Supplementary materials to "Integrating quantitative information from ChIP-chip experiments into motif finding"

Heejung Shim[1] and Sündüz Keleş[1,2*]
[1] Department of Statistics,
[2] Department of Biostatistics and Medical Informatics,
1300 University Avenue, University of Wisconsin, Madison, WI 53705.
Tel: (608) 263-4533. Fax: (608) 262-0032.

March 30, 2007

1 Stat1 and c-Jun ChIP-chip data.

Stat1 data can be downloaded from the ENCODE `Yale Stat1 Sites` track using the `Yale 50-38 Sites` table of the UCSC Genome Browser (`http://genome.ucsc.edu`). c-Jun data are available at the ENCODE `Yale ChIP Sites` track using the `Yale cJun` table. There are a total of 345 Stat1 bound regions and 200 c-Jun bound regions identified by the Snyder Lab at the Yale University. Their analysis report a quantitative score for each probe by using a 501 bps sliding window centered on each oligonucleotide probe and computing the pseudomedian signal of all log2 ratios of treatment and control measurements within the window. In our application, peaks are ranked according to their ChIP-chip score which is computed as the mean of probe specific summary measurements within a peak.

2 CTCM analysis of the Stat1 data

In the analysis of Stat1 data, the best performance is achieved with the CTCM logistic regression model M2. For CTCM(M3), we observe that for the sample size of $N = 20$, increasing the flanking sequence length to 300 bases from 100 bases leads to a worse position weight matrix estimate. Examining the ChIP-chip scores for several of these 20 sequences, we note that the variability in ChIP-chip scores does not increase significantly when considering 300 bases rather than 100

*Corresponding author: keles@stat.wisc.edu

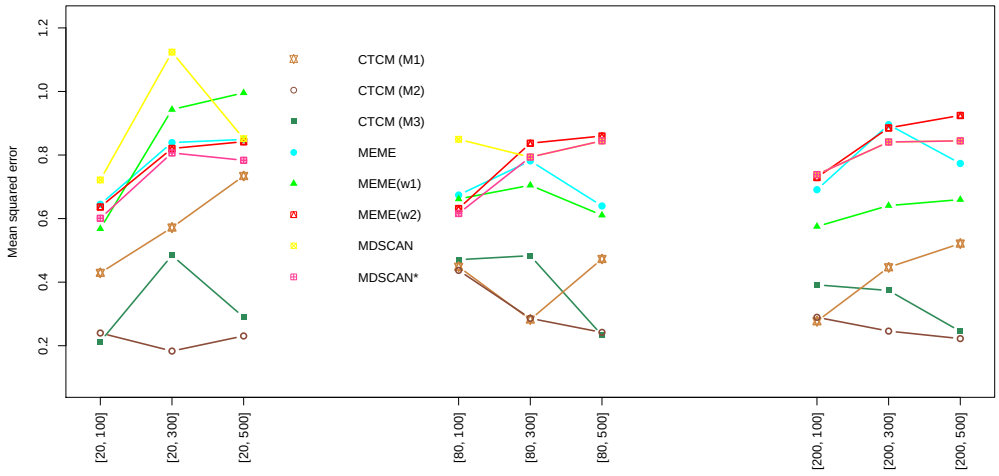

Figure 1: *Stat1: Illustration of the sequence length and sample size effect in the motif analysis of ChIP-chip identified regions.* $[N, L]$, $N = 20, 80, 100$ and $L = 100, 300, 500$, refer to N highest scoring peak regions and L flanking bases to the right and left of the mid point of the peak. Mean squared error is calculated by averaging the squared distance between the components of the true and estimated position weight matrices. MEME(wk), $k = 1, 2$ are two weighted versions of MEME. MDSCAN* refers to use of the best position weight matrix estimate in the mean squared error sense among the top 5 reported by MDSCAN. M1, M2, and M3 refer to beta prior, logistic regression, and piecewise constant linear model formulations of CTCM.

whereas it increases significantly when the flanking sequence length is 500 bases. An example of this phenomenon is displayed in Figure 2, where we plot ChIP-chip scores for a particular peak as a function of the bases around the mid point of the peak. The lack of variability among the ChIP-chip scores up to 300 bases around the mid-point of the peak is apparent from this plot. At the flanking sequence length of 500 bases, we start to observe more variability in ChIP-chip scores. This is utilized by CTCM thereby leading to a decrease in the distance criteria.

3 JASPAR position weight matrices

Figure 3 displays the histogram of information contents of 111 position weight matrices in JASPAR. The triangles indicate the information content of the position weight matrices used in the simulations.

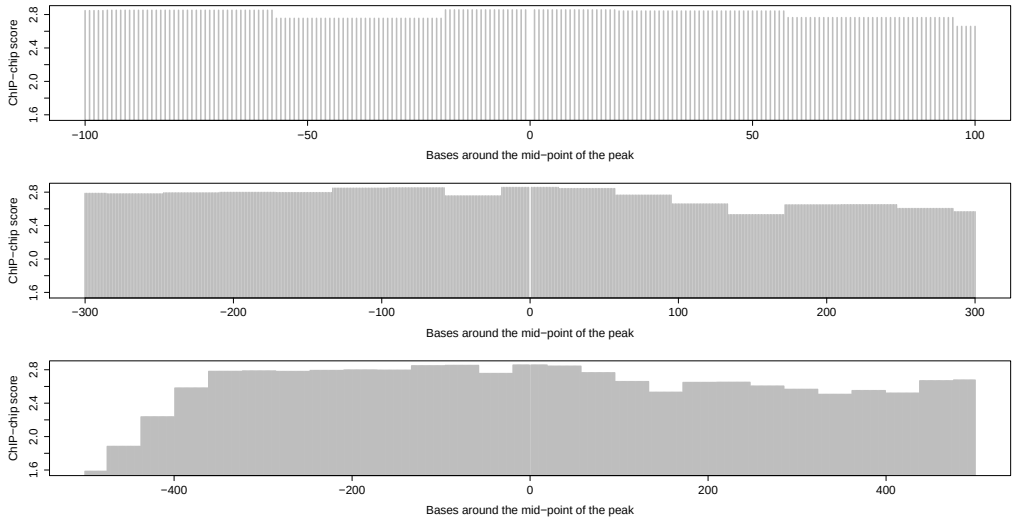

Figure 2: *Stat1: ChIP-chip scores along a peak on chromosome 13 with mid point position at 112814184 bases.*

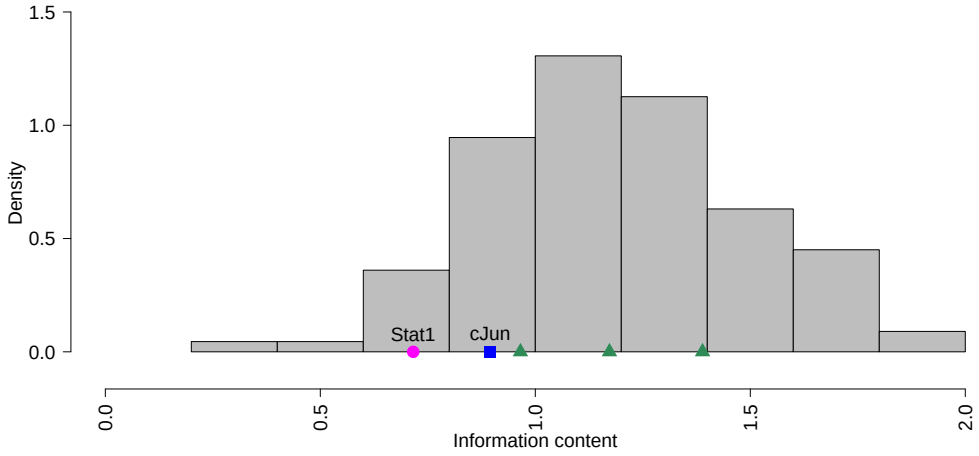

Figure 3: *Histogram of the information contents of the position weight matrices in the JASPAR database.* The triangles depict the three position weight matrices used in the simulation studies.

Factor		a = 0.1	a = 0.3	a = 0.5	a = 0.7	a = 1.3
GABPA	df	3	4	6	6	6
	β_0	-19	-18	-19	-19.3	-19.3
	β_1	1	0.95	0.9	0.925	1
BC2	df	3	4	6	6	6
	β_0	-19.1	-18	-19.45	-19.3	-19
	β_1	1	0.95	0.9	0.925	1
TT3	df	3	4	6	6	6
	β_0	-19	-18	-19	-19.3	-19.3
	β_1	1	0.95	0.9	0.925	1.05

Table 1: *Simulation parameters for the logistic regression model of* $Pr(Z_{il} = 1 \mid T_{il})$.

4 Simulation model

The simulation studies are tailored towards investigating the conditional formulation of the TCM model. We generated 50 sequences of length $L = 800$ from a multinomial model with equal nucleotide probabilities and implanted motif occurrences from each position weight matrix based on the simulated ChIP-chip data. The ChIP-chip score corresponding to each probe of length 25 base pairs is simulated from a chi-square distribution. We decided along each sequence whether a motif starts at a particular site with a probability from a logistic regression model. The logistic regression model was chosen as the simulation model since it fitted the case study data considered in Section 4.1 of our paper. After implanting a motif, we slide to the site next to the end of the motif and continue deciding whether or not to implant a new motif. This process exactly mimics the TCM random process as described in Bailey (1995). The motif abundance parameter is adjusted by controlling the degrees of freedom of the corresponding chi-square distribution and parameters β_0 and β_1. These parameters for all the simulations are summarized in Table 1.

5 Summary of the simulation results

The results are summarized for each position weight matrix and abundance combination in Figures 4, 5, 6.

6 A note on the existence of the maximum likelihood estimator in the logistic regression (M2) CTCM model with small sample sizes

Theorem 1 *Consider the conditional two component mixture model with the logistic regression prior model* $Pr(Z_k = 1 \mid T_k) = logit(\beta_0 + \beta_1 T_k)$ *for integrating the quantitative ChIP-chip information. Assume that at t-th iteration of the M-step step, we have* $\zeta_k = Pr(Z_k = 1 \mid X_k, T_k, \hat{\Psi}^t) \in \{0, 1\}$, $\forall k$, *where* $\hat{\Psi}^t$ *represents the parameters estimates from the previous M-step. If there exists* $\beta^* = (\beta_0, \beta)$ *such that*

$$\begin{aligned} (\beta_0 + \beta_1 T_k) &\geq 0 \quad \textit{if} \quad \zeta_k = 1, \\ (\beta_0 + \beta_1 T_k) &< 0 \quad \textit{if} \quad \zeta_k = 0, \end{aligned}$$

where ζ_k *represents the posterior probability of k-th subsequence being a motif, then the EM algorithm converges to non-finite estimates for the vector parameter* β^*.

Proof: We note that the M-step of the EM algorithm in our conditional TCM formulation has the following form when we use a logistic regression model to incorporate the quantitative ChIP-chip information:

$$\sum_{i=1}^{N} \sum_{l=1}^{L_i - W + 1} \left\{ \zeta_{i,l} (\beta_0 + \beta_1 T_{i,l}) - \log\left(1 + \exp\left(\beta_0 + \beta_1 T_{i,l}\right)\right) \right\}.$$

Since each subsequence is treated independently in the CTCM model, we will index the $\sum_{i=1}^{N} \sum_{l=1}^{L_i - W + 1}$ observations as $k = 1, \cdots, N'$. Then the objective function involving β_0 and β_1 in the M-step becomes

$$l(\beta_0, \beta 1) = \sum_{k=1}^{N'} \left\{ \zeta_k (\beta_0 + \beta_1 T_k) - \log\left(1 + \exp\left(\beta_0 + \beta_1 T_k\right)\right) \right\}, \tag{1}$$

where $0 \leq \zeta_k \leq 1$. We note that when $\zeta_k \in \{0, 1\}\ \forall k$, e.g., when the posterior probabilities of start sites is exactly 0 or 1, equation (1) is exactly a logistic regression log-likelihood. In this case, it is well known that there may not exist any finite maximum likelihood estimators for β_0 and β_1 (Albert and Anderson, 1984). This is easily seen by considering the following case: Assume that there exists a $\beta^* = (\beta_0^{,}\beta_1^{)}$ such that

$$\begin{aligned} (\beta_0 + \beta_1 T_k) &\geq 0 \quad \text{if} \quad \zeta_k = 1, \\ (\beta_0 + \beta_1 T_k) &< 0 \quad \text{if} \quad \zeta_k = 0. \end{aligned}$$

Then, for any $c > 0$, we have

$$\begin{aligned} l(c\beta_0, c\beta_1) &= \sum_{k=1}^{N'} \{\zeta_k(c\beta_0 + c\beta_1 T_k) - \log\left(1 + \exp\left(c\beta_0 + c\beta_1 T_k\right)\right)\}, \\ &= \sum_{k=1, \zeta_k=1}^{N'} \{(c\beta_0 + c\beta_1 T_k) - \log\left(1 + \exp\left(c\beta_0 + c\beta_1 T_k\right)\right)\}, \\ &\quad - \sum_{k=1, \zeta_k=0}^{N'} \log\left(1 + \exp\left(c\beta_0 + c\beta_1 T_k\right)\right). \end{aligned}$$

As $c \longrightarrow \infty$, the above expression converges to 0. Furthermore, we know that $l(\beta_0, \beta_1) \leq 0$, $\forall(\beta_0, \beta_1)$. Therefore, a finite maximum likelihood estimator does not exist for the vector parameter β^* satisfying the above separability condition. We note that perfect separability is largely a small sample issue and in fact it hardly ever happens with noisy ChIP-chip tiling array data. However, it is a concern when, say, performing simulation studies with small sample sizes. Additionally, as the complete data log-likelihood factorizes into two parts involving sequence model parameters and parameters of the prior distribution, the EM algorithm still provides consistent estimates for the sequence model parameters (e.g., parameters of the position weight matrix and the background).

References

Albert, A. and Anderson, J. (1984). On the existence of maximum likelihood estimates in logistic regression models, *Biometrika* **71**: 1–10.

Bailey, T. L. (1995). *Discovering motifs in DNA and protein sequences: The approximate common*

substring problem, PhD thesis, University of California, San Diego.

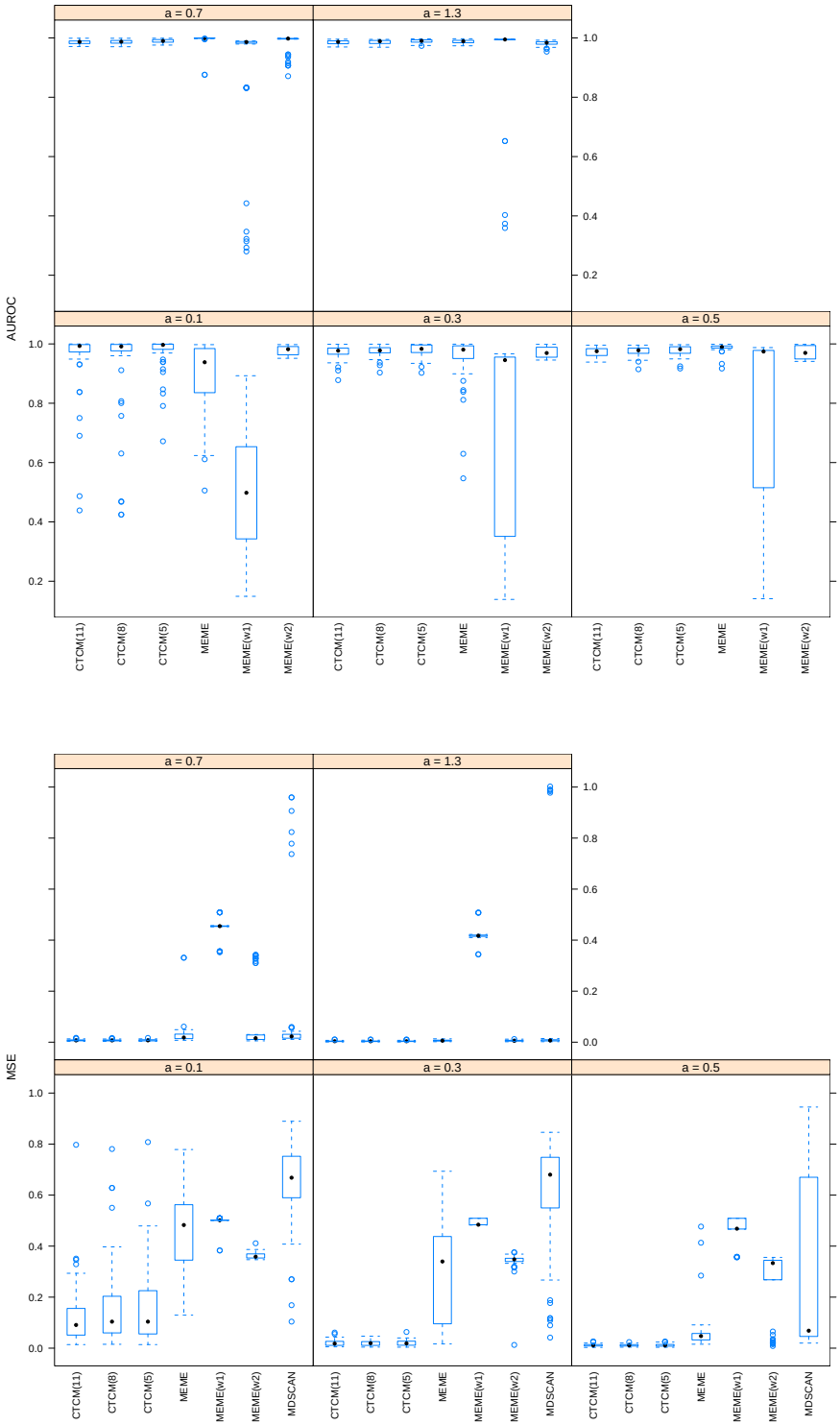

Figure 4: *Simulation results with the GABPA position weight matrix.* Top and bottom panel plots display the boxplots of the area under the ROC curve and mean squared errors for various methods at different abundance levels a. CTCM(b) corresponds to CTCM M3 model with a bin size of b.

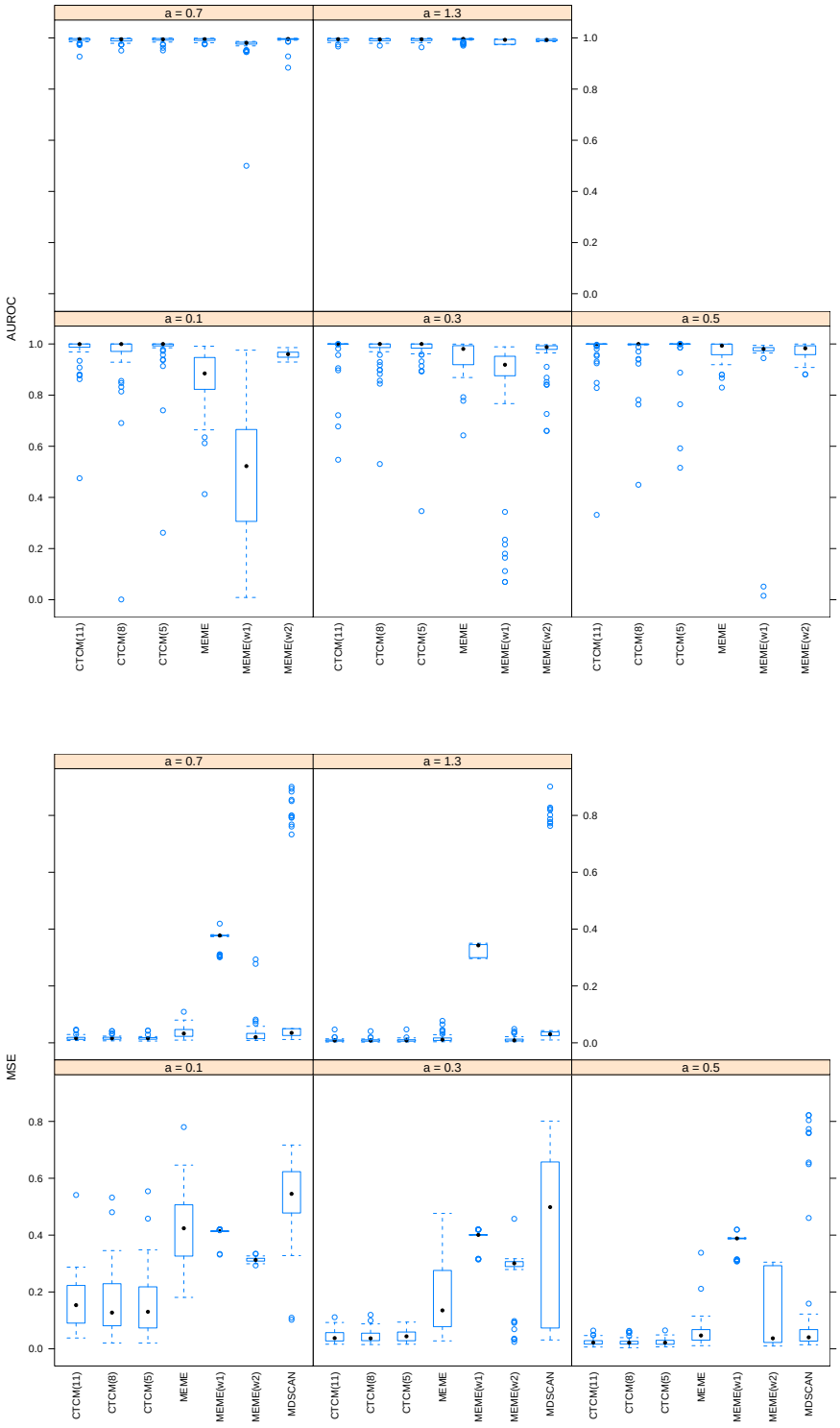

Figure 5: *Simulation results with the TAL1-TCF3 position weight matrix.* Top and bottom panel plots display the boxplots of the area under the ROC curve and mean squared errors for various methods at different abundance levels a. CTCM(b) corresponds to CTCM M3 model with a bin size of b.

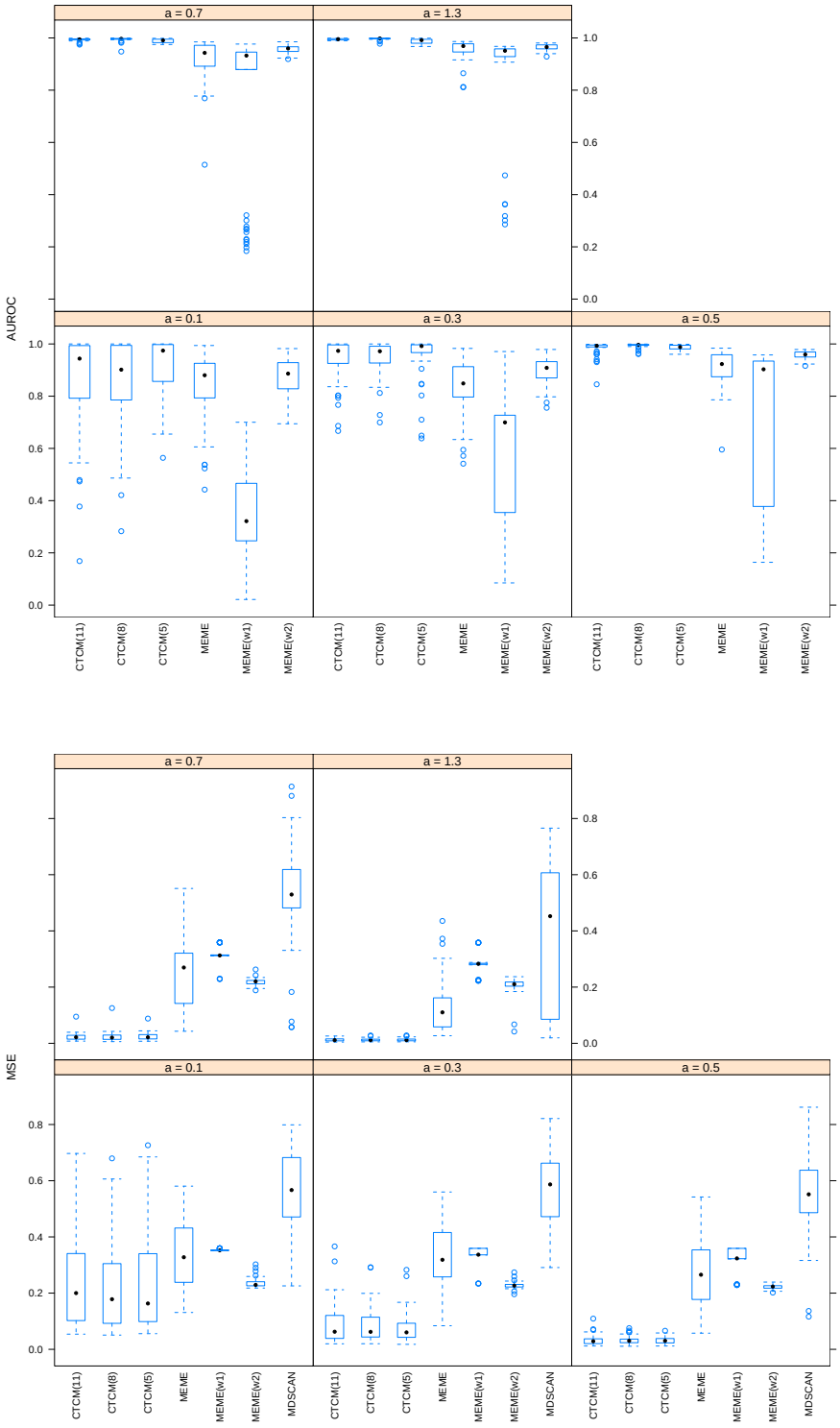

Figure 6: *Simulation results with the Broad-complex_2 position weight matrix.* Top and bottom panel plots display the boxplots of the area under the ROC curve and mean squared errors for various methods at different abundance levels a. CTCM(b) corresponds to CTCM M3 model with a bin size of b.