Supplemenatry Material for Improving MEME via a two-tiered significance analysis

Emi Tanaka, Timothy Bailey, and Uri Keich

1 Methods

1.1 Motif Scores

1.1.1 PWM and site PWM score

MEME uses a PWM to model a motif. This PWM model is estimated from a positional frequency of the residues in the aligned identified sites. That is for a motif of width W, the PWM is

$$\begin{pmatrix} f_{1,a} & f_{2,a} & \cdots & f_{W,a} \\ \vdots & \vdots & \ddots & \\ f_{1,z} & f_{2,z} & \cdots & f_{W,z} \end{pmatrix}$$

where $f_{i,j}$ is the frequency at position i for $1 \leq i \leq W$ and $j \in A$ the set of residues. The site PWM score for a subsequence of W letters $a_1 a_2 ... a_W$ is defined as

$$S(a_1 a_2 ... a_W) = \sum_{i=1}^{W} \log \left(\frac{f_{i,a_i}}{f_{0,a_i}} \right)$$

where $f_{0,a}$ is the background frequency of letter a. This assumes a 0-order Markov or iid (independent identically distributed) model for the background, which is what was used in this paper, but this can be easily extended to any order of Markov model. We used our site scanning program SADMAMA (Keich *et al.*, 2008) to find the best site score in each sequence.

1.1.2 MEME's E-value

The E-value of a motif is the expected number of random alignments with the same dimension and with equal or higher log-likelihood ratio (llr) given the sequences have been generated randomly according to an iid background model.

The information content / relative entropy / llr is defined as

$$\sum_{i=1}^{W} \sum_{j \in A} n_{ij} \log \left(\frac{f_{i,j}}{f_{0,j}} \right), \tag{1}$$

where n_{ij} is the number of occurrences of the j-th letter in the i-th column of the alignment, and $f_{0,a}$ is again the background frequency of letter a in our iid model.

An efficient exact calculation of the E-value is described in (Nagarajan *et al.*, 2005) however MEME's reported E-value is calculated differently: "The E-value reported by MEME is actually an approximation of the E-value of the log likelihood ratio. (An approximation is used because it is far more efficient to compute.) The approximation is based on the fact that the log likelihood ratio of a motif is the sum of the log likelihood ratios of each column of the motif. Instead of computing the statistical significance of this sum (its p-value), MEME computes the p-value of each column and then computes the significance of their product" (MEME's documentation).

Note that computing MEME's approximation of the E-value has a runtime complexity which is cubic in the number of sequences.

1.1.3 Genomic background set

All references in this work to the genomic background file are to the October 2003 version of the S288C strain of *S. cerevisiae* genome which was downloaded from SGD and further processed by removing all sequences with feature type 'gene'.

1.1.4 Null set generation

For all discriminative scores as well as for estimating the 3-Gamma p-value we generate a null set of sequences in the same manner as Ng and Keich (2008b). We bin the regions of the background set by A-T composition. We sample from the bins such that the dimension (length and number of sequences) of the null set of sequences is the same as that of the input set of sequences and the local A-T composition are similar.

1.1.5 Mann-Whitney score (MW)

In OOPS mode we use the PWM reported by MEME, to find the best PWM site score for each input sequence as well as for each null sequence. We then apply the standard MW test to these two sets of scores allowing for ties. Our MW score is the p-value of this MW test.

In ZOOPS mode we use the same method as above except we restrict our attention to the top n best site scores in the input set of sequences and in the null set of sequences where n is the number of occurrences of the motif in the input set as reported by MEME.

1.1.6 Thresholded Mann-Whitney score (tMW)

For the thresholded MW score, we consider only site scores which exceed the threshold. In our experiments we set the threshold as the top 0.1% site scores from all site scores in the genomic background set (Section 1.1.3). The thresholded MW score is then the same as the MW score except that it ignores sequences with site scores less than the threshold. When in ZOOPS mode the last procedure applies only to the top n best sites of each set where again n is the number of sites in the input set as reported by MEME.

1.1.7 Fisher's Exact Test score

Since the Fisher exact test relies on counts we need to set a site defining threshold. We chose this threshold in the same manner as in tMW above by using the 0.999 quantile of the observed sites scores in the background set. Let M denote the number of sequences in the input set, and let m (m_0) denote the number of input (null) sequences that contain a site scoring at or above the site-defining threshold. The Fisher Exact Test score is then the probability of observing at least m scores that exceed the threshold assuming each such a score is equally likely to come from the input set and from the null set. That is,

$$\text{Fisher}_{\text{score}} = \sum_{k=m}^{\min(M,m+m_0)} \frac{\binom{M}{k}\binom{M}{m+m_0-k}}{\binom{2M}{m+m_0}}.$$

In ZOOPS mode we again use only the top n best site scores from each set.

1.1.8 Minimal Hyper-geometric score (MHG)

Let S_k (S_{0k}) be the best site score of the kth sequence in the input (null) set. Let $m(t)$ ($m_0(t)$) be the number of input (null) set of sequences whose best site score is higher than or equal to t, that is $m(t) = \sum_{k=1}^{M} I(S_k \geq t)$ and $m_0(t) = \sum_{k=1}^{M} I(S_{0k} \geq t)$.

The minimal hyper-geometric score is then the minimum over all t of the p-value of the Fisher Exact test applied to $(M, m(t), m_0(t))$.

$$\text{MHG}_{\text{score}} = \min_t \sum_{k=m(t)}^{\min(M,m(t)+m_0(t))} \frac{\binom{M}{k}\binom{M}{m(t)+m_0(t)-k}}{\binom{2M}{m(t)+m_0(t)}}.$$

Note that the term MHG was introduced in Eden *et al.* (2007), Steinfeld *et al.* (2008) and Eden *et al.* (2009) in a more general setting. Namely, given a binary vector we can perform a hypergeometric test to assess the significance of the number of "1"s in the first k entries of the vector. The minimal hypergeometric score is the minimal p-value over all hypergeometric tests, one for each possible value of k. As such, our definition of the MHG score is a special case of this definition: our binary vector is a union of the input (label "1") and null (label "0") sequences and it is sorted according to the score of the optimal motif match in each sequence.

1.1.9 Sign Score

Using the same notation as in Section 1.1.8 for S_k and S_{0k}, let $n = \sum_{k=1}^{M} I(S_k \neq S_{0k})$ and $x = \sum_{k=1}^{M} I(S_k > S_{0k})$. The sign score is then the probability of observing at least x input sequence best site scores larger than the corresponding best null site scores assuming again the scores are randomly assigned to both groups. Note that we disregard sequences for which $S_k = S_{0k}$, therefore:

$$\text{Sign}_{\text{score}} = \sum_{k=x}^{n} \binom{n}{k} 0.5^n.$$

1.1.10 Minimal Sign Score

The minimal sign score is defined analogously to the MHG score: we seek the threshold that minimizes the sign score when applied only to pairs of input-null sequence scores for which at least one of these two scores exceeds the threshold. Specifically, with

$n(t) = \sum_{k=1}^{M} I(S_k \neq S_{0k}) I(\max\{S_k, S_{0k}\} \geq t)$ and $x(t) = \sum_{k=1}^{M} I(S_k > \max\{S_{k0}, t\})$ we define

$$\mathrm{MSign}_{\mathrm{score}} = \min_t \sum_{k=x(t)}^{n(t)} \binom{n(t)}{k} 0.5^{n(t)}.$$

1.1.11 Selective Scores

We can modify any of the above discriminative scores by replacing columns of the PWM with the smallest entropy score with a rigid gap. The entropy or information content of column i of the PWM is defined as $2 + \sum_{j \in A} f_{i,j} \log_2 (f_{i,j})$. The number of columns we select (i.e., that are *not* replaced) is given by $6 + \lfloor \frac{w-6}{3} \rfloor$. The selective versions of the above discriminative scores are then computed as the originally defined version, except that each site score is computed with this gapped PWM.

1.1.12 3-Gamma scores

While we mostly refer to the 3-Gamma p-values in the context of significance assessment, in Section 1.3 of the main paper we look at the performance of the 3-Gamma p-value when used as a motif score to select the optimal motif among several candidates. In this context the 3-Gamma p-value is a point estimate of the p-value of the relative entropy / llr reported by MEME. This estimate is based on the assumption that the null distribution of the llr reported by MEME is well approximated by a 3-Gamma distribution. The latter is not a theoretical result, rather it is one based on extensive simulations (e.g., Figure 1b in Keich and Ng, 2007).

To compute this 3-Gamma p-value we first sample using our null sequence generator n (we use $n = 50$) null sets of sequences of the same dimensions as the input set (Section 1.1.4). MEME is then applied to each of these sets in the exact same way it was applied to the original input set yielding a sample of n null scores. Using these null scores we find the maximum likelihood estimator (MLE) $\hat{\theta} = (\hat{a}, \hat{b}, \hat{\mu})$. The MLE is then plugged into the distribution function when reporting the 3-Gamma p-value of the observed score s: $1 - F_{\hat{\theta}}(s)$, where F_θ is the 3-Gamma distribution function parametrized by θ.

1.2 Null distribution of motif scores by width

To study how the null distribution of each motif score varies with the motif width we constructed histograms of 10,000 empirical null motif scores for each combination of score function and motif width. Specifically, we generated 10,000 null sets of a fixed dimensions: 16 sequences of varying lengths. The first of those sets was generated by random sampling of the *S. cerevisiae* intergenic regions while the remaining sets were sampled from the background set such that local A-T composition is preserved (relative to the first set) as described in Section 1.1.4. MEME was then applied eight times in OOPS mode to each of those null sets with each run specifying a different motif width from 6 to 13 inclusive. Each of these eight MEME reported motifs was then scored by each of the scores described in Section 1.1. The boxplots of the observed scores for each width are shown in Figure 1 in the main paper, and Figures 1, 2 and 3 in this supplementary.

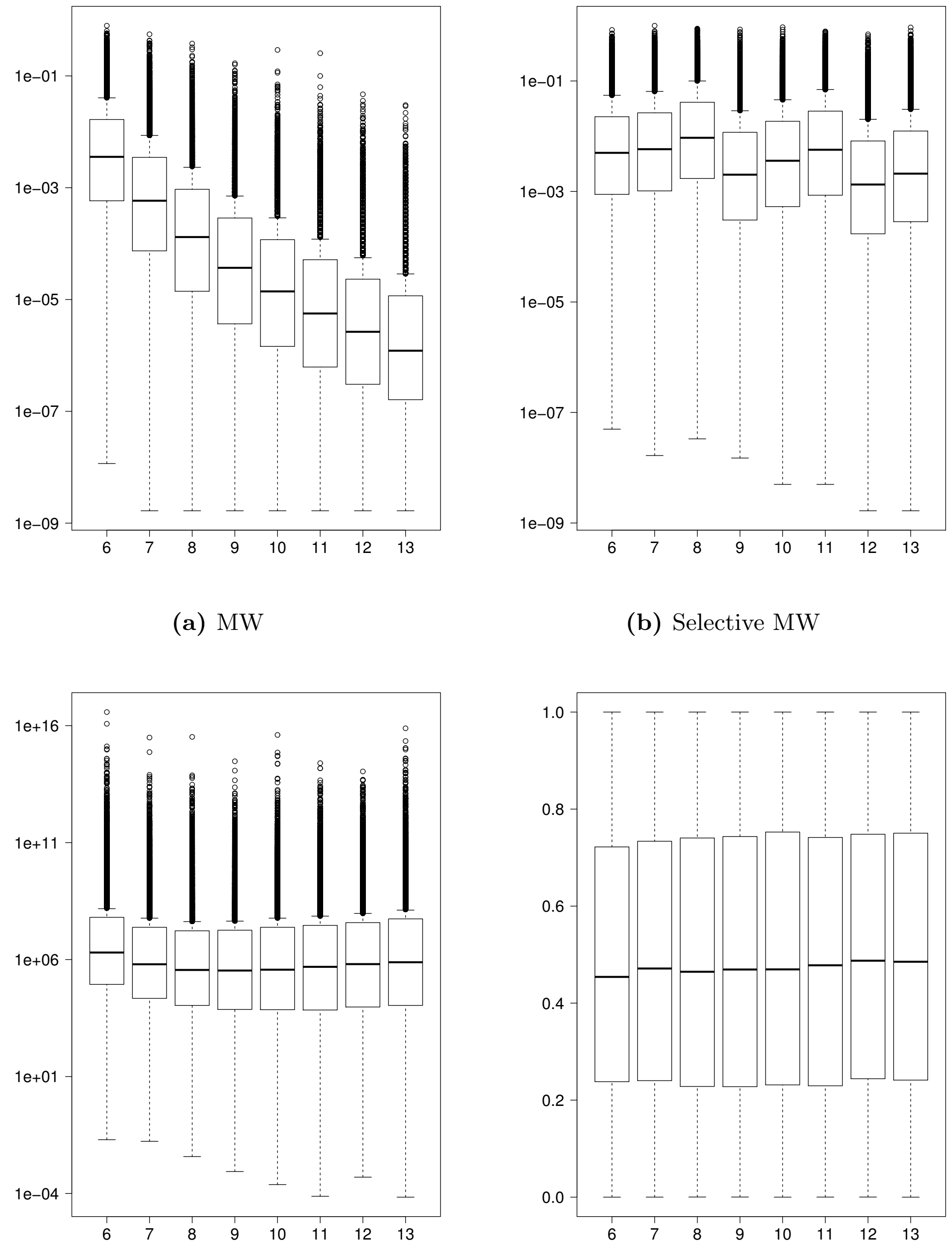
As expected, the null distributions are not uniform because the motifs used to search the sequences are optimised over the input set of sequences and hence the distribution is skewed. These scores are not p-values: the null hypothesis is never satisfied.

Aside from the Fisher score, we find that the non-selective versions of the discriminative scores exhibit a bias for longer motifs. The selective versions of the scores show substantial reduction of this bias. Interestingly, the (non-selective) Fisher score (Figure 2c) does not exhibit as significant a bias as the other methods do especially when considering the longer motifs.

It should be noted that while it is ideal to have similar null distributions across different widths, it is the power of the method to correctly select the best motif when the null is violated that is ultimately more important.

1.3 Comparing the power of the motif scores

To test the retrieval accuracy of our scores, we applied them to two data sets that were constructed with sites derived from real motifs. Each test data set was made of multiple input sets for MEME, each with a varying number of input sequences and lengths. The sequences in each input set were sampled independently from a genomic background file (Section 1.1.3) so they were presumably unrelated to one another. Instances of a motif from our set of transcription factor motifs were then added to the input sequences of each set where each input set was implanted using a single motif. In the OOPS mode every input



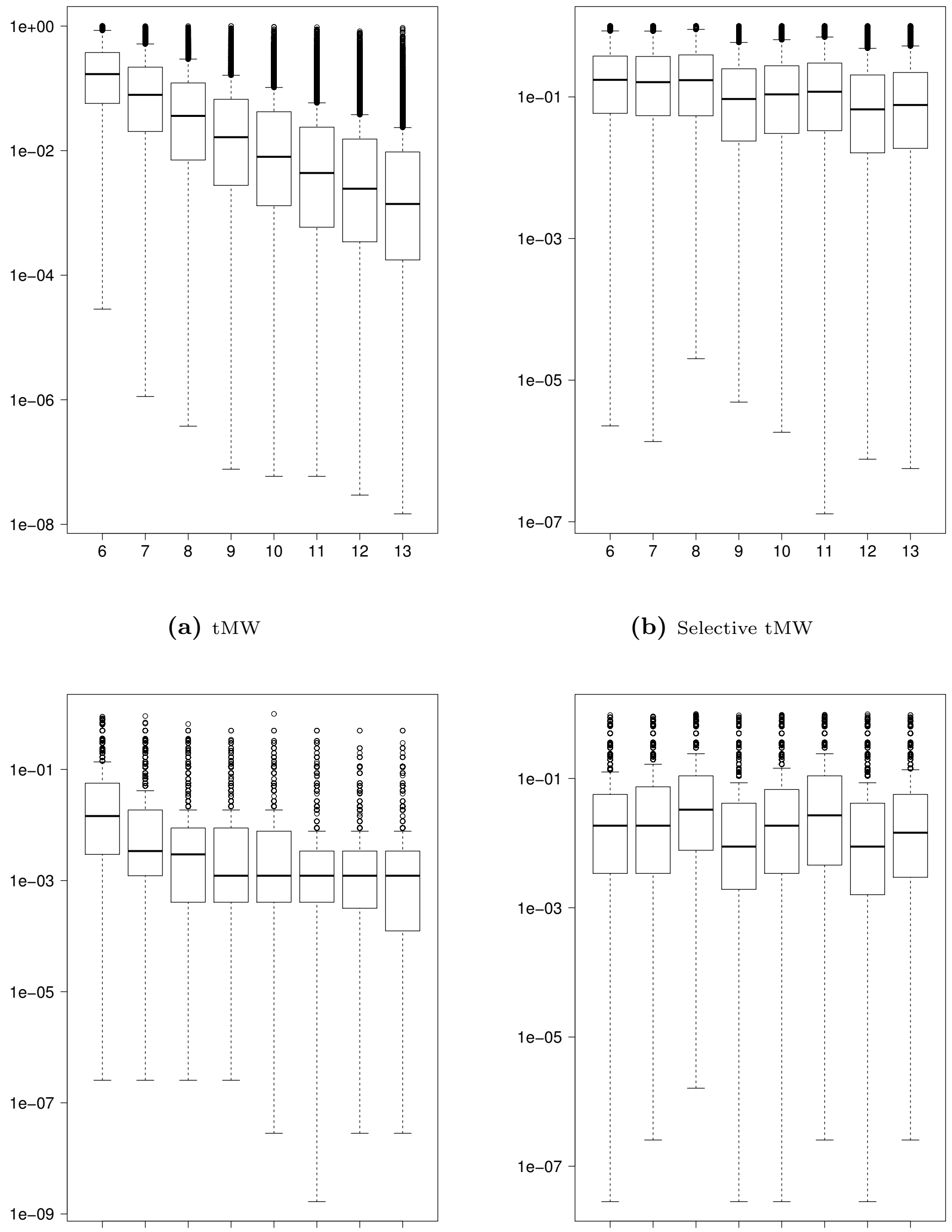
(a) MW

(b) Selective MW

(c) E-value

(d) 3-Gamma

Figure 1: Null distribution of motif scores. The figure shows the boxplots constructed from 10,000 observations of non-selective and selective versions of the MW and the MHG scores and as well as the E-value and 3-Gamma scores for each motif width (ranged from 6 to 13, see Section 1.1 for description of scores). The scores were generated by applying MEME to randomly drawn input set of sequences of a fixed dimensions (number of sequences and their lengths). See Section 1.2 for details.



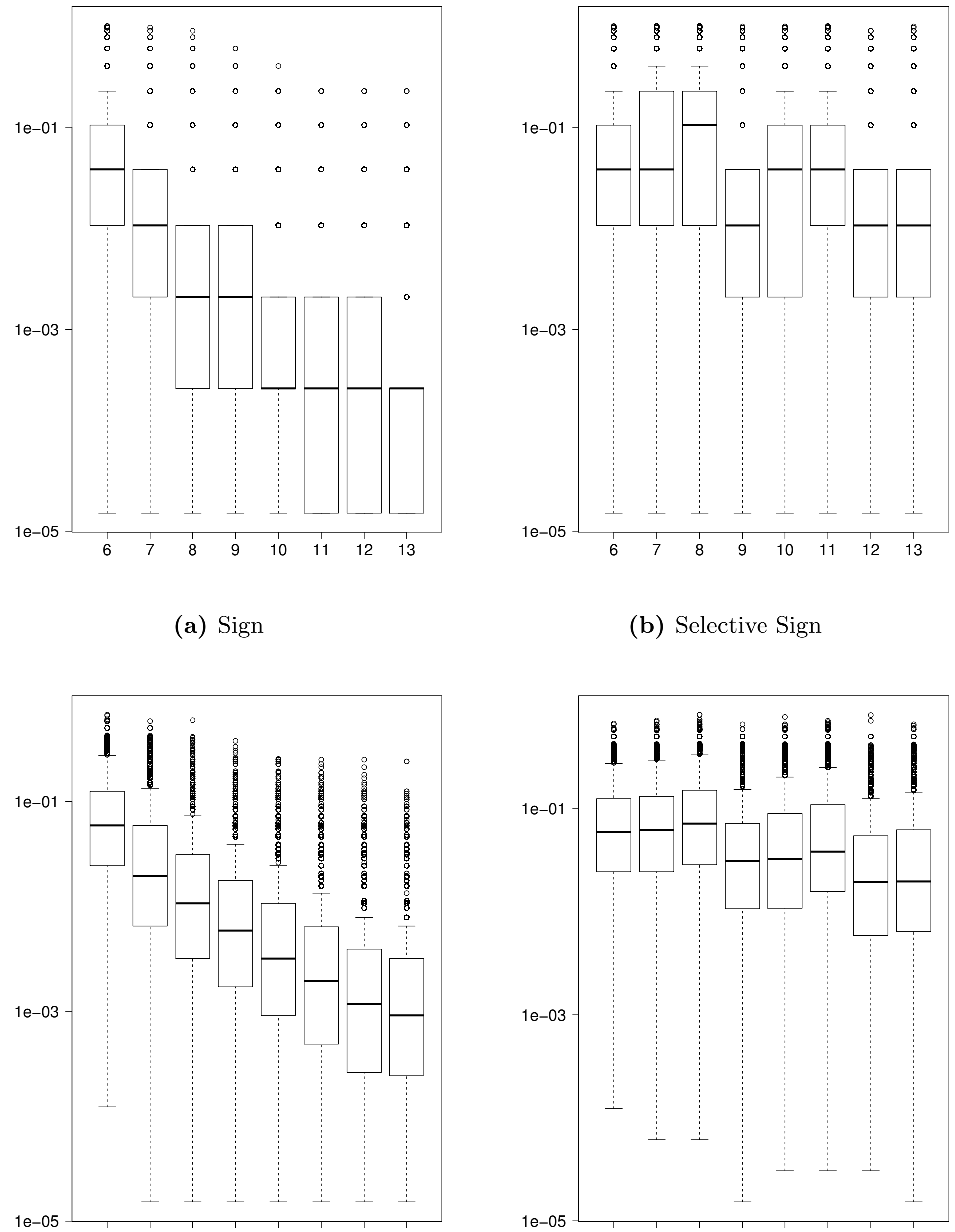
(a) tMW

(b) Selective tMW

(c) Fisher

(d) Selective Fisher

Figure 2: Null distribution of different methods by width. Similar to Figure 1 except for non-selective and selective versions of threshold MW and Fisher scores.



(a) Sign

(b) Selective Sign

(c) MSign

(d) Selective MSign

Figure 3: Null distribution of different methods by width. Similar to Figure 1 except for non-selective and selective versions of Sign and minimal Sign scores.

sequence had a unique instance of the paired motif whereas in ZOOPS mode each input sequence had a single instance of the motif or none.

We applied MEME in the appropriate mode to each of the input sets in each of the two test sets generating multiple candidate motifs for each input set. Each of the considered motif scores was then used to select the best candidate motif for the given input set. That chosen motif was compared with the implanted motif and the similarity level between these two motifs determined whether or not the score function chosen motif matched the implanted motif.

When running this kind of benchmark one has to be somewhat careful when choosing the set of motifs: a small set of motifs might not be sufficiently representative of the known motif space but, on the other hand a larger set might contain many similar motifs which in turn could bias the analysis. Another potentially confounding factor is the difficulty level of the motif finding task: if it is too easy all methods will be successful, and if it is too difficult all will fail. We next describe our experimental setup in more detail concentrating on how we dealt with these problems.

1.3.1 Construction of the spanning set of motifs

To define our set of real motifs we first combined the motifs from the MacIsaac Yeast dataset (MacIsaac *et al.*, 2006) and the Uniprobe Mouse dataset (Newburger and Bulyk, 2009) giving us a total of 510 motifs. While this set of motifs represents a substantial section of known transcription factor motifs it contains many very similar motifs. As such similarities could potentially bias our tests we chose to restrict our attention to a representative spanning set of 100 motifs from the larger set. Specifically we used hierarchical clustering to construct a tree using motif similarities as the distance metric. We then chose a cross section of 100 clusters from this tree and a single motif was selected from each cluster. Technically, the clustering was done using single-linkage agglomerative hierarchical clustering with p-values from TOMTOM defining the distance between a pair of motifs. More specifically, we initially consider each motif as its own cluster. The distance between two clusters is the smallest TOMTOM p-value from all possible motif-motif pairs between the two clusters. Pairs of clusters with the smallest distance were repeatedly merged until we were left with exactly 100 clusters.

A single motif was greedily selected so as to maximize the entropy of the set of selected motif width. That is, if one cluster had more than one motif, then the motif with the width that had the lowest frequency among the set of currently selected motifs was chosen (if the cluster contained several motifs with the same minimal frequency width we randomly chose one of them). The selected motifs have width ranging from 6 to 23. These 100 motifs served as templates to generate sites by using the column frequencies of each template motif as multinomial probabilities.

1.3.2 Mouse And Yeast with OOPS model (MAYO) Set

Our OOPS test set was constructed by pairing each motif from our set of 100 selected motifs with 10 input sets of a varying number of input sequences and lengths giving us a total of 1000 input sets.

In order for our tests to be as discriminative as possible it is important to set the the difficulty of the motif search problem at a level that is neither too high nor too low. Therefore, having chosen the motif we iteratively adjusted the difficulty of the motif search relative to the selected motif by iteratively modifying the dimensions of the input set as explained next.

For each motif we first generated a “maximal” input set by sampling 100 sequences of length 10,000 from our genomic background set of *S. Cerevisiae* intergenic regions (Section 1.1.3). This set was then used as a template for generating 10 more sets of the same dimensions while preserving the original maximal set’s local A-T composition (Section 1.1.4). Note that we sample, not shuffle regions.

We next randomly chose the initial number of sequences N_0 (uniformly between 10 and 20) and each sequence length l_i (uniformly between 100 and 1000) for $i = 1, \ldots, N_0$. These numbers were applied to each of the 10 input sets so if we denote by S_i the initially drawn maximal sequences of one such set then our initial version of this input set consisted of the subsequences $S_i[1 : l_i]$ for $i = 1, \ldots, N_0$.

A single instance of the paired motif generated using a multinomial model was added to each input sequence in each of the 10 input sets and MEME was applied in OOPS mode to each of these 10 input sets using the known paired (implanted) motif width. Since we ran MEME specifying the implanted motif’s correct width there was no issue of selecting the best candidate motif: there was only one candidate motif per dataset, so none of the scoring functions (including the E-value) is involved in selecting the motif. Each of those 10 reported motifs was then compared with the paired motif using TOMTOM where a p-value

≤ 0.05 was considered a success. If the success rate across the 10 sets was 30-70% (inclusive) then we stop our iterative procedure having achieved our target difficulty level. Otherwise, we make the motif search either easier or more difficult according to whether the success rate was below 30% or above 70% respectively.

Technically we adjust the difficulty level by either simultaneously changing the number of sequences in each of our 10 input sets or by simultaneously modifying the length of all the sequences in each set. More specifically, we can make the motif search harder by decreasing the number of sequences and easier by increasing the number of sequences. Similarly, the search becomes more difficult as the sequence lengths increase. Each such change is applied simultaneously to all 10 input sets so that these 10 sets always have the same dimensions and we alternate between changing the number of sequences in the set and changing the sequences lengths. Specifically, we alternately add or drop a single sequence from each input set or extend/clip each of the sequences in each input set by 10%. We naturally make sure a clipped sequence still carries its implanted site.

MEME is applied to the modified set and this iterative process is repeated until we either achieve the desired difficulty level, as described above, or the sets reach their maximal dimensions. Either way, one of the ten sets is then selected as part of the test set. We repeat this process 10 times for each of our 100 motifs yielding a total of 1000 OOPS sets. The median number of sequences per set was 19 and the median of the set-averaged sequence length was 576.

1.3.3 Mouse And Yeast with ZOOPS model (MAYZ) Set

We similarly generated a data set of 1000 input sets using the ZOOPS model. The only differences from the procedure described above for the OOPS model is that:

- We do not implant a site in each sequence of the 10 sets that are used in finding the desired dimension. Rather, a fixed site rate (chosen uniformly between 50-80%) is chosen for all 10 input sets and that rate determines the number of sequences that will carry a single implanted site in each input set. The remaining sequences in each input set do not contain any site.
- Our iterative scheme included a step that increased the ZOOPS factor by 0.1 (to make the motif finding problem easier) or decreased it by 0.1 (to make it harder) subject to the constraint that the ZOOPS factor should stay between 0.5 to 0.8.

The median number of sequences per input set for this MAYZ set was 25 and the median of the set-averaged sequence length was 526.

1.3.4 Motif search analysis: simulated set

For each set in the MAYO set, we used MEME in OOPS mode to search for a single motif of a fixed width from 6 to 13. This approach assures that MEME's E-value metric does not influence the motif it reports. For each reported PWM we assigned a motif score using one of the methods described in Section 1.1. The highest scoring candidate motif was selected as the optimal motif (among the 8 independent runs of MEME). This motif was compared with the original paired motif that was used to generate the set's binding sites using Tomtom. If the Tomtom p-value was less than or equal to 0.05, we labeled this search as a success. Table 1a shows the success rate of each method when applied to the MAYO set.

The same tests were repeated with the MAYZ set but with MEME applied in ZOOPS mode. However, when in ZOOPS mode, MEME internally uses the E-value to choose the best candidate motif among those using different number of occurrences of the motif. Therefore, we replaced this internal E-value selection with fixing the number of sites (`-nsites`) per width. Following MEME's own strategy the values of `-nsites` we considered for each motif width were 2^k for $k \in \{1, 2, \ldots, \lfloor \log_2 N \rfloor\} \cup \{\log_2 N\}$ where N is the number of sequences in the input set.

For each MAYZ input set the selected motif was chosen as the highest scoring one among all candidate motifs generated using the different combinations of width and number of sites. A summary of this test is given in Table 3a. In Table 4 we give a breakdown of the success rate for each fixed motif width by selecting the best motif among candidate motifs generated when varying only the `-nsites` parameter.

Finally, we looked at the power of the different motif scores when the model is misspecified (Table 2). Specifically, we ran the test on the MAYZ but using MEME in OOPS mode. Of course in this case we only varied the motif width to generate the multiple candidates for each input set.

1.3.5 Statistical analysis of the difference in success rates

For all those tests, we assessed whether the observed differences in the number of successes are statistically significant using a two-sided sign test. Specifically, for each pair of motif scores we applied the sign test to compare the number of sets where only one of the two scores was successful in recovering the paired template motif. Under the null hypothesis each such set is equally likely to be discovered by either one of the two considered scores. Therefore the underlying null distribution is binomial and a p-value can be readily assigned.

1.4 3-parameter Gamma fit for motif scores

We previously showed that the 3-parameter Gamma seems to offer a good fit for the score of the best motif reported by the finder for several combinations of finders, scoring functions and null models. Here we settled for a single test based on 100,000 null sampled minimal selective MHG and minimal selective MW scores which were generated as follows. We first generated 100,000 null sets of the same dimension (16 sequences of various lengths) as described in Section 1.2. MEME was then applied eight times in OOPS mode to each of those null sets specifying a different motif width from 6 to 13 with every run. The selective MW and MHG scores were computed for each of the eight reported motifs and the best (minimal) one across all widths was taken as the corresponding minimal selective MW/MHG null score.

Figure 4 shows the probability plot of the 100,000 null sampled minimal selective MW scores[1] against its MLE 3-Gamma fit. Figure 3 in the main paper provides a similar picture of the null distribution of the minimal selective MHG score. In both figures we added as a reference the MLE Normal fit to the data. In both figures we see that the Normal fit deviates considerably from the observations at the tails while the 3-Gamma fit stays fairly close.

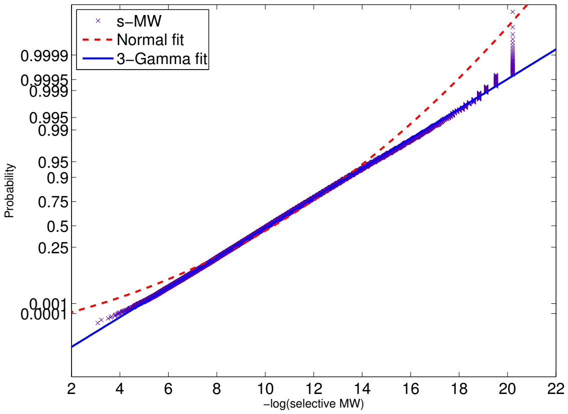


Figure 4: Probability plot of the null selective MW score with normal and 3-Gamma parametric fits. This probability plot compares the empirical null distribution of the selective MW score with the optimal normal as well as 3-Gamma parametric fits. The empirical distribution was generated using 100,000 observations and the parametric fits were estimated using maximum likelihood. See Section 1.4 for details.

[1] minus log of minimum selective scores to be exact

3G	E-value		MW	tMW	Fisher	MHG	Sign	MSign
599	495	Non-selective	558	558	507	546	534	533
		Selective	617	584	548	608	580	580

(a) Success in selecting the best candidate motif.

	3G	E-value	MW	tMW	Fisher	MHG	Sign	MSign
3G		$< 0.0001^-$	0.0055^-	0.0070^-	$< 0.0001^-$	0.0004^-	$< 0.0001^-$	$< 0.0001^-$
E-value	$< 0.0001^+$		$< 0.0001^+$	$< 0.0001^+$	0.4985	0.0005^+	0.0115^+	0.0124^+
MW	0.0055^+	$< 0.0001\ ^-$		1.0000	0.0007^-	0.4175	0.1002	0.0661
tMW	0.0070^+	$< 0.0001^-$	1.0000		0.0022^-	0.4344	0.1019	0.0725
Fisher	$< 0.0001^+$	0.4985	0.0007^+	0.0022^+		0.0185^+	0.1141	0.1166
MHG	0.0004^+	0.0005^-	0.4175	0.4344	0.0185^-		0.4500	0.3853
Sign	$< 0.0001^+$	0.0115^-	0.1002	0.1019	0.1141	0.4500		1.0000
MSign	$< 0.0001^+$	0.0124^-	0.0661	0.0725	0.1166	0.3853	1.0000	
s-MW	0.2174	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$
s-tMW	0.3463	$< 0.0001^-$	0.0828	0.0887	$< 0.0001^-$	0.0128^-	0.0011^-	0.0005^-
s-Fisher	0.0005^+	0.0009^-	0.5459	0.5385	0.0098^-	0.9490	0.3995	0.3611
s-MHG	0.5565	$< 0.0001^-$	0.0003^-	0.0006^-	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$
s-Sign	0.2064	$< 0.0001^-$	0.1274	0.1314	$< 0.0001^-$	0.0176^-	0.0015^-	0.0007^-
s-MSign	0.2086	$< 0.0001^-$	0.1254	0.1294	$< 0.0001^-$	0.0225^-	0.0016^-	0.0004^-

	s-MW	s-tMW	s-Fisher	s-MHG	s-Sign	s-MSign
3G	0.2174	0.3463	0.0005^-	0.5565	0.2064	0.2086
E-value	$< 0.0001^+$	$< 0.0001^+$	0.0009^+	$< 0.0001^+$	$< 0.0001^+$	$< 0.0001^+$
MW	$< 0.0001^+$	0.0828	0.5459	0.0003^+	0.1274	0.1254
tMW	$< 0.0001^+$	0.0887	0.5385	0.0006^+	0.1314	0.1294
Fisher	$< 0.0001^+$	$< 0.0001^+$	0.0098^+	$< 0.0001^+$	$< 0.0001^+$	$< 0.0001^+$
MHG	$< 0.0001^+$	0.0128^+	0.9490	$< 0.0001^+$	0.0176^+	0.0225^+
Sign	$< 0.0001^+$	0.0011^+	0.3995	$< 0.0001^+$	0.0015^+	0.0016^+
MSign	$< 0.0001^+$	0.0005^+	0.3611	$< 0.0001^+$	0.0007^+	0.0004^+
s-MW		0.0007^-	$< 0.0001^-$	0.5455	0.0052^-	0.0035^-
s-tMW	0.0007^+		0.0254^-	0.1019	0.8231	0.8126
s-Fisher	$< 0.0001^+$	0.0254^+		$< 0.0001^+$	0.0266^+	0.0347^+
s-MHG	0.5455	0.1019	$< 0.0001^-$		0.0392^-	0.0314^-
s-Sign	0.0052^+	0.8231	0.0266^-	0.0392^+		1.0000
s-MSign	0.0035^+	0.8126	0.0347^-	0.0314^+	1.0000	

(b) Significance of difference in success rate.

Table 1: MAYO set in OOPS mode. Panel **a** shows the number of successes (out of 1000) of each of the scoring functions described in Section 1 when applied to the MAYO set (see Section 1.3 for the construction of the data set). Panel **b** shows the p-value of the sign test comparing the success rate of the corresponding pair of methods. If the test was significantly better for the column score compared to the row score, there is $^+$ next to the sign test p-value while if it was significantly worse then there is $^-$ sign (we use 5% significance level). See Section 1 for details about each motif score and Section 1.3.4 for the method. The selective version of each score is preceded by the prefix "s-" (e.g. s-MW is the selective version of MW score). If we consistently selected the MEME reported motif of width 6, we would succeed in 503/1000 input sets.

3G	E-value		MW	tMW	Fisher	MHG	Sign	MSign
476	376	Non-selective	436	394	392	402	399	414
		Selective	491	456	457	484	474	465

(a) Success in selecting the best candidate motif.

	3G	E-value	MW	tMW	Fisher	MHG	Sign	MSign
3G		$< 0.0001^{-}$	0.0084^{-}	$< 0.0001^{-}$	$< 0.0001^{-}$	$< 0.0001^{-}$	$< 0.0001^{-}$	0.0001^{-}
E-value	$< 0.0001^{+}$		$< 0.0001^{+}$	0.2643	0.3184	0.0798	0.1093	0.0141^{+}
MW	0.0084^{+}	$< 0.0001^{-}$		0.0003^{-}	0.0010^{-}	0.0098^{-}	0.0052^{-}	0.1028
tMW	$< 0.0001^{+}$	0.2643	0.0003^{+}		0.9486	0.6098	0.7676	0.1352
Fisher	$< 0.0001^{+}$	0.3184	0.0010^{+}	0.9486		0.5139	0.6825	0.1623
MHG	$< 0.0001^{+}$	0.0798	0.0098^{+}	0.6098	0.5139		0.8819	0.3904
Sign	$< 0.0001^{+}$	0.1093	0.0052^{+}	0.7676	0.6825	0.8819		0.2383
MSign	0.0001^{+}	0.0141^{-}	0.1028	0.1352	0.1623	0.3904	0.2383	
s-MW	0.3059	$< 0.0001^{-}$	$< 0.0001^{-}$	$< 0.0001^{-}$	$< 0.0001^{-}$	$< 0.0001^{-}$	$< 0.0001^{-}$	$< 0.0001^{-}$
s-tMW	0.1897	$< 0.0001^{-}$	0.1588	$< 0.0001^{-}$	0.0001^{-}	0.0003^{-}	0.0001^{-}	0.0044^{-}
s-Fisher	0.2064	$< 0.0001^{-}$	0.1477	$< 0.0001^{-}$	$< 0.0001^{-}$	0.0001^{-}	0.0001^{-}	0.0041^{-}
s-MHG	0.6098	$< 0.0001^{-}$	0.0004^{-}	$< 0.0001^{-}$	$< 0.0001^{-}$	$< 0.0001^{-}$	$< 0.0001^{-}$	$< 0.0001^{-}$
s-Sign	0.9458	$< 0.0001^{-}$	0.0059^{-}	$< 0.0001^{-}$	< 0.0001	$< 0.0001^{-}$	$< 0.0001^{-}$	$< 0.0001^{-}$
s-MSign	0.4829	$< 0.0001^{-}$	0.0371^{-}	$< 0.0001^{-}$	$< 0.0001^{-}$	$< 0.0001^{-}$	$< 0.0001^{-}$	0.0001^{-}

	s-MW	s-tMW	s-Fisher	s-MHG	s-Sign	s-MSign
3G	0.3059	0.1897	0.2064	0.6098	0.9458	0.4829
E-value	$< 0.0001^{+}$	$< 0.0001^{+}$	$< 0.0001^{+}$	$< 0.0001^{+}$	$< 0.0001^{+}$	$< 0.0001^{+}$
MW	$< 0.0001^{+}$	0.1588	0.1477	0.0004^{+}	0.0059^{+}	0.0371^{+}
t-MW	$< 0.0001^{+}$	$< 0.0001^{+}$	$< 0.0001^{+}$	$< 0.0001^{+}$	$< 0.0001^{+}$	$< 0.0001^{+}$
Fisher	$< 0.0001^{+}$	0.0001^{+}	$< 0.0001^{+}$	$< 0.0001^{+}$	$< 0.0001^{+}$	$< 0.0001^{+}$
MHG	$< 0.0001^{+}$	0.0003^{+}	0.0001^{+}	$< 0.0001^{+}$	$< 0.0001^{+}$	$< 0.0001^{+}$
Sign	$< 0.0001^{+}$	0.0001^{+}	0.0001^{+}	$< 0.0001^{+}$	$< 0.0001^{+}$	$< 0.0001^{+}$
MSign	$< 0.0001^{+}$	0.0044^{+}	0.0041^{+}	$< 0.0001^{+}$	$< 0.0001^{+}$	0.0001^{+}
s-MW		0.0019^{-}	0.0042^{-}	0.6232	0.2100	0.0464^{-}
s-tMW	0.0019^{+}		1.0000	0.0415^{+}	0.1921	0.5285
s-Fisher	0.0042^{+}	1.0000		0.0465^{+}	0.2567	0.6079
s-MHG	0.6232	0.0415^{-}	0.0465^{-}		0.4713	0.1427
s-Sign	0.2100	0.1921	0.2567	0.4713		0.4478
s-MSign	0.0464^{+}	0.5285	0.6079	0.1427	0.4478	

(b) Significance of difference in success rate.

Table 2: MAYZ set in OOPS mode. The success rate of each method applied to the MAYZ set (1000 sets, see Section 1.3 for details) is listed in Table 2a. In this test the model is misspecified: the MAYZ set was constructed using the ZOOPS model but MEME was applied in OOPS mode. Table 2b shows the p-value of the sign test applied to evaluate the significance of the differences in the success rates between every pair of scores. The method keys are as described as in Table 1.

E-value		MW	tMW	Fisher	MHG	Sign	MSign
368	Non-selective	497	513	376	486	433	548
	Selective	586	580	452	615	549	594

(a) Success in selecting the best candidate motif.

	E-value	MW	tMW	Fisher	MHG	Sign	MSign
E-value		$< 0.0001^+$	$< 0.0001^+$	0.6779	$< 0.0001^+$	0.0003^+	$< 0.0001^+$
MW	$< 0.0001^-$		0.2471	$< 0.0001^-$	0.5032	$< 0.0001^-$	0.0026^+
tMW	$< 0.0001^-$	0.2471		$< 0.0001^-$	0.1154	$< 0.0001^-$	0.0416^+
Fisher	0.6779	$< 0.0001^+$	$< 0.0001^+$		$< 0.0001^+$	0.0004^+	$< 0.0001^+$
MHG	$< 0.0001^-$	0.5032	0.1154	$< 0.0001\ ^-$		0.0006^-	0.0002^+
Sign	0.0003^-	$< 0.0001^+$	$< 0.0001^+$	0.0004^-	0.0006^+		$< 0.0001^+$
MSign	$< 0.0001^-$	0.0026^-	0.0416^-	$< 0.0001^-$	0.0002^-	$< 0.0001^-$	
s-MW	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	0.0319^-
s tMW	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	0.0801
s-Fisher	$< 0.0001^-$	0.0063^+	0.0006^+	$< 0.0001^-$	0.0574	0.2812	$< 0.0001^+$
s-MHG	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	0.0001^-
s-Sign	$< 0.0001^-$	0.0026^-	0.0410^-	$< 0.0001^-$	0.0002^-	$< 0.0001^-$	1.0000
s-MSign	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	$< 0.0001^-$	0.0050^-

	s-MW	s-tMW	s-Fisher	s-MHG	s-Sign	s-MSign
E-value	$< 0.0001^+$	$< 0.0001^+$	$< 0.0001^+$	$< 0.0001^+$	$< 0.0001^+$	$< 0.0001^+$
MW	$< 0.0001^+$	$< 0.0001^+$	0.0063^-	$< 0.0001^+$	0.0026^+	$< 0.0001^+$
tMW	$< 0.0001^+$	$< 0.0001^+$	0.0006^-	$< 0.0001^+$	0.0410^+	$< 0.0001^+$
Fisher	$< 0.0001^+$	$< 0.0001^+$	$< 0.0001^+$	$< 0.0001^+$	$< 0.0001^+$	$< 0.0001^+$
MHG	$< 0.0001^+$	$< 0.0001^+$	0.0574	$< 0.0001^+$	0.0002^+	$< 0.0001^+$
Sign	$< 0.0001^+$	$< 0.0001^+$	0.2812	$< 0.0001^+$	$< 0.0001^+$	$< 0.0001^+$
MSign	0.0319^+	0.0801	$< 0.0001^-$	0.0001^+	1.0000	0.0050^+
s-MW		0.6587	$< 0.0001^-$	0.0734	0.0257^-	0.6643
s-tMW	0.6587		$< 0.0001^-$	0.0373^-	0.0682	0.4219
s-Fisher	$< 0.0001^+$	$< 0.0001^+$		$< 0.0001^+$	$< 0.0001^+$	$< 0.0001^+$
s-MHG	0.0734	0.0373^-	$< 0.0001^-$		$< 0.0001^-$	0.2174
s-Sign	0.0257^+	0.0682	$< 0.0001^-$	$< 0.0001^+$		0.0026^+
s-MSign	0.6643	0.4219	$< 0.0001^-$	0.2174	0.0026^-	

(b) Significance of difference in success rate.

Table 3: MAYZ set in ZOOPS mode. Similar to Table 2 except here the model is correctly specified: MEME was applied in ZOOPS mode. The best candidate motifs were selected among candidate motifs generated by varying the motif width and `-nsites` (the number of site bearing sequences).

Width	E-value		MW	tMW	Fisher	MHG	Sign	MSign
6	377	Non-selective	539	503	460	534	523	510
		Selective			Same as above.			
7	402	Non-selective	542	490	452	535	531	524
		Selective	532	488	455	536	515	505
8	381	Non-selective	543	516	405	515	517	547
		Selective	514	480	416	540	546	516
9	387	Non-selective	504	507	356	505	485	529
		Selective	519	514	367	543	546	537
10	359	Non-selective	465	486	339	471	445	506
		Selective	518	507	361	534	529	533
11	342	Non-selective	409	439	284	401	389	466
		Selective	496	502	308	519	484	516
12	307	Non-selective	344	387	250	351	323	399
		Selective	428	442	257	432	411	467
13	294	Non-selective	308	336	230	314	303	349
		Selective	403	419	252	416	414	446

Table 4: MAYZ set in ZOOPS mode per width. Similar to Table 3 except here for each fixed value of MEME's motif width parameter (6–13) we vary MEME's number of sites parameter `-nsites`. We then select the best candidate motif for that specified motif width using the different motif scoring functions. Note that the actual width of the implanted motif (6–23) is unknown to MEME and is only used by TOMTOM in the evaluation step. What is measured here is the ability of each scoring function to pick the best motif when the `-nsites` is varied–a task that is performed in MEME using the E-value.

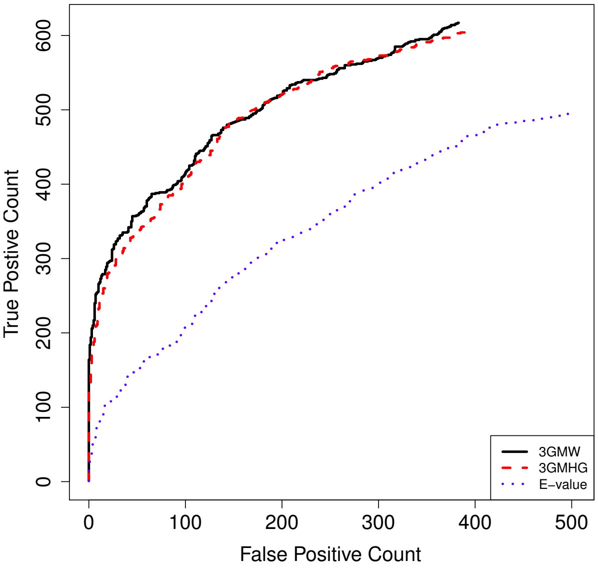


Figure 5: Two-tiered vs. E-value analysis in OOPS mode. For each set, we compare the best candidate motif (from width 6 to 13) as selected by the specified method to the implanted motif using the corresponding Tomtom p-value. If the Tomtom p-value is less than or equal to 0.05, we label the motif as positive otherwise as negative. Varying the significance threshold we plot the number of positive motifs which are deemed significant (TP) vs. the number of negative motifs that are called significant at that level (FP). The significance is determined either by the E-value or by the 3-Gamma point estimate of the p-value (3GMW, 3GMHG). Note that as the optimal motif is selected using different methods: E-value, selective MHG (3GMHG), selective MW (3GMW), the positive/negative label assigned to each set motif might vary with the method. Hence we plot FP vs. TP counts rather than the usual ROC curve. Keep in mind that the significance thresholds are different for the different scores hence while we sort our results according to the reported significance level (E-value or 3-Gamma estimated p-value) we normalize the graphs according to the number of FPs.

Data set	Threshold	3GMW		3GMHG		E-value	
		TP	FP	TP	FP	TP	FP
MAYO	0.001	184	2	179	4	178	76
	0.01	261	10	241	11	192	92
	0.05	322	28	302	32	209	101
MAYZ	0.001	114	2	106	2	94	165
	0.01	185	9	190	5	110	190
	0.05	267	36	274	27	123	207

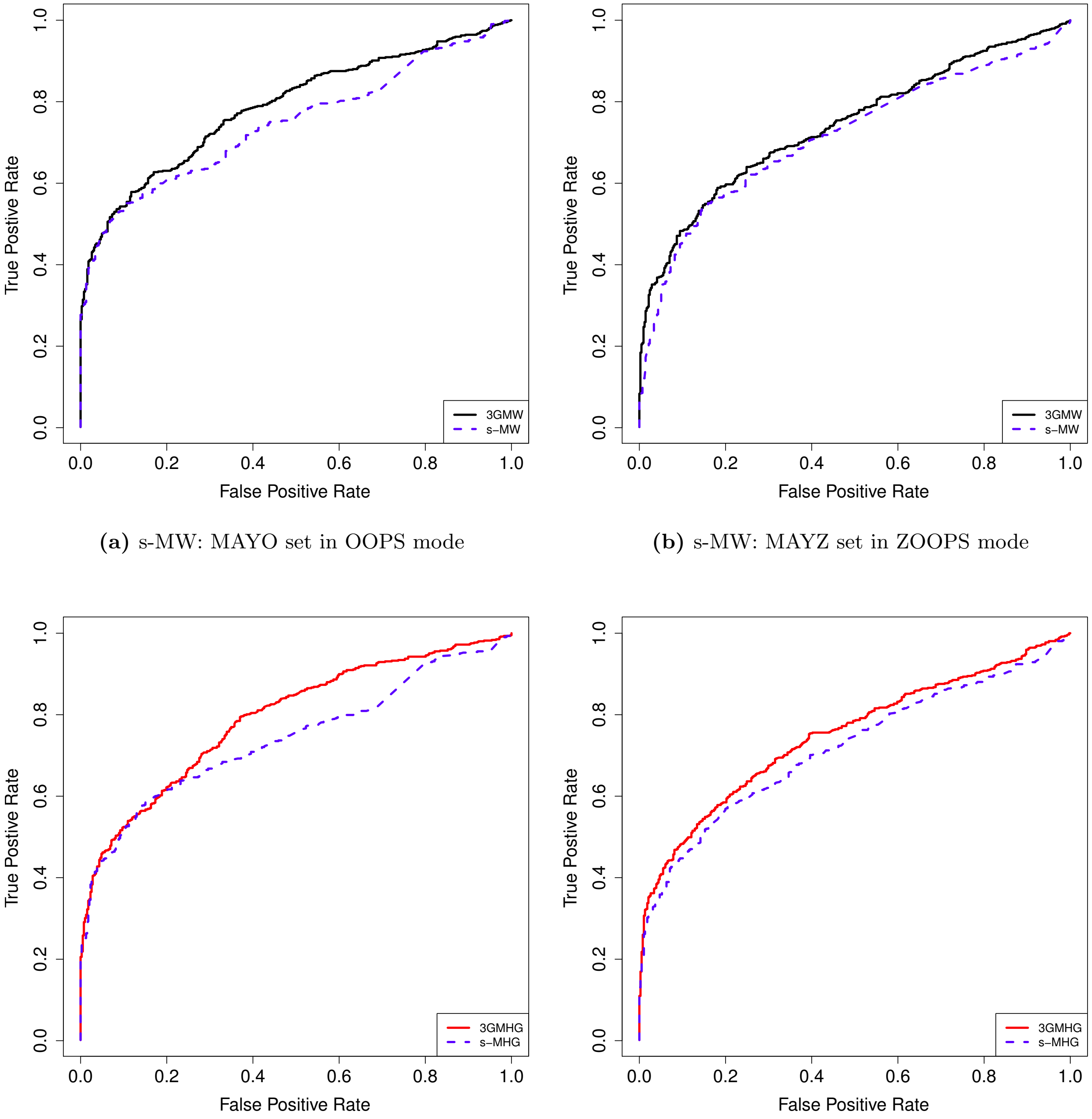
Table 5: The TP and FP counts in Figure 4 in the main paper and Figure 5 for some commonly used thresholds. For each threshold, the number of TP and FP motifs reported by the corresponding method is given. Specifically, the number of TPs is the number of sets with label P with score lower or equal to the threshold and the number of FPs is the number of sets with label N with score lower or equal to the threshold.

Data set	Score	Selective score	3G Significance
MAYO	MW	0.747	0.788
	MHG	0.742	0.792
MAYZ	MW	0.718	0.743
	MHG	0.718	0.750

Table 6: aROC of Figure 6. The aROC of the selective MW and 3GMW as well as the aROC of selective MHG and 3GMHG are given. See Figure 6 for description of method.

Data set	E-value	3GMW	3GMHG
MAYO	0.682	0.815	0.820
MAYZ	0.530	0.839	0.821

Table 7: aROC of Figure 7. The aROC of the different motif scoring methods are given. Here the best motif is always selected using the E-value and the overall motif score is assessed by one of the 3 specified methods. Therefore these aROCs are really a measurement of the score calibration.



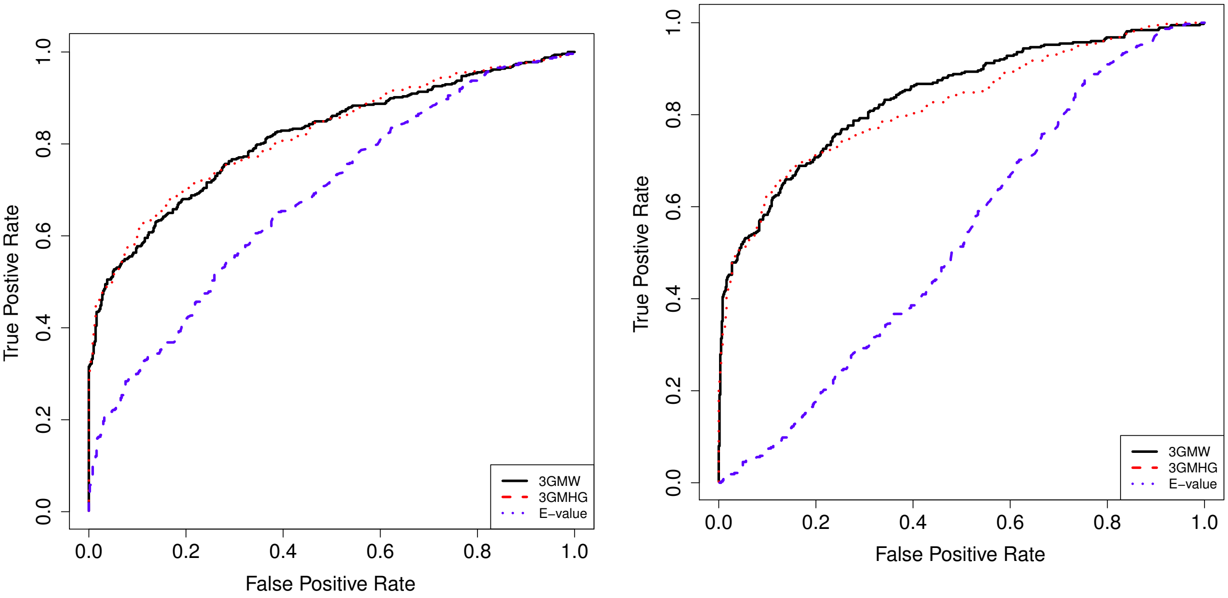
(a) s-MW: MAYO set in OOPS mode

(b) s-MW: MAYZ set in ZOOPS mode

(c) s-MHG: MAYO set in OOPS mode

(d) s-MHG: MAYZ set in ZOOPS mode

Figure 6: Assessing the calibration of the selective discriminative scores. In plots (a) and (b) we compare the classification performance of the selective MW score on its own (s-MW) with its performance when it serves as the basis of our two-tiered analysis (3GMW) in which case the overall motif score is the 3-Gamma p-value. The data sets used for these two plots are described in detail in Section 1.3. Note that only the score of the of the selected motif rather than the motif itself can change between these two methods hence these are standard ROC curves. In particular, the classification performance is essentially a measure of the calibration of these two scores. It is interesting to note that the s-MW score is fairly well calibrated as it is: the difference between these two classifiers is rather small, e.g., the aROCs in (a) are 0.747 (s-MW) and 0.788 (3GMW) (see Table 6). Plots (c) and (d) are the same except for the selective MHG score.



(a) MAYO set in OOPS mode

(b) MAYZ set in ZOOPS mode (the number of site occurrences were chosen internally by MEME using the E-value)

Figure 7: E-value vs. two-tiered calibration. The positive and negative labels are the same as described in Figure 5. In all cases the best candidate motif was selected by using the E-value. The two tiered analysis is used here only to assign an alternative statistical significance to the motif selected using the E-value. Since the motif is selected by the the E-value one could expect that sorting the motifs from different sets according to their E-values should yield optimal results. The figures provide a very different picture showing the E-value is not well calibrated when compared to the 3-Gamma significance. The construction of the data set is described in Section 1.3. See Table 7 for the associated aROCs.

2 Analysis on real data set

2.1 Harbison-Narlikar test set

The "Harbison-Nalikar" test set we use is the same 156 sets of sequences from 80 transcription factors (TFs) used in Ng and Keich (2008a) and Narlikar *et al.* (2007). Specifically this test set was compiled from 310 ChIP-chip experiments of 203 yeast transcription factors in Harbison *et al.* (2004). The consensus sequence of the 80 TFs were obtained from Harbison *et al.* (2004) or MacIsaac *et al.* (2006). These consensus sequence was mapped to a PWM using the same method as Harbison *et al.* (2004).

2.2 Performance on real data set

We compared the performance of five of our motif scores (MW, MHG, selective-MW/MHG, E-value) in choosing the optimal motif among several candidates generated by MEME's EM process applied to the Harbison-Narlikar set. Our analysis here followed the one described in Section 1.3.4: for each set, we used MEME in OOPS mode to search for a motif of a fixed width from 6 to 13. We selected the best candidate motif using either the E-value, MW, MHG, selective MW or selective MHG score. This motif was compared to the consensus PWM (see above) using TomTom. Again we used a cutoff of 0.05 to determine if the search yielded a correct identification.

E-value		MW	MHG
42	Non-selective	48	48
	Selective	54	53

(a) Success in selecting the best candidate motif.

	E-value	MW	MHG	s-MW	s-MHG
E-value		0.2632	0.2379	0.0169	0.0266
MW	0.2632		1.0000	0.1460	0.2668
MHG	0.2379	1.0000		0.1094	0.2266
s-MW	0.0169	0.1460	0.1094		1.0000
s-MHG	0.0266	0.2668	0.2266	1.0000	

(b) Significance of difference in success rate.

Table 8: Harbison-Narlikar set in OOPS mode. Table (a) shows the success rate of each method applied to the Harbison-Narlikar set (156 sets, see Section 1.3 for details). Table (b) shows the significance of the difference between the success rate of a pair of methods. For this test, MEME was applied in OOPS mode.

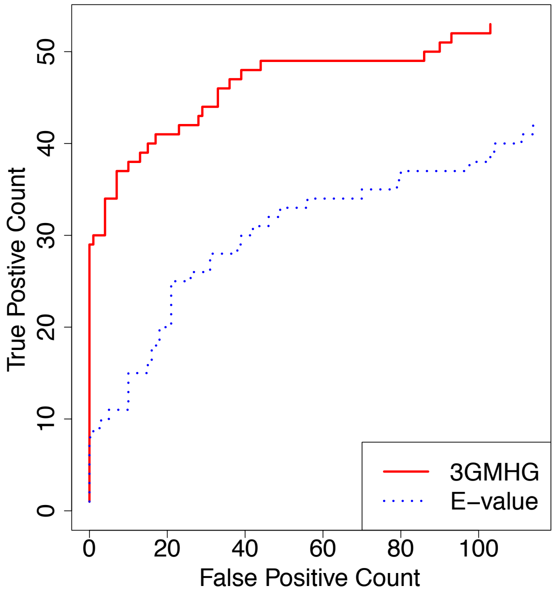


Figure 8: Two-tiered vs. E-value analysis for Harbison-Narlikar test set. Similar to Figure 5 except using Harbison-Narlikar test set and here the plot of TP vs. FP for 3GMHG and E-value are only shown.

The two-tiered analysis was performed only with the selective MHG score (3GMHG) using 50 null sets to assign the 3-Gamma p-value.

Threshold	3GMHG		E-value	
	TP	FP	TP	FP
0.001	30	4	28	34
0.01	35	7	29	39
0.05	40	15	30	41

Table 9: The TP and FP counts in Figure 8 for some commonly used thresholds. The above table is similar to Table 5 except this is for the Harbison-Narlikar test set.

References

Eden, E., Lipson, D., Yogev, S., and Yakhini, Z. (2007). Discovering Motifs in Ranked Lists of DNA Sequences. *PLoS Computer Biology*, **3**(3), e39.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**(1), 48+.

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J., Jennings, E. G., Zeitlinger, J., Pokholok, D. K., Kellis, M., Rolfe, P. A., Takusagawa, K. T., Lander, E. S., Gifford, D. K., Fraenkel, E., and Young, R. A. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**(7004), 99–104+.

Keich, U. and Ng, P. (2007). A conservative parametric approach to motif significance analysis. *Genome informatics International Conference on Genome Informatics*, **19**(1), 61–72.

Keich, U., Gao, H., Garretson, J., Bhaskar, A., Liachko, I., Donato, J., and Tye, B. (2008). Computational detection of significant variation in binding affinity across two sets of sequences with application to the analysis of replication origins in yeast. *BMC Bioinformatics*, **9**(1), 372.

MacIsaac, K., Wang, T., Gordon, D., Gifford, D., Stormo, G., and Fraenkel, E. (2006). An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**(1), 113.

Nagarajan, N., Jones, N., and Keich, U. (2005). Computing the P-value of the information content from an alignment of multiple sequences. *Bioinformatics*, **21**(Suppl 1), i311–i318.

Narlikar, L., Gordân, R., and Hartemink, A. (2007). Nucleosome Occupancy Information Improves *de novo* Motif Discovery. *Proceedings of the 11th Annual International Conference, RECOMB 2007*, **4453**, 107–121.

Newburger, D. and Bulyk, M. (2009). Uniprobe : an online database of protein binding microarray data on protein – dna interactions. *Nucleic Acids Research*, **37**(Database issue), D77–82.

Ng, P. and Keich, U. (2008a). Factoring local sequence composition in motif significance analysis. *Genome Informatics*, **21**, 15–26.

Ng, P. and Keich, U. (2008b). GIMSAN: a Gibbs motif finder with significance analysis. *Bioinformatics*, **24**(19), 2256–2257.

Steinfeld, I., Navon, R., Ardigo, D., Zavaroni, I., and Yakhini, Z. (2008). Clinically driven semi-supervised class discovery in gene expression data. *Bioinformatics*, **24**, i90–i97.