Systematic Discovery of Cofactor Motifs from ChIP-seq Data by SIOMICS

Jun Ding¹, Vikram Dhillon², Xiaoman Li^{2*}, Haiyan Hu^{1*}

¹Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL, 32816, USA

²Burnett School of Biomedical Science, College of Medicine, University of Central Florida, Orlando, FL, 32816, USA

*Corresponding authors

Email addresses:

JD: jding@cs.ucf.edu

VD: dhillonv10@knights.ucf.edu

XL: xiaoman@mail.ucf.edu

HH: haihu@cs.ucf.edu

Abstract

Understanding transcriptional regulatory elements and particularly the transcription factor binding sites represents a significant challenge in computational biology. The chromatin immunoprecipitation followed by massive parallel sequencing (ChIP-seq) experiments provide an unprecedented opportunity to study transcription factor binding sites on the genome-wide scale. Here we describe a recently developed tool, SIOMICS, to systematically discover motifs and binding sites of transcription factors and their cofactors from ChIP-seq data. Unlike other tools, SIOMICS explores the co-binding properties of multiple transcription factors in short regions to predict motifs and binding sites. We have previously shown that the original SIOMICS method predicts motifs and binding sites of more cofactors in more accurate and time-effective ways than two popular methods. In this paper, we present the extended SIOMICS method, SIOMICS_Extension, and demonstrate its usage for systematic discovery of cofactor motifs and binding sites. The SIOMICS tool, including SIOMICS and SIOMICS_Extension, are available at <http://www.eecs.ucf.edu/~xiaoman/SIOMICS/SIOMICS.html>.

Keywords: transcription factor, cofactor, motif, transcription factor binding sites, ChIP-seq, SIOMICS.

1. Introduction

The discovery of transcription factor binding sites (TFBSs) and binding motifs is critical to understand gene transcriptional regulation (1, 2). TFBSs are 5-14 base-pair long nucleotide combinations of DNA that are bound by regulatory proteins called transcription factors (TFs), which modulate the expression of genes near the TFBSs. A TF often binds similar TFBSs, the common pattern of which is called a motif (Figure 1) (3). In eukaryotic DNA sequences, generally multiple TFBSs occur nearby in a short DNA region that determines the expression of genes (4). Such a short region containing multiple TFBSs, normally several hundred base-pair long and rarely more than 2000 base-pair long, is called a cis-regulatory module (CRM) (2, 4, 5) (Figure 1). We define a motif module as a collection of motifs or TFs with their TFBSs co-occurring in a significantly many short DNA regions (i.e., CRMs) (6-9) (Figure 1). Statistically, the likelihood of a motif module occurring in N sequences is much smaller than that of a motif (9). Therefore, the approach of identifying TFBSs and motifs through the prediction of motif modules and CRMs is less error-prone than that of detecting TFBSs from individual TFs.

Chromatin immunoprecipitation followed by massive parallel sequencing (ChIP-seq) experiments are generating enormous amount of data that can be harnessed by computational methods to identify biologically-relevant TFBSs and motifs (10, 11). With the high-quality antibody to a given TF, ChIP-seq experiments identify TF-binding enriched regions for this TF called ChIP-seq peak regions (12). ChIP-seq peak regions are in general several hundred base-pair long and enriched with TFBSs of the TF under consideration and TFBSs of the cofactors of this TF (10, 11). With hundreds or thousands of ChIP-seq peaks, computational methods can then search for motifs and TFBSs in the top ChIP-seq peak regions, all ChIP-seq peak regions, or neighboring regions around ChIP-seq peaks (6, 13-15).

Many computational methods (6, 13, 14, 16-20) have been developed to discover TFBSs and motifs from ChIP-seq data (Table 1). These methods can be broadly classified into two categories. One category consists of methods that discover one motif and its TFBSs at a time (13, 16, 18-20). The other category contains methods that consider multiple motifs simultaneously to discover TFBSs and motifs (6, 14, 17). As indicated above, the second category likely predicts motifs and TFBSs more accurately in a ChIP-seq dataset than the first category. In the following, we focus on one of the methods from the second category, SIOMICS (14). To the best of our knowledge, SIOMICS is the first method that considers multiple motifs simultaneously and de novo discovers motifs and TFBSs in a ChIP-seq dataset. It also enhances the prediction of cofactor motifs and under-represented motifs and TFBSs from ChIP-seq data.

In the following sections, we will first describe the original SIOMICS method (SIOMICS) and its extension (SIOMICS_Extension) in the Material and Methods section. Consequently we discuss the performance of SIOMICS_Extension on ChIP-seq datasets in the Results section. Next, in the Results section, we will discuss how to apply SIOMICS_Extension to next

generation sequencing datasets for motif and TFBS discovery. Finally, we provide more insight about SIOMICS_Extension in the Discussion section.

2. Material and Methods

2.1 *The original SIOMICS method*

The method implemented in the original SIOMICS tool has been delineated in (14). In brief, it starts from a group of input sequences. These sequences come from the extended regions centered around the original ChIP-seq peak regions. The reason for extending the ChIP-seq peak regions is that TFBSs of certain cofactors may not be included in the original ChIP-seq peak regions (14). With the input sequences, SIOMICS ranks all w -mer (w base pair long DNA segment, default $w=8$) patterns in input sequences, based on their likelihood scores (21) from the largest to the smallest. SIOMICS then iteratively chooses top m w -mer patterns to count their co-occurrence frequencies (6) and assess the statistical significance of the co-occurrence frequencies (9). The top m w -mer patterns used in every iteration are different, so that individually insignificant patterns may be identified as motifs because of the significant co-occurrence of their instances with the instances of the predicted motifs so far. Each group of statistically significant co-occurring w -mer patterns discovered in an iteration is presented as a predicted motif module in that iteration, and each w -mer pattern in these predicted motif modules as a predicted motif. The co-occurring instances of the w -mer patterns of a motif module in input sequences are defined as the TFBSs of the motifs in this motif module. Based on these TFBSs, the position weight matrices are constructed to represent each predicted motif. The iteration is repeated until m motifs are discovered or no new motif is found in r consecutive iterations. Here, m and r are parameters input by users. At the end, there will be no more than m predicted motifs, each of which is w base pairs long (default $w=8$).

2.2 *The SIOMICS_Extension method*

The major improvement of the SIOMICS tool in the extended version is its ability to output motifs of variable lengths (Figure 2). In practice, the common motif length varies from 5 to 15 base pairs (22). The original SIOMICS tool may thus cause certain problems by requiring the predicted motif length be always w base pairs long.

To address this issue, SIOMICS_Extension first considers extending these motifs. For each discovered motif, SIOMICS_Extension collects all of its predicted TFBSs by SIOMICS. Note that many segments in input sequences that are exactly the same as the predicted TFBSs of this motif may not be considered as the predicted TFBSs, due to the fact that they do not co-occur with the TFBSs of other motifs in the predicted motif modules containing this motif. With the predicted TFBSs of this motif, SIOMICS_Extension considers the neighboring nucleotides around the predicted TFBSs. The intuition behind this approach is that if this motif is longer than w , it is expected that the neighboring nucleotides around most predicted TFBSs likely share a common nucleotide, on the left side, on the right side, or on both sides. SIOMICS_Extension

applies a binomial test to assess whether the neighboring positions are significantly sharing a type of nucleotides or dinucleotides and then decides whether to extend the motif length. In brief, SIOMICS_Extension identifies the most frequent type of nucleotides at a neighboring position and records its frequency, say K . If no dominant type of nucleotides exists at this position, SIOMICS considers two adjacent neighboring positions. If no dominant type of dinucleotides exists at the two positions, SIOMICS_Extension stops extending from the side under consideration and continues from the other side of the TFBSs for extending. Otherwise, SIOMICS_Extension calculates the binomial tail probability as the sum of $\text{binomial}(n, k, p)$, where n is the number of TFBSs the motif under consideration has, k is from $K+1$ to n , and p is the background frequency of the most frequent type of nucleotides in the input sequences. In the case of considering two adjacent positions, p is the background frequency of the most frequent type of dinucleotides in the input sequences. If this tail probability is smaller than $1\text{E-}10$, SIOMICS_Extension extends the motif by including the nucleotides at this position into the TFBSs. SIOMICS_Extension extends the motif according to the above principle either one or two positions at a time, from the left side to the right side of the currently predicted TFBSs. When no extension can be made or the motif length is already 14, SIOMICS_Extension output the final form of the motif with the corresponding TFBSs.

After extending the motifs, some motifs may be left unchanged. For these motifs, SIOMICS_Extension considers reducing their motif lengths. SIOMICS_Extension compares these unchanged motifs to see whether two unchanged motifs are similar. Two unchanged motifs are similar if they match each other with at most one mismatch, or they share a $w-1$ base pair long substring. If two unchanged motifs are similar, SIOMICS_Extension checks whether these similar motifs form motif modules with a common group of other motifs. If both criteria are satisfied, SIOMICS_Extension merges these similar motifs, and represents them with one unique motif. In the case of sharing a $w-1$ base pair long substring by two unchanged motifs, the resulted merged motif is described by nucleotides inside the matched suffices and prefixes, and the motif length is decreased by 1. In other cases, the motif length is unchanged, while the TFBSs of similar motifs are pooled together to form a new motif.

After extending and reducing motif lengths, SIOMICS_Extension outputs the final motifs, motif modules, and TFBSs (Figure 2). The final motifs are those obtained from the above extension and reduction procedure. The final motif modules are produced by running the SIOMICS_Extension with the final motifs without any iteration. The TFBSs of the final motif modules are output as the final TFBSs. The final predictions are visualized using the STAMP and Cystoscape tools (23, 24).

3. Results

3.1 The performance of SIOMICS_Extension on ChIP-seq datasets

We previously applied the original SIOMICS to 13 real ChIP-seq datasets and 13 random datasets (14). We showed that 76.8% to 100% of the predicted motifs by SIOMICS in the 13 real ChIP-seq datasets were similar to known motifs in public databases (22, 25). The similarity of the predicted motifs to known motifs was measured by using STAMP (23) with the similarity E-value cutoff of $1E-5$. Moreover, many individually underrepresented motifs in the 13 real ChIP-seq datasets were discovered by SIOMICS. We also observed that no motif was predicted in any of the 13 random datasets. Compared with two recent methods for motif discovery on ChIP-seq data (13, 20), we showed that SIOMICS predicted motifs of more cofactors with a comparable or better speed on real datasets.

Here we applied the SIOMICS_Extension to the same 13 real ChIP-seq datasets with the default parameters. Since changing motif lengths may make the predicted motifs more similar to true motifs, we paid attention to how many more known cofactor motifs were predicted and the time cost it took to finish the prediction on each dataset. The known cofactors were extracted from literature as described in the SIOMICS paper (14), and their motifs were obtained from the JASPAR database (25). When a predicted motif was similar to a known cofactor motif with the STAMP similarity E-value smaller than $1E-5$, we claimed that this known cofactor motif was predicted. In Table 2, we listed the names of the known cofactors predicted by the original SIOMICS method and by the extended SIOMICS_Extension method. The SIOMICS_Extension predicted more cofactor motifs in 5 out of 13 datasets and predicted the correct motifs for 10 more cofactors in these datasets. For instance, in the Esrrb dataset, the original SIOMICS predicted motifs of 3 known cofactors of Esrrb together with the Esrrb motif, while the SIOMICS_Extension predicted motifs of 4 more cofactors in addition to the four motifs predicted by the original SIOMICS (Table 2).

We also studied the time cost of the SIOMICS_Extension. All calculations were done on the same computer with the following configuration: Intel Core™ 2 Duo CPU E7500 @ 2.93GHz and 4G RAM. The running time of the SIOMICS_Extension was about 31.4% more than that of the original version on average, with the median increment of 10.5%. For the largest dataset, the Esrrb dataset, the running time was 51412 CPU seconds. This dataset contained 49127 sequences, each of which is 800 base pairs or longer. For a smaller dataset that contained 5347 sequences, the Stat3 dataset, the running time was 8775 CPU seconds (about 2.44 CPU hours). Note that the time cost of the SIOMICS_Extension was still lower than that of Dreme (13) (Figure 3), a popular method for motif discovery on ChIP-seq data, on the corresponding datasets. We also noticed that the increment of the running time became smaller when the size of the datasets increased. For instance, for the aforementioned largest dataset, the time cost only increased by 4.7%. On the other hand, for the aforementioned smaller dataset, the time cost increased by 64.1%.

3.2 The prerequisites to run SIOMICS

We implemented the SIOMICS_Extension method in the latest SIOMICS package, which is a multiplatform tool that comes with a graphical user interface (GUI) and a command line interface along with the extensions to make comparative analyses. In the following, without specification, we used the word “SIOMICS” to mean the SIOMICS tool that implements the SIOMICS_Extension method. The following requirements must be met in order to successfully setup and run the SIOMICS tool:

- (1) Users need to install the Python 2 or Python 3. If the Linux version is desired, in order to enable GUI, users need to install Tkinter (Python 2) or tkinter (Python 3).
- (2) Users need to install Java Runtime Environment.
- (3) Users may also need to configure the Environment Variable for running java and python.

The detailed instructions of how to set up the required environment have been provided in the manual of the SIOMIC tool.

3.3 The input to SIOMICS

The input to the SIOMICS tool is the name of a sequence file together with 7 other parameters. The sequence file must be in the FASTA format, with the sequence length around 800 base pairs. Although SIOMICS can deal with longer sequences, it is recommended that the majority sequences should be shorter than 1500 base pairs.

Table 3 summarizes the major eight parameters in SIOMICS along with their functions. Among these parameters, $-w$ controls the length of motifs before extension and reduction. We recommend use the default $w=8$, based on our experience on motif discovery (8, 14). The parameter $-m$ controls the maximal number of motifs predicted. It also specifies the number of top w -mer patterns considered for motif module discovery in each iteration. Users can start to run SIOMICS with $m=100$. If close to 100 motifs are discovered (for example 95), users can increase m by 50 and try to run SIOMICS again. We rarely see more than 200 motifs predicted in a ChIP-seq dataset using an antibody against a TF. The $-c$ parameter specifies the multiple-comparison-corrected p-value cutoff. The $-s$ parameter tells at least how frequent the co-occurrence of w -mer patterns needs to be for consideration. The smaller $-s$ parameter is, the longer the running time SIOMICS takes. We normally expect that each unknown cofactor bind at least 1% of the ChIP-seq peak regions and thus set the default $-s$ parameter correspondingly. The $-r$ parameter does not have a big impact on the results. With larger $-r$ parameter, users may obtain more predictions with higher time cost. The default $-r$ parameter is also recommended be used.

3.4 The output of SIOMICS

SIOMICS outputs seven files, each of which is briefly described in Table 4. The predicted motifs are included in the file named X.motifs, which provides the motif position weight matrix together with other information that can be used to scan sequences to define putative TFBSs. If users want to know which known motifs are similar to these predicted motifs, they can examine two other files named X.JASPAR.pdf and X.TRANSFAC.pdf. These two files include the

comparison results of the predicted motifs with the known motifs in the JASPAR (25) and TRANSFAC databases (22). The file X.mc contains all predicted motif modules. For each motif module, this file provides the motifs included in this module, the number of the input sequences containing the co-occurring TFBSs of all motifs in this module, and the statistical significance of this module represented as the multiple-comparison-corrected p-value. If users want to explore the global picture of the motif (TF) interaction under the current experimental condition, they can utilize Cytoscape (24) to open the file named X.mc.sif. This file presents the predicted motif modules in a network, in which motifs (TFs) are denoted as nodes and two motifs are connected if they co-occur in at least one predicted motif module. The seventh file is the running log file, which stores the running time and the command used for a dataset under consideration.

3.5 SIOMICS Workflow

Here we provide a brief walk-through for a normal use-case of the SIOMICS application. More details can be found on the software manual page.

1. The data to be processed needs to be in the FASTA format, with each sequence in the following format:
 > Oct4 <Sequence identifier>
 ACGTACT..... <sequence nucleotide information>
 See the software manual page for detailed examples. This FASTA sequence file is recommended be repeat masked at <http://www.repeatmasker.org/> or locally.
2. With the sequences, we can open the SIOMICS with the default parameters to start the motif and TFBS prediction. See the interface opened on the software manual page or at http://www.eecs.ucf.edu/~xiaoman/SIOMICS/man_Extension.html.
3. The GUI window near the bottom will display a message, “Running...” to indicate that the software has begun motif discovery. It will display “Done” after creating all output files.
4. If Cytoscape is installed, the end of processing in SIOMICS will take users directly to the interaction network mapping. Otherwise, users can open the X.mc.sif file using the Cytoscape software (24).

4. Discussion

We developed a useful tool SIOMICS for systematic discovery of motifs in ChIP-seq datasets. Tested on 13 real ChIP-seq datasets and 13 random datasets, we showed that the original SIOMICS tool predicted motifs of known cofactors in each real dataset while no motif in any random dataset (14). We also compared the original SIOMICS tool with two motif discovery methods developed for ChIP-seq data analysis (13, 20). The original SIOMICS tool showed superior or at least comparable performance in terms of the number of motifs of known cofactors predicted in real datasets, false motifs predicted in the random datasets, and the time cost of the predictions (14). With GUI, SIOMICS becomes a useful tool for the systematic study of gene transcription regulation based on ChIP-seq experiments.

In the extension version of the SIOMICS tool, we modified the motif length of a motif based on its predicted TFBSs from all motif modules containing this motif. Alternatively, we can implement similar procedures using all TFBSs of a motif from individual motif modules containing this motif. In this way, we might identify different forms of the same motif when this motif co-occurs with different cofactor motifs. However, given the large number of the predicted motif modules, this process may take a long time. Moreover, there may not be many motifs with different forms when they cooperate with different cofactors. In fact, tested on the 13 real datasets, we rarely found different forms produced from the same w -mer motifs. We thus consider all TFBSs of a motif from all motif modules containing this motif.

Although the SIOMICS tool set (Original SIOMICS and extended SIOMICS) is developed for ChIP-seq data, it also works well on small datasets not from ChIP-seq experiments. We tested SIOMICS on small datasets with around a dozen to several dozen 1000 base-pair long sequences (26-29). SIOMICS identified motifs of several TFs shared by these independent breast cancer related datasets. Note that when the number of sequences is small, we recommend that a relative larger $-s$ parameter should be used to control the false positive rate of the predictions.

In summary, we developed a useful tool for gene transcriptional regulation studies. The developed SIOMICS tool together with its manual, test datasets, and others is freely available at <http://www.cs.ucf.edu/~xiaoman/SIOMICS/SIOMICS.html>. We will continue to keep the updated versions on this site to help researchers fully utilize the potential of the ChIP-seq data.

Acknowledgements

We thank the reviewers for their comments. This work is supported by the National Science Foundation grants 1356524, 1149955, 1125676, and 1218275.

References

1. Birney, E., Stamatoyannopoulos, J. A., Dutta, A., et al. (2007) *Nature* **447**, 799-816.
2. Blanchette, M., Bataille, A. R., Chen, X., et al. (2006) *Genome Res* **16**, 656-68.
3. Stormo, G. D. (2000) *Bioinformatics* **16**, 16-23.
4. Arnone, M. I., and Davidson, E. H. (1997) *Development* **124**, 1851-64.
5. Cai, X., Hou, L., Su, N., et al. (2010) *BMC Genomics* **11**, 567.
6. Ding, J., Cai, X., Wang, Y., et al. (2013) *Pac Symp Biocomput*, 320-31.
7. Ding, J., Hu, H., and Li, X. (2012) *Plant Physiol* **158**, 145-55.
8. Ding, J., Li, X., and Hu, H. (2012) *Plant Physiol* **160**, 613-23.
9. Hu, J., Hu, H., and Li, X. (2008) *Nucleic Acids Res* **36**, 4488-97.
10. Johnson, D. S., Mortazavi, A., Myers, R. M., et al. (2007) *Science* **316**, 1497-502.
11. Robertson, G., Hirst, M., Bainbridge, M., et al. (2007) *Nat Methods* **4**, 651-7.
12. Zhang, Y., Liu, T., Meyer, C. A., et al. (2008) *Genome Biol* **9**, R137.
13. Bailey, T. L. (2011) *Bioinformatics* **27**, 1653-9.
14. Ding, J., Hu, H., and Li, X. (2013) *Nucleic Acids Res* **42**, e35.
15. Wang, J., Zhuang, J., Iyer, S., et al. (2012) *Genome Res* **22**, 1798-812.
16. Hu, M., Yu, J., Taylor, J. M., et al. (2010) *Nucleic Acids Res* **38**, 2154-67.
17. Jin, V. X., Apostolos, J., Nagisetty, N. S., et al. (2009) *Bioinformatics* **25**, 3191-3.
18. Kulakovskiy, I. V., Boeva, V. A., Favorov, A. V., et al. (2010) *Bioinformatics* **26**, 2622-3.
19. Sharov, A. A., and Ko, M. S. (2009) *DNA Res* **16**, 261-73.
20. Thomas-Chollier, M., Herrmann, C., Defrance, M., et al. (2012) *Nucleic Acids Res* **40**, e31.
21. Liu, X. S., Brutlag, D. L., and Liu, J. S. (2002) *Nat Biotechnol* **20**, 835-9.
22. Wingender, E., Dietze, P., Karas, H., et al. (1996) *Nucleic Acids Res* **24**, 238-41.
23. Mahony, S., and Benos, P. V. (2007) *Nucleic Acids Res* **35**, W253-8.
24. Shannon, P., Markiel, A., Ozier, O., et al. (2003) *Genome Res* **13**, 2498-504.
25. Sandelin, A., Alkema, W., Engstrom, P., et al. (2004) *Nucleic Acids Res* **32**, D91-4.
26. Chang, H. Y., Sneddon, J. B., Alizadeh, A. A., et al. (2004) *PLoS Biol* **2**, E7.
27. Wang, Y., Klijn, J. G., Zhang, Y., et al. (2005) *Lancet* **365**, 671-9.
28. Weigelt, B., Hu, Z., He, X., et al. (2005) *Cancer Res* **65**, 9155-8.
29. Li, L., Cheng, A. S., Jin, V. X., et al. (2006) *Bioinformatics* **22**, 2210-6.

Figure legends and Tables

Figure 1. An illustration of CRMs, TFBSs, motifs, and motif modules. A. CRMs are short sequences containing multiple TFBSs. Each line is an input sequence. Each small box on lines represents a TFBS. TFBSs of the same TF are represented with the same shape while with different colors in different sequences; B. TFBSs of the same TF are similar to each other; C. The motif corresponding to TFBSs in B and one motif module containing this motif.

Figure 2. The major steps in SIOMICS_Extension.

Figure 3. Time efficiency comparison between SIOMICS_Extension, SIOMICS and Dreme.

Table 1. Commonly used motif prediction tools for ChIP-seq data analysis.

Program	Notes	URL
DREME	Discriminative motif finder	http://meme.nbcr.net/meme/doc/dreme.html
HMS	incorporates sequencing depth information	http://www.sph.umich.edu/csg/qin/HMS/
ChIPModule	Known Motif weight-matrices as input	http://www.cs.ucf.edu/~xiaoman/ChIPModule/ChIPModule.html
SIOMICS	De-novo motif discovery	http://www.eecs.ucf.edu/~xiaoman/SIOMICS/SIOMICS.html
CisFinder	Position Frequency Matrix	http://lgsun.grc.nia.nih.gov/CisFinder/
RSAT peak-motifs	Combines four methods	http://rsat.ulb.ac.be/peak-motifs_form.cgi
W-ChIPeaks	Probe-based peak enhancement and known motifs	http://motif.bmi.ohio-state.edu/W-ChIPeaks/
ChIPMunk	Fast heuristic motif digger	http://autosome.ru/smbasm/librettos/libretto_chipmunk/chipmunk_details.rhtml

Table 2. Comparison of predicted cofactors by original and SIOMICS_Extension.

Data set	SIOMICS	SIOMIC Extension
Sox2	8/9(Sox2,Klf4,Stat3,Zic3,Hoxa5,Tcf3,Tead1,Oct4)	8/9(Sox2,Klf4,Stat3,Tead1,Oct4,Zic3,Tcf3,Hoxa5)
E2f1	7/10(E2f1,Stat3,Klf4,Fox,Sp1,Nfkb1,Tbp)	8/10(E2f1,Stat3,Klf4,Fox,Sp1,Nfkb1,Creb1,Tbp)
Stat3	6/8(Stat3,Klf4,Sox2,Myc,Sp1,Irf)	6/8(Stat3,Klf4,Sox2,Myc,Sp1,Irf)
Nanog	7/8(Nanog,Sox2,Oct4,Zic3,Klf4,Elf5,Tead1)	8/8(Nanog,Sox2,Oct4,Zic3,Klf4,Esrrb,Elf5,Tead1)
Oct4	8/10(Oct4,Sox2,Klf4,Sox10,Ewsr1,Nanog,Zic,Esrrb)	8/10(Oct4,Sox2,Klf4,Sox10,Ewsr1,Nanog,Esrrb,Tead1)
c-Myc	3/4(Stat3,Egr1,Sp1)	3/4(Stat3,Egr1,Sp1)
Klf4	4/10(Klf4,Stat3,Sox2,Sp1)	5/10(Klf4,Stat3,Sp1,Myc,Sox2)
Ctcf	5/6(Ctcf,Stat3,Gabpa,Yy1,Smad3)	5/6(Ctcf,Stat3,Gabpa,Smad3,Myc)
Zfx	2/4(Zfx,Stat3)	2/4(Zfx,Stat3)
Tcfcp2l1	7/12(Tcfcp2l1,Stat3,Klf4,Sox2,Esrrb,Fox,Sp1)	10/12(Tcfcp2l1,Stat3,Klf4,Sox2,Fox,Sp1,Oct4,Creb,Myc,Tead1)
Esrrb	4/10(Esrrb,Klf4,Rxra,Sp1)	8/10(Esrrb,Klf4,Rxra,Sp1,Ewsr1,Creb,Sox2,Stat3)
n-Myc	2/5(Stat3,Creb)	2/5(Stat3,Creb)
Smad1	5/9(Sox2,Oct4,Esrrb,Klf4,Stat3)	5/9(Sox2,Oct4,Klf4,Stat3,Esrrb)

The fraction in each entry shows the ratio of the number of predicted known cofactor motifs to the number of known cofactor motifs in the corresponding dataset. The predicted cofactor names were listed in the parenthesis.

Table 3. Parameters in SIOMICS

Command line argument	Function
-i	Input sequence file
-o	The output folder for prediction results
-w	Initial motif length to search
-m	Maximum number of motifs to output
-s	Minimum number of sequences needed to contain co-occurring TFBSs of a motif module
-c	Multiple comparison corrected P-value cutoff for motif module prediction
-r	Maximum number of iterations with no new motif discovered allowed

Table 4. Output files from SIOMICS

Output files	Description
X.motifs	Predicted motifs from SIOMICS
X.mc	Predicted motif modules
X.JASPAR.pdf	predicted motifs compared with JASPAR motifs
X.TRANSFAC.pdf	predicted motifs compared with TRANSFAC motifs
X.mc.sif	Interaction network to be viewed in Cytoscape
X.tfbs	TFBSs of the predicted motifs
running.log	Runtime and command log of SIOMICS

Figure 1

A. Input sequences and
CRMs

B. TFBSs of one TF
that binds to CRMs in A

A motif module
C. The motif of the TF
in B and a motif module

Figure 2

Figure 3