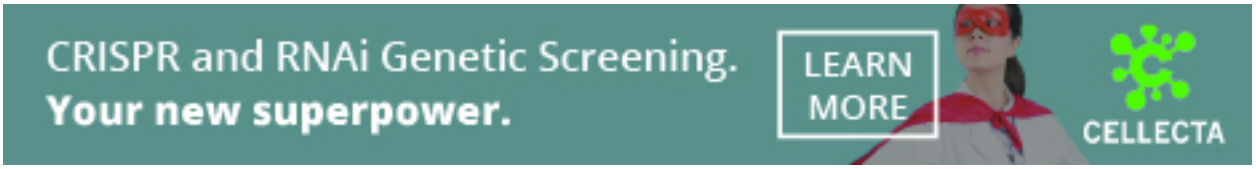
Cross-species de novo identification of *cis*-regulatory modules with GibbsModule: Application to gene regulation in embryonic stem cells

Dan Xie, Jun Cai, Na-Yu Chia, et al.

Genome Res. 2008 18: 1325-1335 originally published online May 15, 2008
Access the most recent version at doi:10.1101/gr.072769.107



Cross-species de novo identification of *cis*-regulatory modules with GibbsModule: Application to gene regulation in embryonic stem cells

Dan Xie,[1] Jun Cai,[1] Na-Yu Chia,[2] Huck H. Ng,[2,3] and Sheng Zhong[1,4,5,6,7]

[1]*Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA;* [2]*Gene Regulation Laboratory, Genome Institute of Singapore, Singapore 138672;* [3]*Department of Biological Sciences, National University of Singapore, Singapore 117543;* [4]*Department of Statistics, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA;* [5]*Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA;* [6]*Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*

We introduce the GibbsModule algorithm for de novo detection of *cis*-regulatory motifs and modules in eukaryote genomes. GibbsModule models the coexpressed genes within one species as sharing a core *cis*-regulatory motif and each homologous gene group as sharing a homologous *cis*-regulatory module (CRM), characterized by a similar composition of motifs. Without using a predetermined alignment result, GibbsModule iteratively updates the core motif shared by coexpressed genes and traces the homologous CRMs that contain the core motif. GibbsModule achieved substantial improvements in both precision and recall as compared with peer algorithms on a number of synthetic and real data sets. Applying GibbsModule to analyze the binding regions of the Krüppel-like factor (KLF) transcription factor in embryonic stem cells (ESCs), we discovered a motif that differs from a previously published KLF motif identified by a SELEX experiment, but the new motif is consistent with mutagenesis analysis. The SOX2 motif was found to be a collaborating motif to the KLF motif in ESCs. We used quantitative chromatin immunoprecipitation (ChIP) analysis to test whether GibbsModule could distinguish functional and nonfunctional binding sites. All seven tested binding sites in GibbsModule-predicted CRMs had higher ChIP signals as compared with the other seven tested binding sites located outside of predicted CRMs. GibbsModule is available at http://biocomp.bioen.uiuc.edu/GibbsModule.

[Supplemental material is available online at www.genome.org.]

Significant advances have been made in the past 15 yr on the computational prediction of transcription factor binding sites (TFBSs) in eukaryote genomes. The accuracy of such predictions has reached a plateau where achieving significant improvements seems difficult. In particular, de novo *cis*-regulatory motif identification algorithms (Bailey and Elkan 1994; Hughes et al. 2000; Liu et al. 2001), capable of locating TFBSs in promoter regions, are expected to have very limited power in the search for TFBSs in distal promoters or enhancers.

Unlike compact genomes such as that of *Saccharomyces cerevisiae*, in which TFBSs typically locate in promoter regions close to the transcription start sites, TFBSs in higher eukaryotic genomes often locate in distal promoters or enhancers that can be tens of thousands of bases (kilobases) of nucleotides away from the transcription start sites (Banerji et al. 1981; Carter et al. 2002; Vokes et al. 2007). Furthermore, the TFBSs that regulate the time and tissue expression domains of a target gene often appear in enhancer rather than promoter regions (Davidson et al. 2002, 2003; Papatsenko and Levine 2005; Davidson 2006). The longer distances between TFBSs and transcription start sites in higher eukaryotes impose a greater computational challenge for de novo motif finding. This is because motif-finding tools are required to search longer sequences, i.e., a larger search space. It is well known that iterative and stochastic searches can easily be trapped into local maxima when the search space is large.

Several experimental and computational strategies have been implemented to identify *cis*-regulatory motifs in both promoter and enhancer regions. Chromatin immunoprecipitation followed by microarray hybridization (ChIP-chip) is used to obtain the DNA-binding regions of a transcription factor, often in a whole genome (Boyer et al. 2005; Venkatesh et al. 2006; Kim et al. 2007). ChIP-chip experiments can narrow enhancer regions down to the range of about 500 base pairs (bp), and ChIP-chip positive regions are good input data for motif searches. The binding sites for a set of interacting transcription factors have the tendency to colocalize into one *cis*-regulatory module (CRM). Assuming all the motifs that constitute a CRM are *known*, researchers have developed algorithms to utilize these motifs to identify enhancer regions (Hallikas et al. 2006; Sinha et al. 2006). Both ChIP-chip and the set of DNA motifs known to constitute a CRM require substantial prior knowledge and experimental efforts on the biological system and the focal process. In exploratory studies for unknown TFBSs, tools for de novo motif discovery from a set of coexpressed genes are still in great demand. Two major directions for de novo identification of *cis*-regulatory elements have been explored.

Comparative genome sequence analysis is one of the best strategies known for finding functional sequences in animal genomes. The basic idea is to look for sequences that are conserved

[7]**Corresponding author.**
E-mail szhong@uiuc.edu; fax (217) 265-0246.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.072769.107.

across species (e.g., Loots et al. 2000; Boffelli et al. 2003; Hardison et al. 2003; Kellis et al. 2003; Margulies et al. 2003; Woolfe et al. 2005). Due to negative (purifying) selection, orthologous sequences that are significantly more similar than what would be expected under some reasonable model of neutral evolution are likely to have critical functional roles (Bejerano et al. 2004; Siepel et al. 2005; Chen et al. 2007; Katzman et al. 2007). Two major approaches have been used to identify sequences that have undergone purifying selection. The first is to look for ultraconserved elements in closely related genomes. Pioneering examples include using high conservation thresholds on mammalian genomes (Bejerano et al. 2004; Chen et al. 2007; Katzman et al. 2007) and Phylogenetic Shadowing on primate genomes (Boffelli et al. 2003; Ovcharenko et al. 2004). The second is to identify elements conserved in genomes spanning a large phylogenetic distance, typically over 400 million years (Siepel et al. 2005; Venkatesh et al. 2006). Phylogenetic footprinting (Blanchette and Tompa 2002, 2003; Blanchette et al. 2002) and its variations (Thomas et al. 2003; Siepel et al. 2005; Sosinsky et al. 2007) are formalized computational methods using this approach.

An orthogonal direction of utilizing homologous sequences for TFBS identification is to use coexpressed genes in the same species. The general problem is to identify *cis*-regulatory motifs in a given set of genes that are likely to be regulated by the same (group of) transcription factor(s) (Zhou and Wong 2004). WordEnumeration (Sinha and Tompa 2002), MEME (Bailey and Elkan 1994), AlignACE (Hughes et al. 2000), and BioProspector (Liu et al. 2001) are among the well-established methods on this direction. Leveraging on the fact that TFBSs are generally both more *overrepresented* across co-regulated genes and more *conserved* across species, a number of recent developments, including CompareProspector (Liu et al. 2004), PhyloCon (Wang and Stormo 2003), PhyloGibbs (Siddharthan et al. 2005), and others (Moses et al. 2004; Sinha et al. 2004; Prakash and Tompa 2005) have shown large improvements compared with methods that use only one of the two properties (conservation and overrepresentation). These methods operate under a predetermined alignment result. Although there are valid arguments regarding the disadvantages of using a predetermined alignment result (Sosinsky et al. 2007; Zhou and Wong 2007), in practice these recently developed methods nevertheless generate the most reliable motifs (Wang and Stormo 2003; Liu et al. 2004; Siddharthan et al. 2005).

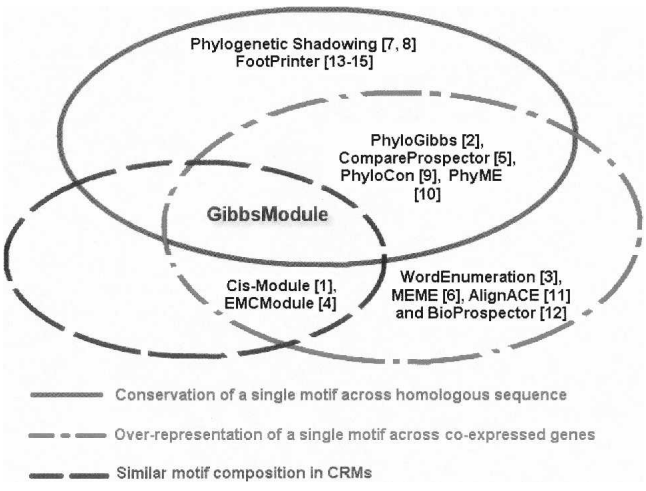


Figure 1. Three information sources for de novo identification of *cis*-regulatory motifs. The three circles represent three information sources that can be utilized for motif and CRM finding. A tool enclosed in a circle indicates that this tool utilizes that information.

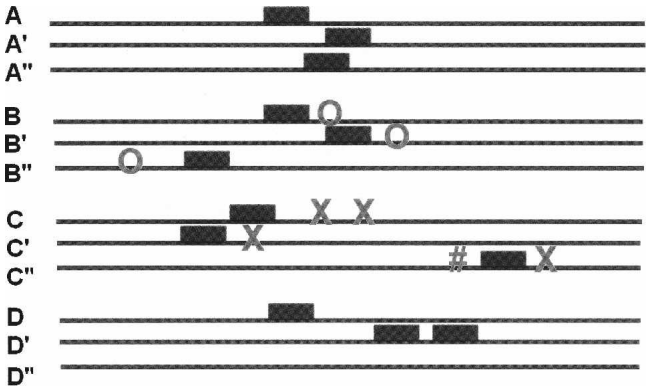


Figure 2. Motifs and CRMs in coexpressed genes and their homologous genes. A, B, C, and D represent coexpressed genes in one species. A′ and A′′ represent the homologous genes to A in two other species, and so on. (X, O, #) TFBSs for different transcription factors.

In eukaryotes, a set of TFBSs that attract interacting transcription factors often colocalize in the genome sequence, forming a CRM. Utilizing this information, joint modeling of TFBSs in CRMs in a single species has demonstrated substantial improvements in de novo motif identification (Zhou and Wong 2004; Gupta and Liu 2005). Figure 1 illustrates the information sources used in motif identification algorithms. It is tempting to combine motif overrepresentation across coexpressed genes and evolutionary conservation of a motif and conservation of CRMs, i.e., all three information sources, to improve de novo motif identification.

Methods

GibbsModule

We present the GibbsModule model and algorithm for de novo *cis*-regulatory motif detection in eukaryotic genomes. GibbsModule takes a list of upstream sequences[8] of coexpressed genes and their homologous sequences as input. We denote the species in which the coexpressed genes are originally obtained as the *target species*, and the other species from which the homologous sequences are retrieved as *assisting species*.

GibbsModule models the coexpressed genes in the target species as sharing a *core cis*-regulatory motif and each homologous gene group as sharing a homologous CRM, characterized by a similar composition of motifs. The core motif is assumed to be overrepresented across coexpressed genes (Fig. 2). Some of these motifs may locate in CRMs. GibbsModule iteratively identifies potential enhancer regions and a core *cis*-regulatory motif within these enhancers. GibbsModule does not require all real TFBSs to locate within conserved CRMs.

GibbsModule utilizes a Gibbs sampling scheme for motif search. A classical Gibbs motif sampler iteratively updates the position specific weight matrix (PSWM) for the motif and

[8]Although we use the term ''upstream sequence'' to describe the method, in practice, the method should be applied to any regions that may contain *cis*-regulatory elements.

samples the locations of TFBSs in the upstream of the group of coexpressed genes (Lawrence et al. 1993). To update TFBS locations in each iteration, a location is sampled on an input sequence from the *posterior* motif distribution on that sequence. A Gibbs sampler has proved to be powerful in detecting common motifs in the input sequences (Lawrence et al. 1993). A drawback of a Gibbs sampler is that it easily falls into local maxima, especially when the input sequences are long. GibbsModule is designed to utilize the conservation of CRMs across species to overcome the drawback of a traditional Gibbs motif sampler. Without using a predetermined alignment result, GibbsModule iteratively traces the homologous CRMs and updates a core motif shared by these CRMs.

In a GibbsModule iteration, instead of sampling one TFBS on each input sequence, it first samples a set of candidate TFBSs on each input sequence as well as on their homologous sequences. Some of the sampled TFBSs may be real sites within CRMs, while the others can be false positives. The CRMs are likely to be more conserved across homologous sequences as compared with neutral sequences. Therefore, GibbsModule assumes the neighboring area of a TFBS in a CRM is more conserved than the neighboring area of a TFBS not in a CRM. In all the experiments described later in this paper, we set GibbsModule to sample three candidate TFBSs on each input sequence and on each of their homologous sequences (Step 2, Fig. 3). GibbsModule then selects one out of the three candidates as the updated TFBS (Step 4, Fig. 3). This selection is judged by which candidate TFBS is most likely to locate within a conserved CRM (Step 3, Fig. 3).

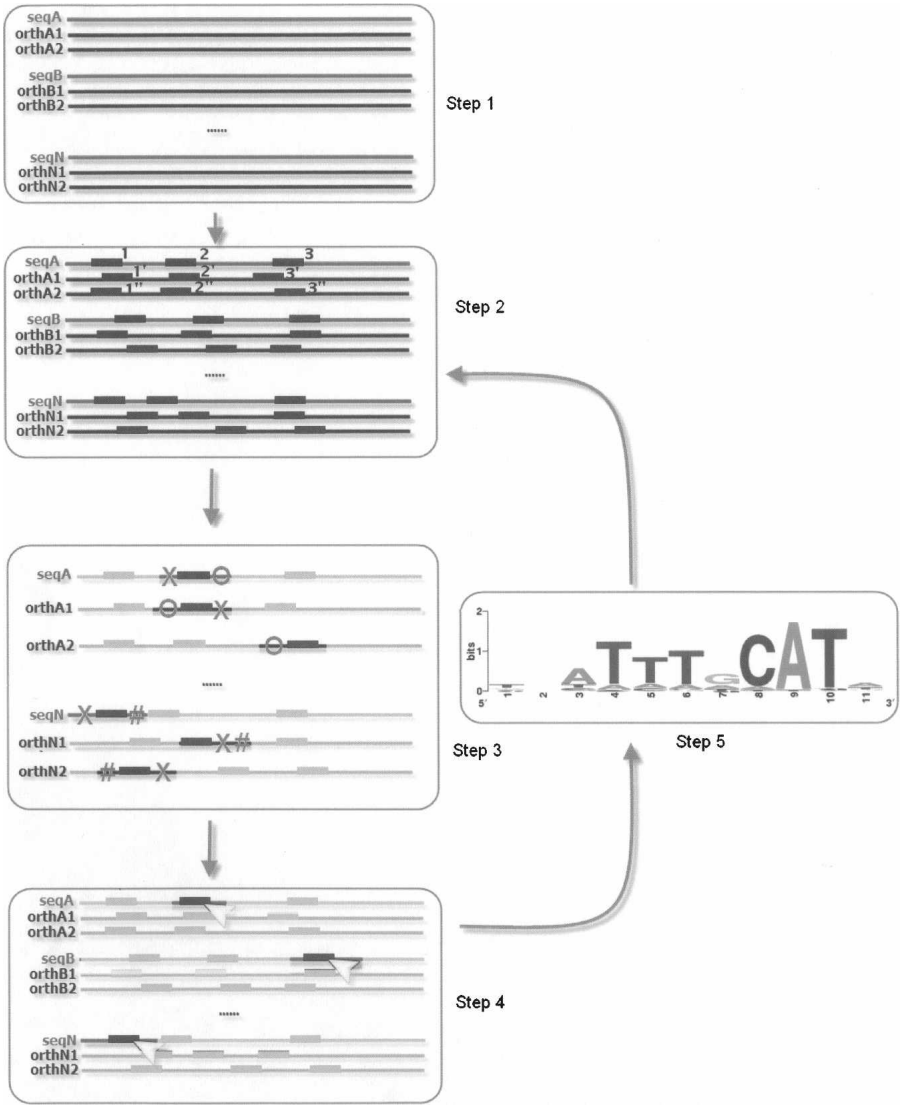


Figure 3. GibbsModule workflow. In Step 1, a random PSWM is initialized. Steps 2–5 are the iterative steps. In Step 2, N candidate binding sites are sampled from every homologous sequence using the same PSWM. In this example, three candidate-binding sites are sampled on each sequence (N = 3). Every sampled binding site defines a candidate CRM, which includes the binding site itself and 100 bp of flanking region on each side. These candidate CRMs are marked 1, 2, 3 on the target sequence, and 1′, 2′, 3′ and 1″, 2″, 3″ on the sequences of two assisting species. In Step 3, Module-Alignment is applied to every candidate CRM on the target sequence and every CRM on the assisting sequences. In the example, the alignments are applied to CRM pairs of (1, 1′), (1, 2′), (1, 3′), (2, 1′), (2, 2′), . . . , (3, 1″), (3, 2″), and (3, 3″). In Steps 3 and 4, a most conserved CRM on the target sequence is picked up by arg $\max_{n}(\max_{n'}(score(n,n')) + \max_{n''}(score(n,n')))$, where n, n', and n'' are indicators of candidate CRMs in homologous sequences SeqA, orthA1, and orthA2, respectively. (X, O, #) Other motifs close to the core motif within a CRM. In Step 5, a new PSWM is calculated from the core motifs in the most conserved CRMs.

In a GibbsModule iteration, instead of sampling one TFBS on each input sequence, it first samples a set of candidate TFBSs on each input sequence as well as on their homologous sequences. Some of the sampled TFBSs may be real sites within CRMs, while the others can be false positives. The CRMs are likely to be more conserved across homologous sequences as compared with neutral sequences. Therefore, GibbsModule assumes the neighboring area of a TFBS in a CRM is more conserved than the neighboring area of a TFBS not in a CRM. In all the experiments described later in this paper, we set GibbsModule to sample three candidate TFBSs on each input sequence and on each of their homologous sequences (Step 2, Fig. 3). GibbsModule then selects one out of the three candidates as the updated TFBS (Step 4, Fig. 3). This selection is judged by which candidate TFBS is most likely to locate within a conserved CRM (Step 3, Fig. 3).

We designed a Module-Alignment algorithm to evaluate the conservation of CRMs (see Module-Alignment). GibbsModule computes a conservation score from Module-Alignment for each candidate TFBS in the target species. For example, if two candidate TFBSs on homologous genes are contained in orthologous CRMs, a high conservation score is expected from Module-Alignment when aligning the neighboring sequences of the two candidates. Because it is uncertain which TFBSs are orthologous, Module-Alignment is applied to all pairs between every candidate TFBS on the target species and every candidate TFBS on every assisting species to compute conservation scores. For example, if there are two homologous sequences for a gene, three candidate TFBSs will be sampled on each homologous sequence. Then Module-Alignment will be applied nine (= 3 × 3) times, and nine conservation scores will be computed. The pair of TFBSs with the largest conservation score is regarded as "orthologous" for this iteration, and their neighboring sequences are supposed to contain orthologous CRMs. The largest pairwise conservation score will be assigned as the conservation score for this TFBS on the target species.

For more than two species, the target species will first be aligned to every assisting species. The conservation score of a TFBS on the target species is the sum

of its conservation scores on every assisting species. The candidate TFBS with the largest conservation score on the target species will be selected as the sampled TFBS for its downstream gene in this iteration (Step 4, Fig. 3). All the sampled TFBSs in the target species are used to update the motif PSWM (Step 5, Fig. 3).

In practice, not all input sequences are guaranteed to contain a TFBS of the shared motif. To allow some input sequences to be devoid of the shared motif, GibbsModule has a built-in sampling threshold that increases at every iteration after burn-in iterations. In the iteration step of sampling TFBSs, GibbsModule computes the posterior probability for each position on a sequence of being a TFBS. GibbsModule completely ignores the positions whose posterior probabilities are below the sampling threshold and does not sample those positions. If all the positions on an input sequence have a posterior probability less than the sampling threshold, the whole sequence will be ignored in that iteration. This strategy of dealing with sequences devoid of TFBSs was first introduced by the authors of BioProspector (Liu et al. 2001).

Hereto, we have described the complete sampling scheme of GibbsModule. To further avoid local maxima, we ask each execution of GibbsModule to perform the sampler described above 50 times. In each run, GibbsModule outputs five locations of putative TFBSs per sequence. For each location on the sequence, we count the cumulative times that it is predicted to be a TFBS location (Supplemental Fig. S3A). If a location has been cumulatively predicted for more than δ times, it is regarded as a location for a TFBS and also as a central location for a CRM. Throughout this paper and as the default in the GibbsModule program, δ is set at 5.

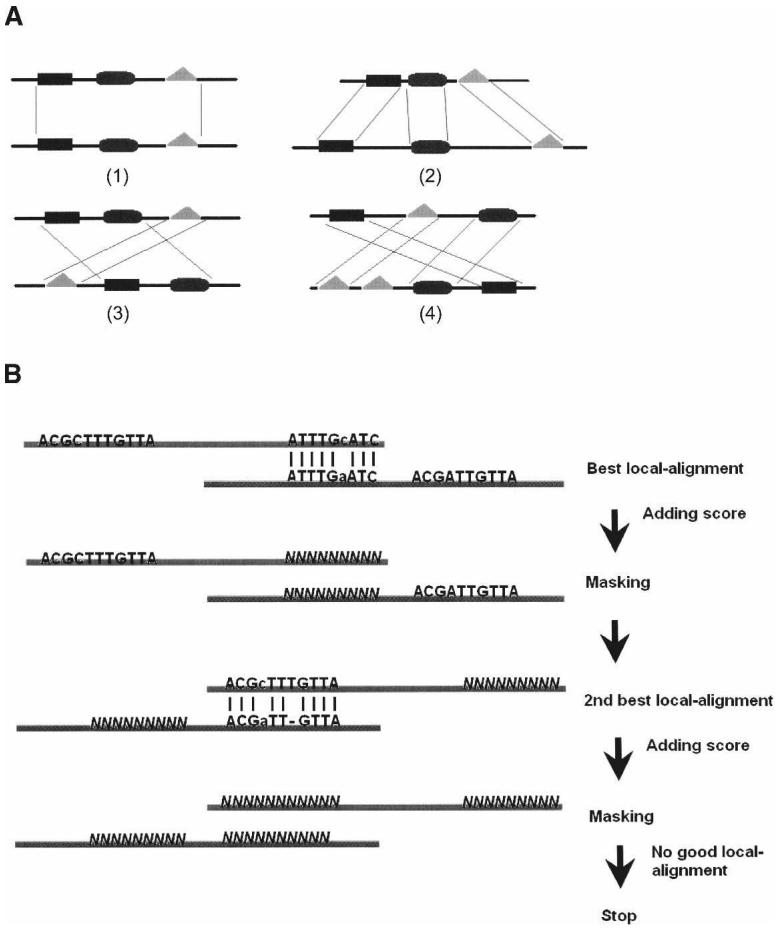


Figure 4. Module-Alignment. (*A*) An illustration of three pairs of orthologous CRMs: All three CRM pairs consist of TFBSs generated from the same motifs (squares, ellipses, and triangles). (*1*) Orthologous CRMs with conserved number, order, and distances. (*2–4*) Orthologous CRMs with different distances, order, and number of TFBSs. (*B*) Workflow of Module-Alignment: Module-Alignment iteratively performs local alignment and masks out conserved regions. The mutations and gaps between the alignable segments on the *upper* and *lower* sequences incur severe penalty so that Smith-Waterman can detect only the best local alignment in the first row. The conservation score from local alignment is the score of the best local alignment. However, the conservation score from Module-Alignment is the sum of the two local alignments scores from the two alignable sequence segments. Module-alignment is not designed to align any two sequences with any arbitrary lengths. Its input sequences should be potentially orthologous CRMs with lengths of several dozen to several hundred base pairs.

Module-Alignment

Module-Alignment is an extension of the Smith-Waterman algorithm for local alignment (Smith and Waterman 1981). Module-Alignment computes a conservation score between any two input sequences. Compared with Smith-Waterman, Module-Alignment would generate a higher conservation score if the input sequences are orthologous CRMs. In other words, Module-Alignment is designed to better differentiate orthologous CRMs and orthologous neutral sequences.

A CRM can be modeled as a set of TFBSs separated by in-module background sequences (Zhou and Wong 2004). The embedded TFBSs are more conserved than in-module background sequences. Orthologous CRMs can perform a conserved function in spite of a class of structural changes (Waterston et al. 2002; Papatsenko and Levine 2005). For example, the following structural changes do not necessarily change the function of a CRM: insertion and deletion on background sequences between TFBSs, change of order of TFBSs, and change of number of TFBSs (Fig. 4A panels B–D)(for examples of these cases, see Waterston et al. 2002 and Papatsenko and Levine 2005). Smith-Waterman is capable of assigning a high conservation score to orthologous CRMs that have conserved TFBS composition, distance, order, and number (Fig. 4A panel A). All of the structural changes in Figure 4A, panels B–D, are penalized, often heavily, in Smith-Waterman. Nevertheless, these changes do not necessarily manifest functional changes, and therefore they may not be under negative selection. Penalizing these changes would suppress the distinction between the conservation of CRMs and that of neutral sequences (Supplemental Fig. S2).

Module-Alignment is designed to compute conservation scores based on the composition of putative TFBSs in CRMs, accommodating potential complex structural changes of CRMs

during evolution. The basic idea is to iteratively use Smith-Waterman to identify putative TFBSs and use the overall conservation level of all the putative TFBSs within a 200-bp window to compute the conservation score.[9] The window length and the scoring system for Smith-Waterman alignment are tuning parameters that can be adjusted. In this paper, all the analyses are run with the window length set to be 200 bp, which is in the range consistent with reported CRMs (Waterston et al. 2002; Zhou and Wong 2004; Gupta and Liu 2005; Papatsenko and Levine 2005). Module-Alignment takes two homologous noncoding sequences and a conservation threshold as input. The conservation threshold is used to determine whether a pair of conserved sequence segments can be regarded putative orthologous TFBSs or orthologous CRMs. The computation procedure is as follows:

(1) Initiation: Set Module-Alignment conservation score as 0.
(2) apply Smith-Waterman, and identify the best local alignment;
(3) compare the alignment score of best local alignment with the conservation threshold;
(4) if the alignment score is larger than the threshold, regard the current best local alignment as orthologous TFBSs or orthologous CRMs. Add the alignment score to Module-Alignment's conservation score. Mask the sequences in the currently aligned regions. Go back to Step (2);
(5) stop the algorithm when no best local alignment satisfies the conservation threshold (Fig. 4B; Supplemental Fig. S1).

Finally, the conservation threshold, similar to mismatch penalty and gap penalty in an alignment algorithm, is a tuning parameter. This tuning parameter can be tuned according to the expected lengths of the TFBSs.

Chromatin immunoprecipitation of POU5F1

Human ES cells (H1 line) were cultured on matrigel in MEF-conditioned media (DMEM/F-12 supplied with 20% KnockOut Serum Replacement [GIBCO], 2 mM L-glutamine, 1.1 mM 2-mercaptoethanol, 1 mM nonessential amino acids, and 8 ng/mL bFGF) (Thomson et al. 1998). A chromatin immunoprecipitation (ChIP) assay was carried out as described previously (Loh et al. 2006). Briefly, cells were cross-linked with 1% formaldehyde for 10 min at room temperature, and formaldehyde was then inactivated by the addition of 125 mM glycine. Chromatin extracts containing DNA fragments with an average size of 500 bp were immunoprecipitated using polyclonal POU5F1 antibody (Ab19857-100, Abcam). Quantitative PCR analyses were performed in real time using the ABI PRISM 7900 sequence detection system and SYBR green master mix. Threshold cycles (Ct) were determined for both immunoprecipitated DNA and a known amount of DNA from input sample for different primer pairs. Relative occupancy values (also known as fold enrichments) were calculated by determining the immunoprecipitation efficiency (ratios of the amount of immunoprecipitated DNA to that of the input sample) and were normalized to the level observed at a control region, which was defined as 1.0.

[9]Throughout this article we use 1, −0.33, −1, −0.33, and 10 as match score, mismatch penalty, gap open penalty, gap extension penalty, and conservation threshold, respectively. This means that we expect to see a 10-bp TFBS to be perfectly conserved, or a 14-bp TFBS to have no more than three mutations.

Results

GibbsModule identifies the locations of a core motif that is shared by a set of CRMs. In this sense, GibbsModule is more of a CRM prediction tool than a motif prediction tool, because the location of one participating TFBS in a CRM is enough to pinpoint the location of this CRM down to a 100-bp resolution. This is the same resolution that other CRM identification tools offer (Zhou and Wong 2004; Gupta and Liu 2005). We define the 200-bp regions centered at the predicted motif locations as GibbsModule-predicted CRMs. GibbsModule does not model and detect all the motifs that constitute its predicted CRMs. However, if the intention is to characterize all TFBSs in the output CRMs, it is a simple task to perform a motif search in the predicted CRMs. Because the predicted CRMs are 200-bp short sequences, motif searches, such as when using MEME (Bailey and Elkan 1994), usually give accurate results.

Synthetic data

We use three synthetic data sets to demonstrate three points: (1) GibbsModule can efficiently pinpoint CRMs; (2) applying MEME to GibbsModule-identified CRMs can identify their participating motifs; (3) directly applying MEME to the full-length sequences usually fails to identify the real motifs. We also applied other motif- and CRM-finding tools including CisModule (Zhou and Wong 2004), CompareProspector (Liu et al. 2004), and PhyloCon (Wang and Stormo 2003) on these synthetic data sets and compared their performances.

The three synthetic data sets are constructed as follows: First, homologous upstream sequences from 22 randomly chosen genes were retrieved from human, mouse, and chicken genomes. All the upstream sequences were retrieved from transcription start sites to 1000 bp upstream. These 22 homologous groups of sequences were used as background sequences in all three data sets. PSWMs of motifs for three transcription factors, POU5F1 (also known as OCT4), SOX2, and FOXD3 were retrieved from the TRANSFAC database (Matys et al. 2006). TFBSs of each transcription factor were generated by the product-multinomial distribution defined by its motif PSWM.

- In data set 1, three TFBSs of each transcription factor were inserted randomly into every background sequence in the first 20 homologous groups. The distance between any two TFBSs is generated from a Poisson distribution with an expected value of 10; i.e., the average distance between any two TFBSs is 10 bp. The order of the TFBSs is conserved across homologous sequences. The last two homologous groups do not contain any TFBSs;
- in data set 2, a random number of TFBSs for each transcription factor were inserted into every background sequence in the first 20 homologous groups. The number of TFBSs for a transcription factor is drawn from a Poisson distribution with an expected value of 1. The distance between any two TFBSs is drawn from a Poisson distribution with an expected value of 10. The order of the TFBSs is conserved across homologous sequences. The last two homologous groups do not contain any TFBSs;
- data set 3 is constructed in the same manner as data set 2 except that the order of the TFBSs is not conserved.

GibbsModule has detected all the CRMs in data set 1 and almost all the CRMs in data sets 2 and 3 with only one false

Table 1. Results on synthetic data sets

Synthetic data sets		1			2			3		
		Real[a]	TP[b]	FP[c]	Real	TP	FP	Real	TP	FP
CRM detection by GibbsModule (20 real CRMs in total)		**20**	**20**	**0**	**20**	**19**	**0**	**20**	**18**	**1**
TFBS detection by applying MEME to GibbsModule detected CRMs	SOX2	60	58	11	29	25	27	30	27	38
	POU5F1	60	37		37	25		31	0	
	FOXD3	60	18		32	0		31	10	
TFBS detection by applying MEME to the original sequences	SOX2	60	54	87	29	0	105	30	0	100
	POU5F1	60	0		37	0		31	0	
	FOXD3	60	0		32	0		31	0	
TFBS detection by PhyloCon	SOX2	60	2	15	29	11	7	30	13	1
	POU5F1	60	1		37	0		31	0	
	FOXD3	60	2		32	1		31	0	
TFBS detection by CompareProspector	SOX2	60	4	21	29	0	7	30	0	20
	POU5F1	60	1		37	0		31	1	
	FOXD3	60	1		32	0		31	1	
TFBS detection by CisModule	SOX2	60	0	270	29	0	207	30	0	217
	POU5F1	60	0		37	0		31	0	
	FOXD3	60	0		32	0		31	0	

[a]Number of real CRMs or TFBSs.
[b]True positive.
[c]False positive.

positive prediction (Table 1). When MEME is applied to GibbsModule-predicted CRMs, all three motifs are recovered from data set 1. POU5F1 and SOX2 motifs are recovered from data set 2. FOXD3 and SOX2 motifs are recovered from data set 3. When MEME is directly applied to the full-length input sequences, it correctly identifies only the SOX2 motif in data set 1, failing to detect the other two motifs. Additionally, MEME completely fails to correctly detect any motifs in data sets 2 and 3 (Table 1). Under default settings, PhyloCon has correctly recovered SOX2 motifs in all three data sets, while it misses almost all POU5F1 and FOXD3 binding sites. CompareProspector and CisModule had very few correct predictions in any of the three data sets. These synthetic data suggest that GibbsModule can successfully identify CRMs, and narrowing the search space down to GibbsModule-predicted CRMs can facilitate the efficiency of motif searches.

Muscle enhancer data

The transcription factors SP1, SRF, TEF and transcription factor families MEF2 and MYF are known to regulate gene expression in muscle cells (Wasserman and Fickett 1998). A set of *cis*-regulatory regions that are sufficient to control skeletal-muscle-specific expression has been experimentally localized to within 200 bp (Wasserman and Fickett 1998). This data set is used as a testing set to calibrate several motif and CRM detection tools (Zhou and Wong 2004, 2007; Gupta and Liu 2005). We extracted the same 20 enhancers as from Zhou and Wong (2004), within which there are 23 CRMs consisting of 15 MEF2, 25 MYF, 21 SP1, 13 SRF, and six TEF experimentally validated TFBSs. To mimic the real situation in which we do not know the locations of these enhancers, we extracted the complete upstream sequences from these enhancers to the transcription start sites of their target genes. The original enhancers were either in human or in mouse. Besides the complete upstream regions covering these enhancers, we also extracted their homologous upstream regions from either human or mouse, and from dog. Repeatmasker was applied to mask repeat regions (A.F.A. Smit, R. Hubley, and P. Green, Repeatmasker at http://www.repeatmasker.org).

GibbsModule predicted a total of 41 CRMs. Of those, 19 of them overlapped with the experimentally verified CRMs (Supplemental Fig. S3A). Therefore, there were 19 true positives and 22 false positives, which translates into a precision of 0.46 and a recall of 0.83 (Table 2). We also executed four other published programs for comparison. CisModule (Zhou and Wong 2004) and EMCModule (Gupta and Liu 2005) jointly utilize the shared TFBS composition in CRMs and the overrepresentation of motifs across genes. CompareProspecter (Liu et al. 2004) and PhyloCon (Wang and Stormo 2003) jointly utilize the overrepresentation and conservation properties of a motif. We applied our best knowledge and expertise to tune the parameters in CisModule, EMCModule, CompareProspecter, and PhyloCon to achieve their best performance. We gave CisModule a tremendous advantage in the test by counting its predicted CRM as a true positive as long as it is within 50 bp of a real CRM (counting from the nearest boundaries). We required GibbsModule-, EMCModule-, CompareProspecter-, and PhyloCon-predicted TFBSs to be contained within real CRMs to be counted as true positives. We also chose the result from CisModule either from one execution or from a cumulative summary of multiple executions (the same strategy as what we implemented in GibbsModule), whichever performed better.

It should be noted that these tools report data that are not completely comparable. GibbsModule, CisModule, and EMCModule predict CRMs (CRM predictors), while CompareProspector and PhyloCon predict motifs (motif predictors). There are two

Table 2. Performance comparison on muscle enhancers

Methods	TP[a]	FP[b]	Recall	Precision
GibbsModule	19	22	**0.83**	**0.46**
CisModule	19	31	0.83	0.38
CompareProspector	6	43	0.26	0.12
PhyloCon	9	16	0.39	0.36
EMCModule	17	46	0.74	0.27

The largest recall and precision values are in bold.
[a]True positive.
[b]False positive.

Table 3. Performance comparison on ESC enhancers

	2K		3K		4K		5K	
Sequence length	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
GibbsModule	**1.00**	**0.74**	**0.97**	**0.53**	**0.87**	**0.36**	**0.85**	**0.27**
CisModule	0.96	0.38	0.94	0.30	0.83	0.15	0.86	0.14
CompareProspector	0.29	0.18	0.06	0.02	0	0	0	0
PhyloCon	0.63	0.53	0.55	0.44	0.35	0.33	n/a	n/a
EMCModule	0.86	0.25	0.91	0.21	**0.87**	0.17	0.79	0.14

The largest recall and precision values are in bold.

strategies to make their results comparable. First, we can degenerate the CRM predictors as motif predictors by ignoring the reported CRMs and using only the individual reported motifs. Second, we can regard the motifs identified by motif-level tools as centers of CRMs, assuming they report CRMs that cover their reported motifs. We adopted the latter strategy because it offers a simple comparable summary to results from all algorithms.

GibbsModule outperformed the other four algorithms in this test. In particular, CompareProspector and PhyloCon have much smaller recalls than that of GibbsModule while their precisions are also smaller. Compared with these two algorithms, the performances of CisModule and EMCModule are closer to that of GibbsModule, probably because of their shared capability of modeling modular information. CisModule achieved a performance closest to GibbsModule. Given the same number of true positives, GibbsModule generates ~50% fewer false positives than CisModule (Table 2). To make sure that this conclusion is not biased by particular thresholds that we set for these algorithms, we performed a more detailed sensitivity analysis (Supplemental Fig. S4). With various thresholds (Supplemental Fig. S3B), GibbsModule has a consistent increase of ~0.1 in precision given the same recall as of CisModule. This is nontrivial because it translates into ~50% fewer false positives. Moreover, the number of true positives in CisModule plateaued at 19, irrespective of the increasing of threshold (vertical drop at recall = 0.83 in Supplemental Fig. S4). GibbsModule can at best detect 21 true positives out of 23 total CRMs (rightmost red dot in Supplemental Fig. S4). At this point, it generates only 32 false positives, giving a precision of 0.40, better than all CisModule results that have at least 13 true positives.

Enhancers regulating gene expression in embryonic stem cells

Embryonic stem cells (ESCs) are derived from early mammalian embryos and can be propagated through apparently unlimited, undifferentiated proliferation (self-renewal) in cultured cell lines (mouse: [Dehal et al. 2002; Evans and Kaufman 1981]; human: [Thomson et al. 1998]). ESCs are capable of differentiating into all derivatives of the three primary germ layers (known as "pluripotency"), and therefore they serve as a powerful in vitro system for studying the processes of differentiation and cell lineage determination. A major challenge in the study of ESCs is to explain how the complex genes network is "wired" to control their properties of pluripotency and self-renewal. Regulation of gene transcription is thought to be a key control mechanism for ESCs to maintain their undifferentiated state (Abeyta et al. 2004; Bhattacharya et al. 2004; Catena et al. 2004; Boyer et al. 2005; Golan-Mashiach et al. 2005; Rodda et al. 2005; Skottman et al. 2005; Bernstein et al. 2006; Lee et al. 2006; Venkatesh et al. 2006). Therefore, the identification of CRMs for ESCs is critical to reveal how the undifferentiated state of ESCs is maintained, and how it can be disrupted to initiate routes of differentiation (Boyer et al. 2005; Venkatesh et al. 2006).

Several key transcription factors are shown to be required for maintaining the pluripotent state of ESCs. They include POU5F1 (Niwa et al. 2000; Matin et al. 2004), SOX2 (Yuan et al. 1995; Kuroda et al. 2005; Rodda et al. 2005), NANONG (Mitsui et al. 2003; Ying et al. 2003), and others (Brandenberger et al. 2004; Ivanova et al. 2006). POU5F1 and SOX2 are known to dimerize and access binding sites that are 0–20 bp apart from each other to regulate ESC gene expression (Hosler et al. 1993; Yuan et al. 1995; Nishimoto et al. 1999; Kuroda et al. 2005; Rodda et al. 2005). In a genome-wide study, Boyer et al. (2005) have shown that the binding regions of NANONG also strongly colocalize with that of POU5F1 and SOX2 in ESCs. These data suggest that POU5F1, SOX2, and NANOG access the same CRMs in ESCs. ChIP-chip analyses revealed genomic distribution of binding regions of POU5F1, SOX2, and NANOG in human (Boyer et al. 2005) and mouse (Venkatesh et al. 2006).[10,11] When the binding regions of POU5F1, SOX2, and NANOG are within 100 bp of each other (counting the nucleotides between the nearest ends of the ChIP-chip-positive regions), we merge these ChIP-chip-positive regions into one big island. Such an island is supposed to contain a CRM. In total, 16 orthologous genes that contain POU5F1, SOX2, and NANOG binding CRMs in both species are identified. This set of 16 pairs of CRM-containing islands form our second biological data set for algorithm assessment. Genomic locations of these islands are listed in Supplemental Table S1. The lengths of these CRM-containing islands range from 112 to 2170 bp.

To mimic real CRM discovery scenarios, we first extended these islands into 2K- (keeping the few sequences that are longer than 2K), 3K-, 4K-, and 5K-bp-long sequences by incorporating their genomic neighboring sequences. Thus, we obtained four test data sets: The first data set contains 2K-long sequences; the second data set contains 3K-long sequences, and so forth. We ran GibbsModule, CisModule, EMCMoudule, CompareProspector, and PhyloCon on each data set. We determined a predicted TFBS to be true positive if it locates within a ChIP-positive island and within 75 bp on either side there is at least one putative TFBS that matches a PSWM of POU5F1, SOX2, or NANOG. The precisions and recalls are summarized in Table 3. GibbsModule again outperformed the other algorithms in all the data sets. CisModule seems to have a closer match to GibbsModule than the other three algorithms. It nevertheless achieves only about half of the

[10]The mouse data are obtained with a ChIP-PET technology, which is similar to ChIP-chip, but instead of hybridizing the ChIP sequences onto a microarray, it uses a sequencing technology to count the immunoprecipitated sequences. In this article, we use the term ChIP-chip to represent both ChIP-chip and ChIP-PET data.

[11]SOX2 ChIP–PET sequences in murine ESCs (H.H. Ng, unpubl.).

precision of GibbsModule, while its recall is also slightly worse than that of GibbsModule. This is not a trivial distinction because it means that given the same number of true positives, GibbsModule reduces >60% of false positives (Supplemental Table S2).

Identification of a Klf motif and evidence for KLF4–SOX2 cooperation in ESCs

KLF2, KLF4, and KLF5 belong to the Krüppel-like factor (KLF) family of evolutionarily conserved zinc finger transcription factors that regulate numerous biological processes, including proliferation, differentiation, development, and apoptosis (McConnell et al. 2007). We have recently demonstrated that Krüppel-like factors are required for the self-renewal of mouse ESCs. Simultaneous depletion of KLF2, KLF4, and KLF5 led to mouse ESC differentiation. KLF2, KLF4, and KLF5 bind to the same binding sites in vitro and in vivo (Jiang et al. 2008). A PSWM for KLF4 was previously derived from SELEX experiment (Shields and Yang 1998). This PSWM, however, has a questionable credibility because its AAAGGAAGG consensus is not consistent with the consensus CCCCACCC, derived by our site-directed mutagenesis analysis (see Fig. 3B in Jiang et al. 2008). To investigate the Klf motif that works in vivo, we analyzed KLF-genomic-binding regions that are identified by ChIP-chip (Jiang et al. 2008). Out of the 205 KLF-binding regions in the mouse genome, we identified 119 homologous loci in the human genome using GenomeVISTA (Couronne et al. 2003). The set of 119 homologous pairs were fed to GibbsModule. GibbsModule-reported CRMs were then subsequently fed to MEME (see Supplemental Table S3 for parameters). Two motifs were reported from this analysis (Fig. 5). One resembles the consensus from mutagenesis analysis (Fig. 5A; Fig. 3B in Jiang et al. 2008) (notice that the reverse complement of GGGT/AGGGG is CCCCA/TCCC), and the other resembles the SOX2 motif from TRANSFAC (Fig. 5A,C). Therefore, the Klf motif derived from ChIP-chip data is consistent with the consensus identified by the mutagenesis analysis, which might be more useful for future in silico analysis than the SELEX motif. Moreover, the identification of a SOX2 motif in the same CRMs as the Klf motif seems to suggest a general cooperation phenomenon between these two transcription factors in ESCs. This is consistent with a recently reported case that KLF4 cooperates with POU5F1 and SOX2 to activate the *Lefty1* core promoter in ESCs (Nakatake et al. 2006).

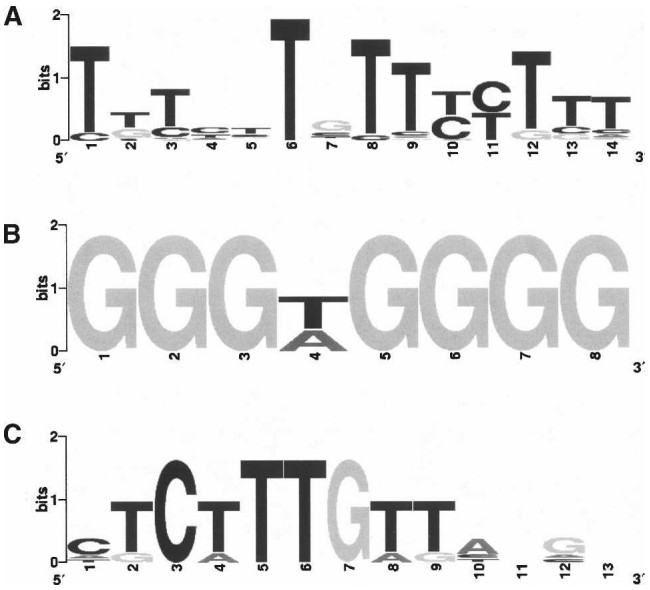


Figure 5. Motifs derived from Klf ChIP-chip and mutagenesis analysis. SOX2 motif (*A*) and KLF motif (*B*) found by GibbsModule and MEME from KLF ChIP-chip data. (*C*) SOX2 motif in TRANSFAC.

Predicting and experimental testing of the binding affinities of in-CRM and outside-CRM binding sites

Although CRMs are often reported in eukaryotes, to date there are few quantitative analyses on the role of CRMs in activation of repression of the downstream genes. In this regard, here we attempt to explore a few fundamental questions. Does residing within a CRM confer better binding affinity of a TFBS to its transcription factor as compared with another TFBS not located in a CRM? Do other motifs in the CRM contribute more, or do neighboring TFBSs of the same kind help more to attract a transcription factor? Although we do not expect a generic answer to these questions, well-deliberated quantitative analyses could provide useful empirical data to study the relationship between regulatory code and gene expression.

We attempted to apply GibbsModule to define the sequences bound by POU5F1 within POU5F1-binding regions that are obtained from ChIP-chip analysis (Boyer et al. 2005). This is an ambitious and experimentally challenging attempt because all putative POU5F1 binding sites in this analysis are located within ChIP-chip peak regions (see Discussion). The ChIP-chip peak regions are typically 1 kb long, which is more than twice as long as most characterized CRMs. The basic idea is to quantitatively compare the binding affinities of POU5F1-binding sites within predicted CRMs and that of other POU5F1-binding sites outside predicted CRMs. We designed a comparative ChIP experiment for this purpose. The ChIP signal is proportional to the precipitated DNA bound to the transcription factor, which is proportional to the time of binding between one transcription factor and its target site in a cell in the equilibrium state, and therefore is proportional to the binding affinity of the TFBS when the cellular condition is unchanged. Hence, the difference of binding affinities (as measured by fold enrichment) between two TFBSs is reflected by the difference in their ChIP signals, given that the ChIP experiments are performed in the same cell population for the same transcription factor, and other proper controls. From the results in the section above, "Enhancers regulating gene expression in embryonic stem cells," we picked seven predicted CRMs that contain putative POU5F1-binding sites (predicted positives) and seven strong putative POU5F1-binding sites that are not in predicted CRMs (predicted negatives) (Supplemental Fig. S5; Supplemental Table S4). Primers are designed to cover ~100-bp regions centered by these 14 putative binding sites. ChIP analysis was performed on each predicted site with three biological replicates (see Methods; Fig. 6; and Supplemental Table S5). To control for other unobserved DNA features, such as chromatin structure and DNA methylation states, we picked five pairs of predicted positive and predicted negative TFBSs from matched ChIP-chip peak regions. Except for the negative prediction on *IRX2*, on which our PCR failed, all five pairs of ChIP analysis had more precipitated DNA bound to the predicted CRMs (*GSH2* [also known as *GSX2*], *RIF1*, *SALL1*, *Lefty1*, and *IRX2* in Fig. 6). The same difference was observed in the four remaining putative binding sites, where at least a twofold increase in ChIP signals was observed for in-CRM putative POU5F1 sites (*CA2*, *EOMES*,

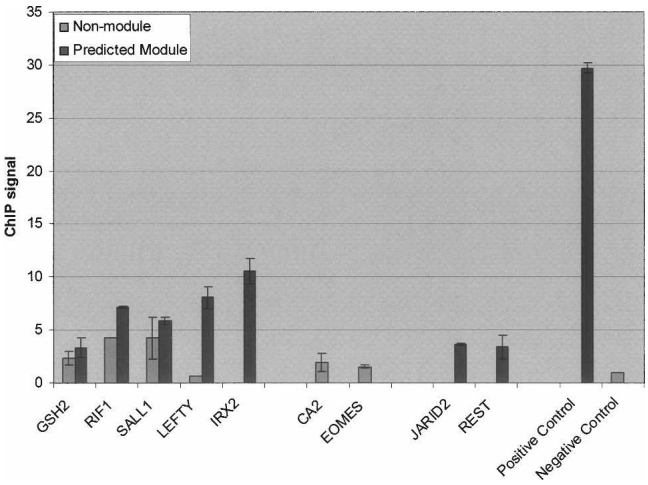


Figure 6. ChIP signals of predicted CRM and non-CRMs that contain putative POU5F1-binding sites.

JARID2, and *REST* in Fig. 6). To test whether the observed differences are due to different distribution of the POU5F1 binding sites themselves in the predicted regions, we adjusted the ChIP signal by the number of putative binding sites in each predicted region. A clear difference between the positive and negative predictions was again observed in the adjusted ChIP signals (Supplemental Fig. S6). Similar differences were also observed when we change the TFBS threshold to either include or eliminate weak POU5F1 binding sites in the analysis (result not shown; see Supplemental Table S5 for the number of POU5F1 biding sites under different thresholds.) These data suggest that GibbsModule identified functional POU5F1 binding sites from ChIP-chip regions and, more importantly, distinguished them from other putative POU5F1 sites.

Discussion

Classical de novo TFBS identification tools typically work well when TFBSs locate in promoter regions close to the transcription start sites, which is often the case in compact genomes such as that of *Saccharomyces cerevisiae*. Higher eukaryote genomes impose a greater computational challenge because TFBSs can locate in enhancers, far away from the target genes. Thus, the algorithms are required to search in a much larger sequence space. With long input sequences, classical TFBS identification tools are typically easily trapped in local maxima and output-only degenerated motifs.

GibbsModule is a multispecies extension of a Gibbs motif sampler that utilizes the conservation of CRMs to help de novo motif discovery. We would like to point out a few important properties of GibbsModule.

First, GibbsModule does not rely on a predetermined alignment result. It has been reported that orthologous CRMs often do not locate within "conserved regions" detected by sequence alignment (Sosinsky et al. 2007). GibbsModule tries to identify orthologous CRMs in its iterations. The sampled TFBSs (Step 2, Fig. 3) serve as "anchor points" to find orthologous CRMs. If there are orthologous CRMs, they likely contain the real TFBSs. If some of the candidates are real, searching for orthologous CRMs around the candidate TFBSs would substantially reduce the search space compared with using local alignment on the whole homologous sequences. Rather, we focus on the neighboring sequences of these candidate TFBSs. Because any candidate TFBS on the target species can be orthologous to any candidate TFBS on a homologous sequence, Module-Alignment needs to be executed for all pairs of candidate TFBSs on homologous sequences between the target species and every assisting species. It nevertheless reduces the total amount of allowed alignments substantially as compared with searching in several thousands of base pairs of upstream sequences for conservation. We call this strategy "anchored search of alignments." We would like to make an analogy between anchored search of alignments and the LAGAN algorithm (Brudno et al. 2003). LAGAN first identifies the most reliably alignable regions. It uses the reliably aligned regions as anchors and then fills in the alignments in not-so-certain regions. The distinction between the anchored search strategy used in GibbsModule and LAGAN is that GibbsModule uses putative TFBSs from motif sampling as anchors.

Second, GibbsModule does not require all real TFBSs to locate within conserved CRMs. When a gene is regulated by a single TFBS, sampling a TFBS from Steps 3 and 4 (Fig. 3) on this gene would be similar to directly sampling a TFBS on the sequence of the target species. Thus, GibbsModule degenerates into a classical Gibbs motif sampler on the genes that are regulated by a single TFBS rather than by a CRM. A major difference between GibbsModule and other multispecies motif detection methods (Wang and Stormo 2003; Liu et al. 2004; Sinha et al. 2004; Siddharthan et al. 2005) is that GibbsModule uses not only the conservation of the putative TFBSs themselves, but the conservation of the neighboring sequences as well.

Third, GibbsModule does not require the CRMs to be shared across coexpressed genes except for the core motif itself. This assumption is very relaxed as compared with CisModule (Zhou and Wong 2004) and EMCModule (Gupta and Liu 2005), which require the module composition to be shared across coexpressed genes. We believe GibbsModule's assumption is more appropriate for the analysis of coexpressed genes determined by microarray data. Microarray data are often noisy, and only a finite number of samples are measured; coexpression of genes may not necessarily imply a very strong co-regulatory mechanism as exemplified by sharing whole CRMs. Another important distinction between GibbsModule and other attempts to identify CRMs (Zhou and Wong 2004; Gupta and Liu 2005; Hallikas et al. 2006; Sinha et al. 2006) is that GibbsModule models and traces only one core motif in a CRM. Pinpointing the location of this core motif in the genome is equivalent to pinpointing the location of a CRM. We think the better performance of GibbsModule as compared with the other algorithms can be primarily attributed to its largely reduced model complexity. In sum, it seems that modeling one rather than multiple motifs relies on a more flexible assumption, reduces model and computation complexity, and generates more accurate results.

For the comparative ChIP analysis, it should be noted that the average length of chromatin DNA for our ChIP assay is 500 bp. Even primers for a nonfunctional POU5F1 site would inevitably get some positive ChIP signal in our experiment, simply because these primers are not too far away (typically 300 bp) from the functional POU5F1 site. Nevertheless, our experimental design appears to have worked, and the data appear to be informative. This result can be appreciated by considering the following scenario. Suppose all precipitated DNA segments cover the functional TFBS and some flanking regions with random lengths. Almost all the precipitated segments could be covered by the

positive primers, but only a fraction of them are possibly covered by the primers for the nonfunctional site. Although the primers for the predicted negatives are inevitably contaminated by some flanking regions, the functional and nonfunctional sites still give a quantitative difference in this comparative ChIP experiment.

A future direction for GibbsModule is to incorporate phylogenetic distance into its model. The use of phylogenetic distance has led to quite a few successful applications (Margulies et al. 2003; Siddharthan et al. 2005); however, it has been difficult to incorporate phylogenetic distance into models that simultaneously model all motifs in CRMs. GibbsModule opens this possibility because it models only the core motif, although it uses the conservation of CRMs. We plan to incorporate phylogenetic distance into the Module-Alignment step and utilize it to improve the conservation score.

Acknowledgments

We thank Dr. H. Rex Gaskins for numerous helpful discussions. We thank David Unger, Dr. Moushumi Sen Sarma, and Feng Hong for proofreading the paper and for helpful suggestions. This work is partially supported by the National Center for Supercomputing Applications and the Illinois Regenerative Medicine Institute.

References

Abeyta, M.J., Clark, A.T., Rodriguez, R.T., Bodnar, M.S., Pera, R.A., and Firpo, M.T. 2004. Unique gene expression signatures of independently derived human embryonic stem cell lines. *Hum. Mol. Genet.* **13:** 601–608.

Bailey, T.L. and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2:** 28–36.

Banerji, J., Rusconi, S., and Schaffner, W. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27:** 299–308.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. 2004. Ultraconserved elements in the human genome. *Science* **304:** 1321–1325.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125:** 315–326.

Bhattacharya, B., Miura, T., Brandenberger, R., Mejido, J., Luo, Y., Yang, A.X., Joshi, B.H., Ginis, I., Thies, R.S., Amit, M., et al. 2004. Gene expression in human embryonic stem cell lines: Unique molecular signature. *Blood* **103:** 2956–2964.

Blanchette, M. and Tompa, M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* **12:** 739–748.

Blanchette, M. and Tompa, M. 2003. FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res.* **31:** 3840–3842.

Blanchette, M., Schwikowski, B., and Tompa, M. 2002. Algorithms for phylogenetic footprinting. *J. Comput. Biol.* **9:** 211–223.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., and Rubin, E.M. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299:** 1391–1394.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., et al. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122:** 947–956.

Brandenberger, R., Wei, H., Zhang, S., Lei, S., Murage, J., Fisk, G.J., Li, Y., Xu, C., Fang, R., Guegler, K., et al. 2004. Transcriptome characterization elucidates signaling networks that control human ES cell growth and differentiation. *Nat. Biotechnol.* **22:** 707–716.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., and Batzoglou, S. 2003. LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13:** 721–731.

Carter, D., Chakalova, L., Osborne, C.S., Dai, Y.F., and Fraser, P. 2002. Long-range chromatin regulatory interactions in vivo. *Nat. Genet.* **32:** 623–626.

Catena, R., Tiveron, C., Ronchi, A., Porta, S., Ferri, A., Tatangelo, L., Cavallaro, M., Favaro, R., Ottolenghi, S., Reinbold, R., et al. 2004. Conserved POU binding DNA sites in the Sox2 upstream enhancer regulate gene expression in embryonic and neural stem cells. *J. Biol. Chem.* **279:** 41846–41857.

Chen, C.T., Wang, J.C., and Cohen, B.A. 2007. The strength of selection on ultraconserved elements in the human genome. *Am. J. Hum. Genet.* **80:** 692–704.

Couronne, O., Poliakov, A., Bray, N., Ishkhanov, T., Ryaboy, D., Rubin, E., Pachter, L., and Dubchak, I. 2003. Strategies and tools for whole-genome alignments. *Genome Res.* **13:** 73–80.

Davidson, E. 2006. *The regulatory genome. Gene regulatory networks in development and evolution*. Academic Press/Elsevier, San Diego, CA.

Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., et al. 2002. A genomic regulatory network for development. *Science* **295:** 1669–1678.

Davidson, E.H., McClay, D.R., and Hood, L. 2003. Regulatory gene networks and the properties of the developmental process. *Proc. Natl. Acad. Sci.* **100:** 1475–1480.

Dehal, P., Satou, Y., Campbell, R.K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D.M., et al. 2002. The draft genome of *Ciona intestinalis*: Insights into chordate and vertebrate origins. *Science* **298:** 2157–2167.

Evans, M.J. and Kaufman, M.H. 1981. Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292:** 154–156.

Golan-Mashiach, M., Dazard, J.E., Gerecht-Nir, S., Amariglio, N., Fisher, T., Jacob-Hirsch, J., Bielorai, B., Osenberg, S., Barad, O., Getz, G., et al. 2005. Design principle of gene expression used by human stem cells: Implication for pluripotency. *FASEB J.* **19:** 147–149.

Gupta, M. and Liu, J.S. 2005. De novo *cis*-regulatory module elicitation for eukaryotic genomes. *Proc. Natl. Acad. Sci.* **102:** 7079–7084.

Hallikas, O., Palin, K., Sinjushina, N., Rautiainen, R., Partanen, J., Ukkonen, E., and Taipale, J. 2006. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124:** 47–59.

Hardison, R.C., Roskin, K.M., Yang, S., Diekhans, M., Kent, W.J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13:** 13–26.

Hosler, B.A., Rogers, M.B., Kozak, C.A., and Gudas, L.J. 1993. An octamer motif contributes to the expression of the retinoic acid-regulated zinc finger gene Rex-1 (Zfp-42) in F9 teratocarcinoma cells. *Mol. Cell. Biol.* **13:** 2919–2928.

Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M. 2000. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **296:** 1205–1214.

Ivanova, N., Dobrin, R., Lu, R., Kotenko, I., Levorse, J., Decoste, C., Schafer, X., Lun, Y., and Lemischka, I.R. 2006. Dissecting self-renewal in stem cells with RNA interference. *Nature* **442:** 533–538.

Jiang, J., Chan, Y.S., Loh, Y.H., Cai, J., Tong, G.Q., Lim, C.A., Robson, P., Zhong, S., and Ng, H.H. 2008. A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat. Cell Biol.* **10:** 353–360.

Katzman, S., Kern, A.D., Bejerano, G., Fewell, G., Fulton, L., Wilson, R.K., Salama, S.R., and Haussler, D. 2007. Human genome ultraconserved elements are ultraselected. *Science* **317:** 915. doi: 10.1126/science.1142430.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. 2007. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128:** 1231–1245.

Kuroda, T., Tada, M., Kubota, H., Kimura, H., Hatano, S.Y., Suemori, H., Nakatsuji, N., and Tada, T. 2005. Octamer and Sox elements are required for transcriptional *cis* regulation of Nanog gene expression. *Mol. Cell. Biol.* **25:** 2475–2485.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C. 1993. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* **262:** 208–214.

Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isono, K., et al. 2006. Control of developmental regulators by polycomb in human embryonic stem cells. *Cell* **125:** 301–313.

Liu, X., Brutlag, D.L., and Liu, J.S. 2001. BioProspector: Discovering

conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.* 127–138.
Liu, Y., Liu, X.S., Wei, L., Altman, R.B., and Batzoglou, S. 2004. Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res.* **14:** 451–458.
Loh, Y.H., Wu, Q., Chew, J.L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., et al. 2006. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* **38:** 431–440.
Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136–140.
Margulies, E.H., Blanchette, M., Haussler, D., and Green, E.D. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13:** 2507–2518.
Matin, M.M., Walsh, J.R., Gokhale, P.J., Draper, J.S., Bahrami, A.R., Morton, I., Moore, H.D., and Andrews, P.W. 2004. Specific knockdown of Oct4 and beta2-microglobulin expression by RNA interference in human embryonic stem cells and embryonic carcinoma cells. *Stem Cells* **22:** 659–668.
Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. 2006. TRANSFAC and its module TRANSCompel: Transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34:** D108–D110. doi: 10.1093/nar/gkj143.
McConnell, B.B., Ghaleb, A.M., Nandan, M.O., and Yang, V.W. 2007. The diverse functions of Kruppel-like factors 4 and 5 in epithelial biology and pathobiology. *Bioessays* **29:** 549–557.
Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S. 2003. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113:** 631–642.
Moses, A.M., Chiang, D.Y., and Eisen, M.B. 2004. Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac. Symp. Biocomput.* 324–335.
Nakatake, Y., Fukui, N., Iwamatsu, Y., Masui, S., Takahashi, K., Yagi, R., Yagi, K., Miyazaki, J., Matoba, R., Ko, M.S., et al. 2006. Klf4 cooperates with Oct3/4 and Sox2 to activate the Lefty1 core promoter in embryonic stem cells. *Mol. Cell. Biol.* **26:** 7772–7782.
Nishimoto, M., Fukushima, A., Okuda, A., and Muramatsu, M. 1999. The gene for the embryonic stem cell coactivator UTF1 carries a regulatory element which selectively interacts with a complex composed of Oct-3/4 and Sox-2. *Mol. Cell. Biol.* **19:** 5453–5465.
Niwa, H., Miyazaki, J., and Smith, A.G. 2000. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat. Genet.* **24:** 372–376.
Ovcharenko, I., Boffelli, D., and Loots, G.G. 2004. eShadow: A tool for comparing closely related sequences. *Genome Res.* **14:** 1191–1198.
Papatsenko, D. and Levine, M. 2005. Quantitative analysis of binding motifs mediating diverse spatial readouts of the Dorsal gradient in the *Drosophila* embryo. *Proc. Natl. Acad. Sci.* **102:** 4966–4971.
Prakash, A. and Tompa, M. 2005. Discovery of regulatory elements in vertebrates through comparative genomics. *Nat. Biotechnol.* **23:** 1249–1256.
Rodda, D.J., Chew, J.L., Lim, L.H., Loh, Y.H., Wang, B., Ng, H.H., and Robson, P. 2005. Transcriptional regulation of nanog by OCT4 and SOX2. *J. Biol. Chem.* **280:** 24731–24737.
Shields, J.M. and Yang, V.W. 1998. Identification of the DNA sequence that interacts with the gut-enriched Kruppel-like factor. *Nucleic Acids Res.* **26:** 796–802.
Siddharthan, R., Siggia, E.D., and van Nimwegen, E. 2005. PhyloGibbs: A Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.* **1:** e67. doi: 10.1371/journal.pcbi.0010067.
Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.
Sinha, S. and Tompa, M. 2002. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.* **30:** 5549–5560.
Sinha, S., Blanchette, M., and Tompa, M. 2004. PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* **5:** 170. doi: 10.1186/1471-2105-5-170.
Sinha, S., Liang, Y., and Siggia, E. 2006. Stubb: A program for discovery and analysis of *cis*-regulatory modules. *Nucleic Acids Res.* **34:** W555–W559. doi: 10.1093/nar/gkl224.
Skottman, H., Mikkola, M., Lundin, K., Olsson, C., Stromberg, A.M., Tuuri, T., Otonkoski, T., Hovatta, O., and Lahesmaa, R. 2005. Gene expression signatures of seven individual human embryonic stem cell lines. *Stem Cells* **23:** 1343–1356.
Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.
Sosinsky, A., Honig, B., Mann, R.S., and Califano, A. 2007. Discovering transcriptional regulatory regions in *Drosophila* by a nonalignment method for phylogenetic footprinting. *Proc. Natl. Acad. Sci.* **104:** 6305–6310.
Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788–793.
Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones, J.M. 1998. Embryonic stem cell lines derived from human blastocysts. *Science* **282:** 1145–1147.
Venkatesh, B., Kirkness, E.F., Loh, Y.H., Halpern, A.L., Lee, A.P., Johnson, J., Dandona, N., Viswanathan, L.D., Tay, A., Venter, J.C., et al. 2006. Ancient noncoding elements conserved in the human genome. *Science* **314:** 1892. doi: 10.1126/science.1130708.
Vokes, S.A., Ji, H., McCuine, S., Tenzen, T., Giles, S., Zhong, S., Longabaugh, W.J., Davidson, E.H., Wong, W.H., and McMahon, A.P. 2007. Genomic characterization of Gli-activator targets in sonic hedgehog-mediated neural patterning. *Development* **134:** 1977–1989.
Wang, T. and Stormo, G.D. 2003. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* **19:** 2369–2380.
Wasserman, W.W. and Fickett, J.W. 1998. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **278:** 167–181.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3:** e7. doi: 10.1371/journal.pbio.0030007.
Ying, Q.L., Nichols, J., Chambers, I., and Smith, A. 2003. BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. *Cell* **115:** 281–292.
Yuan, H., Corbi, N., Basilico, C., and Dailey, L. 1995. Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3. *Genes & Dev.* **9:** 2635–2645.
Zhou, Q. and Wong, W.H. 2004. CisModule: De novo discovery of *cis*-regulatory modules by hierarchical mixture modeling. *Proc. Natl. Acad. Sci.* **101:** 12114–12119.
Zhou, Q. and Wong, W.H. 2007. Coupling hidden Markov models for the discovery of *cis*-regulatory modules in multiple species. *Ann. Appl. Stat.* **1:** 36–65.

Received October 31, 2007; accepted in revised form May 5, 2008.