3154–3164 *Nucleic Acids Research, 2005, Vol. 33, No. 10*
doi:10.1093/nar/gki624

oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes

Shannan J. Ho Sui[1,2], James R. Mortimer[3], David J. Arenillas[1,4], Jochen Brumm[1,5], Christopher J. Walsh[1,2], Brian P. Kennedy[4] and Wyeth W. Wasserman[1,3,*]

[1]Centre for Molecular Medicine and Therapeutics and [2]Genetics Graduate Program, University of British Columbia, Vancouver, BC, Canada, [3]Merck Frosst Centre for Therapeutic Research, Kirkland QC, Canada, [4]Department of Medical Genetics and [5]Department of Statistics, University of British Columbia, Vancouver, BC, Canada

Received April 5, 2005; Revised April 22, 2005; Accepted May 12, 2005

ABSTRACT

Targeted transcript profiling studies can identify sets of co-expressed genes; however, identification of the underlying functional mechanism(s) is a significant challenge. Established methods for the analysis of gene annotations, particularly those based on the Gene Ontology, can identify functional linkages between genes. Similar methods for the identification of over-represented transcription factor binding sites (TFBSs) have been successful in yeast, but extension to human genomics has largely proved ineffective. Creation of a system for the efficient identification of common regulatory mechanisms in a subset of co-expressed human genes promises to break a roadblock in functional genomics research. We have developed an integrated system that searches for evidence of co-regulation by one or more transcription factors (TFs). oPOSSUM combines a precomputed database of conserved TFBSs in human and mouse promoters with statistical methods for identification of sites over-represented in a set of co-expressed genes. The algorithm successfully identified mediating TFs in control sets of tissue-specific genes and in sets of co-expressed genes from three transcript profiling studies. Simulation studies indicate that oPOSSUM produces few false positives using empirically defined thresholds and can tolerate up to 50% noise in a set of co-expressed genes.

INTRODUCTION

DNA microarrays profile patterns of gene expression changes on a genome-wide scale, elucidating sets of genes coordinately expressed under specific conditions. Recent improvements in bioinformatics methods for the analysis of sequences regulating transcription have made it possible to elucidate potential factors involved in key regulatory networks underlying a transcriptional response. The enumeration of such networks, by identifying genes with similar patterns of expression and shared *cis*-regulatory motifs, is crucial to advancing our understanding of biological pathways and processes.

Transcriptional regulation of gene expression is a tightly controlled process that involves the synchronized binding of *trans*-acting transcription factors (TFs) to numerous binding sites in the regions surrounding a gene's transcription start site (TSS), as well as to enhancer regions that mediate gene activation from distal locations. The binding specificities of TFs to their cognate DNA binding motifs are typically modeled using position specific scoring matrices (PSSMs) (1), which are constructed from alignments of binding site sequences that have been characterized experimentally or identified in high-throughput protein–DNA binding assays (2,3). These PSSMs are catalogued in databases such as TRANSFAC (4) and JASPAR (5). The use of PSSMs to detect individual transcription factor binding sites (TFBSs) is well-established (6). However, application of these models typically yields a large number of false positive predictions due to the short, degenerate nature of TFBS motifs. For example, a 6 bp long motif has ~1 in 4000 chance of occurring at random; while tolerance of ambiguity at just one highly variable position can raise the prediction rate to 1 in 1000.

*To whom correspondence should be addressed. Tel: +1 604 875 3812; Fax: +1 604 875 3819; Email: wyeth@cmmt.ubc.ca

Dramatic improvements in the specificity of TFBS prediction are attained by limiting the search space to regions of conserved, non-coding DNA using a comparative genomics approach known as phylogenetic footprinting (7–10). Based on the assumption that functional DNA sequences are subject to greater selective pressure, and therefore, are conserved across moderately diverged organisms, comparison of sequences from orthologous genes can highlight functional, non-coding DNA, providing clues to where regulatory sequences may be located. Phylogenetic footprinting eliminates, on average, 80% of sequence (11), and estimates have placed the proportion of TFBSs occurring within conserved regions when comparing human and mouse sequences at ~70% (11,12). Thus, while the use of phylogenetic footprinting limits our ability to detect binding sites that have evolved in a species-specific manner, the drastic reduction in noise increases specificity, and far outweighs the decrease in sensitivity.

Even with the improved performance conferred by phylogenetic footprinting, most predicted TFBSs are non-functional. By incorporating gene expression data into the analysis procedures, we should improve capacity to discriminate functional binding sites from potential false positive matches. This paper describes a new method, oPOSSUM, which identifies statistically over-represented, conserved TFBSs in the promoters of co-expressed genes. Based on the assumption that some subset of the co-expressed genes is co-regulated by one or more common TFs, we reason that the observed number of binding sites for those TFs should be greater than would be expected by chance. oPOSSUM integrates a pre-computed database of predicted, conserved TFBSs, derived from phylogenetic footprinting and TFBS detection algorithms, with statistical methods for calculating over-representation (Figure 1).

The oPOSSUM system was validated using curated regulatory region collections for genes expressed in a tissue-specific manner, and published targets of the nuclear factor NF-κB. oPOSSUM was then applied to two published transcript profiling data sets, as well as to a new analysis of expression addressing NF-κB inhibition. The results demonstrate that oPOSSUM is able to identify the TFs expected to mediate changes in gene expression through the detection of over-represented TFBSs. Simulations using random sampling gave low false positive rates and revealed tolerance for some noise in the gene sets.

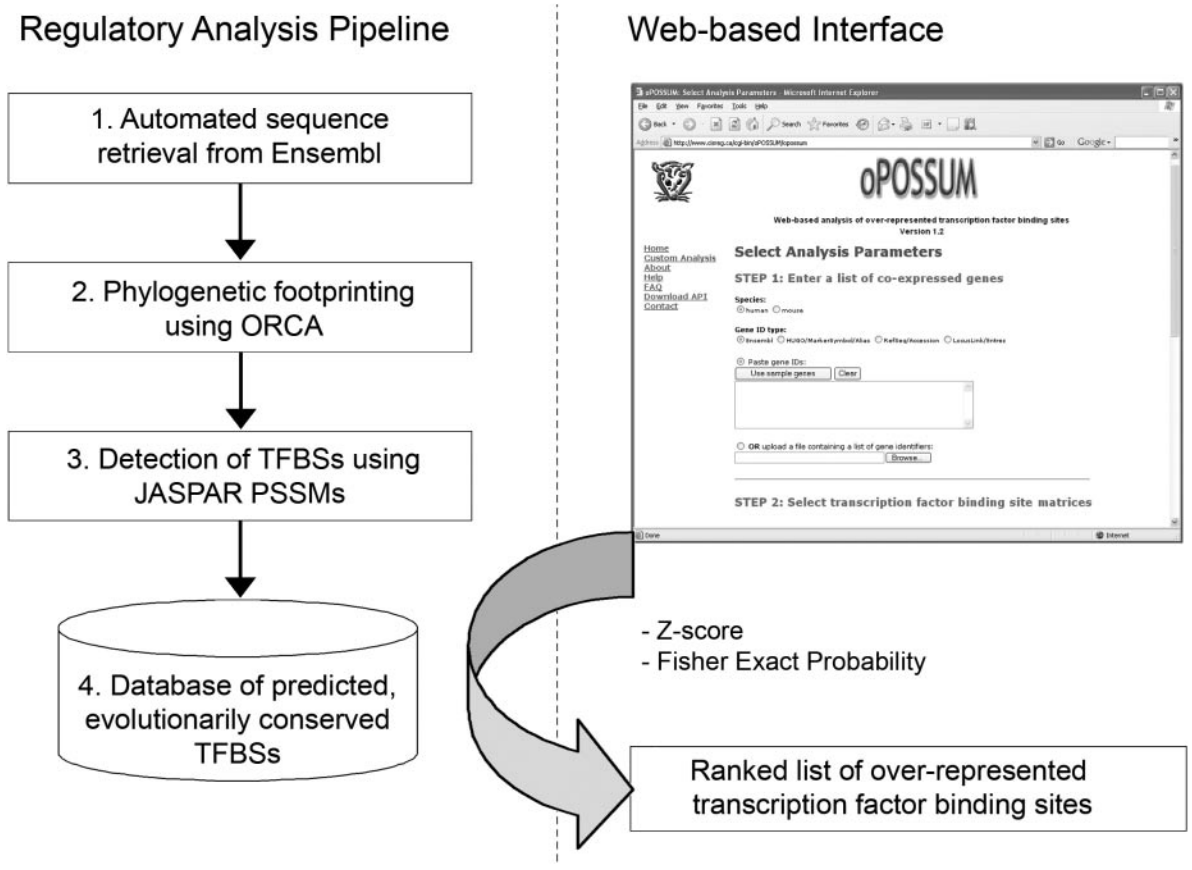

Figure 1. The oPOSSUM system for identifying over-represented TFBSs in sets of co-expressed genes. The system is built upon a database of conserved TFBSs for human–mouse orthologs, derived from an analysis pipeline that combines phylogenetic footprinting with TFBS identification using the JASPAR library of PSSMs. Given a set of human or mouse genes, the pipeline (1) retrieves the genomic DNA sequence for the human and mouse genes plus 5000 bp of upstream sequence, (2) performs an alignment of the orthologous sequences and extracts non-coding DNA subsequences that are conserved above a predefined threshold, (3) searches the subsequences for matches to TFBS profiles contained in JASPAR and (4) stores the results in the oPOSSUM database. Upon querying the web-based interface with a list of co-expressed genes, oPOSSUM retrieves the TFBS counts for each gene in the list and computes two statistics (*Z*-score, Fisher exact test) to measure over-representation of TFBSs in the set relative to a background comprising all genes in the oPOSSUM database.

MATERIALS AND METHODS

Automated retrieval of human–mouse orthologs

The Ensembl software system (13) provides a flexible bioinformatics framework to retrieve sequences and annotations for genes from multiple organisms. Sets of orthologous human and mouse genes are available via EnsMART, a computationally convenient interface to genome annotations. To avoid aligning paralogs (genes which have diverged due to gene duplication in a common ancestor), human genes mapping to more than one mouse gene (and vice versa) are filtered to obtain a set of one-to-one orthologous pairs. For each human–mouse orthologous pair, repeat masked sequences are retrieved encompassing the region 5000 bp upstream of the annotated TSS to either the 3′ end of the gene or, in the case of long genes, 50 000 bp downstream of the TSS. For genes with multiple annotated TSSs, the 5′-most TSS is selected.

Phylogenetic footprinting

Orthologous sequences are aligned using ORCA (D.J. Arenillas and W.W. Wasserman, manuscript under preparation.), a pairwise global progressive alignment algorithm similar to LAGAN (14). ORCA first identifies short segments of high similarity between orthologous genes by performing a local BLASTN alignment using the Bl2Seq algorithm (Version 2.2.5) (15), and then aligns the regions between such segments through the more time-consuming Needleman–Wunsch algorithm (NW) (16) to obtain an overall global alignment of the two sequences. The process is recursive; regions that are too long to align using NW are re-aligned with BLASTN using less stringent parameters. The process comes to a halt when either the regions are short enough to perform NW successfully (the product of the input sequence lengths does not exceed 100 Mb), or the minimum BLASTN word size of seven has been reached. The first iteration of BLASTN was performed using the following parameters: penalty for a nucleotide mismatch = −7; expectation value = 0.10; word size = 15; default values were used for the remaining parameters. For each subsequent run of BLASTN, the nucleotide mismatch score was incremented by two and the word size was decremented by four. NW global alignments used a match score = 3; mismatch score = −1; gap open penalty = 20; and gap extension penalty = 0.

Three dynamically selected and progressively more stringent conservation thresholds are applied. Specifically, each alignment is scanned using a 100 bp sliding window, the percent sequence identity within each window is calculated, and the top 10, 20 and 30% of all windows (excluding those overlapping a coding region) are retained. Minimum identity thresholds of 70, 65 and 60% are required for the high, medium and low conservation levels, respectively. The use of dynamically computed thresholds versus fixed sequence identity cutoffs is motivated by the variable rates of evolution for each gene in each genome.

Detection of TFBSs

The conserved non-coding regions of the promoters are searched for matches to all TFBS profiles in the JASPAR database with information content >8 bits, using the TFBS suite of Perl modules for regulatory sequence analysis (17). Excluding low information content profiles (a measure of the specificity of predictions) eliminates spurious hits. A predicted binding site for a given TF model is reported if the site occurs in both the human and mouse sequences above a threshold PSSM score of 75%, and at equivalent positions in the alignment. Overlapping sites for the same TF are filtered such that only the highest scoring is kept. The location, score, orientation and local sequence conservation level of each TFBS match in the human and mouse genes are stored in the oPOSSUM database.

Discovery of over-represented binding sites

Two statistical measures were calculated to determine which, if any, TFBS were over-represented in the set of promoters for co-expressed genes. These represent two distinct models for counting the occurrences of binding sites.

Z-score calculation for determining TFBS that occur more frequently than expected. The *Z*-score uses a simple binomial distribution model to compare the rate of occurrence of a TFBS in the set of co-expressed genes to the expected rate estimated from the pre-computed background set.

For a given TFBS, let the random variable X denote the number of predicted binding site nucleotides in the conserved non-coding regions of the co-expressed gene set. Let B be the number of predicted binding site nucleotides in the conserved non-coding regions of the background set. Using a binomial model with n events, where n is the total number of nucleotides examined (i.e. the total number of nucleotides in the conserved non-coding regions) from the co-expressed genes, and N is the total number of nucleotides examined from the background gene set, then the expected value of X is $\mu = BC$, where $C = n/N$ (i.e. the ratio of sample sizes). Then, taking $P = B/N$ as the probability of success, the standard deviation is given by $\sigma = \sqrt{nP(1-P)}$.

Now, let x be the observed number of binding site nucleotides in the conserved non-coding regions of the co-expressed genes. By applying the Central Limit Theorem and using the normal approximation to the binomial distribution with a continuity correction, the *Z*-score is calculated as $z = \frac{x-\mu-0.5}{\sigma}$. Thus the *Z*-score indicates a significant difference in the rate of occurrence of sites, and is particularly good for detecting increased prevalence of common sites.

One-tailed Fisher exact probability for determining TFBS that occur in a significant number of the co-expressed genes. In contrast to the *Z*-score, the one-tailed Fisher exact probability compares the proportion of co-expressed genes containing a particular TFBS to the proportion of the background set that contains the site to determine the probability of a non-random association between the co-expressed gene set and the TFBS of interest. It is calculated using the hyper-geometric probability distribution that describes sampling without replacement from a finite population consisting of two types of elements (18). Therefore, the number of times a TFBS occurs in the promoter of an individual gene is disregarded, and instead, the TFBS is considered as either present or absent. A significant value for the Fisher exact probability indicates that there are a significant proportion of genes that contain the site, and is particularly good for rare TFBSs. Fisher exact probabilities were calculated using the *R* Statistics package (http://www.r-project.org).

NF-κB microarray experiment

A list of genes differentially expressed during interruption of the NF-κB pathway by a specific NF-κB inhibitor was obtained from an unpublished microarray experiment. Supplementary Table 1 contains sufficient information to reproduce or challenge the *in silico* promoter analysis described in this text, and the design of the experiment is briefly described here. Human umbilical vein endothelial cells (HUVEC) in the treatment condition were pre-treated with 10 μM of NF-κB inhibitor, followed by stimulation of the NF-κB signaling pathway with 0.1 ng/ml IL-1B 1 h later ($t = 0$ h). A second sample of the same culture was treated with 0.1 ng/ml IL-1B only at $t = 0$. A third sample received only vehicle treatment (0.33% dimethyl sulfoxide) at $t = 0$. From each condition, total RNA was isolated at 6 h using the RNeasy midi kit (Qiagen, USA). The entire paradigm was repeated three times on separate batches of HUVEC, generating nine samples. Equal amounts of RNA were pooled from the three IL-1B treated samples, as the control channel. Each sample, i.e. the three inhibitor treated, the three vehicle treated, and the pool of IL-1B treatment alone, was split in two and labeled with either the Cy3 or Cy5 fluorescent dye (Agilent, USA). Using a two-color microarray system, the labeled cDNA from treatment and control conditions was hybridized to an oligonucleotide microarray representing 23 000 human genes (Agilent, USA) as follows: (i) three replicates of individual vehicle treated samples versus pool of IL-1B samples, (ii) three replicates of pool of IL-1B samples versus individual IL-1B + NF-κB inhibitor treated samples, (iii) same as (i) with fluors reversed and (iv) same as (ii) with fluors reversed. After quantification of the raw data, normalization and combination of the technical, fluor-reversed replicates using the Rosetta Resolver® (version 3.0) gene-expression-data-analysis system (19), an error-weighted ANOVA analysis was performed across replicates in the two groups. Biological replicates were then combined using the Rosetta Resolver® (version 3.0) error model.

For our analysis we focused on a list of 508 sequences showing significantly decreased levels of expression in inhibitor-treated cells, defined by an ANOVA P-value $\leqslant 0.01$, an error-model P-value $\leqslant 0.01$ and a fold-change $\geqslant 1.3$. The 508 sequences were mapped to 326 unique Ensembl gene IDs by identifying gene models from Ensembl V19 (build 34a) which overlapped with the probe sequences. The down-regulated genes were submitted for analysis by oPOSSUM.

Simulations using random sampling

To estimate the false positive rate, we tested oPOSSUM on randomly generated subsets of genes from the oPOSSUM database to determine how frequently TFBSs are identified as over-represented by chance, and to assess the validity of the selected Z-score and Fisher P-value cutoffs of 10 and 0.01, respectively. We created 100 independent sets, each containing 15 genes. These were submitted to oPOSSUM, and the number of TFBSs significantly over-represented during each trial was counted. The number of trials that generated significant TFBSs over 100 independent trials gives us a measure of the false positive rate. We repeated this process for gene lists of 50, 100 and 200 randomly selected genes to see what, if any, effect the number of genes in the list has on the false positive rate.

Next we investigated the amount of noise oPOSSUM can tolerate by adding increasing numbers of randomly selected genes from the oPOSSUM database to our reference gene sets. For the muscle- and liver-specific gene sets, we added 5, 10, 15, 20, 25, 30, 40, 50, 75 and 100 randomly selected genes, and submitted them to oPOSSUM. Additional increments of 150 and 300 genes were tested for the larger set of NF-κB target genes. This process was repeated 100 times for each noise level. The average Z-scores and Fisher P-values for the Mef-2, HNF-1 and NF-κB TFBS profiles over 100 independent trials for each noise level were recorded.

Parameter selection for validation studies

For all of the analyses presented in this study, we examined the promoter region encompassing 5000 bp upstream and 5000 bp downstream of the TSS, used the highest conservation level to extract conserved, non-coding regions (top 10% of conserved regions with a minimum of 70% sequence identity), and required a PSSM score >85% for predicted binding sites, using only the vertebrate-specific PSSMs in JASPAR.

RESULTS

The oPOSSUM database was constructed from an initial set of 14 083 orthologs from human and mouse, obtained by selecting only 'one-to-one' human–mouse orthologs from Ensembl (20). Of these, 4921 (34.9%) of the ortholog sequence pairs failed to produce reasonable alignments of the promoter regions, due largely to an inability to reconcile TSS positions as a result of alternative promoter usage by orthologs, and to a lesser degree, as a consequence of low nucleotide sequence similarity between assigned orthologous gene pairs, genes within genes, and TSSs located within exons of upstream genes on the opposite strand. Attempts to align a subset of the failed promoter pairs using the LAGAN algorithm produced similar results (not shown). An additional 456 (3.2%) ortholog pairs successfully aligned but did not contain conserved, non-coding regions (minimum of 100 bp with >60% identity) in the target region spanning from 5000 bp upstream of the TSS to 5000 bp downstream of the TSS. Of the remaining 8706 genes with conserved promoters, 8698 contained matches to one or more TFBS profiles (PSSM cutoff of 75%), producing 2.4×10^6, 3.3×10^6 and 4.1×10^6 conserved predicted binding sites at the high, medium and low conservation levels, respectively (See Materials and Methods).

Validation using reference gene sets

The muscle and liver regulatory region collections catalogue experimentally verified TFBSs that confer muscle- and liver-specific gene expression, respectively (21,22). We searched the literature for additional experimentally verified sites in human and mouse, adding eight liver-specific and five muscle-specific promoters to these collections (available at http://www.cisreg.ca/tjkwon/). In addition to these tissue-specific genes, we compiled a list of 61 known targets of the nuclear factor NF-κB (23). We used these reference sets to assess oPOSSUM's ability to discriminate functionally relevant TFBSs and to empirically determine appropriate thresholds for our scoring measures.

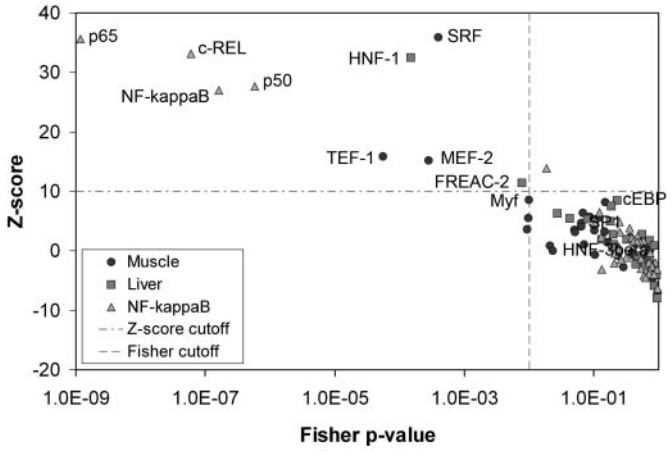

Figure 2. Relationship between the Fisher *P*-values and *Z*-scores for the muscle, liver and NF-κB reference sets. Based on the distribution of scores for the reference sets, a *Z*-score cutoff of 10 and a Fisher *P*-value cutoff of 0.01 were empirically selected as threshold levels to be used for testing. TFBSs that have functional relevance are labeled.

oPOSSUM calculates two statistical measures for binding site over-representation, one at the gene level (Fisher exact test) and the other based on the ratio of TFBSs to nucleotides (*Z*-score). Figure 2 shows the correlation between the scores for each reference set. Clearly, scores for the majority of TFBSs cluster at the bottom right corner of the graph for all reference sets, with *Z*-scores ranging from −10 to 10 and Fisher *P*-values ranging from 0.02 to 1. For each reference set, we also ranked the top 10 binding sites, ordered by *Z*-score, along with associated Fisher *P*-values (Table 1). In each case, the TFs were further investigated for experimentally verified evidence in the given tissue or system.

Muscle-specific regulatory region collection. Studies of skeletal muscle expression have revealed five primary classes of TFs that contribute to skeletal muscle-specific expression: Myf (MyoD), Mef-2, SRF, TEF-1 and Sp-1 (24). Submission of the 25 genes of human, mouse or rat origin in the muscle regulatory collection resulted in 14 pairs of orthologs being analyzed. oPOSSUM ranked SRF, TEF-1, Mef-2 and Myf as the top four most significant profiles (Table 1). In fact, all of these TFs had Fisher *P*-values <0.01 and with the exception of Myf, had *Z*-scores >10, considerably higher than for all other TFs (Figure 2). Sp-1 was ranked tenth but without sufficiently convincing scores to discriminate it from the remainder of the TFBSs (Figure 2); this is not surprising given that it is a ubiquitous activator of numerous genes in the human genome (25).

Liver-specific regulatory region collection. Based on a collection of genes expressed either exclusively in liver hepatocytes or in a small number of tissues including liver hepatocytes, previous studies have found that hepatocyte-specific gene expression can be governed by the combined action of four primary TFs: HNF-1, HNF-3, HNF-4 and c/EBP (26). (There are additional regulatory programs that are controlled independently of these factors in hepatocytes.) Using this established list of 22 genes, we were able to analyze 11 orthologous gene pairs. Predicted HNF-1 sites were the most significantly over-represented TFBSs in the promoters of genes from the liver collection using both the *Z*-score and Fisher measures (Table 1). In fact, with a *Z*-score of 32.5, which is almost three times greater than the next most significant TFBS profile from JASPAR, and a Fisher *P*-value of 1.5×10^{-4}, HNF-1 clearly segregates from the remaining TFBS profiles in this reference set (Figure 2). c/EBP ranked third, but was not sufficiently over-represented to exceed the significance cutoffs of 10 and 0.01 for the *Z*-score and Fisher measures, respectively.

Table 1. Statistically over-represented TFBSs in reference gene sets

	Rank	*Z*-score	Fisher *P*-value
A. Muscle-specific (25 input; 14 analyzed)			
SRF[a]	1	35.93	3.93×10^{-4}
TEF-1[a]	2	15.84	5.48×10^{-5}
MEF2[a]	3	15.26	2.77×10^{-4}
Myf[a]	4	8.585	9.81×10^{-3}
S8	5	8.168	1.49×10^{-1}
Yin-Yang	6	6.396	6.79×10^{-2}
RORalfa-2	7	5.697	8.70×10^{-2}
deltaEF1	8	5.514	9.76×10^{-3}
Nkx	9	5.492	1.17×10^{-1}
SP1[a]	10	4.671	6.34×10^{-2}
B. Liver-specific (22 input; 11 analyzed)			
HNF-1[a]	1	32.46	1.51×10^{-4}
FREAC-2	2	11.46	7.68×10^{-3}
cEBP[a]	3	8.477	2.29×10^{-1}
FREAC-4	4	7.522	1.86×10^{-1}
c-FOS	5	6.286	2.73×10^{-2}
HLF	6	5.454	4.20×10^{-2}
Chop-cEBP	7	5.313	7.93×10^{-2}
SRY	8	5.000	1.78×10^{-1}
Tal1beta-E47S	9	4.338	6.30×10^{-1}
Hen-1	10	3.117	5.19×10^{-1}
C. Known NF-κB targets (61 input; 33 analyzed)			
p65[a]	1	35.60	1.18×10^{-9}
c-REL[a]	2	33.14	5.94×10^{-8}
p50[a]	3	27.62	5.74×10^{-7}
NF-κB[a]	4	27.00	1.62×10^{-7}
SPI-B	5	13.92	1.92×10^{-2}
Irf-2	6	12.88	8.69×10^{-2}
NRF-2	7	6.468	1.22×10^{-1}
Evi-1	8	5.959	2.04×10^{-1}
Elk-1	9	4.912	2.49×10^{-1}
MZF_5–13	10	4.908	1.06×10^{-1}

TFBSs detected by oPOSSUM with the top ten mostly highly ranked *Z*-scores or with Fisher *P*-value < 0.01.

[a]TFs with experimentally-verified sites in the reference sets. The number of genes used as input and the number of genes analyzed by oPOSSUM (i.e. genes that have an unambiguous mouse ortholog) are shown in brackets. See Materials and Methods for how the *Z*-score and Fisher *P*-values were calculated.

Known NF-κB target genes. The NF-κB/Rel family of TFs, which includes RELA (p65), NF-κB1 (p50, p105), NF-κB2 (p52, p100), c-REL and RELB, plays a central role in regulating the immune response (27). oPOSSUM was applied to a set of 61 known NF-κB-regulated genes (23), which include a large number of cytokines and immunoreceptors, and to a lesser extent, antigen presentation proteins, cell adhesion molecules, acute phase proteins, stress response genes and TFs. Of the 61 human genes submitted to oPOSSUM, 33 were mapped to mouse orthologs and subsequently analyzed. The NF-κB, c-REL, p65 and p50 binding sites, which are all members of the NF-κB-family of TFs, ranked as the top four most over-represented TFBSs, using either the Z-score or Fisher *P*-values (Table 1). Figure 2 shows that they were

indeed the only TFBSs with significant scores discriminating them from other sites, with *Z*-scores as high as 35.6 and Fisher *P*-values as low as 1.2×10^{-9}.

Based on the results obtained from the three reference gene sets, we decided empirically to use a *Z*-score cutoff of 10 and Fisher *P*-value cutoff of 0.01 to identify TFBSs for each of our test sets.

Application to transcript profiling data

The reference collections used above are curated sets of genes. In contrast, high-throughput transcript profiling studies typically produce clusters of hundreds of co-expressed genes, of which only a small subset is likely to be co-regulated by a given factor. We assessed oPOSSUM's performance on three sets of genes derived from transcript profiling experiments, and report the results in Table 2. For each set of co-expressed genes, we list the top ten over-represented TFBSs, as determined by the Z-score, as well as any additional TFBSs with significant Fisher *P*-values ($P < 0.01$).

Table 2. Statistically over-represented TF binding sites in gene expression data sets

	TF class	Rank	*Z*-score	Fisher *P*-value	No. genes
A. c-Myc-induced genes (53 input; 30 analyzed)					
Myc-Max[a]	bHLH-ZIP	1	32.41	1.17×10^{-4}	7
ARNT	bHLH	2	23.82	1.56×10^{-4}	12
Max	bHLH-ZIP	3	21.89	7.40×10^{-3}	9
SP1	ZN-finger, C2H2	4	20.90	2.04×10^{-2}	14
USF	bHLH-ZIP	5	17.01	2.39×10^{-2}	10
MZF_1–4	ZN-finger, C2H2	6	14.96	1.35×10^{-1}	20
Staf	ZN-finger, C2H2	7	11.12	7.59×10^{-2}	2
Ahr-ARNT	bHLH	8	10.87	2.05×10^{-1}	12
SAP-1	ETS	9	10.41	2.57×10^{-3}	9
n-MYC	bHLH-ZIP	10	9.821	4.71×10^{-1}	11
B. c-Fos-induced genes (150 input; 98 analyzed)					
c-FOS[a]	bZIP	1	11.01	2.94×10^{-2}	40
CREB	bZIP	2	8.728	2.45×10^{-1}	11
SP1	ZN-finger, C2H2	3	8.015	1.14×10^{-2}	38
E2F	Unknown[b]	4	3.995	1.12×10^{-1}	15
Myc-Max	bHLH-ZIP	5	3.898	3.21×10^{-1}	5
HLF	bZIP	6	3.249	1.84×10^{-1}	10
Pbx	HOMEO	7	2.878	1.38×10^{-1}	6
FREAC-2	FORKHEAD	8	1.763	6.35×10^{-2}	20
HLF	bZIP	9	1.632	3.32×10^{-2}	32
Myc-Max	bHLH-ZIP	10	1.314	5.53×10^{-1}	12
C. Genes downregulated by the NF-κB inhibitor (326 input; 170 analyzed)					
p65[a]	REL	1	27.73	7.78×10^{-11}	46
NF-κB[a]	REL	2	24.11	8.76×10^{-8}	49
c-REL[a]	REL	3	21.31	3.76×10^{-7}	58
p50[a]	REL	4	15.60	9.71×10^{-5}	19
Irf-2	TRP-CLUSTER	5	13.30	1.30×10^{-2}	3
Irf-1	TRP-CLUSTER	6	12.59	1.50×10^{-3}	22
SPI-B	ETS	7	12.45	9.06×10^{-4}	117
FREAC-4	FORKHEAD	8	11.05	2.55×10^{-4}	71
SRY	HMG	9	10.52	2.81×10^{-4}	85
Pbx	HOMEO	10	9.79	8.58×10^{-2}	10
Sox-5	HMG	12	9.00	2.50×10^{-4}	72
cEBP	bZIP	13	8.26	2.63×10^{-4}	44
c-FOS	bZIP	14	7.52	2.70×10^{-3}	71
HFH-2	FORKHEAD	15	7.36	1.68×10^{-3}	46
Nkx	HOMEO	16	6.74	4.57×10^{-3}	106
HNF-3beta	FORKHEAD	28	2.77	4.70×10^{-3}	46
deltaEF1	HMG	40	1.38	1.16×10^{-3}	149

TFBSs detected by oPOSSUM with the top ten mostly highly ranked *Z*-scores or with Fisher *P*-value < 0.01.

[a]TFs over-expressed or inhibited in gene expression studies. The number of genes used as input and the number of genes analyzed by oPOSSUM (i.e. genes that have an unambiguous mouse ortholog) are shown in brackets.

[b]Although E2F is annotated as 'unknown' in the JASPAR database, it is structurally defined as a member of the 'winged helix' class of proteins.

c-Myc SAGE experiment. The c-Myc TF, which dimerizes with the Max protein, is a key regulator of cell proliferation, differentiation and apoptosis (28,29). Using serial analysis of gene expression (SAGE), Menssen and Hermeking (29) identified 216 different SAGE tags corresponding to unique mRNAs that were induced after adenoviral expression of c-Myc in HUVEC. The induction of 53 genes was confirmed using microarray analysis and RT–PCR. We analyzed the 53 genes with oPOSSUM and found that the binding sites of Myc–Max heterodimers are indeed the most significantly over-represented (Table 2); Myc–Max sites were identified in seven of the genes. Matches to the binding profile for homogeneous Max dimers, c-Myc's interacting partner, were also highly over-represented (present in nine genes, giving a high *Z*-score of 21.9). The binding profile for a related protein, n-Myc, ranked amongst the top ten most over-represented profiles.

c-Fos microarray experiment. In a study examining the role of transcriptional repression in oncogenesis, Ordway *et al*. (30) used microarrays to compare the gene expression profile of 208F fibroblasts transformed by c-Fos against the profiles for the parental 208F rat fibroblast cell line. We mapped the list of 252 induced genes to 150 human orthologs, which were submitted to oPOSSUM. As expected, the c-Fos TFBS was ranked as the most over-represented TFBS in the promoters of the induced genes, with a *Z*-score of 11.0 and a Fisher *P*-value of 2.9×10^{-2} (Table 2). c-Fos sites were identified in 40 of the co-expressed genes.

NF-κB microarray experiment. In HUVEC cells, interleukin 1B treatment precipitates an inflammatory response observable as an induction of mRNA expression. This response can be modulated by the inhibition of the NF-κB signaling pathway (31). We assessed oPOSSUM's performance on 326 genes that showed decreased levels of expression in interleukin-1B-stimulated HUVEC cells treated with an NF-κB inhibitor as compared to IL-1B-stimulated HUVEC cells. Binding sites for the NF-κB/Rel family of TFs were the most over-represented (present in ~50 genes) in the inhibitor-modulated genes (Table 2). Other over-represented TFBSs included the immune-related genes Irf-1, Irf-2 and SPI-B.

Specificity assessment

Based on the reference gene sets and expression data, oPOSSUM successfully identifies TFBSs that play a functional role in the regulation of sets of co-expressed genes. In the majority of cases, a *Z*-score >10 and a Fisher *P*-value <0.01 effectively discriminated the known sites within each set of reference genes. To assess how many of the over-represented TFBSs may be expected by chance and ascertain if the qualitatively observed thresholds are appropriate, we tested oPOSSUM on randomly generated subsets of genes from the oPOSSUM database.

In Figure 3, we show the percentage of trials that produced TFBS predictions for random sets of genes, providing a measure of the false positive rate. For a set of 15 genes, using the *Z*-score alone, 23% of the trials produced one false positive prediction, 19% produced two false positives, and so forth, for

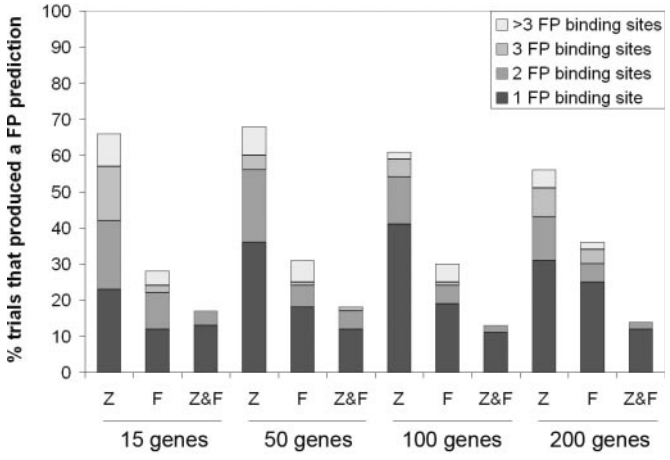

Figure 3. Percentage of trials that produced false positive (FP) predictions. Sets containing 15, 50, 100 and 200 randomly selected genes were generated and submitted to oPOSSUM (100 trials each). Each segment of the bar represents the percentage of trials where *n* TFBSs were over-represented by chance using the *Z*-score and Fisher *P*-value cutoffs. Symbols: *Z* = *Z*-score > 10; *F* = Fisher < 0.01; *Z*&*F* = *Z*-score > 10 and Fisher < 0.01.

an overall false positive rate of 66%. Using only the Fisher exact test for a set of 15 genes, we obtain an overall false positive rate of 28%. Thus, when used in isolation, each of the scoring measures result in surprisingly high false positive rates (average of 63% for the *Z*-score and 31% for the Fisher test), which are dramatically reduced by combining the scores. By applying both the *Z*-score and the Fisher *P*-value cutoffs to the randomly selected sets, we observed an average false positive rate of 15%. The specificity when using the combination of scores (*Z* and *F*) appears consistent across gene sets of different sizes. Thus, with sets as large as 100–200 genes, which is typical of clustered expression data, ~86% of the time no spurious results are observed.

Noise tolerance

Next we performed simulations to investigate the amount of noise oPOSSUM can tolerate. To do this, we added from 5 to 300 randomly selected genes to the reference gene sets, and applied oPOSSUM to determine what proportion of the sets could be noise before losing our ability to elucidate the TFBSs mediating tissue-specific and pathway-specific expression. We considered the Mef-2, HNF-1 and NF-κB binding site profiles to be representative of each set, and plotted their average *Z*-scores and Fisher *P*-values over 100 trials against the proportion of noise in the set. The muscle, liver and NF-κB data sets can tolerate up to 60% of the gene list being noise using the *Z*-score (Figure 4A) and up to 50% using the Fisher *P*-value (Figure 4B). There is significant variation in the degree of noise tolerance amongst the three sets of genes: the NF-κB set is able to tolerate up to 80% of the set being noise versus only 50% for the muscle set. Figure 4 shows that the *Z*-score decreases quadratically and the Fisher *P*-value increases logarithmically with increasing noise for all three sets of genes.

Web implementation

The approach described for the detection of over-represented conserved TFBSs in sets of co-expressed genes has been

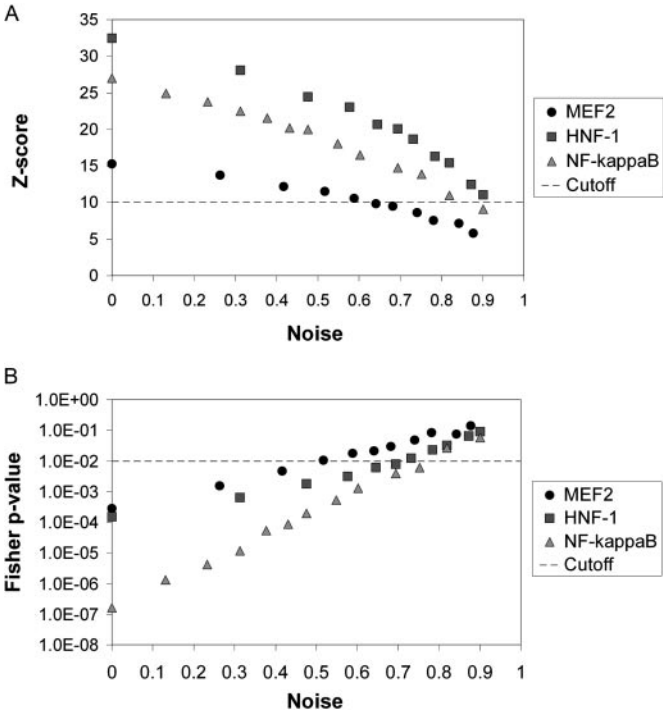

Figure 4. Noise tolerance. Increasing numbers of randomly selected genes were added to the muscle, liver and NF-κB reference sets to assess the effect of noise on (**A**) the *Z*-score and (**B**) Fisher exact probability statistical measures. The amount of noise is represented as the fraction of all genes in the set that were randomly selected. Average *Z*-scores and Fisher *P*-values for MEF2, HNF-1 and NF-κB over 100 trials for each noise level are shown to represent the muscle, liver and NF-κB reference sets, respectively. Suggested cutoffs for the *Z*-score and Fisher *P*-value are shown by the dotted grey lines.

Table 3. Predefined values for phylogenetic footprinting and TFBS detection available in oPOSSUM's default mode

Level	Conservation[a]	PSSM score (%)	Promoter region[b]
1	Top 30th percentile (minimum 60%)	75	−5000 to +5000
2	Top 20th percentile (minimum 65%)	80	−2000 to +2000
3	Top 10th percentile (minimum 70%)	85	−2000 to 0

[a]Conservation thresholds based on percentiles are determined by first calculating the amount of sequence identity for all windows of size 100bp, removing coding regions, and then finding the value above which the top *x*% of scores reside.
[b]Relative to the TSS.

implemented as a flexible, user-friendly website available from www.cisreg.ca. The implementation allows for analysis in default and custom modes. In the default mode, conserved human and mouse TFBS counts have been pre-calculated and stored using combinations of pre-defined values for the following three parameters: (i) the amount of sequence relative to the TSS to be included in the analysis, (ii) the level of interspecies conservation required and (iii) the PSSM score required for a hit to be reported (Table 3). Users simply select a pre-defined set of parameters, select a set of TFBS to be included in the analysis, and submit a list of gene identifiers (Ensembl, GenBank, RefSeq or LocusLink are presently supported) for analysis. oPOSSUM retrieves the TFBS hits matching the specified criteria for each gene in the list,

A

Analysis Results

Selected Parameters

Conservation level:	Top 10.0% of conserved regions (mi
Matrix match score:	80.0%
Upstream sequence length:	5000
Downstream sequence length:	5000
Taxonomic supergroup(s):	vertebrate
Number of genes submitted:	25
Number of genes included:	14
Number of genes excluded:	11

B

MEF2 associated genes:

Ensembl ID	Start	End	Strand	Score
ENSG00000081189	2677	2686	-	0.8201
	1348	1357	+	0.8663
	1221	1230	+	0.8438
	1113	1122	+	0.8123
	980	989	-	0.8413
	-1344	-1335	-	0.8411
ENSG00000092054	-2827	-2818	-	0.8048
ENSG00000104879	-868	-854	+	0.8104
ENSG00000111046	-380	-371	+	0.8377
	-73	-64	-	0.9010

Target Genes

Analyzed: ENSG00000081189 ENSG00000092054 ENSG00000104879 ENSG00000108556 ENSG00000111046 ENSG00000122180 ENSG00000129152 ENSG00000135902 ENSG00000138435 ENSG00000143632 ENSG00000159173 ENSG00000168530 ENSG00000170175 ENSG00000175084

Excluded: ENSG00000007314 ENSG00000100306 ENSG00000109063 ENSG00000114854 ENSG00000125414 ENSG00000132438 ENSG00000135906 ENSG00000149925 ENSG00000159251 ENSG00000166094 ENSG00000170379

One-tailed Fisher Exact Probability Analysis

TF	JASPAR ID	Class	Supergroup	IC	Background gene hits	Background gene non-hits	Target gene hits	Target gene non-hits	Z-score	Fisher P-value
MEF2	MA0052	MADS	vertebrate	15.709	2090	6616	10	4	17.61	2.424e-04
SRF	MA0083	MADS	vertebrate	17.965	294	8412	4	10	26.03	1.020e-03
c-MYB_1	MA0100	TRP-CLUSTER	vertebrate	9.883	4632	4074	13	1	8.143	1.954e-03
MZF_5-13	MA0057	ZN-FINGER, C2H2	vertebrate	9.400	4744	3962	13	1	3.96	2.602e-03
Tal1beta-E47S	MA0091	bHLH	vertebrate	14.070	2030	6676	8	6	6.354	6.659e-03

Z-score Analysis

TF	JASPAR ID	Class	Supergroup	IC	Background TFBS hits	Target TFBS hits	Background TFBS rate	Target TFBS rate	FisherP-value	Z-score
SRF	MA0083	MADS	vertebrate	17.965	325	7	0.0006	0.0057	1.020e-03	26.03
MEF2	MA0052	MADS	vertebrate	15.709	3330	22	0.0048	0.0149	2.424e-04	17.61
deltaEF1	MA0103	ZN-FINGER, C2H2	vertebrate	8.305	60849	188	0.0527	0.0762	6.791e-02	12.74
c-MYB_1	MA0100	TRP-CLUSTER	vertebrate	9.883	11921	40	0.0138	0.0216	1.954e-03	8.143
p50	MA0105	REL	vertebrate	15.627	1521	7	0.0024	0.0052	2.666e-02	6.813

Figure 5. The oPOSSUM result report for the identification of over-represented TFBSs in sets of co-expressed genes. (**A**) Results report showing the selected parameters, genes included and excluded in the analysis, and summary tables containing the Fisher exact probability scores and *Z*-scores for each TFBS (only the first five results are shown for each statistical test in this figure). (**B**) Pop-up window displaying genes that contain a particular TFBS (in this case, MEF2), as well as the site locations and scores.

calculates a Fisher exact probability and *Z*-score for the classes of TFBSs found in the set of genes, and returns ranked lists of TFBSs for each statistical test (Figure 5A). This operation is fast (<30 s for each of the reference sets) due to the pre-calculation of background frequencies. Pop-up windows for each TFBS display the genes in which the site has been located, as well as the site's co-ordinates and score (Figure 5B). Furthermore, the TFBSs are linked to the JASPAR database for easy access to information regarding the binding site profiles.

In the custom mode, users are not restricted to the predefined parameter values for the PSSM score and promoter region, and are given the option to supply user-defined background sets. Users might be motivated to introduce their own background sets if there is prior biological evidence linking sequence composition to expression in the tissue or condition studied. The customization option provides users with more control, and results in more variable processing speeds depending on the size of the background set and the parameters selected.

The oPOSSUM application programming interface (API)

The oPOSSUM API, based on a set of object-oriented Perl modules, provides an interface to the oPOSSUM database and defines data objects for facilitating statistical (Fisher and

Z-score) analysis. A set of modules at the top level of the API tree model each of the data objects in the oPOSSUM database. Briefly, the current version of the API includes modules for connecting and retrieving gene indices, orthologous gene pairs, conserved region information, TFBS matches, and other types of data from the oPOSSUM database, running the *Z*-score and Fisher analyses, and storing the input and output from these analysis modules. The API with accompanying documentation is available through the oPOSSUM website.

DISCUSSION

Regulatory analysis of the promoters of co-expressed genes can give rise to hypotheses about the factors, TFBSs, and putative pathways involved in generating the observed expression patterns. Our integrated approach to regulatory analysis incorporates public data sources, cross-species conservation and complementary statistical methods to identify over-represented motifs. We validate the method using updated reference sets of muscle-specific and liver-specific regulatory regions, and a new set of NF-κB-regulated genes. We demonstrate the utility of this technique for analyzing experimental data with three independent gene expression studies. We show the robustness of this method through computationally assessed rates of false-positives and noise tolerance. The procedure has been implemented as a user-friendly, flexible website called oPOSSUM and as a Perl API. In short, we illustrate herein that oPOSSUM is a novel, validated, useful, robust, user-friendly means for analysts to explore potential regulatory mechanisms in their expression experiments.

Performance

In the case of the muscle regulatory collection, oPOSSUM ranked four of the five documented TFBSs as the four most over-represented sites. In fact, the only three profiles to surpass the specified *Z*-score and Fisher *P*-value cutoffs were those for the muscle-specific TFs SRF, TEF-1 and Mef-2. Similar results were obtained for the extended liver regulatory collection, which contains genes with experimentally verified HNF1/3/4 and c/EBP binding sites. oPOSSUM analysis resulted in two over-represented TFBSs, including the top-ranked HNF-1, followed by forkhead related activator 2 (FREAC-2), a member of the forkhead box family of eukaryotic DNA binding proteins, which includes FREAC-2, FREAC-4, HNF-3β and HNF-4. c/EBP, though not considered significantly over-represented based on our empirical cutoffs, ranked third. The JASPAR database does not currently contain a binding profile for HNF-4, and so this TF could not be included in the analysis. The liver set illustrates how the absence of high-quality PSSM profiles to model all TFs in the human genome represents a key limitation to this method for the entire field.

Application of oPOSSUM to a set of known targets of NF-κB resulted in all four NF-κB-related profiles ranking at the top of the list of over-represented TFBS profiles, with markedly significant scores. Within the ranked list, though not exceeding the thresholds, we observed other immune response-related TFBS profiles. For example, the interferon regulatory factors Irf-1 and Irf-2 (ranked numbers 8 and 6, respectively) are known regulators of the host defense in response to viral infection or cytokine stimulation; they regulate interferon (IFN) and IFN-inducible genes, and also form interactions with SPI-1 and SPI-B (ranked number 5) to induce the activity of various cytokines (32–34). It is worth noting that the default threshold values are simply suggestions and less conservative cutoffs may yield valuable insights as well.

While the system behaved well for the validation collections, we desired to assess the utility for the analysis of larger, more heterogeneous experimental data. In each of the two published gene expression data sets, derived from the ectopic expression of c-Myc and c-Fos respectively, oPOSSUM clearly and appropriately ranked the corresponding TFBSs as being the most significantly over-represented. Further evidence of oPOSSUM's utility in analyzing gene expression data was presented by applying oPOSSUM to a set of genes that showed decreased expression in an experiment examining the effect of a known inhibitor of the NF-κB signaling pathway. This set of genes is distinct from the NF-κB reference set in that the experiment examines an interleukin-induced immune–response in a cellular system, and potentially contains a large number of mRNAs that are independent of NF-κB signaling. Still, oPOSSUM identified the NF-κB binding sites (NF-κB, c-REL, p50 and p65) as being significantly over-represented, as well as the same TFs involved in the immune response that were identified in the NF-κB test set (Irf1, Irf2 and SPI-B). Taken together, the three experimental analyses illustrate the power of promoter sequence analysis to identify the TFs governing gene expression changes observed in heterogeneous microarray and SAGE data.

Challenges

A common problem for promoter analysis is circularity. Binding sites that have been experimentally verified in genes are used to construct binding site profiles, which in turn, are used to search for binding sites in sets containing the original genes. In this study, the Mef-2, SRF, c-REL, p50, p65, Myc–Max and c-Fos binding site profiles were constructed based on SELEX experiments, in which *in vitro* binding experiments are used to isolate suitable binding sites for a particular TF from random oligonucleotides (35). Thus, we can be sure that at least for the SELEX-based profiles, we have avoided any circularity.

At present, a key limitation to the oPOSSUM analysis is the scarcity of annotated binding site profiles. JASPAR, the underlying database supporting oPOSSUM, contains 111 high-quality binding site profiles representing 25 structural classes. When analyzing expression data sets, it is worth keeping in mind that although the TF mediating the observed response may not be present in the sparse JASPAR database, it is possible that a TF that recognizes a similar motif may be identified as being over-represented. For example, looking at the results for the c-Myc experiment in Table 2, it is evident that the over-represented TFBSs we observe are predominately bound by TFs containing the basic helix–loop–helix (bHLH) domain, and in particular, by TFs within the bHLH-ZIP (basic helix–loop–helix/leucine zipper) structural class. In fact, four of the five TFs in JASPAR that belong to the bHLH-ZIP class rank amongst the top ten profiles. This is also true for the NF-κB-related data sets where we see a clear over-representation of TFs belonging to the Rel class of TFs (Tables 1 and 2). It is important to consider whether a match

may be indicative of a member of a structural class, rather than the specific profiled TF. A major challenge, however, is that zinc-finger proteins make up the largest class of TF proteins, comprising ~47% of the estimated 1445 TFs identified in mammalian genomes (36). Cys_2–His_2 zinc-fingers are the most versatile of the DNA-recognition domains, and variations in amino acid sequence enable them to bind to a diverse range of DNA sequences. In addition, zinc-finger proteins in mammalian genomes use multiple, tandem fingers to interact with arrays of subsites, providing a degree of modularity and exceptional adaptability (37). JASPAR currently contains only 17 zinc-finger binding profiles. We will continue our ongoing efforts to expand the JASPAR collection and incorporate new information as it becomes available.

Analysis of the false positive rates using random sampling revealed that, when used in isolation, the Z-score and Fisher tests result in high false positive rates that can be reduced by combining the two scoring measures. While it's true that applying a multiple testing correction could possibly improve performance for the Fisher measure, the *Z*-scores we obtain are extremely large, such that a correction for the 100 or so TFBS profiles being tested has negligible impact on the *Z*-score results. Instead, we have opted to empirically derive threshold cutoffs based on our reference data. Furthermore, our experience with oPOSSUM suggests that it is the ranks of the binding site profiles rather than the specific values of the scores that are indicative of functionally relevant TFBSs. For these reasons, and in light of the binding similarities within factor families, we have abstained from making a Bonferroni correction to adjust for multiple testing. In the future, we may introduce an option for users to make this adjustment that is based on an improved statistical model.

oPOSSUM is the first integrated, web-based tool for analyzing sets of co-expressed genes that incorporates cross-species comparisons, PSSM-based promoter motif detection, and statistical methods for the identification of over-represented TFBSs with a pre-computed database. Other resources are available for detecting and visualizing binding sites within the conserved regions of human genes [Consite (11), rVISTA (38), dbTSS (39), CONREAL (40), CORG (41)], as well as for identifying statistically over-represented motifs in the promoters of related sequences [Clover (42), OTFBS (43), PRIMA (44)]. Other comparable tools that integrate all of these approaches include the Toucan workbench for regulatory sequence analysis (45), CONFAC (46), and CRÈME (47). Unlike Toucan, oPOSSUM employs a pre-computed database of conserved TFBSs, eliminating the need for long processing times involved in retrieving sequences, performing alignments, and detecting motifs via PSSMs. Furthermore, the use of two complementary statistical tests to determine over-represented TFBSs is unique to oPOSSUM, and attempts to address the inherent problems involved in analyzing conserved regions of promoters for TFBSs, which include variation in conservation properties from one orthologous gene pair to another and multiple occurrences of a particular TFBS in the promoter of a single gene.

The oPOSSUM system is under continued development. As new information accumulates, we intend to expand the orthology mapping, increase the number of TFBS profiles supported in JASPAR, include the option for users to specify alternative promoters (TSSs), and improve the over-representation analysis. We believe that this approach to regulatory analysis will be helpful to researchers hoping to elucidate transcriptional pathways from gene expression data.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

Thank you to Yves Boie, Ernest Asante-Appiah, Jimmy Fourtounis and Chris Roberts (Merck) for supplying the NF-κB microarray data. This work was supported by the CIHR/MSF Strategic Training Program in Bioinformatics, the National Science and Engineering Research Council of Canada (NSERC) and Merck-Frosst, as well as a grant from the Canadian Institutes of Health Research (WWW). Funding to pay the Open Access publication charges for this article was provided by Merck-Frosst research funding to the CMMT (WWW).

Conflict of interest statement. None declared.

REFERENCES

1. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
2. Pollock,R. and Treisman,R. (1990) A sensitive method for the determination of protein–DNA binding specificities. *Nucleic Acids Res.*, **18**, 6197–6204.
3. Bulyk,M.L., Gentalen,E., Lockhart,D.J. and Church,G.M. (1999) Quantifying DNA–protein interactions by double-stranded DNA arrays. *Nat. Biotechnol.*, **17**, 573–577.
4. Wingender,E., Dietze,P., Karas,H. and Knuppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.
5. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
6. Wasserman,W.W. and Krivan,W. (2003) *In silico* identification of metazoan transcriptional regulatory regions. *Naturwissenschaften*, **90**, 156–166.
7. Koop,B.F. (1995) Human and rodent DNA sequence comparisons: a mosaic model of genomic evolution. *Trends Genet.*, **11**, 367–371.
8. Hardison,R.C., Oeltjen,J. and Miller,W. (1997) Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.*, **7**, 959–966.
9. Duret,L. and Bucher,P. (1997) Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.*, **7**, 399–406.
10. Wasserman,W.W., Palumbo,M., Thompson,W., Fickett,J.W. and Lawrence,C.E. (2000) Human-mouse genome comparisons to locate regulatory sites. *Nature Genet.*, **26**, 225–228.
11. Lenhard,B., Sandelin,A., Mendoza,L., Engstrom,P., Jareborg,N. and Wasserman,W.W. (2003) Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.*, **2**, 13.
12. Dermitzakis,E.T. and Clark,A.G. (2002) Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.*, **19**, 1114–1121.
13. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, **31**, 38–42.
14. Brudno,M., Do,C.B., Cooper,G.M., Kim,M.F., Davydov,E., Green,E.D., Sidow,A. and Batzoglou,S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.*, **13**, 721–731.
15. Tatusova,T.A. and Madden,T.L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.*, **174**, 247–250.

16. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
17. Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.
18. Fleiss,J.L. (1981) *Statistical methods for Rates and Proportions*. John Wiley, New York.
19. Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
20. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, **31**, 38–42.
21. Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.
22. Krivan,W. and Wasserman,W.W. (2001) A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.*, **11**, 1559–1566.
23. Baeuerle,P.A. and Baichwal,V.R. (1997) NF-kappa B as a frequent target for immunosuppressive and anti-inflammatory molecules. *Adv. Immunol.*, **65**, 111–137.
24. Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.
25. Emami,K.H., Burke,T.W. and Smale,S.T. (1998) Sp1 activation of a TATA-less promoter requires a species-specific interaction involving transcription factor IID. *Nucleic Acids Res.*, **26**, 839–846.
26. Krivan,W. and Wasserman,W.W. (2001) A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.*, **11**, 1559–1566.
27. Li,Q. and Verma,I.M. (2002) NF-kappaB regulation in the immune system. *Nature Rev. Immunol.*, **2**, 725–734.
28. Amati,B., Dalton,S., Brooks,M.W., Littlewood,T.D., Evan,G.I. and Land,H. (1992) Transcriptional activation by the human c-Myc oncoprotein in yeast requires interaction with Max. *Nature*, **359**, 423–426.
29. Menssen,A. and Hermeking,H. (2002) Characterization of the c-MYC-regulated transcriptome by SAGE: identification and analysis of c-MYC target genes. *Proc. Natl Acad. Sci. USA*, **99**, 6274–6279.
30. Ordway,J.M., Williams,K. and Curran,T. (2004) Transcription repression in oncogenic transformation: common targets of epigenetic repression in cells transformed by Fos, Ras or Dnmt1. *Oncogene*, **23**, 3737–3748.
31. Epinat,J.C. and Gilmore,T.D. (1999) Diverse agents act at multiple levels to inhibit the Rel/NF-kappaB signal transduction pathway. *Oncogene*, **18**, 6896–6909.
32. Marecki,S., Riendeau,C.J., Liang,M.D. and Fenton,M.J. (2001) PU.1 and multiple IFN regulatory factor proteins synergize to mediate transcriptional activation of the human IL-1 beta gene. *J. Immunol.*, **166**, 6829–6838.
33. Meraro,D., Gleit-Kielmanowicz,M., Hauser,H. and Levi,B.Z. (2002) IFN-stimulated gene 15 is synergistically activated through interactions between the myelocyte/lymphocyte-specific transcription factors, PU.1, IFN regulatory factor-8/IFN consensus sequence binding protein, and IFN regulatory factor-4: characterization of a new subtype of IFN-stimulated response element. *J. Immunol.*, **168**, 6224–6231.
34. Taniguchi,T., Ogasawara,K., Takaoka,A. and Tanaka,N. (2001) IRF family of transcription factors as regulators of host defense. *Annu. Rev. Immunol.*, **19**, 623–655.
35. Pollock,R. and Treisman,R. (1990) A sensitive method for the determination of protein–DNA binding specificities. *Nucleic Acids Res.*, **18**, 6197–6204.
36. Gray,P.A., Fu,H., Luo,P., Zhao,Q., Yu,J., Ferrari,A., Tenzen,T., Yuk,D.I., Tsung,E.F., Cai,Z. *et al.* (2004) Mouse brain organization revealed through direct genome-scale TF expression analysis. *Science*, **306**, 2255–2257.
37. Urnov,F.D. and Rebar,E.J. (2002) Designed transcription factors as tools for therapeutics and functional genomics. *Biochem. Pharmacol.*, **64**, 919–923.
38. Loots,G.G., Ovcharenko,I., Pachter,L., Dubchak,I. and Rubin,E.M. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.*, **12**, 832–839.
39. Suzuki,Y., Yamashita,R., Nakai,K. and Sugano,S. (2002) DBTSS: dataBase of human transcriptional start sites and full-length cDNAs. *Nucleic Acids Res.*, **30**, 328–331.
40. Berezikov,E., Guryev,V., Plasterk,R.H. and Cuppen,E. (2004) CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome Res.*, **14**, 170–178.
41. Dieterich,C., Wang,H., Rateitschak,K., Luz,H. and Vingron,M. (2003) CORG: a database for comparative regulatory genomics. *Nucleic Acids Res.*, **31**, 55–57.
42. Frith,M.C., Fu,Y., Yu,L., Chen,J.F., Hansen,U. and Weng,Z. (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res.*, **32**, 1372–1381.
43. Zheng,J., Wu,J. and Sun,Z. (2003) An approach to identify over-represented *cis*-elements in related sequences. *Nucleic Acids Res.*, **31**, 1995–2005.
44. Elkon,R., Linhart,C., Sharan,R., Shamir,R. and Shiloh,Y. (2003) Genome-wide *in silico* identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res.*, **13**, 773–780.
45. Aerts,S., Thijs,G., Coessens,B., Staes,M., Moreau,Y. and De Moor,B. (2003) Toucan: deciphering the *cis*-regulatory logic of coregulated genes. *Nucleic Acids Res.*, **31**, 1753–1764.
46. Karanam,S. and Moreno,C.S. (2004) CONFAC: automated application of comparative genomic promoter analysis to DNA microarray datasets. *Nucleic Acids Res.*, **32**, W475–W484.
47. Sharan,R., Ovcharenko,I., Ben Hur,A. and Karp,R.M. (2003) CREME: a framework for identifying *cis*-regulatory modules in human–mouse conserved segments. *Bioinformatics.*, **19** (Suppl. 1), i283–i291.