Genomics 93 (2009) 314–322

Contents lists available at ScienceDirect

Genomics

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

# Discovering sequences with potential regulatory characteristics

Minou Bina [a,*], Phillip Wyss [a], Sheryl A. Lazarus [a], Syed R. Shah [a], Wenhui Ren [b], Wojciech Szpankowski [b], Gregory E. Crawford [c], Sang P. Park [d], Xiaohui C. Song [d]

[a] Department of Chemistry, Purdue University, West Lafayette, IN 47907, USA
[b] Department of Computer Sciences, Purdue University, West Lafayette, IN 47907, USA
[c] Institute for Genome Sciences and Policy, Duke University, Durham, NC, 27708, USA
[d] Rosen Center for advanced computing, Purdue University, West Lafayette, IN 47907, USA

ARTICLE INFO

*Article history:*
Received 15 October 2007
Accepted 17 November 2008
Available online 30 December 2008

*Keywords:*
Human genome
Gene regulation
Regulatory signals
Codes in human DNA
Sequence context
Transcription factor binding sites
Genetic vocabulary

ABSTRACT

We developed a computational model to explore the hypothesis that regulatory instructions are context dependent and conveyed through specific 'codes' in human genomic DNA. We provide examples of correlation of computational predictions to reported mapped DNase I hypersensitive segments in the HOXA locus in human chromosome 7. The examples show that statistically significant 9-mers from promoter regions may occur in sequences near and upstream of transcription initiation sites, in intronic regions, and within intergenic regions. Additionally, a subset of 9-mers from coding sequences appears frequently, as clusters, in regulatory regions dispersed in noncoding regions in genomic DNA. The results suggest that the computational model has the potential of decoding regulatory instructions to discover candidate transcription factor binding sites and to discover candidate epigenetic signals that appear in both coding and regulatory regions of genes.



## Introduction

A relatively long history supports the idea that genomic DNA represents a text, or a language, and that the order and the location of 'words' in that text would define the genetic information [1–3]. In fact, the determination of the genomic DNA sequences has brought linguistic metaphors to new heights: referring to DNA as a language and to the human genome as the 'book of life' [1,2]. Support for description of information in DNA as a text has emerged from the formulation of codons (words) for specifying the amino acid sequence of proteins [4,5].

Regulatory signals are generally assumed to not occur in the coding regions of genes. However, a theoretical model proposes that in addition to coding for proteins, the exons of genes may include information for other biologically meaningful signals such as binding sites for regulators of transcription [6,7].

*A priori* one could expect that regardless of their position in genomic DNA, regulatory signals may share common characteristics. For example, functional transcription factor binding sites have been localized not only near the beginning of genes but also in control regions within intronic sequences and in regulatory regions localized far upstream and far downstream of transcription start sites [8]. In addition to binding sites for transcription factors, regulatory segments might also include underlying signals for controlling other aspects of gene expression. For example, the arrangement of A-tracts in regulatory regions might provide signals for bending DNA to influence the three-dimensional architecture of the sequences in these regions of chromosomes [9]. Occurrences of CpG containing elements may provide epigenetic information and signals for methylation of DNA [10,11].

Several high-throughput procedures have been developed for mapping the position of regulatory regions of genes. Localization of DNase I hypersensitive (HS) sites in chromatin has emerged as a powerful experimental tool for mapping the regulatory regions that are poised for activation of gene expression [12–15]. This method has a long history and has withstood the test of extensive validations [14,15].

Computational models have also aimed at predicting the position of regulatory regions in genomic DNA. These models are often based on specific hypotheses. Examples include the hypothesis that clustering of transcription factor binding sites in a given region of genomic DNA reflects the presence of a regulatory segment: for example see [16,17], reviewed in [18]. Another model is based on the hypothesis that functional regulatory sequences could be subject to evolutionary selection, leaving a signature that could be detected in alignments of genomic DNA sequences from several species: for example see [19–21].

*Abbreviations:* CDSs, Coding sequences; DNase I, Deoxyribonuclease I; HS, Hypersensitive; HSSs, Hypersensitive sites; TFBS, Transcription factor binding sites; TSSs, Transcription start sites.

* Corresponding author. Fax: +1 765 494 0239.
*E-mail address:* Bina@Purdue.edu (M. Bina).

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.11.008

Our hypothesis is that the human genome has evolved to produce a well-defined "language" for conveying regulatory information in the DNA. To explore this idea, previously we examined characteristics of 9-mers collected from proximal-promoters of protein-coding genes [22,23]. Based on experimental data, we assumed that regulatory signals should occur frequently in regions preceding the TSSs. We chose 9-mers because they seemed to be a relatively "good" length for discovering the genomic context of regulatory signals including transcription factor binding sites [22]. We chose constant length DNA in order to reduce computational burden due to short sequences that appeared frequently in genomic DNA [22,24]. The computational model assumes that complementary 9-mers are equivalent. This assumption is based on studies showing that TFBSs can exert control on gene expression irrespective of their orientation in DNA: for example see [25–27].

In this report, we present a computational model (weighted density plots) for identifying the genomic DNA regions that include statistically significant occurrences of 9-mers collected from promoter and coding regions of human genes. We find that, in these plots, specific peaks often correlate with experimentally mapped regulatory regions in genomic DNA.

## Results

### *Sampling of characteristics of 9-mers from human promoter and coding sequences*

We determined the frequency of occurrences of complementary 9-mers in three groups of human DNA: a set of promoters, defined with respect to the 5′ end of ESTs [22]; coding regions in cDNAs obtained from GenBank [28]; and sequences corresponding to a draft of total genomic DNA [22]. The initial goal was to use 9-mers from CDSs as a control for 9-mers derived from the promoter regions. However, as detailed in subsequent sections, unexpectedly we found that a subset of 9-mers from CDSs also appeared frequently in regulatory regions of genes.

Furthermore, previous methods for identifying over-represented n-mers for *de novo* pattern detection [29] have not addressed the problem of sequences that appear frequently in genomic DNA. To resolve this problem, we normalized the frequency of the 9-mers in promoter regions and in coding sequences with respect to their corresponding occurrences in total genomic DNA [22].

Ranking of frequencies provides a statistical measure of the relative abundance of 9-mers in promoters or CDSs, with respect to their corresponding occurrences in total genomic DNA. For example, in the statistical scheme, ranking of 1 corresponds to those 9-mers that appear equally in promoters and total genomic DNA. Thus, rankings greater than 1 statistically could be significant: a ranking of 3.08 had a $p$ (or $\beta$) value of about $10^{-27}$; a ranking of 7 had a $\beta$ value of about $10^{-50}$ [22].

The computational model uses the collected 9-mers to produce weighted density plots to predict the position of potential regulatory signals in human genomic DNA. The program employs a specified window to scan the human genomic DNA. The program examines all possible 9-mers in each window and then finds their computed ranks. The program uses the ranks to determine a weighted sum. As the window slides along a genomic DNA, the weighted sums would produce intensity values at each nucleotide position.

In addition to ranks, we wished to apply specific criteria to distinguish the 9-mers that occurred preferentially in non-coding regions from those that appeared frequently in CDSs. Towards this goal, we created three types of density plots (Fig. 1). We constructed a plot (CDS_Hits) to view the weighted density of matched of genomic DNA sequences with 9-mers collected from coding sequences. The ranking procedure excluded 9-mers that appeared frequently in genomic DNA. We imposed specific criteria (see discussion and methods) to construct a plot (Reg_Signal Pred1) to display the weighted density of matches of genomic DNA sequences with 9-mers collected from the promoter regions of genes. Additional criteria (see methods) were imposed to construct a plot (Reg_Signal Pred2) to distinguish the 9-mers with "high" non-coding context from those with a relatively high coding context. In that plot, intensities greater than one reflect normalized values of 9-mers that appear more frequently in promoters than in coding regions of genes.

### *Analysis of the HOXA locus on human chromosome 7*

To evaluate the computational model, we analyzed several relatively long genomic DNA segments selected to include many genes. As an example, we highlight the results obtained for the HOXA cluster of genes on human chromosome 7. The cluster is relatively long and includes HOXA1, HOXA2, HOXA3, HOXA4, HOXA5, HOXA6, HOXA7, HOXA9, HOXA11, HOXA13, and EVX1 (Supplemental Fig. 1). We evaluate the intensities of predicted signals in the context of genomic positions of mapped DNase I HS segments [15,30,31].

Hypersensitivity to DNase I provides a relatively robust measure of the chromosomal regions that have an "open" chromatin structure [13]. A relatively large body of experimental data indicates that the DNase I HS segments in genomic DNA are either nucleosome-free or contain modified nucleosomal structures [10,14]. Accessibility of these segments is thought to expose the control signals in the DNA, for recognition by the regulators of gene expression [14]. Evidence indicates that the results of high-throughput methods are likely to be accurate since the methods have correctly identified the HS segments mapped by conventional techniques [30].

To compare the position of DNase I HS segments to the predictions of the computational model, we display the results in custom tracks in the genome browser at UCSC. Initially, we will qualitatively compare the position of the peaks in the density plots to the mapped HS segments: Fig. 1, tracks appearing under Duke/NHGRI DNase I-hypersensitivity and under UW/Regulome QCP DNase I Sensitivity [15]. A subsequent section provides statistical evidence for the correlations.

Supplemental Fig. 1 gives an overview of the position of HS segments and the organization of the genes in the HOXA locus. We analyzed the entire locus, using a sliding window of 30 bp. Results show that peaks in the weighted density plots are primarily localized within the gene-rich segments (Supplemental Fig. 1). In the custom tracks, intensity of peaks is displayed as pixilated bars, in order to produce condensed plots (for example, see Fig. 1). Fig. 2 shows an example of full-display of density plots with respect to landmarks including potential TFBSs.

Fig. 1 shows an expanded view of the genomic DNA region that includes HOXA1. Three custom tracks display the predictions. The displayed predictions provide typical results showing that statistically ranked 9-mers from promoter regions may appear as clusters in the vicinity of transcription start sites and in sequences further upstream (Fig. 1, track labeled Reg_Signal Pred1). Predictions also provide typical results showing that statistically ranked 9-mers from CDSs appear to cluster not only in exonic regions but also in non-coding sequences: in that example, upstream of TSSs of HOXA1 (Fig. 1, track labeled CDS_Hits).

A comparison of the tracks shown in Fig. 1 reveals correspondence of the positions of predicted regulatory signals to the experimentally determined DNase I HS segments that include the 5′ end of HOXA1. Also, there is a correspondence between predicted regulatory signals in a region upstream of the transcription start site (~3700 bp) and HS segments in chromatin isolated from several cell lines (Fig. 1, tracks labeled GM069, CD4, HeLa, and $CaCO_2$). Fig. 1 shows that the mapped DNase I HS segments may include the transcribed untranslated region and the coding region of a gene.

Supporting information gives additional examples of correlation of predicted regulatory signals within DNase I HS segments in both

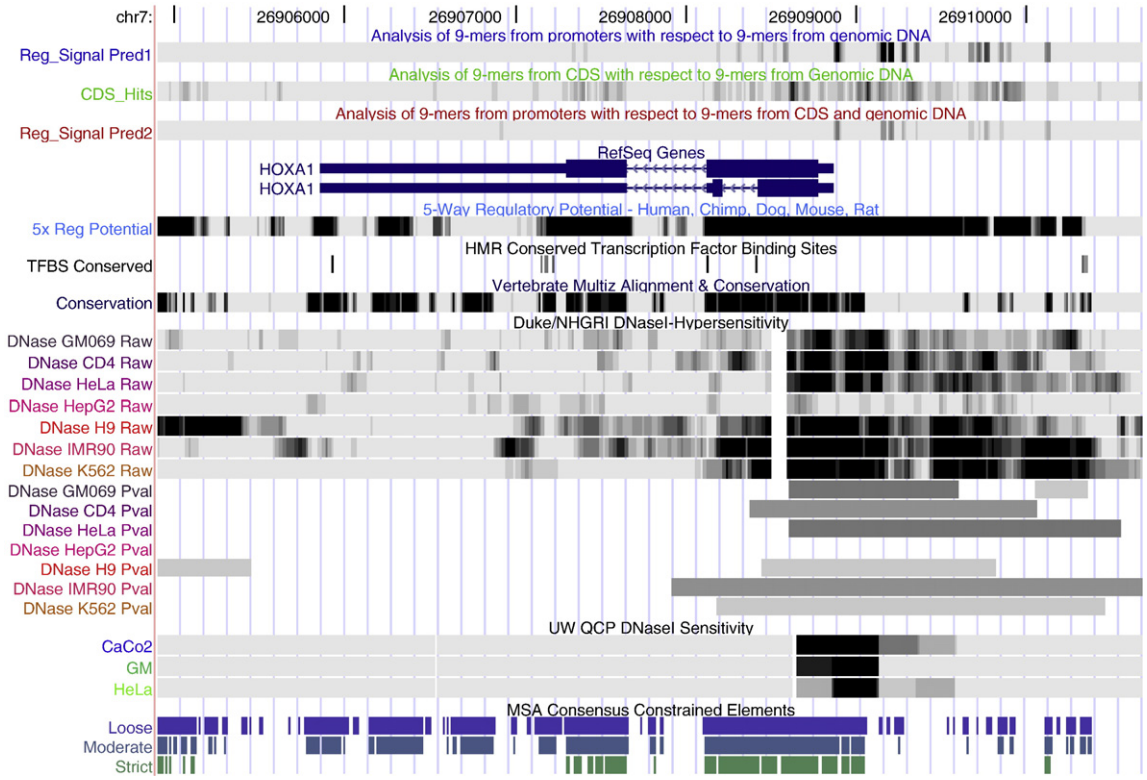


**Fig. 1.** Predictions obtained for the HOXA1 region (human chr7:26,904,901-26,910,680, built hg_17). The top three tracks display condensed views of the weighed density plots. From the options at the UCSC browser [48], we selected the following keys as landmarks: a track providing the genomic position of known genes [49]; a track (5× Reg Potential) displaying regulatory potential scores, computed from alignments of human, chimpanzee, mouse, rat, and dog genomic DNA sequences [20]; a track (TFBS conserved) displaying the inferred location and score of potential transcription factor binding sites conserved in the human/mouse/rat sequence alignments; a track (conservation) providing a measure of evolutionary conservation in 17 vertebrates, including mammalian, amphibian, bird, and fish species, based on a phylogenetic hidden Markov model, phastCons [50]; and tracks corresponding to experimentally determined DNase I HSSs and HS segments [15].

coding and noncoding regions of genes in the HOXA locus: HOXA2 (Supplemental Fig. 2); HOXA3 (Supplemental Fig. 3); HOXA4 (Supplemental Fig. 4); HOXA5 (Supplemental Fig. 5); HOXA6 (Supplemental Fig. 6); HOXA7 (Supplemental Fig. 7); HOXA9 (Supplemental Fig. 8); HOXA10 (Supplemental Fig. 9); HOXA11 (Supplemental Fig. 10); HOXA13 (Supplemental Fig. 11); and EVX1 (Supplemental Fig. 12).

For example, the plot of the HOXA5 region shows that a cluster of 9-mers from promoter regions occurs downstream of the gene (~2000 bp), in a region that includes several mapped DNase I HSSs (Supplemental Fig. 5). This noncoding segment also encompasses many clusters of 9-mers from coding sequences (Supplemental Fig. 5, compare the track-labeled CDS_Hits with the tracks labeled DNase GM069 and DNase CD4). The plot of a region that includes HOXA7 provides another example of occurrences of 9-mers collected from coding sequences in a DNase I HS segment localized upstream of a gene (Supplemental Fig. 7).

Inspection of the plots also provides examples of regions that do not include sequences with potential regulatory and/or CDS characteristics. This can be seen in the overview plot that includes the predictions for the entire HOXA locus. The plot shows that predicted regulatory signals occur very sporadically in segments beyond the gene-rich regions in the locus (Supplemental Fig. 1).

### *Examples of complete "words" and "sentences" with complex lexical features*

We imagined that 9-mers collected from promoter regions may include the symbolic features that impart the characteristics as well as the context of the regulatory signals in genomic DNA [22]. Natural-language and transcribed speech are prime examples of sequential symbolic data that are context dependent [32]. Emerging linguistic models also support important roles that sequence context could play in the selection of transcription factors that regulate gene expression [23,33]. Furthermore, linguistic paradigms imply that a comprehensive vocabulary could help with capturing syntactic and contextual features of clauses and sentences that describe the regulatory information in genomic DNA, for examples see [22,32].

To detect syntactic and contextual features, the algorithm for the density plots was designed to reflect two parameters: the computed statistical rank of each 9-mer in the sliding window and occurrences that map to overlapping positions. Consequently, both the rank and overlapping occurrences of 9-mers contribute to the intensity of the signal observed at each nucleotide position in the genomic DNA.

*A priori*, we expected three scenarios: (1) statistically significant 9-mers might include a complete "word"; (2) a short segment containing several overlapping 9-mers might correspond to composite "words" or "words" that were longer than 9 base-pairs; and (3) extensive overlapping occurrences might define "clause" and "sentences"; i.e. regulatory modules and possibly regulatory regions [22].

Since TFBSs are the most notable examples of regulatory words in genomic DNA, a previous study has examined whether isolated 9-mers include binding sites for known transcription factors [22]. Within a subset of the highly ranked 9-mers, the study identified binding sites for several known transcription factor families including CREB, ETS, EGR-1, SP1, KLF, MAZ, HIF-1, and STATs [22].

Similarly, we find that in some cases, peaks in the density plots correspond to potential TFBSs. For example, Fig. 2 provides an expanded view of density plots obtained for a 630 bp segment localized downstream of HOXA13 and far upstream of HOXA11 (the boxed region in Supplemental Fig. 11). The shown segment includes four potential

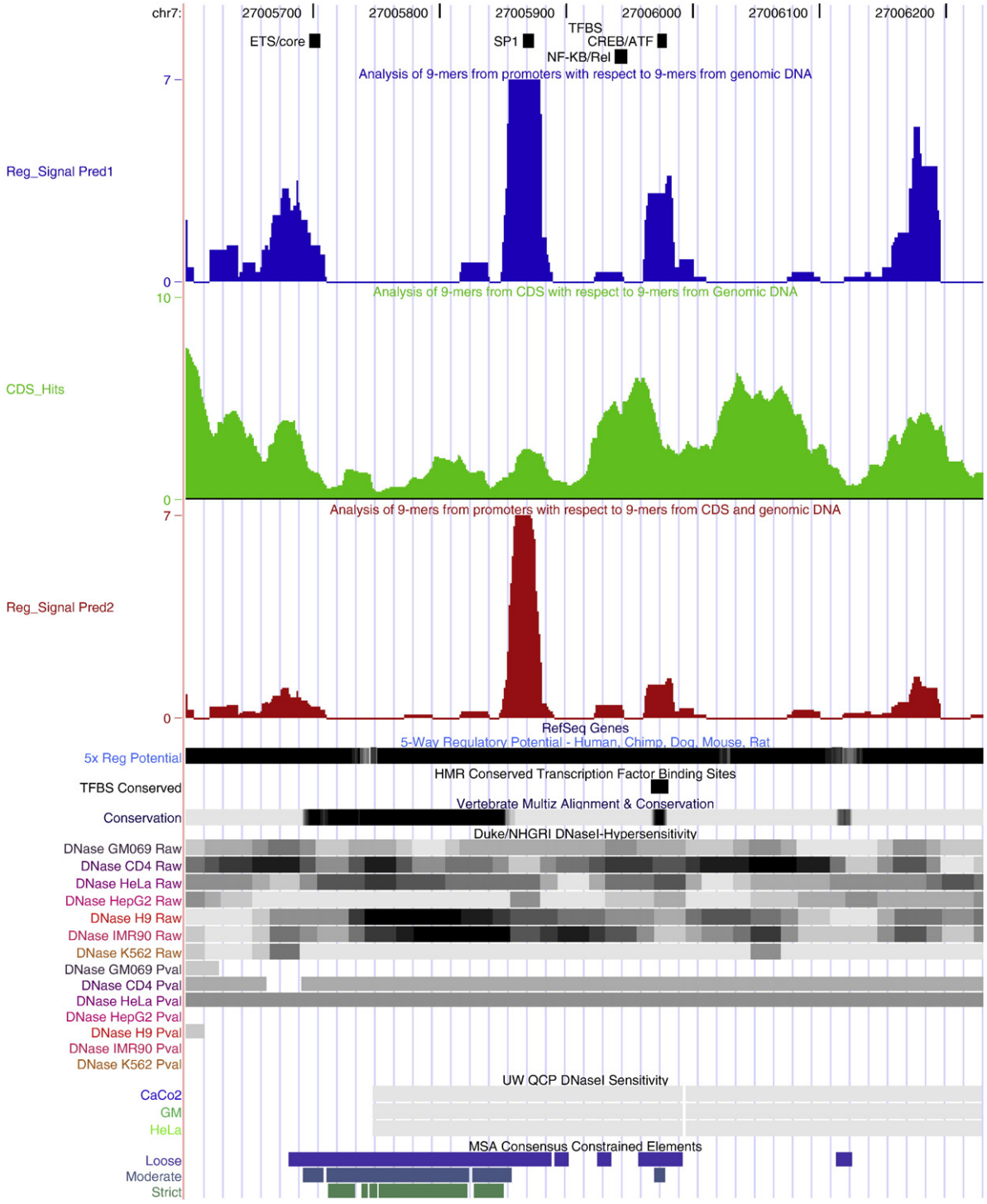


**Fig. 2.** Predictions obtained for a 630 bp intergenic region downstream of HOXA13 and far upstream of HOXA11 (chr7:27,005,601-27,006,230). The figure shows an expanded view of the density plots. The track above the blue plot shows the position of potential TFBSs. The track labeled TFBS conserved was created at UCSC by M. Weirauch and B. Raney, to provide computed matrices and scores for binding sites selected from TRANSFAC, using conservation as a criterion. Three tracks, at the bottom of the figure, display consensus elements generated by the ENCODE Multi-Species Analysis group using three sequence alignment methods (TBA, MLAGAN, and MAVID) and nine different combinations of three conservation algorithms (phastCons, binCons, and GERP).

TFBs: a site for members of the ETS family, a site for the SP1 family, a site for the NF-κB/Rel family, and a site for the CREB/ATF family (Fig. 2). We find a good correspondence of the position of the CREB/ATF site to a peak in the plot of 9-mers derived from promoter regions (Supplemental Fig. 13, the blue plot). This site falls within a conserved TFBS (Supplemental Fig. 13, the tracks labeled TFBS conserved and conservation). We also find a good correspondence of the position of the SP1 site to a peak in the plot of 9-mers derived from promoter regions (Supplemental Fig. 14, the blue and the red plots). The SP1 site is not within but near a relatively conserved region. The ETS site falls between the tail-end of a broad peak in the density plot and a region that is conserved in human, chimp, rhesus, rabbit, dog, and cow DNA but not in the sequences of other species (Supplemental Fig. 15).

Supplemental Fig. 16 provides another example of density plots that include frequent occurrences of high-ranking 9-mers from promoter regions (Reg_Signal Pred 1, the blue plot) and numerous occurrences of high-ranking 9-mers from coding sequences (CDS_Hits, the green plot). The figure displays a region that includes the TSS of HOXA1 and sequences further upstream. Within the expanded density plot of Reg_Signal Pred 1 and Reg_Signal Pred 2, we observe several well-resolved peaks (labeled 1–11).

The position of three peaks correlates with potential TFBSs: MAZ (peaks 3 and 11); AP2 (peak 7). Three peaks fall within highly conserved sequences (peaks 1, 9, and 11). There are also peaks that correlate with sequences that are loosely conserved (examine peaks 2, 3, 4, 6, 10 with respect to the tracks that appear under MSA, Consensus Constrained Elements, created by the ENCODE project). Overall, the plots include many examples of predicted signals that correlate with sequences conserved in primates but not in other species. For example, Supplemental Fig. 17 shows two predicted regulatory signals in a region far downstream of HOXA1. The sequence in this region appears to be conserved in human, chimp, and rhesus DNA but not in other species. One of the predicted signals encompasses a potential site for binding E2F and related proteins.

Generally, it is difficult to create a non-redundant, complete, and accurate list of known transcription factor binding sites for testing computational models. The TRANSFAC database is not suitable for that purpose because it includes numerous redundant entries. To resolve this problem, we compiled a non-redundant list. However, we found that as expected, short or poorly defined binding sites appeared frequently in genomic DNA. Removing these sites from the list would create datasets producing false-negatives. The challenge represents a recurrent and unresolved problem for localization of TFBSs [29]. Nonetheless, positive correlations validate the predictive potential of our approach. Furthermore, predictions that do not include a known binding site may reflect regulatory signals currently unknown.

If statistically significant 9-mers represent words describing regulatory signals (i.e. TFBSs), we could expect to find instances in which overlapping 9-mers would produce sentences and clauses consisting of regulatory instructions in the genomic DNA. To explore this idea, we scanned the density plots of HOXA locus to find high intensity peaks that spanned a stretch of DNA sequence. To enhance the statistical significance of the findings, we imposed the criterion that the signal intensity should be greater than 2 (Supplemental Tables 1–3). For 9-mers derived from promoter regions (Reg_Predict 1), we found nearly 275 instances of peaks produced from overlapping 9-mers (Supplemental Table 1). A number of the overlaps created segments consisting of 30 or more nucleotides. The longest overlapping run (44 bp) appeared in the genomic region corresponding to an intron in the long transcript of HOXA13.

The number of overlapping runs increased when we reduced the threshold intensity to 1. Three histograms display the frequency of length distributions for that threshold. For Reg_Pred1, the length of overlaps reached nearly 75 base pairs. The most frequent runs were about 23 base pairs and appeared both within and outside the mapped HS segments (Supplemental Fig. 18). For Reg_Pred2, promoter 9-mers that were ranked with respect to both genomic DNA and CDSs, we obtained a somewhat skewed distribution for runs that appeared outside and within HS segments (Supplemental Fig. 19). As expected, for CDS_Hits, we obtained relatively long segments produced from overlapping 9-mers in the exonic regions (Supplemental Fig. 20). The histogram indicates that a small fraction of coding sequences appear in HS segments in chromatin (Supplemental Fig. 20).

We have attempted to examine whether we could extract contextual features, from regions that contain runs of overlapping 9-mers. Supplemental Fig. 21 provides an example. This example displays a close-up view of density plots of a region within an intron of HOXA13. To determine contextual features of that segment, we interrogated our database [34] to identify the promoter sequences that included the 9-mers that appeared in HOXA13 intron. From the output of the queries, we selected genes that exert control on gene expression. Table 1 summarizes the results. Close inspection of the list reveals that a significant fraction of the genes falls into two categories: genes that function in chromatin remodeling and epigenetic control (i.e. HDAC1, HDAC3, HDAC6, HDAC11, EZH2, SIRT1, SIRT3, SUV39H2, CITED1, CITED2, SMARCD3, and DNMT3L); and genes for transcription factors that regulate gene expression during development (i.e. PAX6, PITX1, PITX3, OTX1, HOXC13, POU2F2, LHX4, MEIS2, GATA1, GATA3, HES2, HES6, HEYL, SOX9, SOX10, MEF2B, KLF5, NEUROD3, RARA, NR1H3, and NR2F1). From that listing, we predict that the segment in the intron of HOXA13 includes overlapping regulatory words (*cis*-elements) that function in the networks that exert control on both development and chromatin remodeling. Note that listings of this type are often created to identify genes that might be co-regulated through specific *cis*-elements in promoter regions of genes, for example see [35]. We propose that such elements also occur in regions distal to TSSs and

**Table 1**
A listing of genes whose promoter contains the 9-mers in the intron of HOXA13 (chr7:27,011,750-27,011,777)

| 9-mer | Rank | Gene |
|---|---|---|
| GGGCTGGGG | 4.00 | PITX1, MXD4, HIF1A, **HDAC3**, **SIRT3**, NRBP1, DNMT3L, **SMARCD3**, NFKB2, SPIB, HES6 |
| GGCTGGGGA | 2.49 | **CITED1**, PITX1, NFKBIA, **SIRT1**, MED27, SOX10, RARA, IRF1, TRIP4 |
| GCTGGGGAG | 2.67 | TBPL1, GRLF1 |
| CTGGGGAGG | 1.62 | MED27, GATA1 |
| TGGGGAGGG | 2.02 | PITX3, SOX9 |
| GGGGAGGGC | 4.27 | MEF2B, NR1H3, NFE2L2, TRIP6, ELF3, MED6, OTX1, TAF6L, MYCL1, NFX1, NFIB |
| GGGAGGGCG | 15.99 | MEF2B, **HDAC1** |
| GGAGGGCGG | 14.20 | **CITED2**, MEF2B, **HDAC1**, NAP1L1, MYCN, MED6, TAF6L |
| GAGGGCGGG | 24.36 | **CITED2**, MEF2B, OTX1, ATF4, RARA, TAF6L, PITX3, MYCL1, MYCN |
| AGGGCGGGG | 17.87 | **CITED2**, MEF2B, OTX1, ATF4, RARA, TAF6L, PITX3, MYCL1, NFE2L2, CREB3, **HDAC11**, TRIP13, NCOA4, SOX10, NFKB2, TAF9 |
| GGGCGGGGC | 49.86 | **CITED2**, MEF2B, PITX3, HES2, HOXC13, NFYB, NFKBIE, EZH2, NFE2L2, NFATC3, NRBP1, HEYL, MYBL2, KLF5, IRF1, LHX4, PAX6 |
| GGCGGGGCG | 61.13 | **CITED2**, **HDAC1**, MXD4, HES2, HOXC13, NFYB, NFKBIE, EZH2, **SUV39H2**, **HDAC6**, POU2F2, MYBL2, NEUROD3, LHX4, TAF10 |
| GCGGGGCGC | 63.24 | MXD4, XBP1, **SIRT3**, NR2F1 |
| CGGGGCGCA | 44.20 | NR2F1, NFIB |
| GGGGCGCAG | 23.01 | NR2F1, **SUV39H2**, BTF3, TFAP4 |
| GGGCGCAGA | 18.67 | RARA, NRBF2, NCOA4, ETV5, HIF1A |
| GGCGCAGAG | 19.69 | RARA, NRBF2, NCOA4, ETV5, HIF1A |
| GCGCAGAGG | 17.17 | MEIS2 |
| CGCAGAGGG | 10.27 | GATA3, MDFI |
| GCAGAGGGA | 2.20 | NFKBIA, MYC, STAT1 |

Column1: the sequence of 9-mers. Column2: their corresponding statistical ranks. Column 3: promoter of genes that contain the 9-mer. Shown in bold are genes that function in chromatin remodeling. The sequence of overlapping run is GG**GCTGGGGAGGGCGGGGCGCAGAGGG**A. See Supplemental Table 4 for a listing of definitions for the gene names obtained from GenBank.

represent components of a genetic vocabulary that is context dependent.

*Statistical evaluation*

To evaluate the hypothesis that the predicted regulatory signals occur preferentially in DNase I HS segments, we tested a null hypothesis proposing that the predictions are distributed randomly in genomic DNA. The null-hypothesis can be rejected if the evaluations reveal an improbable number of predicted signals in the HS segments, with a significance level of 0.01. The *p*-values for all tests were less than $10^{-24}$.

In the evaluations, we defined the number of hits as the number of genomic positions at which signal intensity is greater than a specified value (threshold or cutoff). From a total number of observed hits ($h$), in a human genomic DNA of length $N$, we calculated the probability that $k$ hits from the population would randomly occur in HS segments without replacement. We determined the number of hits, in the HOXA locus, as a function of cutoff values chosen with respect to the baselines (zero) in the density plots. The goal is to determine whether we observe a trend in the statistical evaluations. In the density plots, signal intensities are in arbitrary units and are obtained from the displayed predictions. In plots of the HOXA locus, the intensities vary: between 0 and 7, for Reg_pred1; between 0 and 10, for CDS_Hits; and between 0 and 7, for Reg_pred2. We noted that the HS segments (obtained for various cell types [30]) mapped to overlapping positions. To eliminate redundancy, we removed the overlaps computationally to create the data set used for statistical evaluations.

At a relatively low cutoff intensity (0.5 above the baseline), we found that 56% of the hits obtained for Reg_Pred1 are localized in the mapped HS segments (Fig. 3). The percentage increases at cutoff values greater than 2 and reaches a plateau at about 75%. Statistical evaluations reject the null hypothesis when the number of hits in HS segments exceeds 20–25% (Fig. 3).

Similarly, at a relatively low cutoff (0.5), a significant fraction (about 53%) of the hits obtained for Reg_Pred2 occurred in the mapped HS segments (Fig. 3). This fraction steadily increased at higher cutoff values. Statistical evaluations reject the null hypothesis when the number of hits in HS segments exceeds 19–26% (Fig. 3).

If we choose 0.5 for cutoff intensity, we find that 26.3% of the CDS_Hits occur in the HS segments. The null hypothesis can be rejected if the number is greater than 19.3%. If we choose 2 for cutoff intensity, selecting for 9-mers that occur relatively infrequently in total genomic DNA, we find that nearly 67% of the hits occur in HS segments; 19.7% would reject the null hypothesis. The fraction of CDS_Hits in HS segments reaches nearly 75% at intensity cutoffs greater than 3.

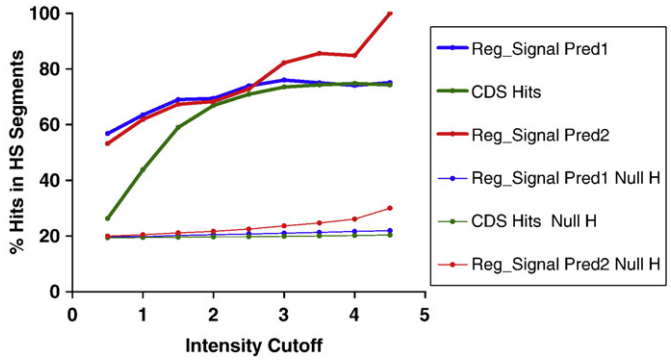


**Fig. 3.** Trends in statistical tests. The *X*-axis displays the intensity cutoffs selected with respect to the baselines, in the density plots. Hits (on the *Y* axis) represent the number of nucleotide positions that have intensity values above the cutoff. The upper three curves show the trend obtained for the %Hits that fall within the HS segments. The lower three plots show the computed limits (for %Hits) that would reject the null hypothesis, with a significance level of 0.01.

## Discussion

While protein coding sequences account for a relatively small fraction of human genomic DNA [3], as much as a third of the genome, a remarkable one billion base pairs, might correspond to regions that control chromosome replication, condensation, pairing, and segregation, and gene expression [8]. Furthermore, the prediction is that the human genome must contain vast amounts of *cis*-regulatory elements to direct the developmental, spatial, and temporal patterns of gene expression [36]. However, the context and the characteristics of sequences that regulate expression of human genes remain largely unknown.

We hypothesized that as the coding information, specific vocabulary might encode the regulatory information in genomic DNA. This idea is supported by studies showing that 9-mers and 8-mers from promoter regions of human genes are enriched in recognition sites for transcription factors that play central roles in various aspects of regulatory mechanisms [22,37,38]. Furthermore, in exploratory experiments, we noted that the occurrences of 9-mers collected from promoter regions created a regulatory module in a region far upstream of the transcription start site of the human NF-IL6 gene [23]. This module was conserved in human and chimp DNA but corresponded to a deletion in rat DNA [23].

Weighted density plots revealed that in relatively long genomic DNA segments, 9-mers from promoter regions occurred not only near TSSs but also upstream of transcription start sites, in intronic regions, and within intergenic regions (see for example Fig. 1, Fig. 2, and Supplemental Figs. 3–12). An earlier study imposed very stringent filtering criteria to detect signals superimposed on relatively noisy background [22]. Additional data evaluations showed that the background noise was, in part, due to CpG-rich 9-mers that occurred both in promoter and in coding regions of genes.

Therefore, we closely inspected the occurrences of 9-mers collected from CDSs in relatively long genomic DNA segments. Weighted density plots of CDS 9-mers (CDS_Hits) unraveled a complex and unexpected picture. Initially, as expected, we found that statistically significant 9-mers collected from CDSs occurred in exons. However, we found that these 9-mers also frequently appeared as clusters, in noncoding regions: near promoter regions; upstream of transcription start sites; in intronic regions; and within intergenic regions (see for example Fig. 1, and Supplemental Figs. 2–12). The results imply that DNA-sequence-based regulatory mechanisms are more complex than currently presumed.

In attempts to uncouple overlapping regulatory signals, we examined the occurrences of various 9-mers in genomic DNA in context of three types of density plots: Reg_Signal Pred1, CDS_Hits, and Reg_Signal Pred2. To construct Reg_Signal Pred1, we imposed the following constraints: the rank of 9-mers from promoter regions should be greater than 2; the corresponding rank of these 9-mers in CDSs should be less than 5; and the ratio of the rank of a promoter 9-mer with respect to its corresponding rank in CDS should be greater than 2. Using these criteria, in the density plots we obtained relatively well-defined signals that were distinguishable from the signals observed in CDS_Hits (for example compare the blue plot and the green plot in Fig. 2 and Supplemental Figs. 13–17). To take the filtering criteria a step further, we created another set of plots (named Reg_Signal Pred2). The signal intensity in these plots reflected whether the 9-mers were found more frequently in promoters or in CDSs (for example see the red plot in Fig. 2 and Supplemental Figs. 13–17). In Reg_Signal Pred2, signal intensities greater than one defined genomic regions that were enriched in 9-mers found more frequently in promoters than in coding regions of genes.

Importantly, the filtering criteria facilitated investigating whether the peaks in the regulatory signal predictions corresponded to TFBSs. In some cases, we observed clear correspondence to potential TFBSs (see for example Fig. 2, and Supplemental Figs. 13 and 14). We speculate that in the other cases, the peaks define TFBSs that currently might be unknown.

We inspected the sequences within several mapped regulatory segments to determine whether our predictions correlated with highly conserved regions in genomic DNA. There is ample evidence that supports the hypothesis that conserved noncoding regions may include transcription factor binding sites, see for example [20,23,39]. Furthermore, multi-species analyses have discovered numerous regulator regions and enhancers of gene expression (see for example [21]. However, while powerful, conservation-based predictions might miss regulatory information that could be species-specific. In fact, accelerated evolution of conserved non-coding sequences may produce traits that distinguish humans from other species [40,41].

We find that while the predicted regulatory signals occasionally appear in conserved sequences (see for example Supplemental Figs. 13 and 16), in most cases the signals primarily correlate with sequences conserved in primates. In predicted regulatory signals, we found one example of a sequence that appeared in chimp and rhesus but not in human DNA (Supplemental Fig. 14, the underlined G near the potential SP1 site). The finding appeared intriguing, in light of intensive efforts directed at identifying differences in the genomic sequences of primates to obtain clues about the molecular basis of speciation [41,42].

The idea of context dependence of regulatory signals can be extended to include overlapping sequences that might create composite and complex regulatory instructions. We imagine that long runs of sequences, produced from overlapping 9-mers, might correspond to combination of regulatory "vocabulary" with which complex instructions are written in genomic DNA. For examples, in intron of HOXA13, overlapping 9-mers produced a relatively broad peak in plots of regulatory signal predictions (Supplemental Fig. 21). Analyses of these 9-mers revealed that they were derived, in part, from promoter regions of genes with functions ranging from modulation of gene expression to chromatin remodeling and epigenetic gene silencing (Table 1 and Supplemental Table 4).

From the results of the analyses, we deduce that superimposed on signals defining TFBSs, the regulatory segments of human genes contain specific CpG-containing sequences that also occur in the coding regions of genes. The results indicate that the phenomenon is widespread and includes not only sequences in promoter regions but also regulatory regions that are distal to TSSs (see for example Fig. 2). The finding that specific CpG-rich sequences appear in both CDSs and non-coding regulatory regions suggests that there is an interconnection among the epigenetic signals that regulate the expression of human genes. Overall, we are tempted to conclude that our computational model has the potential of decoding regulatory instructions to discover sequences that interact with transcription factors and to discover epigenetic signals in both coding and regulatory regions of genes.

## Materials and methods

### *Sequences*

The sequence of HOXA locus was from GenBank (built hg_17). It was retrieved from the genome browser at UCSC. For testing the null hypothesis, we obtained the coordinates of HS segments from Xi et al. [30].

For human promoters (−500 to +50), we retrieved the sequences derived from alignments of full-length cDNAs with respect to a draft of human genomic DNA [43]. From the retrieved promoter sequences, we removed those that were incomplete: i.e. sequences that contained *N* and other IUPAC ambiguity codes, instead of specific nucleotides. From the set, we also removed sequences that appeared to be redundant. For coding sequences, we extracted the region annotated as CDS in RefSeq files from GenBank [28]. We obtained total genomic DNA from the genome browser at UCSC [22]. For data collection, we created a database in MySQL [44].

### *Determination of rankings*

We computed the ranking of the 9-mers in the datasets with respect to their frequencies in promoter sequences (Ei), coding regions (Di), and total human genomic DNA (Gi). Subsequently, the frequencies were normalized to obtain:

$$Ei/E; Di/D; Gi/G (\text{where}, E = \sum Ei; D = \sum Di; G = \sum Gi).$$

For promoter context, the ranking of the 9-mer of type *i* is:

$$R_{(pc)i} = GEi/EGi.$$

For CDS context, the ranking of the 9-mer of type *i* is:

$$R_{(d)i} = GDi/DGi.$$

The rankings provided probability thresholds ($\beta$ values) using the principle of large deviations, as previously described [22]. Briefly, in a typical case, probability achieves its largest value around the mean. Probability decays with a Gaussian tail within a square root distance from the mean, and finally decays exponentially further away from the mean. When estimating probability of rare events, one resorts to large deviations. Large deviations deal with events of exponentially small probability, far away from the mean [45].

### *Density plots*

The program scanned genomic DNA with a user defined window ($w$) to calculate weighted sums. At each nucleotide position, $\text{CDS\_Hits} = \sum_1^w (R_{(d)i})/w$.

To calculate Reg_Signal Pred1 and Reg_Signal Pred2, we imposed several constraints:

$$R_{(pc)i} > 2$$
$$R_{(d)i} < 5$$
$$\left[(R_{(pc)i})/R_{(d)i}\right] > 2$$

Based on the filtering criteria, at each nucleotide position:

$$\text{Reg\_Signal Pred1} = \sum_1^w (R_{(pc)i})/w$$
$$\text{Reg\_Signal Pred2} = \sum_1^w \left[(R_{(pc)i})/R_{(d)i}\right]/w.$$

### *Custom tracks and TFBSs*

We converted the computational predictions to "bed" files for display at the genome browser at UCSC. The home page of the browser provides detailed instructions for creating bed files and custom tracks. To create a listing of TFBSs, whenever possible, we reduced redundancy by grouping the transcription factors according to the structure of their DNA binding domains [16,46]. To eliminate redundancy, we also identified the various names given to transcription factors. To reduce the number of false positives, from the listing we removed sites that occurred frequently in genomic DNA. These included the site for GATA, CRX, and members of the C/EBP family.

### *Test of null-hypothesis*

In a genomic DNA with *N* nucleotide positions, we have a population of *h* observed hits and *N*–*h* non-hits. If we randomly fill

the HS segments of size $R$, with this population without replacement, the probability of exactly $k$ hits in the segments follow the well-known hypergeometric distribution with the discrete probability function:

$$Pr(X=k)=\frac{\binom{h}{k}\binom{N-h}{R-k}}{\binom{N}{R}}.$$

The sample space for the probabilistic model ranges from max $[0,R-(N-h)]$ (one must use some hits if there are not enough non-hits in the population to fill the region) to min$[R,h]$ (one cannot have more hits than the size of the region nor can have more hits than the size of the population). To determine $R$, we eliminated from HS segments [30] those that were redundant. The output produced a single non-redundant union of the mapped HS segments.

To test the null hypothesis, we computed a rejection limit $L$ (the minimum number of hits in HS segments) that would disprove the hypothesis at a selected significance, $S$. That is, $L$ is the minimum over $K$ such that

$$\sum_{k=0}^{K}\Pr(X=k)\geq 1-S.$$

We estimated the $p$-values by computing an upper bound of the sum over the upper tail of the hypergeometric distribution, as previously described [47]:

$$\Pr(X\geq \text{observed hits}|h,N,R)=\sum_{k=\text{observed hits}}^{R}\Pr(X=k|h,N,R)$$

$$\leq\left(\left(\frac{p}{p+t}\right)^{p+t}\left(\frac{1-p}{1-p-t}\right)^{1-p-t}\right)^{R}$$

where $p=\frac{h}{N}$ and $t=\frac{k}{R}-p$, with the condition that $t\geq 0$.

## Acknowledgment

We wish to thank reviewer #2 for insightful suggestions and recommendations.

The bed files for the HOXA locus and TFBSs can be downloaded from the following website: http://www.chem.purdue.edu/bina/Data.htm.

From another site (http://bina-grid.chem.purdue.edu/scripts/form_9mer_ID_Seq_Genes.php), you can obtain information on specific 9-mers; for details, see [34].

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2008.11.008.

## References

[1] D.B. Searls, The language of genes, Nature 420 (2002) 211–217.
[2] F.S. Collins, E.D. Green, A.E. Guttmacher, M.S. Guyer, A vision for the future of genomics research, Nature 422 (2003) 835–847.
[3] International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome, Nature 431 (2004) 931–945.
[4] G. Gamow, M. Ycas, Statistical correlation of protein and ribonucleic acid composition, Proc. Natl. Acad. Sci. U. S. A. 41 (1955) 1011–1019.
[5] F.H. Crick, J.S. Griffith, L.E. Orgel, Codes without commas, Proc. Natl. Acad. Sci. U. S. A. 43 (1957) 416–421.
[6] S. Itzkovitz, U. Alon, The genetic code is nearly optimal for allowing additional information within protein-coding sequences, Genome Res. 17 (2007) 405–412.
[7] T. Bollenbach, K. Vetsigian, R. Kishony, Evolution and multilevel optimization of the genetic code, Genome Res. 17 (2007) 401–404.
[8] M. Levine, R. Tjian, Transcription regulation and animal diversity, Nature 424 (2003) 147–151.
[9] A. Barbic, D.P. Zimmer, D.M. Crothers, Structural origins of adenine-tract bending, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 2369–2373.
[10] C.B. Schaefer, S.K. Ooi, T.H. Bestor, D. Bourc'his, Epigenetic decisions in mammalian germ cells, Science 316 (2007) 398–399.
[11] M.M. Suzuki, A. Bird, DNA methylation landscapes: provocative insights from epigenomics, Nat. Rev. Genet. 9 (2008) 465–476.
[12] C. Wu, The 5′ ends of *Drosophila* heat shock genes in chromatin are hypersensitive to DNase I, Nature 286 (1980) 854–860.
[13] S.C. Elgin, DNAase I-hypersensitive sites of chromatin, Cell 27 (1981) 413–415.
[14] D.S. Gross, W.T. Garrard, Nuclease hypersensitive sites in chromatin, Annu. Rev. Biochem. 57 (1988) 159–197.
[15] ENCODE Project Consortium, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, Nature 447 (2007) 799–816.
[16] E.M. Crowley, K. Roeder, M. Bina, A statistical model for locating regulatory regions in genomic DNA, J. Mol. Biol. 268 (1997) 8–14.
[17] E.M. Crowley, A Bayesian method for finding regulatory segments in DNA, Biopolymers 58 (2001) 165–174.
[18] D. Papatsenko, M. Levine, Computational identification of regulatory DNAs underlying animal development, Nat. Methods 2 (2005) 529–534.
[19] L.A. Pennacchio, E.M. Rubin, Genomic strategies to identify mammalian regulatory sequences, Nat. Rev. Genet. 2 (2001) 100–109.
[20] D.C. King, J. Taylor, L. Elnitski, F. Chiaromonte, W. Miller, R.C. Hardison, Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences, Genome Res. 15 (2005) 1051–1060.
[21] L.A. Pennacchio, N. Ahituv, A.M. Moses, S. Prabhakar, M.A. Nobrega, M. Shoukry, S. Minovitsky, I. Dubchak, A. Holt, K.D. Lewis, I. Plajzer-Frick, J. Akiyama, S. De Val, V. Afzal, B.L. Black, O. Couronne, M.B. Eisen, A. Visel, E.M. Rubin, In vivo enhancer analysis of human conserved non-coding sequences, Nature 444 (2006) 499–502.
[22] M. Bina, P. Wyss, W. Ren, W. Szpankowski, E. Thomas, R. Randhawa, S. Reddy, P.M. John, E.I. Pares-Matos, A. Stein, H. Xu, S.A. Lazarus, Exploring the characteristics of sequence elements in proximal promoters of human genes, Genomics 84 (2004) 929–940.
[23] E.I. Pares-Matos, J.S. Milligan, M. Bina, Exploring transcription factor binding properties of several non-coding DNA sequence elements in the human NF-IL6 gene, J. Mol. Biol. 357 (2006) 732–747.
[24] G.B. Hutchinson, The prediction of vertebrate promoter regions using differential hexamer frequency analysis, Comput. Appl. Biosci. 12 (1996) 391–398.
[25] T. Wirth, D. Baltimore, Nuclear factor NF-kappa B can interact functionally with its cognate binding site to provide lymphoid-specific promoter function, EMBO J. 7 (1988) 3109–3113.
[26] P. Rorth, C. Nerlov, F. Blasi, M. Johnsen, Transcription factor PEA3 participates in the induction of urokinase plasminogen activator transcription in murine keratinocytes stimulated with epidermal growth factor or phorbol-ester, Nucleic Acids Res. 18 (1990) 5009–5017.
[27] S. Richard, H.H. Zingg, Identification of a retinoic acid response element in the human oxytocin promoter, J. Biol. Chem. 266 (1991) 21428–21433.
[28] K.D. Pruitt, T. Tatusova, D.R. Maglott, NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic Acids Res. 33 (2005) D501–D504.
[29] L. Elnitski, V.X. Jin, P.J. Farnham, S.J. Jones, Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques, Genome Res 16 (2006) 1455–1464.
[30] H. Xi, H.P. Shulha, J.M. Lin, T.R. Vales, Y. Fu, D.M. Bodine, R.D. McKay, J.G. Chenoweth, P.J. Tesar, T.S. Furey, B. Ren, Z. Weng, G.E. Crawford, Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome, PLoS Genet. 3 (2007) e136.
[31] P.J. Sabo, M.S. Kuehn, R. Thurman, B.E. Johnson, E.M. Johnson, H. Cao, M. Yu, E. Rosenzweig, J. Goldy, A. Haydock, M. Weaver, A. Shafer, K. Lee, F. Neri, R. Humbert, M.A. Singer, T.A. Richmond, M.O. Dorschner, M. McArthur, M. Hawrylycz, R.D. Green, P.A. Navas, W.S. Noble, J.A. Stamatoyannopoulos, Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays, Nat. Methods 3 (2006) 511–518.
[32] Z. Solan, D. Horn, E. Ruppin, S. Edelman, Unsupervised learning of natural languages, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 11629–11634.
[33] L. Segal, M. Lapidot, Z. Solan, E. Ruppin, Y. Pilpel, D. Horn, Nucleotide variation of regulatory motifs may lead to distinct expression patterns, Bioinformatics 23 (2007) i440–i449.
[34] M. Bina, P. Wyss, S.R. Shah, A database of 9-mers from promoter regions of human protein-coding genes, Methods Mol. Biol. 338 (2006) 129–134.
[35] S. Sinha, A.S. Adler, Y. Field, H.Y. Chang, E. Segal, Systematic functional characterization of cis-regulatory motifs in human core promoters, Genome Res. 18 (2008) 477–488.
[36] B. Lemon, R. Tjian, Orchestrated response: a symphony of transcription factors for gene control, Genes Dev. 14 (2000) 2551–2569.
[37] P.C. FitzGerald, A. Shlyakhtenko, A.A. Mir, C. Vinson, Clustering of DNA sequences in human promoters, Genome Res. 14 (2004) 1562–1574.
[38] L. Marino-Ramirez, J.L. Spouge, G.C. Kanga, D. Landsman, Statistical analysis of over-represented words in human promoter sequences, Nucleic Acids Res. 32 (2004) 949–958.
[39] X. Xie, T.S. Mikkelsen, A. Gnirke, K. Lindblad-Toh, M. Kellis, E.S. Lander, Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 7145–7150.
[40] S. Prabhakar, J.P. Noonan, S. Paabo, E.M. Rubin, Accelerated evolution of conserved noncoding sequences in humans, Science 314 (2006) 786.

[41] H. Liang, Y.S. Lin, W.H. Li, Fast evolution of core promoters in primate genomes, Mol. Biol. Evol. 25 (2008) 1239–1244.
[42] Rhesus Macaque Genome Sequencing and Analysis Consortium, Evolutionary and biomedical insights from the rhesus macaque genome, Science 316 (2007) 222–234.
[43] N.D. Trinklein, S.J. Aldred, A.J. Saldanha, R.M. Myers, Identification and functional analysis of human transcriptional promoters, Genome Res. 13 (2003) 308–312.
[44] P. Wyss, S.A. Lazarus, M. Bina, A program toolkit for the analysis of regulatory regions of genes, Methods Mol. Biol. 338 (2006) 135–152.
[45] W. Szpankowski, Average Case Analysis of Algorithms on Sequences, Wiley, New Yor, 2001.
[46] A. Sandelin, W.W. Wasserman, Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics, J. Mol. Biol. 338 (2004) 207–215.
[47] V. Chvátal, The tail of the hypergeometric distribution, Discrete Math. 25 (1979) 285–287.
[48] D. Karolchik, R.M. Kuhn, R. Baertsch, G.P. Barber, H. Clawson, M. Diekhans, B. Giardine, R.A. Harte, A.S. Hinrichs, F. Hsu, K.M. Kober, W. Miller, J.S. Pedersen, A. Pohl, B.J. Raney, B. Rhead, K.R. Rosenbloom, K.E. Smith, M. Stanke, A. Thakkapallayil, H. Trumbower, T. Wang, A.S. Zweig, D. Haussler, W.J. Kent, The UCSC Genome Browser Database: 2008 update, Nucleic Acids Res. 36 (2008) D773–D779.
[49] F. Hsu, W.J. Kent, H. Clawson, R.M. Kuhn, M. Diekhans, D. Haussler, The UCSC known genes, Bioinformatics 22 (2006) 1036–1046.
[50] A. Siepel, G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.W. Hillier, S. Richards, G.M. Weinstock, R.K. Wilson, R.A. Gibbs, W.J. Kent, W. Miller, D. Haussler, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes, Genome Res. 15 (2005) 1034–1050.