Published online 6 January 2012

Nucleic Acids Research, 2012, Vol. 40, No. 7 ***e50***
doi:10.1093/nar/gkr1135

A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information

Xiaotu Ma[1], Ashwinikumar Kulkarni[1], Zhihua Zhang[1], Zhenyu Xuan[1], Robert Serfling[2] and Michael Q. Zhang[1,3,*]

[1]Department of Molecular and Cell Biology, Center for Systems Biology, [2]Department of Mathematics, University of Texas at Dallas, 800 W. Campbell Road, Richardson, TX 75080, USA and [3]Division of Bioinformatics, Center for Synthetic and Systems Biology, TNLIST, Tsinghua University, Beijing 100084, China

Received August 16, 2011; Revised October 28, 2011; Accepted November 8, 2011

ABSTRACT

Identification of DNA motifs from ChIP-seq/ChIP-chip [chromatin immunoprecipitation (ChIP)] data is a powerful method for understanding the transcriptional regulatory network. However, most established methods are designed for small sample sizes and are inefficient for ChIP data. Here we propose a new *k*-mer occurrence model to reflect the fact that functional DNA *k*-mers often cluster around ChIP peak summits. With this model, we introduced a new measure to discover functional *k*-mers. Using simulation, we demonstrated that our method is more robust against noises in ChIP data than available methods. A novel word clustering method is also implemented to group similar *k*-mers into position weight matrices (PWMs). Our method was applied to a diverse set of ChIP experiments to demonstrate its high sensitivity and specificity. Importantly, our method is much faster than several other methods for large sample sizes. Thus, we have developed an efficient and effective motif discovery method for ChIP experiments.

INTRODUCTION

Decoding the transcriptional regulatory network is a challenging task in molecular biology (1,2). In human, despite the estimated 1391 sequence-specific DNA-binding transcription factors, only ~60 of them have been experimentally verified for both DNA-binding and regulatory functions (3). As of January, 2011, there were only 75 matrix models describing the binding motifs of human transcription factors in the JASPAR database (4). With the rapid development of high-throughput DNA sequencing technology, it is now popular to experimentally map the genome-wide binding regions of transcription factors using chromatin immunoprecipitation (ChIP) coupled with massively parallel sequencing technology (ChIP-seq) or microarray (ChIP-chip) (1,2). For example, binding regions of 23 worm transcription factors (5) and 103 fly transcription factors (6) have been studied in single projects. Identification of functional DNA-motifs from such data may provide valuable resources for modeling the transcription regulatory networks. Although the resolution of binding regions identified from ChIP-seq can be a few hundred base pairs (2), it has been found (7) that existing iterative motif discovery methods, e.g. MEME (8), do not have the computational efficiency required to process the huge amount of data from ChIP-seq/ChIP-chip experiments.

On the other hand, modeling and discovery of DNA motifs from a set of DNA sequences have been a major research focus in computational biology (1,2,9). In the earlier works (8–18), it is generally assumed that the underlying DNA motifs to be discovered are enriched in certain regions (e.g. promoters of co-expressed genes) without any positional preference. Since some transcription factors are known to bind DNA regions close to 5′ transcription start site (TSS) of their target genes (19), Linhart *et al*. (20) introduced a binomial test to determine if a given DNA motif tends to appear in certain bins of the 5′TSS regions of genes. Kim *et al*. (21) introduced a Bayesian model to incorporate positional bias of transcription factor binding sites (TFBSs) in promoter regions to discover DNA motifs. Narang *et al*. (22) introduced a spatial confinement score combined with an overrepresentation score and relative entropy score to discover DNA motifs. Using positional preference for DNA motif discovery was most recently revisited by Keilwagen *et al*. (23).

*To whom correspondence should be addressed. Tel: 516 367 8393; Fax: 516 367 8461; Email: michael.zhang@utdallas.edu; mzhang@cshl.edu

With the ChIP-seq/ChIP-chip technique, positional information is more evident in such data sets. For example, peak intensity profiles were used a priori (24) to accelerate the optimization process. Such intensity profiles were also used to score PWMs (25). Although the above-mentioned motif discovery tools have achieved success in many scenarios, positional information has not been fully exploited for motif discovery. For example, it was found that the underlying DNA motifs are distributed more frequently around the summits of peaks than in the flanking regions of the peaks (26). While the intensity profiles used by Hu *et al*. (24) and Kulakovskiy *et al*. (25) contain positional information, estimation accuracy of the intensity profiles at peaks with low-read coverage was error prone. In addition, both fragment length and distribution of the underlying DNA motif may affect the peak intensity profile (27) that often renders the determination of 'peak segments' for motif discovery *ad hoc*. Also, it is unknown how to optimally specify the start and end points of the detected ChIP peaks for most currently available motif discovery software. As a result, 'foreground' sequences are often determined using arbitrary thresholds. For example, in Jothi *et al*. (7) and Hu *et al*. (24), a 200-bp region centered on the peak summit is used, while a region having a 1000-bp length is used in Corbo *et al*. (28). It is thus interesting to ask how the positional information can be best utilized for DNA motif discovery.

In this work, we first noticed that in a typical ChIP experiment for a sequence-specific transcription factor, the functional DNA motif interacting with the studied protein tends to cluster around the peak summit (26). Based on this observation, we propose a Gaussian-uniform mixture model to describe the positional patterns of *k*-mers relative to the peak summit. A scoring method is also proposed to quickly rank and discover *k*-mers from ChIP data. A positional information guided motif discovery software, termed POSMO, is then implemented. In the following, using both simulated and real data sets, we will demonstrate the higher effectiveness and efficiency of POSMO than available software tools.

MATERIALS AND METHODS

In the following, we assume that the ChIP-seq/ChIP-chip experiments are for sequence-specific DNA-binding transcription factors. It may not be suitable for ChIP-seq/ChIP-chip experiments for non-specific DNA binding proteins, such as histones.

Data sets used in this work

To demonstrate the practical use of our method, we obtained ChIP-seq data for STAT1 (29), CRX (28), CTCF (30), NRSF (31) and FOXA2 (32). To validate the performance of our method for ChIP-chip data, we obtained data for CTCF in human (33). To validate the performance of our method on other species, we obtained ChIP-seq data for CAD, KNI, KR1, KR2, BCD, HB1 and HB2 of *Drosophila melanogaster* (34). We also demonstrated the usefulness of our method on a large cohort of core transcription factors involved in mouse embryonic stem cells (35). Binding motifs of the above transcription factors are either documented in the JASPAR database (4) or in the original publications, facilitating the comparison between our method and other available methods.

Model for motif discovery

Naturally, after peak calling on ChIP-seq/ChIP-chip data, we have a set of chromosomal positions about the potential binding events of a given transcription factor. Typically, peak calling software also reports a 'summit' for each peak [e.g. MACS (36), SISSRs (7), for a recent comparison see (37)]. Therefore, these peaks can be uniquely aligned according to the location (μ_0) of their respective summits. Next, based on the nature of the ChIP-seq/ChIP-chip experiment (7,29,36), we assume that a functional DNA motif exists in the vicinity of peak summits. The DNA motif is a *k*-mer with unknown *k*. In practice, many such *k*-mers may exist as a result of degeneracy and a word-clustering algorithm is employed to group them together. Location (X) of such motifs is assumed to follow a Gaussian distribution: $X \sim N(\mu_0, \sigma^2)$, where μ_0 is the peak summit and σ is an unknown parameter related to the binding nature of the transcription factor being studied, the noise level (e.g. antibody specificity) of the ChIP experiment as well as the noise in the sequencing step. In addition, σ is small compared to the flanking regions of each candidate peak. This assumption is quite reasonable since the underlying functional DNA motifs are generally enriched in peak summit regions with length of a few hundred base pairs or less (7,29,36), and we can freely increase the length (e.g. $\pm$5000 bp was used in this work; denoted as m) of flanking sequences of each peak. Finally, as a background model, the given *k*-mer is assumed to be uniformly distributed in the flanking regions of the identified peaks. With these assumptions, we have

$$X \sim \alpha \times N(\mu_0, \sigma^2) + (1-\alpha) \times U[\mu_0 - m, \mu_0 + m] \quad (1)$$

where α is an unknown enrichment parameter between 0 and 1 and is specific to the *k*-mer. Inferences on α and σ for a given *k*-mer can be made when there is a sufficient number of observations. In principle, a maximum likelihood estimation of α and σ can be obtained by optimization methods (38). However, since a majority of *k*-mers are usually unrelated to the transcription factor, directly solving the above mixture model is not computationally efficient. In fact, we introduced a novel statistic to score and rank each *k*-mer as follows.

Since the peak sequences with exactly the same lengths are aligned by the summit (Supplementary Figure S1A), we can count the appearance frequency of each *k*-mer at a particular position relative to the peak summit across all aligned peak sequences. In other words, we have an appearance frequency profile ($A_1, A_2, \ldots, A_{2m}$) for each *k*-mer (Supplementary Figure S1B), where A_i is the total times of observing a given *k*-mer at position i of all peaks. According to the above mixture model, the appearance frequency profile will be relatively higher when the position index i is close to peak summit μ_0, provided

that the corresponding k-mer is the binding motif of the investigated transcription factor. Clearly, a significant jump must be observed around the summit region (Supplementary Figure S1C) when the appearance frequency profile of a k-mer related to the studied transcription factor is converted into a cumulative appearance frequency profile (CAFP). We thus adopted scoring such a jump. Since a higher jump corresponds to larger area between the observed CAFP and the diagonal line (corresponding to $\alpha = 0$), we used the area (R) between the CAFP and the diagonal (see Supplementary Text 1 for detailed definition) to score each k-mer, and the obtained score is hereinafter referred to as POSMO R score. We allow this area (R) to have a negative value, which corresponds to cases where a given k-mer is depleted around the summit region, but enriched in flanking regions. As shown in Supplementary Text 1, we proved that the POSMO R score asymptotically follows a normal distribution $N(0, 1/48T)$ for large T (say >60), where T represents total occurrences of a given k-mer. With this distribution, we can evaluate the statistical significance of each k-mer by POSMO R score. Such significance scores, termed POSMO Z scores, are then used to rank k-mers (Supplementary Figure S1D).

Although the above POSMO Z score is highly effective in ranking the true k-mers on top, it obviously does not efficiently account for σ in Equation (1). Thus, for the purpose of efficiency, we proposed an approximate solution to estimate σ using linear models. As can be seen in Supplementary Figure S1C, the CAFP has two linear components of the same slope in the flanking regions. In other words, we can model the linear components in the flanking regions by:

$$\text{Left flanking region: } Y_i = aX_i + b \quad i = 1, 2, ..., m$$

$$\text{Right flanking region: } Y_j = aX_j + c \quad j = m+1, m+2, ..., 2m$$

Estimators $\hat{a}$, $\hat{b}$ and $\hat{c}$ can be derived using least square methods for the above model. We then approximately estimate σ of the Gaussian component by checking the residual of the linear fitting of the two flanking regions. Since the profile in Supplementary Figure S1C is monotonic, we can calculate the residual at each position using the linear model on left flanking region and right flanking region. The start (s) and end (e) positions where the residual first exceeds $0.2 \times (\hat{c} - b)$ are estimated by:

$$\hat{s} = \arg\min\{i : Y_i - \hat{a}X_i - \hat{b} > 0.2 \times (\hat{c} - \hat{b})\}$$

$$\hat{e} = \arg\max\{i : Y_i - \hat{a}X_i - \hat{c} > 0.2 \times (\hat{b} - \hat{c})\}$$

The distance between these two positions ($D = \hat{e} - \hat{s}$; hereinafter termed D score) is a second filter in addition to the above POSMO Z score. A true k-mer will have a relatively smaller positive D score than other k-mers. Clearly, the complexity to calculate the above POSMO R score and D score is linear with the length of the flanking regions.

Thresholds

Obviously, a majority of the 4^k k-mers are unrelated to the transcription factor being studied. With the above POSMO Z score and D score for each k-mer, we next want to detect 'significant k-mers' for further analysis. For this purpose, we only consider positive D scores that correspond to k-mers enriched in peak summit regions. A k-mer will be retained if its D score is small, but non-negative:

$$0 \leq D < \mu_D + t_D \times \sigma_D 0$$

where t_D is set to 1.645 (corresponding to one-sided P-value of 0.05) in this work.

Next, we checked the population mean (μ_Z) and standard deviation (σ_Z) of the POSMO Z scores of each k-mer. A k-mer will be filtered out if its POSMO Z score is small:

$$Z < \mu_Z + t_Z \times \sigma_Z$$

where t_Z is set to 2.33 (corresponding to one-sided P-value of 0.01) in this work. k-mers satisfying the above two filters are called 'significant k-mers' and will be subject to word clustering, as described in the next section. We note that the above thresholds (t_D and t_Z) are quite arbitrary and could be fine-tuned for specific data sets. However, in the present work, we found that our method is robust to these parameters (Supplementary Table S3).

Word clustering

Although the desired k-mers are generally within the significant k-mer list, it is difficult to manually inspect the whole list of significant k-mers. In addition, different variations of the same binding motif exist in the list of significant k-mers due to degeneracy. Thus, a traditional PWM representation may be more informative. This raised a question of k-mer clustering, which is still an open problem in bioinformatics. Different methods for clustering k-mers have been proposed. For example, Schones *et al.* (39) found that the chi-square statistics and Fisher-Irwin test are good measurements of PWM similarity. Mahony *et al.* (40) studied the effectiveness of different similarity metrics and tree-building methods on grouping k-mers from an analysis of variance (ANOVA) perspective. They found that the Pearson correlation coefficient is a good similarity measure between PWMs. However, we found that their metric on automatic determination of the number of clusters is generally not satisfactory for our PWMs, i.e. the CH_{log} statistic by Mahony *et al.* (40) does not always generate reasonable global minima to determine optimal cluster numbers (data not shown). Thus, we developed a simpler, yet effective, method to group the significant k-mers by considering context, as described below.

Specifically, with the significant k-mers, we rescan the whole data set of peak sequences, e.g. (−5, 5 kb) flanking region of the peak summits. Each significant k-mer as defined in section 'Thresholds' is assigned its POSMO Z score, while insignificant k-mers are set to zero. For each

continuous sequence segment with score >0, we extract its flanking regions (e.g. ±20 bp) surrounding the k-mer (w_0) with highest score and label this sequence segment using corresponding k-mer (w_0). This step is particularly useful when our k is smaller than the optimal k_0, in which case several k-mers are highly significant because they are simple shifts of the same DNA motif. DNA segments are then grouped according to their labeling k-mers, respectively. For each of these labeling k-mers, a position specific frequency matrix (PWM) is obtained. Next, each PWM is converted into a $4 \times w$ vector where w ($>k$) is a predetermined length (here we used $w = k+2$). Different PWMs are compared using Pearson's correlation coefficient over its region with maximum information content, where the information content is defined as:

$$\frac{1}{w}\sum_{i,k} p_{i,k}\log_2(4p_{i,k})$$

where $p_{i,k}$ is the frequency of observing nucleotide k at position $I = 1,2,\ldots w$. Different offsets are tried to best match a pair of PWMs. At each step of our hierarchical clustering, the pair of PWMs with highest similarity score is joined, and a new PWM is constructed in the clustering tree. We iterate this process until only one node is left. We found that the similarity score between joined nodes in each step decreases when the tree level is close to the root node. With this observation, we determine the final number of clusters by the tree level which is closest to the root node and which has similarity score over a predefined threshold T_0. We found that setting T_0 between (0.8, 0.9) generally gives reasonably good results. A sequence logo was, in turn, generated using the seqLogo package in the BioConductor open source software package, as was done in Hu *et al.* (24).

Comparison with other algorithms on ranking k-mers by POSMO *Z* score

We first set out to compare our algorithm with other motif-discovery methods. For this purpose, we compared our algorithm with DME (41), as a representative of enumerative methods, and MEME (8), as a representative of iterative optimization methods. The comparison was carried out with simulated data using the following procedure:

(1) simulate 2000 sequences each with length L (e.g. 10 000 bp);
(2) half (here 1000) of the sequences are randomly chosen as foreground data and another half (1000) as control data;
(3) a k-mer w_0 is generated to be planted into 50 sequences [corresponding to $\alpha = 0.24$ in Equation (1), as the expected number of a given 8-mer in background will be $1000 \times 10\,000/4^8 \approx 153$ and $50/(153+50) = 0.24$], foreground sequences in step 4;
(4) for each of the 50 sequences (see above) selected as foreground, a random integer x (>0 and $<10\,000$-k) is sampled from Gaussian distribution $N(\mu, \sigma^2)$, where μ (here 5000) controls the location of the real peaks (to be discovered by ChIP experiments), and σ controls the spread of the peaks. The k-mer at position x is replaced by the k-mer in step 3);
(5) all the foreground sequences are processed by POSMO;
(6) for each foreground and background sequence, the substrings from position $5000-l$ to $5000+l$ are extracted to compile a new foreground data set and a new background data set, where l is set as 100 bp. These two data sets are input to DME and MEME as foreground and background, respectively;
(7) the rank of the known k-mer in step (3) by POSMO, DME and MEME is recorded; and
(8) steps (1) through (6) are repeated many (e.g. 500) times to summarize the rank distribution of the known k-mers. A lower rank of the target k-mer is better.

Implementation

We have implemented our algorithm using C++. The POSMO program can be freely downloaded from http://cb.utdallas.edu/Posmo/index.html.

RESULTS

Simulation study

We first used simulation to compare our POSMO algorithm with established methods, such as exhaustive enumeration [e.g. YMF (42), DWE (43) and DME (41)] and iterative optimization [e.g. MEME (8)]. We decided to compare our algorithm with DME and MEME as representatives of the above two categories. As it turned out, our method performs in a manner similar to DME and MEME in ranking the target k-mers on top when the underlying distribution of the target motif is well correlated with the ChIP peak (Supplementary Figure S2A). In a typical ChIP experiment, the cross-linked DNA sequences are sheared into desired length, and these smaller DNA segments are then sequenced. However, some transcription factors may interact with co-factors, which, in turn, lead to sharper or flatter peaks. These unknown parameters will affect the spread of the motif distribution under the ChIP peaks, as modeled by σ in Equation (1). As can be seen from Supplementary Figure S2B, performance of POSMO is more robust to larger σ as compared to that of MEME and DME. In addition, the ChIP peaks may have systematic shift from the true binding site (36,44). As can be seen from Supplementary Figure S2C, POSMO is much more robust against such a systematic error than either MEME or DME. This is not surprising since the performance of both MEME and DME is dependent on the accuracy of the foreground sequences. In this sense our simulation method is in favor of our POSMO methods, since MEME and DME do not consider the positional preferences at all. On the other hand, POSMO is able to find the target motif without the need of explicitly specifying foreground and background sequences since it implicitly contrasts the 'peak center' with the flanking region. Based on

Table 1. Top five *k*-mers ranked by POSMO are related to the underlying transcription factor-DNA interaction

k-mer	*Z*	PWM	*k*-mer	*Z*	PWM	*k*-mer	*Z*	PWM	*k*-mer	*Z*	PWM
STAT1 (29)			**FOXA2** (32)			**NRSF** (31)			**CRX** (28)		
TCCTGGAA	24	13	GTAAACAA	12	3	GGTGCTGA	31	13	CTAATCCC	9	12
TTCCAGGA	24	13	TTGTTTAC	12	3	TCAGCACC	31	13	GGGATTAG	9	12
TCCAGGAA	24	13	AGTAAACA	11	15	GTGCTGAA	26	13	GCTAATCC	8	−6
TTCCTGGA	24	13	TGTTTACT	11	15	TTCAGCAC	26	13	GGATTAGC	8	−6
TCCGGGAA	20	12	TGTAAACA	11	15	CGCTGTCC	25	12	GAGGATTA	8	−5
CTCF (30)			**CTCF** (33)			**GT** (34)			**BCD** (34)		
AGGGGGCG	23	11	AGAGGGCG	23	11	TTGCGCAA	7	9	GGATTA	10	12
CGCCCCCT	23	11	CGCCCTCT	23	11	TGACGCAA	5	8	TAATCC	10	12
GCGCCCCC	22	10	CAGAGGGC	18	11	TTGCGTCA	5	8	CTAATC	5	−10
GGGGGCGC	22	10	GCCCTCTG	18	11	TGACGTAA	5	12	GATTAG	5	−10
GAGGGCGC	22	10	ACCAGGGG	15	11	TTACGTCA	5	12	AAGCCG	3	−14
KR1 (34)			**KR2** (34)			**HB1** (34)			**HB2** (34)		
AAAGGGTT	17	14	AAAGGGTT	16	14	GTAAAAAA	12	13	GTAAAAAA	12	13
AACCCTTT	17	14	AACCCTTT	16	14	TTTTTTAC	12	13	TTTTTTAC	12	13
AAGGGTTA	15	1	AAGGGTTA	14	1	GCAAAAAA	12	12	GCAAAAAA	12	12
TAACCCTT	15	1	TAACCCTT	14	1	TTTTTTGC	12	12	TTTTTTGC	12	12
AGGGTTAA	12	−29	AGGGTTAA	11	−29	AAAAAACG	11	7	AAAAAACG	11	7
KNI (34)			**CAD** (34)								
CAATATTG	2	−2	CAGGTAG	13	−8						
GTCCGCAC	1	−4	CTACCTG	13	−8						
GTGCGGAC	1	−4	GCAGGTA	10	−13						
ACGGCCGT	1	−11	TACCTGC	10	−13						
CCGGATCG	1	1	CCAGGTA	7	−9						

Dark shaded cells represent either significant *k*-mers called by our POSMO algorithm (*Z* columns), or a *k*-mer with a significant PWM score (PWM columns; $>\mu + 3\sigma$ criterion over all 4^k *k*-mers). Light-gray shaded cells represent *k*-mers which are shifts of the genuine motif, thus having an insignificant PWM score. POSMO *Z* score is the average POSMO *Z* score of a *k*-mer and its reverse complementary *k*-mer. *k*-mers for each transcription factor are sorted according to POSMO *Z* score (Z columns).

the above simulation results, we proceeded to apply our method to real ChIP data sets in the next sections.

Application on real data

ChIP-seq on STAT1, NRSF, CTCF, CRX and FOXA2. We first applied our POSMO algorithm on ChIP peaks identified by Jothi *et al*. (7) on STAT1 (29). A refined motif of STAT1 was recently reported by Hu *et al*. (24). We therefore studied the performance of POSMO in ranking *k*-mers for STAT1-binding motif, using the PWM of STAT1 by Hu *et al*. (24) as the gold standard. By focusing on the 4741 top peaks of STAT1 ChIP data [NumTags >50 in Jothi *et al*. (7); see Table 4 for robustness of our method against the number of top peaks used], we found that 8-mers directly related to STAT1 binding are indeed ranked on top (Table 1 and Supplementary Table S1).

We next asked whether POSMO could generally rank *k*-mers related to studied transcription factors on top. To accomplish this, we collected ChIP-seq data for CRX (28), CTCF (30), NRSF (31) and FOXA2 (32). As can be seen from Table 1 and Supplementary Table S1, POSMO successfully ranked the desired *k*-mers on top for all studied factors, indicating the effectiveness of our method in ranking functional *k*-mers for ChIP-seq data.

ChIP-chip on CTCF. We also asked if POSMO could process ChIP-chip data, which has less resolution than that of ChIP-seq data (2). For this purpose, we obtained the 13 720 peaks of CTCF binding determined using ChIP-chip (33). As shown in Table 1 and Supplementary Table S1, the top five 8-mers found by POSMO are all related to the CTCF binding motif. This result suggested that our method is applicable to ChIP-chip data.

ChIP-seq data from D. melanogaster. We next asked if our method could process ChIP-seq data from species other than human. To address this question, we obtained ChIP peaks for transcription factors BCD, HB1, HB2, CAD, KNI, GT, KR1 and KR2 of *D. melanogaster* (34). Table 1 and Supplementary S1 clearly show that POSMO is able to rank the target *k*-mers on top for GT, KR1, KR2, BCD, HB1 and HB2. POSMO failed to rank the known motif of CAD and KNI on top. However, we found that both MEME and DME also failed to discover the correct DNA motifs for CAD and KNI (Table 5 and Supplementary Table S2), indicating that the underlying nature of this data set may not permit us to find the desired motifs.

ChIP-seq data from core transcriptional networks in mouse ES cells

We also investigated the performance of our algorithm on the 13 sequence-specific transcription factors involved in the core transcriptional networks in mouse ES cells (35). As shown in Supplementary Table S1, POSMO successfully ranked the desired *k*-mers on top for 9 of the 13 factors (c-MYC, n-MYC, CTCF, ESRRB, KLF4, OCT4, SOX2, STAT3, ZFX and E2F1). Although the *k*-mer with highest PWM score is ranked 34th for TCFCP2L1, most of the top-ranked *k*-mers are just shift of the known motif (light gray cells in Supplementary

Table S1). For E2F1, no motifs were found in the original work by Chen *et al.* (35). In Bailey (45), the top two motifs found for E2F1 are in the form of GGAA and ATGGCG. On the other hand, the top *k*-mers found by POSMO contain the sequence TTCCGG, which is partly similar to the *in vitro* E2F1 motif documented in JASPAR. As a confirmation, we found that our top *k*-mers, especially motif TTCCGG (Supplementary Table S1), also appeared in independent E2F1 ChIP-seq data (46). Interestingly, in this independent E2F1 ChIP-seq data (46), a motif TTGGCGC with rank 14 is partly similar to the E2F1 motif documented in JASPAR. For SMAD1, our algorithm did not find any significant motifs. However, POSMO identified a motif (Supplementary Table S2) weakly similar to the known SMAD1 motif when using ±200 bp flanking region of the peak summits, indicating that the length of flanking regions may be further optimized for POSMO.

In summary, the above results clearly indicated that POSMO is highly effective for both ChIP-chip and ChIP-seq data from human, mouse and other species, such as fly. In the next section, we will try to group these significant *k*-mers into PWM representation.

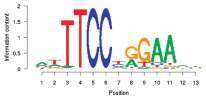
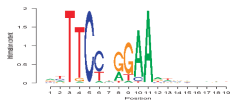
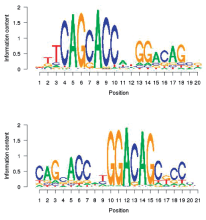
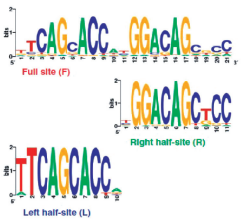
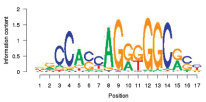
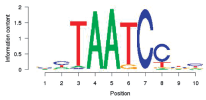
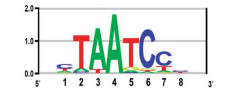
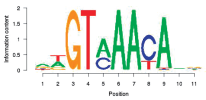
Word clustering on real data

Even though our POSMO software is generally able to rank the desired DNA *k*-mers on top, it is difficult to manually inspect them. Consequently, we next applied our novel word clustering algorithm on the sequence contexts of significant *k*-mers in the genome to obtain PWM representations (see 'Materials and Methods' section). As can be seen in Table 2, motifs reported for CTCF, STAT1, NRSF, CRX and FOXA2 by our word clustering method are highly similar to those reported in the literature.

In the case of *D. melanogaster*, the motifs found by our method for BCD, HB1, HB2, KR1, KR2 and GT (Supplementary Table S2) are highly similar to the motifs cataloged in JASPAR. For transcription factor CAD and KNI, the known motif was not found by either our algorithm or other algorithms (Supplementary Table S2). Interestingly, for CAD, our algorithm reported a motif with consensus CAGGTA, which is implicated in the regulation of early transcribed genes during *Drosophila* development (48).

For the core transcription factors involved in mouse ES cells, motifs reported by POSMO are also highly similar to the known motifs for CTCF, n-MYC, c-MYC, STAT3, KLF4, SOX2, OCT4, ZFX, TCFCP2L1 and ESRRB (Supplementary Table S2). Similar to a recent study using an improved version of MEME by Bailey (45), our algorithm did not report any motif for SMAD1. However, POSMO identified a motif (Supplementary Table S2) weakly similar to the known SMAD1 motif when using ±1000 bp flanking region of the peak summits, indicating that the length of flanking regions may be further optimized for POSMO. The motif for E2F1 was not found by either Chen *et al.* (35) or other methods including POSMO. However, our algorithm found a motif with the pattern CCGGAAG (reverse complement is CTTCCGG; see Supplementary Table S5), which is partly similar to the known motif TTT[G/C][G/C]CGC documented in JASPAR. Interestingly, we note that CC GGAAG is highly similar to the binding motifs of ETS transcription factors (49), which might indicate the interaction between E2F1 and ETS transcription factors. Notably, this motif is also found in an independent E2F1 ChIP-seq data (46) (data now shown).

Table 2. Sequence motifs discovered by POSMO

Transcription factor	Result by POSMO	Literature
STAT1 (29)		 (24)
NRSF (31)		 (7)
CTCF (30)		 (7)
CRX (28)		 (28)
FOXA2 (32)		 (47)

The DNA motifs after word clustering are listed. As a comparison, the DNA motifs from the literature are also listed. For NRSF, two motifs are reported by POSMO, which correspond to the left and right half-sites reported by Hu *et al.* (24).

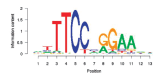
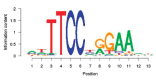
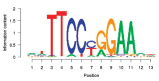
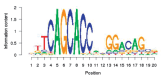
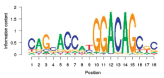
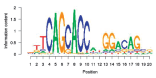
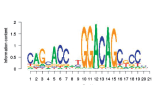
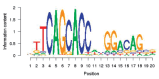
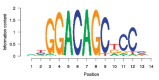
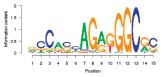
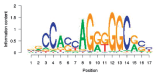
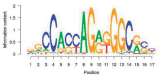
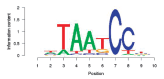
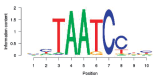
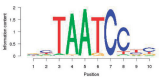
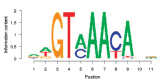
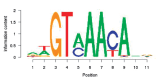
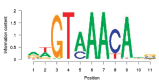
POSMO is robust to input parameters

One general concern for motif finders using the *k*-mer enumeration method is the determination of *k*. We thus asked how *k* affects the performance of POSMO. For this purpose, we ran our POSMO algorithm for 7-, 8- and 9-mer on STAT1, NRSF, CTCF, CRX and FoxA2 data. As can be seen from Table 3, different *k* values do not greatly affect the obtained DNA motifs. Apparently, the closer the specified word pattern is to the truth, the better the results will be. Since it is difficult to know this parameter *a priori*, it may be helpful to try several parameters in real applications. However our results on long motifs of CTCF and NRSF suggest that in general we

do not need very large k to discover long motifs. This is due to our heuristic word clustering method by considering context, as described in 'Materials and Methods' section.

We also asked how input parameters t_D and t_Z (see 'Materials and Methods' section) of POSMO could affect the performance of POSMO. We tried a series of parameters on t_D and t_Z and found that our method is not sensitive to them (Supplementary Table S3). Thus, this fact renders our method easy to use in practice.

Table 3. Input pattern lengths of 7, 8 and 9 are compared (POSMO is robust to input parameter k)

Transcription factor	7-mer	8-mer	9-mer
STAT1 (29)			
NRSF (31)	 	 	 
CTCF (30)			
CRX (28)			
FOXA2 (32)			

POSMO performs well for large sample sizes

As discussed in Schmid and Bucher (26), the percentage of top peak sequences containing the known motif generally decreases as the threshold relaxes. Therefore, we asked how the total number of peak sequences would affect our motif finding results. We ranked the identified peaks for STAT1, CTCF, NRSF and FOXA2 according to the peak height, and we used the top 1000, 2000, 5000 and 10 000 peaks as input for POSMO. As can be seen in Table 4, POSMO is effective for all tested parameters. In particular, POSMO was found to be effective for input data sets containing 10 000 peaks for all studied transcription factors, suggesting the superior performance of POSMO for large sample sizes. Since ChIP-seq experiments typically produce thousands of peak sequences, we conclude that POSMO has broad applicability for ChIP experiments.

POSMO is more effective than available methods for ChIP data

We next compared the overall performance of POSMO with that of DME (41), MEME (8), ChIPMunk (25), HMS (24) and DREME (45) on motif discovery. For this purpose, we checked the rank of the known motifs among all discoveries in DME (5 motifs to be reported), MEME (only top 500 peaks are used as a result of running speed constraint and 5 motifs to be reported), ChIPMunk (1 motif to be reported), HMS (ChIP-seq intensity profile under the peaks were also compiled for applicable

Table 4. POSMO is robust to large sample sizes

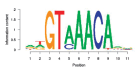
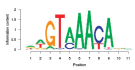
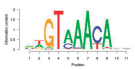
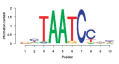
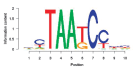
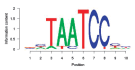
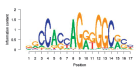
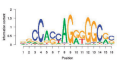
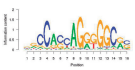
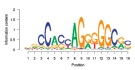
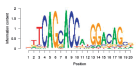
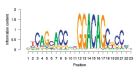
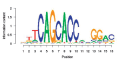
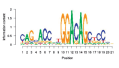
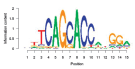
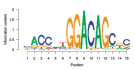
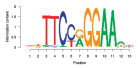
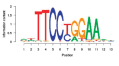
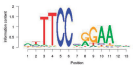
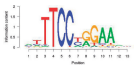
	1000 peaks	2000 peaks	5000 peaks	10 000 peaks
FOXA2 (32)				
CRX (28)				
CTCF (30)				
NRSF (31)	 	 	 	Only 5752 peaks
STAT1 (29)				

Table 5. Performance comparison of POSMO, MEME, DME, ChIPMunk, HMS and DREME on ChIP data

Transcription factor	Rank by POSMO	Rank by MEME	Rank by DME	Rank by ChIPMunk	Rank by HMS	Rank by DREME
STAT1 (29)	1	1	1	1[a]	1	1
NRSF (31)	1, 2	1, 2	1, 2	1	1	1
CTCF (30)	1	1	1	1	1	1
CTCF (33)	1	3	1	1[b]	NA	1
FOXA2 (32)	1	1	1	1	1	1
CRX (28)	1	1	1	1	No match	1
BCD (34)	1	4	4	1	No match	2
CAD (34)	No match	No match	No match	No match	No match	No match
HB1 (34)	1	1	1	1	No match	2
HB2 (34)	1	1	1	1	No match	3
KR1 (34)	1	1	1	1	No match	3
KR2 (34)	1	1	1	1	No match	3
KNI (34)	No match	No match	No match	No match	No match	No match
GT (34)	1	3	7	1	No match	No match
c-MYC (35)	1	1	1	1	1	1
n-MYC (35)	1	1	1	1	1	1
CTCF (35)	1	1	1	1	No match	1
ESRRB (35)	1	1	1	1	1	1
STAT3 (35)	1	1	2	1	No match	1
OCT4 (35)	1	1	1	1	1	1
SOX2 (35)	1	1	1	1	1	3
KLF4 (35)	1	1	2	1	1	1
E2F1 (35)	No match	No match	No match	No match	No match	No match
TCFCP2L1 (35)	1	1	1	1	1	1
ZFX (35)	1	1	1	1	No match	1
NANOG (35)	1	1	1	1[c]	No match	1
SMAD1 (35)	1[d]	No match	No match	1	1	1
Total successes	24/27	23/27	23/27	24/27	12/26	23/27
Average rank	1	1.30	1.47	1	1	1.43

Among the Top 5 motifs found by MEME, DREME and DME, the rank (per *P*-values) of the known binding motif is listed. For NRSF, there are two known motifs and their ranks are counted separately. No match: the software did not report any motif similar to the known motif.
[a]Top 3 peaks removed to get the correct motif.
[b]Triangle intensity profile used.
[c]Motif length of 20 used.
[d]±200 bases flanking peak summit.

transcription factors), DREME and POSMO (details on the motifs found by each method can be found in Supplementary Table S2). For POSMO, our algorithm reported the total number of occurrences of each motif in the ChIP-seq/ChIP-chip data, which could be used to rank all the discovered motifs. As can be seen in Table 5, POSMO performs in a manner similar to DME, MEME and DREME, though with better average rank of the discovered motifs. Motifs discovered by POSMO are similar to that discovered by ChIPMunk. However, the top 3 extremely high scoring peaks must be removed for ChIPMunk to discover the correct motif for STAT1 (Table 5 and Supplementary Table S2). In addition, POSMO is better than HMS for fly transcription factors and several mammal transcription factors including CRX, STAT3 and ZFX, suggesting the high effectiveness of our method. POSMO did not find motif for SMAD1 using our default settings; however, we noted that POSMO correctly identified the motif for SMAD1 when shorter flanking regions (±200 bp) are used (Table 5 and Supplementary Table S2). This result suggests that flanking length may be further optimized for motif finding. Interestingly, the true motif found by POSMO is always ranked in first place, again indicating that PSOMO is more effective than other tools. This property is particularly useful to assign DNA motifs to a newly investigated transcription factor for which no prior motif information is available.

POSMO is more efficient than available methods

We also compared the running time of POSMO with other established methods. As established in Keilwagen *et al.* (23), DME by Smith *et al.* (41) is one of the fastest algorithms for large sample sizes. We therefore compared the running time of our method with that of DME using various numbers of peaks ranging from 500 to 10 000. As can be seen in Figure 1, our method is significantly faster than DME for large sample sizes, where the running time of our method scales linearly with the number of peak sequences, with a typical running time of only a few minutes. In addition, a comparison on the real ChIP data sets revealed that POSMO is significantly more efficient than ChIPMunk, HMS and DREME (Supplementary Table S4). Thus, we conclude that our method is highly efficient for motif discovery from large sample sizes of ChIP experiments. Clearly, the efficiency of our method will quickly decrease with large k. However, as was demonstrated in (50), 77% of the transcription factor binding motifs have <11 informative positions. Most importantly, as can be seen in Table 3, our method is robust to different ks for tested transcription factors. In particular, our method works well

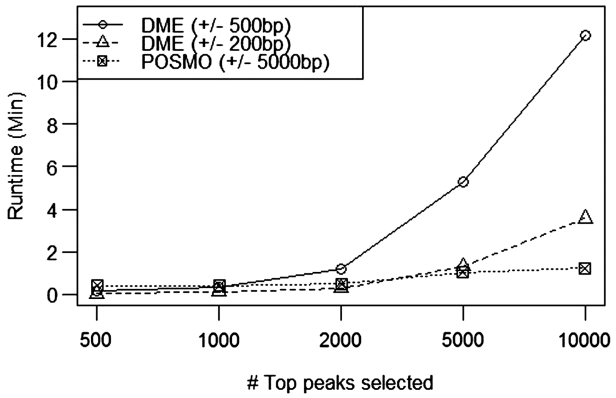


Figure 1. POSMO is more efficient than DME for large sample sizes. Shown in the *y*-axis is the time spent for a given number of top peaks shown in the *x*-axis. Results for POSMO (dashed line with boxes) and DME (dashed line with triangles for a smaller peak window and solid line with circles for a larger peak window) are shown. Here $k = 8$ for both POSMO and MEME.

for CTCF and NRSF motif, which have a long motif >13 bp. This result indicates that we do not need a very large k to find a long motif, partly due to our heuristic word clustering method. Thus, the efficiency of our algorithm is generally guaranteed.

Co-motif finding

In principle, different DNA motifs may co-localize to perform regulatory functions by forming protein complexes. Therefore, it will be very interesting to see if we can discover co-motifs from such high-throughput ChIP-seq data. In fact, sophisticated method targeting this question is already proposed in SpaMo by Bailey and colleagues (51). Though our POSMO is not specifically designed for the purpose of finding co-motifs, we still asked if it can find some of the known co-motifs. For a few transcription factors such as STAT1, CRX, E2F1 and n-Myc (see Supplementary Table S5 for details), POSMO reported some of the known co-motifs as identified by Bailey and colleagues (51). Interestingly, POSMO found co-motif CAGGTA for many fly transcription factors. However, we note that our POSMO is not purposely designed to find co-motifs; therefore, POSMO reported much less co-motifs than SpaMo did. An extension of POSMO specifically designed to detect co-motifs is under development.

DISCUSSION

ChIP-seq/ChIP-chip is a popular experimental method to map *in vivo* binding sites of transcription factors. DNA motif discovery from such data is a necessary step toward understanding gene regulation. However, available motif finding tools are mostly designed to find DNA motifs in sequence segments by optimizing alignments, which renders the optimization process inefficient for large sample sizes. Recently, a few methods have been developed to utilize signal intensity to accelerate the discovery process. In this work, we have introduced a new k-mer enumeration method, POSMO, to predict transcription factor binding motifs. Using simulation, we found that our method is more robust against the information spread and systematic errors in peak locations than available methods in terms of ranking the target k-mer. The high prediction accuracy is further confirmed using a diverse set of real ChIP-seq/ChIP-chip data sets on human, mouse and fly. We also developed a novel word clustering algorithm by checking the sequence context of each significant k-mer. We found that our word clustering method can generate motif representation consistent with reports found in the literature. We found that motifs discovered by POSMO is consistent with that discovered by DME and MEME, though our method always gives the true motif highest rank in all tested data sets. This property could be very important when there is no prior knowledge on the binding motifs of a newly investigated transcription factor. Thus, our method is more effective for motif discovery.

On the other hand, since estimation of peak summits is more accurate than estimation of the exact 'peak regions' from ChIP-chip/ChIP-seq data, our method provides better usability. In addition, since POSMO essentially contrasts far flanking sequences with sequences under the peak summit, our method does not require explicit 'background' to normalize the k-mer appearance frequency, which is generally recommended for many motif discovery methods (i.e. to also construct background data set). This property also better mimics the biology of transcription factor-DNA interactions: instead of optimizing the binding affinity between the target DNA motif and many other genome-wide 'background' sequences, a transcription factor is actually searching the target DNA motifs from the pool of surrounding local DNA sequences. Our results suggested that these local sequences can be better approximated by flanking sequences of ChIP-peak regions than by other 'control' sequences.

Most importantly, since our method is essentially a k-mer enumeration method where hypothesis testing procedures are extensively used, it is very efficient, with a typical running time of only a few minutes for thousands, or even more, ChIP-seq peaks for word length <10. This is in clear contrast to most established methods, such as MEME, which utilize extensive optimization techniques that can take up to hours for a few hundred ChIP-seq peaks (7). Thus, we believe our method will be a useful alternative to quickly study the binding sites of transcription factors.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online: Supplementary Text 1, Supplementary Figures 1 and 2, Supplementary Tables 1–5.

ACKNOWLEDGEMENTS

We thank Joe Corbo for providing PWM of CRX. We also thank Hongyu Zhao for valuable suggestions. The authors are grateful to the anonymous reviewers for their excellent suggestions.

FUNDING

National Institute of Health (HG001696 to M.Q.Z.); National Basic Research Program of China (2012CB316503 to M.Q.Z.); National Natural Science Foundation of China (91019016, 31061160497 to M.Q.Z.); National Science Foundation (DMS-1106091 to R.S.) and UTD Startup Fund (to Z.X.). Funding for open access charge: NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
2. Stormo,G.D. and Zhao,Y. (2010) Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.*, **11**, 751–760.
3. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
4. Portales-Casamar,E., Thongjuea,S., Kwon,A.T., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W.W. and Sandelin,A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.
5. Gerstein,M.B., Lu,Z.J., Van Nostrand,E.L., Cheng,C., Arshinoff,B.I., Liu,T., Yip,K.Y., Robilotto,R., Rechtsteiner,A., Ikegami,K. *et al.* (2010) Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science*, **330**, 1775–1787.
6. Roy,S., Ernst,J., Kharchenko,P.V., Kheradpour,P., Negre,N., Eaton,M.L., Landolin,J.M., Bristow,C.A., Ma,L., Lin,M.F. *et al.* (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, **330**, 1787–1797.
7. Jothi,R., Cuddapah,S., Barski,A., Cui,K. and Zhao,K. (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, **36**, 5221–5231.
8. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
9. Zhang,M.Q. (2007) Inferring Gene Regulatory Networks. In: Lengauer,T. (ed.), *Bioinformatics - From Genomes to Therapies*. Wiley-VCH GmbH, Weinheim, Germany, pp. 807–828.
10. Buhler,J. and Tompa,M. (2002) Finding motifs using random projections. *J. Comput. Biol.*, **9**, 225–242.
11. Eskin,E. and Pevzner,P.A. (2002) Finding composite regulatory patterns in DNA sequences. *Bioinformatics*, **18(Suppl. 1)**, S354–S363.
12. Ettwiller,L., Paten,B., Ramialison,M., Birney,E. and Wittbrodt,J. (2007) Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nat. Methods*, **4**, 563–565.
13. Fratkin,E., Naughton,B.T., Brutlag,D.L. and Batzoglou,S. (2006) MotifCut: regulatory motifs finding with maximum density subgraphs. *Bioinformatics*, **22**, e150–e157.
14. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
15. Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.
16. Marsan,L. and Sagot,M.F. (2000) Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J. Comput. Biol.*, **7**, 345–362.
17. Pavesi,G., Mereghetti,P., Mauri,G. and Pesole,G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.*, **32**, W199–W203.
18. Roth,F.P., Hughes,J.D., Estep,P.W. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.
19. Vardhanabhuti,S., Wang,J. and Hannenhalli,S. (2007) Position and distance specificity are important determinants of *cis*-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Res.*, **35**, 3203–3213.
20. Linhart,C., Halperin,Y. and Shamir,R. (2008) Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome Res.*, **18**, 1180–1189.
21. Kim,N.K., Tharakaraman,K., Marino-Ramirez,L. and Spouge,J.L. (2008) Finding sequence motifs with Bayesian models incorporating positional information: an application to transcription factor binding sites. *BMC Bioinformatics*, **9**, 262.
22. Narang,V., Mittal,A. and Sung,W.K. (2010) Localized motif discovery in gene regulatory sequences. *Bioinformatics*, **26**, 1152–1159.
23. Keilwagen,J., Grau,J., Paponov,I.A., Posch,S., Strickert,M. and Grosse,I. (2011) De-novo discovery of differentially abundant transcription factor binding sites including their positional preference. *PLoS Comput. Biol.*, **7**, e1001070.
24. Hu,M., Yu,J., Taylor,J.M., Chinnaiyan,A.M. and Qin,Z.S. (2010) On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res.*, **38**, 2154–2167.
25. Kulakovskiy,I.V., Boeva,V.A., Favorov,A.V. and Makeev,V.J. (2010) Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, **26**, 2622–2623.
26. Schmid,C.D. and Bucher,P. (2010) MER41 repeat sequences contain inducible STAT1 binding sites. *PLoS One*, **5**, e11425.
27. Ji,H., Jiang,H., Ma,W., Johnson,D.S., Myers,R.M. and Wong,W.H. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, **26**, 1293–1300.
28. Corbo,J.C., Lawrence,K.A., Karlstetter,M., Myers,C.A., Abdelaziz,M., Dirkes,W., Weigelt,K., Seifert,M., Benes,V., Fritsche,L.G. *et al.* (2010) CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. *Genome Res.*, **20**, 1512–1525.
29. Robertson,G., Hirst,M., Bainbridge,M., Bilenky,M., Zhao,Y., Zeng,T., Euskirchen,G., Bernier,B., Varhol,R., Delaney,A. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.
30. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
31. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
32. Wederell,E.D., Bilenky,M., Cullum,R., Thiessen,N., Dagpinar,M., Delaney,A., Varhol,R., Zhao,Y., Zeng,T., Bernier,B. *et al.* (2008) Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res.*, **36**, 4549–4564.
33. Kim,T.H., Abdullaev,Z.K., Smith,A.D., Ching,K.A., Loukinov,D.I., Green,R.D., Zhang,M.Q., Lobanenkov,V.V. and Ren,B. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, **128**, 1231–1245.
34. Bradley,R.K., Li,X.Y., Trapnell,C., Davidson,S., Pachter,L., Chu,H.C., Tonkin,L.A., Biggin,M.D. and Eisen,M.B. (2010) Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. *PLoS Biol.*, **8**, e1000343.
35. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.
36. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.*

(2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
37. Wilbanks,E.G. and Facciotti,M.T. (2010) Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One*, **5**, e11471.
38. Dean,N. and Raftery,A.E. (2005) Normal uniform mixture differential gene expression detection for cDNA microarrays. *BMC Bioinformatics*, **6**, 173.
39. Schones,D.E., Sumazin,P. and Zhang,M.Q. (2005) Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, **21**, 307–313.
40. Mahony,S., Auron,P.E. and Benos,P.V. (2007) DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput. Biol.*, **3**, e61.
41. Smith,A.D., Sumazin,P. and Zhang,M.Q. (2005) Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc. Natl Acad. Sci. USA*, **102**, 1560–1565.
42. Sinha,S. and Tompa,M. (2003) YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, **31**, 3586–3588.
43. Sumazin,P., Chen,G., Hata,N., Smith,A.D., Zhang,T. and Zhang,M.Q. (2005) DWE: discriminating word enumerator. *Bioinformatics*, **21**, 31–38.
44. 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.
45. Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.
46. Cao,A.R., Rabinovich,R., Xu,M., Xu,X., Jin,V.X. and Farnham,P.J. (2011) Genome-wide analysis of transcription factor E2F1 mutant proteins reveals that N- and C-terminal protein interaction domains do not participate in targeting E2F1 to the human genome. *J. Biol. Chem.*, **286**, 11985–11996.
47. Tuteja,G., White,P., Schug,J. and Kaestner,K.H. (2009) Extracting transcription factor targets from ChIP-Seq data. *Nucleic Acids Res.*, **37**, e113.
48. Liang,H.L., Nien,C.Y., Liu,H.Y., Metzstein,M.M., Kirov,N. and Rushlow,C. (2008) The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. *Nature*, **456**, 400–403.
49. Wei,G.H., Badis,G., Berger,M.F., Kivioja,T., Palin,K., Enge,M., Bonke,M., Jolma,A., Varjosalo,M., Gehrke,A.R. *et al.* (2010) Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J.*, **29**, 2147–2160.
50. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. III and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
51. Whitington,T., Frith,M.C., Johnson,J. and Bailey,T.L. (2011) Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res.*, **39**, e98.