MOCCS: clarifying DNA-binding motif ambiguity using ChIP-Seq data

Haruka Ozaki[a,1,*], Wataru Iwasaki[a,b,c,*]

[a] *Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwanoha 5-1-5, Kashiwa, 277-8568, Chiba, Japan*

[b] *Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, 113-0032, Tokyo, Japan*

[c] *Atmosphere and Ocean Research Institute, The University of Tokyo, Kashiwanoha 5-1-5, Kashiwa, 277-8564, Chiba, Japan*

Abstract

Background: As a key mechanism of gene regulation, transcription factors (TFs) bind to DNA by recognizing specific short sequence patterns that are called DNA-binding motifs. A single TF can accept ambiguity within its DNA-binding motifs, which comprise both canonical (typical) and non-canonical motifs. Clarification of such DNA-binding motif ambiguity is crucial for revealing gene regulatory networks and evaluating mutations in *cis*-regulatory elements. Although chromatin immunoprecipitation sequencing (ChIP-seq) now provides abundant data on the genomic sequences to which a given TF binds, existing motif discovery methods are unable to directly answer whether a given TF can bind to a specific DNA-binding motif.

Results: Here, we report a method for clarifying the DNA-binding motif ambiguity, MOCCS. Given ChIP-Seq data of any TF, MOCCS comprehensively analyzes and describes every *k*-mer to which that TF binds. Analysis of simulated datasets revealed that MOCCS is applicable to various ChIP-Seq datasets, requiring only a few minutes per dataset. Application to the ENCODE ChIP-Seq datasets proved that MOCCS directly evaluates whether a given TF binds to each DNA-binding motif, even if known position weight matrix models do not provide sufficient information on DNA-binding motif ambiguity. Furthermore, users are not required to provide numerous parameters or background genomic sequence models that are typically unavailable. MOCCS is implemented in Perl and R and is freely available via https://github.com/yuifu/moccs.

Conclusions: By complementing existing motif-discovery software, MOCCS will contribute to the basic understanding of how the genome controls diverse cellular processes via DNA-protein interactions.

Keywords: DNA binding motifs, ChIP-Seq, Transcription factors

Background

The binding of proteins to DNA is key to the control of almost all cellular processes. Most importantly, transcription factors (TFs) bind to *cis*-regulatory DNA regions that contain characteristic short sequence patterns (DNA-binding motifs) and control coordinated expression of the targeted genes. Thanks to recently invented high-throughput sequencing technologies, especially chromatin immunoprecipitation sequencing (ChIP-Seq), greater than thousands of genomic positions bound by a given TF can now be identified by a single experiment. To date, numerous ChIP-Seq experiments have been performed on various TFs under various conditions, as exemplified by the ENCODE [1] and the Roadmap Epigenomics projects [2]. A precise understanding of DNA-binding motifs based on ChIP-Seq data is indispensable not only for revealing gene regulatory networks behind diverse cellular functions but also for evaluating mutations within *cis*-regulatory regions.—Recent personal genomics studies are revealing many variations within non-coding regions of the human genome [3]. Furthermore, such knowledge would form a basis for analyzing cellular differentiation patterns in regenerative medicine and for enabling targeted manipulation of DNA-binding proteins in synthetic biology.

A single TF may bind to multiple motifs, includ-

*Correspondence: harukao.cb@gmail.com (H.O.), iwasaki@bs.s.u-tokyo.ac.jp (W.I.)

Email addresses: harukao.cb@gmail.com (Haruka Ozaki), iwasaki@bs.s.u-tokyo.ac.jp (Wataru Iwasaki)

[1] Present address: Bioinformatics Research Unit, Advanced Center for Computing and Communication, RIKEN, 2-1 Hirosawa, Wako, 351-0198, Saitama, Japan.

ing both canonical (typical) and non-canonical DNA-binding motifs [4][5][6]. To precisely understand the binding specificity of a given TF, such ambiguity regarding its DNA-binding motifs must be comprehensively described. In bioinformatics, the *motif discovery problem* constitutes one of the most studied branches, for which many software programs exist [7][8]. Given a set of DNA sequences, these programs search for characteristic motifs using two major approaches: profile-based and consensus-based methods. In particular, the recent increase in data size due to the ChIP-Seq technique led to the development of methods that can accept greater than thousands of DNA sequences [9][10][11][12][13][14][15][16][17]. Most of these software programs focused on reductions in computational time, for example, by subsampling input data (e.g., MEME-ChIP [14]), accelerating expectation-maximization steps in profile optimization (e.g., ChIPMunk [12] and STEME [15], which use a greedy approach and suffix array, respectively), or using enriched sequences as starting points for the motif search (e.g., DREME [13], cERMIT [18], and HOMER [11]).

Despite the successful development of these motif-discovery methods, the comprehensive description of DNA-binding motifs based on ChIP-Seq data still poses a difficult problem. Whereas existing methods can partly represent motif ambiguity, these methods are unable to directly answer whether a given TF binds to a specific sequence pattern. For example, the most popular software program MEME [19] and other recently developed software programs [15][20], which adopt the expectation-maximization algorithm, iteratively enrich DNA sequences that contain possible DNA-binding motifs and often converges to a local optimum. Given the nature of this algorithm, discovered DNA-binding motifs can miss non-canonical but significant motifs that were removed from the enriched dataset during the computation. Another problem is that neither the degenerate consensus sequence nor the position weight matrix (PWM) model can consider dependencies between different nucleotide positions within DNA-binding motifs [7].

Here, we describe motif centrality analysis of ChIP-Seq (MOCCS), a method that clarifies the DNA-binding motif ambiguity using ChIP-Seq data. Given the protocol of ChIP experiments and data size of high-throughput sequencing, the frequency distribution of any sequence pattern that is bound by a given TF exhibits a unimodal shape around TF-binding sites (TFBSs) that can be massively identified by ChIP-Seq (Figure 1A). MOCCS evaluates whether the frequency distribution of each sequence pattern shows a unimodal shape. This approach is rather simple, yet effective in discriminating sequence patterns bound by the TF from those not. Indeed, in our previous study with experimental biologists [21], a pilot version of the MOCCS software was applied to ChIP-Seq data of the CLOCK protein, and we successfully identified and experimentally confirmed three novel non-canonical DNA-binding motifs, although CLOCK had already been extensively studied as a key TF in mammalian circadian clockworks. In this study, we implemented MOCCS as a freely available software program, and evaluated it using simulated and real datasets.

Results

Methods overview

MOCCS enumerates all k-mer sequences that are bound by a given TF. Any k-mer bound by the TF frequently appears around the TFBSs (identified by ChIP-Seq), and its frequency distribution exhibits a unimodal shape (Figure 1A). To take advantage of this fact, MOCCS examines the shape of every k-mer's frequency distribution within a w-bp range around all TFBSs (For non-palindromic k-mers, frequency distributions of reverse complementary k-mers are combined). To quantify the sharpness of each unimodal distribution, a cumulative relative frequency curve against distance from TFBSs is generated, and the area under the curve (AUC) is calculated (Figure 1B; note that the AUC becomes larger if the shape becomes sharper). Then, MOCCS identifies every k-mer whose AUC is above a threshold value as a DNA-binding motif (Figure 1C). MOCCS ignores k-mers whose total frequencies within the w-bp range are smaller than the expected frequency in random sequences because such rare k-mers are prone to noise and can exhibit very large AUCs by chance.

Notably, the time complexity of MOCCS is linear in the number of TFBSs and in w. More importantly, MOCCS does not require users to specify background sequence characteristics, such as GC% or short tandem repeats, requirements for which often circumvents successful motif discovery using existing methods. Method details are described in the Materials and Methods section.

Simulation analysis on user-provided parameters

Although the number of user-provided parameters is smaller than most existing methods, MOCCS requires the parameters k (length of DNA-binding motif) and w (range of frequency distribution analysis). Thus, we

evaluated the influence of these parameter settings on the performance using simulated data.

Simulated data were prepared by embedding a "true" k-mer within α% of a randomly generated sample of w-bp DNA sequences, where $\alpha = 1, 3, 5, 10, 20, 40$. The k-mer sequence represents a specific short sequence that is bound by a given TF and was randomly determined in each simulation. When the k-mer is a canonical (or mainly targeted) DNA-binding motif, α is expected to be large. On the other hand, α is reduced for a non-canonical motif. The distribution of the k-mer-embedding positions followed a Gaussian distribution whose mean was the center of the DNA sequence (i.e., the position of TFBS) and whose standard deviation was set to $\sigma = 80$, which is a typical value estimated from public data (discussed later). The number of input DNA sequences, which corresponds to the number of TFBSs, was set to $N = 6,000$. It should be noted that a degenerate motif is regarded as a union of multiple k-mers with different α and σ and can be analyzed by MOCCS in the same way, although only a single k-mer was considered to be true in this simulation setting.

In the first analysis, we set $k = 8$ because this is an upper limit of typical DNA-binding motif lengths of monomeric eukaryotic TFs [13], and we tested $w = 101, 201, 301, 401, 501, 601, 701, 801, 1001, 2001$. We calculated the rank of AUC of the true k-mer, which is expected to reach 1 if MOCCS successfully works. For each parameter set, simulation was repeated 100 times, and the ranks were averaged. Figure 2A depicts how the rank of the true k-mer was affected by w. The rank increased (i.e., better performance) with increasing w. Even if $\alpha = 3$%, the rank was 1.02 when $w = 701$. Thus, we set $w = 701$ bp as the default value, which was adopted in all of the following analyses.

In the second analysis, we tested $k = 5, 6, 7, 8$ to evaluate the influence of k on the performance, and the other conditions were the same as those used in the first analysis. Figure 2B depicts how the rank of the true k-mer was affected by k. The rank increased with increasing k. Even if $\alpha = 3$%, the rank was 1.02 when $k = 8$. On the other hand, if $\alpha \geq 10$%, the rank was 1.00 for $k = 5, 6, 7, 8$. These results indicate that MOCCS can identify even 5-mer DNA-binding motifs if they appear at approximately greater than one-tenth of TFBSs, and 8-mer motifs are successfully identified when they appear around only 3% of TFBSs.

Simulation analysis on TF- and experiment-specific variables

In contrast to the user-provided parameters k and w, the parameters N (the number of input DNA sequences or identified TFBSs), α (the fraction of input sequences that contain TF-bound k-mers), and σ (the sharpness of the frequency distribution of the TF-bound k-mer around TFBSs) depend on the TFs and the ChIP-Seq experimental conditions (e.g., protocols, antibody quality, and existence of cofactors). Note that, within a single ChIP-Seq experiment, non-canonical DNA-binding motifs are expected to have smaller α and larger σ compared with canonical motifs.

Figure 3A presents simulation results similar to those of Figure 2, with the exception of $N = 2000$, 4000, 6000, 8000, 10000, $\alpha = 1, 3, 5, 10, 20, 40$ and $k = 5, 6, 7, 8$. In all cases, the rank increased with increasing N, as expected. Even if $\alpha = 3$%, the rank was 1.02 when $N = 4000$ and $k = 8$. On the other hand, if $\alpha \geq 10$%, the rank was 1.00 when $N \geq 4000$ for all k. These results suggest that if a few thousand TFBSs are provided, MOCCS can identify non-canonical TF-binding 8-mers that are contained in a very limited subset of ChIP-ed sequences as well as 5-mers that are present in greater than one-tenth of those sequences.

In addition to N and α, σ also affects the performance of the method (larger σ leads to worse performance). To obtain a biologically meaningful range of σ, we analyzed ChIP-Seq data from the ENCODE project [1]. We estimated TFBSs using *peak regions* provided in the datasets, calculated frequency distributions of DNA-binding motifs using known PWMs, fit Gaussian distributions, and obtained their standard deviations (Supplementary Table 1). Most data exhibited standard deviations (σ) within a range of 40 to 120. From this observation, we set $\sigma = 40, 60, 80, 100, 120$ and ran the simulation with $\alpha = 1, 3, 5, 10, 20, 40$, $N = 6000$, and $k = 8$ (Figure 3B). Even if $\sigma = 120$ and $\alpha \geq 3$%, the rank was 1.20, suggesting that MOCCS is applicable to cases that frequency distributions are not very sharp.

Collectively, these simulation results demonstrate that MOCCS can be applied to a wide range of TFs and ChIP-Seq data.

Application to real ChIP-Seq data

By setting $w = 701$ and $k = 8$ based on the simulation analysis, we applied MOCCS to several real ChIP-Seq datasets from the ENCODE project [1]. To avoid redundancy, when any two identified k-mers overlap each other for $k - 1$ bp, only the k-mer with the greater AUC is presented.

USF1

USF1 is a TF that regulates genes involved in lipid and glucose metabolism [22] and is known to bind

to a canonical E-box motif CACGTG (palindromic) [23][24] and a non-canonical motif CATGTG (or CACATG in reverse complement) [25].

Figure 4A presents the results of MOCCS on USF1 ChIP-Seq data from K562 cells. As expected, an 8-mer with the greatest AUC was GTCACGTG (or CACGTGAC), which contains the canonical motif. The second highest 8-mer was GTCACATG (or CATGTGAC), which contains the reported non-canonical motif. The fact that both 8-mers sharply and clearly concentrated around the USF1-binding sites (Figure 4B) confirmed that MOCCS successfully identified these known USF1-binding motifs. Furthermore, our analysis proved that USF1 also binds to other non-canonical motifs, GTCAGGTG (or CACCTGAC) and TCAGCTGA (palindromic), which were the third and fourth highest 8-mers, respectively. Although the AUCs and total frequencies of these two 8-mers were reduced compared with the first and second 8-mers, their frequency distributions clearly demonstrated that these 8-mers are genuine USF1-binding motifs (Figure 4B). A subsequent literature survey identified cases where CAGGTG (or CACCTG) and CAGCTG (palindromic) were bound by USF1 at *Gata3* locus [23] and bound by a USF1/USF2 complex [26], respectively. These results indicate that MOCCS identifies various non-canonical motifs. It should be noted that the reported PWMs of the USF1 DNA-binding motif (Figures 4C and 4D, retrieved from the JASPAR database [27] and factorbook [28], respectively) do not clearly indicate whether USF1 binds to these non-canonical motifs, whereas MOCCS does.

GABP

The GA-binding protein (GABP) is a TF known to bind to a GGA(A/T) motif [29][30]. We applied MOCCS to GABP ChIP-Seq data from K562 cells, and the 8-mer with the highest AUC was ACTTCCGG (or CCGGAAGT), which contains GGAA (Figure 5A). To exhaustively examine which nucleotide position is ambiguous in the recognition of this 8-mer, we focused on all 8-mers that are located within a Hamming distance of at most 1 (Figure 5B). Interestingly, whereas substitution of C at the second position (G at the seventh in reverse complement) to T or G maintained the AUCs of the substituted 8-mers above the threshold, the substitution to A resulted in an AUC below the threshold. This DNA-binding motif ambiguity among only the bases C, T, and G is also clearly illustrated in the shapes of the frequency distributions, where only the substitution to A diminished the unimodal shape (Figure 5C). It should be noted that this ambiguity among the bases C, T, and G cannot be directly assessed using the PWMs retrieved from the JASPAR database (Figure 5D) and factorbook (Figure 5E), and one estimated by DREME [13] (Figure 5F). In particular, the binding of GABP to AGTTCCGG (or CCGGAACT) is represented by none of the PWMs, highlighting that MOCCS is able to exhaustively evaluate non-canonical motifs.

SRF

Serum response factor (SRF) is a TF that mediates mitogen-activated protein (MAP) kinase signaling [31]. SRF belongs to the MADS box superfamily of TFs [32] and is known to recognize $CC(A/T)_6GG$, which is referred to as the CarG-box motif [32]. We applied MOCCS to SRF ChIP-Seq data from K562 cells, and the 8-mers with the highest and second-highest AUCs were ATATATGG (or CCATATAT) and CATATAAG (or CTTATATG), respectively (Figure 6A). These data are consistent with its PWMs retrieved from the JASPAR database (Figure 6B) and factorbook (Figure 6C).

Surprisingly, the third 8-mer TGACGTCA (palindromic) was not similar to the above two 8-mers. The shape of the frequency distribution suggests that this 8-mer clearly gathers around TFBSs (Figure 6D); thus, MOCCS predicted the existence of a different binding mechanism. This sequence pattern is known as the cAMP response element (CRE) [33], which is recognized by various TFs, including CRE-binding protein (CREB) [33]. Given that previous studies reported cooperative gene activation by SRF and CREB [34][35], MOCCS likely identified an indirect interaction between SRF and CRE via CREB.

The fourth 8-mer, TTGCGTCA (or TGACGCAA) was at a Hamming distance of 2 from the third 8-mer. Within CRE, G-to-T substitution at the second position was experimentally shown using the CASTing technique to have a small effect on CREB binding [36]. In addition, the A-to-G substitution at the third position was reported to maintain CREB binding to the promoter of the murine *PCNA* gene [37]. Thus, we propose that the simultaneous substitution of the two bases would also maintain the binding activity of CREB, at least in this specific genomic context.

The fifth 8-mer, ACTTCCGG (or CCGGAAGT), was not similar to the four 8-mers described above. Instead, this pattern constitutes a DNA-binding motif of GABP, which is the TF examined in the previous subsection. The existence of a relationship between these two TFs is supported by a study that compared ChIP-Seq data of both TFs and reported the proximity of their binding sites [38]. We also confirmed that 31% of the SRF-binding sites were located within ±100 bp of the GABP-

binding sites in K562 cells. However, we note that these results might also reflect an interaction between SRF and ELK4 because ELK4 (also known as SAP1) is known to bind to ACTTCCGG (Figure 6E presents the PWM retrieved from the JASPAR database) and forms a complex with SRF [39][40]. In either case, our results indicate that MOCCS identifies DNA-binding motifs that are bound by TFs via co-factors.

Evaluation of computational time

We compared the speed of MOCCS with those of the popular motif-finding programs MEME-ChIP [14] and DREME [13] using simulated and real datasets. MEME-ChIP adopts the profile-based approach and is a ChIP-Seq optimized version of MEME [19], whereas DREME adopts the consensus-based approach and searches for DNA-binding motifs that are represented by the International Union of Pure and Applied Chemistry (IUPAC) regular expression.

We applied the three programs to simulated datasets under conditions $N = 2000, 4000, 6000, 8000, 10000$ and $k = 5, 6, 7, 8$. Figure 7A presents the computational time, which increased with increasing N and k. Because MEME-ChIP required an especially long time, it was assessed only for $k = 5$. MOCCS was the fastest under many conditions, whereas DREME was the fastest in cases where k was large and N was small. We also applied MOCCS and DREME to the ENCODE dataset, which contained various ChIP-Seq data with different N values. Notably, MOCCS was considerably faster than DREME, and its computation finished within a few minutes for every condition (Figure 7B). The relatively long computational time of DREME on real datasets appeared to be attributed to its algorithm. DREME iteratively searches for DNA-binding motifs until no new motif with E-value less than a threshold is identified [13]. Because real ChIP-Seq data contain more than one motif, DREME may have iterated those processes numerous times. On the other hand, MOCCS evaluates every k-mer regardless of the number of DNA-binding motifs; thus, MOCCS would be particularly superior for real datasets.

Discussion

In this paper, we described MOCCS, which clarifies DNA-binding motif ambiguity using ChIP-Seq data. The analyses of user-provided and ChIP-Seq-dependent parameters using simulated and real datasets allow users to choose suitable parameters and datasets. MOCCS directly evaluates whether a given TF binds to each canonical or non-canonical DNA-binding motif and is applicable to many ChIP-Seq datasets without huge computational costs. Contrary to subsampling-based methods such as MEME-ChIP [14], MOCCS takes full advantage of the entire dataset to identify non-canonical DNA-binding motifs that are missed by popular methods such as DREME [13]. The key idea is the use of shapes of frequency distributions, where similar ideas are adopted by ChIPMunk [12] and SEME [20] but not for comprehensive description of DNA-binding motifs. Whereas DNA-binding motif ambiguity *in vitro* can be experimentally analyzed by protein-binding microarrays [41] and high-throughput SELEX methods [42], ambiguity *in vivo* can be efficiently and effectively analyzed by MOCCS once ChIP-Seq data are obtained.

In the current implementation, MOCCS examines frequency distributions of k-mers around TFBSs. Although this process is what makes MOCCS quite efficient, TFs with more complex DNA-binding motifs (e.g., motifs containing internal spacer sequences [27]) are also noted. Although other DNA-binding motif-identification software programs also suffer in the analysis of such DNA-binding motifs, MOCCS may be extended for application to more complex DNA-binding motifs. Because simply increasing the k value can result in worse performance due to less frequent observation of each k-mer, a possible solution may be to identify short and long k-mers simultaneously and combine those results, taking advantage of the short running time of MOCCS.

Another perspective is improvement of the threshold of rare k-mers. In the current implementation, MOCCS ignores k-mers whose total frequencies within the w-bp range are less than a threshold value that was based on a random sequence model. Alternatively, the threshold value can be calculated by considering GC% and/or first- and higher-order Markov properties of the background genomic sequences, which may lead to the discovery of rare DNA-binding motifs that are otherwise missed.

Finally, we aim at applying MOCCS to the analysis of other types of protein-nucleic acid interactions in addition to those interactions between TFs and DNA. In theory, MOCCS is applicable to any set of sequences that are collected based on their protein-binding affinities given a sufficiently large dataset. Recently, high-throughput sequencing technologies have been applied to various types of protein-nucleic acid interactions that involve RNA or epigenetically modified DNA, producing a massive amount of sequence data [43][44]. We envision that MOCCS will also contribute to clarifying ambiguity in those interactions (at least at the sequence

level) and to comprehensively describing molecular networks involved in diverse cellular processes.

Materials and Methods

Algorithm

The inputs to MOCCS include a set of w-bp DNA sequences around all TFBSs on the reference genome sequence, where the positions of TFBSs need to be determined by any ChIP-Seq peak-calling program in advance. Note that w is the range of frequency histogram analysis. In addition, a user-provided parameter k (length of DNA-binding motif) is required.

Given these data and parameters, MOCCS counts relative coordinates of every k-mer around every TFBS within the range of $\pm d$ bp, where $d = \lceil \frac{w-k+1}{2} \rceil$. For non-palindromic k-mers, the frequency distributions of reverse complementary k-mers are combined. The frequency distribution function $f(x)\{x \in \mathbf{N} : 1 \leq x \leq w-k+1\}$ is obtained for each k-mer (Note that x is the leftmost coordinate of each k-mer occurrence). In addition, k-mers whose total frequencies were less than an expected frequency $C_{N,k} = \frac{N(w-k+1)}{4^k}$ were ignored hereinafter.

We defined the distance from the TFBS positions as follows:

$$\text{distance}(x) = \begin{cases} center_l - x + 1, & 1 \leq x \leq center_l \\ x - center_r + 1, & center_r \leq x \leq w-k+1, \end{cases}$$

where

$$center_l = d$$

and

$$center_r = \begin{cases} d, & \text{if } w \text{ and } k \text{ are both even or odd} \\ d+1, & \text{otherwise.} \end{cases}$$

The cumulative relative frequency distribution $F(x)$ of each k-mer is calculated as follows:

$$F(x) = \frac{\sum_{i \in \{1 \leq \text{distance}(i) \leq x\}} f(i)}{\sum_{j \in \{1 \leq \text{distance}(j) \leq d\}} f(j)} \quad \{x \in \mathbf{N} : 1 \leq x \leq d\}.$$

Then, the AUC is calculated as follows:

$$AUC = \sum_{1 \leq x \leq d} \left(F(x) - \frac{x}{d} \right).$$

Finally, MOCCS outputs k-mers that have AUCs larger than a threshold value. The threshold value was set to $5 \cdot \sigma_{\text{AUC\_bg}}$, where $\sigma_{\text{AUC\_bg}}$ is the standard deviation of AUCs of all k-mers in background sequences. We adopted w-bp upstream sequences of all transcription start sites as the background for application to the real ChIP-Seq data in this study (Supplementary Table 2). In the simulation analysis, only the rank of "true" k-mer was considered, and the thresholding step was skipped.

Simulation analysis

In each simulation, we randomly generated N w-bp DNA sequences and a "true" k-mer. α% of the N sequences were subsequently edited so that they contain the true k-mer, where their positions followed a Gaussian distribution whose mean was the center of the sequences and whose standard deviation was σ. For each set of tested parameters, we repeated the simulation 100 times and averaged the AUC ranks of the true k-mer.

Datasets used in real data analysis and estimation of σ

The human genome sequence (hg19) was obtained from UCSC Genome Browser (http://genome.ucsc.edu/) [45] as the reference genome sequence. ChIP-Seq peak data were obtained from the ENCODE project [1] (Supplementary Table 1) and TFBSs were defined as the centers of each peak region described in broadPeak or narrowPeak files. As described in Results, w was set to 701 bp. For real data analysis of MOCCS, the repeat-masked sequences were used. To calculate $\sigma_{\text{AUC\_bg}}$ for the threshold-value setting, we obtained w-bp repeat-masked upstream sequences of all transcription start sites of annotated genes in GENCODE Basic v19 [46] via UCSC Table Browser [47].

To evaluate the biologically meaningful range of σ, which represents the sharpness of the TF-bound k-mer frequency distributions around TFBSs, we downloaded position frequency matrices of the ENCODE-analyzed TFs from the JASPAR database [27]. The frequency distributions were obtained by applying MOODS v1.0.2.1 [48] to ±250-bp sequences around the TFBSs with the threshold $p < 0.0001$. Gaussian distributions were fit to the frequency distributions using the norm.fit function of the scipy.stats module (version 0.12.0). The ELK4 motif was based on ChIP-Seq, while the USF1, GABP, and SRF motifs were based on SELEX in JASPAR. For additional comparison regarding the USF1, GABP, and SRF motifs, sequence logos that are based on sequences of the top 500 peaks of the ENCODE ChIP-Seq data were downloaded from factorbook [28].

Motif finding by DREME and computational time evaluation

Motif finding using DREME v4.10.4 [13] on the ENCODE ChIP-Seq data was conducted with default parameters except for $k = 8$ and $w = 201$ bp. Note that w was set much smaller than the value used in MOCCS because input sequences of DREME should contain true DNA-binding motifs at much higher density than those of MOCCS, which utilizes the shapes of the distributions. The output motifs were visualized by sequence logos using the R package 'seqLogo' (version 1.34.0) [49].

The computational times of MOCCS, DREME, and MEME-ChIP v4.10.4 [14] were recorded on Intel Xeon E5-2670 and 64 GB of main memory. w was set to 201 bp for DREME and MEME-ChIP, and k was set to the length of the true k-mer for MOCCS and DREME. Other parameters of DREME and MEME-ChIP were set default. We set $\sigma = 60$ for generating simulated dataset. Each program was applied to each of 100 simulated datasets under the same condition, and the computational times were averaged.

Implementation

We implemented MOCCS in Perl and R languages. The source codes are available at:
https://github.com/yuifu/moccs.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

HO and WI designed the study. HO performed data analysis. HO and WI wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Hikari Yoshitane, Hideki Terajima, and Yoshitaka Fukada at the University of Tokyo for discussing the biological interpretation of the MOCCS results. We also thank Takaho A. Endo at IMS-RIKEN for providing constructive comments. This work was supported by the Japan Science and Technology Agency (CREST), the Japan Society for the Promotion of Science (KAKENHI 23710231), and the Ministry of Education, Culture, Sports, Science, and Technology in Japan (KAKENHI 221S0002). Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

References

[1] The ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (2012) 57–74. doi:10.1038/nature11247.

[2] Roadmap Epigenomics Consortium, et al., Integrative analysis of 111 reference human epigenomes, Nature 518 (7539) (2015) 317–330. doi:10.1038/nature14248.

[3] F. W. Albert, L. Kruglyak, The role of regulatory variation in complex traits and disease, Nature Publishing Group 16 (February) (2015) 197–212. doi:10.1038/nrg3891.

[4] S.-H. Yoo, C. H. Ko, P. L. Lowrey, E. D. Buhr, E.-j. Song, S. Chang, O. J. Yoo, S. Yamazaki, C. Lee, J. S. Takahashi, A noncanonical E-box enhancer drives mouse Period2 circadian oscillations in vivo., Proceedings of the National Academy of Sciences of the United States of America 102 (7) (2005) 2608–13. doi:10.1073/pnas.0409763102.

[5] J. J. Jordan, D. Menendez, A. Inga, M. Nourredine, D. Bell, M. A. Resnick, Noncanonical DNA motifs as transactivation targets by wild type and mutant p53, PLoS Genetics 4 (6). doi:10.1371/journal.pgen.1000104.

[6] Y. Kumaki, M. Ukai-Tadenuma, K.-i. D. Uno, J. Nishio, K.-h. Masumoto, M. Nagano, T. Komori, Y. Shigeyoshi, J. B. Hogenesch, H. R. Ueda, Analysis and synthesis of high-amplitude Cis-elements in the mammalian circadian clock., Proceedings of the National Academy of Sciences of the United States of America 105 (39) (2008) 14946–51. doi:10.1073/pnas.0802636105.

[7] G. D. Stormo, DNA binding sites: representation and discovery., Bioinformatics (Oxford, England) 16 (1) (2000) 16–23. doi:10.1093/bioinformatics/16.1.16.

[8] F. Zambelli, G. Pesole, G. Pavesi, Motif discovery and transcription factor binding sites before and after the next-generation sequencing era., Briefings in bioinformatics 14 (2) (2013) 225–37. doi:10.1093/bib/bbs016.

[9] A. A. Sharov, M. S. H. Ko, Exhaustive search for over-represented DNA sequence motifs with cisfinder, DNA Research 16 (2009) 261–273. doi:10.1093/dnares/dsp014.

[10] L. Li, GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery., Journal of computational biology : a journal of computational molecular cell biology 16 (2) (2009) 317–329. doi:10.1089/cmb.2008.16TT.

[11] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, C. K. Glass, Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities., Molecular cell 38 (4) (2010) 576–89. doi:10.1016/j.molcel.2010.05.004.

[12] I. V. Kulakovskiy, V. a. Boeva, a. V. Favorov, V. J. Makeev, Deep and wide digging for binding motifs in ChIP-Seq data, Bioinformatics 26 (20) (2010) 2622–2623. doi:10.1093/bioinformatics/btq488.

[13] T. L. Bailey, DREME: Motif discovery in transcription factor ChIP-seq data, Bioinformatics 27 (12) (2011) 1653–1659. doi:10.1093/bioinformatics/btr261.

[14] P. Machanick, T. L. Bailey, MEME-ChIP: Motif analysis of large DNA datasets, Bioinformatics 27 (12) (2011) 1696–1697. doi:10.1093/bioinformatics/btr189.

[15] J. E. Reid, L. Wernisch, STEME: Efficient EM to find motifs in large data sets, Nucleic Acids Research 39 (18). doi:10.1093/nar/gkr574.

[16] X. Ma, A. Kulkarni, Z. Zhang, Z. Xuan, R. Serfling, M. Q. Zhang, A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information., Nucleic acids research 40 (7). doi:10.1093/nar/gkr1135.

[17] H. Hartmann, E. W. Guthöhrlein, M. Siebert, S. Luehr, J. Söding, P-value-based regulatory motif discovery using positional weight matrices, Genome Research 23 (2013) 181–194. doi:10.1101/gr.139881.112.

[18] S. Georgiev, A. P. Boyle, K. Jayasurya, X. Ding, S. Mukherjee, U. Ohler, Evidence-ranked motif identification., Genome biology 11 (2010) R19. doi:10.1186/gb-2010-11-2-r19.

[19] T. L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers., in: R. Altman, D. Brutlag, P. Karp, R. Lathrop, D. Searls (Eds.), Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, Vol. 2, Department of Computer Science and Engineering, University of California at San Diego, La Jolla 92093-0114, USA., AAAI Press, 1994, pp. 28–36.

[20] Z. Zhang, C. W. Chang, W. Hugo, E. Cheung, W.-K. Sung, Simultaneously Learning DNA Motif Along with Its Position and Sequence Rank Preferences Through Expectation Maximization Algorithm, Journal of Computational Biology 20 (3) (2013) 237–248. doi:10.1089/cmb.2012.0233.

[21] H. Yoshitane, H. Ozaki, H. Terajima, N.-H. Du, Y. Suzuki, T. Fujimori, N. Kosaka, S. Shimba, S. Sugano, T. Takagi, W. Iwasaki, Y. Fukada, CLOCK-controlled polyphonic regulation of circadian rhythms through canonical and noncanonical E-boxes., Molecular and cellular biology 34 (10) (2014) 1776–87. doi:10.1128/MCB.01465-13.

[22] P. Pajukanta, H. E. Lilja, J. S. Sinsheimer, R. M. Cantor, A. J. Lusis, M. Gentile, X. J. Duan, A. Soro-Paavonen, J. Naukkarinen, J. Saarela, M. Laakso, C. Ehnholm, M.-R. Taskinen, L. Peltonen, Familial combined hyperlipidemia is associated with upstream transcription factor 1 (USF1)., Nature genetics 36 (4) (2004) 371–376. doi:10.1038/ng1320.

[23] J. M. Grégoire, P. H. Roméo, T-cell expression of the human GATA-3 gene is regulated by a non-lineage-specific silencer., The Journal of biological chemistry 274 (10) (1999) 6567–6578. doi:10.1074/jbc.274.10.6567.

[24] M. L. Read, a. R. Clark, K. Docherty, The helix-loop-helix transcription factor USF (upstream stimulating factor) binds to a regulatory sequence of the human insulin gene enhancer., The Biochemical journal 295 (Pt 1 (1 993) (1993) 233–237.

[25] K. Yasumoto, K. Yokoyama, K. Shibata, Y. Tomita, S. Shibahara, Microphthalmia-associated transcription factor as a regulator for melanocyte-specific transcription of the human tyrosinase gene., Molecular and cellular biology 14 (12) (1994) 8058–8070. doi:10.1128/MCB.14.12.8058.Updated.

[26] K. Takahashi, C. Nishiyama, M. Nishiyama, K. Okumura, C. Ra, Y. Ohtake, T. Yokota, A complex composed of USF1 and USF2 activates the human FcepsilonRI alpha chain expression via a CAGCTG element in the first intron., European journal of immunology 31 (2) (2001) 590–599.

[27] A. Mathelier, X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, D. J. Arenillas, S. Buchman, C. Y. Chen, A. Chou, H. Ienasescu, J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin, W. W. Wasserman, JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles, Nucleic Acids Research 42 (2014) D142–147. doi:10.1093/nar/gkt997.

[28] J. Wang, J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. J. Rando, E. Birney, R. M. Myers, W. S. Noble, M. Snyder, Z. Weng, Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors., Genome research 22 (9) (2012) 1798–1812. doi:10.1101/gr.139105.112.

[29] N. Bannert, a. Avots, M. Baier, E. Serfling, R. Kurth, GA-binding protein factors, in concert with the coactivator CREB binding protein/p300, control the induction of the interleukin 16 promoter in T lymphocytes., Proceedings of the National Academy of Sciences of the United States of America 96 (4) (1999) 1541–1546. doi:10.1073/pnas.96.4.1541.

[30] H.-H. Xue, J. Bollenbacher, V. Rovella, R. Tripuraneni, Y.-B. Du, C.-Y. Liu, A. Williams, J. P. McCoy, W. J. Leonard, GA binding protein regulates interleukin 7 receptor alpha-chain gene expression in T cells., Nature immunology 5 (10) (2004) 1036–1044. doi:10.1038/ni1117.

[31] A. J. Whitmarsh, P. Shore, A. D. Sharrocks, R. J. Davis, Integration of MAP kinase signal transduction pathways at the serum response element., Science (New York, N.Y.) 269 (5222) (1995) 403–407. doi:10.1126/science.7618106.

[32] S. Arsenian, B. Weinhold, M. Oelgeschläger, U. Rüther, a. Nordheim, Serum response factor is essential for mesoderm formation during mouse embryogenesis., The EMBO journal 17 (21) (1998) 6289–6299. doi:10.1093/emboj/17.21.6289.

[33] B. Mayr, M. Montminy, Transcriptional regulation by the phosphorylation-dependent factor CREB., Nature reviews. Molecular cell biology 2 (8) (2001) 599–609. doi:10.1038/35085068.

[34] S. Ramirez, S. Ait Si Ali, P. Robin, D. Trouche, A. Harel-Bellan, The CREB-binding protein (CBP) cooperates with the serum response factor for transactivation of the c-fos serum response element, Journal of Biological Chemistry 272 (49) (1997) 31016–31021. doi:10.1074/jbc.272.49.31016.

[35] C. a. Herndon, N. Ankenbruck, B. Lester, J. Bailey, L. Fromm, Neuregulin1 signaling targets SRF and CREB and activates the muscle spindle-specific gene Egr3 through a composite SRF-CREB-binding site, Experimental Cell Research 319 (5) (2013) 718–730. doi:10.1016/j.yexcr.2013.01.001.

[36] D. M. Benbrook, N. C. Jones, Different binding specificities and transactivation of variant CRE's by CREB complexes., Nucleic acids research 22 (8) (1994) 1463–1469. doi:10.1093/nar/22.8.1463.

[37] D. J. Orten, J. M. Strawhecker, S. D. Sanderson, D. Huang, M. B. Prystowsky, S. H. Hinrichs, Differential effects of monoclonal antibodies on activating transcription factor-1 and cAMP response element binding protein interactions with DNA., The Journal of biological chemistry 269 (51) (1994) 32254–32263.

[38] A. Valouev, D. S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. M. Myers, A. Sidow, Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data., Nature methods 5 (9) (2008) 829–834. doi:10.1038/nmeth.1246.

[39] G. Buchwalter, C. Gross, B. Wasylyk, Ets ternary complex transcription factors, Gene 324 (1-2) (2004) 1–14. doi:10.1016/j.gene.2003.09.028.

[40] S. J. Cooper, N. D. Trinklein, L. Nguyen, R. M. Myers, Serum response factor binding sites differ in three human cell types, Genome Research 17 (2) (2007) 136–144. doi:10.1101/gr.5875007.

[41] M. F. Berger, A. a. Philippakis, A. M. Qureshi, F. S. He, P. W. Estep, M. L. Bulyk, Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities., Nature biotechnology 24 (11) (2006) 1429–1435. doi:10.1038/nbt1246.

[42] A. Jolma, T. Kivioja, J. Toivonen, L. Cheng, G. Wei, M. Enge, M. Taipale, J. M. Vaquerizas, J. Yan, M. J. Sillanpaa, M. Bonke, K. Palin, S. Talukder, T. R. Hughes, N. M. Luscombe, E. Ukkonen, J. Taipale, Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities, Genome Research 20 (6) (2010) 861–873. doi:10.1101/gr.100552.109.

[43] J. König, K. Zarnack, N. Luscombe, J. Ule, ProteinRNA interactions: new genomic technologies and perspectives, Nature Re-

views Genetics 13 (2) (2012) 77–83.

[44] T. S. Furey, ChIPseq and beyond: new and improved methodologies to detect and characterize proteinDNA interactions, Nature Reviews Genetics 13 (December) (2012) 840–852. doi:10.1038/nrg3306.

[45] K. R. Rosenbloom, J. Armstrong, G. P. Barber, J. Casper, H. Clawson, M. Diekhans, T. R. Dreszer, P. a. Fujita, L. Guruvadoo, M. Haeussler, R. a. Harte, S. Heitner, G. Hickey, A. S. Hinrichs, R. Hubley, D. Karolchik, K. Learned, B. T. Lee, C. H. Li, K. H. Miga, N. Nguyen, B. Paten, B. J. Raney, A. F. a. Smit, M. L. Speir, A. S. Zweig, D. Haussler, R. M. Kuhn, W. J. Kent, The UCSC Genome Browser database: 2015 update., Nucleic acids research 43 (Database issue) (2015) D670–81. doi:10.1093/nar/gku1177.

[46] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. Van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, T. J. Hubbard, GENCODE: The reference human genome annotation for the ENCODE project, Genome Research 22 (9) (2012) 1760–1774. doi:10.1101/gr.135350.111.

[47] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, W. J. Kent, The UCSC Table Browser data retrieval tool., Nucleic acids research 32 (Database issue) (2004) D493–D496. doi:10.1093/nar/gkh103.

[48] J. Korhonen, P. Martinmäki, C. Pizzi, P. Rastas, E. Ukkonen, MOODS: fast search for position weight matrix matches in DNA sequences., Bioinformatics (Oxford, England) 25 (23) (2009) 3181–2. doi:10.1093/bioinformatics/btp554.

[49] SeqLogo, Bembom O. seqLogo: Sequence logos for DNA sequence alignments. R package version 1.34.0.

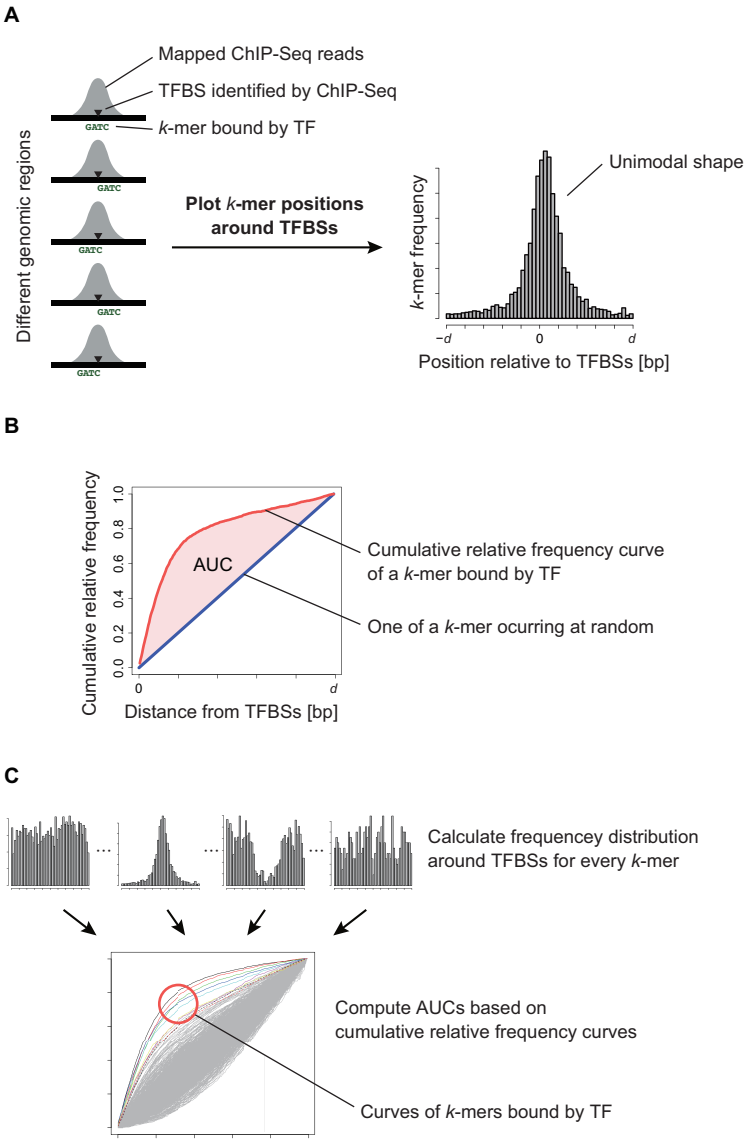


Figure 1: MOCCS workflow. (A) In ChIP-Seq experiments, frequency distributions of TF-binding *k*-mers around TFBSs exhibit unimodal shapes. (B) The sharpness of a unimodal distribution is quantified as an AUC of the cumulative relative frequency curve. Note that unsigned distances are used. (C) AUCs are calculated for all *k*-mers, and those with large AUCs are output as TF-binding *k*-mers.

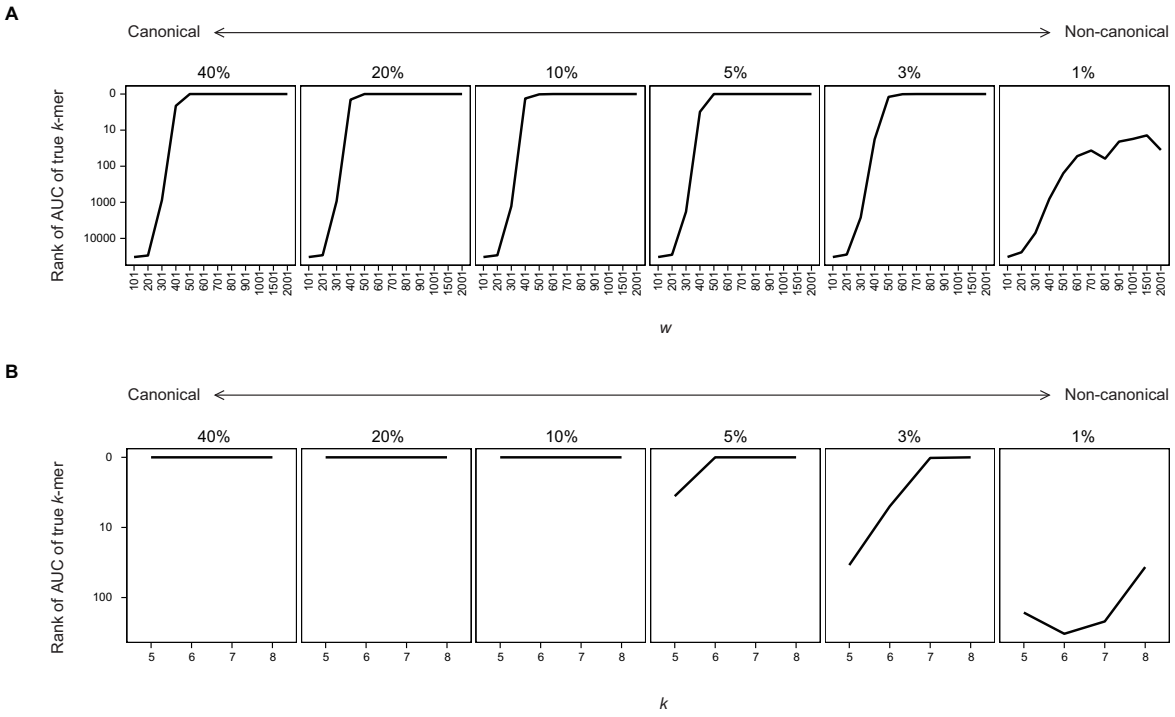


Figure 2: Simulation analysis of user-provided parameters. Averaged ranks of true k-mers for various (A) w and (B) k values are plotted. Several α values (fraction of input sequences that contain the true k-mer) were examined.

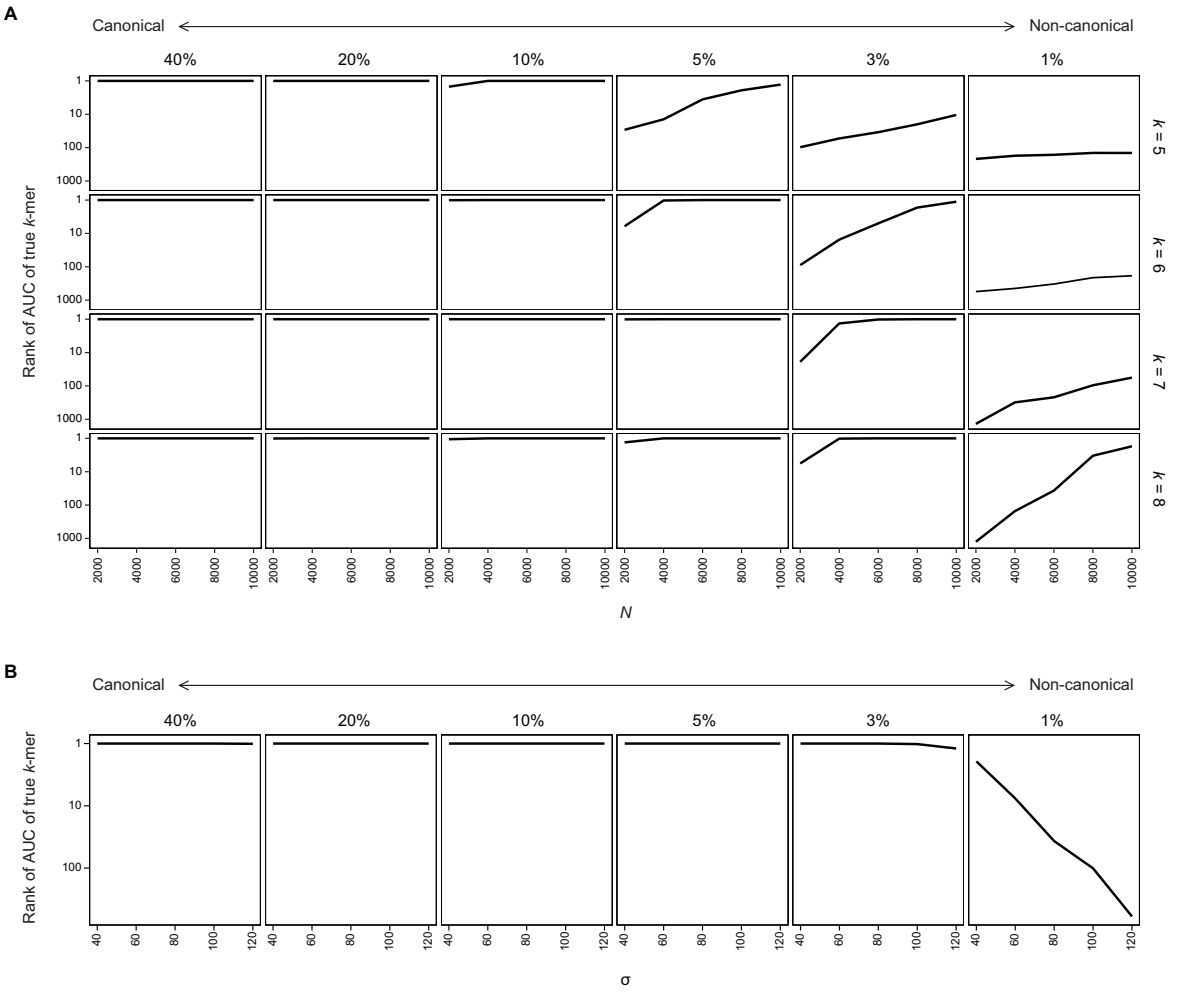
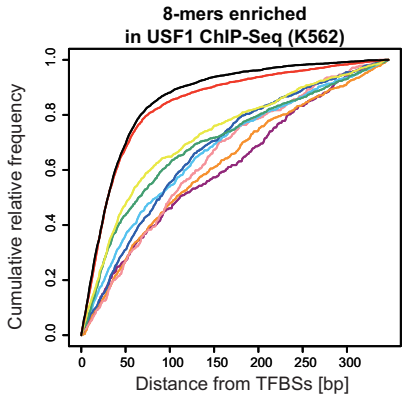


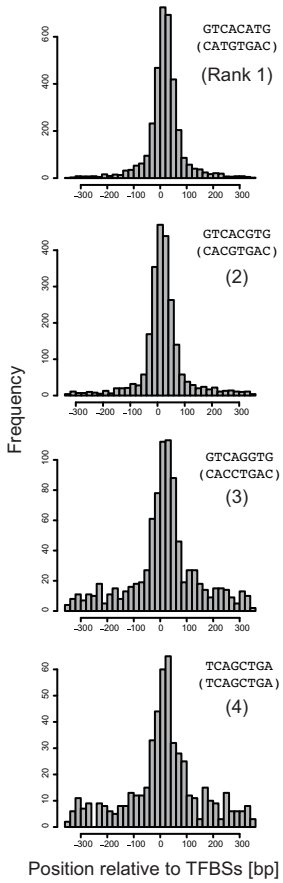
Figure 3: Simulation analysis of TF- and experiment-specific variables. (A) Averaged ranks of true k-mers for various values of N and α. $k = 5, 6, 7, 8$ were examined. (B) Averaged ranks of true k-mers for various values of σ and α. k was set to 8.

A

Rank	Line color	8-mer	AUC	Count
1		GTCACGTG	125.39	3447
2		GTCACATG	118.42	2367
3		GTCAGGTG	78.68	896
4		TCAGCTGA	69.28	523
5		ACTTCCGG	61.73	648
6		TGATTGGC	60.08	669
7		CGCATGCG	51.49	279
8		CCAATGGG	42.20	489
9		GTGATAAG	38.61	293



B



C USF1 motif in JASPAR

D USF1 motifs in factorbook

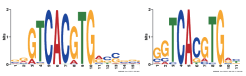


Figure 4: Analysis of USF1 ChIP-Seq data. (A) Identified USF1-binding 8-mers and their cumulative relative frequency curves. (B) Frequency distributions around TFBSs of four 8-mers that showed the largest AUCs. (C) A sequence logo representing the DNA-binding motif of USF1 in JASPAR database (MA0093.1). (D) Sequence logos representing the DNA-binding motifs of USF1 in factorbook.

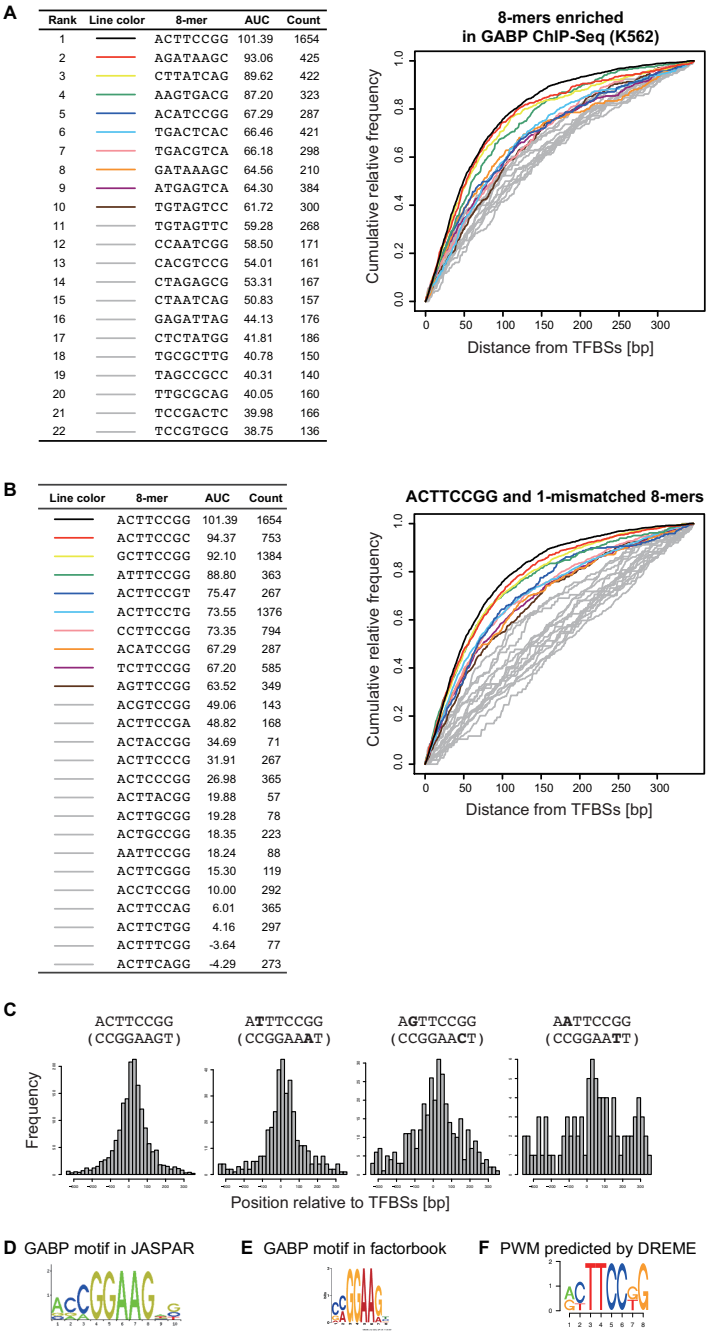


Figure 5: Analysis of GABP ChIP-Seq data. (A) Identified GABP-binding 8-mers and their cumulative relative frequency curves. (B) The cumulative relative frequency curves of GABP-binding 8-mers that are within Hamming distance of at most 1 from the largest AUC 8-mer, ACTTCCGG (or CCGGAAGT in reverse complement). (C) Frequency distributions around TFBSs of the largest AUC 8-mer and those that have a substitution of C at the second position (G at the seventh position in reverse complement). (D) A sequence logo representing the DNA-binding motif of GABP in the JASPAR database (MA0093.1). (E) A sequence logo representing the DNA-binding motif of GABP in factorbook. (F) A sequence logo representing a DNA-binding motif of GABP predicted by DREME.

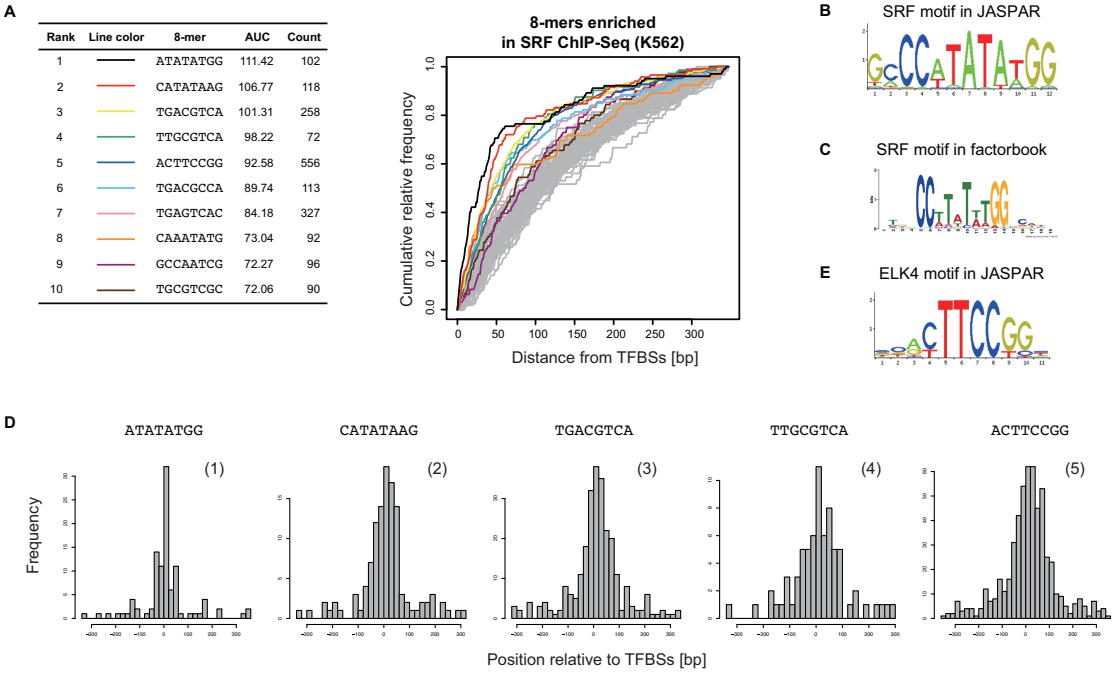


Figure 6: Analysis of SRF ChIP-Seq data. (A) Identified SRF-binding 8-mers and their cumulative relative frequency curves. Only the top ten 8-mers are presented in the list. (B) A sequence logo representing the DNA-binding motif of SRF in the JASPAR database (MA0083.1). (C) A sequence logo representing the DNA-binding motif of SRF in factorbook. (D) Frequency distributions around TFBSs of five 8-mers that exhibited the largest AUCs. (E) A sequence logo representing the DNA-binding motif of ELK4 in the JASPAR database (MA0076.2).

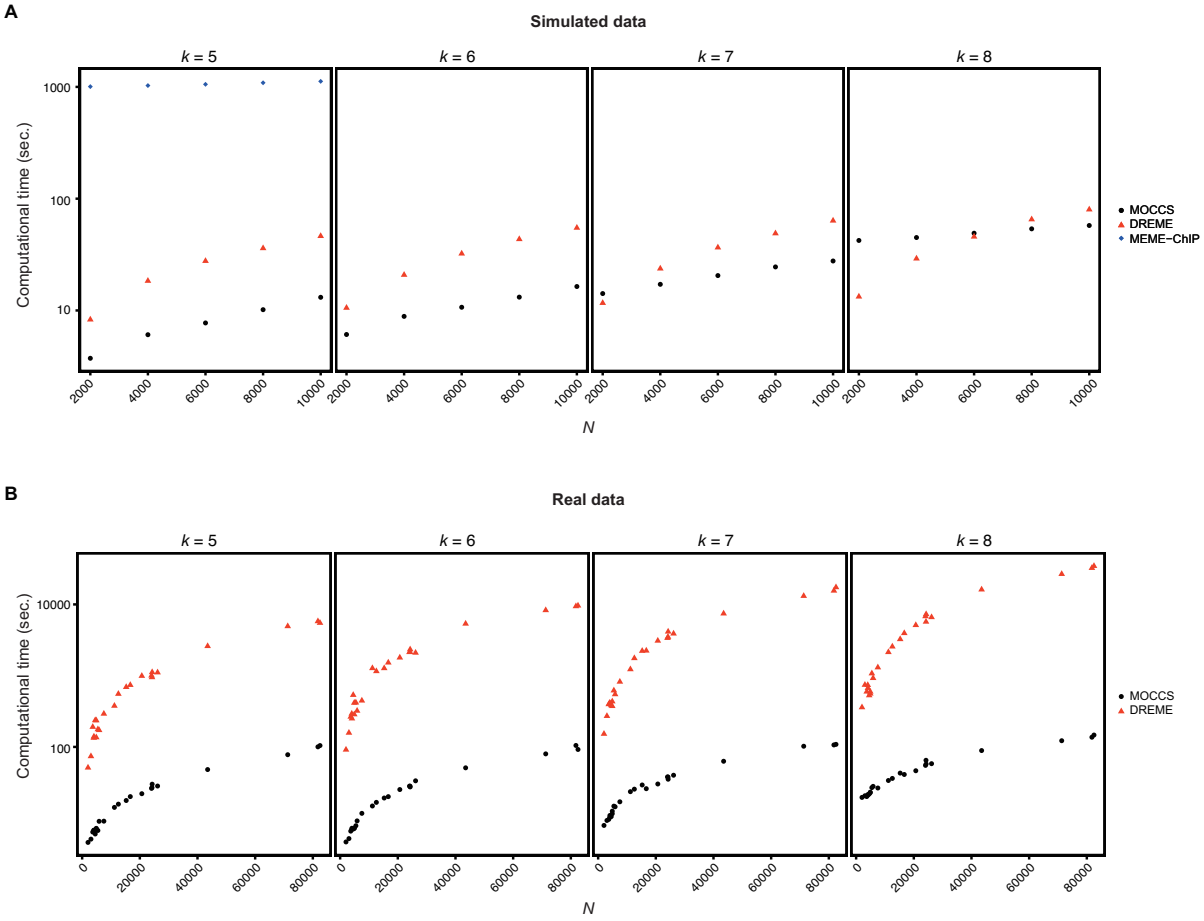


Figure 7: Evaluation of computational time on simulated and real data. Computational times on (A) simulated and (B) real ChIP-Seq data are plotted. The black circles, red triangles, and blue diamonds represent computational times of MOCCS, DREME, and MEME-ChIP, respectively.

Supplementary table 1 ENCODE ChIP-Seq datasets used for estimating σ

ChIP-seq data	TF	Cell type	# of sequences	PWM	# of MOODS hits (p<0.00001)	Mean	Standard deviation
wgEncodeOpenChromChipGm12878CmycPk.narrowPeak	cMyc	Gm12878	24277	MA0147.1	1990	-1.16	113.27
wgEncodeOpenChromChipH1hescCmycPk.narrowPeak	cMyc	H1hesc	20692	MA0147.1	1029	-5.63	131.49
wgEncodeOpenChromChipK562CmycPk.narrowPeak	cMyc	K562	43514	MA0147.1	3616	-0.87	107.83
wgEncodeOpenChromChipGm12878CtcfPkRep1.narrowPeak	CTCF	Gm12878	71290	MA0139.1	33779	-5.49	69.26
wgEncodeOpenChromChipH1hescCtcfPk.narrowPeak	CTCF	H1hesc	82485	MA0139.1	29452	-5.22	75.35
wgEncodeOpenChromChipK562CtcfPk.narrowPeak	CTCF	K562	81765	MA0139.1	37809	-5.01	69.23
wgEncodeHaibTfbsGm12878GabpPcr2xPkRep1.broadPeak	Gabp	Gm12878	4533	MA0062.2	1941	5.53	84.93
wgEncodeHaibTfbsGm12878GabpPcr2xPkRep2.broadPeak	Gabp	Gm12878	4897	MA0062.2	1978	3.22	88.79
wgEncodeHaibTfbsH1hescGabpPcr1xPkRep1.broadPeak	Gabp	H1hesc	16695	MA0062.2	2391	6.18	102.79
wgEncodeHaibTfbsH1hescGabpPcr1xPkRep2.broadPeak	Gabp	H1hesc	4900	MA0062.2	1333	9.20	94.29
wgEncodeHaibTfbsK562GabpV0416101PkRep1.broadPeak*	Gabp	K562	12559	MA0062.2	3219	8.91	91.19
wgEncodeHaibTfbsK562GabpV0416101PkRep2.broadPeak	Gabp	K562	15276	MA0062.2	3306	5.83	108.00
wgEncodeHaibTfbsGm12878SrfPcr2xPkRep1.broadPeak	Srf	Gm12878	3677	MA0083.1	584	4.66	66.92
wgEncodeHaibTfbsGm12878SrfPcr2xPkRep2.broadPeak	Srf	Gm12878	4086	MA0083.1	907	6.50	62.26
wgEncodeHaibTfbsH1hescSrfPcr1xPkRep1.broadPeak	Srf	H1hesc	4031	MA0083.1	2308	10.31	46.47
wgEncodeHaibTfbsH1hescSrfPcr1xPkRep2.broadPeak	Srf	H1hesc	3049	MA0083.1	1366	9.42	52.91
wgEncodeHaibTfbsK562SrfV0416101PkRep1.broadPeak*	Srf	K562	5453	MA0083.1	1260	6.39	62.95
wgEncodeHaibTfbsK562SrfV0416101PkRep2.broadPeak	Srf	K562	2028	MA0083.1	773	11.69	45.38
wgEncodeHaibTfbsGm12878Usf1Pcr2xPkRep1.broadPeak	Usf1	Gm12878	7531	MA0093.1	1916	15.29	61.78
wgEncodeHaibTfbsGm12878Usf1Pcr2xPkRep2.broadPeak	Usf1	Gm12878	5883	MA0093.1	1430	15.47	73.13
wgEncodeHaibTfbsH1hescUsf1Pcr1xPkRep1.broadPeak	Usf1	H1hesc	24226	MA0093.1	3377	14.05	68.59
wgEncodeHaibTfbsH1hescUsf1Pcr1xPkRep2.broadPeak	Usf1	H1hesc	26157	MA0093.1	3340	14.90	74.16
wgEncodeHaibTfbsK562Usf1V0416101PkRep1.broadPeak*	Usf1	K562	24059	MA0093.1	4255	12.55	68.88
wgEncodeHaibTfbsK562Usf1V0416101PkRep2.broadPeak	Usf1	K562	11187	MA0093.1	2259	13.93	73.97

Mean and Standard deviation are the mean and standard deviation of fitted Gaussian distribution, respectively.
* Data used for real data analysis

Supplementary table 2 Standard deviation of AUCs in background genomic sequences

k	Mean	Standard deviation	Mean (no mask)	Standard deviation (no mask)
5	-0.086	3.373	0.005	3.385
6	0.251	4.192	0.332	4.061
7	0.165	5.214	0.221	4.896
8	0.544	7.675	0.574	7.008

Mean and standard deviation of AUCs for all *k*-mers. The values calculated from datasets with and without repeat masking are shown.