

MotifOrganizer: a scalable model-based motif clustering tool for mammalian genomes

Zhaohui S. Qin^{1,2,3}, Misha Bilenky⁴, Gang Su⁵, Steven J. M. Jones⁴

¹Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta GA 30322, USA, ²Department of Biomedical Informatics, Emory University School of Medicine, Atlanta GA 30322, USA. ³Center for Comprehensive Informatics, Emory University, Atlanta GA 30322, USA. ⁴British Columbia Cancer Agency Genome Sciences Centre, Vancouver, BC, V5Z 4E6, Canada. ⁵Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor MI 48109, USA.

TABLE OF CONTENTS

1. Abstract
2. Introduction
3. Material and Methods
 - 3.1. Input Data
 - 3.2. BMCES
 - 3.3. MotifOrganizer
 - 3.4. Distance-based clustering approaches
 - 3.5. Quality and uncertainty measures
 - 3.6. Clustering accuracy
 - 3.7. Motif matching tools
4. Results
 - 4.1. JASPAR
 - 4.2. TRANSFAC
 - 4.3. cisRED
5. Discussion
6. Acknowledgement
7. References

1. ABSTRACT

Assembling a comprehensive catalog of all transcription factors (TFs) and the genes that they regulate (regulon) is important for understanding gene regulation. The sequence-specific conserved binding profiles of TFs can be characterized from whole genome sequences with phylogenetic approaches, and a large number of such profiles have been released. Effective mining of these data sources could reveal novel functional elements computationally. Due to the variability of the binding sites, it is necessary to generalize profiles pertinent to the same TF by clustering. The summarized familial profile is effective in identifying unknown binding sites, thus lead to gene co-regulation prediction. Here we report MotifOrganizer, a scalable model-based clustering algorithm designed for grouping motifs identified from large scale comparative genomics studies on mammalian species. The new algorithm allows grouping of motifs with variable widths and a novel two-stage operation scheme further increases the scalability. MotifOrganizer demonstrated favorable performance comparing to distance-based and single-stage model-based clustering tools on simulated data. Tests on approximately 150k motifs from the cisRED human database demonstrated that MotifOrganizer can effectively cluster whole genome sets of mammalian motifs.

2. INTRODUCTION

Assembling a comprehensive catalog of all transcription factors (TFs) and the genes that they regulate-regulon is important in understanding how gene expression is regulated. An important step towards achieving this goal involves identifying the regulatory elements that are bound by TFs. These regulatory elements, also referred to as motifs, are short DNA segments typically 6 to 30 bp in length. Success in finding these regulatory elements would contribute to our understanding of TF-regulon relationships by providing insight into the mechanism of transcriptional regulation. The rapid accumulation of completely sequenced genomes, the combination of *ab initio* motif searching algorithms, and comparative genomics strategies constitutes a powerful approach for identifying novel regulatory elements (1-5). Several recent studies carried out genome-wide motif discovery using sequence specificity and phylogenetic conservation information to identify motifs that are conserved at a higher-than-expected rate in non-coding regions (6-10). Many online databases such as TRANSFAC (11) and JASPAR (12), compiled large volumes of annotated motifs.

In one such study, Robertson *et al.* described the construction of cisRED, a database that contains

predictions from whole-genome discovery of cis-regulatory elements in mammals and other eukaryotes (9). Applying a series of sophisticated bioinformatics analyses, including multiple motif discovery methods, Robertson *et al.* cataloged more than 200,000 motifs in their database (<http://www.cisred.org>). Many of the motifs in the cisRED database are believed to be regulatory elements that play vital functional roles because they occur in multiple promoter regions of orthologous genes across several vertebrate species. As the majority of enriched modules in cisRED remain unclassified, such a large collection constitutes a great resource for mining novel biological knowledge.

In particular, since regulatory proteins bind to DNA in a sequence specific manner (13, 14), similarity in motif sequences discovered in promoter regions of different genes is indicative of co-regulation (5, 15, 16). A natural follow-up of large-scale motif discovery efforts like cisRED is to cluster similar motifs in the database into groups. This is a critical step in translating the conserved motifs identified by comparative genomics methods into a putative model of regulatory elements, which not only represent functional aggregation but can also be further utilized for unknown function site prediction (17, 18). As Robertson *et al.* pointed out, the large size of mammalian genomes makes it challenging to conduct such a cisRED-type analysis (9). This is especially true when attempting to cluster the hundreds of thousands of motifs from databases like cisRED. Most of the available clustering approaches used for motif clustering are pair-wise distance-based (19), which are conceptually simple and readily available for data sets of moderate size. However, performing distance-based clustering on a large number of motifs requires a huge pair-wise distance matrix which is extremely costly to compute and maintain. Such computational jobs can only be performed on servers with a large amount of memory.

The goal of this project is to develop an efficient clustering strategy that can analyze a large motif dataset (such as the motif collections from the cisRED database) on a typical laptop computer. To do that, the clustering strategy must be highly scalable. Given the limitation of the distance-based clustering methods, we chose the alternative model-based clustering strategy. Model-based motif clustering methods (20, 21) assume that motifs in a cluster share the same product multinomial distribution in nucleotide composition. Clustering is achieved by calculating the probability of each motif belonging to each of the existing clusters. Therefore, no explicit pair-wise distance calculation between all pairs of motifs is needed. Consequently, model-based methods are more scalable than distance-based approaches in practice. More importantly, distance-based approaches treat the input motifs as bona fide patterns, thus ignoring uncertainty in the input. In contrast, model-based approaches explicitly model the uncertainty using probability distribution, which is a more sensible way for representing the highly-variable DNA regulatory elements.

Previously we developed Bayesian Motif Cluster (BMC), a model-based clustering approach named and

successfully applied it to investigate co-regulation in bacterial species including *Escherichia coli* (20). We discovered both novel regulatory elements and proposed new hypotheses on regulatory relationships. In this manuscript, we report the development of a new algorithm called Bayesian Motif Cluster for Eukaryotic Species (BMCES), as well as a two-stage, divide-conquer-combine clustering scheme, MotifOrganizer. The new algorithm and scheme are better suited to mammalian applications than BMC because they allow clusters, as well as a motif and the cluster that it joins, to have different widths. The two-stage scheme adopted by MotifOrganizer can increase the scalability of BMC by three orders of magnitude under the same computation environment. As a result, MotifOrganizer is able to cluster motif collections in the hundreds of thousands, as is typically produced from the results of motif identification efforts conducted by whole genome comparative genomics methods on mammalian species. The overall operation scheme of our methods is illustrated in Figure 1.

3. MATERIALS AND METHODS

3.1. Input Data

The basic units in the input data for BMCES or MotifOrganizer are motifs, each of which is a stack of aligned short DNA sequences that had been identified by multiple, probabilistic, de novo comparative genomics motif discovery methods (3, 22). For example, cisRED motifs were identified in ~2kb promoter regions of sets of orthologous genes in multiple vertebrates species (9) (<http://www.cisred.org>). Each sequence in a discovered motif is assumed to be a phylogenetic counterpart of the other sequences. As a special case, we allowed a motif to contain as few as one sequence. These motifs are represented by position-specific weight matrices (PWMs) in our model-based methods, which are used to calculate Bayes ratios for iterative cluster assignments. See Figure 2 for an illustration of motifs and motif clusters.

3.2. BMCES

BMC (20) assumes that motifs that belong to a cluster follow the same product multinomial distribution (23), and implements a Gibbs sampler procedure (24, 25) to iteratively infer cluster membership. By using only the middle part of each motif, the original algorithm allows the width of input motifs to be different, but requires that all clusters have the same width. Because the widths of mammalian transcription factor binding sites (TFBSs) vary substantially (from 6 bp to >30 bp in JASPAR and TRANSFAC), such a constraint is too restrictive for clustering mammalian motifs. Recently, Jensen and Liu proposed an extended Bayesian model that treats cluster widths as random variables (26). We adopted an alternative strategy in BMCES by allowing motifs and clusters with different widths to be grouped together. Our strategy allows flexible alignment between motifs and clusters. Specifically, during the process that reassigns motifs to clusters, if a cluster's width is larger than a motif's width, we use a sliding window whose width is equal to the motif width to determine which part of the cluster pattern best fits the motif, and use this best match to calculate the

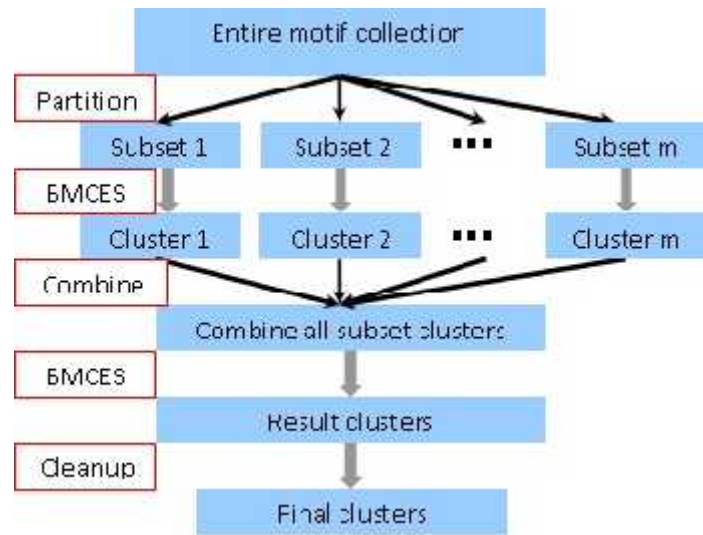


Figure 1. Illustration of phylogenetic footprinting technique and the motif clustering strategy, as well as examples of motifs and motif clusters. B.

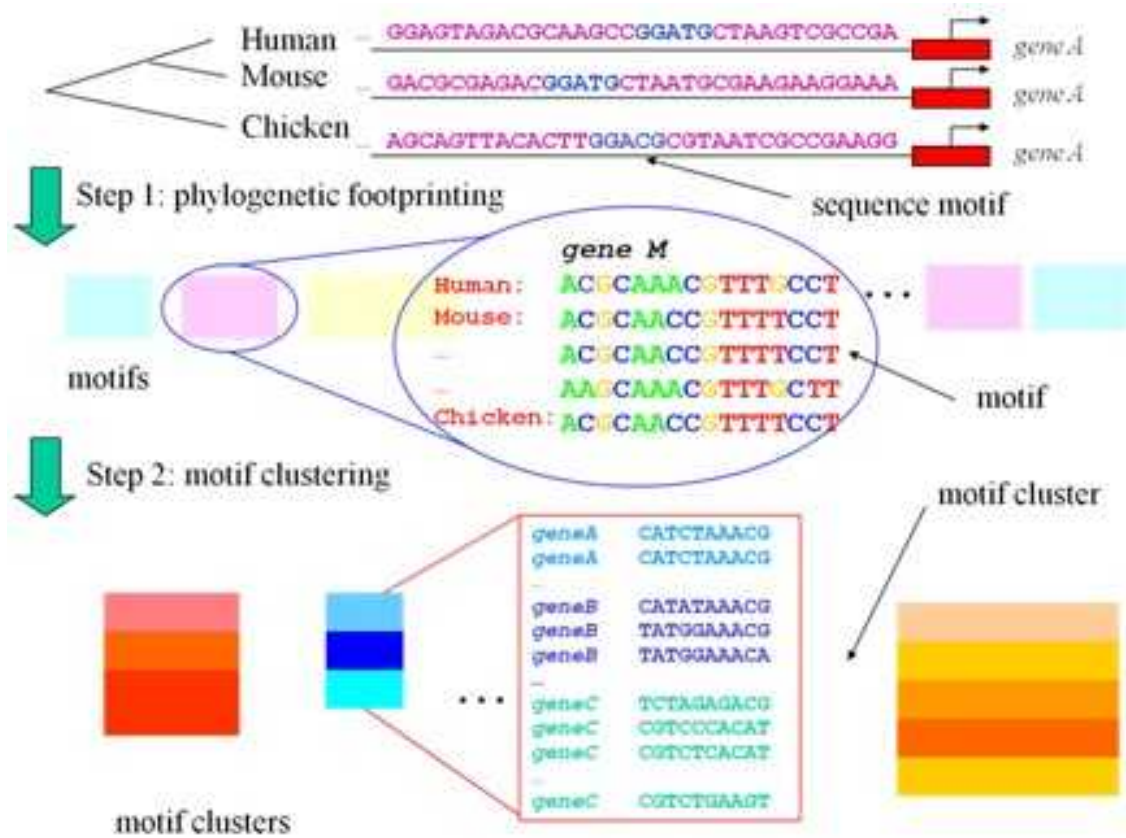


Figure 2. A diagram of the two-stage divide-conquer-combine scheme MotifOrganizer proposed to enable model-based clustering to be performed on a large motif collection.

likelihood that the motif will join the cluster. Conversely, when the motif is wider than the cluster, we use a sliding window with a width equal to the cluster width to

determine which part of the motif best fits the cluster, and use only this subset of the motif to calculate the fit likelihood for this cluster. After a cluster has been

identified as accommodating a motif, if the cluster width is smaller than the motif width, then only the aligned part of the motif is added to the cluster, and the rest of the motif is removed from consideration. Conversely, if the cluster width is greater than the motif width, then the aligned cluster pattern will be trimmed to the motif width, and only the part that aligned to the added motif is kept. Such a strategy allows us to generate clusters of different widths, and is able to group motifs of different widths into a common cluster.

3.3. MotifOrganizer

Because genome-wide collections of mammalian motifs contain hundreds of thousands of motifs which is too voluminous for BMC to cluster, we sought to increase BMC's scalability while retaining its favorable performance. To accomplish this, we devised a novel, two-stage, divide-conquer-combine scheme (Figure 1). In the first stage, we first partition the motif set into subsets, either randomly or by assigning motifs of different widths to different subsets. After partitioning, BMCES algorithm is applied independently to each subset. We then combine all output clusters from this stage to form a new input motif set, in which each input motif becomes a motif cluster from the first-stage output. Since both motifs and motif clusters are represented by PWMs, the BMCES algorithm could again be applied to group these first-stage clusters into final clusters. Because the number of clusters output by the first stage clustering is typically much smaller than the total number of original motifs ($\log n$), the overall demand on computer resources is substantially reduced, making the two-stage strategy highly scalable. Since clustering runs in the first stage can be carried out independently, it is straightforward to implement MotifOrganizer to take advantage of the increasingly available parallel computing environment, which will further reduce the computation time.

From our experience, MotifOrganizer converges rather rapidly. Typically, 100,000 reassignments (100 cycles for a total of 1,000 motifs or 50 cycles for a total of 2,000 motifs) produce satisfactory results. As few as 10,000 reassignments sometimes produce stable results. To avoid being trapped in a local mode, which is a common problem for complex sample spaces, one can choose to run multiple independent chains in MotifOrganizer (simply by specifying an input parameter), each with different initial setting, to better explore the entire sample space. We chose the result that achieves the highest posterior likelihood among all the chains as the final reported clustering result.

3.4. Distance-based clustering approaches

For comparison purposes, we also ran distance-based clustering methods on the JASPAR and TRANSFAC data. There are many existing distance-based clustering methods, such as the most popular hierarchical clustering and the K-means clustering. CLARA, or Clustering Large Applications (27) is designed to handle larger datasets than the classical PAM (Partitioning Around Medoids) method. CLARA is a two-step procedure, in the first step, a sample is drawn from the original dataset and then PAM is applied on the sample; in the second step, each of the un-sampled

data point is assigned to its nearest medoid. The merit of the clustering is measured by the average distance between each object and its medoid. The CLARA is often repeated multiple times and the best clustering result is reported as the final outcome.

3.5. Quality and uncertainty measures

Model-based clustering methods allow formal statistical inference to be performed. BMCES and MotifOrganizer take advantage of this fact to produce a set of significance measures. One measure is the ratio of the probability that all motifs follow the same product multinomial distribution versus the probability that each motif follow its own. This serves as a measure of the cluster's compactness. A higher ratio indicates that individual motif sequences in a cluster more closely resemble each other. In BMCES and MotifOrganizer, since the widths of motifs may differ between clusters, we define normalized Bayes ratio (NBR) for a cluster by dividing B by the width of the cluster. BMC also calculates and reports the posterior assignment probability (PAP) for each motif in a cluster. This is the probability that a motif belongs to its current cluster, conditional on the assignments of all other motifs. A higher PAP indicates a better motif-to-cluster fit. These quality measurements can be used to prioritize clustering results. By sorting clusters from the most significant to the least significant according to NBR, one can focus on the top ones for experimental validation. One can also remove loosely fit motifs from its cluster if its PAP is lower than a pre-specified threshold. NBR and PAP also provide measures of uncertainties of clusters and clustering assignment of motifs.

3.6. Clustering accuracy

To evaluate and compare the performance of different clustering approaches, we adopted the Adjusted Rand Index (ARI) (28, 29), which measures the degree of agreement between two different partitions of multiple objects. It is able to provide an objective assessment of the closeness between our clustering result and the true motif family membership. The values of ARI lie between 0 and 1, with a higher value indicating better agreement. Among many statistics proposed, Milligan and Cooper recommended ARI as the measure of agreement based on extensive empirical studies (29). The detailed formula on how to calculate ARI can be found in the Supplementary material of Qin *et al.* (30). Since model-based clustering is not a deterministic procedure, results from different runs might be slightly different. We therefore measured the performance of all model-based clustering algorithms by repeating the clustering procedure 100 times and taking the average.

3.7. Motif matching tools

To annotate the clustering results, we compared each cluster to motif models in JASPAR 4 (12) and TRANSFAC 9.3 (11) databases. We used MatCompare (31) and MACO (32) to assess the similarity of motif pairs. For MatCompare, we used the default distance measure, which is the minimum Kulback-Leibler (KL) divergence between matched fragments in two motifs; motifs with divergence per column less than 1.0 are regarded as very

similar. For MACO, we determined the threshold value for the matching correlation coefficient scores based on the empirical distribution obtained from 3000 random matrices.

4. RESULTS

In order to evaluate the performance of MotifOrganizer comprehensively, we tested it on three different datasets: we first compared the performance of our new BMCES algorithm with BMC in a dataset containing 1128 motifs from 59 TF binding site (TFBS) models selected from the JASPAR 4 database (12). We then evaluated the advantages of using the two-stage clustering method motifOrganizer over the BMCES algorithm using 5452 motifs from 319 TFBS models selected from the TRANSFAC 9.3 database (11). Finally, we applied motifOrganizer to ~30,000 motifs selected from the cisRED human v.2 database (9).

In the first two motif sets, since the group membership (motif model) of each motif is known, we were able to directly assess how accurate the model-based approach could recapitulate the partitioning in the input model sets, and to characterize the performance of different clustering algorithms. With the third motif set, we compared the clustering result to known motif models from JASPAR and TRANSFAC databases. Since model-based clustering is not a deterministic procedure, results from different runs might be slightly different. Given this, we measured the performance of all model-based clustering algorithms by repeating the clustering procedure 100 times and taking the average. Clustering performance is measured by ARI.

4.1. JASPAR

We compiled a set of 1128 motifs that represented 59 mammalian TFBS models for which individual sequences were available from the JASPAR 4 database (12). Motif widths ranged from 5 to 22 bp, and the number of sequences in each model ranged from 3 to 48, averaging ~19. Clustering was performed using BMCES, the maximum allowable width difference between a motif and a cluster is set to be 2 bp (we used the same setting throughout this study). For comparison, we also tested an advanced distance-based clustering method CLARA (Kaufman and Rousseeuw, 1990). We used the edit distance similarity metric and determined the number of clusters by doing runs over a range of target cluster numbers and selecting the result set that had the maximum average silhouette widths as recommended by Kaufman and Rousseeuw (27). We found that BMCES achieved a higher ARI than CLARA (0.55 versus 0.44) and returned a number of clusters that was closer (54 vs. 34) to the number of input TFBS models--59. BMCES also required less than a third of the memory, and was at least three times faster, depending on how many CLARA runs were used to identify the optimal solution.

Next we took advantage of the quality measures produced by model-based clustering and performed a "cleanup" on the clustering result. To be specific, we used the normalized Bayes ratio (NBR), which reflects a

cluster's tightness; and the posterior assignment probability (PAP), which measures how well a motif fits the cluster to which it has been assigned. Further details about these measures can be found in the Method section. Using filtering thresholds of NBR = 0 and PAP = 0.5, On average, 835 (74%) motifs were retained after cleanup. Using filtering thresholds of NBR = 0 and PAP = 0.5, the average ARI improved from 0.55 to 0.71, while under the same conditions, CLARA showed a more modest ARI increase, from 0.44 to 0.51. These results suggested that NBR and PAP are indeed effective quality indicators and can be used to prioritize clustering results for future experimental validation.

Figure 3 illustrates the clustering performance and relationships between TFBS models and structural classes. Its membership map shows two types of clustering errors: 'combining' errors (indicated by green cell), in which motifs that belong to different TFBS models were clustered together, and 'splitting' errors (indicated by blue cell), in which motifs belonging to same TFBS model were assigned to different clusters. Combining errors were ~2.5 times more frequent than splitting errors. It is also evident from the two plots that the result obtained after cleanup (Figure 3B) indeed shows better clustering quality than before (Figure 3A).

At least two reasons may contribute to the fact that combining errors were ~2.5 times more frequent than splitting errors. First, the model-based algorithm BMC and BMCES are based on is Dirichlet process mixture model which favors small number of clusters and relatively large individual clusters. This may cause more combining errors than splitting errors. Second, the 'gold standard' of JASPAR may be over-specific so that the members of a family can actually be combined together to form a family of higher hierarchy. Nevertheless, the high ARI demonstrates the homogeneity of the clusters and effectiveness of our approach.

Similarities between binding models for structurally related TFs have been summarized by "familial binding profiles" (33-35). Consistent with this, when BMCES's clusters contained motifs from more than one model, the models typically belonged to the same structural class. For example, in the top 15 filtered clusters, which contained more than half of all motifs, three clusters contained motifs from different TF models. One of these clusters was mainly MA0101 (c-REL), MA0105 (p50) and MA01007 (p65) sequences from the REL structural class; another was mainly MA0040 (HFH-1), MA0041 (HFH-2) and MA0047 (HNF-3beta) sequences from the FORKHEAD structural class. At the same time, we expect that sequences for some models may be dispersed across clusters, given that a mixture model (36, 37) or enhanced PWM (38) may better represent variability over binding sequences for a transcription factor than a single PWM.

4.2. TRANSFAC

The second dataset consisted of 5452 motifs from 319 TRANSFAC 9.3 mammalian TFBS models (11). The

Table 1. Comparison of clustering performance between regular model-based clustering approach BMC and the proposed two-stage clustering scheme MotifOrganizer using the 5452 motif TRANSFAC dataset

PAP	NBR 0	NBR 2	NBR 5
0.0	0.363 (0.027)	0.440 (0.032)	0.704 (0.054)
Two-stage	0.563 (0.028)	0.586 (0.018)	0.755 (0.029)
0.5	0.556 (0.043)	0.597 (0.044)	0.786 (0.052)
Two-stage	0.736 (0.026)	0.736 (0.026)	0.793 (0.030)
0.8	0.582 (0.043)	0.626 (0.045)	0.818 (0.052)
Two-stage	0.786 (0.022)	0.786 (0.045)	0.819 (0.023)

Study was performed under various quality measure threshold settings. Cluster performance is measured by Adjust Rand Index (ARI). Both clustering procedures were performed 100 times under each setting. Both average ARI and standard deviation (in parentheses) were reported

Table 2. Comparison of number of clusters generated with the actual number of motif profiles between regular model-based clustering approach BMC and the proposed two-stage clustering scheme MotifOrganizer using the 5452 motif TRANSFAC dataset

PAP	Method		NBR 0	NBR 2	NBR 5
0.0		truth	295 (0)	283 (5)	154 (18)
		inferred	381 (18)	174 (11)	43 (5)
	Two-stage	truth	295 (0)	294 (2)	197 (6)
		inferred	325 (10)	275 (8)	115 (5)
0.5		truth	219 (8)	211 (9)	110 (15)
		inferred	162 (9)	132 (9)	42 (5)
	Two-stage	truth	259 (3)	258 (3)	162 (4)
		inferred	248 (7)	243 (7)	113 (5)
0.8		truth	187 (9)	178 (10)	90 (12)
		inferred	136 (8)	109 (8)	38 (5)
	Two-stage	truth	212 (5)	211 (5)	142 (4)
		inferred	197 (6)	195 (5)	109 (4)

Study was performed under various quality measure thresholds. Both clustering procedures were performed 100 times under each setting. Both average cluster number and standard deviation (in parentheses) were reported.

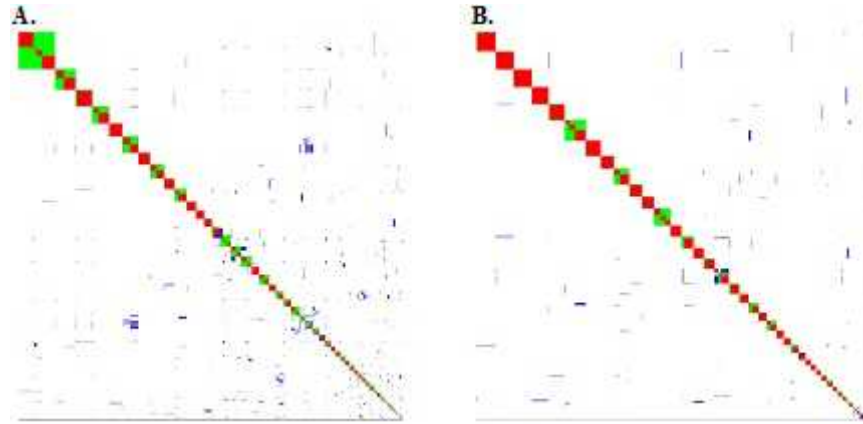








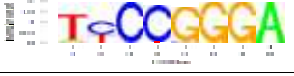



Figure 3. Membership map that summarizes BMCES clustering performance on 1152 motifs from 59 TFBS models from the JASPAR 4 database. Rows represent TFBS models in JASPAR, columns represent clusters generated from BMCES. For each cell (i,j), a) red indicates cases in which motifs i and j were clustered together by BMCES and belonged to the same JASPAR model, b) green indicates cases in which motifs i and j were clustered together by BMCES but belonged to different JASPAR models, c) blue indicates cases in which motifs i and j were not clustered together by BMCES but belonged to the same JASPAR model, and d) white indicates cases in which motifs i and j were neither in the same BMCES cluster nor belonged to the same JASPAR model. A. Membership map for all 1152 motifs in the input dataset. Adjust Rand Index (ARI) is 0.55. B. Membership map for remaining motifs after filtering with thresholds NBR = 0 and PAP = 0.5. ARI is 0.71.

number of motifs in each model ranged from 1 to 169, averaging 18.5. Motif widths ranged from 6 to 30 bps. Our primary goal was to compare performance between single-stage (BMCES) and two-stage (MotifOrganizer) model-based clustering approaches. Using MotifOrganizer, we first partitioned the whole motif set into disjoint motif subsets by width, then applied BMCES in each subset

separately, and concatenated all resulting clusters as input to the second stage clustering. More details about the two-stage scheme can be found in the Method section). Clustering performances measured by average ARI were plotted in Figure 4 and summarized in Tables 1 and 2. Compared to the one-stage approach BMC, one can see that the two-stage approach yielded a higher average ARI

Table 3. Examples of motif clusters that match to known TRANSFAC/JASPAR motif patterns

Cluster ID	Motif cluster logo	NBR	TRANSFAC/JASPAR name	TFBS profile logo	Score 1 ¹	Score 2 ²
		25.7	M00473 FOXO1		0.095	0.998
49		24.7	M00437 CHX10		0.072	0.999
95		20.9	M00179 CRE-BP1		0.542	0.942
157		17.9	MA0063 Nkx2-5		0.148	0.991
159		17.7	MA0137 STAT1		0.580	0.923

¹Score 1 is the Kulback-Leibler divergence used in MatCompare, ²Score 2 is the MACO score, These cisRED motif clusters were reported by both MatCompare and MACO as similar to at least one JASPAR CORE or TRANSFAC TFBS models. The clusters are sorted by descending Normalized Bayes Ratio (NBR)

(green lines are always on top of the corresponding red lines in Figure 4A), estimated the number of clusters more accurately (green lines are always closer to the x-axis than the corresponding red lines in Figure 4B). In addition, Figure 4 showed that cluster quality improved with increasing threshold values of the two quality measures--NBR and PAP. On the other hand, the memory consumption is about half when using the two-stage scheme, and the computing time of the two-stage clustering scheme is less than one third than that of the regular one-stage clustering approach. Running first stage clustering jobs in parallel will further reduce the running time.

4.3. cisRED

We applied MotifOrganizer to a subset of 29,490 conserved DNA sequence motifs from the cisRED human v.2 database. These motifs had been identified using genome-wide comparative genomics approaches that involved combining results from multiple probabilistic de novo discovery methods (Robertson, *et al.*, 2006). We selected the subset of motifs that had p-values < 0.001 and widths between 6 and 20 bp. We applied MotifOrganizer to this dataset. The partition before the first stage is based on the motif width, such that each subset contains motifs of the same width which ranges from 6 to 20. At the end, a total of 8396 clusters were produced from MotifOrganizer. Among which, 4865 clusters consist of 15330 motifs passed the cleanup step with NBR cutoff of 0 and PAP cutoff of 0.5.









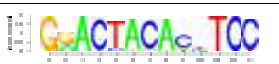

Unlike JASPAR or TRANSFAC, there is no “gold standard” partition of the cisRED motif set. In order to gauge the effectiveness of MotifOrganizer, we compared all 4865 filtered clusters to a large set of high-quality known motif models: 108 models from JASPAR CORE, and 398 models from TRANSFAC 9.3. We used two published comparison tools, MatCompare (31) and MACO (32), to identify “match” between predicted motif clusters and those known motif models. For MatCompare, a

“match” is called if the default distance measure, Kulback-Leibler (KL) distance between the PWMs of two motifs is less than 1. For MACO, we call two motifs “match” each other if their correlation coefficient score is close to 1. In the end, MatCompare identified 558 matches between predicted clusters and known motif models (one cluster may map to multiple motif models); MACO produced 753 matches. 171 matches were identified by both methods. Table 3 contains five such matches. The full list of all matches identified by both MatCompare and MACO, with motif logo plots (39), can be found at our website. Among all 4865 clusters identified by MotifOrganizer, 660 (14%) were similar to one of 506 known motif models using at least one method. Among all the 506 known motif models, 294 (58.1%) matched to at least one predicted cluster.

We then compared our clusters with JASPAR PHYLOFACTS motifs. This database consists of 174 conserved motifs identified in a large scale mammalian comparative genomics study (10).. MatCompare identified 167 matches and MACO identified 166. Seventy matches were identified by both methods. Table 4 contains five such matches. The full list of all matches identified by both MatCompare and MACO, with motif logo plots, can be found at our website. Of the 174 PHYLOFACT motifs, 108 (62.1%) matched to at least one predicted cluster.

We are encouraged that many of the predicted motif clusters identified by MotifOrganizer match to known motif models. And more than half of the known motif models match to at least one of the clusters predicted. We also found that clusters with higher NBR are more likely to match to known motif models in JASPAR CORE, JASPAR PHYLOFACTS or TRANSFAC databases. Of the top 1000 clusters out of 4865 in total, 162 matched to known motif models according to MatCompare, while in clusters 1001 – 2000, 2001 – 3000, 3001 – 4000, 4001 – 4865, only 28, 25, 17 and 7 clusters matched to known motif models. MACO results were similar (data not

Table 4. Examples of motif clusters that match to known JASPAR PHYLOFACTS motif patterns

Cluster ID	Motif cluster logo	NBR	JASPAR PHYLOFACTS name	TFBS profile logo	Score 1 ¹	Score 2 ²
2		36.9	PF0074		0.749	0.774
5		33.4	PF0024		0	1
15		29.8	PF0056		0	1
49		24.7	PF0023		0	1
101		20.5	PF0074		0.631	0.887

¹Score 1 is the Kulback-Leibler divergence used in MatCompare, ²Score 2 is the MACO score. These cisRED motif clusters were reported by both MatCompare and MACO as similar to at least one JASPAR PHYLOFACTS motifs

shown). This suggests that clusters with higher quality scores were more likely to be bona fide functional elements. However, some highly ranked clusters, and overall more than 80% of all identified clusters, matched no motifs in JASPAR CORE, JASPAR PHYLOFACTS or TRANSFAC databases. Table 5 contains motif patterns from 12 such clusters. Most positions in these motifs were highly conserved, and many of the motifs were palindromic, which is typical of homodimer DNA binding proteins. This suggests that these novel motifs may represent binding sites for uncharacterized TFs that mediate expression levels of the genes with which they are associated.

5. DISCUSSION

Large amount of evolutionarily conserved DNA elements has been discovered with fast accumulation of sequenced genomes. Build on the hypothesis that sequence similarity implies functional conservation binding by the same regulatory protein, clustering these motifs will lead to the translation of putative elements into regulatory information. It has been shown that the ability of modeling uncertainties explicitly give model-based clustering approaches advantages over distance-based approaches. However, existing model-based approaches such as BMC are unable to handle large scale motif sets collected from mammalian species on a regular personal computer. In this study, we proposed a novel two-stage model-based algorithm for clustering motifs identified from mammalian species genome-wide. Our new algorithm allows motifs of variable different widths to be clustered together and is capable of handling large scale input motif sets. Comparison studies indicated that our new approach retained and surpassed the clustering accuracy achieved by the single-stage model-based approaches, while reducing computation time and memory requirements to levels that permit clustering genome-wide sets of mammalian motifs on today's commodity computer systems. To further demonstrate MotifOrganizer's scalability, we tested it on 150K motifs from the cisRED human v.2 database. Using

the default parameter setting, the entire two-stage clustering process took about four days to complete on a regular shared cluster computer server, and the peak memory consumption was only about 250MB. The scalability of MotifOrganizer demonstrated in this study is quite promising. As the number of cis regulatory regions may currently be underestimated, we anticipate a persistent need for highly scalable clustering tools to analyze the large motif sets.

In BMCES and motifOrganizer, we employed the same probability model used in BMC, which models motif columns as multinomial with Dirichlet priors. Under this model, nucleotide counts from each species are treated identically as independent observations. However, from the point of view of evolution theory, aligned sequences from closely-related species, like human and chimpanzee, are far from independent. A more desirable model should take into account of the phylogenetic distances among species and weigh the contributions from different species accordingly. Various evolution models and techniques traced back to Felsenstein (40) may be applied. We believe such an improved model will enhance the performance of BMCES and motifOrganizer and will be pursued in our future work.

We were encouraged that many of the predicted motif clusters identified by MotifOrganizer were similar to known motif models, that more than half of known motif models tested matched at least one of the predicted clusters, and that highly-ranked clusters were more likely to be similar to known motifs. On the other hand, most of our filtered clusters appeared to differ from cluster patterns reported by previous large-scale studies, and some clusters with high NBR rank matched to no known motif pattern. These results suggest that regulatory motifs are highly diverse and that a substantial number of new regulatory elements have yet to be discovered and validated.

We seek to create a comprehensive catalog of mammalian cis regulatory motifs that, by facilitating dimension reduction and pattern discovery, and so

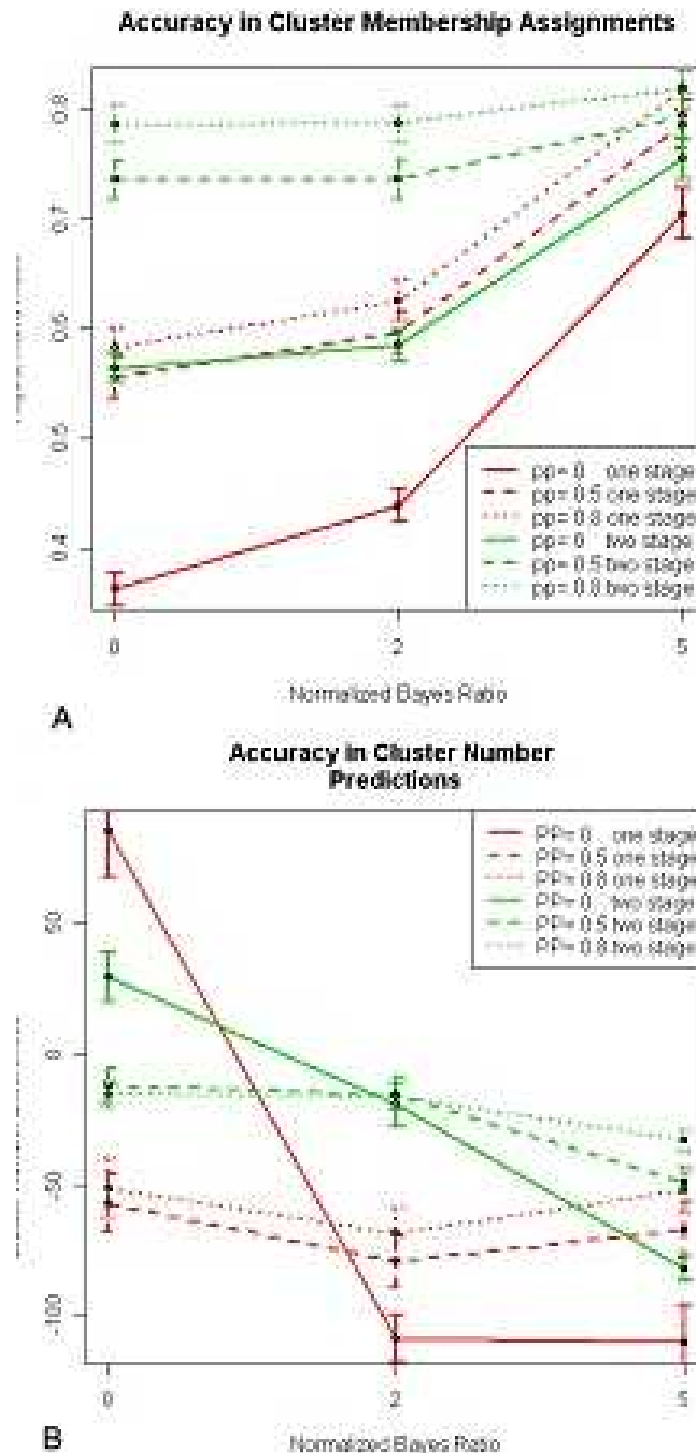
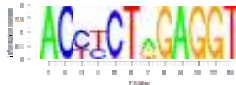








Figure 4. MotifOrganizer clustering results for TRANSFAC 9.3 data. This dataset contains 5452 motifs that belonged to 295 TRANSFAC TFBS models. To evaluate using quality measures to filter clustering results, we tested nine different threshold settings, combining NBR thresholds of 0, 2, and 5 for clusters and PAP thresholds of 0, 0.5 and 0.8 for cluster assignments (solid, dashed and dotted lines respectively). Red lines correspond to one-stage BMCES clustering and green lines to two-stage motifOrganizer clustering. A. Clustering performance measured by ARI; B. Differences between numbers of clusters and number of TRANSFAC models in the remaining motif sets after filtering. See Tables 1 and 2 for numerical values.

Table 5. Selected motif clusters from the result of clustering the 30K motif cisRED Human v.2 dataset using motifOrganizer

Cluster ID	Motif cluster logo	NBR
71		22.9
78		22.4
83		22.1
94		20.9
151		18.1
159		17.7
187		16.3
196		16.0
218		15.3
228		15.1
232		14.9
238		14.7

The motif pattern of these clusters does not match to any JASPAR CORE, PHYLOFACTS or TRANSFAC motif model according to both MatCompare and MACO software. The clusters are sorted by normalized Bayes ratio (NBR)

functional annotation, will contribute to understanding modules and networks in mammalian transcriptional regulation. We anticipate improving MotifOrganizer’s performance by extending it to include parameters that address more aspects of eukaryotic transcriptional regulation. For example, clustering may be more effective when it integrates additional data types like co-factors, DNA and chromatin structure, and histone modifications. A number of such data types, including mammalian TF binding regions, appear to be cost-effectively

characterizable by ChIP-Seq technologies (41-44). As for ChIP-chip (45, 46) and other types of ChIP-sequencing (references 1-10 in (44)), motifs can be identified in bound or enriched regions identified from the target genome (47) using newly developed programs such as HMS (48). However, approaches that seek to combine motifs from regions identified by ChIP-Seq with deep genome-wide comparative genomics methods are likely to continue to require scalable ways of identifying both conserved motifs and groups of similar motifs. We anticipate that

MotifOrganizer and its extensions will serve as an important resource for such work. MotifOrganizer package written in C++ and Perl (source code included) can be freely downloaded from <http://userwww.service.emory.edu/~zqin4/motif/>.

6. ACKNOWLEDGEMENT

We are grateful to the editor and two anonymous reviewers for their insightful comments. We thank Dr. Dustin Schones for assistance with MatCompare program. We thank Christian Mehta for critical reading of the manuscript. And we thank Dr. Gordon Robertson and the cisRED team at BCGSC for their help with the cisRED dataset. This work is partially supported by National Institutes of Health grant R01HG005119 (to ZSQ).

7. REFERENCES

1. R. Gordan, L. Narlikar and A. J. Hartemink: Finding regulatory DNA motifs using alignment-free evolutionary conservation information. *Nucleic Acids Res*, 38(6), e90 (2010)
2. G. Li, B. Liu, Q. Ma and Y. Xu: A new framework for identifying cis-regulatory motifs in prokaryotes. *Nucleic Acids Res*, 39(7), e42 (2011)
3. K. D. MacIsaac and E. Fraenkel: Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput Biol*, 2(4), e36 (2006)
4. E. Valen, A. Sandelin, O. Winther and A. Krogh: Discovery of regulatory elements is improved by a discriminatory approach. *PLoS Comput Biol*, 5(11), e1000562 (2009)
5. E. van Nimwegen: Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics*, 8 Suppl 6, S4 (2007)
6. O. Elemento and S. Tavazoie: Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol*, 6(2), R18 (2005)
7. M. Kellis, N. Patterson, M. Endrizzi, B. Birren and E. S. Lander: Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937), 241-54 (2003)
8. M. Pritsker, Y. C. Liu, M. A. Beer and S. Tavazoie: Whole-genome discovery of transcription factor binding sites by network-level conservation. *Genome Res*, 14(1), 99-108 (2004)
9. G. Robertson, M. Bilenky, K. Lin, A. He, W. Yuen, M. Dagpinar, R. Varhol, K. Teague, O. L. Griffith, X. Zhang, Y. Pan, M. Hassel, M. C. Sleumer, W. Pan, E. D. Pleasance, M. Chuang, H. Hao, Y. Y. Li, N. Robertson, C. Fjell, B. Li, S. B. Montgomery, T. Astakhova, J. Zhou, J. Sander, A. S. Siddiqui and S. J. Jones: cisRED: a database

system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res*, 34(Database issue), D68-73 (2006)

10. X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander and M. Kellis: Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031), 338-45 (2005)
11. V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele and E. Wingender: TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 31(1), 374-8. (2003)
12. A. Sandelin, W. Alkema, P. Engstrom, W. W. Wasserman and B. Lenhard: JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32(Database issue), D91-4 (2004)
13. M. L. Bulyk: Computational prediction of transcription-factor binding site locations. *Genome Biol*, 5(1), 201 (2003)
14. G. D. Stormo: DNA binding sites: representation and discovery. *Bioinformatics*, 16(1), 16-23 (2000)
15. P. Perco, A. Kainz, G. Mayer, A. Lukas, R. Oberbauer and B. Mayer: Detection of coregulation in differential gene expression profiles. *Biosystems*, 82(3), 235-47 (2005)
16. G. Terai, T. Takagi and K. Nakai: Prediction of co-regulated genes in *Bacillus subtilis* on the basis of upstream elements conserved across three closely related species. *Genome Biol*, 2(11), RESEARCH0048 (2001)
17. G. Bejerano, A. C. Siepel, W. J. Kent and D. Haussler: Computational screening of conserved genomic DNA in search of functional noncoding elements. *Nat Methods*, 2(7), 535-45 (2005)
18. M. M. Mwangi and E. D. Siggia: Genome wide identification of regulatory motifs in *Bacillus subtilis*. *BMC Bioinformatics*, 4, 18 (2003)
19. J. D. Hughes, P. W. Estep, S. Tavazoie and G. M. Church: Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol*, 296(5), 1205-14 (2000)
20. Z. S. Qin, L. A. McCue, W. Thompson, L. Mayerhofer, C. E. Lawrence and J. S. Liu: Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nat Biotechnol*, 21(4), 435-9. (2003)
21. S. T. Jensen, L. Shen and J. S. Liu: Combining phylogenetic motif discovery and motif clustering to

- predict co-regulated genes. *Bioinformatics*, 21(20), 3832-9 (2005)
22. P. D'Haeseleer: How does DNA sequence motif discovery work? *Nat Biotechnol*, 24(8), 959-61 (2006)
23. R. Chen and J. S. Liu: Predictive Updating Methods With Application to Bayesian Classification. *Journal of the Royal Statistical Society Series B-Methodological*, 58(2), 397-415 (1996)
24. A. E. Gelfand and A. F. M. Smith: Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409 (1990)
25. J. S. Liu: Monte Carlo Strategies in Scientific Computing. Springer Verlag, Berlin (2001)
26. S. T. Jensen and J. S. Liu: Bayesian Clustering of Transcription Factor Binding Motifs. *Journal of American Statistical Association*, 103, 188-200 (2008)
27. L. Kaufman and P. J. Rousseeuw: Finding groups in data : an introduction to cluster analysis. Wiley, New York (1990)
28. L. Hubert, Arable, P.: Comparing partitions. *Journal of Classification*, 2, 193-218 (1985)
29. G. W. Milligan and M. C. Cooper: A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavior Research*, 21, 441-458 (1986)
30. Z. S. Qin: Clustering microarray gene expression data using weighted Chinese restaurant process. *Bioinformatics*, 22(16), 1988-97 (2006)
31. D. E. Schones, P. Sumazin and M. Q. Zhang: Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, 21(3), 307-13 (2005)
32. G. Su, B. Mao and J. Wang: MACO: a gapped-alignment scoring tool for comparing transcription factor binding sites. *In Silico Biol*, 6(4), 307-10 (2006)
33. S. Mahony, P. E. Auron and P. V. Benos: DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput Biol*, 3(3), e61 (2007)
34. A. Sandelin and W. W. Wasserman: Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol*, 338(2), 207-15 (2004)
35. M. Suzuki and N. Yagi: DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. *Proc Natl Acad Sci U S A*, 91(26), 12357-61 (1994)
36. B. Georgi and A. Schliep: Context-specific independence mixture modeling for positional weight matrices. *Bioinformatics*, 22(14), e166-73 (2006)
37. S. Hannenhalli and L. S. Wang: Enhanced position weight matrices using mixture models. *Bioinformatics*, 21 Suppl 1, i204-12 (2005)
38. N. I. Gershenzon, E. N. Trifonov and I. P. Ioshikhes: The features of Drosophila core promoters revealed by statistical analysis. *BMC Genomics*, 7, 161 (2006)
39. T. D. Schneider and R. M. Stephens: Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20), 6097-100 (1990)
40. J. Felsenstein: Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17(6), 368-76 (1981)
41. A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev and K. Zhao: High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4), 823-37 (2007)
42. D. S. Johnson, A. Mortazavi, R. M. Myers and B. Wold: Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, 316(5830), 1497-502 (2007)
43. T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T. K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander and B. E. Bernstein: Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153), 553-60 (2007)
44. G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder and S. Jones: Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, 4(8), 651-7 (2007)
45. V. R. Iyer, C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder and P. O. Brown: Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, 409(6819), 533-8. (2001)
46. B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell and R. A. Young: Genome-wide location and function of DNA binding proteins. *Science*, 290, 2306-2309 (2000)
47. Z. S. Qin, J. Yu, J. Shen, C. A. Maher, M. Hu, S. Kalyana-Sundaram, J. Yu and A. M. Chinnaiyan: HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics*, 11, 369 (2010)
48. M. Hu, J. Yu, J. M. Taylor, A. M. Chinnaiyan and Z. S. Qin: On the detection and refinement of transcription factor

A scalable model-based motif clustering approach

binding sites using ChIP-Seq data. *Nucleic Acids Res*, 38(7), 2154-67 (2010)

Abbreviations: TF: transcription factor; BMC: Bayesian motif cluster; BMCES: Bayesian motif cluster for eukaryotic species; PWM: position-specific weight matrices; TFBS: transcription factor binding sites; CLARA: clustering large applications; PAM: partition around medoids; NBR: normalized Bayes ratio; PAP: posterior assignment probability; ARI: adjust Rand index.

Key Words: Model-based clustering, Transcription factor binding site, Motif, Bayesian, Scalability.

Send correspondence to: Zhaohui S. Qin, Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, GCR338, 1518 Clifton Road, Atlanta, GA 30322, USA, Tel: 404-712-9576, Fax: 404-727-1370, E-mail: zhaohui.qin@emory.edu