

Supplementary Materials for “A Bayesian Approach to Joint Modeling of Protein-DNA Binding, Gene Expression and Sequence Data”

by Yang Xie, Wei Pan, Kyeong S. Jeong, Guanghua Xiao, and Arkady B. Khodursky

1 Computation Details

1.1 The prior distributions for hyperparameters for joint modeling:

$$\tau_{0x}^2 \sim IG(0.01, 0.01)$$

$$\tau_{1x}^2 \sim IG(0.01, 0.01)$$

$$\lambda_x \sim \logNormal(0, 100)$$

$$p_x \sim Beta(100, 900)$$

$$\tau_{0y}^2 \sim IG(0.01, 0.01)$$

$$\tau_{1y}^2 \sim IG(0.01, 0.01)$$

$$p_{y0}^{\vec{}} \sim Dirichlet(1, 1, 1)$$

$$p_{y1}^{\vec{}} \sim Dirichlet(1, 1, 1)$$

$$\lambda_y \sim \logNormal(0, 100)$$

$$\tau_{0z}^2 \sim IG(0.01, 0.01)$$

$$\tau_{1z}^2 \sim IG(0.01, 0.01)$$

$$p_{z0} \sim Beta(1, 1)$$

$$p_{z1} \sim Beta(1, 1)$$

$$\lambda_{z1} \sim \logNormal(0, 100)$$

$$\lambda_{z2} = \lambda_{z1} + d_z$$

$$d_z \sim \logNormal(0, 100).$$

1.2 The conditional posterior distributions:

$$\begin{aligned}
\mu_{ix}|I_{ix} = 0, \cdot &\sim N\left(\frac{n\tau_{0x}^2\bar{x}_{ij}}{n\tau_{0x}^2 + \sigma_{ix}^2}, \frac{n\tau_{0x}^2\sigma_{ix}^2}{n\tau_{0x}^2 + \sigma_{ix}^2}\right) \\
\mu_{ix}|I_{ix} = 1, \cdot &\sim N\left(\frac{n\tau_{1x}^2\bar{x}_{ij} + \lambda_x\sigma_{ix}^2}{n\tau_{1x}^2 + \sigma_{ix}^2}, \frac{n\tau_{1x}^2\sigma_{ix}^2}{n\tau_{1x}^2 + \sigma_{ix}^2}\right) \\
\sigma_{ix}^2|\cdot &\sim IG\left(\frac{n}{2} + \alpha, \beta + \frac{1}{2}\sum_{j=1} n(x_{ij} - \mu_{ix})^2\right) \\
I_{ix}|\cdot &\sim Ber(p_{ix}) \\
\left\{ \begin{aligned} p_{ix} &= \frac{A}{A+B} \\ A &= p_x(\tau_{1x}^2)^{-\frac{1}{2}} \exp\left(-\frac{(\mu_{ix} - \lambda_x)^2}{2\tau_{1x}^2}\right) p_{y01}^{I_{iY}=0} p_{y11}^{I_{iY}=1} p_{y21}^{I_{iY}=2} p_{z1}^{I_{iz}=1} (1 - p_{z1})^{I_{iz}=0} \\ B &= (1 - p_x)(\tau_{0x}^2)^{-\frac{1}{2}} \exp\left(-\frac{\mu_{ix}^2}{2\tau_{0x}^2}\right) p_{y00}^{I_{iY}=0} p_{y10}^{I_{iY}=1} p_{y20}^{I_{iY}=2} p_{z0}^{I_{iz}=1} (1 - p_{z0})^{I_{iz}=0} \end{aligned} \right\} \\
\tau_{0x}^2|\cdot &\sim IG\left(\frac{G_{0x}}{2} + \alpha, \beta + \frac{1}{2}\sum_{i=1}^G \mu_{ix}^2 1(I_{ix} = 0)\right) \\
\tau_{1x}^2|\cdot &\sim IG\left(\frac{G_{1x}}{2} + \alpha, \beta + \frac{1}{2}\sum_{i=1}^G (\mu_{ix} - \lambda_x)^2 1(I_{ix} = 1)\right) \\
p_x|\cdot &\sim Beta(G_{1x} + 0.1 \times G, G_{x0} + 0.9 \times G) \\
\mu_{iy}|I_{iy} = 0, \cdot &\sim N\left(\frac{n_y\tau_{0y}^2\bar{y}_{ij}}{n_y\tau_{0y}^2 + \sigma_{iy}^2}, \frac{n_y\tau_{0y}^2\sigma_{iy}^2}{n_y\tau_{0y}^2 + \sigma_{iy}^2}\right) \\
\mu_{iy}|I_{iy} = 1, \cdot &\sim N\left(\frac{n_y\tau_{1y}^2\bar{y}_{ij} + \lambda_y\sigma_{iy}^2}{n_y\tau_{1y}^2 + \sigma_{iy}^2}, \frac{n_y\tau_{1y}^2\sigma_{iy}^2}{n_y\tau_{1y}^2 + \sigma_{iy}^2}\right) \\
\mu_{iy}|I_{iy} = 2, \cdot &\sim N\left(\frac{n_y\tau_{1y}^2\bar{y}_{ij} - \lambda_y\sigma_{iy}^2}{\tau_{1y}^2\bar{y}_{ij} + \sigma_{iy}^2}, \frac{n_y\tau_{1y}^2\sigma_{iy}^2}{\tau_{1y}^2\bar{y}_{ij} + \sigma_{iy}^2}\right) \\
\sigma_{iy}^2|\cdot &\sim IG\left(\frac{n_y}{2} + \alpha, \beta + \frac{1}{2}\sum_{j=1} n_y(y_{ij} - \mu_{iy})^2\right) \\
I_{iy}|\cdot &\sim Multinomial(p_{iy0}, p_{iy1}, p_{iy2}) \\
\left\{ \begin{aligned} p_{iy0} &= \frac{A}{A+B+C} \\ A &= (1 - p_y)(\tau_{0y}^2)^{-\frac{1}{2}} \exp\left(-\frac{\mu_{iy}^2}{2\tau_{0y}^2}\right) p_{y00}^{I_{ix}=0} p_{y01}^{I_{ix}=1} \\ B &= p_y(\tau_{1y}^2)^{-\frac{1}{2}} \exp\left(-\frac{(\mu_{iy} - \lambda_y)^2}{2\tau_{1y}^2}\right) p_{y10}^{I_{ix}=0} p_{y11}^{I_{ix}=1} \end{aligned} \right.
\end{aligned}$$

$$\begin{aligned}
C &= p_y(\tau_{1y}^2)^{-\frac{1}{2}} \exp\left(-\frac{(\mu_{iy} + \lambda_y)^2}{2\tau_{1y}^2}\right) p_{y20}^{I_{ix}=0} p_{y21}^{I_{ix}=1} \quad \} \\
\tau_{0y}^2 | \cdot &\sim IG\left(\frac{G_{0y}}{2} + \alpha, \beta + \frac{1}{2} \sum_{i=1}^G \mu_{iy}^2 1(I_{iy} = 0)\right) \\
\tau_{1y}^2 | \cdot &\sim IG\left(\frac{G_{1y}}{2} + \alpha, \beta + \frac{1}{2} \sum_{i=1}^G (\mu_{iy} - \lambda_y)^2 1(I_{iy} = 1) + \frac{1}{2} \sum_{i=1}^G (\mu_{iy} + \lambda_y)^2 1(I_{iy} = 2)\right) \\
p_{y0} | \cdot &\sim \text{Dirichlet}\left(\sum_{i=1}^G 1(I_{ix} = 0, I_{iy} = 0), \sum_{i=1}^G 1(I_{ix} = 0, I_{iy} = 1), \sum_{i=1}^G 1(I_{ix} = 0, I_{iy} = 2)\right) \\
p_{y1} | \cdot &\sim \text{Dirichlet}\left(\sum_{i=1}^G 1(I_{ix} = 1, I_{iy} = 1), \sum_{i=1}^G 1(I_{ix} = 1, I_{iy} = 1), \sum_{i=1}^G 1(I_{ix} = 1, I_{iy} = 2)\right) \\
I_{iz} | \cdot &\sim \text{Ber}(p_{iz}) \\
\{ \quad p_{iz} &= \frac{A}{A+B} \\
A &= p_x(\tau_{1z}^2)^{-\frac{1}{2}} \exp\left(-\frac{(z_i - \lambda_{z1})^2}{2\tau_{1z}^2}\right) p_{z0}^{I_{ix}=0} p_{z1}^{I_{ix}=1} \\
B &= (1 - p_z)(\tau_{0z}^2)^{-\frac{1}{2}} \exp\left(-\frac{(z_i - \lambda_{z0})^2}{2\tau_{0z}^2}\right) (1 - p_{z0})^{I_{ix}=0} (1 - p_{z10})^{I_{ix}=1} \quad \}
\end{aligned}$$

$$\begin{aligned}
\tau_{0z}^2 | \cdot &\sim IG\left(\frac{G_{0z}}{2} + \alpha, \beta + \frac{1}{2} \sum_{i=1}^G (z_i - \lambda_{z0})^2 1(I_{iz} = 0)\right) \\
\tau_{1z}^2 | \cdot &\sim IG\left(\frac{G_{1z}}{2} + \alpha, \beta + \frac{1}{2} \sum_{i=1}^G (z_i - \lambda_{z1})^2 1(I_{iz} = 1)\right) \\
p_{z0} | \cdot &\sim \text{Beta}\left(\sum_{i=1}^G 1(I_{iz} = 1, I_{ix} = 0), \sum_{i=1}^G 1(I_{iz} = 0, I_{ix} = 0)\right) \\
p_{z1} | \cdot &\sim \text{Beta}\left(\sum_{i=1}^G 1(I_{iz} = 1, I_{ix} = 1), \sum_{i=1}^G 1(I_{iz} = 0, I_{ix} = 1)\right)
\end{aligned}$$

where $\theta | \cdot$ means the distribution of θ condition on all other parameters and the data; $\alpha = 0.01$, $\beta = 0.01$, $G = 3924$ the total number of genes, $n = 5$ the number of replicates for binidng data, $n_y = 6$ the number of replicates for expression data, $G_{0x} = \sum_{i=1}^G 1(I_{ix} = 0)$, $G_{1x} = \sum_{i=1}^G 1(I_{ix} = 1)$, $G_{0y} = \sum_{i=1}^G 1(I_{iy} = 0)$, $G_{1y} = \sum_{i=1}^G 1(I_{iy} = 1)$, $G_{2y} = \sum_{i=1}^G 1(I_{iy} = 2)$, $G_{0z} = \sum_{i=1}^G 1(I_{iz} = 0)$, $G_{1z} = \sum_{i=1}^G 1(I_{iz} = 1)$.

Simulation	$\widehat{Pr}(I_{ix}=1)$	$\widehat{Pr}(DE I_{ix}=0)$	$\widehat{Pr}(DE I_{ix}=1)$	$\widehat{Pr}(I_{iz}=1 I_{ix}=0)$	$\widehat{Pr}(I_{iz}=1 I_{ix}=1)$
truth	0.2	0.05	0.8	0.0	1.0
prior 1	0.214±0.003	0.004±0.001	0.865±0.021	0.152±0.031	0.966±0.006
prior 2	0.155±0.004	0.008±0.004	0.923±0.023	0.167±0.038	0.981±0.005

Table 1: The parameter estimations in simulation study 1. Prior 1: $p_x \sim Beta(200, 800)$; Prior 2: $p_x \sim Beta(100, 900)$.

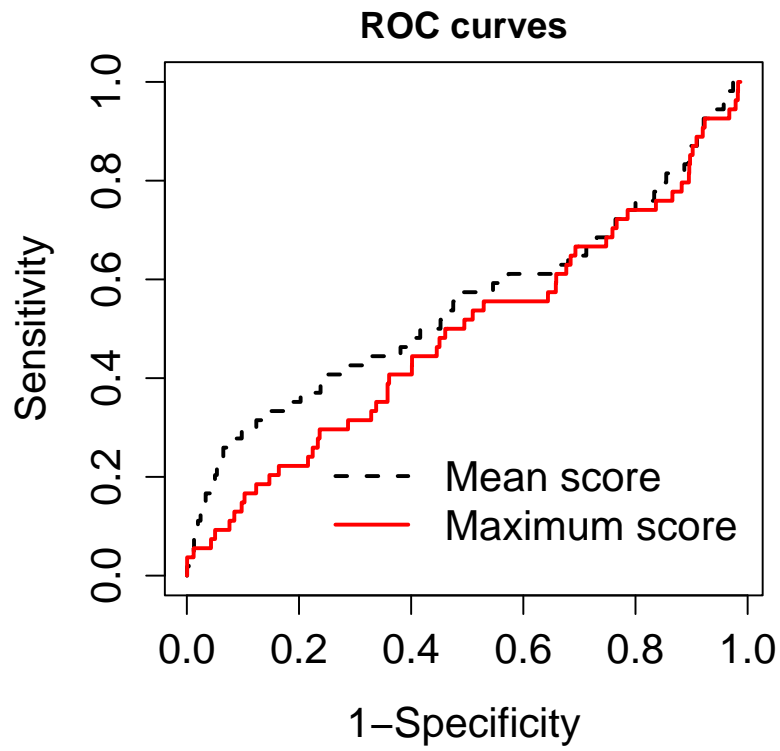


Figure 1: Comparing ROC curves for identifying known Lrp target genes by using mean sequence scores and maximum sequence scores in Lrp data.

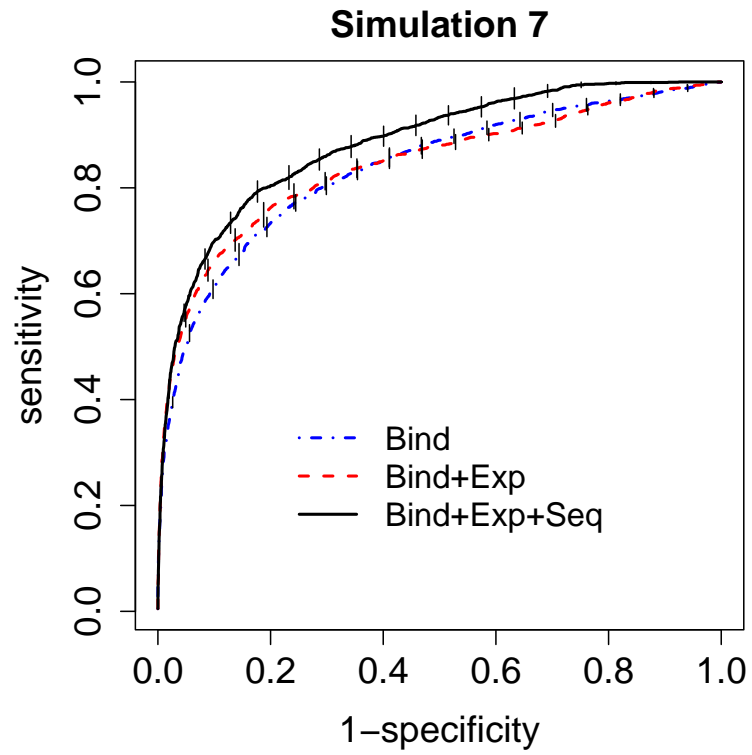


Figure 2: ROC curves for simulation study 7. The mean and standard errors of sensitivities (the vertical tick marks on the line) were used to evaluate the performance of each model.

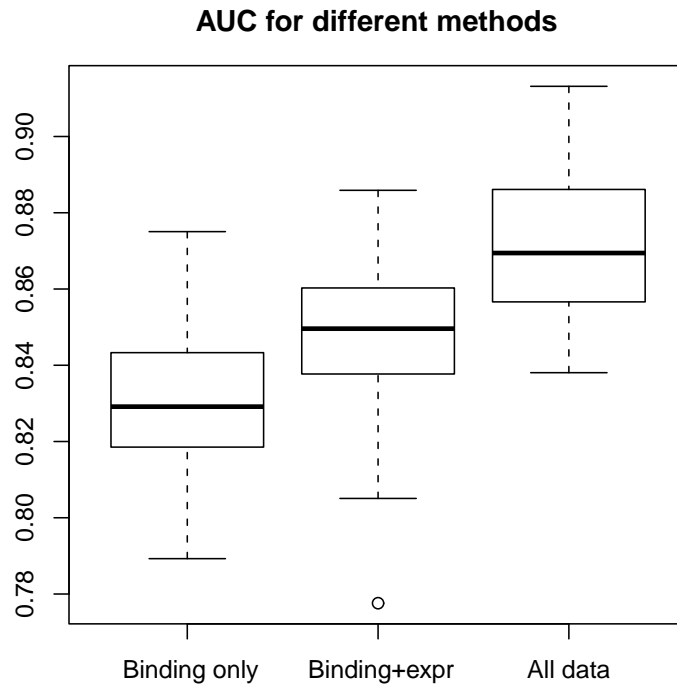


Figure 3: Boxplots of Area Under Curves (AUC) for the simulated data sets using simulation set-up 7, the simulation was repeated 50 times.

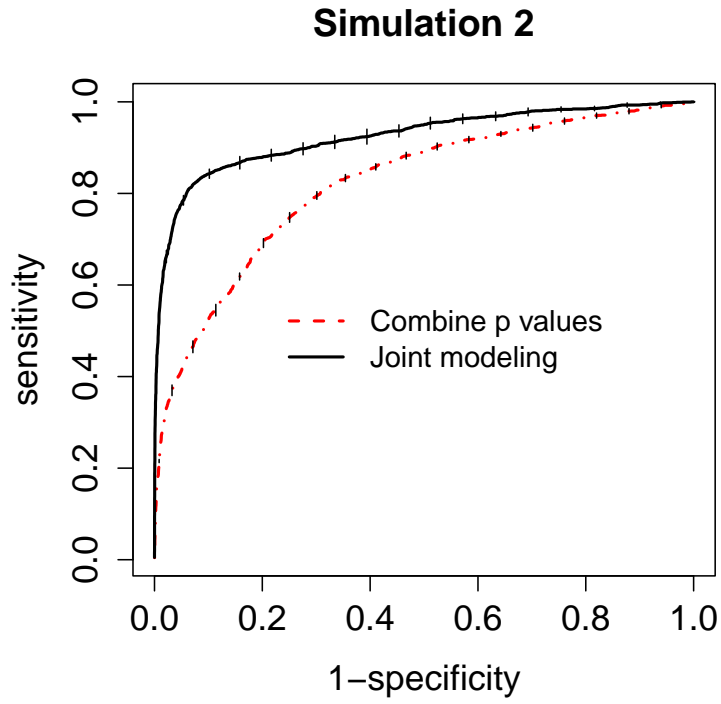


Figure 4: ROC curves for comparing joint modeling and combining p-value approaches in Simulation set-up 2 .

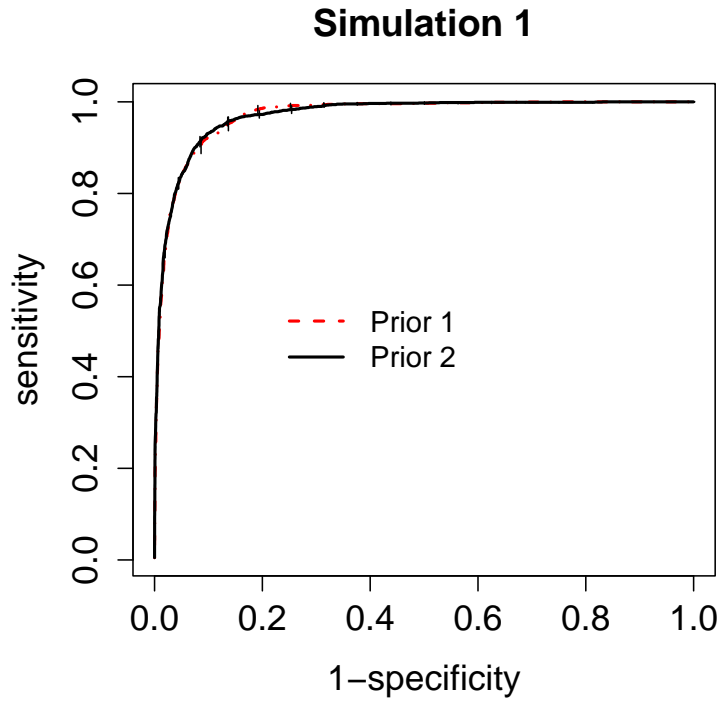


Figure 5: ROC curves for using different priors for $Pr(I_x) = 1$ in simulation 1. Prior 1: $p_x \sim \text{Beta}(200, 800)$; Prior 2: $p_x \sim \text{Beta}(100, 900)$.