

# Supplementary Material for

## MotifOrganizer: A scalable two-stage model-based clustering approach for grouping conserved non-coding elements in mammalian genomes

Zhaohui Qin<sup>1</sup>, Gordon Robertson<sup>2</sup>, Misha Bilenky<sup>2</sup>, Gang Su<sup>3</sup>, Steven Jones<sup>2</sup>

<sup>1</sup> Department of Biostatistics, University of Michigan, Ann Arbor MI 48109

<sup>2</sup> BC Cancer Agency Genome Sciences Centre, Vancouver BC Canada

<sup>3</sup> Bioinformatics Program, University of Michigan, Ann Arbor MI 48109

### 1. BMC and BMC2 models

We have previously developed a Bayesian model-based clustering algorithm called BMC (Bayesian Motif Cluster)<sup>1</sup>. In this model, we assumed that columns of the aligned motifs are independent, and nucleotides in each column follow a multinomial distribution. Therefore motifs that belong to the same cluster follow the same product multinomial distribution characterized by a common  $4 \times W$  probability matrix  $\Theta$  ( $W$  is the motif length). The overall likelihood of observing all the motifs, denoted as  $Y$ , can then be expressed as:

$$P(Y|\Theta_1, \dots, \Theta_K, E) \propto \prod_{k=1}^K \prod_{j=1}^W \prod_{i=1}^4 \theta_{ijk}^{L_{ij}^k},$$

where  $n$  is the total number of motifs,  $E = (e_1, \dots, e_n)$  is the cluster indicator variable,  $L_{ij}^k$  denote the  $(i,j)^{\text{th}}$  entry of the PSWM for the  $k^{\text{th}}$  motif cluster and  $K$  is the current number of clusters.

The prior distribution for each column of  $\Theta$  is Dirichlet distribution

$$P(\theta_1, \theta_2, \theta_3, \theta_4) = \frac{\Gamma(\beta_1 + \beta_2 + \beta_3 + \beta_4)}{\Gamma(\beta_1) \times \Gamma(\beta_2) \times \Gamma(\beta_3) \times \Gamma(\beta_4)} \theta_1^{\beta_1-1} \theta_2^{\beta_2-1} \theta_3^{\beta_3-1} \theta_4^{\beta_4-1}.$$

We set the  $\beta_j$ 's to 1 for all positions and for the background model, providing non-specific information about the motifs.

The prior distribution for  $E$  is defined individually and sequentially.

$$P(e_i | E_{-i}) = \begin{cases} \frac{1}{K+q} & \text{if } e_i = 1, 2, \dots, K, \\ \frac{q}{K+q} & \text{Otherwise.} \end{cases}$$

$K$  represents the current number of clusters, and  $q$  is a tuning parameter, which is fixed at 1 throughout this study.

Our goal is to infer cluster membership vector  $E$ . For ease of computation, we apply the predictive updating technique<sup>2</sup> to integrate out nuisance parameters  $\Theta_k$  analytically to improve the computation efficiency. The marginal likelihood is

$$P(Y | E) = \left( \prod_{k=1}^K \prod_{j=1}^W \frac{\prod_{i=1}^4 \Gamma(L_{ij}^k + \beta_i)}{\Gamma(L^k + \beta)} \right) \left( \frac{\Gamma(\beta)}{\prod_{i=1}^4 \Gamma(\beta_i)} \right)^{KW}.$$

Where  $\beta = \sum_{i=1}^4 \beta_i$  and  $L^m = \sum_{i=1}^4 L_{ij}^m$  (the subscript  $j$  has been omitted for  $L^m$  since column sums are identical).

We first randomly assign the  $n$  motifs into an arbitrary number of clusters. An iteration procedure follows, in which all motifs are reassigned in turn. First, a selected motif is taken from its current assigned cluster, and treated as a new object. Then, given the current assignment of all other motifs, the algorithm reassigns this motif by sampling from a multinomial distribution with probabilities  $Q_l = P(e_i = l | e_0, \dots, e_{i-1}, e_{i+1}, \dots, e_n)$ . We have

$$Q_l \propto \frac{\left( \prod_{j=1}^W \prod_{i=1}^4 \left( \frac{\Gamma(L_{ij}^{*k} + \beta_i)}{\Gamma(L_{ij}^k + \beta_i)} \right) \right) \left( \frac{\Gamma(L^k + \beta)}{\Gamma(L^{*k} + \beta)} \right)^W}{\left( \prod_{j=1}^W \frac{\prod_{i=1}^4 \Gamma(l_{ij}^m + \beta_i)}{\Gamma(l^m + \beta)} \right) \left( \frac{\Gamma(\beta)}{\prod_{i=1}^4 \Gamma(\beta_i)} \right)^W}$$

When  $l > 0$  and  $Q_l \propto q$ , where  $l_{ij}^m$  is the entry of the  $i^{\text{th}}$  row and  $j^{\text{th}}$  column in motif  $m$ 's profile matrix,  $l^m = \sum_{i=1}^4 l_{ij}^m$  is the column total of the  $m^{\text{th}}$  motif's profile matrix, and  $L_{ij}^k$  is the  $(i, j)$  entry of the  $k^{\text{th}}$  mega-profile matrix and  $L^k = \sum_{i=1}^4 L_{ij}^k$ . An asterisk in  $L_{ij}^k$  and  $L^k$  means the updated version after motif  $m$  is included.

We also update the cluster number  $K$  depending on whether a new cluster is preferred at an assignment, which increases the cluster number by one, or annihilates an existing cluster. The iteration is continued until convergence as indicated by the stabilization of the likelihood. The final cluster assignments are those with the highest likelihood.

The new BMC 2 algorithm we proposed in this study allows cluster widths to be different and a motif can join a cluster of different width. We made two changes in the formulas from Section 1: replace  $w$  by  $w_k$  to allow different cluster widths, and introduce a new indicator variable  $o_s$  for each motif  $s$  which is the shift position within that motif. And we have  $L_{ij}^k = \sum_{e(s)=k} l_{i+o_s, j}^s$ .

Assume the motif width is  $u$ , and the cluster width is  $v$ . If  $u \geq v$ , slide the cluster across all  $u - v + 1$  possible alignment start positions in the motif,  $o_i = 1, 2, \dots, u - v + 1$ . If  $u < v$ , slide the motif across all  $u - v + 1$  possible alignment start positions in the cluster,

$o_i = -1, -2, \dots, u - v + 1$ . Both  $e_i$  and  $o_i$  will be determined by sampling from a multinomial distribution with all combinations of clustering assignments and alignment start positions as possible outcomes.

## 2. MotifOrganizer implementation

We devised a novel two-stage divide-conquer-combine strategy in MotifOrganizer to further improve its scalability. The data structure and overall operation scheme is illustrated in Figure S1.

At the beginning, users provided the initial number of clusters which can be specified at random. motifOrganizer randomly assigns motifs to these clusters. From our experience, using a larger number of initial clusters tends to speed up convergence. It is also a good strategy to use multiple runs each with different initial number of clusters to better explore the whole sample space, retaining the output with the highest likelihood.

Our model differs slightly from a Dirichlet process mixture model. The prior probability for a motif to join each existing cluster is assumed to be equal, whereas in a Dirichlet process mixture model, such prior probabilities are assumed to be proportional to the size of those existing clusters. Our experience with simulated and real data suggested that such "uniform" prior assignment probabilities tend to generate "pure" better quality, albeit more numbers of clusters, and the number of clusters generated using such a "uniform" prior is more accurate than the Dirichlet process prior.

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. The results presented here were sometimes obtained with fewer than 100,000 reassignments. 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 a input parameter), each with different initial setting, to better explore the entire sample space.

There is a tuning parameter in our model, which affects the prior probability of a motif forming a new cluster with itself as the sole member. Our comparison studies (data not shown) suggested that the clustering results are robust to this tuning parameter, as was the case for BMC. Therefore we fixed this parameter to be the default value of 1.0 throughout this study as we did for BMC.

## 3. Quality measure

Model-based clustering methods allow formal statistical inference to be performed. BMC, BMC2 and MotifOrganizer take advantage of this to produce a set of significance measures. One measure is the ratio of 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 BMC2, since the widths of motifs may differ between clusters, we define normalized Bayes ratio (NBR) for a cluster by dividing  $B$  by cluster size (number of motifs in this 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.

#### 4. Distance-based clustering approaches

We compared the performance of MotifOrganizer with distance-based clustering methods. There are two basic components in distance-based clustering methods: a distance metric and a clustering strategy. Two different distance metrics were used, depending on the type of data to be clustered. For motifs containing multiple sequences, we adopted the pairwise Kullback-Leibler (KL) distances, which is the default distance defined between PWMs in a commonly used motif comparison program MatCompare<sup>3</sup>. For motifs that represented by only one motif sequence, we used the pairwise Levenshtein or edit distance, which measures the similarity between two strings as the least number of deletions, insertions, or substitutions required to transform one into the other<sup>4</sup>. Two types of clustering strategies were applied in this study. For datasets with less than 1000 entries, we clustered the dissimilarity matrices using Partitioning Around Medoids (PAM)<sup>5</sup> as in MatCompare<sup>3</sup>. When the number of objects was larger than 1000, we choose CLARA<sup>5</sup> which is similar to PAM but is faster and requires less memory. Kaufman and Rousseeuw recommended CLARA over PAM when the number of objects is large. In this study, we calculated the KL distance using the MatCompare software<sup>3</sup> and calculated edit distance using the Perl script in Tisdall 2003. Both PAM and CLARA were performed using the ‘pam’ and ‘clara’ functions in the R ‘cluster’ library<sup>6</sup>. Following advice from Kaufman and Rousseeuw<sup>5</sup>, we determined the number of clusters by repeated PAM or CLARA clustering using a range of target cluster numbers,  $K$ , and selecting the value of  $K$  that maximized the average silhouette width for the whole dataset.

#### 5. Clustering accuracy

The following section is adapted from the document *Details of the Adjusted Rand index and Clustering algorithms Supplement to the paper “Principal Component Analysis for clustering gene expression data”*<sup>7</sup>. For details, please refer to their original document, which can be found at <http://faculty.washington.edu/kayee/pca/supp.pdf>.

To evaluate and compare the performance of different clustering approaches, we need to quantify the agreement between different clustering results. There are numerous ways to compare the agreement between two different partitions among objects in clustering analysis. Milligan and Cooper recommended Adjusted Rand Index (ARI) among many different indices after extensive empirical comparisons and evaluations with different

cluster sizes. ARI is derived from the Rand index<sup>8</sup>, which is defined as the number of pairs that are either in the same groups in both partitions or in different groups in both partitions, divided by the total number of pairs of objects. The values of ARI lie between 0 and 1, with a higher value indicating better agreement. The expected value in the case of random partitions is 0.

Assume we are interested in comparing two different partitions of  $N$  genes:

$P = (p_1, p_2, \dots, p_{K_p})$  and  $Q = (q_1, q_2, \dots, q_{K_q})$ . The agreement can be measured by the ARI, which measure the degree of overlap between these two partitions. Let  $n_{ij}$  represents the common elements in the  $i$ th cluster of partiton 1 and the  $j$ th cluster in partition 2.

Clusters	1	2	...	$q$	Sums
1	$n_{11}$	$n_{12}$	...	$n_{1q}$	$n_{1.}$
2	$n_{21}$	$n_{22}$	...	$n_{2q}$	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$p$	$n_{p1}$	$n_{p2}$	...	$n_{pq}$	$n_{p.}$
Sums	$n_{.1}$	$n_{.2}$	...	$n_{.q}$	$n_{..} = N$

The ARI is calculated by the following formula:

$$\frac{\sum_{i,j} \binom{n_{ij}}{2} - \left[ \sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2} \right] / \binom{N}{2}}{\frac{1}{2} \left[ \sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2} \right] - \left[ \sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2} \right] / \binom{N}{2}}.$$

The ARI ranges from 0 to 1. The higher the index, the better agreement between these two partitions.

## 6. Motif matching tools

To annotate the clustering results, we compare each cluster to motif patterns stored in public databases including JASPAR<sup>9</sup> and TRANSFAC<sup>10</sup>. We used two different programs to assess the similarity of two motif patterns.

### MatCompare

MatCompare<sup>3</sup> quantifies similarity between two PWMs using one of several distance measures. We used the default distance measure which is the minimum KL divergence between matched fragments between two motifs. One can also calculate the p-values from Fisher's exact test on a contingency table or Chi-square approximation. Motifs with divergence per column less than 1.0 are regarded as very similar. We follow the default threshold which set to call all divergences less than 1.0 as a match.

### MACO

MACO<sup>11</sup> computes Pearson correlation coefficient score on the basis of consensus-letter based alignment between two motif PWMs. The threshold value for the matching scores was determined from the empirical distribution obtained using 3000 random matrices and 506 library matrices.

## 7. Data

### JASPAR

JASPAR<sup>9</sup> (<http://jaspar.genereg.net/>) is a collection of transcription factor DNA-binding preferences, saved as PSWMs. The latest version of JASPAR contains three sub-databases. JASPAR CORE, JASPAR FAM and JASPAR phyloFACTS. The original 1716 JASPAR CORE motif sequence data in FASTA format is downloaded from the JASPAR website. A non-redundant subset that contains 1152 motif sequences was used in this study.

### TRANSFAC

TRANSFAC<sup>10</sup> (<http://www.gene-regulation.com/pub/databases.html#transfac>) is a database on eukaryotic *cis*-acting regulatory DNA elements and *trans*-acting factors. In this study, we downloaded all 7641 sequence motifs from the TRANSFAC website. The selected subset of 5452 motif sequences (each motif belong to only one profile) was used in this study.

### *cisRED*

*cisRED*<sup>12</sup> (<http://www.cisred.org>) is a database for conserved DNA sequence motifs and co-occurring motif patterns that are identified and ranked by a genome-scale computational system. The system applies multiple, complementary *de novo* motif discovery methods to such inputs to identify conserved sequence motifs with a typical motif width of 6-30bp.

For every gene in human (mouse) genome, the system creates sequence inputs that currently include over 40 vertebrate (mostly mammalian) species (see Table S1). It utilizes UCSC multiple sequence alignments<sup>13</sup>, and adds targeted sequence data generated in the ENCODE project<sup>14</sup>. Sequence data from partially assembled genomes, and trace file sequence data from ‘low coverage’ genomes<sup>15</sup>. Motifs are predicted using multiple discovery methods, and motif significance is estimated by applying discovery and post-processing methods to randomized sequence sets that are adaptively derived from target sequence sets. The latest Human v.2 database contains ~200K motifs blocks. For promoter regions (that cover 1.5Kb upstream of transcription start site (TSS) net of most repeats) of ~18K human genes, the current

The dataset used in this study contains 29,490 motifs. Total number of sequence motifs is 194,577. All three datasets used in this study can be found at <http://www.sph.umich.edu/csg/qin/motif>.

### Additional References:

- 1      Qin ZS, McCue LA, Thompson W, Mayerhofer L, Lawrence CE, Liu JS: Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nat Biotechnol* 2003; **21**: 435-439.
- 2      Chen R, Liu JS: Predictive Updating Methods With Application to Bayesian Classification. *Journal of the Royal Statistical Society Series B-Methodological* 1996; **58**: 397-415.
- 3      Schones DE, Sumazin P, Zhang MQ: Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics* 2005; **21**: 307-313.
- 4      Levenshtein VI: Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 1966; **10**: 707-710.
- 5      Kaufman L, Rousseeuw PJ: Finding groups in data : an introduction to cluster analysis. New York, Wiley, 1990.
- 6      Struyf A HM, Rousseeuw P: Clustering in an Object-Oriented Environment. *Journal of Statistical Software* 1996; **1**.
- 7      Yeung KY, Ruzzo WL: Principal component analysis for clustering gene expression data. *Bioinformatics* 2001; **17**: 763-774.
- 8      Rand WM: Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 1971; **66**.
- 9      Vlieghe D, Sandelin A, De Bleser PJ *et al*: A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res* 2006; **34**: D95-97.
- 10     Matys V, Kel-Margoulis OV, Fricke E *et al*: TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 2006; **34**: D108-110.
- 11     Su G, Mao B, Wang J: MACO: a gapped-alignment scoring tool for comparing transcription factor binding sites. *In Silico Biol* 2006; **6**: 307-310.
- 12     Robertson G, Bilenky M, Lin K *et al*: cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res* 2006; **34**: D68-73.
- 13     Kuhn RM, Karolchik D, Zweig AS *et al*: The UCSC genome browser database: update 2007. *Nucleic Acids Res* 2007; **35**: D668-673.
- 14     ENOCODE: Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007; **447**: 799-816.
- 15     Bainbridge MN, Warren RL, He A, Bilenky M, Robertson AG, Jones SJ: THOR: targeted high-throughput ortholog reconstructor. *Bioinformatics* 2007; **23**: 2622-2624.

**Table S1. The 30 species used in cisRED that contributed to the 30K motif cisRED Human v.2 dataset.**

Species	Common Name
<i>Mus musculus</i>	Mouse
<i>Pan troglodytes</i>	chimpanzee
<i>Macaca mulatta</i>	rhesus macaque
<i>Canis familiaris</i>	dog
<i>Gallus gallus</i>	chicken
<i>Bos taurus</i>	cow
<i>Monodelphis domestica</i>	opossum
<i>Rattus norvegicus</i>	rat
<i>Tetraodon nigroviridis</i>	freshwater pufferfish
<i>Danio rerio</i>	zebrafish
<i>Loxodonta africana</i>	elephant
<i>Oryctolagus cuniculus</i>	rabbit
<i>Dasypus novemcinctus</i>	armadillo
<i>Echinops telfairi</i>	hedgehog
<i>Callithrix jacchus</i>	marmoset
<i>Cavia porcellus</i>	guinea pig
<i>Choloepus hoffmanni</i>	sloth
<i>Equus caballus</i>	horse
<i>Felis catus</i>	cat
<i>Macropus eugenii</i>	tammar wallaby
<i>Myotis lucifugus</i>	little brown bat
<i>Ornithorhynchus anatinus</i>	platypus
<i>Otolemur garnettii</i>	bush baby
<i>Ovis aries</i>	sheep
<i>Papio anubis</i>	olive baboon
<i>Papio cynocephalus</i>	yellow baboon
<i>Pongo pygmaeus</i>	orangutan
<i>Procavia capensis</i>	hyrax
<i>Rhinolophus ferrumequinum</i>	bat
<i>Sorex araneus</i>	common shrew
<i>Sus scrofa</i>	pig



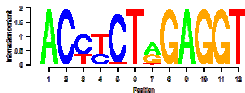
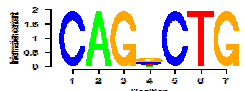
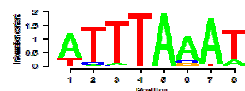
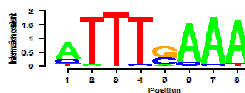
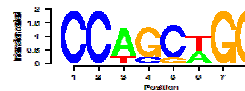
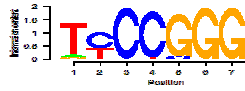
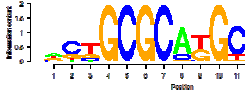
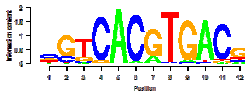
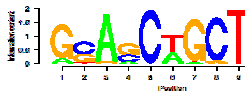
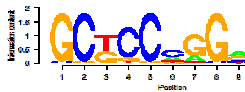
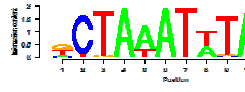
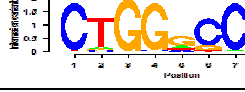
**Table S2. Compare clustering performance between regular model-based clustering approach BMC and the proposed two-stage clustering scheme MotifOrganizer using the ~5000 motif TRANSFAC dataset. Study was performed under various quality measure threshold settings. Cluster performance is measured by Adjust Rand Index (ARI). Both clustering procedures were performed 100 times under each setting. Both average ARI and standard deviation (in parentheses) were reported.**

PAP		$\text{NBR} \geq 0$	$\text{NBR} \geq 2$	$\text{NBR} \geq 5$
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)

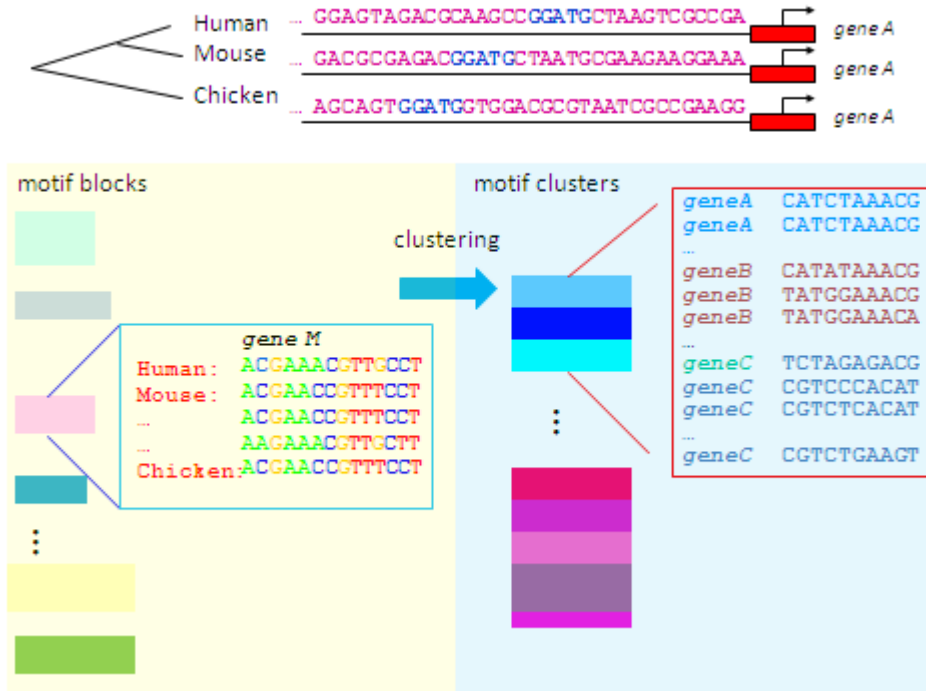
**Table S3. Compare number of clusters generated with the actual number of motif profiles between regular model-based clustering approach BMC and the proposed two-stage clustering scheme MotifOrganizer using the ~5000 motif TRANSFAC dataset. Study was performed under various quality measure thresholds. Both clustering procedures were performed 100 times under each setting. Both average cluster number and standard deviation (in parentheses) were reported.**

PAP	Method		$\text{NBR} \geq 0$	$\text{NBR} \geq 2$	$\text{NBR} \geq 5$
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)

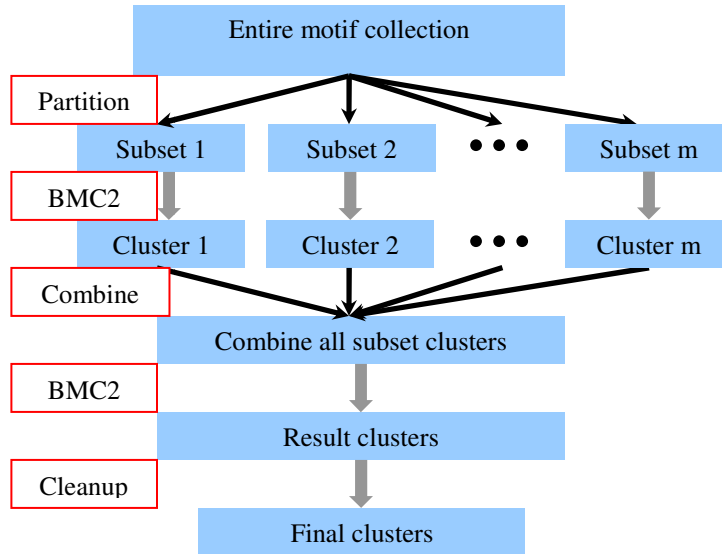
**Table S4. Selected motif clusters from the result of clustering the 30K motif cisRED Human v.2 dataset using motifOrganizer. The motif pattern of these clusters does not match to any JASPAR CORE, phyloFACTS or TRANSFAC TFBS models according to both MatCompare and MACO software. The clusters are sorted by normalized Bayes ratio (NBR).**

Cluster ID	Motif cluster logo	NBR	PHYLOFACTS model logo
71		22.9	DUSP8, NFKBIA
78		22.4	GGTLA1, DBH
83		22.1	C14orf103, NP_060229.2
94		20.9	MID1, CLDN17, VRK1
151		18.1	MC1R, DBH, COL2A1
159		17.7	NP_079167.1, C9orf82, HNRPR
187		16.3	SAFB, MYBBP1A, SDCCAG8, EIF2S2, ZC3H8
196		16.0	HSF2, RAB5A, LMTK2, BRDT, LATS2, OSBPL8, PHF20L1
218		15.3	Q8NBE0_HUMAN, BRUNOL4
228		15.1	GRM5, WNT3A, FAM49A, ANKRD5, AP3S2, HOXD11
232		14.9	TTN, PICALM, ESR1, KBTBD5
238		14.7	Q6ZT51_HUMAN, ZNF385, ARHGAP23, CNM1

A.



B.



**Figure S1. A.** Illustration of phylogenetic footprinting technique and the motif clustering strategy, as well as examples of motifs and motif clusters. **B.** 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.