Published online 20 May 2011

Nucleic Acids Research, 2011, Vol. 39, No. 15 e98
doi:10.1093/nar/gkr341

Inferring transcription factor complexes from ChIP-seq data

Tom Whitington[1], Martin C. Frith[2], James Johnson[1] and Timothy L. Bailey[1,*]

[1]Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia and [2]Computational Biology Research Center, Institute for Advanced Industrial Science and Technology, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan

Received January 22, 2011; Revised April 13, 2011; Accepted April 17, 2011

ABSTRACT

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) allows researchers to determine the genome-wide binding locations of individual transcription factors (TFs) at high resolution. This information can be interrogated to study various aspects of TF behaviour, including the mechanisms that control TF binding. Physical interaction between TFs comprises one important aspect of TF binding in eukaryotes, mediating tissue-specific gene expression. We have developed an algorithm, spaced motif analysis (SpaMo), which is able to infer physical interactions between the given TF and TFs bound at neighbouring sites at the DNA interface. The algorithm predicts TF interactions in half of the ChIP-seq data sets we test, with the majority of these predictions supported by direct evidence from the literature or evidence of homodimerization. High resolution motif spacing information obtained by this method can facilitate an improved understanding of individual TF complex structures. SpaMo can assist researchers in extracting maximum information relating to binding mechanisms from their TF ChIP-seq data. SpaMo is available for download and interactive use as part of the MEME Suite (http://meme.nbcr.net).

INTRODUCTION

Chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing (ChIP-seq) has proven to be a powerful and high-resolution method for mapping the *in vivo* locations of individual transcription factors (TF) proteins, genome-wide in higher eukaryotes (1). In the ChIP methodology, DNA is first covalently cross-linked to bound proteins in a specific tissue. The cross-linked DNA is next broken into small fragments. An antibody for the TF of interest is then used to isolate the population of DNA bound to the feature. High-throughput sequencing of short tags from the resulting DNA population, followed by mapping to a reference genome allows the original genomic binding locations of the TF to be inferred.

Computational analysis is necessary in order to extract biologically relevant information from a transcription factor's ChIP-seq data. Previous TF ChIP-seq studies have employed several common analytical steps. Although the extent of DNA binding by a TF is, in fact, a continuous variable, defining a set of discrete binding regions simplifies subsequent analyses and is therefore a common preliminary step. Once a set of binding regions have been declared, existing computational tools can be employed to investigate the mechanisms by which the TFs bind at those genomic loci.

Some TFs directly interact with DNA via DNA-binding domains (DBDs), with sequence specificity of binding determined by the structure of the DBD. A TF 'motif' models the sequence specificity of the TF's DBD. Given a set of binding regions determined by ChIP-seq, *ab initio* motif discovery tools such as MEME (2) and GLAM (3) can discover the motif corresponding to the TF's DBD, while sometimes identifying additional motifs corresponding to co-regulatory TFs. Motif enrichment analysis tools such as Clover (4) and PASTAA (5) can identify motifs for co-regulatory TFs with increased sensitivity, by considering a restricted set of TF motifs hypothesised to be enriched in the peak regions.

Many TFs physically interact with specific partner TFs when binding to genomic DNA (6,7). These combinatorial interactions are critical to TF biology, as they provide a means by which the cell can integrate diverse signals (8), as well as increasing the sensitivity of transcriptional rates to TF concentration, and allowing non-functional random occurrences of individual motifs to be left unbound (9). Protein–protein interactions between TFs can occur

*To whom correspondence should be addressed. Tel: +617 3346 2614; Fax: +617 3346 2101; Email: t.bailey@imb.uq.edu.au directly, yielding dimers, such FOS/JUN (10) and MYC/MAX (11). Alternatively, interactions between TFs can occur via intermediate bridging proteins, resulting in a DNA-binding complex of multiple TFs, such as GATA-1/SCL/E47/Ldb1 (12). In both cases, spacing of motifs for the DNA-binding TFs is often inflexible, as addition or removal of base pairs disrupts the protein–protein interactions between the TFs (6).

Existing motif analysis tools do not harness the knowledge that spacing of motifs is often tightly restricted in TF complexes. Unlike existing motif enrichment analysis tools, SpaMo detects enrichment of motif *spacings* rather than enrichment of motif occurrences. By employing this approach, SpaMo is able to detect TF complexes from ChIP-seq data with a high positive predictive value. The resulting high resolution information can facilitate prediction of 3D complexes, given X-ray crystal structures of the component TFs. We demonstrate this on 39 prior ChIP-seq data sets, successfully inferring known TF complexes as well as identifying high-confidence novel TF complexes.

MATERIALS AND METHODS

Input data sets

Sequence data sets. We use 39 human and mouse ChIP-seq datasets derived from 7 prior publications and the ENCODE project (13). The complete list of data sets is given in Supplementary Table S1. We process each input ChIP-seq data set in preparation for running SpaMo. For each ChIP-seq peak declared in a given data set, we extract 500 bp of DNA sequence centred on the centre of the declared peak. We use the UCSC table browser tool (14) to extract genomic sequences.

Our null model assumes the sequences corresponding to ChIP-seq peaks are independent, so it is important to remove homologous sequences and repeat regions. Therefore, we filter out highly similar sequences, and we use repeat masking (http://www.repeatmasker.org) to convert repeat regions to the information-less character 'N'. To remove similar sequences from a data set, we align the sequences (without gaps) on the primary motif occurrence and randomly remove a sequence that is Hamming distance 150 or less from some other sequence. We repeat this until no sequence can be removed.

Primary motifs. We assign primary motifs to the input ChIP-seq data sets as shown in Supplementary Table S1.

Secondary motif database. Input secondary motifs include all motifs from the JASPAR CORE (15) and Uniprobe (16) databases, supplemented with custom motifs, as described in Supplementary Table S4. This database contains 645 motifs.

We trim all motifs to eliminate low information content (IC; see definition below) flanking columns prior to running SpaMo. We remove all columns with $\mathrm{IC} \leq 0.25$ bits from both sides of the motif. Failure to trim low IC flanking columns can result in significant spacings not being detected (Supplementary Figure S3).

The IC of an individual column in a motif is defined as:

$$\mathrm{IC} = \sum_{i=1}^{4} p_i \times \log_2(4 \times p_i) = 2 + \sum_{i=1}^{4} p_i \times \log_2(p_i),$$

where p_i is the probability of observing the i-th letter in the given column under the motif model, and the i-th letter is specified by element i in the array [A, C, G, T].

Identifying locations of primary and secondary motifs

FIMO (2) is used to perform motif scans. The best match to a motif of length w in a given double-stranded sequence is defined as the position and strand that yields the highest log-likelihood ratio (LLR), considering all possible substrings of length w in either strand of the sequence. Ties are broken by randomly choosing a single match from all equal best matches.

The LLR of a given genomic position Q is defined as:

$$\mathrm{LLR} = \log_2 \frac{Pr(Q|\mathrm{motif})}{Pr(Q|\mathrm{bg})},$$

where motif is the motif model of binding, and bg is a zero-order background Markov model of the DNA. A single background model is compiled using all the sequences in given input sequence data set.

The primary motif scan excludes 150 bp at either end of each 500 bp input sequence. The secondary motif scan is performed over the 300 bp region centred on the primary motif, and excludes all positions overlapping any part of the primary motif occurrence. Thus, the 'trimmed' length of the sequences is 300 bp plus the width of the primary motif.

Sequences with a maximum primary or secondary LLR less than a specified bit threshold are discarded. In the case of the primary motif, the rationale is that such sequences may not have bound the TF of interest directly. In the case of the secondary motif, discarding sequences reduces noise in the statistical analysis. We used a score threshold of 7 bits for all analyses, except for analysis of the E2F1 ChIP-seq input data set. In the case of the E2F1 ChIP-seq input data, we applied a less strict threshold of 4 bits, as no statistically significant results were obtained using a threshold of 7 bits for this data set.

The distance, D, between the best primary and secondary motif occurrences is defined as the number of nucleotides occurring between the closest edge of the primary motif and the closest edge of the secondary motif. The offset, f, between the primary and secondary motifs is defined as $f = -(D+1)$ if the secondary motif occurs 5′ of the primary motif, and $f = (D+1)$ otherwise (Supplementary Figure S1).

Assessing the significance of motif spacings

The displacement of a given secondary motif site is written as $d = (s, f)$, where $s \in \{\text{same, opposite}\}$ is the strand of the secondary site, and $f \in [-r, \ldots, -1, +1, \ldots, +r]$ is the offset of the secondary site.

r is given by:

$$r = \frac{m - w_p}{2} - w_s,$$

where, m is the length of trimmed sequences, following centering on the primary motif, w_s is the width of the secondary motif and w_p is the width of the primary motif (Supplementary Figure S2).

We assume that every value of d is equally probable under the null hypothesis of no spatial relationship between the motifs. Therefore, the probability of a given displacement value is $\frac{1}{4r}$ under the null model. For an interval of integer size x, we define the probability of a single sequence having observed spacing contained in the given interval as:

$$q = Pr(d \in \text{interval}) = \frac{x}{4r}$$

Therefore, if there is no spatial relationship between the given primary and secondary motifs, the number of sequences, s, with observed displacements in a given interval should follow a binomial distribution $\mathrm{Bin}(s, N, q)$, where s is the number of successes, N is the number of trials, and q is the probability of success. The number of trials, N, is the total number of filtered sequences yielding a secondary–primary motif displacement value. Hence, we use the cumulative distribution function for $\mathrm{Bin}(s, N, q)$ to calculate the probability of observing a displacement value in the given interval for s or more sequences by chance. The resulting value is an uncorrected P-value for the given interval. When applying the algorithm to our input data sets, we consider only intervals of size 1, although SpaMo can also consider larger intervals.

Multiple-testing correction

To reduce the number of independent tests (and, hence, to improve our ability to detect significant results), we only test spacing enrichment for each integer displacement value in the range [−20, +20], ignoring any enrichment in the rest of the potential range, $[-r, \ldots, -1, +1, \ldots, +r]$. For each primary–secondary motif pair, we independently test for enrichment where the motifs are on the same or opposite DNA strands, resulting in a total of $40 \times 2 = 80$ separate binomial tests. Therefore, we perform a Bonferroni correction to correct for the 80 separate intervals tested and the 645 secondary motifs considered. Thus, motif spacing P-values we report are the binomial P-values multiplied by 51 600.

Redundancy reduction

Many motifs included in the secondary motif input database are similar, and hence yield highly similar results. To facilitate easier interpretation of the output data, we perform a redundancy reduction on the output for each ChIP-seq data set.

To determine the degree of similarity between results obtained for two secondary motifs, we measure the overlap in the sequences exhibiting significant enrichment, instead of measuring similarity between the motifs themselves. We define the fractional overlap f_{ij}, between the results yielded by secondary motifs i and j as:

$$f_{ij} = \frac{|s_i| \bigcap |s_j|}{\min(|s_i|, |s_j|)}$$

Here, s_x is the set of sequences whose secondary–primary motif displacement value shows statistically significant enrichment and vertical bars are the set-size operator.

For each secondary motif with at least one interval enrichment $P < 0.05$, the lowest P-value for that motif is identified among the 80 intervals tested. We refer to this value as the 'best P-value'. Secondary motifs are sorted according to their best P-value. Then, proceeding from the secondary motif with the most significant best P-value to the secondary motif with the least significant best P-value (which is still <0.05), for each secondary motif i we consider each motif j with best P-value greater than that of motif i. We calculate f_{ij} between the two motifs, and if f_{ij} exceeds 0.25, then we mark motif j as being redundant with motif i. We then only report results for secondary motifs that are found to be non-redundant (i.e. have a more significant best P-value than all motifs with significant fractional overlap).

TF complex structure prediction

We manually performed superimposition of structures using PyMOL (17). The simulated sequence structure was generated using the Nucleic Acids Builder tool, with default parameters (18). We visualized the resulting structure by hiding all atoms in the original DNA structures.

RESULTS

The SpaMo algorithm

SpaMo analyses the genomic DNA sequences of a set of TF binding site loci estimated by ChIP-seq for a given TF. The algorithm attempts to identify enriched motif spacing patterns indicative of specific transcription factor complexes.

Inputs to the algorithm comprise a set of DNA sequences corresponding to the genomic regions bound by a specific TF, a primary motif that describes the DNA binding specificity of that transcription factor and a database of secondary motifs. SpaMo uses the primary motif to predict the exact location of a binding in each ChIP-seq peak region. For each secondary motif, SpaMo tests the hypothesis that there is enriched spacing of predicted binding sites with respect to the primary motif sites. An individual test corresponds to the following question: 'Does TF A tend to bind DNA at a fixed distance from TF B?'. If the answer is 'yes', it suggests that A is likely to form a complex with B.

SpaMo uses motifs defined as position weight matrices (19), but could easily be adapted to used lookup-table based motifs derived from protein-binding microarrays (20). SpaMo scans each of the input sequences with the primary motif and finds the best match ('hit') to the motif in each sequence as defined by position-weight matrix score. Each sequence is then trimmed to identical

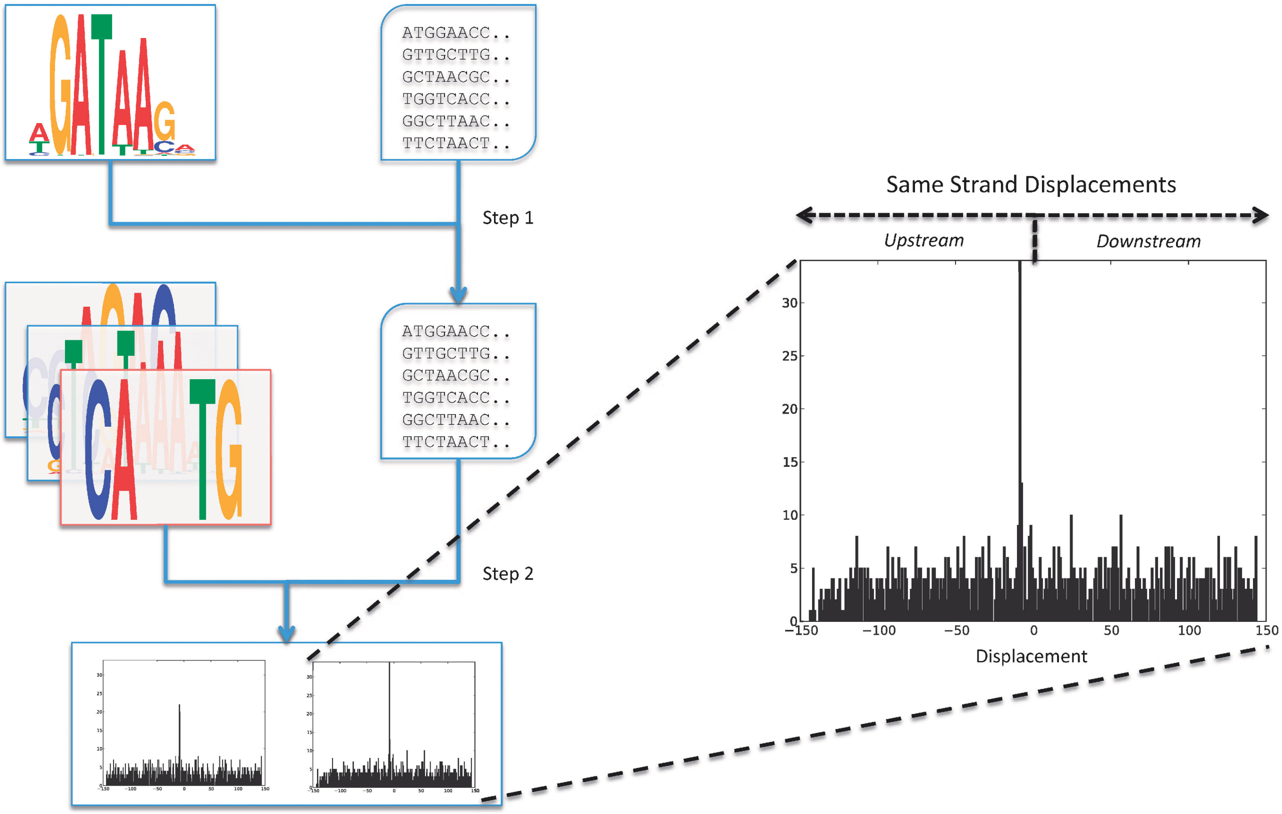


Figure 1. Algorithm procedure and output. Step 1. The strongest match to the specified primary motif is identified in each ChIP-seq peak region genomic sequence. Each sequence is centred on the motif occurrence and trimmed to a consistent length. Step 2. A library of secondary motifs is considered. For a given secondary motif, the processed sequences are scanned to identify the strongest match in each sequence, and the displacements from the primary hit to the secondary hit are recorded. Output. Same-strand and opposite-strand histograms are produced. For the example output shown, the primary and secondary motifs are Gata6 and Ebox, respectively, and the input ChIP-seq data set is human GATA1 in the K562 cell line. The same-strand displacement histogram indicates a clear enrichment of sequences with a secondary–primary displacement of −8 bp.

length, centred on the primary motif occurrence (Step 1, Figure 1). For each secondary motif, SpaMo finds the strongest secondary motif hit, and calculates the displacement from the secondary motif hit to the primary motif hit (Step 2, Figure 1). With motif hits defined in this way, a reasonable null model for the distance between the primary and secondary hits in a single sequence is the uniform distribution.

Each interval of displacements in a user-specified range is tested to determine whether the observed number of sequences with displacements in that range exceeds the number expected by chance. Under the assumption of a uniform distribution of secondary–primary displacements, the expected number of sequences with displacements scoring in a given interval should follow a binomial distribution. Thus, SpaMo uses a binomial test to assess significance for each interval of interest. A Bonferroni correction is then applied to each P-value, to correct for both the number of intervals and the number of secondary motifs considered.

SpaMo produces two histograms of secondary–primary motif site displacements: one for sequences where the primary and secondary motifs occur on the same strand, and a separate histogram for sequences where the motifs occur on opposite strands (Figure 1). Each histogram displays motif site displacement, indicating whether the secondary motif site is upstream or downstream of the primary motif site. Visualizing the resulting four categories of displacement (same strand/opposite strand, upstream/downstream) separately is of practical importance, as they correspond to distinct physical placements of the corresponding TFs.

Evaluating performance of SpaMo

We evaluated the performance of SpaMo on 39 human and mouse ChIP-seq data sets. These data sets were derived from seven prior publications and the ENCODE project (13). For the c-Fos ChIP-seq (comprising two data sets), we performed the analysis twice, using two distinct primary motifs. Thus, we performed a total of 41 analyses (Supplementary Table S1). The complete set of 87 significant motif spacing results at P-value threshold of 0.01 are provided (Supplementary Table S3).

We examined our strongest predictions ($P < 0.001$) to estimate what fraction represent true *in vivo* complex formation. In 20 of our 41 analyses, at least one significant spacing was detected at this more stringent P-value threshold. To validate our approach, we evaluated the single most statistically significant result for each input data set

Table 1. Positive predictive value of top predictions

TF/tissue	Primary motif	Secondary motif	Likely partner	Lowest *P*-value	Evid.
Esrrb/ESC	C Esrrb	C Esrrb	Esrrb	4.23×10^{-56}	S (22)
STAT1/HeLa Stim.	C Stat3	J YY1	YY1	1.52×10^{-29}	
GABP/Jurkat	U Gabpa i	U Fhl1	?	7.95×10^{-28}	
cFos/Gm12878	C NFYA	J NFYA	C/EBP	2.87×10^{-23}	S (23)
cFos/K562	C NFYA	U Cbf1 b	?	8.62×10^{-21}	
Jund/Gm12878	U Jundm2 ii	U Irf4 i	Irf4	2.02×10^{-16}	
GATA1/K562b	U Gata6 i	C Ebox	SCL	2.76×10^{-16}	(12)
cJun/K562	U Jundm2 ii	J SPIB	PU.1	3.49×10^{-16}	(24)
cFos/K562	U Jundm2 ii	J SPIB	PU.1	9.24×10^{-14}	(24)
Tcfcp2l1/ESC	C Tcfcp2l1	C Tcfcp2l1	Tcfcp2l1	9.24×10^{-14}	S
GATA1/G1EER4	U Gata6 i	U Ascl2 i	SCL	1.32×10^{-10}	(12)
STAT1/HeLa Stim.	C Stat3	J YY1	YY1	9.70×10^{-10}	
Srebp1a/Hepg2	C Srebp	U Rsc30	?	3.58×10^{-8}	
Klf4/ESC	U Klf7 i	U Zfp740 i	Klf4	4.35×10^{-7}	S
Nfe2/K562	C Nfe2	U Jundm2 ii	Nfe2	1.08×10^{-5}	S
cMyc/K562	J Mycn	J bZIP910	?	6.30×10^{-5}	
Sox2/ESC	C Oct4	U Sry ii	Sox2	1.33×10^{-4}	(25)
Tcf4/Hct116	U Tcf3 i	U Jundm2 ii	c-Jun	3.12×10^{-4}	(26)
SRF/Jurkat	U Srf i	J ETS1	SAP-1	3.99×10^{-4}	(27)
E2F1/ESC	J E2F1	J YY1	YY1	9.39×10^{-4}	(28)

For each input dataset that yielded one or more results at a *P*-value threshold of 0.001, the single most significant result is presented. In the first column, the TF tissue and reference for the ChIP-seq data set is given. The 'primary motif' indicates the motif used during the first step of the algorithm. The 'secondary motif' indicates the motif found to exhibit the significant spacing. Summary names are provided for both motifs, where 'J' indicates a JASPAR (15) motif, 'U' indicates a Uniprobe (16) motif, 'C' indicates a custom motif. Corresponding sequence logos (29) are shown in Supplementary Table S4. The 'Likely partner' column indicates the TF that we manually assigned to the secondary motif, with '?' indicating we could not assign a likely partner. The *P*-value corresponds to the single most significant spacing interval. The 'Evid.' column states evidence validating the given prediction, with references indicating literature confirmation, and 'S' indicating that the primary and secondary motifs are highly similar.

using primary–secondary motif similarity and literature evidence, as described below.

SpaMo identifies secondary motifs that exhibit enriched spacing with respect to the specified primary motif. Many TFs have paralogs sharing the same DBD and hence the same DNA-binding specificity. For example, there are 17 known KLF family members in mammals, as defined by the presence of a DBD consisting of three highly conserved Cys_2His_2 zinc fingers, which bind to a CACC-box motif (21). Since multiple TFs can bind to the same motif, knowledge of relevant TFs in an individual system must be applied in order to identify the TF corresponding to an observed secondary motif spacing enrichment. Therefore, in order to evaluate our results, we have manually assigned a likely TF to each spatially enriched secondary motif where possible. In some cases, we were unable to assign a likely binding partner corresponding to the given secondary motif (Table 1, '?' in 'Likely Partner' column). This is expected, as many TFs currently have no binding motif.

We searched prior publications for evidence of formation of a complex involving the primary and secondary TFs each of the top 20 results. Ten of the top 20 results are supported by prior publications (Table 1). These studies employed X-ray crystallography, electrophoretic mobility shift assays (EMSAs), immunoprecipitation, yeast two-hybrid and luciferase assays to demonstrate formation of complexes involving our predicted TF pairs (Supplementary Table S2). In addition to the 10 predictions with clear support for complex formation in the literature, we found partial support for our prediction of a complex involving JUND and IRF4. Specifically, IRF4 interacts with both PU.1 (30) and NFAT (31), which are also binding partners of the Jund/c-Fos heterodimer, AP-1, suggesting that our predicted interaction between JUND and IRF4 is plausible.

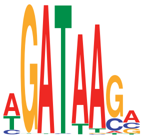
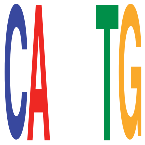
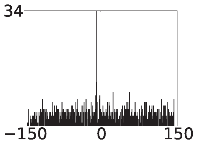
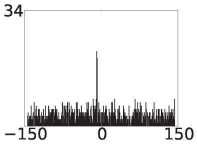
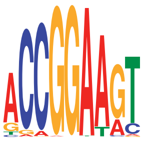
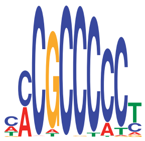
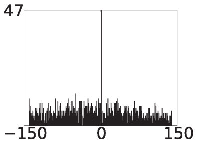
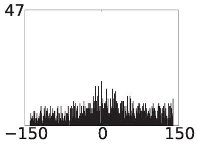
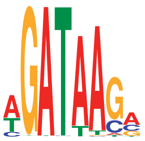
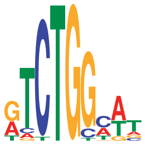
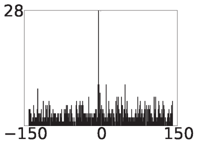
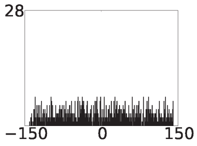
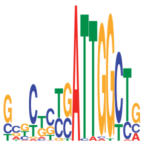
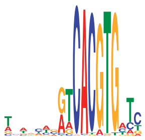
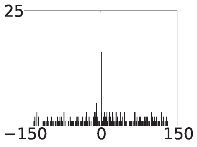
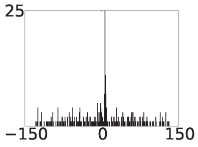
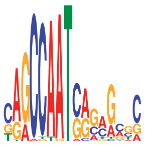
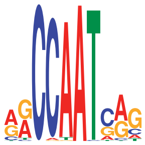
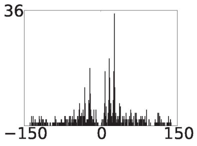
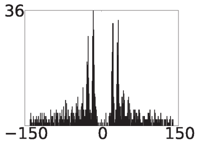
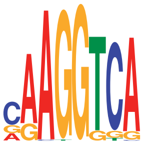
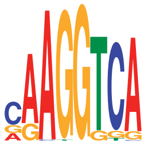
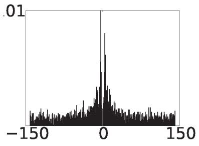
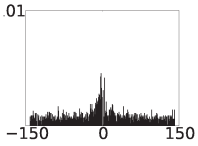
For five of our top 20 results (Table 1, 'S' in 'Evid.' column), the primary and secondary motifs are highly similar or identical. High similarity between the primary and secondary motifs is unlikely to occur by chance. Instead, a likely explanation is homodimer formation, whereby two copies of a TF bind to each other to form a ternary complex with the DNA.

A total of 13 of the top 20 TF complex predictions made by SpaMo are supported by either literature evidence or primary–secondary motif similarity, yielding an estimate of 65% for positive predictive value. The very signficant *P*-values of the remaining seven top predictions (including the JUND/IRF4 interaction) suggest that they are also likely to be relevant to the given TF's binding mechanism. In particular, the interactions predicted by SpaMo between TFs Stat1 and YY1, and between Gabp1 and Fhl1, are very high confidence (*P*-values of 1.52×10^{-29} and 7.95×10^{-28}, respectively). We, therefore, estimate that SpaMo can yield predictions for about half input ChIP-seq data sets, with the majority of such predictions likely to be correct.

Biological interpretation of motif spacing classes

Biologically relevant distinctions are apparent among the observed motif spacings. Meaningful classes can be derived by considering the degree of similarity between

Table 2. Classes of motif spacing

Input	Primary motif	Secondary motif	Same Disps	Opp. Disps	Sig. Interval	Evid.
hg18 GATA1 K562b	 U Gata6 i	 C Ebox			–9, Same 2.76×10^{-16}	1
hg18 GABP Jurkat	 U Gabpa i	 U Sp4 i			+1, Same 1.43×10^{-21}	2
hg18 GATA1 K562b	 U Gata6 i	 J Hand1::Tcfe2a			–5, Same 7.02×10^{-13}	1,3 #
hg18 cFos K562	 C NFYA	 U Cbf1 b			+6, Opp. 8.62×10^{-21}	
hg18 cFos K562	 C NFYA (rc)	 J NFYA			–17, Opp. 6.71×10^{-20}	S
mm8 Esrrb ESC	 C Esrrb	 C Esrrb			–4, Same 4.23×10^{-56}	S

In the first column, the genome assembly, TF, tissue and reference for the input ChIP-seq data set is given. For 'Primary motif' and 'Secondary motif' columns, the sequence logos and summary names are provided. Same strand and opposite strand displacement histograms are shown in columns three and four. The *X*-axis of each histogram shows the motif displacement value. The *Y*-axis shows the number of sequences that exhibited the given secondary–primary motif displacement value, and is scaled linearly with the origin corresponding to zero. The 'Sig. Interval' specifies the displacement value and strand for the single most significant interval, with 'Opp.' indicating opposite strand. The corrected *P*-value of that interval is given. The 'Evid.' column is described in Table 1. '#': the cited studies demonstrate that GATA1 and Tcfe2a (Tcf3; E2A; E47) form at least two distinct DNA-binding complexes. While neither of these complexes correspond to our predicted 'U Gata6 i'/'JHand1::Tcfe2a' motif spacing, they do support our predicted association between GATA1 and Tcfe2a. The reverse complement of the 'C NFYA' motif is shown in row 5 in order to exhibit similarity with the secondary motif 'J NFYA'. Literature evidence is as follows: **1** = (12), **2** = (34), **3** = (35).

the primary and secondary motifs, the breadth of intervals enriched and the distance between the primary and secondary motif sites.

We observed statistically significant motif spacings in which the primary motif is dissimilar to the secondary motif (Table 2, rows 1–4), and others in which the primary motif is similar or identical to the secondary motif (Table 2, rows 5 and 6). The distinction between these two classes is biologically important, as TFs can form homodimers comprising two occurrences of the same TF, or they can form complexes involving distinct TFs. Statistically significant spacings involving highly similar primary and secondary motifs suggest binding of homodimers in the ChIP-seq peak regions. In contrast, spacings involving dissimilar primary and secondary motifs are potentially due to complexes involving the TF of interest and one or more distinct TFs. We detected a statistically significant ($P < 0.01$) spacing suggestive of homodimer formation in 10 (24%) of 41 ChIP-seq data set analyses we performed.

We found some motif spacing enrichments occurred over very tight intervals of 1–2 bp (Table 2, rows 1–4), while others occurred over broader intervals (Table 2, rows 5 and 6). The occurrence of tight motif spacing enrichment can be parsimoniously explained by the binding of TF complexes in the ChIP-seq regions. Previous work on the MAT**a**1/MATα2 TF complex showed that modifying the distance between the respective binding motifs abolishes binding of the complex (32). The fact that TF complex formation requires highly specific motif spacing explains the tight restriction of spacings observed in many of the significant results. Of the 87 results obtained at a *P*-value threshold of 0.01, the majority (84%) exhibit tight spacing enrichment [exactly one displacement value declared significant (Supplementary Figure S6)], which is consistent with TF complex formation.

The relatively small number of results with broad spacing enrichment could be due to the occurrence of multiple adjacent but independent *in vivo* binding sites, rather than adjacent cooperative sites. Clustering of independent binding sites in some cases arises due to selection for a specific response of transcriptional rate to TF concentration (33). However, the two example broad spacings shown (Table 2, rows 5 and 6) are likely to indicate complex formation, as narrow peaks are clear within the broader intervals of enrichment. A periodicity is clear for the broad NFYA-NFYA motif spacing enrichment detected in the c-Fos data set. The periodicity is 10 bp in length, which corresponds to approximately one turn of the DNA double helix, suggesting that the orientation of the two TFs relative to the DNA is important.

Most of the observed significant spacings involved a small gap of <2 bp between the primary and secondary motifs, while a minority of results showed larger gaps (Supplementary Figure S4). Large gaps between the primary and secondary motifs can indicate TF complexes containing bridging molecules. For example, the Gata-Ebox motif spacing shows a gap of 9 bp between the two motifs (Table 2, row 1), consistent with previous CASTing experiments (12). The relatively large gap of 9 bp is due to the formation of a multi-protein/DNA complex in which GATA1 binds to a Gata motif, SCL binds to an Ebox, while E47, Ldb1 and Lmo2 comprise a molecular bridge between GATA1 and SCL. In contrast, small gaps between the primary and secondary motifs suggest dimer formation via direct protein–protein interactions. For example, Gabpa and Sp1 are known to interact directly (34), which corroborates our observation that the Gabpa/Sp motif spacing involves no gap between the two motifs (Table 2, row 2).

Identification of multiple partners for a single TF

SpaMo is capable of identifying more than one significant and distinct secondary motif association for a given input data set. For example, using the Rozowsky *et al.* (36) Stat1 ChIP-seq data set as input, we identified 11 secondary motif interactions at a *P*-value threshold of 0.01 (Supplementary Table S3). The secondary motifs with the five most statistically significant corrected *P*-values are clearly distinct from one another (Table 3).

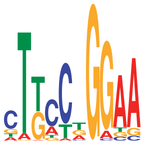
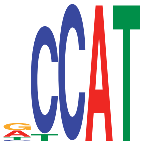
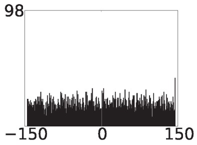
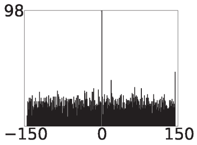
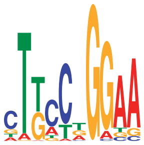
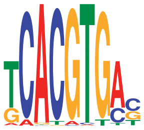
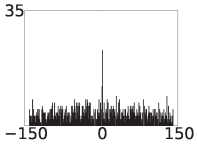
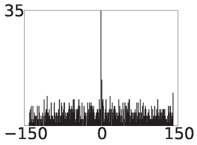
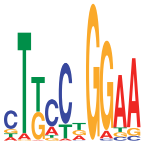
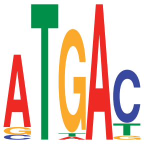
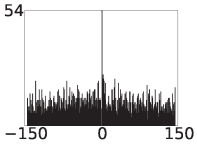
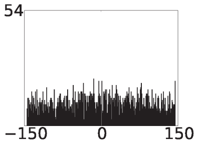
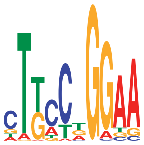
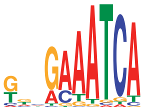
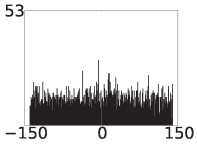
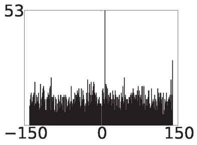
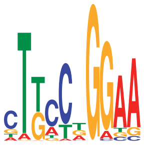
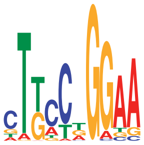
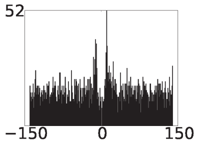
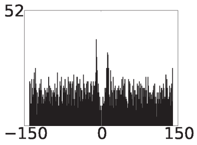
The observed distribution of the number of predicted partners suggests that most TFs have very few interactions, while a small fraction of transcription factors (such as Stat1) possess a relatively large number of interacting partners (Supplementary Figure S5). This might change as more motifs become known. The GATA1, c-Fos, GABP, Stat1 and Tcfcp2l1 input data sets all yield two or more extremely high confidence, distinct secondary motif associations (Supplementary Table S3), and are possible hub nodes in the network of physically interacting TFs. It is noteworthy that we do not detect any high-confidence predicted partners for STAT1 in unstimulated HeLa cells. This is consistent with previous observations that interferon-gamma stimulation causes Janus kinase (Jak) to phosphorylate the polymerization domain of STAT proteins, enabling them to interact with other proteins and bind DNA cooperatively (37).

Predicting 3D transcription factor complex structures

The detailed motif spacing information produced by SpaMo can facilitate prediction of the 3D structure of the corresponding TF complexes. For example, our results show that GABP and CREB1 motifs are enriched at a spacing of 1–2 bp on opposite strands. Based on this information, we construct a model for the structure of the GABP–CREB1–DNA ternary complex, using two separate GABP and CREB1 TF X-ray crystal structures from the Protein Data Bank (PDB) (38) (accessions 1AWC and 1DH3, respectively).

We used the Nucleic Acids Builder tool (18) to generate an artificial structure for a double-stranded DNA sequence consisting of the GABP consensus sequence (5′ACCGGAAGT3′) followed by the reverse complement of the CREB1 consensus sequence (5′ACGTCAGCATG3′), in accordance with the spacing enrichment shown (Figure 2A). We aligned the DNA structures in the original X-ray crystal structures with the corresponding regions of the simulated DNA structure, by manually overlaying the positions of the corresponding nucleotides in the two structures. The resulting predicted ternary GABP/CREB1/DNA complex contains no major steric

Table 3. Discovery of multiple distinct spacings for a single TF

Input	Primary motif	Secondary motif	Same Disps	Opp. Disps	Lowestp-val	Evid.
hg18 STAT1 HeLaStim	 C Stat3	 J YY1			+1, Opp. 1.52×10^{-29}	
hg18 STAT1 HeLaStim	 C Stat3	 U Bhlhb2 i			−1, Opp. 2.35×10^{-16}	1
hg18 STAT1 HeLaStim	 C Stat3	 J NFE2L1::MafG			+1, Same 1.03×10^{-14}	
hg18 STAT1 HeLaStim	 C Stat3	 U Hdx			+7, Opp. 7.89×10^{-12}	
hg18 STAT1 HeLaStim	 C Stat3	 C Stat3			+10, Same 9.44×10^{-08}	S

See Table 2 caption for explanation of columns. 1: This observation is supported by evidence from ref. (40).

hindrances between the GABP and CREB1 proteins, although the protein structures are located close to each other in the model. The model predicts two interactions between the proteins, with CREB1 contacting both the GABPα and GABPβ subunits of the GABP protein. First, the N-terminus of one CREB1 alpha-helical subunit contacts an alpha helix of GABPα in the major groove of the DNA. Second, the N-terminus of the remaining CREB1 subunit is positioned close to a loop in GABPβ (Figure 2A). These results are consistent with the findings of Sawada *et al.* (39), who identified an interaction between CREB1 and hGABPα, and found that hGABPβ increases the affinity of the GABP–CREB1 interaction. The putative interactions, lack of steric hindrance and literature conformity all support the accuracy of the model.

As a further illustration of SpaMo's ability to yield correct TF complex structure information, we have compared the inferred SRF/ETS motif spacing with the known ternary complex involving SRF, SAP-1 (which binds to the ETS motif) and the *c-fos* promoter DNA determined by Mo *et al.* (26) using X-ray crystallography (PDB accession 1K6O). The known structure shows an interaction between the N-terminal loops of the two DBDs at a distance of 4 Å, over the minor groove of the DNA. The *c-fos* promoter DNA sequence used in the structure has a spacing of zero nucleotides between the SRF and ETS motif occurrences (the sequence is

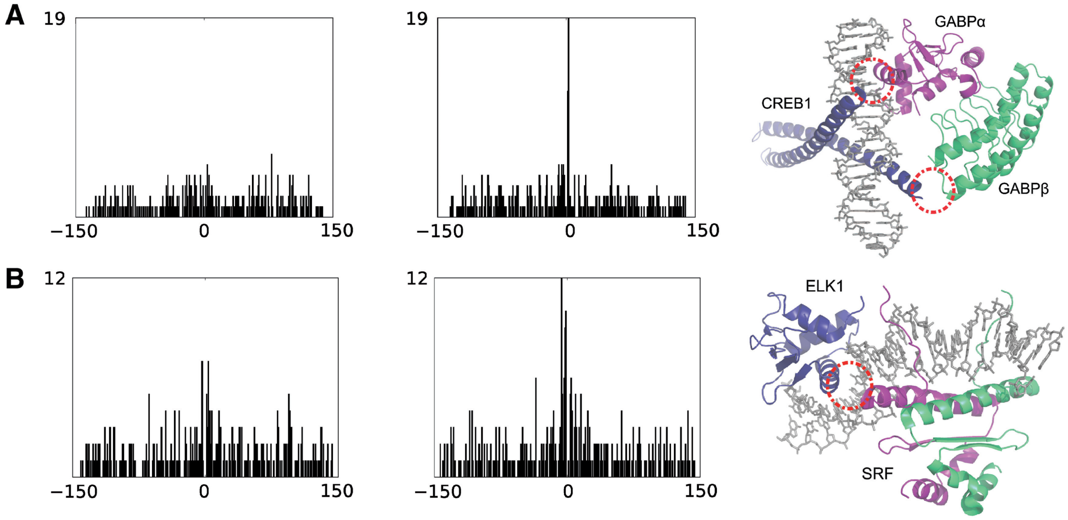


Figure 2. Ternary complex structure elucidation. (**A**) Displacement histograms for GABP/CREB1, with corresponding predicted GABP/CREB1/DNA ternary complex structure. The distance indicated by the red dotted line is 6.8 Å. This is the minimum distance between any pair of GABP and CREB1 atoms at this estimated contact point. (**B**) Displacement histograms for SRF/ETS, with corresponding known SRF-ELK1 ternary complex structure (PDB accession 1K6O).

5′TGTCCTAATATGG*ACATCC*3′, with the SRF and ETS motif occurrences underlined and italicised, respectively). The enriched motif displacement interval identified by SpaMo includes this spacing, and suggests that SRF-ETS motif distances of 1–6 are commonplace. Additionally, the observed enriched spacings indicate that interactions can occur on either side of SRF. Both observations are consistent with the known crystal structure, as the interaction domains of both proteins are flexible loops, with SRF exhibiting an identical structure at both ends of its DBD (Figure 2B).

DISCUSSION

The strength of our method is its ability to infer TF complexes from ChIP-seq data with a high positive predictive value. Various computational studies have addressed the issue of cooperative TF binding (41,42). However, previous methods are not targeted at inferring the presence of TF complexes from ChIP-seq data sets, instead of aiming to extract motif associations from unfocused genomic sequence data (43,44). SpaMo is developed specifically to harness the power and resolution provided by ChIP-seq data, and yields information specific to the input transcription factor and tissue in which ChIP-seq was carried out.

Motif enrichment analysis (MEA) has previously been applied to ChIP-seq data to identify TFs that co-regulate gene expression with the TF of interest (45). MEA assesses whether individual motifs occur more frequently than expected by chance in the input DNA sequences (4,5). When MEA is applied to ChIP-seq data, enrichment of motifs other than the ChIP-ed motif does not necessarily imply the presence of a physical TF complex since the definition of enrichment does not require any particular spatial relationship between the ChIP-ed motif and the secondary motif. In contrast, we focus our analysis on a primary motif that is known to be relevant to the TF (e.g. the motif for the TF's DBD), and we assess whether individual secondary motifs exhibit enriched spacing with respect to the primary motif. This approach specifically identifies TF complexes, which we have demonstrated by detecting known and high confidence novel TF complexes, using existing ChIP-seq data sets.

The mammalian two-hybrid (M2H) system was recently employed to detect protein–protein interactions between TFs, from a comprehensive set of human and mouse TFs (7). This application of the M2H approach was subject to three limitations that are overcome by our method. First, M2H was employed to study direct interactions between TFs. Thus, a complex between two TFs that occur indirectly via a bridging protein will not be detected. For example, the authors do not report a complex between GATA-1 and SCL, presumably because GATA-1 and SCL interact indirectly, via LDB1 and E47 in the known GATA-1/SCL/E47/Ldb1 complex (12). Our method is able to identify the complex between GATA-1 and SCL in both the human and mouse GATA-1 ChIP-seq data sets. Second, the M2H analysis measures binding between TFs without considering the role of DNA in stabilizing the interaction between the two TFs. For some TF complexes, the DNA may play a critical role in reducing the free energy of complex formation. Third, M2H can identify physical complexes, but cannot identify the genomic regions at which those complexes bind *in vivo*. In contrast, SpaMo infers the likely genomic loci of complex formation, as it isolates the sequences containing the enriched motif spacing.

In 39 of our 41 analyses, the primary motif represents the DNA-binding specificity of the DBD for the TF investigated with ChIP-seq. The remaining two analyses are alternative analyses of c-Fos ChIP-seq data sets, in which we employed a primary motif derived by running *ab initio* motif discovery on c-Fos ChIP-seq data. This motif does not represent the known binding specificity of c-Fos itself. However, by employing this motif as the primary in our analysis, we obtained a distinct set of high-confidence TF complex predictions, compared with results obtained using the c-Fos DBD motif. This demonstrates that it can be worthwhile repeating SpaMo analysis using alternative biologically relevant motifs as the primary, in addition to using a motif based on DBD specificity.

In this study, we have used SpaMo with a width parameter of 1 bp to predict numerous TF complexes exhibiting tight motif spacing patterns. In contrast, we identified relatively few broad motif spacings, which suggest clusters of independent binding sites. Clusters of inconsistently spaced binding sites have been observed in various systems, and can mediate a specific rate at which transcription responds to TF concentration (33). Using a larger width parameter with SpaMo should increase the sensitivity of SpaMo to detecting these clusters, although that is not the primary goal of the algorithm.

ChIP-seq technology facilitates high-resolution estimates of TF binding. In combination with complementary methods such as MEA and *ab initio* motif discovery, motif spacing analysis with SpaMo should assist researchers with maximizing biological knowledge extracted from ChIP-seq data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Michael Tallack, Andrew Perkins, Michael Landsberg and members of the Bailey group for useful discussions and comments. We thank Jane Ren and Charles Grant for help integrating SpaMo into the MEME Suite. T.W. developed the SpaMo algorithm. T.L.B. and M.C. refined the methodology. T.W. implemented the algorithm, ran SpaMo on the ChIP-seq data sets and processed the results. J.J. reimplemented the algorithm in C as part of the MEME Suite. T.W. and T.L.B. wrote the manuscript. M.C. edited the manuscript.

FUNDING

Australian Postgraduate Award (to T.W.); Japanese National Institute of Advanced Industrial Science and Technology general research funding grant; National Institute of Advanced Industrial Science and Technology (AIST) general research funds (to M.C.F.); National Institutes of Health (NIH) grant (R0-1 RR021692-05 to T.L.B. and J.J). Funding for open access charge: National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Valouev,A., Johnson,D.S., Sundquist,A., Medina,C., Anton,E., Batzoglou,S., Myers,R.M. and Sidow,A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829–834.
2. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
3. Frith,M.C., Hansen,U., Spouge,J.L. and Weng,Z. (2004) Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res.*, **32**, 189–200.
4. 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.
5. Roider,H.G., Manke,T., O'Keeffe,S., Vingron,M. and Haas,S.A. (2009) PASTAA: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics*, **25**, 435–442.
6. Wolberger,C. (1999) Multiprotein-DNA complexes in transcriptional regulation. *Ann. Rev. Biophys. Biomol. Struct.*, **28**, 29–56.
7. Ravasi,T., Suzuki,H., Cannistraci,C.V., Katayama,S., Bajic,V.B., Tan,K., Akalin,A., Schmeier,S., Kanamori-Katayama,M., Bertin,N. *et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**, 744–752.
8. Chen,L., Glover,J., Hogan,P., Rao,A. and Harrison,S. (1998) Structure of the DNA binding domains from NFAT, Fos and Jun bound specifically to DNA. *Nature*, **392**, 42–48.
9. Ptashne,M. and Gann,A. (2002) *Genes and Signals*. Cold Spring Harbor Lab Press, Cold Spring Harbor, New York.
10. Karin,M., Liu,Z. and Zandi,E. (1997) AP-1 function and regulation. *Curr. Opin. Cell Biol.*, **9**, 240–246.
11. Nair,S. and Burley,S. (2003) X-ray structures of Myc-Max and Mad-Max recognizing DNA: molecular bases of regulation by proto-oncogenic transcription factors. *Cell*, **112**, 193–205.
12. Wadman,I.A., Osada,H., Grtz,G.G., Agulnick,A.D., Westphal,H., Forster,A. and Rabbitts,T.H. (1997) The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *EMBO J.*, **16**, 3145–3157.
13. ENCODE Project Consortium. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
14. Karolchik,D., Hinrichs,A., Furey,T., Roskin,K., Sugnet,C., Haussler,D. and Kent,W. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
15. Sandelin,A., Alkema,W., Engström,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.
16. Newburger,D.E. and Bulyk,M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **37**, D77–D82.
17. DeLano,W. (2002) *The PyMOL Molecular Graphics System*. DeLano Scientific, San Carlos, CA, USA.
18. Brown,R.A. and Case,D.A. (2006) Second derivatives in generalized born theory. *J. Comput. Chem.*, **27**, 1662–1675.
19. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
20. Berger,M.F. and Bulyk,M.L. (2009) Universal protein-binding microarrays for the comprehensive characterization of the dna-binding specificities of transcription factors. *Nat. Protoc.*, **4**, 393–411.
21. Pearson,R., Fleetwood,J., Eaton,S., Crossley,M. and Bao,S. (2008) Kruppel-like transcription factors: a functional family. *Int. J. Biochem. Cell Biol.*, **40**, 1996–2001.
22. Horard,B., Castet,A., Bardet,P., Laudet,V., Cavailles,V. and Vanacker,J. (2004) Dimerization is required for transactivation by estrogen-receptor-related (ERR) orphan receptors: evidence from amphioxus ERR. *J. Mol. Endocrinol.*, **33**, 493–509.
23. Cai,D.H., Wang,D., Keefer,J., Yeamans,C., Hensley,K. and Friedman,A.D. (2008) C/EBP alpha: AP-1 leucine zipper heterodimers bind novel DNA elements, activate the PU.1

promoter and direct monocyte lineage commitment more potently than C/EBP alpha homodimers or AP-1. *Oncogene*, **27**, 2772–2779.

24. Wei,P., Taniguchi,S., Sakai,Y., Imamura,M., Inoguchi,T., Nawata,H., Oda,S., Nakabeppu,Y., Nishimura,J. and Ikuyama,S. (2005) Expression of adipose differentiation-related protein (ADRP) is conjointly regulated by PU.1 and AP-1 in macrophages. *J. Biochem.*, **138**, 399–412.
25. Williams,D., Cai,M. and Clore,G. (2004) Molecular basis for synergistic transcriptional activation by Oct1 and Sox2 revealed from the solution structure of the 42-kDa Oct1 center dot Sox2 center dot Hoxb1-DNA ternary transcription factor complex. *J. Biol. Chem.*, **279**, 1449–1457.
26. Nateri,A., Spencer-Dene,B. and Behrens,A. (2005) Interaction of phosphorylated c-Jun with TCF4 regulates intestinal cancer development. *Nature*, **437**, 281–285.
27. Mo,Y., Ho,W., Johnston,K. and Marmorstein,R. (2001) Crystal structure of a ternary SAP-1/SRF/c-Fos SIRE DNA complex. *J. Mol. Biol.*, **314**, 495–506.
28. Schlisio,S., Halperin,T., Vidal,M. and Nevins,J. (2002) Interaction of YY1 with E2Fs, mediated by RYBP, provides a mechanism for specificity of E2F function. *EMBO J.*, **21**, 5775–5786.
29. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
30. Escalante,C., Shen,L., Escalante,M., Brass,A., Edwards,T., Singh,H. and Aggarwal,A. (2002) Crystallization and characterization of PU.1/IRF-4/DNA ternary complex. *J. Struct. Biol.*, **139**, 55–59.
31. Rengarajan,J., Mowen,K., McBride,K., Smith,E., Singh,H. and Glimcher,L. (2002) Interferon regulatory factor 4 (IRF4) interacts with NFATc2 to modulate interleukin 4 gene expression. *J. Exp. Med.*, **195**, 1003–1012.
32. Jin,Y., Mead,J., Li,T., Wolberger,C. and Vershon,A. (1995) Altered DNA recognition and bending by insertions in the alpha-2 tail of the yeast a1/alpha-2 homeodomain heterodimer. *Science*, **270**, 290–293.
33. Giorgetti,L., Siggers,T., Tiana,G., Caprara,G., Notarbartolo,S., Corona,T., Pasparakis,M., Milani,P., Bulyk,M.L. and Natoli,G. (2010) Noncooperative interactions between transcription factors and clustered DNA binding sites enable graded transcriptional responses to environmental inputs. *Mol. Cell*, **37**, 418–428.
34. Rosmarin,A., Resendes,K., Yang,Z., McMillan,J. and Fleming,S. (2004) GA-binding protein transcription factor: a review of GABP as an integrator of intracellular signaling and protein-protein interactions. *Blood Cells Mol. Dis.*, **32**, 143–154.
35. Vilaboa,N., Bermejo,R., Martinez,P., Bornstein,R. and Cales,C. (2004) A novel E2 box-GATA element modulates Cdc6 transcription during human cells polyploidization. *Nucleic Acids Res.*, **32**, 6454–6467.
36. Rozowsky,J., Euskirchen,G., Auerbach,R.K., Zhang,Z.D., Gibson,T., Bjornson,R., Carriero,N., Snyder,M. and Gerstein,M.B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.
37. Vinkemeier,U., Moarefi,I., Darnell,J.E. and Kuriyan,J. (1998) Structure of the amino-terminal protein interaction domain of stat-4. *Science*, **279**, 1048–1052.
38. Berman,H., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T., Weissig,H., Shindyalov,I. and Bourne,P. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
39. Sawada,J., Simizu,N., Suzuki,F., Sawa,C., Goto,M., Hasegawa,M., Imai,T., Watanabe,H. and Handa,H. (1999) Synergistic transcriptional activation by hGABP and select members of the activation transcription factor/cAMP response element-binding protein family. *J. Biol. Chem.*, **274**, 35475–35482.
40. Muhlethaler-Mottet,A., Di Berardino,W., Otten,L. and Mach,B. (1998) Activation of the MHC class ii transactivator CIITA by interferon-gamma requires cooperative interaction between Stat1 and USF-1. *Immunity*, **8**, 157–166.
41. Frith,M.C., Saunders,N.F.W., Kobe,B. and Bailey,T.L. (2008) Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput. Biol.*, **4**, e1000071.
42. GuhaThakurta,D. and Stormo,G. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics*, **17**, 608–621.
43. Hannenhalli,S. and Levy,S. (2002) Predicting transcription factor synergism. *Nucleic Acids Res.*, **30**, 4278–4284.
44. Yokoyama,K.D., Ohler,U. and Wray,G.A. (2009) Measuring spatial preferences at fine-scale resolution identifies known and novel cis-regulatory element candidates and functional motif-pair relationships. *Nucleic Acids Res.*, **37**.
45. Yu,M., Riva,L., Xie,H., Schindler,Y., Moran,T.B., Cheng,Y., Yu,D., Hardison,R., Weiss,M.J., Orkin,S.H. *et al.* (2009) Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Mol. Cell*, **36**, 682–695.