Gene 436 (2009) 12–22

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Repetitive DNA elements, nucleosome binding and human gene expression

Ahsan Huda [a], Leonardo Mariño-Ramírez [b,c], David Landsman [b], I. King Jordan [a,*]

[a] School of Biology, Georgia Institute of Technology, 310 Ferst Drive, Atlanta, GA 30332, USA
[b] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
[c] Computational Biology and Bioinformatics Unit, Biotechnology and Bioindustry Center, Corporacion Colombiana de Investigacion Agropecuaria - CORPOICA, Km. 14 Via a Mosquera, Bogota, Colombia

ARTICLE INFO

Article history:
Received 22 January 2009
Accepted 23 January 2009
Available online 5 February 2009

Received by M. Batzer

Keywords:
Transposable elements
Nucleosome binding
Epigenetics
Simple sequence repeats
Gene regulation
Promoter architecture
Human genome

ABSTRACT

We evaluated the epigenetic contributions of repetitive DNA elements to human gene regulation. Human proximal promoter sequences show distinct distributions of transposable elements (TEs) and simple sequence repeats (SSRs). TEs are enriched distal from transcriptional start sites (TSSs) and their frequency decreases closer to TSSs, being largely absent from the core promoter region. SSRs, on the other hand, are found at low frequency distal to the TSS and then increase in frequency starting ~150 bp upstream of the TSS. The peak of SSR density is centered around the −35 bp position where the basal transcriptional machinery assembles. These trends in repetitive sequence distribution are strongly correlated, positively for TEs and negatively for SSRs, with relative nucleosome binding affinities along the promoters. Nucleosomes bind with highest probability distal from the TSS and the nucleosome binding affinity steadily decreases reaching its nadir just upstream of the TSS at the same point where SSR frequency is at its highest. Promoters that are enriched for TEs are more highly and broadly expressed, on average, than promoters that are devoid of TEs. In addition, promoters that have similar repetitive DNA profiles regulate genes that have more similar expression patterns and encode proteins with more similar functions than promoters that differ with respect to their repetitive DNA. Furthermore, distinct repetitive DNA promoter profiles are correlated with tissue-specific patterns of expression. These observations indicate that repetitive DNA elements mediate chromatin accessibility in proximal promoter regions and the repeat content of promoters is relevant to both gene expression and function.

1. Introduction

The prevalence of repetitive DNA sequences in mammalian genomes has been appreciated since the classic re-association kinetic (COT-curve) experiments of the late nineteen-sixties (Britten and Kohne, 1968). The completion of the human genome projects at the turn of the millennium further underscored the extent to which the human genome sequence is made up of repetitive DNA elements (Lander et al., 2001; Venter et al., 2001). There are several distinct categories of repetitive sequence elements in the human genome. Interspersed repeat sequences, also known as transposable elements (TEs), make up at least 45% of the euchromatic genome sequence, and novel human TE families continue to be discovered and characterized (Wang et al., 2005; Nishihara et al., 2006). Simple sequence repeats (SSRs) consist of tandem repeats of exact or nearly exact units of length k (k-mers), with $k=1$–13 corresponding to microsatellites and $k=1$–500 for minisatellites. Analysis of the human genome sequence showed that ~3% of the euchromatic sequence was made up of SSRs, and both SSRs and TEs are thought to be far more abundant in heterochromatin. Segmental duplications of 1–200 kb were initially shown to account for ~3% of the human genome sequence (Lander et al., 2001), and more recent results reveal that copy number variants populate the genome to an even greater extent (Kidd et al., 2008).

The evolutionary significance and the functional role that repetitive genomic elements, TEs in particular, play has long been a matter of speculation and inquiry. Once regarded as selfish, or parasitic, genomic elements with little or no phenotypic relevance (Doolittle and Sapienza, 1980; Orgel and Crick, 1980), it has since become apparent that TEs make substantial contributions to the structure, function and evolution of their host genomes (Kidwell and Lisch, 2001). Perhaps the most significant functional effect that TEs have had on their host genomes is manifest through the donation of regulatory sequences that control the expression of nearby genes (Feschotte, 2008). Studies of TE regulatory effects have focused, for the most part, on discrete well characterized regulatory elements such as transcription factor binding sites (Jordan et al., 2003; van de Lagemaat et al., 2003; Wang et al., 2007), enhancers (Bejerano et al., 2006) and alternative promoters (Dunn et al., 2003; Conley et al., 2008). A number of recent studies have also outlined the contributions of TEs to regulatory RNA genes (Smalheiser and Torvik, 2005; Borchert et al., 2006; Piriyapongsa and Jordan, 2007; Piriyapongsa et al., 2007). For this study, we sought to analyze the contribution of

Abbreviations: TE, transposable element; SSR, simple sequence repeat; TSS, transcriptional start site; GNF2, Novartis mammalian gene expresssion atlas 2.

* Corresponding author. Tel.: +1 404 385 2224; fax: +1 404 894 0519.
E-mail address: king.jordan@biology.gatech.edu (I.K. Jordan).

0378-1119/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2009.01.013

repetitive DNA to epigenetic aspects of gene regulation, specifically the relationship between repetitive DNA elements and the chromatin environment of human promoter sequences.

Genomic DNA in eukaryotes is wrapped around histone proteins and packaged into repeating subunits of chromatin called nucleosomes (Kornberg and Lorch, 1999). The importance of specific genomic sequences in determining the binding locations of nucleosomes has recently been confirmed (Segal et al., 2006). A number of factors point to a relationship between repetitive DNA elements, the local chromatin environment and epigenetic gene regulation. Densely compact heterochromatin is enriched for both TEs and SSRs in a number eukaryotic organisms (Dimitri and Junakovic, 1999). Heterochromatin functions to mitigate potentially deleterious effects associated with TEs by repressing both element transcription and ectopic recombination between dispersed element sequences (Grewal and Jia, 2007). In fact, it has been proposed that heterochromatin originally evolved to serve as a genome defense mechanism by silencing TEs (Henikoff and Matzke, 1997; Henikoff, 2000). In the plant Arabidopsis, *de novo* heterochromatin formation can be caused by insertions of TEs into euchromatin, and TEs are able to epigenetically silence genes when they are inserted nearby or inside them (Lippman et al., 2004). In other words, TEs have been shown to cause specific *in situ* changes in the chromatin environment that can spread locally and regulate gene expression in a way that is region-specific but sequence-independent (*i.e.* epigenetic).

The previously established connections between genome repeats, chromatin environment and gene regulation for model organisms, taken together with the repeat-rich nature of the human genome, suggest that repetitive sequence elements may play a role in regulating human gene expression by modulating the local chromatin environment. Specifically, we hypothesized that gene regulatory related differences in nucleosome binding at human promoter sequences are mediated in part by repetitive genomic elements. We evaluated the relationship between nucleosome binding, repetitive element promoter distributions and human gene expression to test this idea. Human proximal promoter sequences were characterized with respect to both their repetitive DNA architectures and predicted nucleosome binding affinities, and the repetitive DNA environment of the promoters was considered with respect to patterns of gene expression.

2. Materials and methods

2.1. Promoter sequence analysis

Our analysis focused on proximal promoter sequence regions, which we define for a gene as ranging from −1 kb at the 5′ end to the transcription start (TSS) at the 3′ end. We relied on the Database of Transcriptional Start Sites (DBTSS) to identify experimentally characterized TSS, based on aligned full-length cDNA sequences, in the human genome (Suzuki et al., 2002). These TSS were mapped to the March 2006 human genome reference sequence (NCBI Build 36.1) and used to extract 1 kb proximal promoter sequences as described previously (Marino-Ramirez et al., 2004; Tharakaraman et al., 2005). This procedure was used to ensure analysis of the most accurate set of human proximal promoter sequences possible. For the additional three mammalian species analyzed – chimpanzee (*Pan troglodytes*), mouse (*Mus musculus*) and rat (*Rattus norvegicus*) – the locations of proximal promoter sequences were determined based on the 5′ most position of NCBI Refseq gene models (Pruitt et al., 2007). These positions were used to download 1 kb proximal promoter sequences from the latest respective genome builds for each organism from the UCSC Genome Browser (Karolchik et al., 2003): chimpanzee $n=24{,}170$, mouse $n=20{,}589$ and rat $n=8737$.

The program RepeatMasker (Smit et al., 1996–2004) was used to detect and annotate repetitive elements in the proximal promoter sequences. RepeatMasker was run using 500 bp of flanking sequence on either end of the proximal promoter regions analyzed to avoid edge effects in the detection of repeats. Repetitive elements detected by RepeatMasker were broken down into two main categories: interspersed repeats, also known as transposable elements (TEs), and simple sequence repeats (SSRs). SSRs may be annotated as low complexity sequences and correspond to runs of repeating *k*-mers where $k=1$–13 bp for microsatellites and $k=14$–500 for minisatellites. TEs were further divided into specific classes: LINEs, SINEs, LTR and DNA as well as specific families L1 and Alu.

Proximal promoter sequences, including 500 bp flanks, were analyzed using the Nucleosome Prediction software developed by the Segal lab (Segal et al., 2006). This software was used to calculate the probability of each nucleotide being occupied by a nucleosome in all promoter sequences. These nucleosome occupancy probabilities are based on the periodicity of dinucleotides – AA/TT/TA – that are a characteristic of genomic sequences that have been experimentally isolated as bound to nucleosomes. Predictions for the relative placement of nucleosomes along genomic sequence are further informed by a thermodynamic stability model. The nucleosome prediction model used in our analysis is based on experimentally characterized nucleosome bound sequences reported for chicken (Satchwell et al., 1986). The chicken model has been proven accurate when used on other vertebrate genomes (Segal et al., 2006). For sets of promoter sequences, nucleosome occupancy averages were calculated over each position of the 1 kb proximal promoter regions and these average values were taken as the position-specific nucleosome binding affinities (nba) reported here.

Two sets of promoter sequence randomizations were done and position-specific nucleosome binding affinities were re-calculated on the randomized sequence sets. The first randomization consisted of randomly shuffling entire 1 kb proximal promoter sequences. This has the effect of maintaining overall nucleotide composition of the promoter sequences while changing the dinucleotide composition as well as any regional nucleotide biases along the promoters. The second randomization procedure consisted on randomly shuffling non-overlapping 100 bp windows along the promoter sequences in place. This has the effect of maintaining both overall and local nucleotide compositions of the promoters while changing the dinucleotide composition.

2.2. Repeat-based promoter clustering

Human proximal promoter sequences were clustered solely based on their repetitive DNA architectures. To do this, we generated 1000-unit vectors that represent the position-specific repeat content for each promoter sequence. A discrete value was assigned to each promoter sequence position (nucleotide) in the following manner:

$$X_i = \begin{cases} 1 & \text{if the nucleotide is part of a TE sequence} \\ -1 & \text{if the nucleotide is part of a SSR sequence} \\ 0 & \text{if the nucleotide is part of a non}-\text{repetitive sequence} \end{cases}$$

where X_i represents the nucleotide at position *i*.

Promoter sequence repeat vectors were then clustered using a combination of *k*-means clustering ($k=5$, 10, 20) and Self Organized Mapping using the program Genesis (Sturn et al., 2002). We found that using *k*-means clustering with $k=5$ followed by a Self Organized Map generated the most coherent clusters in terms of the repeat content of the vectors.

2.3. Gene expression analysis

We used version 2 of the Novartis mammalian gene expression atlas (GNF2), which provides replicate Affymetrix microarray data for 44,775 probes across 79 human tissues (Su et al., 2004). GNF2 expression data, in the form of Affymetrix signal intensity values, were obtained from the UCSC Table Browser (Karolchik et al., 2004), and Affymetrix probes were mapped to NCBI Refseq identifiers using the UCSC Table Browser tools. For each gene, the average, maximum and breadth of expression were computed across the 79 tissues in the GNF2 data set. Expression breadth is taken as the number of tissues where the gene has a signal intensity value of >350. Co-expression between gene pairs was measured by computing the Pearson correlation coefficient (r) between pairs of gene-specific expression signal intensity vectors:

$$g_i = [t_1, t_2 \dots t_{79}]$$

where g_i is the ith gene and t_n is the expression level for that gene in the nth tissue.

For each repeat-specific promoter cluster, the average r-value for all pairwise comparisons between genes in the cluster was computed. In addition, the difference (*diff*) between the cluster-specific r-value averages (cluster-r) and all possible pairwise r-values between genes (all-r) was computed for each cluster:

$$diff = \text{cluster}-r-\text{all}-r.$$

The significance of these differences was computed using the normal deviate:

$$z = diff / se_{diff}$$

where se_{diff} is the standard error of the difference.

2.4. Probabilistic analysis of promoter repeats

We used a probabilistic representation of the repeat content of the human proximal promoter sequence clusters in order to derive gene (promoter)-specific similarity scores that indicate the probability that any human gene (promoter) belongs to a specific repeat cluster. To do this, each proximal promoter sequence (1 kb upstream of the TSS) in a cluster was divided into 20 non-overlapping windows of 50 bp each. For each window (w), the probability (p) of the occurrence of a TE nucleotide, or SSR nucleotide or a non-repetitive (NR) nucleotide was calculated separately using the following formula:

$$p(b, w) = \frac{f_{b,w} + s(b)}{N + \sum_{b' \in \{T,S,N\}} s(b')}$$

where $f_{b,w}$ = counts of base b in window w and b represents counts of either TE nucleotides, or SSR nucleotides or non-repetitive nucleotides, N = number of sites in the window (50) and $s(b)$ = a pseudocount function. The probabilities thus calculated for each window were averaged for all promoters in the cluster. This procedure was repeated to yield repetitive DNA probabilistic representation models for each of the six promoter clusters.

All the proximal promoter sequences analyzed were then scored against each of the six cluster-specific probabilistic models using a log-likelihood ratio approach illustrated as follows:

$$LL_{b,w} = \ln \sum_{TE,SSR,NR} f_{b,w} \ln \frac{f_{b,w}}{f_b}$$

where $f_{b,w} = p_{b,w} \times 50$, which is the model frequency used as background. Promoter-specific scores (S) were then computed as the sum of log-likelihood ratios over the 20 windows of 50 bp each:

$$S = \sum_{w=1}^{20} LL_{b,w}.$$

Using this method, we scored all genes (promoters) against each of the six cluster models to generate six cluster-specific gene (promoter) score vectors. This modeling and scoring method is a modification of the approach used to score sequence motifs, such as transcription factor binding sites, based on motif-characteristic position-weight matrices (Wasserman and Sandelin, 2004).

In order to relate promoter sequence repetitive DNA architecture to tissue-specific gene expression, the gene (promoter)-specific probabilistic repeat cluster scores were correlated with tissue-specific gene expression signal intensity values for each of the 79 tissues in GNF. This was repeated with gene (promoter)-specific scores assigned to each gene for each of the six repeat clusters. For example, for the cluster1 ($c1$) versus tissue1 ($t1$) comparison:

$$c1 = [S_{g1}, S_{g2} \dots S_{g7913}] \text{x } t1 = [e_{g1}, e_{g2} \dots e_{g7913}]$$

where g_i is the ith gene, S is the score for the cluster1 model and e is the expression level for that gene in tissue1. In other words, each gene analyzed is assigned a repeat probability score for each of the six clusters, and these six sets of repeat probability promoter scores are individually correlated with the GNF2 tissue-specific expression values for the genes. This procedure resulted in a 6×79 matrix of correlation values.

2.5. Gene Ontology (GO) analysis

GO annotation terms (Ashburner et al., 2000) for human genes were obtained from the Gene Ontology Annotation database (http://www.ebi.ac.uk/GOA/). GO terms were further mapped to higher level GO slim categories. Expected versus observed frequencies of GO slim terms were compared using χ^2 tests for each promoter repeat cluster, as well as for the combined TE− and TE+ groups, in order to look for over-represented GO slim categories. The pairwise similarity between GO terms was computed using modified semantic similarity method (Lord et al., 2003; Azuaje et al., 2005) as described previously (Marino-Ramirez et al., 2006; Tsaparas et al., 2006). The GO similarity difference (*GOdiff*) was calculated between the average pairwise similarity for GO terms from pairs of genes within TE groups (*e.g.* TE+) and the average pairwise GO similarity for all possible pairs of genes:

$$GOdiff = GOsim-(TE+)-GOsim-(all).$$

The significance of the difference was measured using the normal deviate as described for the gene expression analysis.

2.6. Statistical analysis

Standard statistical tests were used to compare population means for pairwise (Student's t-test) and for multiple comparisons (ANOVA), to correlated vectors of nucleosome binding affinities, TE and SSR densities, expression and promoter score values (Pearson correlation coefficient), to control for the confounding effects of multiple variables on correlation values obtained (partial correlation) and to evaluate the difference between observed and expected GO terms (χ^2) (Zar, 1999).

3. Results and discussion

3.1. Repetitive DNA and nucleosome binding affinity

Experimentally characterized human gene proximal promoter sequences ($n = 7913$) were taken from the Database of Transcriptional Start Sites (DBTSS) (Suzuki et al., 2002) and analyzed with respect to their repetitive DNA content and nucleosome binding affinities. The locations of repetitive DNA elements along promoter sequences were determined by the RepeatMasker program and nucleosome binding affinities were predicted using the method of (Segal et al. (2006). Two classes of repetitive DNA were analyzed separately: interspersed repeats, also known as transposable elements (TEs) and simple sequence repeats (SSRs), which are made up of runs of exact or nearly exact repeating *k*-mers. For each promoter position, from 1 kb upstream to the transcriptional start site (TSS), the average TE and SSR densities over all promoter sequences were calculated as the fraction of sequences for which that position was occupied by a TE or SSR. Average nucleosome binding affinities across promoter positions were calculated as the fraction of sequences for which a given position was predicted to be occupied (bound) by a nucleosome. Average nucleosome binding affinities and the average TE density follow parallel trends along the proximal promoter regions (Fig. 1a). Nucleosomes bind more tightly and TEs are found more frequently distal to the TSS, whereas nucleosomes bind promoter sequences most proximal to the TSS with lower affinity and TEs are rarely found close to the TSS. SSRs show a distinctly different trend with a higher density close to the TSS that corresponds to the decrease in nucleosome binding affinity. The SSR density matches the nucleosome binding even more closely than the TE density just upstream of the TSS. Nucleosome binding affinities decrease steadily from distal regions until ~35 bp upstream of the TSS, then the nucleosome binding affinity increases towards the TSS. Similarly, the SSR density increases to the same point and then drops off as the nucleosome binding affinity increases (Fig. 1a). This core promoter region where nucleosome binding affinity is at its lowest and SSR density is at its highest corresponds to the location where the basal transcriptional machinery assembles, and RNA polymerase II binds, to initiate transcription.

The correlations between nucleosome binding affinities with TE and SSR densities along human proximal promoter regions are robust and highly statistically significant (Fig. 1b). Previously, we observed that nucleotide composition changes markedly along human proximal promoter sequences with an increase in CpG frequency close to the TSS (Marino-Ramirez et al., 2004), while the nucleosome binding

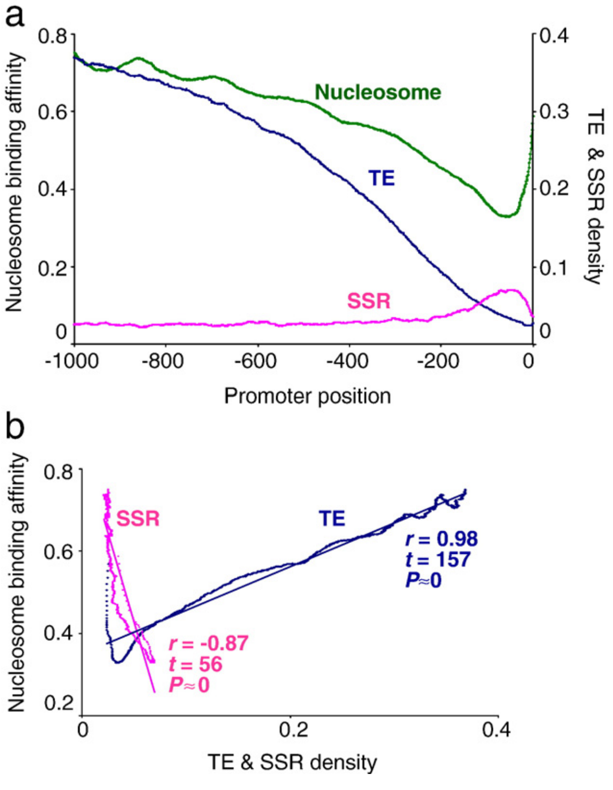


Fig. 1. Repetitive DNA density and nucleosome binding affinity along human proximal promoter sequences. (a) Average nucleosome binding affinities (green line, values on left *y*-axis) along with average TE densities (blue line, values on right *y*-axis) and average SSR densities (pink line, values on right *y*-axis) over 7913 human proximal promoter sequences are plotted over each promoter position starting from − 1000 bp upstream and progressing to the transcriptional start site (TSS at position 0). (b) Linear trends and correlations relating position-specific nucleosome binding affinities (*y*-axis) to TE (blue) and SSR (pink) densities (*x*-axis) are shown. Statistical significance levels of the *r*-values are based on the Student's *t*-distribution with $df = n - 2 = 998$ where $t = r^* \mathrm{sqrt}((n - 2) / (1 - r^2))$.

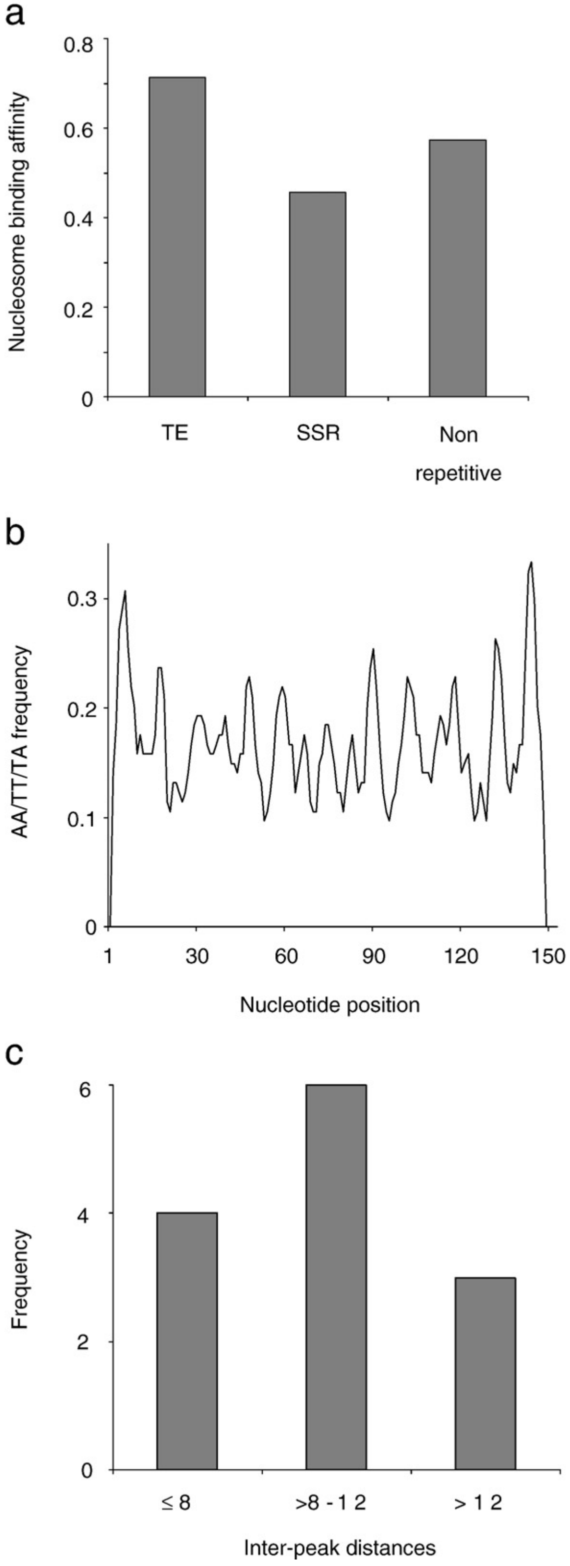


Fig. 2. Nucleosome binding properties for repetitive versus non-repetitive DNA. (a) Average predicted nucleosome binding affinities are shown for TE, SSR and non-repetitive human promoter sequences. (b) Periodicity of the nucleosome binding (wrapping) characteristic dinucleotides AA/TT/TA are shown for 39 experimentally characterized nucleosome bound TE sequences from chicken. (c) Histogram showing the inter-peak distances for AA/TT/TA dinucleotides.

Table 1
Average* nucleosome binding affinities for TE classes (families)

TE class (family)[a]	Avg nba ± s.e.[b]
L1	0.849 ± 6.8e–4
LINE other	0.805 ± 7.6e–4
Alu	0.510 ± 5.2e–4
SINE other	0.789 ± 7.0e–4
LTR	0.807 ± 7.9e–4
DNA	0.802 ± 9.8e–4

[a] TEs are broken down by class (family) using RepeatMasker. The L1 and Alu families are considered separately from all other LINEs and SINEs respectively. All LTR and DNA elements are considered together as classes.

[b] Average nucleotide binding affinities ± standard errors.

* All differences are statistically significant (ANOVA, $F = 2.8e4$, $P \approx 0$).

prediction method we employed in this analysis relies on the periodicity of AT-rich dinucleotides (Segal et al., 2006). Thus, it is possible that the high (low) nucleosome binding affinity of TE (SSR) sequences in proximal promoter regions is a corollary effect of local differences in nucleotide composition. We attempted to control for this possibility in several ways. First of all, average nucleosome binding affinities were computed for all TE, SSR and non-repetitive sequences irrespective of their locations along proximal promoter regions. On average, TE sequences bind nucleosomes most tightly, followed by non-repetitive DNA and SSRs, which have the lowest nucleosome affinities (Fig. 2a); all differences are highly statistically significant (ANOVA, $F = 4.5e11$, $P \approx 0$).

In addition to the binding affinity observations that are based on the nucleosome prediction software, we also analyzed the nucleosome wrapping characteristic AA/TT/TA dinucleotide frequencies along experimentally characterized nucleosome bound sequences from chicken (Satchwell et al., 1986) that we identified as being derived from TEs ($n = 39$). The chicken TE sequences show the characteristic AA/TT/TA dinucleotide periodicity expected of nucleosome bound sequences (Fig. 2b); in fact, the average distance between dinucleotide peaks for these TE sequences is ~10.3 bp, which is close to the expected distance of 10.2 bp corresponding to one turn of the DNA helix (Fig. 2c). This is significant because DNA sequences are thought to wrap around nucleosomes by bending sharply at each repeating turn of the DNA helix, and this sharp bending is facilitated by the specific AA/TT/TA dinucleotides (Widom, 2001).

We also attempted to control for nucleotide composition effects by randomizing promoter sequences and re-calculating nucleosome binding affinities. First, entire 1 kb promoter sequences were randomized and nucleosome binding affinities were re-calculated. This control procedure has the effect of eliminating both native dinucleotide occurrences and local nucleotide composition biases. The average nucleotide binding affinity for such randomized promoter sequences (nba = 0.16) is ~3× lower than seen for the observed promoter sequences (nba = 0.49), and the difference between random and observed affinities is highly significant ($t = 23$, $P = 5.3e{-}100$). In addition to differences in the magnitude of the nucleosome binding affinities, the relative affinity trends along the promoters were compared for the random versus observed sets. Partial correlation was used to control for the effects of the random sequences on the observed relationship between nucleosome binding affinity with TE and SSR densities along proximal promoters. The positive (negative) correlations between nucleosome binding for TE (SSR) do not change when the correlations between random sequences and nucleosome binding along the promoters are accounted for [$r_{\mathrm{nba \cdot TE|random1}} = 0.99$ and $r_{\mathrm{nba \cdot SSR|random1}} = 0.85$].

A second randomization procedure was done to account for local differences in nucleotide composition along proximal promoter sequences. In this case, sequences were randomized within non-overlapping 100 bp windows along the promoters. This had the effect of eliminating native dinucleotide occurrences while maintaining local nucleotide composition. As with the complete sequence randomization procedure, the locally randomized sequences have significantly lower nucleosome binding affinities (nba = 0.23) than the observed sequences (nba = 0.49), and this 2.1× difference is highly statistically significant ($t = 17$, $P = 5.0e{-}55$). Clearly, local nucleotide composition alone cannot explain the observed nucleosome binding affinities. However, the relative trends in nucleosome binding show different

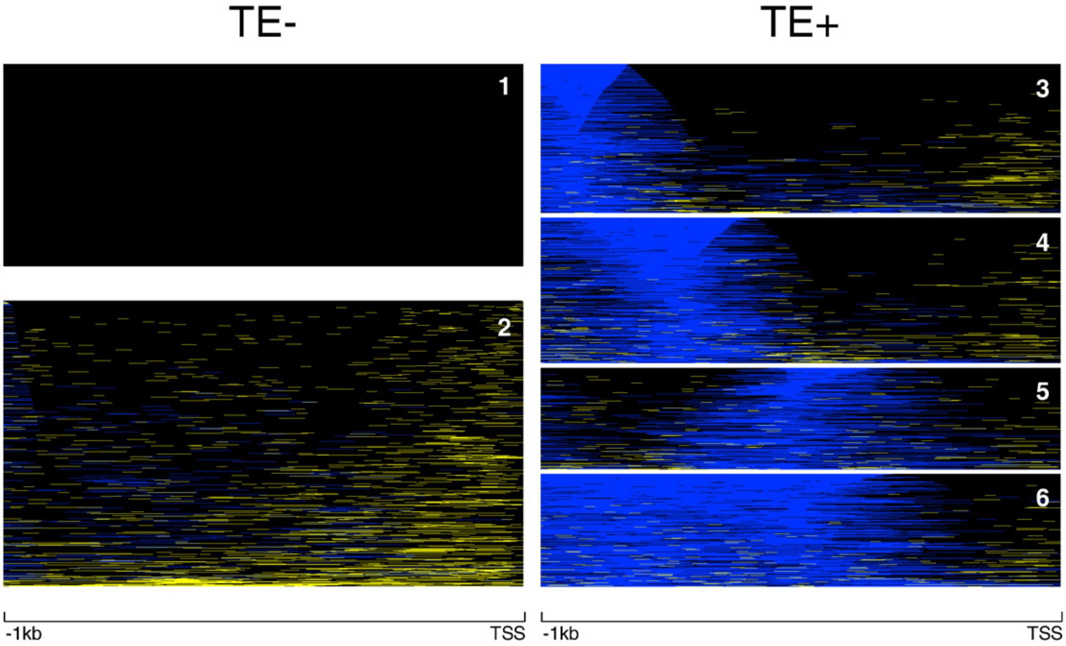


Fig. 3. Clusters of human proximal promoters based on their repetitive DNA sequence distributions. Proximal promoter sequences are represented left-to-right from position −1000 bp upstream to the transcriptional start site (TSS). Promoter sequences are color coded according to their repeat element distributions. Individual promoter nucleotide positions occupied by TEs are shown in blue, SSR positions are shown in yellow and non-repetitive positions are shown in black. The vertical size of the clusters corresponds to the number of sequences in each cluster. There are two (*c*1 and *c*2) clusters that contain promoters largely devoid of TE sequences (TE−), and the promoter sequences of the remaining four clusters (TE+, *c*3–*c*6) contain increasing numbers of TEs.

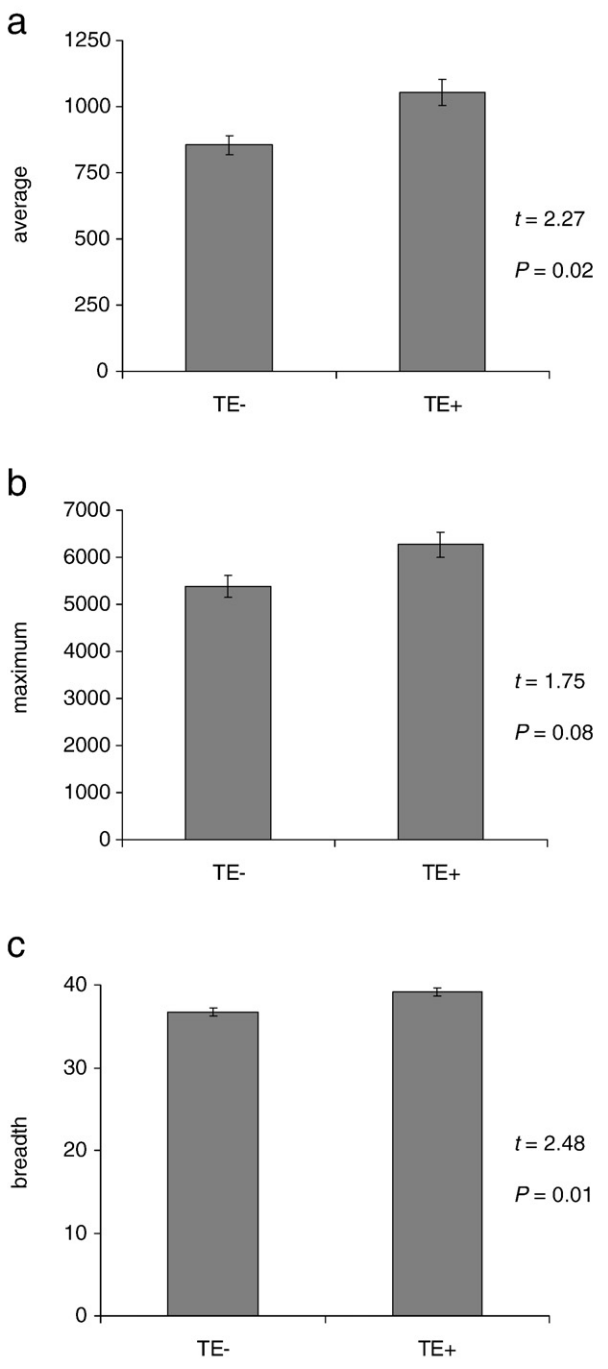


Fig. 4. Gene expression comparison for TE− versus TE+ promoter clusters. Human gene expression data are from the Novartis mammalian gene expression atlas version 2 (GNF2). (a) Average level of expression, (b) maximum level of expression and (c) breadth of expression across 79 human tissues (cells) are compared for genes that have TE− versus TE+ promoter sequences. Statistical significance levels are based on the Student's *t*-test.

local nucleotide composition effects for TEs versus SSRs. The partial correlation controlling for the effects of local nucleotide composition on the relationship between TE density and nucleosome binding eliminates the positive correlation seen across the entire promoter for the observed data [$r_{\text{nba}\cdot\text{TE|random2}} = -0.14$]. This suggests that local nucleotide composition bias influences the decreasing trend in nucleosome binding affinities along proximal promoters irrespective of TE density. Interestingly, this same mitigating effect of local nucleotide composition is not seen for the relationship between SSRs and nucleosome binding [$r_{\text{nba}\cdot\text{SSR|random2}} = -0.53$]. This suggested the possibility that most of the local nucleotide composition bias effect on the relationship between TEs and nucleosome binding may be confined to the region closest to the TSS where TEs are largely absent and SSRs are at their most dense (Fig. 1a). Indeed, when partial correlation controlling for local nucleotide bias is done excluding 150 bp upstream of the TSS, the positive correlation between TEs and nucleosome binding affinity remains [−1000 to −150 $r_{\text{nba}\cdot\text{TE|random2}} = 0.76$]. In other words, positive TE effects on nucleosome binding are most evident away from the TSS, while the SSRs that inhibit nucleosome binding act closest to the TSS.

Taken together, these data suggest the intriguing possibility that the human genome utilizes repetitive DNA content along promoter regions to tune nucleosome binding in such a way as to facilitate maximum access of the basal transcriptional machinery just upstream of TSS. Furthermore, different classes of repeats play distinct roles in this process; TEs bind nucleosomes tightly yielding compact less accessible DNA, while SSRs extrude nucleosomes creating a relatively open chromatin environment.

3.2. Cross-species comparison

In addition to the human genome analysis, the relationship between nucleosome binding and repetitive DNA content of proximal promoter regions was evaluated for four additional mammalian species with complete genome sequences available: chimpanzee (*P. troglodytes*), mouse (*M. musculus*) and rat (*R. norvegicus*). For these species, NCBI Refseq gene models were used to define TSS and proximal promoter regions, while TE and SSR repeats and nucleosome binding were analyzed as was done for the human genome. The trends observed for human are highly similar to those seen for the other mammalian species (Supplementary Fig. 1). In chimpanzee, mouse and rat, nucleosome binding affinities decrease steadily along the proximal promoter region until the core promoter, <50 bp from the TSS, where nucleosome binding begins to increase. For these three species, TE density drops precipitously and steadily along the proximal promoter while SSR density increases sharply at first in the core promoter near the TSS and then drops off again as nucleosome binding affinity increases. Thus, repeat-rich mammalian genomes appear to use repetitive DNA elements to tune nucleosome binding and core promoter accessibility in similar ways. The conservation of the relationship between repetitive DNA and nucleosome biding in core promoters of several mammalian species suggests that this mechanism may have evolved early in the mammalian radiation as repetitive

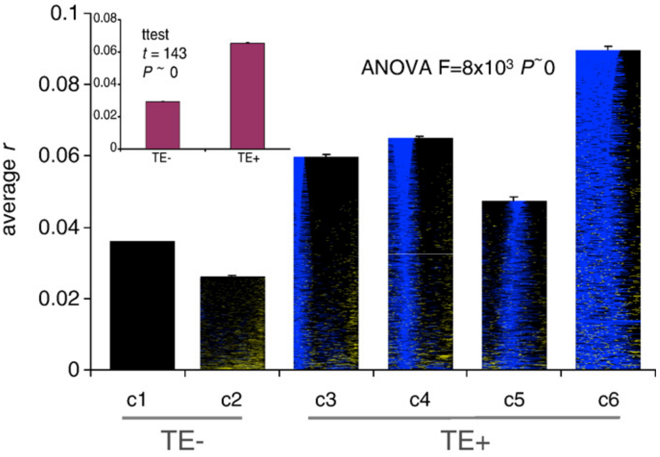


Fig. 5. Gene co-expression for repeat-specific proximal promoter clusters. Average pairwise Pearson correlation coefficients (*r*) for gene expression across 79 human tissues are shown for clusters 1–6 (see Fig. 3) as well as for the TE− versus TE+ clusters (inset). Statistical significance levels are based on ANOVA for multiple comparisons and on the Student's *t*-test for the TE− versus TE+ comparison.

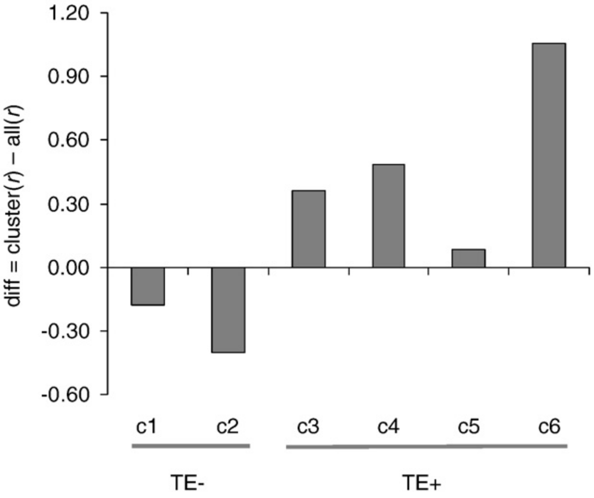


Fig. 6. Differences in gene co-expression between cluster-specific gene pairs versus all possible pairs of genes. Average pairwise Pearson correlations (r) for gene expression across 79 human tissues were measured for all possible gene pairs and this value was subtracted from the average pairwise r-values for genes within each repeat-specific cluster ($c1$–$c6$). A negative value indicates that genes within the cluster have less similar co-expression than background, whereas a positive value indicates that genes within a cluster are more highly co-expressed than expected.

elements were proliferating within genomes. However, many of the repetitive elements that yield these patterns evolve rapidly and are lineage-specific. Accordingly, there may be an ongoing dynamic between repeat generation by mutation and/or transposition followed by selection based on the promoter location of the repeat and specific requirements for chromatin accessibility. For TEs in particular, this could simply mean that the elements are eliminated from core promoter regions close to the TSS by purifying selection. Indeed, negative selection against TE insertions closest to TSS would seem to be the easiest way to explain the observed pattern of TE density (Fig. 1a and Supplementary Fig. 1). However, our analysis of gene expression data, described in following sections, suggests that this is not the case. SSRs, on the other hand, appear to be favored in core promoter regions.

3.3. TE-specific effects on nucleosome binding affinity

The Repbase library of repetitive DNA elements used by the program RepeatMasker can be used to annotate TEs into different classes and families (Jurka et al., 2005; Kapitonov and Jurka, 2008). Using this approach, human TE sequences were divided into LINEs, (L1 and other LINES), SINEs (Alu and other SINEs), LTR retrotransposons, and DNA transposons to determine if different classes (families) of elements show differential nucleosome binding affinities (Table 1). In general, LINEs, LTR retrotransposons and DNA transposons have higher affinities for nucleosomes compared to SINEs. Specifically, L1 elements exhibit the highest nucleosome binding affinities while Alu elements display the lowest affinity for nucleosomes. All differences are statistically significant (Table 1, ANOVA).

The differences in nucleosome binding affinities between L1 and Alu are consistent with their respective nucleotide compositions and perhaps also relevant to their genomic distributions. L1 elements, and LINEs in general, are more AT-rich than Alus (SINEs), and AT-rich sequences are more likely to bind nucleosomes tightly as discussed previously. L1 elements are also biased towards intergenic regions in their distribution, while Alu elements are found primarily in gene rich regions. In fact, it has been shown that Alus are preferentially retained in GC- and gene-rich regions of the genome, and this has been taken to suggest that they may be selectively fixed therein by virtue of some gene-related function that they play (Lander et al., 2001). Our data showing lower nucleosome binding for Alu elements suggests that they may be retained in gene regions by virtue of their ability to maintain a relatively open chromatin environment. Conversely, L1 elements may help to maintain compact chromatin structure characteristic of intergenic regions.

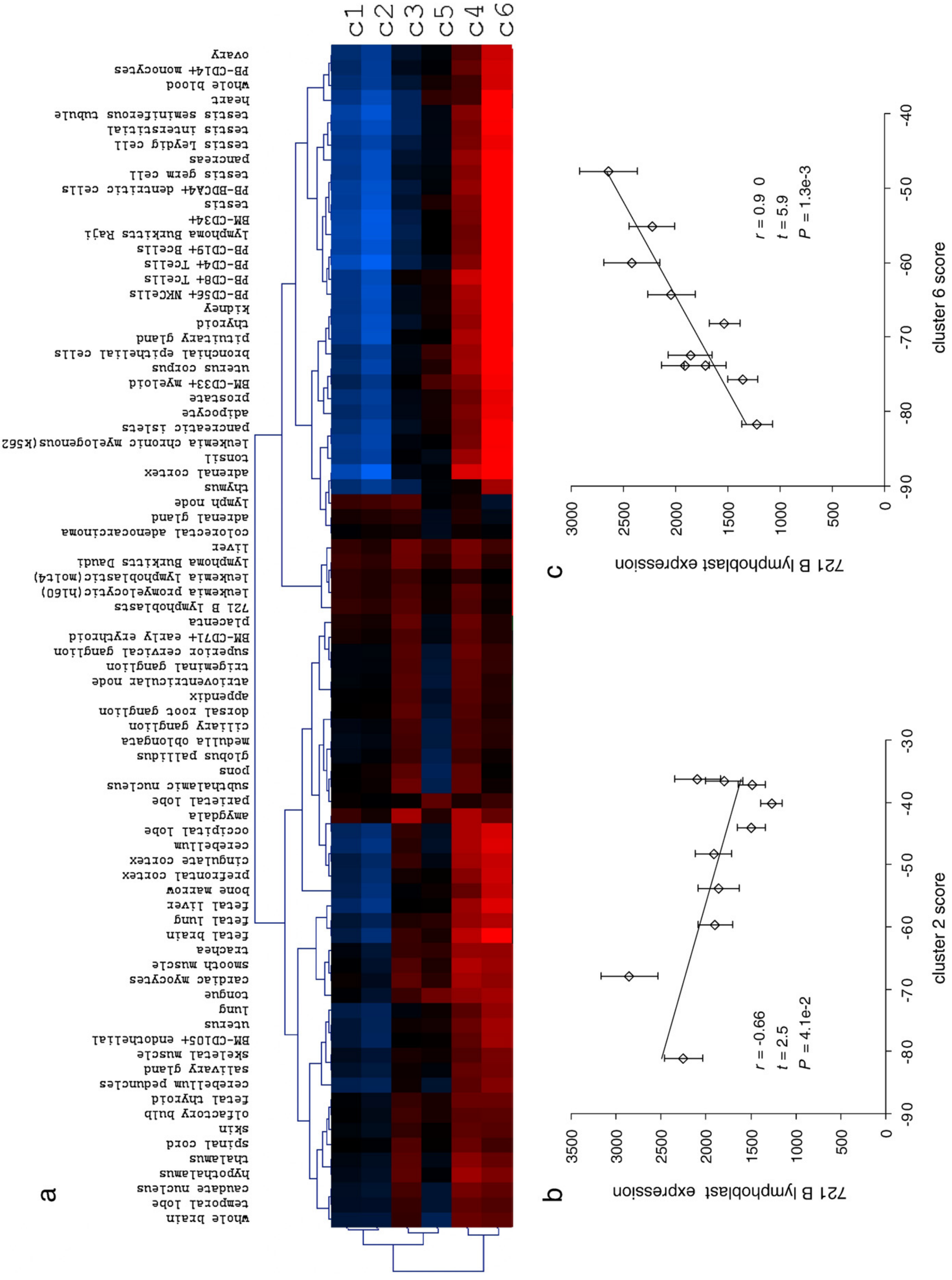
3.4. Promoter repeat architecture and gene expression levels

In light of the observed relationship between repetitive DNA elements and nucleosome binding, we used the repetitive DNA content of proximal promoter regions to group human genes into related clusters. The gene expression and functional properties of the clusters were then compared to their characteristic repeat architectures. To cluster human genes using their promoter repeat distributions, proximal promoter sequences were represented as 1000-unit vectors with each position in a sequence-specific vector receiving a score indicating whether that particular nucleotide is part of a TE, SSR or non-repetitive sequence. These gene-specific promoter repeat vectors were then compared using a distance metric and clustered as described (Materials and methods). This approach ensured that the clusters reflect both the abundance, or lack thereof, and the location of distinct repetitive DNA elements in human promoter sequences. In other words, this scheme relates human genes solely by virtue of their promoter repeat distributions.

We obtained six repeat-specific clusters of human genes in this way (Fig. 3), each cluster representing a distinct overall pattern of TE and/or SSR content and distribution. Two of these clusters ($c1$ and $c2$, TE−) consist of genes that are largely devoid of TEs, while four consist of genes with increasing TE densities ($c3$–$c6$, TE+). $c1$ does not contain any repetitive DNA, while $c2$ is enriched in SSR sequences and has very low TE content. $c3$–$c6$ have progressively more TE content with locations shifting slightly towards the TSS.

The gene expression properties of the human genes in these clusters were analyzed using version 2 of the Novartis mammalian gene expression atlas (GNF2) (Su et al., 2004). This data set consists of Affymetrix microarray experiments, performed in replicate, on 79 different human tissue (cell) samples. For each human gene, over 79 tissues, we computed the average expression level, maximum expression level and breadth of expression as described (Materials and methods); cluster-specific averages for each of these parameters were then compared (Fig. 4). We were surprised to find that clusters that contain TEs (c3–c6, TE+) have higher average, maximum and breadth of expression than clusters that are largely devoid of TEs ($c1$ and $c2$, TE−). Gene expression levels are known to correlate with a

Fig. 7. Promoter repetitive DNA architecture and tissue-specific gene expression. Probabilistic models were used to represent the repetitive DNA architectures of each repeat-specific cluster (see Fig. 3 and Supplementary Fig. 2). Cluster-specific probabilistic models were used to score individual promoter sequences in terms of how closely they resemble a given cluster (Materials and methods). Vectors of cluster-specific gene scores were correlated with vectors of gene expression values specific human tissues. (a) A heat map illustrating the relative correlation values between gene (promoter)-specific scores for each cluster and tissue-specific gene expression values for the 79 tissues in the Novartis gene expression atlas version 2 (GNF2). Relatively high (positive) correlations between gene-cluster scores and gene expression levels are shown in red and low (negative) correlations are shown in blue. Two specific examples of such correlations are shown in panels b and c. (b) Gene (promoter)-specific scores based on the probabilistic model for cluster 2 are negatively correlated with gene expression levels in a B lymphoblast cell line. (c) Gene (promoter)-specific scores based on the probabilistic model for cluster 6 are positively correlated with gene expression levels in a B lymphoblast cell line. In other words, genes with repetitive DNA promoter profiles that most closely resemble cluster 6 are more highly expressed in the B lymphoblast cell line, whereas genes with repetitive DNA promoter profiles that resemble cluster 2 have lower levels of B lymphoblast expression.


a
C1
C2
C3
C5
C4
C6
ovary
PB-CD14+ monocytes
whole blood
heart
testis seminiferous tubule
testis interstitial
testis Leydig cell
pancreas
testis germ cell
PB-BDCA4+ dentritic cells
testis
BM-CD34+
lymphoma burkitts Raji
PB-CD19+ Bcells
PB-CD4+ Tcells
PB-CD8+ Tcells
PB-CD56+ NKCells
kidney
thyroid
pituitary gland
bronchial epithelial cells
uterus corpus
BM-CD33+ myeloid
prostate
adipocyte
pancreatic islets
leukemia chronic myelogenous(k562)
tonsil
adrenal cortex
thymus
lymph node
adrenal gland
colorectal adenocarcinoma
liver
lymphoma burkitts Daudi
leukemia lymphoblastic(molt4)
leukemia promyelocytic(hl60)
721 B lymphoblasts
placenta
BM-CD71+ early erythroid
superior cervical ganglion
trigeminal ganglion
atrioventricular node
appendix
dorsal root ganglion
ciliary ganglion
medulla oblongata
globus pallidus
pons
subthalamic nucleus
parietal lobe
amygdala
occipital lobe
cerebellum
cingulate cortex
prefrontal cortex
bone marrow
fetal liver
fetal lung
fetal brain
trachea
smooth muscle
cardiac myocytes
tongue
lung
uterus
BM-CD105+ endothelial
skeletal muscle
salivary gland
cerebellum peduncles
fetal thyroid
olfactory bulb
skin
spinal cord
thalamus
hypothalamus
caudate nucleus
temporal lobe
whole brain
b
721 B lymphoblast expression
0
500
1000
1500
2000
2500
3000
3500
cluster 2 score
-90
-80
-70
-60
-50
-40
-30
r = -0.66
t = 2.5
P = 4.1e-2
c
721 B lymphoblast expression
0
500
1000
1500
2000
2500
3000
cluster 6 score
-90
-80
-70
-60
-50
-40
r = 0.9 0
t = 5.9
P = 1.3e-3

number of measures of gene 'importance' such as sequence and phylogenetic conservation, fitness effects, numbers of protein interactions, etc. (Duret and Mouchiroud, 2000; Pal et al., 2001; Krylov et al., 2003; Zhang and Li, 2004; Wolf et al., 2006). In other words, genes that are more highly and broadly expressed are under greater purifying selection than genes with lower expression levels. If TEs are eliminated from proximal promoter sequences by purifying selection, then one may expect that TE+ promoters would have lower, and not higher as we observe, levels of gene expression than TE− promoters. In other words, our analysis of repeat cluster gene expression levels argues against the straightforward interpretation that the paucity of TEs in proximal promoter sequences, and their decreasing frequency closer to TSS, is a result of purifying selection against disruptive insertions in core promoters.

On the other hand, one may expect that genes with more restricted and more tightly regulated expression, such as developmental genes, would have more TE sensitive promoters than genes that are highly and broadly expressed. In fact, developmental genes are known to have promoters that are largely devoid of TEs (Simons et al., 2006, 2007). This may reflect the fact that such genes are more finely and tightly regulated and accordingly contain more complex promoters with higher numbers of cis-regulatory elements. If this is indeed the case, then the paucity of TEs in proximal promoter regions may still be explained, to some extent, by purifying selection against disruptive insertions. Discrimination between these two hypotheses regarding the selective elimination, or lack thereof, of proximal promoter TE sequences awaits further analysis.

3.5. Promoter repeat architecture and tissue-specific gene co-expression

In addition to analyzing repeat cluster gene expression levels, we also evaluated the relationship between the tissue-specific expression patterns of genes across the 79 tissues from GNF2 and their promoter repeat content. To do this, gene-specific vectors of expression levels across tissues were compared using the Pearson correlation coefficient (r); positive values of r indicate gene pairs that are co-expressed across tissues. For each cluster, average r-values were computed based on all pairwise comparisons within the cluster (Fig. 5). Higher average r-values are associated with increasing TE promoter content of the clusters. For instance, there is a positive ($R=0.77$), albeit marginally significant ($z=1.72$, $P=0.1$), rank correlation between cluster TE content and co-expression. In addition, all four TE+ clusters have greater average co-expression than either of the TE− clusters, and the average r-value for TE+ clusters together is significantly greater than seen for the combined TE− clusters (Fig. 5).

The possibility of gene co-regulation within repeat clusters was also evaluated by taking the difference between the average r-value for all pairwise comparisons within clusters to average pairwise r-value for all gene comparisons (Materials and methods) (Fig. 6). If genes within clusters are co-regulated, then the value of this difference should be positive, whereas no co-regulation will yield a negative difference value. The TE− clusters 1 and 2 have negative difference values indicating that genes with no TEs in their promoters are less co-expressed with other genes possessing a similar lack of repeats than they are with all genes. On the other hand, the TE+ clusters 3–6 all have positive difference values further demonstrating that genes with similar repetitive DNA profiles in their promoters are more closely co-expressed than random pairs of genes. The difference values for each cluster are statistically significant ($7.3>z>100.6$, 1.4e−13<P<0).

Taken together, these observations on gene co-expression also argue against the notion that TE insertions in proximal promoter sequences are basically disruptive or deleterious, since the presence of similar TE promoter distributions implies a higher level of gene co-regulation than the absence of TEs does. This is not to say that the majority of *de novo* TE insertions in and around functional promoter sequences are not deleterious, clearly they are. However, the repeat sequences that have been fixed in proximal promoter sequences do appear to make functionally relevant contributions to chromatin accessibility and help to regulate levels and specific patterns of gene expression.

3.6. Probabilistic analysis of promoters and gene expression

Given the relationship between gene expression and the repetitive DNA architecture of human promoters we observed, we wanted to further evaluate the propensity of human genes to be expressed in specific tissues based on the repetitive DNA content of their promoters. To do this, we used a probabilistic representation of cluster-specific promoter architectures together with the GNF2 expression data. This involved partitioning 1 kb proximal promoter sequences into 20 non-overlapping windows of 50 bp each, and for a given cluster, representing the probability of observing TE, SSR or non-repetitive nucleotides in each window (Materials and methods). The probabilistic representation of promoter repeat architectures we employed is mathematically analogous to the probabilistic representations of position weight matrices (PWMs) used to summarize position-specific residue frequencies among collections of sequence motifs such as transcription factor biding sites (Wasserman and Sandelin, 2004). Accordingly, promoter repeat profiles can be represented as sequence logos showing the probability and distribution for sites of different repeat classes (Supplementary Fig. 2). The cluster-specific promoter repeat profiles can then be used to score individual promoter sequences just as PWM representations can be used to score putative motif sequences. Connecting these cluster- and position-specific promoter repeat profiles to tissue-specific gene expression profiles was done in a way that is similar to the methodology used to connect the presence of transcription factor binding site motifs to specific gene expression patterns (Conlon et al., 2003).

For each of the 79 tissues in GNF2, each promoter sequence was given six cluster-specific scores, and for each cluster, the gene-specific scores were correlated with the tissue-specific gene expression levels (Materials and methods). This resulted in a 6-by-79 matrix of cluster-by-tissue correlations (Fig. 7). The TE+ clusters 4 and 6 show particularly high correlations with a number of tissues, such as B lymphoblasts (Figs. 7b and c), whereas the TE− clusters 1 and 2 show low correlations with the same tissues and lower correlations overall. This indicates that certain repeat-rich promoter architectures play a role in driving tissue-specific expression, while repeat poor promoters have less coherent regulatory properties. In addition, the differences in promoter score-expression level correlations across tissues and between clusters indicate that different repeat contexts are likely to have tissue-specific regulatory functions. Hierarchical clustering of the tissues and the clusters, according to the promoter score-expression level correlations, group related tissues together including reproductive tissues, immune related cells and cancer samples (Fig. 7a). This indicates that TE-rich promoters may help to regulate genes that function specifically in these tissues further underscoring the biological significance of promoter sequence repetitive DNA profiles.

3.7. Gene Ontology analysis

Having established a connection between repetitive DNA promoter architectures and gene regulation, we wondered whether genes with similar promoter repeat distributions encoded proteins with related functions. In order to test this, we used analysis of Gene Ontology (GO) terms for genes within and between the TE− versus the TE+ repeat-specific promoter clusters (Fig. 3). A modified version of the GO semantic similarity measure (Lord et al., 2003; Azuaje et al., 2005) was used to compare the similarities between GO terms within clusters versus the background GO similarity among all pairs of genes. As described previously (Marino-Ramirez et al., 2006; Tsaparas et al.,

Table 2
Over-represented* GO slim[a] terms for repeat-specific promoter clusters

Group[b]	Molecular function[c]	Cellular component[c]	Biological process[c]
TE−	GO:0030528: transcription regulator activity	–	GO:0007154: cell communication GO:0007275: multicellular organismal development GO:0050789: regulation of biological process
TE+	GO:0003824: catalytic activity GO:0016491: oxidoreductase activity	GO:0005737: cytoplasm	GO:0006810: transport GO:0007154: cell communication
C1	GO:0005198: structural molecule activity	–	–
C2	GO:0016301: kinase activity GO:0016491: oxidoreductase activity GO:0030528: transcription regulator activity	–	GO:0007154: cell communication GO:0007275: multicellular organismal development GO:0007610: behavior GO:0030154: cell differentiation GO:0050789: regulation of biological process
C3	–	–	–
C4	GO:0003824: catalytic activity	GO:0005737: cytoplasm	GO:0006944: membrane fusion GO:0009056: catabolic process
C5	GO:0004872: receptor activity GO:0005215: transporter activity GO:0022857: transmembrane transporter activity	GO:0009986: cell surface	GO:0050896: response to stimulus
C6	GO:0003824: catalytic activity	GO:0005622: intracellular GO:0005737: cytoplasm	GO:0008152: metabolic process GO:0009058: biosynthetic process

[a] GO slim categories provide a high level view of GO functions and subsume a number of lower (more granular) GO functional annotation categories.
[b] Repeat-specific clusters 1–6 along with the combined TE+ and TE− groups (see Fig. 3).
[c] GO functional annotation categories.
[*] Statistical significance for over-represented terms was evaluated using with χ^2 tests with at least $\chi^2>4.2$, $P<0.04$.

2006), the GO semantic similarity approach measures the pairwise similarity between annotation terms along the GO directed acyclic graph in order to evaluate the functional similarity between pairs of genes. For TE− and TE+ genes, the GO similarity difference (*GOdiff*) is equal to the average GO similarity for all gene pairs within clusters minus the average GO similarity for all possible gene pairs (Materials and methods). Negative values of *GOdiff* indicate that gene pairs are more similar within clusters than for all possible pairs. Both the TE− and TE+ gene sets encode proteins that are significantly more functionally similar than the background comparison set [TE− = −3.4e–3, $z=34$, $P\approx 0$; TE+ = −7.9e–3, $z=11$, $P=4.8$e–3]. However, within the TE+ clusters, pairs of genes encode proteins that are significantly more functionally similar, on average, than the pairs of genes found within the TE− clusters ($t=5.8$, $P=6.4$e–9). This is consistent with the stronger signal of gene co-regulation seen for clusters of promoter sequences that are enriched for TEs and underscores the potential biological significance of repeat-rich promoter sequences in the human genome.

Given the functional coherence of repeat-specific clusters demonstrated by the GO similarity analysis, we wanted to evaluate whether certain GO functional categories are over-represented within specific clusters. To do this, we traced the GO terms represented in the dataset to GO slim terms (Table 2). GO slim categories provide a higher level view of more granular individual GO annotations in order to provide an overview of the kinds of functions that may be over-represented in different groups. The observed counts of GO slim categories for each of the six repeat-specific clusters, as well as for the combined TE− and TE +, groups were compared to their expected values based on the background GO slim frequencies across all clusters to look for over-represented terms. Genes in the electron transport, cytoplasm, catalytic activity and oxidoreductase activity categories were found to be over-represented in TE+ clusters and accordingly under-represented in the TE− clusters, whereas genes in cell communication, multicellular organismal development, regulation of biological process and transcription regulator activity categories are over-represented in TE− clusters and under represented in TE+ clusters. Evaluation of over-represented GO terms in individual clusters reveals coherence across the three categories of GO terms: molecular function, cellular component and biological process. For instance, the TE+ cluster 5 has an over-represented receptor and transporter activities in the molecular function category that agree with the cell surface cellular component term and the response to stimulus biological process term. The over-represented catalytic activity molecular process term for the most TE-rich cluster 6 corresponds to a cytoplasmic cellular component term along with metabolic and biosynthetic biological process terms. In a general sense, the coherence of GO functional annotations within repeat-specific clusters and the differences between clusters are consistent with biological significance of the regulatory differences seen for these clusters.

4. Conclusion

We have uncovered a connection between repetitive DNA sequences and nucleosome binding in human proximal promoter regions along with an influence of repetitive DNA promoter sequences on specific patterns of gene expression. Interestingly, different classes of repetitive elements function differently to mediate nucleosome binding; TEs bind nucleosomes tightly and are generally excluded from core promoter regions, while SSRs have a low affinity for nucleosomes and are enriched just upstream of TSSs. Thus, it appears that repetitive sequence elements are differentially utilized to tune the accessibility to promoter sequences by transcription factors, particularly the basal transcriptional machinery that assembles just upstream of the TSS, via changes in the local chromatin environment.

Acknowledgement

This research was supported in part by the Intramural Research Program of the NIH, NLM, NCBI. LMR is supported by Corporacion Colombiana de Investigacion Agropecuaria - CORPOICA. IKJ was supported by an Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839). The authors would like to thank Lee S. Katz and Jittima Piriyapongsa for helpful discussions and technical advice.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2009.01.013.

References

Ashburner, M., et al., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29.

Azuaje, F., Wang, H., Bodenreider, O., 2005. Ontology-driven similarity approaches to supporting gene functional assessment. Proc ISMB SIG meeting on Bio-ontologies 2005, 9–10.

Bejerano, G., et al., 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441, 87–90.

Borchert, G.M., Lanier, W., Davidson, B.L., 2006. RNA polymerase III transcribes human microRNAs. Nat. Struct. Mol. Biol. 13, 1097–1101.

Britten, R.J., Kohne, D.E., 1968. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. Science 161, 529–540.

Conley, A.B., Piriyapongsa, J., Jordan, I.K., 2008. Retroviral promoters in the human genome. Bioinformatics 24, 1563–1567.

Conlon, E.M., Liu, X.S., Lieb, J.D., Liu, J.S., 2003. Integrating regulatory motif discovery and genome-wide expression analysis. Proc. Natl. Acad. Sci. U. S. A. 100, 3339–3344.

Dimitri, P., Junakovic, N., 1999. Revising the selfish DNA hypothesis: new evidence on accumulation of transposable elements in heterochromatin. Trends Genet. 15, 123–124.

Doolittle, W.F., Sapienza, C., 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature 284, 601–603.

Dunn, C.A., Medstrand, P., Mager, D.L., 2003. An endogenous retroviral long terminal repeat is the dominant promoter for human beta1,3-galactosyltransferase 5 in the colon. Proc. Natl. Acad. Sci. U. S. A. 100, 12841–12846.

Duret, L., Mouchiroud, D., 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. 17, 68–74.

Feschotte, C., 2008. Transposable elements and the evolution of regulatory networks. Nat. Rev. Genet. 9, 397–405.

Grewal, S.I., Jia, S., 2007. Heterochromatin revisited. Nat. Rev. Genet. 8, 35–46.

Henikoff, S., 2000. Heterochromatin function in complex genomes. Biochim. Biophys. Acta 1470, O1–8.

Henikoff, S., Matzke, M.A., 1997. Exploring and explaining epigenetic effects. Trends Genet. 13, 293–295.

Jordan, I.K., Rogozin, I.B., Glazko, G.V., Koonin, E.V., 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. 19, 68–72.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110, 462–467.

Kapitonov, V.V., Jurka, J., 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat. Rev. Genet. 9, 411–412 author reply 414.

Karolchik, D., et al., 2003. The UCSC Genome Browser Database. Nucleic Acids Res. 31, 51–54.

Karolchik, D., et al., 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32, D493–D496.

Kidd, J.M., et al., 2008. Mapping and sequencing of structural variation from eight human genomes. Nature 453, 56–64.

Kidwell, M.G., Lisch, D.R., 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution Int. J. Org. Evolution 55, 1–24.

Kornberg, R.D., Lorch, Y., 1999. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. Cell 98, 285–294.

Krylov, D.M., Wolf, Y.I., Rogozin, I.B., Koonin, E.V., 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. 13, 2229–2235.

Lander, E.S., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Lippman, Z., et al., 2004. Role of transposable elements in heterochromatin and epigenetic control. Nature 430, 471–476.

Lord, P.W., Stevens, R.D., Brass, A., Goble, C.A., 2003. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19, 1275–1283.

Marino-Ramirez, L., Spouge, J.L., Kanga, G.C., Landsman, D., 2004. Statistical analysis of over-represented words in human promoter sequences. Nucleic Acids Res. 32, 949–958.

Marino-Ramirez, L., Bodenreider, O., Kantz, N., Jordan, I.K., 2006. Co-evolutionary rates of functionally related yeast genes. Evol. Bioinform Online 2, 295–300.

Nishihara, H., Smit, A.F., Okada, N., 2006. Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Res. 16, 864–874.

Orgel, L.E., Crick, F.H., 1980. Selfish DNA: the ultimate parasite. Nature 284, 604–607.

Pal, C., Papp, B., Hurst, L.D., 2001. Highly expressed genes in yeast evolve slowly. Genetics 158, 927–931.

Piriyapongsa, J., Jordan, I.K., 2007. A family of human microRNA genes from miniature inverted-repeat transposable elements. PLoS ONE 2, e203.

Piriyapongsa, J., Marino-Ramirez, L., Jordan, I.K., 2007. Origin and evolution of human microRNAs from transposable elements. Genetics 176, 1323–1337.

Pruitt, K.D., Tatusova, T., Maglott, D.R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35, D61–D65.

Satchwell, S.C., Drew, H.R., Travers, A.A., 1986. Sequence periodicities in chicken nucleosome core DNA. J. Mol. Biol. 191, 659–675.

Segal, E., et al., 2006. A genomic code for nucleosome positioning. Nature 442, 772–778.

Simons, C., Pheasant, M., Makunin, I.V., Mattick, J.S., 2006. Transposon-free regions in mammalian genomes. Genome Res. 16, 164–172.

Simons, C., Makunin, I.V., Pheasant, M., Mattick, J.S., 2007. Maintenance of transposon-free regions throughout vertebrate evolution. BMC Genomics 8, 470.

Smalheiser, N.R., Torvik, V.I., 2005. Mammalian microRNAs derived from genomic repeats. Trends Genet. 21, 322–326.

Smit, A.F.A., Hubley, R. and Green, P. (1996–2004), RepeatMasker. p. http://repeatmasker.org.

Sturn, A., Quackenbush, J., Trajanoski, Z., 2002. Genesis: cluster analysis of microarray data. Bioinformatics 18, 207–208.

Su, A.I., et al., 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. U. S. A. 101, 6062–6067.

Suzuki, Y., Yamashita, R., Nakai, K., Sugano, S., 2002. DBTSS: database of human transcriptional start sites and full-length cDNAs. Nucleic Acids Res. 30, 328–331.

Tharakaraman, K., Marino-Ramirez, L., Sheetlin, S., Landsman, D., Spouge, J.L., 2005. Alignments anchored on genomic landmarks can aid in the identification of regulatory elements. Bioinformatics 21 (Suppl. 1) i440-8.

Tsaparas, P., Marino-Ramirez, L., Bodenreider, O., Koonin, E.V., Jordan, I.K., 2006. Global similarity and local divergence in human and mouse gene co-expression networks. BMC Evol. Biol. 6, 70.

van de Lagemaat, L.N., Landry, J.R., Mager, D.L., Medstrand, P., 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet. 19, 530–536.

Venter, J.C., et al., 2001. The sequence of the human genome. Science 291, 1304–1351.

Wang, H., et al., 2005. SVA elements: a hominid-specific retroposon family. J. Mol. Biol. 354, 994–1007.

Wang, T., et al., 2007. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proc. Natl. Acad. Sci. U. S. A. 104, 18613–18618.

Wasserman, W.W., Sandelin, A., 2004. Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet. 5, 276–287.

Widom, J., 2001. Role of DNA sequence in nucleosome stability and dynamics. Q. Rev. Biophys. 34, 269–324.

Wolf, Y.I., Carmel, L., Koonin, E.V., 2006. Unifying measures of gene function and evolution. Proc. Biol. Sci. 273, 1507–1515.

Zar, J.H., 1999. Biostatistical Analysis, Fourth ed. Prentice-Hall, Upper Saddle River.

Zhang, L., Li, W.H., 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol. Biol. Evol. 21, 236–239.