

Supplementary Materials for “A Markov Random
Field-Based Approach for Joint Estimation of
Differentially Expressed Genes in Mouse
Transcriptome Data”

Zhixiang Lin, Mingfeng Li, Nenad Sestan, and Hongyu Zhao

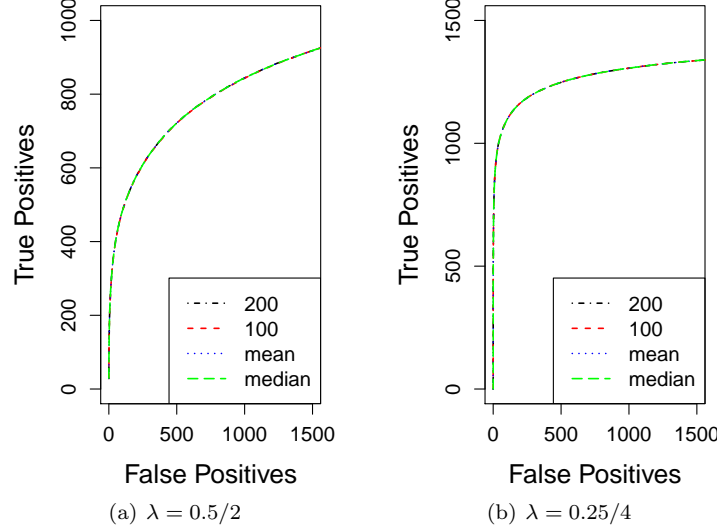


Figure 1S: ROCs comparing different settings of the EM algorithm.

1 Diagnosis for the EM algorithm

We performed simulation studies, where the latent state array was generated as in *Simulation setting 1*. The read counts were then simulated from a negative binomial distribution same as that in the main text. We considered two settings for the fold change of mean read count, where $\lambda = 0.25/4$ or $\lambda = 0.5/2$. We first ran 200 iterations of the EM algorithm. With the parameters fixed, a Gibbs sampler with 20,000 total iterations and 10,000 as burn-in was implemented. We considered four sets of parameters when implementing the Gibbs sampler: a) 100, estimated parameters at the 100th iteration of the EM algorithm; b) 200, estimated parameters at the 200th iteration; c) mean, mean of estimated parameters between the 100th and 200th iterations; d) median, median of estimated parameters between the 100th and 200th iterations. The ROCs comparing the four setting of the EM algorithm are shown in Figure 1S. Boxplots for the AUC are shown in Figure 2S. To calculate AUC, we truncated the ROC at the point where the number of false positives equals the total number of true positives. The results from the four settings are similar. In practice, we use the estimated parameters at the 200th iteration. Trace plots for the estimated parameters in the mouse brain RNA-Seq data are shown in Figure 3S.

2 Hierarchical clustering of the samples

The hierarchical clustering is based on Euclidean distance of $\log_2(RPKM + 1)$.

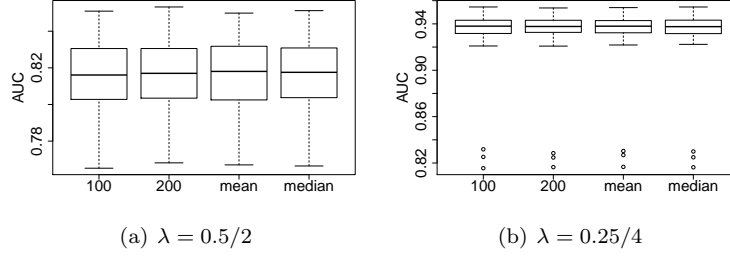


Figure 2S: Boxplots for the AUC comparing different settings of the EM algorithm.

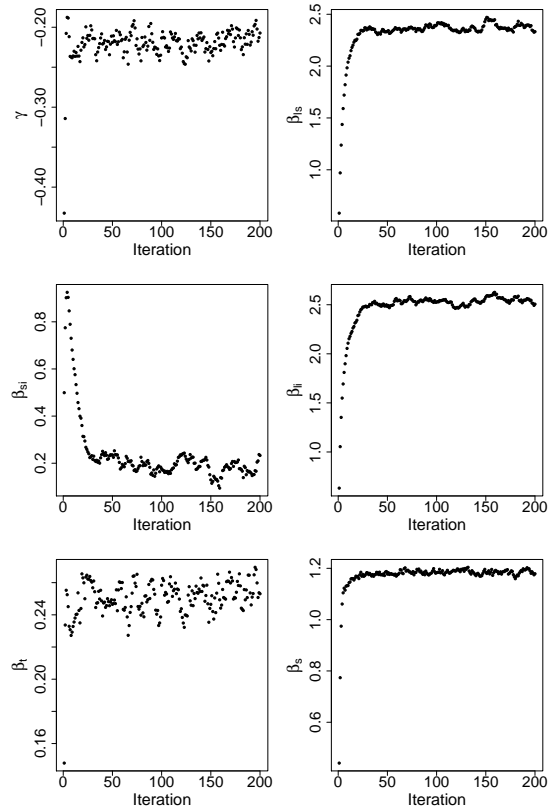


Figure 3S: Trace plots for the estimated parameters in the mouse brain RNA-Seq data.

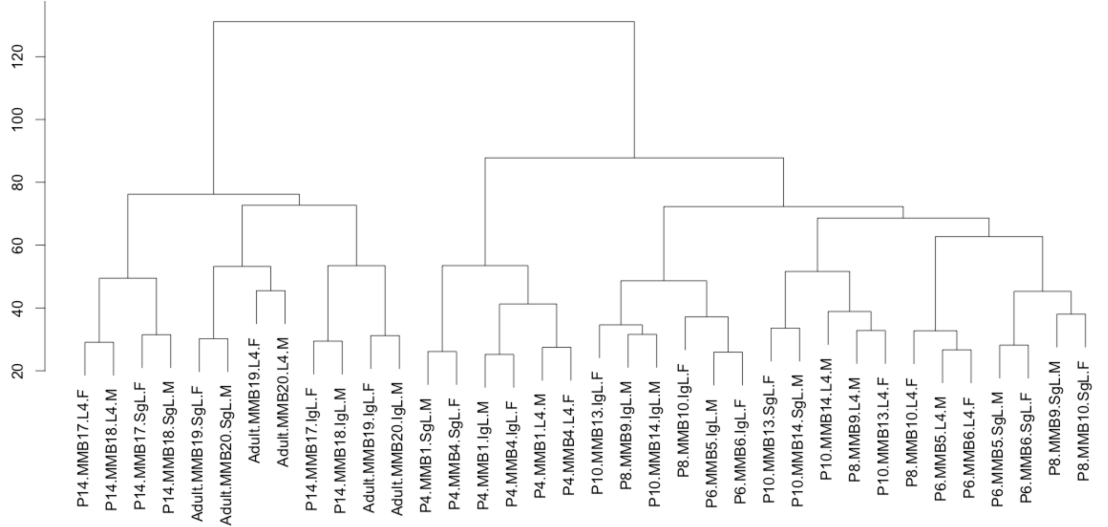


Figure 4S: Hierarchical clustering of the samples.

3 Comparison of parameter estimation

We perform simulation studies to compare the following two methods for parameter estimation: Monte Carlo EM algorithm (MCEM) and EM algorithm with mean field-like approximation (EM-mf). The estimated parameters for Simulation setting 1 MRF and Simulation setting 2 HMM(0.1) are shown in Table 1S. In Table 1S, the numbers in bracket represent the standard deviation in the last 20 iterations.

4 More simulation results

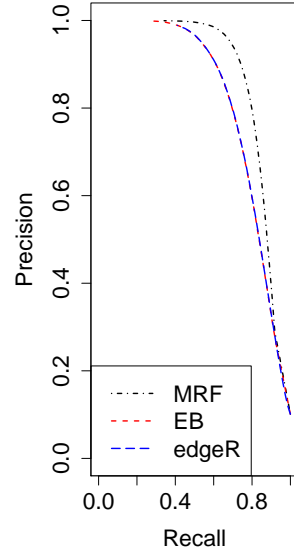
Figures 5S and 6S: more classification measures for the simulation results of 500 genes, including precision-recall curve, and the accuracy curve.

Figure 7S: the achieved false discovery rate vs. power curves.

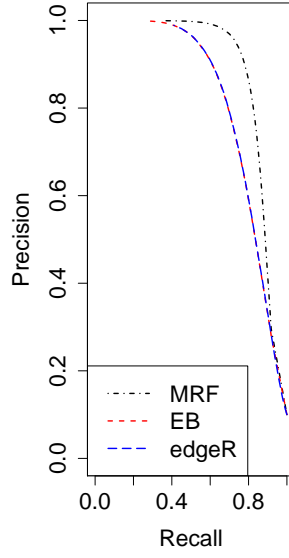
Figure 8S: simulation results of 10,000 genes.

Figure 9S: simulation results with $\sim 20\%$ differentially expressed genes.

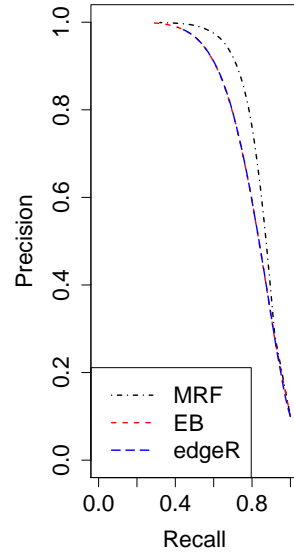
Figure 10S: simulation results with no replicates. To mimic the brain data setting where there are no replicates, we first simulated data with 2 replicates to estimate dispersion; then we used only 1 replicate to estimate differential expression.



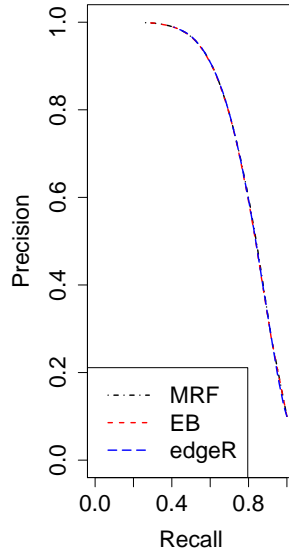
(a) Setting 1



(b) Setting 2: HMM(0.1)

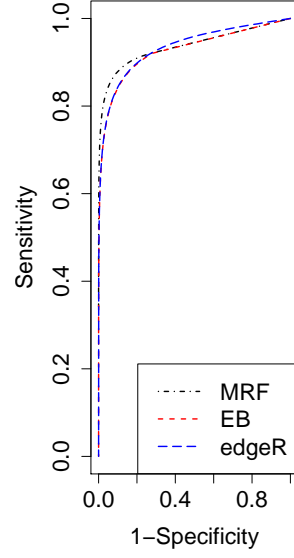


(c) Setting 2: HMM (0.2)

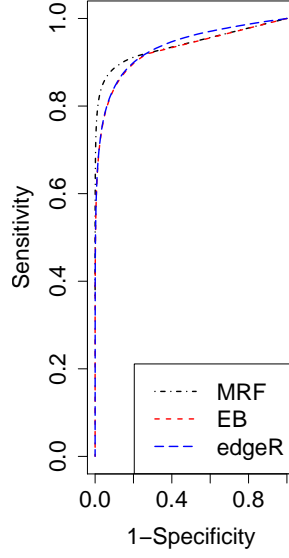


(d) Setting 2: HMM (0.5)

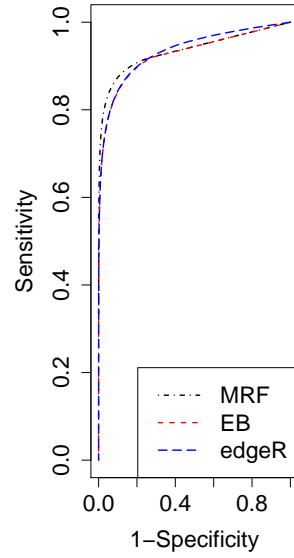
Figure 5S: The precision-recall graph comparing the proposed MRF model, empirical bayesian (EB) and edgeR.



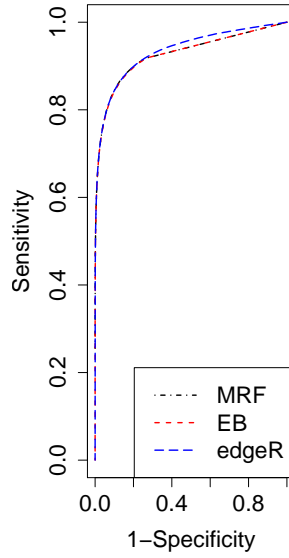
(a) Setting 1



(b) Setting 2: HMM(0.1)



(c) Setting 2: HMM (0.2)



(d) Setting 2: HMM (0.5)

Figure 6S: The accuracy curve comparing the proposed MRF model, empirical bayesian (EB) and edgeR.

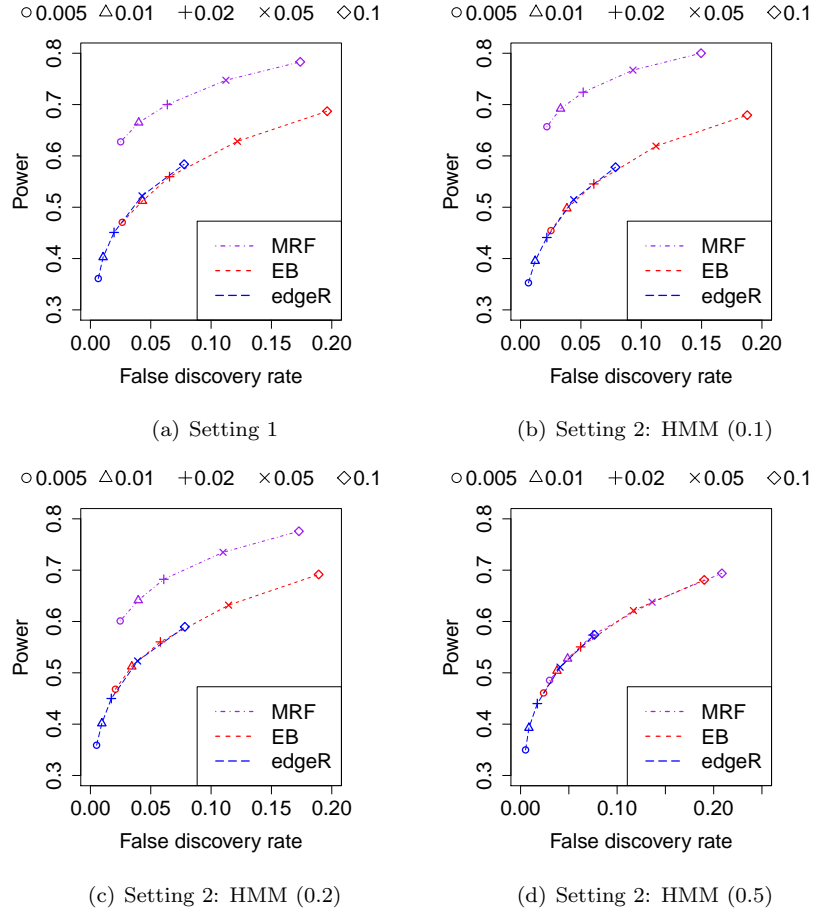
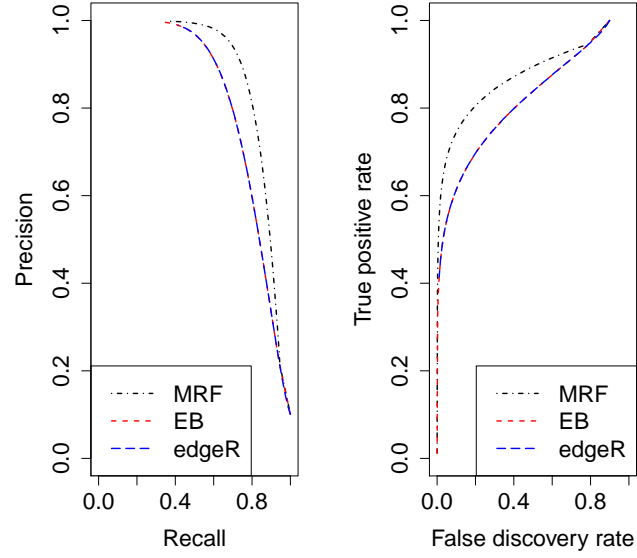
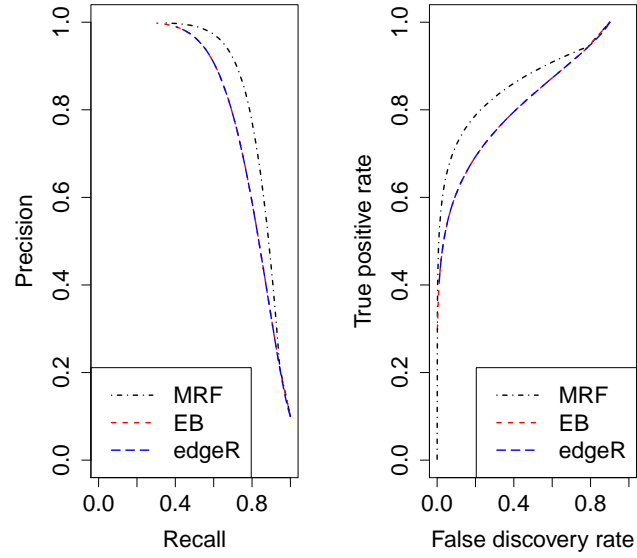


Figure 7S: The achieved false discovery rate vs. power curves. The values of the nominal α , represented by different shapes, are shown on the top of each plot.



(a) Setting 1

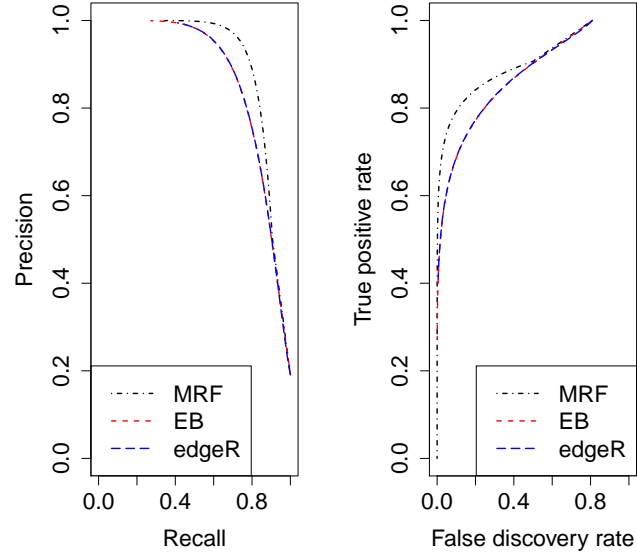
(b) Setting 1



(c) Setting 2: HMM (0.2)

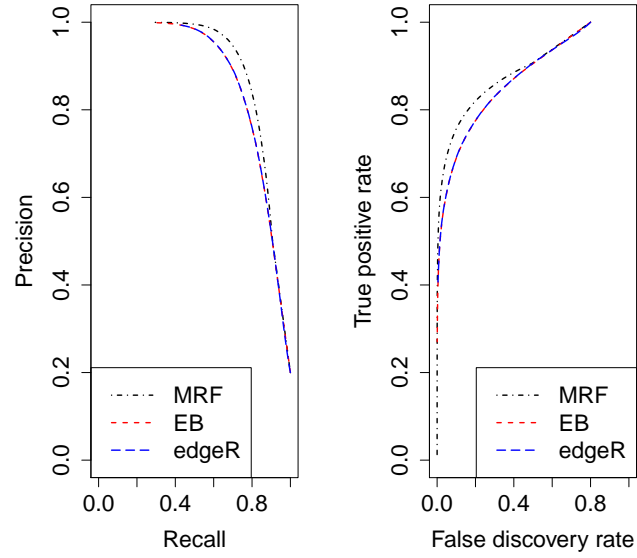
(d) Setting 2: HMM (0.2)

Figure 8S: Simulation with 10,000 genes, comparison of the proposed MRF model, empirical bayesian (EB) and edgeR.



(a) Setting 1

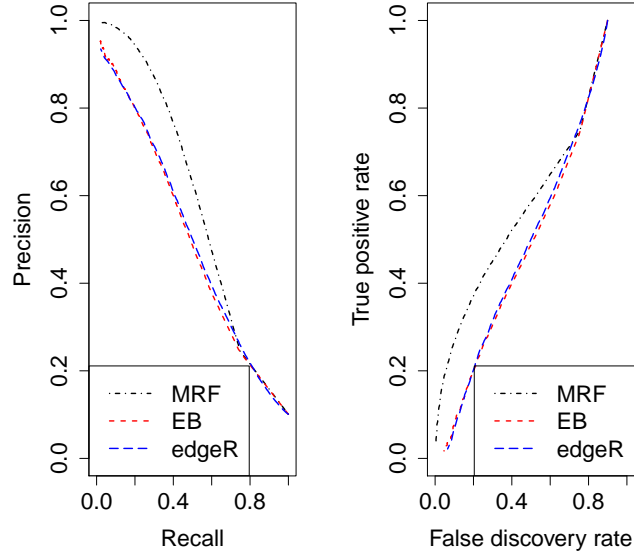
(b) Setting 1



(c) Setting 2: HMM (0.2)

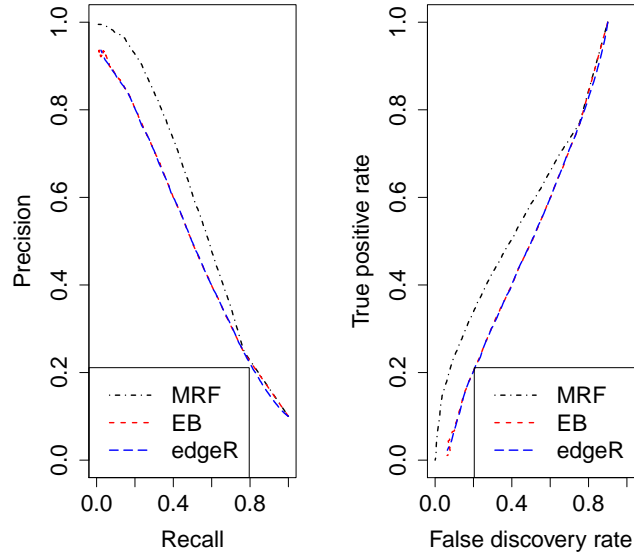
(d) Setting 2: HMM (0.2)

Figure 9S: Simulation with $\sim 20\%$ DE genes (500 genes in total), comparison of the proposed MRF model, empirical bayesian (EB) and edgeR.



(a) Setting 1

(b) Setting 1



(c) Setting 2: HMM (0.2)

(d) Setting 2: HMM (0.2)

Figure 10S: Simulation with no replicates (500 genes in total), comparison of the proposed MRF model, empirical bayesian (EB) and edgeR.

Table 1S: Comparison of MCEM and EM-mf.

Simulation	Parameter	MCEM	EM-mf
MRF	γ	-0.233(0.003)	-0.232(0.038)
	β_{ls}	1.761(0.004)	1.766(0.046)
	β_{si}	0.271(0.006)	0.259(0.041)
	β_{li}	1.686(0.054)	1.711(0.048)
	β_t	0.341(0.002)	0.343(0.024)
	β_s	0.983(0.002)	0.990(0.025)
HMM(0.1)	γ	-0.309(0.002)	-0.313(0.019)
	β_{ls}	0.746(0.002)	0.746(0.016)
	β_{si}	0.687(0.003)	0.685(0.027)
	β_{li}	0.824(0.002)	0.816(0.023)
	β_t	0.479(0.002)	0.482(0.018)
	β_s	0.898(0.002)	0.896(0.014)