5906–5914 *Nucleic Acids Research, 2006, Vol. 34, No. 20* *Published online 24 October 2006*
doi:10.1093/nar/gkl773

'Genome design' model and multicellular complexity: golden middle

Alexander E. Vinogradov*

Institute of Cytology, Russian Academy of Sciences, St Petersburg 194064, Russia

Received August 1, 2006; Revised September 13, 2006; Accepted September 28, 2006

ABSTRACT

Human tissue-specific genes were reported to be longer than housekeeping genes (both in coding and intronic parts). The competing neutralist and adaptationist models were proposed to explain this observation. Here I show that in human genome the longest are genes with the intermediate expression pattern. From the standpoint of information theory, the regulation of such genes should be most complex. In the genomewide context, they are found here to have the higher informational load on all available levels: from participation in protein interaction networks, pathways and modules reflected in Gene Ontology categories through transcription factor regulatory sets and protein functional domains to amino acid tuples (words) in encoded proteins and nucleotide tuples in introns and promoter regions. Thus, the intermediately expressed genes have the higher functional and regulatory complexity that is reflected in their greater length (which is consistent with the 'genome design' model). The dichotomy of housekeeping versus tissue-specific entities is more pronounced on the modular level than on the molecular level. There are much lesser intermediate-specific modules (modules overrepresented in the intermediately expressed genes) than housekeeping or tissue-specific modules (normalized to gene number). The dichotomy of housekeeping versus tissue-specific genes and modules in multicellular organisms is probably caused by the burden of regulatory complexity acted on the intermediately expressed genes.

INTRODUCTION

Human tissue-specific genes were reported to be longer than housekeeping genes, both in coding and intronic parts. The competing models were proposed to explain this observation: selection for economy (in housekeeping genes), mutation bias and 'genome design' (i.e. functional complexity) (1–11). The first two models assume a neutralist (permissive) interpretation of the accumulation of DNA in eukaryotic genomes. In contrast, the 'genome design' model suggests that the length of genomic elements is mostly determined by their functional load. In particular, the greater amount of intra- and intergenic noncoding DNA, in which the tissue-specific genes are embedded, may be involved in the more complex regulation and chromatin-mediated suppression of these genes, whereas the greater length of coding sequences may be related to more complex protein functional architectures. From the standpoint of information theory, the regulation of intermediately expressed genes should be most complex (Figure 1). Here, I test this suggestion and investigate in the genomewide context both the length (using the updated databases and a finer expression breadth scale) and the informational load (using a novel approach) of human genes with different expression pattern. The informational load is extensively studied on all available levels: from gene participation in protein interaction networks, pathways and modules reflected in Gene Ontology categories through transcription factor regulatory sets and protein functional domains to amino acid tuples (words of fixed size) in encoded proteins and nucleotide tuples in introns and promoter regions.

MATERIALS AND METHODS

Gene sequences and expression

Human gene sequences were extracted from the RefSeq database (12). The data on gene expression were taken from the last version of Gene Expression Atlas standardized with the MAS5 algorithm (13). They present the results of oligonucleotide microarray experiments performed uniformly with 72 normal human tissues. The signals from probes on the chip corresponding to the same gene were averaged; the replicates representing the same tissue were also averaged. As recommended (13), a gene was regarded as expressed if its signal level exceeded the dataset median. The complete sets of structural and expression data were obtained for 15 726 genes. The genes were divided into seven groups (bins) differing in the among-tissues expression breadth, with a roughly equal number of genes in each group. In the part

*Tel.: +78 122975310; Fax: +78 122970341; Email: aevin@mail.cytspb.rssi.ru

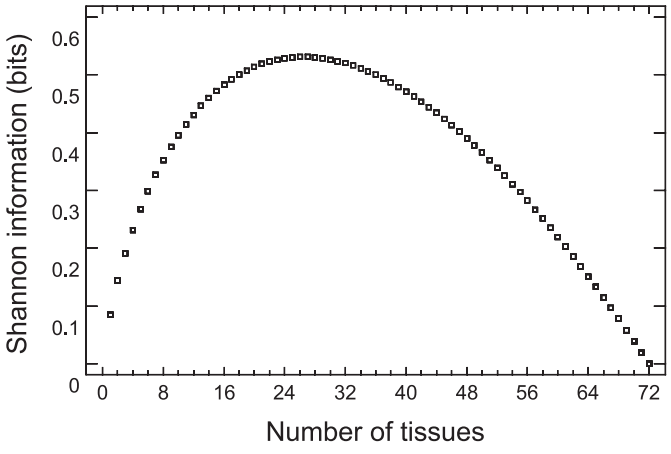

Figure 1. The information (uncertainty) of switch-on/off transition in genes expressed in different numbers of tissues, according to the Shannon formula ($-P*\log_2 P$). (Probability of expression is defined as the ratio of the number of tissues where a given gene is expressed to the total number of tissues studied.)

of analyses of intronic sequences, the gene groups were normalized to a roughly equal number of nucleotide tuples by randomly removing genes from the groups with a relatively greater total intron length. For analysis of intronic sequence, only the sum of internal introns (that reside within the coding sequence) was taken for consistency (because the complete mRNAs may not be known for all genes). There were 14 470 intron-containing genes in the dataset. In a part of analyses, introns were masked for lineage-specific repeats (that were inserted after the human–mouse split) or for all known repeats using the standalone RepeatMasker and DateRepeats programs (A.F.A. Smit, R. Hubley and P. Green; http://repeatmasker.org).

Genomic objects

For genes with references to the SwissProt (UniProt) database (13 273 proteins were found), the functional domains in the encoded proteins were estimated using the SwissPfam (for non-overlapping domains) and InterPro (the compilation of all known domain definitions from different databases, with redundancy) databases (14,15). The sets of genes regulated by different transcription factors were taken from the Molecular Signature Database (MsigDb) (16). The pathway gene sets were compiled using the KEGG (17) and Reactome (18) databases [using Entrez Gene mapping (19)], and HumanCyc (20). In the case of Gene Ontology categories (21), I collected for each category all its subcategories (separately for Biological Processes, Molecular Functions and Cellular Components) using GO graphs, and a gene was regarded as belonging to a given category if it was mapped to any of its subcategories in Entrez Gene. (If only the explicit Entrez Gene mapping of a given gene was used, the picture was similar.) The information on protein interactions was taken from the STRING database (22). All pairwise interactions of a given protein were taken to form the protein interaction set. The gene promoter regions were extracted from the database of experimentally determined exact transcriptional start sites (DBTSS): from 1000 nt upstream to 200 nt downstream of transcription start site (the standard promoter region length presented in the DBTSS) (23). The frequencies of amino acid and nucleotide tuples of different sizes were calculated using a sliding frame of a given size (with 1-letter step) for each gene group. The reduced amino acid alphabets were taken from the work by Li *et al*. (24). Similar to reduced amino acid alphabets, the more evolutionarily stable 2-letter purine/pyrimidine alphabet was used for testing intronic tuples. Thus, using repeats as markers of intronic sequence, it was estimated in regard to repeat ancestor copies (using the RepeatMasker program) that transitions (i.e. mutations from purine to purine or from pyrimidine to pyrimidine) occur roughly twice more frequently than transversions (i.e. mutations from purine to pyrimidine or vice versa).

Estimation of information

The Shannon information (uncertainty) was estimated on the basis of probability of occurrence of a given object (protein domain, transcription factor, pathway, protein interaction, GO category) in a given gene expression group in regard to the total dataset (i.e. the under- or overrepresentation of a given object in the total dataset), using a gene set corresponding to this object and the hypergeometric probability distribution. In other words, the expected count (number of occurrences) of the genomic object in a given gene expression group was estimated on the ground of the count of this object in the total dataset. Then, the probability of the deviation of the observed count from the expected was estimated using the hypergeometric test. If an object was overrepresented in a given group, the probability of equal or higher frequency was taken, if underrepresented, the probability of equal or lower frequency. Only those objects were taken that occur more than thrice in the total dataset (with the higher cutoff values, the picture was similar). This condition gives 1176 InterPro domains, 615 transcription factor sets, 274 pathways, 11 397 protein interaction sets, 1612 GO Biological Processes, 1034 GO Molecular Functions, and 358 GO Cellular Components (with the explicit Entrez Gene mapping, there were 710 Biological Processes, 634 Molecular Functions and 224 Cellular Components). For amino acid and nucleotide tuples (where there were much higher counts), the probability of occurrence of each tuple in a given gene group (in regard to the total dataset) was estimated using the chi-square distribution (with Yates correction).

The information (uncertainty) of each genomic object was calculated using the Shannon formula ($-P * \log_2 P$) (25), where the probability value was taken either from hypergeometric or chi-square test (as said above). Then the average information was determined for each gene expression group, summing the information across the entire set of objects of a given type (e.g. GO Biological Processes) and dividing it by the number of objects in the set.

For revealing the number of over- and underrepresented modules (Tables 1 and 2), I used hypergeometric probability distribution (as said above), and then estimated false discovery rate (q-value) for correction for multiple comparisons using P-value (obtained in hypergeometric test) and the

Table 1. The number of Gene Ontology categories overrepresented in genes expressed in different number of tissues (with correction for multiple tests made using estimation of false discovery rate, q-value < 0.05)

	0–5 tissues	6–18 tissues	19–37 tissues	38–54 tissues	55–66 tissues	67–71 tissues	72 tissues
GO Biological Processes	32	7	0	0	0	7	101
GO Molecular Functions	24	0	0	1	0	0	64
GO Cellular Components	15	6	3	0	0	16	69

Table 2. The number of Gene Ontology categories underrepresented in genes expressed in different number of tissues (with correction for multiple tests made using estimation of false discovery rate, q-value < 0.05)

	0–5 tissues	6–18 tissues	19–37 tissues	38–54 tissues	55–66 tissues	67–71 tissues	72 tissues
GO Biological Processes	40	22	3	0	0	1	35
GO Molecular Functions	6	1	0	0	0	7	26
GO Cellular Components	27	15	6	0	0	4	12

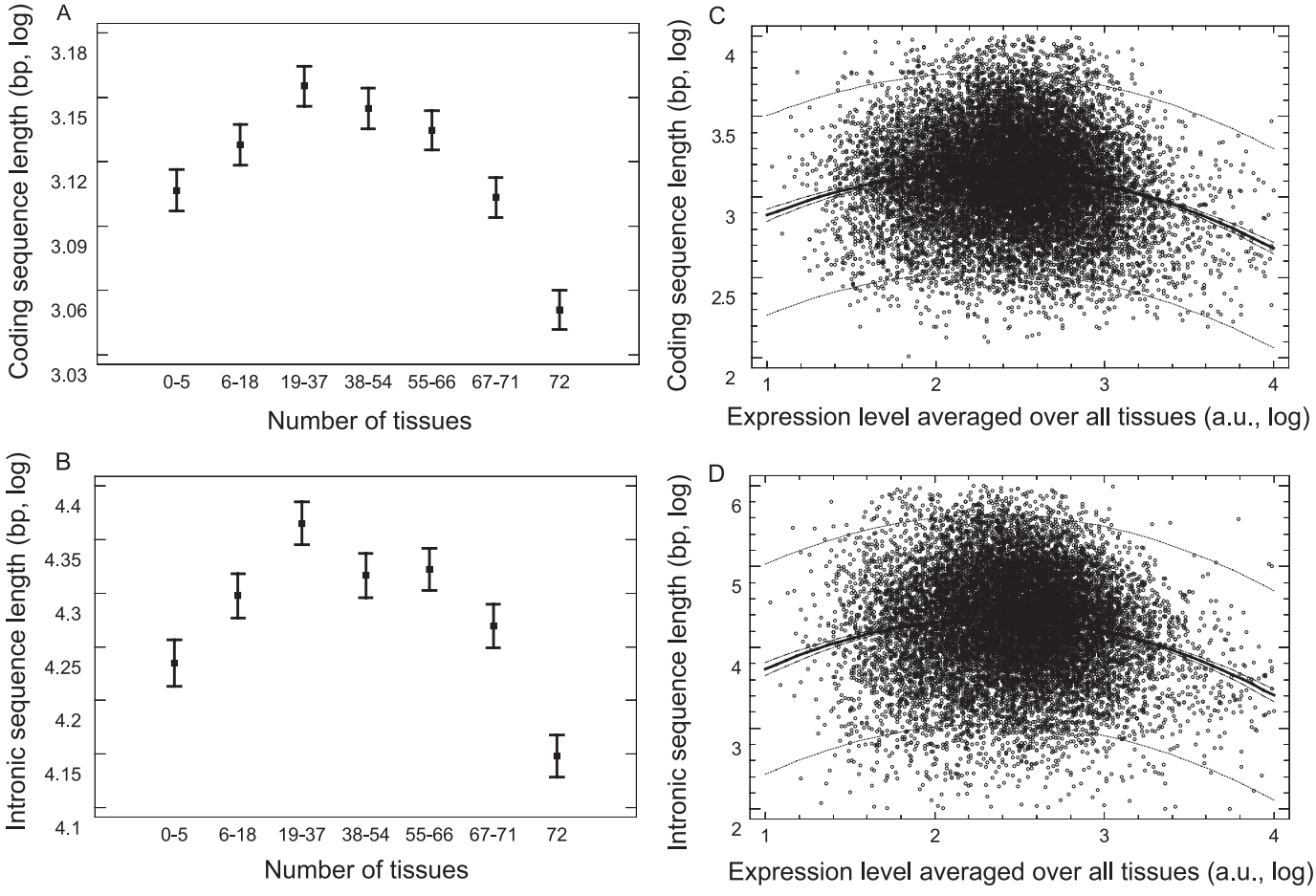

Figure 2. The length of coding and intronic sequences in human genes expressed in different numbers of tissues (**A** and **B**) and with different expression levels (averaged over all tissues) (**C** and **D**). A and B denote mean values with LSD intervals (ANOVA and Kruskal–Wallis, $P < 10^{-12}$ in both cases); C and D denote nonlinear polynomial regression (for the second-order polynomial term, which manifests the nonlinearity, $P < 10^{-12}$ in both cases; dashed lines, confidence limits; dotted lines, prediction limits). (In B and D only genes with introns were taken.)

'q-value' program (26). The conventional statistical analyses (ANOVA and Kruskal–Wallis tests, polynomial regression) were done using the Statgraphics Plus (Statistical Graphics Co.) software package. The star plot (Figure 6C) was done using the Statistica (StatSoft, Inc.) package.

RESULTS

General picture

The intermediately expressed human genes are longer both in coding and intronic part (Figure 2A and B).

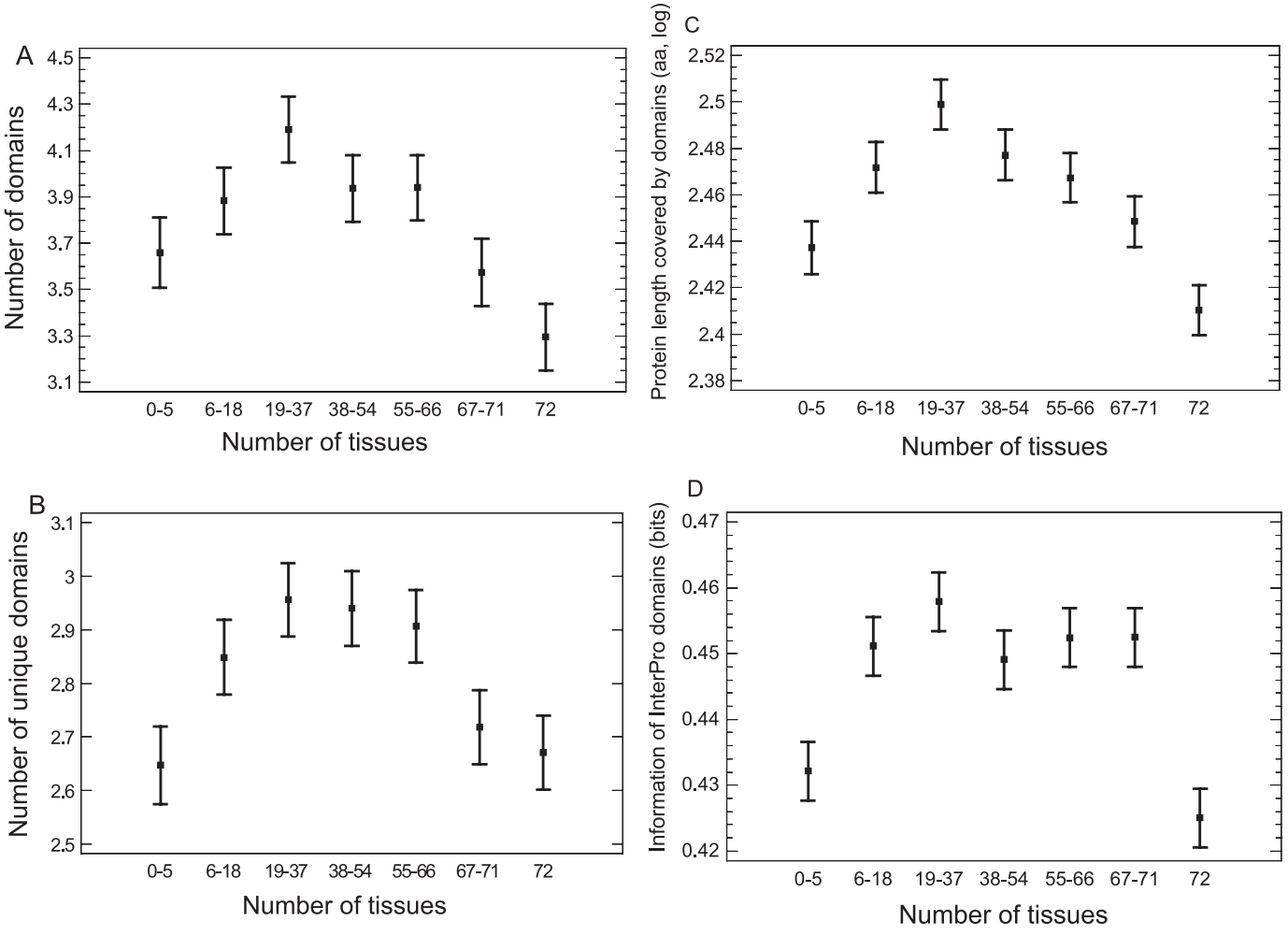

Figure 3. The average number of Pfam domains (**A**), the average number of unique Pfam domains (**B**), the average protein length covered by Pfam domains (**C**), and the average genomewide-contextual Shannon information (uncertainty) of InterPro domains (the picture was similar for Pfam domains) (**D**) in human genes expressed in different numbers of tissues. [ANOVA and Kruskal–Wallis: (A) $P < 10^{-9}$, (B) $P < 10^{-7}$, (C) $P < 10^{-12}$ and (D) $P < 10^{-4}$.]

The tissue-specific genes are generally longer than housekeeping genes, in consistence with the previous reports (1–4). (It should be noted that the previous reports were based on the smaller number of studied tissues and the older versions of gene expression databases.) There is a strong correlation between the number of tissues where a given gene is expressed and its expression level averaged over all tissues, which is similar for different expression thresholds (thus, for thresholds in the range of 0.5–2.0 dataset medians, Spearman $r > 0.91$, $P < 10^{-12}$). This fact allows confirming the effect shown in Figure 2A and B without the use of arbitrary expression threshold and gene grouping (Figure 2C and D).

The 'genome design' model suggests that the length of a gene (including its intronic part) is roughly proportional to its functional load (4,11). Whether the intermediately expressed genes indeed have a higher complexity? First of all, the number of encoded protein functional domains is greater in them (Figure 3A), which suggests that the increased length of coding sequence is not just a 'junk' accumulated because of relaxation of selection for economy and/or mutation bias (as was assumed in the neutralist interpretations) but is related to functional load. The number of unique domains is also greater in the intermediately expressed genes (Figure 3B), which suggests not only the intensification of the same function (through accumulation of identical domains) but also the real increase in functional complexity of protein architectures. The protein length covered by functional domains is also greater in the intermediately expressed genes (Figure 3C), which indicates that their higher number of domains is not associated with a lower domain size and that there is a real increase in protein length loaded with function. Notably, it was recently argued that a lower evolutionary rate of highly expressed (i.e. mostly housekeeping) proteins is not due to a greater functional density of their sequence (27), which is consistent with the present data (Figure 3).

Similar to the case of coding sequence, the greater length of intronic sequence in the intermediately expressed genes cannot be explained by selection for economy and/or mutation bias. Because of the (above-mentioned) strong correlation between average expression level and among-tissues

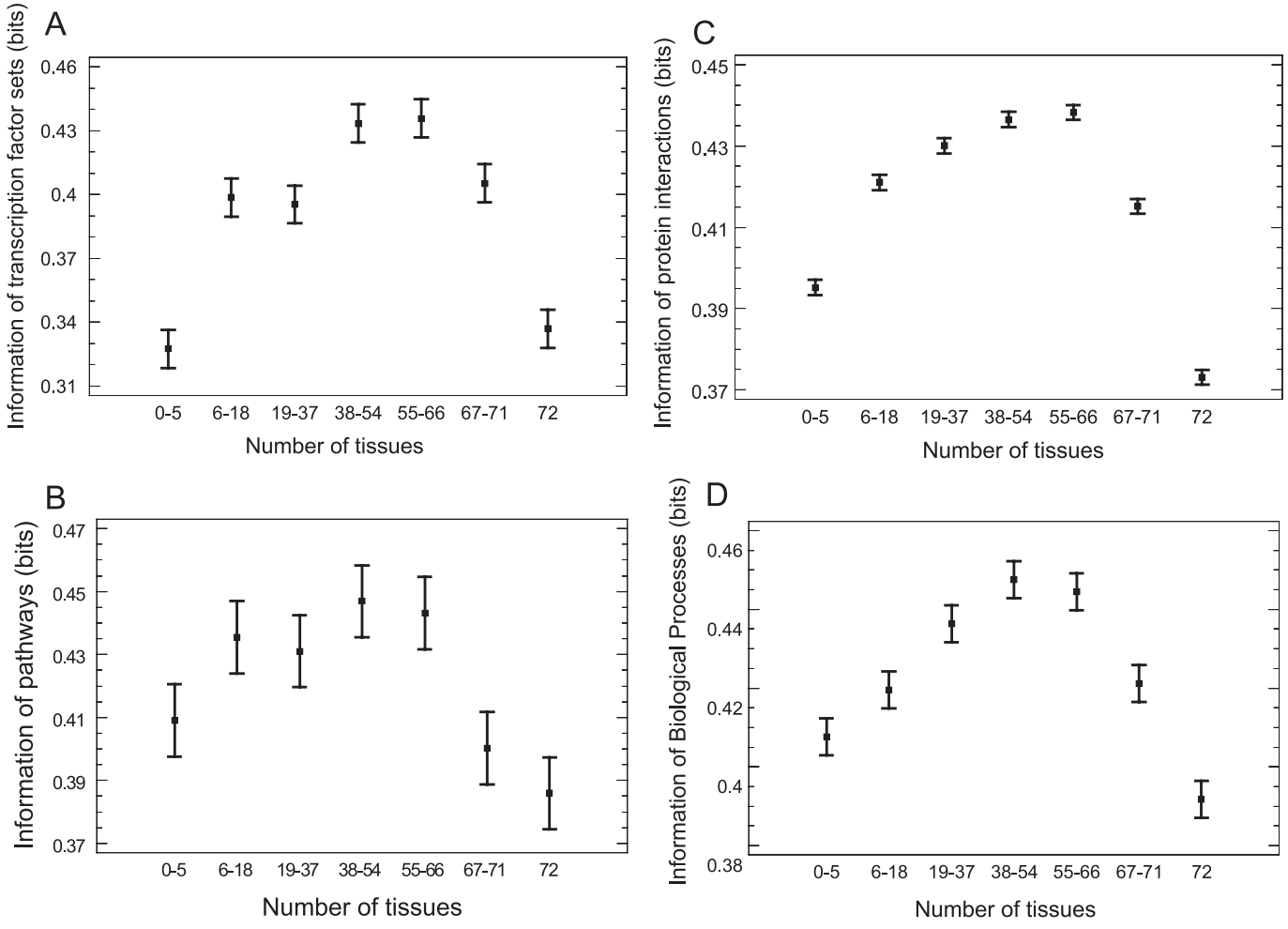

Figure 4. The average genomewide-contextual Shannon information (uncertainty) of transcription factor sets (**A**), pathways (**B**), protein interactions (**C**), Gene Ontology Biological Processes (**D**) in human genes expressed in different numbers of tissues [ANOVA and Kruskal–Wallis: (A,C,D) $P < 10^{-12}$; (B) $P < 10^{-5}$]. (The picture was similar for Gene Ontology Molecular Functions and Cellular Components; see Supplementary Figure 1.)

expression breadth, the economy selection should be more effective in the intermediately expressed genes compared with the narrower expressed (more tissue-specific) genes. Therefore, in the case of selection for economy intronic length should decrease monotonically with expression breadth. Were the mutation bias associated with expression level and/or expression breadth [because transcription can increase mutation and recombination rate (28,29)], the effect of mutation bias should also change monotonically with the change of the latter parameters.

Informational approach

The Shannon information theory defines information as a measure of surprise (uncertainty) of a message estimated using the prior probability of this message (25). This approach allows estimating information of any within-genome object in the genomewide context, which can be used for calculation of the prior probability (see Materials and Methods). The average information (uncertainty) of protein functional domains is greater in the intermediately expressed genes (Figure 3D). In other words, there is the overrepresentation of certain domains (and underrepresentation of other domains) both in housekeeping and tissue-specific genes, whereas in the intermediately expressed genes various domains occur more homogeneously (i.e. they are more diversified).

From the standpoint of information theory, regulation of intermediately expressed genes should involve a higher informational load compared with both housekeeping and tissue-specific genes because of a more complex choice of switch-on/off transition (Figure 1). In general agreement with this theoretical expectation, the intermediately expressed genes show the higher average information (uncertainty) of their transcription factor regulatory sets (Figure 4A). The same is found for their involvement in protein interactions, pathways and modules reflected in Gene Ontology categories (Figure 4B–D; Supplementary Figure 1A and B). In consistence with this finding, there are many over- and underrepresented GO categories (corrected for multiple tests) in genes expressed in 0–5 or 72 tissues, and almost none in genes expressed in 19–37, 38–54, 55–66 tissues (Tables 1 and 2). (It should be emphasized that there are roughly equal numbers of genes in each gene expression group.)

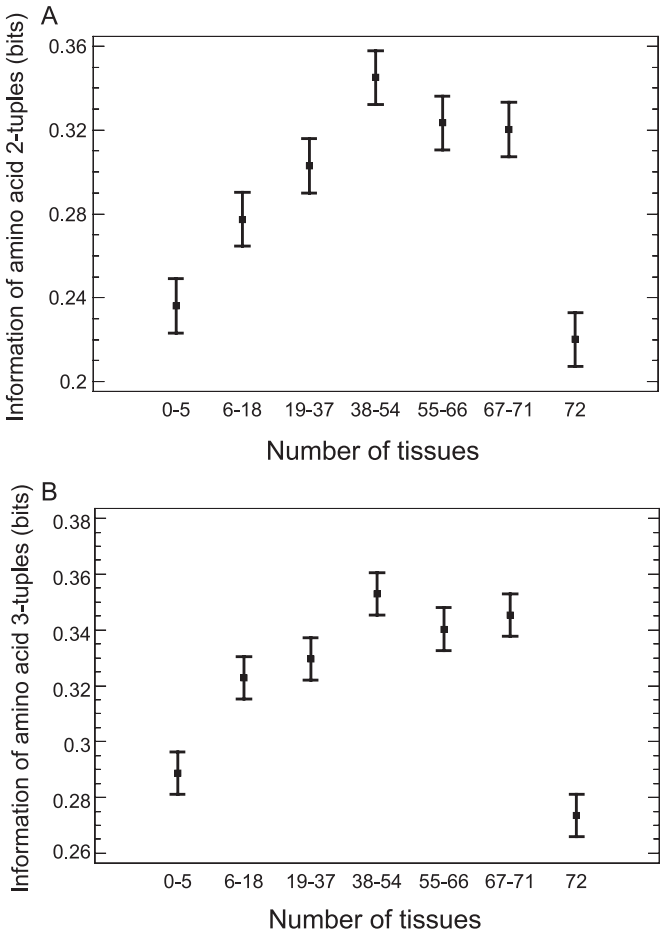

Figure 5. The average genomewide-contextual Shannon information (uncertainty) of amino acid 2-tuples of complete 20-letter alphabet (**A**) and 3-tuples of reduced 10-letter alphabet (**B**) in encoded proteins of human genes expressed in different numbers of tissues (ANOVA and Kruskal–Wallis: $P < 10^{-12}$ in both cases). (The picture was similar for tuples and alphabets of different sizes: 2- to 3-tuples of complete alphabet, 2- to 4-tuples of 10-letter alphabet, 4- to 6-tuples of 5-letter alphabet and 7- to 9-tuples of 3-letter alphabet were tested; see Supplementary Figure 3.)

On the protein sequence level, the intermediately expressed genes show the higher average information of amino acid tuples of different sizes (Figure 5A). To ensure that this effect is not due to a greater total number of tuples in the groups of intermediately expressed genes (because of their longer proteins), the gene groups were normalized to the roughly equal total number of tuples by randomly removing genes from groups with a relatively greater total tuple number. The picture remained similar (Supplementary Figure 2). The effect also holds for tuples of the reduced amino acid alphabets (Figure 5B; Supplementary Figure 3A and B), which reflect more evolutionarily stable protein properties (24,30).

In introns, there is a similar picture with nucleotide tuples (Figure 6A). The effect remains after normalization to an equal total tuple number in each gene expression groups (Supplementary Figure 4). If, by analogy with the reduced amino acid alphabets, the more evolutionarily stable 2-letter purine/pyrimidine alphabet was taken [thus also avoiding the variation in GC content and CpG dinucleotide frequency, which can influence gene expression (31–33)], the effect holds (Figure 6B). It can be seen even without the calculation of information that the distribution of (genomewide-normalized) nucleotide tuple frequencies in the star plot is more homogeneous (and thus have higher uncertainty) in the intermediately expressed genes (Figure 6C). If both intronic DNA strands were taken (and thus excluding also the among-introns variation in purine content), the picture was similar (Supplementary Figure 5A). If the intronic sequences were masked for primate-specific (versus mouse) or all-known repeats, the intermediately expressed genes still showed the higher average information, although there was a relative increase of information in the tissue-specific genes (Supplementary Figure 5B). For the gene promoter regions (from −1000 to +200 nt of transcription start site), the uncertainty of nucleotide tuples is also higher in the intermediately expressed genes for both the complete and 2-letter purine/pyrimidine alphabet (Supplementary Figure 6), which is consistent with the uncertainty of transcription factor regulatory sets (Figure 4A).

However, on the level of sequence tuples (in contrast to explicitly functional objects such as protein functional domains, transcription factor sets, pathways, protein interactions, Gene Ontology categories), there is a problem of discerning information from noise caused by possible redundancy (degeneracy) of the sequence level (especially, in the case of intronic sequence). The use of reduced alphabets with more evolutionarily stable letters (i.e. reflecting those sequence properties that are more tightly linked to function) should reduce this noise. Thus, for amino acid tuples, the picture was similar even with 3-letter alphabet (legend to Figure 5, and Supplementary Figure 3), which reduces the most part of sequence variability (even with a part of information). [It was reported that at least 10-letter alphabet is necessary for description of protein properties; refs (24,30)]. It is impossible to make such a deep alphabet reduction with DNA sequence. For stricter testing of introns, the information (uncertainty) of intronic tuples of 2-letter (purine/pyrimidine) alphabet in both DNA strands was calculated in regard to the equal prior probability of all tuples (i.e. as a non-genomewide-contextual uncertainty). (As said above, sequence variation in GC content and CpG dinucleotide frequency, and its deviation from the equal purine/pyrimidine content were excluded in the case of this 2-letter alphabet and both DNA strands.) This non-genomewide-contextual uncertainty did not differ significantly in the intermediately expressed genes (Figure 6D), which indicates that their increase in the genomewide-contextual uncertainty (Figure 6B) indeed reflects the increase in (genome-specific) information.

DISCUSSION

The whole picture can be summarized as follows. It was argued that the overtaking growth of the number of genes coding for transcription factors over the total number of genes limited the growth of prokaryotic genomes (34). The problem of regulatory complexity turns out to be even more severe for the eukaryotic genomes (35,36). The most complex regulatory problems should appear in the case of

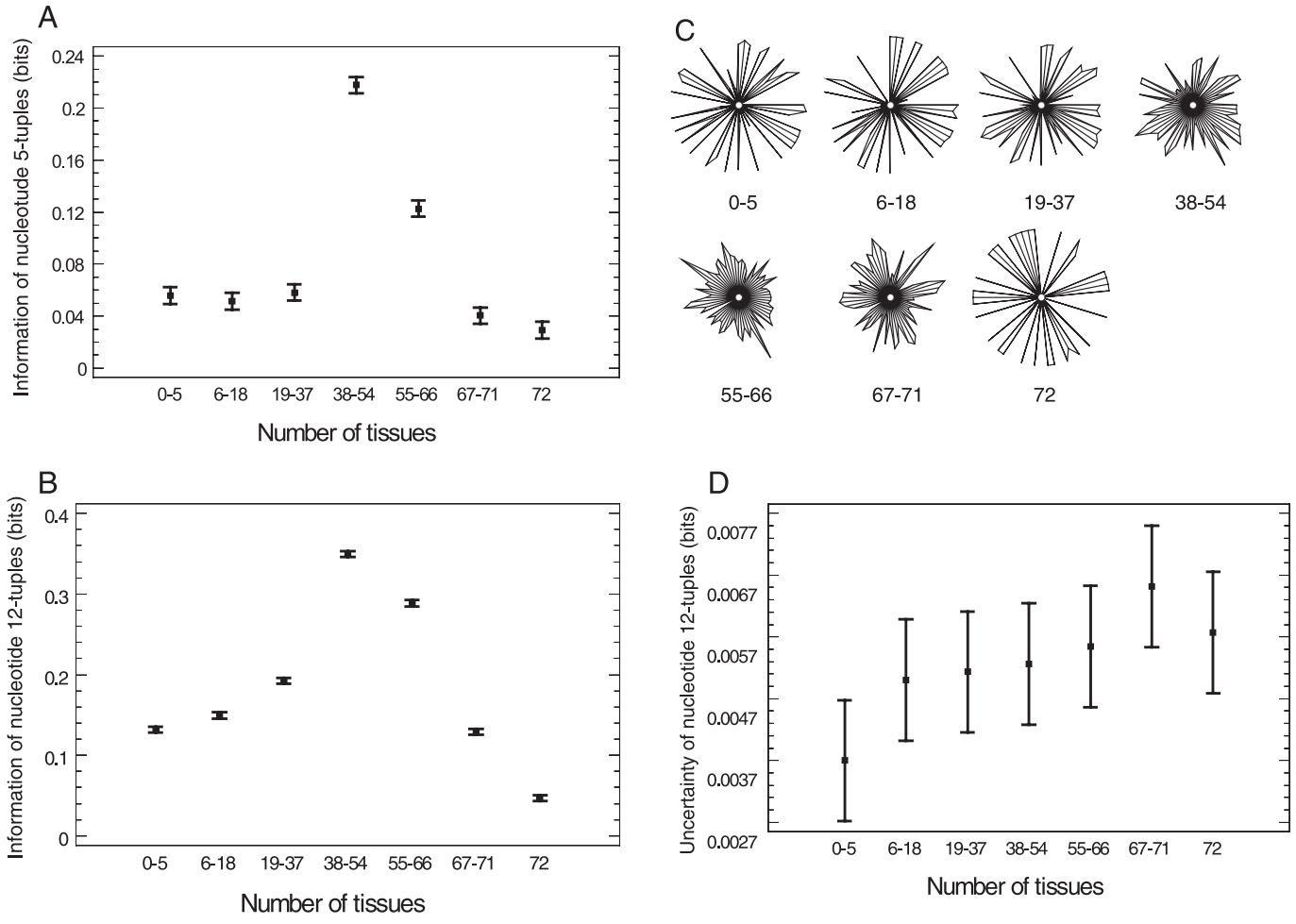

Figure 6. Parameters of intronic nucleotide tuples in human genes expressed in different numbers of tissues. The average genomewide-contextual Shannon information (uncertainty) of nucleotide 5-tuples of complete 4-letter alphabet (**A**) and 12-tuples of 2-letter purine/pyrimidine alphabet (**B**), the star plot of the genomewide-normalized frequencies of 6-tuples of 2-letter purine/pyrimidine alphabet (**C**), the average non-genomewide-contextual uncertainty of 12-tuples of 2-letter purine/pyrimidine alphabet in both DNA strands (**D**). In the star plot (**C**), the relative frequency of each tuple is plotted along one of star rays (in the same order in each plot), the larger the difference among the ray lengths, the more heterogeneous is the distribution of tuple frequencies. Note that in genes expressed in 0–5, 6–18, 19–37 and 72 tissues, some rays are so small that they are even invisible on this scale [i.e. they look as empty sectors of the star). (ANOVA and Kruskal–Wallis: (A and B) $P < 10^{-12}$]. (The picture was similar for tuples of different sizes: 2- to 6-tuples of complete alphabet, and 4- to 14-tuples of 2-letter alphabet were tested; see also Supplementary Figure 5.)

intermediately expressed genes (Figure 1). Therefore, the burden of regulatory complexity might force the dichotomy of housekeeping versus tissue-specific genes in the multicellular organisms [which can be seen in the histogram of genes expressed in different numbers of tissues: e.g. figure 1 in (37); figure 2 in (38)]. The occurrence of protein functional domains, participation in transcription factor regulatory sets, pathways, protein interactions, biological processes, molecular functions and cellular components also reflect this dichotomy, showing the maximum genomewide-contextual uncertainty (and thus, informational load) in the intermediately expressed genes. In other words, there are much less intermediate-specific modules (if any) than housekeeping and tissue-specific modules (Tables 1 and 2). This effect is observed notwithstanding the fact that gene expression bins are normalized to the roughly equal numbers of genes. Thus, the dichotomy of housekeeping versus tissue-specific entities is much more pronounced on the modular level than on the molecular level. The intermediately expressed genes possibly connect housekeeping and tissue-specific modules. In any case, they have the higher functional and regulatory complexity reflected in their greater length, which is consistent with the 'genome design' model.

The domain architecture is considered the most important level of protein functional complexity, especially in the eukaryotic genomes (39–41). Proteins encoded by the intermediately expressed genes are shown here to consist of a greater number of various domains and therefore may perform more complex and diverse functions. The higher complexity of the intermediately expressed genes is also reflected in the frequency of amino acid tuples in encoded proteins and nucleotide tuples in introns and promoter regions. A possible functional load of introns is discussed in (11,42–47). Briefly, introns can harbor a plethora of regulatory elements acting in multiple ways: the interaction with transcription factors (as enhancers and suppressors), the regulation mediated by splicing, and the action of noncoding RNAs located in introns (to say nothing of alternative

splicing which alters the protein structure). It is noteworthy that first intron, which is more often known to contain regulatory elements (48,49), is longer in the intermediately expressed genes (Supplementary Figure 7). The situation is further complicated by participation of introns in chromatin organization and interplay of the latter with transcriptional regulation (e.g. 10,50). It is interesting that in the yeast, introns are longer in the highly expressed genes (51), which contradicts the 'selection for economy' model. [The same is probably true for some other unicellular organisms, judging by the correlation between intron length and frequency of optimal codons (51)]. This fact indicates that in introns of unicellular organisms, the amount of activating elements outweigh the amount of suppressing ones. The fraction of non-housekeeping (i.e. generally suppressed) genes is much lower in unicellular organisms, therefore there should be a lower amount of suppressing elements in their introns. The maximum chromatin condensation is 5-fold lower in yeast when compared with mammals (52), which suggests that yeast introns should be less loaded with chromatin-condensation function.

It should be noted that the 'selection for economy' model comes in two flavors: 'energy economy' and 'time economy', which were contrasted in the case of human bi-directional genes (6,7). The former was rejected in favor of the latter because antisense genes expressed are both shorter and narrower than corresponding sense genes (6,7). However, the antisense genes can be miniaturized because they should be accommodated within the loci of the sense genes, which is consistent with the 'genome design' model (11). (Also, their shorter length may be adequate for their function.) Moreover, in contrast to the energy economy, time economy is not additive in a piecemeal way [as in 'beanbag genetics' (53)]. In other words, the speed of an intracellular event probably cannot be changed without corresponding changes in other parts of the system. (Imagine an electronic circuit where some events are accelerated without adjustment of the others.) Therefore, time economy is closer in sense to 'genome design' because in this case genomic structure should be selected as a system [for timing design (54)].

The combinatorial control of gene expression involving cooperation of multiple transcription factors is now an emerging theme (50,55,56). Due to the most complex choice of switch-on/off transition in the case of intermediately expressed genes (according to the information theory), regulation of these genes should be more complex. Therefore, it may involve a greater amount of multiple regulatory factors (and their binding sites). Finally, evolutionary design becomes a recurrent theme in systems biology of gene and protein networks (54,57–61). It may have a counterpart in the blueprint of these networks (genomic structure).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

I thank two anonymous reviewers 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' (MCB RAS). The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Castillo-Davis,C.I., Mekhedov,S.L., Hartl,D.L., Koonin,E.V. and Kondrashov,F.A. (2002) Selection for short introns in highly expressed genes. *Nature 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. (2004) Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet*., **20**, 248–253.
5. Vinogradov,A.E. (2004) Evolution of genome size: multi-level selection, mutation bias or dynamical chaos? *Curr. Opin. Genet. Devel*., **14**, 620–626.
6. Chen,J., Sun,M., Hurst,L.D., Carmichael,G.G. and Rowley,J.D. (2005) Human antisense genes have unusually short introns: evidence for selection for rapid transcription. *Trends Genet*., **21**, 203–207.
7. Chen,J., Sun,M., Rowley,J.D. and Hurst,L.D. (2005) The small introns of antisense genes are better explained by selection for rapid transcription than by 'genomic design'. *Genetics*, **171**, 2151–2155.
8. Cohen-Gihon,I., Lancet,D. and Yanai,I. (2005) Modular genes with metazoan-specific domains have increased tissue specificity. *Trends Genet*., **21**, 210–213.
9. Sironi,M., Menozzi,G., Comi,G.P., Cagliani,R., Bresolin,N. and Pozzoli,U. (2005) Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences. *Hum. Mol. Genet*., **14**, 2533–2546.
10. Vinogradov,A.E. (2005) Noncoding DNA, isochores and gene expression: nucleosome formation potential. *Nucleic Acids Res*., **33**, 559–563.
11. Vinogradov,A.E. (2006) 'Genome design' model: evidence from conserved intronic sequence in human-mouse comparison. *Genome Res*., **16**, 347–354.
12. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **33**, D501–D504.
13. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
14. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res*., **32**, D138–D141.
15. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al*. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res*., **33**, D201–D205.
16. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
17. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*., **34**, D354–D357.
18. Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G.R., Wu,G.R., Matthews,L. *et al*. (2005)

Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*., **33**, D428–D432.
19. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **33**, D39–D45.
20. Romero,P., Wagg,J., Green,M.L., Kaiser,D., Krummenacker,M. and Karp,P.D. (2005) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol*., **6**, R2.
21. The Gene Ontology Consortium (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res*., **34**, D322–D326.
22. von Mering,C., Jensen,L.J., Snel,B., Hooper,S.D., Krupp,M., Foglierini,M., Jouffre,N., Huynen,M.A. and Bork,P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*., **33**, D433–D437.
23. Suzuki,Y., Yamashita,R., Sugano,S. and Nakai,K. (2004) DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Res*., **32**, D78–D81.
24. Li,T., Fan,K., Wang,J. and Wang,W. (2003) Reduction of protein sequence complexity by residue grouping. *Protein Eng*., **16**, 323–330.
25. Shannon,C. (1948) A mathematical theory of communication. *Bell System Techn. J*., **27**, 379–423.
26. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
27. Drummond,D.A., Bloom,J.D., Adami,C., Wilke,C.O. and Arnold,F.H. (2005) Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA*, **102**, 14338–14343.
28. Aguilera,A. (2002) The connection between transcription and genomic instability. *EMBO J*., **21**, 195–201.
29. Comeron,J.M. (2004) Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics*, **167**, 1293–1304.
30. Fan,K. and Wang,W. (2003) What is the minimum number of letters required to fold a protein? *J. Mol. Biol*., **328**, 921–926.
31. Zhang,L., Kasif,S., Cantor,C.R. and Broude,N.E. (2004) GC/AT-content spikes as genomic punctuation marks. *Proc. Natl Acad. Sci. USA*, **101**, 16855–16860.
32. Vinogradov,A.E. (2005) Dualism of gene GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth. *Trends Genet*., **21**, 639–643.
33. Kudla,G., Lipinski,L., Caffin,F., Helwak,A. and Zylicz,M. (2006) High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol*., **4**, e180.
34. Mattick,J.S. and Gagen,M.J. (2005) Accelerating networks. *Science*, **307**, 856–858.
35. Claverie,J.M. (2001) Gene number. what if there are only 30,000 human genes? *Science*, **291**, 1255–1257.
36. Szathmary,E., Jordan,F. and Pal,C. (2001) Molecular biology and evolution. Can genes explain biological complexity? *Science*, **292**, 1315–1316.
37. Vinogradov,A.E. (2003) Isochores and tissue-specificity. *Nucleic Acids Res*., **31**, 5212–5220.
38. Jongeneel,C.V., Delorenzi,M., Iseli,C., Zhou,D., Haudenschild,C.D., Khrebtukova,I., Kuznetsov,D., Stevenson,B.J., Strausberg,R.L., Simpson,A.J. and Vasicek,T.J. (2005) An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Res*., **15**, 1007–1014.
39. Vogel,C., Bashton,M., Kerrison,N.D., Chothia,C. and Teichmann,S.A. (2004) Structure, function and evolution of multidomain proteins. *Curr. Opin. Struct. Biol*., **14**, 208–216.
40. Orengo,C.A. and Thornton,J.M. (2005) Protein families and their evolution—a structural perspective. *Annu. Rev. Biochem*., **74**, 867–900.
41. Lin,K., Zhu,L. and Zhang,D.Y. (2006) An initial strategy for comparing proteins at the domain architecture level. *Bioinformatics*, **22**, 2081–2086.
42. Le Hir,H., Nott,A. and Moore,M.J. (2003) How introns influence and enhance eukaryotic gene expression. *Trends Biochem. Sci*., **28**, 215–220.
43. Nott,A., Meislin,S.H. and Moore,M.J. (2003) A quantitative analysis of intron effects on mammalian gene expression. *RNA*, **9**, 607–617.
44. Pozzoli,U. and Sironi,M. (2005) Silencers regulate both constitutive and alternative splicing events in mammals. *Cell Mol. Life Sci*., **62**, 1579–1604.
45. Mattick,J.S. (2004) RNA regulation: a new genetics? *Nature Rev. Genet*., **5**, 316–323.
46. Fedorova,L. and Fedorov,A. (2005) Puzzles of the human genome: why do we need our introns? *Curr. Genomics*, **6**, 589–595.
47. Pang,K.C., Frith,M.C. and Mattick,J.S. (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet*., **22**, 1–5.
48. Majewski,J. and Ott,J. (2002) Distribution and characterization of regulatory elements in the human genome. *Genome Res*., **12**, 1827–1836.
49. Keightley,P.D. and Gaffney,D.J. (2003) Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc. Natl Acad. Sci. USA*, **100**, 13402–13406.
50. Barrera,L.O. and Ren,B. (2006) The transcriptional regulatory code of eukaryotic cells—insights from genome-wide analysis of chromatin organization and transcription factor binding. *Curr. Opin. Cell Biol*., **18**, 291–298.
51. Vinogradov,A.E. (2001) Intron length and codon usage. *J. Mol. Evol*., **52**, 2–5.
52. Russell,P. and Nurse,P. (1986) *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae*: a look at yeasts divided. *Cell*, **45**, 781–782.
53. Crow,J.F. (2001) The beanbag lives on. *Nature*, **409**, 771.
54. Zaslaver,A., Mayo,A.E., Rosenberg,R., Bashkin,P., Sberro,H., Tsalyuk,M., Surette,M.G. and Alon,U. (2004) Just-in-time transcription program in metabolic pathways. *Nature Genet*., **36**, 486–491.
55. Ogata,K., Sato,K. and Tahirov,T.H. (2003) Eukaryotic transcriptional regulatory complexes: cooperativity from near and afar. *Curr. Opin. Struct. Biol*., **13**, 40–48.
56. Remenyi,A., Scholer,H.R. and Wilmanns,M. (2004) Combinatorial control of gene expression. *Nature Struct. Mol. Biol*., **11**, 812–815.
57. Alon,U. (2003) Biological networks: the tinkerer as an engineer. *Science*, **301**, 1866–1867.
58. Powell,K. (2004) All systems go. *J. Cell Biol*., **165**, 299–303.
59. Kashtan,N. and Alon,U. (2005) Spontaneous evolution of modularity and network motifs. *Proc. Natl Acad. Sci. USA*, **102**, 13773–13778.
60. Zhang,L.V., King,O.D., Wong,S.L., Goldberg,D.S., Tong,A.H., Lesage,G., Andrews,B., Bussey,H., Boone,C. and Roth,F.P. (2005) Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network. *J. Biol*., **4**, 6.
61. Yu,H., Xia,Y., Trifonov,V. and Gerstein,M. (2006) Design principles of molecular networks revealed by global comparisons and composite motifs. *Genome Biol*., **7**, R55.