Physiol Genomics 31: 96–103, 2007.
First published June 5, 2007; doi:10.1152/physiolgenomics.00183.2006.

Abundance of dinucleotide repeats and gene expression are inversely correlated: a role for gene function in addition to intron length

Vineet K. Sharma, Naveen Kumar, Samir K. Brahmachari, and Srinivasan Ramachandran

G. N. Ramachandran Knowledge Centre for Genome Informatics, Institute of Genomics and Integrative Biology, Delhi, India

Submitted 22 August 2006; accepted in final form 30 May 2007

Sharma VK, Kumar N, Brahmachari SK, Ramachandran S. Abundance of dinucleotide repeats and gene expression are inversely correlated: a role for gene function in addition to intron length. *Physiol Genomics* 31: 96–103, 2007. First published June 5, 2007; doi:10.1152/physiolgenomics.00183.2006.—High and broad transcription of eukaryotic genes is facilitated by cost minimization, clustered localization in the genome, elevated G+C content, and low nucleosome formation potential. In this scenario, illumination of correlation between abundance of $(TG/CA)_{n\geq 12}$ repeats, which are negative *cis* modulators of transcription, and transcriptional levels and other commonly occurring dinucleotide repeats, is required. Three independent microarray datasets were used to examine the correlation of $(TG/CA)_{n\geq 12}$ and other dinucleotide repeats with gene expression. Compared with the expected equi-distribution pattern under neutral model, highly transcribed genes were poor in repeats, and conversely, weakly transcribed genes were rich in repeats. Furthermore, the inverse correlation between repeat abundance and transcriptional levels appears to be a global phenomenon encompassing all genes regardless of their breadth of transcription. This selective pattern of exclusion of $(TG/CA)_{n\geq 12}$ and $(AT)_{n\geq 12}$ repeats in highly transcribed genes is an additional factor along with cost minimization and elevated GC, and therefore, multiple factors govern high transcription of genes. We observed that even after controlling for the effects of GC and average intron lengths, the effect of repeats albeit somewhat weaker was persistent and definite. In the ribosomal protein coding genes, sequence analysis of orthologs suggests that negative selection for repeats perhaps occurred early in evolution. These observations suggest that negative selection of $(TG/CA)_{n\geq 12}$ microsatellites in the evolution of the highly expressed genes was also controlled by gene function in addition to intron length.

transcription; microarray; regulation; $(TG/CA)_n$ repeats

EUKARYOTIC TRANSCRIPTION IS a slow, costly, and complex process involving the interaction of several proteins and regulatory sequences (8). About 18–25 nucleotides are transcribed per second incurring an expenditure of at least two ATP molecules per transcribed nucleotide (22–24). Given this inherently demanding process, several intrinsic factors operating at the level of template DNA have been identified that serve to facilitate the attainment of high transcriptional levels in eukaryotes. Among these, gene length appears to play a significant role. It was observed that highly and broadly transcribed genes (across a wide variety of tissues) are generally shorter compared with weakly transcribed genes (7, 11, 47, 48). These observations suggest that selection for short length in highly transcribed genes is perhaps driven by a necessity to minimize energy expenditure during transcription. Highly and broadly transcribed genes generally carry out housekeeping functions essential for maintenance of cellular physiology. Furthermore, Akashi and Gojobori (3) pointed out that ribosomal proteins, which are housekeeping, use higher frequency of less costly amino acids. Thus minimization of the energy budgets invested in gene expression occurs at all levels in a cell (47).

Although length is an important determinant of gene transcription, this by itself is insufficient. Eukaryotic DNA is hierarchically organized, and therefore other factors in addition to length must play important roles in controlling the level of gene transcription. It was observed that broadly transcribed genes tend to be clustered in the genome (25–26). Similarly, Lercher et al. (26) and Marin et al. (30) observed that highly transcribed genes have elevated (G+C) content. Since a negative correlation was observed between the nucleosome formation potential and the (G+C) content of genes (48), these observations taken together suggest that genes that are clustered and highly and broadly transcribed are located in regions of "open chromatin."

The length, G+C content and the nucleosome formation potential of gene sequences are intrinsic properties stemming from sequence patterns and organization. Another important intrinsic property of the sequence of a gene is its spatial structure. Several years ago it was pointed out that transcription and supercoiling of the template DNA are interrelated, and therefore, the topology adopted by the template DNA is likely to affect transcription (17, 27, 49, 51). The normal structure of DNA in the cell is the B-form, a right-handed helix with 10.4 base pairs per turn. But small segments consisting of unusual compositional characteristics such as alternating purine-pyrimidine simple sequences $(CG)_n$ or $(TG/CA)_n$ display propensity to adopt a Z-form, a left-handed helix with 12 base pairs per turn under conditions close to physiological (18, 31). Such a transition from a B-helical to a Z-form occurs in the negatively supercoiled DNA due to presence of these repeats and affects the movement of RNA polymerase (Fig. 1) (34). In vitro experiments show that this type of inhibition of transcription is most pronounced at CG type sequence motifs (34).

Among all the dinucleotide repeats, $(CG)_n$ repeats are under represented in the human genome, whereas $(TG/CA)_n$ repeats are the most abundant (50%) followed by $(AT)_n$ repeats (35%) and $(GA/TC)_n$ repeats (15%) (12, 13, 39). The remaining dinucleotide repeats show very low distribution. In recent years, experimental evidence has been accumulating on the role of uninterrupted $(TG/CA)_n$ repeats as *cis*-modulators of transcription (2, 10, 16, 31, 32, 35, 39–43). The direction and extent of transcriptional modulation by $(TG/CA)_n$ repeats vary among genes. However, in most cases, it was observed that $(TG/CA)_n$ repeats of length $n \geq 12$ units exert a downregulatory effect on transcription. Moreover, this negative modula-

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: S. Ramachandran, G. N. Ramachandran Knowledge Centre for Genome Informatics, Inst. of Genomics and Integrative Biology, Mall Road, Delhi 110 007, India (e-mail: ramu@igib.res.in; ramuigib@gmail.com).

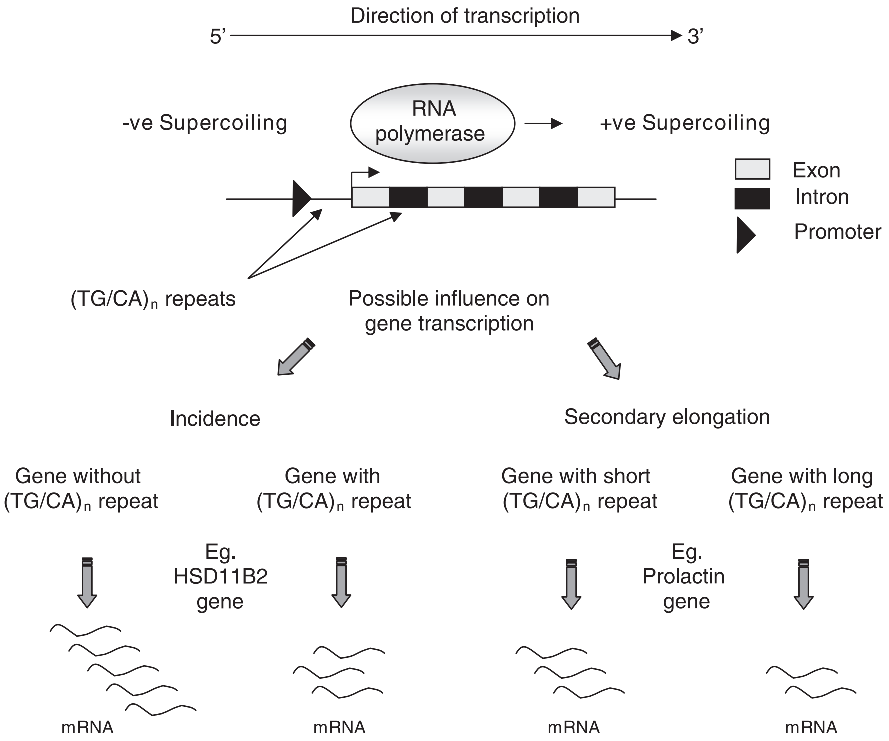

Direction of transcription
5'
3'
-ve Supercoiling
RNA polymerase
+ve Supercoiling
Exon
Intron
Promoter
$(TG/CA)_n$ repeats
Possible influence on gene transcription
Incidence
Secondary elongation
Gene without $(TG/CA)_n$ repeat
Gene with $(TG/CA)_n$ repeat
Gene with short $(TG/CA)_n$ repeat
Gene with long $(TG/CA)_n$ repeat
Eg. HSD11B2 gene
Eg. Prolactin gene
mRNA
mRNA
mRNA
mRNA

Fig. 1. Modulation of transcription due to incidence or secondary elongation of $(TG/CA)_{n \geq 12}$ repeats.

tory effect increases with increasing length of repeats (Fig. 1). The influence of $(TG/CA)_n$ repeats most likely arises from the adoption of a non-B form of the template DNA undergoing transcription. Among the other dinucleotide repeats, $(TA)_n$ and $(GA/TC)_n$ repeats are also shown to have influence on gene expression (5, 6, 9, 28). In this scenario, we focus on the examination of association or abundance of four basic types of dinucleotide repeats $(TG/CA)_n$, $(AT)_n$, $(GA/TC)_n$, and $(GC)_n$ in human genes expressed at different levels. The remaining dinucleotide repeat types are equivalent to these four basic types (13). Such an analysis can illuminate another intrinsic feature of the template DNA, namely, the correlation of transcriptional activity with abundance of uninterrupted dinucleotide repeats.

METHODS

Microarray Datasets

Dataset A. Normalized human gene expression data obtained from the Affymetrix array (HG U95Av2) experiments using blood leukocytes from 13 human individuals including monozygotic twins is available at Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo) under following accession numbers: GSM14477, GSM14478, GSM14479, GSM14480, GSM14481, GSM14482, GSM14483, GSM14485, GSM20645, GSM29053, GSM29054, GSM29055, GSM29056, GSM29057, and GSM29058 (38). Only genes with present "P" call were considered.

Dataset B. Normalized human gene expression data from the Affymetrix array (HG U95A) experiments were obtained from the Gene Expression Atlas database (http://expression.gnf.org), which contains information on gene expression from 46 different human tissues, organs, and cell lines (44). Only genes with average difference values >200 units were considered (44).

Dataset C. Normalized human gene expression data from Affymetrix array (HG U133A) experiments were obtained from the GNF SymAtlas database (http://symatlas.gnf.org/SymAtlas/), which contains information on gene expression from 79 different human tissues, organs, and cell lines (45). Only genes with present "P" call were considered.

Filtering and Binning

A sum of 212 genes from *dataset A*, 192 genes from *dataset B*, and 188 genes from *dataset C*, whose expression varied in monozygotic twins (38), were filtered out. Five additional genes were also removed from *dataset B* because of ambiguous annotations. The remaining *dataset A* had 5,015 genes, *dataset B* had 6,650 genes, and *dataset C* had 11,017 genes. The signal intensities (*datasets A, C*) and average difference values (*dataset B*) were transformed by taking logarithms to the base 10. The log-transformed values were averaged across all the samples. This value, corresponding to log of geometric mean, was considered as the average transcriptional level of a given gene (47). Genes were arranged in increasing order of transcriptional levels, and 5% of the total number of genes was sliced out for each bin. These bins were numbered in an increasing order of bin average of average transcriptional levels of genes. In *dataset A*, the bins from *1–15* contained 251 genes each, and the remaining *bins 16–20* contained 250 genes each. In *dataset B*, *bins 1–10* contained 333 genes each, and the remaining *bins 11–20* contained 332 genes. In *dataset C*, the *bins 1–17* contained 551 genes each, and the remaining *bins 18–20* contained 550 genes. The bin average value of transcription of genes in each bin was used for further analysis.

Housekeeping Genes

The datasets of human housekeeping genes were obtained from (HuGEIndex) (19) and Eisenberg and Levanon (11). Out of 451 housekeeping genes from HuGEIndex, 418 had known HGNC gene symbols. Out of 575 housekeeping genes identified by Eisenberg and Levanon, 565 had known HGNC gene symbols. We repeated the binning exercise after removing the housekeeping genes belonging to the three datasets. A total of 285, 279, and 306 HuGEIndex housekeeping genes were removed from *datasets A, B, and C,* respectively. Similarly a total of 407, 408, and 412 housekeeping genes of the

Eisenberg and Levanon dataset were removed from *datasets A, B, and C*, respectively.

Sequence Retrieval and Mapping of (TG/CA)$_{n\geq 12}$ Repeats

Sequences of human genes (latest updated as of Nov.–Dec. 2006) were retrieved from National Center for Biotechnology Information Entrez (50). The start and end of a gene are considered as the nucleotide sequence from the 5′-end of first exon to the 3′-end of last exon including untranslated regions at 5′-end and 3′-end (UTRs). If alternate splicing was reported, the gene length considered was the longest, including all alternatively spliced products for that gene (48).

Uninterrupted (TG/CA)$_{n\geq 12}$ repeats (type II and type III) were identified as described previously (39, 41). Repeats of these types have been shown to modulate transcription (2, 10, 16, 31, 32, 35, 39–43). Furthermore, repeats of this length have also been shown to have preferential binding to nuclear factors compared with short repeats (14) and can also stimulate mRNA splicing (20–21). Similarly, uninterrupted (AT)$_{n\geq 12}$ and (GA/TC)$_{n\geq 12}$ repeats were also identified to analyze the association of these repeats with gene expression (5, 6, 9, 28). All repeats were scored in the intragenic region (exons, introns, and UTRs) only.

Functional Classification of Gene Families for Comparative Analysis

The genes belonging to bins that showed significantly higher or lower than expected proportion of (TG/CA)$_{n\geq 12}$ repeats were aligned against the KOG database (46) by stand-alone version of BLASTX (50), and the best hit for each gene (E $< 10^{-6}$) was selected for annotation with KOG ID. Using this KOG ID, we classified the genes into 25 KOG functional categories and then grouped them into six functional classes namely, "information," comprising A, J, K, and L categories of KOG; "cell cycle," comprising B and D categories of KOG; "metabolism," comprising C, E, F, G, H, I, P, and Q categories of KOG; "signaling and communication," comprising O, T, and U categories of KOG; "immune and related functions," comprising V category of KOG; and "structure and motility," comprising M, N, W, Y, and Z categories of KOG. These six broad functional classes were defined based on the scheme suggested by Adams et al. (1) and Andrade et al. (4) (also see Refs. 37–39, 41). Two KOG categories (R, S) for "general function prediction only" and "unknown" functions were not included in the functional analysis.

Statistical Methods

Tests of significance for the differences between the proportions of genes containing dinucleotide repeats in different bins compared with global distribution of these repeats in the datasets (*A, B,* or *C*) were carried out by the binomial proportions test. The observed proportion in each bin was tested against the expected proportion, which was computed assuming no preference with respect to bin average transcriptional levels. The null hypothesis was "under neutral model," an equi-distribution of proportion of genes with repeats is expected across all bins. Pearson's product-moment correlation (R) was computed to examine the relationship between bin average transcription and *1*) proportion of genes with repeats in different bins, *2*) average intron and exon lengths of genes in bins, and *3*) average GC content of genes in bins. The Pearson's product-moment correlations (R) between average intron lengths and proportion of genes with repeats were also examined. The R statistical package (36) was used to perform the statistical tests. Partial correlations and their significance values between average transcription and repeats controlling for GC content and average intron lengths were computed with pcor and pcor.test from the ggm library of the R statistical package.

RESULTS

We used three publicly available human microarray gene expression datasets, which we designate "A" (38), "B" (44), and "C" (45). *Datasets A* and *B* were produced using HG U95A arrays, and *dataset C* was produced using HG U133A arrays from Affymetrix (http://www.affymetrix.com). *Dataset A* was obtained from blood leucocytes drawn from 13 normal human individuals (38); *dataset B* was obtained from 46 different human tissues, organs, and cell lines (44); and *dataset C* was obtained from 79 human tissues, organs, and cell lines (45). These datasets taken together offer gene expression data from both natural (from living individuals, *dataset A*) and artificially preserved states (from stored tissue samples, *datasets B* and *C*).

We applied two procedures to reduce or eliminate confounding effects due to random noise, since stochastic noise is an inherent property of gene expression in living systems (15). First, we considered the logarithmically transformed geometric mean values of signal intensities or average difference values to represent the transcriptional level of a given gene (47). Second, we removed genes whose transcriptional levels varied randomly due to environmental causes (38). Application of this sieving procedure to any dataset will likely reduce random noise due to extrinsic environmental fluctuations (38). The resulting filtered datasets ("*A*": 5,015 genes, "*B*": 6,650 genes, and "*C*": 11,017 genes) would be more suitable for examining the role of intrinsic factors of DNA such as the dinucleotide repeats including the most abundant (TG/CA)$_{n\geq 12}$ repeats. Furthermore, genes were partitioned into bin sizes each of 5% of the total numbers of genes in each dataset and numbered in increasing order of bin average transcriptional levels of genes in each bin. This strategy holds the potential to unravel clear correlation patterns as opposed to using individual genes (26, 47). All three datasets compare well with respect to the transcriptional levels in all bins (*dataset A* and *B*: R = 0.93, $P <$ 0.0001; *dataset A* and *C*: R = 0.99, $P < 0.0001$; *dataset B* and *C*: R = 0.97, $P < 0.0001$; Supplementary Fig. S1[1]) and therefore can be used as independent resource datasets for examination of hypotheses.

Examination of the proportion of genes with uninterrupted intragenic (TG/CA)$_{n\geq 12}$ repeats in each bin and the bin average transcription revealed an inverse relationship (Fig. 2). A significant negative correlation was observed in all datasets (*dataset A*: R = −0.95, $P < 0.0001$; *dataset B*: R = −0.93, $P <$ 0.0001; *dataset C*: R = −0.93, $P < 0.0001$). These results show that highly transcribed genes are poorly populated with (TG/CA)$_{n\geq 12}$ repeats. The bins with highly transcribed genes (*dataset A*: 15, 18–20; *dataset B*: 16–20; *dataset C*: 17–20) had lower than expected proportion of genes with (TG/CA)$_n$ repeats ($P < 0.05$ to $P < 0.0001$). Conversely, bins with weakly transcribed genes (*dataset A*: 1–4, 6; *dataset B*: 1, 2, 6, 7, 9; *dataset C*: 1–5, 8) had higher than expected proportion of genes with (TG/CA)$_{n\geq 12}$ repeats ($P < 0.04$ to $P < 0.0001$). It is apparent that an overall agreement exists between the three datasets with respect to the general trend. It is to be noted that in the remaining bins, the differences between the expected and observed proportion of genes with (TG/CA)$_{n\geq 12}$ repeats did not show a clear statistically significant difference.

To examine whether such a pattern is specific to (TG/CA)$_n$ repeats we also analyzed the proportion of other dinucleotide repeats (GA/TC)$_{n\geq 12}$, (AT)$_{n\geq 12}$, and (GC)$_{n\geq 12}$ in the three

[1] The online version of this article contains supplemental material.

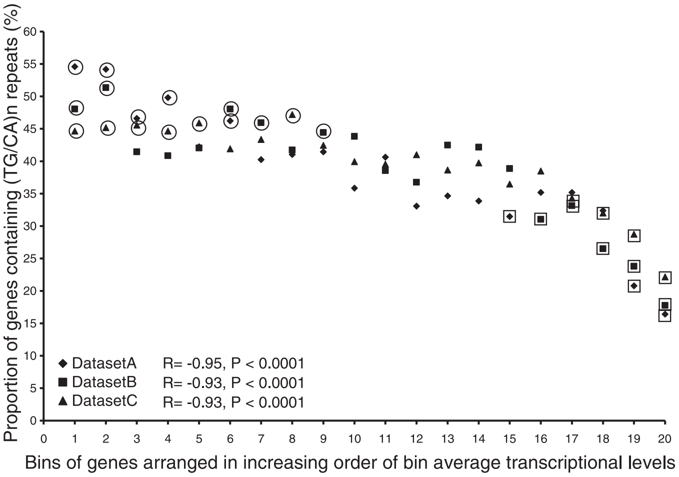


Fig. 2. Inverse relationship between the proportion of genes with uninterrupted intragenic $(TG/CA)_{n\geq12}$ repeats and gene transcriptional levels. The correlation values (R) and the associated confidence values are displayed. Ensquared points: bins of genes with statistically significant lower than expected proportion of genes with repeats ($P < 0.05$ to $P < 0.0001$); encircled points: bins of genes with statistically significant higher than expected proportion of genes with repeats ($P < 0.04$ to $P < 0.0001$).

datasets. The $(TG/CA)_n$ repeats are most abundant followed by $(AT)_n$, $(GA/TC)_n$, and $(GC)_n$(13). The $(GC)_n$ repeats showed a very scarce distribution in all groups of the three datasets and were not considered for statistical analysis. Among the remaining two types of repeats, proportion of genes with $(AT)_{n\geq12}$ repeats also showed a strong negative correlation with average transcriptional levels (*dataset A*: $R = -0.94$, $P < 0.0001$; *dataset B*: R = -0.95, $P < 0.0001$; *dataset C*: R = -0.90, $P < 0.0001$). The bins with highly transcribed genes (*dataset A*: 14, 16, 18–20; *dataset B*: 16–20; *dataset C*: 17–20) had lower than expected proportion of genes with $(AT)_{n\geq12}$ repeats ($P < 0.05$ to $P < 0.0001$). Conversely, bins with weakly transcribed genes (*dataset A*: 1–5; *dataset B*: 1, 2, 5, 6; *dataset C*: 1, 3, 5, 8) had higher than expected proportion of genes with $(AT)_{n\geq12}$ repeats ($P < 0.04$ to $P < 0.0001$). Comparatively, the proportion of genes with $(GA/TC)_{n\geq12}$ repeats showed weaker negative correlation with average transcriptional levels (*dataset A*: R = -0.87, $P < 0.0001$; *dataset B*: R = -0.76, $P < 0.0002$; *dataset C*: R = -0.54, $P < 0.02$).

In our previous analysis, we observed that human housekeeping genes with $(TG/CA)_{n\geq12}$ repeats had lower average transcriptional levels compared with those without repeats (41). Housekeeping genes are compact, highly transcribed, and poor in $(TG/CA)_{n\geq12}$ repeats (11, 41). In all *datasets A, B,* and *C*, we observed that most of the housekeeping genes were located in bins of highly transcribed genes. This pattern is in accordance with previous observations (11, 25).

To resolve whether the observed inverse relationship between transcriptional levels and incidence of $(TG/CA)_{n\geq12}$ and other dinucleotide repeats is a global phenomenon or is biased due to presence of highly transcribed housekeeping genes, we repeated the analysis after removing the housekeeping genes. The inverse relationship was persistent, and the strength of this relationship was maintained even when the housekeeping genes from Eisenberg and Levanon (11) and from Hsiao et al. (19) were removed respectively from *datasets A, B,* and *C*. The respective correlations are: *dataset A*: R = -0.91, $P < 0.0001$ and R = -0.93, $P < 0.0001$; *dataset B*: R = -0.93, $P < 0.0001$ and R = -0.93, $P < 0.0001$; *dataset C*: R = -0.92, $P < 0.0001$ and R = -0.92, $P < 0.0001$. The $(AT)_{n\geq12}$ repeats also displayed similar pattern (*dataset A*: R = -0.88, $P < 0.0001$ and R = -0.96, $P < 0.0001$; *dataset B*: R = -0.91, $P < 0.0001$, R = -0.93, $P < 0.0001$; *dataset C*: R = -0.88, $P < 0.0001$, R = -0.87, $P < 0.0001$). Such a persisting trend was much less evident in the case of $(GA/TC)_n$ repeats (*dataset A*: R = -0.50, $P < 0.03$ and R = -0.84, $P < 0.0001$; *dataset B*: R = -0.61, $P < 0.004$, R = -0.71, $P < 0.0004$; *dataset C*: R = -0.47, $P < 0.04$, R = -0.45, $P < 0.04$). It is apparent from Supplementary Tables S1, S2, and S3 that $(TG/CA)_{n\geq12}$ repeats and $(AT)_{n\geq12}$ repeats are positively associated with weakly expressed genes and negatively associated with highly expressed genes in all three datasets regardless of the breadth of transcription. However, $(GA/TC)_{n\geq12}$ repeats showed a much weaker correlation.

The relationship between average intron lengths of genes and their bin average transcriptional levels displayed negative correlation (Fig. 3), reconfirming the previous observations (7) that highly transcribed genes have short introns (*dataset A*: $R = -0.97$, $P < 0.0001$; *dataset B*: R = -0.93, $P < 0.0001$; *dataset C*: R = -0.86, $P < 0.0001$). The average exon lengths of genes and bin average transcriptional levels in all datasets were also negatively correlated (*dataset A*: R = -0.9, $P <0.0001$; *dataset B*: $R = -0.95$, $P < 0.0001$; *dataset C*: R = -0.87, $P < 0.0001$).

A positive correlation was observed between average intron lengths and proportion of genes with uninterrupted intragenic $(TG/CA)_{n\geq12}$ repeats (Supplementary Fig. S2, *dataset A*: R = 0.94, $P < 0.0001$; Supplementary Fig. D3, *dataset B*: R = 0.90, $P < 0.0001$; Supplementary Fig. D4, *dataset C*: R = 0.96, $P < 0.0001$). Similarly, a positive correlation was observed between average intron lengths and proportion of genes with uninterrupted intragenic $(AT)_{n\geq12}$ repeats (*dataset A*: R = 0.90, $P < 0.0001$; *dataset B*: R = 0.95, $P < 0.0001$; *dataset C*: R = 0.93, $P < 0.0001$) and $(GA/TC)_{n\geq12}$ repeats (*dataset A*: R = 0.96, $P < 0.0001$; *dataset B*: R = 0.77, $P < 0.0001$; *dataset C*: R = 0.78, $P < 0.0001$).

A strong positive correlation was also observed between the average GC [%(G+C)] content of genes in each bin and bin average transcriptional levels in all datasets (Fig. 4, *dataset A*:

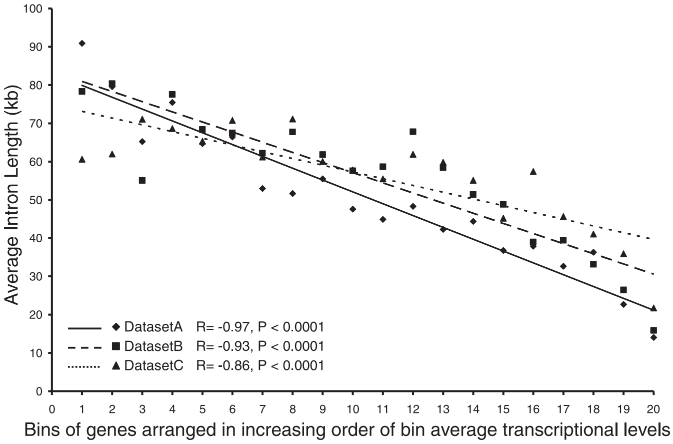


Fig. 3. Negative correlation between average intron lengths and bin average transcriptional levels.

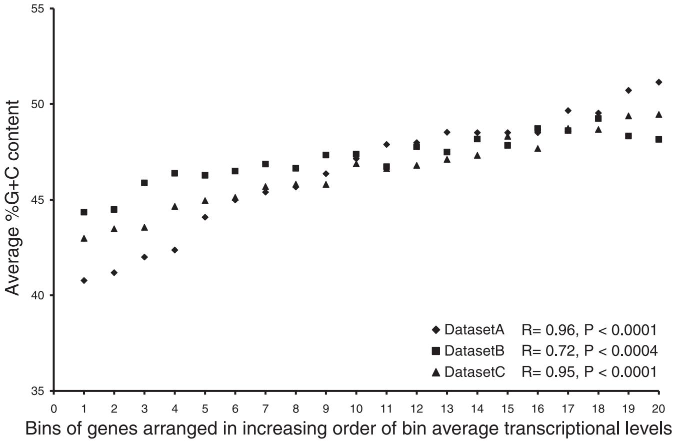


Fig. 4. Positive correlation between average GC content (%) and bin average transcription levels.

R = 0.96, $P < 0.0001$; *dataset B*: R = 0.72, $P < 0.0004$; *dataset C*: R = 0.95, $P < 0.0001$), which is consistent with the previous observations that highly transcribed genes have high GC content (25, 30). Furthermore, an inverse relationship was observed between average GC content and proportion of genes with $(TG/CA)_{n\geq12}$ repeats in each bin (*dataset A*: R = −0.93, $P < 0.0001$; *dataset B*: R = −0.73, $P < 0.0002$; *dataset C*: R = −0.86, $P < 0.0001$). An inverse relationship was also observed between the proportion of genes with $(AT)_{n\geq12}$ repeats and GC content of genes in each bin (*dataset A*: R = −0.94, $P < 0.0001$; *dataset B*: R = −0.83, $P < 0.0001$; *dataset C*: R = −0.86, $P < 0.0001$) showing that GC content of genes and the distribution of $(AT)_{n\geq12}$ repeats are inversely correlated. However, the $(GA/TC)_{n\geq12}$ repeats showed a weaker correlation (*dataset A*: R = −0.87, $P < 0.0001$; *dataset B*: R = −0.73, $P < 0.0003$; *dataset C*: R = −0.53, $P < 0.02$).

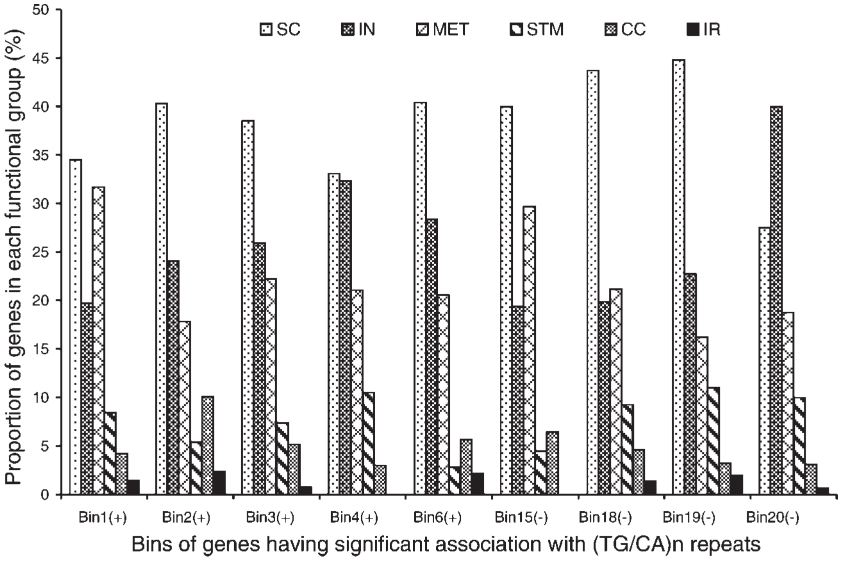
To examine the role of repeats in the context of all three factors, namely, proportion of genes with repeats, average intron length, and GC content on average transcription, we carried out a partial correlation coefficient analysis between average transcription and proportion of genes with repeats by controlling the effects of average intron lengths (most uninterrupted intragenic repeats were located in introns) and %GC content. Partial correlations for each repeat type were computed individually. Even after controlling the effects of both GC content and average intron lengths, we observed significant negative correlation of bin average transcription with $(TG/CA)_{n\geq12}$ repeats in *datasets B* and *C* (*dataset B*: pR = −0.56, $P < 0.02$; *dataset C*: pR = −0.58, $P < 0.02$; Supplementary Table S4). Similarly, proportion of genes with $(AT)_{n\geq12}$ repeats showed significant correlation with bin average transcription (*dataset A*: pR = −0.61, $P < 0.008$; *dataset B*: pR = −0.70, $P < 0.002$). The correlations of proportion of genes with $(GA/TC)_{n\geq12}$ repeats were not statistically significant after controlling for effects of GC content and average intron lengths.

We repeated the foregoing analysis after removing the two sets of housekeeping genes. After the housekeeping gene sets of Hsiao et al. (19) and of Eisenberg and Levanon (11), respectively, were removed, the partial negative correlations between proportion of genes with $(TG/CA)_{n\geq12}$ repeats and bin average transcription were persistent (*dataset B*: pR = −0.61, $P < 0.007$ and *dataset B*: pR = −0.59, $P < 0.02$, *dataset C*: pR = −0.50, $P < 0.04$). Similarly, the partial negative correlation between proportion of genes with $(AT)_{n\geq12}$ repeats and bin average transcription were persistent (*dataset A*: pR = −0.60, $P < 0.009$; *dataset B*: pR = −0.50, $P < 0.04$; and *dataset A*: pR = −0.50, $P < 0.04$). In the case of proportion of genes with $(GA/TC)_{n\geq12}$ repeats the partial correlations after controlling for effects of GC and average intron lengths were not statistically significant.

DISCUSSION

Most (>98%) uninterrupted intragenic $(TG/CA)_{n\geq12}$, $(AT)_{n\geq12}$, and $(GA/TC)_{n\geq12}$ repeats were located in introns. Under the "neutral model," the low abundance of repeats in highly transcribed genes could be due to a secondary consequence of selection either for short introns or for elevated GC composition. The observation that genes with long introns harbor more repeats compared with genes with short introns suggests a strong role for intron length as a controlling factor for abun-

Fig. 5. Functional composition of bins of genes with statistically significant occurrence (high or low) of $(TG/CA)_{n\geq12}$ repeats in *dataset A*. Key: (+) Bins with higher occurrence of repeats; (−) bins with lower occurrence of repeats. CC, cell cycle; IN, information; IR, immune and related functions; MET, metabolism; SC, signaling and communication; STM, structure and motility.



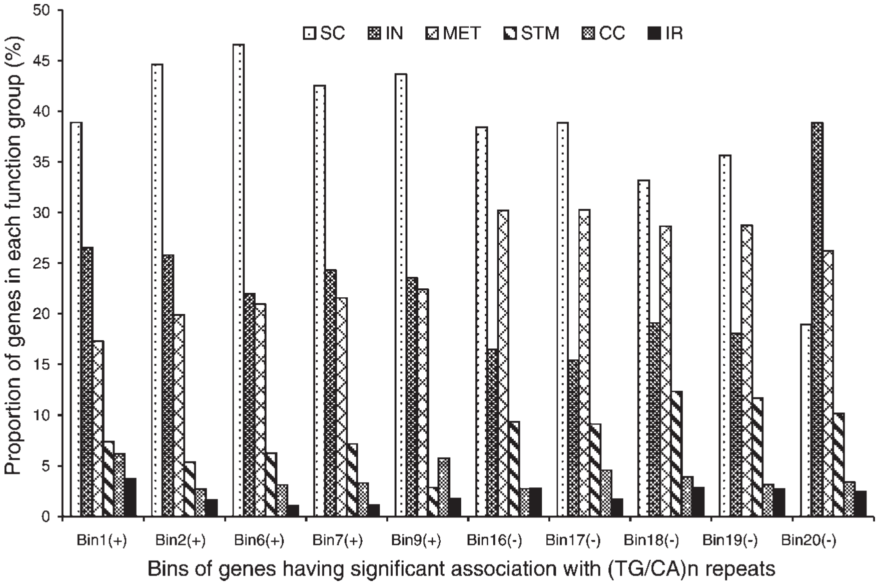


Fig. 6. Functional composition of bins of genes with statistically significant occurrence (high or low) of $(TG/CA)_{n\geq12}$ repeats in *dataset B*. See Fig. 5 legend for abbreviations and symbols.

dance of repeats. However, we found that in case of *datasets A, B,* and *C*, 66, 61, and 56% of the longest introns (≥10 kb) of the topmost highly expressed genes (*bin 19* and *20*) contained no $(TG/CA)_{n\geq12}$ repeats; in contrast to the remaining groups (*bin 1–18*) of *datasets A, B,* and *C*, 56, 54, and 54% of the longest introns (≥10 kb) contained no $(TG/CA)_{n\geq12}$ repeats. Therefore, it appears that although long introns offer more space for accommodation of additional elements, the highly expressed genes tend to have lower proportion of $(TG/CA)_{n\geq12}$ repeats.

It is apparent that the $(TG/CA)_{n\geq12}$ repeats span only a small fraction (0.095–0.145%) of the lengths of introns of genes in different bins in all three datasets. Therefore it is unlikely that the presence of these repeats would increase lengths of introns severely, thereby inflating transcriptional costs compared with large insertions such as transposons whose lengths average 300 bp (7). Therefore, selection for short introns is unlikely to be a prime mover in eliminating microsatellites, which rarely exceed 23 repeat units (46 bp) (12, 39).

In the case of $(GA/TC)_{n\geq12}$ repeats, it has been described that these repeats tend to repress transcription by stabilizing nucleosomes (9, 28). Although the trends in the case of these repeats are less strong compared with $(TG/CA)_{n\geq12}$ and $(AT)_{n\geq12}$ repeats, it appears that the $(GA/TC)_{n\geq12}$ repeats are also negatively selected for in highly expressed genes. The results of the partial correlation analysis showed that even after controlling the effects of GC and average intron length, we found that $(TG/CA)_{n\geq12}$ and $(AT)_{n\geq12}$ repeats displayed significant correlations with average transcription. In other words, although the role of repeats in the context of GC content and average intron lengths appears to be somewhat weaker, their role in transcription is clearly evident. Thus it appears that attainment of high expression is accompanied by the correlated interplay of multiple factors such as intron length, repeat content, and GC content.

Analysis of functional composition of the genes in bins which showed significant association with $(TG/CA)_{n\geq12}$ repeats (Figs. 5–7) showed that, the proportion of genes belong-

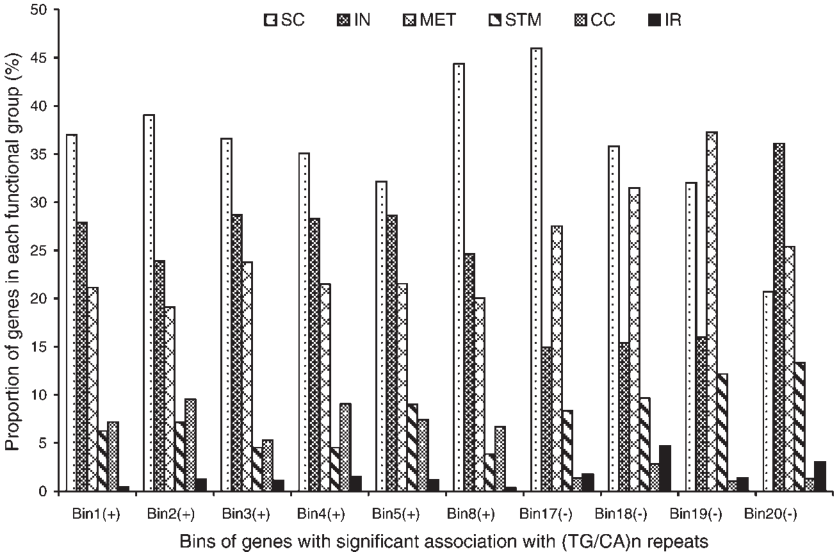


Fig. 7. Functional composition of bins of genes with statistically significant occurrence (high or low) of $(TG/CA)_{n\geq12}$ repeats in *dataset C*. See Fig. 5 legend for abbreviations and symbols.

ing to information and metabolism classes is higher in highly transcribed genes (*bin 20*) compared with genes of the signaling and communication class, which was abundant in the rest of the bins. Recently, we observed that $(TG/CA)_{n\geq12}$ repeats are positively associated with genes of signaling and communication, whereas they were negatively associated with the information class (39, 41). Therefore, it is probable that the low occurrence of $(TG/CA)_{n\geq12}$ repeats and other repeats in highly transcribed human genes is also controlled by gene function (39, 41).

The list of highly transcribed genes includes those coding for ribosomal proteins and structural proteins in accordance with previous observations reported from the analysis of human expressed sequence tags and microarray data from other nonprimate genomes (7). Comparison of the sequences of orthologs of ribosomal protein coding genes revealed that all human ribosomal protein coding genes are devoid of $(TG/CA)_{n\geq12}$, $(AT)_{n\geq12}$, and $(GA/TC)_{n\geq12}$ microsatellites. These genes are perhaps under purifying selection (7) because all of their homologs from the fruit fly *Drosophila melanogaster* or most homologs (93.4%) from the mouse *Mus musculus* spanning >990 million yr are also devoid of these microsatellites. These observations suggest that $(TG/CA)_{n\geq12}$ and other microsatellites were negatively selected for early in the evolution of the highly expressed genes and controlled by gene function in addition to intron length.

GRANTS

V. K. Sharma thanks the Council of Scientific and Industrial Research (CSIR) for a Senior Research Fellowship. S. Ramachandran and S. K. Brahmachari thank the CSIR for Grant CMM0017 under the task force “In Silico Biology for Drug Target Identification.”

REFERENCES

1. **Adams MD, Kerlavage AR, Fleischmann RD, Fuldner RA, Bult CJ, Lee NH, Kirkness EF, Weinstock KG, Gocayne JD, White O, Sutton G, Blake JA, Brandon RC, Chiu M, Clayton RA, Cline RT, Cotton MD, Earle-Hughes J, Fine LD, FitzGerald LM, FitzHugh WM, Fritchman JL, Geoghagen NSM, Glodek A, Gnehm CL, Hanna MC, Hedblom E, Hinkle PS Jr, Kelley JM, Klimek KM, Kelley JC, Liu L, Marmaros SM, Merrick JM, Moreno-Palanques RF, McDonald LA, Nguyen DT, Pellegrino SM, Phillips CA, Ryder SE, Scott JL, Saudek DM, Shirley R, Small KV, Spriggs TA, Utterbach TR, Weidman JF, Li Y, Barthlow R, Bednarik DP, Cao L, Cepeda MA, Coleman TA, Collins E, Dimke D, Feng P, Ferrie A, Fischer C, Hastings GA, He W, Hu J, Huddleston KA, Greene JM, Gruber J, Hudson P, Kim A, Kozak DL, Kunsch C, Ji H, Li H, Meissner PS, Olsen H, Raymond L, Wei Y, Wing J, Xu C, Yu G, Ruben SM, Dillon PJ, Fannon MR, Rosen CA, Haseltine WA, Fields C, Fraser CM, Venter JC.** Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* 377: 3–174, 1995.
2. **Agarwal AK, Giacchetti G, Lavery G, Nikkila H, Palermo M, McTernan C, Bianchi G, Manunta P, Strazzullo P, Mantero F, White PC, Stewart PM.** CA-Repeat polymorphism in intron 1 of *HSD11B2*: effects on gene expression and salt sensitivity. *Hypertension* 36: 187–94, 2000.
3. **Akashi H, Gojobori T.** Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc Natl Acad Sci USA* 99: 3695–3700, 2002.
4. **Andrade MA, Ouzounis C, Sander C, Tamames J, Valencia A.** Functional classes in the three domains of life. *J Mol Evol* 49: 551–557, 1999.
5. **Beutler E, Gelbart T, Demina A.** Racial variability in the UDP-glucuronosyltransferase 1 (UGT1A1) promoter: a balanced polymorphism for regulation of bilirubin metabolism. *Proc Natl Acad Sci USA* 95: 8170–8174, 1998.
6. **Bharaj B, Scorilas A, Giai M, Diamandis EP.** TA repeat polymorphism of the 5alpha-reductase gene and breast cancer. *Cancer Epidemiol Biomarkers Prev* 9: 387–393, 2000.
7. **Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA.** Selection for short introns in highly expressed genes. *Nat Genet* 31: 415–418, 2002.
8. **Coulson RM, Ouzounis CA.** The phylogenetic diversity of eukaryotic transcription. *Nucleic Acids Res* 31: 653–660, 2003.
9. **Croston GE, Kerrigan LA, Lira LM, Marshak DR, Kadonaga JT.** Sequence-specific antirepression of histone H1-mediated inhibition of basal RNA polymerase II transcription. *Science* 251: 643–649, 1991.
10. **Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, Weissenbach J.** A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 380: 152–154, 1996.
11. **Eisenberg E, Levanon EY.** Human housekeeping genes are compact. *Trends Genet* 19: 362–365, 2003.
12. **Ellegren H.** Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet* 16: 551–558, 2000.
13. **Ellegren H.** Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* 5: 435–445, 2004.
14. **Epplen JT, Kyas A, Maueler W.** Genomic simple repetitive DNAs are targets for differential binding of nuclear proteins. *FEBS Lett* 389: 92–95, 1996.
15. **Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB.** Noise minimization in eukaryotic gene expression. *PLoS Biol* 2: e137, 2004.
16. **Gebhardt F, Zanker KS, Brandt B.** Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *J Biol Chem* 274: 13176–13180, 1999.
17. **Giaever GN, Wang JC.** Supercoiling of intracellular DNA can occur in eukaryotic cells. *Cell* 55: 849–856, 1988.
18. **Haniford DB, Pulleyblank DE.** The in-vivo occurrence of Z DNA. *J Biomol Struct Dyn* 1: 593–609, 1983.
19. **Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, Dillon W, Lee KF, Clark KE, Haverty P, Weng Z, Mutter GL, Frosch MP, Macdonald ME, Milford EL, Crum CP, Bueno R, Pratt RE, Mahadevappa M, Warrington JA, Stephanopoulos G, Stephanopoulos G, Gullans SR.** A compendium of gene expression in normal human tissues. *Physiol Genomics* 7: 97–104, 2001.
20. **Hui J, Reither G, Bindereif A.** Novel functional role of CA repeats and hnRNP L in RNA stability. *RNA* 9: 931–936, 2003.
21. **Hui J, Stangl K, Lane WS, Bindereif A.** HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats. *Nat Struct Biol* 10: 33–37, 2003.
22. **Irvine KD, Helfand SL, Hogness DS.** The large upstream control region of the Drosophila homeotic gene Ultrabithorax. *Development* 111: 407–424, 1991.
23. **Izban MG, Luse DS.** Factor-stimulated RNA polymerase II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates. *Biol Chem* 267: 13647–13655, 1992.
24. **Lehninger AL, Nelson DL, Cox MM.** Principles of Biochemistry. New York: Worth, 1982, p. 615–644.
25. **Lercher MJ, Urrutia AO, Hurst LD.** Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet* 31: 180–183, 2002.
26. **Lercher MJ, Urrutia AO, Pavlicek A, Hurst LD.** A unification of mosaic structures in the human genome. *Hum Mol Genet* 12: 2411–2415, 2003.
27. **Liu LF, Wang JC.** Supercoiling of the DNA template during transcription. *Proc Natl Acad Sci USA* 84: 7024–7027, 1987.
28. **Lu Q, Wallrath LL, Granok H, Elgin SC.** (CT)n (GA)n repeats and heat shock elements have distinct roles in chromatin structure and transcriptional activation of the Drosophila hsp26 gene. *Mol Cell Biol* 13: 2802–2814, 1993.
29. **Lunter G, Hein J.** A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics* 20, *Suppl* 1: I216–I223, 2004.
30. **Marin A, Gallardo M, Kato Y, Shirahige K, Gutierrez G, Ohta K, Aguilera A.** Relationship between G+C content, ORF-length and mRNA concentration in *Saccharomyces cerevisiae*. *Yeast* 20: 703–711, 2003.
31. **Meera G, Ramesh N, Brahmachari SK.** Zintrons in rat alpha-lactalbumin gene. *FEBS Lett* 251: 245–249, 1989.
32. **Naylor LH, Clark EM.** $d(TG)_n.d(CA)_n$ sequences upstream of the rat prolactin gene form Z-DNA and inhibit gene transcription. *Nucleic Acids Res* 18: 1595–1601, 1990.
33. **Nordheim A, Rich A.** The sequence (dC-dA)n X (dG-dT)n forms left-handed Z-DNA in negatively supercoiled plasmids. *Proc Natl Acad Sci USA* 80: 1821–1825, 1983.

34. **Peck LJ, Wang JC.** Transcriptional block caused by a negative supercoiling induced structural change in an alternating CG sequence. *Cell*: 129–137, 1985.
35. **Pravica V, Asderakis A, Perrey C, Hajeer A, Sinnott PJ, Hutchinson IV.** In vitro production of IFN-gamma correlates with CA repeat polymorphism in the human IFN-gamma gene. *Eur J Immunogenet* 26: 1–3, 1999.
36. **R Foundation for Statistical Computing.** *R: a Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing, 2006.
37. **Rivera MC, Jain R, Moore JE, Lake JA.** Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci USA* 95: 6239–6244, 1998.
38. **Sharma A, Sharma VK, Horn-Saban S, Lancet D, Ramachandran S, Brahmachari SK.** Assessing natural variations in gene expression in humans by comparing with monozygotic twins using microarrays. *Physiol Genomics* 21: 117–123, 2005.
39. **Sharma VK, Brahmachari SK, Ramachandran S.** (TG/CA)$_n$ repeats in human gene families: abundance and selective patterns of distribution according to function and gene length. *BMC Genomics* 6: 83, 2005.
40. **Sharma VK, Sharma A, Kumar N, Khandelwal M, Mandapati KK, Horn-Saban S, Strichman-Almashanu L, Lancet D, Brahmachari SK, Ramachandran S.** Expoldb: expression linked polymorphism database with inbuilt tools for analysis of expression and simple repeats. *BMC Genomics* 13: 258, 2006.
41. **Sharma VK, Rao CB, Sharma A, Brahmachari SK, Ramachandran S.** (TG/CA)$_n$ repeats in human housekeeping genes. *J Biomol Struct Dyn* 21: 303–310, 2003.
42. **Shimajiri S, Arima N, Tanimoto A, Murata Y, Hamada T, Wang KY, Sasaguri Y.** Shortened microsatellite d(CA)21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. *FEBS Lett* 455: 70–74, 1999.
43. **Streelman JT, Kocher TD.** Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia. *Physiol Genomics* 9: 1–4, 2002.
44. **Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB.** Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci USA* 99: 4465–4470, 2002.
45. **Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB.** A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 101: 6062–6067, 2004.
46. **Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA.** The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 11: 41, 2003.
47. **Urrutia AO, Hurst LD.** The signature of selection mediated by expression on human genes. *Genome Res* 13: 2260–2264, 2003.
48. **Vinogradov AE.** Noncoding DNA, isochores and gene expression: nucleosome formation potential. *Nucleic Acids Res* 33: 559–563, 2005.
49. **Wang JC, Lynch AS.** Transcription and DNA supercoiling. *Curr Opin Genet Dev* 3: 764–768, 1993.
50. **Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequoia E, Suzek TO, Tatusova TA, Wagner L.** Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res* 1: D35–D40, 2004.
51. **Wu HY, Shyy SH, Wang JC, Liu LF.** Transcription generates positively and negatively supercoiled domains in the template. *Cell* 53: 433–440, 1988.