Available online at www.sciencedirect.com
SCIENCE DIRECT
Gene 365 (2006) 104–110
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

# Transposable elements as a significant source of transcription regulating signals

Bartley G. Thornburg [a,1], Valer Gotea [a,b,1], Wojciech Makałowski [a,b,c,*]

[a] Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University, University Park, PA 16802, USA
[b] Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, PA 16802, USA
[c] Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802, USA

Received 6 May 2005; received in revised form 6 September 2005; accepted 27 September 2005
Available online 10 January 2006

**Abstract**

Transposable elements (TEs) are major components of eukaryotic genomes, contributing about 50% to the size of mammalian genomes. TEs serve as recombination hot spots and may acquire specific cellular functions, such as controlling protein translation and gene transcription. The latter is the subject of the analysis presented. We scanned TE sequences located in promoter regions of all annotated genes in the human genome for their content in potential transcription regulating signals. All investigated signals are likely to be over-represented in at least one TE class, which shows that TEs have an important potential to contribute to pre-transcriptional gene regulation, especially by moving transcriptional signals within the genome and thus potentially leading to new gene expression patterns. We also found that some TE classes are more likely than others to carry transcription regulating signals, which can explain why they have different retention rates in regions neighboring genes.


*Keywords:* Transposable elements; Gene regulation; Promoters; Transcription factor binding sites

## 1. Introduction

More than four years after the publication of the first draft of the human genome (Lander et al., 2001), scientists continue to face unsolved mysteries related to its structure. Among these, the abundance of transposable elements (TEs), which contributed to about half of the human genome (Makalowski, 2001; Lander et al., 2001), has no immediate rational explanation. There are many successful organisms with compact genomes, e.g. all prokaryotes, *Takifugu rubripes* among vertebrates, or *Arabidopsis thaliana* among flowering plants, and as a consequence, many scientists regarded these elements as "junk" (Ohno, 1972), unnecessary ballast, genomic burden, selfish DNA or parasites (Doolittle and Sapienza, 1980; Orgel and Crick, 1980; Hickey, 1982; Schmid, 2003). It was through the progress of the human genome project that knowledge about function of different genomic components increased significantly, including knowledge about origin and role of non-coding sequences (Hardison, 2000). More and more biologists started to regard repetitive elements as a genomic treasure (Brosius, 1991; Makalowski, 1995; Britten, 1996b; Brosius, 1999; Makalowski, 2003), as objects worthy of biological studies. Recent years witnessed accelerated progress in understanding genomic dynamics, and it appears that different mobile elements play a significant role in this process (Makalowski, 1995; Britten, 1996a; Lorenc and Makalowski, 2003; Brosius, 2005).

One of the most direct influences of transposable elements on the host genome is their role in modulating the structure and expression of "resident" genes. After discovery that long terminal repeats (integral parts of some retroelements) carry promoter and enhancer motifs it became clear that integration of such elements in proximity of a host gene must have an influence on this gene expression (Sverdlov, 1998). Many TEs have been described in the last decade that can add a variety of functions to their targeted genes. These include polyadenylation sites, promoters, enhancers, and silencers (Makalowski, 1995). It seems

*Abbreviations:* TE, transposable element; SINE, Short INterspersed Element; LINE, Long INterspersed Element; LTR, long terminal repeat; bp, base pair; pol, polymerase

* Corresponding author. Department of Biology, Pennsylvania State University, 514 Mueller Lab, University Park, PA 16802, USA.
*E-mail address:* wojtek@psu.edu (W. Makałowski).
[1] These authors contributed equally to the work.

0378-1119/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2005.09.036

that a sizable fraction of eukaryotic, gene-associated regulatory elements arose in this modular fashion by insertion of TEs, and not only by point mutations of static neighboring sequences. When a TE is inserted upstream from a gene, a few short motifs can be conserved if they were subjected to selective pressure as promoters or enhancers of transcription. Even though the rest of the TE sequence might evolve beyond recognition due to absence of functional constraints, TEs are hence exapted into a novel function (Brosius and Gould, 1992). A recent survey that analyzed 846 functionally characterized *cis*-regulatory elements from 288 genes, showed that 21 of those elements (~2.5%) from 13 genes (~4.5%) reside in TE-derived sequences (Jordan et al., 2003). The same study showed that TE-derived sequences are present in many more (~24%) promoter regions, defined as ~500 bp located 5' of functionally characterized transcription initiation site. Similarly, van de Lagemaat et al. showed that the 5′ UTRs of a large proportion of mammalian mRNAs contain TE fragments, suggesting that they play a role in regulation of gene expression (van de Lagemaat et al., 2003). One should note that the TE influence on gene regulation upon insertion in promoter regions is only due to chance similarity of TE sequence to various *cis*-regulatory elements, or to the presence of regulatory elements that were active in regulating the transcription of the TE itself. To evaluate their content in such elements, we scanned TE sequences located in promoter regions of all annotated human genes for their content in putative transcription regulating signals. We found that not all regulatory signal classes are over-represented in TE-derived sequences as compared to randomly generated sequences of similar length and GC content, and that different TE classes greatly differ in their potential to fortuitously deliver regulatory signals upon insertion in gene promoter regions. Nevertheless, it is clear that all TEs have a potential to alter gene regulation given their mobility, with possible significant long term evolutionary consequences.

## 2. Materials and methods

### 2.1. Finding TEs in promoter sequences

For the purpose of this study we used the July 2003 assembly of the human genome available from the Golden Path at the University of California Santa Cruz (http://genome.ucsc.edu/goldenPath/hg16/), and corresponding gene annotation (we used the refFlat files which contain annotation for RefSeq and predicted genes). For every gene, we extracted 2000 nucleotides upstream from the annotated transcription start coordinate. The 20,193 excised promoter sequences were then scanned for occurrence of TEs using the May 15, 2002 version of RepeatMasker (http://www.repeatmasker.org) with default options, but ignoring simple repeats and low complexity regions ("-nolow" parameter).

### 2.2. Identification of transcription signals

TRANSFAC database of transcription factor binding sites, maintained by Biobase (http://www.biobase.de), was used as a source of verified transcription signals. We relied upon the MATCH program (Kel et al., 2003) from the same software suite for finding such putative signals in human promoter regions. MATCH uses predefined positional weight matrices (PWM), which we chose based on the TRANSFAC classification of transcription factor binding sites (http://www.gene-regulation.com/pub/databases/transfac/cl.html). Representative high-quality matrices were chosen for each class

Table 1
Representative position weight matrices (PWM) from TRANSFAC database used for identifying transcription factor binding sites in human promoter regions

| Class | Factor name | Matrix ID | Quality | Matrix similarity cutoff[a] |
|---|---|---|---|---|
| *Superclass: basic domains* | | | | |
| Leucine zippers | XBP-1 | V$AP1_C | High | 0.98 |
| | CRE-BP1 | V$CREBP1_Q2 | High | 0.96 |
| | C/EBPα | V$CEBP_C | High | 0.93 |
| Helix–loop–helix | E12 | V$E12_Q6 | High | 0.97 |
| | MyoD | V$MYOD_01 | High | 0.94 |
| Helix–loop–helix/ leucine zipper | USF | V$USF_Q6 | High | 0.95 |
| | c-Myc | V$MYCMAX_01 | High | 0.97 |
| RF-X | RF-X2 | V$RF-X1_01 | High | 0.94 |
| Helix–span–helix | AP-2γ | V$AP2_Q6_01 | Low | 0.92 |
| *Superclass: zinc-coordinating domains* | | | | |
| Zinc finger–nuclear receptor | GR | V$GRE_C | High | 0.92 |
| | ER | V$ER_Q6 | High | 0.94 |
| | HNF-4α1 | V$HNF4_01 | High | 0.86 |
| Cys4 zinc fingers | GATA-1 | V$GATA1_02 | High | 0.97 |
| | GATA-3 | V$GATA_C | High | 0.96 |
| Cys2His2 zinc fingers | YY1 | V$YY1_02 | High | 0.92 |
| | Egr-1 | V$EGR1_01 | High | 0.96 |
| *Superclass: helix–turn–helix* | | | | |
| Homeo domain | HNF-1A | V$HNF1_01 | High | 0.90 |
| | Oct-2B | V$OCT_C | High | 0.93 |
| Paired box | Pax-6 | V$PAX6_01 | High | 0.88 |
| | Pax-5 | V$PAX_Q6 | High | 0.86 |
| Fork head/winged helix | HNF3-α | V$HNF3B_01 | High | 0.94 |
| | E2F-1 | V$E2F_Q6 | High | 0.91 |
| Tryptophan clusters | c-ETS-1 p54 | V$ETS1_B | High | 0.94 |
| | IRF-1 | V$IRF1_01 | High | 0.97 |
| *Superclass: beta-scaffold factors* | | | | |
| Rel homology region | p50 | V $NFKAPPAB_01 | High | 0.96 |
| | p65 | V$NFKB_Q6_01 | High | 0.91 |
| STAT | p91 | V$STAT_01 | High | 0.97 |
| MADS box | MEF-2A | V$MEF2_02 | High | 0.93 |
| | SRF | V$SRF_C | High | 0.93 |
| TATA binding proteins | TBP | V$TATA_C | High | 0.95 |
| HMG | Sox-9 | V$SOX9_B1 | High | 0.95 |
| | SSRP1 | V$TCF4_Q5 | High | 0.98 |
| Heteromeric CCAAT factors | CP1B | V$NFY_Q6 | High | 0.96 |
| Grainyhead | CP2 | V$CP2_02 | High | 0.93 |
| Runt | AML-3 | V$AML_Q6 | High | 0.97 |

[a] The matrix similarity cutoff corresponds to a false negative rate of 50% (FN50).

of transcription factors (Table 1). If high-quality matrices were not available, low-quality matrices were chosen only if they were based on more than ten experimentally characterized binding sites, in order to reduce the false positive identification rate (Qiu et al., 2002). The MATCH profile was created using matrix similarity cutoff values corresponding to a false negative rate of 50% (FN50 values). While this setting potentially excludes half of the biologically significant transcription factor binding sites, it drastically reduces the number of false positive matches. Only binding sites completely overlapping with TE sequences were kept for further analysis.

### 2.3. Analysis of binding site over-representation on TE sequences

Because transcription factor binding sites are short, they can be found by chance on any DNA sequence, including TEs. We wanted to learn whether certain binding sites are over-represented on TE sequences as compared to other DNA sequences. For this purpose, we created sets of randomly generated sequences mimicking the number, length, and GC content of individual elements found by RepeatMasker for every TE class (Table 2). Even though controversial, the choice of randomly generated sequences offers a non-biased dataset for the means of comparing the content of different TE classes in regulatory elements. Non-repetitive intergenic sequences have the big disadvantage of unknown origin. It is widely accepted that much more than 50% of the human genome originated in TEs, thus a data set of randomly selected non-repetitive intergenic sequences can be in fact an uncontrollable mix of TE-derived sequences. Moreover, the annotation of regulatory elements is incomplete, making impossibly to avoid selecting already functional regulatory sequences. Putative binding sites were identified using MATCH with the same settings used for the set of real sequences. Normal distributions of binding site occurrences per element were generated using 1000 samples of size 100 for every combination of binding site–TE class. The sample unit was the number of binding sites found by MATCH on real TEs or randomly generated sequences, and therefore the size of datasets was different for every TE class (see Table 2 for number of TEs found for every TE class). The significance of difference was assessed with Student's *t*-test, assuming equal variances. A stringent significance threshold of 0.0001 was set in order to reduce the number of false positive findings.

## 3. Results and discussion

### 3.1. Representation of different TE classes in promoter regions

Among the 20,193 gene promoter regions analyzed, we found that 16,665 (~83%) contain TE-derived sequences. This represents a much higher percentage than the ~24% previously reported (Jordan et al., 2003), and it is probably due to the longer 5′ upstream region analyzed (2000 bp instead of 500 bp). It should be noted, however, that about half of these TE-derived sequences do not carry putative regulatory elements (Fig. 1), and consequently, transcriptional regulation of an additional 3377 genes (~17%) would remain unaffected after the insertion of TE fragments in their regulatory regions (Fig. 2). The remaining 13,288 (~64%) is still a significant number of genes whose transcriptional regulation could be fortuitously influenced by TEs. In reality, the number is probably much smaller because many of these putative signals would likely be non-functional. Nonetheless, the evolutionary implications are considerable. A summary of RepeatMasker findings in the 20,193 promoter sequences is presented in Table 2. Miscellaneous repetitive elements, such as SVA and SVA2 SINE-like elements, were not included because they are unlikely to influence gene regulation at genomic scale due to their very low frequency within promoter regions (only 35 occurrences were detected).

It is interesting to note that the number, as well as the fraction of total sequence, of SINE elements found in promoter regions, is almost three fold larger than that of LINE elements. This is in agreement with previous reports based on smaller datasets (Jordan et al., 2003), but in contrast to the proportion of SINE/LINE elements within the human genome. While LINE elements account for the largest fraction of the human genome among TEs (20.42% vs. only 13.14% for SINE elements), they are only twice less frequent (~0.8 million vs. ~1.5 million copies) as compared to SINE elements (Lander et al., 2001). Our observation is not surprising because a genome wide survey already showed that SINE elements are more frequent in GC-rich regions, while LINE elements are more frequent in AT-rich regions (Korenberg and Rykowski, 1988). SINE density in AT-rich regions tends to be higher near genes (Smit, 1999). The reason for this appears not to be due to insertion site preference, because both SINE and LINE elements seem to insert randomly in the genome (Arcot et al., 1998; Ovchinnikov et al., 2001). One hypothesis that may explain this pattern proposes that SINE elements are subjected to differential retention rates influenced by their ability to regulate protein translation if readily transcribed from open chromatin such as is found

Table 2
Summary of RepeatMasker findings on human promoter sequences

| TE class | Number of TEs detected | Number of promoter regions containing each TE class | % of total sequence | Minimum TE length (bp) | Average TE length (bp) | Maximum TE length (bp) | GC content (%) |
|---|---|---|---|---|---|---|---|
| SINE | 30,271 | 13,759 | 15.72 | 11 | 210.7 | 427 | 51.29 |
| LINE | 11,356 | 7165 | 6.18 | 11 | 220.8 | 2000 | 39.59 |
| LTR | 5534 | 3633 | 3.79 | 11 | 277.8 | 2000 | 45.42 |
| DNA | 4311 | 3137 | 1.79 | 11 | 168.5 | 1880 | 39.62 |

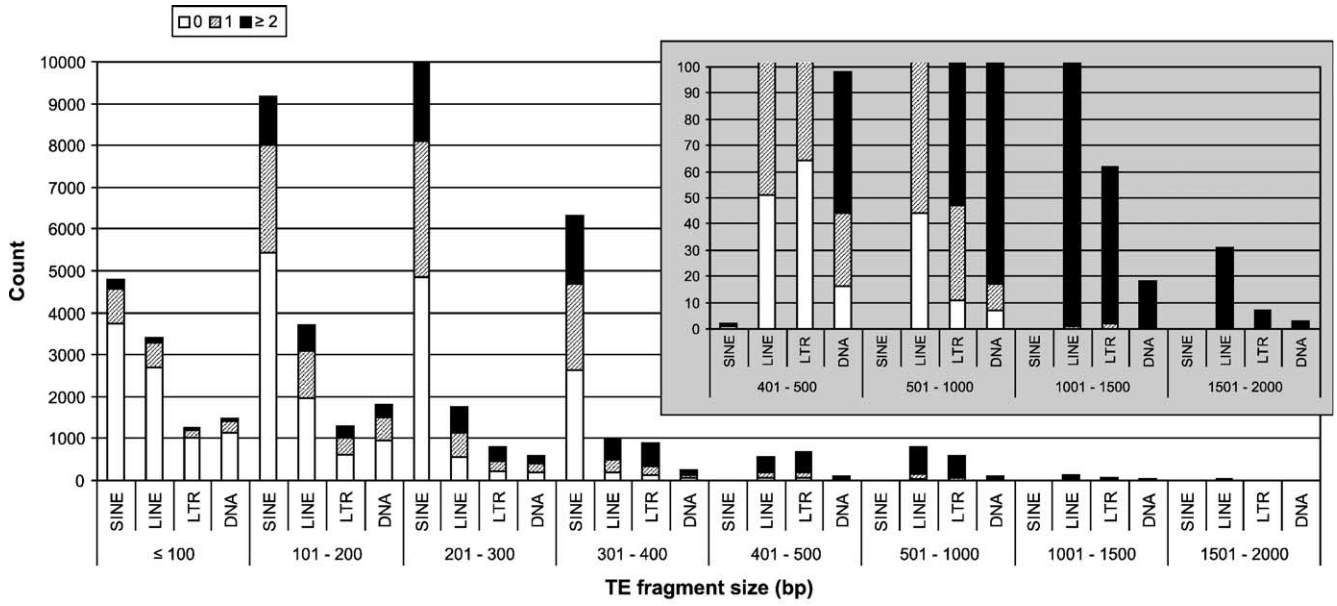


Fig. 1. Size distribution of TE fragments found in gene promoter regions and distribution of transcription factor binding site occurrence on each TE size subclass. Note that the last four bins are 500-bp bins, which, for clarity, are presented in higher-resolution scale as well. Open boxes indicate no occurrence of putative regulatory signals in TE fragments, hashed boxes indicate one occurrence, and solid boxes indicate two or more such occurrences. The maximum number of putative regulatory signals we found in different TE classes was 11 for SINE elements (on a 306-bp AluJO element), 17 for LINE elements (on a 1471-bp L1 element), 18 for LTR elements (on a 2000-bp HERVE element), and 16 for DNA elements (on a 1880-bp Tigger1 element).

near genes (Liu et al., 1995; Chu et al., 1998; Schmid, 1998). Our findings suggest that differential retention of different TE classes might be also determined by their content in transcription factor binding sites (see Section 3.2).

One can also note that the number and proportion of LTR and DNA TEs in promoter regions is the lowest among the four classes (Table 2, Fig. 1). This is probably due to their higher divergence and fragmentation, which makes their detection harder, or even impossible, with current similarity searching techniques. It is known that LTR elements are remnants of more ancient retroviruses, and many of the DNA elements, such as MER75 or Charlie8, are labeled as "fossils". Additionally, the insertion of still active, younger, and more abundant SINE and LINE elements such as Alu and LINE1, respectively, may have gradually replaced older elements from the 2000 nucleotide promoter regions analyzed by us. Consequently, we also observe fewer cases in which LTR and DNA elements occupy the entire or most of the promoter region than we observe for LINE elements (Fig. 1).

### 3.2. TE content in transcription regulating signals

The number of potential transcription factor binding sites found in promoter-residing TEs using MATCH is shown in Table 3. Using the sampling technique described above, we inferred whether TEs contain significantly more binding sites

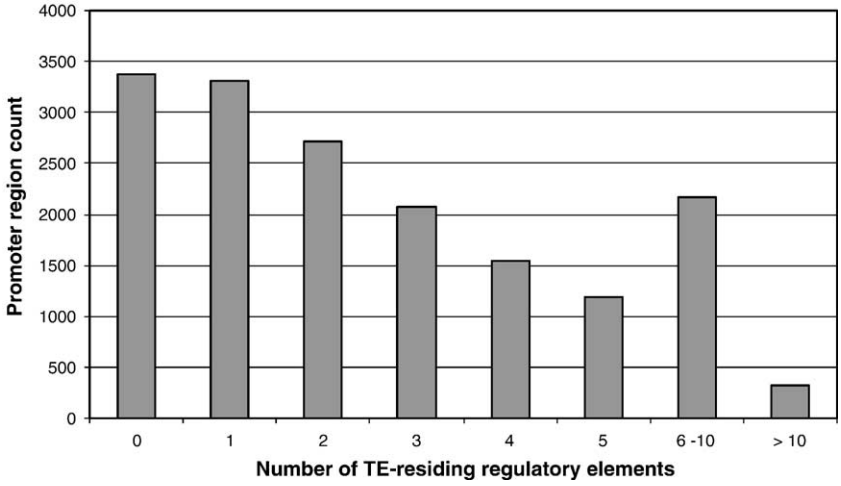


Fig. 2. Distribution of promoter regions based on their content in transcription regulating signals contributed by TE-derived sequences. Promoter regions that do not contain TE-derived sequences are not included in this distribution. The maximum number of TE-contributed putative regulating signals we found in a single promoter region is 32.

Table 3
Comparison of number of putative transcription factor binding sites identified by MATCH on real TE and randomly generated sequences

| Binding site class | TE class | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SINE | | | LINE | | | LTR | | | DNA | | |
| | Obs.[a] | Rnd.[b] | p-value[c] | Obs.[a] | Rnd.[b] | p-value[c] | Obs.[a] | Rnd.[b] | p-value[c] | Obs.[a] | Rnd.[b] | p-value[c] |
| *Superclass: basic domains* | | | | | | | | | | | | |
| Leucine zipper | 1335 | 1363 | 0.493 | 472 | 530 | *4.87·10$^{-06}$* | 565 | 364 | 1.20·10$^{-97}$ | 206 | 153 | 1.62·10$^{-29}$ |
| Helix-loop-helix | 2368 | 1385 | 2.43·10$^{-122}$ | 703 | 258 | 1.46·10$^{-210}$ | 469 | 265 | 2.15·10$^{-119}$ | 234 | 73 | 5.29·10$^{-247}$ |
| Helix-loop-helix / leucine zipper | 873 | 1780 | *2.12·10$^{-120}$* | 158 | 277 | *4.92·10$^{-37}$* | 320 | 268 | 1.04·10$^{-16}$ | 109 | 77 | 3.58·10$^{-19}$ |
| RF-X | 188 | 836 | *1.25·10$^{-219}$* | 183 | 200 | 0.00012 | 159 | 169 | 0.0028 | 75 | 54 | 7.05·10$^{-16}$ |
| Helix-span-helix | 2515 | 3572 | *3.67·10$^{-117}$* | 622 | 278 | 7.19·10$^{-171}$ | 733 | 393 | 5.07·10$^{-221}$ | 245 | 82 | 7.40·10$^{-264}$ |
| *Superclass: zinc domains* | | | | | | | | | | | | |
| Zinc finger / nuclear receptor | 2764 | 2628 | 0.0004 | 1259 | 871 | 7.99·10$^{-102}$ | 961 | 649 | 8.36·10$^{-172}$ | 233 | 260 | *1.06·10$^{-06}$* |
| Zinc finger / GATA | 2402 | 6440 | *0* | 2776 | 3336 | *3.17·10$^{-70}$* | 1892 | 1826 | 8.50·10$^{-07}$ | 641 | 969 | *3.67·10$^{-180}$* |
| Zinc finger / Cis2His2 | 134 | 64 | 2.43·10$^{-15}$ | 33 | 16 | 1.10·10$^{-14}$ | 45 | 17 | 3.91·10$^{-51}$ | 5 | 6 | *4.92·10$^{-06}$* |
| *Superclass: helix turn helix* | | | | | | | | | | | | |
| Homeo domain | 430 | 244 | 3.89·10$^{-36}$ | 385 | 303 | 7.15·10$^{-18}$ | 136 | 109 | 2.46·10$^{-10}$ | 86 | 80 | 0.0003 |
| Homeo / paired box | 575 | 771 | *3.07·10$^{-12}$* | 416 | 339 | 1.47·10$^{-20}$ | 325 | 201 | 4.09·10$^{-102}$ | 123 | 86 | 5.92·10$^{-34}$ |
| Fork head / winged helix | 4155 | 3558 | 1.12·10$^{-17}$ | 1579 | 1470 | 0.00014 | 571 | 852 | *2.73·10$^{-160}$* | 468 | 392 | 8.08·10$^{-27}$ |
| Tryptophan clusters | 338 | 351 | 0.0117 | 293 | 107 | 2.48·10$^{-121}$ | 276 | 81 | 2.79·10$^{-275}$ | 42 | 34 | 1.68·10$^{-09}$ |
| *Superclass: beta-scaffold factors* | | | | | | | | | | | | |
| Rel homology regions | 349 | 771 | *2.11·10$^{-59}$* | 301 | 165 | 1.64·10$^{-44}$ | 406 | 163 | 2.44·10$^{-173}$ | 60 | 54 | 0.842 |
| STAT | 44 | 305 | *8.69·10$^{-124}$* | 141 | 103 | 6.91·10$^{-16}$ | 180 | 78 | 3.54·10$^{-154}$ | 53 | 34 | 1.07·10$^{-23}$ |
| MADS box | 144 | 74 | 1.20·10$^{-13}$ | 94 | 69 | 2.29·10$^{-08}$ | 56 | 38 | 1.80·10$^{-08}$ | 20 | 25 | *5.02·10$^{-06}$* |
| TATA binding proteins | 783 | 440 | 2.11·10$^{-59}$ | 895 | 488 | 2.57·10$^{-175}$ | 391 | 184 | 1.22·10$^{-220}$ | 230 | 133 | 6.36·10$^{-117}$ |
| HMG | 791 | 2238 | *0* | 1487 | 1338 | 5.15·10$^{-19}$ | 742 | 709 | 1.69·10$^{-07}$ | 391 | 369 | *2.52·10$^{-05}$* |
| Heteromeric CCAAT / histone fold | 134 | 935 | *9.67·10$^{-299}$* | 356 | 488 | *3.64·10$^{-31}$* | 869 | 285 | 0 | 83 | 131 | *4.88·10$^{-66}$* |
| Grainyhead | 259 | 577 | *1.16·10$^{-91}$* | 147 | 95 | 7.12·10$^{-22}$ | 317 | 84 | 0 | 57 | 22 | 3.49·10$^{-64}$ |
| Runt | 367 | 278 | 6.34·10$^{-10}$ | 195 | 120 | 3.78·10$^{-38}$ | 129 | 74 | 1.78·10$^{-63}$ | 49 | 32 | 1.18·10$^{-15}$ |

[a] Number of putative transcription factor binding sites identified on promoter residing TEs.
[b] Number of putative transcription factor binding sites identified on randomly generated sequences.
[c] Significance of difference between observations in the two sets of sequences is given by *p*-values (see Methods). They are highlighted or *italicized* if the number of putative transcription factor binding sites in real TEs as compared to randomly generated sequences is over- or under-represented, respectively. Significance level was set to 0.0001.

than random sequences. While statistical significance does not imply biological function, it shows that TEs, when inserted into promoter regions, have an increased potential to alter gene expression in a manner specific to the signals they contain as compared to random sequences.

Unlike the other three TE classes, LTR elements are likely to carry almost all of the binding site classes (Table 3). This might be a consequence of their original function of providing regulatory elements for retrovirus protein coding genes (Sverdlov, 1998). The fork-head/winged helix binding site is the only one being under-represented in LTR elements, the RF-X binding site is significantly over-represented only at 0.01 error level, but the remaining 18 classes are over-represented in LTR elements. LINE elements, and particularly LINE1 elements, are known to contain YY1 pol II (Athanikar et al., 2004) and antisense promoters that have been shown to influence the transcription of adjacent genes (Speek, 2001). We found that LINE elements are likely to carry 14 over-represented classes of binding sites, double the number of transcriptional signals over-represented in SINE elements, which do not contain pol II promoters. Active SINE elements carry, however, pol III A and B boxes (Schmid and Rubin, 1995), which, on the other hand, do not influence the transcription of protein coding genes. The difference in over-represented signals might offer an alternative hypothesis for different retention rates near genes observed for different TE classes. Carrying more transcription regulating signals can cause in more case alteration of gene expression, thus with more possible deleterious effects. Consequently, elements with more regulatory signals are subjected to negative selection, and elements with fewer gene regulatory signals are more likely to be tolerated and fixed as they are less likely to disrupt the regulation of genes upstream of which they are inserted. The comparison is particularly interesting between SINE and LINE elements, the former being the most abundant TE class in promoter regions, as discussed in Section 3.1. The fact that Alu elements are primate specific, and not present in rodents, for example, might indicate that they are indeed, at least for the most part, tolerated rather than positively selected. While it seems reasonable to have fewer promoter–residing LTR elements, other factors, such as the age and genomic abundance, might explain why DNA transposons are the least present in promoter regions.

Another interesting observation is that all 20 binding site classes are likely to be over-represented in at least one TE class,

but only three (helix–loop–helix, TATA binding proteins, runt) are over-represented in all four TE classes. These are all transcription factor binding sites that control the expression of many genes (Beltran et al., 2005; Kitayner et al., 2005; Zukunft et al., 2005), while binding sites over-represented in only one of the TE classes (Table 3), RF-X (Hasegawa et al., 1991), zinc finger/GATA (Pikkarainen et al., 2004; Liew et al., 2005), and heteromeric CCAAT factors/histone folds (Linhoff et al., 1997), appear to have more specific functions. We should reemphasize that our findings do not necessarily imply biological significance, in spite of statistical significance. Two reasons might be invoked here. One is the fact that 2000-nucleotide long 5′ flanking regions are admittedly not perfect substitutes for verified promoter sequences. Our approach is supported by the fact that 5′ flanking regions were shown to be enriched in promoter sequences (Suzuki et al., 2002), and several studies have successfully used the same 5′ flanking region to study influence of promoter sequences (Wang et al., 2001; Zukunft et al., 2005). Secondly, further studies are necessary to show what proportion of TE-residing transcription factor binding sites are in fact functional. MATCH findings should be taken as "potential" until experimental evidence can be provided, in spite of stringent criteria being used for defining over-representation (FN50 values in finding potential binding sites, 0.0001 statistical significance cutoff). Specific examples of such regulatory elements are known to reside in Alu elements (Hamdi et al., 2000; Clarimon et al., 2004; Oei et al., 2004), for example, but we would like to know what is the gene regulation influence of TEs at genomic scale. What our study emphasizes, however, is the *potential* of these TE-residing signals to act as currently unknown regulatory elements or to gain function when carried into new genomic locations by their host TEs.

### 3.3. Conclusions

1. SINE elements are the most abundant TEs in promoter regions, in agreement with conclusions based on smaller datasets. They are three times more numerous and voluminous than LINE elements, in spite of different genomic scale proportion for SINE/LINE occurrence.
2. LTR elements are likely to carry binding sites from all classes. LINE, DNA, and SINE elements carry respectively fewer over-represented binding sites. In addition to previous hypotheses, this might explain why different TE classes have different retention rates in promoter regions.
3. Occupying half of the human genomes, TEs have a big potential to influence gene regulation at genomic scale by carrying potential transcription regulating signals. When inserted in promoter regions, they can alter gene expression patterns by contributing transcription factor binding sites previously not present in promoters of specific genes.

## Acknowledgements

We would like to thank two anonymous reviewers for their critical observations and valuable suggestions for the improvement of the manuscript.

## References

Arcot, S.S., et al., 1998. High-resolution cartography of recently integrated human chromosome 19-specific Alu fossils. J. Mol. Biol. 281, 843–856.

Athanikar, J.N., Badge, R.M., Moran, J.V., 2004. A YY1-binding site is required for accurate human LINE-1 transcription initiation. Nucleic Acids Res. 32, 3846–3855.

Beltran, A.C., Dawson, P.E., Gottesfeld, J.M., 2005. Role of DNA sequence in the binding specificity of synthetic basic-helix–loop–helix domains. ChemBioChem 6, 104–113.

Britten, R.J., 1996a. Cases of ancient mobile element DNA insertions that now affect gene regulation. Mol. Phylogenet. Evol. 5, 13–17.

Britten, R.J., 1996b. DNA sequence insertion and evolutionary variation in gene regulation. Proc. Natl. Acad. Sci. U. S. A. 93, 9374–9377.

Brosius, J., 1991. Retroposons—seeds of evolution. Science 251, 753.

Brosius, J., 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238, 115–134.

Brosius, J., 2005. Echoes from the past—are we still in an RNP world? Cytogenet. Genome Res. 110, 8–24.

Brosius, J., Gould, S.J., 1992. On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". Proc. Natl. Acad. Sci. U. S. A. 89, 10706–10710.

Chu, W.M., Ballard, R., Carpick, B.W., Williams, B.R., Schmid, C.W., 1998. Potential Alu function: regulation of the activity of double-stranded RNA-activated kinase PKR. Mol. Cell. Biol. 18, 58–68.

Clarimon, J., Andres, A.M., Bertranpetit, J., Comas, D., 2004. Comparative analysis of Alu insertion sequences in the APP 5′ flanking region in humans and other primates. J. Mol. Evol. 58, 722–731.

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

Hamdi, H.K., Nishio, H., Tavis, J., Zielinski, R., Dugaiczyk, A., 2000. Alu-mediated phylogenetic novelties in gene regulation and development. J. Mol. Biol. 299, 931–939.

Hardison, R.C., 2000. Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet. 16, 369–372.

Hasegawa, S.L., Sloan, J.H., Reith, W., Mach, B., Boss, J.M., 1991. Regulatory factor-X binding to mutant HLA–DRA promoter sequences. Nucleic Acids Res. 19, 1243–1249.

Hickey, D.A., 1982. Selfish DNA: a sexually-transmitted nuclear parasite. Genetics 101, 519–531.

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.

Kel, A.E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., Wingender, E., 2003. MATCH: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res. 31, 3576–3579.

Kitayner, M., Rozenberg, H., Rabinovich, D., Shakked, Z., 2005. Structures of the DNA-binding site of Runt-domain transcription regulators. Acta Crystallogr., D Biol. Crystallogr. 61, 236–246.

Korenberg, J.R., Rykowski, M.C., 1988. Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. Cell 53, 391–400.

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

Liew, C.K., et al., 2005. Zinc fingers as protein recognition motifs: structural basis for the GATA-1/friend of GATA interaction. Proc. Natl. Acad. Sci. U. S. A. 102, 583–588.

Linhoff, M.W., Wright, K.L., Ting, J.P., 1997. CCAAT-binding factor NF-Y and RFX are required for in vivo assembly of a nucleoprotein complex that spans 250 base pairs: the invariant chain promoter as a model. Mol. Cell. Biol. 17, 4589–4596.

Liu, W.M., Chu, W.M., Choudary, P.V., Schmid, C.W., 1995. Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. Nucleic Acids Res. 23, 1758–1765.

Lorenc, A., Makalowski, W., 2003. Transposable elements and vertebrate protein diversity. Genetica 118, 183–191.

Makalowski, W., 1995. SINEs as a genomic scrap yard: an essay on genomic evolution. In: Maraia, R.J. (Ed.), The Impact of Short Interspersed Elements (SINEs) on the Hpst Genome. RG Landes, Austin, TX, pp. 81–104.
Makalowski, W., 2001. The human genome structure and organization. Acta Biochim. Pol. 48, 587–598.
Makalowski, W., 2003. Genomics. Not junk after all. Science 300, 1246–1247.
Oei, S.L., Babich, V.S., Kazakov, V.I., Usmanova, N.M., Kropotov, A.V., Tomilin, N.V., 2004. Clusters of regulatory signals for RNA polymerase II transcription associated with Alu family repeats and CpG islands in human promoters. Genomics 83, 873–882.
Ohno, S., 1972. So much 'junk' DNA in our genome. Brookhaven Symp. Biol. 23, 366–370.
Orgel, L.E., Crick, F.H., 1980. Selfish DNA: the ultimate parasite. Nature 284, 604–607.
Ovchinnikov, I., Troxel, A.B., Swergold, G.D., 2001. Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. Genome Res. 11, 2050–2058.
Pikkarainen, S., Tokola, H., Kerkela, R., Ruskoaho, H., 2004. GATA transcription factors in the developing and adult heart. Cardiovasc. Res. 63, 196–207.
Qiu, P., Ding, W., Jiang, Y., Greene, J.R., Wang, L., 2002. Computational analysis of composite regulatory elements. Mamm. Genome 13, 327–332.
Schmid, C.W., 1998. Does SINE evolution preclude Alu function? Nucleic Acids Res. 26, 4541–4550.
Schmid, C.W., 2003. Alu: a parasite's parasite? Nat. Genet. 35, 15–16.
Schmid, C.W., Rubin, C.M., 1995. Alu: what's the use? In: Maraia, R.J. (Ed.), The Impact of Short Interspersed Elements (SINEs) on the Hpst Genome. RG Landes, pp. 105–123.
Smit, A.F., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. 9, 657–663.
Speek, M., 2001. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol. Cell. Biol. 21, 1973–1985.
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.
Sverdlov, E.D., 1998. Perpetually mobile footprints of ancient infections in human genome. FEBS Lett. 428, 1–6.
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.
Wang, L., et al., 2001. Analyses of p53 target genes in the human genome by bioinformatic and microarray approaches. J. Biol. Chem. 276, 43604–43610.
Zukunft, J., et al., 2005. A natural CYP2B6 TATA box polymorphism (−82T→C) leading to enhanced transcription and relocation of the transcriptional start site. Mol. Pharmacol. 67, 1772–1782.