic and intergenic sequences (both upstream and downstream) were found for 6874 and 5104 genes, respectively. In cases of alternative-splicing variants the longest CDS was used. The average length

Corresponding author: Alexander E. Vinogradov (aevin@mail.cytspb.rssi.ru).

 0168-9525/$ - see front matter doi:10.1016/j.tig.2004.03.006

and GC content of upstream- and downstream-intergenic sequences were taken as the corresponding values of the local intergenic spacer for a given gene. [For >95% of the tested genes, the untranslated regions (UTRs) are known and therefore included in the genic sequence. It is possible that some unknown UTRs might occur in the intergenic sequence but they should constitute only a minor fraction of this sequence and thus, can not affect the results significantly.] For those genes that have links to the SwissProt database (http://us.expasy.org/sprot/), the number and size of the functional domains in the corresponding proteins were analyzed using the SwissPfam database [15] (5579 proteins were found).

For comparative purposes, the genomes of the nematode worm (*C. elegans*) and the fly (*D. melanogaster*) were studied. The data on nematode gene expression [experiments were conducted using oligonucleotide microarrays (Affymetrix) and gene expression in 12 stages of embryonic development was examined] were from Ref. [16]. The absolute calls (i.e. absence or presence) of gene expression provided by the authors were used. The intronic and intergenic sequences were found for 6248 genes. The data on *Drosophila* gene expression were taken from Ref. [17]; the authors present the results of experiments using oligonucleotide microarrays (Affymetrix) and examined gene expression at 34 categories (stages) of embryonic development and at 53 categories of experimental conditions in adult flies. Because the most significant changes in gene expression were observed between the embryonic and adult categories [17], I composed the groups of categories for analysis taking this into account. Because the authors did not provide absolute calls, I processed their raw data similar to the human dataset in the Gene Expression Atlas (http://expression.gnf.org/cgi-bin/index.cgi) [14] and used the same 200-unit threshold for decisions on gene expression. Intronic and intergenic sequences were found for 11 653 genes.

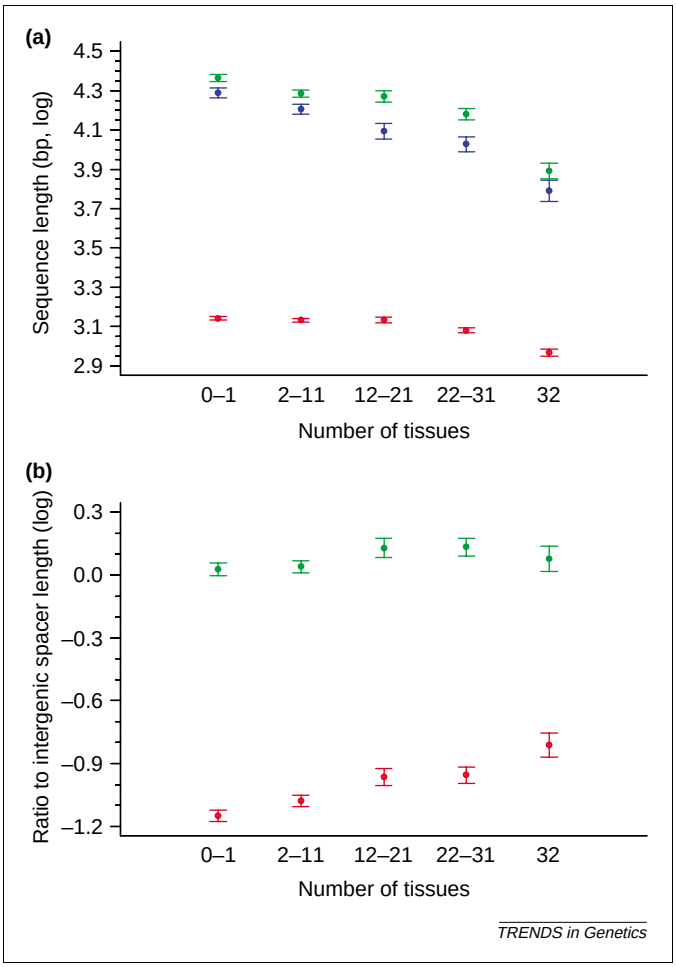

Figure 1. The structural genomic parameters of genes expressed in a different number of human tissues. **(a)** The lengths of coding (red), intronic (green) and intergenic (blue) sequences. **(b)** The ratio of coding (red) and intronic (green) sequence length to the local intergenic sequence length. The mean values with LSD (least significant differences) intervals are shown (ANOVA and Kruskal-Wallis; for intronic to intergenic sequence ratio $P < 0.05$ but there was inconsistent direction of differences, for all other cases $P < 10^{-8}$).

Housekeeping genes are short – but so are the intergenic spacers

Both the CDS and intronic sequence (intervening sequence; IVS) of broadly expressed human genes are shorter than those in tissue-specific genes (Figure 1a). A similar picture was observed with the UTR length (not shown). However, the intergenic spacers (which are presumably not transcribed) show the same, or an even more regular, trend (Figure 1a). If gene length is expressed as a ratio to local intergenic spacer length, the negative correlation between gene length and expression breadth changes direction in the case of CDS (and UTRs; not shown) and disappears in the case of IVS (Figure 1b). Similarly, the negative correlation between gene length and expression level (averaged over all tissues) changes sign in the case of CDS and UTR lengths and disappears in the case of IVS length, if they are expressed as ratios to the local intergenic spacer length (Table 1). This indicates that the length of intronic sequence is proportional to the length of intergenic sequence (which is reminiscent of the 'among-genomic' proportionality between the amounts of intragenic and intergenic non-coding DNA [18–20]). It seems unlikely that the shortness of intergenic spacers around the housekeeping genes is a secondary trait (byproduct) caused by the widespread expression of these genes. (Transposable elements might integrate more frequently in the intergenic sequences that are located in the decondensed chromatin neighboring the transcribed genes, thereby increasing the length of their intergenic spacers.)

Table 1. Correlation between structural genomic parameters and gene expression level averaged over all tissues studied (all parameters are log-transformed)[a]

Length parameter	Correlation coefficient
CDS	-0.14 ($P < 10^{-6}$)
UTRs	-0.11 ($P < 10^{-6}$)
Intronic sequence	-0.22 ($P < 10^{-6}$)
Intergenic spacer	-0.16 ($P < 10^{-6}$)
CDS and intergenic spacer	0.09 ($P < 10^{-5}$)
UTRs and intergenic spacer	0.04 ($P < 0.01$)
Intronic and intergenic spacer	Not significant
Ratio of total non-coding to CDS	-0.14 ($P < 10^{-6}$)
Percent of protein length covered by domains	Not significant

[a]Abbreviations: CDS, coding sequence; UTRs, untranslated regions.

Gene nests

It was long argued that introns (and intergenic sequences) might be necessary for correct chromatin structure [21,22]. The nucleotide-sequence signatures indicate that the nucleosome-formation sites are stronger in introns compared with those in exons [23–26]. It was shown in several cases that after removing the introns experimentally, genes lose the ability to form nucleosomes [27,28]. Sequences involved in gene-specific regulation are sometimes found in introns [29]. The intergenic DNA is also important for gene regulation: enhancers, silencers and insulators of human tissue-specific genes are located in the intergenic sequences upstream and downstream of a gene and can work over distances of >100 kb [29]. It was also shown that intergenic sequences participate in the chromatin-mediated 'sectorial' repression that involves blocks of genes [30–33]. A greater amount of intragenic and intergenic non-coding DNA, in which the tissue-specific genes are embedded ('gene nest'), might facilitate suppression of gene activity in most tissues by the shear bulk of condensed chromatin and might also contain more gene-specific regulatory elements.

It is noteworthy that among strictly tissue-specific genes (i.e. those expressed in none or one tissue of the dataset), there is a weak positive correlation between expression level and length parameters (CDS, not significant; UTRs, $r = 0.08$, $P < 0.01$; intronic sequences, $r = 0.11$, $P < 10^{-4}$; intergenic spacer, $r = 0.08$, $P < 0.01$; ratio of non-coding to coding DNA, $r = 0.11$, $P < 10^{-4}$). This does not support the economy explanation. From the epigenetic standpoint of chromatin-mediated gene suppression, the strictly tissue-specific genes with a higher expression should be more strongly suppressed in other tissues (where they are not to be expressed), which can explain a greater amount of non-coding DNA surrounding them. This suggestion is supported by the negative correlation between expression level and GC content in the strictly tissue-specific genes (CDS, $r = -0.23$, $P < 10^{-6}$; third codon position, $r = -0.25$, $P < 10^{-6}$; intronic sequences, $r = -0.21$, $P < 10^{-6}$; intergenic sequences, $r = -0.16$, $P < 10^{-5}$). By contrast, for the whole dataset, there is a positive correlation between the level and breadth of gene expression and GC content [11]. According to the epigenetic hypothesis put forward here, this is because, in general, the highly and broadly expressed genes should be located in the more open chromatin.

GC content

If GC content is added as covariate into multifactor ANOVA or general linear model (GLM), the correlation between the ratio of non-coding to coding DNA lengths and the breadth of gene expression disappears (Figure 2). It is noteworthy that the GC content of both the intronic and the intergenic sequences was used for the correlation to disappear completely (Figure 2). Both parameters of GC content participate in the model independently (at least partially), each are highly significant ($P < 10^{-4}$) and have a negative sign (i.e. the higher the GC content, the lower the ratio of non-coding to CDS lengths). This fact suggests that the GC content of intron and intergenic sequences taken together is a better reflection of the isochore

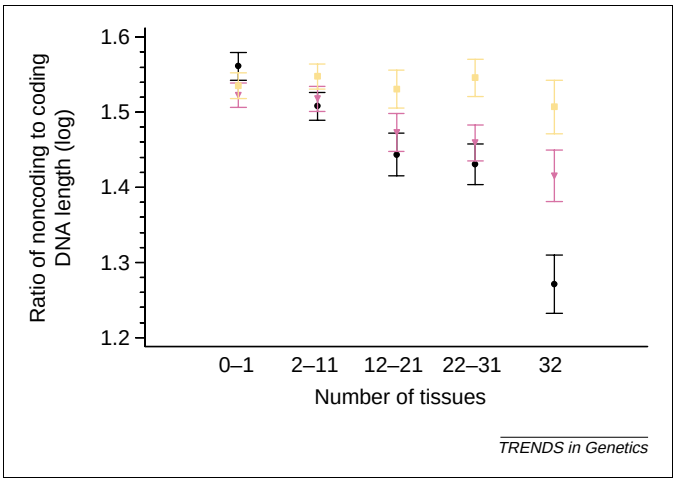

Figure 2. The ratio of non-coding (i.e. intronic plus the local intergenic sequence) to coding sequence lengths in genes expressed in a different number of human tissues. The mean values with LSD (least significant differences) intervals are shown: no correction (black circles; ANOVA and Kruskal-Wallis, $P < 10^{-8}$); corrected in multifactor ANOVA or general linear model (GLM) for intronic GC content (purple triangles; ANOVA and GLM, $P < 10^{-3}$); corrected both for intronic and intergenic GC content (yellow squares; ANOVA and GLM, $P > 0.6$). (If the untranslated regions are included in the non-coding DNA, the results are similar.)

affiliation of a given gene (i.e. the regional effect) or that the local variation of GC content among intronic and intergenic sequences is also an important property of the 'gene nest'. In any case, it is clear that the length and the GC content of non-coding DNA are closely intertwined in their effects on gene expression.

Tissue-specific proteins are longer but not because of 'junk'

The length of CDS is special because protein sequences should be under stronger selective pressure for function than the more dispensable non-coding DNA of introns. But translation is more costly than transcription [34] and presents an addition to transcriptional cost, with each mRNA molecule being translated many times. Therefore, the suggestion that economy selection on the proteins of broadly expressed genes would shorten their CDS is not unreasonable. However, instead of the coding regions of housekeeping genes being shorter for reasons of economy, another explanation is possible: tissue-specific proteins might be longer because of their functional and regulatory complexity. Were the tissue-specific proteins under a relaxed selection for economy, functionally significant regions would cover a lower percent of their length. This is true: functional domains cover a higher percent of length in housekeeping proteins than in tissue-specific proteins (Figure 3a). However, this difference is not sufficient (~3%) to explain the ~1.5-fold increase in the average length of CDS in tissue-specific genes (Figure 1a). At the same time, the increase in the number of domains correlates with the increase in CDS length: it is also ~1.5-fold (Figure 3b). The same trend is observed if only the unique domains are counted (Figure 3b), which shows that there is not just a repetition of domains but that the diversity of domains is also higher in tissue-specific proteins. There is no correlation between the average expression level and the percentage of protein length covered by domains (Table 1), which does not support the

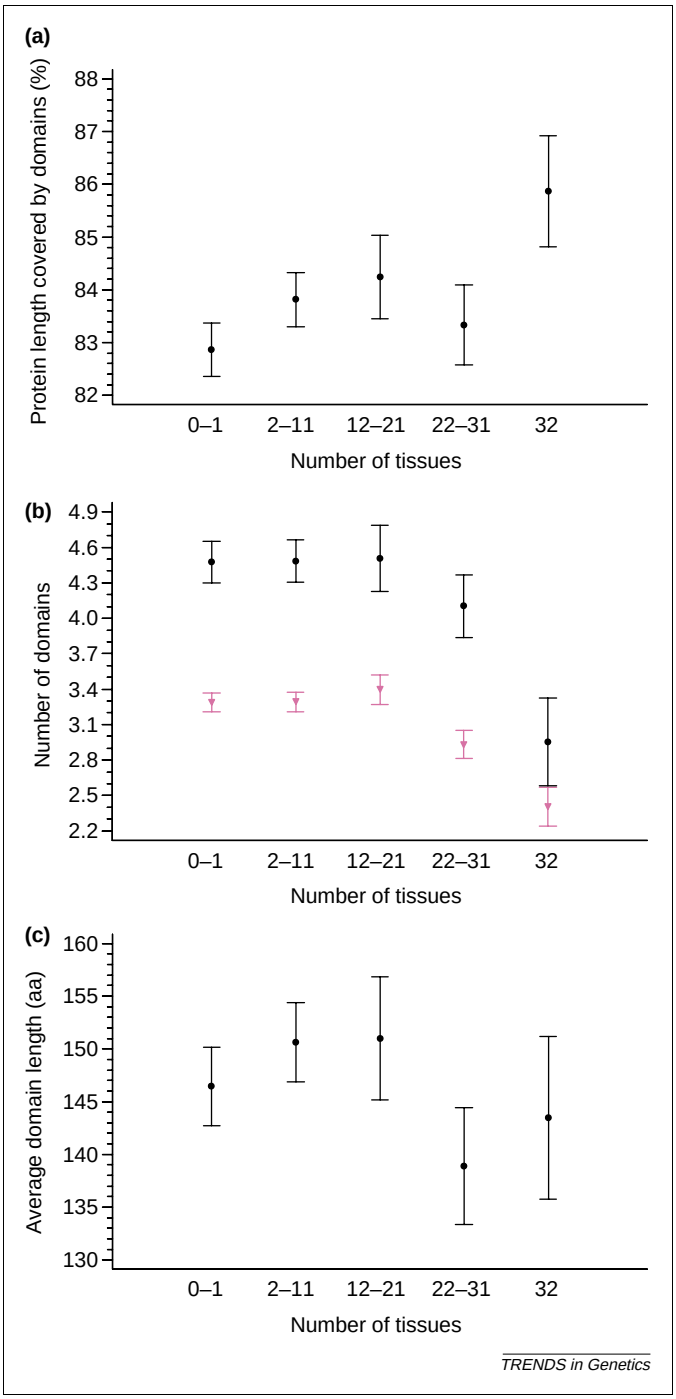

Figure 3. The functional domains in proteins expressed in a different number of human tissues (according to the SwissProt and Pfam databases). The mean values with LSD (least significant differences) intervals are shown. **(a)** The percentage of protein length that is covered by domains (ANOVA and Kruskal-Wallis, $P < 0.01$). **(b)** The number of domains (black circles) and number of unique domains (purple triangles) in a protein (ANOVA and Kruskal-Wallis, in both cases $P < 10^{-8}$). **(c)** The average domain length (ANOVA and Kruskal-Wallis, $P > 0.1$).

economy hypothesis. Although domains in the housekeeping proteins appear to be slightly shorter, this difference is not statistically significant (Figure 3c). These facts show that the higher length of tissue-specific proteins is due to their higher number of (diverse) domains, which is an indication of their more-complex functional architectures.

Multicellular invertebrates

In the nematode, the CDS is longer in broadly expressed genes (Figure 4a); this was shown previously using transcript abundances in the EST database (http://www.ncbi.nlm.nih.gov/dbEST/) [5]. This observation does not concur with the 'economy' hypothesis. The IVS are shorter in broadly expressed genes (Figure 4a), which confirms the previous observation [1], using a more recent experimental dataset [16]. However, the intergenic spacers of broadly expressed genes show even a more regular and stronger length reduction (Figure 4a). As a consequence, the ratio of non-coding to coding DNA lengths ('gene nest' proportion) in the nematode genome negatively correlates with the breadth of gene expression (Figure 4b), similar to the picture observed in the human genome.

In *Drosophila*, the CDSs are shorter in the most broadly expressed (strictly housekeeping) genes (Figure 5a), whereas the IVS show a general trend for reduction, which is not consistent (Figure 5a). The intergenic spacers show a regular stepwise reduction and seem to approach a plateau in the strictly housekeeping genes (Figure 5a). This plateau is probably related to the fact that *Drosophila* has one of the smallest genomes among insects. Therefore, the intergenic distances in this genome might approach

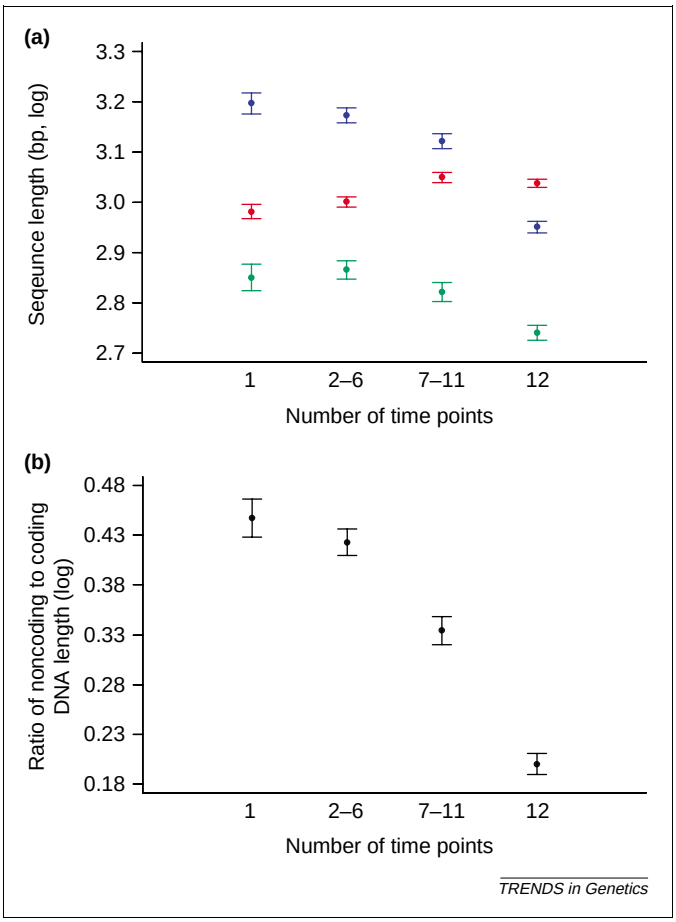

Figure 4. The structural genomic parameters of genes expressed in the nematode. **(a)** The lengths of coding (red), intronic (green) and intergenic (blue) sequences and **(b)** the ratio of non-coding (intronic plus local intergenic sequence) to coding sequence lengths in genes expressed at a different number of time points during *Caenorhabditis elegans* embryonic development. The mean values with LSD (least significant differences) intervals are shown (ANOVA and Kruskal-Wallis, in all cases $P < 10^{-8}$).

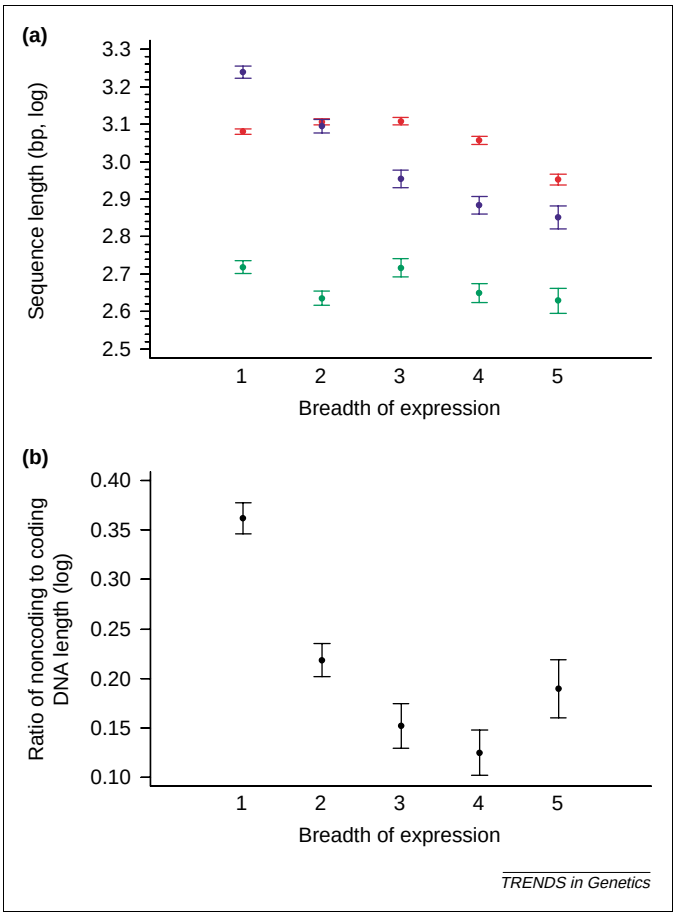

Figure 5. The structural genomic parameters of genes expressed in the fly. **(a)** The lengths of coding (red), intronic (green) and intergenic (blue) sequences and **(b)** the ratio of non-coding (intronic plus local intergenic sequence) to coding sequence lengths in genes expressed at different stages of embryonic development (time points) and experimental conditions with adults (*Drosophila melanogaster*). 1 – expressed in 0–1 stages of embryonic development or in the adult; 2 – expressed in 0–1 stages of embryonic and many (2–53) conditions in the adult or in 0–1 experimental conditions in the adult and many (2–34) stages of embryonic development; 3 – expressed in many stages in embryonic development (2–34) and in adult (2–53); 4 – expressed in all stages of either embryonic development (34) or in the adult (53); 5 – expressed in all stages of embryonic development (34) and adult (53). The mean values with LSD (least significant differences) intervals are shown (ANOVA and Kruskal-Wallis, for introns $P < 10^{-5}$, for other sequences $P < 10^{-8}$).

their lower limit for a given organization level. This plateau, together with the fact that strictly housekeeping proteins of *Drosophila* are short, compared with less broadly expressed proteins (Figure 5a), can explain some elevation of the ratio of non-coding to CDS lengths in strictly housekeeping genes (Figure 5b).

Thus, although CDS and IVS of these multicellular invertebrates do not always change consistently with an increase of gene-expression breadth, in the genomes of *Drosophila* and *C. elegans*, the intergenic spacers show a regular, stepwise length reduction, which is in agreement with the 'genomic design' interpretation.

Concluding remarks

In summary, the previously reported negative correlation between gene length and expression breadth in the human genome disappears, if gene length is expressed as a ratio to the length of the local intergenic spacer, which indicates the within-genomic proportionality between the amounts of intragenic and intergenic non-coding DNA in regard to gene expression. As a result, the ratio of non-coding (introns plus intergenic spacer) to CDS length ('gene nest' proportion) negatively correlates with expression breadth. In addition, human tissue-specific proteins contain a higher number of functional domains than broadly expressed proteins. These facts undermine the 'economy' explanation and provide support for the 'genomic design' interpretation of the negative correlation between gene length and expression breadth. The greater amount of intragenic and intergenic non-coding DNA, in which the tissue-specific genes are embedded, might be related to chromatin-mediated suppression and more-complex regulation of tissue-specific genes, whereas the greater length of their CDS is probably due to more-complex functional architectures of tissue-specific proteins.

The 'economy' hypothesis implicitly assumes the neutralist (permissive) interpretation of the accumulation of non-coding DNA in the eukaryotic genome: the incessantly transcribed genes are supposed to 'slim down' (selection condition), whereas those that work less intensively 'get fat' (permissive condition). Therefore, the rejection of the 'economy' hypothesis suggests an adaptive interpretation of the evolution of non-coding DNA. It does not mean that all non-coding DNA in the human genome is 'epigenetic' (i.e. related to the suppression of non-housekeeping genes). It could be, partly, a result of the activity of transposable elements – behaving as selfish darwinian units – although their propagation is not always adaptive for the organism [35–37]. It could also have a role of nuclear 'skeleton', regulating the nuclear-cytoplasmic ratio and cell size [38] or perform a 'buffering' function (i.e. an energy-independent attenuation of fluctuations in the intracellular medium) [39]. The 'buffering' function might be intertwined with the 'epigenetic' function. It was shown that a few seconds after cell membrane damage, the chromatin of vertebrate cells undergoes drastic decondensation [40]. The process is governed, at least in part, by a difference in the extracellular versus intracellular solute composition, the variation of which represents a permanent challenge to the chromatin-condensation state [40]. Therefore, the higher 'gene nest' proportion of tissue-specific genes might secure their suppression because a bulk of condensed chromatin mass surrounds the gene, which should buffer the surges of chromatin decondensation caused by fluctuations of solute composition, thereby reducing transcriptional noise.

Acknowledgements

I thank three anonymous referees and the editor for helpful comments. This work was supported by the Russian Foundation for Basic Research (RFBR) and by the Programme of the Presidium of the Russian Academy of Sciences 'Molecular and Cellular Biology'.

References

1 Castillo-Davis, C.I. *et al.* (2002) Selection for short introns in highly expressed genes. *Nat. Genet.* 31, 415–418
2 Eisenberg, E. and Levanon, E.Y. (2003) Human housekeeping genes are compact. *Trends Genet.* 19, 362–365
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 Vinogradov, A.E. (2001) Intron length and codon usage. *J. Mol. Evol.* 52, 2–5

5 Duret, L. and Mouchiroud, D. (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4482–4487
6 Duret, L. *et al.* (1995) Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J. Mol. Evol.* 40, 308–317
7 Bernardi, G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene* 241, 3–17
8 International Human Genome Sequencing Consortium, (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
9 Mouchiroud, D. *et al.* (1987) Compositional compartmentalization and gene composition in the genome of vertebrates. *J. Mol. Evol.* 26, 198–204
10 Pesole, G. *et al.* (1999) Isochore specificity of AUG initiator context of human genes. *FEBS Lett.* 464, 60–62
11 Vinogradov, A.E. (2003) Isochores and tissue-specificity. *Nucleic Acids Res.* 31, 5212–5220
12 Saccone, S. *et al.* (2002) Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. *Gene* 300, 169–178
13 Rubin, G.M. *et al.* (2000) Comparative genomics of the eukaryotes. *Science* 287, 2204–2215
14 Su, A.I. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* 99, 4465–4470
15 Bateman, A. *et al.* (2000) The Pfam protein families database. *Nucleic Acids Res.* 28, 263–266
16 Baugh, L.R. *et al.* (2003) Composition and dynamics of the *Caenorhabditis elegans* early embryonic transcriptome. *Development* 130, 889–900
17 Spellman, P.T. and Rubin, G.M. (2002) Evidence for large domains of similarly expressed genes in the Drosophila genome. *J. Biol.* 1, 5
18 Deutsch, M. and Long, M. (1999) Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res.* 27, 3219–3228
19 Vinogradov, A.E. (1999) Intron-genome size relationship on a large evolutionary scale. *J. Mol. Evol.* 49, 376–384
20 McLysaght, A. *et al.* (2000) Estimation of synteny conservation and genome compaction between pufferfish (Fugu) and human. *Yeast* 17, 22–36
21 Zuckerkandl, E. (1981) A general function of noncoding polynucleotide sequences. Mass binding of transconformational proteins. *Mol. Biol. Rep.* 7, 149–158
22 Trifonov, E.N. (1993) Spatial separation of overlapping messages. *Comput. Chem.* 117, 27–31
23 Kiyama, R. and Trifonov, E.N. (2002) What positions nucleosomes?–A model. *FEBS Lett.* 523, 7–11
24 Denisov, D.A. *et al.* (1997) Protective nucleosome centering at splice sites as suggested by sequence-directed mapping of the nucleosomes. *Gene* 205, 145–149
25 Wada-Kiyama, Y. *et al.* (1999) DNA bend sites in the human beta-globin locus: evidence for a basic and universal structural component of genomic DNA. *Mol. Biol. Evol.* 16, 922–930
26 Levitsky, V.G. *et al.* (2001) Nucleosome formation potential of exons, introns, and *Alu* repeats. *Bioinformatics* 17, 1062–1064
27 Lauderdale, J.D. and Stein, A. (1992) Introns of the chicken ovalbumin gene promote nucleosome alignment *in vitro*. *Nucleic Acids Res.* 20, 6589–6596
28 Liu, K. *et al.* (1995) Rat growth hormone gene introns stimulate nucleosome alignment *in vitro* and in transgenic mice. *Proc. Natl. Acad. Sci. U. S. A.* 92, 7724–7728
29 Levine, M. and Tjian, R. (2003) Transcription regulation and animal diversity. *Nature* 424, 147–151
30 Zuckerkandl, E. (1997) Junk DNA and sectorial gene repression. *Gene* 205, 323–343
31 Zuckerkandl, E. (1999) Sectorial gene repression in the control of development. *Gene* 238, 263–276
32 Farkas, G. *et al.* (2000) Chromatin organization and transcriptional control of gene expression in drosophila. *Gene* 253, 117–136
33 Zuckerkandl, E. (2002) Why so many noncoding nucleotides? The eukaryote genome as an epigenetic machine. *Genetica* 115, 105–129
34 Hulbert, A.J. and Else, P.L. (2000) Mechanisms underlying the cost of living in animals. *Annu. Rev. Physiol.* 62, 207–235
35 Doolittle, W.F. and Sapienza, C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284, 601–603
36 Orgel, L.E. and Crick, F.H. (1980) Selfish DNA: the ultimate parasite. *Nature* 284, 604–607
37 Vinogradov, A.E. (2003) Selfish DNA is maladaptive: evidence from the plant Red List. *Trends Genet.* 19, 609–614
38 Cavalier-Smith, T. (1985) Cell volume and the evolution of eukaryote genome size. In *The Evolution of Genome Size* (Cavalier-Smith, T., ed.), pp. 105–184, John Wiley & Sons
39 Vinogradov, A.E. (1998) Buffering: a possible passive-homeostasis role for redundant DNA. *J. Theor. Biol.* 193, 197–199
40 Vinogradov, A.E. (1995) Cell membrane-dependent chromatin condensation. *Cytometry* 19, 183–188