BMC Bioinformatics

Research article



# Predicting combinatorial binding of transcription factors to regulatory elements in the human genome by association rule mining

Xochitl C Morgan†[1], Shulin Ni[2], Daniel P Miranker[2] and Vishwanath R Iyer*[1]

Address: [1]Institute for Cellular and Molecular Biology and Center for Systems and Synthetic Biology, The University of Texas at Austin, Austin, Texas 78712-0159, USA and [2]Department of Computer Science, The University of Texas at Austin, Austin, Texas 78712-0159, USA

Email: Xochitl C Morgan - morganx@mail.utexas.edu; Shulin Ni - shulinn@gmail.com; Daniel P Miranker - miranker@cs.utexas.edu; Vishwanath R Iyer* - vishy@mail.utexas.edu

* Corresponding author †Equal contributors

Published: 15 November 2007

*BMC Bioinformatics* 2007, **8**:445 doi:10.1186/1471-2105-8-445

Received: 9 May 2007
Accepted: 15 November 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/445



## Abstract

**Background:** Cis-acting transcriptional regulatory elements in mammalian genomes typically contain specific combinations of binding sites for various transcription factors. Although some cis-regulatory elements have been well studied, the combinations of transcription factors that regulate normal expression levels for the vast majority of the 20,000 genes in the human genome are unknown. We hypothesized that it should be possible to discover transcription factor combinations that regulate gene expression in concert by identifying over-represented combinations of sequence motifs that occur together in the genome. In order to detect combinations of transcription factor binding motifs, we developed a data mining approach based on the use of association rules, which are typically used in market basket analysis. We scored each segment of the genome for the presence or absence of each of 83 transcription factor binding motifs, then used association rule mining algorithms to mine this dataset, thus identifying frequently occurring pairs of distinct motifs within a segment.

**Results:** Support for most pairs of transcription factor binding motifs was highly correlated across different chromosomes although pair significance varied. Known true positive motif pairs showed higher association rule support, confidence, and significance than background. Our subsets of high-confidence, high-significance mined pairs of transcription factors showed enrichment for co-citation in PubMed abstracts relative to all pairs, and the predicted associations were often readily verifiable in the literature.

**Conclusion:** Functional elements in the genome where transcription factors bind to regulate expression in a combinatorial manner are more likely to be predicted by identifying statistically and biologically significant combinations of transcription factor binding motifs than by simply scanning the genome for the occurrence of binding sites for a single transcription factor.

## Background

Substantial differences of phenotype can be primarily the result of differences in gene expression levels rather than in protein structure. Genes are dynamically regulated, pri-

marily at the transcriptional level, by protein transcription factors that bind DNA at cis-regulatory regions to activate or repress expression. Mammalian cis-regulatory regions range in length from the 60 bp human muSK enhancer [1] to the 450 bp human TGFβ enhancer [2] to the 1100 bp enhancer of murine Pax6 [3], but they are generally a few hundred base pairs in length. Enhancers contain binding sites for transcription factors, sometimes for a single factor and sometimes for many [4]. A detailed understanding of the transcriptional regulatory programs of any organism requires knowledge of the binding sites of transcription factors, the circumstances and cellular conditions under which these transcription factors bind to their targets, and the genes that are regulated by combinations of transcription factors.

Cis-regulatory regions for most of the approximately 20,000 protein-coding genes encoded in the human genome have not yet been characterized [5]. Transcription factor binding sites, and thus cis-regulatory regions, can be identified using high-throughput methods such as ChIP-chip [6-9], but there are more than 2000 transcription factors encoded in the human genome [10,11]. This diversity of transcription factors, coupled with the fact that many are likely to be expressed and to combinatorially regulate target genes in a developmental, cell-, or tissue-specific manner, makes experimental identification of cis-regulatory regions challenging even with genome-wide ChIP-chip. Computational identification of cis-regulatory motifs based on signatures of their presence in the genomic sequence is an attractive alternative.

A major class of computational methods for identifying regulatory elements relies on the occurrence of TF binding sites in close proximity within regulatory elements. For example, the stripe 2 enhancer of the *even-skipped (eve)* gene in *Drosophila melanogaster* has twenty binding sites for four TFs within an area of roughly 600 bp [12]. The *knirps* gene of *Drosophila* is regulated by two enhancers containing six binding sites each for the transcription factors *bicoid* and *caudal* as well as two *hunchback* sites [13]. The HS2 enhancer of the human β-globin locus contains four NF-E1 binding sites and 2 CACC boxes within 250 bp [14], while a 300 bp region near the interleukin 2 transcriptional start site contains multiple binding sites for Ap-1 and Oct1 as well as sites for NFκB and NFAT [15]. Thus, the density of TF binding sites may be used as a means to locate cis-regulatory regions computationally [16].

Computational location of cis-regulatory modules by clustering of transcription factor binding sites has been implemented in genomes ranging from yeast [17] to human [18]. Previous approaches include "sliding window" [19-21] to Hidden Markov models [16,18] to position weight matrix clustering [22-26], while clusters have been defined both homotypically [19,21] and heterotypically [20,27-29]. These computational methods have been used to locate many cis-regulatory regions and novel target genes, notably in *Drosophila*. One limitation of these heterotypic clustering methods is the need to know which combinations of transcription factors should define the heterotypic clusters.

Numerous transcription factors are known to cooperate in certain contexts; for example, it is known that many genes involved in inflammation are regulated by Ap-1 and NFκB [30]. Similarly, interactions between PU.1 and GATA family TFs mediate cell differentiation in B-cell development [31]. Prediction of transcription factor cooperativity has been carried out in yeast [32,33] and human [34], but elucidation of the entire network of transcription factors that cooperate with one another in cis-regulatory regions is far from complete. In order to better define biologically relevant, heterogeneous combinations of transcription factors, we have developed an association rule data mining approach to search genome sequence information and identify over-represented adjacent motifs for transcription factor binding. Predicting transcription factor cooperation by data mining using association rules has previously been attempted in yeast as well as *C. elegans* and human chromosome 22 [35-37] but these attempts have been limited to mining known promoters [35] or repetitive elements such as microsatellites [36,37] rather than applied to the entire human genome.

Association rule data mining [38] was originally used in market basket analysis to determine which items are frequently purchased together. Basket analysis uses a database of transactions in which each tuple is a list of items purchased in one customer's transaction. Mining seeks to discover rules such as "spaghetti ⇒ parmesan cheese," meaning "People who buy spaghetti also often buy parmesan cheese." Association rules can be formally described as follows: [38]

• $I = \{i_1, i_2 ... i_n\}$ is a set of literals called items.

• $D$ is a set of transactions. Each transaction $T$ is a set of items such that $T \subseteq I$.

• A transaction $T$ contains $X$, a set of items in $I$, if $X \subseteq T$.

• An association rule is an implication of $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \varnothing$.

• $C$ is the confidence of a rule $X \Rightarrow Y$ in transaction set $D$ if $c$% of transactions in $D$ that contain $X$ also contain $Y$. It is also known as the conditional probability of Y given X, or P(Y|X).

• $S$ is the support of rule $X \Rightarrow Y$ in set $D$ if $s$% of transactions in $D$ contain both X and Y. It is also known as the joint probability of both X and Y, or $P(X \cap Y)$.

If a rule $X \Rightarrow Y$ has high confidence, it is likely that transactions containing $X$ will likely also contain $Y$. However, the existence of such a rule does not by itself imply any causal relationship between $X$ and $Y$.

Determining over-represented transcription factor partners may help to reveal biological roles for less well-studied transcription factors. Therefore, in our studies, we used data mining to determine whether two transcription factors whose experimentally determined binding motifs were frequently proximal to one another were also likely to have biologically meaningful interactions. For example, the rule "Nuclear Factor Kappa B $\Rightarrow$ Ap-1" would indicate "Where there is a motif for NFκB, there is often also an Ap-1 motif." To allow application of association rules to transcription factor motifs in the human genome, we divided the genome into segments and scored each segment for the presence or absence of each of 83 transcription factor binding motifs (Figure 1). Thus, the set of 83 motifs becomes $I$, each individual transcription factor binding motif becomes an item, and each small segment of genome becomes a transaction $T$ whose contents $X$ are the motifs located within.

## Results

Our major aim was to determine whether a pair transcription factors whose motifs were frequently near one another were more likely to have a biological association than a pair of transcription factors whose motifs were not. In order to test this hypothesis, we located all possible binding sites in the human genome for the position weight matrices (PWMs) of each of 83 transcription factors (Additional file 1). We then divided the genome into 100 bp regions and used association rule data mining to calculate support and confidence for each transcription factor pair in the human genome.

Straightforward association rule mining that simultaneously considers all motif positions discovers high numbers of transcription factor pairs that bind identical or highly similar motifs. For example, two different transcription factors A and B may both bind to the motif "CACGTG", so the confidence $C$ of the rule A $\Rightarrow$ B will be 100%. Similarly, if A binds to "CACGTG" and B binds to "CACGTGA," this high overlap between binding motifs will result in the confidence being very high while the rule is neither interesting nor surprising, although it may still be biologically valid. To avoid discovery of enriched overlapping motifs, for each transcription pair AB, all overlapping binding sites between A and B were removed before calculating support and confidence (Figure 2). We also calculated a $P$-value based on the hypergeometric proba-

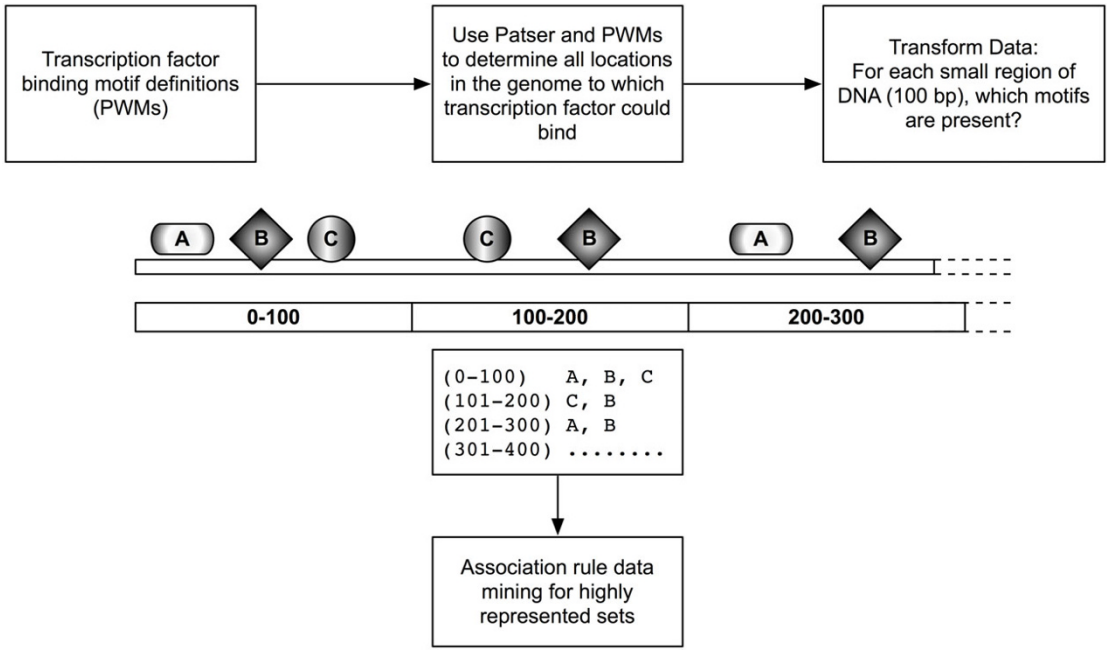


**Figure 1**
**Overview**. Patser is used to map all possible binding sites in the genome for each of 83 position weight matrices (PWMs) from TRANSFAC. The genome is then scored 100 bp at a time for the presence or absence of each PWM, and association rules are used to mine the genome for frequently co-occurring pairs.

bility of observing the association between A and B by chance. We ensured that associations between transcription factor motifs were not an artifact caused by the presence of repetitive DNA, by considering repeat masked regions separately (Methods and Additional file 2). Furthermore, in order to rule out the possibility that associations were generated by nucleotide bias, we ascertained that dinucleotide and trinucleotide frequencies of segments containing motif pairs were not significantly different from segments containing one member of the pair or background (data not shown).

In order to determine whether biologically significant associations between PWMs arise in promoter regions, we applied the same pairwise mining algorithm to the subset of the genome that was 1 kb upstream of the transcriptional start site of all human RefSeq genes [39]. Because transcription factor function is often phylogenetically conserved, we also examined whether the combinations we identified by mining the human genome were identifiable in the mouse genome; we performed identical pairwise mining for significant associations among the same 83 transcription factors on mouse chromosome 1.

### *Identifying meaningful TF pairs*

Due to the size of the human genome and the tendency of PWMs to match at a large number of genomic locations, all TF pairs showed some co-occurrence. This support for possible transcription factor PWM pairs ranged from 9 x $10^{-6}$ to 0.2. Support for the association of a given pair of transcription factors was highly conserved, not only between promoters and the entire genome (Figure 3A),

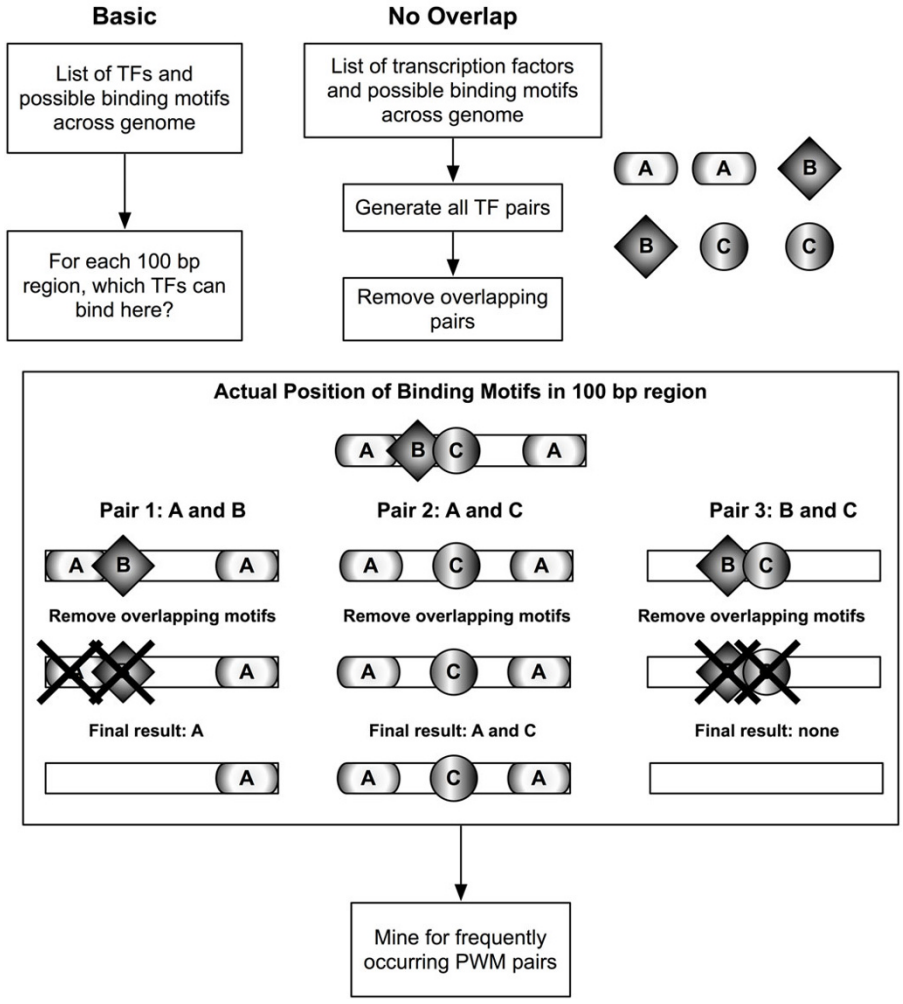


**Figure 2**

**Mining without overlap**. In order to avoid enriching primarily for TF pairs that bind similar motifs, the genome is mined once for each pair AB. All overlapping motifs between A and B are removed before calculating support, confidence, and *P*-values, then restored upon subsequent iterations.

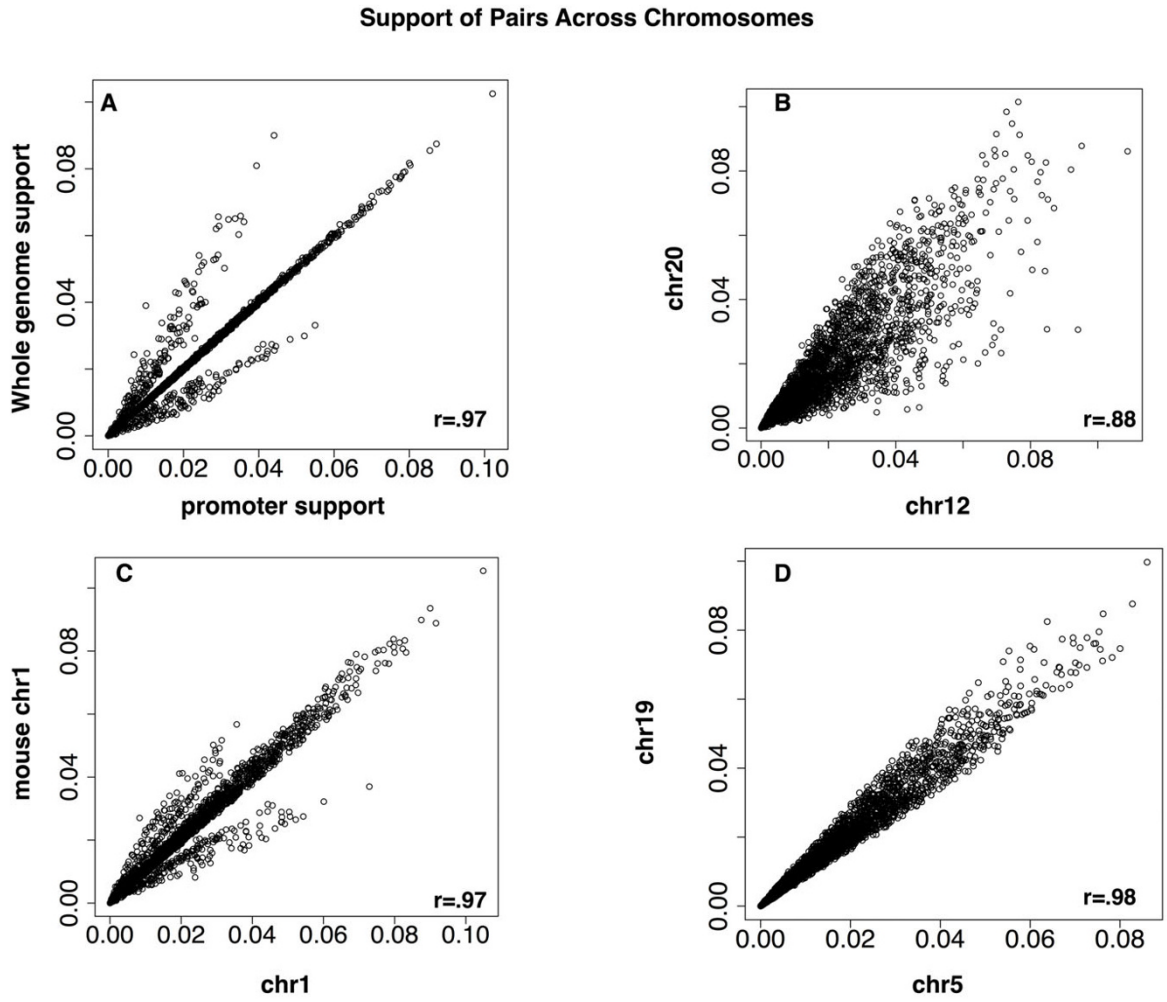


**Figure 3**
**Support of TF pairs across chromosomes**. The support of a given TF pair is highly correlated between chromosomes (B, D). This is also true for support in promoter regions versus the entire human genome (A) as well as support between human and mouse chromosomes (C).

but also between the human chromosomes and mouse chromosome 1 (Figure 3C) and between individual human chromosomes (Figure 3B, Figure 3D), suggesting that the associations revealed by mining are biologically relevant.

Association rules with the highest support and confidence are typically regarded as being interesting; however, if two different transcription factors each have large numbers of independent binding motifs in the genome, they could appear to be associated with high support values merely by chance. To minimize this possibility and to select those TF pairs occurring more frequently than by random chance and thus likely to be biologically meaningful, we also calculated the statistical significance (*P*-value) of observing each TF pair using the hypergeometric probability distribution. We defined the dataset "all" as the complete set of 3403 PWM pairs, and we selected three subsets with high confidence and significance for further analysis: "genomewide," "mouse," and "promoter" (Additional file 3).

The subsets "genomewide", "promoter", and "mouse" were defined as $P < 0.05$, greater than median difference between confidence A $\Rightarrow$ B and confidence B $\Rightarrow$ A. For the subset "genomewide" this was measured on the entire human genome and resulted in 66 TF pairs. For the subset "mouse," this was measured on mouse chromosome 1 and resulted in 184 pairs. For the subset "promoter", this was measured only across regions 1 kb upstream of the transcriptional start site of each RefSeq gene and resulted in 28 pairs.

The subsets of PWM pairs chosen for further inspection were of exceptionally high support and statistical significance. They co-occurred within the same short segment of DNA throughout the human genome much more often than the others, and much more frequently than expected by chance given their individual distributions. Transcription factors binding to the motifs represented by these PWMs were therefore expected to bind and jointly regulate the expression of target genes.

### *Microarray verification*

We hypothesized that high-support, high-significance TF pairs or their target genes might be co-expressed in microarray data more often than other pairs. Therefore, we calculated the Pearson correlations of expression for all genes across 4742 human microarrays from the Stanford Microarray Database, but we saw no difference between the expression correlations of selected TF pairs and all TF pairs and no difference between genes containing both members of a high-support, high-significance motif pair 1 kb upstream of the transcriptional start site and genes without (data not shown).

### *Verification In the literature*

We next manually examined the literature for evidence of biological associations and joint regulation of target genes by the "genomewide" and "mouse" subsets of PWM pairs that were identified by data mining. We found that many of these TF pairs were readily verifiable in the literature as true co-regulators of human and mouse genes (Table 1). For example the subsets "mouse" and "genomewide" both included the pair "Ap-2, Egr1." Genes known to be regulated by these two transcription factors include tumor necrosis factor α [40,41], human phenylethanolamine N-methyltransferase [42], and rat chromogranin B [43]. The subsets "mouse" and "genomewide" contain the pair "Sp1, p53"; each has been shown to regulate ICAM-1[44,45]. A comparison of distributions for all pairs compared to 131 true positives collected from the literature revealed that true positive pairs exhibited higher support and confidence and lower *P*-values than did all pairs (Figure 4), regardless of whether the entire human genome, human promoters, or mouse chromosome 1 were mined. As an exhaustive manual analysis of the literature for all TF pairs was not feasible, we used high-throughput co-citation analysis to further assess the biological relevance of the high-support, high-confidence TF pairs.

**Table 1: High-confidence TF pairs verified in the literature.**

| TF pair | Gene Regulated | Source |
|---|---|---|
| Ap-2, p300 | Mouse CITED4 | [83] |
| Sp1, Gata2 | Human PDGFβ receptor | [84] |
| Sp1, p300 | Human ERK1 | [126] |
| Ap-2, Egr1 | Human tumor necrosis factor α, rat chromogranin B, human PNMT | [41–43, 67] |
| Ap-2, NFκB | Human tumor necrosis factor α | [41] |
| Egr1, Elk1 | Human tumor necrosis factor α | [41, 132] |
| Egr1, Nf1 | Human tissue factor pathway inhibitor 2 | [40] |
| Egr1, p300 | Human tumor necrosis factor α | [67] |
| Egr1, Sp1 | Human TFPI-2, human SOD, human cd95, human TNFα | [40, 74, 97, 132] |
| Sp1, p53 | Human Icam1 | [44, 45] |
| Mzf1, Sp1 | Human N-cadherin | [115] |
| Sp1, Srebp | Porcine LDL receptor, rat FAS | [123, 128] |
| Usf, Sp1 | Rat FAS, human Top3, human liver fructose1,6 biphosphatase | [101, 119, 123] |
| Aml1, NFκB | Human GM-CSF | [66] |
| Aml1, Srebp | Human fatty acid synthase | [64] |
| Elk1, p300 | Human tumor necrosis factor α | [132] |
| Gata2, NFκB | Human erythropoietin | [112] |
| Gata2, Sp1 | Human PDGF receptor | [84] |
| Nf1, NFκB | Human tissue factor pathway inhibitor 2 | [40] |
| NFκB, p300 | Human I-gamma 1, mouse tapasin | [87, 111] |
| Pax5, p300 | Human immunoglobin κ | [133] |
| Ap-1, NFκB | Human interleukin 6, human RANTES, human TNFα, human GM-CSF | [41, 60, 70, 77] |

Examples of high-confidence TF pairs that could be verified in the literature as co-regulators of mammalian genes.

### *High-throughput co-citation*

In order to determine whether the members of a TF pair were co-cited in the literature more often than expected by chance and more often than the pairs that were not significant, we used the CoCiteStats package in R [46] to calculate PubMed co-citation rates for all TF pairs and subsets. For each pair of PWMs, CoCiteStats calculates co-citation by determining the concordance, Jaccard index, and Hubert's Γ, as well as the *P*-values for these indices, which are significant at $P < 0.05$ [47]. Concordance is a straightforward measure of how many papers in PubMed co-cite both genes. The Jaccard index is the ratio of the number of papers containing both genes to the number of papers containing at least one of the two genes. Hubert's Γ measures the degree of association between two binary variables, ranges from -1 to 1, and can be interpreted similarly to the Pearson correlation [47]. Because papers that cite a large number of genes are less likely to contain meaningful information about interactions between any two genes cited in that paper than papers citing fewer genes, CoCiteStats also weights data for paper size (number of genes cited in a paper), gene size (number of papers that cite a gene), and both gene and paper size [47].

Figure 5 shows the fraction of total TF pairs with significant co-citation *P*-values ($P < 0.05$) in each dataset. Asterisks indicate a significant difference between all TF pairs and the selected subset as measured by a Chi square test. All sets indicated by "§" were significant after Bonferroni correction for multiple hypothesis testing. All three sub-

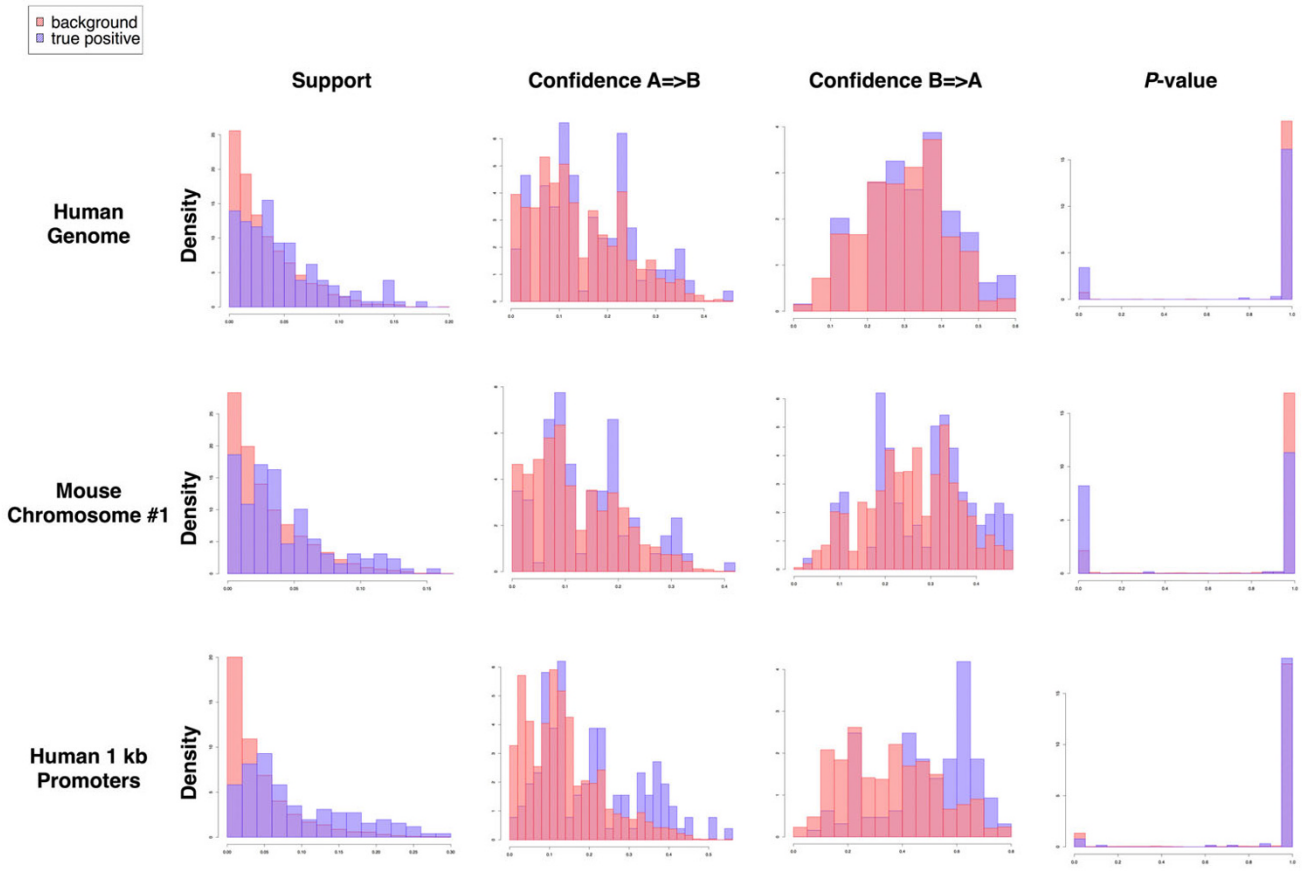


**Figure 4**
**Distributions of support, confidence, and *P*-value for true positives and all pairs**. Distribution histograms of support, confidence, and *P*-value for 131 true positives versus all pairs show higher support and confidence and lower *P*-values for true positives in the entire human genome, human promoter regions, and mouse chromosome 1.

sets showed substantially higher proportions of TF pairs enriched for low co-citation *P*-values in all cases than the set of all pairs, indicating that transcription factors binding to the PWMs that showed substantial association with one another on the genome were more likely to be co-cited in the literature, reflecting a likely biological association between them. This enrichment of "genomewide" was significant for most values at all adjustments. The subset "mouse" was enriched for significant concordances and Jaccard values when unadjusted or adjusted by paper size and was significant for all values when adjusted by both gene and paper size. The subset "promoter" was more significant after adjustments for gene size or both gene and paper size.

## Discussion

Data mining using association rules discovered biologically meaningful cooperating TF pairs. Known true positive TF pairs showed higher support, confidence, and significance than did all pairs. Mined pairs with high significance as measured by the hypergeometric probability distribution and a large difference between confidence A=>B and confidence B=>A were frequently verified in the literature and showed enrichment of low co-citation *P*-values. We found that data mining the entire human genome was a better indicator of biological significance than was mining mouse chromosome 1, as measured by co-citation.

Given that phylogenetically conserved transcription factor binding motifs are thought to be biologically useful [48], it is interesting that 60% of the TF pairs in the subset "genomewide" were also present in the subset "mouse." Comparison of TF pairs for multiple mouse chromosomes or across more than two mammals may lead to even better results. The smaller overlap between "promoter" and "mouse" (14%) and "promoter" and "genomewide" (42%) may be due in part to differences in sequence size and nucleotide frequency; the sequence mined for "promoter" was a tenth the size of the sequence mined for

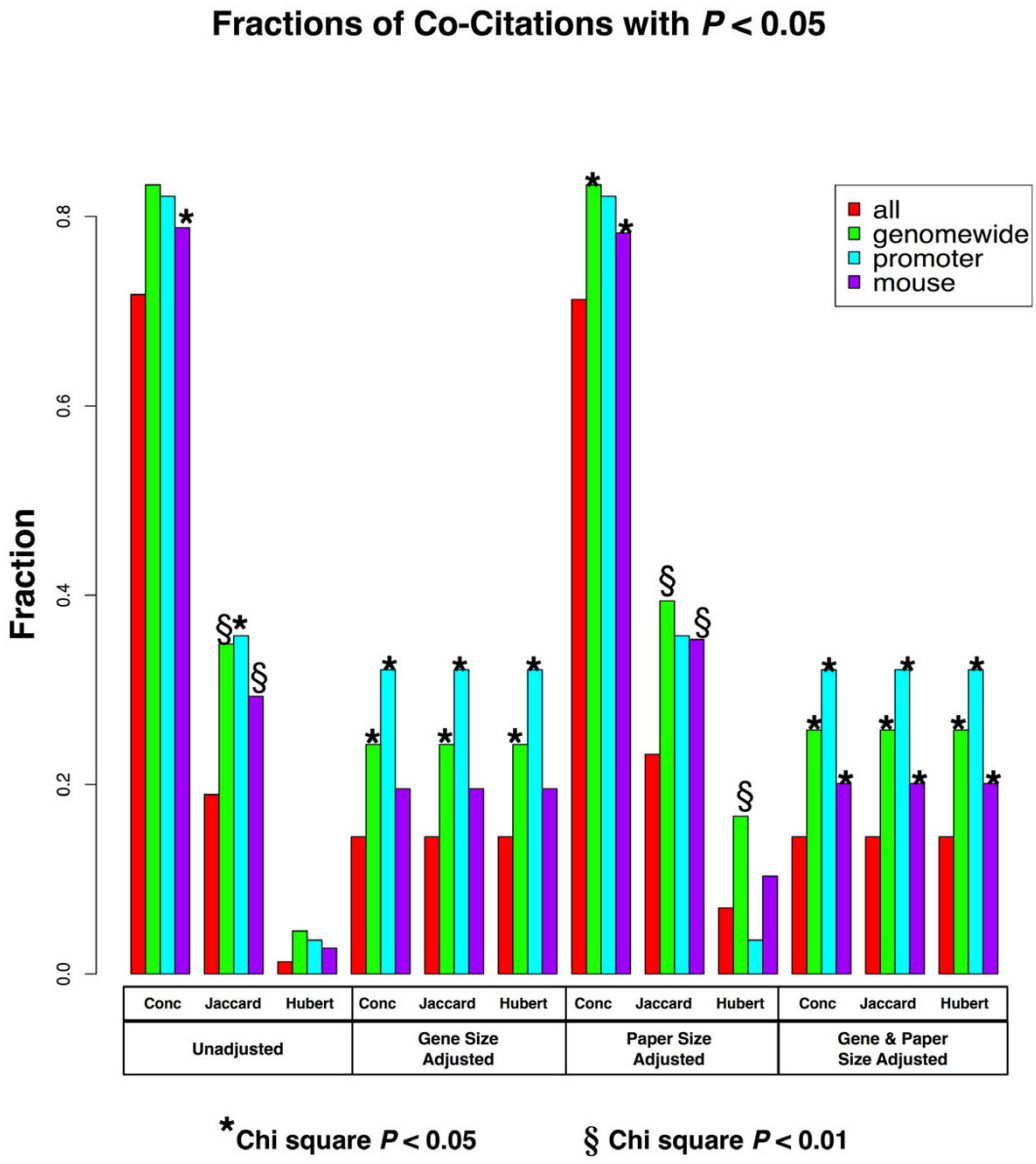


**Figure 5**
**Fractions of TF pairs with significant co-citation *P*-values**. Fractions of TF pairs with significant co-citation *P*-values ($P < 0.05$) in each dataset. Asterisks indicate a significant difference between all pairs and the selected subset as measured by a Chi square test. *P*-values significant after the Bonferroni correction for multiple hypothesis testing are indicated by "§".

"mouse" and ~1/200 the size of the sequence for "genomewide." Furthermore, the "promoter" sequence has a much higher GC content of 53% GC, while the human genome and mouse chromosome 1 are 41% GC; the PWMs used for mining have an average GC content of 46%.

Approximately 2900 of the TF pairs in our analysis were non-significant on mouse chromosome 1, human 1 kb promoter regions, or the human genome. The most confident of the remaining ~550 TF pairs may merit further study. Our estimated error rates for PWM matches located by Patser ranged from 3.5% to 61.5% with an average of 20% and a median of 18% (Additional file 4). Pairs containing a TF with a very high error rate are less likely to be of predictive value, but most TF pairs with high confidence and significance did not have very high Patser error rates.

Our approach is novel, low-cost, and straightforward to implement. The main advantage of this approach is that the signal for the association of transcription factors is

detectable using only the genome sequence and is not limited by lack of prior knowledge about physiological conditions or cell types in which the transcription factor combination may be active. Unlike clustering algorithms, which require items to be assigned to only one cluster, association rules allow items to be members of many groups and may discover these relationships. This algorithm also enables us to analyze a great number of motifs and large amount of sequence data for which Gibbs sampling is not currently feasible. One limitation of our current implementation is that we have applied it to identify only combinations of two distinct transcription factors. Although it is possible to discover associations of multiple transcription factors in the genome sequence through association rule mining, this is more computationally demanding.

As with any computational prediction, the significant challenge is verification of the predicted TF pairs. Co-citation analysis was particularly useful given that expected measures of biological association between the members of predicted TF pairs, such as correlated expression of target genes and network connectivity, were not useful. There are several possible explanations for why we did not observe correlations for significant mined TF pairs in microarray data. First, the activity of transcription factors may not be primarily regulated transcriptionally. Rather, transcription factors may require degradation of chaperones to become active, as does NFκB, or ligand binding may be needed to cause an active receptor to relocalize to the nucleus, as in the case of the estrogen receptor. While some transcription factors, such as targets of immediate early genes, may have similar mechanisms of transcriptional activation, it is likely that many, if not most cooperating transcription factors will have diverse means of transcriptional regulation and will thus not be co-expressed. Furthermore, due to noise and the fact that transcription factors may be inactive in many cell types and experimental conditions, any co-expression signature may be lost in large amounts of microarray data even for transcription factors known to be co-expressed. For example, across the 4247 microarrays we analyzed, the Pearson correlations for Fos with JunB and Jun were -0.11 and 0.146, respectively; the correlation was -0.116 for Gata2 and Gata3 and 0.24 for Sox5 and Sox6. Thus, even for known pairs of transcription factors, there is little detectable coexpression across a large microarray dataset.

We found that genes containing significant pairs of PWMs in their promoters were no more likely to be co-regulated than a background set. One possible explanation is that our list of 4742 microarrays represented a wide variety of experimental conditions, but many of the transcription factors we studied are active only under specific conditions satisfied in only a small number of experiments. Furthermore, the short, degenerate nature of position weight matrices means that thousands of 1 kb upstream regions are likely to contain any given PWM pair. We found that each PWM was present in the upstream regions of 10,948 genes on average, while the promoter region of each gene contained an average of 70 PWMs (data not shown). Thus, any comparisons of subsets became comparisons of most genes versus most genes, making it difficult to detect a change in the distribution of correlation coefficients. Observing correlated expression of the target genes of highly supported TF pairs would be much more likely if target genes could be more rigidly defined and a subset of microarray experiments was chosen to reflect likely conditions for transcription factor activity, but choosing these experiments is nontrivial, particularly for transcription factors that have not been well-studied.

Co-citation is not without drawbacks. The fact that two proteins are cited in a paper does not necessarily mean that they interact with one another. Furthermore, well-studied proteins are likely to be overrepresented while less-studied proteins will be missed. Validation by co-expression, however, requires knowledge of target genes and conditions for transcription factor activity; this may not be known or be feasible for experimental analysis. Future experimental validation of predicted associations could be accomplished by identifying binding targets for these transcription factors by genome-wide chromatin immunoprecipitation analyses and determining joint occupancy of target promoters by predicted combinations of transcription factors. Current maps of human protein-protein interactions [49-53] may not yet define many interactions for human transcription factors or may contain high rates of false positives [54], but they are constantly improving. We anticipate that better human protein-protein interaction maps will eventually provide a superior means of assessing performance of TF pair data mining, allowing this method to be refined to reveal both novel transcription factor interactions and biological context for previously uncharacterized transcription factors.

## Conclusion

Here we have described a novel genomic method for predicting biologically relevant, heterogeneous combinations of cooperating transcription factors by data mining using association rules to search genome information and identify over-represented proximal motifs. Using this approach, we show that that true positive cooperating TF pairs tend to have higher support, confidence, and significance, and that mined TF pairs with high confidence and significance are frequently verified in the literature and enriched for low co-citation *P*-values. Data mining the entire human genome enabled better discovery of biologically meaningful pairs than mining mouse chromosome 1, as measured by co-citation.

## Methods

### *Data transformation*

We collected 163 human position weight matrices (PWMs) from TRANSFAC [55] and removed those which were redundant or could not be mapped to RefSeq genes [39], leaving 83 PWMs for analysis (Additional file 1). We used Patser [56] to map all locations in the human genome assembly hg17 and in the repeat-masked human genome assembly hg18 [57] to which each transcription factor could bind with $P < 0.001$. We then divided the genome into 100 bp regions and scored each region for the presence or absence of each PWM. We chose a region size of 100 bp because it is compatible with the size of known cis-regulatory regions and large enough to contain multiple non-overlapping transcription factor binding motifs. PWMs tend towards large numbers of possible binding sites in the genome; 100 bp regions are small enough to prevent most regions from containing most motifs. We mined this matrix of genomic regions and motifs they contained for frequent itemsets, using association rules to search for $X \Rightarrow Y$ with high support $S$. Support and confidence were highly correlated between hg17 without repeat masking and hg18 with repeat masking (Additional file 2). High-support, high-confidence, significant PWM pairs were comparable between region sizes ranging from 75 bp to 225 bp, although larger region sizes yielded greater numbers of significant pairs.

### *Estimating Patser error rate for PWMs*

We estimated Patser error rates for each position weight matrix by calculating its average *P*-value across the genome as given by Patser, multiplying this by the size of the genome minus the length of masked repeats and then dividing by total number of matches to approximate the number of overestimated Patser matches.

### *Mining without overlap*

In order to avoid enrichment of PWMs with highly similar binding motifs, we mined the human genome without allowing motif overlap, one motif pair at a time. That is, for each TF pair AB (83 transcription factors taken two at a time, or 3403 pairs), after all possible binding motifs for A and B respectively were identified, any overlapping A and B motifs were removed before assigning the remaining non-overlapping sites to their respective 100 bp regions (Figure 2). The full set of matches for each factor was restored at the beginning of each iteration, so overlaps between A and C were unaffected by overlaps between A and B. For example, if transcription factor A had a binding motif of width 5 which was present at positions 100, 130, and 150, while factor B had a binding motif of width 7 present at 102, 160, and 175, the binding sites 100A and 102B would be removed from calculations due to overlap; the remaining binding sites would still allow the region from 100 to 200 to be scored as containing A and B. After scoring each 100 bp region, we calculated association rule support (proportions of regions) for A, B, and AB for each pair on each chromosome, correcting for the proportion of the genome that was repeat-masked. Additionally, we calculated confidence for $A \Rightarrow B$ and $B \Rightarrow A$ and a *P*-value based on the hypergeometric probability of observing the association between A and B by chance, given the individual distributions of their binding motifs in the genome, again correcting for repeat masking. To allow phylogenetic comparison and comparison of promoters versus the entire genome, we performed identical pairwise mining on mouse chromosome 1 and on the subset of the human genome that was 1 kb upstream from the transcriptional start site of all human RefSeq genes.

### *Microarray data*

To determine whether transcription factor pairs with high support and high confidence were highly co-expressed, we downloaded and analyzed a dataset consisting of 4742 human microarrays from the Stanford Microarray Database [58] and calculated the Pearson correlation for each gene pair with 100 or more experimental data points. We defined a list of potential target genes for TF pairs by scanning 1 kb upstream from the transcriptional start site of each RefSeq gene for each PWM.

### *True positives*

From the Compel database [59] and the literature, we collected 131 transcription factor pairs known to co-regulate mammalian genes [40,41,44,45,60-131].

## Software availability and requirements

Project name: Miner

Project home page: http://sourceforge.net/projects/miner/

Operating system(s): All POSIX (Linux/BSD/UNIX-like) operating systems.

Programming language: C++

License: Academic Free License

## Authors' contributions

XM carried out data analysis. XM and SN wrote the computer code. XM and VI wrote the manuscript. DM provided the initial impetus for the design of this project. All authors participated in the design of the study. All authors read and approved the final manuscript.

## Additional material

### Additional file 1

***83 transcription factors from TRANSFAC.*** *A list of the 83 transcription factor position weight matrices from TRANSFAC used for this analysis.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-445-S1.doc]

### Additional file 2

***Effects of repeat masking.*** *Support, confidence A=>B, and confidence B=>A are highly correlated between hg17 without repeat masking and hg18 with repeat masking.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-445-S2.tiff]

### Additional file 3

***The subsets "genomewide", "mouse", and "promoter".*** *"Genomewide", "Promoter", and "Mouse" are defined as top 50% difference between confidence A=>B and confidence B=>A and* $P < 0.05$ *as measured by the hypergeometric distribution. Pairs indicated in bold have been verified in the literature.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-445-S3.doc]

### Additional file 4

***Estimated Patser error rates for PWMs.*** *Approximate overestimation rates of position weight matrices from Patser.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-445-S4.doc]

## Acknowledgements

We thank Orly Alter, Edward Marcotte, Patrick Killion, and Iyer lab members for advice and suggestions. This work was supported in part by a NIAAA Alcohol Training Grant and an ITR grant from the National Science Foundation.

## References

1. Tang H, Veldman MB, Goldman D: **Characterization of a muscle-specific enhancer in human MuSK promoter reveals the essential role of myogenin in controlling activity-dependent gene regulation.** *J Biol Chem* 2006, **281(7):**3943-3953.
2. Shah R, Rahaman B, Hurley CK, Posch PE: **Allelic diversity in the TGFB1 regulatory region: characterization of novel functional single nucleotide polymorphisms.** *Hum Genet* 2006, **119(1-2):**61-74.
3. Kammandel B, Chowdhury K, Stoykova A, Aparicio S, Brenner S, Gruss P: **Distinct cis-essential modules direct the time-space pattern of the Pax6 gene activity.** *Dev Biol* 1999, **205(1):**79-97.
4. Davidson EH: **Genomic regulatory systems : development and evolution.** San Diego , Academic Press; 2001:xii, 261.
5. Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW: **Identification of conserved regulatory elements by comparative genome analysis.** *J Biol* 2003, **2(2):**13.
6. Kim J, Bhinge AA, Morgan XC, Iyer VR: **Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment.** *Nat Methods* 2005, **2(1):**47-53.
7. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA: **Transcriptional regulatory networks in Saccharomyces cerevisiae.** *Science* 2002, **298(5594):**799-804.
8. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J, Brodsky AS, Keeton EK, Fertuck KC, Hall GF, Wang Q, Bekiranov S, Sementchenko V, Fox EA, Silver PA, Gingeras TR, Liu XS, Brown M: **Genome-wide analysis of estrogen receptor binding sites.** *Nat Genet* 2006, **38(11):**1289-1297.
9. Zheng Y, Josefowicz SZ, Kas A, Chu TT, Gavin MA, Rudensky AY: **Genome-wide analysis of Foxp3 target genes in developing and mature regulatory T cells.** *Nature* 2007, **445(7130):**936-940.
10. Tupler R, Perini G, Green MR: **Expressing the human genome.** *Nature* 2001, **409(6822):**832-833.
11. Messina DN, Glasscock J, Gish W, Lovett M: **An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression.** *Genome Res* 2004, **14(10B):**2041-2047.
12. Small S, Blair A, Levine M: **Regulation of even-skipped stripe 2 in the Drosophila embryo.** *Embo J* 1992, **11(11):**4047-4057.
13. Rivera-Pomar R, Lu X, Perrimon N, Taubert H, Jackle H: **Activation of posterior gap gene expression in the Drosophila blastoderm.** *Nature* 1995, **376(6537):**253-256.
14. Philipsen S, Talbot D, Fraser P, Grosveld F: **The beta-globin dominant control region: hypersensitive site 2.** *Embo J* 1990, **9(7):**2159-2167.
15. Rothenberg EV, Ward SB: **A dynamic assembly of diverse transcription factors integrates activation and cell-type information for interleukin 2 gene regulation.** *Proc Natl Acad Sci U S A* 1996, **93(18):**9358-9365.
16. Crowley EM, Roeder K, Bina M: **A statistical model for locating regulatory regions in genomic DNA.** *J Mol Biol* 1997, **268(1):**8-14.
17. Wagner A: **A computational genomics approach to the identification of gene networks.** *Nucleic Acids Res* 1997, **25(18):**3594-3604.
18. Frith MC, Hansen U, Weng Z: **Detection of cis-element clusters in higher eukaryotic DNA.** *Bioinformatics* 2001, **17(10):**878-889.
19. Markstein M, Markstein P, Markstein V, Levine MS: **Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo.** *Proc Natl Acad Sci U S A* 2002, **99(2):**763-768.
20. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome.** *Proc Natl Acad Sci U S A* 2002, **99(2):**757-762.
21. Lifanov AP, Makeev VJ, Nazina AG, Papatsenko DA: **Homotypic regulatory clusters in Drosophila.** *Genome Res* 2003, **13(4):**579-588.
22. Wasserman WW, Fickett JW: **Identification of regulatory regions which confer muscle-specific gene expression.** *J Mol Biol* 1998, **278(1):**167-181.
23. Frech K, Quandt K, Werner T: **Muscle actin genes: a first step towards computational classification of tissue specific promoters.** In *Silico Biol* 1998, **1(1):**29-38.
24. Tronche F, Ringeisen F, Blumenfeld M, Yaniv M, Pontoglio M: **Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome.** *J Mol Biol* 1997, **266(2):**231-245.
25. Kel A, Kel-Margoulis O, Babenko V, Wingender E: **Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells.** *J Mol Biol* 1999, **288(3):**353-376.
26. Kel AE, Kel-Margoulis OV, Farnham PJ, Bartley SM, Wingender E, Zhang MQ: **Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors.** *J Mol Biol* 2001, **309(1):**99-120.
27. Berman BP, Pfeiffer BD, Laverty TR, Salzberg SL, Rubin GM, Eisen MB, Celniker SE: **Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura.** *Genome Biol* 2004, **5(9):**R61.
28. Halfon MS, Grad Y, Church GM, Michelson AM: **Computation-based discovery of related transcriptional regulatory mod-**

**ules and motifs using an experimentally validated combinatorial model.** *Genome Res* 2002, **12(7):**1019-1028.
29. Rebeiz M, Reeves NL, Posakony JW: **SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation.** *Proc Natl Acad Sci U S A* 2002, **99(15):**9888-9893.
30. De Bosscher K, Vanden Berghe W, Haegeman G: **The interplay between the glucocorticoid receptor and nuclear factor-kappaB or activator protein-1: molecular mechanisms for gene repression.** *Endocr Rev* 2003, **24(4):**488-522.
31. Bartholdy B, Matthias P: **Transcriptional control of B cell development and function.** *Gene* 2004, **327(1):**1-23.
32. Beer MA, Tavazoie S: **Predicting gene expression from sequence.** *Cell* 2004, **117(2):**185-198.
33. Das D, Banerjee N, Zhang MQ: **Interacting models of cooperative gene regulation.** *Proc Natl Acad Sci U S A* 2004, **101(46):**16234-16239.
34. Sharan R, Ben-Hur A, Loots GG, Ovcharenko I: **CREME: Cis-Regulatory Module Explorer for the human genome.** *Nucleic Acids Res* 2004, **32(Web Server issue):**W253-6.
35. Brazma A, Vilo J, Ukkonen E, Valtonen K: **Data mining for regulatory elements in yeast genome.** *Proc Int Conf Intell Syst Mol Biol* 1997, **5:**65-74.
36. Horng JT, Huang HD, Jin MH, Wu LC, Huang SL: **The repetitive sequence database and mining putative regulatory elements in gene promoter regions.** *J Comput Biol* 2002, **9(4):**621-640.
37. Horng JT, Lin FM, Lin JH, Huang HD, Liu BJ: **Database of repetitive elements in complete genomes and data mining using transcription factor binding sites.** *IEEE Trans Inf Technol Biomed* 2003, **7(2):**93-100.
38. Agrawal R and Srikant, Ramakrishnan: **Fast Algorithms for Mining Association Rules.** *VLDB 1994* 1994:487-499.
39. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2005, **33(Database issue):**D501-4.
40. Hube F, Reverdiau P, Iochmann S, Cherpi-Antar C, Gruel Y: **Characterization and functional analysis of TFPI-2 gene promoter in a human choriocarcinoma cell line.** *Thromb Res* 2003, **109(4):**207-215.
41. Szlosarek PW, Balkwill FR: **Tumour necrosis factor alpha: a potential target for the therapy of solid tumours.** *Lancet Oncol* 2003, **4(9):**565-573.
42. Wong DL, Siddall BJ, Ebert SN, Bell RA, Her S: **Phenylethanolamine N-methyltransferase gene expression: synergistic activation by Egr-1, AP-2 and the glucocorticoid receptor.** *Brain Res Mol Brain Res* 1998, **61(1-2):**154-161.
43. Mahapatra NR, Mahata M, Ghosh S, Gayen JR, O'Connor DT, Mahata SK: **Molecular basis of neuroendocrine cell type-specific expression of the chromogranin B gene: Crucial role of the transcription factors CREB, AP-2, Egr-1 and Sp1.** *J Neurochem* 2006, **99(1):**119-133.
44. Pazdrak K, Shi XZ, Sarna SK: **TNFalpha suppresses human colonic circular smooth muscle cell contractility by SP1- and NF-kappaB-mediated induction of ICAM-1.** *Gastroenterology* 2004, **127(4):**1096-1109.
45. Gorgoulis VG, Zacharatos P, Kotsinas A, Kletsas D, Mariatos G, Zoumpourlis V, Ryan KM, Kittas C, Papavassiliou AG: **p53 activates ICAM-1 (CD54) expression in an NF-kappaB-independent manner.** *Embo J* 2003, **22(7):**1567-1578.
46. The R Development Core Team: **"R: A Language and Environment for Statistical Computing.".** *R Foundation for Statistical Computing, Vienna, Austria* 2005.
47. Gentleman R: **Bioinformatics and computational biology solutions using R and Bioconductor.** In *Statistics for biology and health* New York , Springer Science+Business Media; 2005:xix, 473.
48. Elemento O, Tavazoie S: **Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach.** *Genome Biol* 2005, **6(2):**R18.
49. Ramani AK, Bunescu RC, Mooney RJ, Marcotte EM: **Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome.** *Genome Biol* 2005, **6(5):**R40.
50. Lehner B, Fraser AG: **A first-draft human protein-interaction map.** *Genome Biol* 2004, **5(9):**R63.
51. Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM: **Probabilistic model of the human protein-protein interaction network.** *Nat Biotechnol* 2005, **23(8):**951-959.
52. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE: **A human protein-protein interaction network: a resource for annotating the proteome.** *Cell* 2005, **122(6):**957-968.
53. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M: **Towards a proteome-scale map of the human protein-protein interaction network.** *Nature* 2005, **437(7062):**1173-1178.
54. Hart GT, Ramani AK, Marcotte EM: **How complete are current yeast and human protein-interaction networks?** *Genome Biol* 2006, **7(11):**120.
55. Wingender E, Dietze P, Karas H, Knuppel R: **TRANSFAC: a database on transcription factors and their DNA binding sites.** *Nucleic Acids Res* 1996, **24(1):**238-241.
56. Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, **15(7-8):**563-577.
57. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ: **The UCSC Genome Browser Database.** *Nucleic Acids Res* 2003, **31(1):**51-54.
58. Demeter J, Beauheim C, Gollub J, Hernandez-Boussard T, Jin H, Maier D, Matese JC, Nitzberg M, Wymore F, Zachariah ZK, Brown PO, Sherlock G, Ball CA: **The Stanford Microarray Database: implementation of new analysis tools and open source release of software.** *Nucleic Acids Res* 2007, **35(Database issue):**D766-70.
59. Kel-Margoulis OV, Kel AE, Reuter I, Deineko IV, Wingender E: **TRANSCompel: a database on composite regulatory elements in eukaryotic genes.** *Nucleic Acids Res* 2002, **30(1):**332-334.
60. Shannon MF, Coles LS, Vadas MA, Cockerill PN: **Signals for activation of the GM-CSF promoter and enhancer in T cells.** *Crit Rev Immunol* 1997, **17(3-4):**301-323.
61. Bertrand-Philippe M, Ruddell RG, Arthur MJ, Thomas J, Mungalsingh N, Mann DA: **Regulation of tissue inhibitor of metalloproteinase 1 gene transcription by RUNX1 and RUNX2.** *J Biol Chem* 2004, **279(23):**24530-24539.
62. Maier H, Ostraat R, Gao H, Fields S, Shinton SA, Medina KL, Ikawa T, Murre C, Singh H, Hardy RR, Hagman J: **Early B cell factor cooperates with Runx1 and mediates epigenetic changes associated with mb-1 transcription.** *Nat Immunol* 2004, **5(10):**1069-1077.
63. Hromas R, Davis B, Rauscher FJ 3rd, Klemsz M, Tenen D, Hoffman S, Xu D, Morris JF: **Hematopoietic transcriptional regulation by the myeloid zinc finger gene, MZF-1.** *Curr Top Microbiol Immunol* 1996, **211:**159-164.
64. Wolf SS, Roder K, Sickinger S, Schweizer M: **The FIRE3-mediated sterol response of the FAS promoter requires NF-Y/CBF as a coactivator.** *Biol Chem* 2001, **382(7):**1083-1088.
65. Libermann TA, Pan Z, Akbarali Y, Hetherington CJ, Boltax J, Yergeau DA, Zhang DE: **AML1 (CBFalpha2) cooperates with B cell-specific activating protein (BSAP/PAX5) in activation of the B cell-specific BLK gene promoter.** *J Biol Chem* 1999, **274(35):**24671-24676.
66. Cockerill PN, Osborne CS, Bert AG, Grotto RJ: **Regulation of GM-CSF gene transcription by core-binding factor.** *Cell Growth Differ* 1996, **7(7):**917-922.
67. Barthel R, Tsytsykova AV, Barczak AK, Tsai EY, Dascher CC, Brenner MB, Goldfeld AE: **Regulation of tumor necrosis factor alpha gene expression by mycobacteria involves the assembly of a unique enhanceosome dependent on the coactivator proteins CBP/p300.** *Mol Cell Biol* 2003, **23(2):**526-533.
68. Falvo JV, Uglialoro AM, Brinkman BM, Merika M, Parekh BS, Tsai EY, King HC, Morielli AD, Peralta EG, Maniatis T, Thanos D, Goldfeld AE: **Stimulus-specific assembly of enhancer complexes on the**

**tumor necrosis factor alpha gene promoter.** *Mol Cell Biol* 2000, **20(6):**2239-2247.
69. Andriamanalijaona R, Felisaz N, Kim SJ, King-Jones K, Lehmann M, Pujol JP, Boumediene K: **Mediation of interleukin-1beta-induced transforming growth factor beta1 expression by activator protein 4 transcription factor in primary cultures of bovine articular chondrocytes: possible cooperation with activator protein 1.** *Arthritis Rheum* 2003, **48(6):**1569-1581.
70. Wickremasinghe MI, Thomas LH, O'Kane CM, Uddin J, Friedland JS: **Transcriptional mechanisms regulating alveolar epithelial cell-specific CCL5 secretion in pulmonary tuberculosis.** *J Biol Chem* 2004, **279(26):**27199-27210.
71. Cohn MA, Hjelmso I, Wu LC, Guldberg P, Lukanidin EM, Tulchinsky EM: **Characterization of Sp1, AP-1, CBF and KRC binding sites and minisatellite DNA as functional elements of the metastasis-associated mts1/S100A4 gene intronic enhancer.** *Nucleic Acids Res* 2001, **29(16):**3335-3346.
72. Johnson BV, Bert AG, Ryan GR, Condina A, Cockerill PN: **Granulocyte-macrophage colony-stimulating factor enhancer activation requires cooperation between NFAT and AP-1 elements and is associated with extensive nucleosome reorganization.** *Mol Cell Biol* 2004, **24(18):**7914-7930.
73. Britos-Bray M, Friedman AD: **Core binding factor cannot synergistically activate the myeloperoxidase proximal enhancer in immature myeloid cells without c-Myb.** *Mol Cell Biol* 1997, **17(9):**5127-5135.
74. Li-Weber M, Krammer PH: **Function and regulation of the CD95 (APO-1/Fas) ligand in the immune system.** *Semin Immunol* 2003, **15(3):**145-157.
75. Debieve F, Thomas K: **Control of the human inhibin alpha chain promoter in cytotrophoblast cells differentiating into syncytium.** *Mol Hum Reprod* 2002, **8(3):**262-270.
76. Ebert SN, Ficklin MB, Her S, Siddall BJ, Bell RA, Ganguly K, Morita K, Wong DL: **Glucocorticoid-dependent action of neural crest factor AP-2: stimulation of phenylethanolamine N-methyltransferase gene expression.** *J Neurochem* 1998, **70(6):**2286-2295.
77. Faggioli L, Costanzo C, Donadelli M, Palmieri M: **Activation of the Interleukin-6 promoter by a dominant negative mutant of c-Jun.** *Biochim Biophys Acta* 2004, **1692(1):**17-24.
78. Zhou L, Nazarian AA, Smale ST: **Interleukin-10 inhibits interleukin-12 p40 gene transcription by targeting a late event in the activation pathway.** *Mol Cell Biol* 2004, **24(6):**2385-2396.
79. Zhou T, Chiang CM: **Sp1 and AP2 regulate but do not constitute TATA-less human TAF(II)55 core promoter activity.** *Nucleic Acids Res* 2002, **30(19):**4145-4157.
80. Yang H, Wang J, Ou X, Huang ZZ, Lu SC: **Cloning and analysis of the rat glutamate-cysteine ligase modifier subunit promoter.** *Biochem Biophys Res Commun* 2001, **285(2):**476-482.
81. Moon SK, Cha BY, Kim CH: **ERK1/2 mediates TNF-alpha-induced matrix metalloproteinase-9 expression in human vascular smooth muscle cells via the regulation of NF-kappaB and AP-1: Involvement of the ras dependent pathway.** *J Cell Physiol* 2004, **198(3):**417-427.
82. Shi Q, Le X, Abbruzzese JL, Wang B, Mujaida N, Matsushima K, Huang S, Xiong Q, Xie K: **Cooperation between transcription factor AP-1 and NF-kappaB in the induction of interleukin-8 in human pancreatic adenocarcinoma cells by hypoxia.** *J Interferon Cytokine Res* 1999, **19(12):**1363-1371.
83. Yahata T, Takedatsu H, Dunwoodie SL, Braganca J, Swingler T, Withington SL, Hur J, Coser KR, Isselbacher KJ, Bhattacharya S, Shioda T: **Cloning of mouse Cited4, a member of the CITED family p300/CBP-binding transcriptional coactivators: induced expression in mammary epithelial cells.** *Genomics* 2002, **80(6):**601-613.
84. Kaneko M, Yang W, Matsumoto Y, Watt F, Funa K: **Activity of a novel PDGF beta-receptor enhancer during the cell cycle and upon differentiation of neuroblastoma.** *Exp Cell Res* 2006, **312(11):**2028-2039.
85. Becker C, Wirtz S, Ma X, Blessing M, Galle PR, Neurath MF: **Regulation of IL-12 p40 promoter activity in primary human monocytes: roles of NF-kappaB, CCAAT/enhancer-binding protein beta, and PU.1 and identification of a novel repressor element (GA-12) that responds to IL-4 and prostaglandin E(2).** *J Immunol* 2001, **167(5):**2608-2618.
86. Braganca J, Eloranta JJ, Bamforth SD, Ibbitt JC, Hurst HC, Bhattacharya S: **Physical and functional interactions among AP-2 transcription factors, p300/CREB-binding protein, and CITED2.** *J Biol Chem* 2003, **278(18):**16021-16029.
87. Dryer RL, Covey LR: **A novel NF-kappa B-regulated site within the human I gamma 1 promoter requires p300 for optimal transcriptional activity.** *J Immunol* 2005, **175(7):**4499-4507.
88. Barski OA, Papusha VZ, Kunkel GR, Gabbay KH: **Regulation of aldehyde reductase expression by STAF and CHOP.** *Genomics* 2004, **83(1):**119-129.
89. Lavrovsky Y, Schwartzman ML, Levere RD, Kappas A, Abraham NG: **Identification of binding sites for transcription factors NF-kappa B and AP-2 in the promoter region of the human heme oxygenase 1 gene.** *Proc Natl Acad Sci U S A* 1994, **91(13):**5987-5991.
90. Mura C, Le Gac G, Jacolot S, Ferec C: **Transcriptional regulation of the human HFE gene indicates high liver expression and erythropoiesis coregulation.** *Faseb J* 2004, **18(15):**1922-1924.
91. Lahlil R, Lecuyer E, Herblot S, Hoang T: **SCL assembles a multifactorial complex that determines glycophorin A expression.** *Mol Cell Biol* 2004, **24(4):**1439-1452.
92. Holzmann C, Schmidt T, Thiel G, Epplen JT, Riess O: **Functional characterization of the human Huntington's disease gene promoter.** *Brain Res Mol Brain Res* 2001, **92(1-2):**85-97.
93. Lin CS, Chow S, Lau A, Tu R, Lue TF: **Identification and regulation of human PDE5A gene promoter.** *Biochem Biophys Res Commun* 2001, **280(3):**684-692.
94. Malakooti J, Memark VC, Dudeja PK, Ramaswamy K: **Molecular cloning and functional analysis of the human Na(+)/H(+) exchanger NHE3 promoter.** *Am J Physiol Gastrointest Liver Physiol* 2002, **282(3):**G491-500.
95. Pocock J, Gomez-Guerrero C, Harendza S, Ayoub M, Hernandez-Vargas P, Zahner G, Stahl RA, Thaiss F: **Differential activation of NF-kappa B, AP-1, and C/EBP in endotoxin-tolerant rats: mechanisms for in vivo regulation of glomerular RANTES/CCL5 expression.** *J Immunol* 2003, **170(12):**6280-6291.
96. Gu JM, Fukudome K, Esmon CT: **Characterization and regulation of the 5'-flanking region of the murine endothelial protein C receptor gene.** *J Biol Chem* 2000, **275(17):**12481-12488.
97. Rojo AI, Salinas M, Martin D, Perona R, Cuadrado A: **Regulation of Cu/Zn-superoxide dismutase expression via the phosphatidylinositol 3 kinase/Akt pathway and nuclear factor-kappaB.** *J Neurosci* 2004, **24(33):**7324-7334.
98. Minc E, de Coppet P, Masson P, Thiery L, Dutertre S, Amor-Gueret M, Jaulin C: **The human copper-zinc superoxide dismutase gene (SOD1) proximal promoter is regulated by Sp1, Egr-1, and WT1 via non-canonical binding sites.** *J Biol Chem* 1999, **274(1):**503-509.
99. Seo SJ, Kim HT, Cho G, Rho HM, Jung G: **Sp1 and C/EBP-related factor regulate the transcription of human Cu/Zn SOD gene.** *Gene* 1996, **178(1-2):**177-185.
100. Kim HT, Kim YH, Nam JW, Lee HJ, Rho HM, Jung G: **Study of 5'-flanking region of human Cu/Zn superoxide dismutase.** *Biochem Biophys Res Commun* 1994, **201(3):**1526-1533.
101. Kim JC, Yoon JB, Koo HS, Chung IK: **Cloning and characterization of the 5'-flanking region for the human topoisomerase III gene.** *J Biol Chem* 1998, **273(40):**26130-26137.
102. Xu Z, Dziarski R, Wang Q, Swartz K, Sakamoto KM, Gupta D: **Bacterial peptidoglycan-induced tnf-alpha transcription is mediated through the transcription factors Egr-1, Elk-1, and NF-kappaB.** *J Immunol* 2001, **167(12):**6975-6982.
103. Yu X, Zhu X, Pi W, Ling J, Ko L, Takeda Y, Tuan D: **The long terminal repeat (LTR) of ERV-9 human endogenous retrovirus binds to NF-Y in the assembly of an active LTR enhancer complex NF-Y/MZF1/GATA-2.** *J Biol Chem* 2005, **280(42):**35184-35194.
104. Han L, Lu J, Pan L, Wang X, Shao Y, Han S, Huang B: **Histone acetyltransferase p300 regulates the transcription of human erythroid-specific 5-aminolevulinate synthase gene.** *Biochem Biophys Res Commun* 2006, **348(3):**799-806.
105. Neish AS, Williams AJ, Palmer HJ, Whitley MZ, Collins T: **Functional analysis of the human vascular cell adhesion molecule 1 promoter.** *J Exp Med* 1992, **176(6):**1583-1593.
106. Da Silva CA, Heilbock C, Kassel O, Frossard N: **Transcription of stem cell factor (SCF) is potentiated by glucocorticoids and interleukin-1beta through concerted regulation of a GRE-**

**like and an NF-kappaB response element.** *Faseb J* 2003, **17(15):**2334-2336.
107. Hermoso MA, Matsuguchi T, Smoak K, Cidlowski JA: **Glucocorticoids and tumor necrosis factor alpha cooperatively regulate toll-like receptor 2 gene expression.** *Mol Cell Biol* 2004, **24(11):**4743-4756.
108. Khan S, Barhoumi R, Burghardt R, Liu S, Kim K, Safe S: **Molecular mechanism of inhibitory aryl hydrocarbon receptor-estrogen receptor/Sp1 cross talk in breast cancer cells.** *Mol Endocrinol* 2006, **20(9):**2199-2214.
109. Manoli I, Le H, Alesci S, McFann KK, Su YA, Kino T, Chrousos GP, Blackman MR: **Monoamine oxidase-A is a major target gene for glucocorticoids in human skeletal muscle cells.** *Faseb J* 2005, **19(10):**1359-1361.
110. Gobin SJ, Biesta P, Van den Elsen PJ: **Regulation of human beta 2-microglobulin transactivation in hematopoietic cells.** *Blood* 2003, **101(8):**3058-3064.
111. Herrmann F, Trowsdale J, Huber C, Seliger B: **Cloning and functional analyses of the mouse tapasin promoter.** *Immunogenetics* 2003, **55(6):**379-388.
112. La Ferla K, Reimann C, Jelkmann W, Hellwig-Burgel T: **Inhibition of erythropoietin gene expression signaling involves the transcription factors GATA-2 and NF-kappaB.** *Faseb J* 2002, **16(13):**1811-1813.
113. Wu CX, Zhao WP, Kishi H, Dokan J, Jin ZX, Wei XC, Yokoyama KK, Muraguchi A: **Activation of mouse RAG-2 promoter by Myc-associated zinc finger protein.** *Biochem Biophys Res Commun* 2004, **317(4):**1096-1102.
114. Biesiada E, Hamamori Y, Kedes L, Sartorelli V: **Myogenic basic helix-loop-helix proteins and Sp1 interact as components of a multiprotein transcriptional complex required for activity of the human cardiac alpha-actin promoter.** *Mol Cell Biol* 1999, **19(4):**2577-2584.
115. Le Mee S, Fromigue O, Marie PJ: **Sp1/Sp3 and the myeloid zinc finger gene MZF1 regulate the human N-cadherin promoter in osteoblasts.** *Exp Cell Res* 2005, **302(1):**129-142.
116. Kang NY, Park YD, Choi HJ, Kim KS, Lee YC, Kim CH: **Regulatory elements involved in transcription of the human NeuAcalpha2,3Galbeta1,3GalNAcalpha2,6-sialyltransferase (hST6GalNAc IV) gene.** *Mol Cells* 2004, **18(2):**157-162.
117. Furlong EE, Rein T, Martin F: **YY1 and NF1 both activate the human p53 promoter by alternatively binding to a composite element, and YY1 and E1A cooperate to amplify p53 promoter activity.** *Mol Cell Biol* 1996, **16(10):**5933-5945.
118. Inoue A, Omoto Y, Yamaguchi Y, Kiyama R, Hayashi SI: **Transcription factor EGR3 is involved in the estrogen-signaling pathway in breast cancer cells.** *J Mol Endocrinol* 2004, **32(3):**649-661.
119. Herzog B, Waltner-Law M, Scott DK, Eschrich K, Granner DK: **Characterization of the human liver fructose-1,6-bisphosphatase gene promoter.** *Biochem J* 2000, **351 Pt 2:**385-392.
120. Xiao S, Marshak-Rothstein A, Ju ST: **Sp1 is the major fasl gene activator in abnormal CD4(-)CD8(-)B220(+) T cells of lpr and gld mice.** *Eur J Immunol* 2001, **31(11):**3339-3348.
121. Golubovskaya V, Kaur A, Cance W: **Cloning and characterization of the promoter region of human focal adhesion kinase gene: nuclear factor kappa B and p53 binding sites.** *Biochim Biophys Acta* 2004, **1678(2-3):**111-125.
122. Schafer H, Diebel J, Arlt A, Trauzold A, Schmidt WE: **The promoter of human p22/PACAP response gene 1 (PRG1) contains functional binding sites for the p53 tumor suppressor and for NFkappaB.** *FEBS Lett* 1998, **436(2):**139-143.
123. Schweizer M, Roder K, Zhang L, Wolf SS: **Transcription factors acting on the promoter of the rat fatty acid synthase gene.** *Biochem Soc Trans* 2002, **30(Pt 6):**1070-1072.
124. Hoffmeister A, Ropolo A, Vasseur S, Mallo GV, Bodeker H, Ritz-Laser B, Dressler GR, Vaccaro MI, Dagorn JC, Moreno S, Iovanna JL: **The HMG-I/Y-related protein p8 binds to p300 and Pax2 transactivation domain-interacting protein to regulate the transactivation activity of the Pax2A and Pax2B transcription factors on the glucagon gene promoter.** *J Biol Chem* 2002, **277(25):**22314-22319.
125. Gordon SJ, Saleque S, Birshtein BK: **Yin Yang 1 is a lipopolysaccharide-inducible activator of the murine 3' Igh enhancer, hs3.** *J Immunol* 2003, **170(11):**5549-5557.
126. Chu BY, Tran K, Ku TK, Crowe DL: **Regulation of ERK1 gene expression by coactivator proteins.** *Biochem J* 2005, **392(Pt 3):**589-599.
127. Ikeda Y, Yamamoto J, Okamura M, Fujino T, Takahashi S, Takeuchi K, Osborne TF, Yamamoto TT, Ito S, Sakai J: **Transcriptional regulation of the murine acetyl-CoA synthetase 1 gene through multiple clustered binding sites for sterol regulatory element-binding proteins and a single neighboring site for Sp1.** *J Biol Chem* 2001, **276(36):**34259-34269.
128. Sekar N, Veldhuis JD: **Involvement of Sp1 and SREBP-1a in transcriptional activation of the LDL receptor gene by insulin and LH in cultured porcine granulosa-luteal cells.** *Am J Physiol Endocrinol Metab* 2004, **287(1):**E128-35.
129. Armelin-Correa LM, Lin CJ, Barbosa A, Bagatini K, Winnischofer SM, Sogayar MC, Passos-Bueno MR: **Characterization of human Collagen XVIII promoter 2: interaction of Sp1, Sp3 and YY1 with the regulatory region and a SNP that increases transcription in hepatocytes.** *Matrix Biol* 2005, **24(8):**550-559.
130. Kawada H, Nishiyama C, Takagi A, Tokura T, Nakano N, Maeda K, Mayuzumi N, Ikeda S, Okumura K, Ogawa H: **Transcriptional regulation of ATP2C1 gene by Sp1 and YY1 and reduced function of its promoter in Hailey-Hailey disease keratinocytes.** *J Invest Dermatol* 2005, **124(6):**1206-1214.
131. Perrotti D, Melotti P, Skorski T, Casella I, Peschle C, Calabretta B: **Overexpression of the zinc finger protein MZF1 inhibits hematopoietic development from embryonic stem cells: correlation with negative regulation of CD34 and c-myb promoter activity.** *Mol Cell Biol* 1995, **15(11):**6075-6087.
132. Tsai EY, Falvo JV, Tsytsykova AV, Barczak AK, Reimold AM, Glimcher LH, Fenton MJ, Gordon DC, Dunn IF, Goldfeld AE: **A lipopolysaccharide-specific enhancer complex involving Ets, Elk-1, Sp1, and CREB binding protein and p300 is recruited to the tumor necrosis factor alpha promoter in vivo.** *Mol Cell Biol* 2000, **20(16):**6084-6094.
133. Maitra S, Atchison M: **BSAP can repress enhancer activity by targeting PU.1 function.** *Mol Cell Biol* 2000, **20(6):**1911-1922.