Web-based Supplementary Materials for "Bayesian Hidden Markov Models to Identify RNA-Protein Interaction Sites in PAR-CLIP"

Jonghyun Yun, Tao Wang and Guanghua Xiao

Web Appendix A

This section elucidates the Dirichlet prior specification of Markov transition and initial probabilities. The Dirichlet prior is assumed on each row of the transition matrix with the following hyperparameters:

$$
\begin{aligned}
&(\kappa_{11}, \kappa_{12}) \sim \text{Dir}(\alpha_{\kappa_{11}}, \alpha_{\kappa_{12}}),\\
&(\kappa_{T,21}, \kappa_{T,22}, \kappa_{T,23}) \sim \text{Dir}(\alpha_{\kappa_{T,21}}, \alpha_{\kappa_{T,22}}, \alpha_{\kappa_{T,23}}),\\
&(\kappa_{T,32}, \kappa_{T,33}) \sim \text{Dir}(\alpha_{\kappa_{T,32}}, \alpha_{\kappa_{T,33}}),\\
&(\kappa_{N,21}, \kappa_{N,22}) \sim \text{Dir}(\alpha_{\kappa_{N,21}}, \alpha_{\kappa_{N,22}}).
\end{aligned}
$$

For the initial distribution, we have

$$
\begin{aligned}
&(\varphi_{T1}, \varphi_{T2}, \varphi_{T3}) \sim \text{Dir}(\alpha_{\varphi_{T,1}}, \alpha_{\varphi_{T,2}}, \alpha_{\varphi_{T,3}}),\\
&(\varphi_{N1}, \varphi_{N2}) \sim \text{Dir}(\alpha_{\varphi_{N,1}}, \alpha_{\varphi_{N,2}}).
\end{aligned}
$$

Web Appendix B

This section describes the Metropolis algorithm that we employ to sample from (6) in Section 3.4. For state $s = 1$, we employ one dimensional random walks as the Metropolis jumping rule on $\text{logit}\,\mu_1^{(t)}$ as follows:

$$\text{logit}\,\mu_1{}^* \sim \text{N}(\text{logit}\,\mu_1{}^{(t)}, \sigma),$$

where σ is chosen to be $\sqrt{0.005}$. For state $s = 2$, two dimensional random walks are employed on $(\text{logit}\,\mu_2{}^{(t)}, \log\eta_2{}^{(t)})$ as follows:

$$[\text{logit}\,\mu_2{}^*, \log\eta_2{}^*] \sim \text{N}([\text{logit}\,\mu_2{}^{(t)}, \log\eta_2{}^{(t)}], \sigma \cdot \mathbf{I}_2),$$

where $\mathbf{I}_2$ is a 2 by 2 identity matrix.

Noting that the proposed sample can be accepted only if it is contained in the constrained parameter space, the acceptance probability for $\text{logit}\,\mu_1{}^*$ is

$$\min\left(1, \frac{\mathbf{1}_{\text{C}_{\theta^{(t)}}}(\mu_1^*)\,\text{N}(\text{logit}\,\mu_1{}^*|\nu_{\mu_1}, \sigma_{\mu_1})\prod_{I_{ij}=1}\text{BG}_c(X_{ij}|\mu_1^*, 0)}{\text{N}(\text{logit}\,\mu_1{}^{(t)}|\nu_{\mu_1}, \sigma_{\mu_1})\prod_{I_{ij}=1}\text{BG}_c(X_{ij}|\mu_1^{(t)}, 0)}\right),$$

and the acceptance probability for $(\text{logit}\,\mu_2{}^*, \log\eta_2{}^*)$ is

$$\min\left(1, \frac{\mathbf{1}_{\text{C}_{\theta^{(t+1)}}}(\mu_2^*, \eta_2^*)\,\text{N}(\text{logit}\,\mu_2{}^*|\nu_{\mu_2}, \sigma_{\mu_2})\,\text{N}(\log\eta_2{}^*|\nu_{\eta_2}, \sigma_{\eta_2})\prod_{s=2}^{3}\prod_{I_{ij}=s}\text{BG}_c(X_{ij}|\mu_2^*, \eta_2^*)}{\text{N}(\text{logit}\,\mu_2{}^{(t)}|\nu_{\mu_2}, \sigma_{\mu_2})\,\text{N}(\log\eta_2{}^{(t)}|\nu_{\eta_2}, \sigma_{\eta_2})\prod_{s=2}^{3}\prod_{I_{ij}=s}\text{BG}_c(X_{ij}|\mu_2^{(t)}, \eta_2^{(t)})}\right).$$

Web Appendix C

Let $\mathbf{Y}_{[1:i]j}$ and $\mathbf{I}_{[1:i]j}$ be the sequence of observations and hidden states, respectively, from the initial to the i-th location of region j. The first step is to evaluate forward probabilities

$$\begin{aligned}\phi_{ijrs} &\propto p(I_{i-1j}=r, I_{ij}=s, \mathbf{Y}_{ij}|\mathbf{N}, \mathbf{Y}_{[1:i-1]j}, \theta, \mathbf{K}, \varphi) \\ &= p(I_{i-1,j}=r|N_{i-1,j}, \mathbf{Y}_{[1:i-1]j}, \theta, \mathbf{K}, \varphi) \\ &\quad \times p(I_{ij}=s|N_{ij}, I_{i-1j}=r) p(\mathbf{Y}_{ij}|N_{ij}, I_{ij}=s, \theta),\end{aligned} \tag{S1}$$

where the normalizing constant is obtained by summing (S1) over r and s. In (S1) above, the second density is the nucleotide-dependent transition probability from r to s, and the third density can be computed from the likelihood function. The first density in (S1) can be evaluated recursively as

$$p(I_{i-1,j}=r|N_{i-1,j}, \mathbf{Y}_{[1:i-1]j}, \theta, \mathbf{K}, \varphi) = \sum_{k=1}^{3} \phi_{i-1jkr}.$$

The second step is to sample I_{ij}'s from the posterior in the backward direction starting from each right terminal position n_j of Markov chains. For $k = 0, \ldots, n_j - 1$, we draw $I_{n_j-k,j}$ from

$$\begin{aligned}&p(I_{n_j-k,j}=r|\mathbf{N}, \mathbf{Y}, \theta, \mathbf{I}_{[n_j-k+1:n_j]j}, \mathbf{K}, \varphi) \\ &\quad = p(I_{n_j-k,j}=r|\mathbf{N}, \mathbf{Y}_{[1:n_j-k+1]j}, \theta, \mathbf{I}_{[n_j-k+1:n_j]j}, \mathbf{K}, \varphi) \\ &\quad \propto \phi_{n_j-k+1,j,r,I_{k+1,j}}.\end{aligned}$$

Web Appendix D

Let $n_{T,rs}$ and $n_{N,rs}$ be the numbers of transitions from r to s when transitions occur onto genomic locations with T nucleotide and with non-T nucleotide, respectively. Also, let $n_{\varphi,Ts}$ and $n_{\varphi,Ns}$ be the number of initial state s at the two types of nucleotide. The conditional distribution of the Markov transition kernels is

$$
\begin{aligned}
& p(\mathbf{K}, \varphi | \mathbf{Y}, \theta, \mathbf{I}) \\
& \quad \propto p(\mathbf{I}|\mathbf{K}, \varphi) p(\mathbf{K}) p(\varphi) \\
& \quad \propto \kappa_{11}^{\alpha_{\kappa_{11}}+n_{11}-1} \kappa_{12}^{\alpha_{\kappa_{12}}+n_{12}-1} \\
& \qquad \times \kappa_{T,21}^{\alpha_{\kappa_{T,21}}+n_{T,21}-1} \kappa_{T,22}^{\alpha_{\kappa_{T,22}}+n_{T,22}-1} \kappa_{T,23}^{\alpha_{\kappa_{T,23}}+n_{T,23}-1} \\
& \qquad \times \kappa_{T,32}^{\alpha_{\kappa_{T,32}}+n_{T,32}-1} \kappa_{T,33}^{\alpha_{\kappa_{T,33}}+n_{T,33}-1} \\
& \qquad \times \kappa_{N,21}^{\alpha_{\kappa_{N,21}}+n_{N,21}-1} \kappa_{N,22}^{\alpha_{\kappa_{N,22}}+n_{N,22}-1} \\
& \qquad \times \varphi_{T1}^{\alpha_{\varphi_{T1}}+n_{\varphi,T1}-1} \varphi_{T2}^{\alpha_{\varphi_{T2}}+n_{\varphi,T2}-1} \varphi_{T3}^{\alpha_{\varphi_{T3}}+n_{\varphi,T3}-1} \\
& \qquad \times \varphi_{N1}^{\alpha_{\varphi_{N1}}+n_{\varphi,N1}-1} \varphi_{N2}^{\alpha_{\varphi_{N2}}+n_{\varphi,N2}-1}.
\end{aligned}
$$

Web Appendix E

Trace plots of MCMC sequences of eight parameters are presented in Figure S1. All the sequences quickly reach approximate stationary distributions.

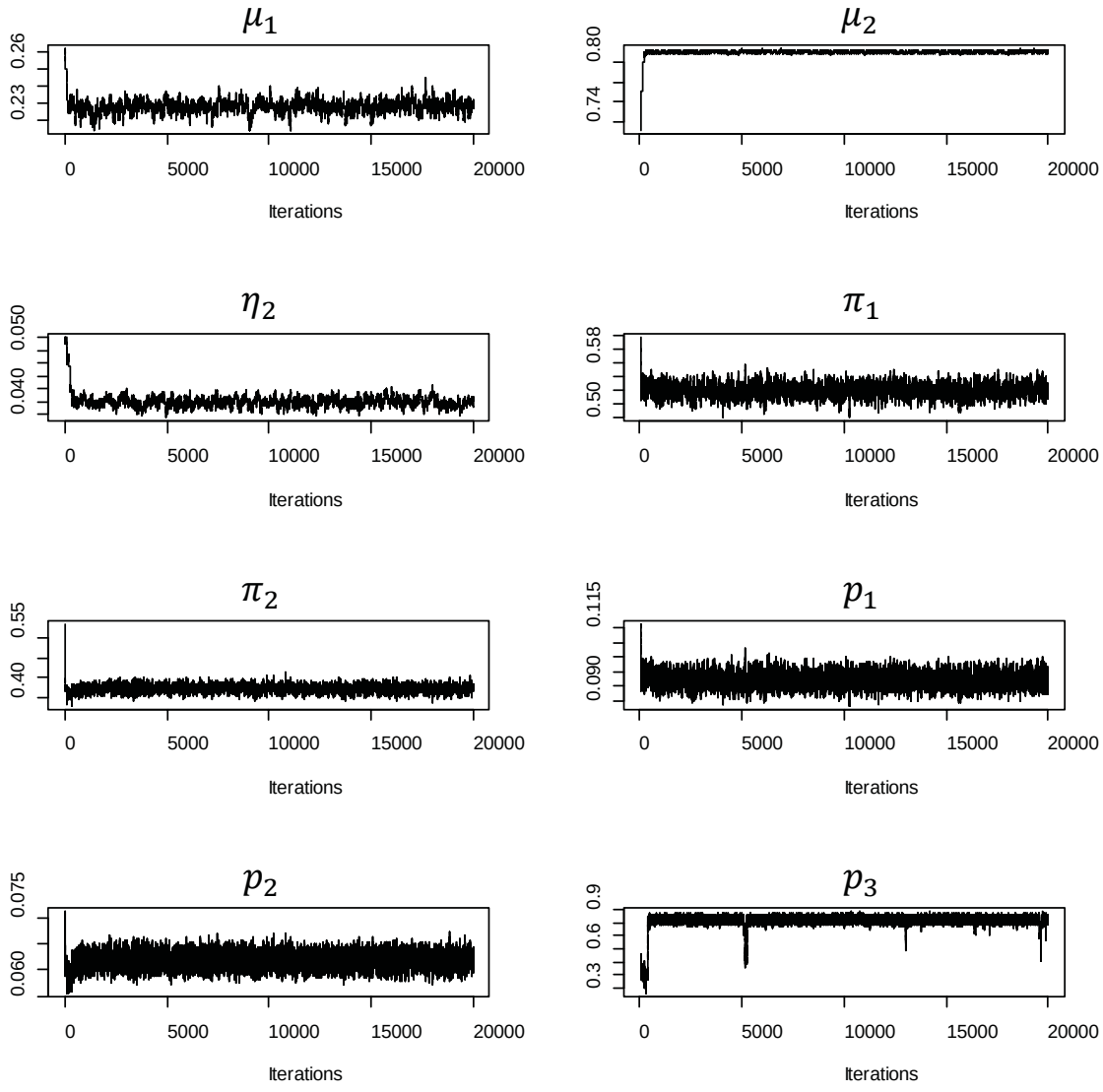

Figure S1: Trace plots for MCMC sequences of eight parameters.

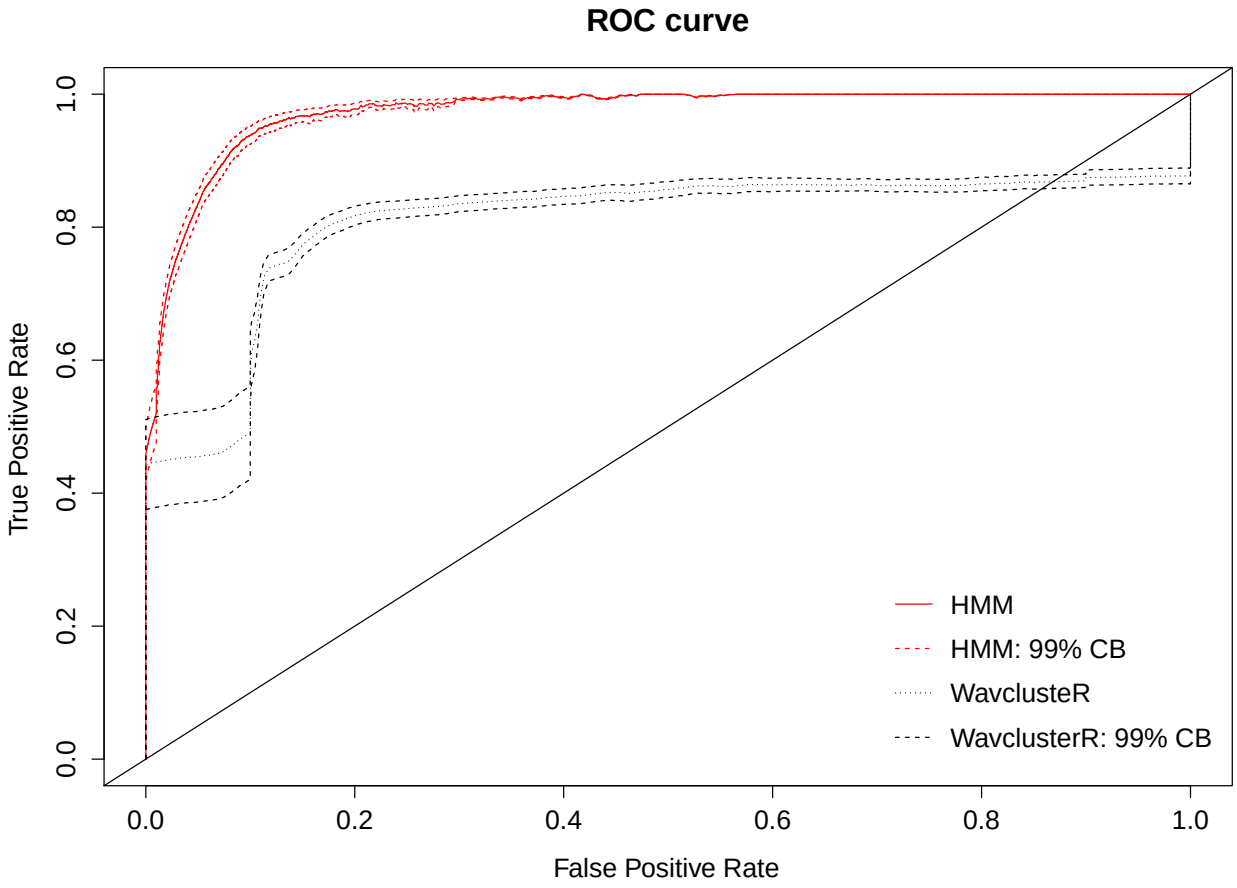

Figure S2: The smoothed ROC curves (based on sites with $M_{ij} > 0$) for the HMM (solid line) and the wavClusteR (dotted line) to identify binding sites based on 100 simulated datasets. The dashed lines represent 99 % confidence bands of each curve. The local constant regressions with truncated Gaussian kernels are used for smoothing.

Web Appendix F

The 2nd stage of the wavClusteR employs the wavelet analysis on regions containing non-experimentally induced mutation sites identified by the 1st stage. The 2nd stage filters out regions with no peaks, and identifies binding regions. In the simulation study, however, we compare the performance of identifying binding sites, not regions containing binding sites. Also, no method to rank the identified regions has been proposed by Sievers et al., which makes the comparison in this study unclear.

Regarding to Simulation Study 1 in Section 4.1, the smoothed ROC curves based on sites with $M_{ij} > 0$ are presented in Figure S2. Based on all simulated observations ($M_{ij} \geq 0$),

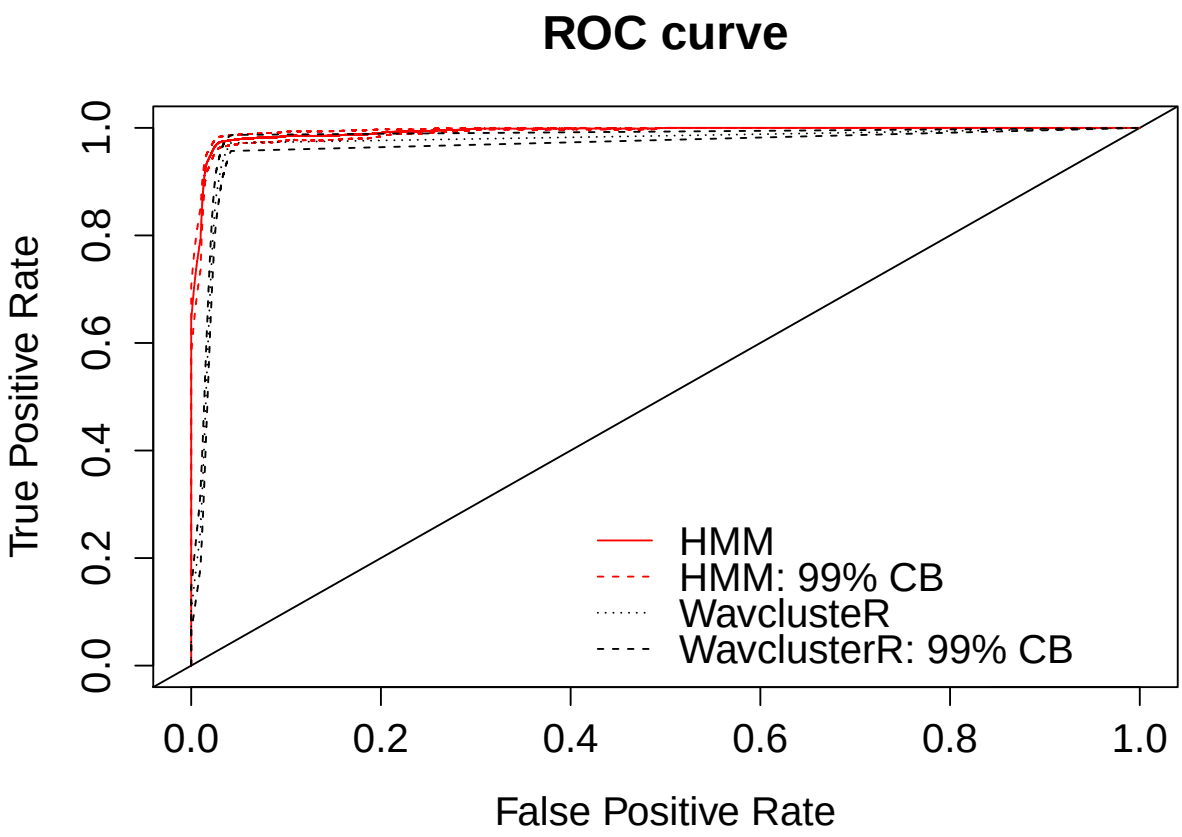

Figure S3: The raw ROC curves (based on all observations) for the HMM (solid line) and the wavClusteR (dotted line) to identify binding sites based on 100 simulated datasets. The dashed lines represent 99 % confidence bands of each curve. Local constant regressions with truncated Gaussian kernels are used for smoothing.

the average AUC of the HMM, the wavClusteR and the difference between the two methods ($\text{AUC}_{\text{HMM}} - \text{AUC}_{\text{WCR}}$) are 0.9937 (s.e. 0.0005), 0.9783 (s.e. 0.0013) and 0.0153 (s.e. 0.0014), respectively. The p-value of the Wilcoxon rank test with an one-sided alternative is less then 0.0001. The smoothed raw ROC curves based on all sites are presented in Figure S3.

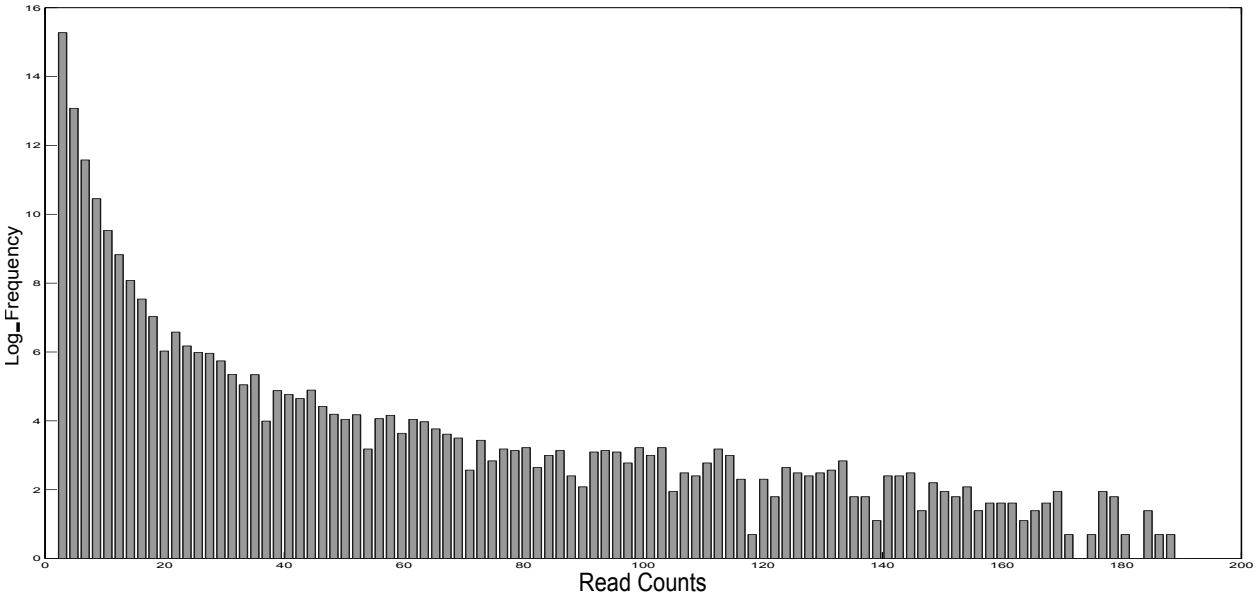

Figure S4: A histogram of read counts $X_{ij} \geq 1$ in the data described in Section 2. Log-transformed frequencies are used to avoid the scale problem in raw frequencies.

Web Appendix G

Often, the NB likelihood is interpreted as an expansion of the Poisson likelihood to take the overdispersion in the count data into account. In our empirical study, The NB likelihood outperforms the Poisson likelihood for fitting the read count X_{ij}. Along with the PPC under the HMM framework, furthermore, we demonstrate that the BG likelihood fits X_{ij} better than the NB likelihood by evaluating the goodness of fit for the three likelihoods based on the real dataset described in Section 2.

We present a histogram of read counts $X_{ij} \geq 1$ with log-frequencies (Figure S4) and the goodness of fit for the BG, NB and Poisson models (Table S1). The data described in Section 2 is used. For all truncation cutoffs we consider, the BG model achieves the greatest likelihood among the three models.

Table S1: Log-likelihoods for three models (the BG, NB and Poisson) on $X_{ij} \geq c$ are presented with the truncation cutoff $c = 1, \ldots, 5$.

	$c=1$	$c=2$	$c=3$	$c=4$	$c=5$
BG (df=2)	-5166142	-2074198	-1010287	-559104	-339918
NB (df=2)	-5236154	-2116632	-1037521	-576601	-351437
Poisson (df=1)	-6784937	-7066748	-2295065	-1377220	-790234

Web Appendix H

We measure the goodness of fit for two-component mixture models equipped with Poisson, negative binomial (NB) and beta geometric (BG) likelihoods on the read count $X_{ij} > c$ in the dataset described in Section 2. We fix the shape parameter $\eta_1 = 0$ in the BG mixture model. As shown in Table S2, the NB mixture outperforms the Poisson mixture in all settings, and the BG mixture outperforms the NB mixture in settings with $c \leq 5$, which preserve a majority of samples. The capability of the BG distribution that controls the high-densities on small counts and the tail thickness enhances the mixture BG model to fit the read count distribution.

Table S2: The log-likelihood (top), AIC (middle) and BIC (bottom) with $c = 1, \ldots, 6$ are presented for three models: two-component Poisson, NB and BG mixture models. Each read-marked performance criterion indicates models of best performance for each c and performance criteria.

c		Poisson (df=3)	NB (df=5)	BG (df=4)
	Log-like	-309854	-223392	-223434
6	AIC	619714	446794	446877
	BIC	619743	446841	446915
	log-like	-454145	-340061	-340065
5	AIC	908296	680133	680139
	BIC	908326	680183	680179
	Log-like	-714917	-559266	-559067
4	AIC	1429841	1118543	1118143
	BIC	1429873	1118596	1118186
	Log-like	-1219542	-1010715	-1010045
3	AIC	2439090	2021440	2020098
	BIC	2439124	2021496	2020143
	Log-like	-2355468	-2075873	-2074269
2	AIC	4710943	4151757	4148546
	BIC	4710977	4151818	4148595
	Log-like	-5552976	-5167805	-5163827
1	AIC	11105959	10335621	10327663
	BIC	11105999	10335688	10327717

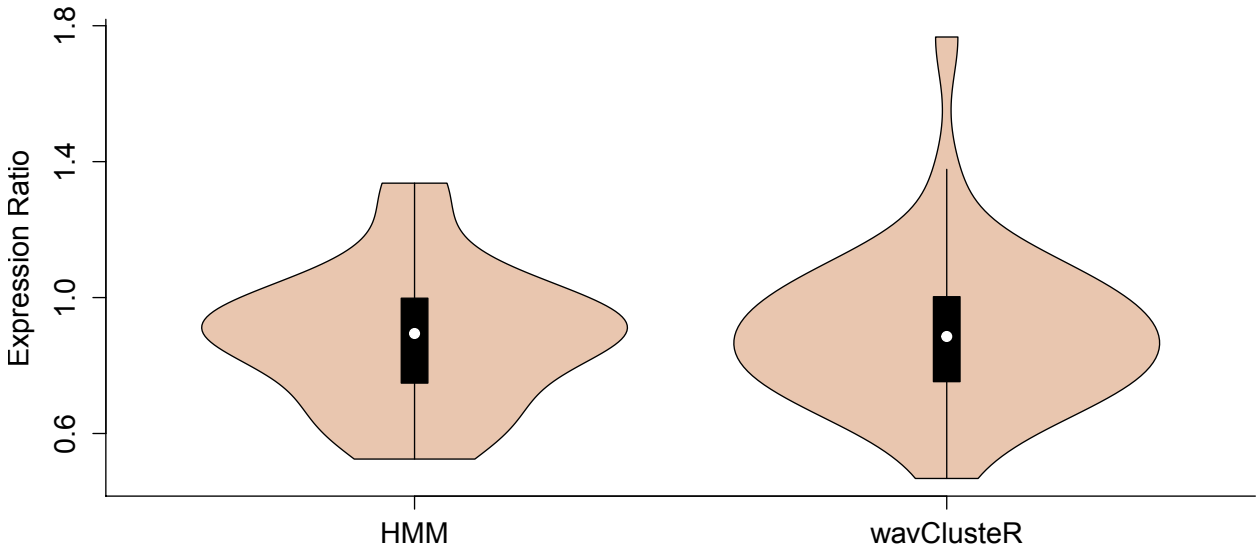

Figure S5: Violin plots of expression ratios are presented. For each method, top 35 binding sites which are found in the expression ratio matrix are used.

Web Appendix I

Another comparison is made by examining the expression ratios of top 35 identified binding sites whose expression ratios are found in the expression ratio matrix. The average expression ratios of these sites are 0.8847 (s.e. 0.0331) for the HMM and 0.8983 (s.e. 0.1518) for the wavClusteR, respectively. The t-tests are carried out on these sites to see if the mean expression ratios are less than 1. The p-value based on our result is 0.0007, whereas the p-value for the wavClusteR is 0.2537. The list of the 35 binding sites and the ratios are presented in Table S3, and violin plots of the ratios are presented in Figure S5. Overall, our method provides the identification results with smaller expression ratios (< 1), which implies our results show stronger evidence of the RNA-RBP interaction.

In Section 5, the posterior mean of read counts on enriched locations is 11.5, and the posterior mean of the mutation probability on true mutation sites is 0.8224. For enriched locations with T nucleotide, the posterior mean of the mutation-read ratio is 0.0248. Our method identifies enriched sites and true mutation sites simultaneously. However, peaks without enough mutations will not be identified as binding sites. A region containing the identified binding site is presented in Figure S6. Among enriched locations, our method

Table S3: The expression ratios of top identified binding sites which are found in the expression ratio matrix. Chromosome (Chr), genomic position (GP) and expression ratio (ER) are presented for our method (HMM) and the wavClusteR.

HMM			wavClusteR		
Chr	GP	ER	Chr	GP	ER
chr20	5991931	0.8971	chr11	65273253	1.7669
chr2	133013144	1.3367	chr1	193054752	1.0269
chr16	47538729	0.9963	chr12	1310444	0.9377
chr8	70602341	1.0679	chr5	170423513	0.8175
chr6	30582101	0.9503	chr7	99228102	0.8284
chr13	79894144	0.8779	chrX	84502511	0.8366
chr1	243998148	0.5829	chr7	135321862	0.9640
chr8	21779574	0.9459	chr12	69667727	0.8985
chr8	25364928	0.9042	chr5	65291534	1.0876
chr13	20298318	0.8792	chr1	145509180	0.9674
chr8	100135779	0.8927	chr15	99281763	1.1299
chr9	33919729	0.8154	chr4	39734173	0.7533
chr3	48975677	0.6432	chrX	103435815	0.9465
chr5	115204870	0.5248	chr8	26190602	0.8477
chr2	133012147	1.3367	chr19	39221097	0.9310
chr17	30852610	1.1166	chr10	27457826	0.6747
chr5	10265006	0.9001	chr19	44601427	0.7396
chr13	50648655	1.0000	chr10	13252859	0.7518
chr10	22974076	0.7302	chr18	57569092	0.6678
chr19	44937168	1.0000	chr3	160123468	0.6795
chr7	156990101	1.0235	chr11	120215430	0.7185
chr5	140700257	0.9321	chr14	55513272	0.7201
chr13	50650083	1.0000	chr18	13030777	0.7789
chr14	55755090	0.8749	chr5	56553359	0.9051
chr3	152099252	0.8828	chr3	128489715	0.7648
chr2	181865037	0.6673	chr2	153616778	0.5900
chr18	54278124	0.7665	chr8	33357258	0.8855
chr4	139086944	0.8943	chr17	34155410	1.3324
chr1	33151486	0.6511	chr12	4665775	0.4676
chr14	97004355	0.6306	chr4	108835835	1.0826
chr20	25597193	1.2463	chr7	116617491	1.0677
chrX	84499094	0.8366	chr4	108835574	1.0826
chr12	101013504	0.5541	chr21	38800124	0.8610
chr14	60761098	0.6947	chr19	36719613	1.0349
chr6	21594140	0.9142	chr12	69233706	0.9780

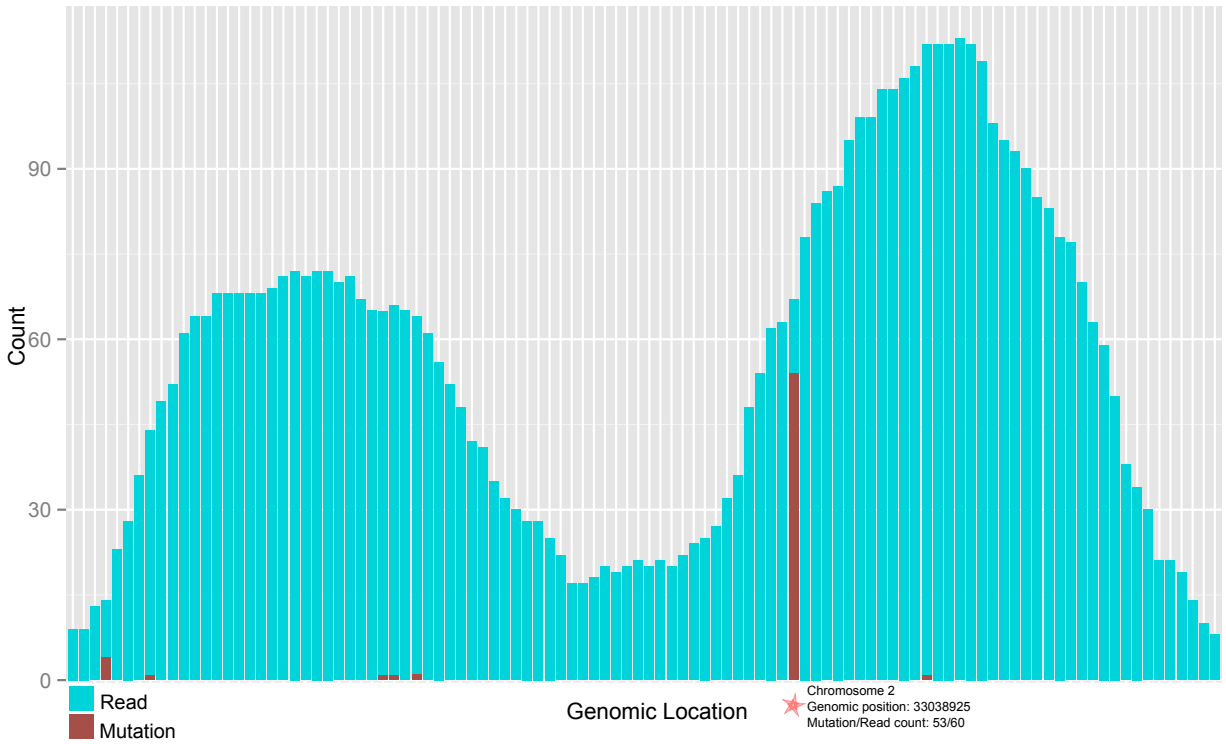

Figure S6: A barplot of mutation and read counts on the region containing one binding site (marked with a star, genomic location: 33038925 in chromosome 2).

pinpoints the location (marked with a star in the figure) whose mutation-read ratio is close enough to the true mutation probability, and identifies it as the binding site.