12 Gray, M.W. *et al.* (2001) The origin and early evolution of mitochondria. *Genome Biol.* 2, 1018
13 Gray, M.W. *et al.* (1999) Mitochondrial evolution. *Science* 283, 1476–1481
14 Lang, B.F. *et al.* (1999) A comparative genomics approach to the evolution of eukaryotes and their mitochondria. *J. Eukaryot. Microbiol.* 46, 320–326
15 Henze, K. and Martin, W. (2001) How do mitochondrial genes get into the nucleus? *Trends Genet.* 17, 383–387
16 Chinnery, P.F. (2003) Searching for nuclear-mitochondrial genes. *Trends Genet.* 19, 60–62
17 Emanuelsson, O. *et al.* (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300, 1005–1016
18 Nakai, K. and Horton, P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* 24, 34–36
19 Bannai, H. *et al.* (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics* 18, 298–305
20 Karlberg, O. *et al.* (2000) The dual origin of the yeast mitochondrial proteome. *Yeast* 17, 170–187
21 Marcotte, E.M. *et al.* (2000) Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 97, 12115–12120
22 Germot, A. *et al.* (1997) Evidence for loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in *Nosema locustae*. *Mol. Biochem. Parasitol.* 87, 159–168
23 Katinka, M.D. *et al.* (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* 414, 450–453
24 Martin, W. *et al.* (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl. Acad. Sci. U. S. A.* 99, 12246–12251
25 Millar, A.H. *et al.* (2001) Analysis of the *Arabidopsis* mitochondrial proteome. *Plant Physiol.* 127, 1711–1727
26 Taylor, S.W. *et al.* (2003) Global organellar proteomics. *Trends Biotechnol.* 21, 82–88
27 Leister, D. (2003) Chloroplast research in the genomic age. *Trends Genet.* 19, 47–56

0168-9525/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S0168-9525(03)00137-9

Human housekeeping genes are compact

Eli Eisenberg and Erez Y. Levanon

Compugen Ltd, 72 Pinchas Rosen Street, Tel Aviv 69512, Israel

We identify a set of 575 human genes that are expressed in all conditions tested in a publicly available database of microarray results. Based on this common occurrence, the set is expected to be rich in 'housekeeping' genes, showing constitutive expression in all tissues. We compare selected aspects of their genomic structure with a set of background genes. We find that the introns, untranslated regions and coding sequences of the housekeeping genes are shorter, indicating a selection for compactness in these genes.

The amazing diversity of the human body stems from the different expression patterns of genes in different tissues. Although most genes show constitutive expression in only a subset of tissues, some gene products are required for the maintenance of the basal cellular function and are constitutively found in all human cells. These genes are called housekeeping genes (HK genes) [1]. HK genes can be used to calibrate measurements of gene expression [2]. They might also help to define the minimal gene complement needed for a human cell [1]. Several attempts have been made recently to define the complete set of HK genes [3,4].

Microarrays are often used to identify sets of genes that are expressed either ubiquitously or in specific tissues or conditions. However, the technique is technically demanding and prone to artifacts, so independent evidence is often required to confirm the results. In principle, identifying the set of HK genes using microarray data is straightforward; one need only look for genes that are expressed in all tissues and all experimental conditions. Employing such an approach has so far resulted in two lists of HK genes [3,4]. However, problems in probe design, measurement noise and other artifacts introduce inevitable errors in such lists. Because a northern blot experiment for each gene in each tissue is impractical, an independent test is needed to validate any list of HK genes. Here, we report a validation test that uses a recently discovered property of highly expressed genes.

The transcription process is both slow and costly; it takes 50 milliseconds [5,6] and two ATP molecules [7] approximately to transcribe a nucleotide. This might be expected to provide selective pressure to make genes as short as functionally possible. The more copies of a gene required for the organism, the stronger this pressure should be. The first demonstration of this principle [8] showed that genes with a large number of expressed sequence tags (ESTs) in public libraries (and hence most mRNAs) have a significantly shorter average intron length than those with fewer ESTs.

Here, an implication of this principle is used to validate a set of HK genes. The HK genes, which are transcribed in all somatic cells and under all circumstances, are by nature highly expressed, and therefore should be selected to have shorter introns. We used a recently published database of microarray experiments [9] to identify a set of HK genes. As a further validation step, we checked the Gene Ontology (GO) annotation of these genes. We compared the structure of the HK genes with all other genes, and not only the introns, but all parts of the HK genes were found to be, on average, shorter than other genes. In particular, the untranslated regions and the translated proteins are all shorter in the HK genes.

Corresponding author: Eli Eisenberg (elie@compugen.co.il).

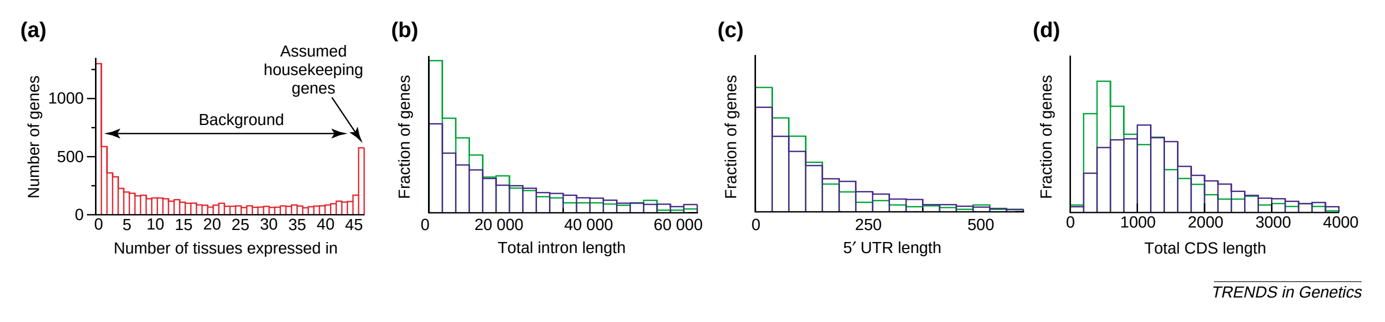

Fig. 1. (a) The distribution of 7500 RefSeq genes represented on the microarray as a function of the number of tissues they express in. Each bin gives the number of genes expressed in *M* out of 47 different tissues. The $M = 47$ bin corresponds to the housekeeping genes, expressed in all tissues. (b–d) Histograms comparing different parameters of gene structure. Green bars, HK genes; blue bars, non-HK genes. (b) Total length of introns. (c) Length of the 5′ untranslated region (UTR). (d) Length of coding sequence (CDS).

Assignment of housekeeping genes

A recently published database provides microarray expression data for Affymetrix U95A chip, containing 12 600 probes, and hybridized to 101 different samples [9] from 47 different human tissues and cell lines. These samples are mainly from the normal human physiological state, and therefore this dataset provides a description of the normal mammalian transcriptome.

We calculated the distribution of the number of different tissues in which a gene is expressed. Discarding probes for which the associated gene was not represented in the RefSeq database [10], and unifying all probes measuring the same gene (ignoring the potential differences among splice variants) yielded probes representing 7500 human genes. The experiments measuring replicates of the same biological condition were averaged to reduce the measurement noise, resulting in 47 data points per probe. We considered that a probe was expressed in a certain condition if its average reading was above a certain cutoff value. The results were not sensitive to the exact cut-off value, and we chose 200 standard Affymetrix average-difference units, considered to be a conservative cut-off value for determining gene presence [9]. This is also the trimmed average expression level in each tissue in accordance with the standard Affymetrix normalization procedure [11,12]. Thus, our HK genes are expressed in all tissues at an above-average level.

A histogram (Fig. 1a) of the number of genes expressed in exactly *M* of the 47 tissues shows a clear tendency for frequency to decrease as *M* increases. However, a substantial number of genes (575), belong to the class of genes that are expressed in all tissues. Because their number is far greater than expected based on the general trend described above, we assumed this class to be rich in HK genes, and considered it to be the set of HK genes.

It is noteworthy that the genes in our HK list tend to have an average expression significantly higher than other genes; the geometric mean expression of our HK genes is 1200 in Affymetrix average difference units, whereas that of other genes is 150. The difference cannot be accounted for by the cutoff used to define the HK genes, and is not a result of a bias due to inclusion of genes expressed in a few tissues only (data not shown).

Two additional tests were conducted to validate this set. First, a study of the GO annotation [13] of these genes revealed the set is rich in metabolic proteins (24%) and RNA-interacting proteins (19%, mostly ribosomal proteins). Second, we compiled a list of 18 well-established HK genes commonly used for quantitative PCR calibration [14,15], and checked our list against it. We found 13 of the 18 genes in our list, and the other five were not represented on the microarray (see Table in Supplementary Information at http://www.compugen.co.il/supp_info/Housekeeping_genes.html).

Length analysis of HK genes

Table 1 compares the lengths of various parts of the HK genes and the background genes. The alignment data was taken from the UCSC genome browser (http://genome.ucsc.edu) [16]. We excluded 322 genes that do not have a unique alignment, as well as 1242 genes that were not expressed in any tissue (to avoid potential problems because of defective probes). This left 532 HK genes and 5404 non-HK genes. The histograms in Fig. 1b–d compare HK genes with the other genes by total intron length, 5′ UTR length and coding sequence length. Remarkably, there was a statistically significant difference between HK and non-HK genes in all aspects of gene structure. Average intron length is shorter for the HK genes than for the background genes (2573 bp versus 5025 bp, respectively); total gene length is shorter (21 050 bp versus 53 418 bp); average exon length is shorter (212 bp versus 240 bp); average lengths of both 3′ and 5′ untranslated regions (UTRs) are shorter (5′: 135 bp versus 173 bp; 3′: 599 bp versus 846 bp); and, most notably, the translated proteins are shorter as well (403 amino acids versus 590 amino acids). Accordingly, the number of introns bp per unit of coding sequence length is lower for the HK genes (20 versus 32). We studied the structure of each gene as a function of the number of tissues it is expressed in and verified that the results are not due to bias of the non-HK genes by tissue-specific genes (data not shown).

The pronounced statistical characteristics of the HK gene set further supports their assignment as a unique set. Our findings confirm and extend previous research, showing that the introns of highly expressed genes are shorter [5]. As mentioned above, the HK genes expression levels are high, and the fact that they have to be expressed in all cells at all times makes them even more costly to transcribe. Previously [8], the high abundance of a certain gene in EST libraries was an indication the gene was highly expressed in the human body. It was pointed out [8],

Table 1. Human housekeeping genes are compact

	HK genes (*n* = 532)	Non-HK (*n* = 5404)	*P*-value
Average intron length	2573 ± 145 (*n* = 4353) 672	5025 ± 71 (*n* = 57447) 1365	4×10^{-130}
Total intron length	21050 ± 1781 9293	53418 ± 1425 20804	7×10^{-28}
Average exon length	212 ± 5 (*n* = 4885) 128	240 ± 2 (*n* = 62851) 132	9×10^{-5}
5′ UTR length	135 ± 8 79	173 ± 3 106	4×10^{-7}
3′ UTR length	599 ± 30 333	846 ± 13 552	3×10^{-13}
Coding sequence length	1211 ± 44 928	1770 ± 26 1322	3×10^{-26}
Number of exons	8.2 ± 0.3 6	10.6 ± 0.2 8	6×10^{-7}
Intron bps per coding bp	20 ± 2 9.9	31.8 ± 0.8 15.6	2×10^{-11}

Comparison of structure of housekeeping (HK) genes versus non-HK genes. For each case the first line gives the average value ± s.e.m, and the second line gives the median. For the average intron and exon lengths, all introns and exons belonging to the relevant set were included; the number appears in parentheses. The *P*-value was calculated using the Mann–Whitney test. UTR, untranslated region.

however, that this method is prone to bias due to the inclusion of normalized and tumor libraries and over-representation of certain tissues. Our approach overcomes this difficulty and confirms the previous result. Moreover, we find here that UTRs and even the encoded proteins are shorter for the HK genes. The magnitude of the difference is greater for the introns than for the exons and proteins (Table 1), which makes sense because the coding sequences and the UTRs are less susceptible to change.

It should be mentioned that intronless genes were included in our analysis after verifying that their inclusion or exclusion had no effect on the results. It also must be noted that the UTRs are not always fully sequenced, and thus their actual lengths might be longer. This bias was found to have no effect on the length of the coding sequences, and in any case the effect would be the same for both HK and non-HK genes.

It has been noted that codon usage bias in non-mammalian organisms is correlated with the expression level and with the gene length [17–19]. These results led to the conjecture of selective pressure on highly expressed genes resulting in shorter proteins [19]. However, no evidence for this selection was found [18], possibly because of a lack of high quality databases for these organisms. Recent works have suggested that there is no selection for codon usage bias in humans [20], and thus our results demonstrate that the expression–length correlation is not related to the expression–codon bias correlation.

It could be argued that selection towards shorter genes should have eliminated the introns in highly expressed genes. However, it is known that introns do have important roles, such as splicing regulation. Therefore, there is a balance between the advantageous contribution of the introns and the selective pressure for shortening.

Finally, when we compared our results with two (largely overlapping) published sets of HK genes, we found that roughly half of the genes in the intersection of those sets were present in our set. We used the genomic structure to test the remaining genes, and found a statistically significant difference between them and our HK gene set. The differences between our results and those of earlier studies [3,4] could be due to the fact that the database we used was based on more advanced chip technology and included many more different tissues, giving it more discriminative power to identify HK genes.

In conclusion, we have identified a set of HK genes. The set is publicly available at http://www.compugen.co.il/supp_info/Housekeeping_genes.html and can be used for calibration of microarrays, toxicity evaluation and quantitative PCR experiments. Furthermore, we show that HK genes have shorter introns, UTRs and coding sequences, attesting to the strong selection for compactness in these genes.

Acknowledgements

We thank Andrew Su for helpful discussion and for providing us with the RefSeq mapping. Gady Cojocaru and Rotem Sorek are acknowledged for comments on the manuscript and insightful discussion.

References

1 Butte, A.J. *et al.* (2001) Further defining housekeeping, or 'maintenance,' genes focus on 'a compendium of gene expression in normal human tissues'. *Physiol. Genomics* 7, 95–96

2 Gibson, U.E. *et al.* (1996) A novel method for real time quantitative RT–PCR. *Genome Res.* 6, 995–1001

3 Warrington, J.A. *et al.* (2000) Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol. Genomics* 2, 143–147

4 Hsiao, L.L. *et al.* (2001) A compendium of gene expression in normal human tissues. *Physiol. Genomics* 7, 97–104

5 Ucker, D.S. and Yamamoto, K.R. (1984) Early events in the stimulation of mammary tumor virus RNA synthesis by glucocorticoids. Novel assays of transcription rates. *J. Biol. Chem.* 259, 7416–7420

6 Izban, M.G. and Luse, D.S. (1992) Factor-stimulated RNA polymerase II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates. *J. Biol. Chem.* 267, 13647–13655

7 Lehninger, A.L. *et al.* (1982) Principles of biochemistry. *Biochemistry*, 615–644

8 Castillo-Davis, C.I. *et al.* (2002) Selection for short introns in highly expressed genes. *Nat. Genet.* 31, 415–418

9 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

10 Pruitt, K.D. *et al.* (2000) Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.* 16, 44–47

11 Lockhart, D.J. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14, 1675–1680

12 Wodicka, L. *et al.* (1997) Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* 15, 1359–1367

13 Gene Ontlology Consortium, (2001) Creating the gene ontology resource: design and implementation. *Genome Res.* 11, 1425–1433

14 Hamalainen, H.K. *et al.* (2001) Identification and validation of endogenous reference genes for expression profiling of T helper cell differentiation by quantitative real-time RT-PCR. *Anal. Biochem.* 299, 63–70

15 Lee, P.D. (2002) Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome Res.* 12, 292–297

16 Karolchik, D. *et al.* (2003) The UCSC genome browser database. *Nucleic Acids Res.* 31, 51–54

17 Akashi, H. (2001) Gene expression and molecular evolution. *Curr. Opin. Genet. Dev.* 11, 660–666

18 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

19 Moriyama, E.N. and Powell, J.R. (1998) Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic Acids Res.* 26, 3188–3193

20 Urrutia, A.O. and Hurst, L.D. (2001) Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics* 159, 1191–1199

0168-9525/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S0168-9525(03)00140-9

Strand misalignments lead to quasipalindrome correction

Vera van Noort[1], Peder Worning[2], David W. Ussery[2], William A. Rosche[3] and Richard R. Sinden[4]

[1]Nijmegen Center for Molecular Life Sciences, P/A Center for Molecular and Biomolecular Informatics, Nijmegen, The Netherlands
[2]Center for Biological Sequence Analysis, The Technical University of Denmark, DK–2800 Lyngby, Denmark
[3]Department of Biological Science, The University of Tulsa, Tulsa, Oklahoma 74104–3126, USA
[4]Laboratory of DNA Structure and Mutagenesis, Center for Genome Research, Institute of Biosciences and Technology, Texas A and M University System Health Sciences Center, Houston, TX 77030, USA

Quasipalindromes, or imperfect inverted repeats, undergo spontaneous mutation to more-perfect inverted repeats. These mutations have been observed in many organisms, ranging from bacteria to humans, where they are associated with mutations leading to disease. We determined the relative frequency of quasipalindromes and perfect palindromes in more than 100 sequenced prokaryotic genomes. In nearly all cases, perfect palindromes were relatively more frequent than quasipalindromes, suggesting that quasipalindrome correction is a general mechanism for mutation in prokaryotes.

Apart from simple misincorporation mutations, primer–template misalignments are probably the predominant cause of spontaneous mutations, and can lead to different types of mutation, such as frameshifts, deletions, duplications, inversions and complex mutations [1,2]. Misalignment requires sequence complementarity, such as direct or inverted repeats. Simple misalignment can occur along a linear template, and complex misalignment can be directed by DNA secondary structure. Imperfect inverted repeats, or quasipalindromes, can undergo spontaneous mutation to form a perfect inverted repeat (Fig. 1). Such mutations have been observed in bacteriophage T4, yeast and prokaryotes [3]. In addition, they have been associated with several human genetic diseases, including hereditary angioneurotic oedema, Duchenne muscular dystrophy, osteogenesis imperfecta, Lesch–Nyhan syndrome, and familial hypertension [4]. Here we provide evidence that a complex mutation, the correction of a quasipalindrome (an imperfect inverted repeat) to a palindrome (a perfect inverted repeat), occurs frequently in prokaryotes.

In 1982, Ripley [5] proposed two models for the correction of quasipalindromes to perfect inverted repeats: the intramolecular strand-switch model (also known as the hairpin-correction model), and the intermolecular strand-switch model (Fig. 2). In the latter model, the unpaired 3′ end of the nascent strand pairs with the quasipalindrome in the opposite template strand; that is, hybridization from the leading to lagging template strand. We have demonstrated that quasipalindrome correction in *Escherichia coli* occurs from a misalignment that is caused by an intermolecular strand switch, preferentially during leading strand replication [6]. Other quasipalindrome correction mutations might occur by an intramolecular (hairpin) replication mechanism [7].

Frequency of quasipalindromes within complete genomes

One would expect that correction of quasipalindromes over a period of time should result in an increase in the frequency of perfect palindromes within the bacterial genome. The repetitiveness of genomes has been investigated before [8], as well as frequencies of specific repeats in single genomes [9,10]. Here we test sequenced bacterial genomes for the frequencies of quasipalindromes and perfect palindromes, and compare them with expected values. In Fig. 3, we show that the relative frequencies of perfect palindromes are generally higher than the relative frequencies of quasipalindromes. In particular, the relative frequency of perfect palindromes in *E. coli* (orange triangle) is much higher than that of quasipalindromes.

Corresponding author: David W. Ussery (dave@cbs.dtu.dk).