

# Regulatory motif discovery in human promoters: motif evaluation, sequence space exploration, and functional validation.

Xiaohui Xie<sup>1</sup>, Eric S. Lander<sup>1,2</sup>, Manolis Kellis<sup>1,3</sup>

xhx@broad.mit.edu, lander@broad.mit.edu, manoli@mit.edu

(1) Broad Institute of MIT and Harvard, Cambridge MA 02139

(2) Whitehead Institute for Biomedical Research, Cambridge MA 02139

(3) MIT Computer Science and Artificial Intelligence Laboratory, Cambridge MA 02139

## 1. Introduction

Comprehensive identification of all functional elements encoded in the human genome is a fundamental need in biomedical research. Comparative genomics provides a powerful approach for the systematic discovery of functional elements, by virtue of their evolutionary conservation across related species<sup>1-9</sup>. Here, we present the computational methods underlying a comparative analysis of the human, mouse, rat, and dog genomes to systematically discover a dictionary of conserved regulatory motifs<sup>10</sup>. We present statistical methods for evaluating motif conservation, algorithms for rapid exploration of motif sequence space, and computational methods for the validation of discovered motifs. The methods presented led to the discovery of 174 candidate promoter motifs, including most previously known transcription factor binding sites and 105 novel motifs. We present several tests for the validation of regulatory motifs based on expression information and positional bias from the transcription start. The methods presented are general, applicable in a range of species, and scalable as additional genomes become available.

## 2. Evaluating regulatory motif conservation

Recent studies have revealed that the majority of functional elements are non-protein-coding, and are likely to include regulatory signals, RNA genes, and structural elements. However, only the largest and most conserved elements are readily identified, and the vast majority of non-coding functional elements remains unknown. It has been particularly difficult to recognize short regulatory sequences, such as the DNA binding sites for transcription factors. However, it should be possible to define the motifs themselves with many fewer species based on their multiple conserved occurrences in the genome. Thus, we evaluate the genome-wide conservation of regulatory motifs, based on their frequent conservation across multiple instances.

We chose a motif representation that allows enumeration and avoids over-fitting. We represent regulatory motifs as consensus sequences (profiles), over an alphabet of 11 characters, consisting of the four nucleotides A,C,G,T, the six two-fold degenerate characters S=[CG], W=[AT], Y=[CT], R=[AG], M=[AC], K=[GT], and the four-fold degenerate character N=[ACGT]. An occurrence of a motif  $m$  is a sequence (over the alphabet ACGT) which matches the consensus of motif  $m$  at every position, namely contains one of the nucleotides allowed by the degenerate code at that position. A conserved occurrence of a motif  $m$  is an instance of the motif in the human genome, for which an exact match to the motif is present in each of the four species. For fully specified motifs, this implies that the sequences are identical across the four species. For motifs with degenerate positions (containing ambiguity codes), all sequences need to match the motif, but they do not need to be identical to each other; the four species can contain different variants of the degenerate positions. We define the conservation rate of a motif  $m$  as the number of human occurrences of  $m$  which are conserved across all four species, divided by the total number of human occurrences of  $m$ .

We evaluate the Motif Conservation Score of a motif  $m$  of given length and degeneracy, by comparing its conservation rate  $p$  to the expected rate  $p_0$ , estimated using similar random motifs of the same length and degeneracy (see below). Given the rate  $p_0$ , we evaluate the binomial probability of observing  $K$  conserved instances out of total  $N$  instances in the human sequence for motif  $m$ . We report the MCS as a Z-score defined as  $MCS = (K - Np_0) / [Np_0(1-p_0)]^{1/2}$ , which measures the number of standard deviations of conserved instances away from what is expected by chance when the null model is assumed to be binomial. Motifs with high motif conservation scores, are both highly conserved and frequently occurring, resulting in both an increased rate, and sufficient statistical significance given the large counts. To estimate the conservation rate  $p_0$  expected for a motif  $m$  of given length and redundancy, we observe the average conservation rate of 1000 random motifs of the same length and redundancy. To account for nucleotide compositional biases in the human genome, we generate these motifs by sampling the human genome. Namely, we select 1000 loci in the four-way species alignment, and extract the human sequences for each of these loci. Based on the degeneracy levels of  $m$ , we generate a motif for each of these sequences, selecting a degeneracy code for each position matching the sequence of the human locus, and the degeneracy level of  $m$  at that position. For example, if the first character of  $m$  is two-fold degenerate and the first nucleotide at the selected locus is A, we pick a two-fold degenerate base containing A (W, R or M), and so on for every character of  $m$ . We then evaluated, for every locus, whether the resulting random motif is conserved in the other three species, and summed across the 1000 loci. This total number of conserved motifs, divided by the 1000 randomly constructed motifs, was used to estimate the expected conservation rate  $p_0$ , under a random model.

## 3. Sequence space exploration for motif discovery

With the ability to enumerate motifs, we took an exhaustive search approach to motif discovery, and developed a method for rapid enumeration and testing of short sequence patterns. We enumerated all motifs of length between 6 and 26, over an alphabet of 11 characters (the four bases A, C, G, T, the six two-fold degenerate IUB codes R=[AG], Y=[CT], K=[GT], M=[AC], S=[GC], W=[AT], and the four-fold degenerate character N=[ATCG]). The number of motifs that can be formed by combining the 11 letters with various lengths is enormous, but it was still possible to screen most of them because only a small subset of them actually occurred in the database. We started by hashing the positions of all 6-mer motifs, possibly with gaps, and then searched and computed the MCS score for all possible extensions of these 6-mers. The method consisted of the following steps:

- (a) We first search and index all positions in the human genome containing a fully-specified 6-mer seed, possibly with a central gap between 0 and 10 non-specified bases. These seeds are of the form UVW-gap-XYZ, where U,V,W,X,Y,Z can be any nucleotide. This resulted in a total number 45,056 six-mers.

- (b) For each of these seeds, we extracted the four-way aligned sequence containing the aligned seeds and their neighboring sequences extending 5 nucleotides on each end.
- (c) We then enumerated all motifs that contain one of these seeds and have more than one instance in the aligned genomes.
- (d) We finally tested the conservation statistics of each of the resulting motifs and selected all motifs with MCS above 6.0

#### 4. Motif gene set enrichment analysis for expression data

We evaluated the tissue-specificity of each regulatory motifs by calculating the tissue-specificity of its target gene set, in a gene expression atlas of 75 human tissues. We first preprocessed the expression data by normalizing the expression of each gene across all tissues to be mean zero and variance 1. We then ranked the genes based on their normalized expression values for each tissue, giving rise to 75 ranked gene lists.

For each motif  $m$ , we generated three gene sets: a target gene set  $S_1$ , and two control gene sets  $S_2$  and  $S_3$ , with the same number of genes.

- (1) We first generated the motif gene set  $S_1$  of 'conserved instances', consisting of the inferred target genes for each motif. This set consisted of all genes whose promoters contained at least one conserved instance of the motif  $m$ .
- (2) We then generated a control gene set  $S_2$  of 'non-conserved instances', by randomly sampling from genes containing non-conserved instances of the motif, until  $S_2$  contained the same number of genes as  $S_1$ .
- (3) We also generated a second control gene set  $S_3$  of 'shuffled conserved instances', by randomly sampling genes from the union of all conserved gene sets ( $S_1$ ), for all motifs.

We used the two control gene sets to evaluate the statistical significance of the tissue enrichment observed in the target gene set  $S_1$ , as compared to two similar but random gene sets with the same cardinality  $S_2$  and  $S_3$ .

We evaluated the enrichment of a motif  $m$  in a given tissue, as the enrichment of its gene set  $S$  in the ranked list for that tissue. We used the Mann-Whitney rank sum statistic to evaluate the non-randomness of the ranks of  $S$ , in the list  $L$  specific to that tissue. We sum the ranks of genes in  $S$  that appear in list  $L$ . The significance of the rank sum is tested against rank sums of random subsets of the list  $L$ , randomly permuted. Let  $\mu$  and  $\sigma^2$  be the mean and variance of the control rank sums. We define the Motif Gene Set Enrichment (MGSE) score to be  $(\mu-S)/\sigma$ , that is, the number of standard deviations smaller than the mean. This statistic is strongest when the items in  $S$  are ranked at the top of the list  $L$ .

For each motif, we computed the MGES for  $S_1$ ,  $S_2$ , and  $S_3$  in all 75 tissue-specific ranked gene lists. For the motif target list  $S_1$ , the best MGES among all tissues is annotated in Table 2 (if the score was above 4.0 SD). We also computed the best MGES scores for the two control sets  $S_2$  and  $S_3$ , and we found that their scores were indeed much less than the target gene sets  $S_1$  (Fig. S2). Only a few non-conserved control sets  $S_2$  in the beginning of the motif list show enrichment score significantly higher than those from randomly permuted sets. The motifs corresponding to those sets have consistently high conservation rates. It is likely that the consensus sequences of these motifs are specific enough to indicate functionality, regardless of conservation.

#### 5. Motif positional bias in promoters

For every motif  $m$ , we tested the presence of a positional bias in the distance distribution between its instances and the TSS. We identified all sites where a motif occurred in human promoters (without requiring conservation) and recorded their positions relative to TSS. We then divided the region (-2000, 2000) bp around TSS into 100 bins, and counted the number of sites located in each of the bins. We computed the mean and variance on the distribution of the number of sites in different bins, and converted the number of sites in each bin to a Z-score measuring the number of standard deviations away from the mean. Positional clustering of the motif was counted as significant if there existed a bin with Z-score above 5.0, in which case the biased position was determined by the location of the bin.

#### 6. Conclusion

Our comparative genome analysis of four mammalian species has provided an initial systematic catalog of human regulatory motifs in promoters. The approach automatically rediscovered many known motifs and discovered many novel ones that appear to be functional based on multiple criteria; many of these show tissue specific expression, distance constraints with respect to the TSS. The next challenge will be to develop systematic methods to discern the specific functions of these motifs in a genome-wide fashion.

The results here are, of course, only a step toward a comprehensive inventory of human regulatory motifs to serve as a foundation for understanding cellular circuitry and its role in health and disease. Our analysis here employed stringent thresholds to focus on the most abundant and most conserved motifs. With sequence from a few dozen additional mammals<sup>11</sup>, it should be possible to create a complete dictionary of such common functional elements.

1. Gumucio, D. L. et al. Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. *Mol Cell Biol* **12**, 4919-29 (1992).
2. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. Human-mouse genome comparisons to locate regulatory sites. *Nat Genet* **26**, 225-8 (2000).
3. Dubchak, I. et al. Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res* **10**, 1304-6 (2000).
4. Pennacchio, L. A. et al. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* **294**, 169-73 (2001).
5. Boffelli, D. et al. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391-4 (2003).
6. Sandelin, A., Wasserman, W. W. & Lenhard, B. ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res* **32**, W249-52 (2004).
7. Sinha, S., Blanchette, M. & Tompa, M. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* **5**, 170 (2004).
8. Dermitzakis, E. T. et al. Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res* **14**, 852-9 (2004).
9. Bejerano, G. et al. Ultraconserved elements in the human genome. *Science* **304**, 1321-5 (2004).
10. Xie, X. et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* (2005).
11. Margulies, E. H. et al. in *Proc Natl Acad Sci U S A (in press)* (2005).