Web-based Supplementary Materials for "A Semi-parametric Bayesian Approach for Detection of Gene Expression Heterosis with RNA-Seq Data" by Ran Bi and Peng Liu

Ran Bi[a] and Peng Liu[a]

[a]Department of Statistics, Iowa State University, Ames, IA, USA

ARTICLE HISTORY
Compiled November 17, 2021

Web Appendix A: Detailed Derivation of Full Conditional Distributions

Detailed derivations of the full conditionals for each parameter are provided below.

(1) Obtain draws of λ_{gij}'s from their full condition distributions.

$$\begin{aligned}
p(\lambda_{g1j}|\cdot) &\propto p(Y_{g1j}|\lambda_{g1j})p(\lambda_{g1j}|\alpha_g,\beta_g)\\
&\propto e^{-S_{1j}\lambda_{g1j}}(S_{1j}\lambda_{g1j})^{Y_{g1j}}\lambda_{g1j}^{\alpha_g-1}e^{-(\lambda_{g1j}\beta_g)}\\
&\propto \lambda_{g1j}^{(Y_{g1j}+\alpha_g-1)}e^{-\lambda_{g1j}(S_{1j}+\beta_g)},\\
p(\lambda_{g2j}|\cdot) &\propto p(Y_{g2j}|\lambda_{g2j})p(\lambda_{g2j}|\alpha_g,\beta_g,\rho_{g1})\\
&\propto e^{-S_{2j}\lambda_{g2j}}(S_{2j}\lambda_{g2j})^{Y_{g2j}}\lambda_{g2j}^{\alpha_g-1}e^{-(\lambda_{g2j}\beta_g\rho_{g1})}\\
&\propto \lambda_{g2j}^{(Y_{g2j}+\alpha_g-1)}e^{-\lambda_{g2j}(S_{2j}+\beta_g\rho_{g1})},\\
p(\lambda_{g3j}|\cdot) &\propto p(Y_{g3j}|\lambda_{g3j})p(\lambda_{g3j}|\alpha_g,\beta_g,\rho_{g2})\\
&\propto e^{-S_{3j}\lambda_{g2j}}(S_{3j}\lambda_{g2j})^{Y_{g3j}}\lambda_{g3j}^{\alpha_g-1}e^{-(\lambda_{g3j}\beta_g\rho_{g2})}\\
&\propto \lambda_{g3j}^{(Y_{g3j}+\alpha_g-1)}e^{-\lambda_{g3j}(S_{3j}+\beta_g\rho_{g2})}.
\end{aligned}$$

Thus, λ_{gij}'s are drawn from

$$\begin{aligned}
\lambda_{g1j}|\cdot &\sim \text{Gamma}(Y_{g1j}+\alpha_g, S_{1j}+\beta_g),\\
\lambda_{g2j}|\cdot &\sim \text{Gamma}(Y_{g2j}+\alpha_g, S_{2j}+\beta_g\rho_{g1}),\\
\lambda_{g3j}|\cdot &\sim \text{Gamma}(Y_{g3j}+\alpha_g, S_{3j}+\beta_g\rho_{g2}).
\end{aligned}$$

CONTACT Peng Liu. Email: pliu@iastate.edu

(2) Generate samples of β_g's from their full conditional distributions.

$$\begin{aligned}
p(\beta_g|\cdot) &\propto \pi(\beta_g)\prod_{j=1}^{n_1} p(\lambda_{g1j}|\alpha_g,\beta_g)\prod_{j=1}^{n_2} p(\lambda_{g2j}|\alpha_g,\beta_g,\rho_{g1})\prod_{j=1}^{n_3} p(\lambda_{g3j}|\alpha_g,\beta_g,\rho_{g2}) \\
&\propto \beta_g^{a_0-1}e^{-\beta_g b_0}\cdot\prod_{j=1}^{n_1} e^{-\lambda_{g1j}\beta_g}\beta_g^{\alpha_g}\cdot\prod_{j=1}^{n_2} e^{-\lambda_{g2j}\beta_g\rho_{g1}}(\beta_g\rho_{g1})^{\alpha_g}\cdot\prod_{j=1}^{n_3} e^{-\lambda_{g3j}\beta_g\rho_{g2}}(\beta_g\rho_{g2})^{\alpha_g} \\
&\propto \beta_g^{(n_1+n_2+n_3)\alpha_g+a_0-1}e^{-\beta_g(\sum_{j=1}^{n_1}\lambda_{g1j}+\sum_{j=1}^{n_2}\lambda_{g2j}\rho_{g1}+\sum_{j=1}^{n_3}\lambda_{g3j}\rho_{g2}+b_0)}.
\end{aligned}$$

Thus, β_g's are drawn from

$$\beta_g|\cdot \sim \text{Gamma}\Big(\alpha_g(n_1+n_2+n_3)+a_0,\ \sum_{j=1}^{n_1}\lambda_{g1j}+\sum_{j=1}^{n_2}\lambda_{g2j}\rho_{g1}+\sum_{j=1}^{n_3}\lambda_{g3j}\rho_{g2}+b_0\Big)$$

(3) Draw samples of α_g's.

$$\begin{aligned}
p(\alpha_g|\cdot) &\propto \pi(\alpha_g)\prod_{j=1}^{n_1} p(\lambda_{g1j}|\alpha_g,\beta_g)\prod_{j=1}^{n_2} p(\lambda_{g2j}|\alpha_g,\beta_g,\rho_{g1})\prod_{j=1}^{n_3} p(\lambda_{g3j}|\alpha_g,\beta_g,\rho_{g2}) \\
&\propto e^{-r\alpha_g}\cdot\prod_{j=1}^{n_1}\frac{\lambda_{g1j}^{\alpha_g-1}\beta_g^{\alpha_g}}{\Gamma(\alpha_g)}\cdot\prod_{j=1}^{n_2}\frac{\lambda_{g2j}^{\alpha_g-1}(\beta_g\rho_{g1})^{\alpha_g}}{\Gamma(\alpha_g)}\cdot\prod_{j=1}^{n_3}\frac{\lambda_{g3j}^{\alpha_g-1}(\beta_g\rho_{g2})^{\alpha_g}}{\Gamma(\alpha_g)} \\
&\propto e^{-r\alpha_g}\cdot\frac{\beta_g^{\alpha_g(n_1+n_2+n_3)}\rho_{g1}^{\alpha_g n_2}\rho_{g2}^{\alpha_g n_3}}{[\Gamma(\alpha_g)]^{n_1+n_2+n_3}}\cdot\prod_{i=1}^{3}\prod_{j=1}^{n_i}\lambda_{gij}^{\alpha_g-1}.
\end{aligned}$$

The full conditional distribution of each α_g has no closed-form. If $p(\alpha_g|\cdot)$ is a log-concave function with respect to α_g, then we could apply the adaptive rejection sampling method proposed by Gilks (1992) to draw samples of α_g's.

$$\begin{aligned}
log\ p(\alpha_g|\cdot) =& -r\alpha_g+\alpha_g(n_1+n_2+n_3)log(\beta_g)+\alpha_g n_2 log(\rho_{g1})+\alpha_g n_3 log(\rho_{g2}) \\
&-(n_1+n_2+n_3)log\ \Gamma(\alpha_g)+\sum_{i=1}^{3}\sum_{j=1}^{n_i}(\alpha_g-1)log(\lambda_{gij}),
\end{aligned}$$

then the first derivative of $log\ p(\alpha_g|\cdot)$ is

$$\begin{aligned}
\frac{\partial\ log\ p(\alpha|\cdot)}{\partial\ \alpha_g} =& -r+(n_1+n_2+n_3)log(\beta_g)+n_2 log(\rho_{g1})+n_3 log(\rho_{g2}) \\
&-(n_1+n_2+n_3)\frac{\partial\ log\ \Gamma(\alpha_g)}{\partial\ \alpha_g}+\sum_{i=1}^{3}\sum_{j=1}^{n_i} log(\lambda_{gij}),
\end{aligned}$$

the second derivative of $log\ p(\alpha_g|\cdot)$ is

$$\frac{\partial^2\ log\ p(\alpha_g|\cdot)}{\partial\ \alpha_g^2} = -(n_1+n_2+n_3)\frac{\partial^2\ log\ \Gamma(\alpha_g)}{\partial\ \alpha_g^2}.$$

We can derive that

$$\frac{\partial^2 \ log\ \Gamma(\alpha_g)}{\partial\ \alpha_g^2} = \sum_{k=0}^{\infty} \frac{1}{(\alpha_g + k)^2} > 0.$$

Therefore,

$$\frac{\partial^2 \ log\ p(\alpha_g|\cdot)}{\partial\ \alpha_g^2} < 0,$$

i.e., $p(\alpha_g|\cdot)$ is log-concave.

(4) Let the Markov chain consist of $(\xi_1, \ldots, \xi_G)$ and $(\rho_1^*, \ldots, \rho_K^*)$. Generate posterior samples for ρ_{g1}'s as below:

(i) Update the configuration vector $(\xi_1, \ldots, \xi_G)$.

- If $\xi = \xi_l$ for some $l \neq g$,

$$\begin{aligned} p(\xi_g = \xi|\boldsymbol{\xi}_{-g},\ \text{rest}) &= cn_{\xi}^{(-g)} \Pi_{j=1}^{n_2} p(\lambda_{g2j}|\alpha_g, \beta_g, \rho_{\xi}^*) \\ &= cn_{\xi}^{(-g)} \Pi_{j=1}^{n_2} \frac{\beta_g^{\alpha_g} (\rho_{\xi}^*)^{\alpha_g}}{\Gamma(\alpha_g)} \lambda_{g2j}^{\alpha_g - 1} exp(-\beta_g \rho_{\xi}^* \lambda_{g2j}). \end{aligned}$$

- Otherwise,

$$\begin{aligned} & p(\xi_g \neq \xi_l \text{ for all } l \neq g|\boldsymbol{\xi}_{-g},\ \text{rest}) \\ &= cM \int \Pi_{j=1}^{n_2} p(\lambda_{g2j}|\alpha_g, \beta_g, \rho) F_0(\rho) d\rho \\ &= cM \int_0^{\infty} \Big[\Pi_{j=1}^{n_2} \frac{\lambda_{g2j}^{\alpha_g - 1} exp(-\beta_g \rho \lambda_{g2j}) (\beta_g \rho)^{\alpha_g}}{\Gamma(\alpha_g)} \cdot (1 - p_0) \frac{\rho^{\alpha_0 - 1} exp(-\beta_0 \rho) \beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \Big] d\rho \\ &\quad + cM \int_0^{\infty} \Big[\Pi_{j=1}^{n_2} \frac{\lambda_{g2j}^{\alpha_g - 1} exp(-\beta_g \rho \lambda_{g2j}) (\beta_g \rho)^{\alpha_g}}{\Gamma(\alpha_g)} \cdot p_0 \delta_{\{1\}} \Big] d\rho \\ &= cM(1 - p_0) \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \frac{\beta_g^{n_2 \alpha_g}}{[\Gamma(\alpha_g)]^{n_2}} \Pi_{j=1}^{n_2} \lambda_{g2j}^{\alpha_g - 1} \int_0^{\infty} \rho^{n_2 \alpha_g + \alpha_0 - 1} exp\Big\{ - \big(\beta_g \sum_{j=1}^{n_2} \lambda_{g2j} + \beta_0\big) \rho \Big\} d\rho \\ &\quad + cMp_0 \Pi_{j=1}^{n_2} \left\{ \frac{\beta_g^{\alpha_g}}{\Gamma(\alpha_g)} \lambda_{g2j}^{\alpha_g - 1} exp(-\beta_g \lambda_{g2j}) \right\} \\ &= cM(1 - p_0) \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \frac{\beta_g^{n_2 \alpha_g}}{[\Gamma(\alpha_g)]^{n_2}} \Pi_{j=1}^{n_2} \lambda_{g2j}^{\alpha_g - 1} \times \frac{\Gamma(n_2 \alpha_g + \alpha_0)}{\Big(\beta_0 + \beta_g \sum_{j=1}^{n_2} \lambda_{g2j}\Big)^{n_2 \alpha_g + \alpha_0}} \\ &\quad + cMp_0 \Pi_{j=1}^{n_2} \left\{ \frac{\beta_g^{\alpha_g}}{\Gamma(\alpha_g)} \lambda_{g2j}^{\alpha_g - 1} exp(-\beta_g \lambda_{g2j}) \right\}. \end{aligned}$$

Here c is an appropriate normalizing constant to ensure that probabilities add up to 1.

(ii) Update $(\rho_1^*, \ldots, \rho_K^*)$.

$$
\begin{aligned}
p(\rho_k^*|.) &\propto \Pi_{\{g:\xi_g=k\}}\Pi_{j=1}^{n_2} p(\lambda_{g2j}|\alpha_g, \beta_g, \rho_{g1}) \cdot F_0(\rho_k^*) \\
&\propto \Pi_{\{g:\xi_g=k\}}\Pi_{j=1}^{n_2} p(\lambda_{g2j}|\alpha_g, \beta_g, \rho_k^*) \cdot F_0(\rho_k^*) \\
&\propto \Pi_{\{g:\xi_g=k\}}\Pi_{j=1}^{n_2} \frac{\lambda_{g2j}^{\alpha_g-1} exp(\beta_g\rho_k^*\lambda_{g2j})(\beta_g\rho_k^*)^{\alpha_g}}{\Gamma(\alpha_g)} \\
&\quad \times \Big\{ p_0\delta_{\{1\}} + (1-p_0)\frac{(\rho_k^*)^{\alpha_0-1} exp(-\beta_0\rho_k^*)\beta_0^{\alpha_g}}{\Gamma(\alpha_0)} \Big\} \\
&\propto p_0 exp\Big\{ -\Big(\sum_{\{g:\xi_g=k\}} \sum_{j=1}^{n_2} \beta_g\lambda_{g2j} \Big) \Big\} \cdot \Big\{ \Pi_{\{g:\xi_g=k\}}\Pi_{j=1}^{n_2} \frac{\lambda_{g2j}^{\alpha_g-1}\beta_g^{\alpha_g}}{\Gamma(\alpha_g)} \Big\} \delta_{\{1\}} \\
&\quad + (1-p_0)(\rho_k^*)^{\sum_{\{g:\xi_g=k\}}\sum_{j=1}^{n_2}\alpha_g+\alpha_0-1} \cdot exp\Big\{ -\Big(\sum_{\{g:\xi_g=k\}} \sum_{j=1}^{n_2} \beta_g\lambda_{g2j} + \beta_0 \Big)\rho_k^* \Big\} \\
&\quad \cdot \Big\{ \Pi_{\{g:\xi_g=k\}}\Pi_{j=1}^{n_2} \frac{\lambda_{g2j}^{\alpha_g-1}\beta_g^{\alpha_g}}{\Gamma(\alpha_g)} \Big\} \cdot \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \\
&\propto p_0 exp\Big\{ -\Big(\sum_{\{g:\xi_g=k\}} \sum_{j=1}^{n_2} \beta_g\lambda_{g2j} \Big) \Big\} \delta_{\{1\}} \\
&\quad + (1-p_0)\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \cdot \frac{\Gamma\big(n_2\sum_{\{g:\xi_g=k\}}\alpha_g + \alpha_0\big)}{\big(\sum_{\{g:\xi_g=k\}}\sum_{j=1}^{n_2}\beta_g\lambda_{g2j} + \beta_0\big)^{n_2\sum_{\{g:\xi_g=k\}}\alpha_g+\alpha_0}} \\
&\quad \cdot \mathrm{Gamma}\Big(n_2 \sum_{\{g:\xi_g=k\}} \alpha_g + \alpha_0, \sum_{\{g:\xi_g=k\}} \sum_{j=1}^{n_2} \beta_g\lambda_{g2j} + \beta_0 \Big) \\
&\propto p_0 exp\Big\{ -\Big(\sum_{\{g:\xi_g=k\}} \sum_{j=1}^{n_2} \beta_g\lambda_{g2j} \Big) \Big\} \delta_{\{1\}} \\
&\quad + c_0 \mathrm{Gamma}\Big(n_2 \sum_{\{g:\xi_g=k\}} \alpha_g + \alpha_0, \sum_{\{g:\xi_g=k\}} \sum_{j=1}^{n_2} \beta_g\lambda_{g2j} + \beta_0 \Big),
\end{aligned}
$$

where $c_0 = (1-p_0)\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \cdot \frac{\Gamma(n_2\sum_{\{g:\xi_g=k\}}\alpha_g+\alpha_0)}{\Big(\beta_0+\sum_{\{g:\xi_g=k\}}\sum_{j=1}^{n_2}\beta_g\lambda_{g2j}\Big)^{n_2\sum_{\{g:\xi_g=k\}}\alpha_g+\alpha_0}}$.

(5) The procedure for obtaining posterior samples for ρ_{g2}'s is similar to ρ_{g1}'s.

Web Appendix B: Table Results for Simulations A and B

Web Table 1.: Results for LPH in Simulation A.

Nominal Level of FDR	Method	Actual FDR	Number of Declared Heterosis Genes	Number of Declared Truly Heterosis Genes	Total Number of Heterosis Genes
0.01	*SBA*	0.0006	402	401	613
	SBA_div	0.0005	400	400	
	eBayes_Laplace	0.2891	723	513	
	eBayes_Normal	0.2882	738	524	
0.05	*SBA*	0.0035	452	451	613
	SBA_div	0.0031	452	450	
	eBayes_Laplace	0.4979	1139	571	
	eBayes_Normal	0.4947	1153	581	
0.1	*SBA*	0.0099	490	485	613
	SBA_div	0.0110	489	484	
	eBayes_Laplace	0.5928	1471	598	
	eBayes_Normal	0.5916	1477	602	
0.2	*SBA*	0.0401	561	539	613
	SBA_div	0.0435	561	536	
	eBayes_Laplace	0.6840	1941	612	
	eBayes_Normal	0.6814	1925	613	

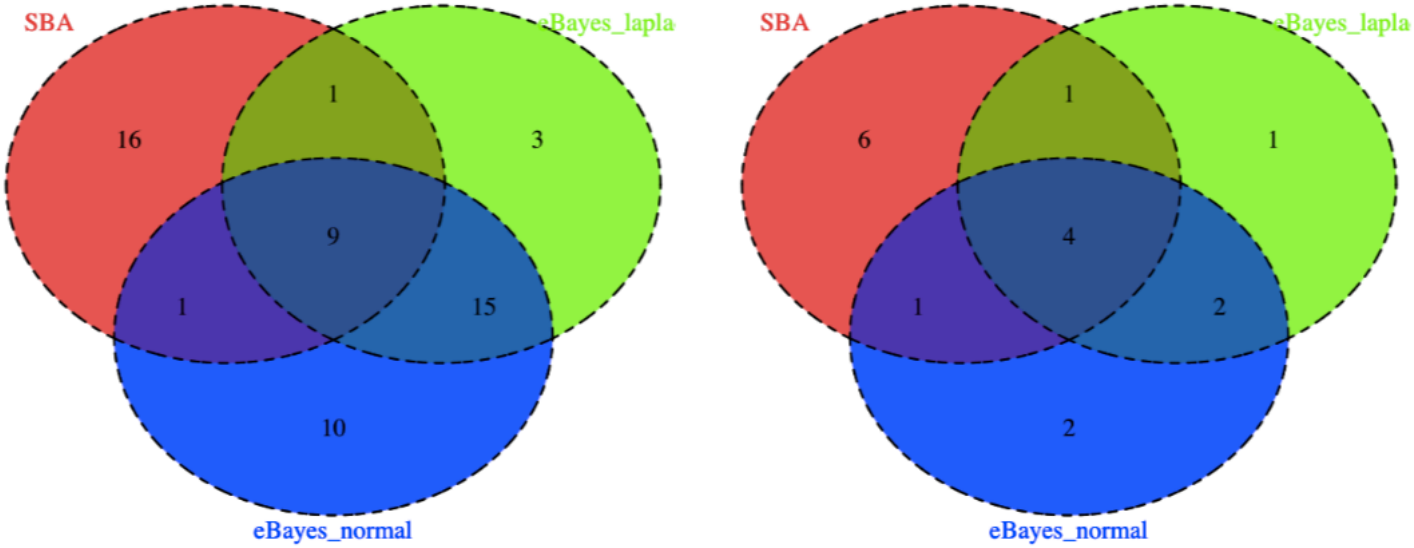
Web Table 2.: Results for HPH in Simulation B.

Nominal Level of FDR	Method	Actual FDR	Number of Declared Heterosis Genes	Number of Declared Truly Heterosis Genes	Total Number of Heterosis Genes
0.01	*SBA*	0.0018	261	261	538
	SBA_div	0.0014	255	255	
	eBayes_Laplace	0.0728	422	391	
	eBayes_Normal	0.0802	432	397	
0.05	*SBA*	0.0096	330	327	538
	SBA_div	0.0093	323	320	
	eBayes_Laplace	0.2214	568	442	
	eBayes_Normal	0.2386	590	449	
0.1	*SBA*	0.0267	373	363	538
	SBA_div	0.0261	363	354	
	eBayes_Laplace	0.3391	705	465	
	eBayes_Normal	0.3564	736	473	
0.2	*SBA*	0.0811	447	410	538
	SBA_div	0.0718	431	400	
	eBayes_Laplace	0.4906	971	494	
	eBayes_Normal	0.5080	1016	499	

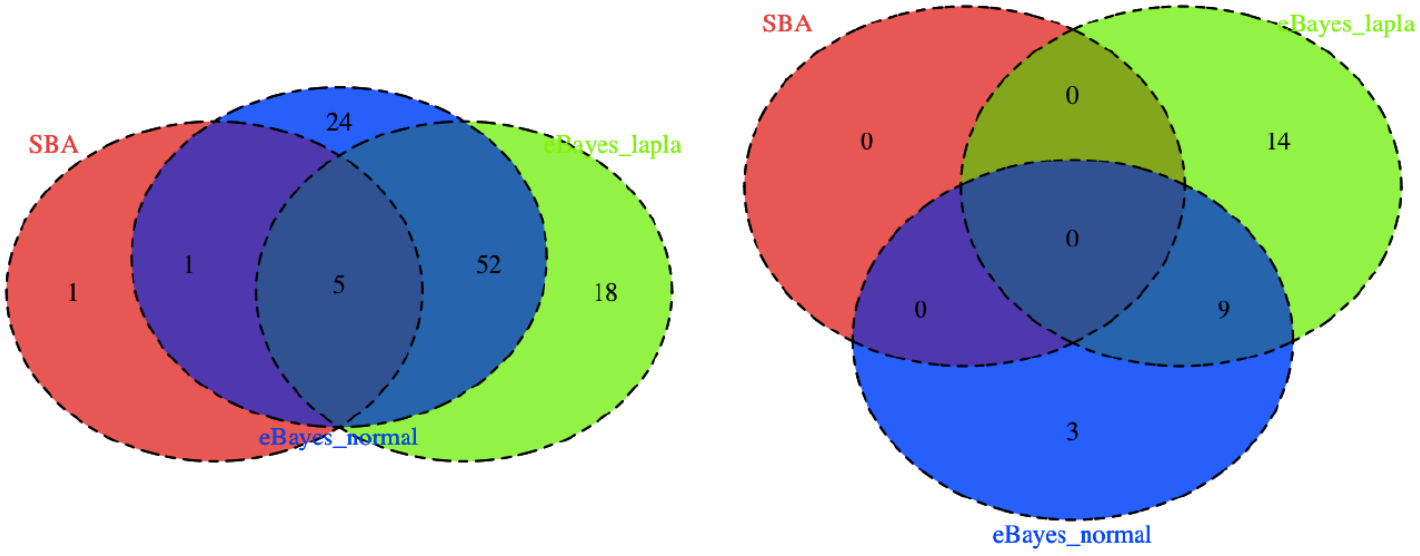
Web Table 3.: Results for LPH in Simulation B.

Nominal Level of FDR	Method	Actual FDR	Number of Declared Heterosis Genes	Number of Declared Truly Heterosis Genes	Total Number of Heterosis Genes
0.01	*SBA*	0.0005	194	194	536
	SBA_div	0.0002	191	191	
	eBayes_Laplace	0.0738	377	349	
	eBayes_Normal	0.0847	390	357	
0.05	*SBA*	0.0040	246	245	536
	SBA_div	0.0040	241	240	
	eBayes_Laplace	0.2356	543	415	
	eBayes_Normal	0.2567	572	425	
0.1	*SBA*	0.0135	280	277	536
	SBA_div	0.0109	273	270	
	eBayes_Laplace	0.3576	701	449	
	eBayes_Normal	0.3786	745	462	
0.2	*SBA*	0.0440	339	324	536
	SBA_div	0.0343	326	315	
	eBayes_Laplace	0.5125	1010	491	
	eBayes_Normal	0.5308	1072	502	

Web Appendix C: Venn Diagrams for Real Data Analysis



Web Figure 1.: Real data analysis results for HPH. The left Venn diagram provides the number of overlapping identified HPH genes from *SBA*, *eBayes_Laplace* and *eBayes_Normal* methods while controlling FDR at 0.1; the right Venn diagram gives corresponding results while controlling FDR at 0.05.



Web Figure 2.: Real data analysis results for LPH. The left Venn diagram provides the number of overlapping identified LPH genes from *SBA*, *eBayes_Laplace* and *eBayes_Normal* methods while controlling FDR at 0.1; the right Venn diagram gives corresponding results while controlling FDR at 0.05.

Web Appendix D: Robustness of Prior $p_0 = 0.5$ and Data Division under Different Simulation Scenarios

In our proposed method SBA, we specify $p_0 = 0.5$ so that no prior preference is given to either differential expression or equivalent expression between hybrid offspring and either parental line. Our simulation results in Section 4 are also based on the setting that half of the 3000 genes have fold change 1 between hybrid offspring and each parent. To investigate the robustness of setting $p_0 = 0.5$ under different simulation scenarios, we vary the true proportion of genes having fold change 1 between hybrid offspring and each parental line to be 0.2, 0.5 and 0.8, and perform similar method comparison as in the manuscript in terms of ranking heterosis genes and FDR control.

More specifically, 3000 genes were drawn from $NB(\mu_g, \phi_g)$, where pairs of μ_g and ϕ_g were sampled from the estimates from the same maize data as in the manuscript. 20% (or 50%, or 80%) of the 3000 genes were randomly selected to have fold changes $\rho_{g1} = 1$ between hybrid and parental line 1. The remaining 80% (or 50%, or 20%) of the 3000 genes were simulated to have fold change parameters ρ_{g1} equal to fixed values or from certain distribution, same as in Section 4 of the manuscript. Then RNA-seq count data for parental lines 1 and 2 were drawn in the same way as in Section 4 of the manuscript.

In addition to our proposed semi-parametric approaches, *SBA* (without data division) and *SBA_div* (with data division), we also evaluate two more strategies similar to *SBA_div* but with different data division method (*SBA_divrho* and *SBA_divcount*), and compare them with the empirical Bayes method in Niemi et al. (2015) (*eBayes_Laplace* and *eBayes_Normal*, depending on the parametric prior assumption). Notice that *SBA_div*, *SBA_divrho* and *SBA_divcount* differentiate only by the way they divide the total of G genes. *SBA_div* randomly divides genes into several groups, where genes can only borrow information from those within the same group. Intuitively, when similar genes are grouped together, the information they borrow could be more reliable. *SBA_divcount* provides the possibility to group similar genes together, according to estimated heterosis status based on point estimation from count data. *SBA_divrho* divides genes into groups based on the true heterosis status, which can only be used in simulation studies to investigate how the division based on count data performs (*SBA_divcount*).

In particular, methods under comparison include:

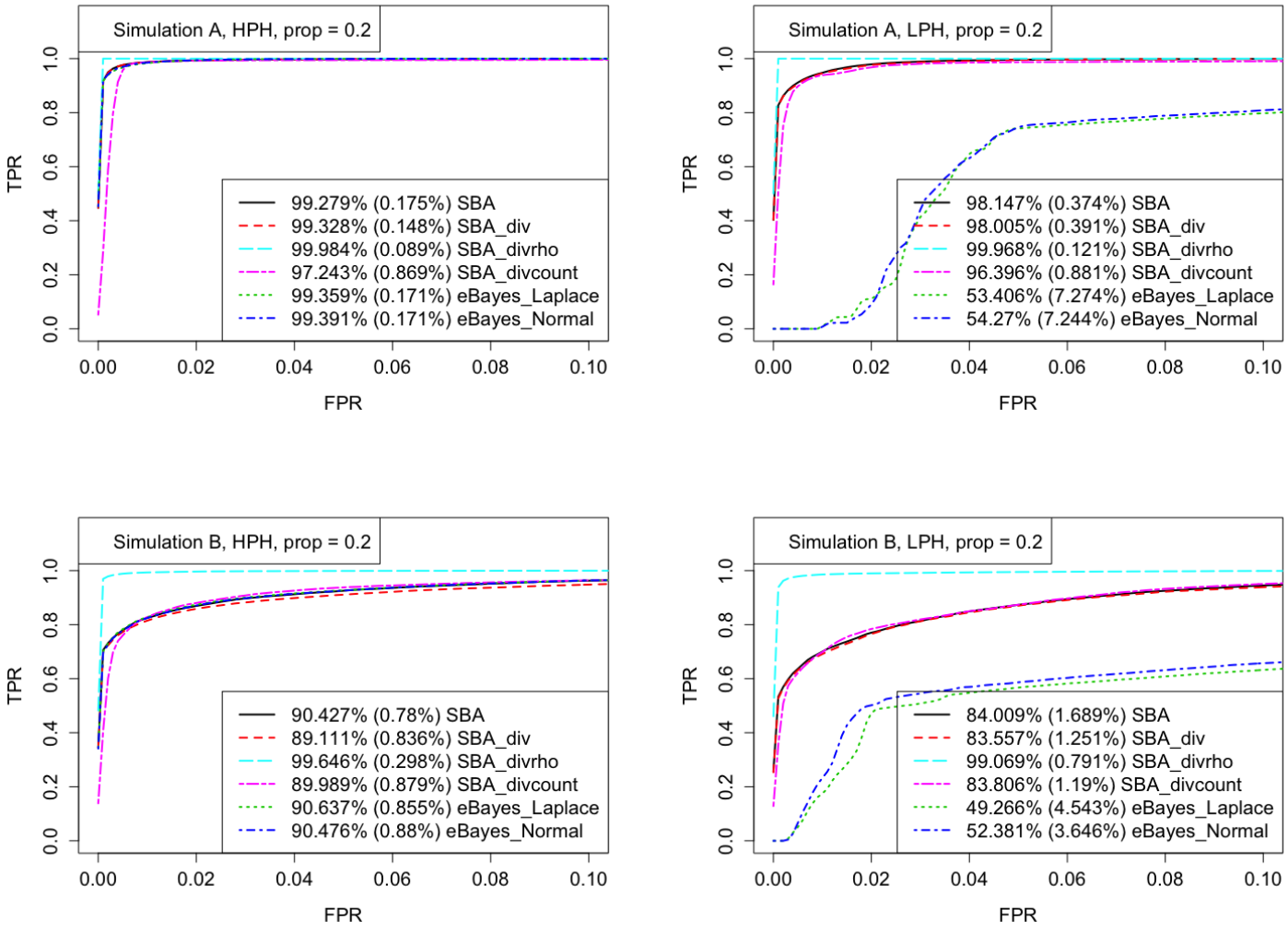
- *SBA* - Our proposed semi-parametric approach without data division.
- *SBA_div* - Our proposed semi-parametric approach with data division, where G genes are randomly divided into 5 groups.
- *SBA_divrho* - Semi-parametric approach with data division, where G genes are divided into 5 groups based on true heterosis status (true ρ_{g1} and ρ_{g2}): genes exhibit HPH ($\rho_{g1} > 1$ and $\rho_{g2} > 1$) in one group, genes exhibit LPH ($\rho_{g1} < 1$ and $\rho_{g2} < 1$) in one group, genes with $\rho_{g1} = \rho_{g2} = 1$ in two groups, others in one group.
- *SBA_divcount* - Semi-parametric approach with data division, where G genes are divided into 5 groups based on estimated heterosis status (estimated $\hat{\rho}_{g1}$ and $\hat{\rho}_{g2}$, calculated by the ratio of mean normalized count of hybrid offspring over mean normalized count of each parental line): genes estimated to be HPH ($\hat{\rho}_{g1} > 1.5$ and $\hat{\rho}_{g2} > 1.5$) in one group, genes estimated to be LPH ($\hat{\rho}_{g1} < 1/1.5$ and $\hat{\rho}_{g2} < 1/1.5$) in one group, genes with both $\hat{\rho}_{g1}$ and $\hat{\rho}_{g2}$ lie in [1/1.5, 1.5] in two groups, others in one group.
- *eBayes_Laplace* and *eBayes_Normal* - Empirical Bayes methods proposed by Niemi et al. (2015).

As indicated in Web Figures 3, 5 and 7, our proposed methods (*SBA* and *SBA_div*) robustly generated higher ROC curves and greater AUC values than the empirical Bayes method proposed in Niemi et al. (2015), under various proportions of genes with fold changes 1 for both simulation settings A and B. When evaluating FDR control, Web Figures 4, 6 and 8 demonstrated that our proposed methods (*SBA* and *SBA_div*) controlled FDR

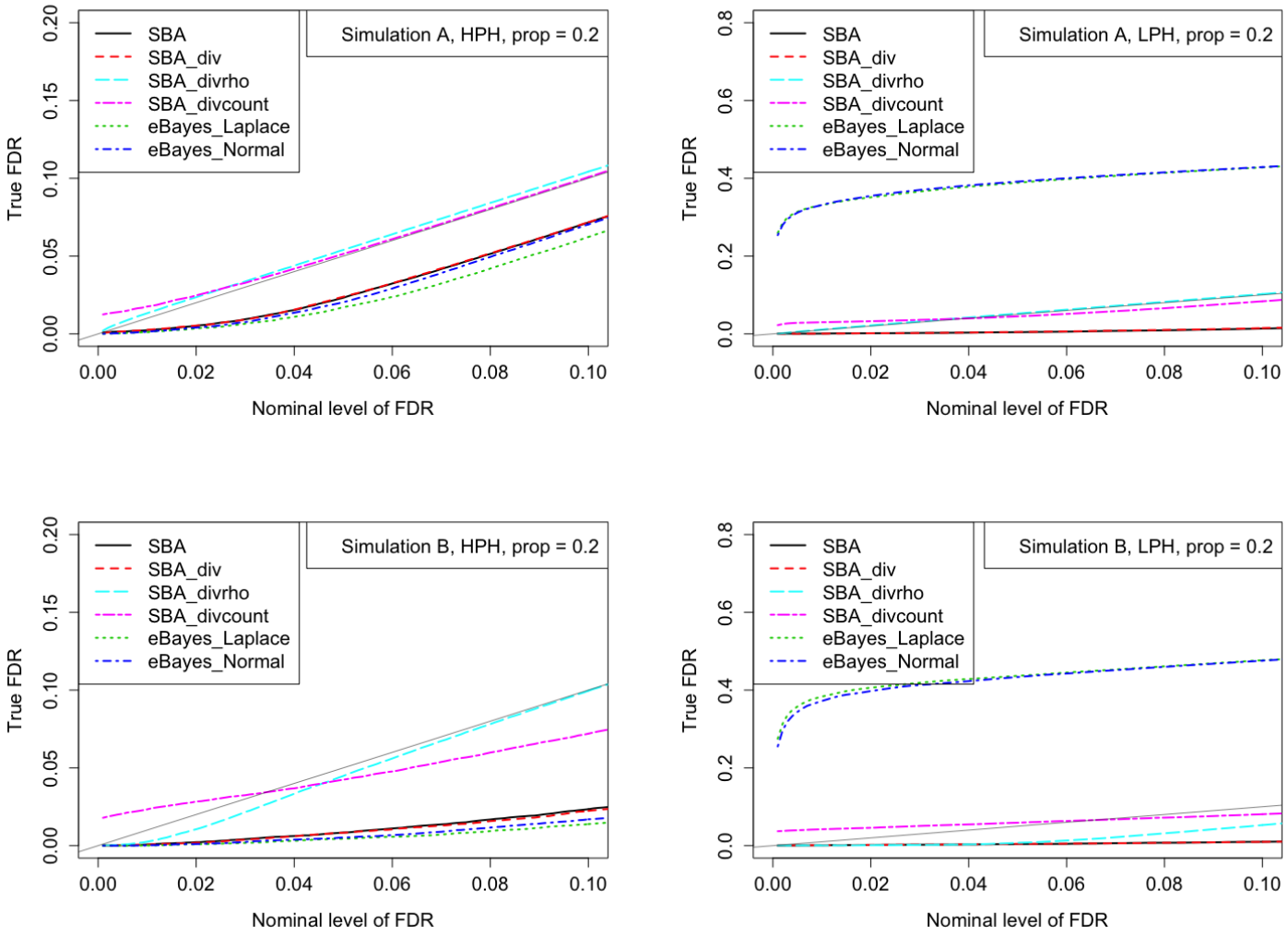
close to the nominal levels, while FDR was not controlled for the empirical Bayes method in Niemi et al. (2015) under almost all scenarios.

It is easy noticing that *SBA_divrho* generated the highest ROC curves and nearly 100% AUC values, as well as FDR control that is close to the nominal level. This verifies our thought that grouping similar genes would improve our method performance. *SBA_divcount* performed quite unstable in terms of ranking heterosis genes, and it generated more false positives than desired. This is due to limited (only 3) biological replicates in each group, thus point estimation of fold changes based on count data may not be reliable.

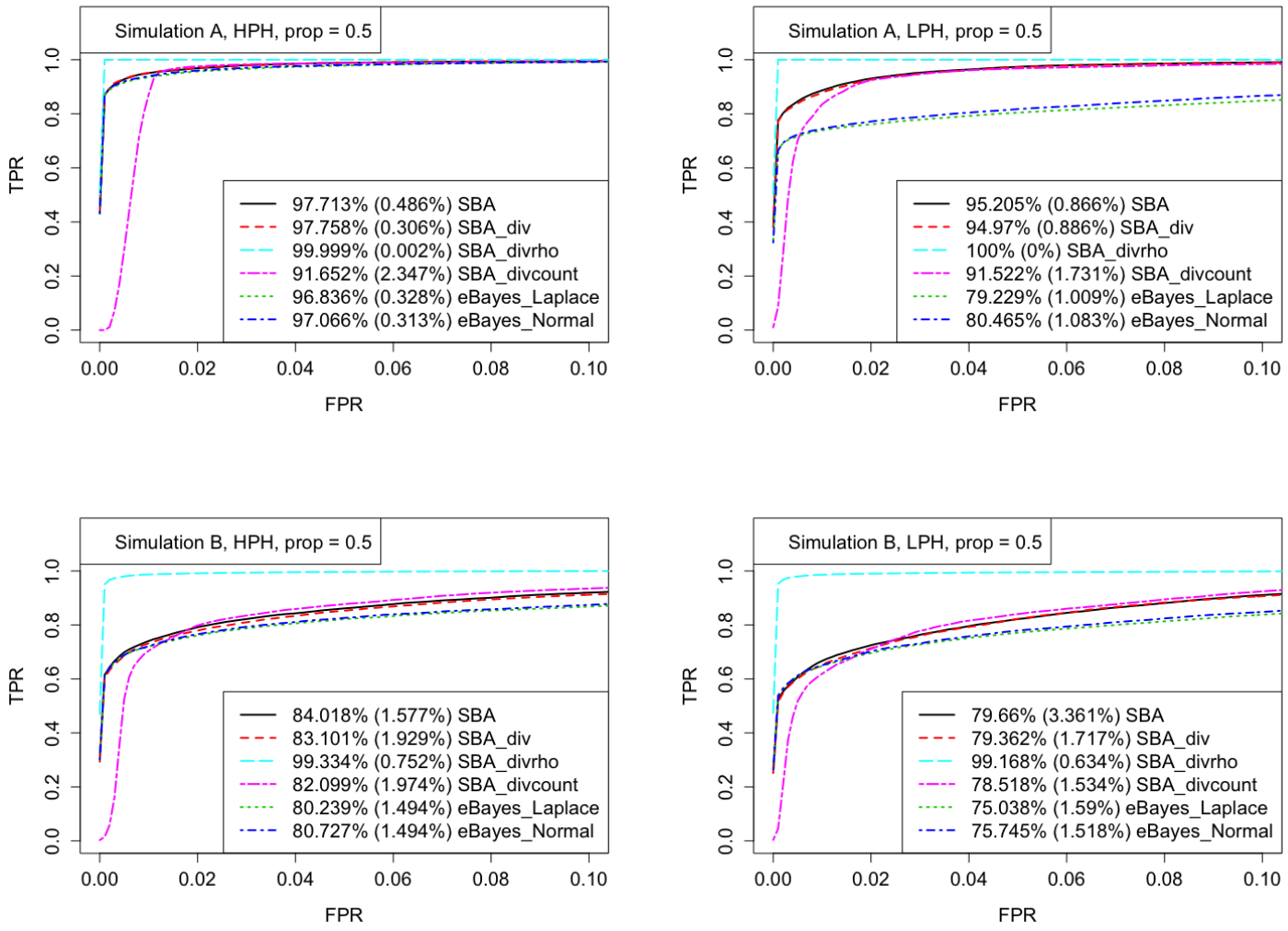
Therefore, *SBA_div* that randomly divides genes into groups would be the simplest and best data division strategy for now. It performs close to *SBA* and better than the empirical Bayes methods in general.



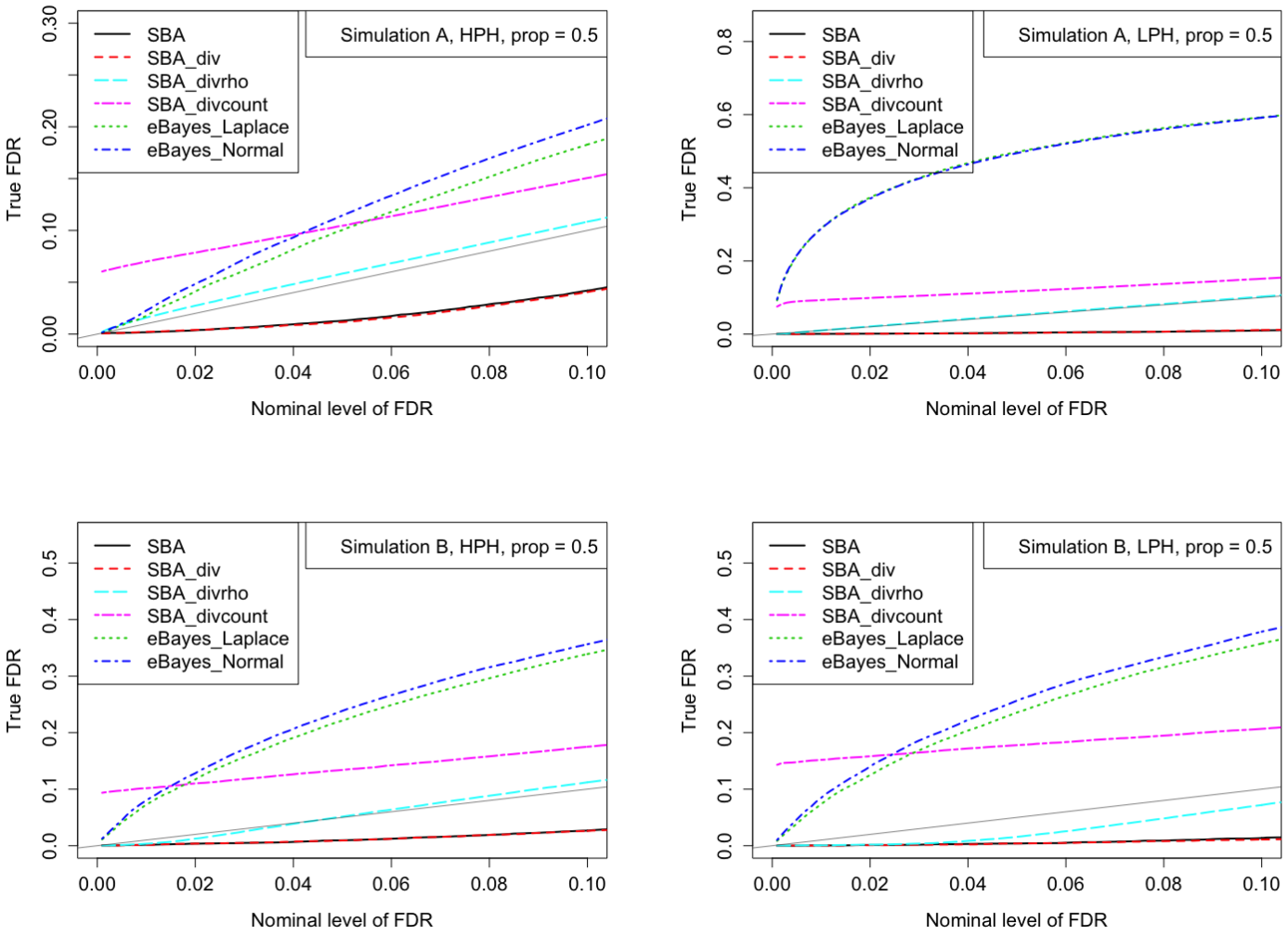
Web Figure 3.: ROC curves for Simulations A and B when 20% of genes have fold changes 1. Given each FPR level, the TPRs were averaged over 32 simulated datasets. The partial AUC values were calculated by averaging the percentages of the total area in the plotted region where FPR is below 0.1, and reported in the legends, with the standard deviations in parentheses.



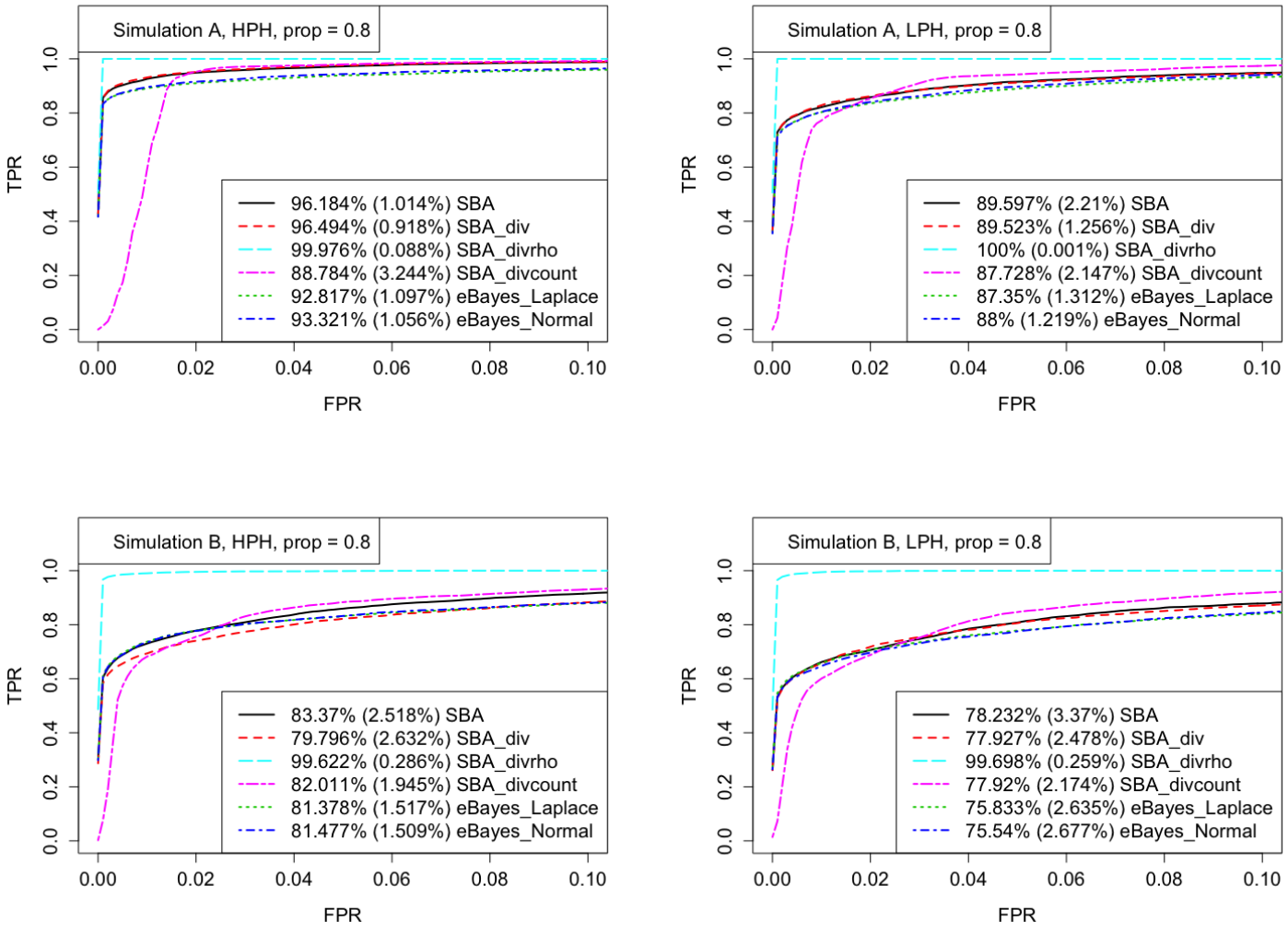
Web Figure 4.: FDR plots for Simulations A and B when 20% of genes have fold changes 1. Given each nominal level of FDR, the actual observed FDRs were estimated by averaging the proportion of false discoveries among declared heterosis genes across 32 simulated datasets. The gray solid lines represent the $Y = X$ line.



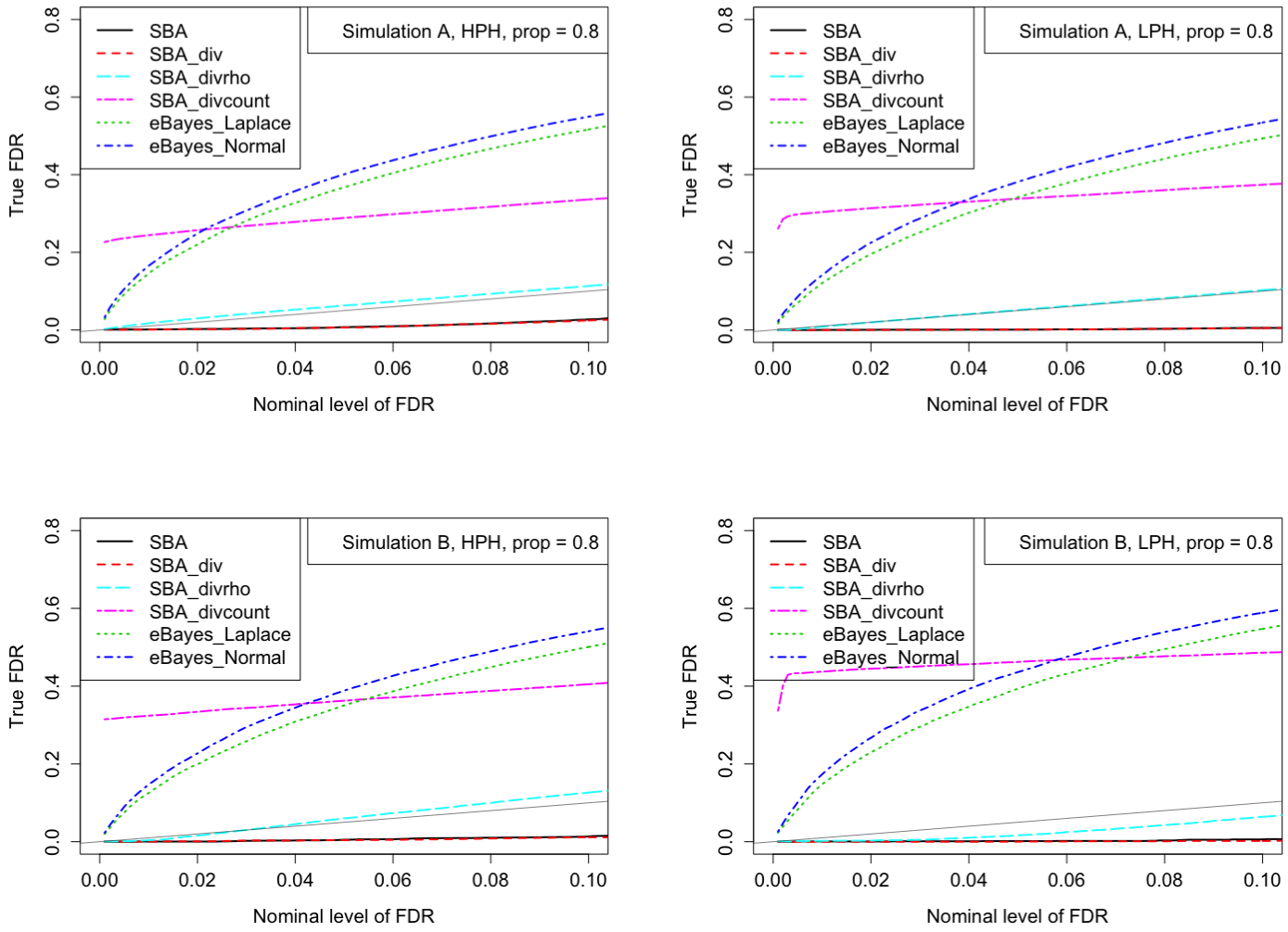
Web Figure 5.: ROC curves for Simulations A and B when 50% of genes have fold changes 1. Given each FPR level, the TPRs were averaged over 32 simulated datasets. The partial AUC values were calculated by averaging the percentages of the total area in the plotted region where FPR is below 0.1, and reported in the legends, with the standard deviations in parentheses.



Web Figure 6.: FDR plots for Simulations A and B when 50% of genes have fold changes 1. Given each nominal level of FDR, the actual observed FDRs were estimated by averaging the proportion of false discoveries among declared heterosis genes across 32 simulated datasets. The gray solid lines represent the $Y = X$ line.



Web Figure 7.: ROC curves for Simulations A and B when 80% of genes have fold changes 1. Given each FPR level, the TPRs were averaged over 32 simulated datasets. The partial AUC values were calculated by averaging the percentages of the total area in the plotted region where FPR is below 0.1, and reported in the legends, with the standard deviations in parentheses.



Web Figure 8.: FDR plots for Simulations A and B when 80% of genes have fold changes 1. Given each nominal level of FDR, the actual observed FDRs were estimated by averaging the proportion of false discoveries among declared heterosis genes across 32 simulated datasets. The gray solid lines represent the $Y = X$ line.

References

W.R. Gilks, *Adaptive Rejection Sampling for Gibbs Sampling*, Applied Statistics, **41** (1992), pp. 337–348.

J. Niemi, E. Mittman, W. Landau, and D. Nettleton, *Empirical bayes analysis of RNA-seq data for detection of gene expression heterosis*, Journal of Agricultural, Biological, and Environmental Statistics, **20(4)** (2015), pp. 614–628.