IDENTIFICATION OF OVER-REPRESENTED COMBINATIONS OF TRANSCRIPTION FACTOR BINDING SITES IN SETS OF CO-EXPRESSED GENES

SHAO-SHAN HUANG,[1,2,*] DEBRA L. FULTON,[1,2,*] DAVID J. ARENILLAS,[1,2,3]
PAUL PERCO,[4] SHANNAN J. HO SUI,[1,2] JAMES R. MORTIMER[5] AND
WYETH W. WASSERMAN[1,2,3,#]

[1] *Centre for Molecular Medicine and Therapeutics,*
[2] *Child and Family Research Institute,*
[3] *Department of Medical Genetics,*
University of British Columbia, Vancouver, Canada
[4] *Department of Nephrology, Medical University of Vienna, Vienna, Austria*
[5] *Merck Frosst Centre for Therapeutic Research, Kirkland QC, Canada*
[*] *These authors contributed equally to this work.*
[#] *Corresponding author. E-mail: wyeth@cmmt.ubc.ca*

Transcription regulation is mediated by combinatorial interactions between diverse trans-acting proteins and arrays of cis-regulatory sequences. Revealing this complex interplay between transcription factors and binding sites remains a fundamental problem for understanding the flow of genetic information. The oPOSSUM analysis system facilitates the interpretation of gene expression data through the analysis of transcription factor binding sites shared by sets of co-expressed genes. The system is based on cross-species sequence comparisons for phylogenetic footprinting and motif models for binding site prediction. We introduce a new set of analysis algorithms for the study of the combinatorial properties of transcription factor binding sites shared by sets of co-expressed genes. The new methods circumvent computational challenges through an applied focus on families of transcription factors with similar binding properties. The algorithm accurately identifies combinations of binding sites over-represented in reference collections and clarifies the results obtained by existing methods for the study of isolated binding sites.

1. Introduction

The interaction between transcription factor (TF) proteins and transcription factor binding sites (TFBS) is an important mechanism in regulating gene expression. Each cell in the human body expresses genes in response to its developmental state (e.g. tissue type), external signals from neighboring cells and environmental stimuli (e.g. stress, nutrients). Diverse regulatory mechanisms have evolved to facilitate the programming of gene expression, with a primary mechanism being TF-mediated modulation of the rate of transcript initiation. Given a finite collection of protein structures capable of binding to specific DNA sequences and the diversity of conditions to which cells must respond, it is logical and well-documented that combinatorial interplay between TFs drives much of the observed specificity of gene expression. The arrays of TFBS at which the interactions occur are often termed cis-regulatory modules (CRM).[1]

The sequence specificity of TFs has stimulated development of computational methods for discovery of TFBS on DNA sequences. Well established methods represent aligned collections of TFBS as position weight matrices (PWM). Sequence specificity of individual PWM profiles can be quantified by information content, and scoring a sequence against the PWM of a TF gives a quantitative measure of the sequence's similarity to the binding profile (for review see Wasserman and Sandelin[16]). Searching for high scoring motifs in putative regulatory sequences with a collection of profiles (for instance, JASPAR[10]) can suggest the binding sites in the sequence and the associated TF. However, this methodology is plagued by poor specificity due to the short and variable nature of the TFBS. Phylogenetic footprinting filters have been demonstrated repeatedly to improve specificity.[6] Such filters are justified by the hypothesis that sequences of biological importance are under higher selective pressure and will thus accumulate DNA sequence changes at a slower rate than other sequences. Based on this expectation, the search for potential TFBS can be limited to the most similar non-coding regions of aligned orthologous gene sequences from species of suitable evolutionary distance. Further, one might expect that genes which are coordinately expressed are under the control of the same TFs, suggesting that over-represented TFBS in the co-expressed genes are likely to be functional. These concepts are implemented by Ho Sui *et al.* in the web service tool oPOSSUM,[3] which, when given a set of co-expressed genes, can identify the TFBS motifs that are over-represented with respect to a background set of genes. This approach has achieved success in finding binding sites known to contribute to the regulation of reference gene sets.

Prior methods that attempt to address the known interplay between TFs at CRMs can be difficult to interpret.[2,5,12] We introduce a new approach rooted in the biochemical properties of TFs, which allows greater computational efficiency and improved interpretation of results. The resulting method is assessed against diverse reference data to demonstrate its utility for the applied analysis of gene expression data. Supplementary information is available at http://www.cisreg.ca/oPOSSUM2/supplement/.

2. Methods

2.1. *Background: the oPOSSUM database*

Ho Sui *et al.*[3] describe the creation of the oPOSSUM database which stores predicted, evolutionarily conserved TFBS to support over-representation analysis of TFBS for single TFs. Briefly, human-mouse orthologs are retrieved from Ensembl. TFBS profiles from the JASPAR database are used to identify putative TFBS within the conserved non-coding regions from 5000 base pairs (bp) upstream to 5000 bp downstream of the annotated transcription start site (TSS) on both strands. The oPOSSUM database stores the start and end positions and the matrix match score ($>$ 70 %) of each site. This data is used by the oPOSSUM II algorithm in searching for over-represented TFBS combinations (described below).

2.2. *Overview and rationale of oPOSSUM II algorithm*

Finding over-represented combinations of TFBS presents several new issues that are not encountered in single site analysis. We address two of the main challenges: computational complexity and TFBS class redundancy. Firstly, the number of possible combinations of size n from m TFBS ($n \leq m$) increases combinatorially with respect to both m and n, which greatly impacts computing time. Secondly, several TFs have similar binding properties, thus subsets of profiles may be effectively redundant. Consequently, an exhaustive search is not an efficient method to find over-represented combinations of patterns.

To address both problems we introduced two approaches. First, we used a novel method to group the profiles into classes. Rather than using protein sequence similarity, a hierarchical clustering procedure was applied to group the profiles into classes according to their quantitative similarity. One representative member was selected from each class for further analysis. We then searched for the occurrences of class combinations in both co-regulated genes (foreground) and a set of background genes. We considered unordered combinations and applied an inter-binding site distance (IBSD) constraint to avoid exhaustive enumeration of all combinations, since many co-operative TFBS are found to occur in clusters without strict ordering constraints.[1] Thus, we only need consider each set of TFBS where all IBSDs satisfy the distance parameter. This approach can dramatically reduce the search space when evaluating any combination size. A scoring scheme was adopted from the Fisher exact test to compare the degree of over-representation of the class combinations. The highly over-represented class combinations were re-assessed using all possible profile combinations within the indicated classes.

The overall scheme of oPOSSUM II analysis is shown in Figure 1. The sections below describe the details of each step.

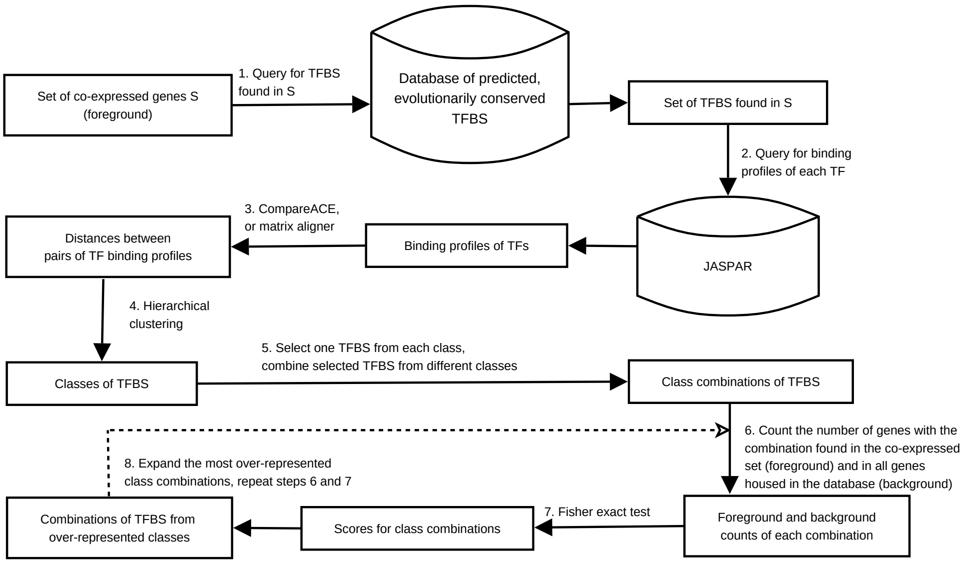
Set of co-expressed genes S (foreground)
1. Query for TFBS found in S
Database of predicted, evolutionarily conserved TFBS
Set of TFBS found in S
2. Query for binding profiles of each TF
JASPAR
Binding profiles of TFs
3. CompareACE, or matrix aligner
Distances between pairs of TF binding profiles
4. Hierarchical clustering
Classes of TFBS
5. Select one TFBS from each class, combine selected TFBS from different classes
Class combinations of TFBS
6. Count the number of genes with the combination found in the co-expressed set (foreground) and in all genes housed in the database (background)
Foreground and background counts of each combination
7. Fisher exact test
Scores for class combinations
Combinations of TFBS from over-represented classes
8. Expand the most over-represented class combinations, repeat steps 6 and 7

Figure 1. Overview of the oPOSSUM II analysis algorithm. Steps are numbered in the order executed. The database of predicted TFBS is identical to that of the oPOSSUM analysis system (Ho Sui *et al.*[3]).

2.3. *TFBS in foreground gene set*

When presented with a set of co-expressed genes S, oPOSSUM II queries the oPOSSUM database for all putative TFBS T present in S within a maximum of 5000 bp upstream and 5000 bp downstream from the TSS on each gene. The analysis may be restricted to those TFs found in selected taxonomic subgroups (plant, vertebrate and insect are currently available), or TFs whose profiles exceed a minimum information content.

2.4. *Classification of TFBS profiles*

Binding profiles for T were retrieved from the JASPAR database. A profile comparison algorithm, either CompareACE[4] (default) or matrix aligner,[11] calculated the pairwise similarity scores of all the profiles using profile alignment methods. The similarity score $s(t_i, t_j)$ between profiles t_i and t_j was converted to distance $d(t_i, t_j)$ by $d(t_i, t_j) = 1 - s(t_i, t_j)$. A distance matrix M was formed from these pairwise distances. From M, an agglomerative clustering procedure produced a hierarchy of clusters (subsets) of T. The complete linkage method was used since it tends to find cohesive classes. Cutting the cluster tree at a specified height thr_H partitioned T into classes.

2.5. *Selection of TFBS and enumeration of combinations*

For each class C, we selected the profile that is the most similar to other profiles in C as the class representative. We chose this approach as we could not identify an adequate procedure that would generate a consensus profile with comparable specificity to the matrices within the class. To identify the class representative, we first calculated the sum of pairwise similarity score σ_i between a profile t_i and other profiles in C, i.e., $\sigma_i = \sum_{t_i, t_j \in C} s(t_i, t_j)$. The profile with the maximum sum of similarity score was chosen. From the selected TFBS, unordered combinations of specified size (*cardinality*) were created. oPOSSUM II then searched the foreground gene set (the co-expressed genes) and the background gene set (default is all the genes in the database) for occurrences of these combinations. Let max_d be the maximum inter-binding site distance. For each gene, the occurrences of the combinations were found using a sliding window of width equal to max_d within the required search region. We counted the number of genes with a combination in both the foreground and background gene sets.

2.6. *Scoring of combinations*

The Fisher exact test detects non-random association between two categorical variables. We adopted the Fisher P-values to rank the significance of non-random association between the occurrence of a combination and the foreground gene set, i.e., over-representation of the combination in the foreground compared to background. For each combination, a two-dimensional contingency table was constructed from the foreground and background count distributions:

	Number of genes with a given combination	Number of genes without a given combination
Foreground	a_{11}	a_{12}
Background	a_{21}	a_{22}

For $i, j = 1, 2$, row sum $R_i = a_{i1} + a_{i2}$ and column sum $C_j = a_{1j} + a_{2j}$, and the total count $N = \sum_i R_i = \sum_j C_j$. From the hypergeometric probability function, the conditional probability P_{cutoff} given the row and column sums is

$$P_{\text{cutoff}} = \frac{(C_1!C_2!)(R_1!R_2!)}{N! \prod_{i,j=1,2} a_{ij}}.$$

We calculated the P-values for all other possible contingency tables with row sums equal to R_i and column sums equal to C_j. The Fisher P-value is the sum of all the P-values less than or equal to P_{cutoff}, which represent equal or greater deviation from independence than the observed table.

Caution must be taken when interpreting these Fisher P-values. First, the foreground and background genes are allowed to overlap, which is a violation of an assumption for the statistical test. Secondly, the Fisher exact test model may not precisely characterize the data sets being analyzed. As a result, the Fisher P-values were used purely as a measure to compare the degree of over-representation between different combinations. We will hereafter refer to the P-values as “scores”. Although the scores do not describe the probabilistic nature of the over-representation, the ranking they provide is shown to be useful.[3]

2.7. *Finding significant TFs from over-represented class combinations*

Let thr_C be the maximum score for which a TFBS combination may be considered significant. Our empirical studies of reference collections suggested that a default maximum score value of 0.01 detects relevant TF combinations. Let x_i be any TFBS class combination with a score less than or equal to thr_C and X is the set of distinct class combinations that satisfy the score threshold: $X = \{x_i | score(x_i) \leq thr_C\}$. For each combination x_i, let each of $C_1,\ C_2,\ \ldots, C_h$ be a set of TFBS profiles that are represented by each of the h class profiles in that combination. Compute the Cartesian product $\mathbb{C}_p$ of $C_1, \ldots, C_h$. We call this “expanding the TFBS classes” from the class representatives. The enumeration and ranking procedures were repeated for the h-tuples in $\mathbb{C}_p$.

2.8. *Random sampling simulations of foreground genes*

oPOSSUM II needs to accommodate input gene sets of different cardinalities, so we wished to investigate the relationship between gene set size and the false positive rate. 100 random samples of r genes were selected from the background and given to oPOSSUM II as foreground genes. For each sample, oPOSSUM II reported the scores for all the class combinations. As these random samples of genes were not expected to be co-regulated, any combination was a false positive. Let $(0, max_s]$ be the interval over which false positives are accumulated. We recorded the number of false positive class combinations for a range of max_s when $r = 20, 40, 60, 80, 100$.

2.9. *Validation*

Three reference sets of human genes were used as input to oPOSSUM II to assess the performance of the algorithm. Two independent sets of skeletal muscle genes were tested. The first set (muscle set 1) was compiled from the reference collection identified by Wasserman and Fickett[15] and updated by a review of recent literature. A second set (muscle set 2) combines the results of microarray studies of Moran *et al.*[8] and Tomczak *et al.*[14] The third set contains smooth muscle-specific genes experimentally verified by Nelander *et al.*[9] All sets were validated with max_d=100, matrix score threshold=75%, and conservation level=1.

As a further comparison to the methods in Kreiman,[5] which were validated in part against the yeast CLB2 gene cluster,[13] the yeast CLB2 cluster was analyzed using the yeast oPOSSUM database (Ho Sui, unpublished).

3. Results

3.1. *TFBS classification*

Since the three reference gene sets were restricted to vertebrates, the first step in oPOSSUM II analysis was to cluster the available vertebrate TFBS. We cut the hierarchical cluster tree at a height of 0.45 ($thr_H = 0.45$) because the majority of resulting clusters correlated well with the structural families defined in JASPAR (cluster tree available in web supplement). Most notably, binding profiles from FORKHEAD, HMG and ETS families were grouped according to classifications. However, as we expected, the zinc finger profiles were dispersed into new groupings due to their divergent binding profile composition. Using this approach, the 68 vertebrate TFBS in JASPAR were partitioned into 32 classes. This step produced a considerable reduction in the search space. For example, in the analysis of pair combinations, the search space was reduced by a factor of four.

3.2. *Validation with reference data sets*

3.2.1. *Yeast CLB2 cluster*

The yeast CLB2 gene cluster contains genes whose transcription peaks at late G2/early M phase of the cell cycle. Transcription of these genes is regulated by the TF FKH, which is a component of the TF SFF, and which interacts with the TF MCM1. Each of the top ten scoring class combinations found by oPOSSUM II contained the binding sites of the ECB class, of which MCM1 is a member. The highest ranked combination was $\{$ECB, FKH1$\}$, which is consistent with the literature and the results of Kreiman.[5] The complete results are available on the supplementary web site.

3.2.2. *Three human reference gene sets*

Figure 2 lists the top five over-represented class combinations for each of the three human reference gene sets. The score values for these combinations were less than 2.0E-3. Also listed are the five most over-represented TFBS classes in the total 32 classes created, as reported by oPOSSUM single site analysis.

Prior studies involving muscle set 1[15] have identified the occurrence of clusters of muscle regulatory sites including MEF2, SRF, Myf/MyoD, SP1 and TEF. The classes that contain MEF2 and SP1 dominated the top combinations in both skeletal muscle sets (Figure 2a and 2b). Yin-Yang modulates SRF-dependent, skeletal muscle expression. Thing1-E47 is a bHLH TF localized to gut smooth muscle in adult mice, therefore, the presence of class

Combination TFBS Pairs	Single TFBS
8 (Bsap); 20 (MEF2*)	1 (Myf*)
8 (Bsap); 29 (SRF*)	8 (Bsap)
1 (Myf*); 31 (Yin-Yang*)	29 (SRF*)
20 (MEF2*); 28 (SP1*)	26 (RREB-1)
20 (MEF2*); 29 (SRF*)	28 (SP1*)

a. Skeletal Muscle Set 1

Combination TFBS Pairs	Single TFBS
20 (MEF2*); 28 (SP1*)	20 (MEF2*)
20 (MEF2*); 32 (Thing1-E47*)	25 (Androgen)
20 (MEF2*); 21 (MZF_5-13)	29 (SRF*)
8 (Bsap); 20 (MEF2*)	1 (Myf*)
1 (Myf*); 20 (MEF2*)	7 (Spz1)

b. Skeletal Muscle Set 2

Combination TFBS Pairs	Single TFBS
28 (SP1*); 29 (SRF*)	29 (SRF*)
21 (MZF_5-13); 29 (SRF*)	26 (RREB-1)
29 (SRF*); 31 (Yin-Yang*)	20 (MEF2*)
29 (SRF*); 7 (Spz1)	7 (Spz1)
29 (SRF*); 32 (Thing1-E47*)	1 (Myf*)

c. Smooth Muscle Set

Figure 2. The top five over-represented pair combinations of TFBS classes reported by oPOSSUM II and over-represented single TFBS sites reported by oPOSSUM for the skeletal and smooth muscle sets. The numbers are the class identifiers and enclosed in parentheses is the name of a TF within that class, which is either known to mediate transcription in the assessed tissue (*) or is a class representative.

32 in the list may be linked to other myogenic factors in the bHLH superfamily (such as Myf). Bsap and MZF are not muscle specific. The Bsap motif is long (20 bp) and exhibits an unusual pattern of low information content distributed across the entire motif, suggesting that it may behave differently than other binding profiles. The inclusion of this profile in the JASPAR database is under review (B. Lenhard, personal communication).

For the smooth muscle genes, the SRF class appeared in each of the top five combinations, consistent with established knowledge.[7] The top combination, $\{$SP1, SRF$\}$, is required for the expression of smooth muscle myosin heavy chain in rat. Yin-Yang can stimulate smooth muscle growth. Spz1 acts in spermatogenesis, and has no known role in muscle expression.

For all three reference sets, the top scoring combinations suggested new classes not found in the single site analysis. In all cases, there were relevant TFBS identified only in the combination analysis.

3.3. *Effect of set size on false positive rate*

The result of random sampling simulation of foreground genes is shown in Figure 3, which plots the rate of false positive predictions for a range of gene set sizes as a function of max_s. The data suggested no dependency of the false prediction rate on set size. We also noted that at low score values, the proportion of false positives is low.

3.4. *Web interface*

oPOSSUM II web service is available at http://www.cisreg.ca/oPOSSUM2/opossum2.php. A user enters a set of putatively co-expressed genes and specifies the parameter values to be used in the analysis. Certain parameter values may produce lengthy runtimes. To accommodate this possibility, our web service will queue the analysis request and will notify the user via e-mail once the analysis is complete.

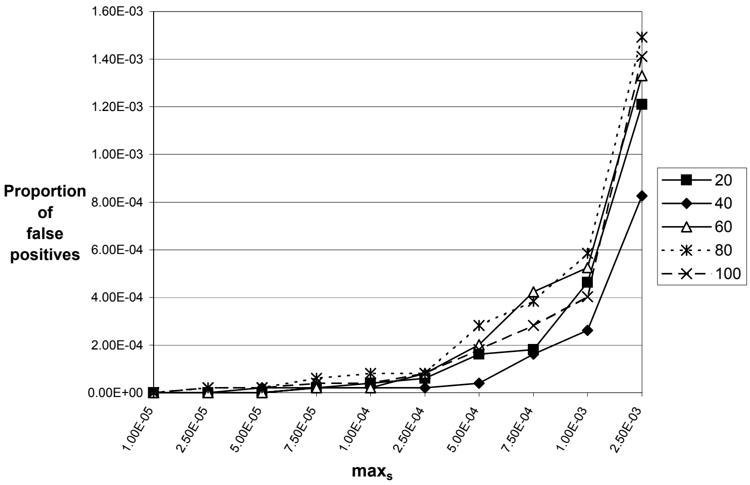

Figure 3. Effect of gene set size on false positive rate observed from pairwise TFBS combinations in randomly generated foreground gene sets.

4. Discussion

The analysis of over-represented combinations of TFBS in the promoters of co-expressed genes is motivated by biochemical and genetic studies which reveal the functional importance of cis-regulatory modules. In contrast to previously described methods which identify single over-represented motifs, the analysis of combinations must solve or circumvent the consequence of a combinatoric explosion, which can precipitate prohibitive runtimes. To reduce the search space, oPOSSUM II restricts its analysis to binding site combinations using biologically justifiable criteria, namely, TF profile similarity.

Our results suggest two important contributions over the existing single-site TFBS over-representation methods. Firstly, in each reference gene set, there is at least one relevant TF class that appears in multiple combinations, an observation that is not immediately obvious in single site analysis. Secondly, the algorithm finds functional TFBS that are not indicated in single site analysis. For instance with the yeast CLB2 gene cluster, members of the top scoring combination, ECB and FKH1, are ranked the first and eleventh in single site analysis. In the smooth muscle reference set, the SRF and SP1 combination is the most significant, but they are ranked the first and fourteenth in single site analysis. These results clearly demonstrate the power of combination site analysis.

Analysis of the microarray-based skeletal muscle reference set correctly implicates the combination of MEF2 and SP1 TFs in myogenesis. This result confirms the utility of high-quality microarray data for regulatory sequence analysis.

While our result for the yeast CLB2 cluster is comparable to that reported by Kreiman,[5] there are significant differences between the methods. Kreiman initially uses a motif discovery algorithm to identify new motif patterns from a gene set and then subsequently looks for over-represented combinations of motifs using both the new motif patterns and a TFBS profile database. In our interpretation, there is circular logic in looking for relevant motifs in a reference gene set and then identifying their over-represented combinations. For the CLB2 cluster, the profiles were taken from an existing database and our results are comparable. For the first skeletal muscle collection, Kreiman reports the top scoring combination as SP1, SRF, TEF and a motif drawn from the promoters of the positive gene set.

Although this paper presents the results for pairs of TFBS, the oPOSSUM II implementation is also able to evaluate combinations of higher cardinality. However, validation of larger combinations is seriously limited by the lack of robust reference data sets that include genes known to be regulated by multiple binding sites.

A few issues remain to be addressed by future research. First, the interpretation of analysis results is confounded by intra-class binding similarity. While this property facilitates the oPOSSUM II algorithm, users must be prepared to consider which proteins in a family are most likely to act within the tissue or under the condition studied. For instance, the fact that an E-box motif is over-represented in the skeletal muscle data does not directly lead the researcher to the MyoD protein; instead the user must consider the entire range of bHLH-domain TFs. Second, inter-class similarity can influence the results. Although oPOSSUM II does not allow overlap between TFBS in the analysis of a given combination, TFBS from different combinations can overlap. Thus two G-rich motifs may be reported as over-represented in different combinations (for instance, the SP1 and MZF motifs in Figure 2c) but highlight the same candidate TFBS within the sequences analyzed. A related issue is the compositional sequence bias in tissue specific genes,[17] which would motivate selection of a more refined background gene set. Finally, the required computing time is prohibitively long for a synchronous web service. Parallelization of the enumeration algorithm is a natural way to improve the running time.

5. Conclusion

oPOSSUM II utilizes putative TFBS identified from comparative genomic analysis, in conjunction with knowledge of co-regulated expression, to search for functional combinations of TFBS that may confer a given gene expression pattern. It uses a novel scheme to classify similar binding site profiles. Using this clustering approach, the oPOSSUM II method is able to circumvent the combinatorial challenge associated with the identification of significant TFBS combinations. Furthermore, the application of an IBSD constraint limits the number of possible combinations to analyze. Validation results suggest that TFBS combination site analysis can provide valuable information that is not available through a single-site analysis.

Acknowledgments

We thank Andrew Kwon for annotation of the muscle reference collections. We acknowledge operating support from the Canadian Institutes for Health Research (CIHR) and Merck Frosst; DF was supported by the CIHR/MSFHR Bioinformatics training program and the Merck Frosst Co-op program; WWW is supported as a Michael Smith Foundation for Health Research Scientist and a New Investigator of the CIHR.

References

1. M. I. Arnone and E. H. Davidson. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124(10):1851–64, 1997.
2. N. Bluthgen, S. M. Kielbasa, and H. Herzel. Inferring combinatorial regulation of transcription in silico. *Nucleic Acids Res*, 33(1):272–9, 2005.
3. S. J. Ho Sui, J. R. Mortimer, D. J. Arenillas, J. Brumm, C. J. Walsh, B. P. Kennedy, and W. W. Wasserman. oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res*, 33(10):3154–64, 2005.
4. J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol*, 296(5):1205–14, 2000.
5. G. Kreiman. Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. *Nucleic Acids Res*, 32(9):2889–900, 2004.
6. B. Lenhard, A. Sandelin, L. Mendoza, P. Engstrom, N. Jareborg, and W. W. Wasserman. Identification of conserved regulatory elements by comparative genome analysis. *J Biol*, 2(2):13, 2003.
7. C. S. Madsen, J. C. Hershey, M. B. Hautmann, S. L. White, and G. K. Owens. Expression of the smooth muscle myosin heavy chain gene is regulated by a negative-acting GC-rich element located between two positive-acting serum response factor-binding elements. *J Biol Chem*, 272(10):6332–40, 1997.
8. J. L. Moran, Y. Li, A. A. Hill, W. M. Mounts, and C. P. Miller. Gene expression changes during mouse skeletal myoblast differentiation revealed by transcriptional profiling. *Physiol Genomics*, 10(2):103–11, 2002.
9. S. Nelander, P. Mostad, and P. Lindahl. Prediction of cell type-specific gene modules: identification and initial characterization of a core set of smooth muscle-specific genes. *Genome Res*, 13(8):1838–54, 2003.
10. A. Sandelin, W. Alkema, P. Engstrom, W. W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32(Database issue):D91–4, 2004.
11. A. Sandelin, A. Hoglund, B. Lenhard, and W. W. Wasserman. Integrated analysis of yeast regulatory sequences for biologically linked clusters of genes. *Funct Integr Genomics*, 3(3):125–34, 2003.
12. R. Sharan, A. Ben-Hur, G. G. Loots, and I. Ovcharenko. CREME: Cis-Regulatory Module Explorer for the human genome. *Nucleic Acids Res*, 32(Web Server issue):W253–6, 2004.
13. P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–97, 1998.
14. K. K. Tomczak, V. D. Marinescu, M. F. Ramoni, D. Sanoudou, F. Montanaro, M. Han, L. M. Kunkel, I. S. Kohane, and A. H. Beggs. Expression profiling and identification of novel genes involved in myogenic differentiation. *FASEB J*, 18(2):403–5, 2004.
15. W. W. Wasserman and J. W. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol*, 278(1):167–81, 1998.
16. W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*, 5(4):276–87, 2004.
17. R. Yamashita, Y. Suzuki, S. Sugano, and K. Nakai. Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. *Gene*, 350(2):129–36, 2005.