4 Skaletsky, H. *et al.* (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423, 825–837
5 Ross, M.T. *et al.* (2005) The DNA sequence of the human X chromosome. *Nature* 434, 325–337
6 Huynh, K.D. and Lee, J.T. (2005) X-chromosome inactivation: a hypothesis linking ontogeny and phylogeny. *Nat. Rev. Genet.* 6, 410–418
7 Turner, J.M. *et al.* (2005) Silencing of unsynapsed meiotic chromosomes in the mouse. *Nat. Genet.* 37, 41–47
8 Turner, J.M. (2007) Meiotic sex chromosome inactivation. *Development* 134, 1823–1831
9 Khil, P.P. *et al.* (2004) The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. *Nat. Genet.* 36, 642–646
10 Emerson, J.J. *et al.* (2004) Extensive gene traffic on the mammalian X chromosome. *Science* 303, 537–540
11 Vinckenbosch, N. *et al.* (2006) Evolutionary fate of retroposed gene copies in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* 103, 3220–3225
12 Potrzebowski, L. *et al.* (2008) Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol.* 6, e80
13 Betran, E. *et al.* (2002) Retroposed new genes out of the X in *Drosophila*. *Genome Res.* 12, 1854–1859
14 Veyrunes, F. *et al.* (2008) Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome Res.* 18, 965–973
15 Wang, P.J. (2004) X chromosomes, retrogenes and their role in male reproduction. *Trends Endocrinol. Metab.* 15, 79–83
16 Wang, P.J. and Page, D.C. (2002) Functional substitution for TAF(II)250 by a retroposed homolog that is expressed in human spermatogenesis. *Hum. Mol. Genet.* 11, 2341–2346
17 Marques, A.C. *et al.* (2005) Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol.* 3, e357
18 Hillier, L.W. *et al.* (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432, 695–716
19 Hedges, S.B. (2002) The origin and evolution of model organisms. *Nat. Rev. Genet.* 3, 838–849
20 van Rheede, T. *et al.* (2006) The platypus is in its place: nuclear genes and indels confirm the sister group relation of monotremes and Therians. *Mol. Biol. Evol.* 23, 587–597

0168-9525/$ – see front matter
doi:10.1016/j.tig.2008.07.006 Available online 5 September 2008

Genome Analysis

On the nature of human housekeeping genes

Jiang Zhu[1,2*], Fuhong He[1,2*], Songnian Hu[1] and Jun Yu[1]

[1] Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China
[2] Graduate University of Chinese Academy of Sciences, Beijing, China

Using a collection of expressed sequence tag (EST) data, we re-evaluated the correlation of tissue specificity with genomic structure, phyletic age, evolutionary rate and promoter architecture of human genes. We found that housekeeping genes are less compact and older than tissue-specific genes, and they evolve more slowly in terms of both coding and core promoter sequences. Housekeeping genes primarily use CpG-dependent core promoters, whereas the majority of tissue-specific genes possess neither CpG-islands nor TATA-boxes in their core promoters.

Expressed sequence tag–based evidence

Housekeeping (HK) genes are ubiquitously expressed in all tissue and cell types and constitute the basal transcriptome for the maintenance of basic cellular functions. Partitioning transcriptomes into HK and tissue-specific (TS) genes and characterizing the two groups of genes in terms of their genomic structure, phyletic age, evolutionary rate and transcriptional regulation are fundamental to understand human transcriptomes. Many studies have revealed the structural [1,2], evolutionary [3,4] and promoter features of HK genes [5,6], but they were largely based on microarray data that tend to underestimate the number of human HK genes [7]. On the basis of publicly available expressed sequence tag (EST) data, we found that a large fraction (40%) of currently-annotated genes are universally expressed [7]. Here we used an EST-based estimate of expression breadth – the number of tissues in which a gene is expressed – to re-evaluate the nature of tissue specificity. We confirmed that HK genes are in general highly expressed [2] and evolve slower in coding sequence (CDS) [3] compared with TS genes. However, our analyses cast doubt on previous observations that HK genes have more compact structure [1,2] and reduced promoter conservation [6] than TS genes. In addition, we showed that HK genes are in general older than TS genes and have very distinct core promoter architecture.

Gene structure

We investigated the breath of expression for 17 288 human RefSeq loci [downloaded from NCBI (June 18, 2007 update)] across 18 human tissues (Ref. [7]; Supplementary Methods), and defined HK and TS genes as those expressed in all 18 tissues and in only 1 tissue, respectively. We observed that genes' length parameters are positively correlated with expression breadth (Figure 1a). The medians of genomic, transcript and CDS lengths are 28.8, 2.8 and 1.4 kb for HK genes and 7.2, 1.6 and 1.0 kb for TS genes, respectively (Table 1); all length parameters of HK genes are significantly longer than those of TS genes (Wilcoxon test, $P < 2.2 \times 10^{-16}$). Moreover, HK genes tend to have a greater number of exons (median of 11) than TS genes (median of 4; Wilcoxon test, $P < 2.2 \times 10^{-16}$). These observations contradict previous microarray-based results that a selection for compactness makes the lengths of intron, untranslated region (UTR) and CDS in HK genes shorter than those in non-HK genes [1,2].

The selection for compactness in HK genes was consistent with the 'selection for economy' hypothesis – highly

Corresponding author: Yu, J. (junyu@big.ac.cn).
* These authors contributed equally to this work

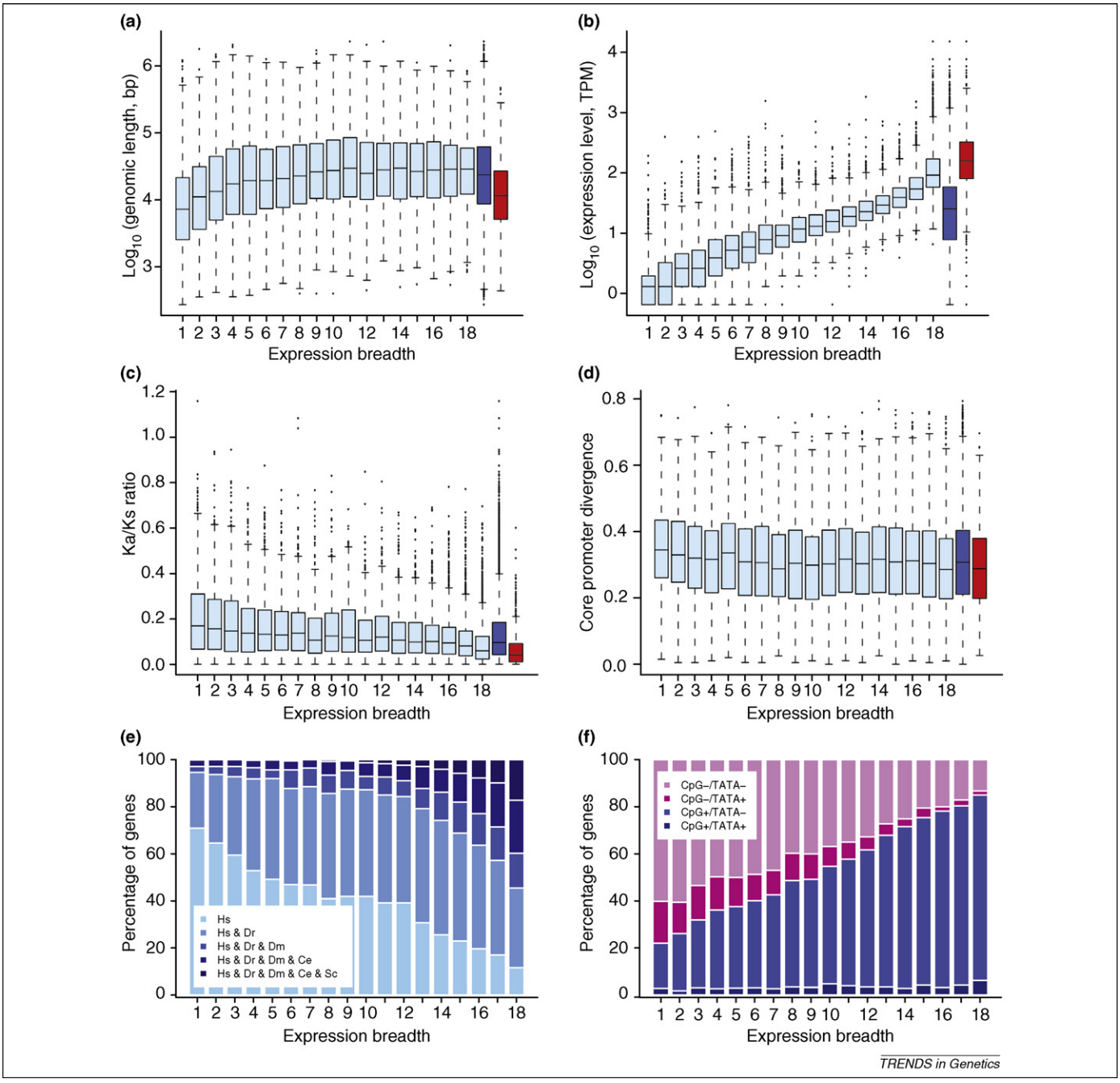

Figure 1. Relationships among gene parameters. **(a)** The genomic length, **(b)** expression level, **(c)** Ka/Ks ratio of coding sequence (CDS) and **(d)** sequence divergence of core promoter are shown as box-plot for genes in each expression breadth group. The boxes depict data between the 25th and 75th percentiles with central horizontal lines representing the median values; extreme values are indicated by dots outside the boxes. Total genes (blue boxes) and microarray-defined housekeeping (HK) genes [1] (red boxes) are also plotted for comparison. Genomic length and expression level are positively correlated with expression breadth, whereas Ka/Ks ratio in CDS and the number of substitutions per site in core promoter are negatively correlated [(a) Spearman $\rho = 0.166$; (b) Spearman $\rho = 0.855$; (c) Spearman $\rho = -0.264$; (d) Spearman $\rho = -0.078$; all $P < 2.2 \times 10^{-16}$]. **(e)** Genes are classified into five phyletic categories approximately representing different evolutionary origins: (i) mammalian-specific, conserved in human (Hs) only; (ii) vertebrate-specific, conserved in human and zebrafish (Dr) but not in other species; (iii) conserved in human, zebrafish and fly (Dm) but not in other species; (iv) metazoan-specific, conserved in human, zebrafish, fly and nematode (Ce) but not in yeast (Sc) and (v) eukaryote-specific, conserved in all five organisms. The fractions of each category are plotted over expression breadths, showing that widely expressed genes are in general older than specifically expressed genes. **(f)** Genes are classified into four groups according to their CpG/TATA content in core promoter, and the fractions of each group are plotted over expression breadths. A majority of widely expressed genes use CpG-dependent core promoters, but most specifically expressed genes have neither CpG-islands nor TATA-boxes.

expressed genes are subjected to intron shortening to reduce the energy burden of transcription [8]. We confirmed that expression level is positively correlated with expression breadth [2] (Figure 1b); the medians of expression level for HK and TS genes are 91.4 and 1.3 transcripts per million (TPM), respectively (Table 1). However, the negative correlation between expression level and length parameters is actually nonlinear. Although highly expressed genes are rarely large, poorly expressed genes show a broad range of variations in gene length (Supplementary Figure S1). Expression level is thus not a meaningful predictor for gene length. Indeed, this negative correlation does not exist in yeasts [9], flies [10] and plants [11]. Genes in these species are almost one magnitude shorter than those in human and

Table 1. Gene parameters for HK[a] and TS genes

Gene parameters[b]	TS genes ($n = 885$)	HK genes ($n = 3140$)	Total genes ($n = 17\,288$)	MA HK genes[c] ($n = 567$)
Genomic length (bp)	7191 (29 113.0)	28 792 (49 945.9)	23 541 (59 964.1)	11 555 (24 867.3)
Transcript length (bp)	1601 (1846.1)	2801 (3363.2)	2517 (3048.4)	1807 (2148.1)
CDS length (bp)	963 (1190.6)	1380 (1833.4)	1314 (1722.5)	978 (1231.6)
Number of exons	4 (5.6)	11 (13.1)	8 (10.9)	8 (9.4)
Expression level (TPM)[d]	1.3 (3.6)	91.4 (188.9)	25.3 (61.0)	157.5 (369.5)
Ka/Ks ratio	0.17 (0.21)	0.06 (0.09)	0.10 (0.13)	0.04 (0.07)
Promoter divergence[e]	0.35 (0.35)	0.29 (0.29)	0.31 (0.31)	0.29 (0.29)

[a]HK, housekeeping; TS, tissue specific; MA, microarray; TPM, transcripts per million.
[b]Median and mean values (in parentheses) of gene parameters are shown.
[c]Previous MA-defined HK genes [1].
[d]Original expressed sequence tag counts in non-normalized cDNA libraries are converted into transcripts per million.
[e]The number of substitutions per site in the [−100, +100] core promoter of human-mouse orthologs is calculated by using PAML package.

other vertebrates (data not shown); therefore, the selection on gene length for economy might only act on the extremely large genes [12,13].

In general, we observed that TS genes are expressed at lower level and are shorter than HK genes (Supplementary Figure S1a). We found that many moderately expressed large genes, although defined as HK genes by EST data, were not categorized as HK genes by previous microarray data [1] (Supplementary Figure S1b). It is this false-negative detection of HK genes that leads to an underestimate of HK gene length in microarray data. Studies using updated microarray data also proposed a 'genome design' model – genes with the intermediate expression breadth have the maximal length because of greater functional and regulatory complexity – to explain the relationship between tissue specificity and gene length [2,14]. However, this model was recently challenged by the observation that HK genes are longer than narrowly expressed genes with similar expression levels [15].

Evolutionary features

The coding sequences of HK genes are thought to evolve more slowly than those of TS genes [3]. The Ka/Ks ratios, calculated from 14 961 human–mouse orthologs, are negatively correlated with expression breadth (Figure 1c), consistent with previous observations. Moreover, we observed that the sequence divergences of [−100, +100] core promoters are also negatively correlated with expression breadth (Figure 1d). The numbers of substitutions per site in HK gene promoters (median of 0.29) are significantly lower than that in TS gene promoters (median of 0.35; Wilcoxon test, $P < 2.2 \times 10^{-16}$; Table 1) These observations implied that purifying selection acts on both CDS and core promoters of HK genes. A recent study reported that distant promoters of HK genes showed reduced sequence conservation compared with those of TS genes [6], but understanding the evolution of distant promoter requires a clear knowledge of the complex mechanism of transcriptional regulation, which remains largely unknown for most human genes.

It has been reported that ancient genes tend to be universally expressed [4]. To validate if HK genes are, on average, older than TS genes, we compiled orthologs among five organisms, *Homo sapiens*, *Danio rerio*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae*. We classified 15 501 (89.7% of total) human genes into five major categories according to their phyletic distribution (Figure 1e and Supplementary Table S1). The phyletic categories approximate the age of genes; the more diverged species genes are conserved in, the more ancient they are. We found that universally expressed genes in general are more ancient in origin compared with specifically expressed genes (Figure 1e); 54.5% of HK genes originate before the separation of vertebrate lineage in contrast to 5.4% of TS genes, whereas only 11.6% of HK genes are unique to mammals in contrast to 70.8% of TS genes. Interestingly, genes conserved between human and zebrafish but not others (vertebrate-specific) are more prevalent in genes with moderate tissue specificity and deficient in both HK and TS genes. These observations are consistent with previous reports that ancient genes are longer and evolve slower [16].

Promoter architecture

TATA-box and CpG-island are two of the most-characterized promoter features related to tissue specificity [5]. TATA-box is thought to be absent in HK genes [17], whereas CpG-island covers the transcription start sites of most HK genes [18]. We classified 16 585 (95.9% of total) human genes according to the presence and absence of CpG-island and TATA-box in the core promoter. We found that CpG+/TATA−, CpG−/TATA−, CpG−/TATA+ and CpG+/TATA+ genes constitute 58.0, 31.4, 6.7 and 3.9% of the total, respectively. The fraction of CpG+/TATA− genes is positively correlated with expression breadth, whereas those of CpG−/TATA− and CpG−/TATA+ are negatively correlated. Surprisingly, CpG+/TATA+ genes appear equally prevalent in all expression breadth groups as if occurring randomly (Figure 1f). More than three quarters (78.7%) of HK genes predominantly have a CpG+/TATA− core promoter, whereas TS genes show less preference; 60.2, 19.2 and 17.7% are associated with CpG−/TATA−, CpG+/TATA− and CpG−/TATA+ core promoters, respectively. The majority of TS genes possess neither CpG-islands nor TATA-boxes in their core promoters. Further molecular experiments are necessary to elucidate the mechanism of transcription initiation specific to TS genes.

CpG-islands typically have multiple GC-boxes bound by the transcription factor SP1, and GC-boxes coupled with Inr elements can initiate transcription in the absence of TATA-box [19]. These observations implied that CpG-dependent core promoters (62% of total genes) represent a more general mechanism of transcription initiation, and

TATA-dependent initiation (11% of total genes) is the exception rather than the rule [20,21]. Moreover, we observed that the fractions of different core promoter types vary gradually among expression breadth groups. Such promoter diversity implied that different types of core promoters might have evolved as functional equivalents to position the transcription machinery [22].

Concluding remarks

Structural and expression parameters concerning gene organization, expression rate, tissue specificity and regulation are correlated to variable extents and convey important information on gene and genome evolution [23,24]. Based on thorough analyses of the public expressed sequence tag data, we re-evaluated some of these parameters related to tissue specificity and compared them with previous results that were largely based on microarray data. We confirmed some relationships, such as the positive and negative correlations of tissue specificity with Ka/Ks ratio and expression level, respectively. However, our results contradicted some of the previous findings: (i) Housekeeping (HK) genes are less, not more, compact than tissue-specific (TS) genes. The genomic, transcript and coding sequence (CDS) lengths, as well as the number of exons, are all greater in HK genes than in TS genes. (ii) HK genes have enhanced, not reduced, sequence conservation in the core promoter region compared with TS genes. Therefore, purifying selection acts on both CDS and core promoters of HK genes. We also demonstrated that HK genes are more ancient than TS genes. More than one half of the HK genes originated before the divergence of vertebrate lineage, but only 5% of TS genes did. Furthermore, we showed that a majority of HK genes use CpG-dependent core promoters, whereas only a minority of TS genes uses TATA-dependent core promoters, and most of them have neither CpG-islands nor TATA-boxes in the core promoters. These observations highlight the distinct role of HK genes in forming the basal human transcriptome and have important implications for understanding the cell-specific transcriptomes in multicellular organisms.

Acknowledgements

The authors thank the anonymous reviewers for critical comments and helpful suggestions. This work was supported by the National Basic Research Program of China (2006CB910404).

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2008.08.004.

References

1 Eisenberg, E. and Levanon, E.Y. (2003) Human housekeeping genes are compact. *Trends Genet.* 19, 362–365
2 Vinogradov, A.E. (2004) Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.* 20, 248–253
3 Zhang, L. and Li, W.H. (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol. Biol. Evol.* 21, 236–239
4 Freilich, S. *et al.* (2005) Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins. *Genome Biol.* 6, R56
5 Schug, J. *et al.* (2005) Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.* 6, R33
6 Farre, D. *et al.* (2007) Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biol.* 8, R140
7 Zhu, J. *et al.* (2008) How many human genes can be defined as housekeeping with current expression data? *BMC Genomics* 9, 172
8 Castillo-Davis, C.I. *et al.* (2002) Selection for short introns in highly expressed genes. *Nat. Genet.* 31, 415–418
9 Vinogradov, A.E. (2001) Intron length and codon usage. *J. Mol. Evol.* 52, 2–5
10 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
11 Ren, X.Y. *et al.* (2006) In plants, highly expressed genes are the least compact. *Trends Genet.* 22, 528–532
12 Wong, G.K. *et al.* (2001) Most of the human genome is transcribed. *Genome Res.* 11, 1975–1977
13 Wang, J. *et al.* (2003) Vertebrate gene predictions and the problem of large genes. *Nat. Rev. Genet.* 4, 741–749
14 Vinogradov, A.E. (2006) 'Genome design' model and multicellular complexity: golden middle. *Nucleic Acids Res.* 34, 5906–5914
15 Li, S.W. *et al.* (2007) Selection for the miniaturization of highly expressed genes. *Biochem. Biophys. Res. Commun.* 360, 586–592
16 Alba, M.M. and Castresana, J. (2005) Inverse relationship between evolutionary rate and age of mammalian genes. *Mol. Biol. Evol.* 22, 598–606
17 Butler, J.E. and Kadonaga, J.T. (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.* 16, 2583–2592
18 Gardiner-Garden, M. and Frommer, M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.* 196, 261–282
19 Smale, S.T. and Kadonaga, J.T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem.* 72, 449–479
20 Carninci, P. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* 38, 626–635
21 Sandelin, A. *et al.* (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat. Rev. Genet.* 8, 424–436
22 Muller, F. *et al.* (2007) New problems in RNA polymerase II transcription initiation: matching the diversity of core promoters with a variety of promoter recognition factors. *J. Biol. Chem.* 282, 14685–14689
23 Rocha, E.P. (2006) The quest for the universals of protein evolution. *Trends Genet.* 22, 412–416
24 Larracuente, A.M. *et al.* (2008) Evolution of protein-coding genes in Drosophila. *Trends Genet.* 24, 114–123

doi:10.1016/j.tig.2008.08.004 Available online 9 September 2008