Published online 14 April 2010 *Nucleic Acids Research, 2010, Vol. 38, No. 15* ***4907–4916***
doi:10.1093/nar/gkq234

Sequence composition similarities with the 7SL RNA are highly predictive of functional genomic features

Yanick Paquet[1,2] and Alan Anderson[1,2,*]

[1]Centre de recherche en cancérologie de l'Université Laval, L'Hôtel-Dieu de Québec, Centre hospitalier universitaire de Québec, Québec G1R 2J6 and [2]Département de biologie, Université Laval, Québec G1K 7P4, Canada

Received September 13, 2009; Revised March 3, 2010; Accepted March 19, 2010

ABSTRACT

Transposable elements derived from the 7SL RNA gene, such as Alu elements in primates, have had remarkable success in several mammalian lineages. The results presented here show a broad spectrum of functions for genomic segments that display sequence composition similarities with the 7SL RNA gene. Using thoroughly documented loci, we report that DNaseI-hypersensitive sites can be singled out in large genomic sequences by an assessment of sequence composition similarities with the 7SL RNA gene. We apply a root word frequency approach to illustrate a distinctive relationship between the sequence of the 7SL RNA gene and several classes of functional genomic features that are not presumed to be of transposable origin. Transposable elements that show noticeable similarities with the 7SL sequence include Alu sequences, as expected, but also long terminal repeats and the 5′-untranslated regions of long interspersed repetitive elements. In sequences masked for repeated elements, we find, when using the 7SL RNA gene as query sequence, distinctive similarities with promoters, exons and distal gene regulatory regions. The latter being the most notoriously difficult to detect, this approach may be useful for finding genomic segments that have regulatory functions and that may have escaped detection by existing methods.

INTRODUCTION

Identification of non-coding functional DNA segments in large genomic sequences is one of the most challenging tasks of whole genome annotation. With literally tens of genomes of a few billion base pairs each sequenced, deciphering and understanding this expanding universe of genomic sequences has become a major challenge. While countless transcriptional enhancers have been studied in great detail, there still is limited understanding of their fundamental properties and modes of action and the detection of their presence in large genomic sequences remains notoriously difficult (1–3).

It has long been known that DNaseI hypersensitive sites (DHS) are associated with enhancers, silencers, insulators, 5′ promoters and LCRs (4). In fact, mapping DHS has traditionally been taken as the gold standard for the experimental identification of DNA sequence elements of regulatory interest (5). In addition, numerous promoters are found within exons (6), and may also be associated with DHS (7).

With the proliferation of entirely sequenced genomes, comparative genomics is proving to be a powerful means of identifying potentially functional well-conserved genomic sequences. Initial efforts to find regulatory sequences in intergenic regions showed that gene deserts can contain enhancers located more than 500 kb from their target gene (8). Revisiting loci that have been the object of thorough experimental investigation with state-of-the-art comparative genomics is also proving to be most informative, as shown by Hugues *et al.* (9) for the α-globin locus. Recently, genome-wide endeavors have uncovered thousands of conserved non-coding elements (CNEs) per human chromosome (10). While a very high fraction of those CNEs can be expected to harbor functions, it cannot be assumed that all functional non-coding genomic sequences will be brought to light by comparative genomics. It is to be expected that lineage- and species-specific regulatory modules will be more difficult to ascertain in this manner.

It is now widely accepted that transposable elements (TEs) have shaped eukaryotic genomes (11,12). Their contribution is well documented relative to both gene structure and regulation (13,14). Indeed, sequencing of the genome of the marsupial *Monodelphis domestica* suggests that TEs have supplied a substantial fraction of CNEs in mammalian genomes, making them major contributors to

*To whom correspondence should be addressed. Tel: + 418 691 5281; Fax: +418 691 5439; Email: alan.anderson@bio.ulaval.ca

the evolution of mammalian gene regulation (15). The contribution of TEs to functional genomic features comes in many forms (16). In primates, exonized Alu sequences have contributed hundreds of lineage- and species-specific exons (17). Moreover, several thousand Alu elements in the human genome are just a few mutations away from exonization (18). Alu elements have also contributed lineage-specific functional gene regulatory elements (19).

Identifiable TEs make up ~45% of the human genome and this is clearly an underestimate because many ancient TEs will have diverged so as to be unrecognizable by DNA sequence alignment methods (16,20). Eucaryotic TEs fall into two classes, retrotransposons (class I) and DNA transposons (class II) (21,22). In humans, they constitute about 42 and 3% of the genome, respectively (23). Mammalian retrotransposons include retroviral-like long terminal repeat (LTR) elements, and two classes of non-LTR elements, short interspersed repetitive elements (SINEs) and long interspersed repetitive elements (LINEs) (23,24). Of the three LINE families in the human genome (L1, L2 and L3), L1-LINEs are the most abundant and the only ones still active (20). Typically, L1-LINEs span about 6 kb and contain two open reading frames (ORFs) flanked by variable 5′- and 3′-untranslated regions (UTRs) (25). L1-LINEs constitute ~17% of the human genome (26). A substantial portion (23%) of DHS in $CD4^+$ human T cells are contained within TEs (27). In addition, a surprisingly high percentage of human regulatory sequences can be recognized as TE-derived by sequence alignment methods (19). For example, about 24% of 2004 human promoter sequences analyzed contained TE-derived sequences (13).

The sequence composition of entire genomes, as assessed in terms of short word frequencies in their DNA sequence, has been used to suggest phylogenetic relationships (28–30). It has been reported, for example, that di- and tri-nucleotide frequencies are alike in closely related species (31). It has even been suggested that word-frequency variation between related species is reflective of differences in their respective processes of DNA modification, replication and repair (32). The snapshot of the sequence composition of an entire genome, however, is more likely to reflect its content in TEs rather than their respective differences in fundamental processes of DNA maintenance, especially for mammalian genomes. The appreciation that the genomes of higher eukaryotes are mosaics of distinct and diverse sequences, of which nearly half are derived from TEs, must be kept in mind when analyzing their sequence composition.

While retracing the origins of some TEs in any given genome is a task of uncertain outcome, several SINEs in mammalian genomes are clearly derived from tRNAs as well as parts the 7SL RNA, and in many cases from both (24). Alu elements in primates and B1 elements of rodents are good examples of prolific 7SL RNA-derived TEs (11,24). In *M. domestica*, a 7SL pseudogene is present in ~16 000 copies [Gentles,A.J. and Jurka,J. P7SL_MD. Direct submission to Repbase Update (33). Rel. 11.02, March 3, 2006].

The knowledge that TEs have made a substantial contribution to functional genomic features involved in gene regulation is reshaping our understanding of the origin and nature of gene regulatory elements. The hypothesis that many sequences of transposable origin in the human genome are no longer recognizable as such raises the possibility that novel regulatory modules could be brought to light with detection methods that are more sensitive than sequence alignments, provided they are sufficiently discriminatory. In the present study, a root word frequency-based approach has been used to accurately detect distal gene regulatory regions, such as enhancers and LCRs, previously identified as DHS or CNEs, by using the sequence of the 7SL RNA gene as query sequences. This approach highlights similarities in the sequence composition profiles of DNA sequences that cannot be observed by sequence alignment-based methods.

MATERIALS AND METHODS

Root word count correlation

In *ab initio* pattern discovery approaches, the generation of string neighbors is used to retrace the hypothetical 'root' string for recurring motifs. In variations of this approach, as in that described by Keich and Pevzner (34), *(l,k)*-motifs (35) are inferred by generating *k*-neighbors from substrings of length *l* found in a sample sequence. In this concept, a *k*-neighbor is generated by allowing at most *k* substitutions from the sample string. Overrepresented words among those generated as neighbors are then considered as potentially relevant motifs. The term root word will be used herein to refer to *k*-neighbors of strings found in sample sequences.

Here, the same concept of root words is used but for different purposes. We assess sequence composition similarities between two DNA segments taken as a whole. For every string found in a query sequence, all *k*-neighbors of equal length are considered in a root word count and registered in a word table that includes all possible non-redundant words of a given length. For the query and subject sequences, every 5′–3′ string on both strands is used to obtain the root word counts. Hence, only non-redundant 5′–3′ words are considered in the word table that registers the counts.

We use a complete root word count in a query sequence as an expression of its sequence composition profile as a whole to find similarities in other DNA segments within a genomic sequence by performing the same counts on the latter. The resulting method thus relies simply on the root word counts in two distinct DNA segments followed by an assessment of correlation between the two data sets using the Pearson correlation coefficient. In this work, the data sets used to calculate the correlation coefficient were the counts for all non-redundant 5′–3′ 5-nt words. From all possible 5-nt words ($n = 1024$), complementary redundant words are removed ($n = 512$). For each of the possible 512 5-nt non-redundant words, a count is registered for each of their occurrences allowing at most one substitution, except when the substitution creates a CG dinucleotide. CG dinucleotides are excluded to avoid the biased

sequence differences that are the consequence of hypermutable CpGs (36) as obsereved for Alu elements detected in the genomic sequences scanned in this study (data not shown). Both strands are simultaneously considered in the counts and the scan is strand indifferent. The application performing the scans is written in Python and correlation coefficients between the root word counts of the query and a window of the subject sequences are calculated using the 'pearsonr' function of the Scipy package (E.Jones *et al*. Scipy: Open source scientific tools for Python, 2001, http://www.scipy.org). As the sliding window acquires and loses entering and exiting strings respectively, the root word counts are updated in a word array by querying the root word dictionary for every word along the subject sequence. At fixed intervals, the correlation coefficient with the counts of the query sequence is reevaluated. In the present work, the Pearson correlation coefficient was calculated at 10-bp intervals.

Sequences and gene annotations

All genomic sequences and gene annotations in this study are from NCBI build 36.3 of the human genome (reference assembly). Repeat Masker (A.F.A.Smit *et al*. *RepeatMasker Open-3.0*, 1996–2004, http://www.repeatmasker.org) was used to identify repeated elements and mask genomic sequences. CNEs are from the Vista enhancer browser database (10).

The 7SL sequence used as query in this work is that of the human 7SL gene.

RESULTS

Sequence composition similarities between the 7SL sequence and DHS

The β-globin locus was used to show specific sequence composition similarities between the 7SL sequence and TEs as well as functional genomic elements not presumed to be TE-derived. This locus has been thoroughly documented and meticulously mapped for DHS (37) which makes it an ideal candidate to test computational approaches to be used for the identification of functional regulatory elements and most likely to reveal false positives.

Initial sequence similarity searches using our root word count correlation approach, with the β-globin HS2 and HS3 sequences used as queries, led to the identification of other regulatory regions, structural features of the β-globin genes and TEs, most noticeably of Alu sequences (data not shown). Based on these observations, we used our root word count correlation approach to investigate the extent of sequence composition similarities between the 7SL sequence, Alu elements and the DHS of the β-globin LCR as well as other genomic features. While no sequence alignment is possible between the β-globin LCR DHS and 7SL-derived sequences, using the 7SL sequence to scan a 15.5-kb sequence that spans the region harboring HS1 to HS5 of the β-globin LCR clearly highlighted unexpected commonalities between the 7SL sequence and certain DHS (Figure 1). The well documented β-globin HS2, HS3 and HS4 (37) showed sequence composition similarities with the 7SL sequence that were clearly distinguishable from the background with respective r values of 0.44 ($P = 1.33 \times 10^{-25}$), 0.57 ($P = 3.62 \times 10^{-45}$) and 0.34 ($P = 1.30 \times 10^{-15}$) and comparable to those of the Alu elements identified in the same region with Repeat Masker (Figure 1).

Given the high GC content of the 7SL sequence (63%), we performed the same scan as that presented in Figure 1 with 100 consecutively generated shuffled versions of the 7SL sequence to assess whether the GC content was the major contributor to the observed similarities. The results presented in Supplementary Figure S1 show that more than 90% of the shuffled 7SL sequences display lower root-word content correlations than for the same DHS intervals with the original 7SL sequence. Some shuffled versions of the 7SL sequence even yield negative Pearson r values for the same intervals, clearly indicating that the GC content alone does not dictate the results.

Relationship between the 7SL sequence and functional genomic features

Analysis of a 300-kb sequence containing the human β-globin locus using the 7SL sequence as query showed striking sequence composition similarities with several TEs in addition to Alu sequences. Segments of LTRs and L1-LINEs generated similarity peaks comparable to those of Alu sequences (Figure 2A). Interestingly, the similarities with the 7SL sequence found in the four L1-LINEs reside in their 5′-UTRs. The projection of L1-LINEs over the graph along with their orientation clearly shows the location of these similarities (Figure 2A). Scanning the same 300-kb sequence masked for repeated elements (Figure 2B) revealed that every DNA segment exhibiting appreciable sequence composition similarities with the 7SL sequence coincides with a functional feature related to gene structure or regulation. In Figure 2B, the β-globin HS–111 (38) is shown in addition to those already mentioned (Figure 1), while all other annotated similarity peaks correspond to exonic sequences, mainly of olfactory receptor (OR) genes. It should be noted that the coding regions of OR genes also contain regulatory elements (39). Data from the same scan show a clear overlap of similarity peaks with promoters and exonic sequences of the β-globin genes in addition to the DHS (Figure 2C). The δ-globin gene, which shows expression levels 40- to 50-fold less than the β-globin gene (40–42), is the only one not to have a 7SL similarity peak in its promoter/first exon area. The peak overlapping HS4 in Figures 1 and 2A is missing from Panels B and C of Figure 2 because its core sequence is masked by Repeat Masker as an LTR4 element, a solo LTR from human endogenous retrovirus ERV3 (43). The sequence alignment between HS4 and the Repbase consensus LTR4 element [A.F.Smit, LTR4. Direct Submission to Repbase Update (FEB-2000)] spans 133 bp with a sequence identity of 88%. The core sequences of the other β-globin DHS do not contain any sequences derived from TEs based on Repeat Masker

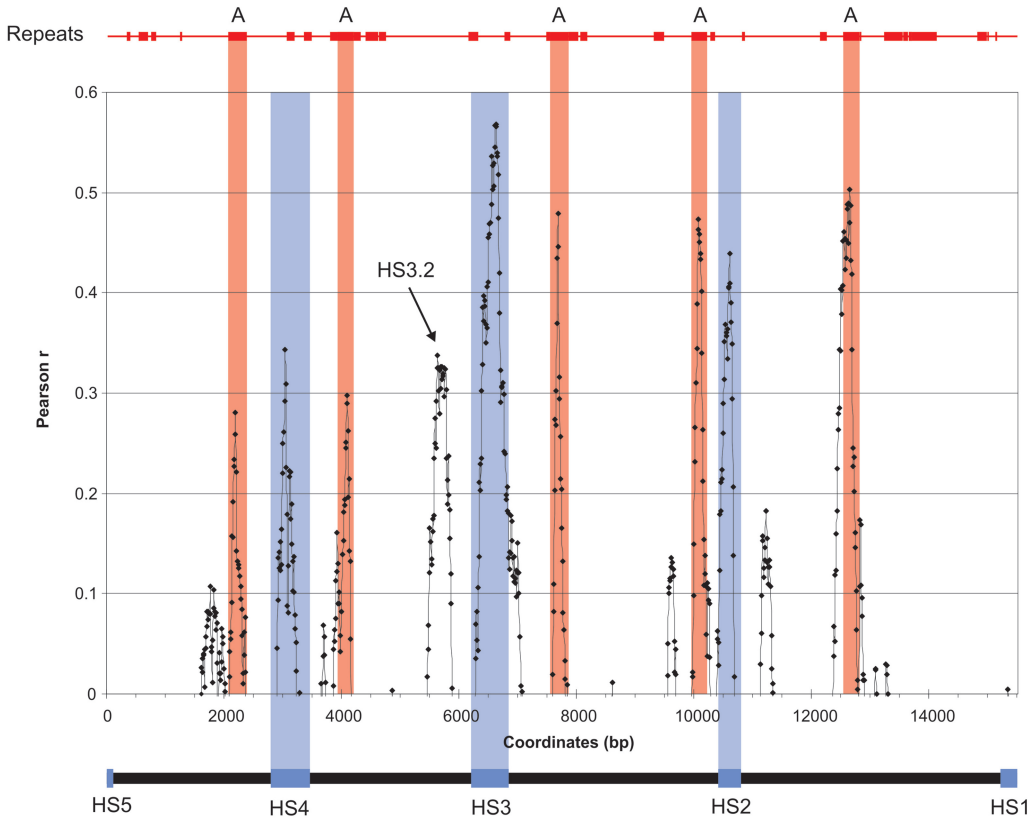

Figure 1. Root word count correlation scan using the 7SL sequence as query. Diamonds represent the center of a 300-bp sliding window. The region shown is that of the human β-globin LCR. The 15.5-kb sequence and annotations of DHS are from NCBI Reference Sequence: NG_000007.3. HS3.2 was described by Molete *et al.* (53). Repeated elements identified by Repeat Masker are shown in red above the graph and Alu sequences ('A') are projected down onto the graph in transparent red. DHS are shown in blue and HS2, HS3 and HS4 are projected upward in transparent blue.

alignments (data not shown). In fact, the sequences of those DHS are unique in the genome as are most CNEs (44), but, as shown here, they are easily identifiable based on sequence composition similarity with the 7SL sequence.

Similarity peaks with the 7SL RNA sequence coincide with DHS

Results obtained from an analysis of the human *TAL1* locus are presented here to further demonstrate the intriguing relationship between 7SL sequence composition similarities and DHS. The *TAL1* locus contains several regulatory elements that have been experimentally characterized (45) as well as several CNEs (10). It was recently mapped for DHS using an array-based approach (7), confirming previous findings, which makes it an appropriate locus to illustrate the capabilities of the method presented in this work. Side by side comparison of experimental DHS mapping data and graphic representation of sequence composition similarities with the 7SL sequence provides a novel perspective on *in silico* DHS predictability.

We show here that the human *TAL1* locus is rich in sequences that display both DNaseI hypersensitivity and sequence composition similarities with the 7SL sequence. Similarity peaks reflecting sequence composition resemblance with the 7SL sequence indentify promoters, exons and DHS with surprising accuracy. Our results show that sequence composition similarity with the 7SL sequence is highly predictive of DHS. The pattern of 7SL similarity peaks in and around the *TAL1* gene (Figure 3) is in striking accordance with the microarray DHS mapping of Follows *et al.* (7) over the same genomic region (see their Figure 4). A careful side by side comparison of their results for the human *TAL1* locus with our assessment of 7SL sequence similarities for the same region reveals a near perfect match with DNaseI hypersensitivity (Supplementary Figure S2). While the 65-kb *STIL* gene has only an isolated peak that coincides with its promoter, the region surrounding the *TAL1* and *PDZK1IP1* genes is densely packed with 7SL sequence similarities, from the −10-kb HS to the +51-kb HS, that faithfully reproduce the pattern of HS/enhancer regions previously described (7). Some of the latter are also documented as CNEs, as in the case of the −4, +20/21, +24 and +51 kb elements (10).

Non-random distribution of 7SL sequence similarities

Figure 4A shows the relative abundance, in a 3.5-Mb region of chromosome 11, of sequences that display sequence composition similarities with the 7SL sequence in both masked (black) with Repeat Masker and

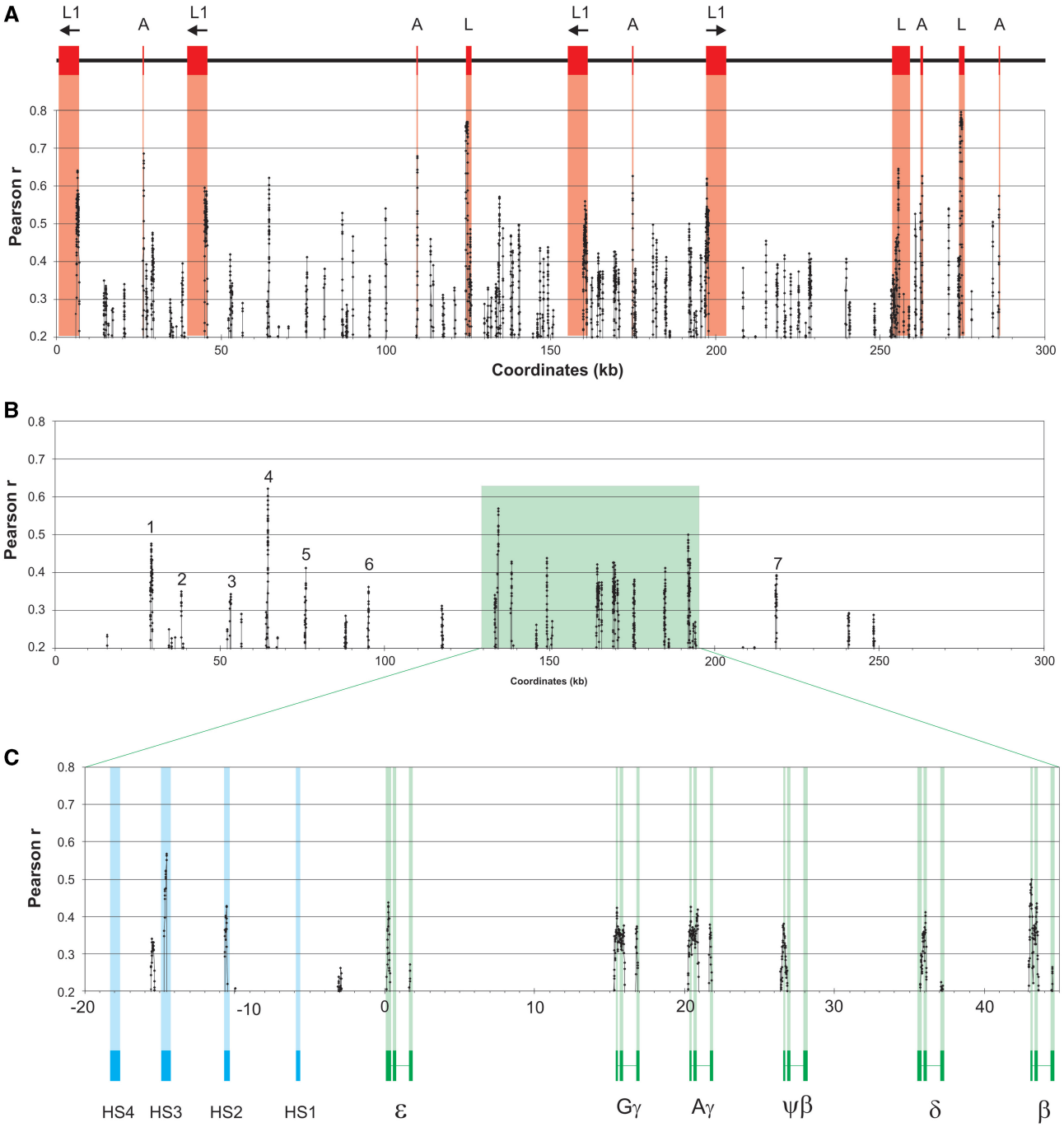

Figure 2. Root word count correlation scan of a 300-kb sequence containing the human β-globin locus using the 7SL sequence as query. Diamonds represent the center of a 300-bp sliding window. The region shown is from human chromosome 11 (chr11: 5 097 000–5 397 000, minus strand). (**A**) The unmasked sequence was analyzed. Relevant transposable elements are shown in red above the graph and projected downward in transparent red. A, Alu sequences and L, LTRs. Arrows indicate orientation of L1-LINEs. In (**B**) and (**C**) sequences masked with a Repeat Masker were analyzed. In (B), relevant similarity peaks outside the region projected in panel C (transparent green) are numbered on the graph: 1, OR51M1; 2, β-globin HS-111; 3, hmm1448204; 4, LOC643745; 5, OR51B5; 6, OR51B2; 7, OR51V1; (**C**) HS1 to HS4 are shown in blue and exons of β-globin genes are shown in green. All features are projected in their respective transparent color onto the graph. Coordinates are relative to the start of exon 1 of the ε-globin gene.

unmasked (red) sequences. Similarity peaks shown on the graph in red are attributable to TEs, mainly Alu elements and L1-LINEs. The similarity peaks shown in black are not presumed to be of transposable origin, as they are not masked by Repeat Masker, and coincide for the most part with exons and promoters. Those peaks show a clustering tendency and reflect local exonic density (data not shown). To illustrate this general observation, a typical 580-kb region taken from the 3.5 Mb sequence of Figure 4A is shown in Figure 4B. It displays patterns of sequence

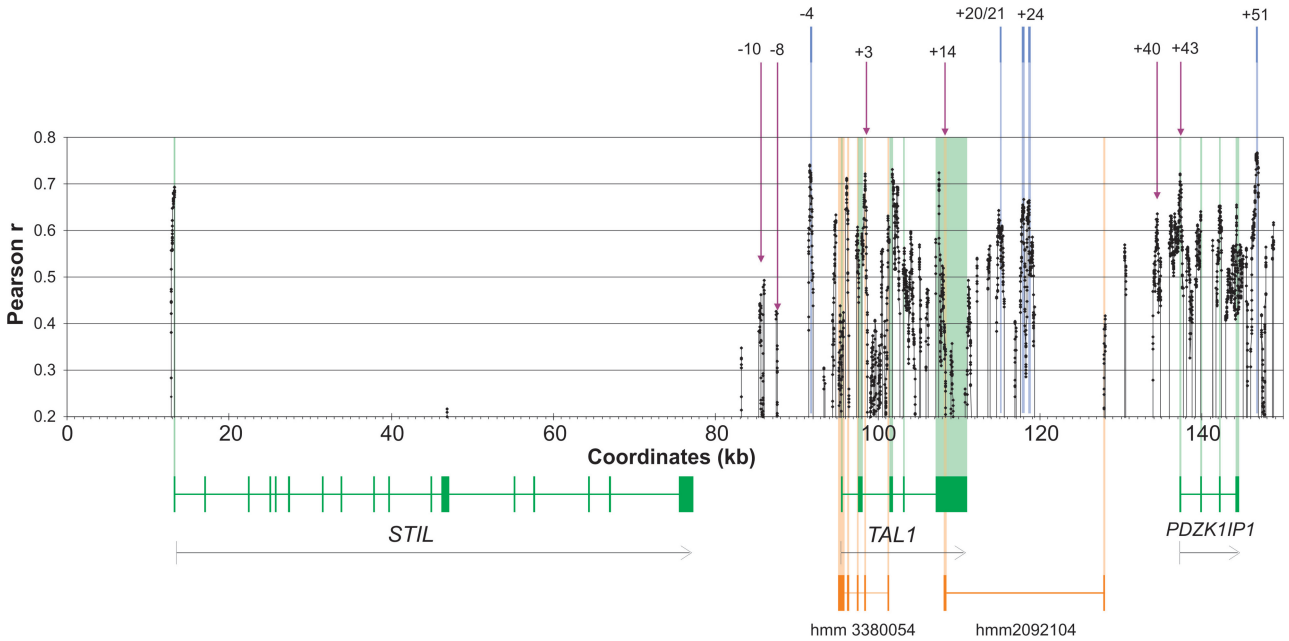

Figure 3. Root word count correlation scan of a 150-kb sequence containing the human *TAL1* locus using the 7SL sequence as query. Diamonds represent the center of a 300-bp sliding window. The region shown is from human chromosome 1. Coordinates are those of Follows *et al.* (7). Exons are in green, exons predicted by Gnomon (NCBI build 36.3) are in orange and CNEs present in the Vista enhancer browser database (10) are in blue. Exons and CNEs are annotated to scale with the graph and projected onto it in their respective transparent colors. Relevant features described by Follows *et al.* (7) are identified relative to the start of exon 1a of the *TAL1* gene from −10 kb to +51 kb directly over the corresponding CNEs, or designated by purple arrows for elements not identified as CNEs.

similarities with the 7SL sequence that are reminiscent of observations made with the β-globin and *TAL1* loci (Figures 2 and 3). 7SL sequence similarity peaks coincide with exons of known or predicted genes (Gnomon annotations, NCBI build 36.3). In regions containing relatively short genes (hmm1068204, hmm6082203, RHOG, FRAG1, *CHRNA10, ART1, ART5* and *TRPC2*), sequence similarities with the 7SL sequence reflect exon density. Note that similarity peaks are only found near the first and last exons in the case of *NUP98* and only in the first exon of *RRM1*. The presence of similarity peaks either at the 5′- or 3′-ends of these genes, as in the case of the *STIL* gene (Figure 3), combined with observations made for the β-globin and TAL1 loci suggests that this is not a random occurrence and is in accordance with the notion that there is a high likelihood of functionality for the genomic segments identified here. Finally, STIM1 is intermediate in that the similarity peaks are concentrated in or near the 5′ and 3′ exons, but are not restricted to the first and last exons.

DISCUSSION

The root word count method presented in this work permits facile detection of sequence composition similarities, which in turn can suggest similar biological properties of unrelated or distantly related DNA segments. It is important, however, to understand that this method is not a pattern discovery approach nor is it a substitute for sequence alignments. The specific use of this method in this work shows the predictability of certain DHS using TEs as query sequences, and raises intriguing questions about the elusive sequence features that may contribute to the high likelihood that such DNA segments will have a functional role either as TEs, coding regions of exons, 5′-UTRs, 3′-UTRs, promoters or distal regulatory regions. While the results presented here show a broad spectrum of functional genomic features with sequence composition similarities with the 7SL sequence, not all such features are detectable. The 7SL sequence, when used as query, reveals the presence of most documented DHS in the β-globin (Figure 2) and *TAL1* (Figure 3) loci and although there is a striking accordance with the experimental DHS mapping of Follows *et al.* (7) for the *TAL1* locus, there is no reason to believe that all DHS can be detected in this manner. Some DHS do not show sequence composition similarities with the 7SL sequence, most notably, HS1 and HS5 of the β-globin LCR (Figure 1). Given the observation that 7SL sequence composition similarities are highly predictive of potentially functional sequence features, some aspects of the approach presented here could be very useful in efforts to perform exhaustive genome wide annotations of functional coding and non-coding sequences. Whereas the annotation of coding sequences is for all intents and purposes complete, there is still limited knowledge—and hence limited annotation—of the gene regulatory features that are subtly spread and scattered across gigabases of genomic sequence. The recent work of Heintzman *et al.* (46), which reports the identification of over 55 000 distal regulatory elements, is a major step

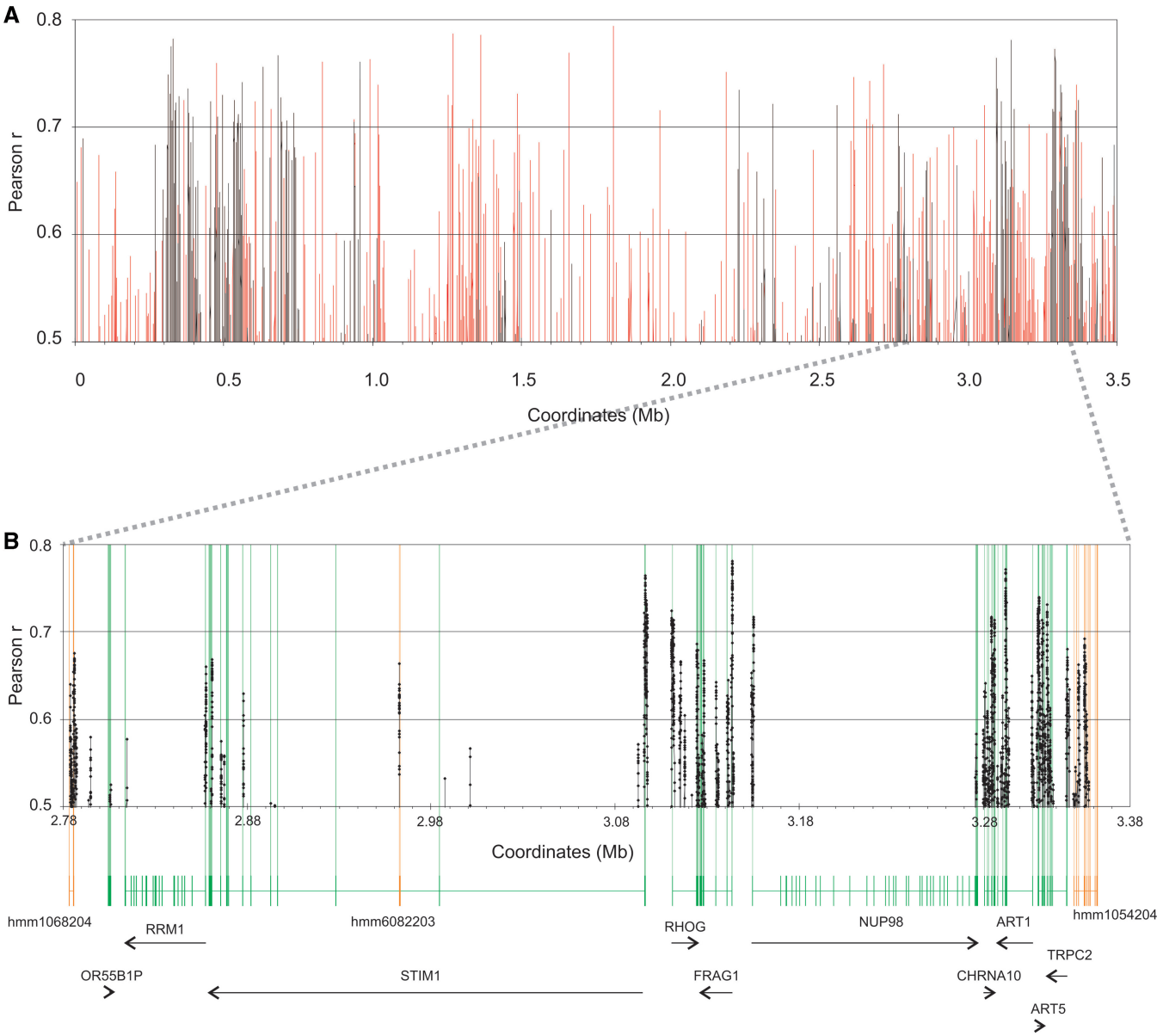

Figure 4. Representative root word count correlation scans using the 7SL sequence as query. Diamonds represent the center of a 300-bp sliding window. Region shown is from human chromosome 11. (**A**) A 3.5-Mb sequence (chr11: 3 430 000–6 930 000, minus strand). Scan results from the unmasked sequence are in red and those from the same sequence masked by Repeat Masker are superimposed in black. (**B**) A 580 000-kb portion of the masked sequence from panel A is shown with gene names and directions of transcription. Exons are in green and exons predicted by Gnomon (NCBI build 36.3) are in orange. Exons are projected onto the graph for all genes except for *RRM1* and *NUP98* for which only the first and last exons are projected to improve clarity.

forward in that regard, but at the same time underlines the work still ahead as this assessment is based on experimental results with only two cell types with the great majority of enhancers being cell-specific. Indeed, with only some 5000 of the enhancers identified in that study that are common to both cell types, it is possible that Heintzman *et al.*'s assessment of cell-specific enhancers in Hela and K562 cells may just be the tip of the iceberg of the total number of enhancers present in the human genome. Their own assessment puts that number somewhere between 10^5 and 10^6. Many such elements may also escape genome-wide experimental detection. The method described here is complementary to such large scale experimental analyses as we cannot assume that any method will be exhaustive.

Distal regulatory segments have proven more difficult to detect than promoters or exons, in essence due to the great diversity of their sequence, internal organization and distance from the genes on which they act. They can reside tens or even hundreds of kb either upstream or downstream of their target gene (47) and are often nested in the introns of neighboring genes (48). Approaches to indentifying such gene regulatory regions by searching for clusters of TF-binding sites (TFBSs),

coupled with gene expression data, have yielded promising results (3). However, knowledge of the *in vivo* binding preferences for most TFs is still incomplete, and therefore cannot be overly relied on when attempting genome-wide identification of gene regulatory regions. Moreover, such an approach is justified on the assumption that the major sequence and functional determinants of distal regulatory regions are the TFBSs that they contain. Ascertaining regulatory regions that are TE-derived on this assumption would be even less reliable given the recent observation that TE-derived TFBSs evolve more rapidly than those which are not repeat-derived sites (49). In such cases, a sequence context based approach, such as the one presented here, may be more successful or serve as a valuable complement to methods that rely on knowledge of TFBSs.

Comparative genomics has been successfully used to identify a large number of CNEs, some of them with *in vivo* confirmations of their functional capacity as exemplified by the Vista enhancer browser database (50). The sequences looked for are either highly conserved (typically $\geq$70% identity over $\geq$100 bp), in distantly-related species such as fish, birds and mammals, or ultraconserved ($\sim$100% identity over $\geq$100 bp), in more closely-related species such as different mammals (1). No preconditions are assumed, except that highly conserved or ultraconserved non-coding sequences are likely to be of regulatory significance. However, the presence of detectable CNEs with regulatory significance inevitably reflects the regulatory nature of conserved processes that have not been subjected to substantial innovations or drift in the sequence composition of their regulatory elements. Therefore, comparative genomics is especially suited and useful for detecting regulatory regions that have been well conserved between species and in cases where gene orthologs can be easily identified in genomic regions with well preserved synteny. Less conserved regulatory regions or more recently acquired lineage- or even species-specific regulatory features may elude these efforts.

In our root word count approach, the starting point is a modular sequence known or suspected to be able to detect unrelated or distantly related regulatory sequences. Other sequences of similar composition are searched for in the same or other genomes. The degree of sequence identity between the query sequence and the identified candidate segments is often <50% making these sequence similarities almost impossible to find by sequence alignment methods. This consideration is especially important as it contrasts with large scale identification of CNEs that does not allow such a low threshold of sequence similarities. It should be noted that the purpose of the work reported here is not to infer common origins for the genomic segments identified but rather commonalities in functional properties.

An important conclusion of this work is that DNA segments exhibiting sequence composition similarities with the 7SL RNA gene, detected using our root word count approach, all harbored functional features of one kind or another. Considering the significant sequence composition similarity, as assessed here (Figure 2A), between the 7SL sequence and segments of L1-LINEs and LTRs, the similarities found in this work with the 7SL sequence in masked sequences raises the possibility that they may be exapted fragments of these retroelements. Moreover, the presumed TEs that could have given birth to these functional features may have been active only in an extinct vertebrate ancestor and it will not be possible to ascertain their alleged contribution unless the right intermediates to do so are found. However, while it is tempting to infer a common origin for the elements identified in this study and the 7SL gene, or any 7SL RNA-derived TE, we should also consider the possibility that the sequence composition similarities highlighted in this work may be products of some form of convergent sequence selection that is characteristic of certain regulatory sequences. Whether these DNA segments are of a common origin or not, they as a whole, share a sequence composition bias that is sufficient to detect their presence in ways that would not be possible by sequence alignments. The usefulness of this approach as a detection method for potentially relevant DHS or other types of functional elements remains, regardless of the true explanation for the observed sequence composition similarities uncovered in this work.

The identification of DHS by Southern blotting in large genomic sequences was a tedious task. New techniques making use of microarray mapping, real-time PCR and massively parallel signature sequencing render possible large scale identification of DHS (5,7,51,52). Genome-wide mapping of such sites has been attempted for a few cell types and has generated thousands of DHS (5,52). A refined determination of the functional core of these sites will allow computational methods, such as the one presented here, to exploit this knowledge to tackle genome-wide mapping of such elements and complement experimental efforts in order to yield better annotations of DHS.

The root word frequency approach that we have developed allows simple assessment of sequence similarity with a query sequence in large genomic sequences with very high sensitivity and a very low apparent false positive rate, although no assessment of true positive versus false positive rates can be presented here since no unified annotation exists for the elements identified in this study. It is rather the findings that a very high fraction, if not all, DNA segments that show strong sequence composition similarities with the 7SL sequence are functional genomic features that should considered foremost in assessing the usefulness of this approach. Our approach complements existing methods by using the sequence composition profile of known regulatory regions or TEs, as a whole, in order to find DNA segments that may harbor telling characteristic sequence features. It brings to light fundamental sequence composition similarities between regulatory DNA segments that are presumed to be unrelated, as most of them are 'unique' sequences in the human genome. Until recently, the major limitation for computational methods in identifying distal regulatory regions was the lack of large experimental training datasets to allow their exhaustive genome-wide detection. The approach presented here opens the door to large scale detection of

distal regulatory regions and DHS using one such element to find many others even in unrelated loci.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Les Instituts de recherche en santé du Canada (grant number MOP-77580).

Conflict of interest statement. None declared.

REFERENCES

1. Visel,A., Bristow,J. and Pennacchio,L.A. (2007) Enhancer identification through comparative genomics. *Semin. Cell Dev. Biol.*, **18**, 140–152.
2. GuhaThakurta,D. (2006) Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Res.*, **34**, 3585–3598.
3. Pennacchio,L.A., Loots,G.G., Nobrega,M.A. and Ovcharenko,I. (2007) Predicting tissue-specific enhancers in the human genome. *Genome Res.*, **17**, 201–211.
4. Gross,D.S. and Garrard,W.T. (1988) Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.*, **57**, 159–197.
5. Crawford,G.E., Holt,I.E., Mullikin,J.C., Tai,D., Blakesley,R., Bouffard,G., Young,A., Masiello,C., Green,E.D., Wolfsberg,T.G. *et al.* (2004) Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proc. Natl Acad. Sci. USA*, **101**, 992–997.
6. Carninci,P., Sandelin,A., Lenhard,B., Katayama,S., Shimokawa,K., Ponjavic,J., Semple,C.A., Taylor,M.S., Engstrom,P.G., Frith,M.C. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.
7. Follows,G.A., Dhami,P., Gottgens,B., Bruce,A.W., Campbell,P.J., Dillon,S.C., Smith,A.M., Koch,C., Donaldson,I.J., Scott,M.A. *et al.* (2006) Identifying gene regulatory elements by genomic microarray mapping of DNaseI hypersensitive sites. *Genome Res.*, **16**, 1310–1319.
8. Nobrega,M.A., Ovcharenko,I., Afzal,V. and Rubin,E.M. (2003) Scanning human gene deserts for long-range enhancers. *Science*, **302**, 413.
9. Hughes,J.R., Cheng,J.F., Ventress,N., Prabhakar,S., Clark,K., Anguita,E., De Gobbi,M., de Jong,P., Rubin,E. and Higgs,D.R. (2005) Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *Proc. Natl Acad. Sci. USA*, **102**, 9830–9835.
10. Visel,A., Minovitsky,S., Dubchak,I. and Pennacchio,L.A. (2007) VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35**, D88–D92.
11. Bohne,A., Brunet,F., Galiana-Arnoux,D., Schultheis,C. and Volff,J.N. (2008) Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome. Res.*, **16**, 203–215.
12. Wheelan,S.J., Aizawa,Y., Han,J.S. and Boeke,J.D. (2005) Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. *Genome Res.*, **15**, 1073–1078.
13. Jordan,I.K., Rogozin,I.B., Glazko,G.V. and Koonin,E.V. (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.*, **19**, 68–72.
14. Makalowski,W. and Toda,Y. (2007) Modulation of host genes by mammalian transposable elements. *Genome Dyn.*, **3**, 163–174.
15. Mikkelsen,T.S., Wakefield,M.J., Aken,B., Amemiya,C.T., Chang,J.L., Duke,S., Garber,M., Gentles,A.J., Goodstadt,L., Heger,A. *et al.* (2007) Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature*, **447**, 167–177.
16. Feschotte,C. (2008) Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.*, **9**, 397–405.
17. Corvelo,A. and Eyras,E. (2008) Exon creation and establishment in human genes. *Genome Biol.*, **9**, R141.
18. Sorek,R., Lev-Maor,G., Reznik,M., Dagan,T., Belinky,F., Graur,D. and Ast,G. (2004) Minimal conditions for exonization of intronic sequences: 5′ splice site formation in *Alu* exons. *Mol. Cell*, **14**, 221–231.
19. Marino-Ramirez,L., Lewis,K.C., Landsman,D. and Jordan,I.K. (2005) Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenet. Genome Res.*, **110**, 333–341.
20. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
21. Capy,P. (2005) Classification and nomenclature of retrotransposable elements. *Cytogenet. Genome Res.*, **110**, 457–461.
22. Bowen,N.J. and Jordan,I.K. (2007) Exaptation of protein coding sequences from transposable elements. *Genome Dyn.*, **3**, 147–162.
23. Wong,L.H. and Choo,K.H. (2004) Evolutionary dynamics of transposable elements at the centromere. *Trends Genet.*, **20**, 611–616.
24. Kramerov,D.A. and Vassetzky,N.S. (2005) Short retroposons in eukaryotic genomes. *Int. Rev. Cytol.*, **247**, 165–221.
25. Matsutani,S. (2006) Links between repeated sequences. *J. Biomed. Biotechnol*, **2006**, 13569.
26. Yang,N. and Kazazian,H.H. Jr (2006) L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nat. Struct. Mol. Biol.*, **13**, 763–771.
27. Marino-Ramirez,L. and Jordan,I.K. (2006) Transposable element derived DNaseI-hypersensitive sites in the human genome. *Biol. Direct.*, **1**, 20.
28. Bush,E.C. and Lahn,B.T. (2006) The evolution of word composition in metazoan promoter sequence. *PLoS Comput. Biol.*, **2**, e150.
29. Dehnert,M., Plaumann,R., Helm,W.E. and Hutt,M.T. (2005) Genome phylogeny based on short-range correlations in DNA sequences. *J. Comput. Biol.*, **12**, 545–553.
30. Fertil,B., Massin,M., Lespinats,S., Devic,C., Dumee,P. and Giron,A. (2005) GENSTYLE: exploration and analysis of DNA sequences with genomic signature. *Nucleic Acids Res.*, **33**, W512–W515.
31. Karlin,S. and Ladunga,I. (1994) Comparisons of eukaryotic genomic sequences. *Proc. Natl Acad. Sci. USA*, **91**, 12832–12836.
32. Karlin,S. and Burge,C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.*, **11**, 283–290.
33. Kohany,O., Gentles,A.J., Hankus,L. and Jurka,J. (2006) Annotation, submission and screening of repetitive elements in Repbase: Repbase Submitter and Censor. *BMC Bioinformatics*, **25**, 474.
34. Keich,U. and Pevzner,P.A. (2002) Finding motifs in the twilight zone. *Bioinformatics*, **18**, 1374–1381.
35. Pevzner,P.A. and Sze,S.H. (2000) Combinatorial approaches to finding subtle signals in DNA sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 269–278.
36. Walser,J.C., Ponger,L. and Furano,A.V. (2008) CpG dinucleotides and the mutation rate of non-CpG DNA. *Genome Res.*, **18**, 1403–1414.
37. Mahajan,M.C., Karmakar,S. and Weissman,S.M. (2007) Control of beta globin genes. *J. Cell Biochem.*, **102**, 801–810.
38. Papachatzopoulou,A., Kaimakis,P., Pourfarzad,F., Menounos,P.G., Evangelakou,P., Kollia,P., Grosveld,F.G. and Patrinos,G.P. (2007) Increased γ-globin gene expression in β-thalassemia intermedia patients correlates with a mutation in 3′HS1. *Am. J. Hematol.*, **82**, 1005–1009.
39. Merriam,L.C. and Chess,A. (2007) *cis*-Regulatory elements within the odorant receptor coding region. *Cell*, **131**, 844–846.
40. Tang,D.C., Ebb,D., Hardison,R.C. and Rodgers,G.P. (1997) Restoration of the CCAAT box or insertion of the CACCC motif activates δ-globin gene expression. *Blood*, **90**, 421–427.
41. Steinberg,M.H. and Adams,J.G. III (1991) Hemoglobin A2: origin, evolution, and aftermath. *Blood*, **78**, 2165–2177.

42. Humphries,R.K., Ley,T., Turner,P., Moulton,A.D. and Nienhuis,A.W. (1982) Differences in human α-, β- and δ-globin gene expression in monkey kidney cells. *Cell*, **30**, 173–183.
43. Cohen,M., Powers,M., O'Connell,C. and Kato,N. (1985) The nucleotide sequence of the env gene from the human provirus ERV3 and isolation and characterization of an ERV3-specific cDNA. *Virology*, **147**, 449–458.
44. Kamal,M., Xie,X. and Lander,E.S. (2006) A large family of ancient repeat elements in the human genome is under strong selection. *Proc. Natl Acad. Sci. USA*, **103**, 2740–2745.
45. Delabesse,E., Ogilvy,S., Chapman,M.A., Piltz,S.G., Gottgens,B. and Green,A.R. (2005) Transcriptional regulation of the SCL locus: identification of an enhancer that targets the primitive erythroid lineage in vivo. *Mol. Cell Biol.*, **25**, 5215–5225.
46. Heintzman,N.D., Hon,G.C., Hawkins,R.D., Kheradpour,P., Stark,A., Harp,L.F., Ye,Z., Lee,L.K., Stuart,R.K., Ching,C.W. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.
47. Levine,M. and Tjian,R. (2003) Transcription regulation and animal diversity. *Nature*, **424**, 147–151.
48. Dean,A. (2006) On a chromosome far, far away: LCRs and gene expression. *Trends Genet.*, **22**, 38–45.
49. Polavarapu,N., Marino-Ramirez,L., Landsman,D., McDonald,J.F. and Jordan,I.K. (2008) Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. *BMC Genomics*, **9**, 226.
50. Pennacchio,L.A., Ahituv,N., Moses,A.M., Prabhakar,S., Nobrega,M.A., Shoukry,M., Minovitsky,S., Dubchak,I., Holt,A., Lewis,K.D. *et al.* (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, **444**, 499–502.
51. Follows,G.A., Janes,M.E., Vallier,L., Green,A.R. and Gottgens,B. (2007) Real-time PCR mapping of DNaseI-hypersensitive sites using a novel ligation-mediated amplification technique. *Nucleic Acids Res.*, **35**, e56.
52. Crawford,G.E., Holt,I.E., Whittle,J., Webb,B.D., Tai,D., Davis,S., Margulies,E.H., Chen,Y., Bernat,J.A., Ginsburg,D. *et al.* (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.*, **16**, 123–131.
53. Molete,J.M., Petrykowska,H., Sigg,M., Miller,W. and Hardison,R. (2002) Functional and binding studies of HS3.2 of the beta-globin locus control region. *Gene*, **283**, 185–197.