1 The Simulation Study

We compare our method to existing ones in a simulation study. We generate thirty datasets using simulation parameters calculated from real data and analyze the pseudo-data using our method and the popular R language packages `edgeR`, `baySeq`, and `ShrinkBayes` [6] [1]. Using ROC (receiver operating characteristic) curves, we rank the methods' abilities to identify heterosis genes.

2 Simulated Data

We begin with a real heterosis RNA-seq dataset from a study by Paschold, Jia, Marcon, and others [5]. We select four libraries from each parent genotype and from the hybrid genotype, totaling twelve libraries for analysis. After library selection, we trim low-count features (genes): that is, we remove all the features with mean expression level below exp(1) or with more than three zero counts, leaving 27888 features. Using the `calcNormFactors()`, `estimateGLMTagwiseDisp()`, and `glmFit()` functions in the `edgeR` package, we calculate normalization factors $c_1, \ldots, c_{12}$, dispersion parameters ψ_f for feature $f = 1, \ldots, 27888$, and main effects $\mu_{f,t}$ for each f and treatment group $t = 1$ (parent 1), 2 (parent 2), 3 (hybrid). These estimates serve as simulation parameters for all of our thirty pseudo-datasets.

To simulate a dataset with N libraries per treatment group ($3N$ total libraries), the count for feature f and library i is drawn from a $\text{NB}(\exp(c_{\lceil 4i/N \rceil} + \mu_{f,\lceil i/N \rceil}),\ \psi_f)$ distribution independently of the other counts. Note that feature f is a heterosis feature if $\mu_{f,3} > \max(\mu_{f,1}, \mu_{f,2})$ or if $\mu_{f,3} < \min(\mu_{f,1}, \mu_{f,2})$. Lastly, we apply the same trimming procedure as before and select a random subset of 25000 of the remaining features. For that dataset, we maintain a "truth vector" $H = (h_1, \ldots, h_{25000})$, where $h_f = 1$ if feature f of the simulated dataset is a heterosis feature and $h_f = 0$ otherwise.

We simulate 30 datasets total: 10 with $N = 4$, 10 with $N = 8$, and 10 with $N = 16$.

3 `edgeR`

`edgeR` is one of the most popular R packages in RNA-sequencing data analysis. Its newest implementation applies a negative binomial loglinear model to the data. It uses a Cox-Reid adjusted profile likelihood to estimate dispersion parameters, and in the case of `estimateGLMTagwiseDisp()`, shrinks the final dispersion estimates towards those of neighboring features on a common trend. It then estimates main effects using a Fisher scoring algorithm [7] [4].

Using the `calcNormFactors()`, `estimateGLMTagwiseDisp()`, and `glmFit()` functions in the `edgeR` package, we calculate normalization factor estimates $\widehat{c}_i$ for $i = 1, \ldots, 3N$, dispersion parameter estimates $\widehat{\psi}_f$ for feature $f = 1, \ldots, 25000$, and main effects $\widehat{\mu}_{f,t}$ for each f and treatment group $t = 1$ (parent 1), 2 (parent 2), 3 (hybrid). Using the `glmLRT()` function, we use likelihood ratio tests to perform the following hypothesis tests.

$$H_{0,f,1} : \mu_{f,3} - \mu_{f,1} = 0 \text{ vs } H_{a,f,1} : \mu_{f,3} - \mu_{f,1} \neq 0$$
$$H_{0,f,2} : \mu_{f,3} - \mu_{f,2} = 0 \text{ vs } H_{a,f,1} : \mu_{f,3} - \mu_{f,2} \neq 0$$

We obtain p-values $p_{f,1}$ and $p_{f,2}$, respectively, from each of the above tests. To translate the results into a test for heterosis for each feature, we compute the following p-values

$$p_{f,\texttt{edgeR}} = \begin{cases} p_{f,1}/2 & \widehat{\mu}_{f,3} < \widehat{\mu}_{f,1} \leq \widehat{\mu}_{f,2} \text{ or } \widehat{\mu}_{f,3} > \widehat{\mu}_{f,1} \geq \widehat{\mu}_{f,2} \\ p_{f,2}/2 & \widehat{\mu}_{f,3} < \widehat{\mu}_{f,2} \leq \widehat{\mu}_{f,1} \text{ or } \widehat{\mu}_{f,3} > \widehat{\mu}_{f,2} \geq \widehat{\mu}_{f,1} \\ 1 & \widehat{\mu}_{f,1} \leq \widehat{\mu}_{f,3} \leq \widehat{\mu}_{f,2} \text{ or } \widehat{\mu}_{f,2} \leq \widehat{\mu}_{f,3} \leq \widehat{\mu}_{f,1} \end{cases}$$

4 ShrinkBayes

ShrinkBayes is based on the inla package, which applies an integrated nested Laplace approximation to fit models in empirical Bayes fashion. ShrinkBayes applies a zero-inflated negative binomial model with normal distributions as priors [8]. In our usage, we make the following reparameterization

$$\begin{aligned}
\phi_f &= \frac{\mu_{f,1} + \mu_{f,2}}{2} && \text{(parental mean)} \\
\alpha_f &= \frac{\mu_{f,2} - \mu_{f,1}}{2} && \text{(half parental difference)} \\
\delta_f &= \mu_{f,3} - \frac{\mu_{f,1} + \mu_{f,2}}{2} && \text{(hybrid effect)}
\end{aligned}$$

We use the ShrinkSeq() and FitAllShrink() functions to fit the model and use inla.make.lincombs(), BFUpdatePosterior(), and SummaryWrap() to calculate posterior probabilities $P(\delta_f + \alpha_f > 0 \mid \text{data})$, $P(\delta_f - \alpha_f > 0 \mid \text{data})$, $P(\delta_f - \alpha_f < 0 \mid \text{data})$, and $P(\delta_f + \alpha_f < 0 \mid \text{data})$, along with estimates of ϕ_f, α_f, and δ_f for $f = 1, \ldots, 25000$. Using this information, we calculate the posterior probability that each feature f is a heterosis feature,

$$p^*_{f,\texttt{ShrinkBayes}} = \begin{cases} 0 & |\widehat{\delta}_f| \le |\widehat{\alpha}_f|. \text{ Otherwise,} \\ P(\delta_f + \alpha_f > 0 \mid \text{data}) & \widehat{\delta}_f > -\widehat{\alpha}_f \ge 0 \\ P(\delta_f - \alpha_f > 0 \mid \text{data}) & \widehat{\delta}_f > \widehat{\alpha}_f \ge 0 \\ P(\delta_f - \alpha_f < 0 \mid \text{data}) & \widehat{\delta}_f < \widehat{\alpha}_f \le 0 \\ P(\delta_f + \alpha_f < 0 \mid \text{data}) & \widehat{\delta}_f < -\widehat{\alpha}_f \le 0 \end{cases}$$

Finally, we let $p_{f,\texttt{ShrinkBayes}} = 1 - p^*_{f,\texttt{ShrinkBayes}}$ be the posterior probability that feature f is not a heterosis feature.

5 baySeq

baySeq uses an empirical Bayes procedure to calculate the posterior probabilities that each feature follows each of the multiple models supplied by the user [2]. In the baySeq framework, a user-supplied model is an an assignment of libraries to treatment groups. In the case of heterosis experiments, it is appropriate to consider the following five models.

$$\begin{aligned}
&M_1 : \mu_{f,1} = \mu_{f,2} = \mu_{f,3} \\
&M_2 : \mu_{f,1} = \mu_{f,2} \\
&M_3 : \mu_{f,1} = \mu_{f,3} \\
&M_4 : \mu_{f,2} = \mu_{f,3} \\
&M_5 : \text{All } \mu_{f,t}\text{'s are distinct.}
\end{aligned}$$

Now, let $p_{f,\texttt{baySeq}}$ be the posterior probability that feature f of a given simulated dataset is not a heterosis feature. We can calculate

$$p_{f,\texttt{baySeq}} = \begin{cases} 1 & \widehat{\mu}_{f,1} \le \widehat{\mu}_{f,3} \le \widehat{\mu}_{f,2} \text{ or } \widehat{\mu}_{f,2} \le \widehat{\mu}_{f,3} \le \widehat{\mu}_{f,1} \\ P(M_1|\text{data}) + P(M_3|\text{data}) + P(M_4|\text{data}) & \text{otherwise} \end{cases}$$

We calculate estimates $\widehat{\mu}_{f,t}$ for $f = 1, \ldots, 25000$ and $t = 1, 2, 3$ using edgeR as described previously.

6 ROC curves

We use receiver operating characteristic (ROC) curves to compare the effectiveness of our method versus `edgeR`, `ShrinkBayes`, and `baySeq`. A ROC curve is a tool for measuring the effectiveness of a binary classifier. It is a graph of the true positive rate (TPR) of detection against the false positive rate (FPR), so a high area under the curve (AUC) is favorable. Landau and Liu [3] describe most of the details of calculation. However, note that in this study, posterior probabilities replace p-values for the Bayesian methods, and we test for heterosis, not differential expression.

7 FDR control procedure 1

We check if our method is controlling the false discovery rate (FDR) of detecting heterosis features. For a given m between 1 and 25000, let

$$\overline{p}_m = \frac{1}{m} \sum_{f=1}^{m} p_{(f)} \qquad \overline{I}_m = \frac{1}{m} \sum_{f=1}^{m} I_{(f)}.$$

We can think of $\overline{p}_m$ as a Bayesian estimate of FDR and $\overline{I}_m$ as the false discovery proportion (FDP). FDR is controlled at $\overline{p}_m$ if $\overline{I}_m \leq \overline{p}_m$. To check FDR control, we plot $\overline{I}_m - \overline{p}_m$ versus $\overline{p}_m$ for $0 \leq \overline{p}_m \leq 0.15$.

8 FDR control procedure 2

This is the same FDR control procedure 1, except that

$$\overline{p}_m = \frac{1}{100} \sum_{f=m}^{m+100} p_{(f)} \qquad \overline{I}_m = \frac{1}{100} \sum_{f=m}^{m+100} I_{(f)}$$

and m is between 1 and 24900.

References

[1] Robert C Gentleman, Vincent J. Carey, Douglas M. Bates, and others. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[2] Hardcastle, T. J. and Kelly, K. A. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(422), 2010.

[3] W. M. Landau and P. Liu. Dispersion Estimation and Its Effect on Test Performance in RNA-seq Data Analysis: A Simulation-Based Comparison of Methods. *Plos One*, 8(12), December 2013.

[4] McCarthy, D. J, Chen, Y., and Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297, January 2012.

[5] Paschold, A., Jia, Y., Marcon, C., and others. Complementation contributes to transcriptome complexity in maixe (Zea mays L.) hybrids relative to their inbred parents. *Genome Research*, 22:2445–2454, October 2012.

[6] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014.

[7] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, October 2009.

[8] van de Weil, M., Neerincx, M., Buffart, T. E., Sie, D., and Verheul, H MW. ShrinkBayes: a versatile R-package for analysis of count-based sequencing data in complex study designs. *BMC Bioinformatics*, 15(116), 2014.